Estimating HIV incidence from multiple sources of data

Francesco Brizzi

Girton College

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2018

Declaration

I hereby declare that this dissertation is the result of my own work, undertaken between October 2013 and October 2017. To my knowledge, all work is original and includes nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. None of the work contained in this dissertation has been submitted or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or at any other University.

Francesco Brizzi
March 2018

Acknowledgements

I would like to express my appreciation and gratitude to my supervisors, Dr. Daniela De Angelis and Dr. Paul Birrell, for the guidance and support over the past four years. I am extremely grateful for the enthusiasm and the invaluable help and time, without which I could have never completed this thesis.

I shall also thank my advisors, Dr Shaun Seaman and Dr Brian Tom, for the insightful comments on the more challenging aspects of my project. Thanks also to Dr Martyn Plummer for his invaluable advice on the use of the `JAGS` software, and for being a lovely host during my research visits in Lyon.

I would also like to thank Public Health London (PHE), and in particular Peter Kirwan, for providing the MSM surveillance dataset used for the application of the methodology developed in this thesis. I acknowledge the financial support of the Medical Research Council and the National Institute for Health Services Research for this project.

Many thanks to my colleagues at the MRC Biostatistics Unit, from who I have learnt a lot, and who made it an enjoyable place to study. A sincere thank to Terry, for his patience and invaluable help on IT matters, and to the PhD students who made me feel at home; without them completing this thesis would have been considerably more difficult.

I would also like to thank all the people that made this four years in Cambridge particularly enjoyable. In particular, thanks to Constantine, Felix, José, Luca, Sam, Stefan, and Andreja. I shall also thank Svea for listening to my complaints. I must also express my gratefulness, for always being there, to my childhood friends: Ale, Alex, Angelo, Aureggi, Fede, Filo, Giorgio, Giovi, Giulia, Isa, Lore, Mencioc, Nino, and Stury. Finally, a special thank to Nico for always listening.

Much love to my parents Elena and Paolo, and thank you for teaching me to be curious and for always encouraging me to learn. Thank you to my brothers Stefano and Andrea, for making me laugh every day. I also need to thank my grandparents (Anna, Cesarina, Enzo and Luciano), uncle and aunt (Antonella and Antonello), cousins (Maria Sole and Francesco),

and extended family members (Ester and Mina) for their continuous support since I was little.

Finally, I need to thank with much love Olympia, not only for thoroughly proofreading and improving this thesis, but most importantly for her love that made this PhD journey much smoother.

Abstract

This thesis develops novel statistical methodology for estimating the incidence and the prevalence of Human Immunodeficiency Virus (HIV) using routinely collected surveillance data. The robust estimation of HIV incidence and prevalence is crucial to correctly evaluate the effectiveness of targeted public health interventions and to accurately predict the HIV-related burden imposed on healthcare services.

Bayesian CD4-based multi-state back-calculation methods are a key tool for monitoring the HIV epidemic, providing estimates of HIV incidence and diagnosis rates by disentangling their competing contribution to the observed surveillance data. Improving the effectiveness of public health interventions, requires targeting specific age-groups at high risk of infection; however, existing methods are limited in that they do not allow for such subgroups to be identified.

Therefore the methodological focus of this thesis lies in developing a rigorous statistical framework for age-dependent back-calculation in order to achieve the joint estimation of age-and-time dependent HIV incidence and diagnosis rates. Key challenges we specifically addressed include ensuring the computational feasibility of proposed methods, an issue that has previously hindered extensions of back-calculation, and achieving the joint modelling of time-and-age specific incidence. The suitability of non-parametric bivariate smoothing methods for modelling the age-and-time specific incidence has been investigated in detail within comprehensive simulation studies.

Furthermore, in order to enhance the generalisability of the proposed model, we developed back-calculation that can admit surveillance data less rich in detail; these handle surveillance data collected from an intermediate point of the epidemic, or only available on a coarse scale, and concern both age-dependent and age-independent back-calculation.

The applicability of the proposed methods is illustrated using routinely collected surveillance data from England and Wales, for the HIV epidemic among men who have sex with men (MSM).

Table of contents

List of figures

List of tables

Abbreviations

AIC Aikake Information Criterion

AIDS Acquired Immunodeficiency Syndrome

CDC Centers for Disease Control and Prevention

CV Cross Validation

GAM Generalized Additive Models

GAMM Generalized Additive Mixed Models

GCV Generalized Cross Validation

GLM Generalised Linear Model

GP Gaussian Processes

GPR Gaussian Process Regression

HAART Highly Active Anti-Retroviral Therapy

HIV Human Immunodeficiency Virus

HMC Hamiltonian Monte Carlo

i.i.d independent identically distributed

INLA Integrated Nested Laplace Approximations

MCMC Markov Chain Monte Carlo

MPES Multi-Parameter Evidence Synthesis

MSE Mean Squared Error

MSM Men who have Sex with Men

NCS Natural Cubic Splines

NUTS No U-Turn Sampler

OLS Ordinary Least Squares

ONS Office for National Statistics

PCP Pneumocystis Carinii pneumonia

PHE Public Health England

P-IRLS Penalised Iteratively Re-weighted Least Squares

PLS Penalised Least Squares

PMSE Predictive Mean Squared Error

PreP Pre-exposure Prophylaxis

RITA Recent Infection Testing Algorithms

SE Squared Exponential

SIV Simian Immunodeficiency Virus

STAR Structured Additive Regression

TPS Thin Plate Splines

UA Unlinked Anonymous

UNAIDS Joint United Nations Programme on HIV/AIDS

Chapter 1

Introduction

1.1 A brief history of the HIV/AIDS epidemic

The AIDS epidemic, caused by HIV virus infection, is one of the greatest and most tragic global epidemics (*i.e.* pandemic) of the 20^{th} and 21^{st} centuries. From its beginning to the end of 2016, around 76 million people have contracted the HIV virus, with approximately 45% of these dying of AIDS-related illnesses (UNAIDS, 2016).

In the summer of 1981, approximately 30 young sexually active homosexual men in California and New York were unexpectedly diagnosed with rare diseases such as Pneumocystis Carinii pneumonia (PCP) and Kaposi's sarcoma (Centers for Disease Control, 1981a,b). These diseases, at the time only observed in elderly men or drug users, were associated with a deficit in CD4 T-helper lymphocytes, a type of white blood cell that plays a key role in the functioning of the immune system. The cases in young homosexuals were formally reported by the Centers for Disease Control and Prevention (CDC) as the emergence of a new disease in 1982. The disease, officially named as Acquired Immunodeficiency Syndrome (AIDS), was defined to be "at least moderately predictive of a defect in cell-mediated immunity, occurring in a person with no known cause for diminished resistance to that disease. Such diseases include Kaposi's sarcoma, PCP, and serious opportunistic infections" (Centers for Disease Control, 1982).

It was only in 1983, that the link between Human Immunodeficiency Virus (HIV) infection and AIDS occurrence was uncovered with HIV separately identified in AIDS infected patients in both France and the US (Barre-Sinoussi et al., 1983; Levy et al., 1984). The first test enabling detection of HIV was licensed in the US and Europe in 1985.

The HIV virus was retrospectively identified in stored blood samples. The earliest known case of HIV dates back to an adult female in 1960 from the Democratic Republic of Congo. Phylogenetic analysis suggests that the HIV virus has been present in central Africa since the early 1900s, but the epidemic only started to spread extensively in the 1960s with growing urbanization (Worobey et al., 2008). Gilbert et al. (2007) suggest that HIV arrived in the US via Haiti in the late 60's, as a considerable number of Haitians was working in Congo at the time.

Despite an enormous amount of effort and funding have been dedicated to HIV-related research, no effective treatment to cure the HIV virus has been found to date. Zidovudine (also known as AZT) was the first anti-retroviral drug released, in 1987, and aimed to attack the HIV virus in order to limit its devastating effect on the immune system. Despite AZT yielding clinical benefit, it was associated with several serious side effects with the HIV virus becoming immune to AZT after prolonged therapy (Richman et al., 1987).

An effective contribution to HIV treatment was only developed in 1995. This is a Highly Active Anti-Retroviral Therapy (HAART), comprising a fixed-dosed combination of multiple anti-retroviral drugs, that has radically changed the dynamics of the HIV epidemic: the number of AIDS-related deaths has plummeted since the mid 90s (Moore and Chaisson, 1999), and now, HIV is seen as a chronic, rather than a fatal condition. Nevertheless, even though the widespread use of HAART has reduced HIV infection, it has not succeeded in eliminating the epidemic.

In 2015, 46% (95% CI, 43%-50%) of the HIV-infected individuals worldwide were under HAART treatment (UNAIDS, 2016). In some regions of the world where HAART coverage is poor the HIV epidemic is still unacceptably severe. For instance, it is estimated that there are approximately 1 million of new HIV infections per year in Africa. This compares to only 73,000 new yearly infections in western and central Europe and North America. Despite relatively small, this number highlights that significant HIV transmission is still occurring in regions where HAART is widely used.

In 2014, a breakthrough Pre-exposure Prophylaxis (PreP) treatment was approved in the US: a new anti-retroviral drug (Truvada) is targeted at people that are not HIV infected, but are at high risk of infection. Clinical trial results are thus far encouraging, as they suggest that PreP is an extremely effective way to reduce the risk of infection (Grant et al., 2010; Baeten et al., 2012; McCormack et al., 2016). In the UK, a new technology assessment of PreP and its implementation are starting soon.

1.2 Biological evolution of HIV/AIDS

The HIV virus is a mutation of Simian Immunodeficiency Virus (SIV), present in apes (chimpanzees, gorillas and sooty mangabeys) but the mechanics of how cross-species transmission occurred remain unclear. The most widely adopted explanation is the "hunter theory", which supports that the exchange of fluids during ape hunting (from a bite, or cut) lead to SIV transmission to humans (Sharp et al., 2001; Keele et al., 2006).

HIV transmission occurs via the exchange of certain fluids such as blood, semen, pre-seminal fluid, rectal and vaginal fluids and breast milk. Hence, HIV transmission typically occurs via sexual intercourse (anal, oral and vaginal), needle and syringe-sharing, blood transfusion, and through infected mothers to their offspring.

HIV is a lentivirus (type of retrovirus) that penetrates the host's body by attaching itself to the surface of the CD4 protein. At this point, the virus starts replicating (*i.e.* the viral load increases) and infecting, among other cells, the CD4 T-helper lymphocytes.

HIV infected individuals typically experience *seroconversion* within 3 weeks of HIV infection, even though this may take up to three months (Horsburgh et al., 1989; Tindall and Cooper, 1991). Seroconversion only lasts for a few weeks, and is characterised by a strong immune system reaction to the virus, including development of antibodies and a sharp decrease in the CD4-count (defined as the number of CD4 T-helper lymphocytes in a fixed volume of blood) with flu-like symptoms (*i.e.* high fever, rashes, swollen lymphs nodes) potentially developing. HIV testing typically involves the detection of antibodies developed during seroconversion, hence the test's results may be unreliable if the test is performed between infection and seroconversion.

After seroconversion, the CD4-count returns to an almost normal level that is, however, followed by a slow, yet sustained, decline of the CD4-count, and a progressive increase in the viral load. In the absence of treatment, this leads in the long term to immunodeficiency (*i.e.* the destruction of the immune system), AIDS-related diseases (*e.g.* Kaposi's sarcoma) and death.

The *incubation period* is defined as the period of time between infection and the development of AIDS. In the absence of treatment, it lasts between 8 and 10 years (Bacchetti and Moss, 1989; Cori et al., 2015). The period of time between infection and seroconversion is known as the *window period*. During this time, infected individuals are highly infective as HIV-specific antibodies have not yet developed, for this reason this is also known as the *acute HIV infection* period.

1.3 Statistical methods for monitoring the HIV/AIDS epidemic

Robust quantification of HIV prevalence and incidence is key to the monitoring of the epidemic and the design of targeted interventions. Prevalence and incidence measure the proportion of the population infected with HIV and the rate of occurrence of new infections respectively. Their estimation is not straightforward: due to the asymptomatic and long natural history of HIV, infections times are typically unobserved and a large proportion of the infected population remains undiagnosed for a long time.

These challenges have motivated the development of a number of original estimation methods (see Brookmeyer and Gail (1994), Foulkes (1998), and Becker and Marschner (2001) for comprehensive reviews), exploiting the progressively increased available information and increasingly synthesising information from different sources, both cross-sectionally and longitudinally.

1.3.1 Estimating prevalence

Studies designed to unbiasedly estimate HIV prevalence in various risk-groups are challenging, if not impossible. Unlinked Anonymous (UA) surveys, based on testing for HIV anonymous blood samples originally collected for other purposes, represent the closest approximation to such ideal surveys. UA surveys have been (and still are) used by the surveillance systems of many developed countries. In England, UA testing has been historically carried out in pregnant women, people who inject drugs, and attendees of sexual health clinics, providing regular estimates of HIV prevalence in these groups. These prevalence estimates have been combined with demographic data to derive population prevalence estimates, using either direct (Karon et al., 1998; McGarrigle et al., 2006), or more sophisticated Multi-Parameter Evidence Synthesis (MPES) methods (Goubar et al., 2008; Conti et al., 2011). The MPES approach, in particular, has been developed to allow the explicit incorporation of information on biases in the data, to detect likely inconsistencies across data sources (Presanis et al., 2008) and has been further extended to jointly estimate HIV prevalence and incidence (Presanis et al., 2011; De Angelis et al., 2014).

1.3.2 Estimating incidence

Cohort studies allow the direct estimation of HIV incidence by longitudinally following up a cohort of uninfected, but at risk of infection, individuals. However, these studies are typically very expensive, and may lack generalisability (Miller et al., 1995; Coovadia et al., 2007; Jansen et al., 2011; and references in Karon et al., 2001) due to their typical small sample size and to the limitations in the design. Inevitably, these studies involve individuals at different risk of infection than the relevant population of interest and with irregular attendance, likely depending on the risk behaviour. This will result in informative missingness and censoring. In the HIV literature, survival analysis techniques have been employed and extended to address these challenges, see De Gruttola and Lagakos (1989), Carstensen (1996), Farrington and Gay (1999), Alioum et al. (2005) and Becker and Marschner (2001) for further details.

Single or serial cross-sectional prevalence surveys have also been used to estimate HIV incidence, by using the relationship between incidence and prevalence (Keiding, 1991). For example, Ades and Medley (1994) and Ades (1995) estimated age-and-time-specific incidence from a series of UA sero-prevalence surveys and modelled the relative inclusion rate, to account for the differential inclusion (in the survey) of infected and uninfected individuals. Incidence estimates can be obtained both within a parametric and a non-parametric framework (Marschner, 1996; Marschner, 1997; Nagelkerke et al., 1999; Williams et al., 2001 and Hallett et al., 2008).

A further example of estimating incidence from a single prevalence study is the "snapshot sampling" proposed by Brookmeyer and Quinn (1995). Here a random sample of individuals is taken and tested for HIV. Prevalence of the p-24 antigen in HIV negative or indeterminate individuals can then be used, together with the information on the average length of the p24 positive period to estimate incidence. This is based on the idea that the p24 antigen is typically detectable only for a certain period prior to seroconversion and hence characterises recent infections. In this case incidence is estimated, through the approximate relationship ”$prevalence = incidence \times mean\ window\ period$”, where prevalence is the p24 prevalence and the mean window period represent the mean duration of the p24 positivity. This approach has been over time extended to other biomarkers of recent infection (Janssen et al., 1998; Karon et al., 2008; Sommen et al., 2011) and considerable work has been devoted to the estimation of their mean window period (Sweeting et al., 2010; Kassanjee et al., 2017; Koulai et al., 2017).

Dynamic compartmental transmission models (Anderson et al., 1992; Garnett, 2002; Grassly and Fraser, 2008) have been predominantly used in the HIV literature to forward simulate the epidemic, under a number of different scenarios, in a deterministic manner (Punyacharoensin et al., 2011). However recent work has focused on estimating the parameters of the infection process. Examples include Alkema et al. (2007) considering a Bayesian melding approach to inference and Presanis et al. (2011), who embedded a dynamic transition model within a MPES framework, in order to better inform parameter estimation using multiple sources of data (Presanis et al., 2011).

Furthermore individual-level simulation models, informed by multiple sources of data, have been extended to estimate HIV incidence using various flavours of rejection sampling (Phillips et al., 2015; Punyacharoensin et al., 2015; Punyacharoensin et al., 2016; Nakagawa et al., 2017). These models require multiple sources of data and are time consuming and very complex. In contrast, back-calculation is a conceptually simple approach which uses all information available through a more parsimonious model.

1.4 Back-calculation

The method of back-calculation was first proposed by Brookmeyer and Gail (1986, 1988) in order to obtain estimates of HIV prevalence and short-term projections of AIDS diagnoses. HIV incidence was originally estimated on the basis of reported AIDS diagnoses and assumed knowledge of the distribution of the incubation time, characterising the time between HIV infection and the AIDS diagnosis. The distribution of the incubation time was then used to relate the estimated incidence to obtain short term predictions of the minimum number of future AIDS diagnoses.

Back-calculation has considerably been developed since then, and still plays a key role in the monitoring of the HIV epidemic globally, with its focus progressively shifting from short-term AIDS diagnoses prediction to the estimation of incidence. The core of back-calculation methods is expressed by the following convolution:

$$d(t) = \int_{t_0}^{t} h(s) f(t-s|s)\, ds \tag{1.4.1}$$

t_0 denotes the starting time of the epidemic, $d(t)$ is the rate of AIDS diagnoses at time t, and $h(s)$ is the rate of new infections at time s. $f(t-s|s)$ is the time-varying incubation distribution and denotes the probability that an individual infected at s experiences an AIDS diagnosis at time t. Knowledge of any two components of the back-calculation convolution, allows estimation of the third one.

Back-calculation is often expressed in discrete times, as diagnosis data are reported on a discrete scale. Let us consider the time period $(t_0, t_T]$ spanning the HIV epidemic. This can be split into T disjoint consecutive intervals $(t_{i-1}, t_i]$, $(i = \{1, \ldots, T\})$. The discrete version of Equation 1.4.1 is then:

$$d_i = \sum_{i_0=1}^{i} h_{i_0} f_{i_0, i-i_0} \qquad (1.4.2)$$

where d_i is the expected number of diagnosis in the interval $(t_{i-1}, t_i]$, h_{i_0} is the expected number of new infections in $(t_{i_0-1}, t_{i_0}]$, and $f_{i_0,i-i_0}$ is the probability that an individual infected in the interval $(t_{i_0-1}, t_{i_0}]$ experiences a diagnosis event in $(t_{i-1}, t_i]$, for $i \geq i_0$ and $i = \{1, \ldots, T\}$. The size of the intervals $(t_{i-1}, t_i]$ is typically determined by the availability of surveillance data (*e.g.* yearly or quarterly) and, for simplicity, the intervals are commonly assumed to be of equal length, despite this is not necessary. Note that a smaller interval length allows for a more refined modelling of the epidemic process, but is more computationally burdensome.

Infections are typically assumed to follow a non-homogeneous Poisson process in the back-calculation literature (Becker et al., 1991; Rosenberg and Gail, 1991). Hence the number of new infections H_{i_0} in the intervals $(t_{i_0-1}, t_{i_0}]$, $i_0 = \{1, \ldots, T\}$, are independent identically distributed (i.i.d) Poisson random variables with means $h_{i_0} = E[H_{i_0}] = \int_{t_{i_0-1}}^{t_{i_0}} \lambda(s)\, ds$. Linear combinations of Poisson random variables are also Poisson-distributed variables. Thus the number of new AIDS diagnoses D_i in each interval $(t_{i-1}, t_i]$, $i = \{1, \ldots, T\}$ is also Poisson distributed, with means d_i given by Equation 1.4.2. Extra-Poisson variation has been previously addressed via the use of over-dispersion parameters (Brookmeyer and Liao, 1990; Rosenberg et al., 1992), or by employing an alternative multinomial framework (Brookmeyer and Gail, 1988; Rosenberg and Gail, 1991; Bellocco and Pagano, 2001).

The validity of back-calculation results principally depends on three components: the diagnosis data available, the assumptions made regarding the distribution of the number of new infections over time, and the chosen incubation distribution (Bacchetti et al., 1993; Mariotti and Cascioli, 1996).

Under-reporting, or under-ascertainment, is a common limitation of surveillance data resulting from changes in reporting conventions, or from reporting delays. Statistical methods have been developed to estimate reporting delays (De Angelis and Gilks, 1994) and under-reporting is typically adjusted for in back-calculation (following Brookmeyer and Damiano, 1989).

Estimating the expected number of new infections over time $\boldsymbol{h} = (h_1, \ldots, h_T)^T$ is challenging for two main reasons. Firstly, there is large uncertainty around estimates in the most recent years: diagnosis data are only weakly informative about recent infections, as only a small proportion of recent infections have had sufficient time to be diagnosed and be reported. Secondly, back-calculation is ill-posed (O'Sullivan, 1986; Lessner, 1998) as T quantities (*i.e.* expected number of new infections) are estimated from T data-points. Despite estimation having a unique solution, the high dimensionality of $\boldsymbol{h}$ leads to high instability in estimation and even small changes in the data may result in substantial changes in results.

In order to address identifiability issues, $\boldsymbol{h}$ is often modelled in terms of infection parameters $\boldsymbol{\theta}$, which are of smaller dimension. Early literature employed parametric models from the exponential family (*e.g.* Day et al., 1989; Aalen et al., 1994), however misspecification of these results in biased estimates and overly narrow confidence intervals. The use of weakly parametric models such as splines (*e.g.* Rosenberg and Gail, 1991) or piecewise-constant step functions (*e.g.* Aalen et al., 1997) has been alternatively proposed. The latter are either assumed constant over long time periods or/and are subject to smoothing constraints (Marschner, 1994).

Modelling the incubation distribution is also challenging. This is separately estimated from cohort studies and is typically assumed as known. However, uncertainty in the estimation of the incubation distribution is typically not accounted for, and there is no guarantee that the cohort used is representative of the population of interest. The incubation distribution has been typically described via parametric distributions. Multi-state models have been alternatively considered to more accurately characterise the different infection stages (*e.g.* Longini et al., 1992; Dietz et al., 1994; Aalen et al., 1997) and inference is highly dependent on the distribution chosen (Rosenberg and Gail, 1990; Bacchetti et al., 1992). The incubation distribution has been further modified in order to account for the effect of treatment, delaying the development of AIDS (Solomon and Wilson, 1990; Brookmeyer, 1991; Rosenberg, 1994). The introduction of HAART complicated the implementation of back-calculation models, which were reformulated to use endpoints other than AIDS diagnoses. The time of HIV diagnosis for AIDS diagnosed individuals, was initially incorporated in back-calculation models (Aalen et al., 1994; Dietz et al., 1994; Farewell et al., 1994; Marschner, 1994; De Angelis et al., 1998). These were subsequently extended to accommodate HIV diagnoses not necessarily followed by AIDS (Aalen et al., 1997; Bellocco and Marschner, 2000; Cui and Becker, 2000; Chau et al., 2003; Sommen et al., 2009). This requires the characterisation of the distribution of the time between infection and HIV diagnosis, but allows the HIV diagnosis rates to be estimated (Marschner, 1994; Dietz et al., 1994; Chau et al., 2003; An et al., 2015).

More recently, biomarker data have been further incorporated within a back-calculation framework. CD4-count data have been employed to better characterise the time between infection and diagnosis and to estimate trends in HIV testing probabilities, via population multi-state models (Sweeting et al., 2005; Birrell et al., 2012). Yan et al. (2011) and Ndawinz et al. (2011) used data on recent infection biomarkers to distinguish between recent and remote in time infections.

From an implementation point of view, inference within a back-calculation is not straightforward. Base-case back-calculation models (Equation 1.4.2) can be expressed as GLMs (Rosenberg and Gail, 1991) so that estimation can be carried out via standard software. However, as previously mentioned, estimation may be highly unstable. Moreover a GLM formulation may not exist for back-calculation models synthesising multiple data sources. Thus, the likelihood of the back-calculation model has been numerically maximized using either the Newton-Raphson (Bacchetti et al., 1993) or the EM (Dempster et al., 1977) algorithms. The latter often incorporates a smoothing step in order to improve the infection parameters identifiability (EMS algorithm, Becker et al., 1991; Marschner, 1994; Yan et al., 2011). Penalised likelihood has also been considered for this purpose (Greenland, 1996; Sommen et al., 2009). Becker and Marschner (1993) showed that the EMS procedure is related to the maximization of a certain penalised likelihood.

Carlin and Gelman (1993) first implemented back-calculation within a Bayesian framework, which was later developed by several authors (Raab et al., 1994; De Angelis et al., 1998; Mezzetti and Robertson, 1999; Sweeting et al., 2005; Birrell et al., 2012). The main advantage of the Bayesian approach is that it allows for the uncertainty in the unknown components of back-calculation models to be accommodated through the inclusion of appropriate prior distributions.

Recent applications of back-calculation models include Birrell et al. (2013), Supervie et al. (2014), Fellows et al. (2015), and van Sighem et al. (2015).

1.5 Age-dependent back-calculation

Back-calculation methods have been extended to derive age-specific estimates of HIV incidence. These estimates can be used to identify sub-groups of the population that are at increased risk of infection, which is necessary if effective targeted public health interventions are to be designed (Becker and Marschner, 1993; Rosenberg, 1995; Verdecchia and Mariotto, 1995; Greenland, 1996; Marschner and Bosch, 1998; Becker et al., 2003). Mezzetti and Robertson (1999) first considered Bayesian age-dependent back-calculation, in application to lung-cancer mortality data. Incorporation of age within back-calculation further allows the incubation distribution to be better characterised; for instance, by accounting for the faster progression for those infected at an older age (Rosenberg and Goedert, 1994). Age also helps to refine the estimation of dates of infections, as it provides a lower bound for the time of infection.

Returning to the back-calculation introduced in the previous Section, let us consider the time period $(t_0, t_T]$ that span the HIV epidemic, and an appropriate age range $(a_0, a_A]$. These can be split into T and A disjoint consecutive time and age intervals, *i.e.* $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$ and $(a_{j-1}, a_j]$, $j = \{1, \dots, A\}$ respectively. If the simplifying assumption that time and age are measured on the same scale is made, the time-and-age specific analogue of the back-calculation in Equation 1.4.2 can be expressed using the following convolution:

$$d_{i,j} = \sum_{i_0=max(1,i-j+1)}^{i} h_{i_0,j-i+i_0} f_{i_0,j-i+i_0,i-i_0} \tag{1.5.1}$$

where $d_{i,j}$ is the expected number of diagnoses in intervals $(t_{i-1}, t_i]$ and $(a_{j-1}, a_j]$ $(i = \{1, \dots, T\}, j = \{1, \dots, A\})$; $h_{i_0,j-i+i_0}$ denotes the expected number of infections in the time interval $(t_{i_0-1}, t_{i_0}]$, and age-interval $(a_{j-i+i_0-1}, a_{j-i+i_0}]$; and $f_{i_0,j-i+i_0,i-i_0}$ denotes the incubation distribution that is dependent on the time-and-age intervals of infection and the number of intervals elapsing between infection and diagnosis.

The bivariate incidence surface (*i.e.* the time-and-age dependent expected number of new infections) has been modelled by many authors using the following multiplicative model: $h_{i,j} = h_i \pi_j$, where the expected number of infections over time h_i, is multiplied by an age-specific risk-factor π_j (Becker and Marschner, 1993; Becker et al., 2003). However this model assumes constant age effects over time, and thus can not capture different time-trends across age-classes. Non-parametric modelling of $h_{i,j}$ has been considered in order to relax this assumption and enhance the model's flexibility: Rosenberg (1995), Marschner and Bosch (1998) and Mezzetti and Robertson (1999) used step-functions.

Back-calculation has been alternatively extended to incorporate birth-cohort data. Greenland (1996) and Wand et al. (2009) employed age-independent back-calculation on diagnosis data stratified by birth-cohort, so that age-dependent incidence estimates were obtained combining the estimates from different cohorts. Verdecchia and Mariotto (1995) instead extended the back-calculation model to include the population susceptible to HIV infection becoming infected according to a rate that is parametrically modelled as a function of calendar time, age and birth-cohort.

All the age-dependent back-calculation models discussed so far solely considered AIDS diagnosis data. Only Becker et al. (2003) examined age-dependent back-calculation including both HIV and AIDS diagnoses.

1.6 Aims of the thesis

This thesis aims to develop a comprehensive extended statistical framework for age-dependent back-calculation, merging existing approaches and addressing limitations of previous work.

As previously discussed, the majority of age-dependent back-calculation models are solely based on AIDS data, which are now increasingly sparse and minimally informative since HAART. The only age-dependent back-calculation approach based on HIV diagnoses used a bivariate incidence curve reliant upon strong multiplicative parametric assumptions (Becker et al., 2003). The main contributions of this thesis are to extend Marschner and Bosch (1998) and Becker et al. (2003) using a more flexible modelling of the time-and-age dependent incidence curve and by including a richer, more informative, array of data.

Rosenberg (1995) used non-parametric bivariate step-functions with broad steps in the time dimension and small steps in the age-dimension. Also Marschner and Bosch (1998) considered a step function, subject to thin plate smoothing, which however rely on an unverified isotropy assumption (*i.e.* equal smoothing in the age-and-time dimensions). We employ tensor product splines (Eilers and Marx, 2003; Wood, 2006b) and Gaussian processes (Rasmussen and Williams, 2006) to continuously model the incidence surface, without defining discrete steps. Furthermore these smoothing models allows us to separately estimate the amount of smoothing required in each dimension. Their properties have been thoroughly investigated and compared within simulation studies.

As discussed by Sweeting et al. (2005) and Birrell et al. (2012) in an age-independent back-calculation framework the incorporation of CD4-count data would allow, via the use of an age-dependent population multi-state model, to estimate age-and-time dependent HIV incidence as well as age-and-time dependent diagnosis probabilities and the prevalence of undiagnosed infection, by age and disease state.

In contrast to earlier age-dependent back-calculation approaches, the introduction of the multiplicity of data-types prevent us from writing back-calculation as a standard Generalised Linear Model (GLM), posing a challenge to the estimation as standard software cannot be used. A maximum penalised likelihood and a Bayesian approach to inference have been developed. The latter approach allows a more coherent propagation of uncertainty and allows for a more straightforward calculation of model-derived quantities (*e.g.* the prevalence of undiagnosed infection). In both frameworks, obtaining an efficient implementation of the model is crucial to ensure that estimation is not hindered by prohibitively lengthy run times.

1.7 Thesis outline

This thesis is structured as follows: Chapter 2 introduces the age-independent CD4-stage back-calculation model (Birrell et al., 2012) that is used as a building block. The base back-calculation model has been extended to consider the epidemic on a reduced time scale, which is crucial when only incomplete surveillance data are available. The model has been further adapted to accommodate surveillance data collected on a coarser time scale to that originally considered. Chapter 3 discusses univariate non-parametric smoothing methods (splines and Gaussian processes) and their implementation within the back-calculation framework introduced in Chapter 2. Chapter 4 compares these non-parametric smoothing methods within a back-calculation framework in a Bayesian simulation study. Chapter 5 extends the back-calculation, discussed in Chapter 2, to age-specific settings. Chapter 6 considers bivariate non-parametric smoothing methods and their implementation to estimate the bivariate (time-and-age specific) incidence surface within the age-specific back-calculation framework described in Chapter 5. Chapter 7 investigates the feasibility of age-specific back-calculation and the appropriateness of the different non-parametric smoothing methods to model bivariate incidence. Chapter 8 illustrates the application of the methods discussed in Chapters 2 and 5 to the HIV epidemic among men who have sex with men in England and Wales. Finally, Chapter 9 summarizes the achievements of this thesis, discusses outstanding issues and suggests ideas and directions for future work.

Chapter 2

Age independent back-calculation

2.1 Introduction

The back-calculation method proposed by Birrell et al. (2012), extending the work of De Angelis et al. (1998) and Sweeting et al. (2005), is routinely used to monitor the status of the HIV-epidemic among Men who have Sex with Men (MSM) in England and Wales (Kirwan et al., 2016).

In this Chapter the motivating surveillance dataset is discussed (Section 2.2) and the back-calculation model is reviewed in Section 2.3. Some extensions of the basic model are then presented in Section 2.4 to handle the more realistic situation when data are not available from the beginning of the epidemic and to deal with coarser data.

2.2 Motivating surveillance dataset

Public Health England (PHE) routinely collects surveillance data to monitor the HIV epidemic in England and Wales. New HIV diagnoses are classified as early (denoted by HIV) or late (denoted by AIDS), depending on whether clinical AIDS symptoms occur within 3 months of diagnosis. CD4-counts taken within 3 months of the first positive test in specialized haematology laboratories are a further source of information available since 1991 (Gupta et al., 2000). These counts are linked by PHE to the registry of HIV/AIDS diagnoses via patient identifiers (Brown et al., 2012).

Let $(t_0, t_T]$ be the time-period spanning the HIV epidemic, split into T disjoint, consecutive intervals $(t_{i-1}, t_i]$, $i = \{1, \ldots, T\}$. The data available include:

- y_i^H, the aggregated number of new HIV diagnoses in $(t_{i-1}, t_i]$.
- y_i^A, the aggregated number of new AIDS diagnoses in $(t_{i-1}, t_i]$.
- A subset of y_i^H of size n_i has an associated CD4-count, taken around diagnosis. These subsets are grouped into K categories, defined by CD4 thresholds: $\boldsymbol{y}_i^{H_C} = (\mathrm{y}_{i,1}^{H_C}, \mathrm{y}_{i,2}^{H_C}, \ldots, \mathrm{y}_{i,K}^{H_C})^T$ is a $K \times 1$ of new HIV diagnoses in $(t_{i-1}, t_i]$, with respective CD4-cell counts being categorized into intervals $[c_1, \infty)$, $[c_2, c_1)$, $\ldots$ and $[0, c_{K-1})$, where $c_1 > c_2 > \cdots > c_{K-1}$.

$\mathbf{y}^H = (y_1^H, \ldots, y_T^H)^T$ and $\mathbf{y}^A = (y_1^A, \ldots, y_T^A)^T$ are $T \times 1$ vectors denoting the number of new HIV and AIDS diagnoses over time respectively and $\mathbf{y}^{H_C} = \{\boldsymbol{y}_1^{H_C}, \ldots, \boldsymbol{y}_T^{H_C}\}$ denotes the collection of CD4-linked diagnoses.

2.3 Model

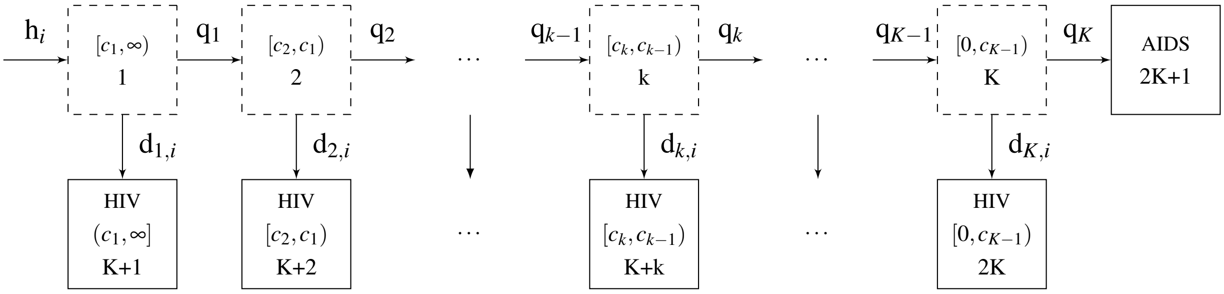


Fig. 2.1 Back-calculation multi-state model, for a general number of undiagnosed states K. Dashed states $\{1, \ldots, K\}$ denote undiagnosed states. Solid states $\{K+1, \ldots, 2K+1\}$ denote states where diagnosis events are observed. q_k denotes the probability of progressing between the latent undiagnosed states k and $k+1$. $d_{k,i}$ represents the probability of being diagnosed from the k^{th} undiagnosed state in the i^{th} time interval.

The basic discrete back-calculation introduced in Chapter 1, is extended to characterise more accurately the HIV epidemic by incorporating all the sources of information described above. Diagnosed individuals are a mixture of individuals with long-standing infections, likely to have been tested as a consequence of HIV-related symptoms, and of recently-infected individuals, likely to have been tested as a consequence of some recent risky behaviour. Hence changes in both HIV transmission and in testing behaviour affect the number of HIV diagnosis. Consequently diagnoses are a result of three distinct interlinked processes: transmission leading to infections, progression leading to HIV-related symptoms and diagnosis. Disentangling the individual contribution of each process to the diagnoses is only possible by reconstructing the complex mechanism underlying the data (Birrell et al., 2013).

For this purpose, a discrete-time back-calculation method, based on a non-homogeneous population-level CD4-count multi-state model (Figure 2.1) is proposed. The multi-state model is characterised by $2K+1$ states: states $\{1,\dots,K\}$ are latent undiagnosed states, describing disease progression through declining CD4-count. The remaining states are absorbing, end-point (or diagnosis) states: AIDS is represented by state $2K+1$, while HIV diagnosis from the respective undiagnosed state is into states $\{K+1,\dots,2K\}$.

Following most of the literature (Section 1.4) the infection process is approximated by a time non-homogeneous Poisson Process with time varying rate $\lambda(u)$, so that the expected size of the infected cohort in $(t_{i_0-1}, t_{i_0}]$ is $h_{i_0} = \int_{t_{i_0-1}}^{t_{i_0}} \lambda(u)du$.

In successive intervals $(t_{i-1}, t_i]$, $i = \{i_0+1,\dots,T\}$ the infected cohort is subject to competing disease progression and diagnosis pressure, represented by movements to undiagnosed states with lower CD4-counts and to the absorbing diagnosis states respectively. By the end of the surveillance period (*i.e.* the end of T^{th} interval) individuals in the infected cohort are either diagnosed with HIV in state $k = \{K+1,\dots,2K\}$, diagnosed with AIDS in state $2K+1$, or remain undiagnosed in one of the latent states $k = \{1,\dots,K\}$.

The progression and diagnosis processes are expressed in terms of probabilities. The progression probabilities are denoted by $\boldsymbol{q} = (q_1,\dots,q_k,\dots,q_K)^T$, where q_k is the probability of progressing from the undiagnosed state k to the state $k+1$ ($k = \{1,\dots,K\}$). $\boldsymbol{q}$ describes the natural history of HIV infections and progression probabilities are assumed to be known from external cohort studies (CASCADE Collaboration, 2000) and to remain constant over calendar time. Diagnosis probabilities are instead denoted by $\boldsymbol{d}_i = (d_{1,i},\dots,d_{k,i},\dots,d_{K,i})^T$, $i = \{1,\dots,T\}$, where $d_{k,i}$ is the probability of being diagnosed (in state $k+K$) from the undiagnosed state k in the i^{th} interval. Diagnosis probabilities are allowed to vary by both calendar time and undiagnosed state as HIV-testing depends on both HIV-related symptoms, that are more likely to occur in lower CD4-counts, and time-varying public health intervention policies.

The inferential problem lies in estimating, in the intervals $(t_{i-1}, t_i]$, $i = \{1,\dots T\}$: the expected number of new infections $\boldsymbol{\mathcal{H}} = \{h_1,\dots,h_T\}$, to which we refer as the *incidence curve* (or *incidence*), and the *diagnosis probabilities* $\boldsymbol{\mathcal{D}} = \{\boldsymbol{d}_1,\dots,\boldsymbol{d}_T\}$. Both $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{D}}$ will be characterised by parameters, *i.e.* $\mathcal{H} \equiv \mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D} \equiv \mathcal{D}(\boldsymbol{\delta})$. In what follows the back-calculation model is described in general terms without focusing on any parameterisation, which will be discussed in Chapter 3. Also for notational convenience, unless otherwise needed, the dependency of $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$ on parameters will be dropped.

Once parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ have been estimated, a number of other population-level quantities of significant epidemiological interest can be derived. The most important consists of the number of undiagnosed infections over time, corresponding to the number of individuals in undiagnosed states 1 to K of the model.

2.3.1 Transition matrices

It is assumed that the time interval $(t_{i-1}, t_i]$ considered is small enough so that individuals can at most experience one event (infection, progression and diagnosis) per interval. New infections in the i$^{\text{th}}$ interval are assumed to occur at the beginning of the i$^{\text{th}}$ interval and are hence not allowed to further progress or to be diagnosed in the interval of infection; both diagnoses and progressions are instead assumed to occur at the end of the i$^{\text{th}}$ interval, the former before the latter.

Transition matrices $\boldsymbol{Q}_i(\boldsymbol{\delta})$ and $\boldsymbol{D}_i(\boldsymbol{\delta})$, also referred to as progression and diagnosis matrices, are functions of progression $\boldsymbol{q}$ and diagnosis probabilities $\mathcal{D}$ respectively, as well as parameters $\boldsymbol{\delta}$. The matrices describe the probabilities of moving between the undiagnosed states $\{1, \dots, K\}$ of the model, and from the undiagnosed to the diagnosis states $\{K+1, \dots, 2K+1\}$ in interval $(t_{i-1}, t_i]$. For notational convenience, $\boldsymbol{Q}_i(\boldsymbol{\delta})$ and $\boldsymbol{D}_i(\boldsymbol{\delta})$ will be denoted $\boldsymbol{Q}_i$ and $\boldsymbol{D}_i$ respectively.

$\boldsymbol{Q}_i$ is a $K \times K$ matrix, whose $(k,l)^{\text{th}}$ entry is defined as:

$$(\boldsymbol{Q}_i)_{k,l} = \begin{cases} (1-d_{k,i})(1-q_k) & \text{if } l = k \\ (1-d_{k,i})q_k & \text{if } l = k+1 \text{ and } k < K \\ 0 & \text{elsewhere} \end{cases} \tag{2.3.1}$$

$\boldsymbol{D}_i$ is a $K \times (K+1)$ matrix, whose $(k,l)^{\text{th}}$ entry is defined as:

$$(\boldsymbol{D}_i)_{k,l} = \begin{cases} d_{k,i} & \text{if } l = k \\ (1-d_{k,i})q_k & \text{if } l = K+1 \text{ and } k = K \\ 0 & \text{elsewhere} \end{cases} \tag{2.3.2}$$

2.3.2 Model dynamics

The model dynamics are described by a system of recursive equations that define the expected behaviour over time, characterised in terms of the incidence curve $\boldsymbol{\mathcal{H}}$ and progression and diagnosis probabilities $\boldsymbol{q}$ and $\boldsymbol{\mathcal{D}}$. Let $\boldsymbol{e}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\boldsymbol{\mu}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ denote the expected number of undiagnosed infections in states 1 to K and the expected number of new diagnoses at the end of the interval $(t_{i-1}, t_i]$ respectively. These will be denoted $\boldsymbol{e}_i = (e_{i,1}, \dots, e_{i,K})^T$ and $\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,K+1})^T$ for notational convenience. Let $\boldsymbol{h}_i = (h_i, 0, \dots, 0)^T$ denote the $K \times 1$ vector of expected new infections in the i^{th} interval. For $i = \{1, \dots, T\}$:

$$\boldsymbol{e}_i = \boldsymbol{Q}_i^T \boldsymbol{e}_{i-1} + \boldsymbol{h}_i^T \tag{2.3.3}$$

$$\boldsymbol{\mu}_i = \boldsymbol{D}_i^T \boldsymbol{e}_{i-1} \tag{2.3.4}$$

where $\boldsymbol{e_0} = (0, \dots, 0)^T$ is a $K \times 1$ vector of zeroes.

2.3.3 Alternative representation of model dynamics

An alternative to the prescription of Equations 2.3.3 and 2.3.4 to represent the model dynamics can be derived by writing a typical back-calculation convolution (as in Section 1.4). The time to endpoints is characterised by the $K \times (K+1)$ matrix $\boldsymbol{P}^{(i_0,i)}(\boldsymbol{\theta}, \boldsymbol{\delta})$, with $(1,k)^{\text{th}}$ entry being the probability for individuals infected in the i_0^{th} interval to be diagnosed in state k in the i^{th} interval. Again for notational convenience, the dependency on parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is suppressed. Also recall (Section 2.3.1) that nor progression nor diagnosis events can occur in the interval of infection, therefore $\boldsymbol{P}^{(i_0,i)}$ has the following structure:

$$\boldsymbol{P}^{(i_0,i)} = \begin{cases} \boldsymbol{D}_{i_0+1} & \text{if } i = i_0+1 \\ \left(\prod_{u=i_0+1}^{i-1} \boldsymbol{Q}_u\right) \boldsymbol{D}_i & \text{if } i > i_0+1 \end{cases} \tag{2.3.5}$$

$\mu_{i,k}$ denotes the k^{th} entry of vector $\boldsymbol{\mu}_i$ (Equation 2.3.4). This can alternatively be defined as:

$$\mu_{i,k} = \sum_{i_0=1}^{i-1} h_{i_0} P_{1,k}^{(i_0,i)} \tag{2.3.6}$$

This formulation of model dynamics is less computationally efficient than the one given in Section 2.3.2 as it requires storing $\frac{T^2}{2}$ $\boldsymbol{P}^{(i_0,i)}$ matrices during computations, against T $\boldsymbol{Q}_i$ and T $\boldsymbol{D}_i$ matrices for the formulation in Section 2.3.2 . This formulation is thus not further pursued.

2.3.4 Likelihood

Given the data introduced in Section 2.2, the likelihood can be formulated based on two fundamental assumptions:

1. Infections arise from a non-homogeneous Poisson process.
2. The distribution of the CD4-counts is representative of the CD4 distribution at diagnosis for all individuals; in other terms, the CD4-counts of unlinked diagnoses are missing at random.

By the first assumption and the properties of the Poisson process (Cox and Isham, 1980), the number of arrivals into each diagnosis state in intervals $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$ forms a set of independent Poisson random variables with means $\boldsymbol{\mu}_i$ (function of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$), obtained from Equation 2.3.4 (or 2.3.6). Hence the distribution of the HIV and AIDS diagnoses (denoted Y_i^H and Y_i^A respectively) is given by independent Poisson variables with means $\mu_t^H = \mu_{i,1} + \dots + \mu_{i,K}$ and $\mu_i^A = \mu_{i,K+1}$ respectively - *i.e.*:

$$Y_i^H \sim Po\left(\mu_i^H\right) \tag{2.3.7}$$

$$Y_i^A \sim Po\left(\mu_i^A\right) \tag{2.3.8}$$

The contribution of the subsample of size n_i of the CD4-linked diagnoses in the i^{th} interval must be accounted for. The number of new HIV diagnoses in the i^{th} interval is the sum of K independent Poisson random variables with means $\mu_{i,1}, \dots, \mu_{i,K}$. By the second assumption, the conditional joint distribution $\boldsymbol{Y}_i^{H_C}$ of the number of individuals diagnosed in states $\{K+1, \dots, 2K\}$ is:

$$\boldsymbol{Y}_i^{H_C} \sim Multinomial(n_i, \boldsymbol{p}_i) \tag{2.3.9}$$

where $\boldsymbol{p}_i = (p_{i,1}, \dots, p_{i,K})$ and $p_{i,k} = \frac{\mu_{i,k}}{\mu_i^H}$, $k = \{1, \dots, K\}$.

Given the observed data $\mathbf{y}^H$, $\mathbf{y}^A$ and $\mathbf{y}^{H_C}$, the overall likelihood $L(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\mu})$ accounts for the contribution of the three sources of information described in Section 2.2 and is:

$$\begin{aligned} L(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\mu}) &= L(\mathbf{y}^{H_C} \mid \mathbf{y}^H, \mathbf{y}^A, \boldsymbol{\mu})\, L(\mathbf{y}^H, \mathbf{y}^A \mid \boldsymbol{\mu}) \\ &\propto \prod_{i=1}^{T} \left(\prod_{k=1}^{K} \left(p_{i,k}\right)^{y_{i,k}^{H_C}} \right) e^{-\mu_i^A} \left(\mu_i^A\right)^{y_i^A} e^{-\mu_i^H} \left(\mu_i^H\right)^{y_i^H} \end{aligned} \tag{2.3.10}$$

where $\mathbf{y}^H$, $\mathbf{y}^A$ and $\mathbf{y}^{H_C}$ are the HIV, AIDS and CD4-linked diagnoses respectively.

2.4 Model customization

The previous Section introduces a base-case back-calculation to be used as building block. This however suffers from a number of limitations (van Sighem et al., 2015). Firstly, this model does not account for under-reporting (see Section 1.4); Section 2.4.1 proposes a modification to address this.

Secondly, the proposed model relies on the complete collection of HIV and AIDS diagnosis data from the beginning of the epidemic, which is typically not the case in countries with less developed surveillance systems (Riedner and Dehne, 1999). Section 2.4.2 considers back-calculation from an intermediate time t_b, assuming that surveillance data are then fully available in $(t_b, t_T]$. Moreover, as briefly shown at the end of this Section, considering the whole epidemic history may be computationally burdensome; the computations can be alleviated by considering only the period $(t_b, t_T]$. As estimates of the recent number of new infections are crucial to inform public health policies and interventions, there is typically little value in fully reconstructing the epidemic history in $(t_0, t_b]$.

Thirdly, in Section 2.3.2, individuals in the multi-state model are conveniently allowed at most one move between the states per interval. This is appropriate when a small time scale is employed. For instance in Birrell et al. (2012) a quarterly interval time scale on quarterly surveillance data is used. However if data are only available at a coarser (*e.g* yearly) time scale, continuing to permit only one movement between the states will not allow patients to be diagnosed quickly enough (*e.g.* using $K = 3$ undiagnosed CD4 states, it would take a minimum time of 3 years to develop AIDS). Section 2.4.3 further modifies back-calculation to address this challenge.

The recursive equations (Section 2.3.2) require $O(T)$ computational time to be evaluated. As $i = \{1, \dots, T\}$, $2T$ matrices (*i.e.* $\boldsymbol{D}_i$, $\boldsymbol{Q}_i$) of size $K \times K + 1$ and $2T$, $K \times 1$ vectors (i.e. $\boldsymbol{e}_i$, $\boldsymbol{h}_i$) and $K + 1 \times 1$ vectors (*i.e.* $\boldsymbol{\mu}_i$) must be stored. Running back-calculation from an intermediate starting point or on a wider time scale, induces a smaller T and thus computational savings. Fewer parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ also need to be estimated for a shorter epidemic period, leading to further computational gains.

Finally, note that the back-calculation methodology discussed suffers from a number of other drawbacks, such as not accounting for mortality or immigration and emigration. These are, however, outside of the scope of this thesis.

2.4.1 Under-reporting

Both HIV and AIDS diagnoses may be subject to under-reporting that if left unaccounted for, will underestimate the true number of infections (De Angelis and Gilks, 1994). The CD4-linked diagnosis data refer to a subsample of the HIV diagnoses. Under-reporting will result in a reduced sample size and accounting for under-reporting is not crucial as long as it is possible to assume that the distribution of CD4 at diagnosis is the same for reported and non-reported diagnoses. For the HIV and AIDS data, is is possible to correct for under-reporting by extending the model to include additional parameters.

Let υ_i^H and υ_i^A denote the proportion of HIV and AIDS diagnoses in the interval $(t_{i-1}, t_i]$ that are reported by the end of the interval $(t_{i-1}, t_i]$; $\boldsymbol{\upsilon}$ denotes the collection of parameters υ_i^H and υ_i^A over time. The expected number of HIV and AIDS diagnoses can then be modified to account for under-reporting as follows:

$$\mu_i^{H'} = \upsilon_i^H \mu_t^H \tag{2.4.1}$$

$$\mu_i^{A'} = \upsilon_i^A \mu_t^A \tag{2.4.2}$$

The likelihood (Equation 2.3.10) can be modified by replacing μ_i^H and μ_i^A by $\mu_i^{H'}$ and $\mu_i^{A'}$ and will further depend on the under-reporting parameters $\boldsymbol{\upsilon}$.

Typically $\mathbf{y}^H$ and $\mathbf{y}^A$ are uninformative with respect to parameters $\boldsymbol{\upsilon}$. However external evidence on under-reporting levels is often available and could be incorporated in the model (Birrell et al., 2012).

2.4.2 Back-calculation over a reduced time period

Formulation of the back-calculation model is considered on a subset $(t_b, t_T]$ of the full epidemic period $(t_0, t_T]$, $t_b > t_0$, requires specification of the state of the model at time t_b. $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$ denotes the expected number of undiagnosed infections, in states $k = \{1, \dots, K\}$, at time t_b.

The choice of t_b is either based on the availability of surveillance data or computational constraints. In both cases, misspecification of $\boldsymbol{\pi}$ might lead to biased incidence and diagnosis probabilities estimates.

The model dynamics in intervals $(t_{b+i-1}, t_{b+i}]$, $i = \{1, \dots, T-b\}$ can be expressed using the recursive Equations 2.3.3 and 2.3.4 and setting $\boldsymbol{e_0} = \boldsymbol{\pi}$ and $i = \{1, \dots, T-b\}$.

2.4.3 Back-calculation on a coarser time scale

Surveillance data may only be available at coarse time scale. In this situation, the challenge is in defining appropriate model dynamics as the simple formulation of the transition and diagnosis matrices (Equations 2.3.3 and 2.3.4), based on at most one transition between the states of the model per time interval, does not allow infected individuals to be diagnosed sufficiently rapidly.

Let $(t_0, t_T]$ be the time-period spanning the HIV epidemic, split into T disjoint sub-intervals $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$. In turn, every interval $(t_{i-1}, t_i]$ is further divided into N_s disjoint sub-intervals $(t_{i,s-1}, t_{i,s}]$, where $s = \{1, \dots, N_s\}$, $t_{i,0} \equiv t_{i-1}$ and $t_{i,N_s} \equiv t_i$. The assumption of at most one move between states holds in the sub-intervals, but not in the larger interval.

Let $h_{i,s}$, $d_{k,i,s}$ and $q_{k,k+1,s}$ denote the expected number of new infections and diagnosis and progression probabilities from state k in $(t_{i,s-1}, t_{i,s}]$ respectively, which must be estimated. The transition and progression matrices, denoted $\boldsymbol{Q}_{i,s}$ and $\boldsymbol{D}_{i,s}$ respectively, can be defined (Equations 2.3.1 and 2.3.2) in $(t_{i,s-1}, t_{i,s}]$ as only one movement between the states is allowed. Further assume that the expected number of new infections and diagnosis probabilities, and thus progression and transition matrices, are constant in the N_s sub-intervals constituting the larger intervals - *i.e.*:

$$
\begin{aligned}
& h_i \equiv h_{i,1} = \cdots = h_{i,N_s}, && i = \{1, \dots T\} \\
& d_{k,i} \equiv d_{k,i,1} = \cdots = d_{k,i,N_s}, && i = \{1, \dots T\},\ k = \{1, \dots, K\} \\
& q_{k,k+1} \equiv q_{k,k+1,1} = \cdots = d_{k,k+1,N_s}, && k = \{1, \dots, K-1\} \\
& \boldsymbol{Q}_i \equiv \boldsymbol{Q}_{i,1} = \cdots = \boldsymbol{Q}_{i,N_s}, && i = \{1, \dots T\} \\
& \boldsymbol{D}_i \equiv \boldsymbol{D}_{i,1} = \cdots = \boldsymbol{D}_{i,N_s}, && i = \{1, \dots T\}
\end{aligned}
$$

The dynamic equations (Equations 2.3.3 and 2.3.4) can be then replaced with a pair of equations providing the expected number of undiagnosed $\boldsymbol{e_i'}$ and newly diagnosed individuals $\boldsymbol{\mu_i'}$ at the end of intervals $(t_{i-1}, t_i]$, $i = \{1, \dots T\}$:

$$\boldsymbol{e_i'} = \boldsymbol{Q}_i'^T \boldsymbol{e}_{i-1}' + \boldsymbol{h}_i' \tag{2.4.3}$$

$$\boldsymbol{\mu_i'} = \boldsymbol{D}_i'^T \boldsymbol{e}_{i-1}' + \boldsymbol{d}_i' \tag{2.4.4}$$

where $\boldsymbol{e_0}$ is a $K \times 1$ vector of zeros. The $K \times K$ transition $\boldsymbol{Q}_i'$ and the $K \times K + 1$ progression matrices $\boldsymbol{D}_i'$ for $(t_{i-1}, t_i]$ are constructed by aggregating the transition and progression matrices defined in the subintervals:

$$\boldsymbol{Q}_i' = \boldsymbol{Q}_i^{N_s} \qquad\qquad \boldsymbol{D}_i' = \sum_{j=0}^{N_s-1} \boldsymbol{Q}_i^j \, \boldsymbol{D}_i \tag{2.4.5}$$

where $\boldsymbol{Q}_i^0$ is a $K \times K$ identity matrix. Equation 2.4.5 could be interpreted as follows: individuals remaining undiagnosed in $(t_{i-1}, t_i]$, must remain undiagnosed in each of the N_s sub-intervals comprising the i^{th} interval. This equates to raising the sub-interval transition matrix $\boldsymbol{Q}_i$ to the power of N_s. If individuals are diagnosed in $(t_{i-1}, t_i]$, the diagnosis must have occurred in one of the comprising N_s sub-intervals. The $\boldsymbol{D}_i'$ formulation allows for the diagnosis to occur in the first sub-interval, with probability $\boldsymbol{D}_i$, in the second sub-interval, with probability $\boldsymbol{Q}_i\boldsymbol{D}_i$, and so on.

Let the $K \times 1$ vector of new expected infections in $(t_{i-1}, t_i]$ be $\boldsymbol{h}_i = (h_i, 0, \ldots, 0)^T$. h_i individuals are expected to be infected at the beginning of each subinterval $(t_{i,s-1}, t_{i,s}]$, $i = \{1, \ldots T\}$ and $s = \{1, \ldots, N_s\}$. New infections either remain undiagnosed $\boldsymbol{h_i'}$ or are diagnosed $\boldsymbol{d_i'}$ by the end of the interval $(t_{i-1}, t_i]$, where:

$$\boldsymbol{h}_i' = \left(\sum_{s=1}^{N_s-1} \boldsymbol{Q}_i^s \right)^T \boldsymbol{h}_i \qquad\qquad \boldsymbol{d}_i' = \left(\sum_{s=1}^{N_s-1} (N_s - s) \boldsymbol{Q}_i^{s-1} \, \boldsymbol{D}_i \right)^T \boldsymbol{h}_i \tag{2.4.6}$$

The number of infected individuals being diagnosed and remaining undiagnosed by the end of the infection interval $(t_{i-1}, t_i]$ can be reconstructed by considering the sub-intervals $(t_{i,s-1}, t_{i,s}]$, $s = \{1, \ldots N_s\}$.

Consider the infected individuals in the beginning of $(t_{i-1,1}, t_{i,1}]$; these can either remain undiagnosed throughout or be diagnosed in one of the successive sub-intervals $(t_{i,s-1}, t_{i,s}]$, $s = \{2, \ldots N_s\}$. The probability to remain undiagnosed is $\boldsymbol{Q}_i^{N_s-1}$. Diagnosis might occur in the second sub-interval $(s = 2)$ with probability $\boldsymbol{D}_i$, in the third sub-interval ($s = 3$, probability $\boldsymbol{Q}_i\boldsymbol{D}_i$) and so on, up to the $s = N_s$ sub-interval (probability $\boldsymbol{Q}_i^{N_s-2}\boldsymbol{D}_i$). Similarly individuals infected in the second sub-interval can remain undiagnosed throughout $(t_{i,s-1}, t_{i,s}]$, $s = \{3, \ldots, N_s\}$ with probability $\boldsymbol{Q}_i^{N_s-2}$. Alternatively, diagnoses can occur in the $s = 3$ subinterval (probability $\boldsymbol{D}_i$) and so on, up to the $s = N_s$ subinterval (probability $\boldsymbol{Q}_i^{N_s-3}\boldsymbol{D}_i$). Summing undiagnosed and diagnosed individuals over all sub-intervals constituting $(t_{i-1}, t_i]$ yields Equation 2.4.6.

Note that Equations 2.4.3 and 2.4.4 can be modified to model the epidemic from an intermediate point t_b, as discussed in Section 2.4.2.

2.5 Summary

In this Chapter the back-calculation model originally proposed for the surveillance of the MSM HIV epidemic in England and Wales has been introduced (Birrell et al., 2012). Together with the motivating dataset (Section 2.2) initial extensions have been proposed in Sections 2.4.2 and 2.4.3. These increase the generalisability of the originally proposed model, and render the model applicable to a wider range of HIV epidemics, where surveillance data are only available less frequently or are not available from the start of the epidemic, or both. A summary of the notation used is available in Appendix B.1.

Plausible parameterisations for the incidence curve $\mathcal{H}(\boldsymbol{\theta})$ and diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$ have not been discussed. The next Chapter proposes a number of non-parametric smoothed models for $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$.

Chapter 3

Univariate non-parametric smoothing methods

3.1 Introduction

Chapter 2 introduced back-calculation in general terms, without specifying parameterisations for the incidence curve $\mathcal{H}(\boldsymbol{\theta})$ and diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$. The estimation problem is ill-posed in the sense that small variations in the data, lead to large variations in the estimates (Lessner, 1998). For HIV back-calculation, due to the relatively long time between infection and diagnosis, estimates of recent infection levels are highly sensitive to the parameterisations of $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$. To limit the reliance on strong parametric assumptions, whilst ensuring appropriate flexibility, non-parametric methods have been considered in the literature. Smoothing is introduced to improve identifiability, and this has been achieved by integrating a kernel smoothing step into the EM algorithm or through the use of penalised likelihood methods (Section 1.4).

Splines and Gaussian processes are instruments for smoothly modelling curves. Early work on splines dates back to Reinsch (1967), Duchon (1977), De Boor (1978), Wahba (1980, 1983, 1990), and Silverman (1985). Excellent books on splines are Green and Silverman (1994), Ruppert et al. (2003), Wood (2006c), and Fahrmeir et al. (2007). Gaussian processes have been employed for decades in geo-statistics (Krige, 1966; Matheron, 1973), under the name of *kriging*, but they became extremely popular in the machine learning community in the late 90's (Williams and Rasmussen, 1996; MacKay, 1998). Both splines and Gaussian processes are still active and widely applied research fields (Marra and Radice, 2010; Hensman et al., 2013; Roberts et al., 2013; Bauer et al., 2016).

Most of the literature focuses on scatter-plot smoothing purposes, where splines and Gaussian processes are used to smooth out a set of observed data points. However, within the back-calculation framework of Chapter 2, these methods need to be extended to model latent processes, as both the incidence curve and the diagnosis probabilities are not directly observed.

This Chapter begins by demonstrating the applicability of splines (Section 3.3) and Gaussian processes (Section 3.4) to scatter-plot smoothing problems (Section 3.2). These are first compared in Section 3.5, and are subsequently (Section 3.6) embedded in a back-calculation framework to model the latent $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$. Note that understanding properties of splines and Gaussian processes is fundamental to appropriately interpret the back-calculation results.

The methodological overview on splines and Gaussian processes presented in this Chapter is mostly based on Wood (2006a), Bowman and Evers (2013), and Rasmussen and Williams (2006).

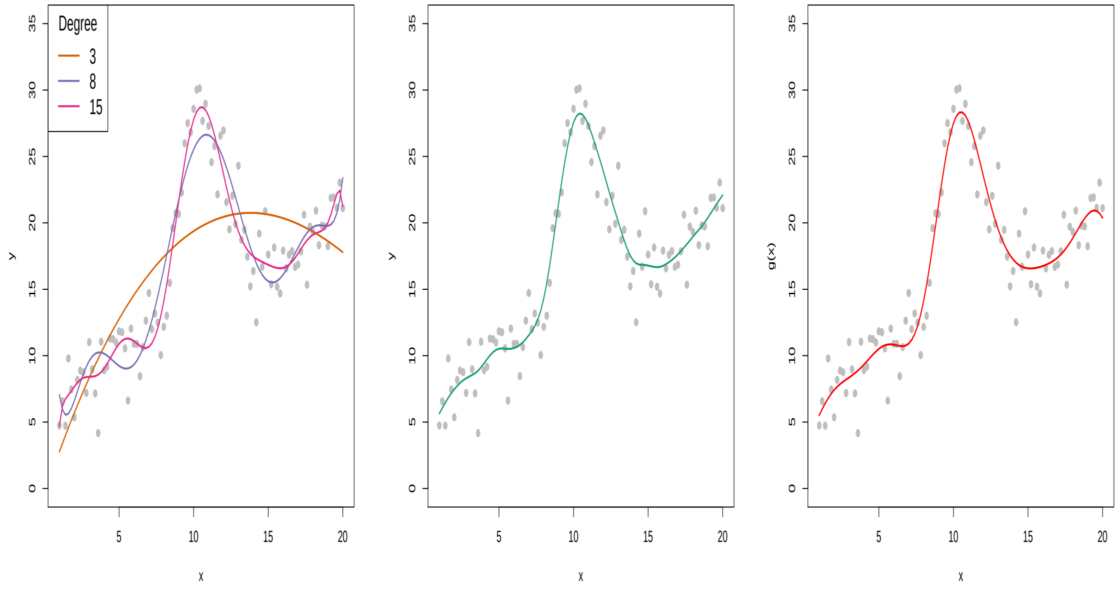
3.2 The problem

Consider data (y_i, x_i), $i = \{1, \ldots, n\}$, where $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ is an $n \times 1$ vector of observations made, conditional upon the $n \times 1$ vector of covariates $\boldsymbol{x} = [x_1, \ldots, x_n]^T$. Linear regression is one of the simplest statistical models, but is often not flexible enough to capture the trend exhibited in the data.

As shown in Figure 3.1a polynomial regression offers greater flexibility, but suffers from a number of drawbacks. A high degree polynomial needs to be considered in order to capture the trend in the data. However this often displays strong oscillations, which are not supported by the data and high curvature at both ends of the data-range. Non-parametric methods, such as splines and Gaussian processes, have the added value of capturing complex trends in the data in an accurate, yet smooth, manner; this is demonstrated in Figures 3.1b and 3.1c.

Non-parametric regression aims to find a smooth curve $g(x)$, often called smooth, so that, for $i = \{1, \ldots, n\}$:

$$y_i = g(x_i) + \varepsilon_i \tag{3.2.1}$$



(a) Polynomial Regression (b) Spline (c) Gaussian Process

Fig. 3.1 Comparison of fits obtained by a polynomial regression, a cubic spline and a Gaussian process respectively, based on an illustrative dataset.

where $g(x)$ is a smooth curve that does not have a pre-restricted shape (*e.g.* a straight line), but instead adjusts its shape to smoothly capture the features of the data. This is also known as the scatter-plot smoothing problem.

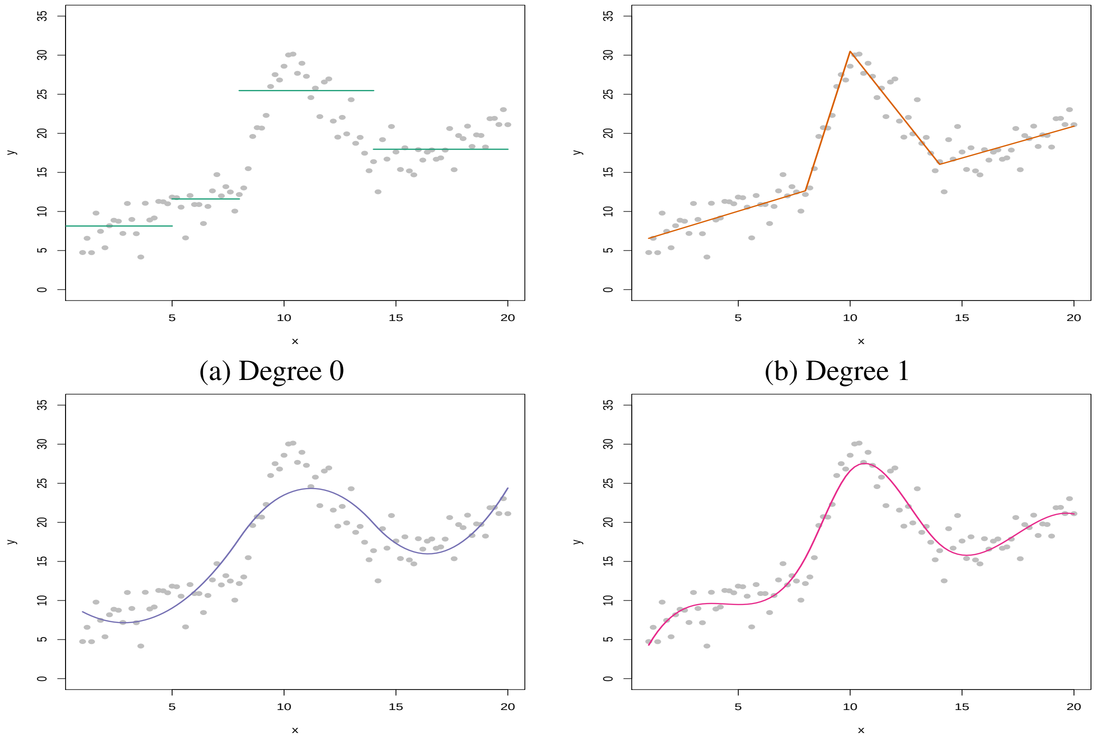
3.3 Splines

3.3.1 Introduction

Definition 3.3.1. A polynomial spline of degree d is a function $g(x) : [a,b] \to \mathbb{R}$, defined on a set of ordered knots $\boldsymbol{\kappa} = \{a = \kappa_1 < \kappa_2 < \cdots < \kappa_k = b\}$, so that:

- $g(x)$ is a polynomial of degree d in intervals (κ_i, κ_{i+1}), $i = \{1,\ldots,k\}$.
- $g(x)$ is $d-1$ times continuously differentiable.

The degree d and the number and location of knots k characterise the polynomial spline, as shown in Figures 3.2 and 3.3.



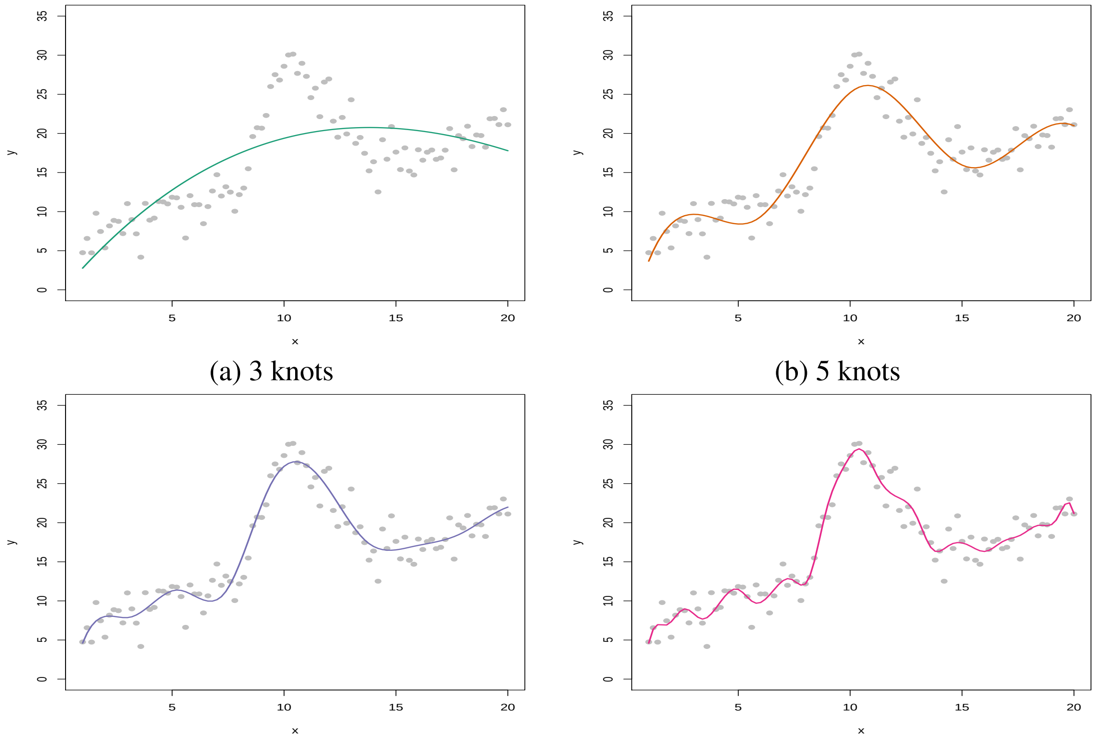
(a) Degree 0

(b) Degree 1

(c) Degree 2

(d) Degree 3

Fig. 3.2 Polynomial splines of different degrees



(a) 3 knots

(b) 5 knots

(c) 10 knots

(d) 20 knots

Fig. 3.3 Cubic polynomial splines with different number of knots.

A spline of degree zero is a discontinuous step function (Figure 3.2a), a spline of degree one is continuous with discontinuous derivatives (Figure 3.2b), a spline of degree two is continuous with continuous first derivative (Figure 3.2c), and finally a third degree (cubic) spline is continuous with continuous first and second derivatives (Figure 3.2d). Splines that do not have enough knots can not capture the features of the data, while splines with a large number of knots are likely to be wiggly, and sensitive to perturbations in the data due to noise (*i.e.* can lead to overfitting). Figure 3.3 illustrates this behaviour for a cubic spline.

In summary, the flexibility of a polynomial spline increases with both the degree and the number of knots. Typically the degree of a spline is fixed (*e.g.* cubic), while the number of knots is allowed to vary. Flexibility is a double edged sword: it is necessary to capture complex trends in the data, but if not carefully employed it may lead to overfitting.

A polynomial spline (Definition 2.1.1) is constructed as a linear combination of appropriately defined basis functions $b(x)$:

$$g(x) = \sum_{j=1}^{B} \beta_j b_j(x)$$

This can be expressed in a regression format, $\boldsymbol{g} = \boldsymbol{X\beta}$, where:

$$\boldsymbol{g}_{[n\times 1]} = \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{bmatrix} \quad \boldsymbol{X}_{[n\times B]} = \begin{bmatrix} b_1(x_1) & \dots & b_B(x_1) \\ \vdots & \vdots & \vdots \\ b_1(x_n) & \dots & b_B(x_n) \end{bmatrix} \quad \boldsymbol{\beta}_{[B\times 1]} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_B \end{bmatrix}$$

the $B \times 1$ vector $[b_1(x), \dots, b_B(x)]^T$ contains the bases of the spline, which depend on the spline chosen (Sections 3.3.4 to 3.3.7).

Penalised regression is typically considered for estimating the parameters of the spline and will be discussed in the following Section.

3.3.2 Penalised regression

A strategy to control the flexibility of a spline consists in specifying a large number of knots, while constraining their influence. A penalty term is introduced so that smooth curves are favoured over wiggly ones, counteracting the overfitting induced by the large number of knots. Penalised Least Squares (PLS) is then employed, consisting of the Ordinary Least Squares (OLS) criterion and an additional penalty term:

$$min\, ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta} \qquad (3.3.1)$$

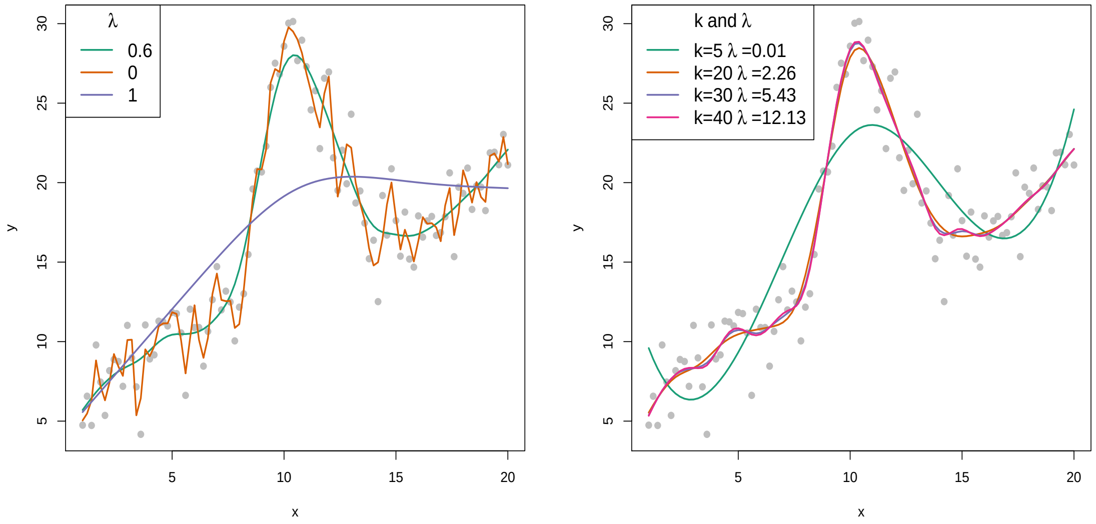
here $\boldsymbol{S}$ is a penalty matrix quantifying the quadratic roughness of parameters $\boldsymbol{\beta}$, which varies depending on the type of spline considered; λ is the smoothing parameter and controls the trade off between goodness of fit, measured by the OLS term, and roughness, as measured by the penalty term.

It can be shown that the vector $\hat{\boldsymbol{\beta}}$ minimizing PLS is given by (Wood, 2006a):

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{S}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{3.3.2}$$

Estimated coefficients $\hat{\boldsymbol{\beta}}$ depend on the value of λ, which tunes smoothness by introducing bias and reducing variance. If λ is equal to 0 the PLS solution is equivalent to the unbiased, but potentially volatile, OLS estimate. Conversely for large λ values, $\hat{\boldsymbol{\beta}}$ leads to an overly smooth curve with little variance, but potentially subject to large bias. Figure 3.4a demonstrates the effect of λ.

The actual number of knots used to construct the spline is not crucial within a penalised regression framework: a larger number of knots can always be specified leading to enhanced



(a) (b)

Fig. 3.4 Implications of the choice of λ and the number of knots. a) Three smoothing parameter choices resulting in a overly smooth, a overly rough, and a suitable curve. b) Splines with increasing number of knots. Once a minimum number of knots is reached to ensure enough flexibility, similar results are obtained for all knots considered as the smoothing parameter guards against overfitting.

overfitting. This can be, however, offset by using a greater value of λ. Figure 3.4b further clarifies this point.

While it is fundamental to choose a large enough number of knots to ensure enough flexibility, it is however wasteful, especially in terms of computation time, to choose a number of knots that is too large. Sensitivity analysis can help to address whether enough knots have been chosen. The number of knots can be increased and if the results are comparable to those obtained with fewer knots, it can be informally concluded that the number of knots chosen is sufficient.

Recall that any spline is defined by a design matrix $\boldsymbol{X}$ and related parameter vector $\boldsymbol{\beta}$. The latter can be estimated via a penalised least square criterion (Equation 3.3.1), after introducing a penalty matrix $\boldsymbol{S}$. Hence defining explicitly $\boldsymbol{\beta}$, $\boldsymbol{X}$, and $\boldsymbol{S}$ allow us to embed splines within the convenient penalised regression framework discussed in this Section. In what follows (Sections 3.3.3 to 3.3.7) we discuss how this can be achieved for different types of splines. It is key to understand how these splines are constructed and their properties, as these will be subsequently considered within a back-calculation framework.

3.3.3 Optimal natural cubic splines

Recall that the $\boldsymbol{S}$ matrix somehow measures the roughness of a curve $g(x) : [a,b] \to \mathbb{R}$. The integrated second derivative squared of $g(x)$ is an appealing measure for quantifying roughness, as high values of the second derivative squared characterise rougher curves. Reinsch (1967) first proposed splines, arising as a solution to the minimization of the criterion, in Equation 3.3.3, in the space of two times continuoulsy differentiable functions:

$$min \sum_{i=1}^{n} (\, y_i - g(x_i) \,)^2 + \lambda \int_a^b \{g''(x)\}^2 dx \tag{3.3.3}$$

This objective function is a special case of penalised regression (Section 3.3.2) and it compromises between goodness of fit, as measured by the residual sum of squares, and the roughness of the curve, as measured by the integrated second derivative of $g(x)$. Again, λ is the smoothing parameter.

Theorem. Equation 3.3.3 is uniquely minimized, in the space of continuously differentiable function in $[a,b]$, by a natural cubic spline with a knot for every unique x_i.

Natural Cubic Splines (NCS) are a particular type of polynomial splines, formally defined as follows:

Definition 3.3.2. A Natural Cubic Spline is a function $g(x) : [a,b] \to \mathbb{R}$, defined on a set of ordered knots $\boldsymbol{\kappa} = \{a = \kappa_1 < \kappa_2 < \cdots < \kappa_k = b\}$, so that:

- $g(x)$ is a cubic polynomial in intervals (κ_i, κ_{i+1}), $i = \{1, \dots, k\}$.
- $g(x)$ is two times continuously differentiable - *i.e.* the first and second derivatives of $g(x)$ are continuous.
- $g''(a) = g''(b) = g'''(a) = g'''(b) = 0$, *i.e.* the second and third derivatives evaluated at the boundary knots are zero.

The first two conditions characterise any polynomial spline of degree three. The last condition ensures that extrapolation of the spline outside of the boundary knots is linear.

Green and Silverman (1994) show that a NCS can be written as follows, even though other parameterisations exist (Wood, 2006a):

Definition 3.3.3. Consider a function $g(x) : [a,b] \to \mathbb{R}$, defined by a set of ordered knots $\boldsymbol{\kappa} = \{a = \kappa_1 < \kappa_2 < \cdots < \kappa_k = b\}$ and parameters $\alpha_1, \alpha_2, \delta_1, \dots \delta_k$:

$$g(x) = \alpha_0 + \alpha_1 x + \frac{1}{12} \sum_{j=1}^{k} \delta_j |x - \kappa_j|^3$$

This is a NCS if the following constraints are satisfied:

$$\sum_{j=1}^{k} \delta_j = \sum_{j=1}^{k} \delta_j \kappa_j = 0$$

Note that the basis for a NCS is: $\left\{1, x, \frac{1}{12}|x-\kappa_1|^3, \dots, \frac{1}{12}|x-\kappa_k|^3\right\}$.

Thus NCS with a knot per observation (referred to as *"optimal NCS"* from now on) are optimal: if roughness is measured using the integrated second derivative squared, there will be no smoother curve than a NCS producing an equally good fit to the data. Optimal NCS are "natural" in the sense that a NCS basis is "naturally" obtained as a solution to a specific smoothing problem, rather than being arbitrarily defined.

The optimal NCS can be formulated within the penalised regression framework discussed in Section 3.3.2, by expressing the smoothing problem in Equation 3.3.3 as the PLS criterion in Equation 3.3.1. This is crucial as it allows us to estimate the spline's parameters using Equation 3.3.2.

This can be achieved by defining parameter vectors $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]^T$ and $\boldsymbol{\delta} = [\delta_1, \dots, \delta_n]^T$, and matrices $\boldsymbol{T}$ and $\boldsymbol{E}$ respectively of dimension $n \times 2$ and $n \times n$:

$$\boldsymbol{T} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \frac{1}{12}|x_1 - x_1|^3 & \cdots & \frac{1}{12}|x_1 - x_n|^3 \\ \vdots & \cdots & \vdots \\ \frac{1}{12}|x_n - x_1|^3 & \cdots & \frac{1}{12}|x_n - x_n|^3 \end{bmatrix} \tag{3.3.4}$$

so that the optimal NCS, with one knot per observation, can be written as follows:

$$\boldsymbol{g} = \boldsymbol{T\alpha} + \boldsymbol{E\delta} \quad \text{s.t} \quad \boldsymbol{T^T\delta} = \boldsymbol{0} \tag{3.3.5}$$

Green and Silverman, 1994 (page 141), demonstrate that the roughness integral (in Equation 3.3.3) can be expressed in quadratic form:

$$\int_a^b \{g''(x)\}^2 dx = \boldsymbol{\delta^T E \delta}$$

Hence, the smoothing criterion in Equation 3.3.3 can be rewritten as the following PLS criterion:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda \boldsymbol{\delta^T E \delta} \quad \text{s.t} \quad \boldsymbol{T^T\delta} = \boldsymbol{0} \tag{3.3.6}$$

After some algebra, the above penalised least squared criterion, which is subject to two constraints, can be reparametrised (using the QR decomposition of $\boldsymbol{T}$, see Horn and Johnson, 2012 and Appendix C.1.1) into the simple PLS criterion $||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$ discussed in Section 3.3.2. The mathematical details are available in Appendix C.1.2.

Optimal NCS require a knot, and therefore a parameter, per observation. As discussed in Section 3.3.2, a large number of knots is often unnecessary, as penalised regression allows to obtain similar estimates with fewer knots, by adjusting the smoothing parameter. *"low-rank"* bases splines, that is with fewer parameters than observations, can also be employed. Despite not having the optimality property of optimal NCS, these work well in practice and are more computationally efficient (*i.e.* faster running time, less prone to numerical error). The following Sections 3.3.4 to 3.3.7 will describe low-rank splines.

3.3.4 Knots-based natural cubic splines

Recall that optimal NCS are a special case of NCS, where a knot is placed at each observation. Low-rank NCS can be defined based on a set of knots that is smaller than the set of observations, expressed within the usual penalised regression framework.

Let $\boldsymbol{\kappa} = \{a = \kappa_1 < \kappa_2 < \cdots < \kappa_k = b\}$ be a set of knots of size $k < n$. Let $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_0, \alpha_1\end{bmatrix}^T$ and $\boldsymbol{\delta} = \begin{bmatrix}\delta_1, \dots, \delta_k\end{bmatrix}^T$ be parameter vectors and let $\boldsymbol{T}$, $\boldsymbol{E}$ and $\boldsymbol{C}$ be matrices of dimension $n \times 2$, $n \times k$ and $2 \times k$ respectively:

$$\boldsymbol{T} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \frac{1}{12}|x_1 - \kappa_1|^3 & \cdots & \frac{1}{12}|x_1 - \kappa_k|^3 \\ \vdots & \cdots & \vdots \\ \frac{1}{12}|x_n - \kappa_1|^3 & \cdots & \frac{1}{12}|x_n - \kappa_k|^3 \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \kappa_1 & \kappa_2 & \cdots & \kappa_k \end{bmatrix}$$

A NCS characterised by a set of knots $\boldsymbol{\kappa}$ (Definition 3.3.2) can be written as:

$$\boldsymbol{g} = \boldsymbol{E\delta} + \boldsymbol{T\alpha} \quad \text{s.t} \quad \boldsymbol{C\delta} = \boldsymbol{0} \tag{3.3.7}$$

The quadratic roughness integral can be approximated as follows:

$$\int_a^b \{g''(x)\}^2 dx \approx \boldsymbol{\delta^T E \delta}$$

Hence the following constrained smoothing objective is considered:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda \boldsymbol{\delta^T E \delta} \quad \text{s.t} \quad \boldsymbol{C\delta} = \boldsymbol{0} \tag{3.3.8}$$

Analogously to Section 3.3.3, after employing suitable reparameterisations, the above constrained PLS criterion can be turn into the unconstrained PLS criterion $||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$ characterising penalised regression (Section 3.3.2) so that parameter estimates can easily be obtained (using Equation 3.3.2). Mathematical details are available in Appendix C.2.

Knots based NCS work well in practice if a suitable number of knots is chosen; usually k knots are placed at the k quantiles of $\boldsymbol{y}$. Undertaking sensitivity analyses to verify that the fitted curve does not depend on the knots location is crucial.

3.3.5 Thin plate regression splines

Wood (2003) introduces thin plate regression splines which are another, more "optimal", low-rank approximation to optimal (or full-rank) NCS. The main idea is to introduce a dimension (or rank) reduction matrix, relating a low-rank NCS to a full-rank NCS. Using the same data, these yield different fitted values and penalise roughness differently. Such differences can be quantified and minimized by an "optimal" dimension reduction matrix.

Recall that an optimal NCS has n parameters $\boldsymbol{\delta}$, subject to two constraints (Section 3.3.3). A dimension reduction matrix $\boldsymbol{\Gamma_k}$ of size $n \times k$ is introduced to map the parameters $\boldsymbol{\delta}$ of the full-rank spline to parameters $\boldsymbol{\delta_k}$ ($k < n$) of the low-rank spline:

$$\boldsymbol{\delta} = \boldsymbol{\Gamma_k \delta_k} \tag{3.3.9}$$

$\boldsymbol{\Gamma_k}$ is a matrix of rank k, the columns of which form a k-dimensional orthonormal basis. Thus $\boldsymbol{\Gamma_k^T \Gamma_k} = \boldsymbol{I_k}$, but $\boldsymbol{\Gamma_k \Gamma_k^T} \neq \boldsymbol{I}_n$, where $\boldsymbol{I}_l$ denotes an identity matrix of size l. The PLS fitting criterion for thin plate regression splines is:

$$min||\boldsymbol{y} - \boldsymbol{E\Gamma_k \delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \delta_k^T \boldsymbol{\Gamma_k^T E \Gamma_k} \delta_k \quad \text{s.t} \quad \boldsymbol{T^T \Gamma_k \delta_k} = 0 \tag{3.3.10}$$

As $\boldsymbol{\Gamma_k^T \Gamma_k} = \boldsymbol{I_k}$ and $\boldsymbol{\delta} = \boldsymbol{\Gamma_k \delta_k}$, the above can be re-written in terms of $\boldsymbol{\delta}$, rather than $\boldsymbol{\delta_k}$:

$$min||\boldsymbol{y} - \widetilde{\boldsymbol{E}}_{\boldsymbol{k}}\boldsymbol{\delta} - \boldsymbol{T\alpha}||^2 + \lambda \delta^T \widehat{\boldsymbol{E}}_{\boldsymbol{k}} \delta \quad \text{s.t.} \quad \boldsymbol{T^T \delta} = 0 \tag{3.3.11}$$

where $\widetilde{\boldsymbol{E}}_k = \boldsymbol{E\Gamma_k \Gamma_k^T}$ and $\widehat{\boldsymbol{E}}_k = \boldsymbol{\Gamma_k \Gamma_k^T E \Gamma_k \Gamma_k^T}$.

As both the low and full rank splines are expressed in terms of parameters $\boldsymbol{\delta}$, the fitting criterion of optimal NCS (Equation 3.3.6) and thin plate regression spline (Equation 3.3.11) can be compared. The fitting criteria differ in both the fitted values $(\boldsymbol{E} - \widetilde{\boldsymbol{E}}_{\boldsymbol{k}})\boldsymbol{\delta}$ and the roughness measure $\boldsymbol{\delta^T}(\boldsymbol{E} - \widehat{\boldsymbol{E}}_{\boldsymbol{k}})\boldsymbol{\delta}$ terms. It is desirable to find $\boldsymbol{\Gamma_k}$ that jointly minimizes changes in goodness of fit and roughness, however Wood (2003) claims that this is not possible. Nonetheless the more loose criteria of "worst possible changes" in fit and roughness can be simultaneously minimized over all values of $\boldsymbol{\delta}$.

Fit: Thin plate regression splines approximate $\boldsymbol{E}$ by $\widetilde{\boldsymbol{E}}_{\boldsymbol{k}}$. Since the fitted values of the reduced-rank spline are $\widetilde{\boldsymbol{E}}\boldsymbol{\delta} + \boldsymbol{T\alpha}$ (and $\boldsymbol{\alpha}$ is the same as for optimal NCS) the worst possible change in fitted values is measured by:

$$\widetilde{e}_k = max_{\boldsymbol{\delta} \neq 0} \frac{||(\boldsymbol{E} - \widetilde{\boldsymbol{E}}_{\boldsymbol{k}})\boldsymbol{\delta}||}{||\boldsymbol{\delta}||}$$

Roughness: The optimal NCS penalty matrix $\boldsymbol{\delta^T E \delta}$ is approximated by $\boldsymbol{\delta^T \widehat{E}_k \delta}$, hence the worst possible change in fitted values is measured by:

$$\widehat{e_k} = max_{\boldsymbol{\delta} \neq 0} \frac{\boldsymbol{\delta}^T(\boldsymbol{E} - \boldsymbol{\widehat{E}_k})\boldsymbol{\delta}}{||\boldsymbol{\delta}||^2}$$

Wood (2003) shows that $\widetilde{e_k}$ and $\widehat{e_k}$ are jointly minimized by the "optimal" dimension reduction basis $\boldsymbol{\Gamma_k} = \boldsymbol{U_k}$, which is obtained from the eigen-decomposition of matrix $\boldsymbol{E}$. Explicitly $\boldsymbol{E}$ (defined in Equation 3.3.4, of size $n \times n$) can be expressed as the matrix product $\boldsymbol{UDU^T}$. $\boldsymbol{D}$ is a diagonal matrix, of dimension $n \times n$, with the absolute values of the eigenvalues of $\boldsymbol{E}$ sorted in ascending order along the main diagonal. $\boldsymbol{U}$ is the $n \times n$ matrix of eigenvectors. Let $\boldsymbol{U_k}$ be the $n \times k$ matrix consisting of the first k columns of $\boldsymbol{U}$ and $\boldsymbol{D_k}$ be the top left $k \times k$ submatrix of $\boldsymbol{D}$.

After some algebra (Appendix C.3), the fitting criterion of Equation 3.3.10, with $\boldsymbol{\Gamma_k} = \boldsymbol{U_k}$, can be expressed as the usual penalised fitting criterion $||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$ in Equation 3.3.1.

Thin plate regression splines avoid choosing the knots location, as they are fully specified by k parameters arising from the truncated eigen-decomposition. Analogously to the number of knots, k needs to be large enough to ensure sufficient flexibity, but not too large to avoid computational waste.

3.3.6 Thin plate regression splines with linear shrinkage

All splines considered so far (optimal NCS, knots-based NCS and thin plate regression splines) have a penalty matrix $\boldsymbol{S}$ and are composed of: a linear term $\boldsymbol{T\alpha}$ (*i.e.* $\alpha_0 + \alpha_1 x$) and a non-parametric term (*i.e.* $\boldsymbol{E\delta}$, often subject to reparameterisation).

As λ increases, the penalty matrix $\boldsymbol{S}$ shrinks the parameters $\boldsymbol{\delta}$ towards zero while the parameters $\boldsymbol{\alpha}$ are not subject to any penalty term (see Equations C.1.7, C.2.2 and C.3.4). Thus greater λ values shrink the spline towards a straight line (*i.e.* the unpenalised term) but not towards zero. Marra and Wood (2011), within a context of variable selection, propose a strategy to penalise the null-space (*i.e.* the space of unpenalised coefficients $\boldsymbol{\alpha}$) allowing shrinkage to zero.

The eigendecomposition $\boldsymbol{S} = \boldsymbol{UDU^T}$ is considered, where $\boldsymbol{U}$ and $\boldsymbol{D}$ are as defined in Section 3.3.5. As there are two unpenalised coefficients (*i.e.* α_1, α_2), the last two eigenvalues are equal to zero. These are replaced by a small portion ε of the minimum strictly positive eigenvalue of $\boldsymbol{D}$, creating matrix $\boldsymbol{D'}$. The original penalty matrix $\boldsymbol{S}$ can be then replaced by $\boldsymbol{S'} = \boldsymbol{UD'U^T}$. This gives $\boldsymbol{S'} \approx \boldsymbol{S}$, so that the null-space is penalised.

Marra and Wood (2011) show that $\varepsilon = 1/10$ works well in practice, thus the penalty imposed on the null space is smaller than the one imposed on the originally penalised space. Hence as λ increases the spline is first penalised towards a straight line, and then, if needed, the straight line is further shrunk to zero.

Splines with shrinkage can be expressed via the usual penalised regression criterion $||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S' \beta}$, where $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{S}$ (used to construct $\boldsymbol{S'}$) depend on the spline considered. Marra and Wood (2011) suggest using thin plate regression splines (Section 3.3.5) with linear shrinkage.

3.3.7 P-splines

Eilers and Marx (1996), building on the work by O'Sullivan (1986), suggest a pragmatic alternative to construct a low-rank spline that is not based on approximations of optimal NCS (as in Sections 3.3.4 to 3.3.6). They consider P-splines, which are B-splines (De Boor, 1978) estimated within a penalised regression framework (Section 3.3.2).

B-splines are low-rank polynomial splines characterised by a local basis, defined by degree d and $\boldsymbol{\kappa} = \{a = \kappa_1 < \cdots < \kappa_k = b\}$ equally spaced internal knots dividing the domain $[a,b]$ into $k-1$ disjoint intervals. The basis function $B_i^d(x)$ is recursively defined:

$$B_i^d(x) = \frac{x - \kappa_{i-d}}{\kappa_i - \kappa_{i-d}} B_{i-1}^{d-1}(x) + \frac{\kappa_{i+1} - x}{\kappa_{i+1} - \kappa_{i+1-d}} B_i^{d-1}(x) \tag{3.3.12}$$

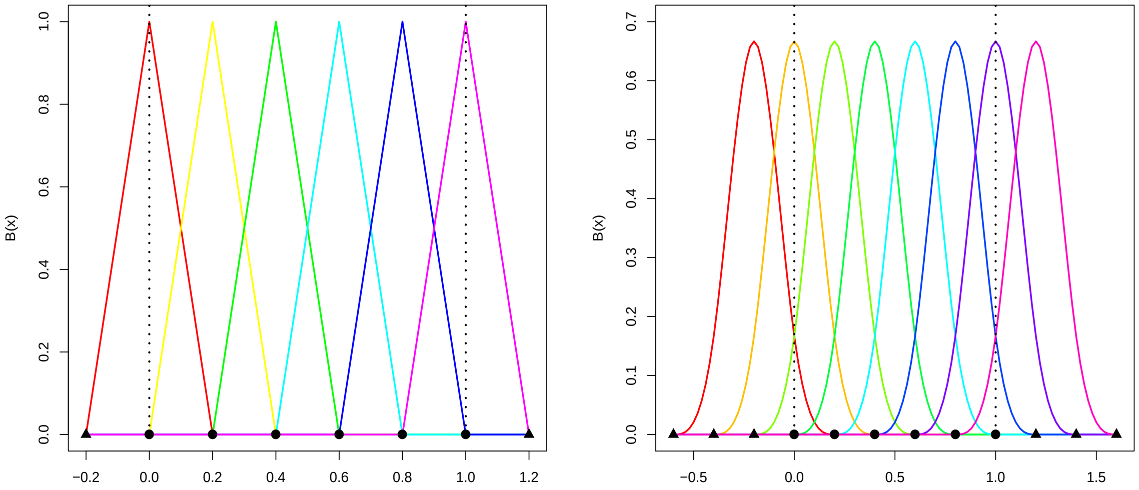
$$B_i^0(x) = \begin{cases} 1 & \text{for } \kappa_i \leq x \leq \kappa_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

A B-spline $g(x) : [a,b] \to \mathbb{R}$ is a sum of B-spline bases:

$$g(x) = \sum_{j=1}^{k-1+d} B_j^d(x)\beta_j \tag{3.3.13}$$

The B-spline basis is: $\left\{B_1^d(x), \ldots, B_{k-1+d}^d(x)\right\}$ and satisfies the following properties:

- $2d$ extra knots outside of $[a,b]$ are necessary for the recursion in Equation 3.3.12 to be valid, yielding a total number of knots of $k+2d$.



(a) 1^{st} degree

(b) 3^{rd} degree

Fig. 3.5 Basis for B-splines of degree 1 and 3, with internal knots (black dots) at $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $2d$ external knots (black triangles). The original domain is $[0,1]$, and is delimited by the vertical dotted lines.

- Comprises $d+1$ polynomials of degree d.
- The polynomials join at the k inner knots.
- The derivative up to order $d-1$ are continuous.
- $k-1+d$ polynomial bases are defined by the recursion.
- For every $x \in (a,b)$ only $d+1$ bases are non-zero.
- For all x $\in [a,b]$: $\sum_{i=1}^{k-1+d} B_i^d(x) = 1$.

Figure 3.5 shows bases of a B-spline of first and third degrees. These are the sum of a series of piecewise linear or cubic polynomials, joined at the knots. Linear (or cubic) B-splines require two (or six respectively) extra knots (denoted by black triangles) to be defined outside the domain $[a,b]$. The basis function is local and ordered, meaning that successive bases (and coefficients) only affect neighbouring parts of the curve.

Eilers and Marx (1996) define P-splines by estimating the coefficients of a B-spline subject to a penalty. If successive coefficients take similar values, the fitted spline is smooth. Consequently a difference matrix $\boldsymbol{D_r}$ of degree r is used as penalty matrix, where $r = 1$ and

$r = 2$ are most commonly used:

$$\boldsymbol{D_1} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \qquad \boldsymbol{D_2} = \begin{bmatrix} -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \end{bmatrix}$$

The penalty matrix $\boldsymbol{S}$ is obtained by multiplying: $\boldsymbol{D_r^T D_r}$. For first and second order difference matrices the penalty term is equivalent to: $\boldsymbol{\beta^T D_1^T D_1 \beta} = \sum_{i=1}^{k-2+d} (\beta_{i+1} - \beta_i)^2$ and $\boldsymbol{\beta^T D_2^T D_2 \beta} = \sum_{i=1}^{k-3+d} (\beta_{i+2} - 2\beta_{i+1} + \beta_i)^2$.

A first order penalty shrinks the coefficients (and thus the spline) towards a common constant, as only two equal successive coefficients are not penalised. In contrast, a second order penalty shrinks coefficients towards a linear trend, as only three neighbouring coefficients forming a linear trend are unpenalised.

As before, P-splines of degree d, with a penalty matrix of order r, can be expressed within the usual penalised regression framework $min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$. Matrices $\boldsymbol{X}$ and $\boldsymbol{S}$ are explicitly defined in Appendix C.4.

3.3.8 Splines beyond scatter-plot smoothing

Note that, so far, no distributional assumptions on the outcome vector $\boldsymbol{y}$ have been made as coefficients $\hat{\boldsymbol{\beta}}$ were obtained by minimizing the PLS criterion (Equation 3.3.2). Akin to linear regression, equivalent estimates for $\hat{\boldsymbol{\beta}}$ can be found by maximising the following penalised log-likelihood criterion, assuming a normally distributed error term $\boldsymbol{\varepsilon}$, *i.e.* $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{I}\sigma^2)$, with $\boldsymbol{X}$ and $\boldsymbol{\beta}$ indicating the design matrix and parameters of a spline, and $\boldsymbol{I}$ an $n \times n$ diagonal matrix:

$$\max_{\boldsymbol{\beta}}\ l(\boldsymbol{y}|\boldsymbol{\beta}) + \lambda \boldsymbol{\beta^T S \beta}$$

where $l(\boldsymbol{y}|\boldsymbol{\beta})$ denotes the log-likelihood of the data.

When employed within a GLM or Generalized Additive Models (GAM) framework, splines can be used to smoothly model more complex outcomes, with the existence of criterion to determine the optimal amount of smoothing required $\hat{\lambda}$ (Appendix C.5.2) and to obtain confidence intervals (Appendix C.5.3). However the back-calculation model introduced in Chapter 2 does not fall into such a framework and we have to look at alternative approaches to inference.

3.3.9 Bayesian inference

The main gain of using a Bayesian approach lies in its flexibility in that it allows for splines to be integrated within more complex models (Wood, 2016), which cannot be expressed in the GLM or GAM framework.

Within a Bayesian framework, the posterior parameter distribution, given the data $f(\boldsymbol{\beta}|\boldsymbol{y})$, is proportional to the likelihood $f(\boldsymbol{y}|\boldsymbol{\beta})$ and the prior distribution $f(\boldsymbol{\beta})$ of the parameters (Wahba, 1983; Silverman, 1985; Wood, 2006c). On the log-scale:

$$l(\boldsymbol{\beta}|\boldsymbol{y}) \propto l(\boldsymbol{y}|\boldsymbol{\beta}) + l(\boldsymbol{\beta})$$

where $l(\boldsymbol{\beta}|\boldsymbol{y}), l(\boldsymbol{y}|\boldsymbol{\beta}), l(\boldsymbol{\beta})$ are the log-posterior, log-likelihood and log-prior respectively.

The penalised likelihood criterion (Equation 3.3.8) can be re-interpreted from a Bayesian perspective, with the penalty term $\lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{\beta}$ corresponding to a multivariate Normal prior for $\boldsymbol{\beta}$, with mean $\mathbf{0}$ and precision matrix $\lambda\boldsymbol{S}$, and λ being implicitly assigned a flat prior.

The multivariate Normal prior is improper as $\boldsymbol{S}$ is rank-deficient for all splines discussed, with the exception of thin plate regression splines with linear shrinkage. The parameters $\boldsymbol{\beta}$ that are not subject to a penalty term (Equations C.1.7, C.2.2, C.3.4, C.4.1) are given flat priors.

Formally, let p be the rank of $\boldsymbol{S}$, and $\boldsymbol{S}_p$ (of size $p \times p$), be the full-rank sub-matrix of $\boldsymbol{S}$. Let $\boldsymbol{\beta} = [\boldsymbol{\beta_p}\ \boldsymbol{\beta_u}]^T$, where $\boldsymbol{\beta_p}$ and $\boldsymbol{\beta_u}$ are the vectors of penalised and unpenalised coefficients, of size p and u respectively. The penalty term can be then re-formulated, within a Bayesian framework, using the priors:

$$\begin{aligned} \boldsymbol{\beta_p} &\sim N_p\left(\mathbf{0}, (\lambda\boldsymbol{S}_p)^{-1}\right) \\ \boldsymbol{\beta_u} &\sim U(-\infty, \infty) \end{aligned} \tag{3.3.14}$$

Despite the Bayesian formulation of splines being established over 30 years ago (Silverman, 1985; Wahba, 1983), the earliest implementations of Bayesian splines only date back to Fahrmeir and Lang (2001) and Lang and Brezger (2004) due to computational challenges, resolved by the development of MCMC.

Splines can be implemented using general purpose Bayesian inference software (Appendix A.4) such as `BUGS` (Lunn et al., 2000), `JAGS` (Plummer, 2003), or `Stan` (Stan Development Team, 2016b) to avoid developing bespoke MCMC algorithms.

In practice, two reparameterisations can be employed to increase the computational efficiency of MCMC sampling. The first reparameterisation, known as centering, allows any spline to be re-written so that the first parameter is an unpenalised global intercept, and the sum of the spline values over the covariates is zero (*i.e.* $\sum_{i=1}^{n} g(x_i) = 0$). The second reparameterisation allows for the penalty matrix $\boldsymbol{S}$ to be re-written as a diagonal matrix with p ones and u zeroes along the main diagonal. For the mathematical details on the two reparameterisations, see Appendix C.6.

Following the sequential application of the two reparameterisations, the spline is characterised by parameters $\boldsymbol{\beta'}$ and design matrix $\boldsymbol{X'}$. The centering reparameterisation, results in an intercept term (β'_1) and the first column of $\boldsymbol{X'}$ being a vector of ones. Given the diagonal structure of the precision matrix $\boldsymbol{S'}$ obtained from the second reparameterisation, i.i.d priors are imposed on penalised coefficients $\beta'_{P_i} \sim N(0, 1/\lambda)$, $i = \{1, \dots, p\}$ and i.i.d flat priors are imposed on unpenalised coefficients $\boldsymbol{\beta}_U$.

`JAGS`, unlike `Stan`, does not allow the specification of improper priors; typically a vague proper prior, usually i.i.d Normal, is instead specified for $\beta'_{U_i} \sim N(0, 1/\lambda_0)$, $i = \{1, \dots, u-1\}$. The intercept term β'_1 is generally given a distinct weakly informative prior. λ_0 is a second smoothing parameter, imposed on the originally unpenalised coefficients. λ_0 can either be fixed or can be estimated, as suggested in Wood (2016). Priors must be then assigned to the smoothing parameters λ and λ_0, with dispersed gamma or log-uniform priors (Wood, 2016), or the Half-Cauchy ot Half-t priors (Gelman et al., 2006) being common choices.

Let $\boldsymbol{y}$ follow any distribution, not necessarily from the exponential family, with $l(\boldsymbol{y}|\boldsymbol{\beta'})$ the respective likelihood. A fully Bayesian spline can be then specified as follows:

$$
\begin{aligned}
&\boldsymbol{y} \sim l(\boldsymbol{y}|\boldsymbol{\beta'}) \\
&\beta'_1 \sim f(\beta') \\
&\beta'_{P_i} \sim N(0, 1/\lambda), \quad i = \{1, \dots, p\} \\
&\beta'_{U_i} \sim N(0, 1/\lambda_0), \ \ i = \{1, \dots, u-1\} \\
&\lambda \sim f(\lambda) \\
&\lambda_0 \sim f(\lambda_0)
\end{aligned}
\tag{3.3.15}
$$

Where $f(\cdot)$ denotes a prior distribution.

Note that λ_0 is only introduced for splines with more than one unpenalised coefficients (*i.e.* $u > 1$) before reparameterisations. As previously mentioned, the centering parameterisation transforms one of the originally unpenalised parameters to a global intercept. The remaining unpenalised coefficients are given a prior characterised by λ_0. Only thin plate regression splines with shrinkage and first order B-splines do not require λ_0, as they have zero and one originally unpenalised coefficients respectively. All the other splines discussed (optimal and knots based NCS, thin plate regression splines and second order B-splines) do require λ_0 as they have two unpenalised coefficients (*i.e.* $u = 2$).

The *jagam* function in the `R` package **mgcv** automatically generates `JAGS` code to estimate a spline in a Bayesian framework, using the parameterisations described in this Section. The flexibility of the Bayesian framework allows to embed splines within more complex models.

3.4 Gaussian processes

3.4.1 Introduction

Splines can be used to construct a smooth non-parametric curve $g(x) : [a,b] \rightarrow \mathbb{R}$ through a flexible basis: in a Bayesian framework, the spline parameters $\boldsymbol{\beta}$ are assigned a prior distribution and a posterior distribution is obtained, after conditioning on the data.

A more general approach for obtaining a smooth non-parametric curve $g(x)$ involves placing a prior over the set of all possible functions in $[a,b]$; this reflects prior beliefs regarding properties of the functions considered, for example smoothness. The posterior distribution over the set of functions $g(x)$ can be then obtained. A prior can be placed on an uncountably infinite set of possible functions using stochastic processes, as these specify, by definition, a probability distribution over a set of functions.

3.4.2 Defining Gaussian processes

A Gaussian Processes (GP) (Rasmussen and Williams, 2006) is a stochastic process, consisting of a (potentially infinite) collection of random variables, any finite number of which have a joint Normal distribution.

A GP is specified via the mean and covariance of the joint multivariate Normal distribution. These are constructed by specifying a mean $m(x_i)$ and a covariance function (often called kernel) $k(x_i, x_j|\boldsymbol{\phi})$, characterised by hyper-parameters $\boldsymbol{\phi}$. The former is unrestricted, whereas the latter must ensure the positive-definiteness of the covariance matrix. Formally:

Definition 3.4.1. A Gaussian Process $\boldsymbol{g} \sim GP(m(x_i), k(x_i, x_j|\boldsymbol{\phi}))$, is a collection of random variables $\boldsymbol{g} = \{g(x_1), \dots, g(x_n)\}$ which have a joint multivariate Normal distribution, for any finite collection of covariates $\boldsymbol{x} = \{x_1, \dots, x_n\}$:

$$[g(x_1), g(x_2), \dots, g(x_n)]^T \sim N_n(\boldsymbol{m}, \boldsymbol{K})$$

where $\boldsymbol{\mu}$ and $\boldsymbol{K}$ are an $n \times 1$ mean vector and a $n \times n$ covariance matrix respectively, with elements $\boldsymbol{m}_i = m(x_i)$ and $\boldsymbol{K}_{ij} = k(x_i, x_j|\boldsymbol{\phi})$.

$k(x_i, x_j|\boldsymbol{\phi})$ relates the random variables $g(x_i)$ and $g(x_j)$ depending on the distance between the covariates x_i and x_j. As numerous distance measures have been considered in the literature, several choices of covariance functions exist (Rasmussen and Williams, 2006, Chapter 4; Duvenaud, 2014). The default choice is the squared exponential function (or kernel) due to its convenient properties. Firstly, it describes smooth functions with infinitely many derivatives. Secondly it is stationary, as $k(x_i, x_j|\boldsymbol{\phi})$ only depends on the squared distance between x_i and x_j, and is hence invariant to translations (*i.e.* the function values are unchanged if all covariates are shifted by a constant).

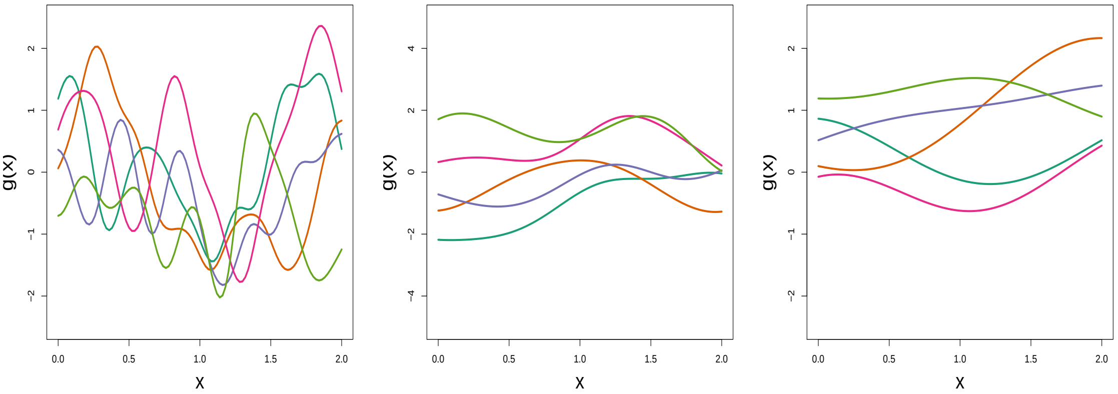
For simplicity, here we start by considering the case where the mean is an $n \times 1$ zero vector, and the covariance function is the squared exponential kernel with hyper-parameters $\boldsymbol{\phi} = \{\eta, \rho\}$:

$$k(x_i, x_j|\eta, \rho) = \eta^2 exp\left(-\frac{(x_i - x_j)^2}{2\rho^2}\right) \tag{3.4.1}$$

ρ is the length-scale, controlling the frequency and hence smoothness of the GP; lower frequencies correspond to rougher functions, while larger frequencies lead to linear functions. η is a scale factor determining the magnitude of the oscillation of the GP away from $m(x_i)$. Figures 3.6 and 3.7 clarify the role of these hyper-parameters.

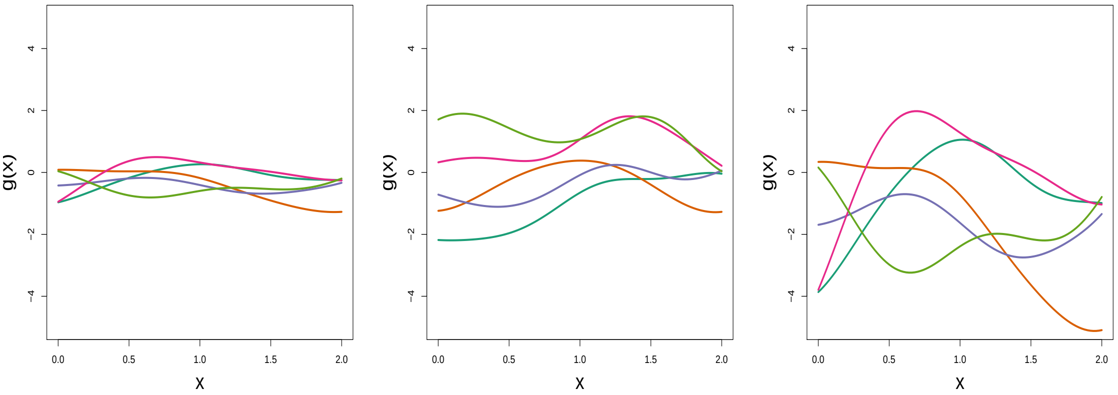
3.4.3 Gaussian process regression

Gaussian Process Regression (GPR) is similar to Bayesian regression, but yields a posterior over a space of functions $\boldsymbol{g}$, instead of parameters $\boldsymbol{\beta}$. Reconsider the smoothing problem posed in Equation 3.2.1 and a GP prior is imposed on $\boldsymbol{g} \sim N_n(\boldsymbol{0}, \boldsymbol{K})$. Observations $\boldsymbol{y}$ are assumed to have i.i.d normally distributed errors, with mean 0 and variance σ^2, so that $\boldsymbol{y}|\boldsymbol{g} \sim N_n(\boldsymbol{g}, \sigma^2\boldsymbol{I})$. The posterior distribution over the space of functions $\boldsymbol{g}$ is obtained by



(a) $\rho = 0.1$ (b) $\rho = 0.15$ (c) $\rho = 0.5$

Fig. 3.6 Realisations of a GP with $\eta = 1$ and different ρ values. The smoothness of the functions increases with ρ.



(a) $\eta = 0.5$ (b) $\eta = 1$ (c) $\eta = 2$

Fig. 3.7 Realisations of a GP with $\rho = 0.5$ and different η values. The range of the functions increases with η.

exploiting the conditioning properties of the multivariate Normal distribution as follows. Let $\boldsymbol{z} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z_1} \\ \boldsymbol{z_2} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Let $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ be of dimension n_1 and n_2 (*i.e* $n = n_1 + n_2$). The conditional density of $\boldsymbol{z_2}$, given $\boldsymbol{z_1}$, is also multivariate Normal:

$$\boldsymbol{z_2}|\boldsymbol{z_1} \sim N_{n_2}\left(\boldsymbol{\mu_2} + \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z_1} - \boldsymbol{\mu_1}),\, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right) \tag{3.4.2}$$

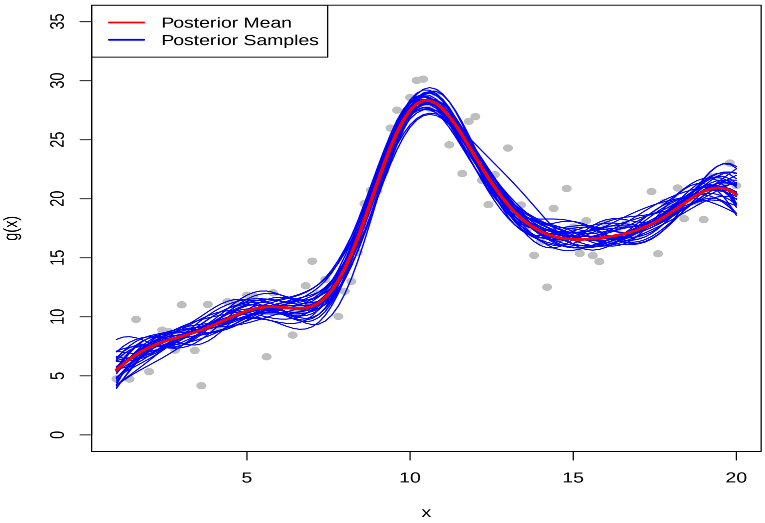


Fig. 3.8 Posterior mean (red line) and 100 samples (blue lines) from the posterior $\boldsymbol{g}|\boldsymbol{y}$, after conditioning on an illustrative dataset.

By the properties of GP, the joint Normal distribution of $(\boldsymbol{g}, \boldsymbol{y})$ is:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{g} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} + \sigma^2 \boldsymbol{I} & \boldsymbol{K} \\ \boldsymbol{K} & \boldsymbol{K} \end{bmatrix} \right) \tag{3.4.3}$$

The posterior distribution $\boldsymbol{g}|\boldsymbol{y}$ is obtained from Equation 3.4.2:

$$\boldsymbol{g}|\boldsymbol{y} \sim N\left(\boldsymbol{K}(\boldsymbol{K}+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{y},\ \boldsymbol{K}-\boldsymbol{K}(\boldsymbol{K}+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}\right) \tag{3.4.4}$$

Despite the prior mean being $\boldsymbol{0}$, the posterior mean is a non-zero function of the covariance matrix $\boldsymbol{K}$; thus the prior mean is typically set to zero in the literature. The posterior covariance is the prior covariance minus a term quantifying the reduction in uncertainty after conditioning on the data. Figure 3.8 plots the posterior mean along posterior distribution $\boldsymbol{g}|\boldsymbol{y}$ samples for an illustrative dataset.

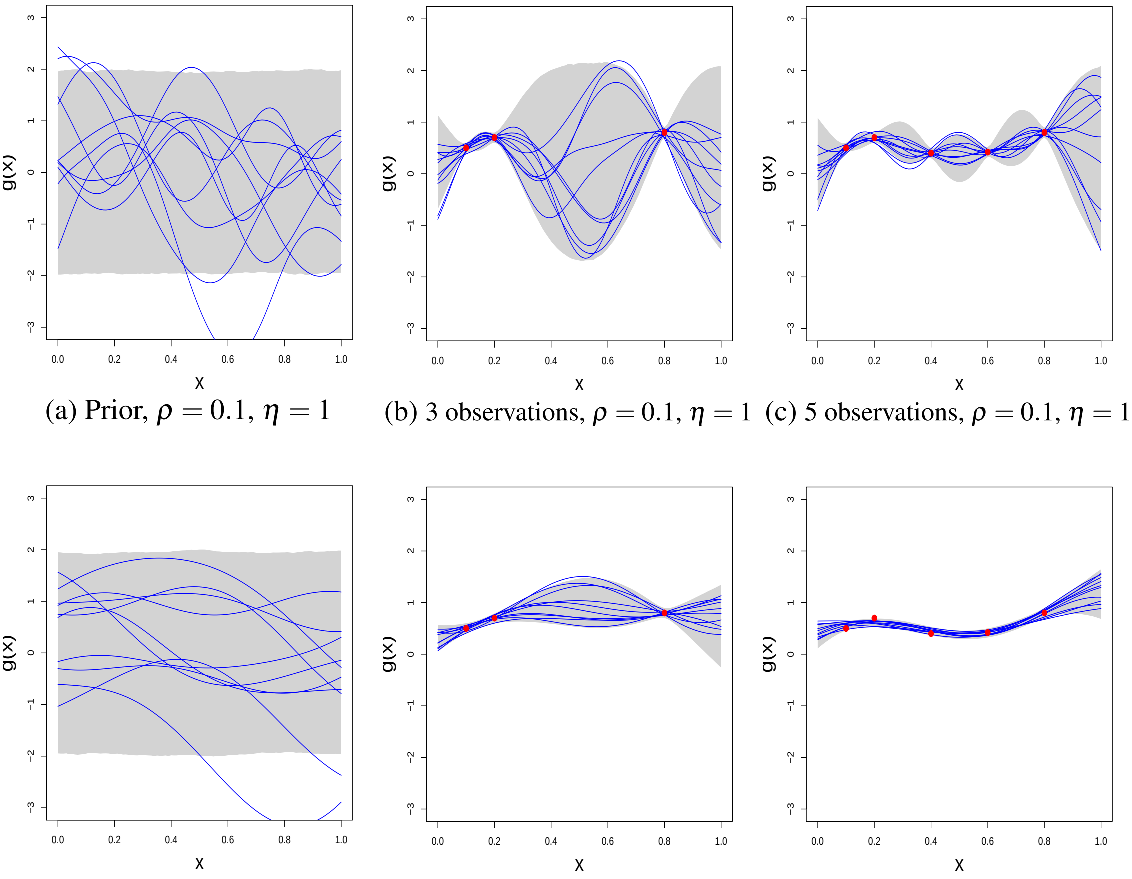
The posterior distribution $\boldsymbol{g}|\boldsymbol{y}$ is only specified at covariates $\boldsymbol{x}$. Prediction at new covariates $\boldsymbol{x}^\star = \{x_1^\star, \ldots x_m^\star\}$ can be achieved by finding the posterior-predictive distribution at $\boldsymbol{x}^\star$, conditional on data $\boldsymbol{y}$. As a GP prior is defined over a set of functions, rather than covariates, it can be extended to include function values $\boldsymbol{g}^\star = [g^\star(x_1), \ldots, g^\star(x_m)]$. The posterior-predictive distribution is derived, as in Equation 3.4.3, by the joint Normality of $(\boldsymbol{y}, \boldsymbol{g}^\star)$:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{g}^\star \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{xx} + \sigma^2 \boldsymbol{I} & \boldsymbol{K}_{xx^\star} \\ \boldsymbol{K}_{xx^\star}^T & \boldsymbol{K}_{x^\star x^\star} \end{bmatrix} \right) \tag{3.4.5}$$

where the (i, j)-entry of any covariance matrix is defined by the covariance functions:

$$(\boldsymbol{K}_{xx})_{i,j} = k(x_i, x_j | \boldsymbol{\phi}) \qquad (\boldsymbol{K}_{xx^\star})_{i,j} = k(x_i, x_j^\star | \boldsymbol{\phi}) \qquad (\boldsymbol{K}_{x^\star x^\star})_{i,j} = k(x_i^\star, x_j^\star | \boldsymbol{\phi})$$

Figure 3.9 gives an illustrative example of how a posterior predictive distribution for unobserved g$(x^\star)$ is sequentially constructed for an increasing number of observations.



(a) Prior, $\rho = 0.1$, $\eta = 1$ (b) 3 observations, $\rho = 0.1$, $\eta = 1$ (c) 5 observations, $\rho = 0.1$, $\eta = 1$

(d) Prior, $\rho = 0.4$, $\eta = 1$ (e) 3 observations, $\rho = 0.4$, $\eta = 1$ (f) 5 observations, $\rho = 0.4$, $\eta = 1$

Fig. 3.9 Illustration of a GP prior (a and d) on functions $\boldsymbol{g}$ and the corresponding posterior-predictive distribution, conditional upon 3 (b and e) and 5 observations (c and f). Two different length-scales ($\rho = 0.1$ - top, $\rho = 0.4$ - bottom) are considered for the covariance function. The shaded grey area and blue lines respectively represent the 95% credible intervals and samples of functions from the GP posterior-predictive.

3.4.4 Inference

Hyperparameters $\boldsymbol{\phi} = \{\rho, \eta\}$ are typically estimated from the data, along with the residual variance σ^2. Even though Gaussian processes are naturally defined within a Bayesian framework, frequentist inference approaches are typically used to avoid computationally expensive MCMC methods (Rasmussen and Williams, 2006, Chapter 5).

The likelihood of the data is obtained by integrating over the set of latent functions $\boldsymbol{g}$:

$$p(\boldsymbol{y}|\boldsymbol{\phi}) = \int p(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{\phi}) p(\boldsymbol{g}|\boldsymbol{\phi}) d\boldsymbol{g} \tag{3.4.6}$$

In the case of Normal data (as in Section 3.4.3), the log-likelihood is equal to $\boldsymbol{y}|\boldsymbol{\phi} \sim N(\boldsymbol{0}, \boldsymbol{K} + \sigma^2 \boldsymbol{I})$, for further details refer to Appendix C.7. If a likelihood with zero-mean $\boldsymbol{y}$ is maximised, then the empirical Bayes estimates will have non-zero mean.

To carry out Bayesian inference, a prior $p(\boldsymbol{\phi})$ is assigned to the hyper-parameters. The posterior distribution is:

$$p(\boldsymbol{\phi}|\boldsymbol{y}) \propto p(\boldsymbol{\phi}) p(\boldsymbol{y}|\boldsymbol{\phi}) = p(\boldsymbol{\phi}) \int p(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{\phi}) p(\boldsymbol{g}|\boldsymbol{\phi}) d\boldsymbol{g}$$

Informative priors can overcome the identifiability issues arising from data being uninformative about $\boldsymbol{\phi}$. Neal (1997) used Hamiltonian Monte Carlo for hyperparameters estimation, as standard Metropolis-Hastings algorithms perform poorly. Recent work has considered Gibbs and slice sampling (Gramacy et al., 2007; Murray and Adams, 2010).

Bayesian inference further integrates uncertainty in the hyperparameters into predictions:

$$p(\boldsymbol{g}^{\star}|\boldsymbol{y}) = \int p(\boldsymbol{g}^{\star}|\boldsymbol{y}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}$$

3.4.5 Gaussian processes beyond scatter-plot smoothing

In the previous Sections, we illustrated how GPR (Section 3.4.3) can be used within a scatter plot smoothing framework. In theory, it is possible to extend this framework to non-normally distributed observations $\boldsymbol{y}$ and inference can be carried as described in Section 3.4.4.

However, in practice this is not straightforward; the posterior distribution $p(\boldsymbol{g}|\boldsymbol{y})$ and likelihood of non-normally distributed outcomes $\boldsymbol{y}$ can not be expressed analytically. This can only be derived via approximations, such as the Laplace approximation (Williams and Barber, 1998) or Expectation Propagation (Minka, 2001).

GP have been extensively used for classification problems, involving binary outcomes (Rasmussen and Williams, 2006, Chapter 3). Chan and Dong (2011) propose Generalized Gaussian Processes to smoothly model the link function $\eta(\mu)$ of a GLM (as in Section 3.3.8) with a GP prior. However the back-calculation model of Chapter 2 can not be expressed as a GLM, and these methods are not applicable.

3.5 Comparing Gaussian processes and splines

Both splines and Gaussian processes are methods for non-parametric regression and can be integrated in a Bayesian regression framework. Consider any linear function in $\boldsymbol{\beta}$:

$$\boldsymbol{g} = \boldsymbol{X\beta}$$

If a multivariate normal prior $\boldsymbol{\beta} \sim N(\boldsymbol{0},\boldsymbol{C})$ is chosen, a Gaussian process prior is implicitly set on the space of functions $\boldsymbol{g} \sim N(\boldsymbol{0},\boldsymbol{XCX^T})$. The Bayesian interpretation of splines (Equation 3.3.14) is a special case of Bayesian regression where $\boldsymbol{g} \sim N(\boldsymbol{0},\boldsymbol{XS}^{-1}\boldsymbol{X}^T)$, *i.e* $\boldsymbol{S}^{-1} \equiv \boldsymbol{C}$. Recall that $\boldsymbol{S}$ may not be of full rank, due to the unpenalised parameters $\boldsymbol{\beta}_U$ having a flat prior. For the previous statement to hold, this is replaced by a wide vague Normal prior $\boldsymbol{\beta}_U \sim N(0,\xi)$ $(\xi \to \infty)$ so that $\boldsymbol{S}$ is of full rank.

Gaussian processes impose a prior over a set of functions $g(x) \sim N(\boldsymbol{0},\boldsymbol{K})$. $\boldsymbol{K}$ is not characterised by basis functions, but by a covariance function $k(x_i,x_j|\boldsymbol{\phi})$. Gaussian processes can, however, be expressed in terms of an infinite number of basis $b(x)$ functions by Mercer's theorem (Williams, 1997), whereas splines are characterised by finite basis functions. On the other hand, splines can be expressed as a GP prior. For instance (Rasmussen and Williams, 2006, Section 6.3) optimal NCS (Section 3.3.3) can be obtained by considering the space of functions $g(x) = \beta_1 + \beta_2 + r(x)$ and imposing a $N(\boldsymbol{0},\boldsymbol{K}_{sp})$ prior on $r(x)$, where $(\boldsymbol{K}_{sp})_{ij} = \frac{|x_i-x_j|v^2}{2} + \frac{v^3}{3}$ and $v = min(x_i,x_j)$. The covariance function is continuous but only once differentiable, hence rougher than the squared exponential kernel. Outside of the first and last data-point of the domain defined by $\boldsymbol{x}$ the posterior mean reverts to the prior mean for the squared exponential kernel, while extrapolates linearly for the spline kernel.

In summary, a Gaussian process is more flexible than a spline: from a Bayesian regression perspective, the latter pre-specifies the covariance matrix while the former allows it to depend upon hyperparametrs. Both Gaussian processes and splines incorporate uncertainty in $g(x)$; however splines assume homoskedastic confidence (or credible) intervals, whereas GP allow for the credible interval size to vary depending on the distance between the $\boldsymbol{x}$ (Figure 3.9).

GP are more computationally demanding: iterative algorithms, such as MCMC or numerical likelihood maximizers, invert the covariance matrix $\boldsymbol{K}$ (with $O(n^3)$ cost) at every iteration. Instead splines pre-specify $\boldsymbol{S}$ and thus require only one matrix inversion.

A number of `R` packages allows fitting splines and GP (see Appendix C.8) within standard models (GLMs, GAMs and scatter-plot smoothing). However these are not flexible enough to fit more complex models, such as the back-calculation discussed in Chapter 2.

3.6 Modelling the latent processes in back-calculation

Chapter 2 introduced multi-state back-calculation, without however specifying parameterisations for the incidence curve $\mathcal{H}(\boldsymbol{\theta})$ and diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$. Sections 3.3 and 3.4 introduced splines and GP within a scatter-plot smoothing framework. This Section shows how these can be embedded within back-calculation, to model the latent $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$ processes. Finally, both penalised likelihood and Bayesian inference are discussed.

3.6.1 Incidence curve

To ensure positiveness of $\mathcal{H}(\boldsymbol{\theta})$, the expected infections are modelled on the log scale. Denote $\gamma_i = log(h_i)$ the log of the expected number of new infections in the i^{th} interval. We will refer to $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)^T$, the vector of log-expected infections, as the log-incidence curve. Three parameterisations are considered for $\boldsymbol{\gamma}$: step functions, splines and GP.

Step functions

Step functions assume that the log-incidence curve is piecewise constant in intervals $(\tilde{t}_{i-1}, \tilde{t}_i]$, where $i = \{1, \ldots, \tilde{T}\}$, $\tilde{T} \leq T$, $\tilde{t}_{\tilde{T}} \equiv t_T$, and $\tilde{t}_0 \equiv t_0$. To improve identifiability, the time intervals $(\tilde{t}_{i-1}, \tilde{t}_i]$ are typically chosen to be large and/or $\boldsymbol{\gamma}$ is subject to smoothing constraints. From a Bayesian perspective smoothness can be incorporated through first and second order random walk priors (Section 1.4). For $i = \{2, \ldots, \tilde{T}\}$:

$$\begin{aligned} \gamma_i &\sim N(\gamma_{i-1}, \sigma_I^2) \\ \gamma_i &\sim N(2\gamma_{i-2} - \gamma_{i-1}, \sigma_I^2) \end{aligned} \tag{3.6.1}$$

A first order random walk behaviour specifies a preference, a priori, for models with a constant number of log-expected infections in successive time periods. A second order random walk favours, a priori, a linear increase (or decrease) in the log-expected number of infections in the i^{th} interval. The variance σ_I^2 controls the smoothing, *i.e.* how much the log-expected infections vary between successive time periods. A prior on γ_1 must be imposed, reflecting beliefs on the log-expected number of infections in $(\tilde{t}_0, \tilde{t}_1]$. Using a piece-wise constant function for $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$, the infection parameters are $\boldsymbol{\theta} = \{\gamma_1, \dots, \gamma_{\tilde{T}}, \sigma_I^2\}$.

Splines

Using a spline model for the log-incidence curve (Section 1.4), let:

$$\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{\beta} \tag{3.6.2}$$

where $\boldsymbol{X}$ is the design matrix of the spline. This parameterisation of the log-incidence curve is characterised by parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \lambda\}$, where λ is the smoothing parameter.

Gaussian processes

If the log-incidence curve is modelled with a Gaussian Process, then:

$$\boldsymbol{\gamma} \sim GP(m(x_i), k(x_i, x_j | \boldsymbol{\phi})) \tag{3.6.3}$$

where $m(x_i)$ is the mean function and $k(x_i, x_j | \boldsymbol{\phi})$ denotes the covariance function, characterised by hyper-parameters $\boldsymbol{\phi}$ (Section 3.4.2). In this case, the parameters are $\boldsymbol{\theta} = \boldsymbol{\phi}$.

3.6.2 Diagnosis process

A model for the diagnosis process $\boldsymbol{\mathcal{D}}(\boldsymbol{\delta})$ needs to be specified. Diagnosis probabilities are considered on a logistic scale $\delta_{k,i} = log\left(\frac{d_{k,i}}{1-d_{k,i}}\right)$ and are assumed to be piece-wise constant in intervals $(\check{t}_{i-1}, \check{t}_i]$, where $i = \{1, \dots, \check{T}\}$, $\check{T} \leq T$, $\check{t}_{\check{T}} \equiv t_T$, and $\check{t}_0 \equiv t_0$.

Logistic regression

One candidate model is logistic regression:

$$\delta_{k,i} = \alpha + \zeta_k + \xi_i + \nu_{k,i} \tag{3.6.4}$$

subject to the identifiability constraint $\zeta_1 = \xi_1 = \nu_{1,i} = \nu_{k,1} = 0$. ζ_k and ξ_i denote the effects of undiagnosed state k and calendar interval $(\check{t}_{i-1}, \check{t}_i]$ respectively. Using a logitic regression parameterisation, parameters $\boldsymbol{\delta} = \{\alpha, \zeta_1, \dots, \zeta_k, \xi_1, \dots, \xi_{\check{T}}, \nu_{1,1}, \dots, \nu_{K,\check{T}}\}$.

Step functions

Alternatively piecewise constant step-functions can be considered. Within a Bayesian framework, independent first order random walk priors for each undiagnosed state can be employed:

$$\delta_{k,i} \sim N(\delta_{k,i-1}, \sigma^2_{D,k}) \tag{3.6.5}$$

The diagnosis parameters are $\boldsymbol{\delta} = \{\delta_{1,1}, \dots, \delta_{1,\check{T}}, \dots, \delta_{K,1}, \dots, \delta_{K,\check{T}}, \sigma^2_{D,1}, \dots, \sigma^2_{D,K}\}$.

Splines and Gaussian Processes

Analogously to the infection process, the logistic diagnosis probabilities could have been modelled with a spline or a GP. To avoid examining an excessive number of parameterisations, in this thesis we only considered the two latter parameterisations, following Birrell et al. (2012). We focus on parameterisations for the incidence curve, rather than for the diagnosis probabilities, as incidence estimates are more important for public health purposes.

3.6.3 Inference

Irrespectively of the parameterisation chosen, inference is not straightforward. Multi-state back-calculation can not be expressed as a GLM; the likelihood (Equation 2.3.10) includes two Poisson and one Multinomial term and a single link function can not be specified. The expected number of diagnoses is a complex non-linear function of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ (Equations 2.3.4 and 2.3.6). This implies that standard results and software allowing to model the expected response of a GLM as a spline or as a Gaussian process can not be employed. Moreover the derivatives of the likelihood are not analytically tractable, thus the expectation-maximization equations of the EM(S) algorithm can not be computed.

We investigated a penalised likelihood approach to inference by modelling $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$ with a spline (Equation 3.6.2) and logistic regression (Equation 3.6.4) respectively. Within a GLM framework (Section 3.3.8) standard algorithms are available to numerically maximise the penalised likelihood and to obtain an optimal smoothing parameter λ (see Appendix C.5.1

and C.5.2). However these are not applicable to multi-state back-calculation. We thus numerically maximized the penalised likelihood using the quasi-Newton BFGS algorithm (Nash, 1990) via the `R` function *optimx* (Nash and Varadhan, 2011). Following Wood et al. (2016), an AIC criterion can be derived to estimate the optimal $\hat{\lambda}$ and confidence intervals can be constructed, as in Appendix Section C.5.3, from a large-sample posterior distribution that is obtained re-interpreting the penalty as a prior. This procedure suffers from a number of drawbacks: firstly, the model has to be refitted for a number of plausible λs, to determine $\hat{\lambda}$, which is computationally intensive. Secondly, only approximate confidence intervals, not accounting for uncertainty in λ, can be obtained.

Note that a Bayesian approach overcomes both limitations, as both a posterior distribution for λ and non-approximated credible intervals can be obtained. Bayesian inference has a number of further advantages. Firstly, it does not require the likelihood to be arranged in a convenient form (*i.e.* GLM). Secondly, latent variables and parameters can both be readily sampled using MCMC hence latent random walks and Gaussian processes, can be easily used to model $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$. Most importantly, a Bayesian framework allows to easily incorporate external sources of information (*e.g.* under-reporting levels) and additional data-sources (*e.g.* the CD4-count data) whilst enabling a coherent propagation of uncertainty. Rather than developing an efficient bespoke MCMC sampler, general purpose Bayesian inference software, such as `JAGS` (Plummer, 2003) and `Stan` (Stan Development Team, 2016b) is used (Appendix A.4). This ensures portability of the model, as the available code can easily be used, and if needed modified, for other applications. Codes are available on Github (https://github.com/frbrz25/Thesis_Codes).

3.7 Summary

This Chapter starts by describing the properties of two non-parametric smoothing methods: splines (Section 3.3) and Gaussian processes (Section 3.4). These are first compared (Section 3.5) and are then embedded within a more sophisticated back-calculation model, introduced in Chapter 2. After describing the back-calculation model, proposing suitable parameterisations and discussing appropriate inferential methods (Section 3.6), in the following Chapter we will fit the back-calculation model.

Chapter 4

Age independent back-calculation simulations

4.1 Introduction

Chapter 2 introduced multi-state back-calculation, which characterised the observed surveillance data as a complex function of the latent incidence curve $\mathcal{H}(\boldsymbol{\theta})$ and diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$. This, along with the ill-posed nature of back-calculation and the large uncertainty characterising the most recent years considered, renders the estimation of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ challenging (Section 3.6). Chapter 3 discussed a number of non-parametric smoothing models for $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$. This Chapter investigates, through a simulation study the properties of these different smoothing methods within a back-calculation framework, to establish whether some are more suitable than others.

This Chapter is structured as follows: first, in Section 4.2, the data-generating mechanism is outlined. Section 4.3 discusses specific parameterisations for the non-parametric models employed for $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$. Sections 4.4 and 4.5 describe the setup and the performance assessment of the simulation study. Finally, results are presented in Section 4.6.

4.2 Data generating mechanism

Here, the back-calculation model described in Section 2.3 is employed with $K = 4$ latent undiagnosed states (Figure 4.1). A quarterly time scale, allowing at most one movement

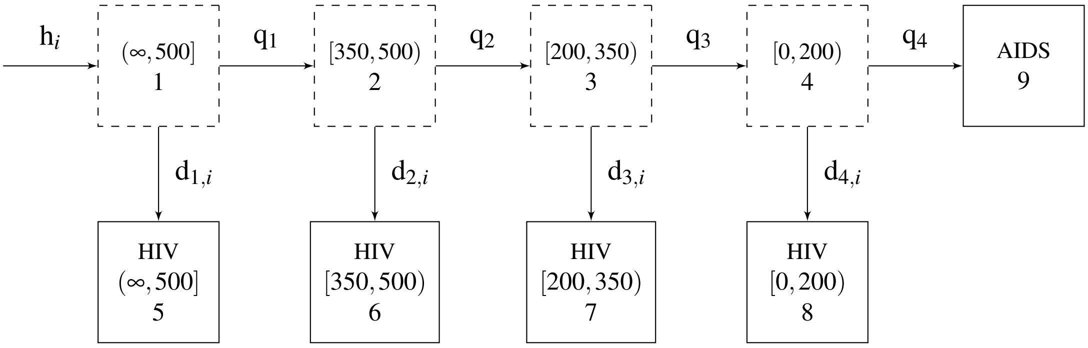


Fig. 4.1 Back-calculation multi-state model, used for this simulation study with $K = 4$ undiagnosed states. Dashed states $\{1,\ldots,4\}$ denote undiagnosed states. Solid states $\{5,\ldots,9\}$ denote diagnosed states.

between states per time interval is employed. Back-calculation is run from an intermediate point of the epidemic (Section 2.4.2) for 80 quarters (*i.e.* $T = 80$), without including underreporting (Section 2.4.1).

The following data-generating mechanism is considered: the true incidence curve $\mathcal{H}^\star$, true diagnoses $\mathcal{D}^\star$ and progression $\mathbf{q}^\star$ probabilities are first specified, along with the expected number of initially undiagnosed infections $\boldsymbol{\pi}^\star$. True progression and diagnosis transition matrices $\boldsymbol{D}_i^\star$ (Equation 2.3.2) and $\boldsymbol{Q}_i^\star$ (Equation 2.3.1) as well as the true expected number of individuals in each diagnosis state $\boldsymbol{\mu}_i^\star$ (Equation 2.3.4) are then implicitly determined, for each interval $(t_{i-1}, t_i]$, $i = \{1, \ldots 80\}$. In turn, this specifies $\mu_i^{H\star}$, $\mu_i^{A\star}$ and $\boldsymbol{p}_i^\star$ the true expected number of HIV diagnoses, AIDS diagnoses and CD4 proportions respectively over time (see Section 2.3.4). HIV and AIDS diagnoses counts in the i^{th} interval, $Y_i^{H\star}$ and $Y_i^{A\star}$, can be simulated from independent Poisson distributions with means $\mu_i^{H\star}$ and $\mu_i^{A\star}$ respectively (Equations 2.3.7 and 2.3.8 respectively). Finally, CD4 diagnosis counts $\boldsymbol{Y}_i^{C\star}$ in the i^{th} interval are simulated from a $Multinomial(n_i^\star, \boldsymbol{p}_i^\star)$ (Equation 2.3.9). The last step requires specification of $n_i^\star$.

The values chosen for $\mathcal{H}^\star$, $\mathcal{D}^\star$, $\mathbf{q}^\star$, $\boldsymbol{\pi}^\star$, and $n_i^\star$ to simulate the data, are realistic values for the MSM-HIV epidemic in England and Wales between 1995 and 2015. These are consistent with findings from previous studies (Aalen et al., 1997; Sweeting et al., 2005; Birrell et al., 2012).

Three plausible true incidence curves $\mathcal{H}^\star$ are considered: an increasing, a decreasing, and a flat one. These are assumed to be the same over the time span of the epidemic, apart from the three most recent years (see Figure 4.2). We are particularly interested in understanding how well incidence can be estimated in most recent years, as diagnosis data are typically only weakly informative regarding recent infections (Section 1.4). The true incidence curves were purposely not obtained from any non-parametric model, not to advantage any of the non-parametric models investigated.

Likewise true diagnosis probabilities $\mathcal{D}^\star$ were not generated from a non-parametric model. These are obtained by adding some zero-mean Gaussian noise to the diagnosis probabilities estimated for the MSM-HIV epidemic in England and Wales, using the age-independent back-calculation model (Chapter 2), with data up to end of 2015 (Kirwan et al., 2016). Figure 4.3 shows that diagnosis probabilities, from each undiagnosed state, are slowly

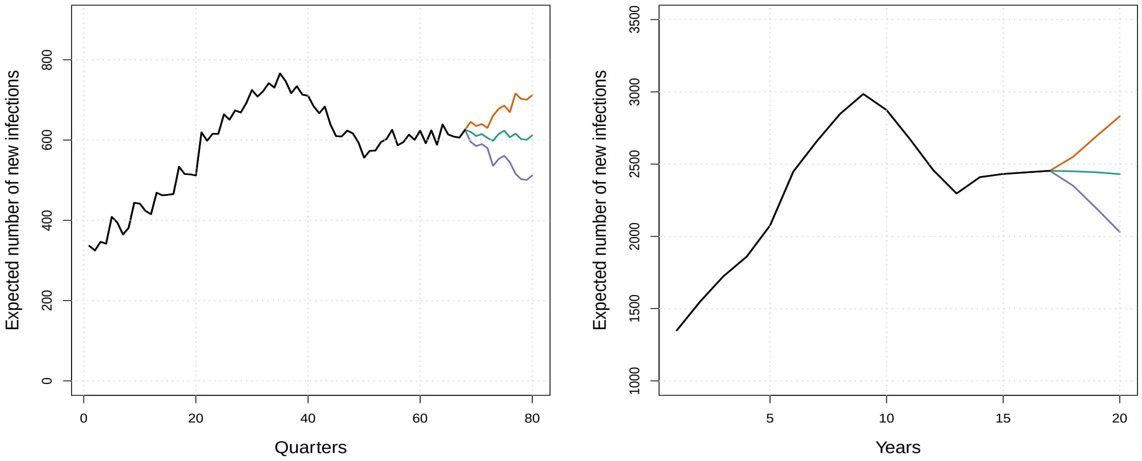


Fig. 4.2 True incidences: increasing (orange), flat (green) and decreasing (grey) plotted quarterly (left) and yearly (right). The section where the curves coincide is plotted in black.

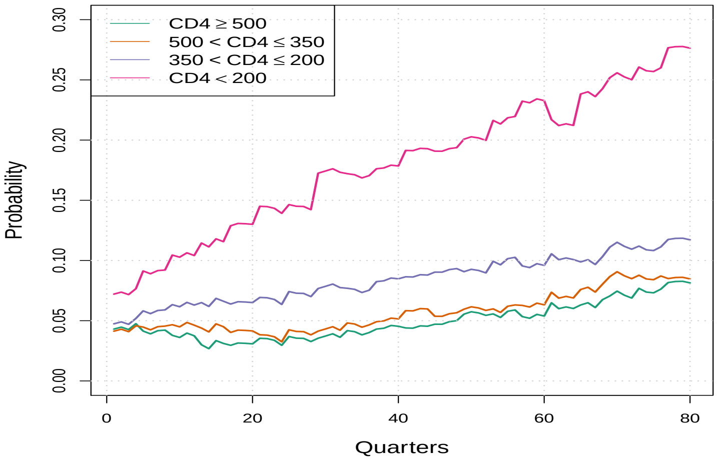


Fig. 4.3 True diagnosis probabilities, by state.

increasing over time. Progression probabilities between undiagnosed states are set to $\mathbf{q}^\star = (0.0976, 0.1150, 0.1160, 0.1480)^T$ following Birrell et al. (2012) and references therein (CASCADE Collaboration, 2000). The number of expected initially undiagnosed infections is set to $\boldsymbol{\pi}^\star = (1710.6796, 1191.2027, 1191.2027, 870.0087)^T$. This is obtained by applying an age-specific extension of the model by Aalen et al. (1997) to the MSM-HIV surveillance data in England and Wales from 1978 to 1994. $n_i^\star$ is chosen so that the percentage of CD4 linked diagnoses over time is the same as in the surveillance dataset for the MSM-HIV epidemic in England and Wales (Chapter 8). For instance if in $(t_0, t_1]$: $Y_1^{H\star} = 100$ and 10% of diagnoses are CD4-linked in the existing surveillance data, then $n_1^\star = 10$.

4.3 Back-calculation parameterisations

As demonstrated in Section 3.6, it is convenient to carry out inference within a Bayesian framework; this Section discusses appropriate priors for the non-parametric methods for the incidence curve $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ and diagnosis probabilities $\boldsymbol{\mathcal{D}}(\boldsymbol{\delta})$ discussed in Sections 3.6.1 and 3.6.2.

4.3.1 Incidence curve

Random walk

The log-incidence $\boldsymbol{\gamma}$ can be modelled with a first or second order random walk (Equation 3.6.1), allowed to vary at every quarter, *i.e.* $(\tilde{t}_{i-1}, \tilde{t}_i]$, $i = \{1, \ldots, 80\}$, following the notation of Section 3.6.1. A prior needs to be imposed on the starting values of the random walk (γ_1 and γ_1, γ_2 for a first and second order random walk respectively) and the variance σ_I^2:

$$\begin{aligned}
\gamma_1 &\sim N(5.52, 0.29)\\
\gamma_2 &\sim N(5.52, 0.29)\\
\sigma_I^2 &\sim \Gamma(1, 32)
\end{aligned}$$

These are weakly informative priors on the level of log-expected infections in $(\tilde{t}_0, \tilde{t}_1]$ and $(\tilde{t}_1, \tilde{t}_2]$ so that the expected number of quarterly infections in these intervals a priori lies in the (140,445) interval. The variance σ_I^2 of the log-random walk is given a weakly informative Gamma prior, so that the standard deviation σ_I lies with 95% prior probability in (0.03,0.335).

Such prior allows the log-random walk to approximately vary by two standard deviations (*i.e.* by at most 70%) between successive time intervals.

Splines

Here we consider the following splines to model $\boldsymbol{\gamma}$: knots based NCS (Section 3.3.4), thin plate regression splines with and without linear shrinkage (Sections 3.3.5 and 3.3.6), and first and second order P-splines (Section 3.3.7). Equation 3.3.15 (here characterised by parameters $\boldsymbol{\beta}$ rather than $\boldsymbol{\beta}'$ and by the standard deviation σ rather than the precision λ, for notational convenience) allows expressing splines in a Bayesian framework. In this simulation study, the following weakly-informative priors are imposed:

$$\begin{aligned}
\beta_1 &\sim N(5, 1.5) \\
\beta_{P_i} &\sim N(0, \sigma^2), \quad i = \{1, \dots, p\} \\
\beta_{U_i} &\sim N(0, \sigma_0^2), \quad i = \{1, \dots, u-1\} \\
\sigma &\sim t_+(4, 20) \\
\sigma_0 &\sim t_+(4, 20)
\end{aligned}$$

where β_1 is a global intercept, describing the average number of log-expected infections per quarter; the Normal prior chosen imposes that this the expected number of quarterly infections lies, with 95% prior probability, in (7,2980).

$\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_U$ are the p penalised and the u unpenalised coefficients of the spline respectively. All splines employed have been considered with 10 parameters (*i.e.* $p+u=10$). The knots locations need to be chosen for knots-based NCS: equidistant knots were placed on the $(\tilde{t}_0, \tilde{t}_{\tilde{T}}]$ range. The normal-prior employed for these coefficients are obtained from a Bayesian re-interpretation of the penalty matrix $\boldsymbol{S}$ (Section 3.3.9).

Finally $\sigma^2 = 1/\lambda$ and $\sigma_0^2 = 1/\lambda_0$ are inverse smoothing parameters related to penalised and unpenalised parameters respectively. Recall that (Section 3.3.9) thin plate regression splines with shrinkage and P-splines of first order do not require the additional smoothing parameter λ_0. $t_+(d,s)$ denotes a half-t distribution, the absolute value of a Student-t distribution; this is defined on the $[0,\infty)$ range, it is monotonically decreasing from zero, and it is characterised by d degrees of freedom and a scale parameter s. Following Gelman et al. (2006) and Stan Development Team (2016b), a $t_+(4,20)$ is chosen as prior for σ and σ_0 as this is a weakly-informative prior constructed so that 95% of the prior density of σ and σ_0 lies within the region [0,40]. Penalised and unpenalised coefficients are given the same prior following

Wood (2016); a priori there is no reason to believe that one of the two sets of parameters should be penalised more heavily than the other.

Gaussian Process

Finally, $\boldsymbol{\gamma}$ is modelled with a GP (Section 3.4):

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{K})$$

The covariates x_i of the GP correspond to the time scaled to the $[0,1]$ range. This is achieved using the following transformation, for $i = \{1,\dots 80\}$:

$$x_i = \frac{t_i - t_1}{t_{80} - t_1}$$

The $(i,j)^{\text{th}}$ entry of $\boldsymbol{K}$ is characterised by a squared exponential covariance function:

$$k(x_i,x_j) = \eta^2 exp\left(-\frac{(x_i - x_j)^2}{2\rho^2}\right) + \mathbb{1}_{x_i=x_j} 0.0001$$

where $\mathbb{1}_{x_i=x_j}$ is an indicator function equal to 1 if $x_i = x_j$ and to 0 otherwise and η and ρ are the magnitude and length-scale hyper-parameters of the GP. Note that an extra small value (0.0001) has been added to the squared exponential function, when $x_i = x_j$. This is to avoid that lack of numerical precision renders the covariance matrix singular and hence not invertible. Hyper-parameters η and ρ are restricted to be positive (to avoid non-identifiability arising from $(-\eta)^2 = \eta^2$ and similarly for ρ) and are assigned priors:

$$\begin{aligned} \eta &\sim N_{+}(4,1) \\ \frac{1}{\rho} &\sim t_{+}(4,1) \end{aligned}$$

where N_{+} denotes a Normal distribution truncated at zero, and t_{+} a half-t distribution.

The magnitude parameter η can be thought of as an approximation of the standard deviation of γ_i, as the diagonal entries of $\boldsymbol{K}$ are $k(x_j,x_j) \approx \eta^2$. Thus, a weakly informative Normal prior can be employed so that γ_i lies in the 90% prior range $[-2\eta, 2\eta]$ a priori. Our prior choice implies that the log-expected number of quarterly infections lie in $[-12,12]$ a priori. Length-scales ρ larger than the covariates' range lead to straight line curves (Stan Development Team, 2016b). As our covariates lie in the $[0,1]$ domain, a half-t prior distribution is chosen with four degrees of freedom and scale parameter equal to one, so that the 90% prior mass of the inverse length-scale $1/\rho$ is concentrated in $[0,2]$ and smoother models are favoured over rougher ones.

4.3.2 Diagnosis probabilities

Finally, diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$ are parameterised: a logistic random walk (Equation 3.6.5) is chosen, allowed to vary at every quarter, *i.e.* $(\check{t}_{i-1}, \check{t}_i]$, $i = \{1, \ldots, 80\}$ using the notation of Section 3.6.2. The initial logistic random walk $\delta_{k,1}$ values, and all variance parameters are given the following prior distributions:

$$\delta_{1,1} \sim N(-3.2, 0.2), \;\; \delta_{1,2} \sim N(-3.2, 0.2), \;\; \delta_{1,3} \sim N(-3, 0.2), \;\; \delta_{1,4} \sim N(-2.5, 0.3)$$

$$\sigma^2_{k,D} \sim \Gamma(1, 32), \;\; k = \{1,2,3,4\}$$

Weakly informative priors are assigned to $\delta_{k,1}$. The prior for $\sigma^2_{k,D}$ is interpreted as the prior for σ^2_I, when specifying a first order random walk for the log-incidence, with the difference that the random walk for diagnosis probabilities is defined on the logistic scale.

4.4 Simulation study setup

A comparison of non-parametric models for log-incidence is now carried out using:

- 50 datasets generated for each of the three true incidence curve options (increasing, decreasing, flat) resulting in a total of 150 datasets. The term *true incidence scenario* will refer to the datasets generated under a specific true incidence (*e.g.* increasing).
- The latent incidence curve is modelled on each dataset based on eight different parameterisations: first and second order random walks (abbreviated *rw1ord* and *rw2ord* respectively), knots based NCS (*tpknotsloc*), thin plate spline without and with shrinkage (*tp* and *ts* respectively), first and second order cubic P-splines (*bsord1* and *bsord2*) and Gaussian Processes (*GP*). The term *incidence model* will refer to a specific parameterisation of the incidence curve (*e.g. bsord1*).
- The term *simulation* describes the combination of a true incidence scenario (*e.g.* increasing), one incidence model (*e.g. bsord1*) and a dataset (*e.g.* dataset number 25); 1200 simulations have been undertaken.

- For each simulation, inference is carried out using `Stan`, which employs Hamiltonian Monte Carlo (HMC) methods (see Appendix A.2.2 and A.4). Three chains of length 2000 and burn-in of 1000 are used, resulting in a posterior sample of size 3000. Default initial values are automatically generated by `Stan`.

4.5 Simulation study performance assessment

This Section covers performance evaluation within the simulation study considered, based on a prominent Bayesian simulation study by Browne and Draper (2006) and various, non-Bayesian, simulation studies on splines by Wood (*e.g* Wood, 2003).

Within a Bayesian framework, parameters (and functions of parameters) have a posterior distribution for each simulation m, conveniently summarized by the posterior mean. For simulation m, denote the posterior-mean of the incidence curve and diagnosis probabilities from state k over time as $\widehat{\mathcal{H}}_m = \{\widehat{h}_1^m, \dots, \widehat{h}_T^m\}$ and $\widehat{\mathcal{D}}_{k,m} = \{\widehat{d}_{k,1}^m, \dots, \widehat{d}_{k,T}^m\}$ respectively. These are the (Bayesian) *estimates* of the true incidence curve and diagnosis probabilities for simulation m. Uncertainty is quantified through $\alpha\%$ credible intervals, corresponding to the region lying within the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution of the parameter (or function of parameter) of interest. For simulation m, the quantile $\frac{\alpha}{2}$ of the incidence curve and diagnosis probabilities from state k are denoted as $\widehat{\mathcal{H}}_m^{\alpha/2} = \{\widehat{h}_1^{m,\alpha/2}, \dots, \widehat{h}_T^{m,\alpha/2}\}$ and $\widehat{\mathcal{D}}_k^{m,\alpha/2} = \{\widehat{d}_{k,1}^{m,\alpha/2}, \dots, \widehat{d}_{k,T}^{m,\alpha/2}\}$.

In simulation studies, performance is typically assessed via the Mean Squared Error (MSE) which measures the bias and the variance of an estimate. In this study, the true incidence curve $\mathcal{H}^\star = \{h_1^\star, \dots, h_T^\star\}$ is not a parameter and is instead characterised by the expected number of true infections in intervals $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$ (see Section 4.2); these are estimated by the posterior mean of the incidence curve $\widehat{\mathcal{H}}_m = \{\widehat{h}_1^m, \dots, \widehat{h}_T^m\}$. Performance is therefore assessed via the Predictive Mean Squared Error (PMSE), which is the mean of squared errors, for the m^{th} simulation:

$$PMSE(\widehat{\mathcal{H}}_m) = \frac{1}{T}\sum_{i=1}^{T}\left(\widehat{h}_i^m - h_i^\star\right)^2 \tag{4.5.1}$$

Higher $PMSE(\widehat{\mathcal{H}}_m)$ values indicate a higher mean squared bias. The mean-PMSE is derived over the simulations as:

$$MPMSE(\widehat{\mathcal{H}}) = \frac{1}{M}\sum_{m=1}^{M} PMSE(\widehat{\mathcal{H}}_m) \tag{4.5.2}$$

$MPMSE(\widehat{\mathcal{H}})$ is calculated for each incidence model, under each true incidence scenario. When comparing two, or more, incidence models (under the same incidence scenario) the one associated with lower $MPMSE(\widehat{\mathcal{H}})$ is considered to more accurately reconstruct the true incidence curve. The distribution of $PMSE(\widehat{\mathcal{H}}_m)$ could be alternatively assessed (*e.g* using a box-plot, see Section 4.6.3).

The same criterion is used to evaluate how well true diagnosis probabilities from state $k = \{1,\ldots,4\}$ are reconstructed. The PMSE for the m^{th} simulation, for diagnosis probabilities from state k is:

$$PMSE(\widehat{\mathcal{D}}_{k,m}) = \frac{1}{T}\sum_{i=1}^{T}\left(\widehat{d}^{m}_{k,i} - d^{\star}_{k,i}\right)^2 \tag{4.5.3}$$

and the mean $MPMSE(\widehat{\mathcal{D}}_k)$ can be calculated as in Equation 4.5.2. It is not necessary to consider all time intervals in the calculation of PMSE. It is crucial to evaluate the performance of estimates in the most recent years, as these are characterised by large uncertainty: Section 4.6.3 considers the distribution of $PMSE(\widehat{\mathcal{H}}_m)$ and $PMSE(\widehat{\mathcal{D}}_{k,m})$ for the last twelve quarters (three years).

Another important performance measure is coverage. For the incidence curve, in a given time interval $(t_{i-1}, t_i]$, the true value $h^{\star}_i$ may or may not lie within credible intervals. The $\alpha\%$-coverage for the m^{th} simulation is defined to be the percentage of intervals in which $h^{\star}_i$ lies within the estimated confidence intervals:

$$Covg_{\alpha}(\widehat{\mathcal{H}}_m) = \frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{h^{\star}_i \in \left[\widehat{h}^{m,\alpha/2}_i,\, \widehat{h}^{m,1-\alpha/2}_i\right]} \tag{4.5.4}$$

$\mathbb{1}_{h^{\star}_t \in [\widehat{h}^{m,\alpha/2}_t,\, \widehat{h}^{m,1-\alpha/2}_t]}$ is an indicator function equal to one when the true expected infection values $h^{\star}_i$ in the i^{th} interval lie within credible intervals. $\alpha\%$-coverage should be equal to the nominal α value, for well specified models. In Section 4.6.3, the distribution of $Covg_{\alpha}(\widehat{\mathcal{H}}_m)$, over the datasets, is analysed. Mean $\alpha\%$-coverage is equal to:

$$MCovg_{\alpha}(\widehat{\mathcal{H}}) = \frac{1}{M}\sum_{m=1}^{M} Covg_{\alpha}(\widehat{\mathcal{H}}_m) \tag{4.5.5}$$

Coverage of true diagnosis probabilities is similarly defined:

$$Covg_{\alpha}(\widehat{\mathcal{D}}_{k,m}) = \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{d^{\star}_{k,i} \in \left[\widehat{d}^{m,\alpha/2}_{k,i},\, \widehat{d}^{m,1-\alpha/2}_{k,i}\right]} \tag{4.5.6}$$

Mean $\alpha\%$-coverage for diagnosis probabilities from state k can be calculated as in Equation 4.5.5.

4.6 Simulation study results

4.6.1 Convergence assessment

As simulations are carried in a Bayesian framework, assessing convergence is key, yet non-trivial. Inspecting the univariate trace plots of MCMC samples from each parameter quickly becomes infeasible when, as in this situation, a large number of parameters in an even larger number of simulations must be checked. Thus convergence is assessed using the $\hat{R}$ statistics (Gelman and Rubin, 1992), specifically a simulation is considered not to have converged if $\hat{R} > 1.05$ for any parameter (refer to Appendix A.3 for more details). Table 4.1 shows the percentage of non-convergent simulations under different incidence models and scenarios. Very few simulations have not converged, and in most of these the inspection of trace plots for the few parameters with $\hat{R} > 1.05$ reveals satisfactory mixing. Defining $\hat{R} > 1.05$ for a single parameter as non-convergence may be too stringent, however to avoid any ambiguity in the simulation study conclusions all non-convergent simulations have been discarded.

%	*bsord1*	*bsord2*	*GP*	*rw1ord*	*rw2ord*	*tp*	*tpknotsloc*	*ts*
Increasing	0	0	0	6	58	0	0	0
Flat	0	0	4	4	42	0	0	0
Decreasing	0	0	14	0	56	0	0	0

Table 4.1 Percentage of simulations that have not converged by true incidence scenarios and models.

%	*bsord1*	*bsord2*	*GP*	*rw1ord*	*rw2ord*	*tp*	*tpknotsloc*	*ts*
Increasing	0	4	0	0	9	84	82	0
Flat	0	12	0	0	24	90	86	0
Decreasing	0	44	0	0	9	82	100	0

Table 4.2 Percentage of simulations that have at least one divergent transitions by true incidence scenarios and models, after removing simulations that have not converged.

Divergent transitions are an important diagnostic tool for HMC; these indicate that some regions of the equilibrium distribution may be hard to explore, resulting in biased estimates (see Appendix A.3 for more details). Table 4.2 shows how many simulations have at least one divergent transition, after excluding non-convergent simulations. This reveals two incidence model "groups", those with few or no divergent transitions (*rw1ord*, *GP*, *ts*, *bs1ord*) and those with multiple divergent transitions (*rw2ord*, *tp*, *tpknotsloc*, *bs2ord*).

Understanding the causes of divergent transitions and non-convergence is key. The *rw2ord* incidence model is first examined, as it suffers from both problems. Figure 4.4 plots posterior samples of the random walk variance σ_I^2 against the corresponding posterior values of the log-posterior; red dots indicate the σ_I^2 values leading to divergent transitions. Note that higher log-posterior values are achieved with variances close to zero; however divergent transitions show that the region of the parameter space where σ_I^2 is near 0 is ill-explored. Trace plots in Figure 4.5 further illustrate slow mixing and high auto correlation for σ_I^2. Finally Figure 4.6 shows the respective trace plots for σ_I^2 for the *rw1ord* incidence model: mixing, despite some auto correlation, is substantially better. Indeed *rw1ord* models do not suffer from non-convergence nor divergent transitions. De Angelis et al. (1998) also highlighted instability issues in estimating the variance parameter of second order random walks, in a similar Bayesian back-calculation framework.

Note that some splines (*tpknotsloc*, *tp*, *bsord2*) exhibit divergent transitions, while other (*ts*, *bsord1*) do not. This is due to the number of smoothing parameters to be estimated. Some splines (*tpknotsloc*, *tp*, *bsord2*) require a second smoothing parameter λ_0 to obtain proper priors (Section 3.3.9). This likely induces convergence issues; the parameter λ_0 is the prior precision of a single parameter β_{U_1} so that there may not simply be enough data for robust estimation. Figures 4.7 and 4.8 provide further evidence: posterior λ_0 values mostly lie around zero, but occasionally wander off to very high values, where divergent transitions occur. The smoothing parameter λ is instead the prior precision of eight parameters $\{\beta_2, \dots, \beta_8\}$, and can thus be estimated from the data as shown in Figure 4.9. *ts* and *bsord1* splines, that only have a single smoothing parameter, do not suffer from divergent transitions (Figure 4.10).

One could attempt to eliminate, or reduce the number of divergent transitions by increasing the HMC resolution (Appendix A), by reparameterising the model or preventing parameters to wander in low-probability regions, typically by imposing more informative priors. However, in this simulation study the HMC resolution is set to the maximum allowed, thus divergent transitions highlight problems in the posterior distribution geometry. The non-centered reparameterisation, which is typically employed in the case of divergent transitions

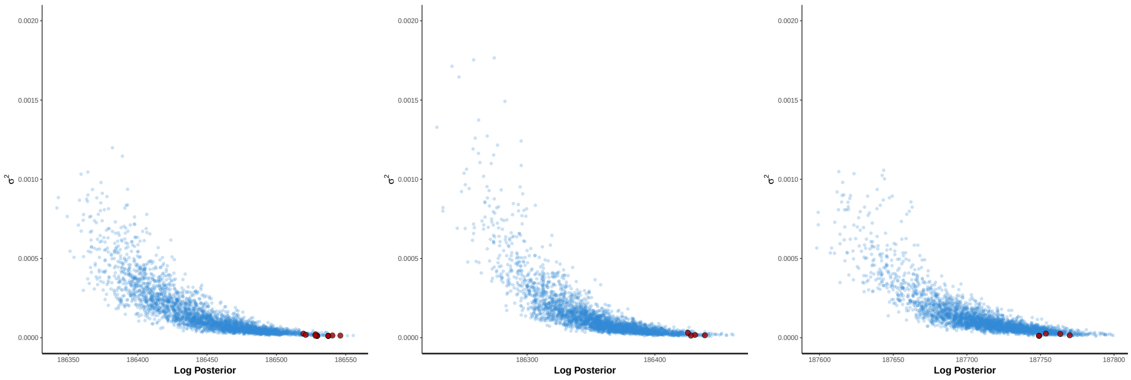


Fig. 4.4 Scatterplots of the posterior variance σ_I^2 values against corresponding log-posterior values for *rw2ord* in three illustrative simulations; red dots denote divergent transitions.

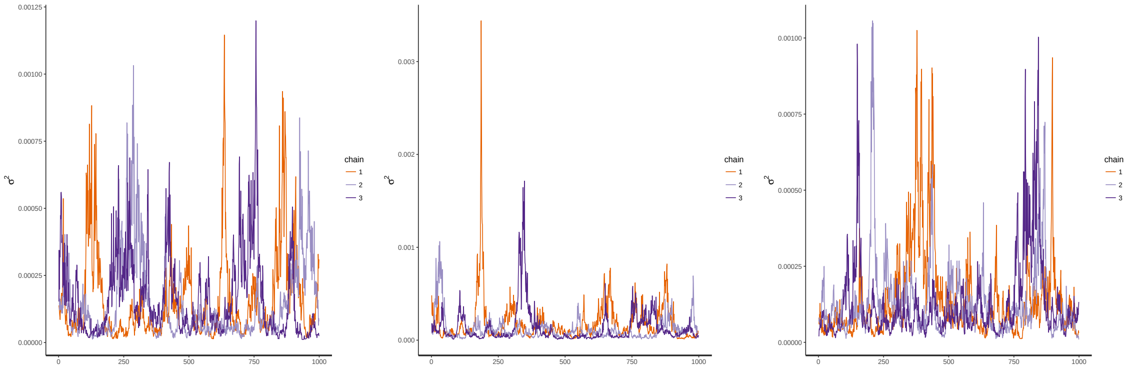


Fig. 4.5 Trace plots of the posterior variance σ_I^2 values for *rw2ord* in three illustrative simulations.

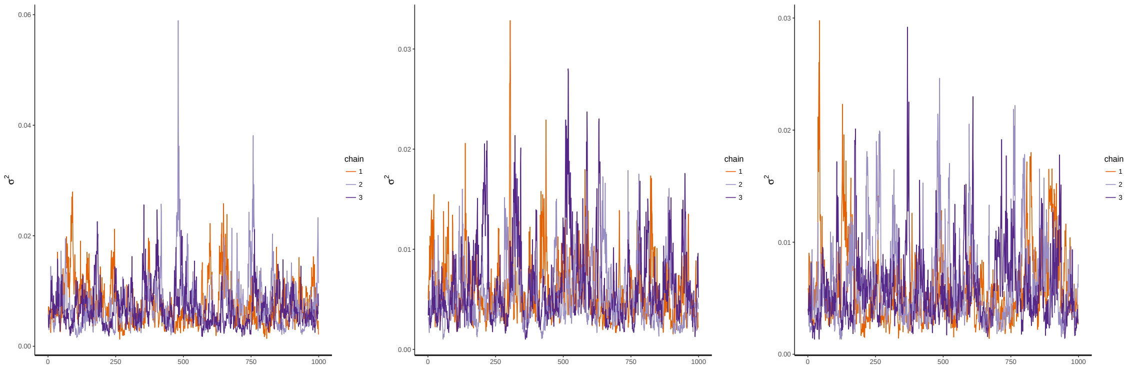


Fig. 4.6 Trace plots of the posterior variance σ_I^2 values for *rw1ord* in three illustrative simulations.

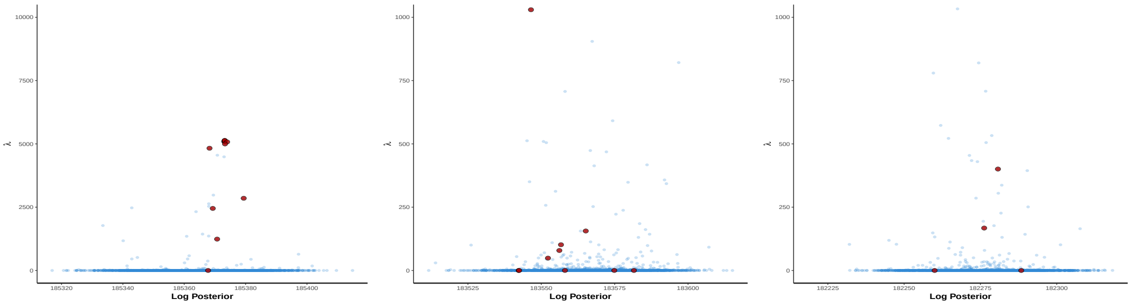


Fig. 4.7 Scatterplot of the posterior smoothing parameter λ_0 values against log-posterior values, for *bsord2* (left), *tp* (center) and *tpknotsloc* (right) in three illustrative simulations. Red dots denote divergent transitions.

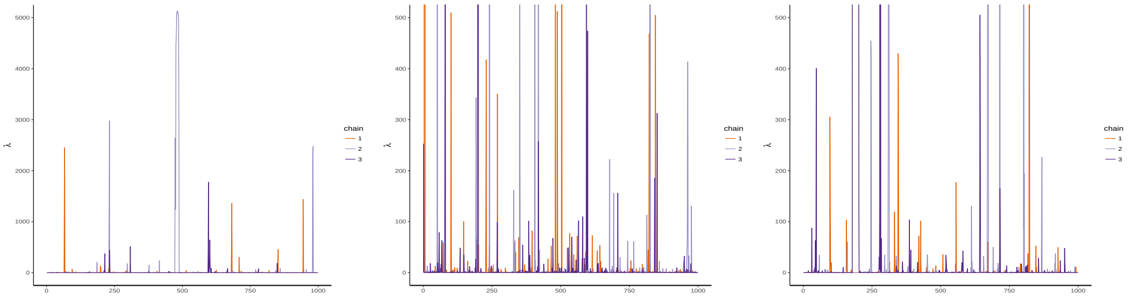


Fig. 4.8 Trace plots for the posterior smoothing parameter λ_0 values for *bsord2* (left), *tp* (center) and *tpknotsloc* (right) in the three illustrative simulations.

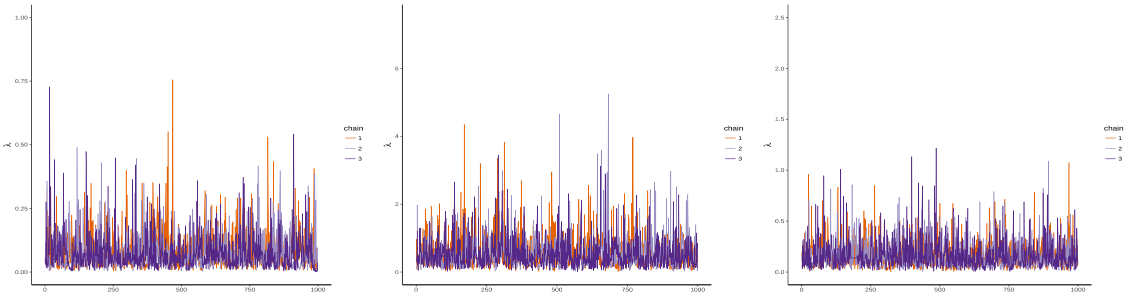


Fig. 4.9 Trace plots for the posterior smoothing parameter λ values, for *bsord2* (left), *tp* (center) and *tpknotsloc* (right) in the three illustrative simulations.

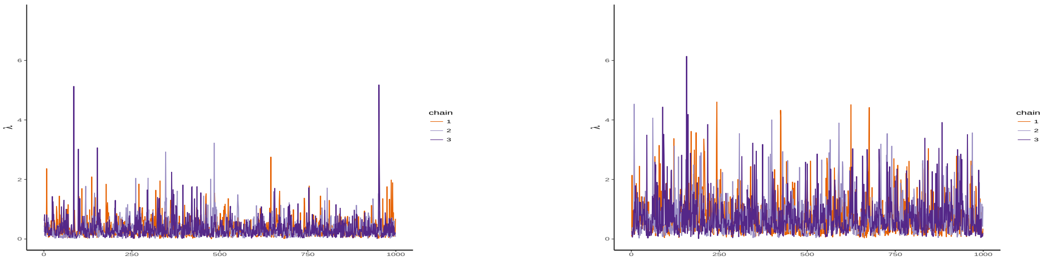


Fig. 4.10 Trace plots for the posterior smoothing parameter λ values, for *bsord1* (left) and *ts* (right) in the three illustrative simulations.

(Betancourt, 2017b), leads to an even higher number of divergent transitions. Finally, most divergent transitions occur in parameter space regions where λ_0 is large; an informative prior forcing λ_0 away from large values, *i.e.* pushing σ_0 (equal to $1/\sqrt{\lambda_0}$) away from 0, could be used. Recall that forcing σ_0 away from 0 reduces the model's smoothness, however it is hard to judge a priori what are the sufficient level of smoothing required. It is more sensible to simply favour, a priori, smoother over complicated models (by employing a half-t prior for σ and σ_0). A number of weakly informative priors for σ and σ_0 (*i.e.* changing the scale parameter of the t_+ prior distribution) have been tested; estimates of σ are robust to prior specifications, while all prior choices for σ_0 result in poor mixing and divergent transitions.

Based on the convergence results, it can be concluded that there is not enough data to estimate the smoothing parameter λ_0 on the null-space. This is not always the case, for instance Wood (2016) successfully estimates such parameter in a scatter-plot smoothing context, rather than a latent process within a hierarchical model. From now onwards, only splines with a single smoothing parameter (*i.e. ts* and *bsord1*) will be considered.

4.6.2 Discussion on the results from the simulation study

The results of the simulation study under the three true incidence scenarios (increasing, flat and decreasing) and four incidence models (*rw1ord*, *ts*, *bsord1*, *GP*) are displayed in Section 4.6.3 (Figures 4.11 to 4.26). Specifically:

1. Figures 4.11, 4.13, 4.17 and 7.10 depict the posterior-mean (*i.e.* the estimates) of the incidence curve $\widehat{\mathcal{H}}_m$ for each simulated dataset.
2. Figures 4.12, 4.14, 4.16 and 4.18 show the posterior-mean (*i.e.* the estimates) of the diagnosis probabilities from state 1 ($\widehat{\mathcal{D}}_{1,m}$) for each simulated dataset.
3. Figures 4.19 to 4.22 depict $PMSE(\widehat{\mathcal{H}}_m)$ and $PMSE(\widehat{\mathcal{D}}_{1,m})$ distribution for the full time-scale considered and for the last twelve quarters (*i.e.* three years) only.
4. Figures 4.23 to 4.26 show the $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ and $Covg_{0.95}(\widehat{\mathcal{D}}_{1,m})$ distribution for the full time-scale considered and for the last twelve quarters (*i.e.* three years) only.

At a first glance, all incidence models reconstruct the three true incidence curves sufficiently well. A closer inspection reveals that the true incidence curve is, in all cases, accurately reconstructed except from the first two and last three years. It seems that diagnosis data in the first years of the epidemic are incorrectly attributed to new infections rather than initially prevalent individuals, resulting in over estimation of the true incidence curve. Eventually all initially prevalent individuals are diagnosed, and incidence estimates become more accurate, as all new diagnoses are then attributed to incidence. Incidence estimates in the earliest years must be interpreted with care, as results heavily rely on the expected number of undiagnosed individuals at the beginning of the epidemic (*i.e.* $\boldsymbol{\pi}$).

Crucially, the results highlight a major pitfall: estimates of the incidence curves in the most recent years are severely biased. In all true incidence scenarios, all infection models greatly over estimate incidence, for almost every simulated dataset. Results are particularly unsatisfactory when true incidence is decreasing as all incidence models, with the exception of Gaussian processes, fail to detect the decreasing trend in incidence in the most recent years. However, it is important to note that the credible intervals almost always include the true incidence values in recent years (see the 95%-coverage plots for the incidence curve in the last three years, Figure 4.25). Credible intervals in the last three years are as expected, very wide (for all true incidence scenarios and incidence models), as recent diagnosis data are not potentially informative of recent infections.

This over-estimation of the incidence curve is likely due to recent diagnoses (primarily diagnosed from state 1 of the model, with CD4 > 500) being incorrectly attributed to an increase in incidence rather than in diagnosis probabilities, resulting in underestimation of the former and overestimation of the latter.

Let us now focus on which non-parametric models are best for estimating the true incidence curves. The Bayesian implementation of different non-parametric incidence models involves different prior assumptions on the shape of the incidence curve. However these seem dominated by the likelihood, as the posterior means of the incidence curves are very similar irrespectively of the parameterisation chosen. Examination of the distribution of $PMSE(\widehat{\boldsymbol{\mathcal{H}}}_m)$ reveals all incidence models perform equally well (Figures 4.19 and 4.21).

Splines and Gaussian processes are smoother processes than random walks hence choosing between the two can be based on a bias-variance tradeoff. Close inspection of the estimates of incidence reveals that random walks accurately capture the peaks of the true incidence curves (in 2004 and 2007), whereas splines and Gaussian processes smooth these out (Figures 4.11, 4.13, 4.17 and 7.10). Thus random walks are associated with marginally less bias, evidenced by the somewhat lower $MPMSE(\widehat{\boldsymbol{\mathcal{H}}})$, but wider credible intervals (*i.e.* greater variance) than splines and Gaussian processes. The latter are associated with higher bias (*i.e.* higher $MPMSE(\widehat{\boldsymbol{\mathcal{H}}})$) but lower variance (*i.e.* smaller credible intervals). Figure 4.23 shows mean-coverage $MCovg_{0.95}(\widehat{\boldsymbol{\mathcal{H}}})$ for random walks of over 95%, whilst for splines and Gaussian processes approximately 85%, suggesting that random walks overestimate, whilst splines and Gaussian processes underestimate uncertainty.

Incidence models can be further distinguished based on the behaviour of the estimated incidence curves in the most recent years, where data are least informative, and consequently the impact of different prior assumptions on the shape of incidence estimates is considerable. Assuming a first order random walk model for the incidence curve implies that models with flat incidence between successive time intervals are a priori preferred; indeed Figure 4.11 shows that incidence in the last years is mostly flat, independently of the true incidence scenario. Recall that first order P-splines (*bs1ord*, Section 3.3.7) and thin plate splines (*ts* in this case, Section 3.3.6) assume flat curves and curves with second derivative equal to 0 (*i.e.* linear functions) to be the smoothest respectively. Hence incidence estimates from *bsord1* splines flatten out (as for random walks), whereas those from *ts* extrapolate linearly in most recent years (Figures 4.13 and 7.10). In absence of data Gaussian processes revert to the prior mean (*i.e.* zero); Figure 4.17 demonstrates this as for several datasets incidence estimates have an artificial hump, in the last years, subsequently tending towards zero. By reverting to the zero prior mean, Gaussian processes a priori favours curves with decreasing incidence and thus outperform other incidence models in the decreasing true incidence scenario.

Chapter 3 introduced a number of splines and we aim to establish whether some of them estimate the true incidence curves better. Splines with two smoothing parameters (*tp*, *tpknotsloc*, *bsord1*) are excluded from the comparison due to either non-convergence or divergent transitions (Section 4.6.1). The simulations results suggest that *bsord1* splines outperform *ts* splines. Despite the distribution of $PMSE(\widehat{\mathcal{H}}_m)$ being overall similar for both splines (Figure 4.19), in the last three years this favours *bsord1* (Figure 4.21). As previously discussed incidence estimates obtained with *ts* extrapolate linearly in most recent years; this may lead to unrealistically high incidence estimates. Finally *bsord1* splines have $MCovg_{0.95}(\widehat{\mathcal{H}})$ closer to 0.95 than *ts* splines (Figure 4.23).

An alternative way of choosing between incidence models for estimating incidence would be to compare model fit on the simulated data. In this study this is not practical, as all incidence models fit the simulated data equally well (see Appendix D.1.2).

This Section only considers diagnosis probabilities from state 1, further details on diagnosis probabilities can be found in Appendix D.1. Recall that recent diagnoses (*i.e.* from state 1) are a result of competing infection and diagnosis processes, which leads to identifiability issues in recent years. Estimates of diagnosis probabilities from states 2, 3 and 4 are typically accurate as they are associated with undiagnosed individuals with long-standing infections, which are relatively insensitive to shifts in recent incidence and diagnosis probabilities.

In summary, this simulation study demonstrates that the proposed back-calculation model estimates the true incidence and diagnosis probabilities reasonably well, despite some identifiability issues in the most recent years. Although there is no single best non-parametric model to estimate incidence, a number of recommendations can be made based on the study findings: Gaussian processes must be used with care as reversion to the prior mean may distort the incidence estimates in the latest years. This issue could be avoided by estimating the mean of the Gaussian process. However, this increases the number of latent parameters to be estimated, resulting in convergence problems. Among splines, first order B-splines appear to be the most suitable for modelling incidence. Among random walks, first order random walks estimate the true incidence curves adequately, while second order random walks suffer from convergence issues. The choice between *rw1ord* and *bsord1* is subjective, as it is based on a variance-bias tradeoff; random walks are associated with slightly less bias and wide credible intervals. Splines are associated with slightly more biased incidence estimates, but with reduced variance and hence narrower credible intervals. Finally note that none of the infection models considered avoids unidentifiability issues in the most recent years.

4.6.3 Plots of the results from the simulation study

Results from the first order random walk incidence model

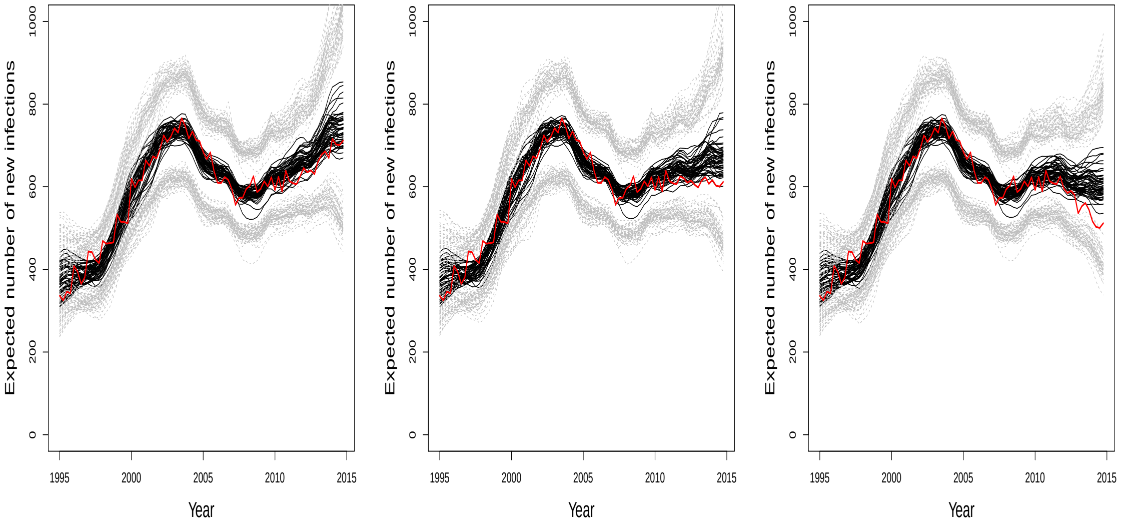


Fig. 4.11 Estimated incidence curves: the red lines depict the three true incidence curves (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence curve obtained using a *rw1ord* to model incidence. Grey lines denote the respective 95% credible intervals.

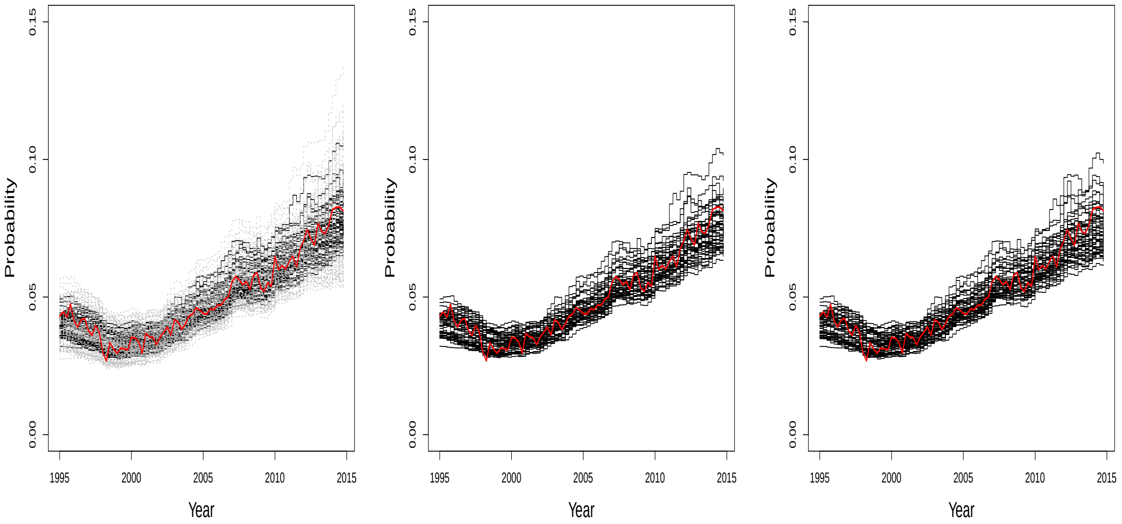


Fig. 4.12 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are only depicted on the left figure (in grey) to demonstrate they overlap with the estimates, rendering the plot hard to interpret.

Results from the thin plate spline with linear shrinkage incidence model

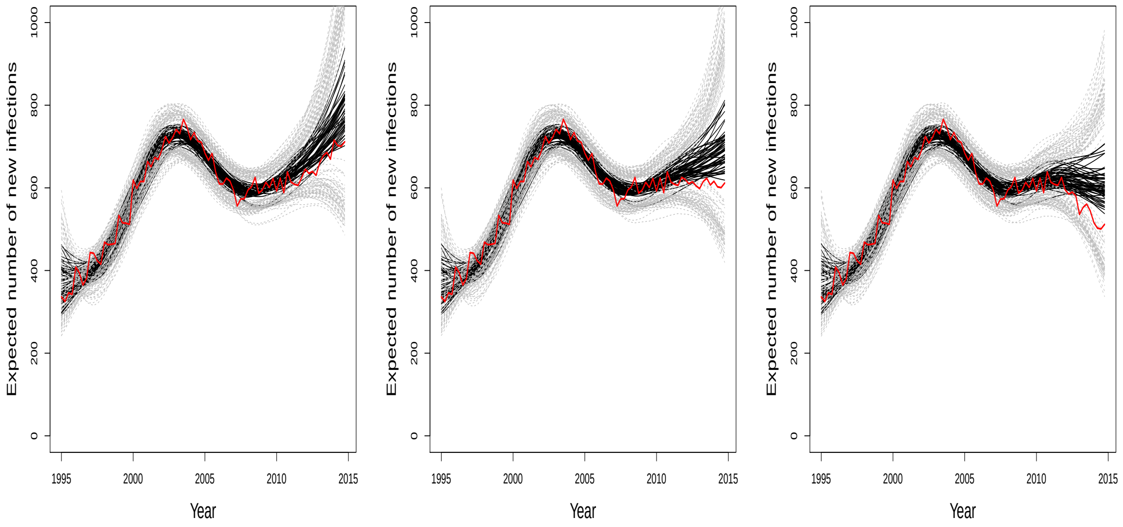


Fig. 4.13 Estimated incidence curves: the red lines depict the three true incidence curves (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence curve obtained using a *ts* spline to model incidence. Grey lines denote the respective 95% credible intervals.

.

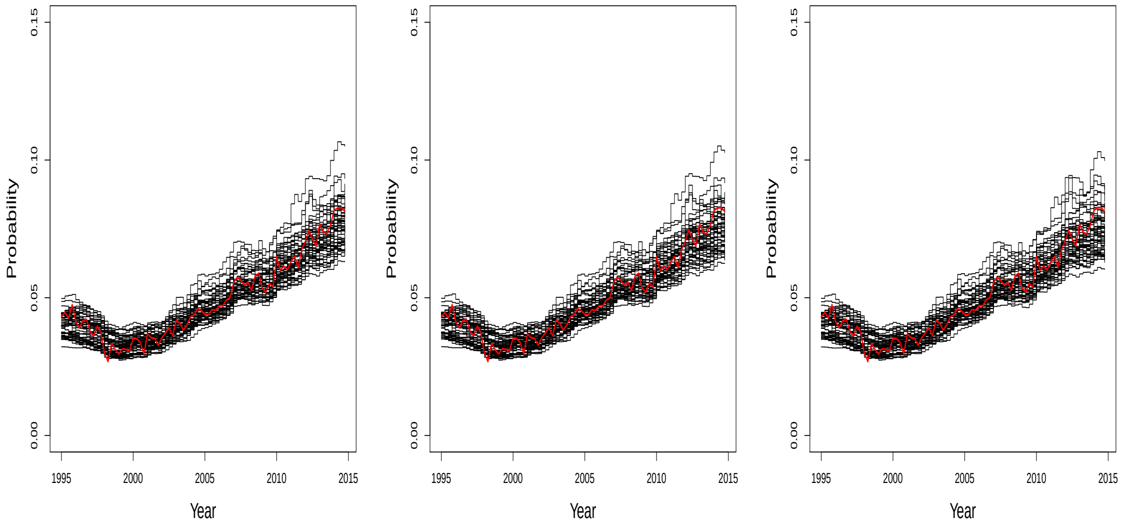


Fig. 4.14 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted, as they overlap with the estimates.

Results from the first order P-spline incidence model

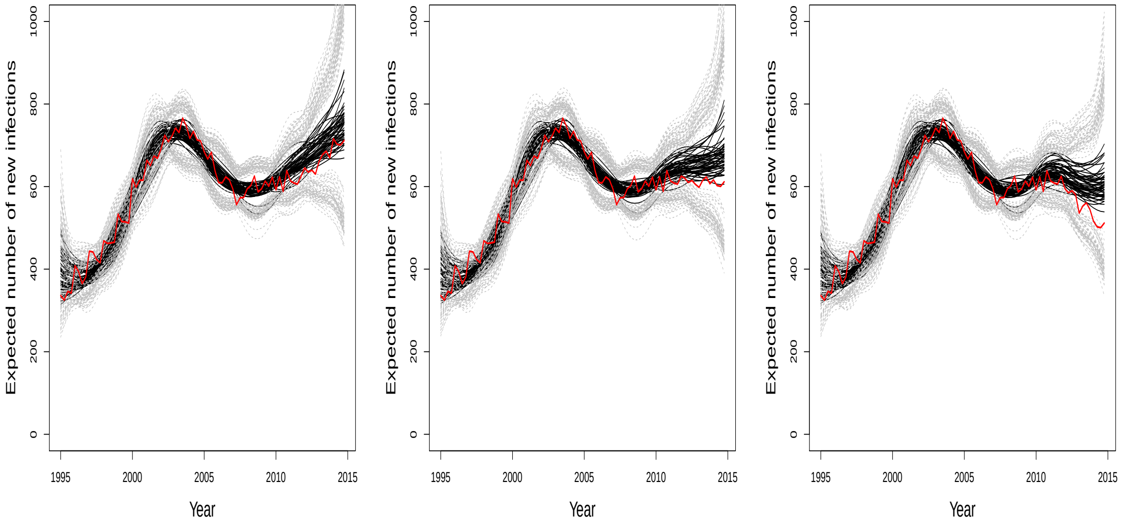


Fig. 4.15 Estimated incidence curves: the red lines depict the three true incidence curves (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence curve obtained using a *bsord1* spline to model incidence. Grey lines denote the respective 95% credible intervals.

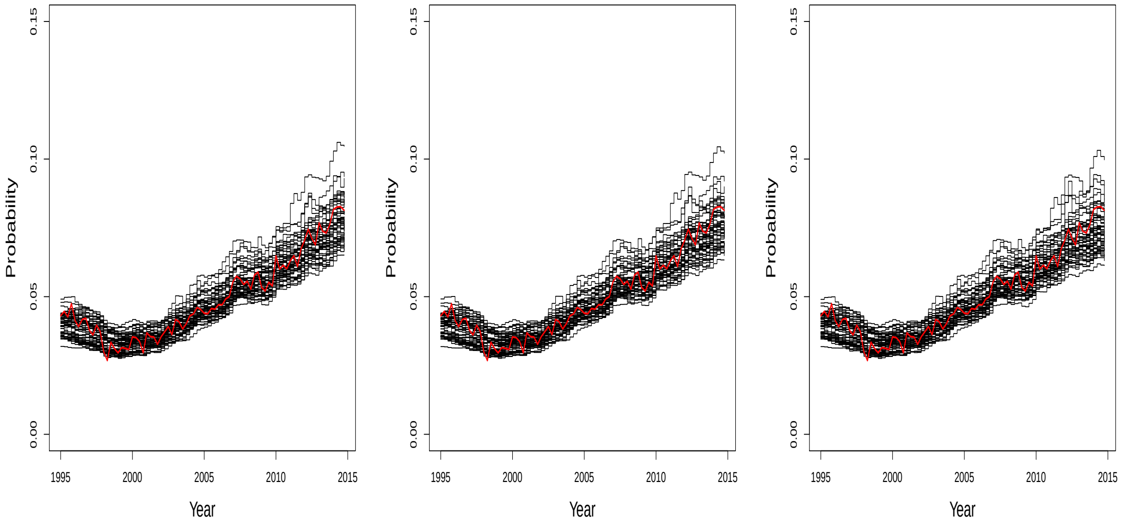


Fig. 4.16 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted, as they overlap with the estimates.

Results from the Gaussian process incidence model

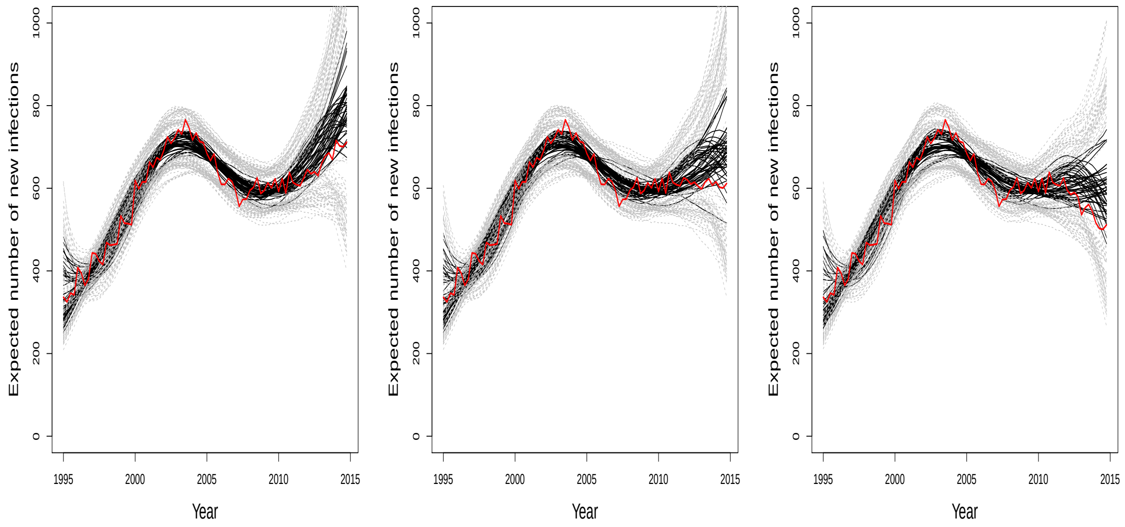


Fig. 4.17 Estimated incidence curves: the red lines depict the three true incidence curves (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence curve obtained using a *GP* to model incidence. Grey lines denote the respective 95% credible intervals.

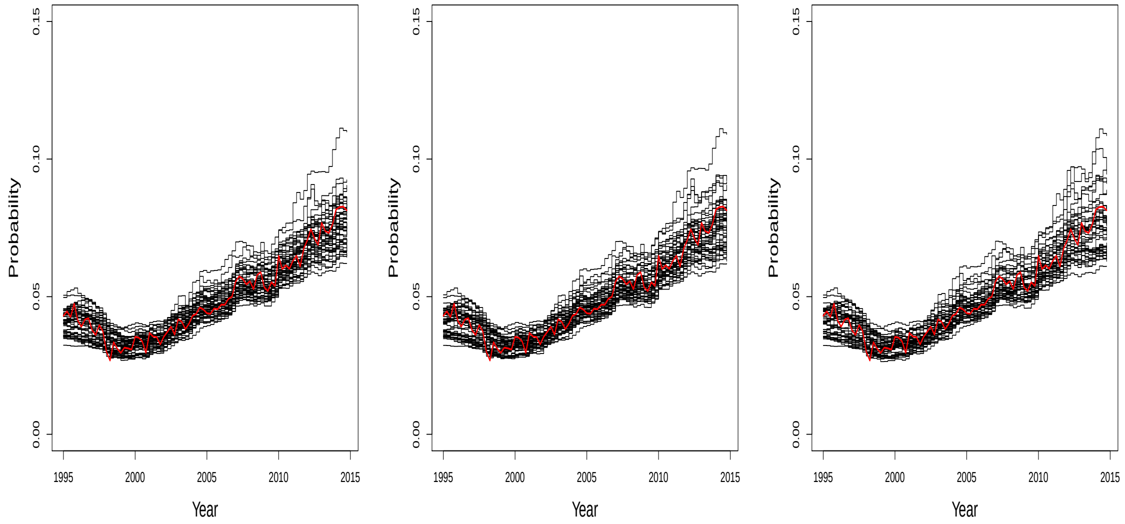


Fig. 4.18 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted, as they overlap with the estimates.

Predicted Mean Squared Error

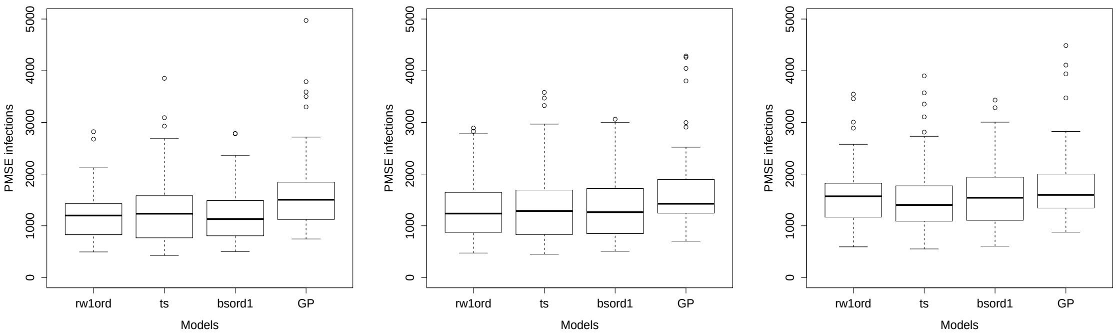


Fig. 4.19 Distribution of PMSE for incidence curve, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

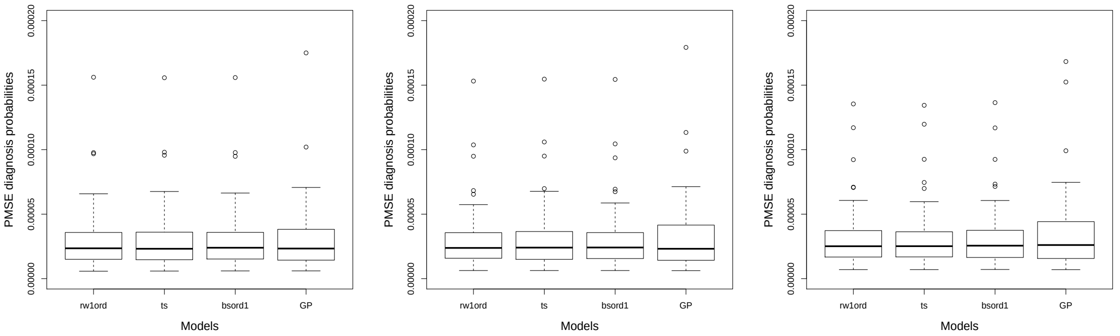


Fig. 4.20 Distribution of PMSE for diagnosis probabilities, under three different true incidence scenarios: increasing (left), flat (center), decreasing (right).

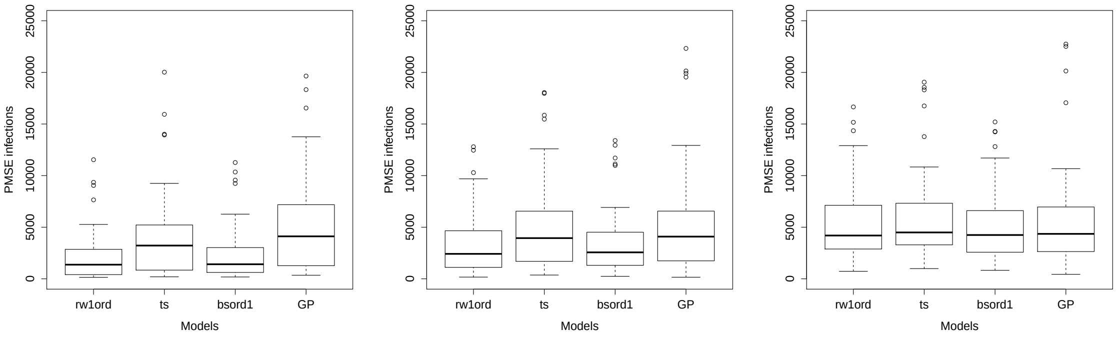


Fig. 4.21 Distribution of PMSE in the last 3 years for incidence curve, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

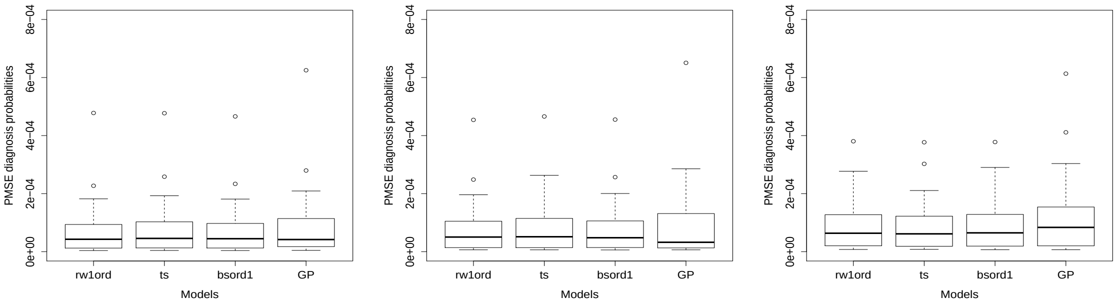


Fig. 4.22 Distribution of PMSE in the last 3 years for diagnosis probabilities, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

Coverage

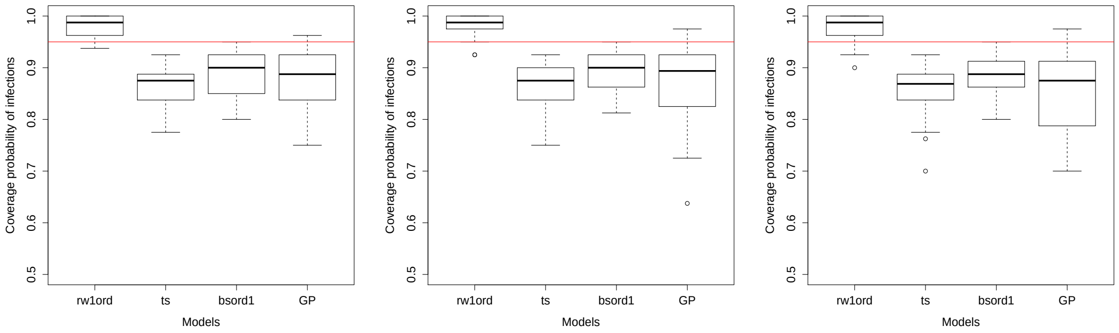


Fig. 4.23 Distribution of 95%-coverage for the incidence curve, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

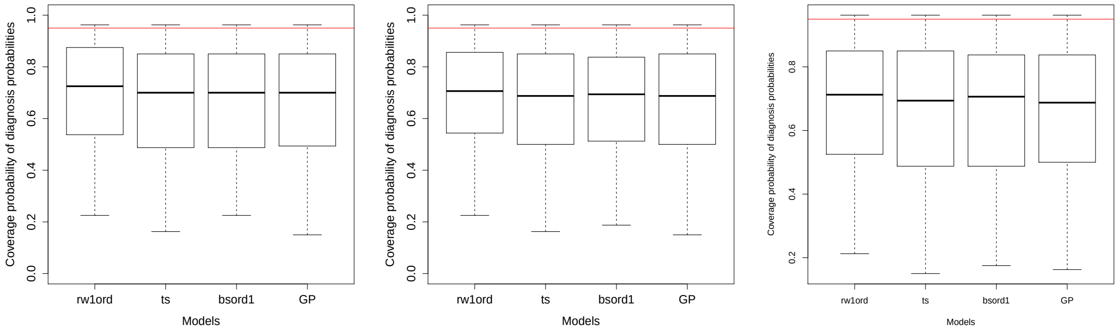


Fig. 4.24 Distribution of 95%-coverage for diagnosis probabilities from state 1, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

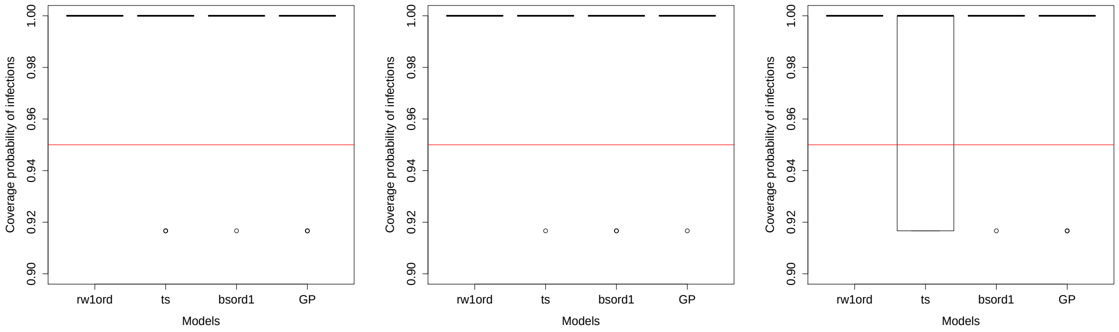


Fig. 4.25 Distribution of 95%-coverage for the incidence curve in the last three years, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

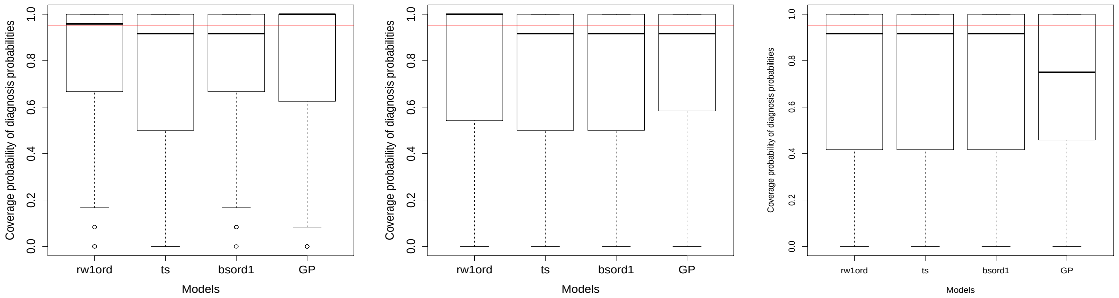


Fig. 4.26 Distribution of 95%-coverage for the last 3 years for diagnoses from state 1, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

4.6.4 Prior sensitivity analysis

The discussion thus far mostly focused on the incidence estimates. Let us take a closer look to the estimated diagnosis probabilities; these are modelled, with a first order random walk for all incidence models and scenarios. Recall that true diagnosis probabilities are underestimated in the most recent years. Figure 4.24 shows that mean-coverage $MCovg_{0.95}(\widehat{\boldsymbol{\mathcal{D}}})$ is approximately 70% suggesting that credible intervals for the diagnosis probabilities are unduly narrow.

This issue is investigated by examining the posterior mean of the variances of the logistic random walk for the diagnosis probabilities from state 1, denoted $\widehat{\sigma}^2_{D,m}$ for the m^{th} dataset. Recall that the true logit-diagnosis probabilities are not generated from a first order random walk (see Section 4.2), hence a "true" variance parameter does not exist. An "approximated true" variance parameter may instead be constructed by fitting a first order random walk to the true logit-diagnosis probabilities from state 1, and taking the posterior mean of the variance to be the "approximated true" variance; this is denoted $\sigma^2_{D,T}$ and is equal to 0.0068.

Figure 4.27 plots the distribution of $\widehat{\sigma}^2_{D,m}$, over the datasets: underestimation of $\sigma^2_{D,T}$ is striking and is pertinent to all scenarios examined. This is likely to cause the estimated diagnosis probabilities not to increase as rapidly as the true ones in most recent years.

Under-estimation of $\sigma^2_{D,T}$ may be due to a poor choice of prior. The sensitivity to prior specifications is investigated under the true increasing incidence scenario, using the *rw1ord* incidence model. We choose six different priors for both the log-infection and logit-diagnoses

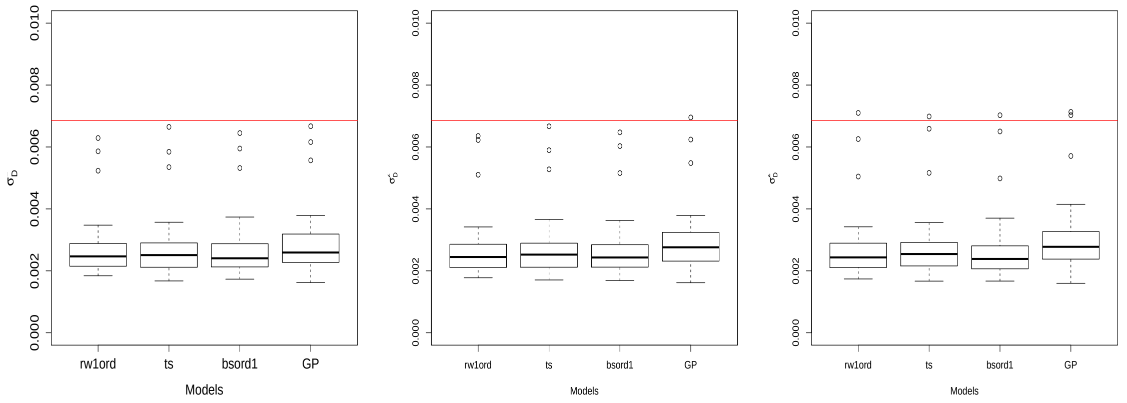


Fig. 4.27 Distribution of the posterior mean variance estimates $\widehat{\sigma}^2_{D,m}$ of diagnosis probabilities for all incidence models, under three true incidence scenarios: increasing (left), flat (center), decreasing (right). The "true approximated" variance parameter $\sigma^2_{D,T}$ is given by the red line.

random walks' variances (denoted σ_I^2 and σ_D^2) and assess their impact on posterior estimates. Let $\Gamma(a,b)$ denote a Gamma distribution with shape parameter a and scale parameter b and $t_+(d,s)$ denote a half-t distribution with d degrees of freedom, and scale parameter s.

Prior Scenarios:

1. $\sigma_D^2 \sim \Gamma(1,32)$ and $\sigma_I^2 \sim \Gamma(1,32)$. This is the reference case (as used in the simulation study); the 90% standard deviation prior range is [0.0008, 0.115].

2. $\sigma_D^2 \sim \Gamma(1,64)$ and $\sigma_I^2 \sim \Gamma(1,64)$. These priors are tighter, and more informative, than the reference priors and shift the prior mass towards zero; the 90% standard deviation prior range is [0.0004, 0.057].

3. $\sigma_D^2 \sim \Gamma(1,8)$ and $\sigma_I^2 \sim \Gamma(1,8)$. These priors are wider, less informative, than the reference priors and shift the prior mass away from zero; the 90% standard deviation prior range is [0.003, 0.46].

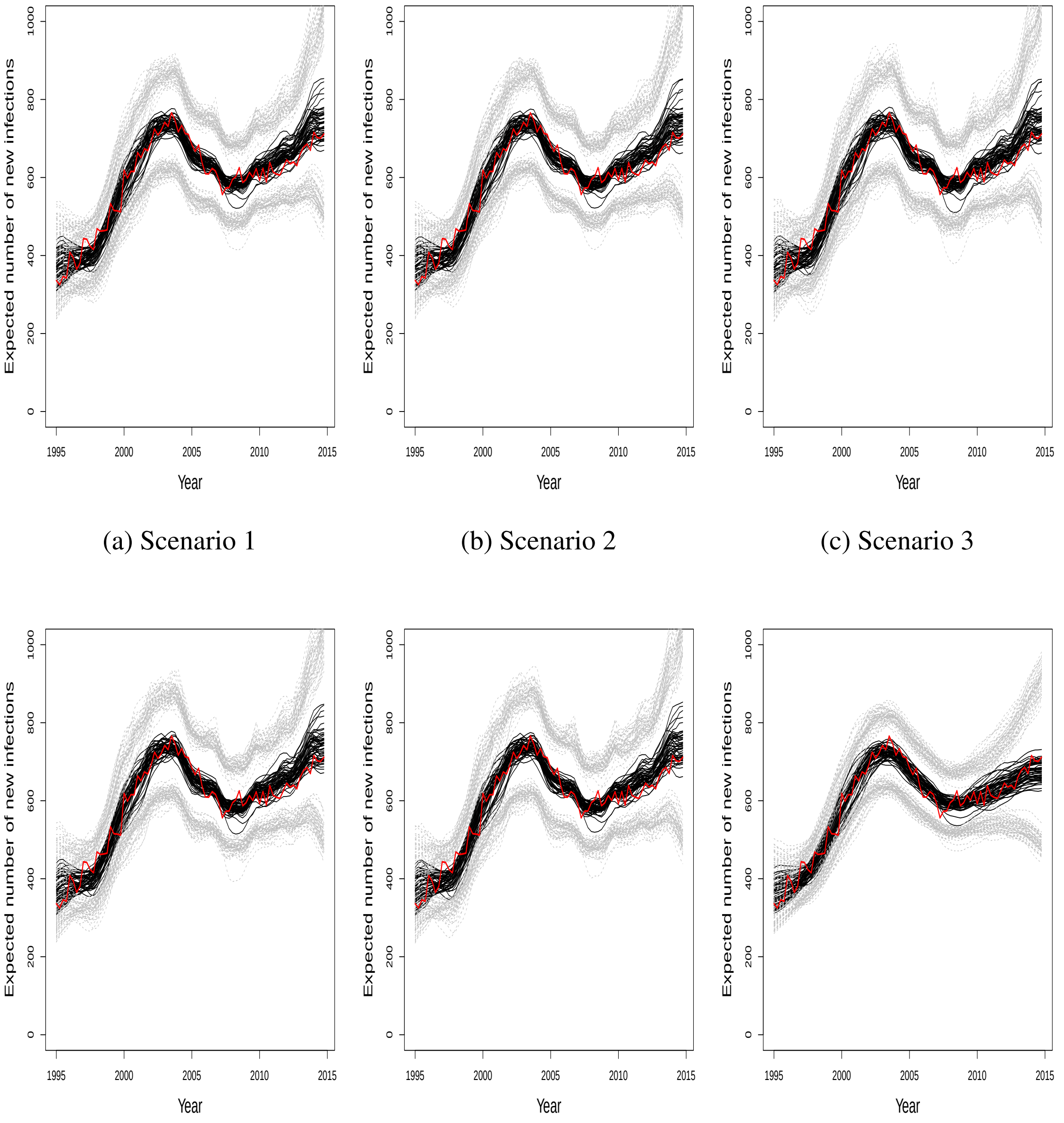
4. $\sigma_D^2 \sim t_+(4,0.1)$ and $\sigma_I^2 \sim t_+(4,0.1)$. The half-t distribution is defined on $[0,\infty)$ and is monotonically decreasing from zero; the 90% prior range for the standard deviation is [0, 0.44]. The t_+ distribution has a heavy right tail and is often used as an uninformative prior, favouring smaller variances a priori (Gelman et al., 2006).

5. $\sigma_D^2 \sim t_+(4,0.5)$ and $\sigma_I^2 \sim t_+(4,0.5)$. These priors are less informative than the priors in Scenario 4, as the 90% standard deviation prior range is [0, 1].

6. $\sigma_D^2 \sim N(0.0068, 0.00001)$ and $\sigma_I^2 \sim N(0.0028, 0.00001)$. This is the "approximately true" scenario; extremely informative priors are specified, concentrating the prior mass of the random walk's variance around "true approximated" variances $\sigma_{D,T}^2$ and $\sigma_{I,T}^2$. Similarly to $\sigma_{D,T}^2$, is the "true approximated" variance for infections $\sigma_{I,T}^2$ (0.0028) is obtained by fitting a random walk model to log-expected infections.

With the exception of scenario 6, the same priors are assigned to both the infection and diagnosis probabilities random walks' variances. Different priors for the variances imply a priori different levels of smoothness for the two processes; since there is no reasons to believe that one process is smoother than the other, such cases are not further considered.

Results from the different prior scenarios are displayed in Figures 4.28 and 4.29. Scenarios 1 to 5 yield very similar incidence and diagnoses estimates. As before, the true incidence curve is overestimated whereas the diagnosis probabilities are underestimated. Unsurprisingly the "true approximate" scenario (6) is the best, yielding unbiased estimates of incidence and diagnosis probabilities also in most recent years.



(a) Scenario 1

(b) Scenario 2

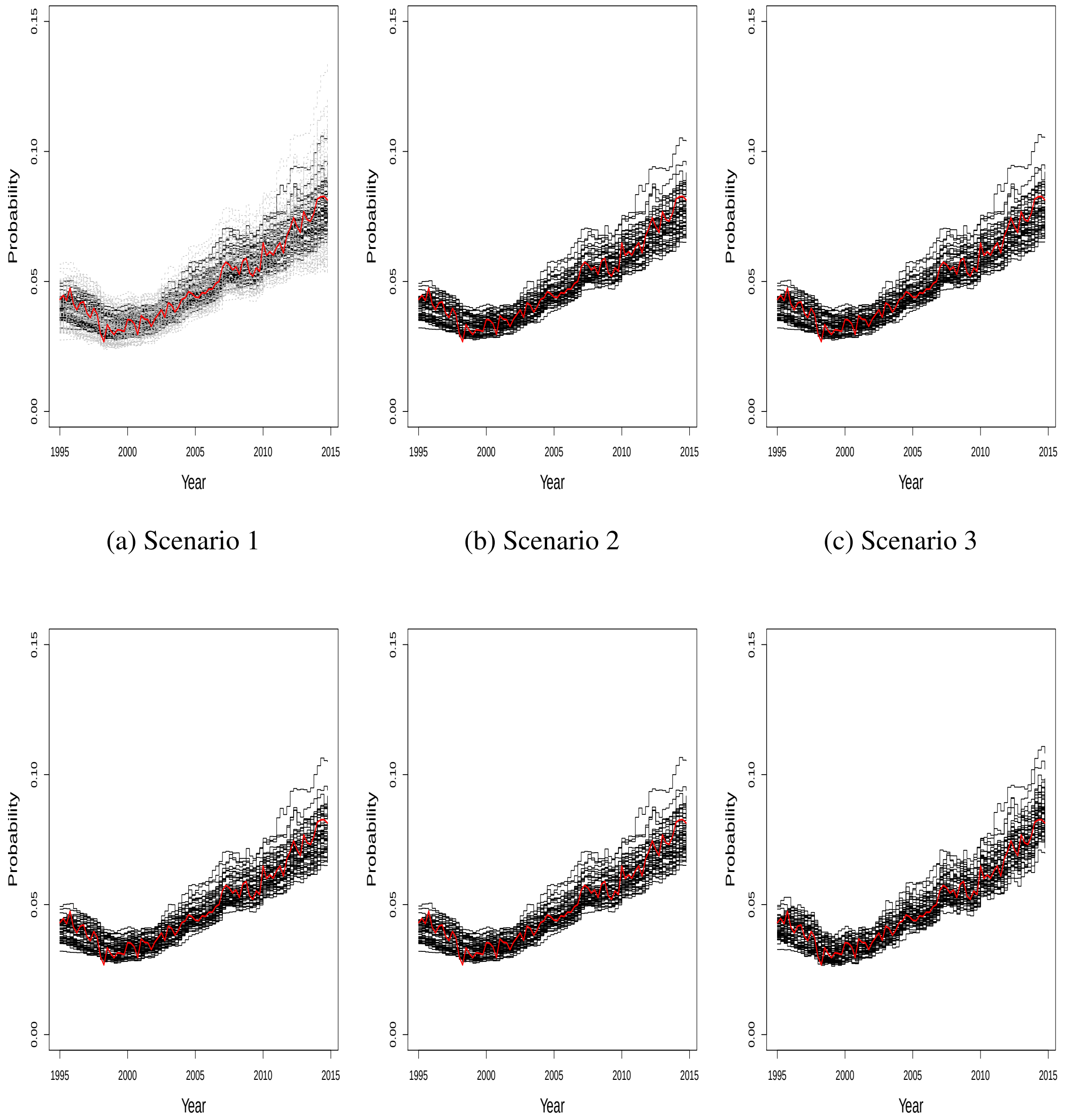
(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Fig. 4.28 Estimated incidence curves: the true increasing incidence curves are plotted in red. The posterior means for each dataset are plotted in black and the associated 95% credible intervals in grey.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Fig. 4.29 Estimated diagnosis probabilities from State1: the true diagnosis probabilities are plotted in red. The posterior means for each dataset are plotted in black. The associated 95% credible intervals are only depicted in Scenario 1, as they overlap with posterior means.

To sum up, in this Section the back-calculation model is fit with five different weakly informative priors, all yielding very similar estimates of incidence and diagnosis probabilities. Estimates are robust to prior specifications, which is reassuring, however they are biased in recent years. Bias reduction can be achieved by specifying strong priors with mass concentrated on the "approximated true" values of the random walks' variances. Clearly in a real life context, true parameters are unknown and this approach can not be employed; thus unidentifiability leads to poor estimates in recent years, irrespectively of prior specifications. Care must be taken when interpreting back-calculation results in the most recent years, taking large uncertainty into account.

4.7 Summary

This Chapter discusses a Bayesian simulation study for the age-independent multi-state back-calculation (Chapter 2), involving a number of non-parametric models for the latent incidence curve (Section 4.3). The simulation study intends to answer two questions:

1. Is back-calculation feasible? What are its strength and limitations?
2. Modeling the latent incidence curve is challenging. Are some non-parametric models better suited than others for this purpose?

The answer to the first question is yes. The true incidence curve and diagnosis probabilities were successfully reconstructed, except from the first and last three years of the epidemic. In the most recent years, incidence and diagnosis probabilities are consistently over and underestimated respectively. As discussed in Section 4.6.4 biased estimates do not appear to be a consequence of poor prior specifications, but instead of unidentifiability issues (Section 4.6.2). Unfortunately, such issues can not be addressed with the available data.

With regards to the second question, some incidence models (*rw1ord* and *ts* splines) are indeed better suited than other to estimate incidence (Section 4.6.2). It is crucial to highlight that none of the incidence models can provide valid estimates of the true incidence curve in the latest years, where the impact of prior assumptions is more pronounced.

The following Chapter introduces an extension of the back-calculation model to age specific settings. As discussed in Section 1.5, the incorporation of age allows to better characterise the HIV epidemic.

Chapter 5

Age dependent back-calculation

5.1 Introduction

The back-calculation model described in Chapter 2 has proven to be, over the years, an important public health tool for monitoring the MSM HIV epidemic in England and Wales (Birrell et al., 2013; Kirwan et al., 2016).

However for a better monitoring of the epidemic, age-specific estimates of HIV incidence are of crucial importance. Thus back-calculation, introduced in Chapter 2, is here extended to age-specific settings. This entails two main challenges: modelling a two dimensional (age-time) infection rate and developing a computationally efficient implementation.

This Chapter is structured as follows: the motivating dataset employed is first described (Section 5.2), followed by a description of the age-dependent back-calculation model (Section 5.3). Finally Section 5.4 discusses possible extensions.

5.2 Motivating surveillance dataset

The motivating surveillance dataset remains the same routinely collected aggregated surveillance data for the MSM-HIV epidemic in England and Wales (see Section 2.2), further stratified by age at diagnosis.

The age-independent back-calculation notation is extended as follows. Let $(t_0, t_T]$ be the time-period spanning the HIV epidemic, which is split into T disjoint, consecutive intervals

$(t_{i-1}, t_i]$, $i = \{1, \ldots, T\}$. Similarly the age-range $(a_0, a_A]$ is split into A disjoint, consecutive intervals $(a_{j-1}, a_j]$, $j = \{1, \ldots, A\}$. The data available include:

- $y^H_{i,j}$, the aggregated number of new HIV diagnoses in the time interval $(t_{i-1}, t_i]$ and age interval $(a_{j-1}, a_j]$.
- $y^A_{i,j}$, the aggregated number of new AIDS diagnoses in the time interval $(t_{i-1}, t_i]$ and age interval $(a_{j-1}, a_j]$.
- A subset of $y^H_{i,j}$, of size $n_{i,j}$, with associated CD4-count, taken within three months of diagnosis. These subsets are grouped into K categories, defined by CD4 thresholds. $\mathbf{y}^{H_C}_{i,j}$ = $(\mathrm{y}^{H_C}_{i,j,1}, \mathrm{y}^{H_C}_{i,j,2}, \ldots, \mathrm{y}^{H_C}_{i,j,K})^T$ is a $K \times 1$ vector containing the number of new CD4-linked diagnoses in the intervals $(t_{i-1}, t_i]$ and $(a_{j-1}, a_j]$, with respective CD4-counts being categorized according to intervals $[c_1, \infty)$, $[c_2, c_1)$, $\ldots$ and $[0, c_{K-1})$, where $c_1 > c_2 > \cdots > c_{K-1}$.

$\mathbf{y}^H = \{y^H_{11}, \ldots, y^H_{1A}, \ldots, y^H_{T1}, \ldots y^H_{TA}\}$ and $\mathbf{y}^A = \{y^A_{11}, \ldots, y^A_{1A}, \ldots, y^A_{T1}, \ldots y^A_{TA}\}$ are $TA \times 1$ vectors denoting the number of new HIV and AIDS diagnosis over time and age respectively. $\mathbf{y}^{H_C} = \{\mathbf{y}^{H_C}_{1,1}, \ldots, \mathbf{y}^{H_C}_{1,A}, \ldots, \mathbf{y}^{H_C}_{T,1}, \ldots, \mathbf{y}^{H_C}_{T,A}\}$ denotes the collection of CD4 diagnoses over time and age.

5.3 Model

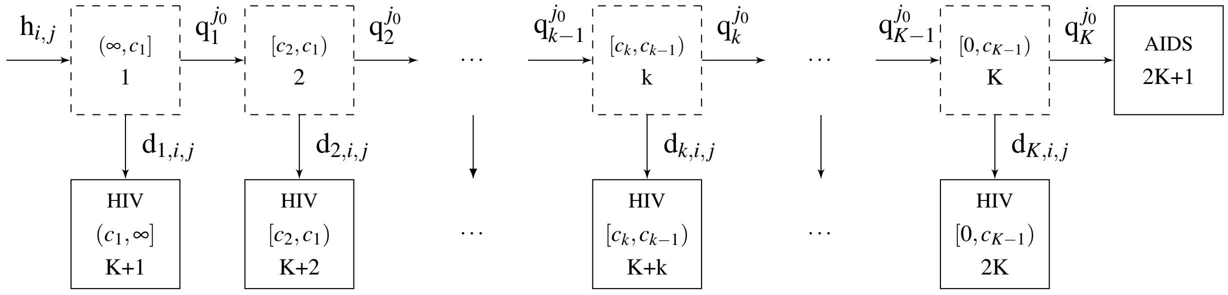


Fig. 5.1 Age-dependent back-calculation multi-state model, for a general number of undiagnosed states K. Dashed states $\{1, \ldots, K\}$ denote undiagnosed states. Solid states $\{K+1, \ldots, 2K+1\}$ denote diagnosed states. $d_{k,i,j}$ denotes the probability of diagnosis from the undiagnosed state k in the i$^{\text{th}}$ time interval and in the j$^{\text{th}}$ age interval. $q_k^{j_0}$ denotes the probability of progression between undiagnosed states k and $k+1$, given that the infection occurred in the j_0^{th} age interval.

The non-homogeneous population-level CD4-count multi-state model in Chapter 2 (Figure 2.1) is extended to age-specific settings (Figure 5.1) by characterising the processes of infection, progression and diagnosis in terms of both time, current age and age at infection.

New infections are now occurring according to a two dimensional non-homogeneous Poisson Process with rate $\lambda(u,v)$. Then the expected number of new infections in the time interval $(t_{i_0-1}, t_{i_0}]$ and in the age interval $(a_{j_0-1}, a_{j_0}]$ is $h_{i_0,j_0} = \int_{t_{i_0-1}}^{t_{i_0}} \int_{a_{j_0-1}}^{a_{j_0}} \lambda(u,v)\, dudv$.

Progression probabilities now depend on the age at infection, as the progression towards AIDS is known to be substantially faster for individuals infected at an older age (CASCADE Collaboration, 2000). Moreover diagnosis probabilities are also dependent on current age, as the propensity to test for HIV may vary with age. Hence, diagnosis and progression probabilities are denoted $\boldsymbol{d}_{i,j} = (d_{1,i,j}, \ldots, d_{K,i,j})^T$ and $\boldsymbol{q}^{j_0} = (q_1^{j_0}, \ldots, q_K^{j_0})^T$ respectively, to stress the dependency on the j_0^{th} age interval at infection.

Analogously to Chapter 2, the aim is to estimate the expected number of new time and age specific infections $\mathcal{H} = \{h_{1,1}, \ldots, h_{T,A}\}$, to which we refer as the *incidence surface* (or simply *incidence*), and the diagnosis probabilities $\mathcal{D} = \{\boldsymbol{d}_{1,1}, \ldots, \boldsymbol{d}_{T,A}\}$, characterised by parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ respectively. From here onwards $h_{i,j}(\boldsymbol{\theta})$ and $\boldsymbol{d}_{i,j}(\boldsymbol{\delta})$ will be written as $h_{i,j}$ and $\boldsymbol{d}_{i,j}$ respectively, for notational convenience. $\mathcal{Q} = \{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^A\}$ denotes the collection of age-at-infection dependent progression probabilities.

5.3.1 Transition matrices

To start with, the time $(t_{i-1}, t_i]$ and age intervals $(a_{j-1}, a_j]$ are assumed to have equal length, *i.e.* $|t_i - t_{i-1}| = |a_j - a_{j-1}|$ $(i = \{1, \ldots, T\},\ j = \{1, \ldots, A\})$. Similarly to Chapter 2, the intervals are assumed to be small enough so that at most one transition event (diagnosis or progression) can occur within any interval and newly infected individuals are not allowed to progress nor to be diagnosed in the time and age intervals of infection. Recall that infections are assumed to occur at the beginning of intervals $(t_{i-1}, t_i]$ and $(a_{j-1}, a_j]$, whereas diagnosis and progression events are assumed to occur at the end of the interval (the former before the latter).

$\boldsymbol{Q}_{i,j}^{j_0}(\boldsymbol{\delta})$ and $\boldsymbol{D}_{i,j}^{j_0}(\boldsymbol{\delta})$ are the progression and diagnosis transition matrices, which are functions of $\boldsymbol{q}^{j_0}$ and $\boldsymbol{d}_{i,j}$. As before, the dependency of the transition matrices on the diagnosis parameters is suppressed for notational convenience.

$\boldsymbol{Q}_{i,j}^{j_0}$ is a $K \times K$ matrix, whose $(k,l)^{\text{th}}$ entry is defined as:

$$\left(\boldsymbol{Q}_{i,j}^{j_0}\right)_{k,l} = \begin{cases} (1-d_{k,i,j})(1-q_k^{j_0}) & \text{if } l = k \\ (1-d_{k,i,j})q_k^{j_0} & \text{if } l = k+1 \text{ and } k < K \\ 0 & \text{elsewhere} \end{cases} \tag{5.3.1}$$

$\boldsymbol{D}_{i,j}^{j_0}$ is a $K \times K+1$ matrix, whose $(k,l)^{\text{th}}$ entry is defined as:

$$\left(\boldsymbol{D}_{i,j}^{j_0}\right)_{k,l} = \begin{cases} d_{k,i,j} & \text{if } l = k \\ (1-d_{k,i,j})q_k^{j_0} & \text{if } l = K+1 \text{ and } k = K \\ 0 & \text{elsewhere} \end{cases} \tag{5.3.2}$$

5.3.2 Model dynamics

The movement of the individuals in the model in the i^{th} time interval is not exclusively determined by the undiagnosed state where individuals are (as in age independent back-calculation), but also by current age (as it characterises the diagnosis process) and age at infection (describing the progression process).

To describe the dynamics of the model, the evolution of new infections in the i_0^{th} time interval and j_0^{th} age interval throughout the states of the model is followed over time intervals $(t_{i-1}, t_i]$ and $(a_{j-1}, a_j]$ $(i = \{i_0+1, \ldots, T\},\ j = \{j_0+1, \ldots, j_0+T-i_0\})$. As time and age intervals have equal length, when a time interval elapses so does an age interval; moreover the relationship $i_0 = i - j + j_0$, linking the indices of the time and age intervals of infection to the indices of the current time and age intervals, holds.

Let $\boldsymbol{e}_{i,j}^{j_0}(\boldsymbol{\theta}, \boldsymbol{\delta})$ be a $K \times 1$ vector, denoting the expected number of individuals in undiagnosed states $\{1, \ldots, K\}$ in the i^{th} time and j^{th} age intervals, from the cohort infected in the j_0^{th} age interval (and thus in the i_0^{th} time interval). Similarly, the $(K+1) \times 1$ vector $\boldsymbol{\mu}_{i,j}^{j_0}(\boldsymbol{\theta}, \boldsymbol{\delta})$ denotes the expected number of new diagnoses in states $\{K+1, \ldots, 2K+1\}$ in the i^{th} time and j^{th} age intervals, from the infected cohort in the j_0^{th} age interval. For notational convenience, the dependency on the parameters is removed, so $\boldsymbol{e}_{i,j}^{j_0} = (e_{i,j,1}^{j_0}, \ldots, e_{i,j,K}^{j_0})^T$ and $\boldsymbol{\mu}_{i,j}^{j_0} = (\mu_{i,j,1}^{j_0}, \ldots, \mu_{i,j,K+1}^{j_0})^T$ are defined as follows, for $i = \{i_0+1, \ldots, T\}$ and $j =$

$\{j_0+1,\ldots,j_0+T-i_0\}$:

$$\boldsymbol{e}^{j_0}_{i,j} = \left(\boldsymbol{Q}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j-1} \tag{5.3.3}$$

$$\boldsymbol{\mu}^{j_0}_{i,j} = \left(\boldsymbol{D}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j-1} \tag{5.3.4}$$

where $\boldsymbol{e}^{j_0}_{i_0,j_0} = (h_{i_0,j_0},0,\ldots,0)^T$,$i_0 = \{1,\ldots,T-1\}$, $j_0 = \{1,\ldots,A-1\}$.

Then, the expected number of individuals in undiagnosed states $k = \{1,\ldots,K\}$ and the expected number of new diagnoses in states $k = \{K+1,\ldots,2K+1\}$ in the i^{th} time and j^{th} age intervals (denoted $\boldsymbol{e}_{i,j}$ and $\boldsymbol{\mu}_{i,j}$ respectively) is obtained by summing the expected number of individuals in the states of the model in the i^{th} time and j^{th} age intervals, infected at different age-intervals j_0:

$$\boldsymbol{e}_{i,j} = \sum_{j_0=max(1,j-i+1)}^{j} \boldsymbol{e}^{j_0}_{i,j} \tag{5.3.5}$$

$$\boldsymbol{\mu}_{i,j} = \sum_{j_0=max(1,j-i+1)}^{j} \boldsymbol{\mu}^{j_0}_{i,j} \tag{5.3.6}$$

Note that the time and age intervals at diagnosis provide a lower bound for the age interval of infection (*e.g.* if someone is diagnosed in the 5^{th} time interval, the age interval of infection can not be smaller than the age interval at diagnosis minus five).

5.3.3 Likelihood

The likelihood is formulated similarly to the age-independent likelihood, on the basis of the same two underlying assumptions (Section 2.3.4). Hence the likelihood of HIV and AIDS diagnoses (described in Section 5.2) is given by the product of independent Poisson random variables (denoted as $Y^H_{i,j}$ and $Y^A_{i,j}$ respectively) with means $\mu^H_{i,j}(\boldsymbol{\theta},\boldsymbol{\delta}) = \mu^H_{i,j} = \mu_{i,j,1}+\cdots+\mu_{i,j,K}$ $\mu^A_{i,j}(\boldsymbol{\theta},\boldsymbol{\delta}) = \mu^A_{i,j} = \mu_{i,j,K+1}$. For $i = \{1,\ldots,T\}$ and $j = \{1,\ldots,A\}$:

$$Y^A_{i,j} \sim Po\left(\mu^A_{i,j}\right) \tag{5.3.7}$$

$$Y^H_{i,j} \sim Po\left(\mu^H_{i,j}\right) \tag{5.3.8}$$

The subsample of CD4-linked diagnoses is instead distributed according to a Multinomial random variable:

$$\boldsymbol{Y}^{H_C}_{i,j} \sim Multinomial(n_{i,j},\boldsymbol{p}_{i,j}) \tag{5.3.9}$$

where $\boldsymbol{p}_{i,j} = (p_{i,j,1}, \ldots, p_{i,j,K})$ and $p_{i,j,k} = \frac{\mu_{i,j,k}}{\mu^H_{i,j}}$, $k = \{1, \ldots, K\}$.

The likelihood, expressed in terms of $\boldsymbol{\mu}$ (and thus of $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$) is proportional to:

$$\begin{aligned} L(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) &= L(\mathbf{y}^{H_C} \mid \mathbf{y}^H, \mathbf{y}^A, \boldsymbol{\theta}, \boldsymbol{\delta})\, L(\mathbf{y}^H, \mathbf{y}^A \mid \boldsymbol{\theta}, \boldsymbol{\delta}) \qquad (5.3.10) \\ &\propto \prod_{i=1}^{T} \prod_{j=1}^{A} \left(\prod_{k=1}^{K} (p_{i,j,k})^{y^{H_C}_{i,j,k}} \right) e^{-\mu^A_{i,j}} \left(\mu^A_{i,j}\right)^{y^A_{i,j}} e^{-\mu^H_{i,j}} \left(\mu^H_{i,j}\right)^{y^H_{i,j}} \end{aligned}$$

5.4 Model customization

The previous Section introduced a base-case age-specific back-calculation model to be used as building block. Analogously to age-independent back-calculation (Section 2.4), this model can be extended in a number of ways to account for under-reporting (Section 5.4.1); and to handle situations where surveillance data are not available from the beginning of the epidemic (Section 5.4.2); data are only available on a coarse scale (Section 5.4.3); or data are available at different time and age scales (Section 5.4.4).

Furthermore age-dependent back-calculation is highly computationally intensive: $O(TA^2)$ operations are required to evaluate the recursive Equations 5.3.3 and 5.3.4. Moreover a large number (TA^2) of $\boldsymbol{Q}^{j_0}_{i,j}$ and $\boldsymbol{D}^{j_0}_{i,j}$ matrices need to be stored. Hence considering back-calculation on a reduce and/or coarse scale may substantially alleviate the computational burden of the model and may achieve implementation within an acceptable computational time.

5.4.1 Under-reporting

Under-reporting can be incorporated within back-calculation, as in Section 2.4.1. However, within an age-dependent framework, it is possible to consider age-dependent under-reporting parameters if the proportion of reported diagnoses is believed to vary with age.

Let parameters $\upsilon^H_{i,j}$ and $\upsilon^A_{i,j}$ denote the proportion of new HIV and AIDS diagnoses in the i^{th} time and j^{th} age intervals, that are actually reported by the end of the intervals. The expected number of HIV and AIDS diagnoses (defined in Section 5.3.3) can then be modified to account for age and time specific under-reporting as follows:

$$\mu^{H'}_{i,j} = \upsilon^H_{i,j} \mu^H_{i,j} \qquad (5.4.1)$$

$$\mu^{A'}_{i,j} = \upsilon^A_{i,j} \mu^A_{i,j} \qquad (5.4.2)$$

The likelihood (Equation 5.3.10) can be appropriately modified, by replacing $\mu^{H}_{i,j}$ and $\mu^{A}_{i,j}$ by $\mu^{H'}_{i,j}$ and $\mu^{A'}_{i,j}$, to depend on the under-reporting parameters.

5.4.2 Back-calculation over a reduced time period

As described in Section 2.4.2, back-calculation can be run on a subset $(t_b, t_T]$ of the full epidemic period $(t_0, t_T]$, $t_b > t_0$ by characterising the expected number of individuals initially undiagnosed in the model at time t_b. For age-dependent back-calculation, this needs to be stratified not only by undiagnosed state $\{1, \ldots, K\}$ (as for age-independent back-calculation) but also by age (interval) at infection and current age at t_b (as these characterise the diagnosis and progression process).

In practice, characterising the number of individuals undiagnosed at t_b by age at infection is hardly feasible, as it requires knowledge of the infection time and age (pre t_b) of these individuals. Hence this requires to somehow model the epidemic before t_b, where data may simply not be available.

Therefore we assume that the individuals initially undiagnosed (at time t_b) progress according to their calendar age at t_b, rather than their age at infection; this might lead to bias, as individuals infected at a older age are assumed to progress faster towards AIDS.

Initially undiagnosed infections are denoted by the vectors $\boldsymbol{\pi}_j = (\pi_{j,1}, \ldots, \pi_{j,K})^T$, $(j = \{1, \ldots, A\})$; the model dynamics, for time intervals $(t_{b+i-1}, t_{b+i}]$ and age intervals $(a_{j-1}, a_j]$ $(i = \{1, \ldots, T-b\},\ j = \{1, \ldots, A\})$, can be expressed simply by modifying the starting value $(e^{j_0}_{i_0,j_0})$ of Equations 5.3.3 and 5.3.4 as follows:

$$\boldsymbol{e}^{j_0}_{i_0,j_0} = \begin{cases} (h_{i_0,j_0} + \pi_{j_0,1}, \pi_{j_0,2}, \ldots, \pi_{j_0,K})^T & \text{if } i_0 = 1 \\ (h_{i_0,j_0}, 0, \ldots, 0)^T & \text{if } i_0 > 1 \end{cases} \tag{5.4.3}$$

Note that $i_0 = 1$ now denotes the time interval $(t_b, t_{b+1}]$, where initially undiagnosed infections must be taken into account.

5.4.3 Back-calculation on a coarser time scale

Epidemic data may only be available on a coarse scale; in this situation allowing at most one transition between the states of the model (as in Section 5.3.1) in an interval does not allow infected individuals to be diagnosed rapidly enough. However the dynamics of the model over the coarse scale, can be reconstructed by considering smaller sub-intervals, where the aforementioned assumption holds (see Section 2.4.3).

The period $(t_0, t_T]$ spanning the epidemic is thus split into T disjoint, consecutive, "large" intervals $(t_{i-1}, t_i]$, $i = \{1, \ldots, T\}$ and the age-range $(a_0, a_A]$ is further split into A disjoint, consecutive "large" intervals $(a_{j-1}, a_j]$, $j = \{1, \ldots, A\}$. Time and age intervals are split into N_s intervals of equal length, denoted $(t_{i,s-1}, t_{i,s}]$ and $(a_{j,s-1}, a_{j,s}]$, where $s = \{1, \ldots, N_s\}$ and $t_{i,0} \equiv t_{i-1}$, $t_{i,N_s} \equiv t_i$ and $a_{j,0} \equiv a_{j-1}$, $a_{j,N_s} \equiv a_j$. As in Section 2.4.3, it is assumed that at most one move between the states of the model is allowed in the sub-intervals; consequently up to N_s move between the states of the model are allowed in the intervals. Note that time and age are measured on the same scale (for both the intervals and the sub-intervals), hence when one time interval elapses so does the age interval.

Let $h_{i_0,j_0,s}$ denote the expected number of new infections $(t_{i_0,s-1}, t_{i_0,s}]$ and $(a_{j_0,s-1}, a_{j_0,s}]$. $d_{k,i,j,s}$ is the diagnosis probability in $(t_{i,s-1}, t_{i,s}]$ and $(a_{j,s-1}, a_{j,s}]$ from undiagnosed state k, whereas $q^{j_0}_{k,s}$ denotes the progression probability from undiagnosed state k for an individual infected in $(a_{j_0,s-1}, a_{j_0,s}]$. The transition (denoted by $\boldsymbol{Q}^{j_0}_{i,j,s}$) and progression matrices ($\boldsymbol{D}^{j_0}_{i,j,s}$) can easily be defined (Equations 5.3.1 and 5.3.2) in $(t_{i,s-1}, t_{i,s}]$ and $(a_{j,s-1}, a_{j,s}]$ as only one movement between the states of the model is allowed. It is further assumed that the expected number of new infections and diagnosis probabilities, and thus progression and transition matrices, are constant in the N_s sub-intervals, *i.e.*:

$$\begin{array}{ll}
h_{i_0,j_0} \equiv h_{i_0,j_0,1} = \cdots = h_{i_0,j_0,N_s}, & i_0 = \{1, \ldots T\},\ j_0 = \{1, \ldots A\} \\
d_{k,i,j} \equiv d_{k,i,j,1} = \cdots = d_{k,i,j,N_s}, & i = \{1, \ldots T\},\ j = \{1, \ldots, A\},\ k = \{1, \ldots, K\} \\
q^{j_0}_{k} \equiv q^{j_0}_{k,1} = \cdots = q^{j_0}_{k,j,N_s}, & j_0 = \{1, \ldots, A\},\ k = \{1, \ldots, K-1\} \\
\boldsymbol{Q}_{i,j,j_0} \equiv \boldsymbol{Q}_{i,j,j_0,1} = \cdots = \boldsymbol{Q}_{i,j,j_0,N_s}, & i = \{1, \ldots T\},\ j = \{1, \ldots, A\},\ j_0 = \{1, \ldots, A\} \\
\boldsymbol{D}_{i,j,j_0} \equiv \boldsymbol{D}_{i,j,j_0,1} = \cdots = \boldsymbol{D}_{i,j,j_0,N_s}, & i = \{1, \ldots T\},\ j = \{1, \ldots, A\},\ j_0 = \{1, \ldots, A\}
\end{array}$$

Now the expected number of undiagnosed infections ($\boldsymbol{e}^{j_0}_{i,j}$) and new diagnoses ($\boldsymbol{\mu}_{i,j,j_0}$) at the end of the i^{th} time and j^{th} age interval, for infection occurring in the j_0^{th} age interval (and implicitly in the i^{th} time interval), can be expressed using the following recursive equations, for $i = \{i_0 + 1, \ldots, T\}$ and $j = \{j_0 + 1, \ldots, j_0 + T - i_0\}$:

$$\boldsymbol{e}^{j_0}_{i,j} = \left(\widetilde{\boldsymbol{Q}}^{j_0}_{i,j}\right)^{\boldsymbol{T}} \boldsymbol{e}^{j_0}_{i-1,j-1} \tag{5.4.4}$$

$$\boldsymbol{\mu}^{j_0}_{i,j} = \left(\widetilde{\boldsymbol{D}}^{j_0}_{i,j}\right)^{\boldsymbol{T}} \boldsymbol{e}^{j_0}_{i-1,j-1} \tag{5.4.5}$$

where:

$$\widetilde{\boldsymbol{Q}}^{j_0}_{i,j} = \left(\boldsymbol{Q}^{j_0}_{i,j}\right)^{N_s} \qquad \widetilde{\boldsymbol{D}}^{j_0}_{i,j} = \sum_{s=0}^{N_s-1} \left(\boldsymbol{Q}^{j_0}_{i,j}\right)^{s} \boldsymbol{D}^{j_0}_{i,j} \tag{5.4.6}$$

and the initial values of the recursion are:

$$\boldsymbol{e}^{j_0}_{i_0,j_0} = \left(\sum_{s=0}^{N_s-1} \left(\boldsymbol{Q}^{j_0}_{i_0,j_0}\right)^{s}\right)^{T} \boldsymbol{h}_{i_0,j_0} \tag{5.4.7}$$

$$\boldsymbol{\mu}^{j_0}_{i_0,j_0} = \left(\sum_{s=1}^{N_s-1} (N_s - s) \left(\boldsymbol{Q}^{j_0}_{i_0,j_0}\right)^{s-1} \boldsymbol{D}^{j_0}_{i_0,j_0}\right)^{T} \boldsymbol{h}_{i_0,j_0} \tag{5.4.8}$$

where the K-vector $\boldsymbol{h}_{i_0,j_0} = (h_{i_0,j_0}, 0, \ldots, 0)^T$ and $\boldsymbol{Q}^{0}_{i,j,j_0}$ is a $K \times K$ identity matrix.

The above equations are equivalent to Equations 2.4.5 and 2.4.6 and can be interpreted as discussed in Section 2.4.3. The only difference is that, as age-dependent back-calculation is considered, the infections vector and the diagnosis and progression transition matrices further depend on the age intervals of infection and diagnosis.

5.4.4 Back-calculation on different age and time scales

So far the assumption that time and age are measured on the same scale has been central to the formulation of age-dependent model-dynamics (in Sections 5.3.2, 5.4.3 and 5.4.3). This assumption is here relaxed, as surveillance data are often available on a larger age scale than time scale (*e.g.* quarterly time scale and yearly age scale for MSM in England and Wales). Data could be aggregated to have equal scales, but this would entail a loss of information.

Intervals $(t_{i-1}, t_i]$, $i = \{1, \ldots T\}$, and $(a_{j-1}, a_j], j = \{1, \ldots, A\}$, are now defined so that the length of N_a time-intervals is equal to the length of one age-interval (*e.g.* $N_a = 4$ for a quarterly time scale and a yearly age scale). All individuals in the model become one age-interval older in the beginning of the $(N_a + 1)^{\text{th}}$, $(2N_a + 1)^{\text{th}}$ intervals etcetera. This implies that the number of age-intervals elapsing between the i_1^{th} and the i_2^{th} time interval is equal to: $\left\lfloor \frac{i_2 - \varepsilon}{N_a} \right\rfloor - \left\lfloor \frac{i_1 - \varepsilon}{N_a} \right\rfloor$, where ε is an infinitely small positive value.

To begin, assume that at most one transition is allowed, between the states of the model, per time interval. The dynamical Equations 5.3.3 and 5.3.4 can be re-written as follows, for

$i = \{i_0+1,\dots,T\}$ and $j_0 = \{j_0, +\left\lfloor \frac{i_0+1-\varepsilon}{N_a} \right\rfloor - \left\lfloor \frac{i_0-\varepsilon}{N_a} \right\rfloor, \dots, j_0 + \left\lfloor \frac{T-\varepsilon}{N_a} \right\rfloor - \left\lfloor \frac{i_0-\varepsilon}{N_a} \right\rfloor\}$:

$$\boldsymbol{e}^{j_0}_{i,j} = \begin{cases} \left(\boldsymbol{Q}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j} & \text{if } i \,\%\, N_a \neq 1 \\ \left(\boldsymbol{Q}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j-1} & \text{if } i \,\%\, N_a = 1 \end{cases} \tag{5.4.9}$$

where $i \,\%\, N_a = 1$ denotes the case where the remainder of the integer division of i by N_a is equal to one, identifying time intervals $i = \{N_a+1, 2N_a+1, \dots, \lfloor T/N_a \rfloor + 1\}$:

$$\boldsymbol{\mu}^{j_0}_{i,j} = \begin{cases} \left(\boldsymbol{D}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j} & \text{if } i \,\%\, N_a \neq 1 \\ \left(\boldsymbol{D}^{j_0}_{i,j}\right)^T \boldsymbol{e}^{j_0}_{i-1,j-1} & \text{if } i \,\%\, N_a = 1 \end{cases} \tag{5.4.10}$$

where $\boldsymbol{e}^{j_0}_{i_0,j_0} = (h_{i_0,j_0}, 0, \dots, 0)^T, i_0 = \{1, \dots, T-1\}$, $j_0 = \{1, \dots, A-1\}$.

Note that the assumption that at most one move per time interval can be also relaxed considering a coarser scale, following the instructions given in Section 5.4.3. In this context, the model dynamics would be obtained by replacing Equations 5.4.9 and 5.4.10 with Equations 5.4.4 and 5.4.5. Hence the above matrices $\boldsymbol{Q}^{j_0}_{i,j}$ and $\boldsymbol{D}^{j_0}_{i,j}$ would be replaced by matrices $\widetilde{\boldsymbol{Q}}^{j_0}_{i,j}$ and $\widetilde{\boldsymbol{D}}^{j_0}_{i,j}$ (defined in Equation 5.4.6) and $\boldsymbol{e}^{j_0}_{i_0,j_0}$ would be defined as in Equation 5.4.7.

Scenarios where the time scale is wider than the age scale could also be easily considered, by reversing the time and age indices used in this Section.

5.5 Summary

In this Chapter an extension of the multi-state back-calculation model (Chapter 2) has been introduced to estimate the number of infections and the diagnosis probabilities by age.

Recall that parameterisations of $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$ have not yet been discussed. Modelling the bivariate incidence surface is particularly challenging, thus the next Chapter will review a number of non-parametric methods, which will subsequently be applied in a novel way, within the proposed age-dependent back-calculation framework. A simulation study will be undertaken to understand the appropriateness of these methods.

Sections 5.3.2, 5.4.3 and 5.4.4 proposed different time-and-age scales for modelling, motivated by both the availability of data and computational considerations: choosing a coarser scale may lead to computational savings, yielding however a rougher approximation of the underlying true epidemic process. The sensitivity of the results obtained on the choice of the scale employed has not been assessed in the literature and thus will be further investigated in this thesis.

Chapter 6

Two dimensional non-parametric smoothing methods

6.1 Introduction

Age-specific back-calculation, as discussed in Chapter 5, requires modelling the latent bivariate incidence surface. Thus far bivariate incidence has only been modelled using strong parametric assumptions, or bivariate step functions (Rosenberg, 1995; Marschner and Bosch, 1998). Here, we investigate extensions of the univariate non-parametric smoothing models, discussed in Chapter 3, to bivariate settings. Consider data $(y_i, \boldsymbol{x}_i)$, $i = \{1, \ldots, n\}$, where $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ is a $n \times 1$ vector of observations and $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are associated covariates, so that $\boldsymbol{x}_i = [x_{i1}, x_{i2}]^T$ is a 2×1 vector. It is of interest finding a smooth surface $g(\boldsymbol{x}) : [a,b] \times [c,d] \rightarrow \mathbb{R}$ so that, for $i = \{1, \ldots, n\}$:

$$y_i = g(\boldsymbol{x}_i) + \varepsilon_i \tag{6.1.1}$$

where ε_i are assumed to be i.i.d zero mean random variables. This is known as the scatter plot smoothing problem.

The structure of this Chapter is similar to the one of Chapter 3: bivariate splines (Section 6.2) and bivariate Gaussian processes (Section 6.3) are first considered in a simple scatter plot smoothing framework, in order to understand their properties. These are subsequently applied within the age-specific back-calculation framework of Section 6.4, where these smoothing methods are extended to model the latent age-and-time specific incidence surface.

6.2 Splines

6.2.1 Penalised regression

As in Chapter 3, bivariate splines are considered within a penalised regression framework, using the following PLS criterion:

$$min\ ||\boldsymbol{y}-\boldsymbol{X\beta}||^2+\sum_{i=1}^{N_p}\lambda_i\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}_i\boldsymbol{\beta} \tag{6.2.1}$$

where $\boldsymbol{X}$ is the design matrix of the spline and $\boldsymbol{\beta}$ the respective coefficients. In contrast to univariate splines (Equation 3.3.1), $\boldsymbol{\beta}$ may be subject to more than a single quadratic penalty matrix and thus there may be multiple smoothing parameters. This occurs when roughness is not equally penalised in both dimensions (see Section 6.2.5), or when certain reparameterisations are employed (see Section 6.2.7). Let N_p be the number of quadratic penalties (denoted $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_{N_p}$) and smoothing parameters (denoted $\lambda_1,\ldots,\lambda_{N_p}$).

It can be shown that the vector $\hat{\boldsymbol{\beta}}$ minimizing the PLS criterion is given by (Wood, 2006a):

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X^TX}+\sum_{i=1}^{N_p}\lambda_i\boldsymbol{S_i}\right)^{-1}\boldsymbol{X^Ty} \tag{6.2.2}$$

Sections 6.2.2 to 6.2.5 will discuss various approaches for constructing bivariate splines and will explicitly define $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_{N_p}$.

6.2.2 Optimal thin plate splines

Reinsch (1967) extends the univariate smoothing problem, posed in Section 3.3.3 to bivariate settings, by measuring roughness in $\mathbb{R}^2$ (*i.e.* in two dimensions) using the Laplacian quadratic integral. Hence the problem of finding a function $g(\boldsymbol{x})$, in the space of twice continuously differentiable functions, minimising the following criterion is considered:

$$min\sum_{i=1}^{n}(\,y_i-g(\boldsymbol{x}_i)\,)^2+\lambda\int_a^b\int_c^d\left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_1^2}\right)^2+2\left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_1\partial x_2}\right)^2+\left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_2^2}\right)^2 dx_1dx_2 \tag{6.2.3}$$

Similarly to the univariate objective function (Equation 3.3.3), the above criterion is a special case of penalised regression (see Section 6.2.1); it is a compromise between goodness of fit, as measured by the residual sum of squares (*i.e.* the first term in Equation 6.2.3), and the curve's roughness, as measured by the Laplacian quadratic integral. Again, the smoothing parameter λ determines the trade-off between goodness of fit and smoothness. The following theorem provides the solution to the minimization problem:

Theorem. In the space of continuously differentiable functions in $[a,b] \times [c,d]$, Equation 6.2.3 is uniquely minimized by a thin plate spline with a knot at every unique $\boldsymbol{x}_i$.

Thin Plate Splines (TPS) are formally defined as follows:

Definition 6.2.1. A thin plate spline is a function $g(\boldsymbol{x}) : [a,b] \times [c,d] \rightarrow \mathbb{R}$ defined on a set of knots $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_1, \cdots, \boldsymbol{\kappa}_k\}$, where $\boldsymbol{\kappa}_i = (\kappa_{i1}, \kappa_{i2})$, so that:

$$g(\boldsymbol{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \sum_{i=1}^{k} \delta_i \nu(||\boldsymbol{x}_i - \boldsymbol{\kappa}_i||)$$

subject to the following constraints:

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \delta_i \kappa_{i1} = \sum_{i=1}^{k} \delta_i \kappa_{i2} = 0$$

where $|| \cdot ||$ denotes the Euclidean distance and $\nu(r)$ is a $\mathbb{R} \rightarrow \mathbb{R}$ distance function:

$$\nu(r) = \begin{cases} \frac{1}{16\pi} r^2 log(r^2) & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases}$$

A TPS basis is thus: $\{1, x_1, x_2, \nu(||\boldsymbol{x} - \boldsymbol{\kappa}_1||), \ldots, \nu(||\boldsymbol{x} - \boldsymbol{\kappa}_k||)\}$.

Thin plate splines with a knot per observation (referred to as *"optimal TPS"* from now on) are optimal as if roughness is measured by the Laplacian integral, there is no smoother spline with equal (or better) goodness of fit. Note that optimal TPS are simply optimal NCS (Section 3.3.3) extended to bivariate settings.

Optimal TPS (Definition 6.2.1) can be formulated within a penalised regression framework (Section 6.2.1), so that parameter estimates can be obtained. Let us start by defining the parameter vectors $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]^T$ and $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_n]^T$, and matrices $\boldsymbol{T}$ and $\boldsymbol{E}$ (of dimension $n \times 3$ and $n \times n$ respectively):

$$\boldsymbol{T} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \nu(||\boldsymbol{x}_1 - \boldsymbol{x}_1||) & \cdots & \nu(||\boldsymbol{x}_1 - \boldsymbol{x}_n||) \\ \vdots & \cdots & \vdots \\ \nu(||\boldsymbol{x}_n - \boldsymbol{x}_1||) & \cdots & \nu(||\boldsymbol{x}_n - \boldsymbol{x}_n||) \end{bmatrix} \tag{6.2.4}$$

It can be shown that the roughness integral can be written in a quadratic form (Green and Silverman, 1994):

$$\int_a^b \int_c^d \left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 g(\boldsymbol{x})}{\partial x_2^2}\right)^2 dx_1 dx_2 = \boldsymbol{\delta}^{\boldsymbol{T}} \boldsymbol{E} \boldsymbol{\delta}$$

Hence, the smoothing criterion (Equation 6.2.3) can be reformulated as:

$$min||\boldsymbol{y} - \boldsymbol{T}\boldsymbol{\alpha} - \boldsymbol{E}\boldsymbol{\delta}||^2 + \lambda \boldsymbol{\delta}^{\boldsymbol{T}} \boldsymbol{E} \boldsymbol{\delta} \quad \text{s.t} \quad \boldsymbol{T}^{\boldsymbol{T}} \boldsymbol{\delta} = \mathbf{0}$$

The objective function above can be further reformulated into the unconstrained PLS criterion of Equation 6.2.1, following the approach discussed for optimal NCS in Section 3.3.3 and Appendix C.1.2. For the mathematical details refer to Appendix F.1.

Similarly to optimal NCS, optimal TPS require one parameter for each data-point. For large data volumes, a large number of correlated parameters must be estimated, which may result in long computational running time and potentially numerical problems. Within a penalised regression framework (Section 6.2.1), the smoothing parameter λ offsets the overfitting effect induced by a large number of knots. Thus a similar fit can be obtained with fewer knots and a lower λ value (see Section 3.3.2 and Figure 3.4b). In practice, low-rank approximations of optimal TPS, that is bivariate splines with fewer parameters than knots, produce very similar results to optimal TPS, despite not having the optimality properties discussed. The following Sections discuss various types of low-rank thin plate splines.

6.2.3 Knots based thin plate splines

In a similar way to NCS (Section 3.3.4), an optimal TPS can be approximated by a low-rank knots-based TPS, that uses a subset of the observations as knots. Let $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_1 < \cdots < \boldsymbol{\kappa}_k\}$ be a set of knots ($k < n$), so that $\boldsymbol{\kappa}_i = (\kappa_{i1}, \kappa_{i2})$. Let $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]^T$ and $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_k]^T$ be parameter vectors and $\boldsymbol{T}$, $\boldsymbol{E}$ and $\boldsymbol{C}$ matrices of dimension $n \times 3$, $n \times k$

and $3 \times k$ respectively:

$$\boldsymbol{T} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \nu(||\boldsymbol{x}_1 - \boldsymbol{\kappa}_1||) & \cdots & \nu(||\boldsymbol{x}_1 - \boldsymbol{\kappa}_k||) \\ \vdots & \cdots & \vdots \\ \nu(||\boldsymbol{x}_n - \boldsymbol{\kappa}_1||) & \cdots & \nu(||\boldsymbol{x}_n - \boldsymbol{\kappa}_k||) \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \kappa_{11} & \kappa_{12} & \cdots & \kappa_{1k} \\ \kappa_{21} & \kappa_{22} & \cdots & \kappa_{2k} \end{bmatrix}$$

Following the same rationale as in Section 3.3.4, a knots based TPS can be expressed using the following PLS criterion:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda \boldsymbol{\delta^T E \delta} \quad \text{s.t} \quad \boldsymbol{C\delta} = \boldsymbol{0}$$

This can be conveniently rearranged in the unconstrained PLS criterion $min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$, characterising the penalised regression framework. For the mathematical details refer to Appendix F.2.

The number of knots chosen must be large enough, to ensure sufficient flexibility, but not excessively large, to avoid computational waste. Sensitivity analysis to the number of knots and their location is essential.

6.2.4 Thin plate regression splines

Similarly to optimal NCS, an "optimal" (according to the definition given in Section 3.3.5) low-rank approximation of optimal TPS can be obtained by extending thin plate regression splines (Section 3.3.5), potentially with shrinkage (Section 3.3.6), to bivariate settings. A thin plate-regression spline can be defined (as in Section 3.3.5) via the following PLS criterion:

$$min||\boldsymbol{y} - \boldsymbol{EU_k\delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \delta_k^T \boldsymbol{U_k^T E U_k} \delta_k \quad \text{s.t} \quad \boldsymbol{T^T U_k \delta_k} = 0 \tag{6.2.5}$$

where $\boldsymbol{E}$ and $\boldsymbol{T}$ are defined in Equation 6.2.4. $\boldsymbol{D_k}$ is a $k \times k$ diagonal matrix containing, in ascending order (*i.e.* from top left to bottom right), the k largest absolute values of the eigenvalues of the matrix $\boldsymbol{E}$. $\boldsymbol{U_k}$ is the $n \times k$ matrix consisting of the first k columns of $\boldsymbol{U}$ (the matrix of eigenvectors of $\boldsymbol{E}$).

The criterion in Equation 6.2.5 can be expressed as the unconstrained criterion $min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}$, details are available in Appendix F.3.

Shrinkage of thin plate regression splines towards zero can be further achieved (as in Section 3.3.6) by imposing a penalty on the null-space (*i.e.* the $\boldsymbol{\alpha}$ coefficients); this defines thin plate regression splines with shrinkage.

Thin plate regression splines are defined in terms of the k largest eigenvectors, rather than a set of knots, avoiding having to explicitly choose the knots location. The number of parameters k shall be chosen to be adequately large to ensure sufficient flexibility, but not excessively large, to avoid computational waste.

6.2.5 Tensor product splines

Tensor product splines were first introduced by De Boor (1978). Eilers and Marx (2003) estimate these within a penalised regression framework, extending their work on P-splines (Section 3.3.7) to bivariate settings. This approach has been further generalised, to consider other type of splines, by Wood (2006b).

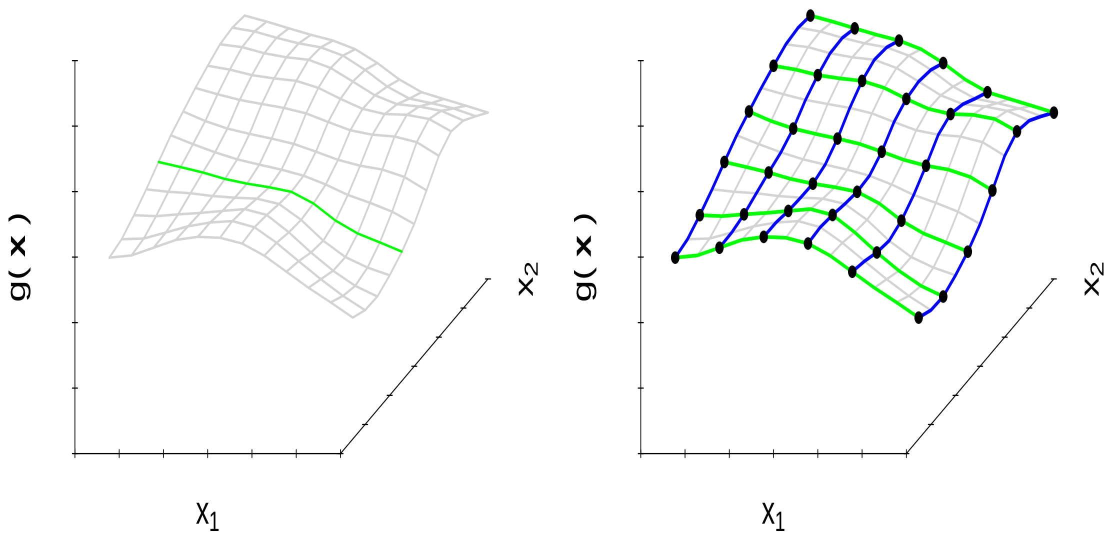
Tensor product splines are obtained by multiplying two univariate spline bases, each separately defined on the marginal dimensions (*i.e.* x_1 and x_2). This is a pragmatic approach for constructing bivariate splines; unlike bivariate TPS splines (Sections 6.2.2 to 6.2.4), these are not motivated by the smoothing criterion in Equation 6.2.3 depending on the Laplacian quadratic integral, which assumes isotropy, *i.e.* roughness is equally penalised in the dimensions x_1 and x_2. Tensor product splines relax this assumption.

We start by considering two univariate splines $g_1(x_1):[a,b]\rightarrow\mathbb{R}$ and $g_2(x_2):[c,d]\rightarrow\mathbb{R}$, with bases $\{B_{11}(x_1),\cdots,B_{1t_1}(x_1)\}$ and $\{B_{21}(x_2),\cdots,B_{2t_2}(x_2)\}$, coefficients $\boldsymbol{\beta_1}=[\beta_{11},\ldots,\beta_{1t_1}]^T$ and $\boldsymbol{\beta_2}=[\beta_{21},\ldots,\beta_{2t_2}]^T$, design matrices $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ (of dimension $n\times t_1$ and $n\times t_2$ respectively), and penalty matrices $\boldsymbol{S}_{(1)}$ and $\boldsymbol{S}_{(2)}$ (of dimension $t_1\times t_1$ and $t_2\times t_2$ respectively). For $i=\{1,\ldots,t_1\}$ and $j=\{1,\ldots,t_2\}$ the tensor product basis is equal to:

$$B_{i,j}(x_1,x_2)=B_{1i}(x_1)B_{2j}(x_2)$$

Hence the tensor product design matrix $\boldsymbol{X}$, of dimension $n\times(t_1t_2)$, is:

$$\boldsymbol{X}=\begin{bmatrix} B_{1,1}(x_{11},x_{12}) & \cdots & B_{t_1,1}(x_{11},x_{12}) & \cdots & B_{1,t_2}(x_{11},x_{12}) & \cdots & B_{t_1,t_2}(x_{11},x_{12}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{1,1}(x_{n1},x_{n2}) & \cdots & B_{t_1,1}(x_{n1},x_{n2}) & \cdots & B_{1,t_2}(x_{n1},x_{n2}) & \cdots & B_{t_1,t_2}(x_{n1},x_{n2}) \end{bmatrix}$$



(a) (b)

Fig. 6.1 Measuring the bivariate roughness of a tensor product spline. The grey lines represent the continuous bivariate spline surface. a) The green line denotes $g(x_1|x_2)$ for a fixed x_2. Its roughness $J_1(g(x_1|x_2))$ (in the x_1 dimension) can easily be evaluated. b) $g(x_1|x_2)$ (green lines) and $g(x_2|x_1)$ (blue lines), for a number of fixed x_1 and x_2 on a regular grid (black dots). The marginal roughnesses $J(g(x_1|x_2))$ and $J(g(x_2|x_1))$ can be evaluated at each fixed x_1 and x_2 respectively.

The i$^{\text{th}}$ row of $\boldsymbol{X}$, denoted $\boldsymbol{X}_{i\cdot}$, can be alternatively obtained from the Kroenecker product of the i$^{\text{th}}$ rows of the marginal P-splines design matrices, $\boldsymbol{X}_{(1)i\cdot}$ and $\boldsymbol{X}_{(2)i\cdot}$ respectively, yielding a $(t_1 t_2) \times 1$ vector:

$$\boldsymbol{X}_{i\cdot} = \boldsymbol{X}_{(1)i\cdot} \otimes \boldsymbol{X}_{(2)i\cdot} \tag{6.2.6}$$

The $(t_1 t_2) \times 1$ parameter vector of the tensor product spline is $\boldsymbol{\beta} = [\beta_{1,1}, \cdots, \beta_{t_1,1}, \cdots, \beta_{1,t_2}, \cdots, \beta_{t_1,t_2}]^T$.

The overall roughness of the tensor product spline can be quantified based on the idea that we know how to measure roughness marginally (via the specification of $\boldsymbol{S}_{(1)}$ and $\boldsymbol{S}_{(2)}$). For a fixed x_2, $J_1(g(x_1|x_2)) = \boldsymbol{\beta}_{\mathbf{1}}^T \boldsymbol{S}_{(1)} \boldsymbol{\beta}_{\mathbf{1}}$ quantifies roughness with respect to the x_1 dimension (Figure 6.1a). Evaluating $J_1(g(x_1|x_2))$ for infinitely many fixed x_2 and taking its average over the x_2 points yields the overall roughness with respect to x_1. Roughness in the x_2 direction is similarly measured, fixing x_1 instead. The overall roughness of $g(x_1, x_2)$ is mathematically expressed as:

$$J(g(\boldsymbol{x})) = \lambda_{(1)} \int_c^d J_1\left(g_1(x_1|x_2)\right) dx_2 + \lambda_{(2)} \int_a^b J_{x_2}\left(g_2(x_2|x_1)\right) dx_1 \tag{6.2.7}$$

This integral can not be analytically evaluated; a discrete approximation can be derived by reparameterising the spline in terms of the values of the functions on a regular grid (depicted by the black dots in Figure 6.1b). A set of equidistant values, denoted as $\boldsymbol{x}_1^\star = \{x_{11}^\star, \ldots, x_{1t_1}^\star\}$ and $\boldsymbol{x}_2^\star = \{x_{21}^\star, \ldots, x_{2t_2}^\star\}$ respectively, is constructed in the x_1 and x_2 dimensions. It can be shown that (Wood, 2006b):

$$\int_c^d J_1\left(g_1(x_1|x_2)\right)dx_2 \approx h_1 \sum_{j=1}^{t_2} J_1\left(g_1(x_1|x_{2j}^\star)\right) = h_1\left(\boldsymbol{\beta}^T\left(\boldsymbol{A}_1^{-T}\boldsymbol{S}_{(1)}\boldsymbol{A}_1\right)\otimes \boldsymbol{I}_{t_2}\boldsymbol{\beta}\right) \qquad (6.2.8)$$

where $\boldsymbol{A}_1$ is a $t_1 \times t_1$ matrix with entries $(\boldsymbol{A}_1)_{ij} = B_{1i}(x_{1j}^\star)$, and h_1 is a constant of proportionality to account for the spacing of $\boldsymbol{x}_1^\star$. The integral for $J_2\left(g_2(x_2|x_1)\right)$ can be similarly approximated, by defining instead h_2 and a $t_2 \times t_2$ matrix $\boldsymbol{A}_2$, with entries $(\boldsymbol{A}_2)_{ij} = B_{2i}(x_{2j}^\star)$.

Tensor product splines can be expressed within the usual penalised regression framework (Equation 6.2.1):

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_1\boldsymbol{\beta}^T\boldsymbol{S}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^T\boldsymbol{S}_2\boldsymbol{\beta} \qquad (6.2.9)$$

where $\boldsymbol{\beta} = [\beta_{1,1}, \cdots, \beta_{t_1,1}, \cdots, \beta_{1,t_2}, \cdots, \beta_{t_1,t_2}]^T$, $\lambda_1 = h_1\lambda_{(1)}$, $\lambda_2 = h_2\lambda_{(2)}$ and:

$$\boldsymbol{S}_1 = \left(\boldsymbol{A}_1^{-T}\boldsymbol{S}_{(1)}\boldsymbol{A}_1\right)\otimes \boldsymbol{I}_{t_2}$$
$$\boldsymbol{S}_2 = \boldsymbol{I}_{t_1}\otimes\left(\boldsymbol{A}_2^{-T}\boldsymbol{S}_{(2)}\boldsymbol{A}_2\right)$$

6.2.6 A comparison of splines

In contrast to the univariate case, there are substantial differences between the TPS family of splines (optimal TPS, knots based TPS and thin plate regression splines, Sections 6.2.2 to 6.2.4) and tensor product splines (Section 6.2.5).

TPS splines are subject to a single penalty, given by the Laplacian quadratic integral. This is invariant to rotation and translation in $\mathbb{R}^2$, but not to the rescaling of coordinates $\boldsymbol{x}$. The Laplacian quadratic integral is also isotropic, which is a desirable property when (x_1, x_2) are measured using the same unit.

Tensor product splines relax the isotropy assumption, allowing for different smoothing levels in different dimensions and ensure that results are unaffected by rescaling and translation of coordinates $\boldsymbol{x}$. However, they are not invariant to rotation of coordinates.

The appropriateness of the type of spline is application specific. Thin plate-splines are typically well-suited for geographic smoothing problems where (x_1, x_2) are measured using the same unit (*e.g.* latitude and longitude in degrees). Rotational invariance ensures that results obtained would be unaffected if the axis were to be rotated. Tensor product splines are also suitable for this type of application, and can be used to informally test whether the isotropy assumption holds; however tensor product splines do not guarantee rotational invariance. On the other hand, tensor product splines are better suited when (x_1,x_2) are measured on different scales (*e.g.* meters and years) as the isotropy assumption does not hold in this case. Moreover the scale invariance property of tensor product splines (that does not hold for TPS) ensures that the same results would be obtained if (x_1,x_2) were rescaled (*e.g.* kilometres and quarters). Finally note that when covariates (x_1,x_2) are measured on the same scale, but on very different ranges (*e.g.* $[0,10]$ for x_1 and $[0,100]$ for x_2), tensor product splines should be used rather than TPS. In fact, it is unfair to equally penalise a unit change in the x_1 dimension ($1/10$) as a unit change in the x_2 dimension ($1/100$). This issue could be circumvented by rescaling both dimensions on the unit space; however using tensor product splines is preferable as rescaling is arbitrary.

6.2.7 Bayesian inference

The advantages of Bayesian inference over penalised likelihood inference were discussed in Section 3.3.9. Recall that the penalty term can be viewed as a prior on the spline's coefficients $\boldsymbol{\beta} \sim N(\mathbf{0},(\lambda \boldsymbol{S})^{-1})$. As splines of the TPS-family are subject to a single penalty term, they can be formulated within a Bayesian framework using i.i.d Normal priors, after a suitable reparameterisation (Section 3.3.9).

Tensor product splines instead have two different smoothing parameters and penalty matrices, corresponding to the prior $\boldsymbol{\beta} \sim N(\mathbf{0},(\lambda_1 \boldsymbol{S_1} + \lambda_2 \boldsymbol{S_2})^{-1})$. However the above precision matrix can not be reparameterised as an identity matrix (to obtain i.i.d priors); reasons are discussed in Appendix F.4.

Depending on the definition of $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$, the matrix $\boldsymbol{S} = \boldsymbol{S_1} + \boldsymbol{S_2}$ may not be of full rank, so that the vector of coefficients $\boldsymbol{\beta}$ can be split into p penalised and u unpenalised parameters $\boldsymbol{\beta} = [\boldsymbol{\beta_p}\ \boldsymbol{\beta_u}]^T$. Let $P = p + u$ be the length of the coefficient vector.

If $\boldsymbol{S}$ is not of full rank, the multivariate normal prior on $\boldsymbol{\beta}$ is improper. To avoid the use of the improper prior, a "small" quadratic penalty $\boldsymbol{S_0}$ is introduced for the unpenalised parameters $\boldsymbol{\beta_u}$, so that an approximate penalty $\tilde{\boldsymbol{S}} = \boldsymbol{S_1} + \boldsymbol{S_2} + \boldsymbol{S_0}$ of full rank can be defined (Marra and Wood, 2011). $\boldsymbol{S_0}$ is constructed starting from the positive-definite symmetric matrix $\boldsymbol{S} = \boldsymbol{S_1} + \boldsymbol{S_2}$ (size $P \times P$ and rank p). Its eigen-decomposition is $\boldsymbol{S} = \boldsymbol{UDU^T}$, where $\boldsymbol{U}$ and $\boldsymbol{D}$

are defined as in Section 3.3.5. The last u eigenvalues of $\boldsymbol{D}$ are zero, and $\boldsymbol{U_0}$ are the columns of $\boldsymbol{U}$ corresponding to the zero eigenvalues, and $\boldsymbol{S_0}$ is defined to be $\boldsymbol{U_0 U_0^T}$. Thus a proper multivariate Normal prior is obtained for $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}) \sim N_P\left(\mathbf{0}, \left(\sum_{i=1}^{2} \lambda_i \boldsymbol{S_i} + \lambda_0 \boldsymbol{S_0}\right)^{-1}\right) \tag{6.2.10}$$

which requires estimating a further smoothing parameter λ_0, ruling the amount of smoothing for the originally unpenalised coefficients $\boldsymbol{\beta_U}$.

The centering reparameterisation (see Section 3.3.9) allows for the tensor product spline to be rewritten in terms of parameters $\boldsymbol{\beta'}$ so that one of the originally unpenalised parameters $\boldsymbol{\beta_u}$ becomes the global intercept, and the sum of the spline values over the covariates is zero (*i.e.* $\sum_{i=1}^{n} g(\boldsymbol{x}_i) = 0$). For notational simplicity, let β_1' be the global intercept and $\boldsymbol{\beta'}_{P-1}$ the remaining $P-1$ parameters $\boldsymbol{\beta'} = [\beta_1', \boldsymbol{\beta'}_{P-1}]^T$.

Let $\boldsymbol{y}$ follow any distribution, not necessarily from the exponential family, with likelihood function $l(\boldsymbol{y}|\boldsymbol{\beta'})$. A Bayesian tensor product spline can be then specified:

$$\begin{aligned} \boldsymbol{y} &\sim l(\boldsymbol{y}|\boldsymbol{\beta'}) \\ \beta_1' &\sim f(\beta') \\ \boldsymbol{\beta'}_{P-1} &\sim N_{P-1}\left(\mathbf{0}, \left(\sum_{i=1}^{2} \lambda_i \boldsymbol{S_i} + \lambda_0 \boldsymbol{S_0}\right)^{-1}\right) \\ \lambda &\sim f(\lambda) \\ \lambda_0 &\sim f(\lambda_0) \end{aligned} \tag{6.2.11}$$

where $f(\cdot)$ denotes a prior distribution.

For tensor product splines, λ_0 is only required if there is, at least, one marginal spline with more than one unpenalised coefficients (*i.e.* $u > 1$) prior to reparameterisation (Section 3.3.9). Thus if first-order B-splines ($u = 1$) or thin plate regression splines with linear shrinkage ($u = 0$) are specified on both dimensions, λ_0 is not required. All other univariate splines considered (second degree B-splines, NCS, knots-based NCS and TPS) require λ_0.

6.3 Gaussian processes

Gaussian processes (GP), introduced in Chapter 3, can be extended to model data $(y_i, \boldsymbol{x}_i)$ where covariates $\boldsymbol{x}_i$ are two-dimensional; by definition this requires specifying a mean $m(\boldsymbol{x}_i)$ and a covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\phi})$. The latter is typically constructed by multiplying together covariance functions defined on each individual covariate (Duvenaud, 2014).

Using a squared exponential covariance function (Section 3.4), the covariance function of a two dimensional GP can be written as follows:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j|\eta, \rho_1, \rho_2) = \eta^2 exp\left(-\frac{1}{2}\sum_{d=1}^{2}\frac{(x_{id}-x_{jd})^2}{\rho_d^2}\right) \tag{6.3.1}$$

where $\boldsymbol{\phi} = \{\eta, \rho_1, \rho_2\}$ are the GP hyper-parameters.

The specification in Equation 6.3.1 is called the Squared Exponential (SE) kernel, and like a tensor product spline, allows for different levels of smoothing (given by ρ_1 and ρ_2) in the x_1 and x_2 dimensions. Large values of ρ_i $(i = \{1,2\})$ indicate that y is independent of the predictor x_i as the exponential term of $k(\boldsymbol{x}_i, \boldsymbol{x}_j|\eta, \rho_1, \rho_2)$ approaches zero (Neal, 1996; Williams and Rasmussen, 1996).

The covariance matrix $\boldsymbol{K}$ (dimension $n \times n$), with $(i, j)^{\text{th}}$ entry $(\boldsymbol{K})_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\phi})$, is obtained via a Kronecker product:

$$\boldsymbol{K} = \boldsymbol{K_1} \otimes \boldsymbol{K_2} \tag{6.3.2}$$

where $\boldsymbol{K_1}$ and $\boldsymbol{K_2}$ are the covariance matrices for each individual covariate, of dimension $n_1 \times n_1$ and $n_2 \times n_2$ respectively, so that $n = n_1 \times n_2$ and with entries $(\boldsymbol{K_1})_{i,j} = k(x_{i1}, x_{j1}|\boldsymbol{\phi})$ and $(\boldsymbol{K_2})_{i,j} = k(x_{i2}, x_{j2}|\boldsymbol{\phi})$.

The following Kronecker product property is key for efficient inference:

$$\boldsymbol{K}^{-1} = \boldsymbol{K_1}^{-1} \otimes \boldsymbol{K_2}^{-1} \tag{6.3.3}$$

Inverting the two individual covariance matrices, $\boldsymbol{K_1}$ and $\boldsymbol{K_2}$, is substantially faster compared to inverting the full covariance matrix $(O(n_1^3) + O(n_2^3)$ versus $O((n_1 n_2)^3)$ operations). Hence, iterative algorithms for GP that require the inversion of the covariance matrix at each iteration, substantially benefit from taking advantage of Kronecker product properties (Saatçi, 2012; Flaxman et al., 2015).

The posterior and predictive distributions are then directly obtained as in Section 3.4.3. Inference can be carried in a frequentist framework, by maximizing the (marginal) likelihood, or in a Bayesian framework (see Section 3.4.4).

6.4 Back-calculation

Chapter 5 introduced age-dependent multi-state back-calculation without, however, specifying parameterisations for the incidence surface $\mathcal{H}(\boldsymbol{\theta})$ and diagnosis probabilities $\mathcal{D}(\boldsymbol{\delta})$.

6.4.1 Incidence surface

To ensure positiveness of $\mathcal{H}(\boldsymbol{\theta})$, log-expected infections $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{TA})^T$ are modelled, where γ_{ij} denotes the log-expected number of new infections $log(h_{ij})$ in the i^{th} time and the j^{th} age intervals. Three non-parametric methods were considered to model the age-independent incidence curve (Section 3.6) and we now illustrate how these methods can be further extended to model the bivariate time and age dependent incidence surface.

Step functions

Step functions, that are equivalent to random walk priors in a Bayesian framework, can be employed to model the time-and-age dependent incidence surface using a multiplicative model (Mezzetti and Robertson, 1999; Becker et al., 2003). This does not allow for the age-profile of infection to vary over time and consequently we cannot assess the impact of public health policies targeted at specific age-groups.

Splines

The incidence surface can be modelled using a bivariate spline model, as follows:

$$\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{\beta} \tag{6.4.1}$$

where $\boldsymbol{X}$ is the design matrix of the spline. The spline parameterisation of the log-incidence surface is characterised by infection parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\lambda}\}$, where $\boldsymbol{\lambda}$ are the smoothing parameters of the spline.

Gaussian processes

A GP can also be employed to model the log-expected number of infections:

$$\boldsymbol{\gamma} \sim GP(\mathbf{0}, k(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{\phi})) \tag{6.4.2}$$

where $m(\boldsymbol{x})$ and $k(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{\phi})$ denote the mean and the covariance functions respectively (Section 6.3). Using the GP parameterisation, the infection parameters are $\boldsymbol{\theta} = \{\boldsymbol{\phi}\}$, where $\boldsymbol{\phi}$ are the hyperparameters of the GP.

6.4.2 Diagnosis process

Diagnosis probabilities are defined on a logistic scale, *i.e.* $\delta_{k,i,j} = log\left(\frac{d_{k,i,j}}{1-d_{k,i,j}}\right)$. To ensure a parsimonious representation, $\delta_{k,i,j}$ are assumed to be piecewise constant in both the time intervals $(\breve{t}_{i-1}, \breve{t}_i]$ $(i = \{1, \ldots, \breve{T}\}$, $\breve{T} \leq T$, $\breve{t}_0 \equiv t_0$, and $\breve{t}_{\breve{T}} \equiv t_T)$, and the age intervals $(\breve{a}_{j-1}, \breve{a}_j]$ $(j = \{1, \ldots, \breve{A}\}$, $\breve{A} \leq A$, $\breve{a}_{\breve{a}} \equiv a_A$, and $\breve{a}_{\breve{a}} \equiv a_A)$.

Logistic Regression

Logistic regression can be considered for $\mathcal{D}(\boldsymbol{\delta})$, for $k = \{1, \ldots, K\}$:

$$\delta_{k,i,j} = \alpha + \zeta_k + \xi_i + \varphi_j + \nu_{k,i} + \varsigma_{i,j} \tag{6.4.3}$$

subject to the identifiability constraint $\zeta_1 = \xi_1 = \varphi_1 = \nu_{1,t} = \nu_{k,1} = \varsigma_{1,a} = \varsigma_{c,1} = 0$. ζ_k, ξ_i and φ_j denote the fixed effects of undiagnosed state k, calendar interval $(\breve{t}_{i-1}, \breve{t}_i]$ and age interval $(\breve{a}_{j-1}, \breve{a}_j]$ respectively. $\nu_{k,i}$ and $\varsigma_{i,j}$ denote the interactions of state k with time interval $(\breve{t}_{i-1}, \breve{t}_i]$, and of the time interval $(\breve{t}_{i-1}, \breve{t}_i]$ with the age interval $(\breve{a}_{j-1}, \breve{a}_j]$. Further interactions could be similarly introduced. Using this parameterisation diagnosis parameters are $\boldsymbol{\delta} = \{\alpha, \zeta_1, \ldots, \zeta_k, \xi_1, \ldots, \xi_{\breve{T}},, \varphi_1, \ldots, \varphi_{\breve{A}}, \nu_{1,1}, \ldots, \nu_{K,\breve{T}}, \varsigma_{1,1}, \ldots, \varsigma_{\breve{T},\breve{A}}\}$.

Step functions

Alternatively piecewise constant step-functions can be considered to model the diagnosis probabilities. Within a Bayesian framework, these correspond to first order random walk priors for each undiagnosed state $k = \{1,\dots,K\}$:

$$\delta_{k,i,j} \sim N(\delta_{k,i-1,j}, \sigma^2_{D,k}) \tag{6.4.4}$$

where $\delta_{k,1,j} \sim N(\alpha_j, \sigma_0)$ or $\delta_{k,1,j} \sim N(\alpha_{j,k}, \sigma_0)$. α_j and $\alpha_{j,k}$ are age and age as well as state specific intercepts respectively which must be estimated, whereas σ_0 are fixed. These intercepts allow the incorporation of an age-effect (which may be undiagnosed state specific) on the diagnosis probabilities in a parsimonious manner. Using a random walk parameterisation, the diagnosis parameters are $\boldsymbol{\delta} = \{\delta_{1,1,1},\dots,\delta_{1,\check{T},\breve{A}},\dots,\delta_{K,1,1},\dots,\delta_{K,\check{T},\breve{A}}, \sigma^2_{D,1},\dots,\sigma^2_{D,K}\}$.

6.4.3 Penalised likelihood inference

Section 3.6 discussed the limitations of frequentist age-independent back-calculation inference. Nonetheless frequentist inference has been considered for age-specific back-calculation and for instance Marschner and Bosch (1998) proposed, despite in a simpler framework, penalised likelihood inference.

Age-specific back-calculation can not be expressed as a GLM, so that standard software can not be employed to fit bivariate splines. However a penalised likelihood criteria can be formulated if splines (Equation 6.4.1) and logistic regression (Equation 6.4.3) are employed to model $\mathcal{H}(\boldsymbol{\theta})$ and $\mathcal{D}(\boldsymbol{\delta})$ respectively. Splines theory (Section 6.2) suggests using the following penalty term for the age-dependent back-calculation likelihood:

$$l_p(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) = l(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) - \frac{1}{2}\sum_{i=1}^{N_p} \lambda_i \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S}_i \boldsymbol{\beta} \tag{6.4.5}$$

where $l(\mathbf{y}^H, \mathbf{y}^A, \mathbf{y}^{H_C} \mid \boldsymbol{\theta}, \boldsymbol{\delta})$ denotes the log of the likelihood in Equation 5.3.10. N_p and $\boldsymbol{S}_i$ depend on type of spline employed. For fixed λ_i, the above maximization problem does not have an analytical solution.

The penalised likelihood criterion can be numerically maximized using the quasi-Newton BFGS algorithm (Nash, 1990) via the R function *optimx* (Nash and Varadhan, 2011). In the back-calculation literature the EMS algorithm has been extensively used; this is not applicable in this situation as the derivatives of the likelihood are not analytically tractable.

Re-interpreting the penalty term as a prior (Section 6.2.7), Wood et al. (2016) show that the following large-sample approximation of the posterior distribution of the spline parameters

$\boldsymbol{\beta}$ holds for any likelihood (*i.e.* not necessarily a GLM):

$$f(\boldsymbol{\beta}|\mathcal{Y}) \sim MVN\left(\widehat{\boldsymbol{\beta}}, \left(\widehat{\boldsymbol{I}}+\boldsymbol{S}^{\lambda}\right)^{-1}\right) \tag{6.4.6}$$

where $\hat{\boldsymbol{\beta}}$ denote the estimated maximum penalised likelihood parameters and $\boldsymbol{S}^{\boldsymbol{\lambda}} = \sum_{i=1}^{N_P} \lambda_i \boldsymbol{S}_i$. $\widehat{\boldsymbol{I}}$ is the expected information matrix (i.e. the negative Hessian) evaluated at $\hat{\boldsymbol{\beta}}$.

The proposed age-specific back-calculation model employs a spline to model the incidence surface, however, Equation 6.4.6 must be appropriately adjusted to account for the D diagnosis parameters $\boldsymbol{\delta}$. These, in contrast to $\boldsymbol{\beta}$, are unpenalised. $\boldsymbol{S}^{\boldsymbol{\lambda}}$ denotes the augmented penalty matrix (size $(P+D)\ \times\ (P+D)$):

$$\boldsymbol{S}^{\boldsymbol{\lambda}} = \begin{bmatrix} (\sum_{i=1}^{N_P} \lambda_i \boldsymbol{S}_i)_{[P\times P]} & \mathbf{0}_{[P\times D]} \\ \mathbf{0}_{[D\times P]} & \mathbf{0}_{[D\times D]} \end{bmatrix}$$

where $\mathbf{0}$ denote zero matrices.

$\boldsymbol{\phi} = \left[\boldsymbol{\beta}, \boldsymbol{\delta}\right]$ (size $P+D$) is the vector of parameters of the age-specific back-calculation model. The large-sample approximation of their asymptotic posterior distribution is:

$$f(\boldsymbol{\phi}|\mathbf{y}^{H}, \mathbf{y}^{A}, \mathbf{y}^{H_C}) \sim MVN\left(\widehat{\boldsymbol{\phi}}, \left(\widehat{\boldsymbol{I}}+\boldsymbol{S}^{\boldsymbol{\lambda}}\right)^{-1}\right) \tag{6.4.7}$$

where $\widehat{\boldsymbol{\phi}}$ denotes the maximum penalised likelihood estimate of $\boldsymbol{\phi}$, for fixed $\boldsymbol{\lambda}$. $\widehat{\boldsymbol{I}}$ is the expected information matrix evaluated at $\widehat{\boldsymbol{\phi}}$.

Confidence intervals are obtained by sampling parameters from the above asymptotic distribution, which however ignores uncertainty in the smoothing parameters' estimates; thus these may be unduly narrow. Uncertainty in the smoothing parameter could be further taken into account by modifying Equation 6.4.7, as in Wood et al. (2016). This has not been considered, as it would involve obtaining further derivatives numerically, rendering the model even more computationally intensive.

As the Hessian of the age-specific back-calculation model can not be analytically derived, it is numerically evaluated using the `R` package **NumDeriv**; this requires however substantial computational effort. Furthermore, due to algorithmic numerical instabilities, there is no guarantee for $\widehat{\boldsymbol{I}}$ to be positive definite; consequently $\widehat{\boldsymbol{I}}+\boldsymbol{S}$ may be non-invertible and confidence intervals are unobtainable.

An adapted version of the Aikake Information Criterion (AIC) is used to select the optimal smoothing parameters. Wood et al. (2016) show that, under an asymptotic large-sample assumption, the AIC is equal to:

$$\mathbf{AIC} = -2l(\widehat{\boldsymbol{\phi}}) + 2\,tr\left(\left(\widehat{\boldsymbol{I}} + \boldsymbol{S}^{\boldsymbol{\lambda}}\right)^{-1}\widehat{\boldsymbol{I}}\right) \tag{6.4.8}$$

The AIC must be evaluated over a grid of plausible $\boldsymbol{\lambda}$ values; the optimal smoothing parameter vector $\widehat{\boldsymbol{\lambda}} = \{\hat{\lambda}_1, \dots, \hat{\lambda}_{N_p}\}$ correspond to the values yielding the smallest AIC.

Alternatively, parametric bootstrapping could have been considered to construct confidence intervals (Efron and Tibshirani, 1994). Given $\hat{\boldsymbol{\phi}}$, the maximum penalised likelihood estimates for the expected number of infections ($\hat{\mu}^H_{i,j}$, $\hat{\mu}^A_{i,j}$), and proportion of diagnoses in each state at each time and age intervals ($\hat{p}_{i,j,k}$) are obtained. N datasets $\{Y_1^{A\star}, Y_1^{H\star}, Y_1^{C\star}, \dots, Y_{N^\star}^{A\star}, Y_{N^\star}^{H\star}, Y_{N^\star}^{C\star}\}$ can be then simulated from the parametric distribution of the diagnosis data (*i.e.* Equations 5.3.7, 5.3.8, 5.3.9) with means $\hat{\mu}^A_{t,a}$,$\hat{\mu}^H_{t,a}$ and $\hat{p}_{k,t,a}$ respectively. Bootstrapped estimates of the parameters $\{\hat{\boldsymbol{\phi}}^\star_1, \dots, \hat{\boldsymbol{\phi}}^\star_{N^\star}\}$ are obtained using maximum penalised likelihood; confidence intervals are derived from their empirical distribution. Parametric bootstrap does not make a large-sample approximation assumption, however it is computationally strenuous as it requires re-fitting the model $N^\star$ times; hence it has not been further pursued.

6.4.4 Discussion on inferential approaches

The penalised likelihood approach discussed in the previous Section has several drawbacks; firstly, a large sample approximation based on a Bayesian re-interpretation of the penalised likelihood is required to determine the optimal smoothing parameters and confidence intervals. Moreover it may not be possible to evaluate the asymptotic distribution covariance matrix due to a lack of numerical precision.

Furthermore inference (and numerically evaluating the Hessian) is very time consuming even when the likelihood function is efficiently coded in the `C++` language, and is called from `R` via the **Rcpp** package (Eddelbuettel and François, 2011). The likelihood must be maximized for a grid of smoothing parameters in order to determine optimal $\hat{\boldsymbol{\lambda}}$; this is particularly burdensome for tensor product splines, characterised by two different smoothing parameters.

For instance, in the application for the MSM-HIV epidemic in England and Wales (Chapter 8) two hours are required to numerically maximize the likelihood and to numerically obtain the Hessian. If a tensor product spline, with a 10×10 grid of plausible $\boldsymbol{\lambda}$ values was considered, 200 hours would be required to estimate $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\lambda}}$; 20 hours would instead be sufficient for thin plate splines that involve a single smoothing parameter. Despite parallelisation reducing almost linearly (in the number of cores) the computing time, a simulation study appropriately comparing thin plate and tensor product splines is infeasible. Nevertheless, fitting tensor product splines is crucial when the isotropy assumption is unjustified (Section 6.2.6).

Most of the aforementioned caveats can be addressed within a Bayesian framework. Firstly, meaningful credible intervals (accounting for uncertainty in the smoothing parameter) are obtained from the posterior distribution, without resorting to an unverifiable large-sample approximation. Secondly, the model does not require refitting over a grid of plausible $\boldsymbol{\lambda}$, as a posterior distribution for $\boldsymbol{\lambda}$ is obtained; splines with two smoothing parameters can thus be fit without (or very marginally) increasing computational times. Furthermore Bayesian approaches allow for increased flexibility and is straightforward to consider (in contrast to frequentist methods) random walks and GPs to model the incidence surface and the diagnosis probabilities. Bayesian inference, on the other hand, requires specification of appropriate priors and MCMC algorithms; these are application-specific and are discussed in the following Chapter. Codes for age-dependent Bayesian back-calculation are available on Github (https://github.com/frbrz25/Thesis_Codes).

6.5 Summary

This Chapter extends the univariate non-parametric smoothing methods introduced in Chapter 3 to bivariate settings. Specifically, bivariate splines and Gaussian processes are described in Section 6.2 and Section 6.3 respectively.

These can be employed (Section 6.4) to parameterise the bivariate latent incidence surface and the diagnosis probabilities characterising the age-specific back-calculation model described in Chapter 5. Section 6.4.4 argues that Bayesian inference is particularly convenient. In the following Chapter the suitability of these parameterisations and the appropriateness of both inferential methods will be investigated.

Chapter 7

Age dependent back-calculation simulations

7.1 Introduction

Section 6.4 described how bivariate splines and Gaussian processes can be employed in an age-dependent multi-state back-calculation framework (Chapter 5) to smoothly model the incidence surface.

This Chapter concerns a simulation study aimed to investigate the feasibility of age-dependent back-calculation (*i.e.* whether the true incidence surface and diagnosis probabilities can be accurately estimated) and the properties of the different non-parametric smoothing methods, in order to establish which parameterisations are most suitable. Both a Bayesian and a frequentist simulation study are undertaken. As previously discussed, Bayesian inference has a number of advantages; however, it is important to demonstrate that similar results can be obtained within a frequentist framework.

This Chapter starts by describing the data generating mechanism for the simulation study (Section 7.2) and then moves on to specify a number of non-parametric models for incidence and diagnosis probabilities Section 7.3. Hence both the Bayesian and the penalised likelihood simulation studies are setup in Section 7.4 and performance assessment is discussed in Section 7.5. Finally, results are presented in Sections 7.6 and 7.7.

7.2 Data generating mechanism

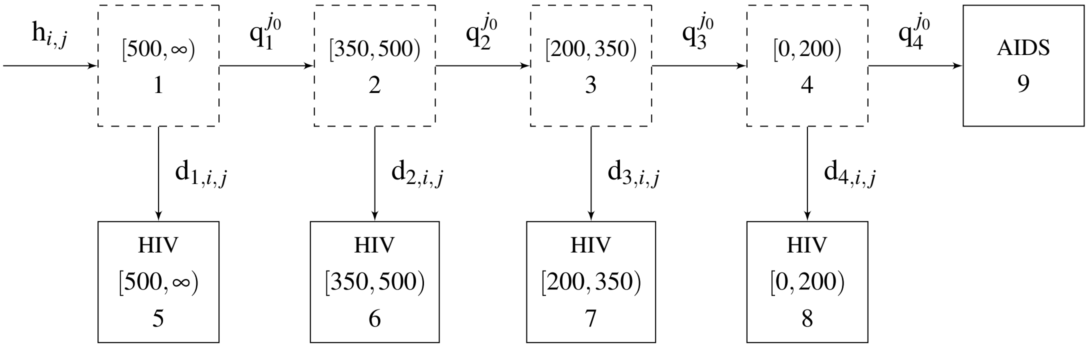
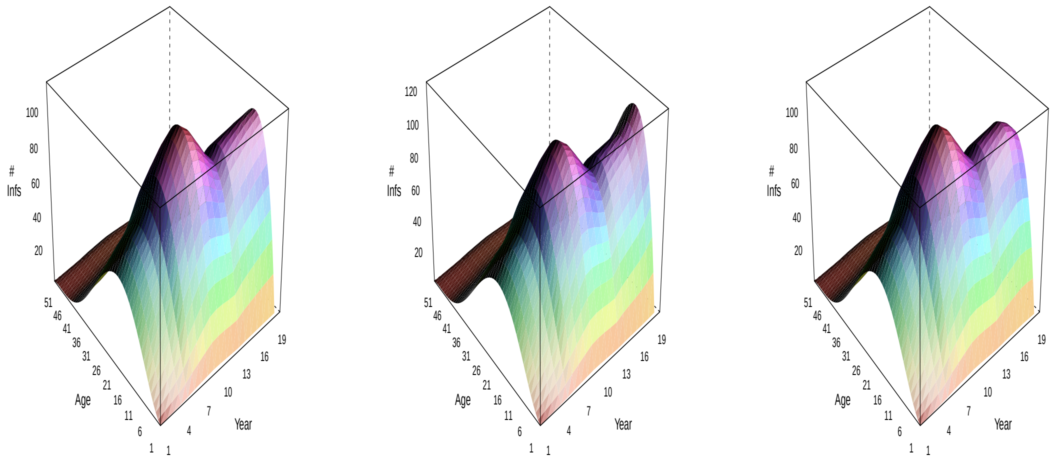


Fig. 7.1 Age-dependent back-calculation multi-state model, used for this simulation study with $K = 4$ undiagnosed states. Dashed states $\{1, \ldots, 4\}$ denote undiagnosed states. Solid states $\{5, \ldots, 9\}$ denote diagnosed states.

The age-dependent back-calculation model described in Section 5.3 is employed with $K = 4$ undiagnosed states. States 1, 2, 3 and 4 have CD4-count $[500, \infty)$, $[350, 500)$, $(200, 350]$ and $(0, 200]$ respectively. For reasons of computational feasibility, a yearly time scale (rather than a quarterly one, as in the age-independent simulations) is employed for both the time and age intervals; yearly dynamics are constructed by aggregating quarterly sub-intervals ($N_s = 4$, see Section 5.4.3). Back-calculation is implemented without including under-reporting (Section 5.4.1), from an intermediate point of the epidemic (Section 5.4.2), for 20 time intervals and 52 age intervals, *i.e.* we consider $(t_{i-1}, t_i]$ and $(a_{j-1}, a_j]$ for $i = \{1, \ldots, 20\}$, $j = \{1, \ldots, 52\}$.

The data generating mechanism is very similar to the one described for age-independent simulations (see Section 4.2 for the details): HIV, AIDS and CD4 diagnoses (denoted $Y_{i,j}^{H\star}$, $Y_{i,j}^{A\star}$ and $\boldsymbol{Y}_{i,j}^{H_C\star}$ respectively) can be generated from the model after specifying a bivariate true incidence surface, diagnosis and progression probabilities, and the expected number of initially undiagnosed infections (*i.e.* $\boldsymbol{\mathcal{H}}^\star$, $\boldsymbol{\mathcal{D}}^\star$ $\boldsymbol{\mathcal{Q}}^\star$ and $\boldsymbol{\pi}^\star$ respectively). As in Section 4.2, the values specified for these quantities are chosen to be realistic for the MSM-HIV epidemic in England and Wales between 1995 and 2015, for individuals aged between 15 and 66 years (these are denoted ages 1-52, to be consistent with the j indices referring to the age intervals).

The true bivariate incidence surface is constructed using a multiplicative model $h_{i,j} = h_i v_{i,j}$; h_i denotes the total number of expected infections in the interval $(t_{i-1}, t_i]$ (*i.e.* the time profile

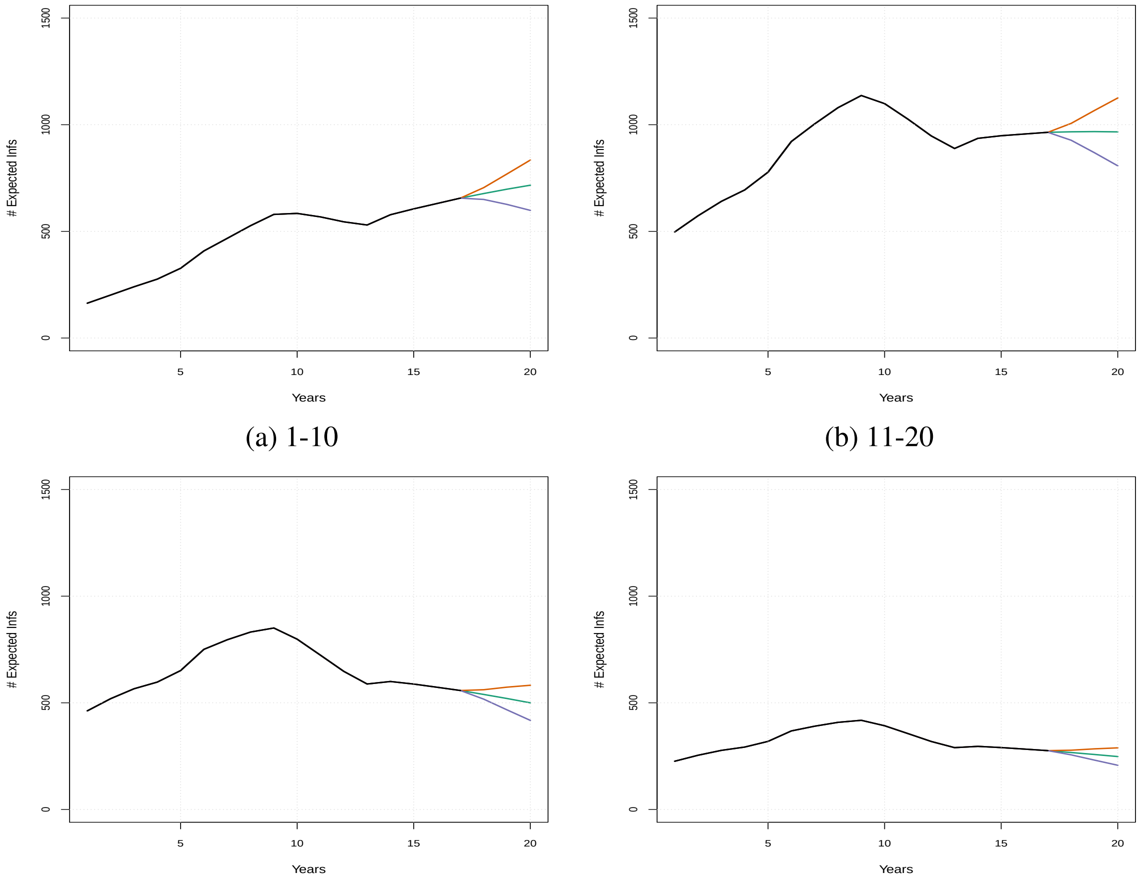


(a) Increasing

(b) Flat

(c) Decreasing

Fig. 7.2 Time and age-specific true incidence surfaces with three different time profiles.



(a) 1-10

(b) 11-20

(c) 21-30

(d) 31-52

Fig. 7.3 Time profiles of the increasing (orange), flat (green) and decreasing (gray) incidence surfaces, stratified by age-class.

of expected infections), whereas $v_{i,j}$ is the proportion of h_i in the j^{th} age interval (note that $\sum_{j=1}^{52} v_{i,j} = 1$).

The increasing, decreasing and flat time profiles considered for age independent back-calculation simulations are again employed for h_i (Figure 4.2, Section 4.2). $v_{i,j}$ are constructed so that the mean age at infection shifts linearly from age 29, in the first time interval, to 19, in the 20th time interval. $v_{i,j}$ is unchanged in each of the three time profile scenarios. The incidence surface obtained is depicted in Figure 7.2. Figure 7.3 depicts the time profile of the three incidence surfaces, stratified by age-class. Note that the flat incidence time profile is masking contrasting trends in different age-classes. This is not the case for the increasing and decreasing time profiles, where the number of expected infections in all age-classes is increasing and decreasing respectively.

For consistency with the age-independent study and to limit the computational burden of the simulation study, the age-independent diagnosis probabilities used in Section 4.2 are used again (Figure 4.3). More precisely, the values in the first quarter of each year were considered, as we are considering a yearly scale. $\boldsymbol{\mathcal{Q}}^{\star}$ and $\boldsymbol{\pi}^{\star}$ involve 52 ages and 4 states and are chosen to have realistic values for the HIV-MSM epidemic (see Appendix G.1). $\boldsymbol{\pi}^{\star}$ denotes the expected number of undiagnosed infections in 1995; these were obtained using a simple age-dependent extension of the Aalen et al. (1997) model, applied to MSM surveillance data from 1978 to 1994. We chose to model incidence with a bivariate step-function, so that incidence is constant within age-classes 1-10, 11-20, 21-30 and 31-52. $\boldsymbol{\mathcal{Q}}^{\star}$ estimates were obtained from the CASCADE cohort study (CASCADE Collaboration, 2000).

7.3 Back-calculation parameterisations

This Section discusses non-parametric models for the incidence surface $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ and diagnosis probabilities $\boldsymbol{\mathcal{D}}(\boldsymbol{\delta})$ introduced in Sections 6.4.1 and 6.4.2. We consider inference within both a Bayesian and a frequentist framework. However, in this Section we predominantly focus on the former inference methods, as the latter method was discussed in Section 6.4.

7.3.1 Incidence

The log-expected number of infections, over time and age, $\boldsymbol{\gamma}$ can be parametrised using the bivariate smoothing methods described in Chapter 6.

Splines

Various types of TPS splines can be used for this purpose (knots-based, thin plate regression splines and thin plate regression splines with shrinkage). As these have a single penalty, they can be specified within a Bayesian framework using i.i.d Normal priors (see Equation 3.3.15 with parameters now denoted β rather than β' for notational simplicity, and priors on the standard deviation rather than precision parameters). The following priors are thus considered:

$$\begin{aligned}
\beta_1 &\sim N(0,30) \\
\beta_{P_i} &\sim N(0,\sigma^2), \quad i=\{1,\dots,p\} \\
\beta_{U_i} &\sim N(0,\sigma_0^2), \quad i=\{1,\dots,u-1\} \\
\sigma &\sim t_+(2,200) \\
\sigma_0 &\sim t_+(2,200)
\end{aligned} \tag{7.3.1}$$

where $t_+(d,s)$ denotes a half-t distribution with d degrees of freedom and scale parameter s. $\sigma^2 = 1/\lambda$ and $\sigma_0^2 = 1/\lambda_0$. Let P be the number of coefficients: p and u of these are penalised and unpenalised respectively ($P = p+u$). Knots based and thin plate regression splines have $u = 3$, whereas for thin plate regression splines with shrinkage $u = 0$. Hence the latter spline does not require an additional smoothing parameter σ_0 (see Section 3.3.9). A diffuse half-t distribution with 2 degrees of freedom and with scale parameter 200 is chosen as prior for σ and σ_0, so that 95% of the prior density lies in the [0,400] region, reflecting a lack of knowledge for the β parameters. This, however, is a choice of prior to which outputs are particularly insensitive.

Tensor-product splines can also be employed to model $\boldsymbol{\gamma}$, defined by the following marginal splines: thin plate regression splines (with and without shrinkage) and first and second order B-splines (with first order and second order difference penalty respectively). Tensor product splines can be expressed in a Bayesian framework (Equation 6.2.11, now characterised by parameters $\boldsymbol{\beta}$ rather than $\boldsymbol{\beta}'$, for notational simplicity, and priors on the standard deviation rather than the precision parameters) without i.i.d Normal priors:

$$\begin{aligned}
\beta_1 &\sim N(0,30) \\
\boldsymbol{\beta}_{P-1} &\sim N_{P-1}\left(\mathbf{0}, \left(\sum_{i=1}^{2} \lambda_i \boldsymbol{S_i} + \lambda_0 \boldsymbol{S_0}\right)^{-1}\right) \\
\sigma &\sim t_+(2,200) \\
\sigma_0 &\sim t_+(2,200)
\end{aligned} \tag{7.3.2}$$

where $\boldsymbol{S}_0$ is a "small" penalty defined on the null space of $\boldsymbol{S} = \boldsymbol{S_1} + \boldsymbol{S_2}$. Recall that when thin plate regression splines with shrinkage or first order B-splines with first order difference penalty are used as marginal splines, $\boldsymbol{S_0}$ (as well as λ_0 and σ_0) is not required (see Section 6.2.7).

β_1 is the global intercept (*i.e.* it describes the mean number of log-expected infections per age and time interval) and is assigned a very weakly informative prior, so that β_1 lies with approximately 95% prior probability in the $[-60,60]$ range. The choice of priors on the parameters $\boldsymbol{\beta}_{P-1}$ are dictated by the penalty term re-interpretation as a precision matrix.

All splines considered have 80 parameters. Knots-based TPS require specification of the knots location: we specified these at intervals of two years in the time dimension, and every 6.5 years in the age dimension (*i.e.* for a total of 10 and 8 knots in the time and age dimension respectively, resulting in 80 unconstrained parameters, see Section 6.2.3). For each of the two marginal splines of a tensor product we specified 10 and 8 parameters in the time and age dimension respectively (for a total of 80 parameters, see Section 6.2.5).

Recall that TPS splines can be estimated within a maximum penalised likelihood framework, as described in Section 6.4. Also tensor product splines could theoretically be estimated within a frequentist framework, however their implementation is hindered by the large running times required.

Gaussian process

$\boldsymbol{\gamma}$ can be further modelled using a Gaussian process (GP) (Section 3.4). The covariates $\boldsymbol{x}_\iota = (x_{\iota,1}, x_{\iota,2})$, $\iota = \{1,\ldots,1040\}$ of the GP are the time and age intervals, rescaled to the

$[0,1]$ range (Flaxman et al., 2015). This is achieved using the following transformation:

$$x_{\iota,1} = \frac{t_\iota - t_1}{t_{20} - t_1} \qquad x_{\iota,2} = \frac{a_\iota - a_1}{a_{52} - a_1}$$

where 20 and 52 are the number of time and age intervals respectively. A zero-mean GP is used to model the bivariate incidence surface:

$$\boldsymbol{\gamma} \sim N_T(\mathbf{0}, \boldsymbol{\Sigma}) \tag{7.3.3}$$

The (l,m)$^{\text{th}}$ entry of the $TA \times TA$ covariance matrix $\boldsymbol{\Sigma}_{l,m}$ is defined by the squared exponential kernel:

$$\boldsymbol{\Sigma}_{l,m} = \eta^2 exp\left(-\frac{(x_{l,1} - x_{m,1})^2}{2\rho_1^2} - \frac{(x_{l,2} - x_{m,2})^2}{2\rho_2^2}\right) + \mathbb{1}_{\boldsymbol{x_l}=\boldsymbol{x_m}} 0.00001$$

where $\mathbb{1}_{\boldsymbol{x}_l=\boldsymbol{x}_m}$ is an indicator function which is equal to 1 if $\boldsymbol{x}_l = \boldsymbol{x}_m$, 0 otherwise. As for age-independent simulations (Section 4.3) a very small positive value value is added to the diagonal entries of $\boldsymbol{\Sigma}$, to ensure its positive-definiteness. The infection parameters to be estimated, $\boldsymbol{\theta} = \{\eta, \rho_1, \rho_2\}$, are restricted to be positive. η is given the prior:

$$\eta \sim N_+(4,1)$$

where $N_+(\mu, \sigma)$ denotes a truncated (at zero) Normal distribution, with mean μ and standard deviation σ. η is approximately the standard deviation of the log-expected number of age and time specific infections (Section 4.3). The weakly informative prior chosen ensures that the 90% prior range for η lies in $[2,6]$ so that the expected number of infections, over time and age, lies in $[exp(-12), exp(12)]$.

Two distinct priors are considered for the length-scales:

1. Half-t priors on the inverse length-scale: $\frac{1}{\rho_1} \sim t_+(4,1)$ and $\frac{1}{\rho_2} \sim t_+(4,1)$
2. Truncated normal priors on the length-scale: $\rho_1 \sim N_+(0.5, 0.5)$ and $\rho_2 \sim N_+(0.5, 0.5)$

The half t-priors $t_+(d,s)$, with d degrees of freedom and scale parameter s, on the inverse length-scales are weakly informative and are equivalent to the prior employed in the age-independent simulation study (Section 4.3). The truncated-normal priors are more informative, as the length-scales ρ_1 and ρ_2 are centered a priori on half of the $\boldsymbol{x}_i$ data-range.

The parameters of a Gaussian process, embedded in a back-calculation framework, can not be easily estimated in a frequentist framework (Section 6.4).

7.3.2 Diagnosis probabilities

Two candidate models are used for the diagnosis probabilities $\mathcal{D}(\boldsymbol{\theta})$. Diagnosis probabilities are modelled on a logistic scale, and are assumed to be independent of age, to be consistent with the data generating mechanism (Section 7.2). The logistic-diagnosis probabilities for state k in interval $(\check{t}_{i-1}, \check{t}_i]$ is denoted $\delta_{k,i} = log\left(\frac{d_{k,i}}{1-d_{k,i}}\right)$ $(i = \{1,\ldots,\check{T}\}$ and $k = \{1,\ldots,4\})$.

Random walk

The logistic diagnosis probabilities can be modelled independently for each of the four undiagnosed states using a yearly (*i.e.* $\check{T} = 20$) first order logistic random walk parameterisation: $\delta_{k,i} \sim N(\delta_{k,i-1}, \sigma^2_{k,D})$. The initial values $\delta_{k,1}$ values and the variance parameters are given the following priors:

$$\delta_{1,1} \sim N(-3.2, 0.2),\ \ \delta_{1,2} \sim N(-3.2, 0.2),\ \ \delta_{1,3} \sim N(-3, 0.2),\ \ \delta_{1,4} \sim N(-2.5, 0.3)$$

$$\sigma^2_{k,D} \sim \Gamma(1, 32),\ \ k = \{1,2,3,4\}$$

The above priors are equivalent to those used in the age-independent simulations (Section 4.3).

Logistic regression

Random walk models can easily be fit within a Bayesian framework, however frequentist inference is not straightforward. In a penalised likelihood back-calculation context, we consider modelling the diagnosis probabilities using a logistic regression. Specifically we employ two-year piecewise-constant time effects (*i.e.* $\check{T} = 10$), and a state effect for each of the undiagnosed states and their interaction. For $i = \{1,\ldots,10\}$ and $k = \{1,\ldots,4\}$:

$$\delta_{k,i} = \alpha + \zeta_k + \xi_i + \nu_{k,i} \tag{7.3.4}$$

Subject to the identifiability constraints: $\zeta_1 = \xi_1 = \nu_{1,i} = \nu_{k,1} = 0$.

7.4 Simulation study setup

A comparison of non-parametric methods for modelling is now pursued using the simulated datasets; both Bayesian and frequentist (penalised likelihood) inferential methods are considered.

An outline of the Bayesian simulation study is given below:

- 50 datasets are generated for each of the three true time profile of the incidence surface options (increasing, decreasing, and flat incidence) resulting in a total of 150 datasets. The term *true incidence scenario* will refer to a dataset generated under a specific true incidence option (*e.g.* increasing).
- Eight different parameterisations of the incidence surface are implemented on each dataset: Gaussian processes (*GP*), a knots-based TPS (*tpknotsloc*), a thin plate regression spline (*tp*), a thin plate regression spline with shrinkage (*ts*) and tensor product splines with four different marginal splines (thin plate regression splines (*ptenstp*), thin plate regression splines with shrinkage splines (*ptensts*), cubic B-splines with first (*ptensbsord1*) and second order (*ptensbsord2*) difference penalties). The term *incidence model* will refer to a specific parameterisation of the incidence surface (*e.g. bsord1*). 80 knots (or parameters) are used for each spline: equispaced knots are located every two years in the time dimension and every six years and a half in the age dimension.
- The term *simulation* describes the combination of a true incidence scenario (*e.g.* increasing), one incidence model (*e.g. bsord1*) and one dataset (*e.g.* dataset number 25); 1200 simulations have been undertaken.
- Inference is carried out using `Stan`, which employs HMC methods (see Appendix A.2.2 and A.4). Each simulation involves three chains of 2000 iterations with burn-in of 1000, resulting in a posterior sample size of 3000 iterations. Default initial values for the HMC algorithm are automatically generated by `Stan`. The approximate running time per simulation is 10 hours.

The penalised likelihood simulation study is described below:

- The same 150 datasets generated for the Bayesian simulation study are used.
- Only knots-based TPS (*tpknotsloc*), thin plate regression splines (*tp*) and thin plate regression spline with shrinkage (*ts*) are considered, with 80 parameters. The knots location chosen, for knots-based TPS, is as in the Bayesian simulation study.

- The diagnosis probabilities are modelled using the logistic regression parameterisation.

- The term *data scenario* describes a specific scenario involving one true incidence scenario (*e.g.* increasing) one incidence model (*e.g.* *bsord1*) and one dataset (*e.g.* dataset number 25); 450 data scenarios have been considered.

- For each *data scenario*, the optimal smoothing parameter $\hat{\lambda}$ is estimated by considering a set of 10 candidate smoothing parameter values $\lambda = \{0, 0.5, 2, 5, 8, 10, 13, 16, 20, 40\}$. The optimal smoothing parameter is chosen by AIC (Equation 6.4.8) minimisation.

- The term *simulation* describes the combination of one true incidence scenario (*e.g.* increasing), one incidence model (*e.g.* *bsord1*), one dataset (*e.g.* dataset number 25) and a plausible smoothing parameter (*e.g.* λ =2); 4500 simulations have been undertaken.

- The numerical maximization algorithm (Section 6.4.3) requires initial values specification for the infection and diagnosis parameters. The former are obtained by fitting a Bayesian spline to the true incidence surface $\mathcal{H}^{\star}$ with added noise; the initial values are chosen to be samples from the posterior distribution of $\boldsymbol{\beta}$. The initial values of the diagnosis parameters are obtained by fitting the logistic regression model in Equation 3.6.4 to the true diagnosis probabilities, and subsequently adding some noise to increase the variability of starting values.

Note that TPS splines require slightly different penalties in a frequentist and in a Bayesian framework; recall that a small penalty is imposed on the unpenalised coefficients of TPS splines to ensure proper priors (Section 3.3.9).

7.5 Simulation study assessment

The simulation study performance evaluation concepts introduced for age-independent back-calculation (Section 4.5) are here extended to age specific settings and to be applied also in a frequentist framework.

For a simulation (denoted by m), the incidence surface and the diagnosis probabilities have an estimate. In a Bayesian context this corresponds to the mean of the posterior distribution of the quantities of interest. In a frequentist framework, the incidence surface and the diagnosis probabilities estimates are obtained by first finding the parameters $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\delta}}$ maximising the penalised likelihood (Equation 6.4.3) and then plugging these estimates in the definitions of $\boldsymbol{\gamma}$ (Equation 6.4.1) and $\boldsymbol{\delta}_{k,i}$ (Equation 7.3.4). The estimates of the incidence surface and diagnosis probabilities from state k are respectively denoted as $\widehat{\boldsymbol{\mathcal{H}}}_m = \{\widehat{h}^m_{1,1}, \ldots, \widehat{h}^m_{1,A}, \ldots, \widehat{h}^m_{T,1}, \ldots \widehat{h}^m_{T,A}\}$ and $\widehat{\boldsymbol{\mathcal{D}}}_{k,m} = \{\widehat{d}^m_{k1}, \ldots, \widehat{d}^m_{kT}\}$. Credible and confidence intervals for the estimates of the incidence surface and the diagnosis probabilities can be obtained both in a Bayesian and in a frequentist framework. For the m^{th} simulation, these are denoted $\widehat{\boldsymbol{\mathcal{H}}}^{\alpha/2}_m = \{\widehat{h}^{m,\alpha/2}_{1,1}, \ldots, \widehat{h}^{m,\alpha/2}_{T,A}\}$ and $\widehat{\boldsymbol{\mathcal{D}}}^{m,\alpha/2}_k = \{\widehat{d}^{m,\alpha/2}_{k,1}, \ldots, \widehat{d}^{m,\alpha/2}_{k,T}\}$. Finally, $\boldsymbol{\mathcal{H}}^{\star}{}_m = \{h^{\star}_{1,1}, \ldots, h^{\star}_{1,A}, \ldots, h^{\star}{}_{T,1}, \ldots h^{\star}_{T,A}\}$ and $\boldsymbol{\mathcal{D}}^{\star}_{k,m} = \{d^{\star}_{k1}, \ldots, d^{\star}_{kT}\}$ denote the true incidence surface and diagnosis probabilities from state k.

PMSE and MPMSE defined in Section 4.5, are useful summary statistics for performance assessment. For the m^{th} simulation, the infections PMSE is:

$$PMSE(\widehat{\boldsymbol{\mathcal{H}}}_m) = \frac{1}{TA} \sum_{i=1}^{T} \sum_{j=1}^{A} \left(\widehat{h}^m_{i,j} - h^{\star}_{i,j} \right)^2 \tag{7.5.1}$$

The mean-PMSE, which is the PMSE averaged across simulations, is obtained as:

$$MPMSE(\widehat{\boldsymbol{\mathcal{H}}}) = \frac{1}{M} \sum_{m=1}^{M} PMSE(\widehat{\boldsymbol{\mathcal{H}}}_m) \tag{7.5.2}$$

$\alpha\%$-Coverage and mean $\alpha\%$-Coverage for the incidence surface are obtained as follows:

$$Covg_{\alpha}(\widehat{\boldsymbol{\mathcal{H}}}_m) = \frac{1}{TA} \sum_{i=1}^{T} \sum_{j=1}^{A} \mathbb{1}_{h^{\star}_{i,j} \in \left[\widehat{h}^{m,\alpha/2}_{i,j}, \widehat{h}^{m,1-\alpha/2}_{i,j}\right]} \tag{7.5.3}$$

$$MCovg_{\alpha}(\widehat{\boldsymbol{\mathcal{H}}}) = \frac{1}{M} \sum_{m=1}^{M} Covg_{\alpha}(\widehat{\boldsymbol{\mathcal{H}}}_m) \tag{7.5.4}$$

As diagnosis probabilities are modelled independently of age (Section 7.3), performance assessment for diagnosis probabilities performance proceeds as in the age-independent simulation study, using the formulae defined in Section 4.5 for $PMSE(\widehat{\boldsymbol{\mathcal{D}}}_{k,m})$, $MPMSE(\widehat{\boldsymbol{\mathcal{D}}}_k)$, $Covg_{\alpha}(\widehat{\boldsymbol{\mathcal{D}}}_{k,m})$ and $MCovg_{\alpha}(\widehat{\boldsymbol{\mathcal{D}}}_k)$.

7.6 Bayesian simulation study results

This Section considers the Bayesian simulation study: convergence assessment is undertaken in Section 7.6.1 and results are subsequently discussed in Section 7.6.2 and plotted in Section 7.6.3.

7.6.1 Convergence assessment

We start by assessing whether simulations have converged, following the same principles as for the age-independent simulations (Section 4.6.1). Recall that a simulation is considered not to have converged if $\hat{R} > 1.05$ for at least one parameter. Note that this is a stringent criterion, as it is possible that when only a few of the parameters have $\hat{R} > 1.05$, the overall mixing is still satisfactory.

Non-convergent simulations, stratified by true incidence scenario and incidence model, are reported in Table 7.1. Convergence is achieved for all splines, with few exceptions. We further inspected the trace plots of the parameters of splines that did not satisfy the set convergence criterion; these still appear to mix satisfactorily. Nevertheless, to avoid any ambiguity in interpretation, we discarded all simulations that did not achieve convergence.

GP on the other hand exhibit convergence issues: *GP* with Normal priors converge in approximately 75% of the simulations and hardly any *GP* with t-priors converge. Consider, for example, the trace plots of the hyperparameters $\{1/\rho_1, 1/\rho_2, \eta\}$ of a *GP* for datasets number one and two, generated under the increasing true incidence scenario. In dataset two convergence is achieved using Normal priors (Figure 7.5), but not t-priors (Figure 7.4), while in dataset one, convergence is not reached irrespective of the prior chosen (Figures 7.6 and 7.7). It is well-known that estimating the parameters of a latent *GP* is non-trivial

%	*tp*	*tp*	*ts*	*ptens*	*ptens*	*ptens*	*ptens*	*GP*	*GP*
	knotsloc			*tp*	*ts*	*bsord1*	*bsord2*	*N-prior* ρ	*t-prior* $1/\rho$
Increasing	0	0	2	0	0	0	0	26	92
Flat	0	0	4	0	0	0	0	14	94
Decreasing	0	0	2	2	0	0	0	22	84

Table 7.1 Percentage of simulations that have not converged by true incidence scenarios and models.

%	*tp*	*tp*	*ts*	*ptens*	*ptens*	*ptens*	*ptens*	*GP*	*GP*
	knotsloc			*tp*	*ts*	*bsord1*	*bsord2*	*N-prior* ρ	*t-prior* $1/\rho$
Increasing	4	0	0	0	0	0	0	0	0
Flat	8	4	0	0	0	0	0	0	0
Decreasing	0	6	2	0	0	0	0	0	0

Table 7.2 Percentage of simulations that have at least one divergent transitions by true incidence scenarios and models, after removing simulations that have not converged.

(Flaxman et al., 2015). Informative priors could be used to improve identifiability and thus mixing. However, informative priors are challenging to specify when little, or no, prior information on the shape of the incidence surface is available; in such cases understanding the impact of the choice of prior on the results is crucial. From now onwards, the term Gaussian Process or *GP* will only refer to GP with Normal priors, as t-priors hardly ever achieve convergence and are therefore no longer considered.

Convergence assessment alone is not sufficient; divergent transitions must be analysed in order to detect unwanted behaviour of the HMC algorithm. Table 7.2 shows the number of simulations which have divergent transitions, after the exclusion of non-convergent simulations. Only simulations with *tp* and *tpknotsloc* have at least one divergent transition. As for age-independent simulations, divergent transitions are caused by the extra smoothing parameter λ_0 imposed on the null space of splines.

Scatter plots of posterior λ_0 values against log-posterior values for *tpknotsloc* and *tp* splines are depicted in Figures 7.8 and 7.9 respectively. Figures 7.8a and 7.9a correspond to datasets without divergent transitions, where all posterior λ_0 values are concentrated around zero. In contrast, Figures 7.8b and 7.9b have a divergent transition (red dot). Even though most λ_0 posterior values are also concentrated around zero, λ_0 occasionally wander off to high values, where divergent transitions may occur; this might suggest that divergent transitions occur when HMC chains are exploring the posterior distribution tails, however analysis of trace-plots for λ_0 (not displayed) demonstrate that chains do not "get stuck" in the posterior distribution tails, hence divergent transitions may simply be due to numerical error. Nonetheless, following a zero-tolerance policy for divergent transitions (Stan Development Team, 2016b), all simulations with divergent transitions were discarded.

In theory, a number of actions could be taken to avoid divergent transitions: increasing the HMC resolution, setting a tighter prior on λ_0, or using the non-centred parameterisation. However, in practice, the non-centred parameterisation leads to substantially more divergent iterations, and the HMC algorithm is already considered with the highest possible resolution. Finally a more informative prior than the current weakly informative half t-prior should force λ_0 away from zero; to achieve this, a prior pushing σ_0 (i.e $1/\sqrt{\lambda_0}$) far from 0 must be chosen. However, quantifying a priori the smoothness of a spline can be difficult.

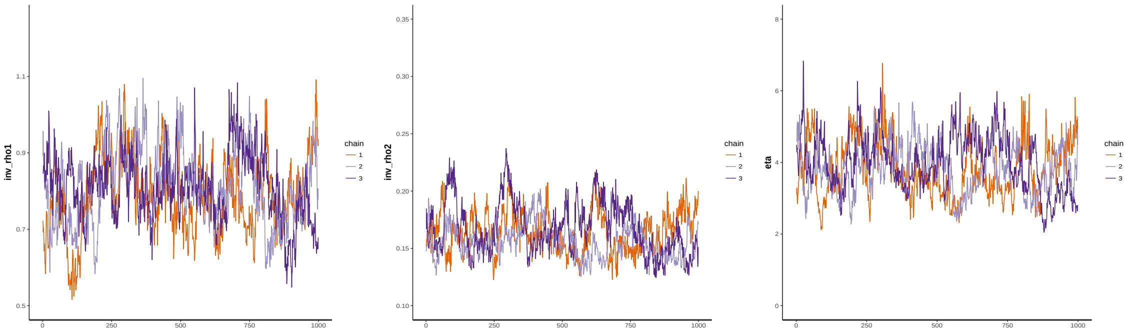


Fig. 7.4 Trace plots for $1/\rho_1$ (left), $1/\rho_2$ (center) and η (right) for GP using t-priors and dataset number 1, increasing incidence.

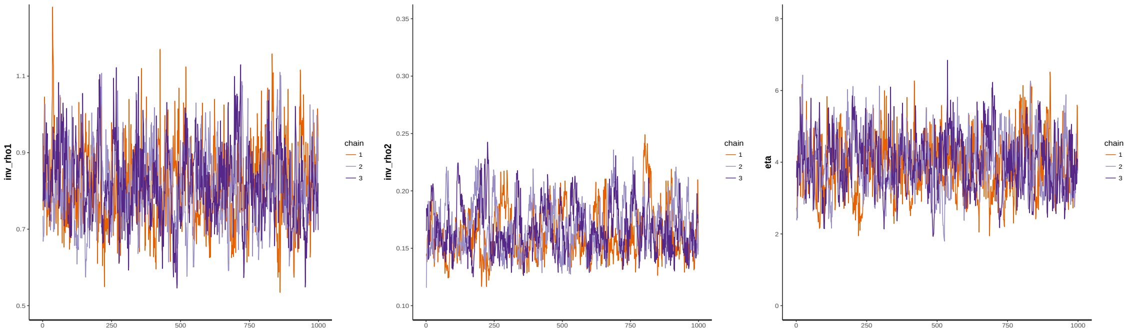


Fig. 7.5 Trace plots for $1/\rho_1$ (left), $1/\rho_2$ (center) and η (right) for GP using Normal priors and dataset number 1, increasing incidence.

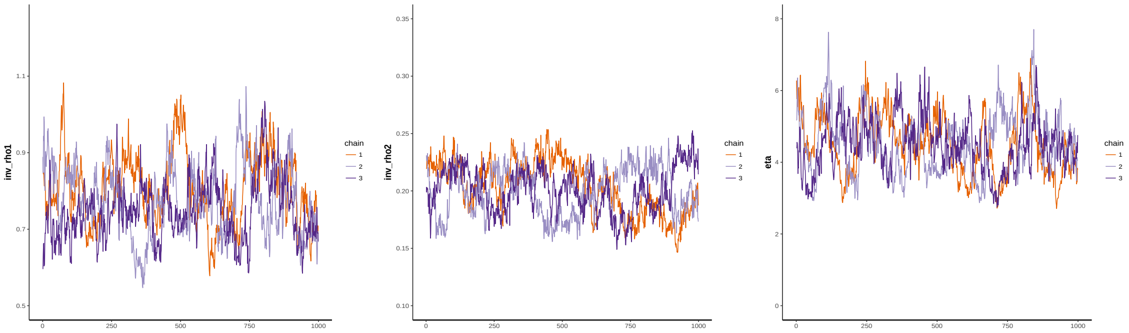


Fig. 7.6 Trace plots for $1/\rho_1$ (left), $1/\rho_2$ (center) and η (right) for GP using t-priors and dataset number 2, increasing incidence.

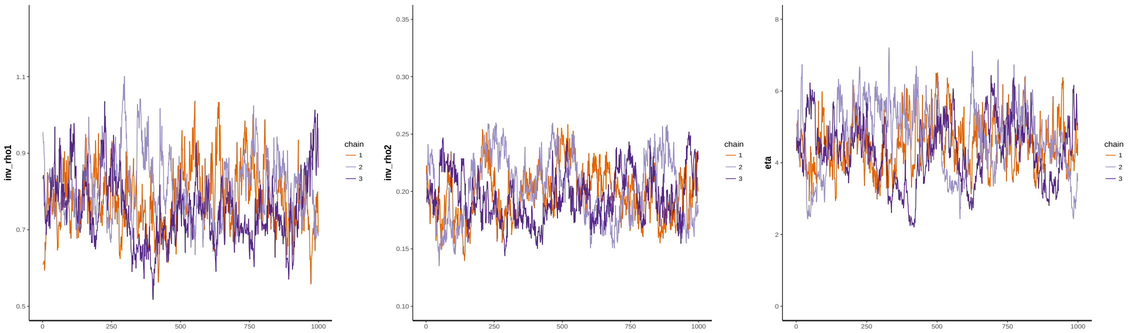
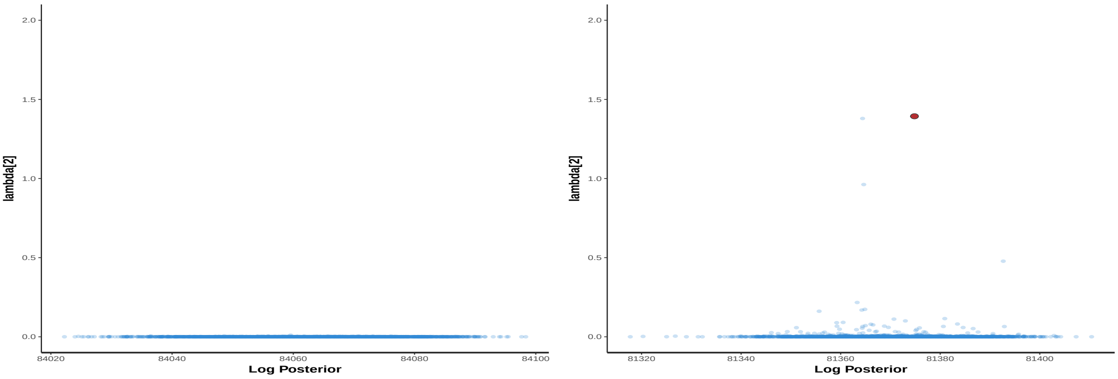


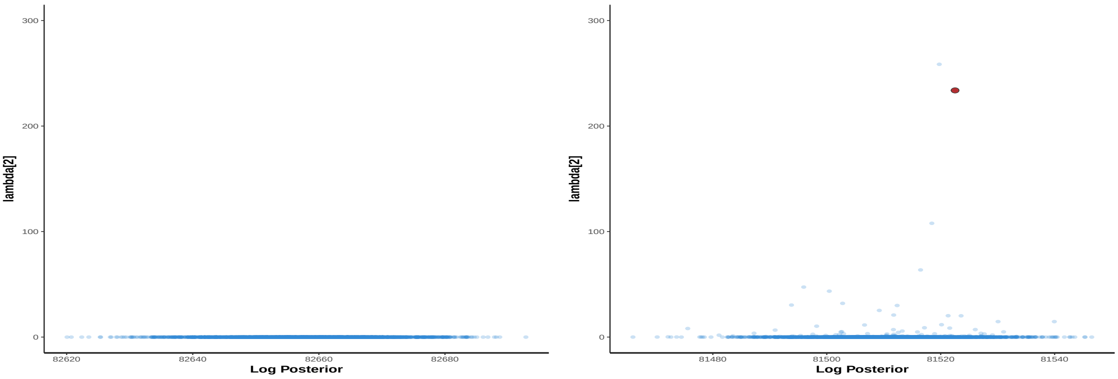
Fig. 7.7 Trace plots for $1/\rho_1$ (left), $1/\rho_2$ (center) and η (right) for GP with Normal priors, dataset number 2, increasing incidence.



(a) No divergent transitions

(b) Divergent transitions

Fig. 7.8 Scatter plot of posterior λ_0 values against log-posterior values for *tpknotsloc* splines in datasets 6 (left) and 8 (right) in the increasing incidence scenario.



(a) No divergent transitions

(b) Divergent transitions

Fig. 7.9 Scatter plot of posterior λ_0 values against log-posterior values for *tp* splines for datasets 5 (left) and 4 (right) in the increasing incidence scenario.

7.6.2 Comments on the results of the simulation study

The results of the simulation study under the three true incidence scenarios (increasing, flat and decreasing) and the eight incidence models (*tpknotsloc*, *tp*, *ts*, *ptenstp*, *ptensts*, *ptensbsord1*, *ptensbsord2 GP*) are displayed in Section 7.6.3. Specifically:

1. Figures 7.10, 7.13, 7.16, 7.19, 7.22, 7.25, 7.28 and 7.31 depict the posterior distribution of the time profile of the incidence surface, for all datasets considered.
2. Figures 7.12, 7.15, 7.18, 7.21, 7.24, 7.27, 7.30 and 7.33 depict the posterior distribution of the time profile of the incidence surface, by age-classes 1-10, 11-20, 21-30, 30-52 for all datasets considered.

3. Figures 7.11, 7.14, 7.17, 7.20, 7.23, 7.26, 7.29, 7.32 plot the posterior distribution of the diagnosis probabilities from undiagnosed state 1 (*i.e.* $\widehat{\mathcal{D}}_{1,m}$) for all datasets considered.

4. Figures 7.34 to 7.37 depict $PMSE(\widehat{\mathcal{H}}_m)$ and $PMSE(\widehat{\mathcal{D}}_{k,m})$ distribution (over the datasets) for the full time-scale considered and for the last three years only.

5. Figures 7.38 to 7.41 show the $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ and $Covg_{0.95}(\widehat{\mathcal{D}}_{1,m})$ distribution (over the datasets) for the full time-scale considered and for the last three years only.

All incidence models, apart from *GP*, appropriately reconstruct the time profile of the true incidence for the three incidence scenarios. Taking a closer look, as for age-independent simulations, the true time profiles of the incidence surface are accurately estimated in all but the first, and last, three years of the epidemic. In the early epidemic stages, diagnosis data may be explained by either initially undiagnosed infections or newly infected individuals being diagnosed; incorrectly attributing diagnoses to either group, leads to bias in the incidence and diagnosis probabilities estimates. The time-profiles estimated are poor, for some datasets, in the early years with no consistent under or over estimation noted. As only incidence estimates in most recent years are of interest for public health purposes, these subpar estimates in early years are not a major concern.

On the other hand, the validity of incidence estimates in recent years is crucial. As for the age-independent case, bias is introduced: in the flat and decreasing true incidence scenarios, the time profiles of the true incidence surface is overestimated by most incidence models. This is more pronounced for the decreasing true incidence scenario, where the negative trend in the most recent years is hardly captured by any incidence model. On the positive side, the credible intervals contain the true time profile of the incidence surface in almost 90% of the datasets in the last 3 years (see Figure 7.40).

The overestimation of incidence in the latest years is induced by incorrectly attributing recent diagnoses, in recent years, to increased incidence rather than increased diagnosis probabilities. In fact, the diagnosis probabilities from state 1 (*i.e.* concerning recent infections) are underestimated in most recent years for all true incidence scenarios and incidence models.

On a positive note, the age-specific time profile of the true incidences are adequately estimated for all age-classes and incidence models, except for *GP*s. Note that age-specific time profile estimates have the same issues as the respective population-level estimates: estimates are inaccurate in earliest and latest years, and tend to be biased upwards. Overall, incidence estimates are fairly accurate, even in the latest years, in the 1-10 and 11-20 age-classes, whilst estimates in the 21-30 and 30+ age-classes are more volatile and imprecise. This may be due to fewer diagnosis occurring in the older age-groups, leading to less precise estimates.

*GP*s deserve a special mention as they are the only incidence model yielding unsatisfactory, overly-smooth, incidence estimates, both at population and at age-specific level. Moreover, incidence is under and over estimated in the youngest (1-10) and oldest (31-52) age-classes respectively. Recall that the *GP* length scale parameters determine its smoothness; incorrect specification of informative priors (Normal-priors) may lead to excessive smoothness, and thus poor estimates. Other kernels (*e.g.* Matérn) may be more appropriate than the squared exponential one, as the latter is known to produce overly smooth estimates.

We further compared the performance of different splines in order to establish whether there is a most suitable model for incidence. Performance is assessed using the $PMSE(\widehat{\mathcal{H}}_m)$ and $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ distributions, depicted in Figures 7.34 and 7.38 respectively. Four spline types, *tp*, *ts*, *ptensbsord1* and *ptensbsord2*, outperform the others; they have smaller $PMSE(\widehat{\mathcal{H}}_m)$ values, and coverage ($Covg_{0.95}(\widehat{\mathcal{H}}_m)$) closer to the nominal 0.95 level. The first two splines belong to the thin plate spline family, whereas the others are tensor product splines. Among thin plate splines, *tp* and *ts* outperform *tpknots*, in agreement with Wood (2003). *ts* splines have a similar coverage than *tp*, but are not associated with simulations with divergent transitions (see Table 7.2), and allow shrinkage towards zero (see Section 6.2.6). Among tensor product splines, *ptensbsord1*, *ptensbsord2* (*i.e.* marginal cubic b-splines, with first and second order penalty respectively) outperform *ptenstp* and *ptensts* (marginal cubic splines). Furthermore *ptensbsord1* is always superior to *ptensbsord2* in terms of coverage.

Hence, the simulation results suggest that *ts* and *ptensbsord1* are the two most appropriate choices for estimating the incidence surface; their $PMSE(\widehat{\mathcal{H}}_m)$ values are comparable, whereas $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ values are closer to the nominal level for *ts*. The incidence time profiles estimated by *ts* and *ptensbsord1* (Figures 7.16 and 7.25 respectively) are very similar, apart from the latest years. Differences can be explained from a variance-bias perspective; *ts* splines are less biased (*i.e.* more accurately reconstruct the true underlying trend, especially for the true decreasing scenario) but more volatile (resulting in poor estimates for some datasets). This contrasting behaviour is attributable to the prior having a considerable impact on the incidence estimates in most recent years, where data are weakly informative. *ptensbsord1* tends to a priori flatten the incidence time profile (Section 3.3.7), whereas *ts* splines favour surfaces with second derivatives equal to 0, resulting in a time profile extrapolating linearly in most recent years. Linear extrapolation often captures better the true incidence, but occasionally results in unrealistically high estimates.

Nevertheless, *ptensbsord1* are preferred over *ts* for their theoretical properties. *ts* are isotropic (Section 6.2.6), which is only desirable if there are valid reasons to assume equal smoothing in the two dimensions; this appears to be the case in this simulation study, as time and age intervals are both considered yearly. However, it is easy to imagine an example where data are collected on an uneven time and age scale. As the isotropy assumption is hardly testable, or justifiable in practice, we recommend using tensor product splines, despite thin plate splines slightly outperform tensor product splines in terms of $MPMSE(\widehat{\boldsymbol{\mathcal{H}}})$ in this study.

An alternative way of determining which incidence models are most suitable to estimate incidence is by examining their fit to simulated data. However, in our case, this was uninformative as all incidence models fit the simulated data (for the three true incidence scenarios) equally well. Goodness of fit plots are available in Appendix G.2.2.

Appendix G.2.1 includes further details regarding the diagnosis probabilities; this Section only considered diagnosis probabilities from state 1, as these characterise recent infections and are poorly identified in most recent years. Diagnosis probabilities from states 2, 3 and 4 are instead typically accurately estimated, as they affect individuals with long standing infections and are not affected by shifts in incidence or diagnosis probabilities from state 1 in recent years.

To sum up, the proposed age-specific back-calculation model adequately estimates the true incidence surface and diagnosis probabilities. *GP* are not recommended for modelling bivariate incidence, whereas all bivariate splines, and in particular *ts* and *ptensbsord1*, accurately reconstruct the true incidence; *ptensbsord1* are recommended as they do not assume isotropy. Finally estimates from the latest years must be interpreted with caution as, identifiability problems often introduce bias.

7.6.3 Plots of results from simulation study

Results for the tpknotsloc spline incidence model

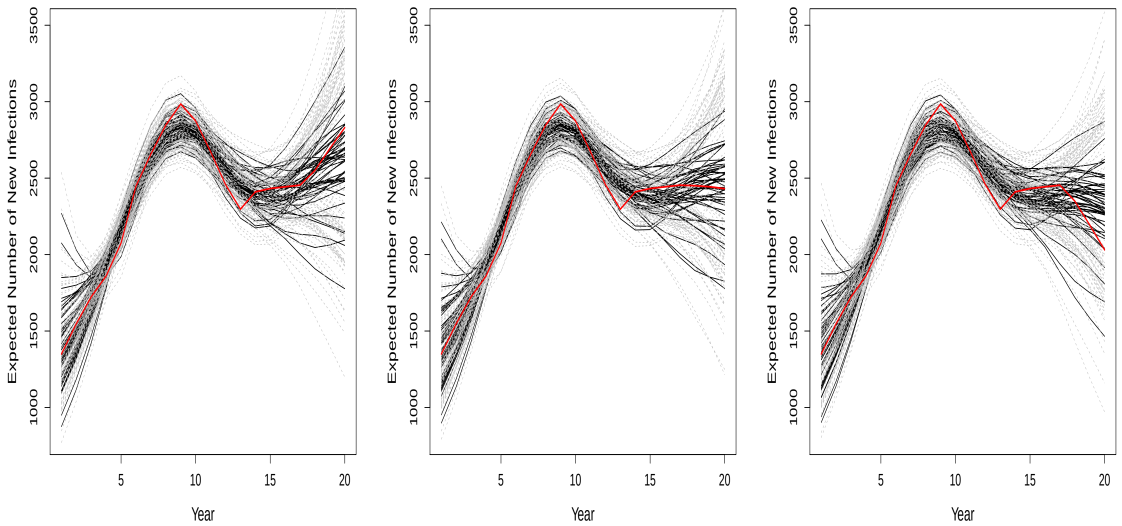


Fig. 7.10 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *tpknotsloc* spline to model incidence. Grey lines denote the respective 95% credible intervals.

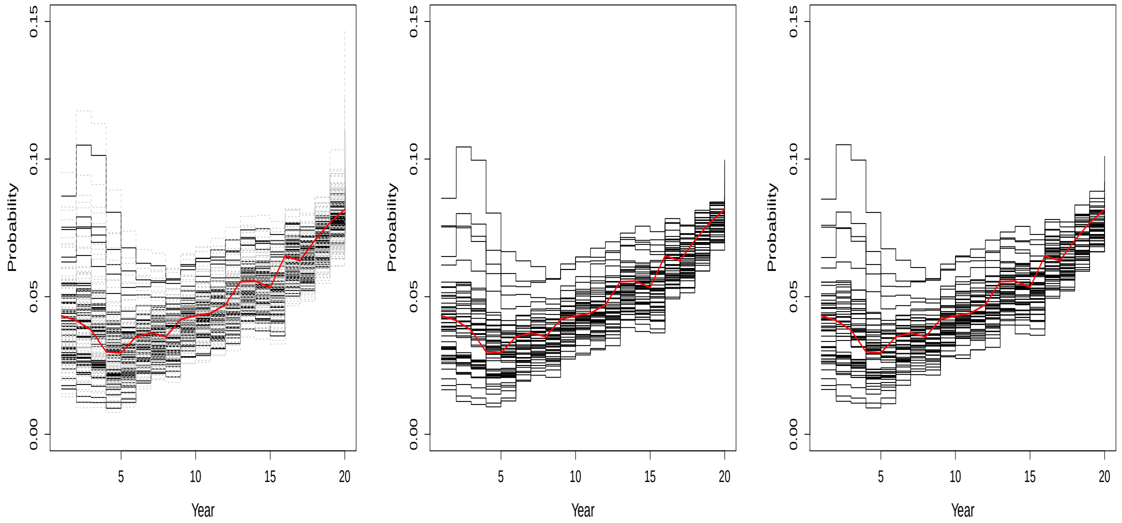
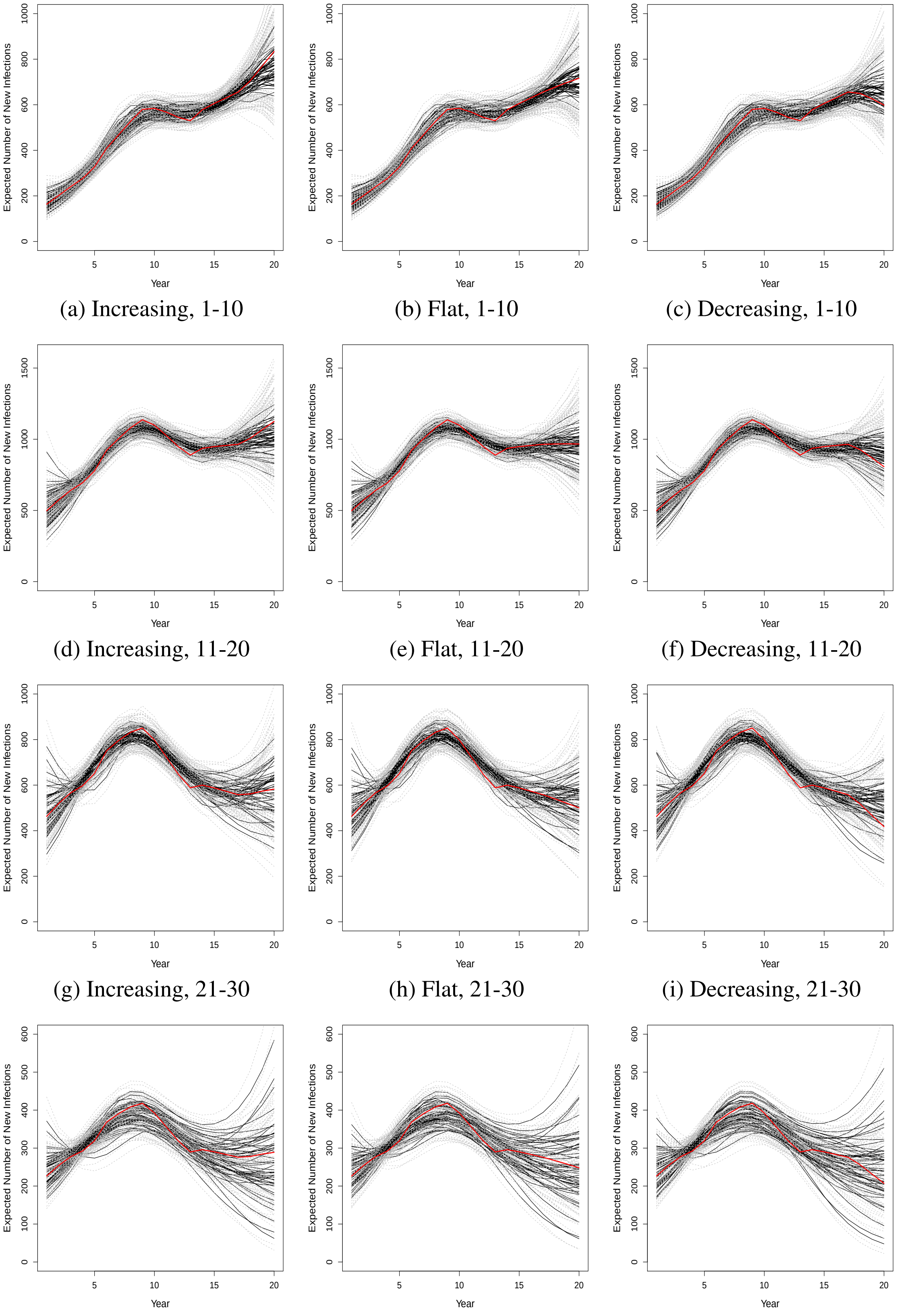


Fig. 7.11 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are only depicted on the left figure (in grey) to demonstrate they overlap with the estimates, rendering the plot hard to interpret.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.12 Estimates (posterior means) of the time profile of the incidence surface, using a *tpknotsloc* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the tp spline incidence model

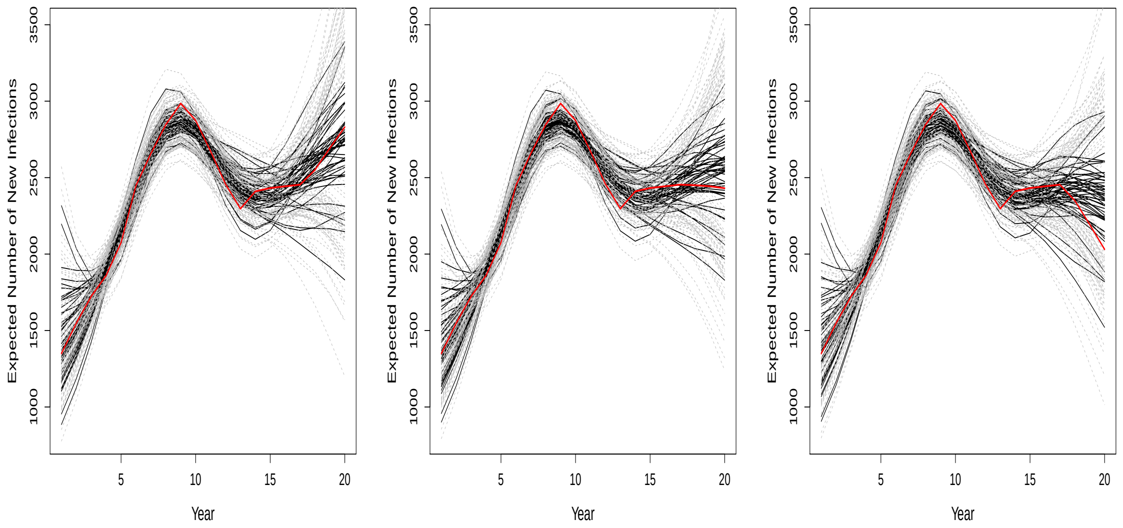


Fig. 7.13 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *tp* spline to model incidence. Grey lines denote the respective 95% credible intervals.

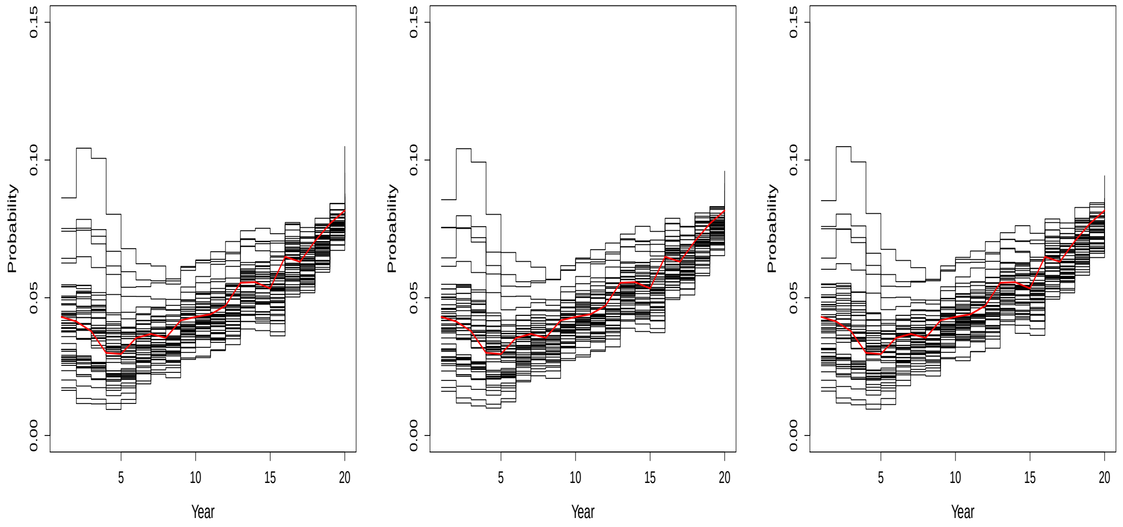
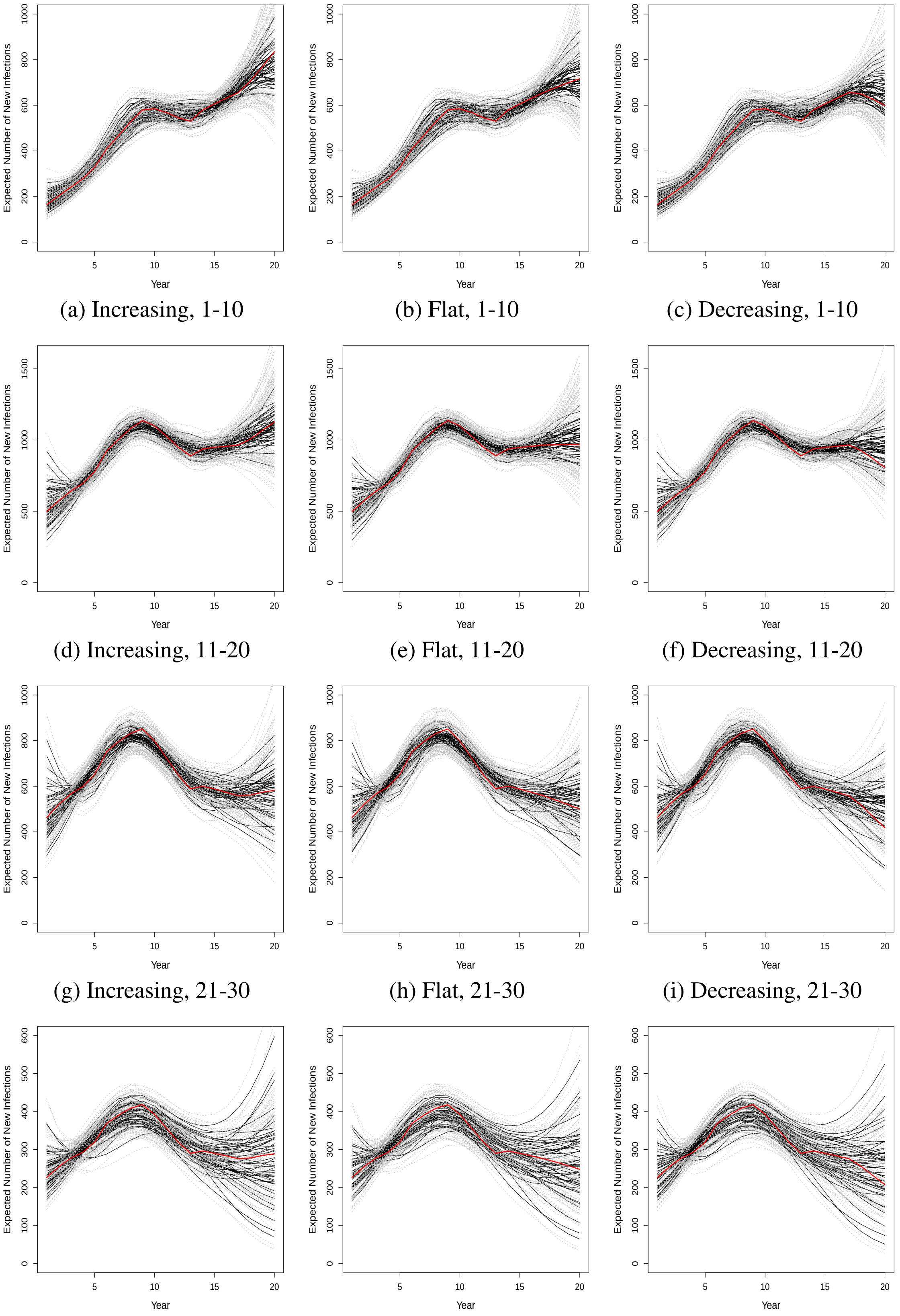


Fig. 7.14 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.15 Estimates (posterior means) of the time profile of the incidence surface, using a *tp* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the ts spline incidence model

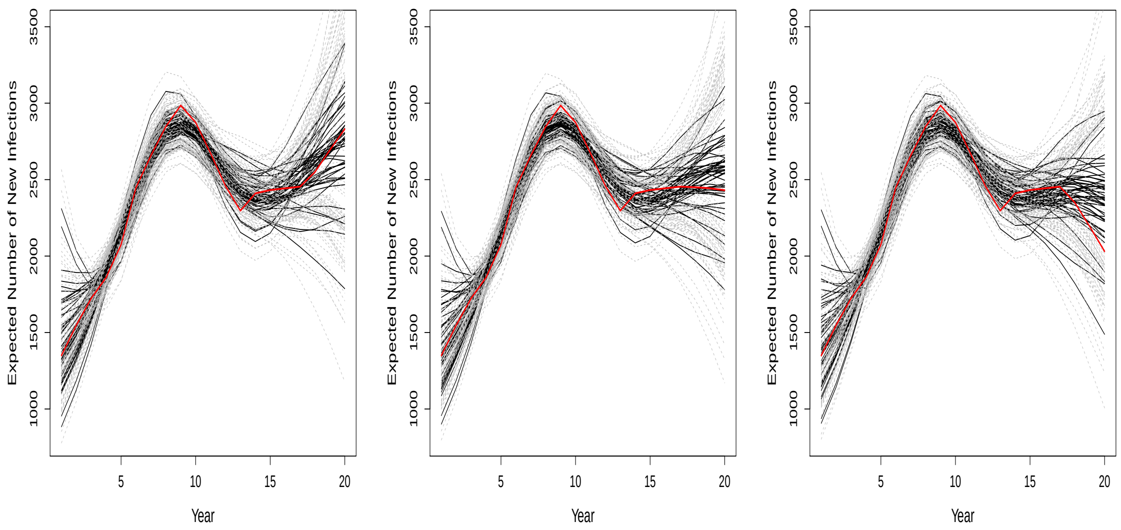


Fig. 7.16 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *ts* spline to model incidence. Grey lines denote the respective 95% credible intervals.

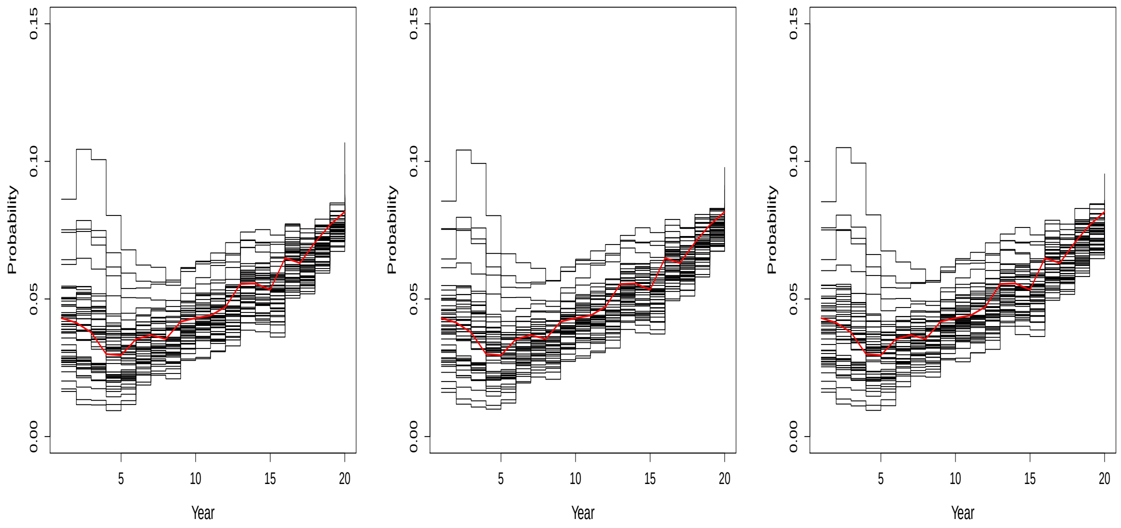
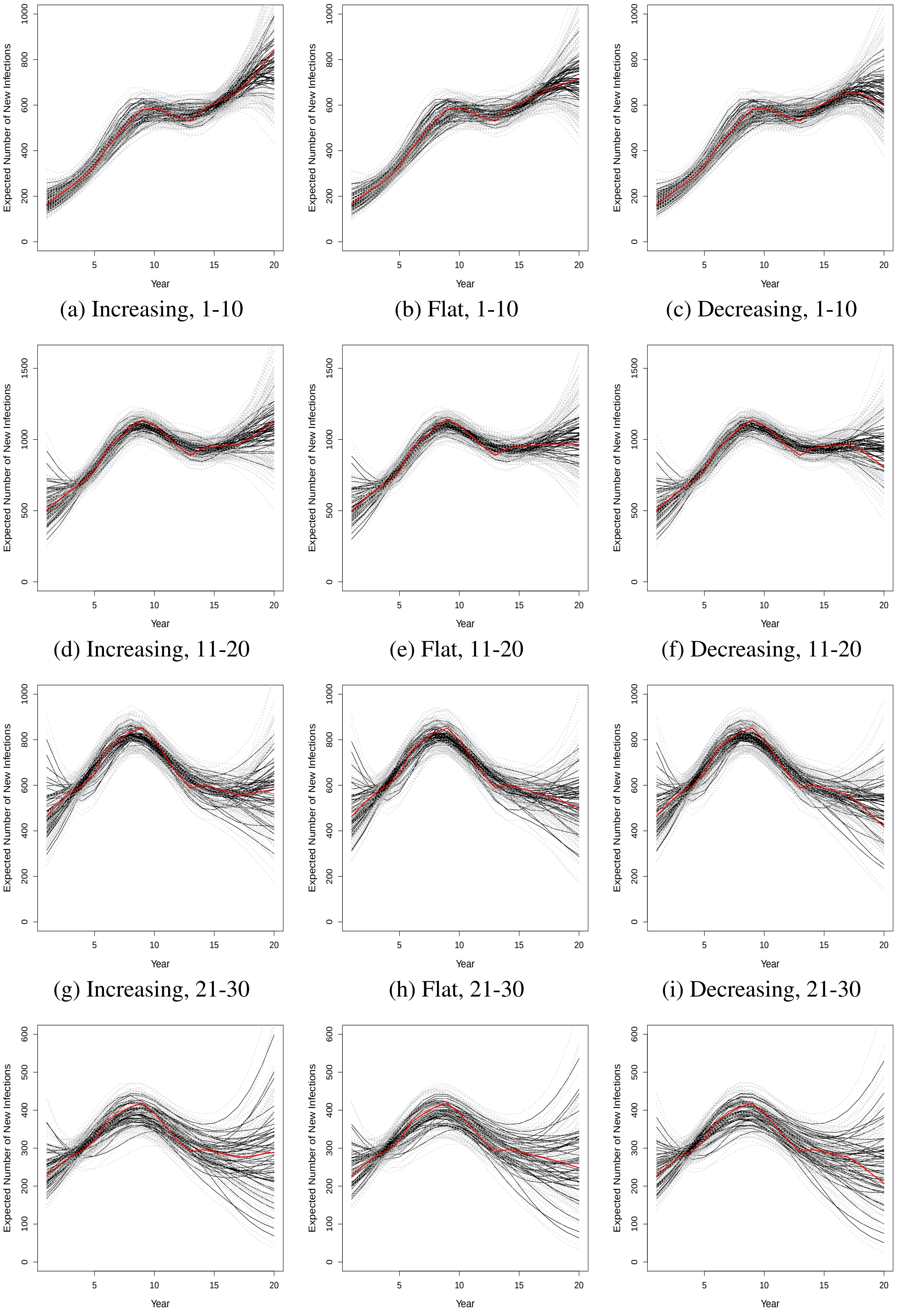


Fig. 7.17 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.18 Estimates (posterior means) of the time profile of the incidence surface, using a *ts* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the ptenstp spline incidence model

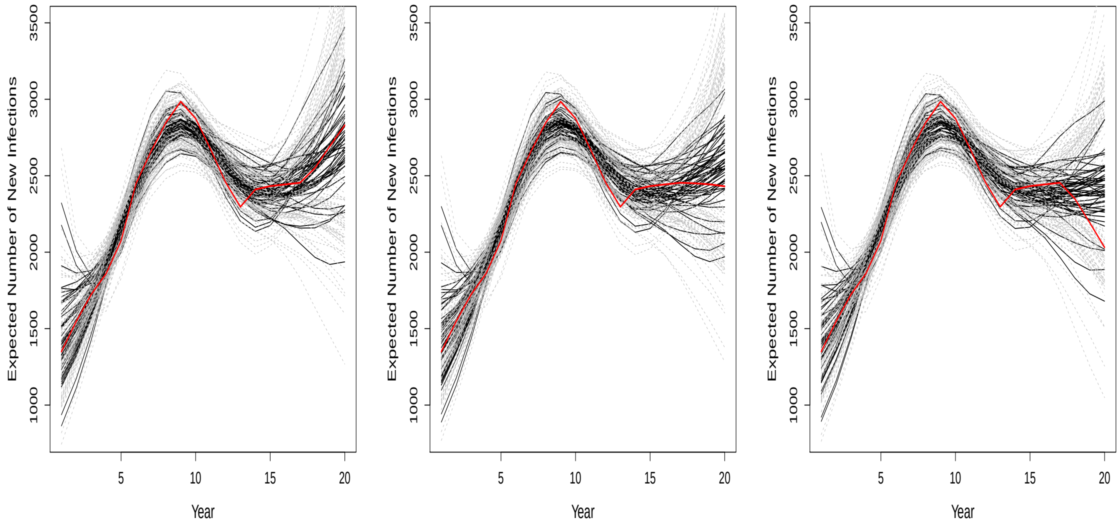


Fig. 7.19 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *ptenstp* spline to model incidence. Grey lines denote the respective 95% credible intervals.

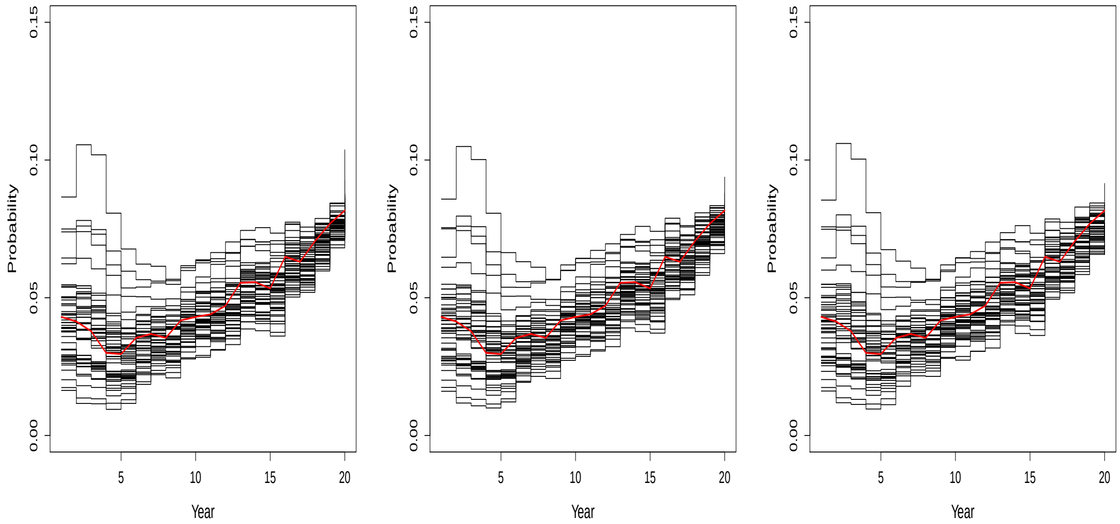
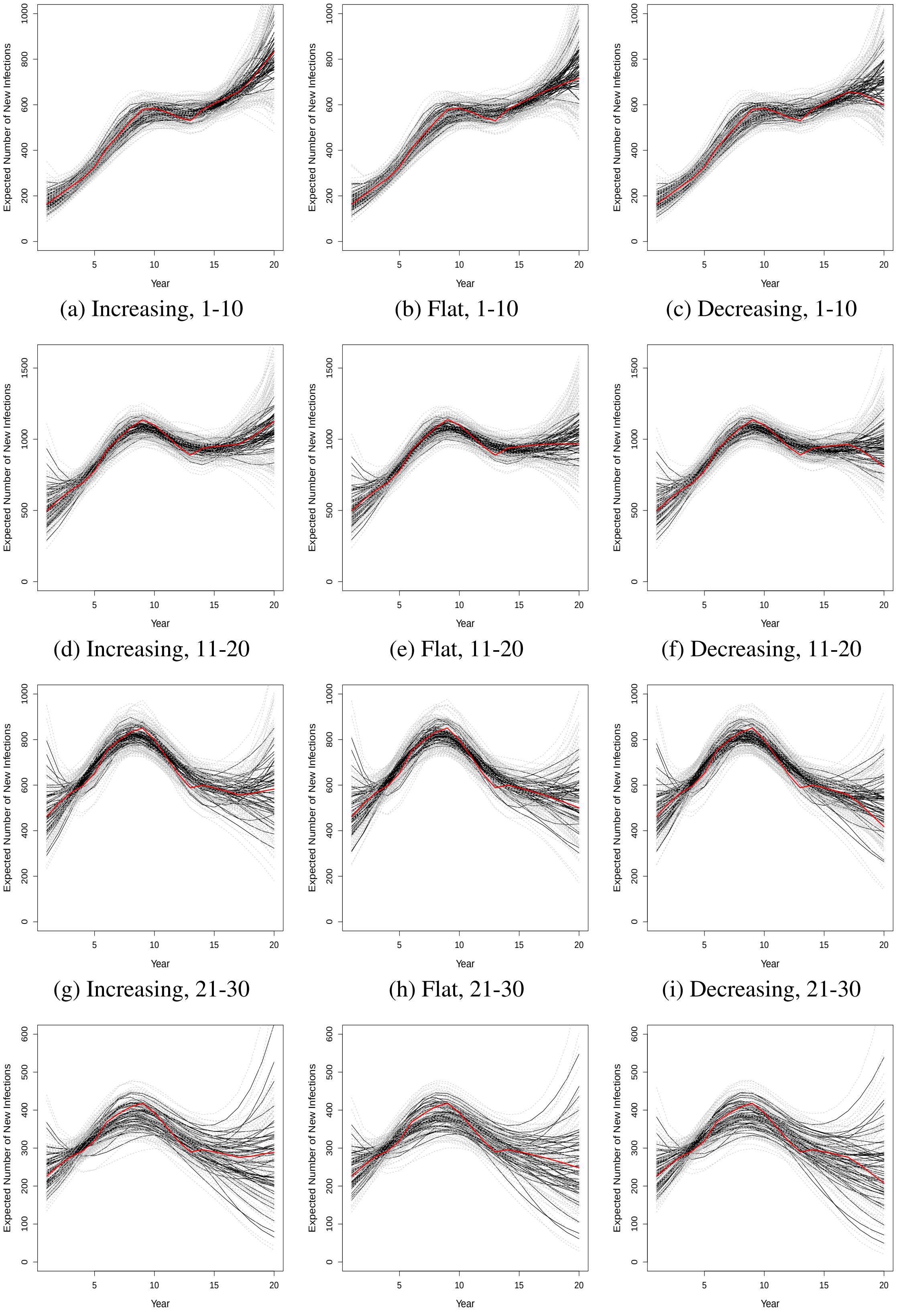


Fig. 7.20 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.21 Estimates (posterior means) of the time profile of the incidence surface, using a *ptenstp* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the ptensts spline incidence model

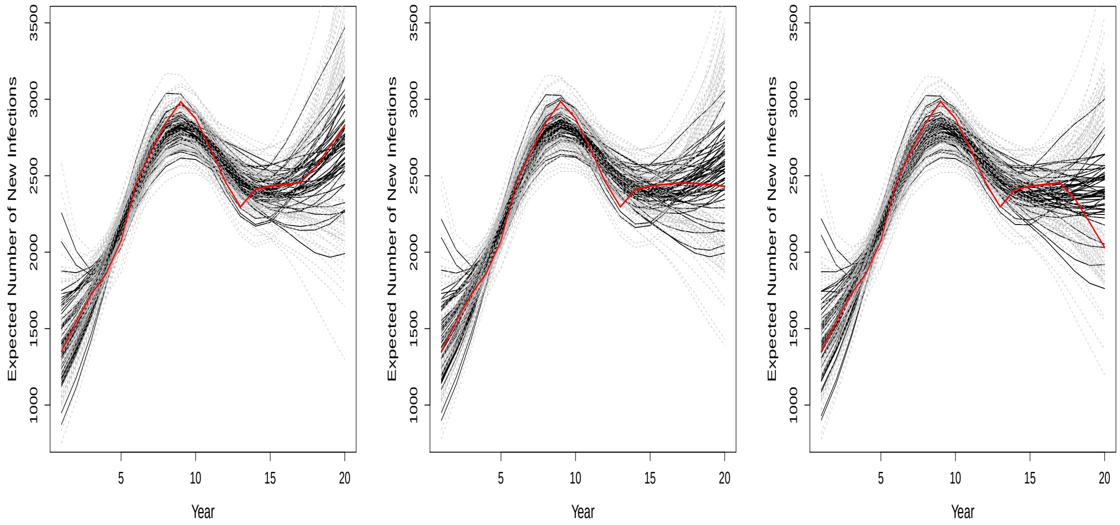


Fig. 7.22 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *ptensts* spline to model incidence. Grey lines denote the respective 95% credible intervals.

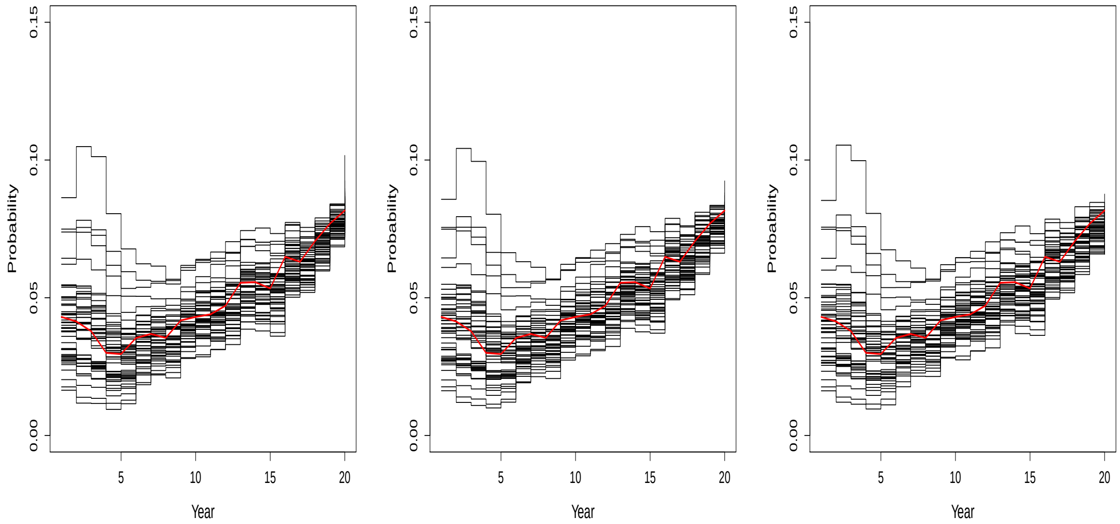
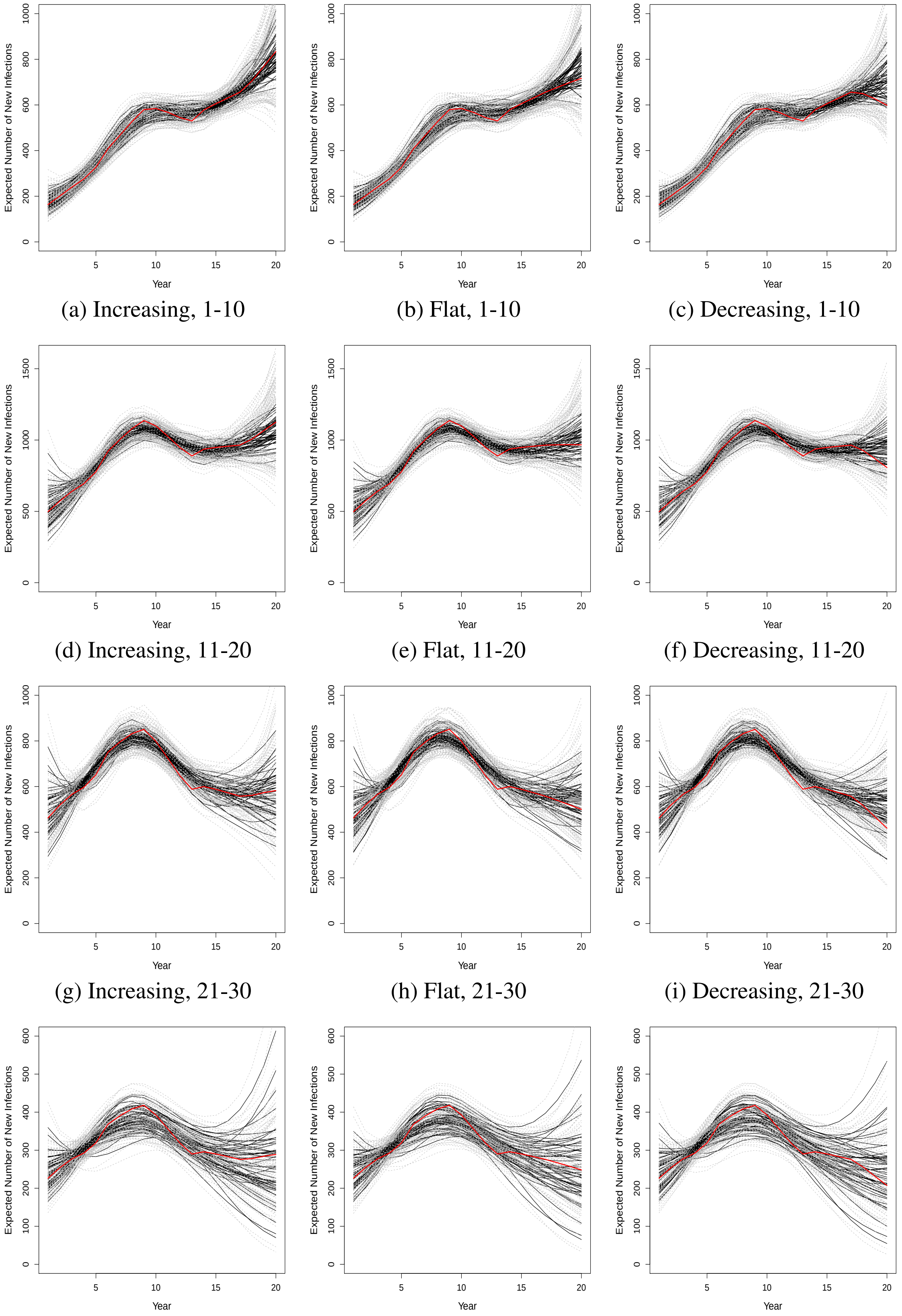


Fig. 7.23 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10 (b) Flat, 1-10 (c) Decreasing, 1-10

(d) Increasing, 11-20 (e) Flat, 11-20 (f) Decreasing, 11-20

(g) Increasing, 21-30 (h) Flat, 21-30 (i) Decreasing, 21-30

(j) Increasing, 31+ (k) Flat, 31+ (l) Decreasing, 31+

Fig. 7.24 Estimates (posterior means) of the time profile of the incidence surface, using a *ptensts* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the ptensbsord1 spline incidence model

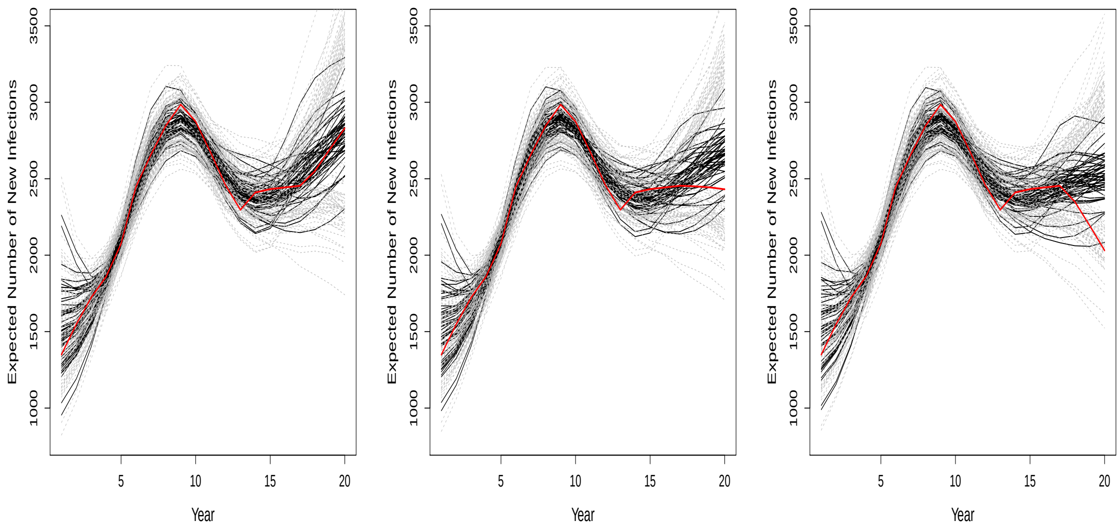


Fig. 7.25 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *ptensbsord1* spline to model incidence. Grey lines denote the respective 95% credible intervals.

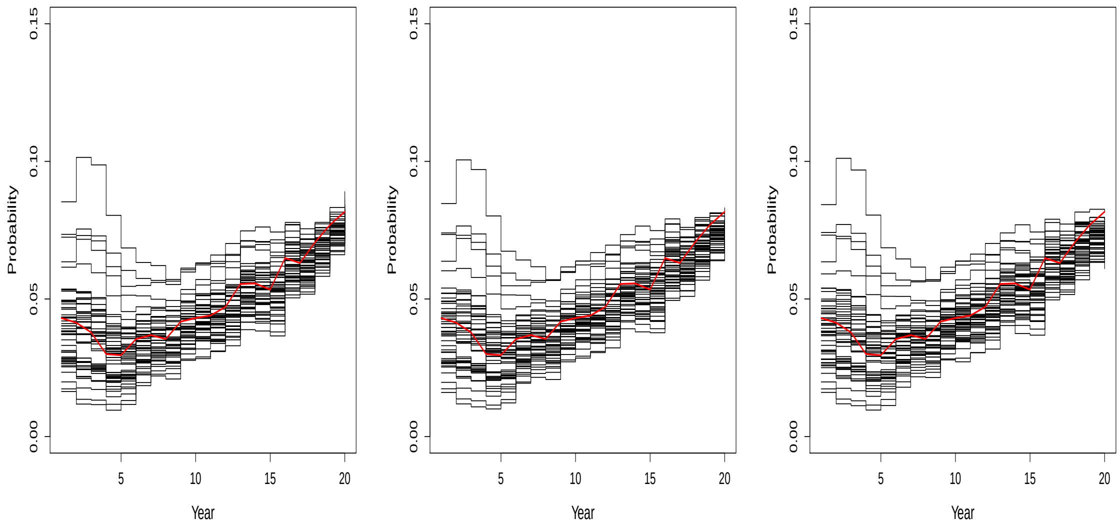
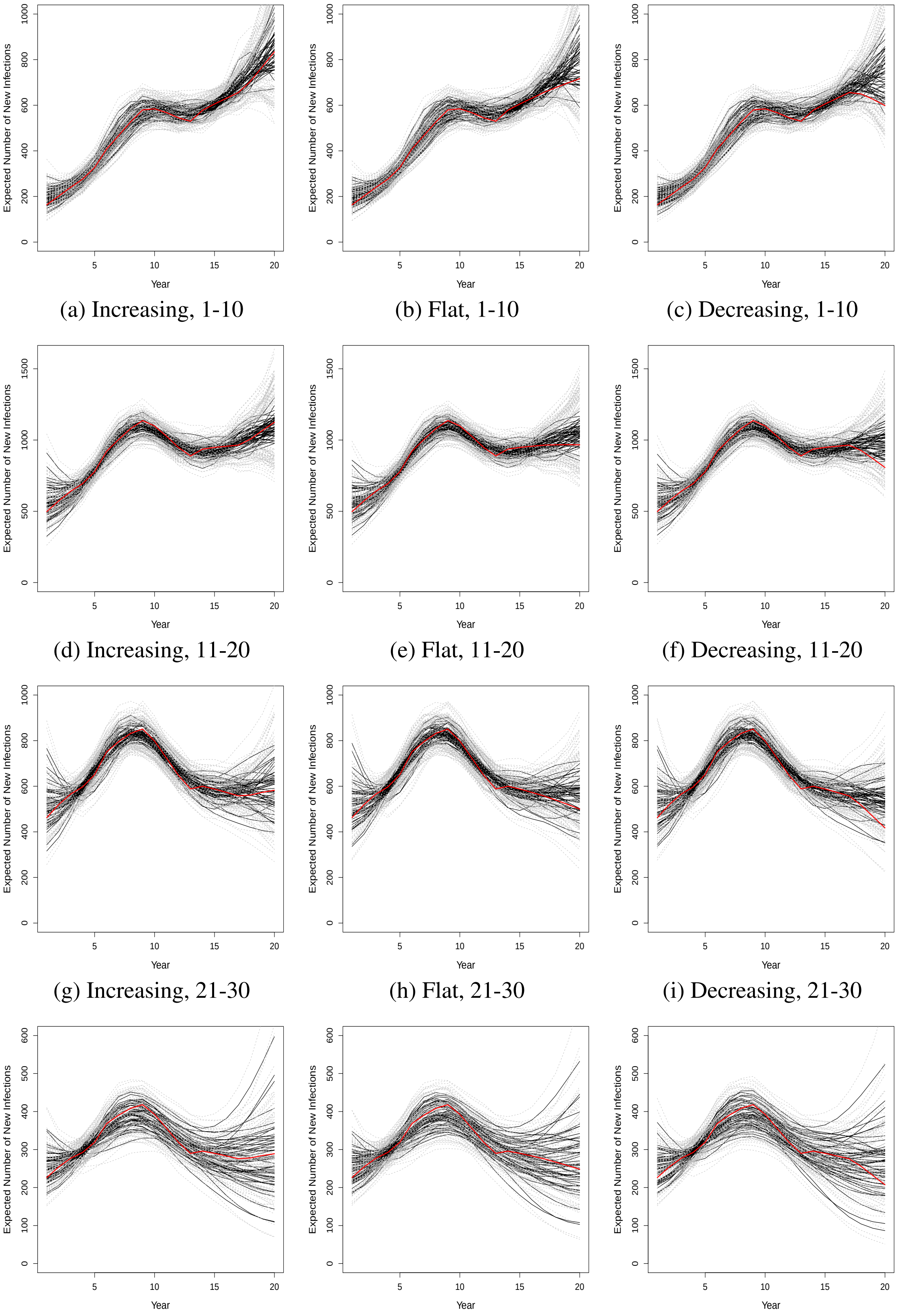


Fig. 7.26 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.27 Estimates (posterior means) of the time profile of the incidence surface, using a *ptensbsord1* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the ptensbsord2 spline incidence model

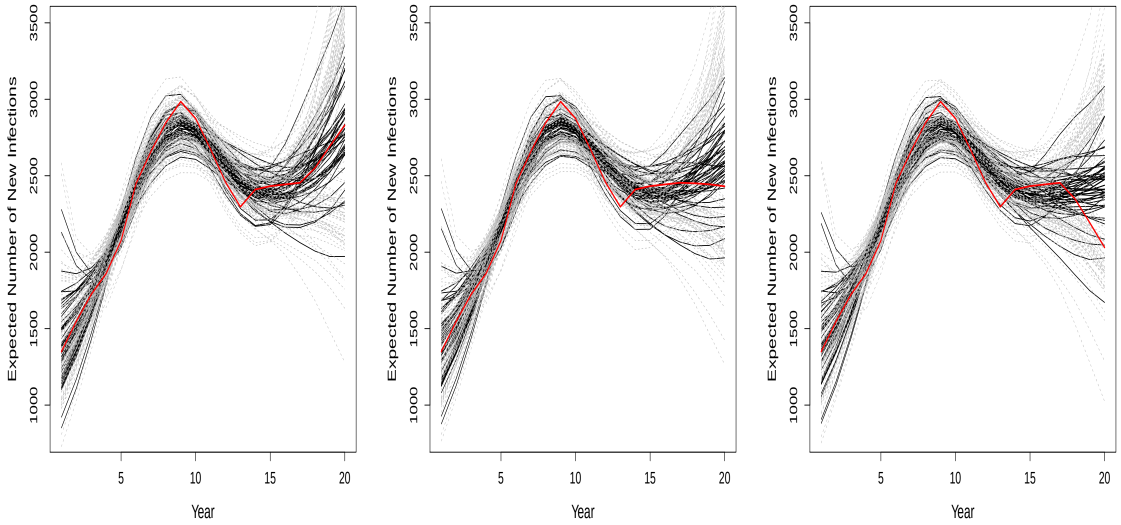


Fig. 7.28 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *ptensbsord2* spline to model incidence. Grey lines denote the respective 95% credible intervals.

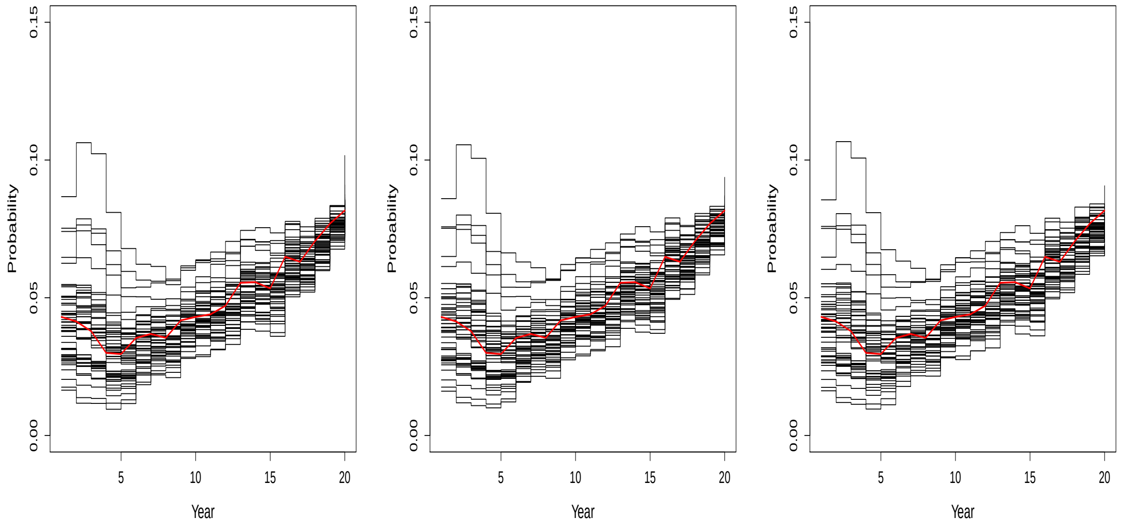
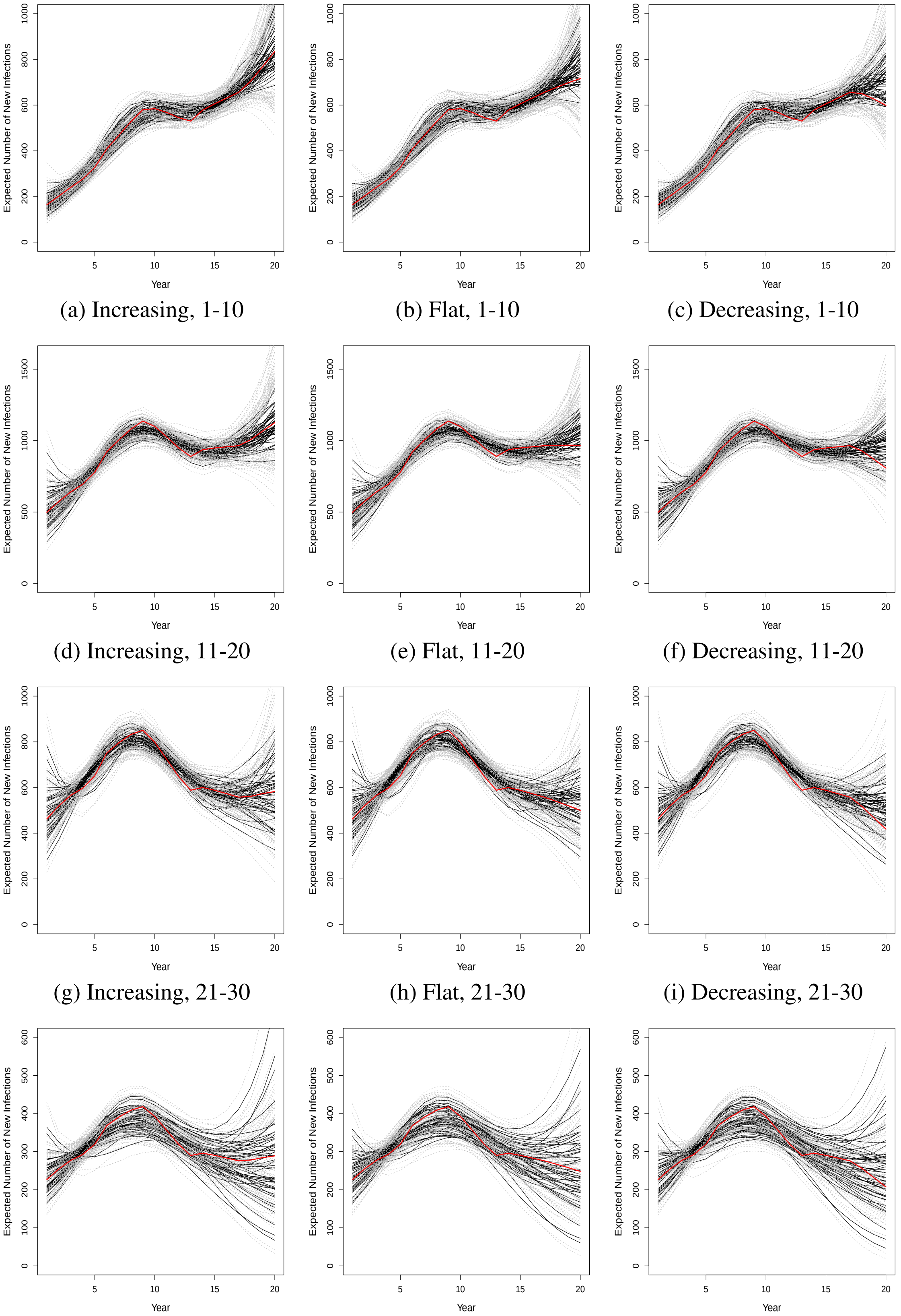


Fig. 7.29 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10 (b) Flat, 1-10 (c) Decreasing, 1-10

(d) Increasing, 11-20 (e) Flat, 11-20 (f) Decreasing, 11-20

(g) Increasing, 21-30 (h) Flat, 21-30 (i) Decreasing, 21-30

(j) Increasing, 31+ (k) Flat, 31+ (l) Decreasing, 31+

Fig. 7.30 Estimates (posterior means) of the time profile of the incidence surface, using a *ptensbsord2* spline incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Results for the Gaussian process incidence model

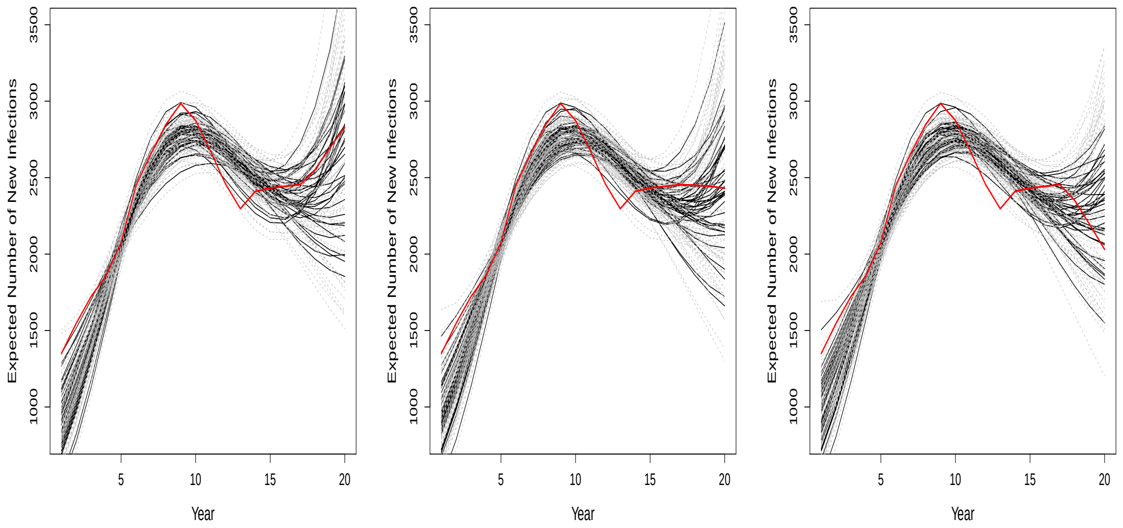


Fig. 7.31 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the incidence surface obtained using a *GP* to model incidence. Grey lines denote the respective 95% credible intervals.

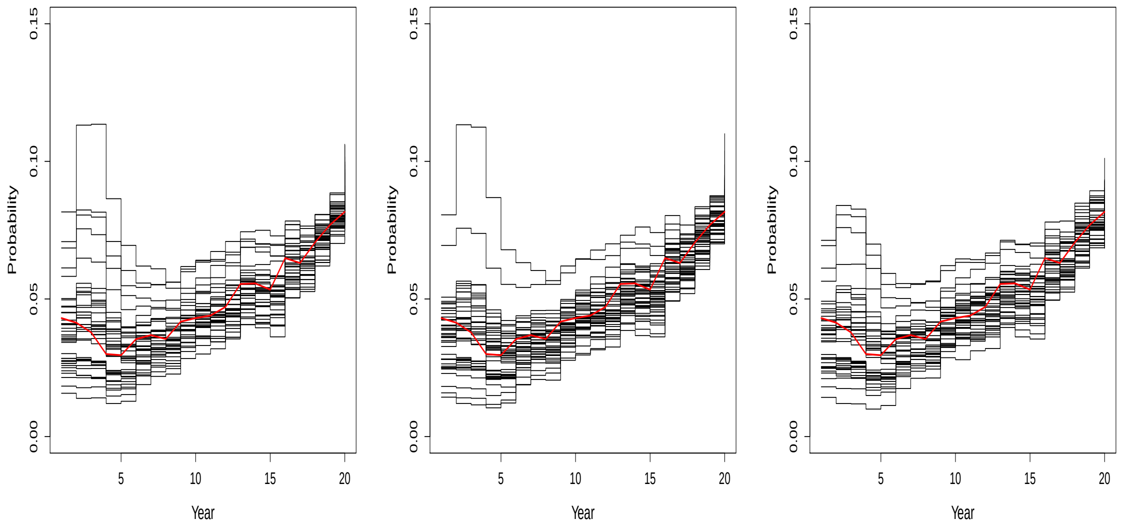
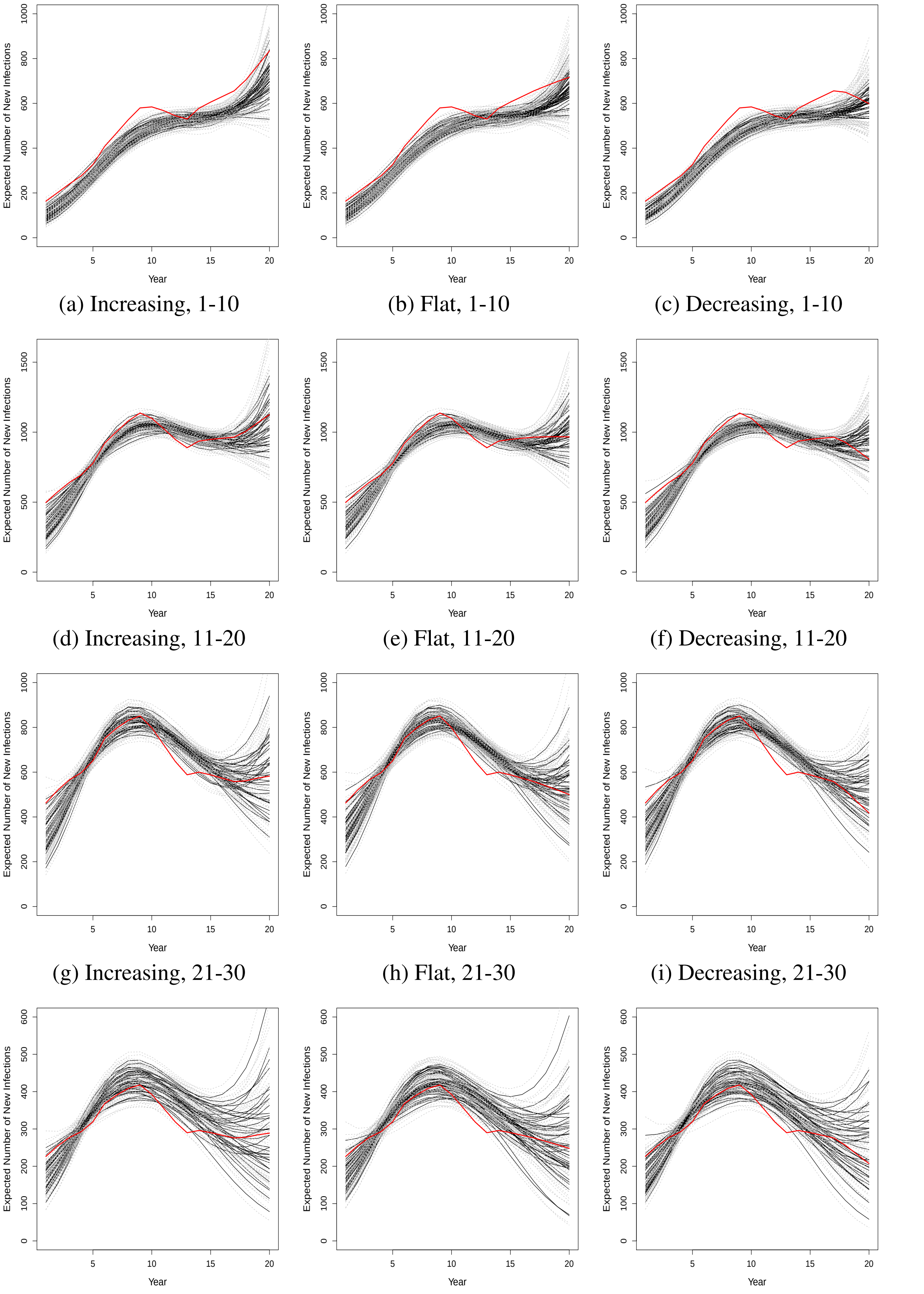


Fig. 7.32 Estimated diagnosis probabilities from state 1: the red lines depict the true diagnois probabilities which are the same in the three true incidence scenarios considered (increasing - left, flat - center, decreasing - right). The black lines represent the estimates (posterior means), for each dataset, of the diagnosis probabilities from state 1 obtained using a first order random walk. Credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.33 Estimates (posterior means) of the time profile of the incidence surface, using a *GP* incidence model, stratified by age-class (1-10, 11-20, 21-30 and 31+ age-classes plotted in the first, second, third and fourth column) under the three true incidence scenarios (increasing - left, flat - centre, decreasing - right). Red lines represent the true time profile and gray lines the respective credible intervals.

Simulations Performance - PMSE

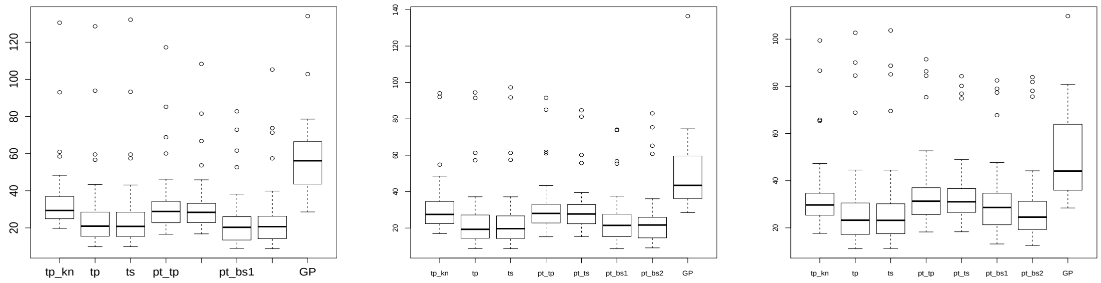


Fig. 7.34 Distribution of $PMSE(\widehat{\mathcal{H}}_m)$ for all incidence models, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

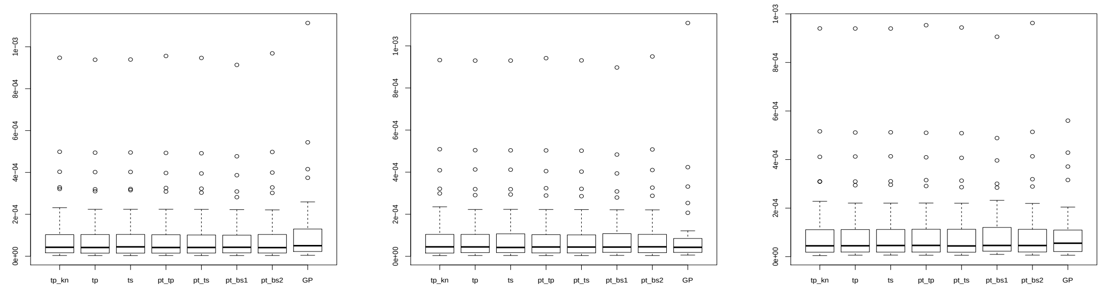


Fig. 7.35 Distribution of $PMSE(\widehat{\mathcal{D}}_{1,m})$ for all incidence models, under three different true incidence scenarios: increasing (left), flat (center), decreasing (right).

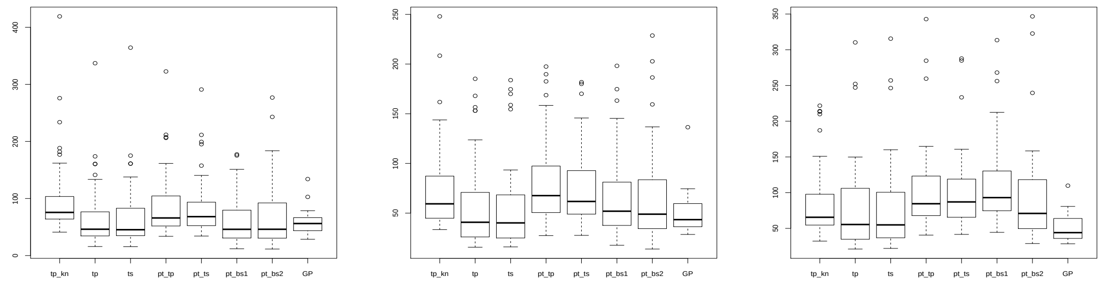


Fig. 7.36 Distribution of $PMSE(\widehat{\mathcal{H}}_m)$ for all incidence models in the last 3 years only, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

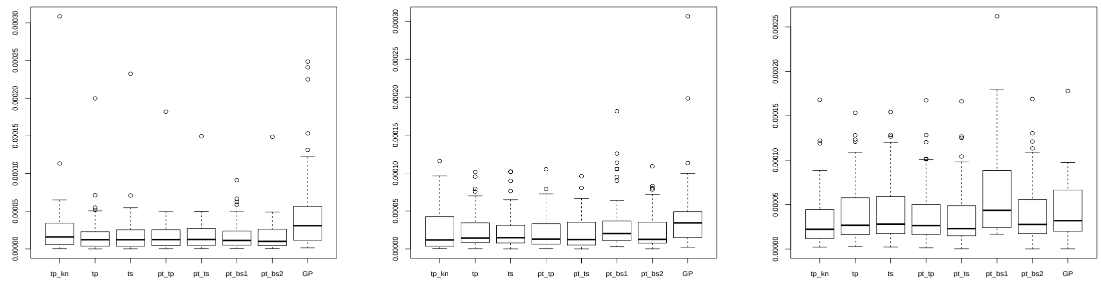


Fig. 7.37 Distribution of $PMSE(\widehat{\mathcal{D}}_{1,m})$ for all incidence models in the last three years only, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

Simulations Performance - Coverage

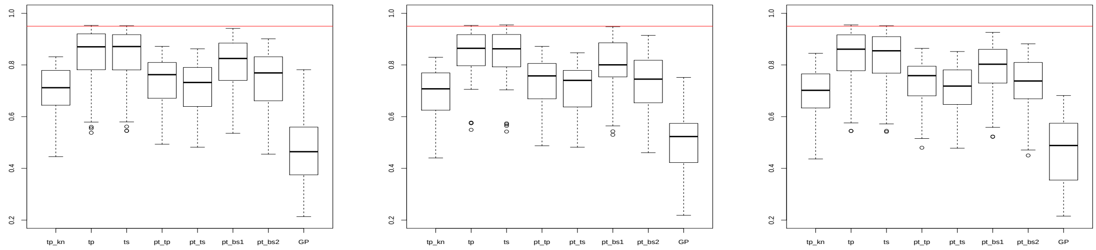


Fig. 7.38 Distribution of $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ for all incidence models, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

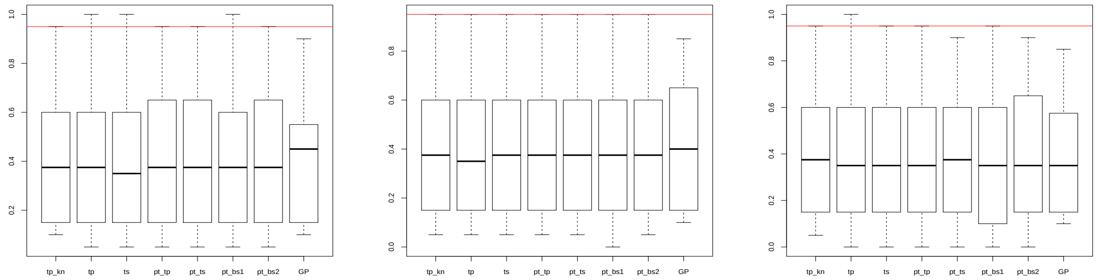


Fig. 7.39 Distribution of $Covg_{0.95}(\widehat{\mathcal{D}}_{1,m})$ for all incidence models, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

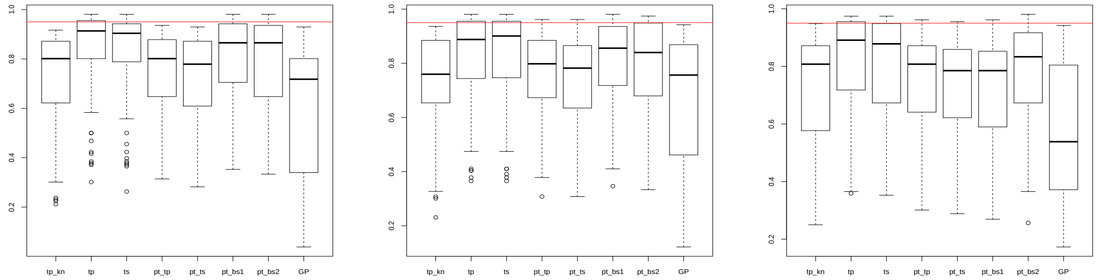


Fig. 7.40 Distribution of $Covg_{0.95}(\widehat{\mathcal{H}}_m)$ for all incidence models in the last three years, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

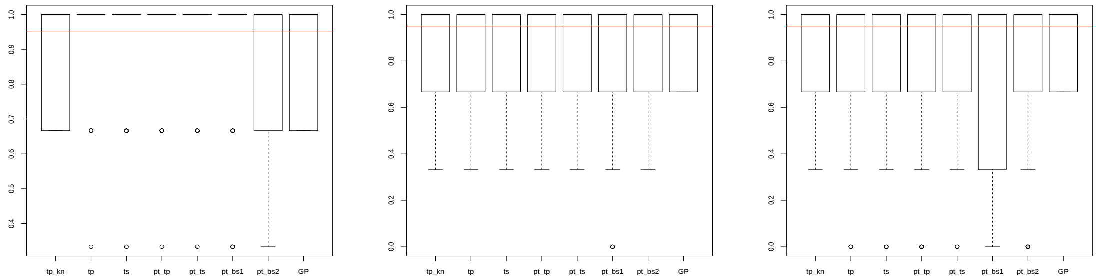


Fig. 7.41 Distribution of $Covg_{0.95}(\widehat{\mathcal{D}}_{1,m})$ for all incidence models for the last three years, under three true incidence scenarios: increasing (left), flat (center), decreasing (right).

7.7 Maximum penalised likelihood simulation study

This Section presents the results of an age-dependent back-calculation simulation study based on maximum penalised likelihood (see Section 6.4.3). Convergence assessment is discussed in Section 7.7.1, then Section 7.7.2 summarizes the study findings and plots of the results obtained are available in Section 7.7.3.

7.7.1 Assessing simulation convergence

Convergence within a frequentist framework involves verifying that the numerical penalised likelihood maximization routine (implemented by the R function optimx) successfully reaches a maximum. Note that we assume that the maximum found is indeed the global maximum of the likelihood. This cannot be formally tested, however we did check that parameters successfully maximised their profile likelihood. Occasionally, the numerical routine did fail to reach a maximum and displayed a related error message.

A further reason the estimation procedure may fail concerns the observed information matrix $\widehat{\boldsymbol{I}}$. This is numerically evaluated (via the **NumDeriv** R package) and hence is not guaranteed to be positive definite; $\widehat{\boldsymbol{I}}+\boldsymbol{S}$ is then non-invertible and hence the confidence intervals and the AIC score are unobtainable.

Tables 7.3 and 7.4 show the number of simulations that encountered convergence issues, stratified by incidence model and by the value of the smoothing parameter λ respectively. Numerical errors occur less often for *ts* splines (in approximately 3% of simulations involving *ts* splines) than for *tp* and *tpknotsloc* splines (5% and 15% of simulations respectively). We further note that numerical errors are more likely when λ is large, which may lead to unrealistically smooth curve and therefore convergence issues.

%	*tp*	*tpknotsloc*	*ts*
No maximum found	0	2	0
(I+S) not positive definite	5	13	3
No error	95	85	97

Table 7.3 Percentage of simulations that encountered convergence issues and that successfully converged, by incidence model.

λ (%)	0	0.5	2	5	8	10	13	16	20	40
No maximum found	0	0	1	0	0	0	0	0	1	2
(I+S) not positive definite	5	5	5	6	6	7	6	8	9	7
No error	85	85	85	84	84	83	84	82	80	81

Table 7.4 Percentage of simulations that encountered convergence issues and that successfully converged, by the value of fixed smoothing parameter λ.

7.7.2 Comments on the results of the simulation study

The evaluation of the simulations' performance is pursued as described in Section 7.5. In Section 7.7.3 results from the *tpknotsloc*, *tp* and *ts* incidence models are plotted. Specifically:

- Figures 7.42, 7.45 and 7.48 depict the time profile of the estimated incidence surface.
- Figures 7.43, 7.46 and 7.49 illustrate the estimated diagnosis probabilities from state 1.
- Figures 7.44, 7.47 and 7.50 show the time profile for the 1-10, 11-20, 21-30, 31+ age-classes.

The *tpknotsloc* splines (Figures 7.42 and 7.44) struggle to estimate the true incidence surface, overall and by age-class, in contrast to the Bayesian setup (Section 7.6.3). The estimated time profile of the true incidence surface is over-smoothed for several datasets. However this is not the case for both *tp* and *ts*, that produce fairly accurate estimates. A head-to-head comparison reveals that the incidence estimates obtained with *ts* splines are more accurate, and thus preferable. This finding is further supported by the Bayesian simulation study in which *ts* splines outperformed all other thin plate splines types.

Overall, time profile estimates of the true incidence surface obtained using *ts* splines within the Bayesian (Figure 7.16) and penalised likelihood framework (Figure 7.48) are comparable; the true incidence is accurately reconstructed but in the earliest and latest years of the epidemic. More in-depth examination reveals that maximum penalised likelihood time profiles estimates are more biased in the recent years, but less variable, compared to the respective Bayesian estimates. The true time profiles of the incidence surface are consistently overestimated in most recent years within a frequentist framework; in a Bayesian framework, respective estimates do not suffer from the same issue but are on occasions extremely poor in most recent years.

Figure 7.51 depicts Bayesian and frequentist quantities of interest obtained using *ts* splines, for an example dataset (number 25) generated under the three true incidence scenarios, in order to illustrate the differences between the two frameworks.

A plausible explanation for the contrasting behaviour of estimates from different frameworks, may concern the $\hat{\lambda}$ estimates. These have a different interpretation in the two frameworks; the optimal $\hat{\lambda}$ is considered to be in a Bayesian framework, the mean of the posterior distribution of λ, whereas in a frequentist framework this is considered to be the λ minimizing the AIC (over a grid of candidate λ values). Figure 7.52 suggests that the larger number of higher $\hat{\lambda}$ values are estimated in the likelihood framework for the *ts* incidence model. Greater λ values enforce a higher level of smoothing, hence stronger bias may be induced to reduce the variability of the estimates. This can be noted when comparing the estimated time profiles of the incidence surfaces (with the *ts* incidence model) within a Bayesian and frequentist framework (Figures 7.16 and 7.48).

A further explanation for the differences in incidence estimates between frameworks concerns the diagnosis probabilities. These are more severely underestimated in a likelihood framework, resulting in a greater bias in the incidence estimates. This is because, in a Bayesian framework, diagnosis probabilities are assumed to be constant within a one-year rather than in a two-years interval as in the frequentist framework. The increased flexibility of the Bayesian set-up guarantees a better fit, but also results to some of the estimates being very poor in the most recent years.

Finally, note that confidence intervals, for the time profile of the incidence surface in the most recent years, are narrower than the respective Bayesian credible intervals. This is probably due to the approximate asymptotic posterior distribution (Equation 6.4.7) from which confidence intervals are derived being an insufficient approximation (Section 6.4.3). Further evidence on the poor performance of the credible intervals is provided by the lower coverage $Covg_{0.95}(\widehat{\mathcal{H}})$.

7.7.3 Plots of results from simulation study

Results for the tpknotsloc spline incidence model

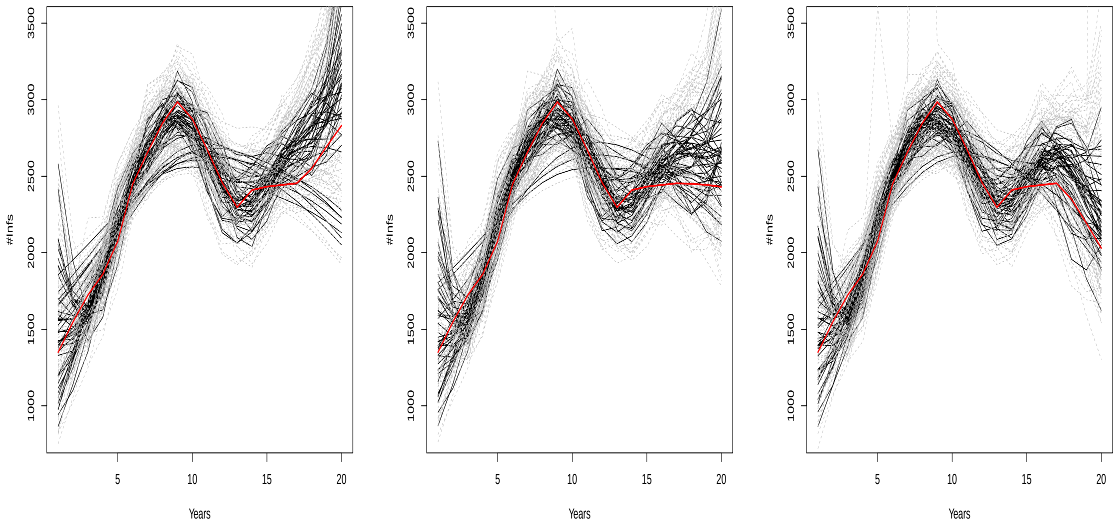


Fig. 7.42 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The estimates of the time profile of the incidence surface for each dataset are plotted in black and their respective 95% credible intervals in gray.

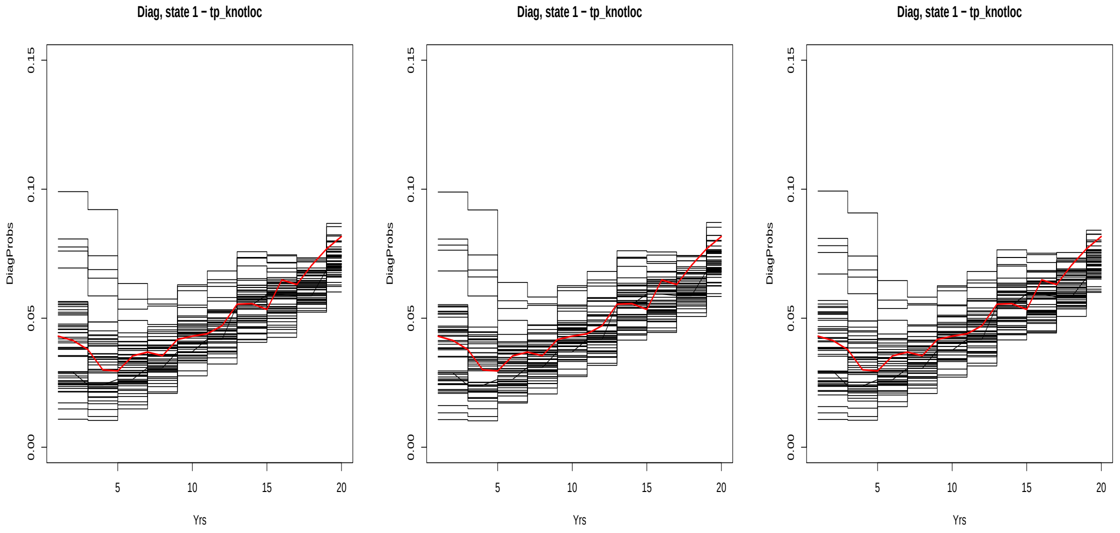
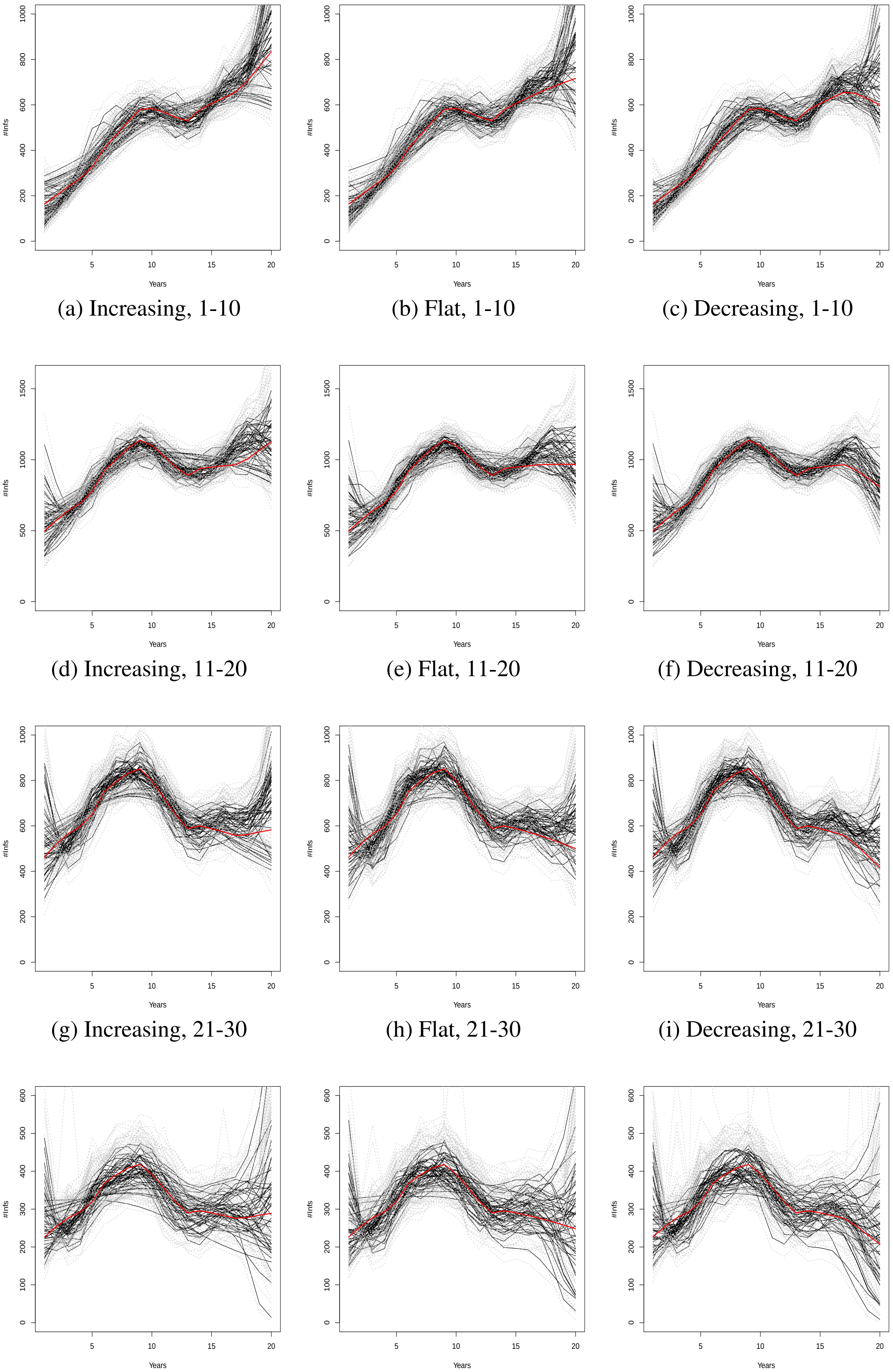


Fig. 7.43 Estimated diagnosis probabilities from state 1: the red lines depict the three true diagnosis curves (the same for all true incidence scenarios). The black lines represent the posterior means for each dataset; credible intervals are only depicted (in grey) on the left figure as these overlap with posterior means, rendering the plot hard to interpret.



(a) Increasing, 1-10 (b) Flat, 1-10 (c) Decreasing, 1-10

(d) Increasing, 11-20 (e) Flat, 11-20 (f) Decreasing, 11-20

(g) Increasing, 21-30 (h) Flat, 21-30 (i) Decreasing, 21-30

(j) Increasing, 31+ (k) Flat, 31+ (l) Decreasing, 31+

Fig. 7.44 Estimated time profile of the incidence surface, stratified by age-class (1-10, 11-20, 21-30 and 31+ classes depicted in the first, second, third and fourth column) under the three incidence scenarios (increasing - left, flat - centre, decreasing - right). True incidence is plotted in red and credible intervals in grey.

Results for the tp spline incidence model

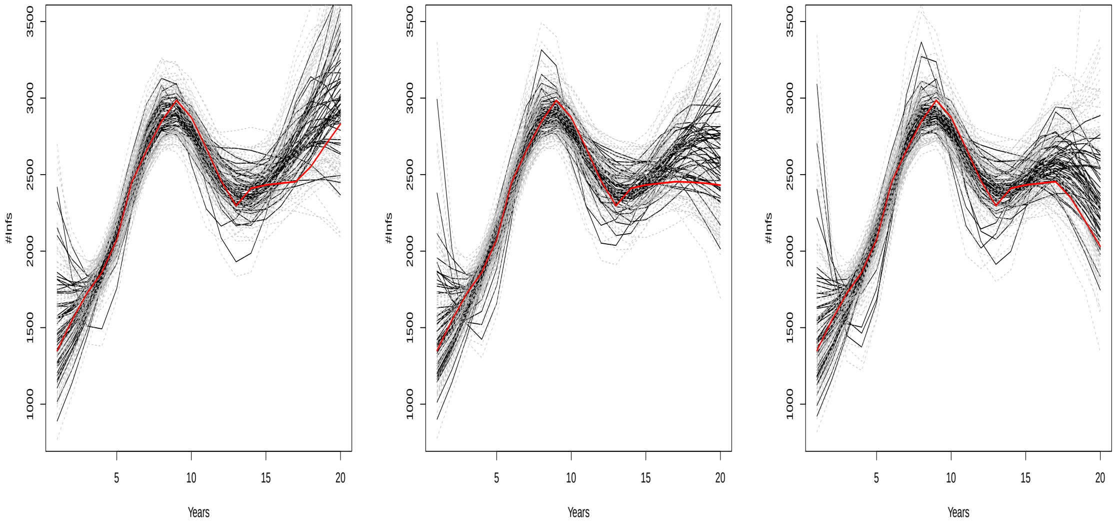


Fig. 7.45 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The estimates of the time profile of the incidence surface for each dataset are plotted in black and their respective 95% credible intervals in gray.

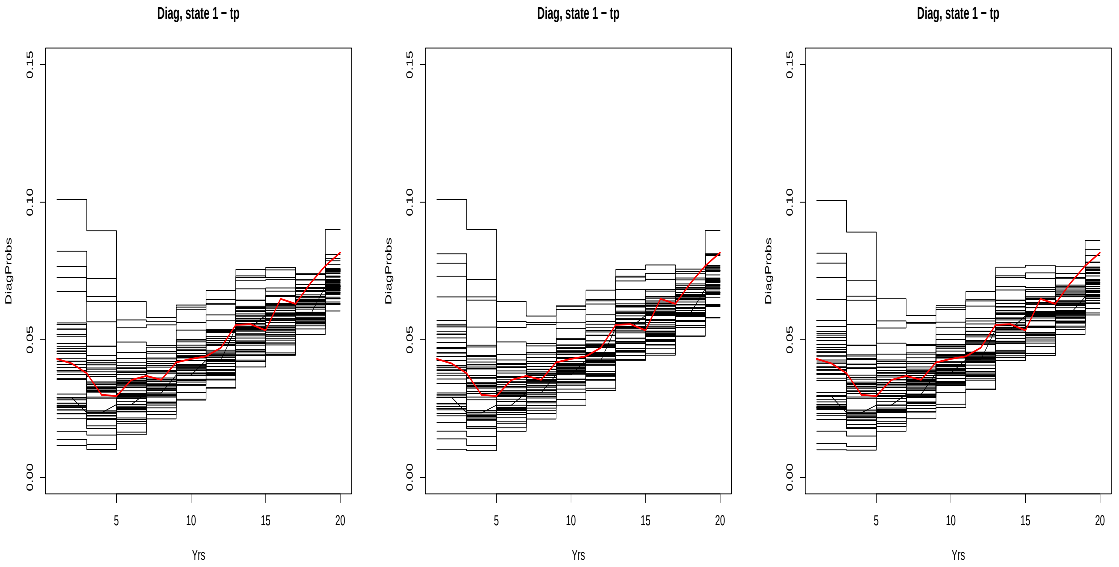
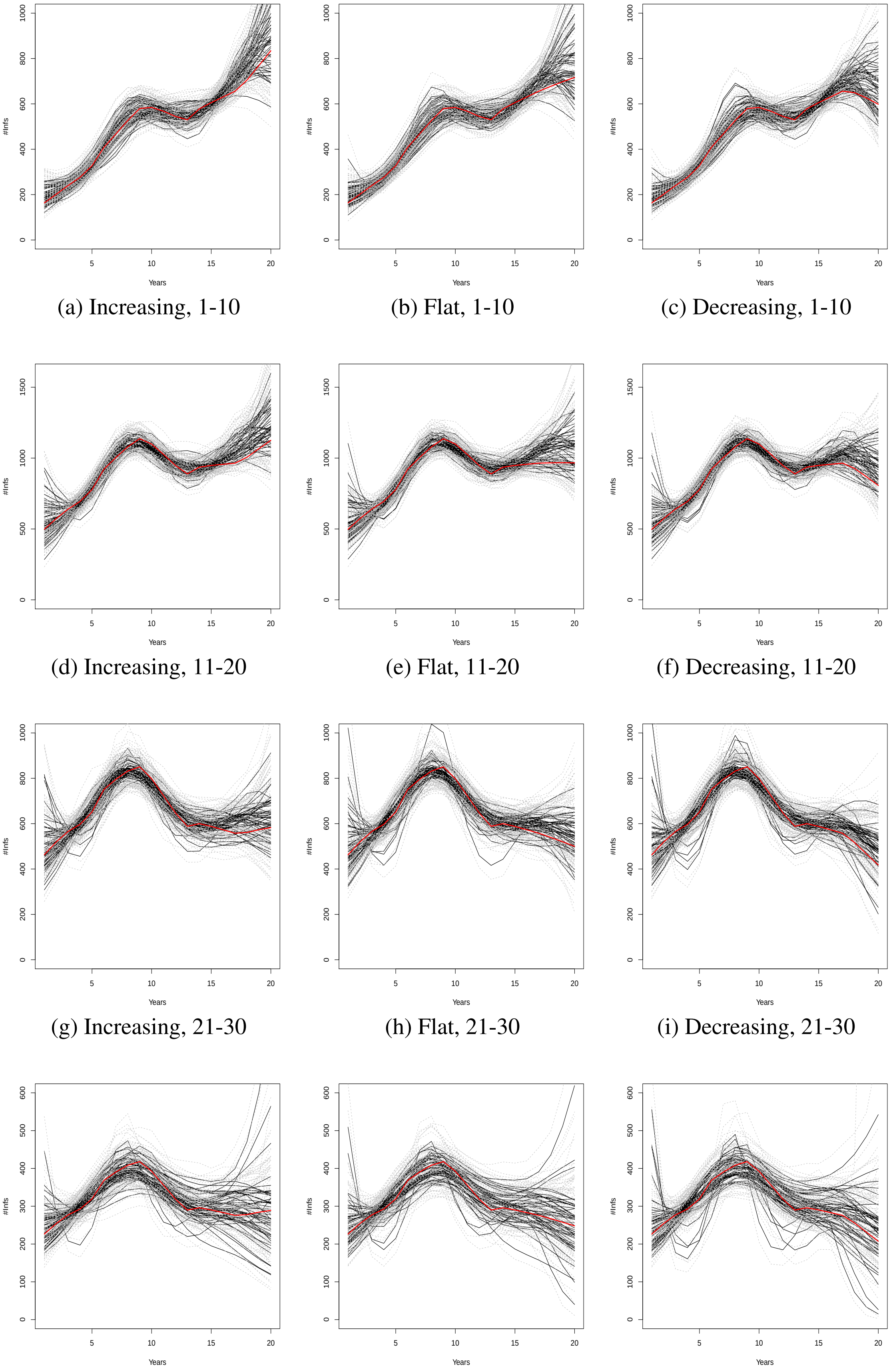


Fig. 7.46 Estimated diagnosis probabilities from state 1: the red lines depict the three true diagnosis curves (that are the same for all true incidence scenarios).The black lines represent the posterior means for each dataset; credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.47 Estimated time profile of the incidence surface, stratified by age-class (1-10, 11-20, 21-30 and 31+ classes depicted in the first, second, third and fourth column) under the three incidence scenarios (increasing - left, flat - centre, decreasing - right). True incidence is plotted in red and credible intervals in grey.

Results for the ts spline incidence model

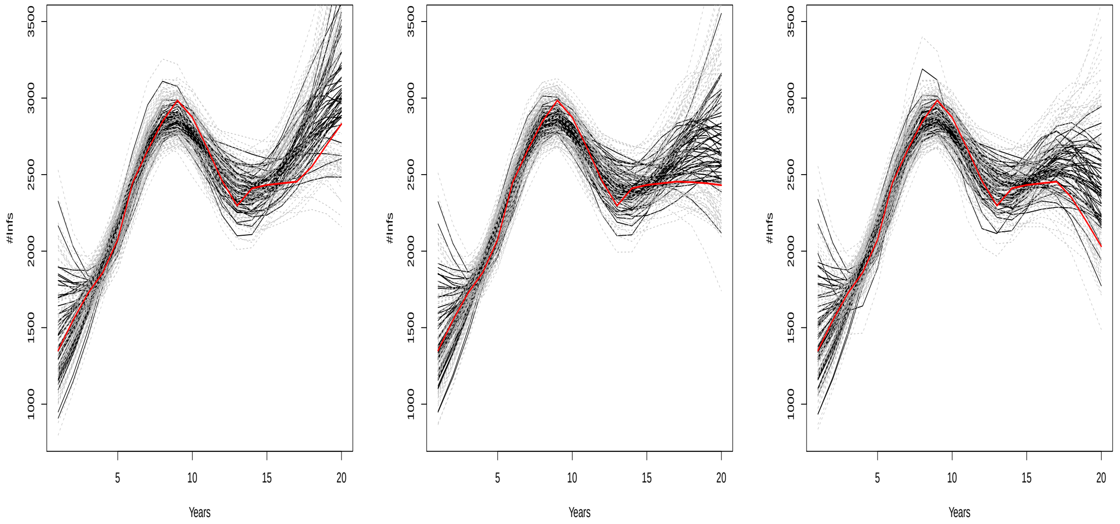


Fig. 7.48 Estimated time profile of the incidence surface: the red lines depict the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The estimates of the time profile of the incidence surface for each dataset are plotted in black and their respective 95% credible intervals in gray.

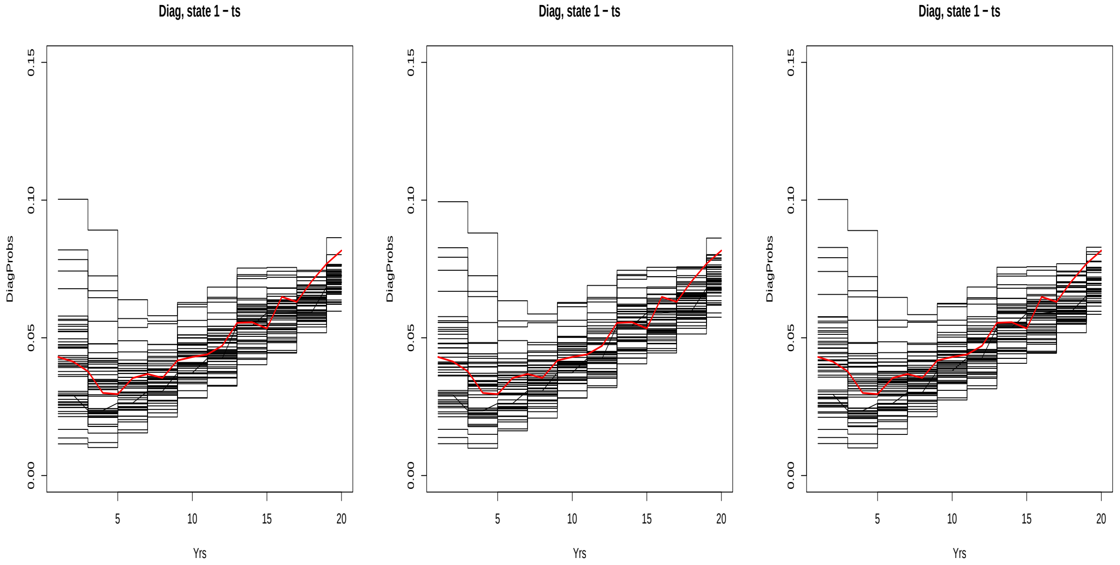
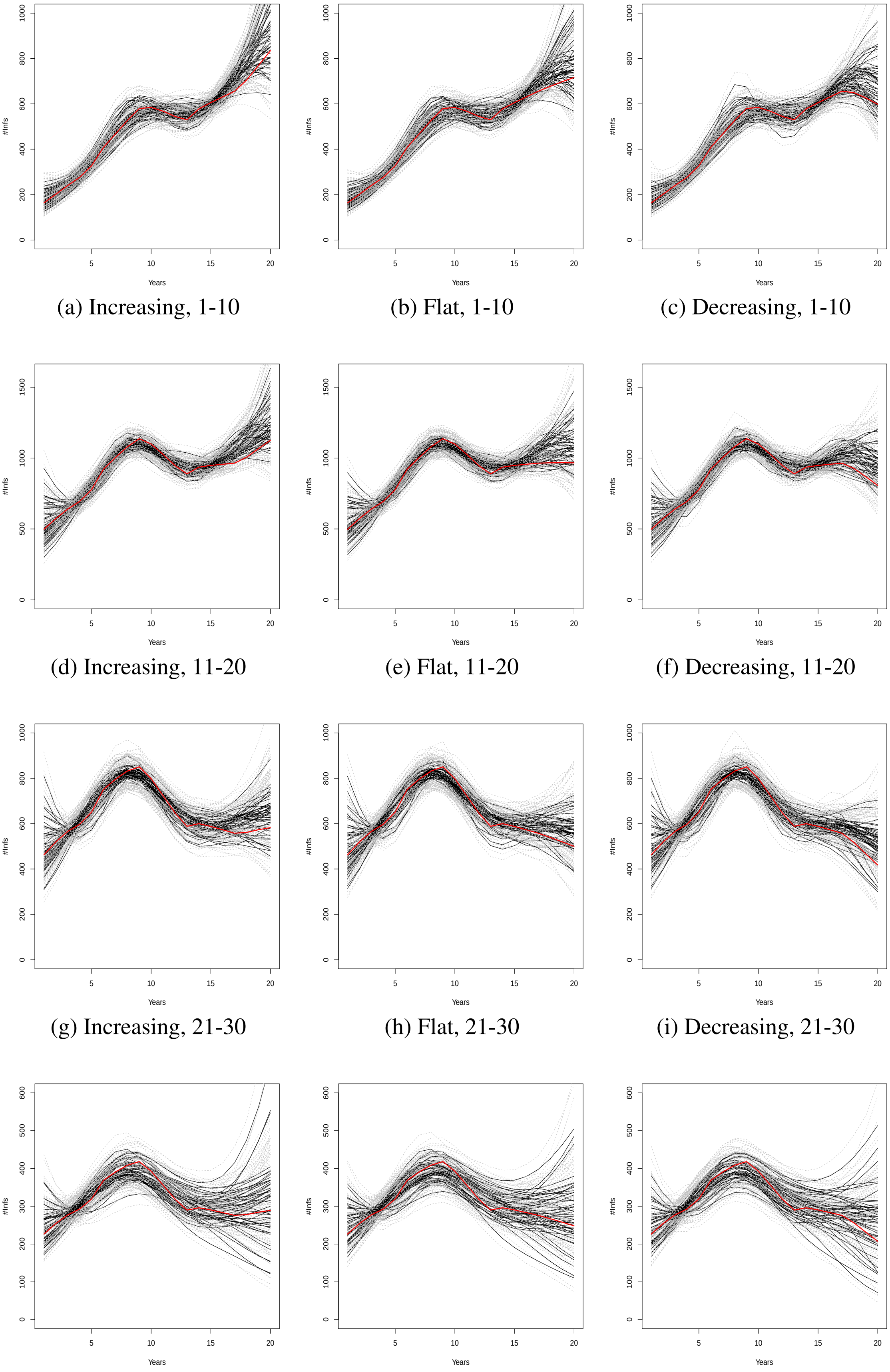


Fig. 7.49 Estimated diagnosis probabilities from state 1: the red lines depict the three true diagnosis curves (that are the same for all true incidence scenarios).The black lines represent the posterior means for each dataset; credible intervals are not depicted.



(a) Increasing, 1-10

(b) Flat, 1-10

(c) Decreasing, 1-10

(d) Increasing, 11-20

(e) Flat, 11-20

(f) Decreasing, 11-20

(g) Increasing, 21-30

(h) Flat, 21-30

(i) Decreasing, 21-30

(j) Increasing, 31+

(k) Flat, 31+

(l) Decreasing, 31+

Fig. 7.50 Estimated time profile of the incidence surface, stratified by age-class (1-10, 11-20, 21-30 and 31+ classes depicted in the first, second, third and fourth column) under the three incidence scenarios (increasing - left, flat - centre, decreasing - right). True incidence is plotted in red and credible intervals in grey.

Comparison of a Bayesian and MPL simulation results using a ts spline

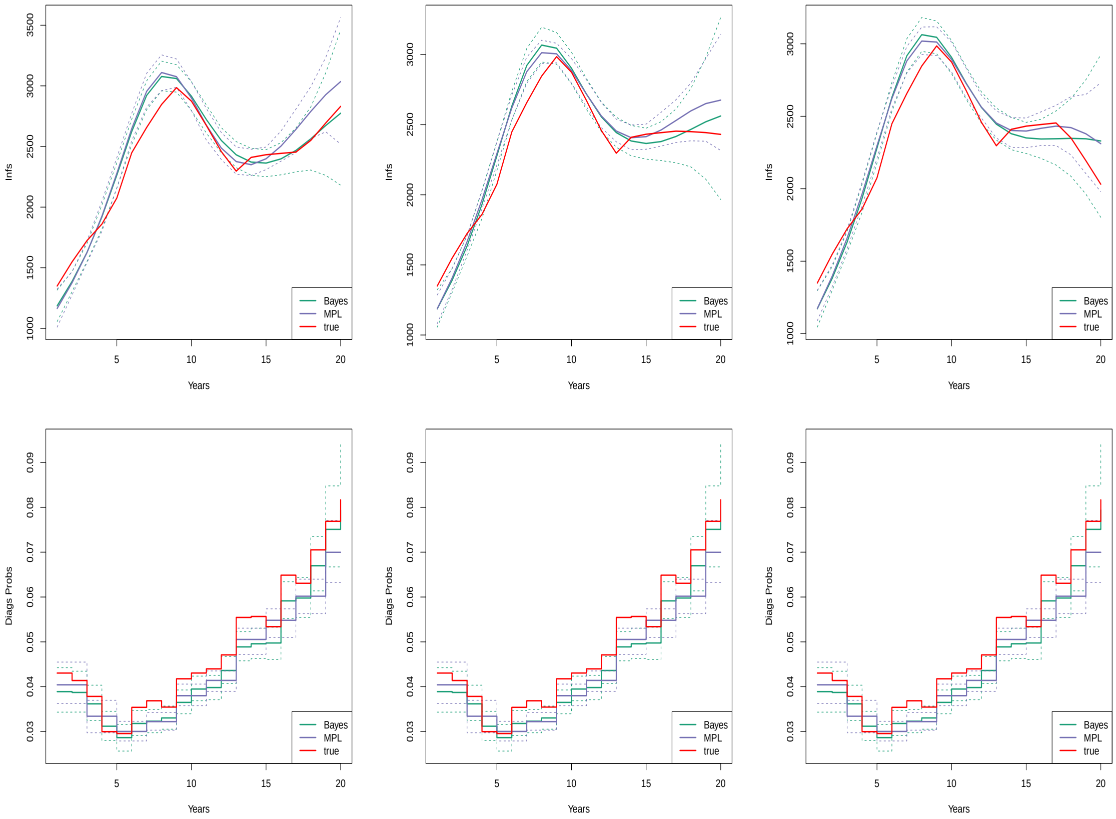


Fig. 7.51 Estimated Time Profile of Incidence Surface (top row) and diagnosis probablities from state 1 (bottom row) for dataset 25 under the three true incidence scenarios (increasing - left, flat - center, decreasing - right). The true time profile of the incidence surface is given in red and the estimated time profile within a penalised likelihood and Bayesian framework is given in blue and green respectively.

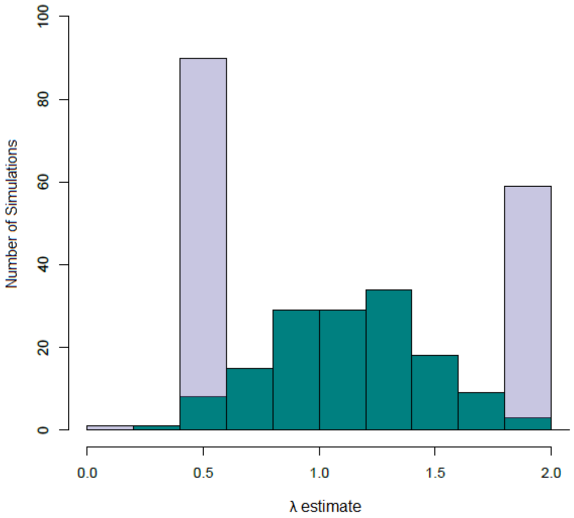


Fig. 7.52 Distribution of the estimated smoothing parameter $\hat{\lambda}$ originating from the penalised likelihood (blue) and Bayesian (green) frameworks.

7.8 Summary

This chapter comprised two main components, a Bayesian and a frequentist back-calculation simulation study. These have been designed to answer the following three questions:

1. Is age-specific back-calculation feasible? What are its strengths and limitations?
2. Can back-calculation's parameters be estimated within both a frequentist and a Bayesian framework?
3. Are some semi-parametric models more appropriate than others to model the latent age-and-time specific incidence surface?

The answer to the first question is yes. In all simulations considered the "true" incidence surface and diagnosis probabilities were satisfactorily reconstructed, except in the most recent years where identifiability issues occur (Section 4.7). Various priors for the smoothing parameters and the variance of the logistic random walks for the diagnosis probabilities have been investigated (results not shown) in order to asses the sensitivity of the estimates on the prior choice (as considered in Section 4.6.4 for age-independent back-calculation). Incidence and diagnosis probabilities estimates are robust to prior specifications.

As far as question two is concerned, both inferential frameworks allow for accurate estimates of the incidence surface and the diagnosis probabilities to be obtained. However Bayesian inference is preferable for a number of reasons, discussed in Section 6.4.4. Furthermore the implementation of the model is superior in terms of computational speed in a Bayesian framework (8 hours versus 15 hours for the frequentist estimate).

Let us now focus on question three. Our Bayesian simulation study considered a number of thin plate and tensor product splines as well as bivariate, quadratic exponential kernel, Gaussian processes for modelling the incidence surface. The latter did not perform well, while *ts* and *ptensbsord1* splines were found to be the most appropriate for this purpose. As discussed in Section 6.2.6, thin plate splines make the assumption of equal smoothing in the time and age dimensions, which is difficult to justify. Consequently we believe that using tensor product splines (*ptensbsord1*) is more appropriate; this further supports adopting a Bayesian perspective, as tensor product splines can not be implemented within a reasonable timeframe in a frequentist framework.

Chapter 8

Application to real data

8.1 Introduction

In this Chapter the back-calculation models developed in Chapters 2 and 5 are applied to surveillance data provided by PHE, on the HIV-MSM epidemic in England and Wales. These data are described in Section 8.2 and the applications of the age-independent and age-dependent back-calculation models are considered in Sections 8.3 and 8.4 respectively.

8.2 Description of the dataset

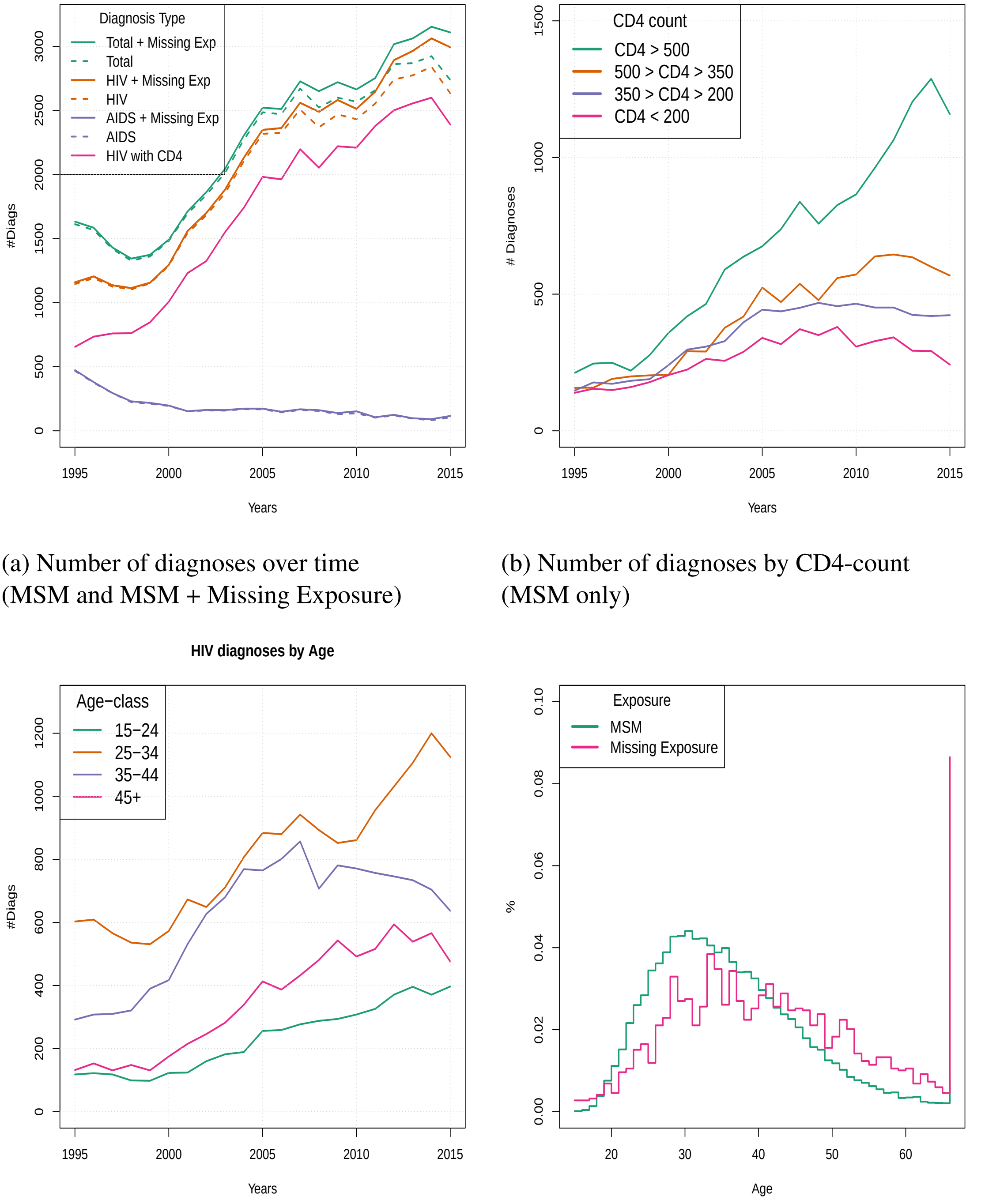
In 2015, approximately 101,000 people were living with HIV in England and Wales, with about 13,000 ($\approx 13\%$) of these being unaware of their infection status (Kirwan et al., 2016). Over 95% of infections are estimated to have occurred via a sexual route, about half of these through homosexual contact. Amongst the heterosexual population, black African men and women are at particular risk of infection; notably, even though they represent only 3.5% of the total UK population, they constitute approximately 30% of the heterosexual population living with HIV.

Monitoring the epidemic is carried out through the collection of relevant surveillance data. The establishment of regular reporting of new AIDS and HIV diagnosis from clinicians and microbiologists dates back to early 1980s, leading to one of the most comprehensive electronic registries in developed countries. New diagnoses are reported quarterly to PHE, together with a range of information, such as ethnicity, country of birth and the age at diagnosis. HIV diagnoses are classified as early (or HIV) or late (AIDS) diagnoses, according to whether clinical AIDS symptoms occur within 3 months of the first positive test.

From 1991, also information on the CD4-cell counts around diagnosis (*i.e.* taken within 3 months of the first positive test and before the uptake of treatment) is available. The national CD4 surveillance dataset (Chadborn et al., 2006), which collects data on all CD4-counts performed by laboratories in England and Wales, is linked to the registry of diagnosis of PHE via unique patient identifiers (Brown et al., 2012). This linkage is not perfect and not all HIV diagnoses have CD4-cell information. However, the fact that the collection of CD4-count data is carried out independently of any other information related to HIV diagnosis justifies the assumption (made in Chapters 2 and 5) that the distribution of the available CD4-count is representative of the distribution of all CD4-counts.

The risk exposure group is unknown for a small percentage of diagnoses; these cases are further investigated by PHE to reconstruct missing exposure information. Missing exposure diagnoses are more common in recent years, as there is a time lag between chasing, obtaining and subsequently linking exposure information. To avoid under estimating the extent of the epidemic, PHE imputes as MSM individuals with unknown risk exposure, based on their empirical distribution. Age-dependent back-calculation further requires specification of age at diagnosis for individuals with missing exposure; we carried out individual-level imputation, using individual-level information available, including age at diagnosis, on the diagnoses with unknown exposure. Note that the distribution of the age at diagnosis for individuals with missing exposure is not similar to the corresponding distribution of age at diagnosis for MSM (Figure 8.1d). Imputing the CD4-count for unknown exposure diagnoses is not necessary, as the CD4-count is assumed to be available for a subsample of the HIV diagnosis only (Section 2.3.4).

A total of 45,972 MSM diagnoses (HIV and AIDS) have been recorded since 1995, six diagnoses did not have age at diagnosis and have been excluded from the analysis. Trends in diagnosis data are shown in Figure 8.1a, aggregated by year. Figure 8.1b displays trends in diagnoses by CD4-category over time, stratified by categories $[500,\infty)$, $[350,500)$, $[200,350)$ and $[0,200)$. Figure 8.1c plots the diagnoses by age-classes 15-24, 25-34, 35-44 and 45+ (without considering the age of individuals with unknown exposure at diagnosis). From Figure 8.1 the total diagnoses have been steadily increasing since 2000, until 2015. Linkage with CD4 information (Figure 8.1a) increased from 60% in 1995 to 90% in 2015, in which there is a substantial increase in diagnoses with CD4 $\geq$ 500, whereas the contribution of AIDS to total diagnoses dropped from 30% in 1995 to 5% in 2015. Age at diagnosis ranges between 15 and 88 years. From 2010 onwards, HIV diagnoses for individuals aged between



(a) Number of diagnoses over time (MSM and MSM + Missing Exposure)

(b) Number of diagnoses by CD4-count (MSM only)

(c) Number of diagnoses by age at diagnosis (MSM only)

(d) Comparison of age-at-diagnosis distribution for diagnosed MSM and diagnosed individuals with unknown exposure

Fig. 8.1 Descriptive plots of diagnosis data.

15 and 34 years steadily increased, with this increase being most marked in the 25-34 age-class. On the other hand there was a decrease in diagnoses at older ages. Consequently the median age at diagnosis dropped from 36 years in 2010 to 33 years in 2015.

This preliminary analysis of the data poses a series of questions: is this observed increase in diagnoses due to an increase in infections or due to increased testing (or both)? In which age-groups is HIV incidence particularly pronounced? Is the decrease in median age at diagnosis due to increasing infections among young people or a consequence of the reduced time to diagnosis?

We will employ back-calculation methods, allowing the estimation of incidence and diagnosis probabilities, in order to answer the above questions. Note that age-independent back-calculation can only address the first question, an extension to a novel age-dependent back-calculation is required to tackle the second and third problems.

8.3 Age independent back-calculation case-study

We start with considering a case study involving the application of the age-independent back-calculation model, discussed in Chapter 2, to the MSM-HIV epidemic in England and Wales.

8.3.1 The model

The epidemic is modelled from 1995 to 2015, employing a quarterly (T=84) back-calculation model with $K = 4$ undiagnosed states, as described in Figure 4.1 (see Section 4.2). The expected number of undiagnosed infections in 1995 and the progression probabilities are chosen to be equal to $\boldsymbol{\pi}^\star$ and $\boldsymbol{q}^\star$ as defined in Section 4.2. The values of $\boldsymbol{q}^\star$ are so that the average time spent in each of the undiagnosed states, ordered from $CD4 \geq 500$ to $CD4 < 200$, is 2.56, 2.17, 2.15, 1.68 years respectively, and that newly infected individuals require, on average, 8.56 years to develop AIDS.

The only difference with respect to the back-calculation model of Section 4.2 lies in the incorporation of under-reporting. This is considered to only affect AIDS diagnoses, to reflect a change in reporting conventions in 2000, resulting in some AIDS diagnoses being considered less relevant for surveillance purposes and not being subsequently reported.

Under-reporting is assumed to be constant over time, as defined in Section 2.4.1:

$$\begin{aligned} \upsilon_i^H &= 1, \quad i = \{1,\dots,84\} \\ \upsilon_i^A &= 1, \quad i = \{1,\dots,20\} \\ \upsilon_i^A &= \upsilon, \quad i = \{21,\dots,84\} \end{aligned} \tag{8.3.1}$$

where υ is the under-reporting parameter to be estimated. Based on expert opinion, approximately two-thirds of AIDS diagnoses have been reported from year 2000 onwards, so we choose an informative $Beta(236,118)$ prior for υ to reflect this.

Chapter 4 investigated non-parametric methods to model the latent log-incidence curve. Results, summarized in Section 4.7, suggested to use first order random walks (*rw*), univariate thin plate regression splines (*ts*) and cubic B-splines with first order difference penalty splines (*bsord1*) or Gaussian Processes (*GP*). Four distinct quarterly-varying first order random walks on the logistic scale $\delta_{k,i} \sim N(\delta_{k,i-1}, \sigma^2_{D,k})$, $i = \{1,\dots,84\}$, $k = \{1,\dots,4\}$ were used to model diagnosis probabilities (Section 4.3).

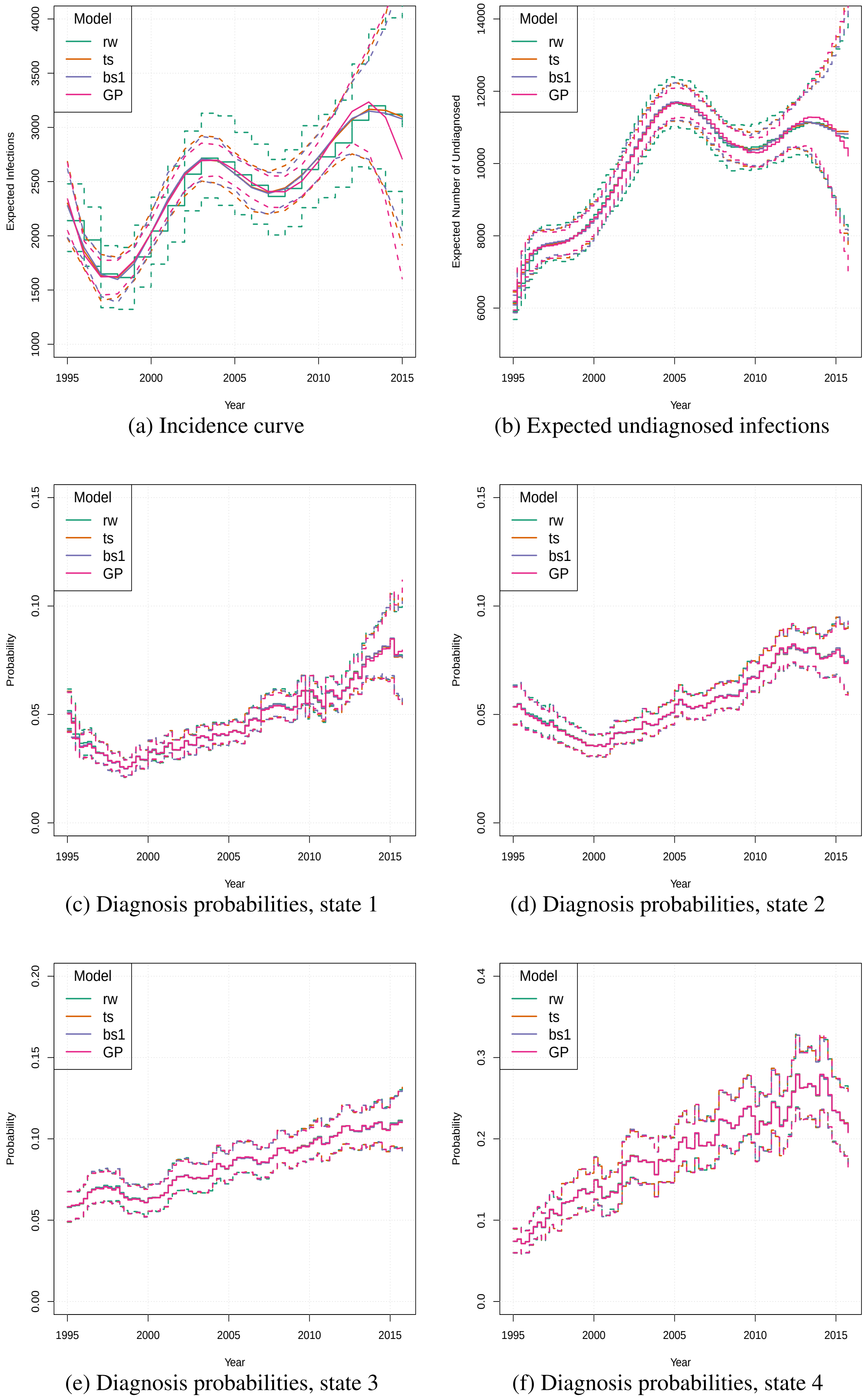
Inference is only carried in a Bayesian framework and implementation is based on `Stan`, with four parallel chains of 2000 iterations, 1000 of which are burn-in, resulting in total of 4000 posterior draws. Results are obtained within five minutes. Again, convergence was assessed using trace plots and R-hat statistics (as shown in Section 4.6.1) and was achieved for all models considered.

Results are first presented in Section 8.3.2. Section 8.3.3 explores the differences between running back-calculation from the beginning or from an intermediate point of the epidemic. Finally, Section 8.3.4 reports a sensitivity analysis on the specification of $\boldsymbol{\pi}^\star$.

8.3.2 Results

Throughout this Chapter, in every Figure, unless otherwise stated, solid lines represent the means of the posterior distributions of interest, and dashed lines denote the corresponding 95% credible intervals (CrI).

Figure 8.2 depicts the results obtained using a first order random walk, *ts* and *bsord1* splines and a Gaussian Process as incidence models respectively. The incidence curve estimates are very similar for all incidence models (Figure 8.2a): the expected number of yearly infections declines from approximately 2250 in 1995 to 1750 in 1999, and subsequently smoothly increases until 2004, to approximately 2750. After 2004, the number of expected yearly



(a) Incidence curve

(b) Expected undiagnosed infections

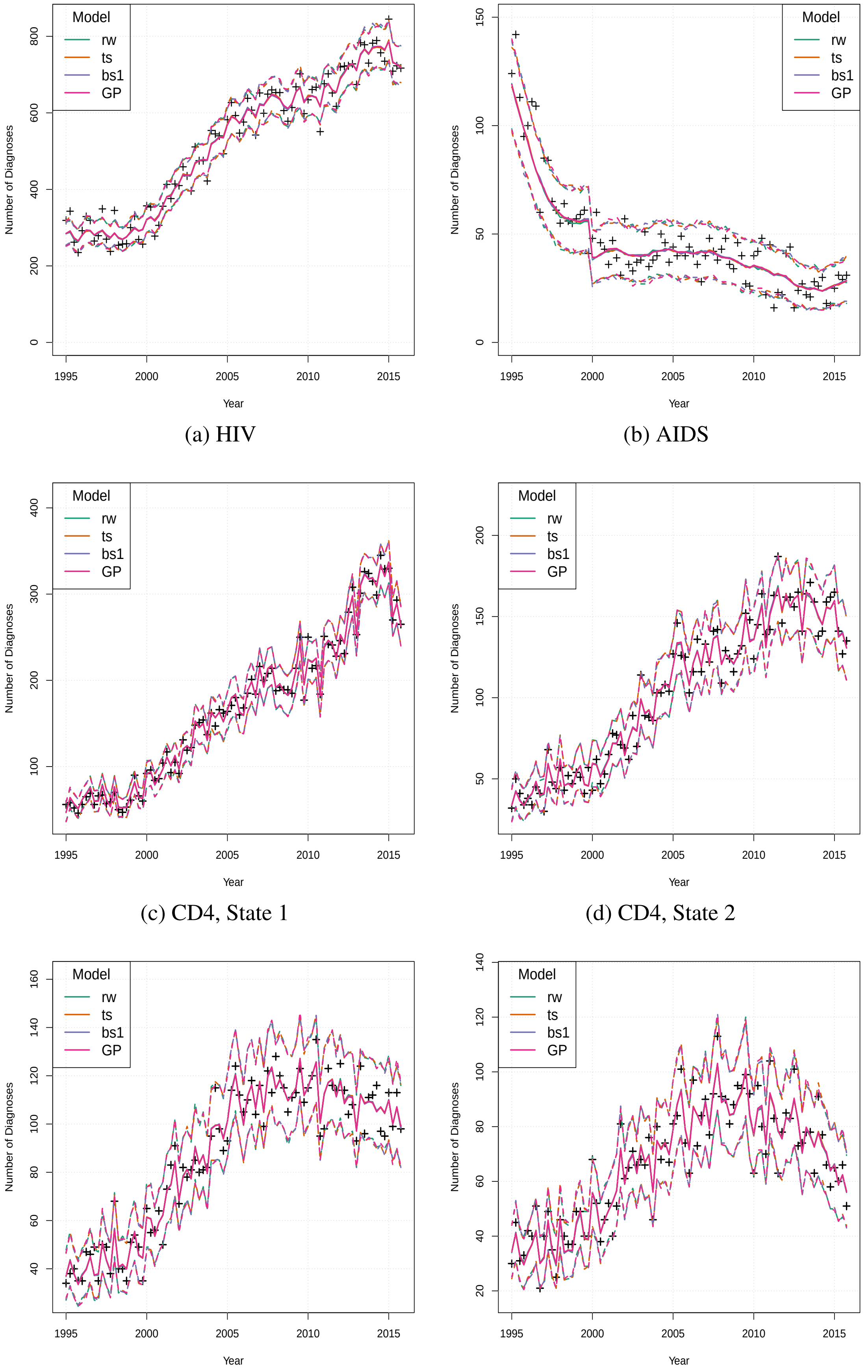
(c) Diagnosis probabilities, state 1

(d) Diagnosis probabilities, state 2

(e) Diagnosis probabilities, state 3

(f) Diagnosis probabilities, state 4

Fig. 8.2 Age-independent back-calculation results; *rw* = first order random walk, *bs* = first degree P-splines, *ts* = thin plate spline with shrinkage and *GP* = Gaussian process. State 1: CD4 $\geq$ 500, State 2: 350 $<$ CD4 $\leq$ 200, State 3: 350 $<$ CD4 $\leq$ 200, State 4: CD4 $<$ 200.



(a) HIV

(b) AIDS

(c) CD4, State 1

(d) CD4, State 2

(e) CD4, State 3

(f) CD4, State 4

Fig. 8.3 Age-independent back-calculation goodness of fit; *rw* = first order random walk, *bs* = first degree P-splines, *ts* = thin plate spline with shrinkage and *GP* = Gaussian process. State 1: CD4 $\geq$ 500, State 2: 350 $<$ CD4 $\leq$ 200, State 3: 350 $<$ CD4 $\leq$ 200, State 4: CD4 $<$ 200.

infections has been slowly decreasing until 2008 and has since been increasing, reaching a plateau of approximately 3000 expected infections per year. In the most recent years, diagnosis data are not informative about recent infections (Section 1.4), resulting in widening credible intervals.

The incidence estimates for the splines (*ts* and *bsord1*) and the first order random walks incidence models are essentially identical, with splines resulting in narrower credible intervals (see Section 4.6.2). In contrast to *rw*, *ts*, and *bsord1*, *GP* suggests that incidence is decreasing. As discussed in Section 4.6.2, this is due to the prior mean (zero) reversion property of Gaussian processes.

For all undiagnosed states, diagnosis probabilities increased over the last 15 years, with a further increase being observed in the last five years (excluding the $CD4 < 200$ state). It is interesting to note that diagnosis probabilities from the $CD4 > 500$ state have substantially increased in the last five years.

Figure 8.2b shows trends in the expected number of individuals living with undiagnosed HIV; this steadily increases from 6080 (95% CrI 5870, 6350) in 1995 to 11750 (95% CrI 11200, 12200) in 2005, due to incidence having increased. Subsequently the expected number of HIV undiagnosed infections drops, reaching a minimum of approximately 10350 (95% CrI 9920, 10800) in 2009 and subsequently stabilizes at approximately 11,000 individuals as a result of increasing diagnosis pressure counteracting the effect of increasing incidence.

To assess goodness of fit, the posterior-predictive distribution for the data is plotted (see Figure 8.3) for the replicated data that could have been observed (Gelman et al., 2014, Chapter 6). All incidence models considered accurately fit the data. The credible intervals for the HIV posterior-predictive distribution are narrower in the early years of the epidemic as uncertainty around the number of undiagnosed infections in 1995 ($\boldsymbol{\pi}$) is ignored. Approximately 85% of the AIDS and HIV diagnosis data are covered by the 95% credible interval; this lack of coverage is particularly severe in the earliest years. In fact, this figure is revised to approximately 92% when considering data from 2000 onwards. All CD4-count data are covered by the posterior-predictive credible intervals at the nominal 95% level, suggesting they are particularly influential. The noisy fit of the CD4-count data in Figures 8.3c to 8.3f suggests that overfitting may be present, which typically leads to poor predictive performance. It is recommended to avoid using back-calculation for predictive purposes, as the incidence estimates obtained, in the most recent years, are highly uncertain and result in highly volatile predictions. However, it is of interest understanding whether overfitting leads to substantial changes in incidence estimates, when small changes in the data are introduced. This is further investigated in Section 8.4.2.

8.3.3 Plausibility of back-calculation on a reduced time period

We here examine the sensitivity of results to running back-calculation from the beginning of the epidemic (*1978-model*) compared to running it from a midpoint (*1995-model*). In this sensitivity analysis, the log-incidence is modelled by a first order random walk.

Results

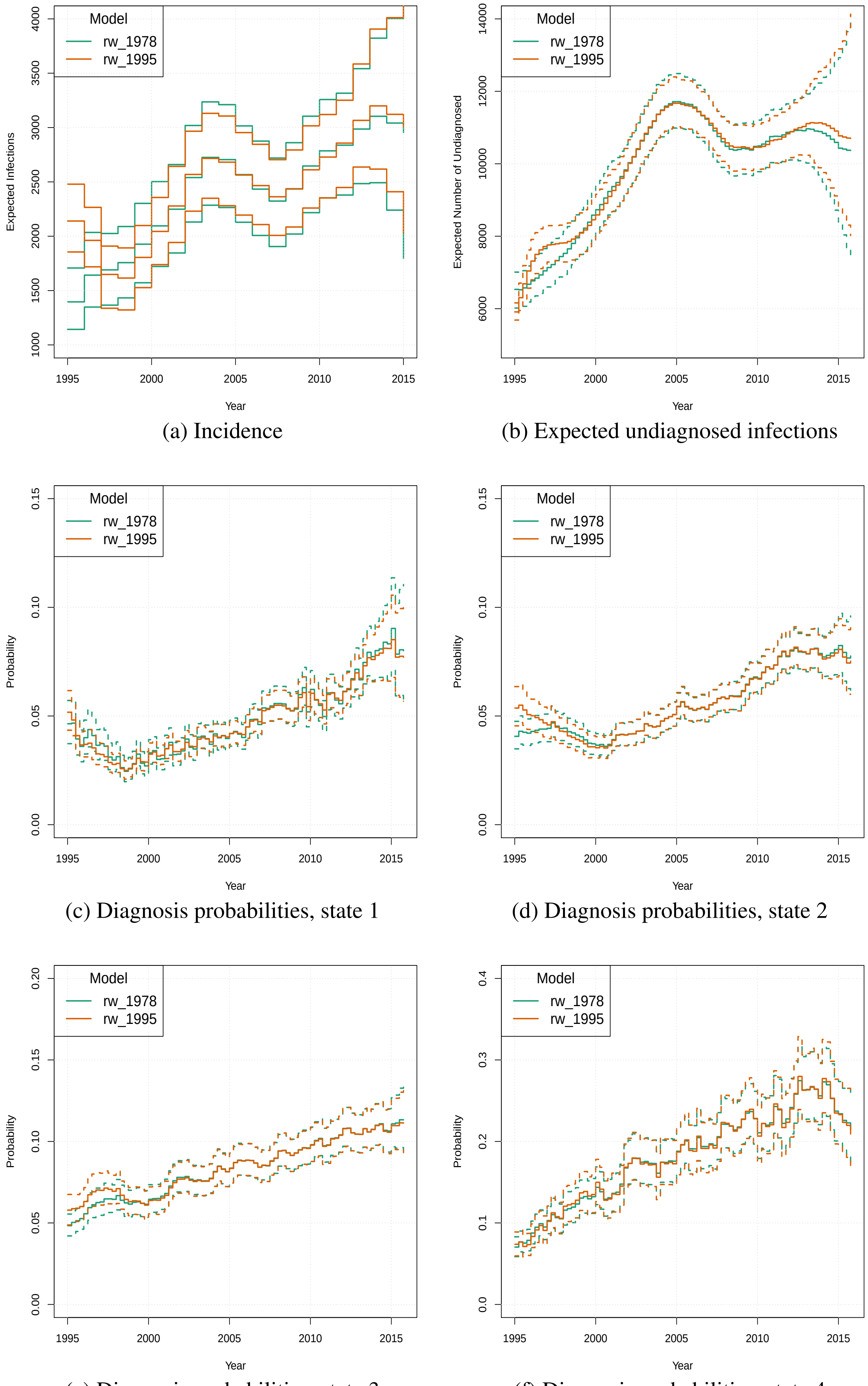
Figure 8.4 shows the results from the *1978-model* and the *1995-model*. Table 8.1 compares the pre-specified expected number of undiagnosed infections in the *1995-model* to the expected number of undiagnosed infections in 1995 estimated with the *1978-model*. Estimates of the incidence curve from the two models substantially differ only in the first two years, as the expected number of undiagnosed infections in 1995 is lower for the *1995-model* (see Table 8.1). Thus, the incidence estimates from the *1995-model* are inflated in the late 90's, to ensure that diagnosis data are accurately captured. Goodness of fit plots (Appendix H.1) confirm that both models fit equally well throughout the epidemic.

Credible intervals are somewhat narrower for the *1995-model*. The estimated incidence curve from the *1978-model* is quite volatile in the 80's but smooths out from the 90's onwards. Since the variance of the log-expected incidence first order random walk is assumed to be constant over time, the greater variability in the 80's is reflected in wider credible intervals.

Estimated diagnosis probabilities from the two models are also very similar, with the exception of the first four years for states 1 to 3. Given that the fixed number of expected undiagnosed infections in the *1995-model* is lower (with the exception of state 4) than the estimate number of undiagnosed people from the *1978-model* (Table 8.1), higher diagnosis probabilities are necessary to accurately fit the data.

Model	State 1	State 2	State 3	State 4
1978-model	2176.89	1539.08	1340.96	838.36
1995-model	1710.67	1191.20	1191.20	870.00

Table 8.1 Comparison between the expected number of undiagnosed infections in the beginning of 1995, as estimated through the *1978-model* and the fixed number used in the *1995-model*.



(a) Incidence

(b) Expected undiagnosed infections

(c) Diagnosis probabilities, state 1

(d) Diagnosis probabilities, state 2

(e) Diagnosis probabilities, state 3

(f) Diagnosis probabilities, state 4

Fig. 8.4 Age-independent back-calculation estimates; *rw*_1978 model run with a first order random walk as incidence model from the beginning of the epidemic. *rw*_1995 model run with a first order random walk as incidence model from an intermediate point (1995) of the epidemic.

8.3.4 Sensitivity analysis to the specification of the expected number of initially undiagnosed infections

A sensitivity analysis to assess the impact of the expected number of initially undiagnosed infections on the estimates of incidence and diagnosis probabilities has been conducted.

We consider $\boldsymbol{\pi}^{\star}$ used so far as the "baseline" value. Two scenarios are examined; in the first scenario, the baseline expected number of undiagnosed infections in each state, and thus the overall expected number of undiagnosed infections $\pi^{\star}_{tot}$, is multiplied by a constant κ. The second scenario considers that the total number of expected undiagnosed infections is the same as for the baseline case ($\pi^{\star}_{tot}$ = 4963.07), but the distribution of individuals across the undiagnosed states is modified; Tables 8.2 and 8.3 provide further details.

Results

Figure 8.5 displays the posterior mean and 95% credible intervals of the posterior distribution of incidence and the expected number of undiagnosed infections for all cases considered

Model	κ	$\pi^{\star}_1$	$\pi^{\star}_2$	$\pi^{\star}_3$	$\pi^{\star}_4$	$\pi^{\star}_{tot}$
Baseline	1	1710.67	1191.20	1191.20	870.00	4963.07
Case A	0.05	85.53	59.56	59.56	43.50	248.15
Case B	0.5	855.33	595.60	595.60	435.00	2481.53
Case C	0.75	496.30	992.61	1488.92	2481.53	5459.36
Case D	1.25	2138.33	1489.00	1489.00	1087.50	6203.83
Case E	1.75	2566.00	1786.80	1786.80	1305.00	7444.60
Case F	1.95	3335.80	2322.84	2322.84	1696.50	9677.98

Table 8.2 Expected number of initially undiagnosed specification, scenario 1

Model	κ_1	κ_2	κ_3	κ_4	$\pi^{\star}_1$	$\pi^{\star}_2$	$\pi^{\star}_3$	$\pi^{\star}_4$	$\pi^{\star}_{tot}$
Baseline	0.34	0.24	0.24	0.18	1710.67	1191.20	1191.20	870.00	4963.07
Case A	0.15	0.25	0.25	0.35	744.48	1240.76	1240.76	1737.07	4963.07
Case B	0.5	0.2	0.2	0.1	2481.53	992.62	992.62	496.30	4963.07
Case C	0.1	0.2	0.2	0.5	496.30	992.62	992.62	2481.53	4963.07
Case D	0.94	0.02	0.02	0.02	4665.29	99.26	99.26	99.26	4963.07
Case E	0.02	0.94	0.02	0.02	99.26	4665.29	99.26	99.26	4963.07
Case F	0.02	0.02	0.94	0.02	99.26	99.26	4665.29	99.26	4963.07
Case G	0.02	0.02	0.02	0.94	99.26	99.26	99.26	4665.29	4963.07

Table 8.3 Expected number of initially undiagnosed specification, scenario 2

under both scenarios; Figure 8.6 depicts the corresponding estimates for the diagnosis probabilities.

Looking at the incidence curve estimates, from the first scenario, it appears that from 2000 onwards the estimates coincide for all values of κ specified. Case A ($\kappa = 0.05$), which amounts to considering expected initial number of undiagnosed infections close to zero, is the only exception: the estimated incidence curve is extremely volatile in this case. In the earliest years considered, different κ values lead to different estimated expected number of infections. Lower κ values assume fewer initially undiagnosed infections, thus a higher incidence estimate is required to appropriately reconstruct the observed diagnosis data.

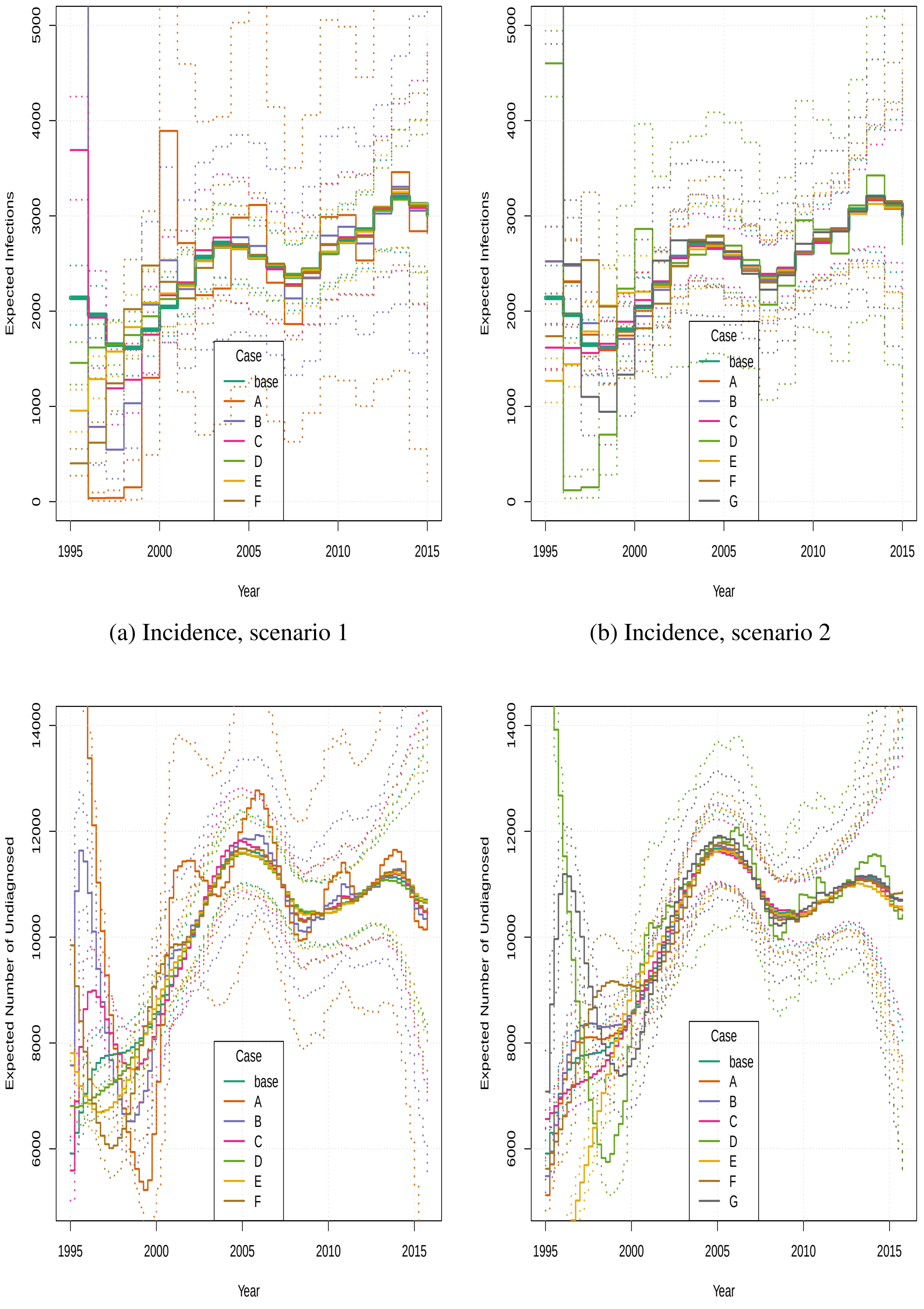
Analogously, estimates of the diagnosis probabilities (Figures 8.6a to 8.6d) only behave similarly from 2002 onwards, with the exception of Case A. Figures 8.6a to 8.6d only feature diagnosis probabilities estimates from undiagnosed states 1 and 4; state 2 and 3 estimates behave similarly (Appendix H.2).

The number of expected undiagnosed infections over time (Figure 8.5c) is also affected by assumptions on $\boldsymbol{\pi}^\star$, as this is a function of incidence and diagnosis probabilities. Estimates agree from 2002 onwards, for all cases, except for Case A.

Note that the choice of κ crucially affects the size of credible intervals for both incidence, diagnosis probabilities, and number of expected undiagnosed infections. The infection and diagnosis processes are modelled by first order random walks with constant variance, thus highly variable estimates in the early years lead to wider credible intervals throughout the epidemic.

Goodness-of-fit plots (see Figure 8.6e) are useful for assessing whether κ, and hence $\boldsymbol{\pi}^\star$, is appropriately specified. Unrealistic κ choices (*e.g.* Case A and Case F) lead to poor data fits in the earliest considered. Instead suitable choices of κ (Cases C to E) result in a satisfactory fit to the data even in the early years considered.

In the second Scenario, the estimates' sensitivity to the pre-specified distribution of $\boldsymbol{\pi}^\star$ is considered. Cases A, B and C only slightly modify the baseline distribution of initially undiagnosed infections into the four undiagnosed states. Cases D, E, F and G are more extreme; the expected number of initially undiagnosed infections is almost fully allocated to a single state. Estimates obtained for incidence, diagnosis probabilities and undiagnosed infections behave as for Scenario 1: the posterior means differ in the first years, whereas from (approximately) 2002 onwards estimates coincide and are robust to $\boldsymbol{\pi}^\star$ specifications. Credible intervals are similarly affected by the variability of estimates in the early years. Case D is an exception as incidence, diagnosis probabilities and expected undiagnosed



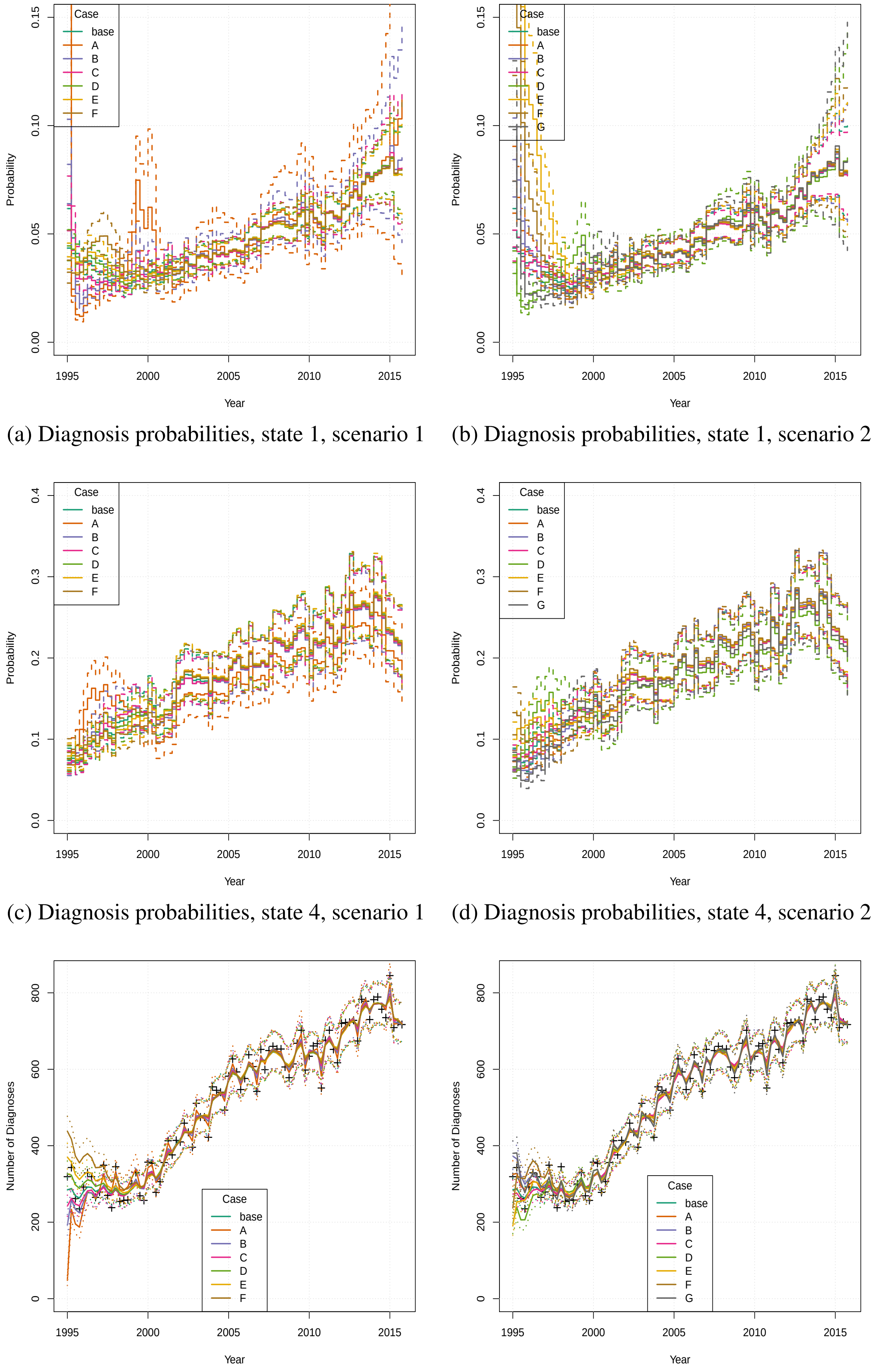
(a) Incidence, scenario 1

(b) Incidence, scenario 2

(c) Expected undiagnosed infections, scenario1

(d) Expected undiagnosed infections, scenario2

Fig. 8.5 Age-independent back-calculation estimates; Scenarios and cases are described in Tables 8.2 and 8.3.



(a) Diagnosis probabilities, state 1, scenario 1

(b) Diagnosis probabilities, state 1, scenario 2

(c) Diagnosis probabilities, state 4, scenario 1

(d) Diagnosis probabilities, state 4, scenario 2

(e) HIV, scenario 1

(f) HIV, scenario 2

Fig. 8.6 Age-independent back-calculation estimates and goodness-of-fit. Scenarios and cases are described in Tables 8.2 and 8.3.

infections estimates differ from the other cases also after 2002. The expected number of initially undiagnosed infections is fully concentrated in State 1 and somewhat unexpectedly the estimated incidence is extremely high in the first year. A large number of diagnoses with CD4-count less than 500 (*i.e.* from undiagnosed states 2, 3, 4) occur in the early years; however, by assumption, only a few initially undiagnosed infections are in such states. Consequently a large number of infections is estimated in the first year so that a sufficiently high number of individuals can be subsequently diagnosed in states 2, 3 and 4. In the late 90's, estimated incidence drops to almost zero, to avoid over-estimation of the diagnosis data. As a result, incidence and diagnosis probabilities estimates are extremely volatile throughout the epidemic.

Goodness-of-fit plots for CD4-count data (Appendix H.2) provide evidence on the appropriateness of $\boldsymbol{\pi}^{\star}$ specification. Cases D to F poorly fit CD4-count data in the early years for at least one state, whereas the baseline and cases A to C correctly describe these data.

In summary, with the exception of the first seven years considered, back-calculation starting at an epidemic mid-point yields estimates of incidence, diagnosis probabilities, and expected undiagnosed infections are robust to $\boldsymbol{\pi}^{\star}$ specifications. If the estimation of these quantities in the first seven years is key, then back-calculation should be either run from the beginning of the epidemic, or from an earlier starting point. In any case estimates from the first seven years must be interpreted with caution, and a buffer bigger than seven years can be practically used for safety reasons. $\boldsymbol{\pi}^{\star}$ misspecification inflates the credible intervals size. Finally, goodness-of-fit plots can be used to informally assess the appropriateness of assumptions on the expected number of initially undiagnosed infections.

8.4 Age-dependent back-calculation

8.4.1 A preliminary model

This Section discusses the application of the age-dependent back-calculation model (Chapter 5) to the surveillance dataset for the MSM-HIV epidemic in England and Wales (Section 8.2), further stratified by age at diagnosis. The underlying multi-state model was discussed in Section 7.2 (Figure 7.1). To begin with, age-dependent back-calculation is considered on a yearly time and age scale, constructed by aggregating quarterly intervals (Section 5.4.3), to keep the computation burden of the model manageable as in Chapter 7.

The year 1995 is assumed to be the starting point of back-calculation so that the expected number of initially undiagnosed infections $\boldsymbol{\pi}^{\star}$ is chosen to be the same as the one used in the age-dependent simulation study (Section 7.2, detail in Appendix G.1). A total of $T = 21$ and $A = 52$ yearly time and age intervals are considered, individuals are assumed to only be infected between 15 ($j = 1$) and 66 ($j = 52$) years of age. In Section 8.3.4 we have seen that the specification of $\boldsymbol{\pi}^{\star}$ only affects incidence and diagnosis probabilities estimates in the early years after the chosen starting point; further investigations (not reported) revealed that similar results hold for age-dependent back-calculation.

Progression probabilities, depending on the age at infection ($\mathcal{Q}$ in Section 5.3) are as set in Section 7.2 (details in Appendix G.1). Table 8.4 shows the mean time spent in each undiagnosed state, stratified by age at infection. Under-reporting is modelled as in Section 8.3.

Simulations (Chapter 7) have demonstrated that both thin plate splines (*ts* splines) and tensor product splines, with marginal B-splines with a first order difference penalty, (*ptensbsord1*) are appropriate to model the latent log-incidence surface. The features of the splines used (*i.e.* the number of parameters, the priors used) are described in Section 7.3. Diagnosis probabilities are modelled independently of age using logistic random walks (as in Section 7.3).

Models are implemented in `Stan`. We run four parallel chains of 2,000 iterations, of which the first 1,000 were discarded as burn-in. The resulting posterior sample of 4,000 iterations was obtained in approximately 8 hours for *ptensbsord1* and 30 hours for *ts* splines. Convergence was achieved for all models and was again assessed using trace plots and R-hat statistics.

Age at inf	State 1	State 2	State 3	State 4	Time to
a_0	($CD4 \geq 500$)	($500 < CD4 \leq 350$)	($350 < CD4 \leq 200$)	($CD4 > 200$)	AIDS
15	2.77	2.13	2.23	1.76	8.90
25	2.70	2.14	2.19	1.73	8.76
35	2.62	2.13	2.13	1.68	8.54
45	2.53	2.08	2.03	1.59	8.23
55	2.5	2.08	1.92	1.47	7.97
> 65	2.30	1.86	1.68	1.31	7.16

Table 8.4 Mean time (in years) spent in each undiagnosed state, stratified by age at infection.

Results

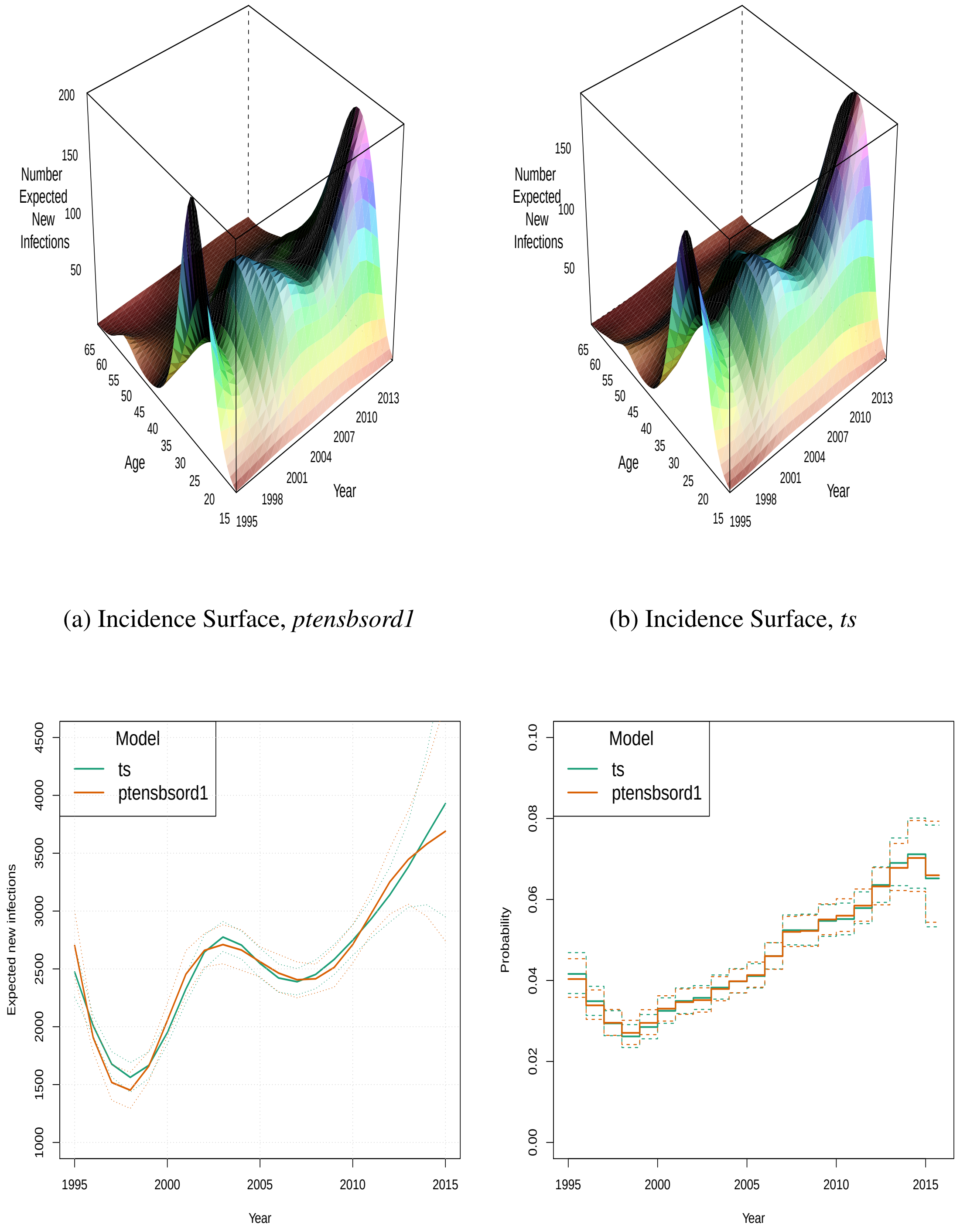
Figures 8.7 and 8.8 show results from the age-dependent back-calculation model, using both a *ptensbsord1* spline and a *ts* spline. The estimated time-and-age dependent incidence surface is depicted in Figures 8.7a and 8.7b, with Figures 8.7c, 8.8a and 8.8b showing the time and age profiles (for selected years) of the surface obtained. The *ts* spline estimates a higher incidence than the *ptensbsord1* splines in the last two years, where diagnosis data are weakly informative about infection levels. This is likely due to the properties of *ptensbsord1* splines, which a priori favour a flat time profile for the incidence surface, whereas *ts* splines favour a linear trend (see Section 7.6.2). The age-profile of infections has slightly shifted towards younger ages since 2000. This is further highlighted in Figures 8.8c and 8.8d depicting the incidence time profile by age-class. Estimated infections within the 25-34 age-class appear to be sharply increasing since 2010, whereas incidence is approximately constant in the most recent years in all other age-classes.

Figure 8.7d plots the estimated diagnosis probabilities from the $CD4 > 500$ undiagnosed state. Estimates obtained using *ts* and *ptensbsord1* splines are very similar and diagnosis probabilities have steadily increased since 2000.

It is well-known that back-calculation produces highly uncertain estimates in the most recent years and these estimates are sensitive to the amount of information available in the model. We investigate the robustness of this model to the sequential inclusion of further years of data.

Figures 8.8e and 8.8f displays the estimates of the time profile of the incidence surface and diagnosis probabilities for these sequential analyses. Denote $\hat{h}_{i,e}$ the estimate of incidence in the i^{th} year, obtained with data up to the end of the e^{th} year.

The estimated number of expected infections decrease by 500 (out of approximately 3500) between $\hat{h}_{12,12}$ and $\hat{h}_{12,13}$, and between $\hat{h}_{13,13}$ and $\hat{h}_{13,14}$. Using data up to the end of 2012 or 2013, the age-dependent model estimates approximately 4000 infections per year in 2012 and 2013. However estimates of incidence in 2012 and 2013 decrease to approximately 2500 using data up to the end of 2015, when data become informative about infection levels in 2012 and 2013. Hence the increasing trend estimated in the most recent years by the age-dependent model, using data up to the end of 2012 and 2013, is potentially artificial.



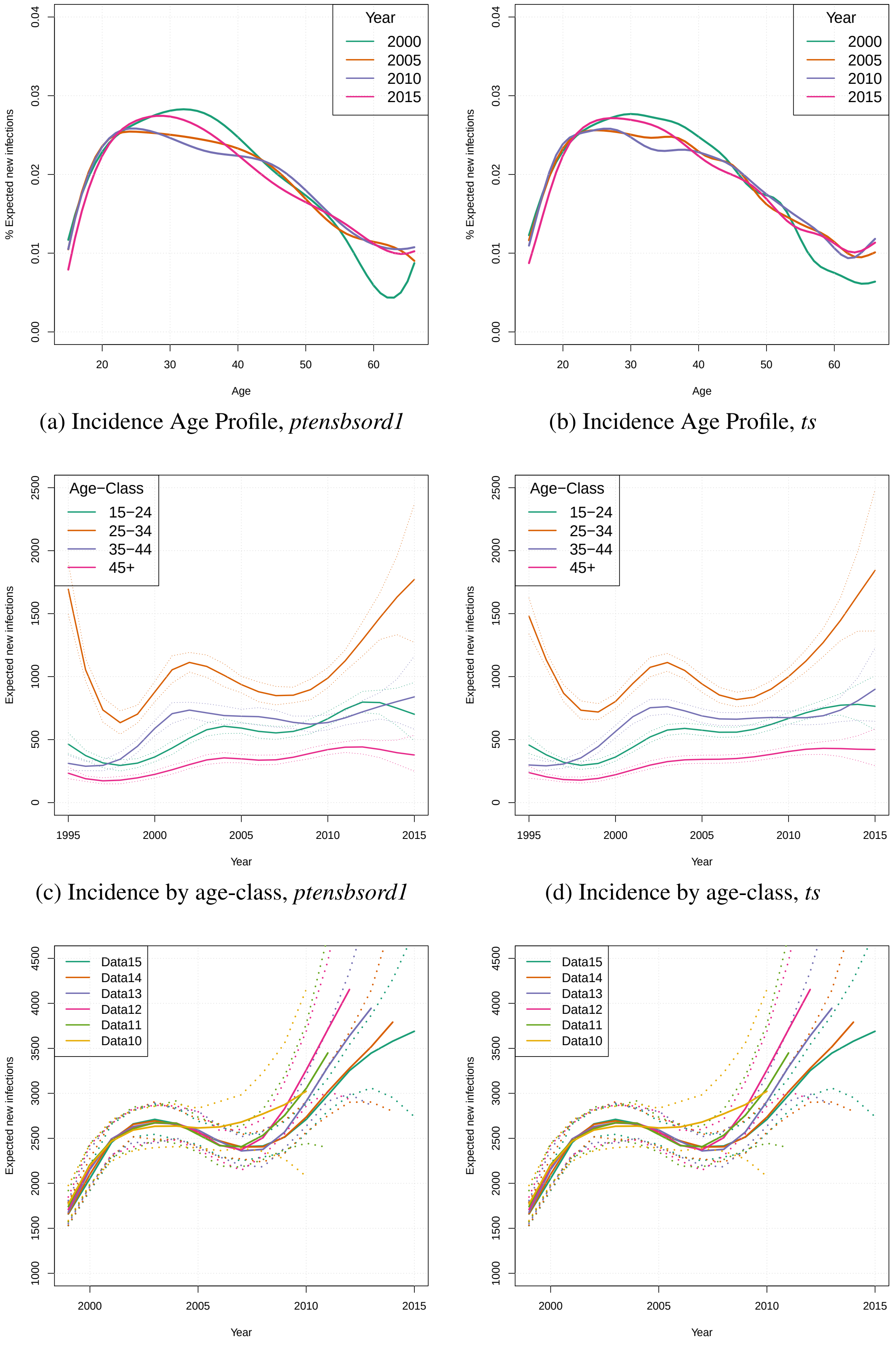
(a) Incidence Surface, *ptensbsord1*

(b) Incidence Surface, *ts*

(c) Incidence Time Profile

(d) Diagnosis probabilities, State 1

Fig. 8.7 Results from age-dependent back-calculation model; *ts* denotes a thin plate spline and *ptensbsord1* denotes tensor product splines.



(a) Incidence Age Profile, *ptensbsord1*

(b) Incidence Age Profile, *ts*

(c) Incidence by age-class, *ptensbsord1*

(d) Incidence by age-class, *ts*

(e) Model robustness, *ptensbsord1*

(f) Model robustness, *ts*

Fig. 8.8 Results from age-dependent back-calculation model; *ts* denotes a thin plate spline and *ptensbsord1* denotes tensor product splines.

8.4.2 Investigating the robustness of the age-dependent back-calculation

It is key to further investigate the lack of robustness of the model in the most recent year. Allowing for enhanced flexibility may allow the model to better adapt to recent changes in the data, but may also result in overfitting. This can be achieved by considering a finer time scale and/or relaxing the assumption that diagnosis probabilities are independent of age.

So far an age-dependent yearly back-calculation model has been considered; this is based on yearly aggregated diagnosis data and its dynamics are constructed by aggregating quarterly intervals. It is assumed that the diagnosis probabilities and the expected number of infection remains constant over the quarters of a year (Section 5.4.3). Quarterly models make a fuller use of surveillance data, by considering quarterly data and allowing incidence and diagnosis probabilities to vary quarterly.

Thus far diagnosis probabilities, for both age-independent and age-dependent back-calculation, have been modelled using time dependent first order logistic random walk for each undiagnosed state - *i.e.* $\delta_{k,i} \sim N(\delta_{k,i-1}, \sigma^2_{D,k})$, $k = \{1,\ldots,4\}$, $i = \{1,\ldots,T\}$ (see Sections 3.6.2 and 6.4). This assumes independence of the diagnosis process on the current age-interval $(a_{j-1}, a_j]$, $j = \{1,\ldots,A\}$. This strong assumption can be relaxed, by allowing the diagnosis probabilities to depend on current age. Independent random walks are considered for each state and age-class, *i.e.* $\delta_{k,i,j} \sim N(\delta_{k,i-1,j}, \sigma^2_{D,k})$ where $\delta_{k,1,j} \sim N(\alpha_j, \sigma_0)$ or $\delta_{k,1,j} \sim N(\alpha_{j,k}, \sigma_0)$. Here α_j and $\alpha_{j,k}$ are age and age-state specific intercepts respectively, which must be estimated, whereas the variance σ_0 is fixed (see Section 6.4). All intercept parameters are given a $N(0,1)$ prior. Diagnosis probabilities are assumed to be constant within four age intervals: $(a_0, a_{10}]$ (*i.e.* age 15-24), $(a_{10}, a_{20}]$ (*i.e.* age 25-34), $(a_{20}, a_{30}]$ (*i.e.* age 35-44), and $(a_{30}, a_{52}]$ (*i.e.* age 45+).

Thus six models can be considered:

1. Yearly time scale ($T = 21$) and yearly age scale ($A = 52$), age-independent diagnosis probabilities (*Yr-AiDx*). This has been discussed in the previous section.
2. Yearly time scale ($T = 21$) and yearly age scale ($A = 52$), age-dependent diagnosis probabilities with $\delta_{k,1,j} \sim N(\alpha_j, \sigma_0)$ intercept (*Yr-AdDx1*).
3. Yearly time scale ($T = 21$) and yearly age scale ($A = 52$), age-dependent diagnosis with $\delta_{k,1,j} \sim N(\alpha_{j,k}, \sigma_0)$ intercept (*Yr-AdDx2*).

4. Quarterly time scale ($T = 84$) and yearly age scale ($A = 52$), age-independent diagnosis probabilities (*Qt-AiDx*).
5. Quarterly time scale ($T = 84$) and yearly age scale ($A = 52$), age-dependent diagnosis probabilities with $\delta_{k,1,j} \sim N(\alpha_j, \sigma_0)$ intercept (*Qt-AdDx1*).
6. Quarterly time scale ($T = 84$) and yearly age scale ($A = 52$), age-dependent diagnosis with $\delta_{k,1,j} \sim N(\alpha_{j,k}, \sigma_0)$ intercept (*Qt-AdDx2*).

The abbreviations given in brackets will be subsequently used to refer to these models.

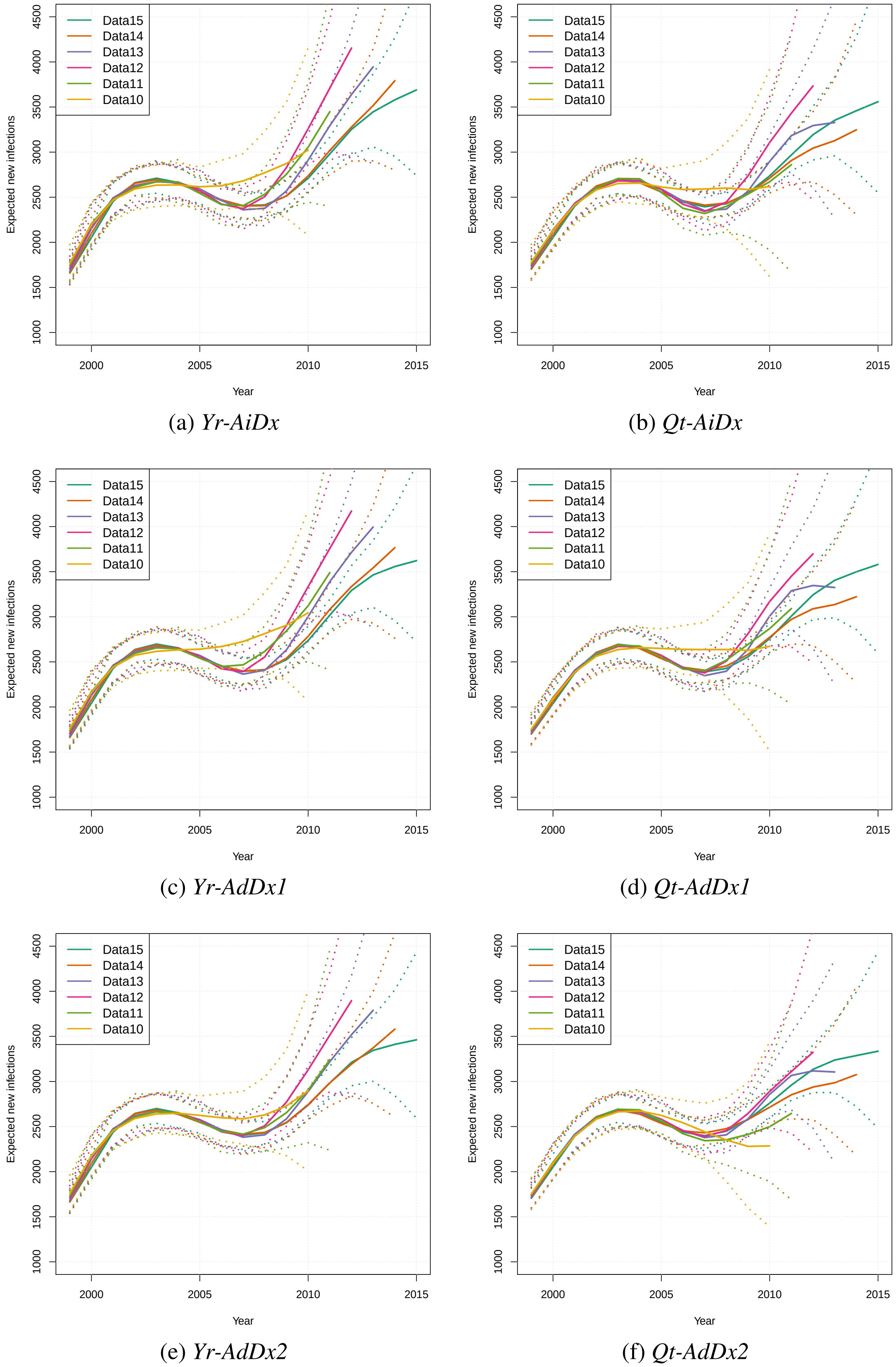
All models assume the same expected number of initially undiagnosed infections $\boldsymbol{\pi}$ and progression probabilities $\boldsymbol{Q}$, as defined in Section 8.4. However there are two main differences: the time scale employed and the parametrization chosen for the diagnosis probabilities.

Incidence is only modelled using a *ptensbsord1* spline, as the isotropic smoothing assumption, underlying the *ts* spline (see Section 7.6.2), is not appropriate when different time and age scales are considered. Note that the *ptensbsord1* spline is defined by 80 parameters for both quarterly and yearly models. Intuitively quarterly models require enhanced flexibility and thus a larger number of parameters as they allow the incidence surface to vary quarterly rather than yearly. However splines, produce similar results for any sufficiently large number of parameters (see Figure 3.4b in Section 3.3.2). In fact, we did considered larger number of parameters without, however, noting any change to the results.

Model implementation was undertaken in `Stan`. We run four parallel chains with 1000 and 2000 iterations, of which 500 and 1000 burn-in for quarterly and yearly models respectively. The posterior sample was obtained in approximately 70 and 8 hours for quarterly and yearly models with age-independent diagnosis probabilities respectively (and in 80 and 12 hours when considering age-dependent diagnosis probabilities). Note that quarterly age-dependent back-calculation models are highly computationally expensive and this may limit their applicability. Convergence was again assessed using trace plots and R-hat statistics and was achieved for all models.

Results

Figure 8.9 displays the sensitivity of the estimated time profile of the incidence surface to the sequential addition of further years of data. $\hat{h}_{12,12} \approx 4200$ for the yearly models *Yr-AiDx* and *Yr-AdDx1*. After the inclusion of data to the end of 2015, the 2012 incidence estimate ($\hat{h}_{12,15}$) is revised to ≈ 3250. From 2012, estimates in the most recent year considered appear



(a) *Yr-AiDx*

(b) *Qt-AiDx*

(c) *Yr-AdDx1*

(d) *Qt-AdDx1*

(e) *Yr-AdDx2*

(f) *Qt-AdDx2*

Fig. 8.9 Sensitivity of the estimated time profile of incidence to the sequential addition of years of data, by model considered.

to be consistently revised to a lower number when a further year of data is added. On the other hand, revised estimates $\hat{h}_{12,13}$, $\hat{h}_{12,14}$ and $\hat{h}_{12,15}$ lie within the wide credible intervals of $\hat{h}_{12,12}$. Note that the *Yr-AdDx2* model produces incidence estimates that are less sensitive to the addition of further years of data. For instance, $\hat{h}_{12,12}$ is ≈ 3750 and this is similarly revised to ≈ 3250 when using data up to the end of 2015.

The robustness of the model is further investigated using quarterly age-dependent back-calculation models. Note that the incidence estimates obtained from the yearly and quarterly model, that use the same data, are similar in the earliest years considered. Nonetheless the quarterly scale mitigates the potentially artificial increase in incidence in most recent years. For instance, as for the age-independent models, $\hat{h}_{12,15} \approx 3250$; however $\hat{h}_{12,12} \approx 3700$ for *Qt-AiDx* and *Qt-AdDx1* models (versus $\hat{h}_{12,12} \approx 4200$ for the respective yearly models *Yr-AiDx* and *Yr-AdDx1*).

Figure 8.9 demonstrates that the *Qt-AdDx2* model is the most robust among the models considered. Estimates of incidence, both at population and at age-specific level (Figure H.10, in Appendix H.4) in the last year are only slightly revised when further years of data are added, suggesting that the estimated trends in incidence are not artificial.

Goodness of fit plots (see Appendix H.4) can be further used for model selection. HIV data are fit equally well by all models, however the *Yr-AdDx2* and *Qt-AdDx2* models improve the fit to the AIDS and CD4-count data, especially in the 15-24 and 45+ age-classes. The 15-24 age-class is the only age-class with an increasing number of CD4-count diagnoses in states 2 and 3, between 2005 and 2010. The enhanced flexibility of the *Yr-AdDx2* and *Qt-AdDx2* models, using a state-and-age dependent intercept for modelling diagnosis probabilities, allows to capture this feature of the data. For all age-classes, the data posterior-predictive of CD4-count data include all data-points, but credible intervals are wide. Despite overfitting may be present, the *Qt-AdDx2* successfully achieves robust incidence estimates.

Hence this investigation shows that increased flexibility, achieved by employing a quarterly time scale and age-dependent diagnosis probabilities (*Qt-AdDx2*), overcomes the lack of robustness and improves goodness of fit of the originally discussed age-dependent back-calculation model (Section 8.4.1). Results from the *Qt-AdDx2* are presented in the following Section.

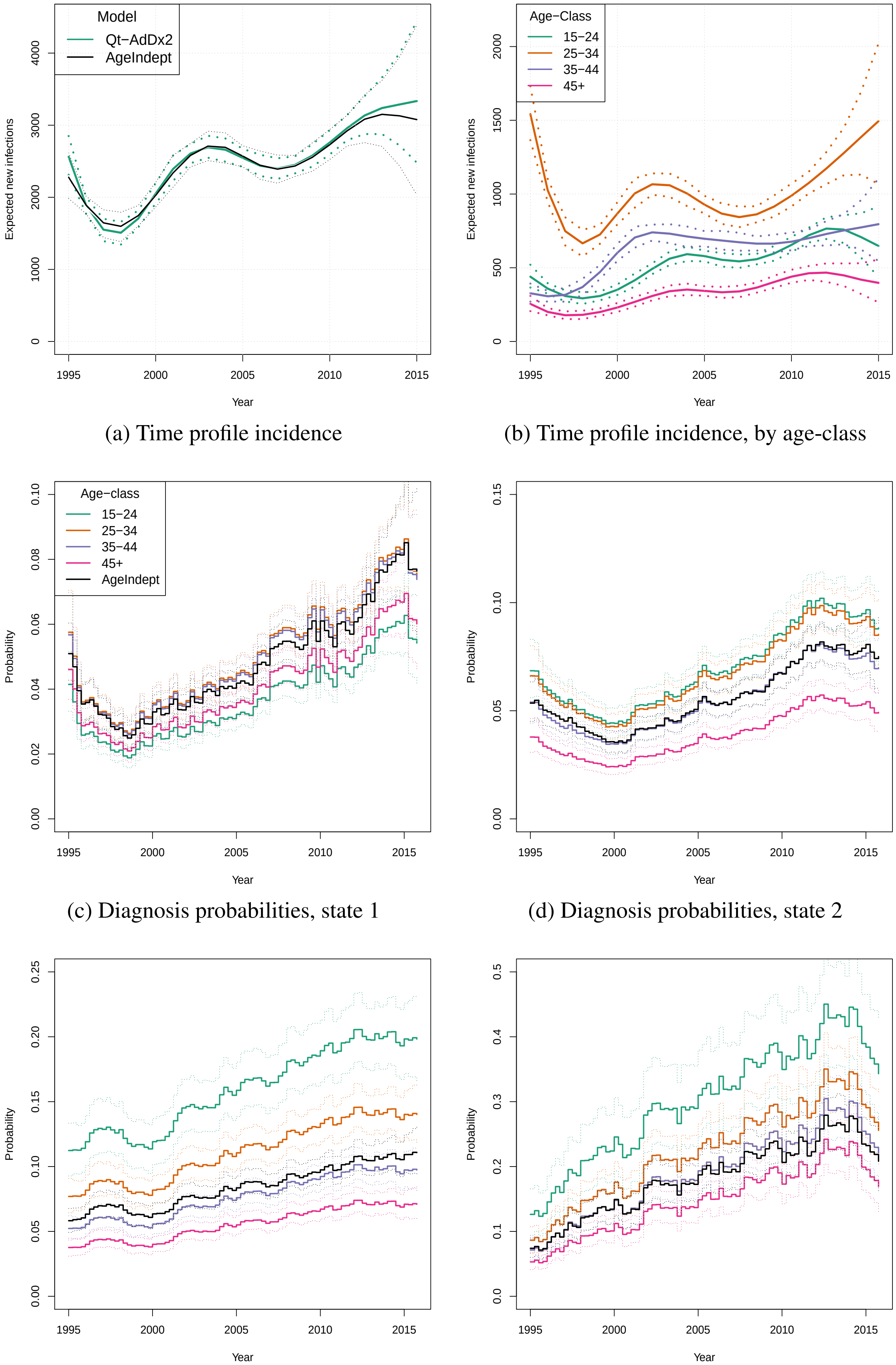
8.4.3 Results from quarterly age-dependent back-calculation with age-dependent diagnosis probabilities (*Qt-AdDx2*)

Figures 8.10 and 8.11 show results from the *Qt-AdDx2* model along with results from the age-independent model, discussed in Section 8.3. Comparing the models is not straightforward as they use different data and model the infection and diagnosis processes in different ways. Nevertheless, it is important to informally verify that the age-dependent and age-independent back-calculation yield similar estimates.

Figure 8.10a plots the expected number of infections over time. After reaching a minimum of $\approx$ 1500 (95% CrI 1350, 1650) yearly expected infections in 1998, incidence increases to $\approx$ 2700 (95% CrI 2550, 2850) expected infections in 2003. Incidence subsequently smoothly decreases to $\approx$ 2400 (95% CrI 2250, 2540) expected infections in 2007. From 2007 onwards expected infections steadily increase, even though incidence appears to reach a plateau in the three latest years. In 2015, 3335 (95% CrI 2480, 4440) new expected infections are estimated. It is reassuring to note that incidence estimates are similar to those obtained using age-independent back-calculation.

Age specific incidence is plotted in Figure 8.10b. Incidence is mostly concentrated within the 25-34 age-class, where it steadily increases from 845 expected infections (95% CrI 775, 915), in 2007, to 1495 (95% CrI 1070, 2020) expected infections in 2015. Incidence in the 15-24 age-class, increases from 540 (95% CrI 500, 590) expected infections in 2007 to 760 (95% CrI 430, 915) expected infections in 2013; despite the decreasing trend in the two last years, credible intervals remain very wide so that the possibility of decreasing incidence can not be precluded. Incidence slowly increases between 2007 and 2015, shifting from 670 (95% CrI 615, 740) to 780 (95% CrI 545, 1100) expected infections for 35-44 year olds and from 340 (95% CrI 300, 375) to 400 (95% CrI 265, 555) expected infections for 45+ year olds. Over the last fifteen years, the age at infection shifted to younger ages: in 2000, 17%, 42%, 30%, 11% of individuals were respectively newly infected in age-classes 15-24, 25-34, 35-44 and 45+, compared to 19%, 45%, 24%, 12% of individuals in 2015.

Estimated diagnosis probabilities for *Qt-AdDx2* have been gradually increasing since 2000 for all CD4 states (Figures 8.10c to 8.10f), with a further sharp increase in state 1 in the last three years. This finding is consistent with the results of the age-independent model and the observed steady increase in HIV diagnoses with a CD4 $>$ 500 since 2010 (Figure 8.1b). The diagnosis probabilities in the 25-34 age-class are higher compared to those for the 35-44 and 45+ age-groups in all undiagnosed states. However diagnosis probabilities for the 15-24 age-class behave unexpectedly: they are the lowest of all age-classes in state 1, whilst they



(a) Time profile incidence

(b) Time profile incidence, by age-class

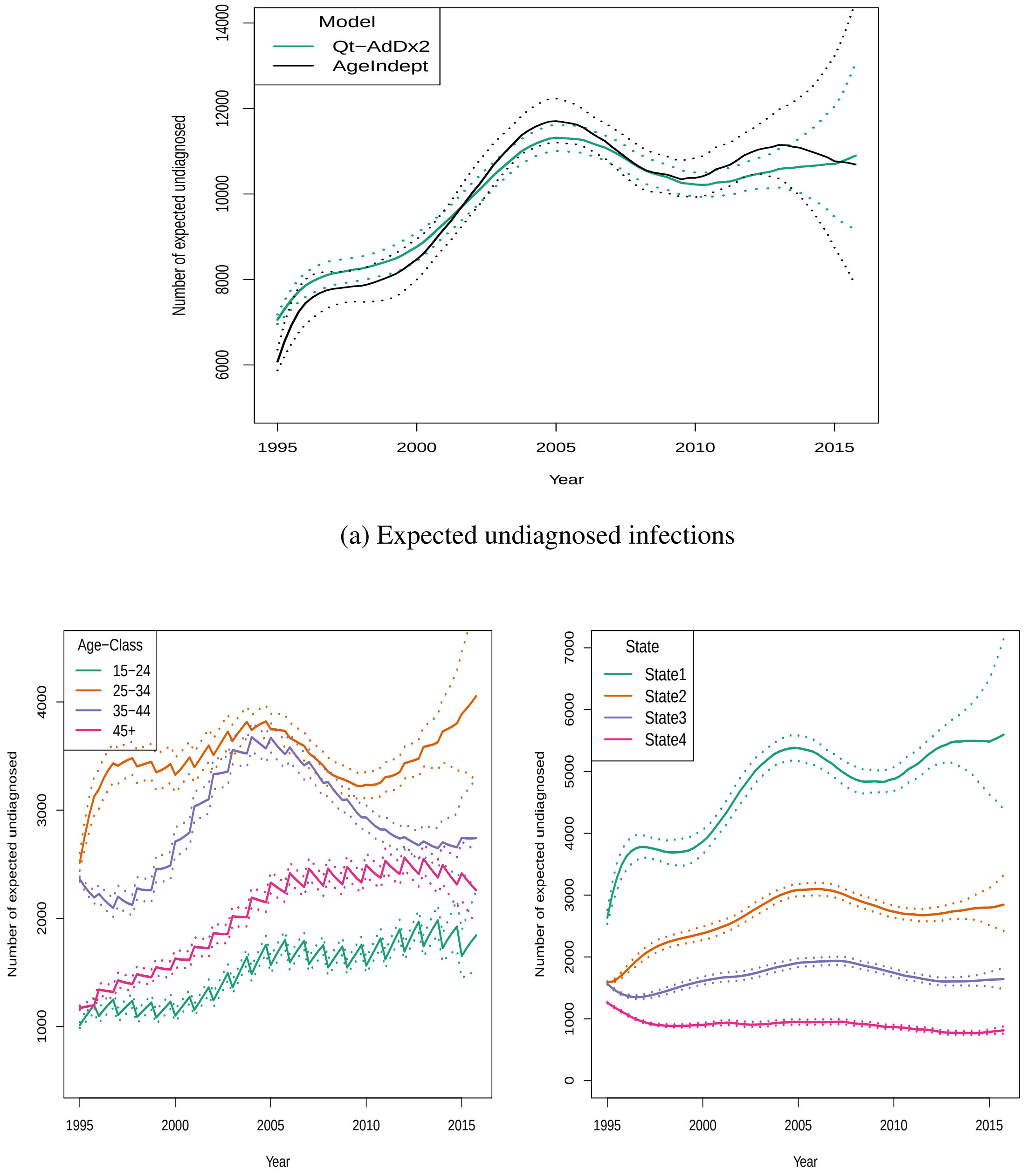
(c) Diagnosis probabilities, state 1

(d) Diagnosis probabilities, state 2

(e) Diagnosis probabilities, state 3

(f) Diagnosis probabilities, state 4

Fig. 8.10 Estimates for the *Qt-AdDx2* model of the time profile of incidence (a), the time profile stratified by age-class (b), and of diagnosis probabilities stratified by state (c to f). State 1: CD4 ≥ 500, State 2: $350 <$ CD4 ≤ 200, State 3: $350 <$ CD4 ≤ 200, State 4: CD4 < 200.



(a) Expected undiagnosed infections

(b) Expected undiagnosed infections, by age

(c) Expected undiagnosed infections, by state

Fig. 8.11 Estimates obtained for the *AdDx2* model. State 1: CD4 $\geq$ 500, State 2: 350 $<$ CD4 $\leq$ 200, State 3: 350 $<$ CD4 $\leq$ 200, State 4: CD4 $<$ 200.

are the highest of all age-class in states 2 to 4. This finding suggests that recently infected individuals aged 15-24 have a small probability of getting diagnosed, that rapidly increases as time from infection increases.

This unusual behaviour could be due to misspecification of progression probabilities. Recall that infected individuals are subject to two competing processes: progression and diagnosis pressure (Section 5.3). It is plausible that the model specification does not allow for aged 15-24 to progress rapidly enough from the first to the second undiagnosed state; hence low diagnosis probabilities are estimated. Conversely, undiagnosed 15-24 year-olds are potentially assumed by the model to progress towards lower CD4-undiagnosed states too rapidly, which in turn requires higher diagnosis probabilities. Another plausible explanation concerns the size of the 15 to 24 age-class; there is substantial heterogeneity in sexual behaviour of individuals aged 15 to 24. Teenagers (18 years of age or less) are substantially less sexually active than individuals in their twenties, and as such these two groups are associated with different risk behaviours and thus diagnosis probabilities.

Diagnosis probabilities are assumed to be constant within the 15-24, 25-34, 35-44 and 45+ age-classes. It would be more appropriate to allow diagnoses to smoothly vary with age, for instance by forming smaller age-classes and imposing some smoothing between neighbouring age-groups. Different trends over time could further be allowed for diagnosis probabilities from different age-groups. These models have not been further pursued as they require a large number of parameters, rendering computations extremely intensive.

The posterior mean of the expected number of undiagnosed infections has been approximately flat between 2010 and 2015 (see Figure 8.11). This suggests that increasing diagnosis pressure counteracts the effect of increasing infections, preventing an increase in the number of undiagnosed infections. Figure 8.11c stratifies undiagnosed infections by latent undiagnosed state. Undiagnosed prevalence mostly concerns recently infected individuals (state 1); within the last 15 years the expected proportion of long-standing undiagnosed infections has decreased, as the percentage of expected undiagnosed infections in CD4-states 1 to 4 shifted from 52%, 28%, 16% and 5% in 2000 to 56%, 27%, 13% and 4% in 2015. The expected number of undiagnosed infections has substantially decreased since 2005 in the 35-44 age-class (Figure 8.11c). However, the expected number of undiagnosed infections in the 25-34 age-class has steadily increased; thus the increase in diagnosis probabilities has not been sufficient to counteract the increase in incidence. The expected number of undiagnosed infections has been constant in the last ten years for the 15-24 and 45+ age-classes. In conclusion, the age of undiagnosed infections has decreased, probably as a consequence of increasing infections amongst younger people.

8.5 Summary

In this Chapter, we have applied age-independent (Section 8.3) and age-dependent (Section 8.4) back-calculation models to estimate HIV incidence and diagnosis probabilities for MSM in England and Wales using routinely collected surveillance diagnoses data.

In Chapters 2 and 5 we extended back-calculation to only consider a subset of the whole epidemic period, after specifying an expected number of undiagnosed infections $\boldsymbol{\pi}^\star$ at an intermediate starting point considered. Thus we carried a sensitivity analysis on the specification of $\boldsymbol{\pi}^\star$, in an age independent context (Sections 8.3.3 and 8.3.4). The results obtained when back-calculation is run from the beginning of the epidemic or from an intermediate point typically match (except for extreme cases) after a number of years from the intermediate starting point considered. In our application, seven years are sufficient. Similar results (despite not reported) have been obtained in an age-dependent context.

We then considered age-dependent back-calculation, starting from the yearly model with age-independent diagnosis probabilities employed for the simulations in Chapter 7. However, when applied to real data, this model shows a lack of robustness to the sequential addition of further years of data; this is likely due to unidentifiability in the most recent years. This drawback was circumvented by considering a more flexible model (*Qt-AdDx2*) based on a quarterly time scale and allowing diagnosis probabilities to depend on age.

Despite it is well known that incidence estimates obtained with back-calculation are highly sensitive to the addition of data, no formal methods to assess model robustness have been considered in the back-calculation literature. Subsequently adding further years of data is a step in this direction.

It would be interesting to formally investigate the differences between the quarterly and the yearly, and between the age-dependent and the age-independent back-calculation models (Section 8.4.2) within a simulation study. However long running times for quarterly models ($\approx$ 80 hours) prevented us from doing so. Despite this is a critical issue when the model has to be fitted a large number of times (*e.g.* when building and testing the model), this is not crucial for public health purposes. In fact HIV is not a rapidly evolving epidemic and surveillance data are typically updated quarterly, thus back-calculation must only be occasionally run.

Chapter 9

Conclusions and further work

Over 35 years have elapsed since the beginning of the HIV pandemic, yet the global fight to control the pandemic is far from over. Despite the encouraging reduction in the number of infections globally, from approximately 3.2 million in 2000, to 2.1 million in 2015, and the 18-fold increase in the HAART uptake from 1 million in 2000 to 18.2 million in 2015, HIV still poses a great challenge for public health (UNAIDS, 2016).

The core target of the Joint United Nations Programme on HIV/AIDS (UNAIDS) is to eradicate the AIDS pandemic by 2030; in absence of a cure, the only way this can be achieved is by eliminating HIV transmission. Therefore, to achieve further improvements, it is crucial that the populations at highest risk of infection are identified, interventions are targeted to these populations, and their efficacy is evaluated. A number of statistical models have been developed in the last three decades to tackle these issues; however existing approaches have several limitations and must be continuously re-evaluated and extended, both to better characterise the HIV epidemic, and to allow for the use of newly available, more informative data sources.

9.1 Thesis contributions

This thesis focuses on extending back-calculation methods, in order to address limitations of currently applied formulations. Back-calculation plays a vital role in monitoring the HIV epidemic, as it can be used to reconstruct historical trends in HIV incidence based on routinely collected registries of confirmed HIV and AIDS diagnoses (Chapter 1). In particular the method can accommodate different national surveillance data, that are increasingly becoming available. The CD4-staged back-calculation methods by Sweeting et al. (2005) and Birrell et al. (2012) improve upon the original formulations of back-calculation (Section 1.4), by incorporating CD4-count data, taken around HIV diagnosis; these act as a marker of the time-since-infection. Together with knowledge of the natural history of HIV infection these contribute to the estimation of HIV incidence at a population-level, whilst also permitting estimating trends in diagnosis probabilities.

Nevertheless, what is crucial for public health purposes, is the estimation of incidence for specific age-groups; research on age-dependent back-calculation is scarce and the few models available suffer, in practice, from important implementation issues (Section 1.5). In this thesis we identified and addressed this methodological gap by developing the first, to our knowledge, Bayesian CD4-based multi-state back-calculation model, allowing the joint estimation of age and time specific HIV incidence, as well as age and time specific diagnosis probabilities. We further propose novel extensions that can handle surveillance data only available on a coarse scale, or from an intermediate point of the epidemic; these enhance the method's applicability by further allowing for its use in countries with less rich surveillance data (Chapter 5).

Estimates of incidence in most recent years are subject to a lack of identifiability due to diagnosis data being only weakly informative, and smoothing methods are essential to improve their identifiability. However, existing age-specific back-calculation models have modelled the incidence surface using strong multiplicative assumptions, non-smoothed and smoothed step functions. This thesis proposes smoothing methods (tensor product splines and Gaussian processes) that, unlike previous approaches, allow for differential smoothing in the time and age-dimension (Chapter 6) and continuous modelling of the incidence surface, yielding age-specific estimates on a fine-level of detail (52 age-classes). Furthermore we put an emphasis on comparing these non-parametric bivariate smoothing methods within an age-dependent back-calculation framework. The appropriateness of these methods was established via extensive simulation studies (Chapter 7).

Despite age-dependent back-calculation being the methodology focus of this thesis, novel extensions of age-independent back-calculation have also been proposed. The current state-of-the-art model by Birrell et al. (2012) is extended so that it can be used even when coarse or incomplete (or both) surveillance data are available (Chapter 2). Several modelling options for the incidence curve are further proposed (Chapter 3); of these only Gaussian processes have not been previously considered in the back-calculation literature. However these methods have only been studied in isolation in the past, and there we undertook a comprehensive simulation study in Chapter 4 to understand whether some methods are more appropriate than others.

Finally, a fundamental objective of this work was to ensure the computational feasibility of the proposed methods, an issue that has previously hindered the development of back-calculation methods. Hence considerable effort has been devoted to implement these computationally intensive models using `Stan` (Stan Development Team, 2016b); this is a widely applicable and flexible probabilistic programming language and avoids the need of writing bespoke MCMC algorithms. The run time for the age-independent model plummeted from eight hours (Birrell et al., 2012) to merely ten minutes. The implementation of the age-dependent back-calculation model is more challenging; for instance, it requires four days to run if the combination of a quarterly time and a yearly age scale is considered. Nevertheless, it is important to highlight that, despite time-consuming, implementation via `Stan` has rendered CD4-based age-dependent back-calculation feasible, which was not possible previously.

9.2 Main thesis findings

This thesis thoroughly investigated the properties of CD4-based multi-state back-calculation, both in age-independent and an age-dependent framework. The implementation of multiple non-parametric smoothing methods in order to model HIV incidence has been been examined in two extensive Bayesian simulation studies and a maximum penalised likelihood one (in an age-dependent context). The main findings are briefly recounted in the following Sections.

9.2.1 Age independent back-calculation methods

Past trends in incidence and diagnosis probabilities are accurately reconstructed, with the exception of the last three years of surveillance data, where incidence and diagnosis probabilities (from the $CD4 > 500$ undiagnosed state, concerning recent infections) are over and underestimated respectively in the last three years. This is due to surveillance diagnosis data typically being uninformative about recent infection levels.

Among all the non-parametric smoothing methods investigated first order random walks, thin plate regression splines with shrinkage, and cubic B-splines with a first order difference penalty are the best-suited to model incidence, whereas second order random walks, knots-based thin plate splines, thin plate regression splines, and cubic B-splines with a second order difference penalty encountered convergence issues. Zero mean Gaussian processes revert to the prior mean in most recent years, yielding a potentially artificial decreasing trend in incidence. We further showed that the specification of priors has little impact on the incidence and diagnosis probabilities estimates obtained.

9.2.2 Age dependent back-calculation methods

Inference has been carried in a frequentist and in a Bayesian framework, and estimates obtained within the two frameworks are comparable. Bayesian inference has a number of advantages (Section 6.4.4): the credible intervals and an estimate of the optimal smoothing parameters can be obtained without resorting to a large-sample approximation. Moreover computational constraints do not hinder (as in the frequentist framework) the computational feasibility of tensor product splines and, finally, the Bayesian framework allows for a coherent propagation of uncertainty for the parameters of interest and model derived quantities.

As with the age-independent case, the time profile of the incidence surface is accurately reconstructed with the exception of the last three years of data, where the time profile of incidence and diagnosis probabilities (from the $CD4 > 500$ undiagnosed state) are over and underestimated respectively in the last three years. Note that the age profile of the incidence surface is accurately captured.

The simulation results suggest thin plate splines with shrinkage and tensor product splines (constructed by using cubic B-splines with a first order difference penalty as marginal splines) are the best-suited for modelling incidence. We suggest, however, to employ tensor-product splines; even though thin plate splines perform better in terms of predictive mean squared error, they require an isotropy assumption that can not be verified and that does not hold when different time and age scales are considered.

9.2.3 Back-calculation in practice

Key findings on the applicability of the proposed methods based on the HIV MSM epidemic in England and Wales, are here discussed.

The age-dependent and age-independent models produce consistent incidence and diagnosis probabilities estimates, with the exception of the last four years of the epidemic, where the age-dependent back-calculation model yields higher incidence and lower diagnosis probabilities (from the $CD4 > 500$ undiagnosed state) than the age-independent model. This difference is likely due to the surveillance data used; in the last five years, the number of HIV diagnoses is sharply increasing in the 25-34 age-class (the largest in size, 45% of all cases), whereas the total number of HIV diagnoses is only overall slowly increasing.

Age-dependent back-calculation has been considered both on a yearly and a quarterly scale, and with age-independent as well as age-dependent diagnosis probabilities; these models yield different estimates in the latest years. Model selection has been (informally) undertaken on the basis of the robustness of incidence estimates in most recent years to the subsequent addition of further years of data in the model, as these have been found to be quite volatile in some cases. The age-dependent model characterised by a quarterly time and a yearly age scale, and diagnosis probabilities with age and state dependent intercept, produced robust estimates (this model was denoted *Qt-AdDx2* in Chapter 8). We recommend the use of this model model and we stress the importance of both taking into account wide credible intervals for the incidence estimates in most recent years and analysing the robustness of such estimates to the addition of further years of data. Age-dependent models on a yearly time scale and on a quarterly time scale without age-dependent diagnosis probabilities should be avoided as they yield unrobust incidence estimates.

Incidence has been increasing since 2010, reaching a plateau of approximately 3000 expected infections per year in the last two years (2014-2015); this finding is supported by both the age-independent and the age-dependent *Qt-AdDx2* model. Rather worryingly, incidence appears to be steadily increasing in the 25-34 age-class, the most prominent in the data, whereas it flattens out in the remaining age-classes. Diagnosis probabilities for recent infections (*i.e.* from the $CD4 > 500$ state) are increasing. Allowing for age-dependent diagnosis probabilities suggests that diagnosis pressure is higher in the 25-34 age-class. The expected number of undiagnosed HIV-infected individuals has been approximately constant since 2010; this is likely a consequence of the increase in the probability of being diagnosed counteracting the increase in HIV incidence.

If back-calculation models are considered on a reduced time scale, the incidence, diagnosis probabilities estimates are sensitive to the specification of the initially undiagnosed number of individuals for the first seven years only. Thus we recommend considering at least the last fifteen years of epidemic data, and estimates obtained for the first seven years must be interpreted with caution.

9.3 Future work

There are two main obstacles to the widespread usage of back-calculation models. The first obstacle concerns data, rather than model limitations. High-quality surveillance data are not always available, or are limited in the amount of information recorded; for instance the CD4-count around diagnosis is not recorded by some surveillance systems. The second obstacle is the complexity of the model proposed; implementation is challenging, and fitting is often computationally burdensome; thus the degree to which these models can be extended is determined by the capabilities and limitations of software currently available.

Possible extensions of back-calculation models, described in the following Sections, include modelling mortality and migration, incorporating newly available data characterising recent infections sources, integration with other statistical methods to better monitor the epidemic, and software developments.

9.3.1 Mortality and migration

The current back-calculation model assumes that newly infected HIV individuals will subsequently be HIV (or AIDS) diagnosed. In reality, HIV-infected individuals may die of outside causes before diagnosis. Similarly, they may migrate. Ignoring these effects may well lead to under-estimation of infection levels. The converse is true when the number of HIV-infected individuals entering the country is sufficiently large. Whilst the effects of migration are considered negligible when working with data from the MSM exposure group, back-calculation is not currently applied to the heterosexual HIV epidemic in England and Wales due to the non-ignorable number of imported infections (mostly from sub-Sahran Africa).

To account for the impact of either requires external information. Mortality could be incorporated with the inclusion of a ‘death’ state in the multi-state model, with transition rates to death informed from mortality registries. However, quantifying the impact of migration on diagnoses is particularly challenging as HIV incidence is highly heterogeneous among different migrant populations (UNAIDS, 2016). Without information on the prevalence of undiagnosed HIV among migrants, the best that can be achieved is to estimate infection rates amongst people who will be eventually diagnosed in England and Wales.

9.3.2 Spatial modelling of the infection process

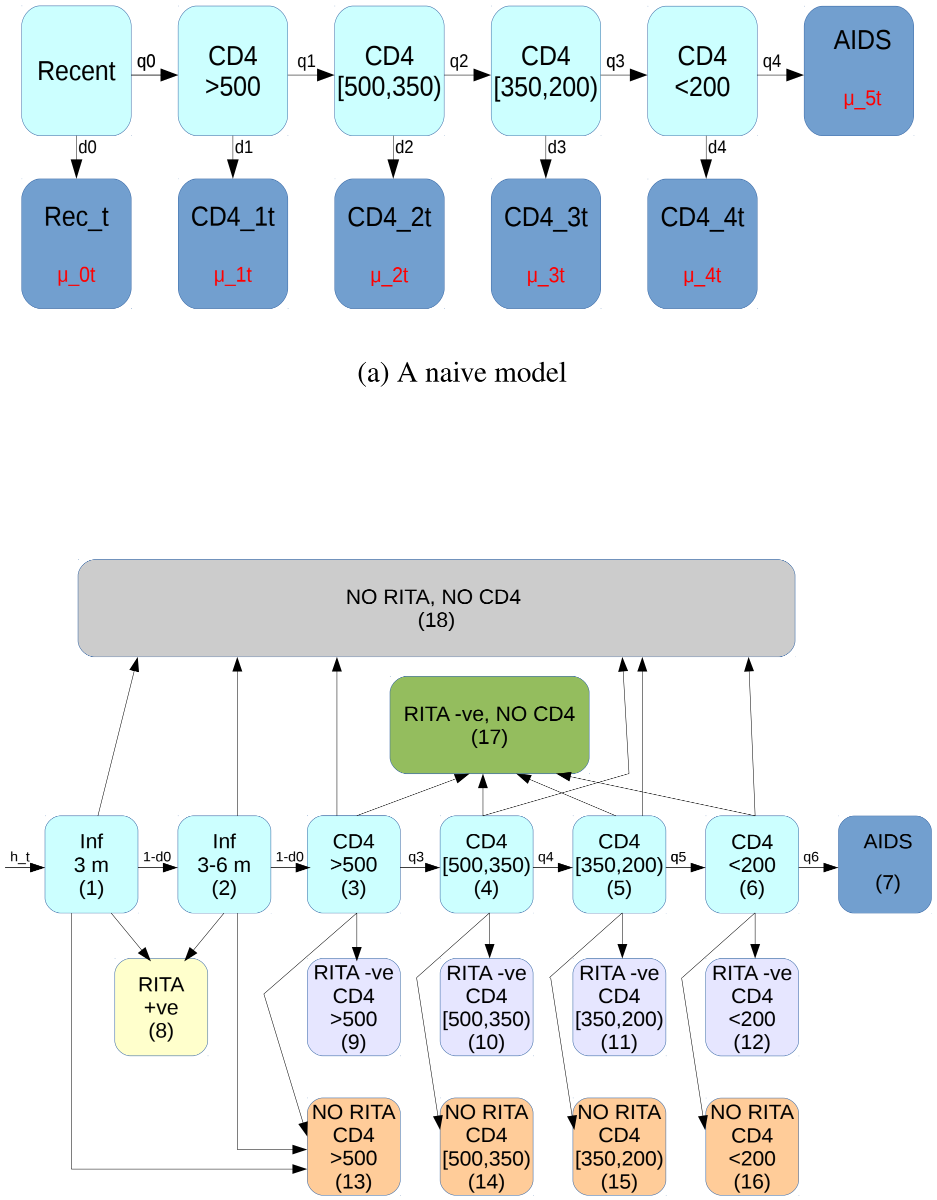
We may be further interested in identifying infection hot-spots, areas where testing rates are low and to assess the targeting of localised public health interventions. This could be achieved through stratifying the diagnosis data by regions and considering spatio-temporal modelling of the infection process. However, the infection and diagnosis events may occur in different regions, as a consequence of the long time elapsing between infection and diagnosis. Thus incorporating information on internal migration (i.e. within the country of interest) becomes crucial. This could be achieved by, for example in England and Wales, the use of Office for National Statistics (ONS) data on internal migration trends (ONS, 2015).

9.3.3 Incorporating new biomarker data on recent infections

To address the limitation of back-calculation whereby incidence in the most recent years is difficult to estimate with any certainty, the proposed back-calculation model could be further extended to incorporate data from tests for the presence of biomarkers whose levels can be indicative of recent infection. Such approaches have been implemented in different healthcare settings (Ndawinz et al., 2011; Yan et al., 2011).

In England and Wales, in 2009, PHE introduced the routine application of Recent Infection Testing Algorithms (RITA) to new HIV diagnoses, allowing for the identification of ‘recent’ infections (i.e. in the 6 months prior to testing, Aghaizu et al., 2014). These data could eliminate much of the uncertainty around recent estimates of incidence. The existing multi-state backcalculation framework is relatively easy to adapt to use such data, through the addition of "recent infection" undiagnosed states for newly infected individuals to pass through prior to reaching the CD4 states of the current model. Figure 9.1a illustrates a naive adaptation of the model. In practice, however, modifying the age-independent back-calculation to allow for the incorporation of RITA data is not so straightforward. RITA and CD4-count data are incomplete, so that only a sample of new HIV diagnoses have an associate RITA result (approximately 25% in 2009, up to 50% in 2012) and only a sub-sample of these has an associated CD4-count. As can be seen from Figure 9.1b, it is the CD4-counts of the non-early diagnoses that are of interest, but we do not know which these are unless a RITA test is also taken.

Model formulations that will fully exploit all the information held in both the RITA testing data and the CD4-counts are currently being explored. This will involve adding additional



(a) A naive model

(b) A more complicated model

Fig. 9.1 Models for incorporating data on recent infections

states to the CD4-structured model, potentially leading to an increased computational demand and complexity. Figure 9.1b illustrates one such a proposed complex model.

9.3.4 Long term developments

As a longer term aim, back-calculation may be integrated within a prevalence-based MPES framework (see Section 1.3 and Presanis et al., 2011). These are two complementary compartmental models, informed by distinct datasets, that provide independent assessments of incidence. A composite model would account for both transmission and disease progression simultaneously, yielding incidence estimates subject to less bias, and consequently an improved understanding of the disease stage at which transmission occurred.

An alternative route that merits further investment in research concerns individual-level back-calculation, or other individual-based simulation methods (De Angelis et al., 1998; Taffe and May, 2008; Fellows et al., 2015; Nakagawa et al., 2017). Since 2010, linkage of the date of the most recently available negative HIV test with the date of the first positive test is undertaken by PHE for new HIV diagnoses. This information provides an extremely valuable lower bound on the time of infection. However these data pose several challenges; missingness is likely to be informative and to be associated with infrequent testers. Further research is required to establish whether infrequent are more likely than frequent testers to have a long-standing infection; if this would be the case the propensity of each individual to be tested would also need to be modelled.

9.3.5 Software

As demonstrated throughout this thesis, the implementation of back-calculation models can be a challenging task particularly for statisticians without relevant expertise, due to both the mathematical and computational complexity of these methods. Hence it is crucial, to increase the wide applicability of these methods, to provide detailed practical guidance and general purpose software to facilitate the implementation of back-calculation models.

As far as software is concerned, to our knowledge, there are two implementations of back-calculation in R; the *backprojNP* function in the **surveillance** package (Meyer et al., 2017) and the **hivbackcalc** package (Fellows, 2017). The former only implements the simple AIDS based back-calculation by Becker et al. (1991), and the latter only implements the individual-level back-calculation method by Fellows et al. (2015) that requires the date of last negative HIV test for all individuals to be available.

Hence there is a gap concerning software for back-calculation. We support, as a key step forward, the construction of an R package accommodating the implementation of back-calculation methods for a wide range of surveillance data. A potential strategy to achieve this would be constructing an R function that would allow users to specify a bespoke back-calculation model (*e.g.* multi-state or simple, quarterly or yearly, age-specific or not). This R function would first automatically generate Stan code for the back-calculation model specified, and subsequently fit the model; these feature of the R function would be appealing to statisticians with limited, or no experience with the Stan language. This idea is central to the **rstanarm** and **brms** packages (Stan Development Team, 2016a; Bürkner, 2017).

Future research should also focus on how computational aspects of the model could be enhanced; parallelisation (potentially using GPUs) and other efficient inference method should be investigated, such as sequential Monte Carlo (Doucet et al., 2001), approximate Bayesian computation (Beaumont et al., 2002) and variational Bayes (Blei et al., 2017).

References

Aalen, O. O., Farewell, V. T., De Angelis, D., and Day, N. E. (1994). The Use of Human Immunodeficiency Virus Diagnosis Information in Monitoring the Acquired Immune Deficiency Syndrome Epidemic. *Journal of the Royal Statistical Society. Series A (Statistics in Society),*, 157(1):3–16.

Aalen, O. O., Farewell, V. T., De Angelis, D., Day, N. E., and Gill, O. N. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in medicine*, 16(January 1996):2191–2210.

Ades, A. (1995). Serial hiv seroprevalence surveys: interpretation, design, and role in hiv/aids prediction. *Journal of acquired immune deficiency syndromes and human retrovirology: official publication of the International Retrovirology Association*, 9(5):490–499.

Ades, A. and Medley, G. (1994). Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: Hiv in new york city. *Statistics in medicine*, 13(18):1881–1894.

Aghaizu, A., Murphy, G., Tosswill, J., DeAngelis, D., Charlett, A., Gill, O., Ward, H., Lattimore, S., Simmons, R., and Delpech, V. (2014). Recent infection testing algorithm (rita) applied to new hiv diagnoses in england, wales and northern ireland, 2009 to 2011. *Euro Surveill*, 19(2):20673.

Alioum, A., Commenges, D., Thiébaut, R., and Dabis, F. (2005). A multistate approach for estimating the incidence of human immunodeficiency virus by using data from a prevalent cohort study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):739–752.

Alkema, L., Raftery, A. E., and Clark, S. J. (2007). Probabilistic projections of hiv prevalence using bayesian melding. *The Annals of Applied Statistics*, pages 229–248.

An, Q., Kang, J., Song, R., and Hall, H. I. (2015). A Bayesian hierarchical model with novel prior specifications for estimating HIV testing rates. *Statistics in Medicine*, (October 2014):n/a–n/a.

Anderson, R. M., May, R. M., and Anderson, B. (1992). *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library.

Bacchetti, P. and Moss, A. R. (1989). Incubation period of aids in san francisco. *Nature*, 338(6212):251–253.

Bacchetti, P., Segal, M. R., and Jewell, N. P. (1993). Backcalculation of hiv infection rates. *Statistical Science*, pages 82–101.

Bacchetti, P. R., Segal, M. R., and Jewell, N. P. (1992). Uncertainty about the incubation period of aids and its impact on backcalculation. In *AIDS Epidemiology*, pages 61–80. Springer.

Baeten, J. M., Donnell, D., Ndase, P., Mugo, N. R., Campbell, J. D., Wangisi, J., Tappero, J. W., Bukusi, E. A., Cohen, C. R., Katabira, E., et al. (2012). Antiretroviral prophylaxis for hiv prevention in heterosexual men and women. *N Engl J Med*, 2012(367):399–410.

Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., and Dauguet, C. (1983). Isolation of t-lymphotropic retrovirus from a patient at risk for acquired immune defficiency syndrome (aids). *Science*, 220(1):868–871.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in medicine*, 35(11):1848–1865.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Becker, N. G., Lewis, J. J. C., Li, Z., and McDonald, A. (2003). Age-specific back-projection of HIV diagnosis data. *Statistics in medicine*, 22(13):2177–2190.

Becker, N. G. and Marschner, I. C. (1993). A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. *Biometrika*, 80:165–178.

Becker, N. G. and Marschner, I. C. (2001). Advances in medical statistics arising from the aids epidemic. *Statistical methods in medical research*, 10(2):117–140.

Becker, N. G., Watson, L. F., and Carlin, J. B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in medicine*, 10(February):1527–1542.

Belitz, C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics & Data Analysis*, 53(1):61–81.

Bellocco, R. and Marschner, I. C. (2000). Joint analysis of hiv and aids surveillance data in back-calculation. *Statistics in medicine*, 19(3):297–311.

Bellocco, R. and Pagano, M. (2001). Multinomial analysis of smoothed HIV back-calculation models incorporating uncertainty in the AIDS incidence. *Statistics in Medicine*, 20(August 1999):2017–2033.

Betancourt, M. (2017a). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M. (2017b). Diagnosing biased inference with divergences. http://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html. Accessed: 2017-09-01.

Birrell, P. J., Chadborn, T. R., Gill, O. N., Delpech, V. C., and De Angelis, D. (2012). Estimating Trends in Incidence, Time-to-Diagnosis and Undiagnosed Prevalence using a CD4-based Bayesian Back-calculation. *Statistical Communications in Infectious Diseases*, 4(1).

Birrell, P. J., Gill, O. N., Delpech, V. C., Brown, A. E., Desai, S., Chadborn, T. R., Rice, B. D., and De Angelis, D. (2013). HIV incidence in men who have sex with men in England and Wales 2001-10: A nationwide population study. *The Lancet Infectious Diseases*, 13(4):313–318.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

Bowman, A. and Evers, L. (2013). Lecture notes on nonparametric smoothing for academy for phd training in statistics (apts).

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, 50(4):967–991.

Brookmeyer, R. (1991). Reconstruction and future trends of the aids epidemic in the united states. *Science(Washington)*, 253(5015):37–42.

Brookmeyer, R. and Damiano, A. (1989). Statistical methods for short-term projections of aids incidence. *Statistics in Medicine*, 8(1):23–34.

Brookmeyer, R. and Gail, M. (1986). Minimum size of the acquired immunodeficiency syndrome (aids) epidemic in the united states. *The Lancet*, 328(8519):1320–1322.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the aids epidemic. *Journal of the American Statistical Association*, 83(402):301–308.

Brookmeyer, R. and Gail, M. H. (1994). *AIDS epidemiology: a quantitative approach*. Oxford University Press on Demand.

Brookmeyer, R. and Liao, J. (1990). Statistical modelling of the aids epidemic for forecasting health care needs. *Biometrics*, pages 1151–1163.

Brookmeyer, R. and Quinn, T. C. (1995). Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American journal of epidemiology*, 141(2):166–172.

Brown, A. E., Kall, M. M., Smith, R. D., Yin, Z., Hunter, A., and Delpech, V. C. (2012). Auditing national hiv guidelines and policies: the united kingdom cd4 surveillance scheme. *The open AIDS journal*, 6(1).

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

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Carlin, J. and Gelman, A. (1993). Assessing uncertainty in backprojection: comment on the paper by bacchetti.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The Stan Math Library: Reverse-Mode Automatic Differentiation in C++. *ArXiv e-prints*.

Carstensen, B. (1996). Regression models for interval censored survival data: application to hiv infection in danish homosexual men. *Statistics in Medicine*, 15(20):2177–2189.

CASCADE Collaboration (2000). Changes in the uptake of antiretroviral therapy and survival in people with known duration of hiv infection in europe: results from cascade. *HIV medicine*, 1:224–31.

Centers for Disease Control (1981a). Kaposis sarcoma and pneumocystis pneumonia among homosexual men–new york city and california. *Morbidity and mortality weekly report*, 30(25):305–8.

Centers for Disease Control (1981b). Pneumocystis pneumonia–los angeles. *Morbidity and mortality weekly report*, 30(21):250–2.

Centers for Disease Control (1982). Update on acquired immune deficiency syndrome (aids)–united states. *Morbidity and mortality weekly report*, 31(37):507.

Chadborn, T. R., Delpech, V. C., Sabin, C. A., Sinka, K., and Evans, B. G. (2006). The late diagnosis and consequent short-term mortality of hiv-infected heterosexuals (england and wales, 2000–2004). *Aids*, 20(18):2371–2379.

Chan, A. B. and Dong, D. (2011). Generalized gaussian process models. In *CVPR*, pages 2681–2688.

Chau, P. H., Yip, P. S. F., and Cui, J. S. (2003). Reconstructing the Incidence of Human Immunodeficiency Virus (HIV) in Hong Kong by Using Data from HIV Positive Tests and Diagnoses of Acquired Immune Deficiency Syndrome. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 52(2):237–248.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Conti, S., Presanis, A. M., van Veen, M. G., Xiridou, M., Donoghoe, M. C., Stengaard, A. R., and De Angelis, D. (2011). Modeling of the hiv infection epidemic in the netherlands: A multi-parameter evidence synthesis approach. *The Annals of Applied Statistics*, pages 2359–2384.

Coovadia, H. M., Rollins, N. C., Bland, R. M., Little, K., Coutsoudis, A., Bennish, M. L., and Newell, M.-L. (2007). Mother-to-child transmission of hiv-1 infection during exclusive breastfeeding in the first 6 months of life: an intervention cohort study. *The Lancet*, 369(9567):1107–1116.

Cori, A., Pickles, M., van Sighem, A., Gras, L., Bezemer, D., Reiss, P., and Fraser, C. (2015). CD4+ cell dynamics in untreated HIV-1 infection: overall rates, and effects of age, viral load, sex and calendar time. *AIDS (London, England)*, 29(18):2435–46.

Cox, D. R. and Isham, V. (1980). *Point processes*, volume 12. CRC Press.

Cui, J. S. and Becker, N. G. (2000). Estimating HIV incidence using dates of both HIV and AIDS diagnoses. *Statistics in Medicine*, 19(9):1165–1177.

Day, N., Gore, S., McGee, M., and South, M. (1989). Predictions of the aids epidemic in the uk: the use of the back projection method. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pages 123–134.

De Angelis, D., Day, N. E., Gilks, W. R., and Day, N. E. (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47:449–498.

De Angelis, D. and Gilks, W. R. (1994). Estimating acquired immune deficiency syndrome incidence accounting for reporting delay. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 31–40.

De Angelis, D., Presanis, A. M., Conti, S., and Ades, A. E. (2014). Estimation of HIV burden through Bayesian evidence synthesis. *Statistical Science*, 29(1):9–17.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to aids. *Biometrics*, pages 1–11.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dietz, K., Seydel, J., and Schwartlander, B. (1994). Back-projection of German AIDS data using information on dates of tests. *Statistics in medicine*, 13(19-20):1991–2008.

Doucet, A., De Freitas, N., and Gordon, N. (2001). Sequential monte carlo methods in practice. series statistics for engineering and information science.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer.

Duvenaud, D. (2014). *Automatic Model Construction with Gaussian Processes*. PhD thesis, Computational and Biological Learning Laboratory, University of Cambridge.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eilers, P. H. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, 66(2):159–174.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, pages 731–761.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2007). *Regression*. Springer.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220.

Farewell, V., Aalen, O., Angelis, D. D., and MRC, N. D. (1994). Estimation of the rate of diagnosis of hiv infection in hiv infected individuals. *Biometrika*, 81(2):287–294.

Farrington, C. and Gay, N. (1999). Interval-censored survival data with informative examination times: parametric models and approximate inference. *Statistics in medicine*, 18(10):1235–1248.

Fellows, I. E. (2017). Hivbackcalc r package. https://github.com/hivbackcalc/package1.0. Accessed: 2017-09-22.

Fellows, I. E., Morris, M., Birnbaum, J. K., Dombrowski, J. C., Buskin, S., Bennett, A., and Golden, M. R. (2015). A new method for estimating the number of undiagnosed hiv infected based on hiv testing history, with an application to men who have sex with men in seattle/king county, wa. *PloS one*, 10(7):e0129551.

Flaxman, S., Gelman, A., Neill, D., Smola, A., Vehtari, A., and Wilson, A. G. (2015). Fast hierarchical gaussian processes. *Manuscript in preparation*.

Foulkes, M. A. (1998). Advances in hiv/aids statistical methodology over the past decade. *Statistics in Medicine*, 17(1):1–25.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing*, 19(4):479–492.

Garnett, G. P. (2002). An introduction to mathematical models in sexually transmitted disease epidemiology. *Sexually transmitted infections*, 78(1):7–12.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

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

Gilbert, M. T. P., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M. (2007). The emergence of hiv/aids in the americas and beyond. *Proceedings of the National Academy of Sciences*, 104(47):18566–18570.

Gilks, W. R., Best, N., and Tan, K. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. CRC press.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Goubar, A., Ades, A. E., De Angelis, D., McGarrigle, C. A., Mercer, C. H., Tookey, P. A., Fenton, K., and Gill, O. N. (2008). Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3):541–580.

Gramacy, R. B. (2007). tgp: An R package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, 19(9):1–46.

Gramacy, R. B. et al. (2007). tgp: an r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, 19(9):6.

Grant, R. M., Lama, J. R., Anderson, P. L., McMahan, V., Liu, A. Y., Vargas, L., Goicochea, P., Casapía, M., Guanira-Carranza, J. V., Ramirez-Cardich, M. E., et al. (2010). Preexposure chemoprophylaxis for hiv prevention in men who have sex with men. *New England Journal of Medicine*, 363(27):2587–2599.

Grassly, N. C. and Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nature reviews. Microbiology*, 6(6):477.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Greenland, S. (1996). Historical HIV incidence modelling in regional subgroups: Use of flexible discrete models with penalized splines based on prior curves. *Statistics in Medicine*, 15(5):513–525.

Gupta, S. B., Gilbert, R. L., Brady, A. R., Livingstone, S. J., Evans, B. G., et al. (2000). Cd4 cell counts in adults with newly diagnosed hiv infection: results of surveillance in england and wales, 1990–1998. *Aids*, 14(7):853–861.

Hallett, T. B., Zaba, B., Todd, J., Lopman, B., Mwita, W., Biraro, S., Gregson, S., Boerma, J. T., Network, A., et al. (2008). Estimating incidence from prevalence in generalised hiv epidemics: methods and validation. *PLoS medicine*, 5(4):e80.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

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

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Horsburgh, C. R., Jason, J., Longini, I., Mayer, K., Schochetman, G., Rutherford, G., SEAGE, G., Ou, C., Holmberg, S., Schable, C., et al. (1989). Duration of human immunodeficiency virus infection before detection of antibody. *The Lancet*, 334(8664):637–640.

Jansen, I. A., Geskus, R. B., Davidovich, U., Jurriaans, S., Coutinho, R. A., Prins, M., and Stolte, I. G. (2011). Ongoing hiv-1 transmission among men who have sex with men in amsterdam: a 25-year prospective cohort study. *Aids*, 25(4):493–501.

Janssen, R. S., Satten, G. A., Stramer, S. L., Rawal, B. D., O'brien, T. R., Weiblen, B. J., Hecht, F. M., Jack, N., Cleghorn, F. R., Kahn, J. O., et al. (1998). New testing strategy to detect early hiv-1 infection for use in incidence estimates and for clinical and prevention purposes. *Jama*, 280(1):42–48.

Kalaitzis, A., Honkela, A., Gao, P., and Lawrence, N. D. (2015). *Gaussian Processes Tool-Kit*. R package version 1.08.

Karon, J. M., Fleming, P. L., Steketee, R. W., and De Cock, K. M. (2001). Hiv in the united states at the turn of the century: an epidemic in transition. *American journal of public health*, 91(7):1060.

Karon, J. M., Khare, M., and Rosenberg, P. S. (1998). The current status of methods for estimating the prevalence of human immunodeficiency virus in the united states of america. *Statistics in medicine*, 17(2):127–142.

Karon, J. M., Song, R., Ron Brookmeyer, H.Kaplan, E., and Hall, H. I. (2008). Estimating HIV incidence in the United States fromHIV/AIDS surveillance data and biomarker HIV test results. *Statistics in medicine*, 27(23)(August):4817–4834.

Kassanjee, R., Angelis, D. D., Farah, M., Hanson, D., Labuschagne, J. P. L., Laeyendecker, O., Vu, S. L., Tom, B., Wang, R., and Welte, A. (2017). Cross-sectional hiv incidence surveillance: A benchmarking of approaches for estimating the 'mean duration of recent infection'. *Statistical Communications in Infectious Diseases*, 9(1).

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic hiv-1. *Science*, 313(5786):523–526.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 371–412.

Kirwan, P. D., Chau, C., Brown, A. E., Gill, O. N., Delpech, V. C., and contributors (2016). Hiv in the uk - 2016 report.

Kneib, T., Heinzl, F., Brezger, A., Bove, D. S., and Klein, N. (2014). *BayesX: R Utilities Accompanying the Software Package BayesX*. R package version 0.2-9.

Koulai, L., Presanis, A., Murphy, G., Suligoi, B., and De Angelis, D. (2017). Quantifying the recency of HIV infection using multiple longitudinal biomarkers. *ArXiv e-prints*.

Krige, D. (1966). *Two-dimensional weighted moving average trend surfaces for ore evaluation*. South African Institute of Mining and Metallurgy Johannesburg.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lessner, L. (1998). Estimating hiv incidence: An ill-posed problem. *Socio-Economic Planning Sciences*, 32(1):45–55.

Levy, J. A., Hoffman, A. D., Kramer, S. M., Kandis, J. A., Shimabururo, J. M., and Oshiro, L. S. (1984). Isolation of lymphocytopathic retroviruses from san francisco patients with aids. *Science*, 225:840–843.

Longini, I. M., Byers, R. H., Hessol, N. a., and Tan, W. Y. (1992). Estimating the stage-specific numbers of HIV infection using a Markov model and back-calculation. *Statistics in medicine*, 11(6):831–843.

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

MacDonald, B., Ranjan, P., and Chipman, H. (2015). GPfit: An R package for fitting a gaussian process model to deterministic simulator outputs. *Journal of Statistical Software*, 64(12):1–23.

MacKay, D. J. (1998). Introduction to gaussian processes. In C.M., B., editor, *Neural Networks and Machine Learning*, pages 84–92. Springer-Verlag.

Mariotti, S. and Cascioli, R. (1996). Sources of uncertainty in estimating hiv infection rates by back-calculation: and application to italian data. *Statistics in medicine*, 15(24):2669–2687.

Marra, G. and Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19(2):107–125.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 5(7).

Marschner, I. C. (1994). Using time of first positive HIV test and other auxiliary data in back-projection of AIDS incidence. *Statistics in Medicine*, 13(19-20):1959–1974.

Marschner, I. C. (1996). Fitting a multiplicative incidence model to age- and time-specific prevalence data. *Biometrics*, 52(2):492–499.

Marschner, I. C. (1997). A method for assessing age-time disease incidence using serial prevalence data. *Biometrics*, pages 1384–1398.

Marschner, I. C. and Bosch, R. J. (1998). Flexible assessment of trends in age-specific HIV incidence using two-dimensional penalized likelihood. *Statistics in Medicine*, 17(April 1997):1017–1031.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in applied probability*, 5(3):439–468.

McCormack, S., Dunn, D. T., Desai, M., Dolling, D. I., Gafos, M., Gilson, R., Sullivan, A. K., Clarke, A., Reeves, I., Schembri, G., et al. (2016). Pre-exposure prophylaxis to prevent the acquisition of hiv-1 infection (proud): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *The Lancet*, 387(10013):53–60.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McGarrigle, C., Cliffe, S., Copas, A., Mercer, C., DeAngelis, D., Fenton, K., Evans, B., Johnson, A., and Gill, O. (2006). Estimating adult hiv prevalence in the uk in 2003: the direct method of estimation. *Sexually transmitted infections*, 82(suppl 3):iii78–iii86.

Meyer, S., Held, L., and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, 77(11):1–55.

Mezzetti, M. and Robertson, C. (1999). A hierarchical Bayesian approach to age-specific back-calculation of cancer incidence rates. *Statistics in Medicine*, 18(8):919–933.

Miller, E., Waight, P., Tedder, R., Sutherland, S., Mortimer, P., and Shafi, M. (1995). Incidence of hiv infection in homosexual men in london, 1988-94. *BMJ*, 311(7004):545.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Moore, R. D. and Chaisson, R. E. (1999). Natural history of hiv infection in the_era of combination antiretroviral therapy. *Aids*, 13(14):1933–1942.

Murray, I. and Adams, R. P. (2010). Slice sampling covariance hyperparameters of latent gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740.

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. *Statistics in Medicine*, 18(3):307–320.

Nakagawa, F. et al. (2017). An epidemiological modelling study to estimate the composition of hiv-positive populations including migrants from endemic settings. *Aids*, 31(3):417–425.

Nash, J. C. (1990). *Compact numerical methods for computers: linear algebra and function minimisation*. CRC press.

Nash, J. C. and Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43(9):1–14.

Ndawinz, J. D. a., Costagliola, D., and Supervie, V. (2011). New method for estimating HIV incidence and time from infection to diagnosis using HIV surveillance data: results for France. *AIDS (London, England)*, 25(15):1905–1913.

Neal, R. M. (1996). Lecture notes in statistics.

Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, pages 370–384.

ONS (2015). Internal migration, england and wales: Year ending june 2015. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/bulletins/internalmigrationbylocalauthoritiesinenglandandwales/yearendingjune2015. Accessed: 2017-09-22.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518.

Phillips, A. N., Cambiano, V., Miners, A., Lampe, F. C., Rodger, A., Nakagawa, F., Brown, A. E., Gill, O. N., De Angelis, D., Elford, J., Hart, G., Johnson, A. M., Lundgren, J. D., Collins, S., and Delpech, V. C. (2015). Potential impact on HIV incidence of higher HIV testing rates and earlier antiretroviral therapy initiation in MSM. *AIDS*, 29(14):1855–62.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC)*, pages 20–22. Vienna.

Presanis, A., De Angelis, D., Spiegelhalter, D., Seaman, S., Goubar, A., and Ades, A. (2008). Conflicting evidence in a bayesian synthesis of surveillance data to estimate human immunodeficiency virus prevalence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(4):915–937.

Presanis, A. M., De Angelis, D., Goubar, A., Gill, O. N., and Ades, A. E. (2011). Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men. *Biostatistics*, 12(4):666–681.

Punyacharoensin, N., Edmunds, W. J., De Angelis, D., Delpech, V., Hart, G., Elford, J., Brown, A., Gill, N., and White, R. G. (2015). Modelling the hiv epidemic among msm in the united kingdom: quantifying the contributions to hiv transmission to better inform prevention initiatives. *Aids*, 29(3):339–349.

Punyacharoensin, N., Edmunds, W. J., De Angelis, D., Delpech, V., Hart, G., Elford, J., Brown, A., Gill, O. N., and White, R. G. (2016). Effect of pre-exposure prophylaxis and combination hiv prevention for men who have sex with men in the uk: a mathematical modelling study. *The lancet HIV*, 3(2):e94–e104.

Punyacharoensin, N., Edmunds, W. J., De Angelis, D., and White, R. G. (2011). Mathematical models for the study of hiv spread and control amongst men who have sex with men. *European journal of epidemiology*, 26(9):695.

Raab, G. M., Gore, S. M., Goldberg, D. J., and Donnelly, C. A. (1994). Bayesian forecasting of the human immunodeficiency virus epidemic in scotland. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 17–30.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183.

Richman, D. D., Fischl, M. A., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., Leedom, J. M., Groopman, J. E., Mildvan, D., and Hirsch, M. S. (1987). The toxicity of azidothymidine (azt) in the treatment of patients with aids and aids-related complex. a double-blind, placebo-controlled trial. *The New England journal of medicine*, 317(4):192–197.

Riedner, G. and Dehne, K. L. (1999). Hiv/aids surveillance in developing countries. experiences and issues.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A*, 371(1984):20110550.

Rosenberg, P., Gail, M., and Carroll, R. (1992). Estimating hiv prevalence and projecting aids incidence in the united states: a model that accounts for therapy and changes in the surveillance definition of aids. *Statistics in medicine*, 11(13):1633–1655.

Rosenberg, P. S. (1994). Backcalculation models of age-specific HIV incidence rates. *Statistics in medicine*, 13(19-20):1975–1990.

Rosenberg, P. S. (1995). Scope of the AIDS epidemic in the United States. *Science (New York, N.Y.)*, 270(5240):1372–1375.

Rosenberg, P. S. and Gail, M. H. (1990). Uncertainty in estimates of hiv prevalence derived by backcalculation. *Annals of epidemiology*, 1(2):105–115.

Rosenberg, P. S. and Gail, M. H. (1991). Backcalculation of flexible linear models of the human immunodeficiency virus infection curve. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 40(2):269–282.

Rosenberg, P. S. and Goedert, J. J. (1994). Effect of age at seroconversion on the natural aids incubation distribution. *Aids*, 8(6):803–810.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Saatçi, Y. (2012). *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., and Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1410):867–876.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–52.

Solomon, P. and Wilson, S. (1990). Accommodating change due to treatment in the method of back projection for estimating hiv infection incidence. *Biometrics*, pages 1165–1170.

Sommen, C., Alioum, A., and Commenges, D. (2009). A multistate approach for estimating the incidence of human immunodeficiency virus by using hiv and aids french surveillance data. *Statistics in medicine*, 28(11):1554–1568.

Sommen, C., Commenges, D., Vu, S. L., Meyer, L., and Alioum, A. (2011). Estimation of the Distribution of Infection Times Using Longitudinal Serological Markers of HIV: Implications for the Estimation of HIV Incidence. *Biometrics*, 67(2):467–475.

Stan Development Team (2016a). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1.

Stan Development Team (2016b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0*.

Sundararajan, S. and Keerthi, S. S. (2000). Predictive approaches for choosing hyperparameters in gaussian processes. In *Advances in neural information processing systems*, pages 631–637.

Supervie, V., Ndawinz, J. D., Lodi, S., and Costagliola, D. (2014). The undiagnosed hiv epidemic in france and its implications for hiv screening strategies. *AIDS (London, England)*, 28(12):1797.

Sweeting, M. J., De Angelis, D., and Aalen, O. O. (2005). Bayesian back-calculation using a multi-state model with application to HIV. *Statistics in Medicine*, 24(August):3991–4007.

Sweeting, M. J., De Angelis, D., Parry, J., and Suligoi, B. (2010). Estimating the distribution of the window period for recent hiv infections: a comparison of statistical methods. *Statistics in medicine*, 29(30):3194–3202.

Taffe, P. and May, M. (2008). A joint back calculation model for the imputation of the date of hiv infection in a prevalent cohort. *Statistics in medicine*, 27(23):4835–4853.

Tindall, B. and Cooper, D. A. (1991). Primary hiv infection: host responses and intervention strategies. *Aids*, 5(1):1–14.

UNAIDS (2016). Fact sheet - latest statistics on the status of the aids epidemic. http://www.unaids.org/en/resources/fact-sheet. Accessed: 2017-09-03.

UNAIDS (2016). Global aids update 2016. *Geneva: Joint United Nations Programme on HIV/AIDS and others*.

van Sighem, A., Nakagawa, F., De Angelis, D., Quinten, C., Bezemer, D., de Coul, E. O., Egger, M., de Wolf, F., Fraser, C., and Phillips, A. N. (2015). Estimating HIV Incidence, Time to Diagnosis, and the Undiagnosed HIV Epidemic Using Routine Surveillance Data. *Epidemiology (Cambridge, Mass.)*, 26(5):653–60.

Verdecchia, A. and Mariotto, A. B. (1995). A back-calculation method to estimate the age and period hiv infection intensity, considering the susceptible population. *Statistics in medicine*, 14(14):1513–1530.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2.

Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wand, H., Wilson, D., Yan, P., Gonnermann, A., McDonald, A., Kaldor, J., and Law, M. (2009). Characterizing trends in HIV infection among men who have sex with men in Australia by birth cohorts: results from a modified back-projection method. *Journal of the International AIDS Society*, 12(1):19.

Williams, B., Gouws, E., Wilkinson, D., and Karim, S. A. (2001). Estimating hiv incidence rates from age prevalence data in epidemic situations. *Statistics in medicine*, 20(13):2003–2016.

Williams, C. K. (1997). *Regression with Gaussian processes*. Kluwer.

Williams, C. K. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Williams, C. K. and Rasmussen, C. E. (1996). Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520.

Wood, S. (2006a). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. (2012). mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(1):95–114.

Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. N. (2006c). On confidence intervals for generalized additive models based on penalized regression splines. 48(4):445–464.

Wood, S. N. (2016). Just another gibbs additive modeller: Interfacing jags and mgcv. *arXiv preprint arXiv:1602.02539*.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, (just-accepted):1–45.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., et al. (2008). Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, 455(7213):661.

Yan, P., Zhang, F., and Wand, H. (2011). Using HIV Diagnostic Data to Estimate HIV Incidence: Method and Simulation. *Statistical Communications in Infectious Diseases*, 3(1):4690.

Appendix A

Bayesian Inference

"Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model" (Gelman et al., 2014).

This Appendix begins by reviewing Bayesian inference and then focuses on Monte-Carlo Methods and available software to implement Bayesian hierarchical models. Finally the implementation of back-calculation using available Bayesian software is discussed. Most of the material discussed in this Appendix is based on Gilks et al. (1996) and Betancourt (2017a).

Note that in this Appendix $\boldsymbol{\theta}$ denote any general parameters, in contrast to the main body of this thesis where $\boldsymbol{\theta}$ denoted the infection parameters.

A.1 A brief introduction to Bayesian inference

Statistical models describe a set of data $\boldsymbol{Y}$ through the means of unknown parameters $\boldsymbol{\theta}$. Within a maximum likelihood framework, parameters are assumed to be fixed, whereas within a Bayesian framework they are assumed to follow a certain distribution (prior). Bayes' rule provides a probabilistic tool for updating our "beliefs" on the prior distribution of parameters $\boldsymbol{\theta}$, conditional on some observed data $\boldsymbol{y}$.

Inference within a Bayesian framework requires the specification of the likelihood function for the data $p(\boldsymbol{y}|\boldsymbol{\theta})$, as well as the prior distribution $p(\boldsymbol{\theta})$. The posterior distribution of the parameters $p(\boldsymbol{\theta}|\boldsymbol{y})$, conditional on the data $\boldsymbol{y}$ being observed, can be then derived via Bayes' rule as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{A.1.1}$$

The posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the quantity of primary interest in Bayesian analysis, and its features (*e.g.* moments, quantiles) can be expressed in terms of posterior expectations of functions $g(\boldsymbol{\theta})$ as follows:

$$\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[g(\boldsymbol{\theta})] = \frac{\int g(\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{A.1.2}$$

In practice, the integrals in the numerator and denominator (and thus $\mathbb{E}[g(\boldsymbol{\theta}]$ itself) are often analytically intractable. However, if we were able to obtain a sample $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ from $p(\boldsymbol{\theta}|\boldsymbol{y})$ (*i.e.* a posterior sample), the expectation in Equation A.1.2 could be approximated by the sample mean:

$$\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[g(\boldsymbol{\theta})] \approx \frac{1}{n}\sum_{i=1}^{n} g(\theta^{(i)}) \tag{A.1.3}$$

Drawing independent samples $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ from $p(\boldsymbol{\theta}|\boldsymbol{y})$ is typically impossible. However, Markov chain Monte Carlo (MCMC) allow to obtain a correlated sample $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}\}$, for which the above Equation still holds.

A.2 Markov chain Monte Carlo methods

The problem introduced in the previous Section for Bayesian inference, can be expressed in a more general for via the following expectation:

$$\mathbb{E}_f[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{A.2.1}$$

where $f(\boldsymbol{\theta})$ is a target distribution (*e.g.* $p(\boldsymbol{\theta}|\boldsymbol{y})$ in Section A.1).

MCMC methods aim to construct a Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}\}$, that converges to some stationary distribution; if the algorithm is correctly specified, the stationary distribution corresponds to the target distribution of interest, so that its sample mean can be used to approximate the integral in Equation A.2.1.

The challenge is in ensuring that the stationary distribution of the Markov chain is indeed the target distribution; this can be achieved by ensuring that the transition kernel of the chain $k(\boldsymbol{\theta}|\boldsymbol{\theta}')$, specifying the transition probability between the states, preserves the target distribution, *i.e.* $f(\boldsymbol{\theta}') = \int f(\boldsymbol{\theta})k(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}$. Satisfying the detailed balance is a sufficient criterion, for further technical details Gilks et al. (1996). The following Sections describe algorithms that construct Markov chains so that these properties are satisfied.

A.2.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (Hastings, 1970) algorithm proceeds as follows: given a starting value $\boldsymbol{\theta}^{(0)}$ for the parameters, updated values $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots\}$ are obtained at each iteration $k+1$ as follows:

1. A candidate $\boldsymbol{\theta'}$ is generated from the transition kernel $k(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(k)})$.
2. The candidate is accepted with probability:

$$\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(k)}) = min\left(1, \frac{f(\boldsymbol{\theta}')k(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta}^{(k)})k(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(k)})}\right) \tag{A.2.2}$$

   If accepted, the updated $\boldsymbol{\theta}^{(k+1)}$ is assigned the candidate value $\boldsymbol{\theta}'$, otherwise the previous value $\boldsymbol{\theta}^{(k)}$ is carried forward.

The transition kernel can have any form and the Markov chain will converge to the target distribution subject to mild regularity conditions (Gilks et al., 1996).

Note that to compute Equation A.2.2 $f(\boldsymbol{\theta})$ must be known only up to a proportionality constant. This is particularly advantageous within a Bayesian inference context (Section A.1), as the (potentially intractable) normalizing constant of $p(\boldsymbol{\theta}|\boldsymbol{y})$ (Equation A.1.1) does not need to be evaluated.

Several Metropolis-Hastings variants exist; for instance, the Metropolis algorithm only considers symmetric proposals (*i.e.* $k(\boldsymbol{\theta}'|\boldsymbol{\theta}) = k(\boldsymbol{\theta}|\boldsymbol{\theta}')$) so that the acceptance probability simplifies to $min(1, f(\boldsymbol{\theta}')/f(\boldsymbol{\theta}^{(k)}))$. One of the most widely used variant is the random-walk Metropolis-Hastings, specifying a normally distributed proposal - *i.e.* $k(\boldsymbol{\theta}'|\boldsymbol{\theta}) = N(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{\Sigma})$.

It is often simpler to update the components of $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ individually rather than simultaneously; this is known as the single-component Metropolis-Hastings. Note the each iteration $k+1$ involves m further updates, assumed (for simplicity) to occur in increasing order of components for $\boldsymbol{\theta}$. Now, a value θ_i' is proposed for the i^{th} component according to a transition kernel $k(\theta_i'|\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)})$, where $\boldsymbol{\theta}_{-i}^{(k)} = \{\theta_1^{(k+1)}, \ldots, \theta_{i-1}^{(k+1)}, \theta_{i+1}^{(k)}, \ldots \theta_m^{(k)}\}$, and the acceptance probability is:

$$\alpha(\theta_i'|\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)}) = min\left(1, \frac{f(\theta_i'|\boldsymbol{\theta}_{-i}^{(k)})k(\theta_i^{(k)}|\theta_i', \boldsymbol{\theta}_{-i}^{(k)})}{f(\theta_i^{(k)}|\boldsymbol{\theta}_{-i}^{(k)})k(\theta_i'|\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)})}\right) \tag{A.2.3}$$

where $f(\theta_i^{(k)}|\boldsymbol{\theta}_{-i})$ is the full conditional distribution for θ_i.

Gibbs Sampling (Geman and Geman, 1984) is a special case of the single-component Metropolis Hastings algorithm where for the i^{th} component, the transition kernel is equal to the full conditional distribution of θ_i (*i.e.* $k(\theta_i'|\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)}) = f(\theta_i'|\boldsymbol{\theta}_{-i}^{(k)})$) rendering the acceptance probability in Equation A.2.3 equal to one.

Finally, Metropolis-Hastings can be also used to update multiple, but not all, components simultaneously; this is known as block-sampling.

A.2.2 Hamiltonian Monte Carlo

It is well known (Betancourt, 2017a) that the efficiency of Metropolis-Hastings algorithms decreases as the dimension of $\boldsymbol{\theta}$ increases. This is because finding a transition kernel, making relevant proposals $\boldsymbol{\theta}'$, becomes challenging and results in poor algorithmic performance. Hamiltonian Monte Carlo (HMC) is a variant of MCMC that explores the target distribution more efficiently, by making proposals based on geometrical properties of the target distribution.

Given a $m \times 1$ vector of parameters $\boldsymbol{\theta}$, HMC introduces an $m \times 1$ vector of auxiliary momentum parameters $\boldsymbol{\rho}$. The joint density for $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ is:

$$f(\boldsymbol{\theta}, \boldsymbol{\rho}) = f(\boldsymbol{\rho}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$

Hamiltonian dynamics are widely used in the fields of physics and differential geometry, and could be also applied in a Monte-Carlo context. In a physics context, the negative joint log-density is referred to as the Hamiltonian function

$$H(\boldsymbol{\theta}, \boldsymbol{\rho}) \equiv -log\, f(\boldsymbol{\theta}, \boldsymbol{\rho}) = -log\, f(\boldsymbol{\rho}|\boldsymbol{\theta}) - log\, f(\boldsymbol{\theta}) = K(\boldsymbol{\rho}|\boldsymbol{\theta}) + U(\boldsymbol{\theta})$$

and $K(\boldsymbol{\rho}|\boldsymbol{\theta})$ and $U(\boldsymbol{\theta})$ are referred to as the kinetic and potential energy respectively. Note that the potential energy is fully specified by the target distribution whereas the kinetic energy must be appropriately chosen.

In physics, Hamiltonian dynamics describe the conservation of energy over time. The same value of the Hamiltonian function $H(\boldsymbol{\theta},\boldsymbol{\rho})$ (*i.e.* Hamiltonian energy level) can be obtained by different $(\boldsymbol{\theta},\boldsymbol{\rho})$ combinations, described by the following set of differential equations (*i.e.* Hamilton's equations):

$$\begin{aligned}\frac{d\boldsymbol{\theta}}{dt} &= +\frac{\partial H}{\partial \boldsymbol{\rho}} &= +\frac{\partial K}{\partial \boldsymbol{\rho}}\\ \frac{d\boldsymbol{\rho}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} &= -\frac{\partial K}{\partial \boldsymbol{\theta}} - \frac{\partial U}{\partial \boldsymbol{\theta}}\end{aligned}$$

Given an initial value $\boldsymbol{\theta}^{(0)}$ and a conditional distribution for the momentum $f(\boldsymbol{\rho}|\boldsymbol{\theta})$, an ideal HMC sampler would construct a Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots\}$ as follows at the $k+1^{\text{th}}$ iteration:

1. A momentum $\boldsymbol{\rho}$ is sampled from its conditional distribution - *i.e.* $\boldsymbol{\rho} \sim f(\boldsymbol{\rho}|\boldsymbol{\theta}^{(k)})$.
2. The given Hamiltonian energy level is explored, by integrating the Hamilton's equations for a certain time t from the starting point $(\boldsymbol{\rho},\boldsymbol{\theta}^{(k)})$. This yields $(\boldsymbol{\rho}',\boldsymbol{\theta}')$.
3. $\boldsymbol{\theta}^{(k+1)}$ is set to $\boldsymbol{\theta}'$

In practice, the integral of Hamilton's equations can not be solved analytically, and are thus numerically approximated by a leap-frog algorithm; this is not numerically perfect and introduces bias. To ensure that the detailed balance is satisfied a Metropolis acceptance step is included (Betancourt, 2017a) by modifying step 2 and step 3 of the algorithm as follows:

2. The integral of Hamilton's equations for a certain time t from the starting point $(\boldsymbol{\rho},\boldsymbol{\theta}^{(k)})$ is approximated using the leap-frog algorithm. This yields $(\boldsymbol{\rho}',\boldsymbol{\theta}')$.
3. The candidate $\boldsymbol{\theta}^{(k)}$ is accepted with probability:

   $$\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(k)}) = min(1, exp(H(\boldsymbol{\theta},\boldsymbol{\rho}) - H(\boldsymbol{\theta}',\boldsymbol{\rho}'))$$

   If accepted, the updated $\boldsymbol{\theta}^{(k+1)}$ is assigned the new candidate value $\boldsymbol{\theta}'$,otherwise the previous value $\boldsymbol{\theta}^{(k)}$ is carried forward

Note that the performance of HMC depends on the fine-tuning of the leap-frog algorithm and on the distribution chosen for $f(\boldsymbol{\rho}|\boldsymbol{\theta})$. The time limit of the integral evaluated by the leap-frog algorithm must be large enough to ensure that the new proposal made is sufficiently different from the current value; once a "sufficiently far" proposal is made, based on a pre-determined criterion, the algorithm proceeds to step 3. In practice $f(\boldsymbol{\rho}|\boldsymbol{\theta})$ is typically multivariate normal $N(\mathbf{0},\boldsymbol{\Sigma})$, so that $f(\boldsymbol{\rho}|\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$. $\boldsymbol{\Sigma}$ shall be appropriately chosen, so that the sampled momentum avoids the algorithm from getting stuck in regions characterised by low $\boldsymbol{\theta}$ mass. Riemaniann HMC is an extension of HMC, that considers $f(\boldsymbol{\rho}|\boldsymbol{\theta})$ not to be independent of $\boldsymbol{\theta}$ (Girolami and Calderhead, 2011).

A.3 Convergence

In the previous Sections we constructed a Markov Chain so that it asymptotically tends to its stationary (*i.e.* target) distribution; however convergence may take very long, especially when the algorithms are initiated in low-probability regions of the stationary distribution. Thus the first *burn-in* iterations from MCMC algorithms are typically discarded.

Convergence assessment is a key yet non-trivial task. Diagnosing convergence is non-trivial, as no formal method of establishing convergence exists. A number of convergence tests do exist that provide evidence (but not guarantee) of achieved convergence. A widely used test involves inspecting univariate trace plots of MCMC samples of the parameters of interest and if these resemble a random scatter-plot (*i.e.* a "fat hairy caterpillar") one may conclude that a lack of convergence cannot be detected. The $\hat{R}$ statistics (Gelman and Rubin, 1992) provides another test for convergence, based on analysis of variance for multiple chains with over-dispersed starting points. The variance between single chains should be comparable to the variance within all chains: the test assesses whether the difference in these is such to suggest lack of convergence. Values of $\hat{R}$ greater than 1.05 are taken, as a rule of thumb, to indicate non-convergence.

Divergent transitions (or divergences) are another helpful tool to diagnose lack of convergence, which is only available for HMC. These indicate that the Markov chain has encountered regions of high curvature in the target distribution that cannot be adequately explored, resulting in biased estimates of the parameters of interest (Betancourt, 2017b).

As discussed in Appendix A.2.2, HMC simulate Hamiltonian dynamics using a leap-frog algorithm, which is governed by the resolution or *step size*. The resolution rules how far the proposal can be from the initial state, in order to make "optimal" proposals. If the resolution chosen is too large the leap-frog algorithm may fail in regions of high-curvature of the posterior distribution. An excellent analogy for resolution is[1]: if one walks down a mountain by taking very large steps, falling is natural. The only way to explore the mountain is by taking smaller steps. When the resolution is too big, the leap-frog proposal diverges to infinity resulting in numerical errors. The obvious solution is picking a smaller resolution, but this does not always solve the problem and makes the algorithm slower. The geometry of the posterior distribution is often too complex to be adequately explored and requires a model reparameterisation (*e.g.* the non-centered parameterisation). Unfortunately re-parameterisations are not guaranteed to have nice geometries. A zero-tolerance policy for divergent transitions is typically employed in `Stan` (Stan Development Team, 2016b).

A.4 JAGS and Stan

`JAGS` (Plummer, 2003) and `Stan` (Stan Development Team, 2016b) are two probabilistic programming language for Bayesian analysis. Both software are designed to take a user's description of a hierarchical model and returning an MCMC sample of the posterior distribution. These are black-box software, avoiding users having to explicitly code MCMC sampler. Despite `JAGS` and `Stan` have similar objectives, their implementation substantially differs.

`JAGS` is an acronym for Just Another Gibbs Sampler and has been developed and is maintained by Dr. Martyn Plummer. It was first released in 2003 and version 4.3 is the latest available. `JAGS` is based on the `BUGS` language, that was originally developped for `WinBUGS` and `OpenBUGS`. The `BUGS` language is declarative; the user is only requested to specify the relationships among the variables, in terms of probabilistic or deterministic functions. Based on the user's specification of the prior and the likelihood of a model, `JAGS` builds a directed acyclic graph, which expresses a hierarchical model.

From a directed acyclic graph, the the posterior distribution can be derived as well as the full conditional distributions $f(\theta_i'|\boldsymbol{\theta}_{-i}^{(k)})$. Gibbs sampler is then applied, sometimes in combination with other samplers including Metropolis–Hastings algorithm (Chib and Greenberg, 1995), Slice sampling (Neal, 2003), and the Adaptive Rejection sampling (Gilks et al., 1995).

`Stan` is named after Stanislaw Ulam, a mathematician who was one of the pioneers of Monte Carlo methods in the 40's; it has been developed and is maintained by the Stan Development

[1] http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup

Team, lead by Dr. Andrew Gelman and Dr. Bob Carpenter. Stan version 1.0 was released in 2012 and version 2.16 is the latest available.

As briefly discussed in Section A.2.2, a number of drawbacks prevent the widespread practical implementation of HMC; partial derivatives of the negative log-likelihood $U(\boldsymbol{\theta})$ must be taken with respect to each parameter θ_i, $i = \{1,\ldots,m\}$. Moreover tuning the leap-frog integrator and choosing the momentum's distribution covariance matrix $\boldsymbol{\Sigma}$ appropriately is not straightforward.

Stan provided solutions to these problem by implementing automatic differentiation (Carpenter et al., 2015) and by using the No U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014); this is a highly optimized HMC algorithm which achieves fine tuning of the leap-frog iterator and $\boldsymbol{\Sigma}$.

Stan and JAGS are both state-of-the-art software for Bayesian analysis. The former tends to be more efficient and faster than JAGS (in terms of effective sample size, for the same hierarchical model). However, unlike JAGS, Stan does not allow inference for discrete parameters as the underlying differential equations require parameter to be differentiable and thus continuous.

A.5 Application to back-calculation

Age-independent and age-dependent back-calculation have been implemented using both JAGS and Stan. Despite the results obtained with the two software are very similar, back-calculation is, in my opinion, simpler to implement in Stan, as equivalent results are obtained quicker than in JAGS and requires less ad-hoc specifications.

The only "trick" that was required for efficiently implementing back-calculation in Stan was considering both the centred and non-centred parameterisation for the parameters of interest (Betancourt, 2017b). For the diagnosis parameter the latter was more efficient than the former. Recall the logit diagnosis parameters are modelled with a first order logistic random walk - i.e. $\delta_{k,i} \sim N(\delta_{k,i-1}, \sigma^2_{D,i})$, where k denotes a progression state and i denotes the i^{th} time interval. Note that this is equivalent to $\delta_{k,i} = \delta_{k,i-1} + \sigma_{D,k}\, z_{k,i}$, where $z_{k,i} \sim N(0,1)$, $i = \{1,\ldots,T\}$, $k = \{1,\ldots,K\}$. Sampling $\{z_1,\ldots,z_T\}$, yields a higher effective sample size for $\{\delta_{k,1},\ldots,\delta_{k,T}\}$, than if we were to sample the $\delta_{k,i}$ directly.

Implementing the model with JAGS required a much higher level of ad-hoc knowledge and could have not been achieved without the help of Dr. Martyn Plummer.

The log-expected number of individuals diagnosed in the i^{th} time and j^{th} age intervals in state k can be written as a sum of latent variables statified by the age j_0^{th} (and implicitly the time interval j_0^{th}) of infection. The latent variables are denoted $\varepsilon_{i,j,k}^{j_0}$ ($i = \{1,\ldots,T\}$, $j = \{1,\ldots,A\}$, $j_0 = \{1,\ldots,j\}$, $k = \{1,\ldots,K\}$) and can be formulated as a log-linear model $\varepsilon_{i,j,k}^{j_0} = \gamma_{i_0,j_0} + log(p_{1,k}^{(j_0,j,i)})$, where $p_{1,k}^{(j_0,j,i)}$ denotes the probability of being infected in the i_0^{th} time and j_0^{th} age intervals and being diagnosed in state k in the i^{th} time and j^{th} age intervals. These correspond to the k^{th} entry of the vector defined in Equation 5.3.4. Recall that this log-linear formulation (discussed in Section 2.3.3 for age-dependent back-calculation only) is less efficient than the recursive equations.

JAGS updates the parameters of a log-linear Poisson regression models (*i.e.* the infection parameters for back-calculation) using the latent variable based auxiliary mixture sampling algorithm (Frühwirth-Schnatter et al., 2009). Diagnosis parameters are modelled using standard Metropolis-Hastings. The main challenge lies in ensuring that the latent variables are updated while remaining consistent with the data. JAGS' sum sampler is subsequently employed to sample with replacement pairs of latent variables; this adds and subtracts x (sampled from a geometrical distribution) to the first and second latent variable respectively, so that the sum of the latent variables is unchanged. Note that the introduction of the sum sampler in version 4.0 of JAGS was also motivated by the need to implement age-specific back-calculation models.

Appendix B

Appendix for Chapter 2

B.1 Notation

Recall that $(t_0, t_T]$ is the time-period spanning the HIV epidemic, which is split into T disjoint, consecutive intervals $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$:

- The i$^{\text{th}}$ interval refers to $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$.
- The i_0^{th} interval usually denotes$(t_{i_0-1}, t_{i_0}]$, $i_0 = \{1, \dots, T\}$, when infections occur.

A summary of the notation introduced in Chapter 2 is here presented: appropriate definitions are given in the main body, the list below only serves as reference:

- K is the number of undiagnosed states in the proposed multi-state model.
  States 1 to K denote latent undiagnosed states, state $2K+1$ corresponds to AIDS diagnosis and states $K+1$ to $2K$ identify HIV diagnosis with a certain CD4-count.
  $k = \{1, \dots, 2K+1\}$ denotes one of the states.
- $h_i \equiv h_i(\boldsymbol{\theta})$ represents the expected number of infections occurring in the i^{th} interval.
  $\mathcal{H} \equiv \mathcal{H}(\boldsymbol{\theta}) = \{h_1, \dots, h_T\}$ denotes the incidence curve, parametrised by $\boldsymbol{\theta}$.
- $\boldsymbol{d}_i \equiv \boldsymbol{d}_i(\boldsymbol{\delta}) = (d_{1,i}, \dots, d_{K,i})$ refers to the diagnosis probabilities in the i$^{\text{th}}$ interval from undiagnosed states k=$\{1, \dots, K\}$.
  $\mathcal{D} \equiv \mathcal{D}(\boldsymbol{\delta}) = \{\boldsymbol{d}_1, \dots, \boldsymbol{d}_T\}$ refers to the collection of diagnosis probabilities over time, parametrised by $\boldsymbol{\delta}$.
- $\boldsymbol{q} = (q_1, \dots, q_K)$ denotes undiagnosed probabilities between undiagnosed states $k = \{1, \dots, K\}$, within any interval i, as these are assumed constant over time.

- $\boldsymbol{Q}_i \equiv \boldsymbol{Q}_i(\boldsymbol{\delta})$, of size $K \times K$, and $\boldsymbol{D}_i \equiv \boldsymbol{D}_i(\boldsymbol{\delta})$, of size $K \times K+1$ are the transition and progression matrices of the multi-state model. They describe the transition probabilities in the i$^{\text{th}}$ interval between undiagnosed states $k = \{1, \ldots, K\}$ and from undiagnosed to diagnosis states $k = \{K+1, \ldots, 2K+1\}$ respectively.

- $\boldsymbol{e}_i \equiv \boldsymbol{e}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a $K \times 1$ vector, containing the expected number of individuals in undiagnosed states $k = \{1, \ldots, K\}$ in the end of the i$^{\text{th}}$ interval.

- $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a $(K+1) \times 1$ vector, describing the expected number of new arrivals in diagnoses states $k = \{K+1, \ldots, 2K\}$ at the end of the i$^{\text{th}}$ interval.
  $\mu_i^A \equiv \mu_i^A(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\mu_i^H \equiv \mu_i^H(\boldsymbol{\theta}, \boldsymbol{\delta})$ respectively refer to the expected number of new AIDS and HIV diagnoses at the end of the i$^{\text{th}}$ interval.

- $\boldsymbol{p}_i \equiv \boldsymbol{p}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ describes the expected proportion of diagnoses in each of the HIV-states $k = \{K+2, \ldots, 2K+1\}$ at the end of the i$^{\text{th}}$ interval.

- $\boldsymbol{\upsilon} = \{\upsilon_1^H, \ldots \upsilon_T^H, \upsilon_1^A, \ldots, \upsilon_T^A\}$ denotes the collection of under-reporting parameters over time-intervals.

- $\boldsymbol{\pi}$, a $K \times 1$ vector, that refers to the initial expected number of undiagnosed infections in states $k = \{1, \ldots, K\}$.

Appendix C

Appendix for Chapter 3

C.1 Further details for Section 3.3.3

C.1.1 The QR decomposition

This Appendix provides the details of the QR decomposition. For further details refer to Wood, 2006a, page 46 and 334.

Let $\boldsymbol{X}$ be a $n \times m$ matrix, where $n \geq m$. $\boldsymbol{X}$ can always be decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_1 \boldsymbol{R}_1 \tag{C.1.1}$$

Where $\boldsymbol{R}$ is an upper triangular $m \times m$ matrix, $\boldsymbol{0}$ is a $(n-m) \times m$ matrix and $\boldsymbol{Q}$ is a $n \times n$ orthogonal matrix. $\boldsymbol{Q}$ can be further split into $\boldsymbol{Q}_1$ (dimension $n \times m$) and $\boldsymbol{Q}_2$ (dimension $n \times n - m$), both with orthogonal columns.

Now consider the specific case where a linear (or spline) model $\boldsymbol{X\beta}$ is subject to the constraint $\boldsymbol{C\beta} = 0$. $\boldsymbol{C}$ is a $c \times m$ matrix, imposing c distinct constraints.

The aim is to to reparametrize the linear (or spline) model in terms of $\boldsymbol{\beta}'$ containing $m - c$ free parameters, rather than m parameters subject to c constraints. To do so it is necessary finiding a matrix $\boldsymbol{Z}$, of dimension $m \times (m-c)$, so that:

$$\boldsymbol{\beta} = \boldsymbol{Z\beta}' \tag{C.1.2}$$

$$\boldsymbol{CZ} = 0 \tag{C.1.3}$$

After re-parameterisation in Equation C.1.2, the constraint term becomes $\boldsymbol{CZ\beta'} = 0$. Then the condition in Equation C.1.3 allows $\boldsymbol{\beta'}$ to take any value, while satisfying the constraint.

Finally, $\boldsymbol{Z}$ is constructed using the QR decomposition of $\boldsymbol{C^T}$.

$$\boldsymbol{C^T} = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix} \tag{C.1.4}$$

Now set $\boldsymbol{Z} = \boldsymbol{Q_2}$ and consider:

$$\boldsymbol{CZ} = \begin{bmatrix}\boldsymbol{R^T} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}_1^{\boldsymbol{T}}\\ \boldsymbol{Z^T}\end{bmatrix}\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{R^T} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{I}\end{bmatrix} = \boldsymbol{0} \tag{C.1.5}$$

$\boldsymbol{Q_1}$ and $\boldsymbol{Q_2}$ span different columns of the same orthogonal matrix, thus the dot product characterising each entry of $\boldsymbol{Q_1^TQ_2} = \boldsymbol{Q_1^TZ}$ is made by orthogonal vectors and is equal to zero.

C.1.2 Mathematical details for reformulating Equation 3.3.6 as Equation 3.3.1

Recall that the smoothing objective can be rewritten as follows for an optimal NCS:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda\boldsymbol{\delta^T E\delta} \quad \text{s.t} \quad \boldsymbol{T^T\delta} = \boldsymbol{0}$$

Note that $\boldsymbol{T^T\delta} = \boldsymbol{0}$ imposes two constraints on $\boldsymbol{\delta}$, thus it would be simpler working with the vector $\boldsymbol{\delta'}$ having $n-2$ free parameters. $\boldsymbol{\delta}$ and $\boldsymbol{\delta'}$ can be linked via an orthogonal dimension (or rank) reduction matrix $\boldsymbol{Z}$, of size $n \times n-2$, so that $\boldsymbol{\delta} = \boldsymbol{Z\delta'}$ and the constraint term becomes $\boldsymbol{T^TZ\delta'} = \boldsymbol{T^T\delta} = 0$. By constraining $\boldsymbol{T^TZ}$ to be $\boldsymbol{0}$, the constraint is satisfied for any value $\boldsymbol{\delta'}$, which then becomes an unconstrained vector. The matrix $\boldsymbol{Z}$ is obtained via a QR decomposition (of $\boldsymbol{T}$), for further details see Appendix C.1. Equation 3.3.5 can be reformulated using this reparameterisation as:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{EZ\delta'}||^2 + \lambda\boldsymbol{\delta'^T Z^T EZ\delta'} \tag{C.1.6}$$

Finally, the above can be rewritten as in Equation 3.3.1:

$$min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda\boldsymbol{\beta^T S\beta} \tag{C.1.7}$$

Where $\boldsymbol{\beta}_{[n\times 1]} = \begin{bmatrix} \alpha_0 & \alpha_1 & \delta'_1 & \cdots & \delta'_{n-2} \end{bmatrix}^T$ and :

$$\boldsymbol{X}_{[n\times n]} = \begin{bmatrix} \boldsymbol{T}_{[n\times 2]} & \boldsymbol{E}_{[n\times n]}\boldsymbol{Z}_{[n\times n-2]} \end{bmatrix} \qquad \boldsymbol{S}_{[n\times n]} = \left[\begin{array}{c|c} \mathbf{0}_{[2\times 2]} & \mathbf{0}_{[2\times n-2]} \\ \hline \mathbf{0}_{[n-2\times 2]} & \boldsymbol{Z^T E Z}_{[n-2\times n-2]} \end{array}\right]$$

C.2 Further details for Section 3.3.4

Recall that Equation 3.3.8 describes a knots based thin plate spline as follows:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda \boldsymbol{\delta^T E \delta} \quad \text{s.t} \quad \boldsymbol{C\delta} = \mathbf{0} \tag{C.2.1}$$

Thus knots-based NCS are defined by $k+2$ parameters, subject to two constraints. As for optimal NCS, a QR decomposition (of $\boldsymbol{C}^T$) can be employed to find $\boldsymbol{Z}$ (of size $k\times k-2$) so that $\boldsymbol{\delta} = \boldsymbol{Z\delta'}$ and $\boldsymbol{CZ} = \mathbf{0}$; this makes $\boldsymbol{\delta'}$ unconstrained. The PLS criterion (Equation 3.3.1) for knots-based NCS is:

$$min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta} \tag{C.2.2}$$

where $\boldsymbol{\beta}_{[k\times 1]} = \begin{bmatrix} \alpha_0 & \alpha_1 & \delta'_1 & \cdots & \delta'_{k-2} \end{bmatrix}^T$ and:

$$\boldsymbol{X}_{[n\times k]} = \begin{bmatrix} \boldsymbol{T}_{[n\times 2]} & \boldsymbol{E}_{[n\times k]}\boldsymbol{Z}_{[k\times k-2]} \end{bmatrix} \qquad \boldsymbol{S}_{[k\times k]} = \left[\begin{array}{c|c} \mathbf{0}_{[2\times 2]} & \mathbf{0}_{[2\times k-2]} \\ \hline \mathbf{0}_{[k-2\times 2]} & \boldsymbol{Z^T E Z}_{[k-2\times k-2]} \end{array}\right]$$

C.3 Further details for Section 3.3.5

This Appendix demonstrates the missing algebra steps, necessary to achieve the usual PLS criterion (Equation 3.3.1) from the PLS criterion for thin plate regression splines (Equation 3.3.10).

Recall that $\boldsymbol{U_k}$ is obtained from the eigen-decomposition of matrix $\boldsymbol{E}$. The following identities hold:

$$\begin{aligned} \boldsymbol{EU_k} &= \boldsymbol{U_k D_k} \\ \boldsymbol{D_k} &= \boldsymbol{U_k^T E U_k} \end{aligned}$$

Building upon linear algebra results, Wood (2003) shows that $\widetilde{e_k}$ and $\widehat{e_k}$ are jointly minimized by the "optimal" reduction basis $\boldsymbol{\Gamma_k} = \boldsymbol{U_k}$. It follows that:

$$\begin{aligned} \widetilde{E_k} &= \boldsymbol{E U_k U_k^T} = \boldsymbol{U_k D_k U_k^T} \\ \widehat{E_k} &= \boldsymbol{U_k U_k^T E U_k U_k^T} = \boldsymbol{U_k D_k U_k^T} \end{aligned}$$

Thus the "optimal" low-rank fitting criterion in Equation 3.3.11 becomes:

$$min||\boldsymbol{y} - \boldsymbol{U_k D_k U_k^T \delta} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta^T U_k D_k U_k^T} \delta \quad \text{s.t.} \quad \boldsymbol{T^T \delta} = 0 \tag{C.3.1}$$

Given that $\boldsymbol{\delta_k} = \boldsymbol{U_k^T \delta}$, this can be re-expressed in terms of parameters $\boldsymbol{\delta_k}$ matching Equation 3.3.10:

$$min||\boldsymbol{y} - \boldsymbol{E U_k \delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta_k^T D_k \delta_k} \quad \text{s.t.} \quad \boldsymbol{T^T U_k \delta_k} = 0 \tag{C.3.2}$$

The constrained minimisation problem can be turned into an unrestricted one by finding (via the QR decomposition) an orthogonal matrix $\boldsymbol{Z}$ (size $k \times k-2$) so that $\boldsymbol{T^T U_k Z} = 0$ and $\boldsymbol{\delta_k} = \boldsymbol{Z\delta'}$:

$$min||\boldsymbol{y} - \boldsymbol{E U_k Z \delta'} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta'^T Z^T D_k Z \delta'} \tag{C.3.3}$$

This can be expressed within the usual penalised regression framework (Equation 3.3.1):

$$min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta} \tag{C.3.4}$$

Where: $\boldsymbol{\beta}_{[k\times 1]} = \begin{bmatrix} \alpha_1 & \alpha_2 & \delta'_1 & \cdots & \delta'_{k-2} \end{bmatrix}^T$ and:

$$\boldsymbol{X}_{[n\times k]} = \begin{bmatrix} \boldsymbol{T}_{[n\times 2]} & \boldsymbol{E}_{[n\times n]} \boldsymbol{U_k}_{[n\times k]} \boldsymbol{Z}_{[k\times k-2]} \end{bmatrix} \quad \boldsymbol{S}_{[k\times k]} = \left[\begin{array}{c|c} \boldsymbol{0}_{[2\times 2]} & \boldsymbol{0}_{[2\times k-2]} \\ \hline \boldsymbol{0}_{[k-2\times 2]} & \boldsymbol{Z^T D_k Z}_{[k-2\times k-2]} \end{array} \right]$$

C.4 Further details for Section 3.3.7

P-splines of degree d, with a penalty matrix of order r, can be expressed within the following penalised regression framework.

$$min\ ||\boldsymbol{y} - \boldsymbol{X_d \beta}||^2 + \lambda \boldsymbol{\beta^T S_r \beta} \tag{C.4.1}$$

Where:

$$\boldsymbol{\beta}_{[k\times 1]} = \begin{bmatrix}\beta_1 & \cdots & \beta_{k-1+d}\end{bmatrix}^T \quad \boldsymbol{X_d}_{[n\times(k+1-d)]} = \begin{bmatrix} B_1^d(x_1) & \cdots & B_{k-1+d}^d(x_1) \\ \vdots & \vdots & \vdots \\ B_1^d(x_n) & \cdots & B_{k-1+d}^d(x_n)\end{bmatrix}$$

$$\boldsymbol{S_r}_{[(k-1+d)\times(k-1+d)]} = \left[\begin{array}{c|c} \mathbf{0}_{[r\times r]} & \mathbf{0}_{[r\times(k-1+d-r)]} \\ \hline \mathbf{0}_{[(k-1+d-r)\times r]} & (\boldsymbol{D_r^T D_r})_{[(k-1+d-r)\times(k-1+d-r]} \end{array}\right]$$

d and r are suppressed for notational simplicity, letting $\boldsymbol{X} = \boldsymbol{X_d}$ and $\boldsymbol{S} = \boldsymbol{S_r}$.

C.5 Further details for Section 3.3.8

C.5.1 GLMs and GAMs

Generalised Linear Models (GLMs) first introduced by Nelder and Wedderburn (1972) offer an extension to ordinary linear regression (McCullagh and Nelder, 1989). These allow for the outcome variable $\boldsymbol{y}$ to be from any distribution in the exponential family $f_\psi(\boldsymbol{y})$ and use a link function $\eta(\boldsymbol{\mu})$ to linearly relate the expected response of $\boldsymbol{Y}$ with the unknown model parameters. If the identity function is specified as link function a simple linear model is obtained. Mathematically a GLM is defined as follows:

$$y_i \sim f_\psi(y_i) = exp\left(\frac{y_i\psi - b(\psi)}{a(\phi)} + c(y_i, \phi)\right)$$
$$E[\boldsymbol{y}] = \boldsymbol{\mu} = \eta^{-1}(\boldsymbol{X\beta})$$

A list of distributions from the exponential family and related parameters ϕ, functions $a(\phi)$, $b(\psi)$, $c(y_i, \phi)$ and link functions $\eta(\boldsymbol{\mu})$ can be found in Wood (2006a), page 61.

Splines can be employed within a GLM framework to model the link function $\eta(\boldsymbol{\mu}) = \boldsymbol{X\beta}$ in a smooth rather than linear fashion. Parameter estimates are obtained by maximizing a penalised log-likelihood criterion:

$$\max_{\boldsymbol{\beta}}\ l(\boldsymbol{y}|\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S\beta} \tag{C.5.1}$$

The penalty term is the same as for the PLS criterion (Equation 3.3.1) while the residual sum of squares, quantifying goodness of fit, is replaced by the log-likelihood $l(\boldsymbol{y}|\boldsymbol{\beta})$.

For a fixed λ, Equation C.5.1 only has an analytic solution for normally distributed outcomes $\boldsymbol{y}$; this is equivalent to the PLS one (Equation 3.3.2).

Generalised Additive Models (GAMs) enhance the flexibility of GLMs, by allowing the link function $\eta(\boldsymbol{\mu})$, of data arising from the exponential function, to be written as a sum of smooth functions (Hastie and Tibshirani, 1990; Wood, 2006a, Chapter 3). An example of a link function for data $\boldsymbol{y}$ (arising from the exponential family) modelled with a GAM is:

$$\eta(\boldsymbol{\mu}) = \alpha + \boldsymbol{X}\boldsymbol{\psi} + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) + \ldots \tag{C.5.2}$$

The above is made of three components: the intercept α, a strictly parametric term $\boldsymbol{X}\boldsymbol{\psi}$ and smooth non-parametric functions f_i, typically specified by splines. Parameter estimates can be obtained within the aforementioned penalised regression framework, by using the Penalised Iteratively Re-weighted Least Squares (P-IRLS) algorithm to obtain $\hat{\boldsymbol{\beta}}$ numerically, for a given λ (Wood, 2006a).

C.5.2 Smoothing parameter selection criterion

Increasing values of λ reduce the effective number of degrees of freedom for $\hat{\boldsymbol{\beta}}$. Wood, 2006a, Section 4.4 defines these as the trace of the hat matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{S})^{-1}\boldsymbol{X}^{\boldsymbol{T}} \tag{C.5.3}$$

Where $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the estimated mean and $\boldsymbol{W}$ be a diagonal matrix of weights, with entries $\boldsymbol{W}_{ii} = (\eta'(\hat{\mu}_i)V(\hat{\mu}_i))^{-1}$.

The optimal amount of smoothing $\hat{\lambda}$ is generally estimated via Cross Validation (CV), which is a measure of the model's predictive performance: for a given λ value, the model is fitted using all but one data point. The fitted model is subsequently used to predict the data point which was originally excluded. This procedure is repeated for all data points. In practice, it is typically unnecessary to re-fit the model sequentially excluding all data points, as, for GLM, CV can be expressed analytically. However, CV is not scale invariant: different parameterisations of $\boldsymbol{X}$ lead to the same values of $\hat{\boldsymbol{\beta}}$ but to different CV scores. Generalized Cross Validation (GCV) is an extension of CV that is scale-invariant; the optimal $\hat{\lambda}$ is chosen to minimize the GCV score. To avoid naively evaluating GCV over a grid of plausible λ values, Wood (2006a) Sections 4.6, 4.7, proposes a Newton-Raphson algorithm to minimize GCV with respect to λ, in combination with P-IRLS to maximize $\hat{\beta}$ conditional on λ. GCV is equal to:

$$GCV(\lambda) = \frac{m||\sqrt{\boldsymbol{W}}(\boldsymbol{z}-\boldsymbol{X}\hat{\boldsymbol{\beta}})||^2}{(m-tr(\boldsymbol{H}))^2} \tag{C.5.4}$$

In the above formula $\boldsymbol{z}$ are "pseudo-data", $z_i = \eta'(\hat{\boldsymbol{\mu}})(y_i - \hat{\mu}_i) + \boldsymbol{X_i}\hat{\boldsymbol{\beta}}$. $\boldsymbol{H}$ is the hat matrix defined above.

C.5.3 Confidence intervals

Within a GLM framework, confidence intervals for parameters are typically constructed by exploiting the asymptotic normality property of estimators: $\hat{\boldsymbol{\beta}} \sim N(E[\hat{\boldsymbol{\beta}}], V[\hat{\boldsymbol{\beta}}])$, where $E[\hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\beta}}$ and $V[\hat{\boldsymbol{\beta}}] = \mathcal{I}^{-1}$ (the inverse of Fisher's information matrix). However, within a penalised regression (or likelihood) framework, bias is introduced to reduce variance (see Section 3.3.2). $\boldsymbol{E}[\hat{\boldsymbol{\beta}}]$ is no longer equal to $\hat{\boldsymbol{\beta}}$ and thus asymptotic confidence intervals have poor coverage.

An empirical Bayesian approach is used to construct approximate confidence intervals, based on the Bayesian re-interpretation of penalised likelihood, described in Section 3.3.9. Equation 3.3.14 yields priors for penalised ($\boldsymbol{\beta_p}$) and unpenalised ($\boldsymbol{\beta_u}$) coefficients that are equivalent, from a Bayesian perspective, to the penalty matrix **S**.

Approximate confidence intervals are constructed from a large-sample approximation of the posterior distribution (Wood, 2006c), given in the Equation below, to circumvent the use of Markov Chain Monte Carlo (MCMC) which is computationally cumbersome.

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\hat{\boldsymbol{\beta}}, (\boldsymbol{X^T W X} + \lambda \boldsymbol{S})^{-1}\phi\right) \tag{C.5.5}$$

Despite the prior on $\boldsymbol{\beta_U}$ being improper, the resulting posterior distribution is proper. The above expression is exact if observations $\boldsymbol{y}$ are normally distributed. $\boldsymbol{W}$ is a diagonal matrix of weights and is distribution specific (see Appendix C.5.2). Confidence intervals are obtained by sampling from the approximate posterior distribution; these have been shown to have satisfactory coverage properties (Wood, 2006c).

Confidence intervals could be further modified to account for uncertainty in λ via a simulation based method, proposed by Wood (2006c). However this is computationally intensive and has hardly been used in practice.

C.6 Further details for Section 3.3.9

In the two following Subsections the mathematical details of the centering and prior-precision reparameterisation are presented. The starting spline, which is subject to re-parameterisations, is characterised by parameters $\boldsymbol{\beta}$ (size k) and design $\boldsymbol{X}$ and penalty $\boldsymbol{S}$ matrices, respectively of size $n \times k$ and $k \times k$.

C.6.1 Centering re-parameterisation

Any spline, as defined above, can be made subject to the constraint that the sum of the spline values over the coordinates of the data is equal to zero (*i.e.* $\sum_{i=1}^{n} g(x_i) = 0$). In matrix notation this is equivalent to

$$\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} = 0 \tag{C.6.1}$$

Where $\mathbf{1}$ is a $n \times 1$ vector. The above constraint can be integrated using the usual QR decomposition (of $\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}$) approach. An orthogonal matrix $\boldsymbol{Z}$, of dimension $k \times (k-1)$, is found so that $\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\beta}'$ and $\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{Z} = 0$. After reparameterisation, the quadratic penalty matrix for $\boldsymbol{\beta}'$, of size $(k-1) \times (k-1)$, is $\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{Z}$.

Notice that in integrating such constraint results in loosing a degree of freedom. This is compensated by the introduction of a global intercept α. Thus the resulting reparametrized spline has k parameters $\tilde{\boldsymbol{\beta}} = \begin{bmatrix}\alpha & \boldsymbol{\beta}'\end{bmatrix}$, design matrix $\tilde{\boldsymbol{X}} = \begin{bmatrix}\mathbf{1}^T & \boldsymbol{X}\boldsymbol{Z}\end{bmatrix}$ of size $k \times k$ and finally $\tilde{\boldsymbol{S}} = \begin{bmatrix}\mathbf{0}^T \\ \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{Z}\end{bmatrix}$, where $\mathbf{0}$ is a vector of zeroes of size k.

Further details about this re-parameterisation are available in Wood, 2006a, Section 4.2.

C.6.2 Prior-precision re-parameterisation

To increase the efficiency of MCMC sampling, splines with a single penalty matrix $\boldsymbol{S}$ (*i.e.* all univariate splines) can be reparametrized to have an identity matrix as penalty matrix.

Recall that $\boldsymbol{S}$ is reinterpreted, within a Bayesian framework, as the precision matrix of a multivariate normal prior on coefficients $\boldsymbol{\beta}$. Hence having $\boldsymbol{S} = \boldsymbol{I}$ leads to iid Normal priors for the $\boldsymbol{\beta}$ components. This reparameterisation is only undertaken for computational purposes, as having a multivariate Normal prior on $\boldsymbol{\beta}$, rather than iid Normal priors on the component of $\boldsymbol{\beta}$, leads to slower MCMC updating, in my experience.

Consider any $(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{S})$ spline where $\rho \leq k$ is the rank of $\boldsymbol{S}$. If some of the $\boldsymbol{\beta}$ coefficients are unpenalised, then $\rho < k$. Denote $\boldsymbol{\beta_U}$ and $\boldsymbol{\beta_P}$, of size $k-\rho$ and ρ, unpenalised and penalised coefficients.

Apply the eigendecomposition $\boldsymbol{S} = \boldsymbol{UDU^T}$. $\boldsymbol{D}$ is a diagonal matrix, with entries being the eigenvalues of $\boldsymbol{S}$ sorted in ascending order and $\boldsymbol{U}$ is the matrix of corresponding eigenvectors. Due to positive semi-definiteness of $\boldsymbol{S}$ all eigenvalues are positive, with the exception of $k-\rho$ zero eigenvalues. Further let $\boldsymbol{\Lambda}$ be a diagonal matrix, with entries $\Lambda_{ii} = \sqrt{D_{ii}}$ so that $\boldsymbol{D} = \boldsymbol{\Lambda^T \Lambda}$. The penalty becomes:

$$\lambda \boldsymbol{\beta^T S \beta} = \lambda \boldsymbol{\beta^T U D U^T \beta} = \lambda \boldsymbol{\beta^T U \Lambda^T \Lambda U^T \beta} \tag{C.6.2}$$

Where:

$$\boldsymbol{U\Lambda^T \Lambda U^T} = \left[\begin{array}{c|c} \boldsymbol{I}_{[\rho\times\rho]} & \boldsymbol{0}_{[\rho\times k-\rho]} \\ \hline \boldsymbol{0}_{[k-\rho\times k-\rho]} & \boldsymbol{0}_{[k-\rho\times k-\rho]} \end{array}\right]$$

Now let $\boldsymbol{\beta' = \Lambda U^T \beta}$ and notice that the penalty term is equivalent to:

$$\boldsymbol{\beta^T U \Lambda^T \Lambda U^T \beta} = \boldsymbol{\beta_P^T I \beta_P}$$

By orthogonality of $\boldsymbol{U}$ it holds that:

$$\boldsymbol{\beta} = (\boldsymbol{\Lambda U^T})^{-1}\boldsymbol{\beta'} = (\boldsymbol{U^T})^{-1}(\boldsymbol{\Lambda})^{-1}\boldsymbol{\beta'} = (\boldsymbol{U}^{-1})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta'} = \boldsymbol{U\Lambda}^{-1}\boldsymbol{\beta'}$$

Hence the original design and penalty matrices of the spline are now reparametrized as:

$$\boldsymbol{X\beta} = \boldsymbol{XU\Lambda}^{-1}\boldsymbol{\beta'} \qquad \lambda\boldsymbol{\beta^T S\beta} = \lambda\boldsymbol{\beta_P^T I \beta_P}$$

Further details about this reparameterisation are available in Wood (2006a). `JAGS` requires specifying a proper prior distribution for all parameters, including originally unpenalised $\boldsymbol{\beta_U}$. It is suggested to use a vague proper prior, usually iid Normal i.i.d for $\beta'_{Uj} \sim N(0, 1/\lambda_0)$. Hence the original penalty is now replaced by the following approximation:

$$\lambda \boldsymbol{\beta^T S \beta} \approx \lambda \boldsymbol{\beta_P^T I \beta_P} + \lambda_0 \boldsymbol{\beta_U^T I \beta_U} \tag{C.6.3}$$

C.7 Further details for Section 3.4.4

If a GP (Normal) prior is assumed over a set of functions, modelling normally distributed data, then the log-likelihood is equal to:

$$-\frac{1}{2}log|\boldsymbol{K}_y| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y} - \frac{n}{2}log(2\pi) \tag{C.7.1}$$

where $\boldsymbol{K}_y = \boldsymbol{K} + \sigma^2\boldsymbol{I}$. The (marginal) log-likelihood depends on both hyper-parameters $\boldsymbol{\phi}$ and covariates $\boldsymbol{x}$, as these define $\boldsymbol{K_y}$ via a covariance function (see Section 3.4.2).

The terms of the log-likelihood have a precise interpretation. As the length-scale ρ increases and the magnitude η decreases, the model becomes smoother (hence simpler) but less flexible, resulting in a poorer model fit. The $-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y}$ term relates the hyperparameters to data-fit, as it increases for larger ρ and smaller η. The $-\frac{1}{2}log|\boldsymbol{K}_y|$ term is a negative complexity penalty, which counteracts overfitting by increasing for increasing ρ and decreasing η. The last term is a normalizing constant.

An expression for hyper-parameters $\hat{\boldsymbol{\phi}}$, maximizing the likelihood can not be derived analytically. The likelihood may instead be numerically maximized by gradient-descent techniques (Rasmussen and Williams, 2006, Section 5.4) as partial derivatives of the covariance matrix with respect to each parameter can be typically derived. These techniques may suffer from multiple local optima, especially for a large number of hyperparameters $\boldsymbol{\phi}$. Alternative parameter estimation methods via predictive measures, *e.g.* cross validation, have been explored (Sundararajan and Keerthi, 2000).

C.8 Further details for Section 3.5

Splines can be used for scatter-plot smoothing purposes, for smoothly modelling the expected response of a GLM, or as part of more complex models such as: GAM, Generalized Additive Mixed Models (GAMM) (Wood, 2006a) and Structured Additive Regression (STAR) (Belitz and Lang, 2008; Brezger and Lang, 2006; Fahrmeir et al., 2004).

To date, there are two main R packages for splines: **BayesX** (Kneib et al., 2014) and **mgcv** (Wood, 2012). The former builds upon the P-splines work of Eilers and Marx (2003, 1996) and considers STAR models in a penalised regression, Bayesian and empirical Bayesian (mixed model) frameworks. **mgcv** is based upon the extensive work of Wood (2006a, 2003, 2016) and implements all splines discussed (and some others too...) for scatter-plot smoothing, GLM, GAM and GAMM models within a penalised regression framework.

The *jagam* function in the **mgcv** package automatically generates `JAGS` code for a spline estimation in a Bayesian framework, with the parameterisations described in Section 3.3.9 (Wood, 2016). **BayesX** is more efficient for standard GLM, GAMM and STAR model estimation. However, the flexibility of `JAGS` can be exploited to extend such model, including them as a part of a more complex stochastic model.

A number of packages are available for GP: **gpfit** (MacDonald et al., 2015), **gptk** (Kalaitzis et al., 2015), **tgp** (Gramacy, 2007); the first two implement likelihood inference, whereas **tgp** focuses on Bayesian estimation. `Stan` can be also used for Bayesian inference (Flaxman et al., 2015; Stan Development Team, 2016b). Rue et al. (2009) propose a fast approximation to MCMC for latent Gaussian models (which include both GP and splines), based on Integrated Nested Laplace Approximations (INLA).

Appendix D

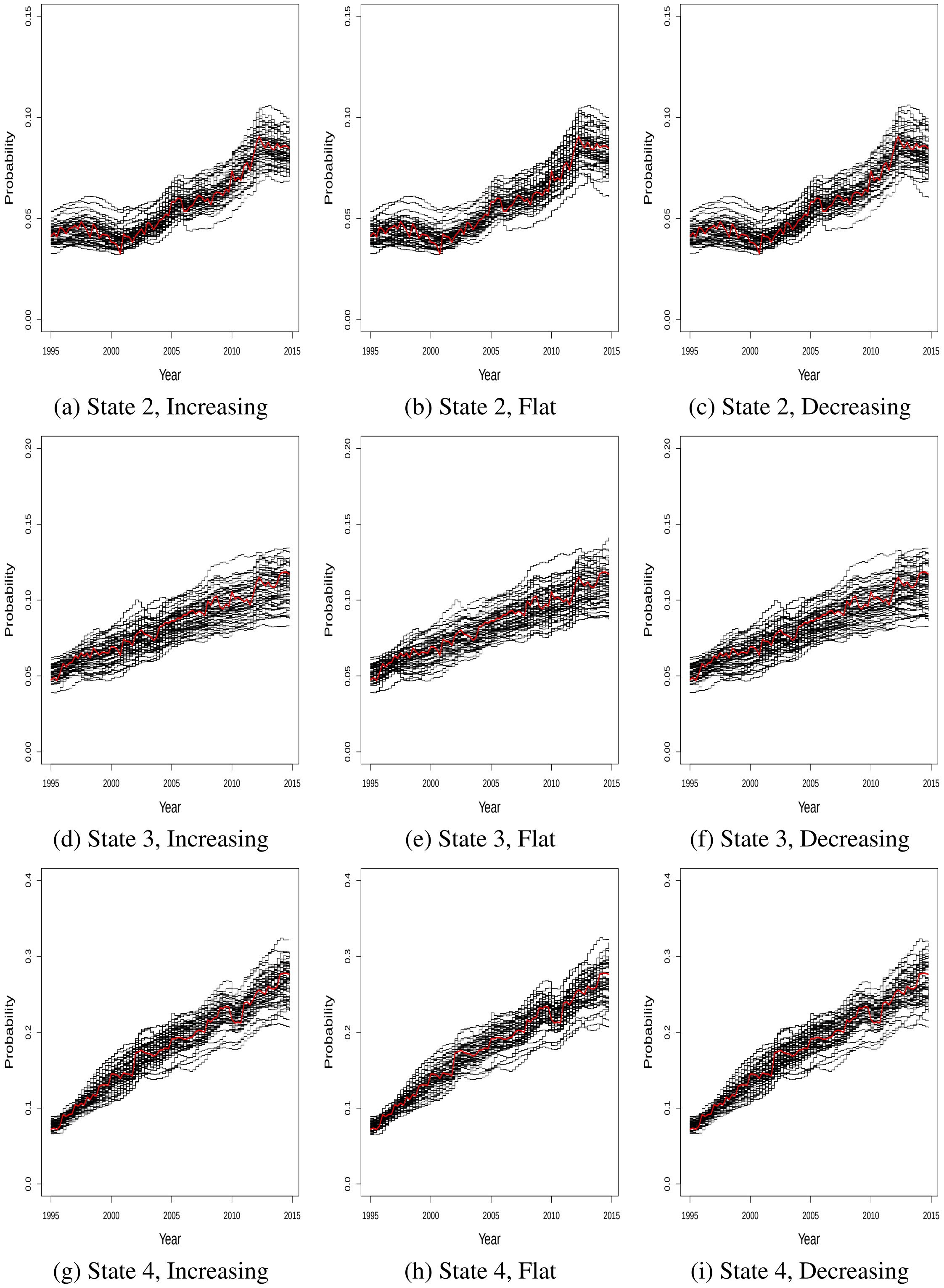
Appendix for Chapter 4

D.1 Further details for Section 4.6.2

D.1.1 Estimated diagnosis probabilities from states 2, 3, and 4

In this Section posterior means (estimates) of diagnosis probabilities from states 2, 3 and 4 for all incidence models, under the three true incidence scenarios, are presented. Recall that diagnosis probabilities (from all states) are constant in the three incidence scenarios. It can be observed that, for all incidence models and under all incidence scenarios, diagnosis probabilities are reasonably well estimated and, unlike diagnosis probabilities from state 1, no consistent bias appears in the most recent years.

rword1 incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

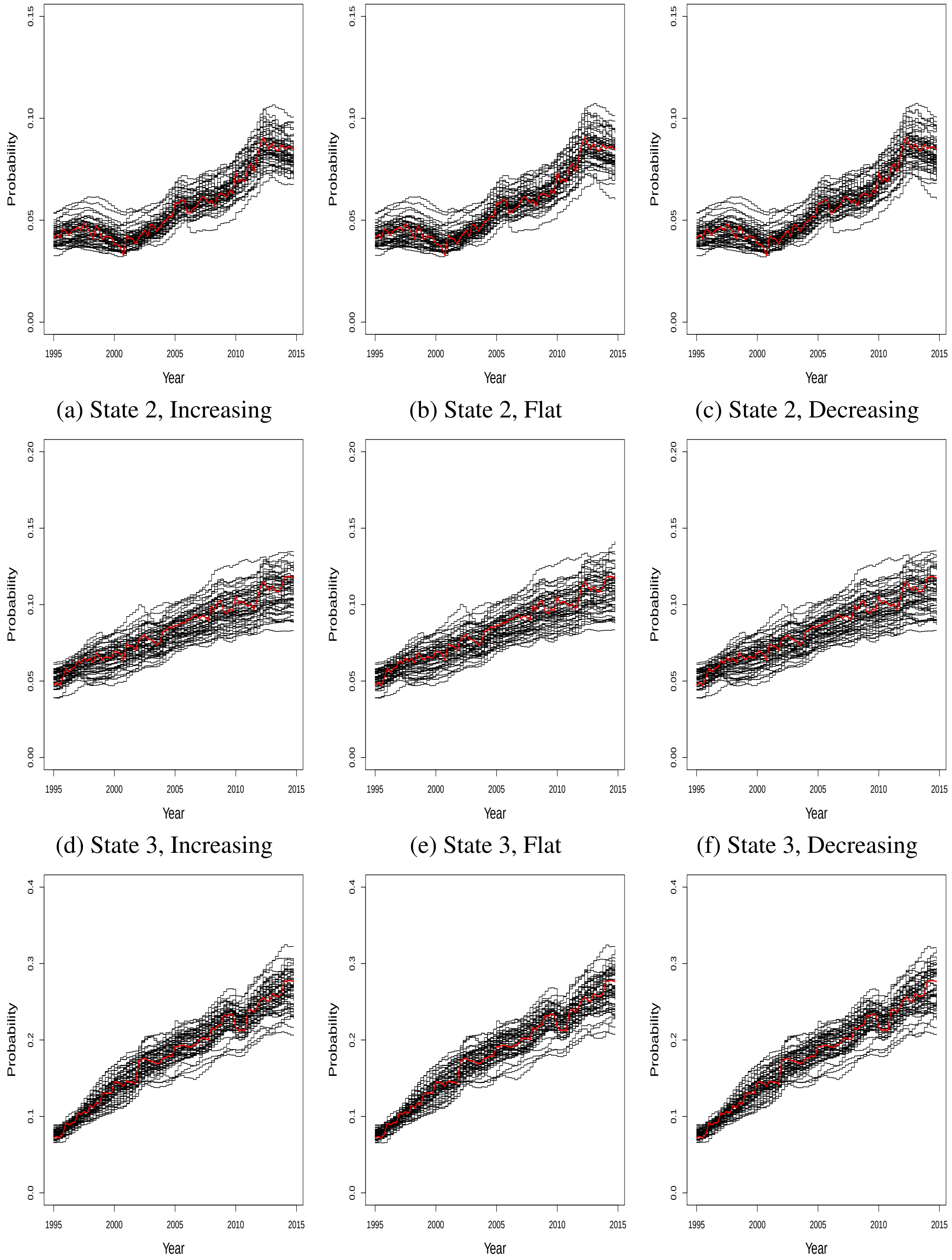
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. D.1 Estimated diagnosis probabilities from states 2, 3 and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

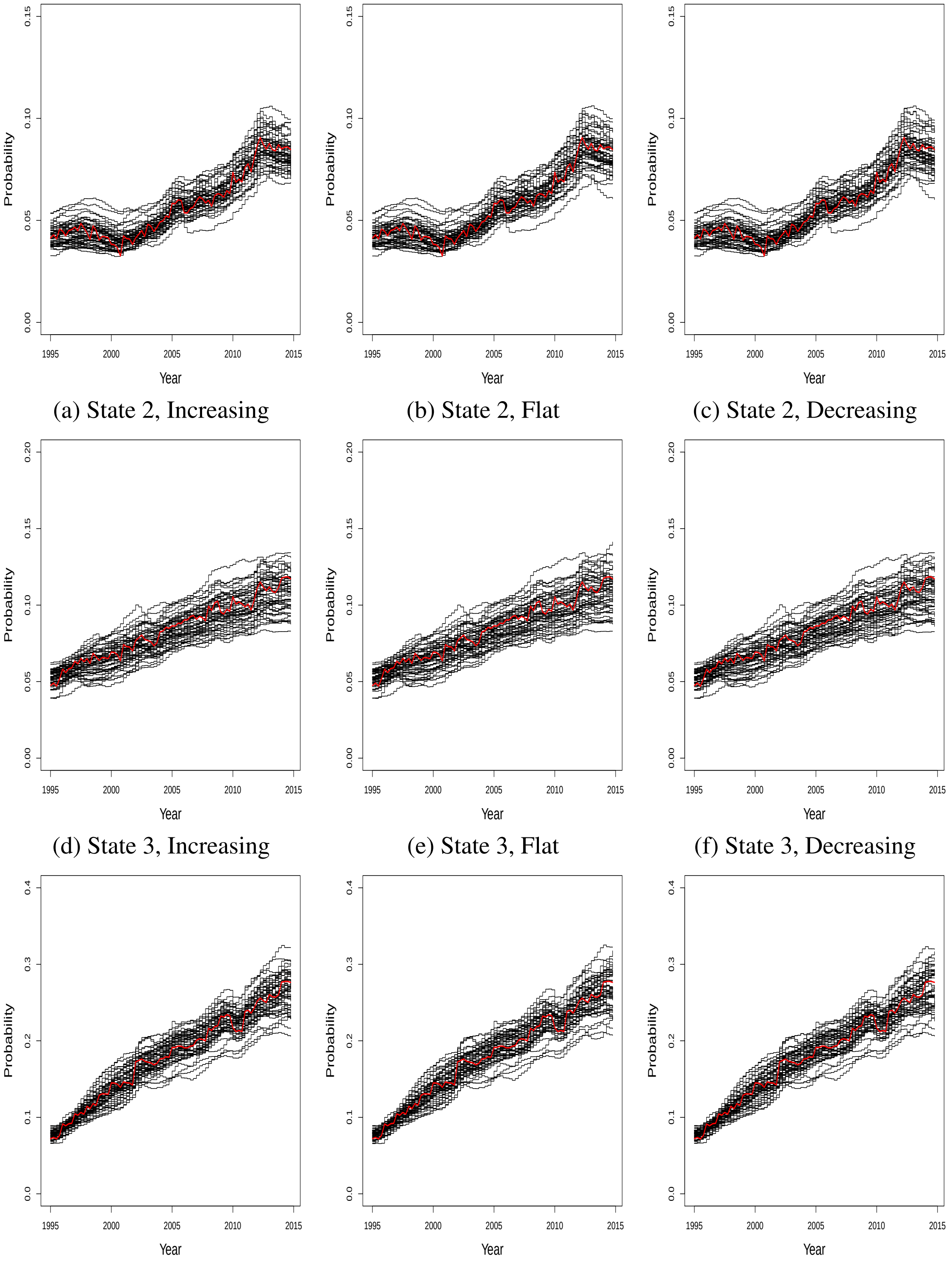
ts incidence model



(a) State 2, Increasing
(b) State 2, Flat
(c) State 2, Decreasing
(d) State 3, Increasing
(e) State 3, Flat
(f) State 3, Decreasing
(g) State 4, Increasing
(h) State 4, Flat
(i) State 4, Decreasing

Fig. D.2 Estimated diagnosis probabilities from states 2, 3 and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

bsord1 incidence model



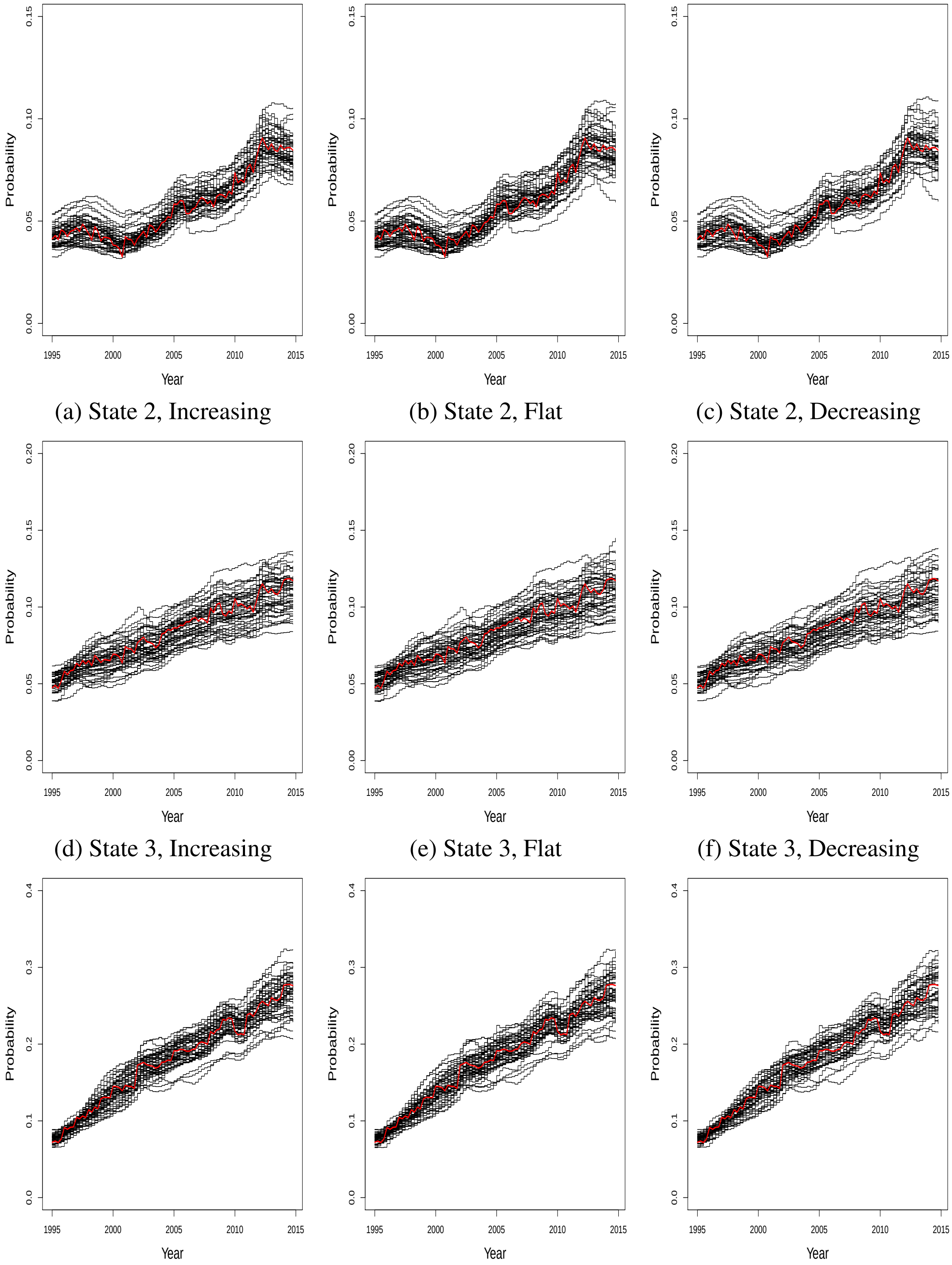
(a) State 2, Increasing (b) State 2, Flat (c) State 2, Decreasing

(d) State 3, Increasing (e) State 3, Flat (f) State 3, Decreasing

(g) State 4, Increasing (h) State 4, Flat (i) State 4, Decreasing

Fig. D.3 Estimated diagnosis probabilities from states 2, 3 and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

GP incidence model



(a) State 2, Increasing (b) State 2, Flat (c) State 2, Decreasing

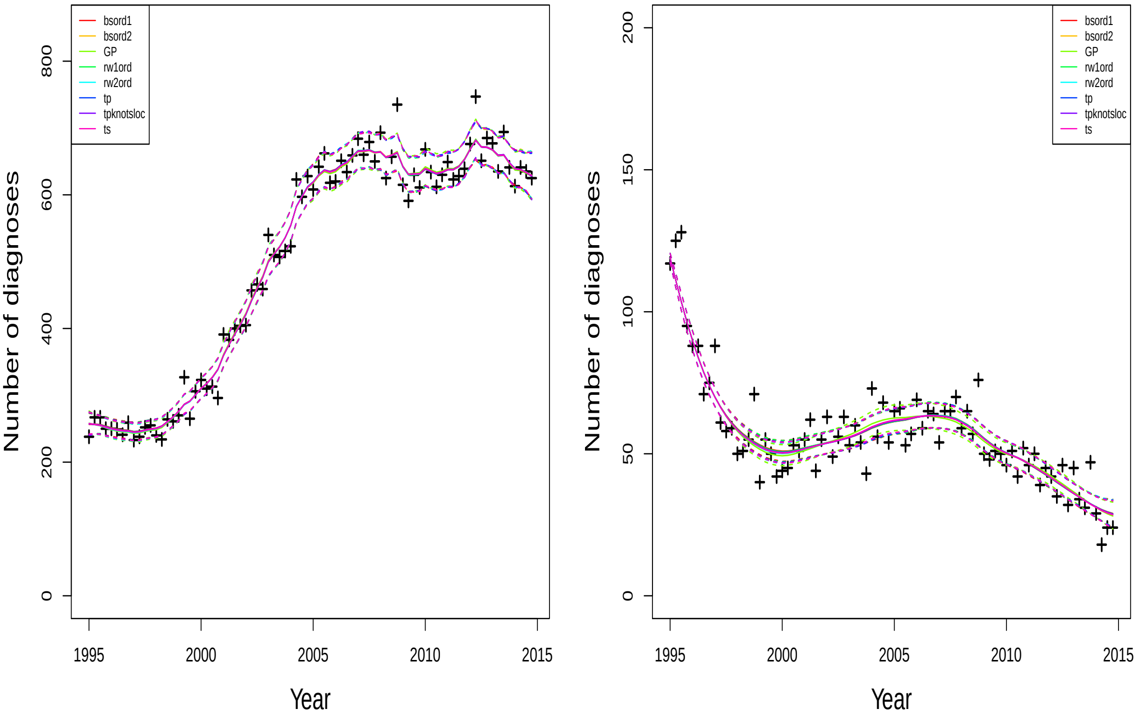
(d) State 3, Increasing (e) State 3, Flat (f) State 3, Decreasing

(g) State 4, Increasing (h) State 4, Flat (i) State 4, Decreasing

Fig. D.4 Estimated diagnosis probabilities from states 2, 3 and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

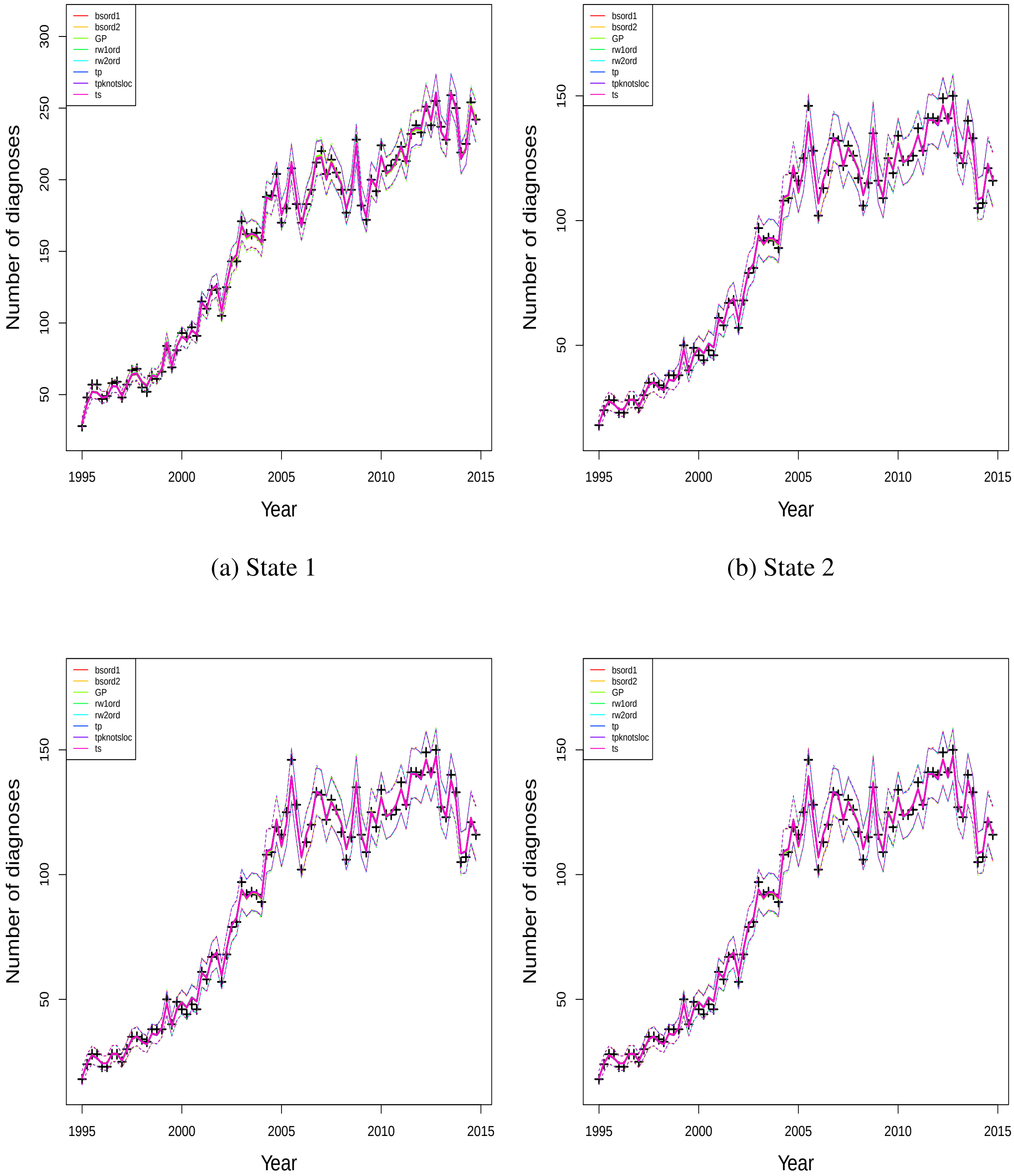
D.1.2 Goodness of fit plots

In this Section, goodness of fit plots are presented, for the dataset number 25 generated under the true flat incidence scenario. This comprises three data sources: HIV, AIDS and CD4. After estimating the parameters in a Bayesian framework, fitted values for the three different data-sources are obtained, using four different incidence models (*rw*, *bsord1*, *ts*, *GP*). All incidence models considered provide very similar and very satisfactory fit for all data-sources considered.



(a) HIV (b) AIDS

Fig. D.5 Goodness of fit plots for HIV and AIDS diagnoses in dataset 25 (generated with flat incidence). The coloured lines and crosses respectively represent all incidence models and the simulated data points



(a) State 1

(b) State 2

(c) State 3

(d) State 4

Fig. D.6 Goodness of fit plots for CD4 diagnoses, stratified by state as indicated in captions, for data 25 (generated with flat incidence). The coloured lines and crosses respectively represent all incidence models and the simulated data points.

Appendix E

Appendix for Chapter 5

E.1 Notation

Recall that $(t_0, t_T]$ is the time-period spanning the HIV epidemic, which is split into T disjoint, consecutive intervals $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$. Moreover $(a_0, a_A]$ is the age-range spanning the HIV epidemic, which is split into A disjoint, consecutive intervals $(a_{j-1}, a_j]$, $j = \{1, \dots, A\}$. Age and time intervals are typically assumed to have equal lengths, with the exception of Section 5.4.3.

- The i$^{\text{th}}$ interval refers to $(t_{i-1}, t_i]$, $i = \{1, \dots, T\}$.
- The j$^{\text{th}}$ interval refers to $(a_{j-1}, a_j]$, $j = \{1, \dots, A\}$.
- The i_0^{th} interval usually denotes $(t_{i_0-1}, t_{i_0}]$, $i_0 = \{1, \dots, T\}$, when infections occur.
- The j_0^{th} interval usually denotes $(a_{j_0-1}, a_{j_0}]$, $j_0 = \{1, \dots, A\}$, when infections occur.

A summary of the notation introduced in Chapter 5 is here presented: appropriate definitions are given in the main body, the list below only serves as reference:

- K is the number of undiagnosed states in the proposed multi-state model.
  States 1 to K denote latent undiagnosed states, state $2K + 1$ corresponds to AIDS diagnosis and states $K + 1$ to $2K$ identify HIV diagnosis with a certain CD4-count.
  $k = \{1, \dots, 2K + 1\}$ denotes one of the states.
- $h_{i_0, j_0} \equiv h_{i_0, j_0}(\boldsymbol{\theta})$ represents the expected number of infections occurring in the i_0^{th} time and j_0^{th} age intervals.

$\mathcal{H} \equiv \mathcal{H}(\boldsymbol{\theta}) = \{h_{11}, \ldots, h_{TA}\}$ denotes the *incidence surface*, *i.e.* the expected number of infection over time and age, parametrised by $\boldsymbol{\theta}$.

- $\boldsymbol{d}_{i,j} \equiv \boldsymbol{d}_{i,j}(\boldsymbol{\delta}) = (d_{1,i,j}, \ldots, d_{K,i,j})$ refers to the diagnosis probabilities in the ith time and jth age intervals from undiagnosed states k=$\{1, \ldots, K\}$.

  $\mathcal{D} \equiv \mathcal{D}(\boldsymbol{\delta}) = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T\}$ refers to the collection of diagnosis probabilities over time, parametrised by $\boldsymbol{\delta}$.

- $\boldsymbol{q}^{j_0} = (q^{j_0}_{1,2}, \ldots, q^{j_0}_{K,K+1})$ denotes progression probabilities between undiagnosed states $k = \{1, \ldots, K\}$, for an individual infected in the j_0^{th} age interval (and hence implicitly in the i_0^{th} time interval).

  $\mathcal{Q} = \{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^A\}$ denotes the collection of age-at-infection dependent progression probabilities, which are known from external cohort studies.

- $\boldsymbol{Q}^{j_0}_{i,j} \equiv \boldsymbol{Q}^{j_0}_{i,j}(\boldsymbol{\delta})$, of size $K \times K$, and $\boldsymbol{D}^{j_0}_{i,j} \equiv \boldsymbol{D}^{j_0}_{i,j}(\boldsymbol{\delta})$, of size $K \times K+1$ are the transition and progression matrices of the multi-state model. They describe the transition probabilities between the undiagnosed states $k = \{1, \ldots, K\}$ and from the undiagnosed to diagnosis states $k = \{K+1, \ldots, 2K+1\}$ respectively, in the ith time and jth age intervals for individuals infected in the j_0^{th} interval (and hence implicitly in the i_0^{th} time interval).

- $\boldsymbol{e}^{j_0}_{i,j} \equiv \boldsymbol{e}^{j_0}_{i,j}(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a $K \times 1$ vector, containing the expected number of individuals in undiagnosed states $k = \{1, \ldots, K\}$ in the end of the ith time and jth age intervals for individuals infected in the j_0^{th} interval.

- $\boldsymbol{\mu}^{j_0}_{i,j} \equiv \boldsymbol{\mu}^{j_0}_{i,j}(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a $(K+1) \times 1$ vector, describing the expected number of new arrivals in diagnoses states $k = \{K+1, \ldots, 2K\}$ at the end of the ith and jth age intervals for individuals infected in the j_0^{th} interval.

  $\mu^A_{i,j} \equiv \mu^A_{i,j}(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\mu^H_{i,j} \equiv \mu^H_{i,j}(\boldsymbol{\theta}, \boldsymbol{\delta})$ respectively refer to the expected number of new AIDS and HIV diagnoses at the end of the ith time and the jth age intervals.

- $\boldsymbol{p}_{i,j} \equiv \boldsymbol{p}_{i,j}(\boldsymbol{\theta}, \boldsymbol{\delta})$ describes the expected proportion of diagnoses in each of the HIV-states $k = \{K+1, \ldots, 2K\}$ at the end of the ith time and jth age intervals.

- $\boldsymbol{\upsilon} = \{\upsilon^H_{11}, \ldots \upsilon^H_{TA}, \upsilon^A_{11}, \ldots, \upsilon^A_{TA}\}$ denotes the collection of under-reporting parameters over time-intervals.

- $\boldsymbol{\pi}_j$ is a $K \times 1$ vectors, that refers to the initial expected number of undiagnosed infections in states $k = \{1, \ldots, K\}$ for individuals aged j.

Appendix F

Appendix for Chapter 6

F.1 Further details for Section 6.2.2

Hence, the smoothing criterion (Equation 6.2.3) can be reformulated as:

$$min||\boldsymbol{y} - \boldsymbol{T\alpha} - \boldsymbol{E\delta}||^2 + \lambda\boldsymbol{\delta^T E\delta} \quad \text{s.t} \quad \boldsymbol{T^T\delta} = \boldsymbol{0}$$

The TPS parameters $\boldsymbol{\delta}$ are subject to three constraints. A $n-3$ vector of unconstrained parameters $\boldsymbol{\delta'}$ is defined by letting $\boldsymbol{\delta} = \boldsymbol{Z\delta'}$, where $\boldsymbol{Z}$ is a $n \times n-3$ matrix, obtained via a QR decomposition of $\boldsymbol{T}$, so that $\boldsymbol{T^T Z} = \boldsymbol{0}$. Hence Equation 6.2.3 can be rewritten within the usual penalised regression framework:

$$min\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda\boldsymbol{\beta^T S\beta} \tag{F.1.1}$$

where $\boldsymbol{\beta}_{[n\times 1]} = \begin{bmatrix}\alpha_0 & \alpha_1 & \alpha_2 & \delta'_1 & \cdots & \delta'_{n-3}\end{bmatrix}^T$ and

$$\boldsymbol{X}_{[n\times n]} = \begin{bmatrix}\boldsymbol{T}_{[n\times 3]} & \boldsymbol{E}_{[n\times n]}\boldsymbol{Z}_{[n\times n-3]}\end{bmatrix} \qquad \boldsymbol{S}_{[n\times n]} = \left[\begin{array}{c|c}\boldsymbol{0}_{[3\times 3]} & \boldsymbol{0}_{[2\times n-3]} \\ \hline \boldsymbol{0}_{[n-3\times 3]} & \boldsymbol{Z^T E Z}_{[n-3\times n-3]}\end{array}\right]$$

F.2 Further details for Section 6.2.3

We aim to express a knots-based TPS within the usual PLS criterion (Equation 6.2.1) for penalized regression:

$$min\,||\boldsymbol{y}-\boldsymbol{X\beta}||^2+\lambda\boldsymbol{\beta^T S\beta} \tag{F.2.1}$$

where $\boldsymbol{\beta}_{[k\times 1]}=\begin{bmatrix}\alpha_0 & \alpha_1 & \alpha_2 & \delta'_1 & \cdots & \delta'_{k-3}\end{bmatrix}^T$ and:

$$\boldsymbol{X}_{[n\times k]}=\begin{bmatrix}\boldsymbol{T}_{[n\times 3]} & \boldsymbol{E}_{[n\times k]}\boldsymbol{Z}_{[k\times k-3]}\end{bmatrix} \qquad \boldsymbol{S}_{[k\times k]}=\left[\begin{array}{c|c}\mathbf{0}_{[3\times 3]} & \mathbf{0}_{[3\times k-3]}\\ \hline \mathbf{0}_{[k-3\times 3]} & \boldsymbol{Z^T E Z}_{[k-3\times k-3]}\end{array}\right]$$

where $\boldsymbol{Z}$ is a $k\times k-3$ matrix, obtained via a QR decomposition of $\boldsymbol{C}$, so that $\boldsymbol{\delta}=\boldsymbol{Z\delta'}$ and $\boldsymbol{CZ}=\mathbf{0}$. $\boldsymbol{Z}$ links the vector $\boldsymbol{\delta}$ of k coefficients (subject to three constraints) to the vector $\boldsymbol{\delta'}$ of $k-3$ unconstrained coefficients.

F.3 Further details for Section 6.2.4

We aim to express a thin-plate regression spline, within the usual PLS criterion (Equation 6.2.1) for penalized regression:

$$min\,||\boldsymbol{y}-\boldsymbol{X\beta}||^2+\lambda\boldsymbol{\beta^T S\beta} \tag{F.3.1}$$

where $\boldsymbol{\beta}_{[k\times 1]}=\begin{bmatrix}\alpha_1 & \alpha_2 & \delta'_1 & \cdots & \delta'_{k-2}\end{bmatrix}^T$ and:

$$\boldsymbol{X}_{[n\times k]}=\begin{bmatrix}\boldsymbol{T}_{[n\times 3]} & \boldsymbol{E}_{[n\times n]}\boldsymbol{U_k}_{[n\times k]}\boldsymbol{Z}_{[k\times k-3]}\end{bmatrix} \quad \boldsymbol{S}_{[k\times k]}=\left[\begin{array}{c|c}\mathbf{0}_{[3\times 3]} & \mathbf{0}_{[3\times k-3]}\\ \hline \mathbf{0}_{[k-3\times 3]} & \boldsymbol{Z^T D_k Z}_{[k-3\times k-3]}\end{array}\right]$$

where $\boldsymbol{E}$ and $\boldsymbol{T}$, $\boldsymbol{D_k}$ and $\boldsymbol{U_k}$ are defined in Sections 6.2.3 and 6.2.4. The n-vector of $\boldsymbol{\delta}$ parameters defining the optimal thin plate spline is linked to a $k-3$ vector of unconstrained parameters $\boldsymbol{\delta}=\boldsymbol{U_k Z\delta'}$. $\boldsymbol{Z}$ is a $k\times k-3$ matrix so that $\boldsymbol{T^T U_k Z}=0$, obtained via the QR decomposition of $\boldsymbol{T}$.

F.4 Further details for Section 6.2.7

In Appendix C.6 the following decompostion was considered so that the penalty/precision matrix can be reparametrized to have an identity matrix as prior considering the eigen-decomposition of $\boldsymbol{S}$.

$$\lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} = \lambda \boldsymbol{\beta}^T \boldsymbol{U} \boldsymbol{D} \boldsymbol{U}^T \boldsymbol{\beta} = \lambda \boldsymbol{\beta}^T \boldsymbol{U} \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \boldsymbol{U}^T \boldsymbol{\beta} \tag{F.4.1}$$

Letting $\boldsymbol{\beta}' = \boldsymbol{\Lambda} \boldsymbol{U}^T \boldsymbol{\beta}$, the penalty term is equivalent to:

$$\boldsymbol{\beta}^T \boldsymbol{U} \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \boldsymbol{U}^T \boldsymbol{\beta} = \boldsymbol{\beta}_P^T \boldsymbol{I} \boldsymbol{\beta}_P$$

Where $\boldsymbol{\beta}_P$ are the originally penalized parameters of the spline.

Unfortunately splines with two (or more) penalty terms can not be re-parametrized as above. This is because:

$$\lambda \boldsymbol{\beta}^T \boldsymbol{S}_1 \boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{S}_2 \boldsymbol{\beta} = \lambda \boldsymbol{\beta}^T \boldsymbol{U}_1 \boldsymbol{\Lambda}_1^T \boldsymbol{\Lambda}_1 \boldsymbol{U}_1^T \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^T \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^T \boldsymbol{\Lambda}_2 \boldsymbol{U}_2^T \boldsymbol{\beta}$$

Reparametrizing as $\boldsymbol{\beta}' = \boldsymbol{\Lambda} \boldsymbol{U}^T \boldsymbol{\beta}$ is not possible because $\boldsymbol{\Lambda}_1 \neq \boldsymbol{\Lambda}_2$ and thus the penalty term can not be rewritten as above.

Appendix G

Appendix for Chapter 7

G.1 Further details for Section 7.2

The table below reports the expected number of initially undiagnosed individuals $\boldsymbol{\pi}$ chosen for the simulation study, for each CD4 undiagnosed state and age at time 1.

Age at t_1	Stage1	Stage2	Stage3	Stage4
	$(CD4 \geq 500)$	$(500 < CD4 \leq 350)$	$(350 < CD4 \leq 200)$	$(CD4 > 200)$
1	8.36	1.75	1.75	0.31
2	13.14	3.90	3.90	1.16
3	15.86	5.91	5.91	2.32
4	17.41	7.56	7.56	3.54
5	19.84	11.06	11.06	6.64
6	21.24	13.74	13.74	9.34
7	22.04	15.74	15.74	11.56
8	22.44	17.08	17.08	13.17
9	22.62	17.97	17.97	14.30
10	22.74	18.81	18.81	15.43
11	46.25	24.55	24.55	17.30
12	59.51	31.31	31.31	20.54
13	66.91	37.09	37.09	23.91
14	71.01	41.83	41.83	27.44
15	71.96	43.54	43.54	28.93
16	72.42	44.86	44.86	30.23

Age at t_1	Stage1	Stage2	Stage3	Stage4
	$(CD4 \geq 500)$	$(500 < CD4 \leq 350)$	$(350 < CD4 \leq 200)$	$(CD4 > 200)$
17	72.61	45.83	45.83	31.28
18	72.60	46.48	46.48	32.04
19	72.51	46.91	46.91	32.56
20	72.33	47.06	47.06	32.74
21	72.13	47.19	47.19	32.89
22	71.93	47.28	47.28	33.01
23	71.72	47.33	47.33	33.07
24	71.50	47.37	47.37	33.11
25	71.28	47.37	47.37	33.09
26	71.05	47.36	47.36	33.07
27	70.82	47.34	47.34	33.03
28	70.59	47.32	47.32	33.00
29	70.35	47.29	47.29	32.96
30	70.11	47.26	47.26	32.91
31	39.80	40.21	40.21	31.53
32	23.02	31.72	31.72	28.02
33	13.74	24.04	24.04	23.40
34	8.59	17.83	17.83	18.68
35	5.78	13.19	13.19	14.53
36	4.18	9.72	9.72	11.02
37	3.27	7.18	7.18	8.21
38	2.78	5.51	5.51	6.23
39	2.53	4.42	4.42	4.87
40	2.37	3.62	3.62	3.82
41	2.26	2.99	2.99	2.97
42	2.19	2.50	2.50	2.28
43	2.17	2.42	2.42	2.16
44	2.15	2.35	2.35	2.06
45	2.14	2.33	2.33	2.03
46	2.12	2.31	2.31	2.01
47	2.11	2.30	2.30	2.00
48	2.10	2.28	2.28	1.98
49	2.08	2.27	2.27	1.97
50	2.07	2.26	2.26	1.95
51	2.06	2.24	2.24	1.94
52	5.88	11.43	11.43	11.48

The table below reports the progression probabilities $\mathcal{Q}$ values used in the simulation study, for each CD4 undiagnosed state and age at time 1.

Age at inf	Stage1	Stage2	Stage3	Stage4
a_0	$(CD4 \geq 500)$	$(500 < CD4 \leq 350)$	$(350 < CD4 \leq 200)$	$(CD4 > 200)$
1	0.09	0.12	0.11	0.14
2	0.09	0.12	0.11	0.14
3	0.09	0.12	0.11	0.14
4	0.09	0.12	0.11	0.14
5	0.09	0.12	0.11	0.14
6	0.09	0.12	0.11	0.14
7	0.09	0.12	0.11	0.14
8	0.09	0.12	0.11	0.14
9	0.09	0.12	0.11	0.14
10	0.09	0.12	0.11	0.14
11	0.09	0.12	0.11	0.14
12	0.09	0.12	0.11	0.14
13	0.09	0.12	0.11	0.15
14	0.09	0.12	0.11	0.15
15	0.09	0.12	0.12	0.15
16	0.09	0.12	0.12	0.15
17	0.09	0.12	0.12	0.15
18	0.09	0.12	0.12	0.15
19	0.09	0.12	0.12	0.15
20	0.10	0.12	0.12	0.15
21	0.10	0.12	0.12	0.15
22	0.10	0.12	0.12	0.15
23	0.10	0.12	0.12	0.15
24	0.10	0.12	0.12	0.15
25	0.10	0.12	0.12	0.15
26	0.10	0.12	0.12	0.15
27	0.10	0.12	0.12	0.15
28	0.10	0.12	0.12	0.15
29	0.10	0.12	0.12	0.16
30	0.10	0.12	0.12	0.16

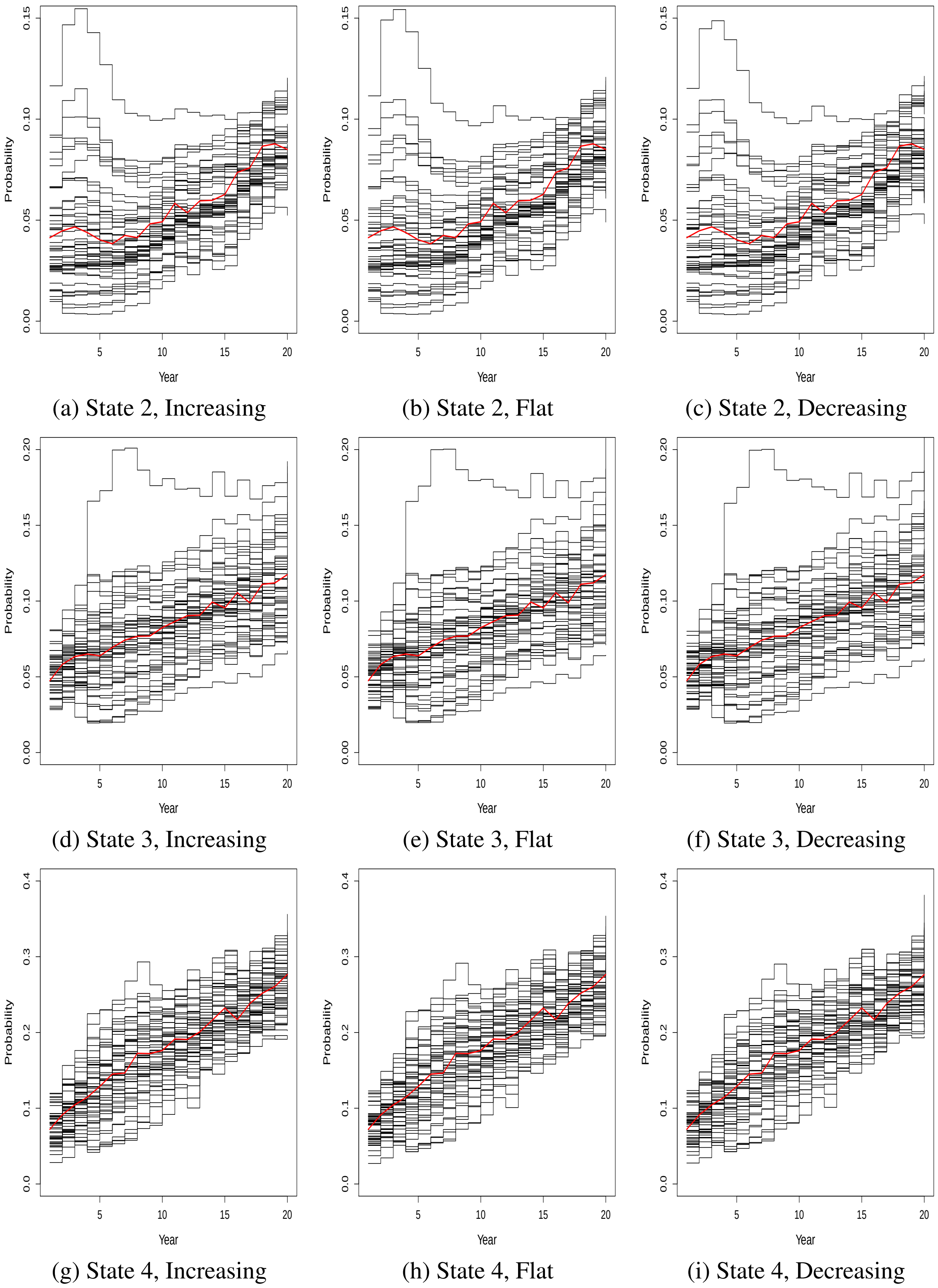
Age at inf	Stage1	Stage2	Stage3	Stage4
a_0	$(CD4 \geq 500)$	$(500 < CD4 \leq 350)$	$(350 < CD4 \leq 200)$	$(CD4 > 200)$
31	0.10	0.12	0.12	0.16
32	0.10	0.12	0.12	0.16
33	0.10	0.12	0.12	0.16
34	0.10	0.12	0.13	0.16
35	0.10	0.12	0.13	0.16
36	0.10	0.12	0.13	0.16
37	0.10	0.12	0.13	0.16
38	0.10	0.12	0.13	0.16
39	0.10	0.12	0.13	0.17
40	0.10	0.12	0.13	0.17
41	0.10	0.12	0.13	0.17
42	0.10	0.13	0.13	0.17
43	0.10	0.13	0.14	0.17
44	0.10	0.13	0.14	0.17
45	0.10	0.13	0.14	0.18
46	0.11	0.13	0.14	0.18
47	0.11	0.13	0.14	0.18
48	0.11	0.13	0.14	0.18
49	0.11	0.13	0.14	0.18
50	0.11	0.13	0.15	0.19
51	0.11	0.13	0.15	0.19
52	0.11	0.13	0.15	0.19

G.2 Further details for Section 7.6.2

G.2.1 Estimated diagnosis probabilities from states 2, 3, and 4

In this Section the posterior means (estimates) of diagnosis probabilities from states 2, 3 and 4 are depicted for all incidence models, under the three true incidence scenarios. Recall that the true diagnosis probabilities (from all states) are constant in the three true incidence scenarios. It can be observed that, for all incidence models and under all incidence scenarios, diagnosis probabilities are reasonably well estimated and, unlike diagnosis probabilities from state 1, no consistent bias appears in the most recent years.

***tpknotsloc* incidence model**



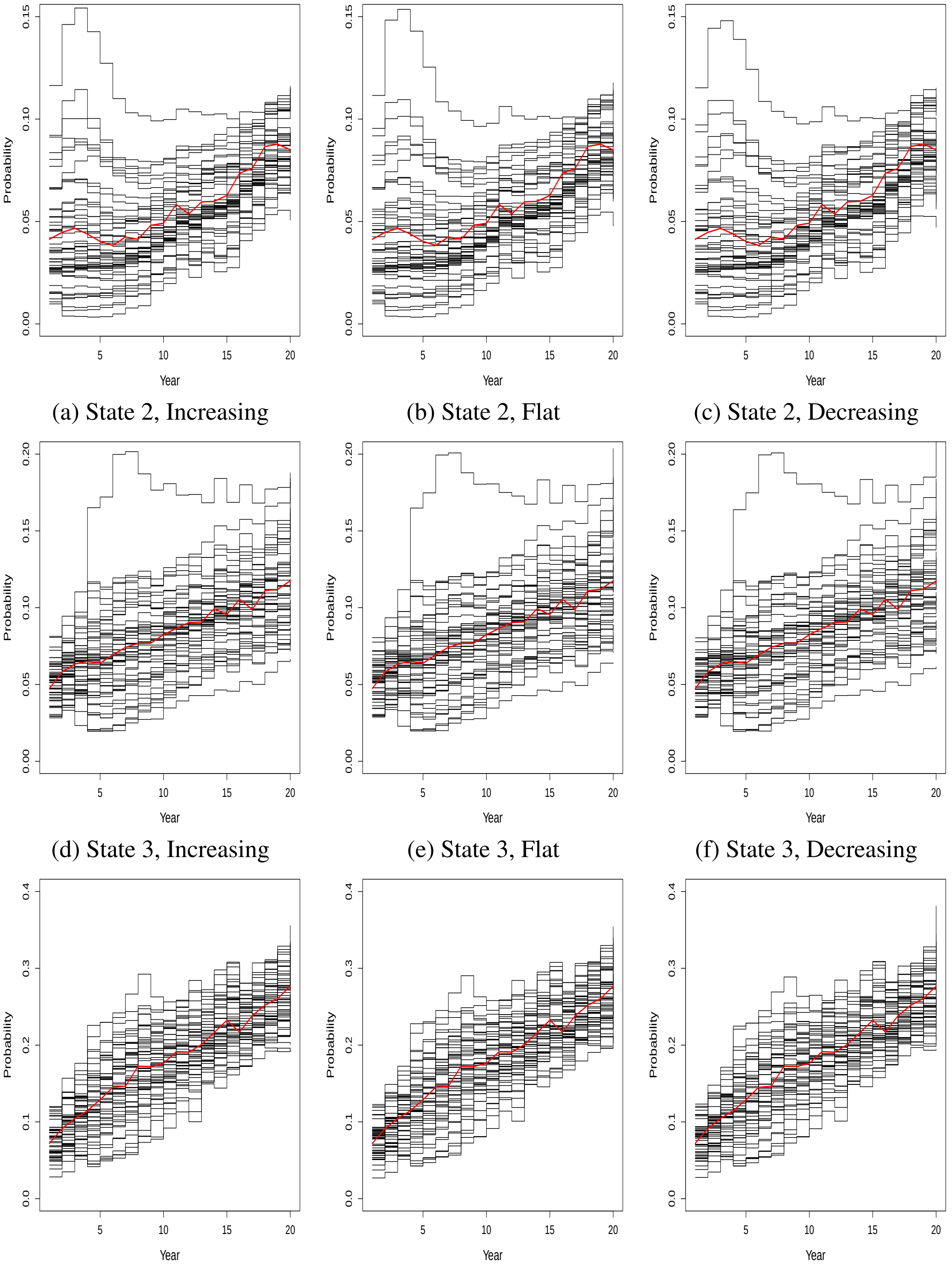
(a) State 2, Increasing (b) State 2, Flat (c) State 2, Decreasing

(d) State 3, Increasing (e) State 3, Flat (f) State 3, Decreasing

(g) State 4, Increasing (h) State 4, Flat (i) State 4, Decreasing

Fig. G.1 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

tp incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

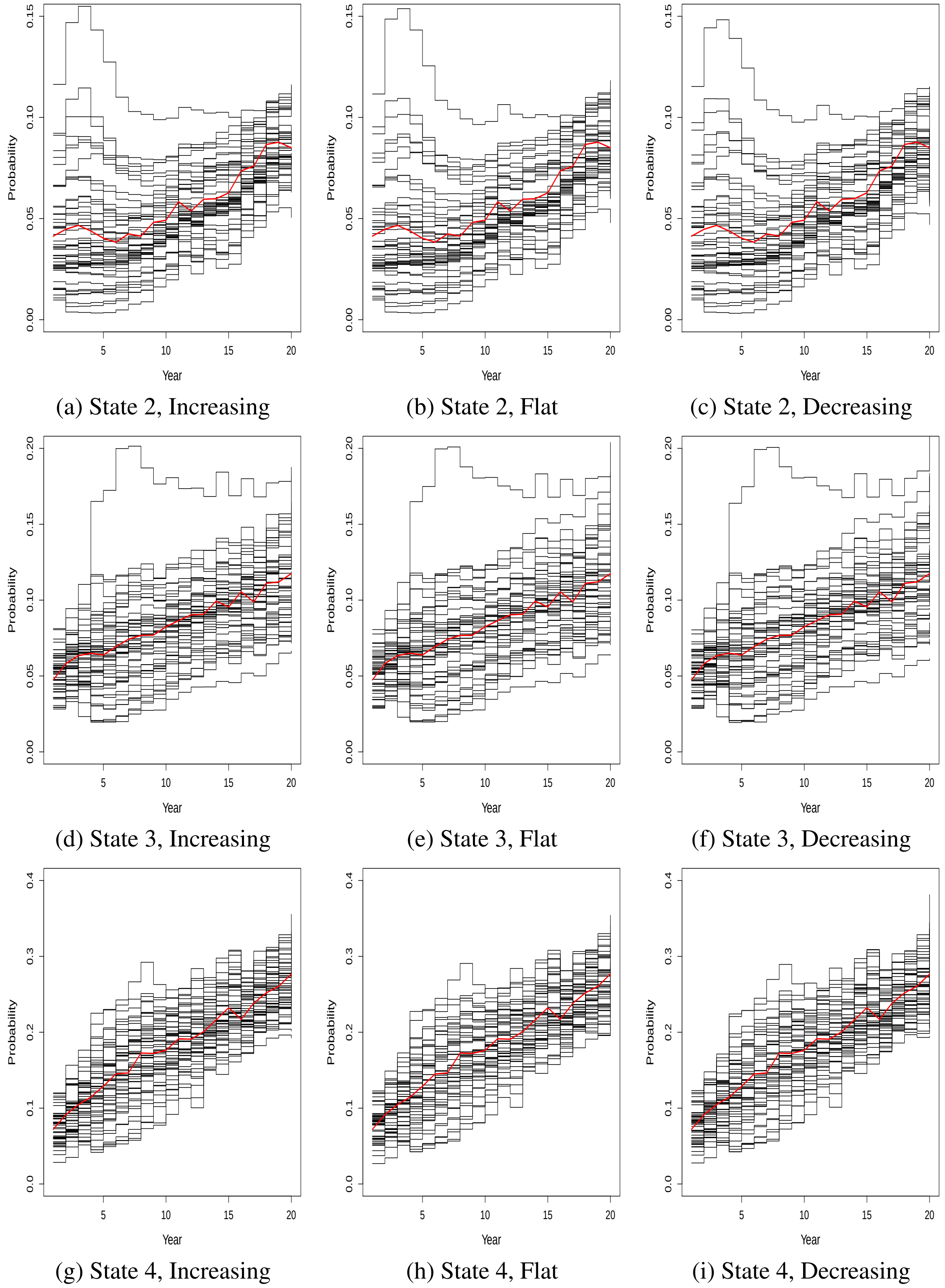
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.2 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

ts incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

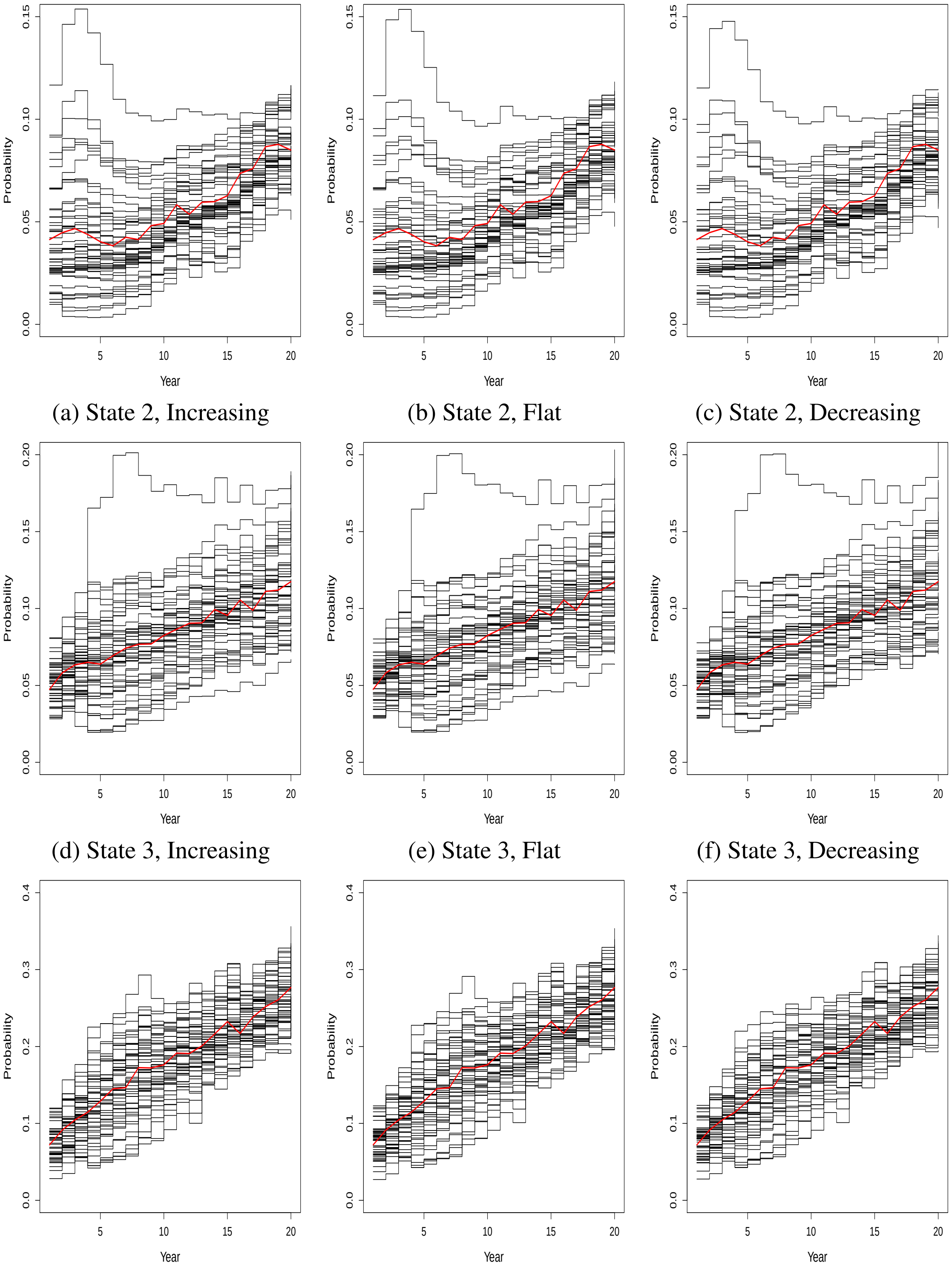
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.3 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

ptenstp incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

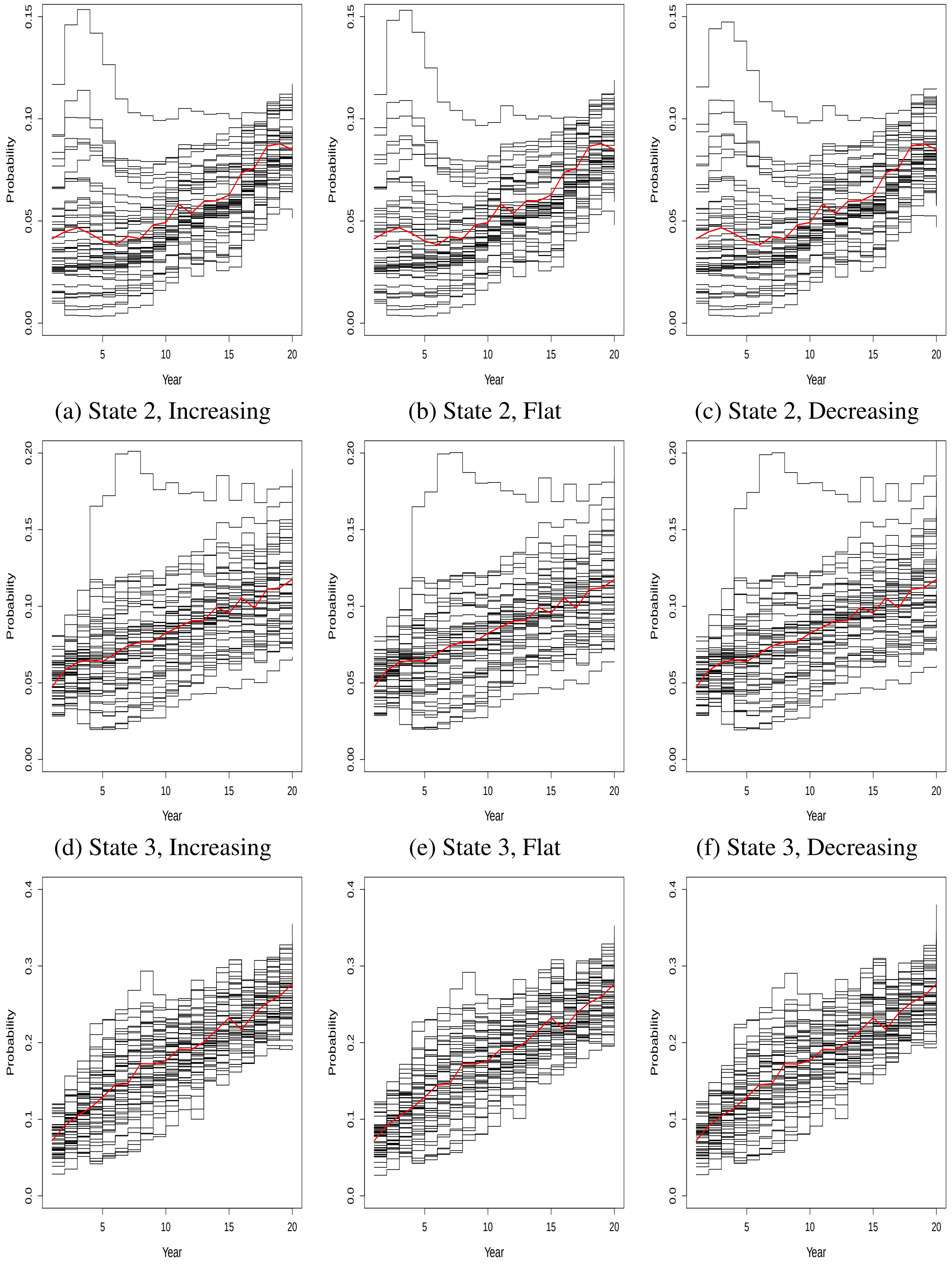
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.4 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

ptensts incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

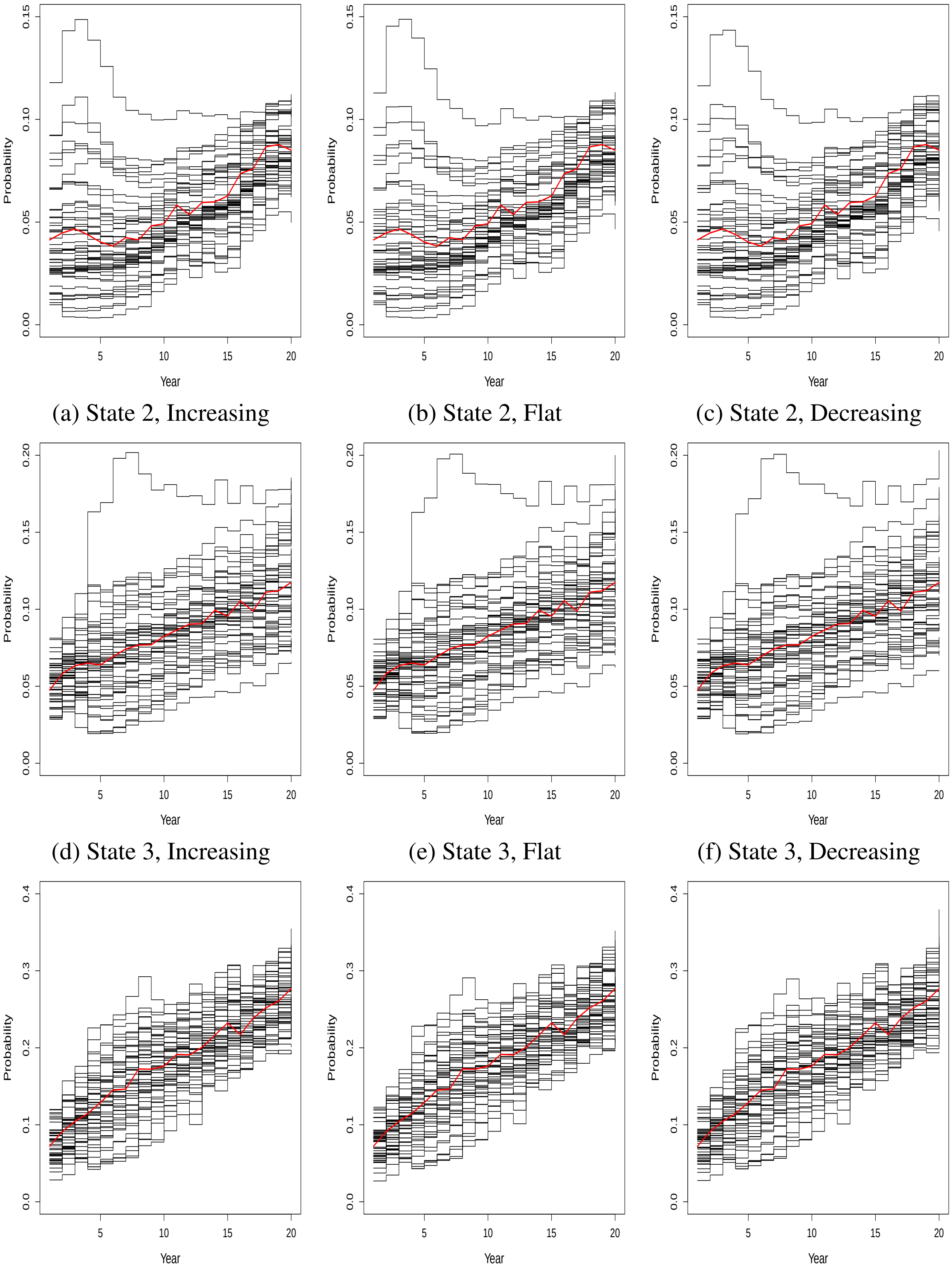
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.5 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

ptensbsord1 incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

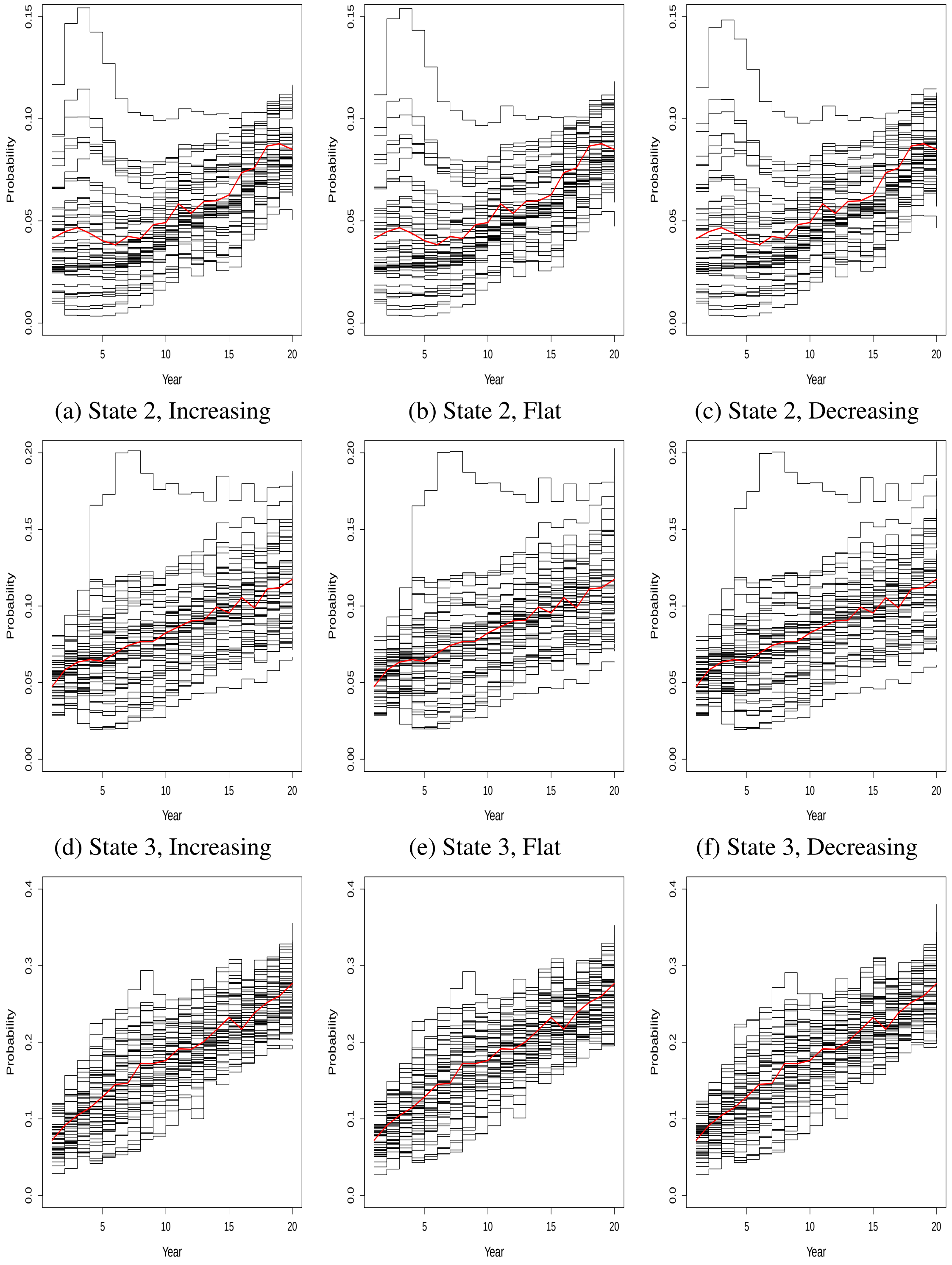
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.6 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

ptensbsord2 incidence model



(a) State 2, Increasing

(b) State 2, Flat

(c) State 2, Decreasing

(d) State 3, Increasing

(e) State 3, Flat

(f) State 3, Decreasing

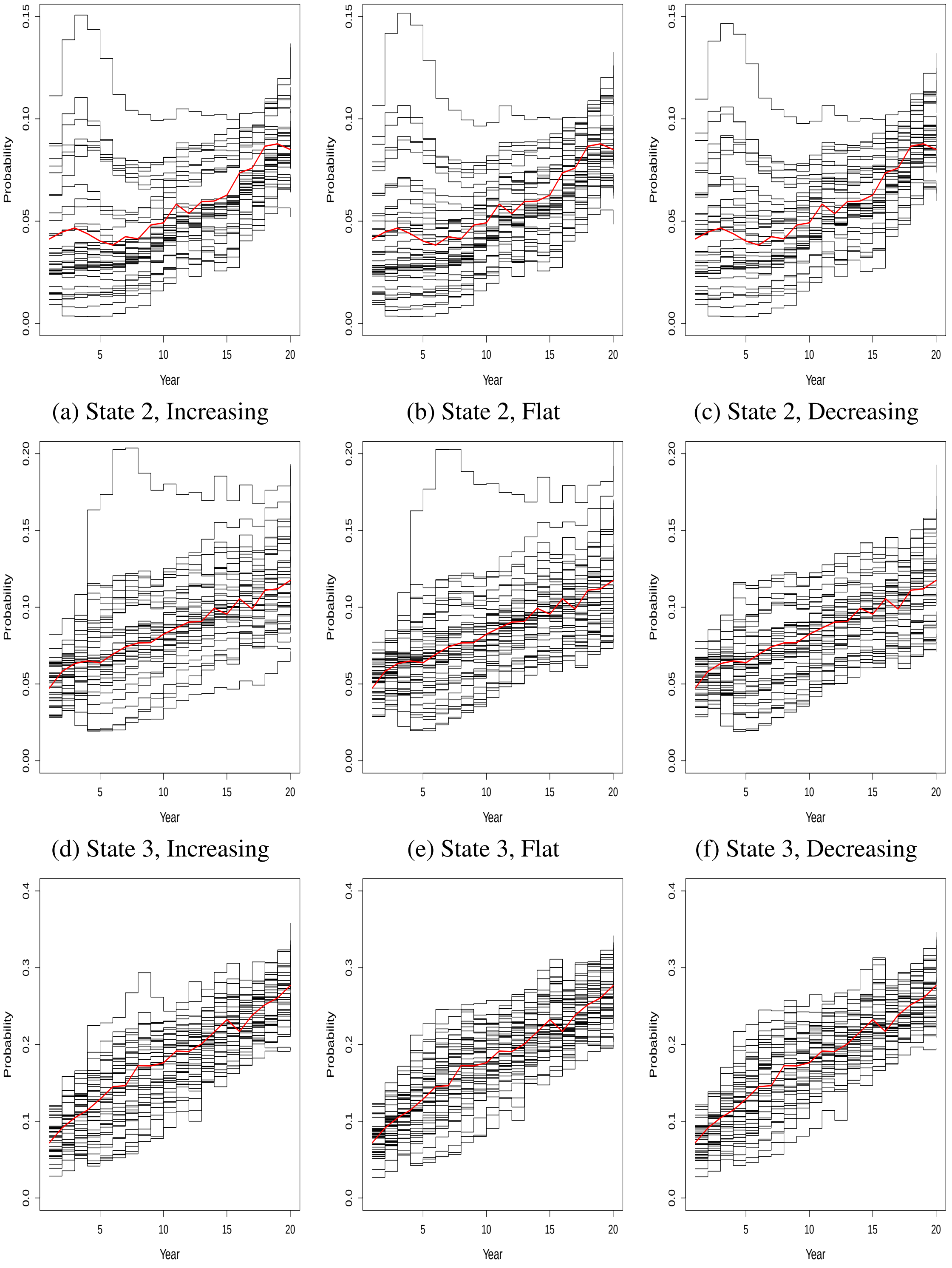
(g) State 4, Increasing

(h) State 4, Flat

(i) State 4, Decreasing

Fig. G.7 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

GP incidence model



(a) State 2, Increasing (b) State 2, Flat (c) State 2, Decreasing

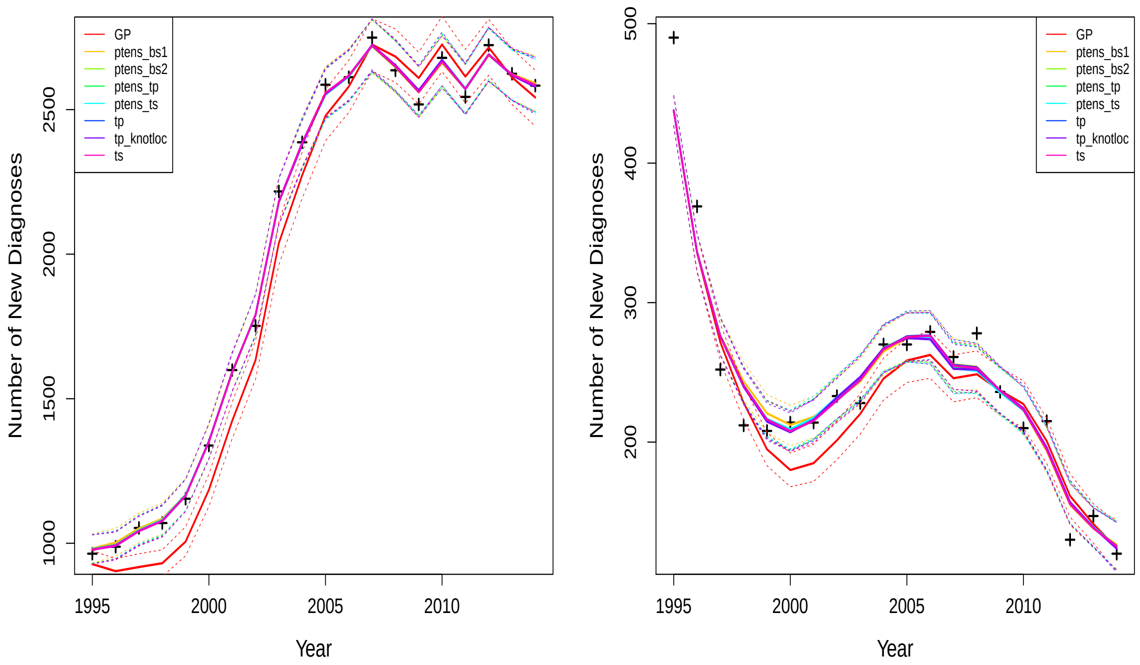
(d) State 3, Increasing (e) State 3, Flat (f) State 3, Decreasing

(g) State 4, Increasing (h) State 4, Flat (i) State 4, Decreasing

Fig. G.8 Estimated diagnosis probabilities from states 2, 3, and 4 under the three different incidence scenarios, as indicated by captions. Red and black lines depict the true diagnoses curves and the posterior means for each dataset, credible intervals are not depicted.

G.2.2 Goodness of fit plots

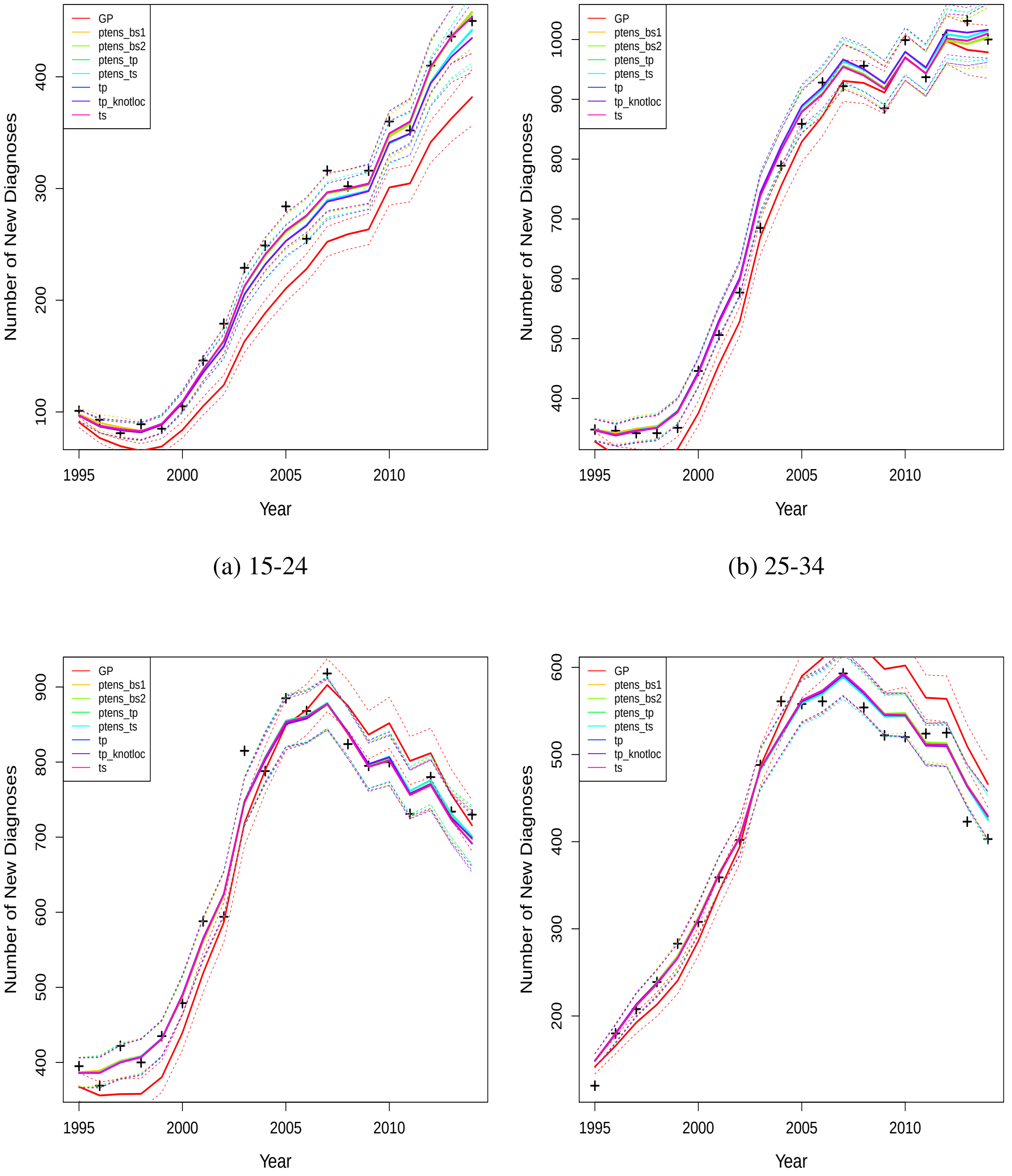
In this Section, goodness of fit plots are presented. Dataset 25 generated under the true flat incidence scenario is considered. This comprises three data sources: HIV, AIDS and CD4. For each data source, goodness of fit is considered at population level and is further stratified within the usual four age-classes: 15-24, 25-34, 35-44, 45+. All incidence models considered, with the exception of *GP* (red lines), provide very similar goodness of fit for all data-sources considered. Data are always extremely well fitted, both at population and at age-specific level. Only CD4-count data stratified by age-class are not particularly well fitted.



(a) HIV

(b) AIDS

Fig. G.9 Goodness of fit plots for HIV and AIDS diagnoses in data 25 (generated with flat incidence). The different colored lines and crosses respectively represent all incidence models and the simulated data points.



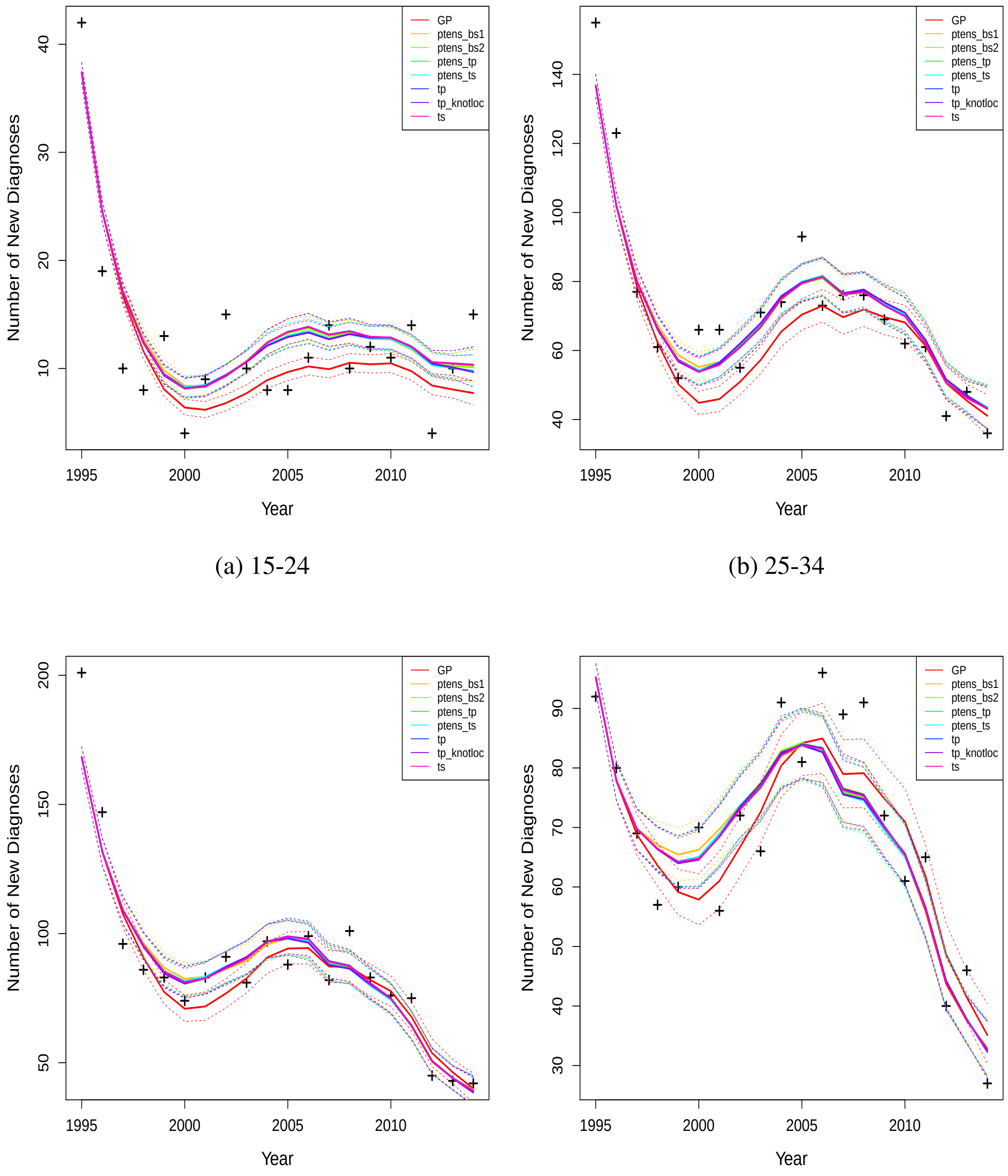
(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. G.10 Goodness of fit plots for HIV diagnoses, stratified by age-class as indicated in captions, for data 25 (generated with flat incidence). The different colored lines and crosses respectively represent all incidence models and the simulated data points.



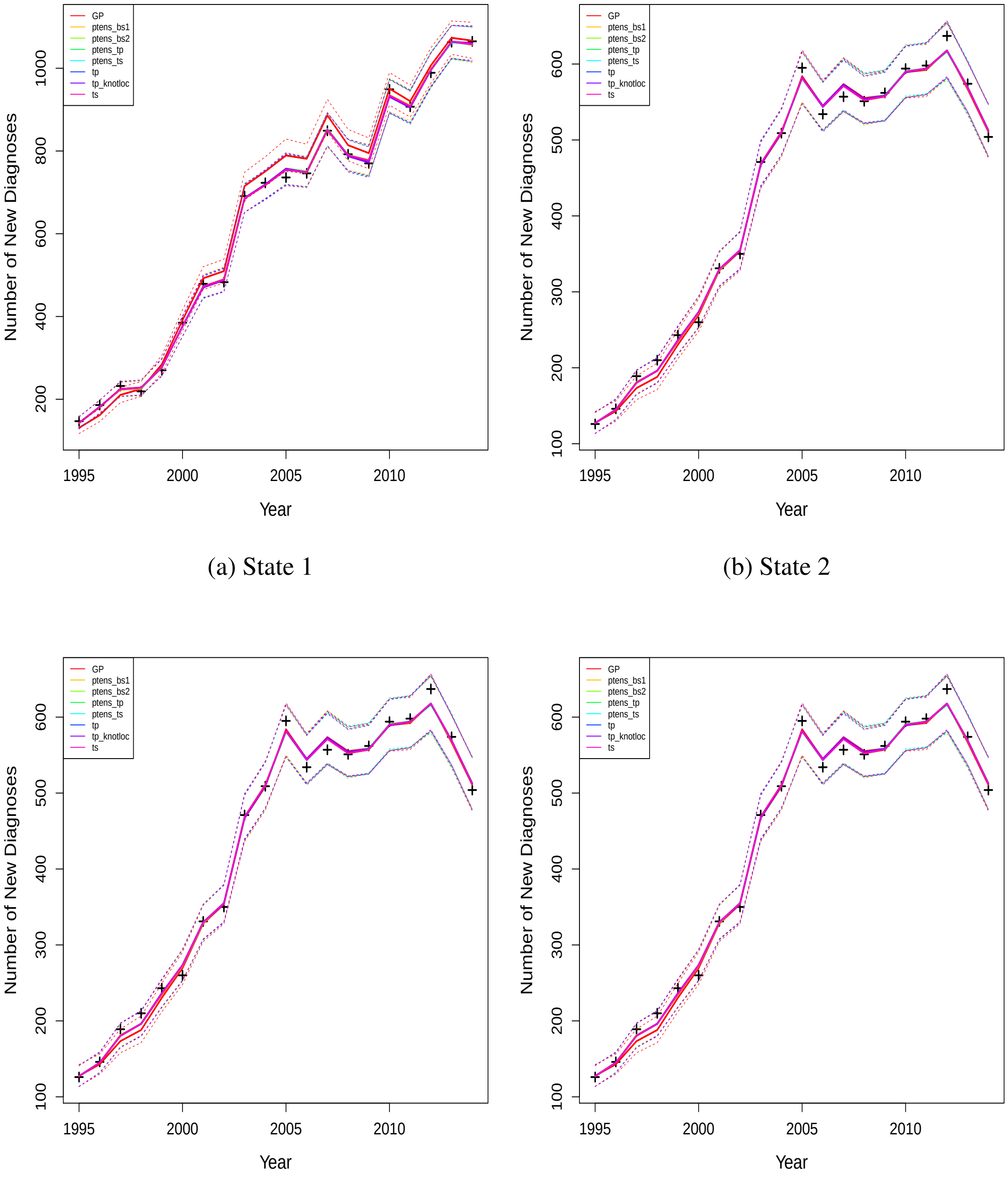
(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. G.11 Goodness of fit plots for AIDS diagnoses, stratified by age-class as indicated in captions, for data 25 (generated with flat incidence). The different colored lines and crosses respectively represent all incidence models and the simulated data points.



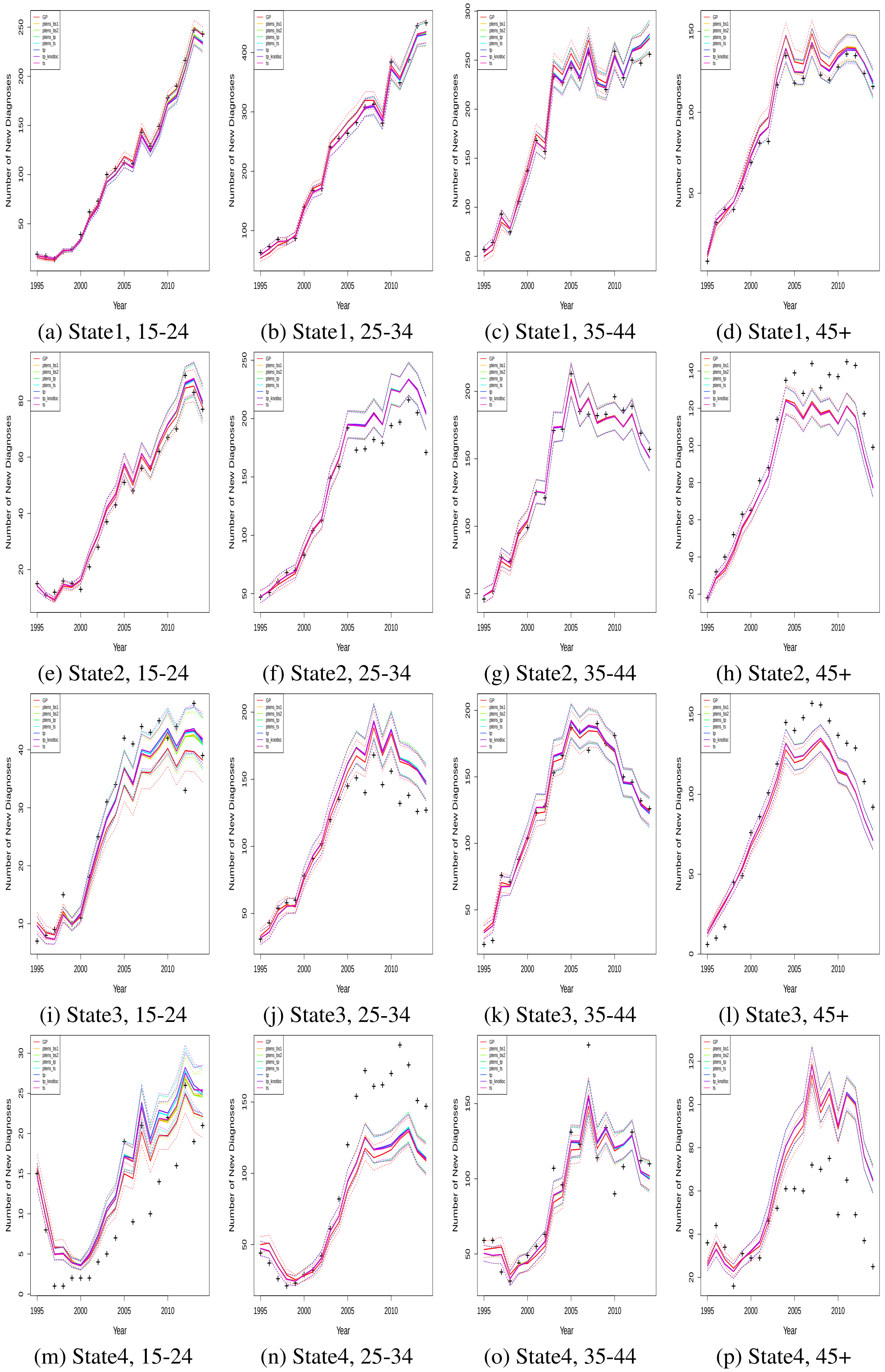
(a) State 1

(b) State 2

(c) State 3

(d) State 4

Fig. G.12 Goodness of fit plots for CD4 diagnoses, stratified by state as indicated in captions, for data 25 (generated with flat incidence). The different colored lines and crosses respectively represent all incidence models and the simulated data points.



(a) State1, 15-24 (b) State1, 25-34 (c) State1, 35-44 (d) State1, 45+

(e) State2, 15-24 (f) State2, 25-34 (g) State2, 35-44 (h) State2, 45+

(i) State3, 15-24 (j) State3, 25-34 (k) State3, 35-44 (l) State3, 45+

(m) State4, 15-24 (n) State4, 25-34 (o) State4, 35-44 (p) State4, 45+

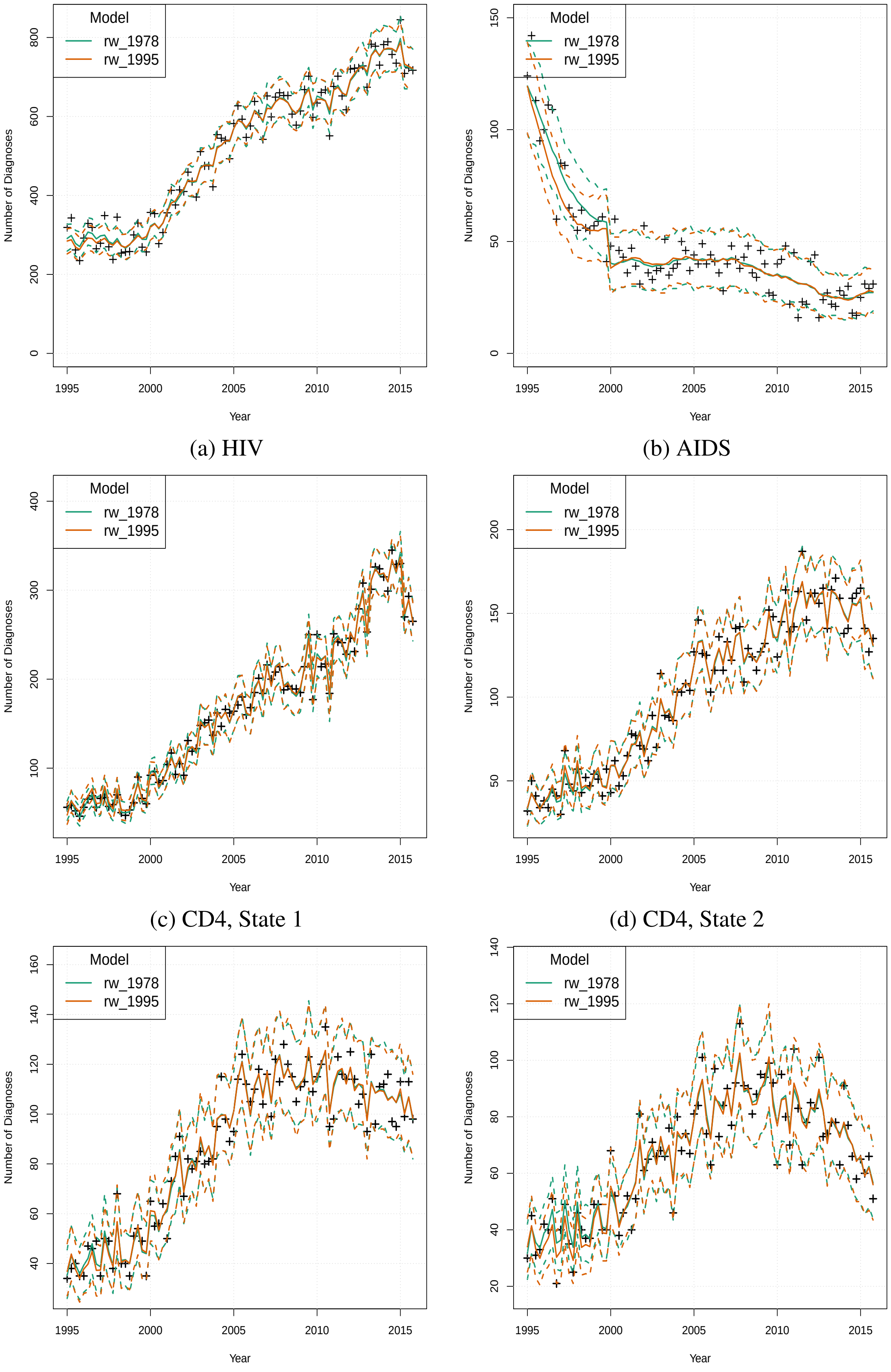
Fig. G.13 Goodness of fit for age and state specific CD4 diagnoses (as indicated in legends) for dataset 25 (generated with flat incidence). The different colored lines and crosses respectively represent all incidence models and the simulated data points.

Appendix H

Appendix for Chapter 8

H.1 Further details for Section 8.3.3

In this Section goodness of fit plots for the *1978-model* and *1995-model*, discussed in Section Section 8.3.3, are depicted. The posterior-predictive distribution, with 95% credible intervals, is considered. Goodness of fit is very similar for both models, for all data considered.



(a) HIV

(b) AIDS

(c) CD4, State 1

(d) CD4, State 2

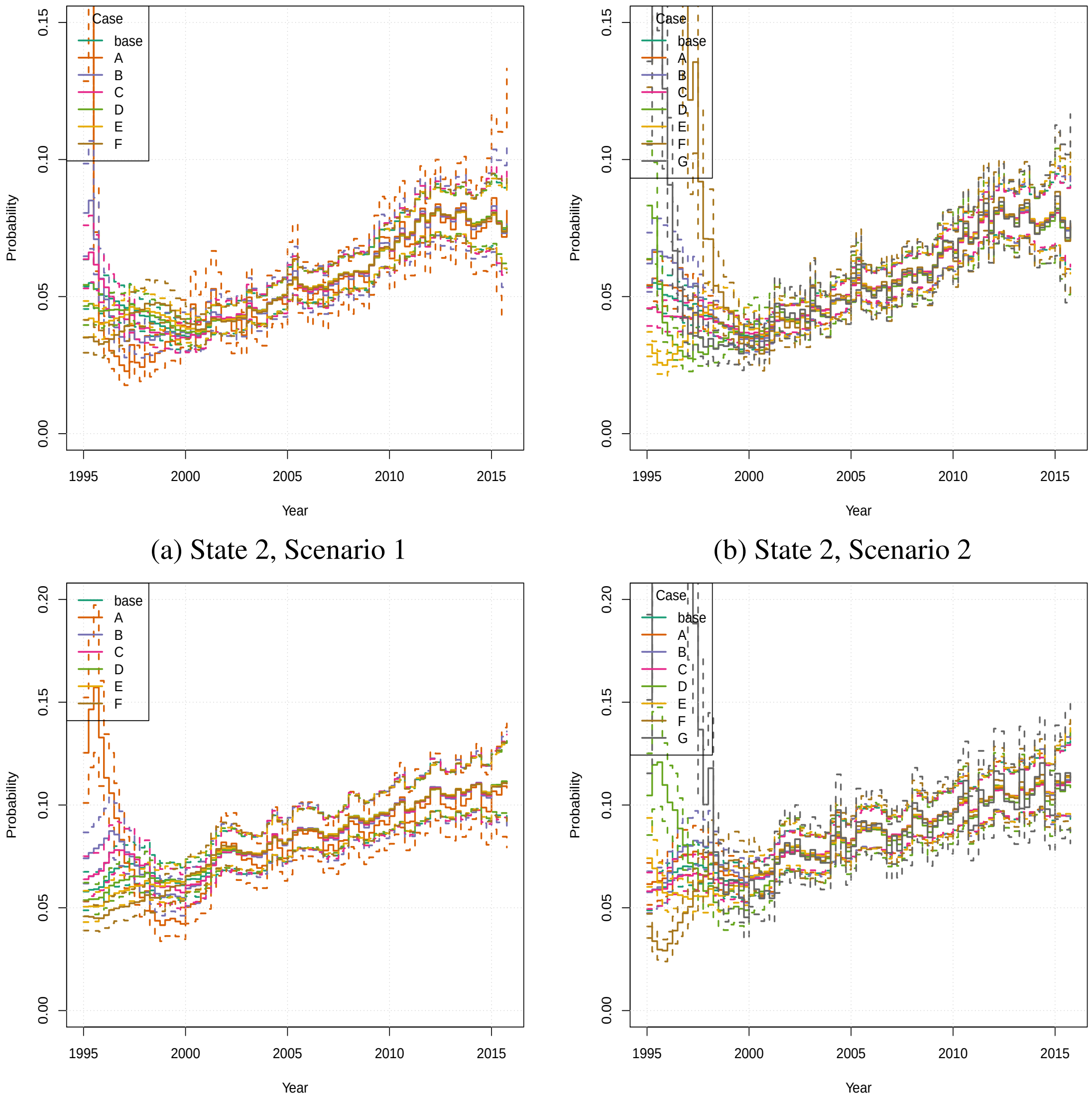
(e) CD4, State 3

(f) CD4, State 4

Fig. H.1 Goodness of fit for the age-independent back-calculation model, as indicated by the sub-captions.

H.2 Further details for Section 8.3.4

Below the plots of the estimated diagnosis probabilities from undiagnosed states 2 and 3, for both scenarios, are depicted. Solid lines represent the means of the posterior distribution and dashed lines represent 95% credible intervals.



(a) State 2, Scenario 1

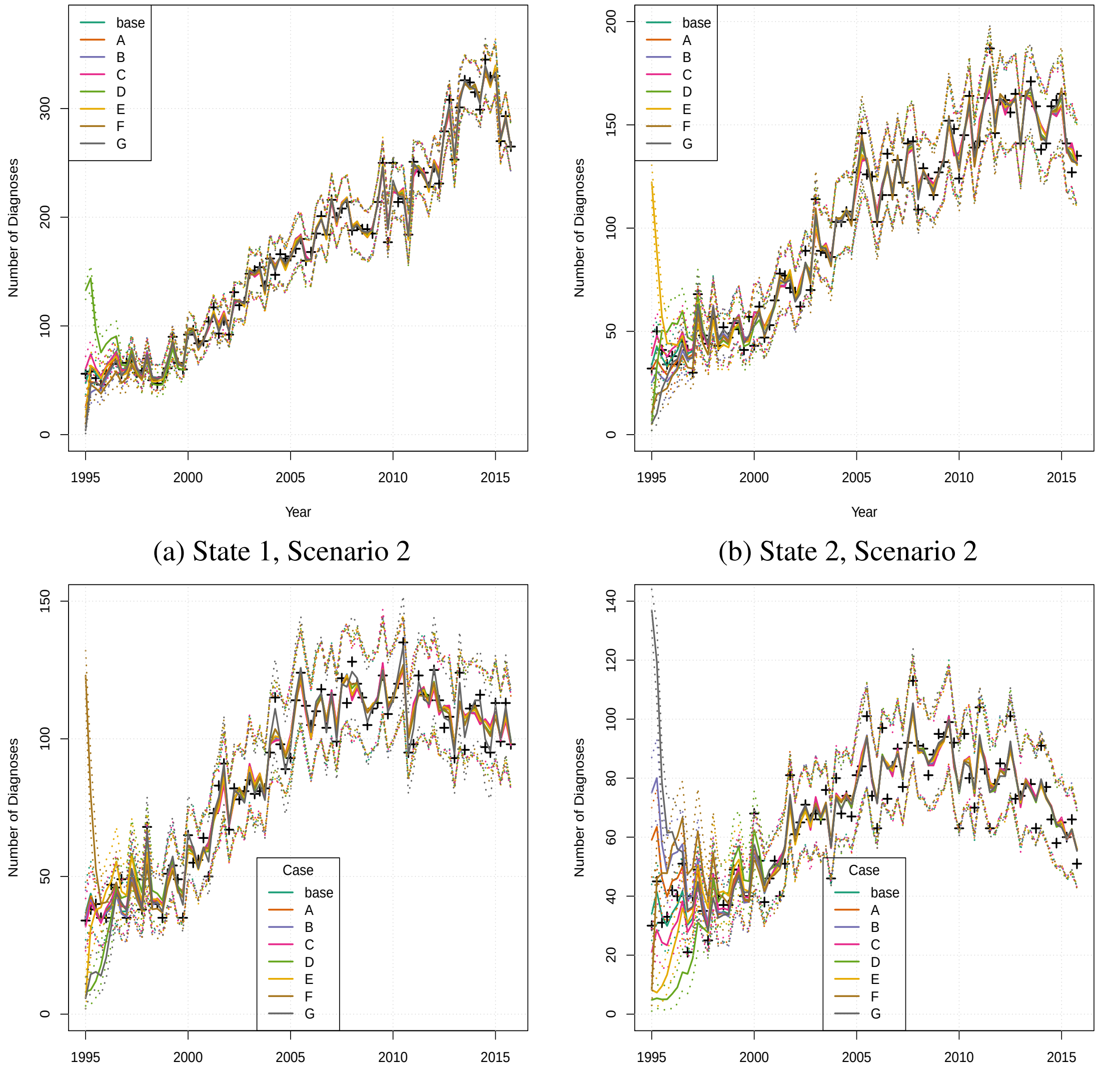
(b) State 2, Scenario 2

(c) State 3, Scenario 1

(d) State 3, Scenario 2

Fig. H.2 Diagnosis probabilities, by state and scenario.

Below CD4-count data goodness of fit for scenario 2 is plotted, for the four different latent stages considered, can be found. The posterior-predictive distribution for the data is depicted: solid lines show posterior means and dashed lines 95% credible intervals.



(a) State 1, Scenario 2

(b) State 2, Scenario 2

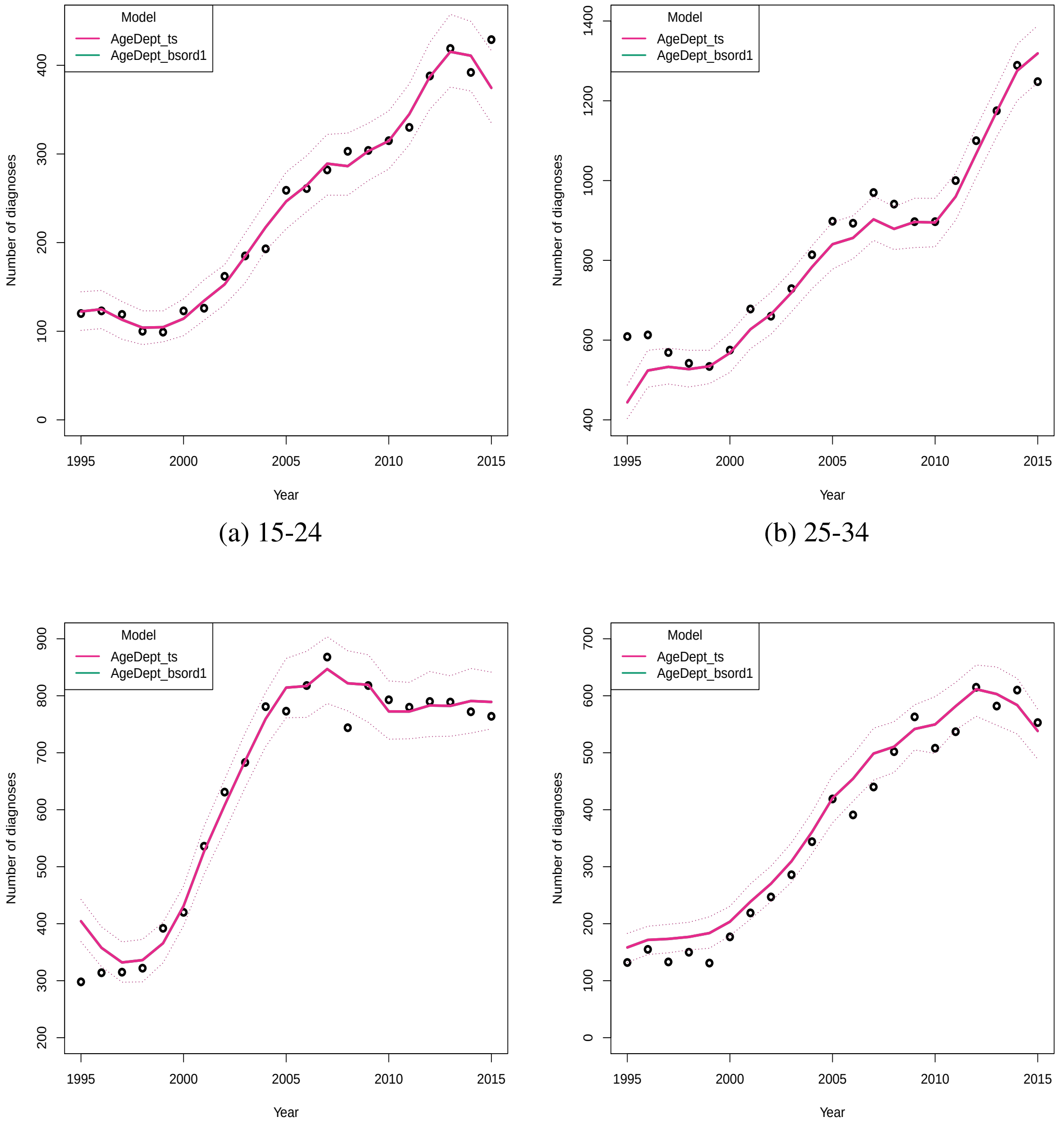
(c) State 3, Scenario 2

(d) State 4, Scenario 2

Fig. H.3 CD4 diagnoses goodness of fit, by state and scenario.

H.3 Further details for Section 8.4.1

Below age-specific goodness of fit plots for age-specific back-calculation models are reported, using the *ts* and *ptensbsord1* incidence models. The posterior-predictive distribution for the data is depicted: solid lines show posterior means and dashed lines 95% credible intervals.



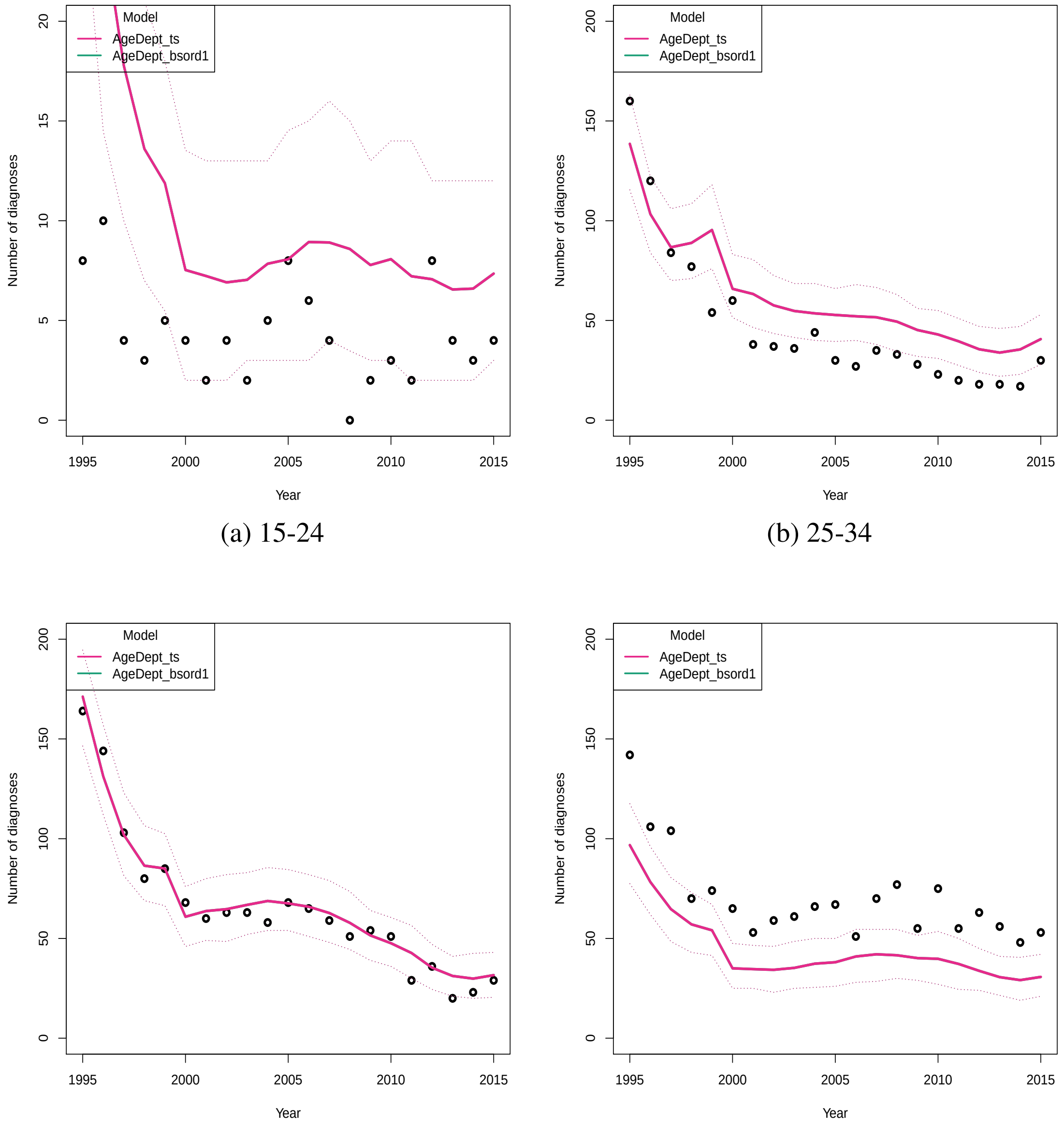
(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. H.4 HIV diagnoses goodness of fit, stratified by age-classes.



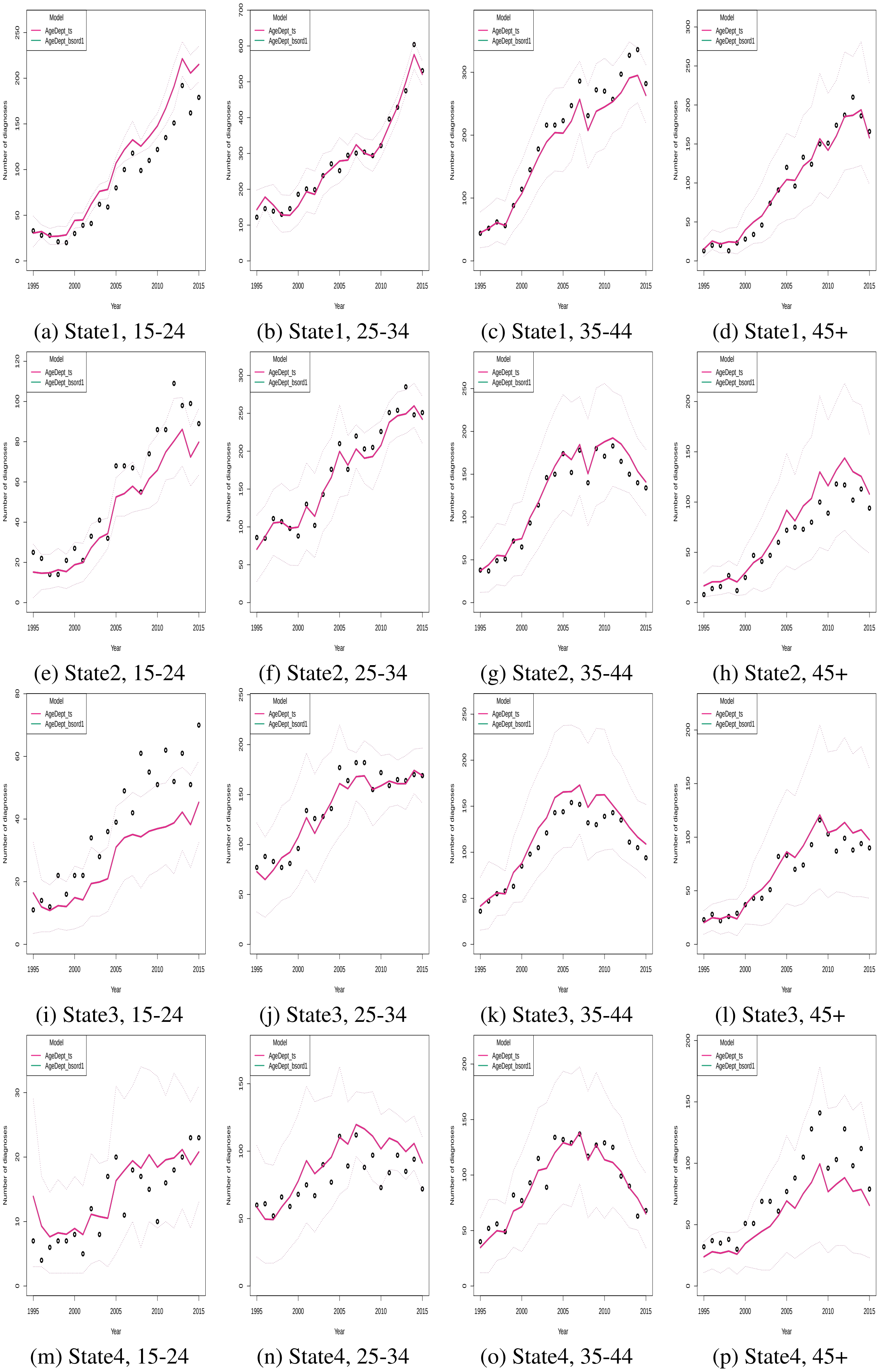
(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. H.5 AIDS diagnoses goodness of fit, stratified by age-classes.



(a) State1, 15-24 (b) State1, 25-34 (c) State1, 35-44 (d) State1, 45+

(e) State2, 15-24 (f) State2, 25-34 (g) State2, 35-44 (h) State2, 45+

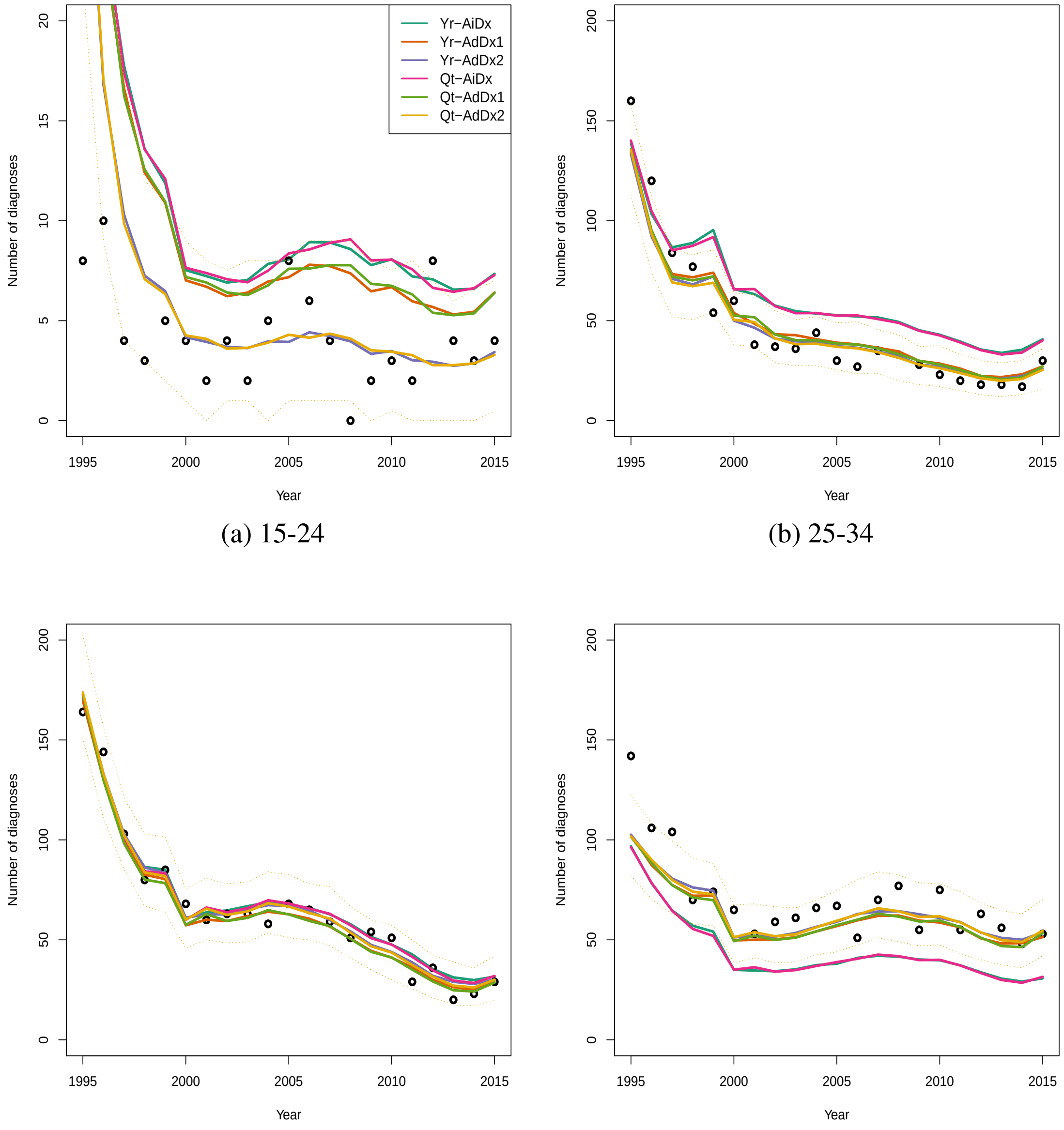
(i) State3, 15-24 (j) State3, 25-34 (k) State3, 35-44 (l) State3, 45+

(m) State4, 15-24 (n) State4, 25-34 (o) State4, 35-44 (p) State4, 45+

Fig. H.6 Goodness of fit for age and state specific CD4 diagnoses. The pink and green lines represent the *ts* and *ptensbsord1* models and the points the observed data.

H.4 Further details for Section 8.4.2

Goodness of fit plots for the models considered in Section 8.4.2 are displayed. For goodness of fit to be comparable, the yearly (rather than quarterly) data posterior-predictive distribution for the quarterly models are displayed. Solid lines show posterior means and dashed lines 95% credible intervals. The latter have only been reported for the *Qt-AdDx2* Only age-specific goodness of fit plots for AIDS and CD4-count data are shown, as for HIV data all models produce the same fit.



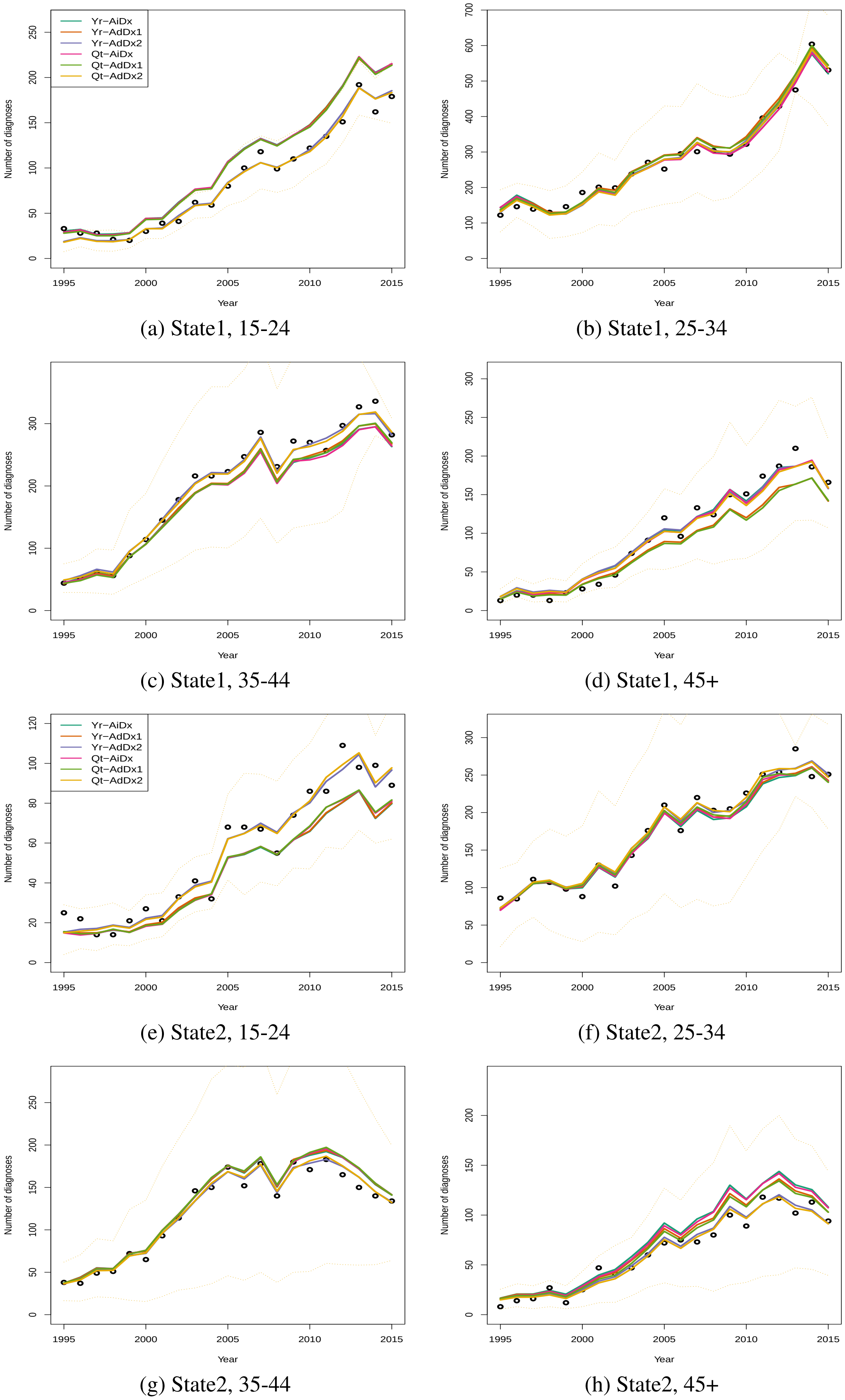
(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. H.7 AIDS diagnoses goodness of fit, stratified by age-class.



(a) State1, 15-24

(b) State1, 25-34

(c) State1, 35-44

(d) State1, 45+

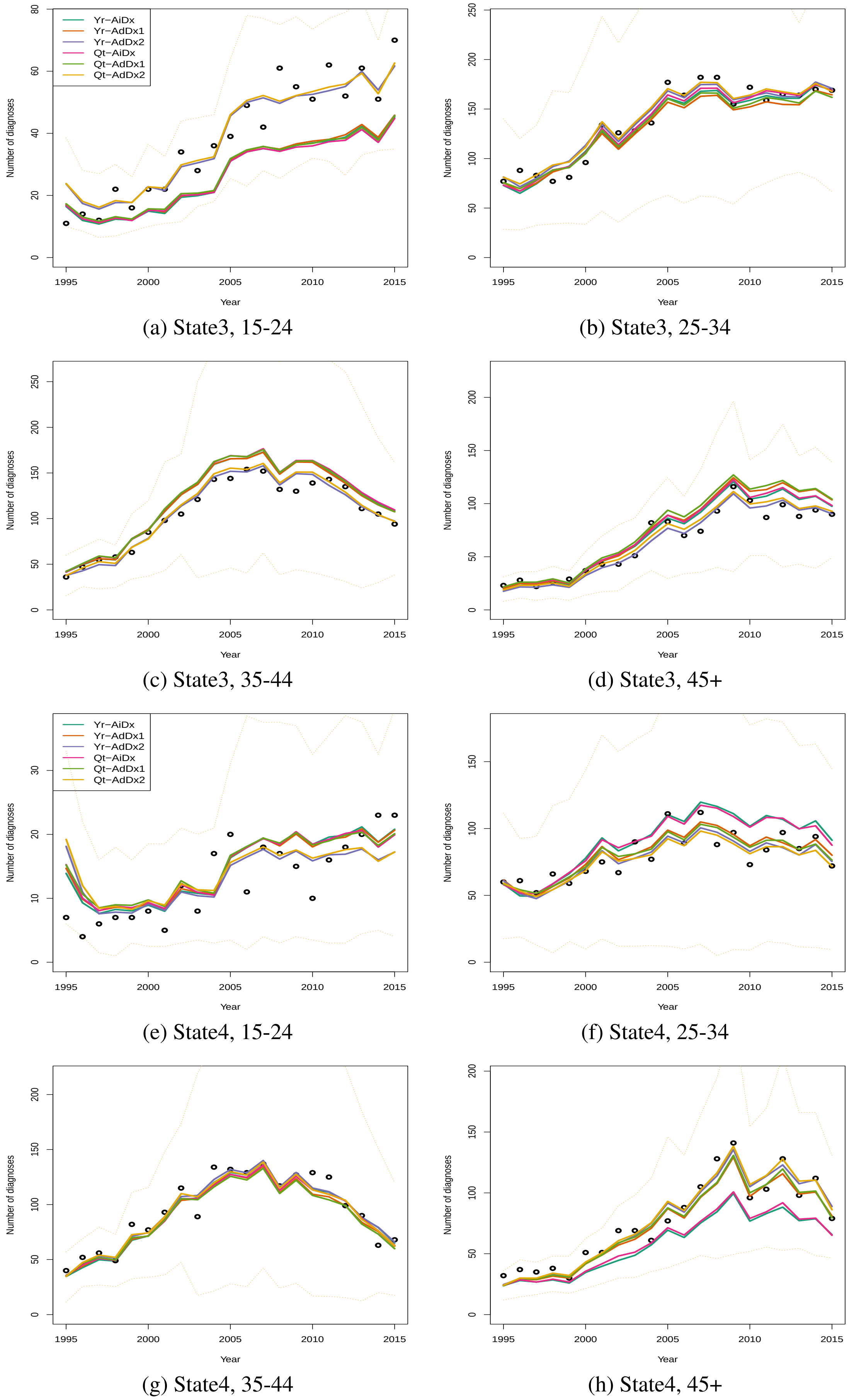
(e) State2, 15-24

(f) State2, 25-34

(g) State2, 35-44

(h) State2, 45+

Fig. H.8 CD4 diagnoses goodness of fit, stratified by age-class, states 1 and 2.



(a) State3, 15-24

(b) State3, 25-34

(c) State3, 35-44

(d) State3, 45+

(e) State4, 15-24

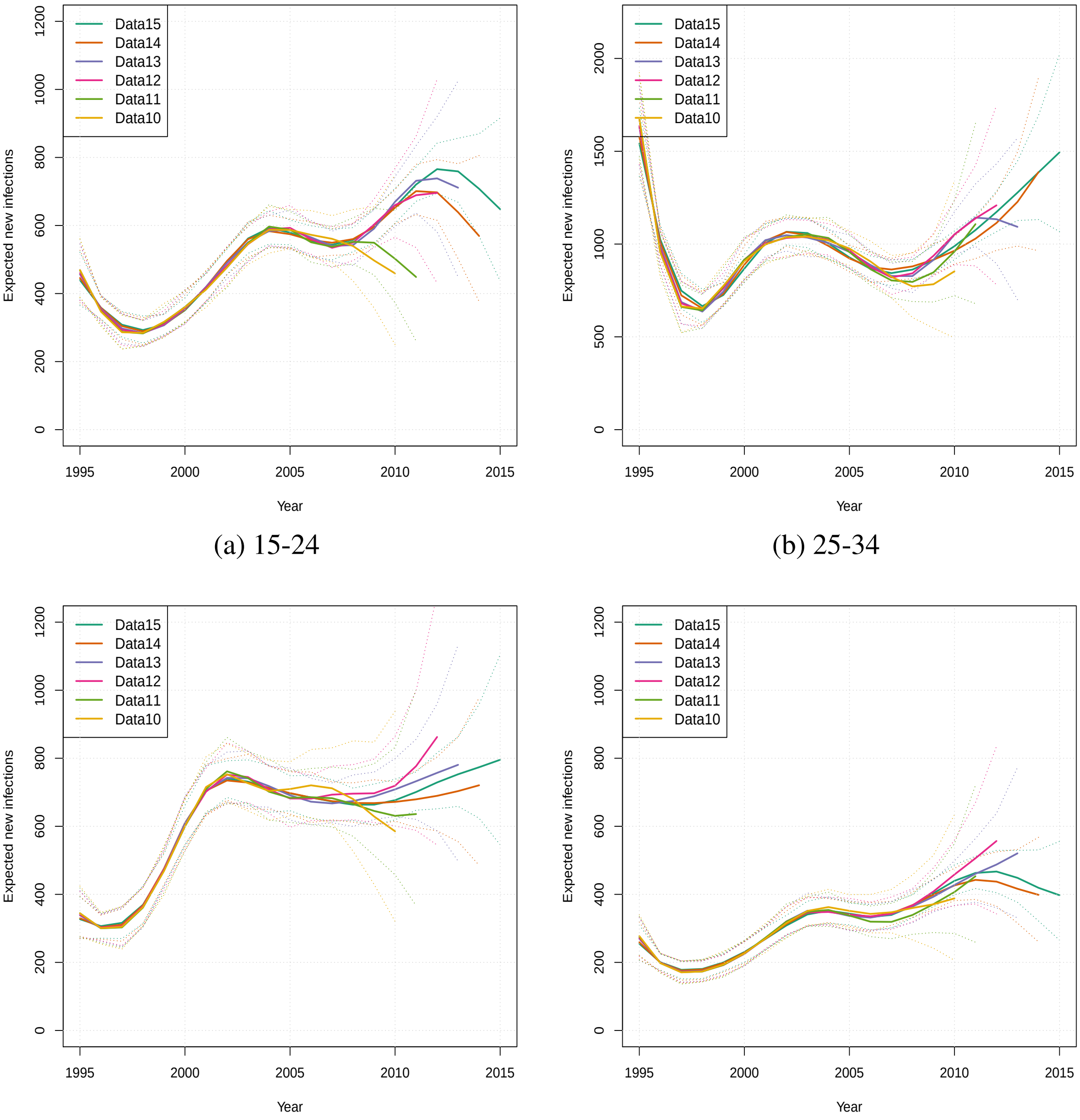
(f) State4, 25-34

(g) State4, 35-44

(h) State4, 45+

Fig. H.9 CD4 diagnoses goodness of fit, stratified by age-class, states 3 and 4.

The plots below show that also age-specific incidence estimates from the *Qt-AdDx2* model are robust to the sequential addition of further years of data.



(a) 15-24

(b) 25-34

(c) 35-44

(d) 45+

Fig. H.10 Sensitivity of the estimated time profile of incidence to the sequential addition of years of data, for *Qt-AdDx2* stratified by age-class.