Inferring the time-varying transmission rate and effective reproduction number by fitting semi-mechanistic compartmental models to incidence data

Inferring the time-varying transmission rate and effective reproduction number by fitting semi-mechanistic compartmental models to incidence data

By Greg Forkutza B.Sc. M.Sc.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Science

"What is your aim in Philosophy?

To show the fly the way out of the fly-bottle."

—Ludwig Wittgenstein

McMaster University
Master of Science (2024)
Hamilton, Ontario (Mathematics & Statistics)

TITLE: Inferring the time-varying transmission rate and effective reproduction number by fitting semi-mechanistic compartmental models to incidence data
AUTHOR: Greg Forkutza (McMaster University)
SUPERVISOR: Ben Bolker
NUMBER OF PAGES: x, 75

Lay Abstract

This thesis explores a new way to model how diseases spread using a deterministic mathematical framework. We focus on estimating the changing transmission rate and the effective reproduction number, key factors in understanding and controlling disease outbreaks. Our method, incorporated into the `macpan2` software, uses advanced techniques to smoothly estimate these changing rates over time. We first prove the effectiveness of our approach with simulations and then apply it to real data from Scarlet Fever, COVID-19, and Measles. We also compare model performance using statistical criteria and different smoothing methods. Our results show that this flexible and user-friendly approach is a valuable tool for modelers working on disease dynamics.

Abstract

This thesis presents a novel approach to ecological dynamic modeling using non-stochastic compartmental models. Estimating the transmission rate (β) and the effective reproduction number (R_t) is essential for understanding disease spread and guiding public health interventions. We extend this method to infectious disease models, where the transmission rate varies dynamically due to external factors. Using Simon Wood's partially specified modeling framework, we introduce penalized smoothing to estimate time-varying latent variables within the `R` package `macpan2`. This integration provides an accessible tool for complex estimations. The efficacy of our approach is first validated via a simulation study and then demonstrated with real-world datasets on Scarlet Fever, COVID-19, and Measles. Additionally, we infer the effective reproduction number (R_t) using the estimated β values, providing further insights into the dynamics of disease transmission. Model fit is compared using Akaike Information Criterion (AIC), and we evaluate the performance of different smoothing bases derived using the `mgcv` package. Our findings indicate that this methodology can be extended to various ecological and epidemiological contexts, offering a versatile and robust approach to parameter estimation in dynamic models.

Acknowledgements

Acknowledgements go here. TBD

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Greg Forkutza, declare that this thesis titled, *Inferring the time-varying transmission rate and effective reproduction number by fitting semi-mechanistic compartmental models to incidence data* and the work presented in it are my own. I confirm that:

I did most of the research.

Also the writting.

Sometimes I pulled my hair out.

But mostly I had fun.

Chapter 1

Introduction

Ecological dynamic models describe how ecological processes drive populations to change over time. These models often account for both process error and observation error. State-space models allow for the estimation of parameters of the deterministic process, observation error, and process error from a single time series [1]. However, estimating parameters for state-space models is challenging. One simplification is to assume that the dynamic model only has observation error. This can be done by starting with the initial conditions of the system and computing the entire trajectory over the domain at once. This approach assumes there is no uncertainty in the predicted values of the states at each time step, meaning there is no process error. Since the trajectory at every time step is determined solely by the starting parameters and initial conditions, the model is deterministic. The only error accounted for is the difference between observed and predicted values.

To fit the model to data and compute the maximum likelihood estimate (MLE) for a given set of parameters, one assumes independent observations and sums the likelihood for each observation based on the chosen model of observation error. Using a quasi-Newton method, parameter estimates are updated based on the current iteration of trajectory matching until convergence is achieved. If the observation error is normally distributed with constant variance, this process simplifies to least squares fitting.

When there are unobserved variables in the model, a simple approach is to treat these variables as fixed parameters. In modeling infectious diseases, for example, one may be interested in estimating the unobserved transmission rate between an infectious individual and a susceptible one. Assuming a fixed value for the transmission rate at every time step is often inaccurate because many diseases have dynamic transmission rates influenced by factors such as seasonality or non-pharmaceutical interventions (e.g., social distancing, masking, and changes in mobility patterns). Therefore, it is reasonable to assume that the transmission rate is a time-varying parameter, computed at each observed data point.

A standard method for dealing with time-varying parameters is to assume a stochastic model of the latent variable, introducing a parametric model and estimating its parameters as part of the MLE process. However, imposing a particular shape on the unknown function describing the latent variable, based on a specific statistical distribution, is

often a phenomenological characterization rather than one derived from known biological mechanisms. This can lead to mismatches between the model and data not due to biological inaccuracies but due to the chosen parametric form of the latent variable.

Fitting non-stochastic dynamic ecological models with time-varying latent variables is challenging. Instead of specifying the functional form of the unknown function a priori, we can use a method that allows for more flexible function estimation, minimizing incidental assumptions and model misspecification. Simon Wood [2] describes a methodology using penalized smoothing to estimate time-varying latent variables for non-stochastic dynamic ecological models. He introduces what he calls the partially specified modeling framework, also known as semi-mechanistic models. The basic idea is to use deterministic specifications and parametric models for components backed by known biological mechanisms while employing non-parametric methods for flexible function estimation when insufficient information is available to justify a structured parametric form.

Implementing Wood's methodology requires expertise in statistical inference, nonlinear optimization, spline literature, numerical analysis, and statistical computing, making it complex for many domain-specific modelers. Our goal was to develop a user-friendly way for modelers to estimate time-varying latent variables in non-stochastic compartmental models. We integrated Simon Wood's methodology for smoothing parameter estimation into the general-purpose compartmental modeling tool `macpan2`, which utilizes `TMB` for optimization. We use the `mgcv` package in R to obtain low-rank smoothing matrices for the basis and penalty, resulting in a software module within `macpan2` that allows users to easily formulate and fit semi-mechanistic models.

Being able to quickly and easily estimate the transmission rate at each observation is crucial for infectious disease modelers. The transmission rate is integral to calculating other epidemiological quantities, such as the effective reproduction number (R_t), and provides the basis for many downstream applications, including evaluating intervention strategies, forecasting disease dynamics, identifying high-risk periods and populations, validating models, understanding transmission dynamics, assessing the impact of variants, and developing and testing hypotheses about factors influencing disease spread.

In Chapter 2, we review the basic theory of univariate smoothing in the context of Gaussian regression and fitting compartmental models to data. We discuss the construction of linear smoothers, review different bases available in the R package `mgcv`, explore general penalized likelihood methods, and examine the relationship between smooths and random effects, including a Bayesian perspective. Chapter 3 covers the technical details required to implement the smoothing parameter estimation methodology in `macpan2`. Chapter 4 presents the results of the simulation study and applications to Scarlet Fever, COVID-19, and Measles datasets. Finally, Chapter 5 discusses the results, the assumptions and simplifications used in our modeling methodology, and future research avenues.

Chapter 2

Background

This chapter contains two sections. The first section outlines the theory required to understand how to use univariate Gaussian regression smoothers in statistical modeling. Most importantly, this section culminates in describing the relationship between Gaussian random effects and smoothing splines within a Bayesian framework. Understanding this relationship is crucial for fitting semi-mechanistic models to data.

The second section provides a brief introduction to the basic concepts of using compartmental models in the statistical modeling of infectious diseases.

2.1 Smoothers

When using linear model methods, there's an assumption that the true function $f(\mathbf{X})$ is linear in $\mathbf{X}$. To extend beyond this linearity, one approach is to employ a linear transformation of $\mathbf{X}$. This leads to:

$$f(\mathbf{X}) = \sum_{m=1}^{M} \beta_m h_m(\mathbf{X}),$$

which is a linear basis expansion of $\mathbf{X}$, where $h_m(\mathbf{X}) : \mathbb{R}^p \to \mathbb{R}$ represents the m-th transformation of $\mathbf{X}$.

A notable class of these transformations is restriction methods, where the class of functions that $f(\mathbf{X})$ can assume is limited. A common example within this class is *splines*, which define m basis functions $\beta_m h_m(\mathbf{X})$ as local polynomial representations. The domain is divided into contiguous intervals, each represented by a separate polynomial function. The boundaries of these intervals are known as *knots*. An order-M spline with knots ξ_j, $j = 1, \ldots, K$, is a piecewise-polynomial of order M and has continuous derivatives up to order $M - 2$.

When the location, derivative order, and number of knots are predetermined, the technique is referred to as regression splines. Natural cubic splines are a specific type where the spline function is composed of cubic polynomial segments and it is linear

beyond the outermost knots. This spline is represented by K basis functions $\beta_m h_m(\mathbf{X})$, one for each specified knot.

The complexity of the fit can be adjusted by incorporating regularization to manage the trade-off between data fidelity and smoothness of the curve fit. This is achieved by minimizing a residual sum of squares with an additional penalty term. The penalty term includes a parameter that controls this trade-off, allowing the fit to range between the extremes of pure interpolation and a linear least squares fit. Splines used in this context are known as smoothing splines.

For a higher-level overview of this topic, see [3]. For a more detailed exposition, refer to [4]. Additionally, a comprehensive mathematical treatment of splines and their applications can be found in [5].

2.1.1 Natural cubic splines

Suppose we have a dataset consisting of points (x_i, y_i) for $i = 1, \ldots, n$, with each x_i strictly less than x_{i+1}. A *natural cubic spline* $g(x)$ is defined as a function that smoothly interpolates between these points. This spline is constructed from cubic polynomial segments, where each segment corresponds to an interval $[x_i, x_{i+1}]$. These polynomial segments are connected such that $g(x)$, $g'(x)$, and $g''(x)$ are all continuous across the entire domain.

Furthermore, the spline satisfies the interpolation condition $g(x_i) = y_i$ for each point, with the additional constraint that the second derivatives at the endpoints of the domain, x_1 and x_n, are set to zero. This constraint ensures that the spline is linear beyond the boundary points, contributing to its 'natural' behavior.

Among all functions that are continuous over $[x_1, x_n]$, possess absolutely continuous first derivatives, and interpolate the given data points (x_i, y_i), the natural cubic spline $g(x)$ is uniquely the smoothest in terms of minimizing the integral:

$$J(f) = \int_{x_1}^{x_2} f''(x)^2 \, dx, \tag{2.1}$$

which quantifies the overall curvature of the function. Minimizing this integral promotes a function with less curvature and smoother transitions between the interpolated points.

A detailed proof and further discussion on this property of natural cubic splines can be found in [6], which explores the mathematical underpinnings and optimizations that define these splines.

2.1.2 Cubic smoothing splines

With a natural cubic spline, $g(x)$, we have two main options: we can interpolate the data by setting each $g(x_i) = y_i$, or we can smooth the data by treating the values $g(x_i)$

as variables to be optimized, as in *cubic smoothing splines*. Smoothing is achieved by minimizing the following objective function:

$$J_2(f) = \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(x)^2\, dx, \tag{2.2}$$

Here, λ serves as a tuning parameter that balances the fidelity to the data with the smoothness of the function g. A higher λ value places greater emphasis on minimizing the integral of the squared second derivative, which encourages a smoother curve for g. Conversely, a lower λ value focuses on closely matching the actual data points, minimizing the sum of squared differences $\sum_{i=1}^{n}(y_i - g(x_i))^2$.

The formulation given in Equation 2.2 is flexible and does not depend on a predefined set of basis functions. Instead, the model itself dictates the structure, leading to the derivation of optimal basis functions based on the specified terms for data fidelity and smoothness.

Solving Equation 2.2 entails tackling a variational problem where the basis functions for the cubic spline are derived using the Euler-Lagrange equation. While a comprehensive derivation of this process is provided in [7], in appendix A.1 we briefly outline some key aspects of the proof to understand why these basis functions take their particular form. We present this outline for two primary reasons: historically, cubic smoothing splines were among the first types of smoothers to be thoroughly investigated. Furthermore, the general approach of penalized likelihood methods can be effectively abstracted from the solution to the minimization problem presented by cubic smoothing splines. For a broader definition of penalized splines, refer to subsection 2.1.8

2.1.3 Cubic regression splines

Using the results from appendix A.1, we can define the cubic spline function $f(x)$ with k knots $x_1, \ldots, x_k$ as follows:

$$f(x) = a_j(x)f(x_j) + b_j(x_j)f(x_{j+1}) + c_j(x)f''(x_j) + d_j(x_j)f''(x_{j+1}) \quad \text{if} \quad x_j \leq x \leq x_{j+1}. \tag{2.3}$$

This formulation ensures continuity at each knot, as the continuity conditions A.5 imply that the derivatives at the knots must match:

$$\mathbf{B} \cdot (f''(x_2), \ldots, f''(x_{k-1}))^T = \mathbf{D} \cdot (f(x_2), \ldots, f(x_{k-1}))^T$$

By integrating these conditions into Equation 2.3, and redefining $f(x)$ in terms of the basis functions, we arrive at:

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i, \tag{2.4}$$

where $\beta_i = f(x_i)$ and b_i represents the transformed basis functions obtained by applying $\mathbf{B^{-1}D}$ to the original basis functions A.1. This defines a *cubic regression spline*. The full details of how to derive equation 2.4 can be found in [8].

This structure effectively maps the spline basis to the spline evaluated at a specified set of knots. Furthermore, a computationally efficient form of Equation 2.1 in terms of these basis functions and matrix elements is:

$$\int_{x_1}^{x_k} f''(x)^2\, dx = \beta^T \mathbf{D}^T \mathbf{B}^{-1} \mathbf{D} \beta = \beta^T \mathbf{S} \beta,$$

where $\mathbf{S} \equiv \mathbf{D}^T \mathbf{B}^{-1} \mathbf{D}$ is the called the *penalty matrix* for this basis.

In subsection 2.1.8, we show how the method of penalizing cubic splines lead to a more general method of penalized regression and the abstract notion of penalized likelihood methods.

In the following sections we investigate some other penalized regression splines that are are used in chapter 4 as smoothers (available as basis and penalty matrices in the R package `mgcv`) for estimating time-varying parameters in discrete time deterministic compartmental models.

2.1.4 B-splines

B-splines offer another basis for representing polynomials that are used in constructing smoothing splines. Detailed explanations of B-spline basis functions and their mathematical properties can be found in [3] and [5]. For a concise definition of B-spline basis functions, refer to [4].

B-splines allow an $(m+1)^{\text{th}}$ order spline to be expressed uniquely using the B-spline basis functions $B_i^m(x)$:

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i,$$

where each $B_i^m(x)$ is defined by a local combination of knots. The means any curve can be uniquely represented as a linear combination of B-splines and hence as a linear smoother.

2.1.5 P-splines

P-splines are a low rank (the number of knots is less than the length of the input vector) smoother that adapt a B-spline basis with a penalty to control wiggliness by penalizing adjacent $\beta_i = f(x_i)$ as

$$\mathbf{P} = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2.$$

Let matrix $\mathbf{P}$ be defined as follows:

$$\mathbf{R} = \begin{bmatrix} -1 & 1 & 0 & \cdots & \\ 0 & -1 & 1 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \end{bmatrix}$$

so that

$$\begin{bmatrix} \beta_2 - \beta_1 \\ \beta_3 - \beta_2 \\ \vdots \end{bmatrix} = \mathbf{R}\beta.$$

Then you can write the penalty term for adjacent basis coefficients as:

$$\mathbf{P} = \beta \mathbf{R}^T \mathbf{R} \beta = \beta \begin{bmatrix} 1 & -1 & 0 & \cdots & \\ -1 & 2 & -1 & \cdots & \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & & \ddots & \ddots & \end{bmatrix} \beta.$$

2.1.6 Thin plate regression splines

So far, the discussion has focused on smoothing methods that apply to a single predictor variable. While univariate smoothing is useful, it's worth noting that there are multivariate smoothing techniques available as well. One such technique is the thin plate splines (TPS). This method can be used even with just one predictor variable. We follow the method given by Simon Wood in [9].

In the general case, a *thin plate spline* (TPS) is defined for a number of predictor variables d and a penalty of degree m such that $2m > d$. The TPS penalty, J_{md}, is given by:

$$J_{md} = \int_{\mathbb{R}^d} \sum_{\nu_1 + \cdots + \nu_d = m} \frac{m!}{\nu_1! \cdots \nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \cdots \partial x_d^{\nu_d}} \right)^2 dx_1 \cdots dx_d,$$

where $\nu_1, \nu_2, \ldots, \nu_d$ are non-negative integers representing the orders of the partial derivatives with respect to each predictor variable x_i.

The function that minimizes this penalty is expressed as:

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(\|x - x_i\|) + \sum_{j=1}^{M} \alpha_j \phi_j(x),$$

where δ and α are vectors of coefficients to be estimated. The δ coefficients must satisfy the linear constraints $\mathbf{T}^T\delta = 0$, with $T_{ij} = \phi_j(x_i)$. Here, $M = \binom{m+d-1}{d}$, representing the number of linearly independent polynomials $\phi_i(x)$ that span the space of polynomials in $\mathbb{R}^d$ of degree less than m. These polynomials form the null space of J_{md}.

The remaining basis functions $\eta_{md}(r)$ are defined as:

$$\eta_{md}(r) = \begin{cases} (-1)^{m+1+d/2} \frac{2^{2m-1}\pi^{d/2}(m-1)!}{\Gamma(m-d/2)!} r^{2m-d} \log(r), & \text{if } d \text{ is even,} \\ \Gamma(d/2-m) \frac{2^{2m}\pi^{d/2}(m-1)!}{\Gamma(m)} r^{2m-d}, & \text{if } d \text{ is odd.} \end{cases}$$

Therefore thin plate splines area an expansion δ_i in radial basis functions $\eta_{md}(r)$.

In the univariate case with derivative penalty of degree two, i.e $d = 1$ and $m = 2$, equation 2.7 is equivalent to that of the cubic smoothing spline, with objective function equation 2.2.

In this case, the minimizing function has the form:

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i |x - x_i|^2 \log(|x - x_i|) + \alpha_0\phi_1(x) + \alpha_1\phi_2(x),$$

where the basis function $\eta_{21}(r) = |x-x_i|^2 \log(|x-x_i|)$ is a specific radial basis function for the univariate TPS. The basis functions ϕ_i are:

$$\phi_1(x) = 1, \quad \phi_2(x) = x_1.$$

Here, δ_i and α_j are coefficients to be estimated, with α_0 and α_1 accounting for the polynomial part of the spline, which form the familiar "linear" null space of the univariate Gaussian objective function we have seen so far.

The expression of a TPS as a linear mixed model, with a separation of basis functions into null and non-null spaces, provides a clear understanding of how the smoothness penalty operates.

We construct the model matrix for the basis functions $\mathbf{E}$ for the radial basis functions $\mathbf{E}_{ij} = |x_i - x_j|^2 \log(|x_i - x_j|)$ and $\mathbf{T}$ for the polynomial basis functions $\mathbf{T}_{ij} = \phi_j(x_i)$.

By expressing the function $\hat{f}(x)$ in terms of δ and α, the minimization problem becomes:

$$\text{minimize} \quad \|\mathbf{y} - \mathbf{E}\delta - \mathbf{T}\alpha\|^2 + \lambda\delta^T\mathbf{E}\delta \quad \text{subject to} \quad \mathbf{T}^T\delta = 0,$$

where $\mathbf{y}$ is the vector of observed data points y_i, $\mathbf{E}\delta$ captures the non-null space (penalized) part of the basis functions, $\mathbf{T}\alpha$ captures the null space (unpenalized) part of the basis functions and the constraint $\mathbf{T}^T\delta = 0$ ensures that the non-null space coefficients δ do not interfere with the polynomial terms.

See subsection 2.1.9 for background on thinking about smoothers as random effects in a linear mixed model framework.

Now Simon Wood constructs a low rank version of TPS called *thin plate regression spline* (TPRS) by leaving the α parameter space unchanged and instead finding a truncated basis in the δ parameter space. Details of this construction are done in [10]. TPRS are the low rank approximation to TPS that are used in the R package `mgcv`, that we use to obtain the basis and penalty matrices for our smoothers. See chapter 3 for more details on how `mgcv` is utilized in constructing compartmental model using smoothers to estimate time varying latent variables.

2.1.7 Cyclic regression splines

If a smooth function has the same value and first few derivatives at its upper and lower boundaries its is called *cyclic*. This means that $f_1''(\mathbf{x}) = f_k''(\mathbf{x})$ and $f_1(\mathbf{x}) = f_k(\mathbf{x})$. In the case of the cubic regression spline the matrices $\mathbf{B}$ and $\mathbf{D}$ are defined similarly and result in the same quadratic matrix expression of the second derivative penalty.

P splines and thin plate splines also can be constructed with cyclic basis. These basis are available in the `mgcv` package and are used by us in chapter 4.

2.1.8 General definition of a penalized spline

The penalized regression problem, formulated to minimize:

$$\|y - \mathbf{X}\beta\|^2 + \lambda\beta^T\mathbf{S}\beta, \tag{2.5}$$

leads to a solution for the *smoothing coefficients* β expressed as:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T y, \tag{2.6}$$

with the corresponding hat matrix:

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T,$$

often called the *smoother matrix*. In a typical data model $y = f(x) + \epsilon$, the estimate $\hat{y}$ is computed as:

$$\hat{y} = \mathbf{L}y,$$

where $\mathbf{L}$ is an $n \times n$ matrix dependent on λ, influencing the degree of smoothing. Although $\mathbf{L}$ renders the smoother non-linear in nature due to its dependency on λ, it can be treated as linear for fixed λ, simplifying the use of penalized regression splines in practical applications.

Now, consider the general case of flexible function estimation within a statistical model defined over some domain $\mathcal{X}$. The functional can be expressed as:

$$S(f) = L(f|\text{data}) + J(f), \tag{2.7}$$

where $L(f|\text{data})$ represents the likelihood of the function f given the data—essentially the deterministic part of the model—and $J(f)$ is a functional that defines what it means for functions f on the domain $\mathcal{X}$ to be smooth.

In the examples discussed previously, such as cubic smoothing splines, P-splines, B-splines, and thin-plate regression splines, each represents a specialized case of this general equation tailored for Gaussian regression. Typically, the data model is assumed to be $y = f(x)+\epsilon_i$, where ϵ_i follows a Gaussian distribution. This assumption is why the likelihood function $L(f)$ adopts the form of a least squares functional, and the roughness penalty $J(f)$ employs a quadratic penalty. However, one could alternatively assume a different distribution (e.g., from the exponential family) for the residual errors in the data model, which would modify $L(f)$ accordingly. The choice of a quadratic functional as the roughness penalty is also based on this assumption. Specializing to Gaussian Regression simplifies equation 2.7 to the specific form seen in equation 2.2.

By focusing on Gaussian regression within the framework of penalized likelihood estimation, we derive an objective function that integrates both a stochastic component,the least squares functional and a roughness penalty component, the quadratic functional.

2.1.9 The duality of smooths and random effects

It is possible to conceptually link penalized smoothing splines with linear mixed models (LMMs). This connection was first explored in-depth in [11]. In this framework, the smooth function in flexible function estimation is represented by a high-dimensional basis. The penalty on spline coefficients, represented as $\beta^T\mathbf{S}\beta$, can be interpreted as an apriori distribution, thus yielding a mixed linear model through the following steps:

1. Transform the smoothing coefficients into the constraint space. Here the constraint matrices is such that $\mathbf{C}\beta = 0$, as in the sum to zero constraint (see section 3.3). Perform QR decomposition $\mathbf{C}^T = \mathbf{QR}$ and define $\mathbf{Z}$ to be the matrix which is $\mathbf{Q}$ less its first n_c

columns, where n_c is the number of rows of $\mathbf{C}$. Now transform the basis matrix into constraint space by setting $\beta_Z = \mathbf{Z}\beta$.

2. Perform the eigendecomposition $\mathbf{Z}^T\mathbf{S}\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ of the penalty matrix $\mathbf{S}$, where $\mathbf{S}$ is rank-deficient due to the non-zero dimension of the null space of $\mathbf{P}$. This transformation aims to align the basis matrices and the coefficient vector with the eigenspace of $\mathbf{S}$.

3. Arrange the eigenvalues in decreasing order along the diagonal and form the submatrix $\mathbf{D}^+$ by including only the non-zero eigenvalues. Then, multiply the coefficients $\beta_U = \mathbf{U}^T\beta_Z$ and transform the basis matrix $\mathbf{X}_U = \mathbf{X}\mathbf{Z}\mathbf{U}$ to align these matrices along the eigenvectors of the penalty matrix and the constraint space. This step separates the effects of the penalty into components associated with non-zero eigenvalues and those that fall into the null space (associated with zero eigenvalues).

4. Partition β_U into $[\beta_u, \beta_F]$ and $\mathbf{X}_U$ into $[\mathbf{X}_u, \mathbf{X}_F]$ based on the aforementioned eigenvalue arrangement. Define $b = \sqrt{\mathbf{D}^+}b_u$ and $\mathbf{X}_R = \mathbf{X}_U(\sqrt{\mathbf{D}^+})^{-1}$, where $\sqrt{\mathbf{D}^+}$ is derived from the Cholesky decomposition of $\mathbf{D}$. The objective function then becomes:

$$s = \|y - \mathbf{X}\beta - \mathbf{X}_\mathbf{R}b\|^2 + \lambda b^T b. \tag{2.8}$$

5. Given the data $\mathbf{y}$, the estimates of $\mathbf{b}$ and β_F that result from minimizing s correspond to the expected values under the mixed model:

$$\mathbf{y} = \mathbf{X}_F\beta_F + \mathbf{X}_R\mathbf{b} + \epsilon, \tag{2.9}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, $\mathbf{b} \sim \mathcal{N}(0, \tau^2\mathbf{I})$, and $\lambda = \frac{\sigma^2}{\tau^2}$. Here, σ^2 and τ^2 are the variances of the observation error terms and random effects, respectively.

In this model, the smooth coefficients $\mathbf{b}$ are tied under a common distribution, allowing them to share information. The columns of $\mathbf{X}_F$ form a basis for the null space of the smoothing penalty, and the columns of $\mathbf{X}_R$ form a basis for its range space. For a deeper exploration of this duality, see [9].

The primary advantage of using the mixed model formulation for penalized splines is that the computational methods developed to estimate random effects in software like `lme4` can also be employed to estimate smooth functions. However, it is possible to write the objective function out explicitly with the penalty term expressed as a quadratic form using the penalty matrices computed via the `mgcv` package. Optimization of the objective function is handled by `macpan2`, which uses template model builder as the optimization engine. Therefore in this case it is not necessary to write out the model in the mathematical form of a mixed model. However, should one choose to adopt this modeling methodology with another optimization engine, detailing the model in the form of a mixed model could be essential. These aspects are further explored in chapter 3.

2.1.10 A Bayesian perspective of smoothing

Considering the quadratic penalty as a legitimate a priori distribution, we derive the likelihood for independent observations (x_i, y_i), $i = 1, \ldots, n$, within the Linear Mixed Model framework:

$$Y \mid u \sim \mathcal{N}(\mathbf{X}\beta + \mathbf{X}_R b, \sigma^2 \mathbf{I}_n), \quad u \sim \mathcal{N}(0, \lambda^{-1}\sigma^2 \mathbf{D}^{-1}),$$

with the marginal likelihood defined as:

$$Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{V}_\lambda),$$

where $\mathbf{V}_\lambda = \mathbf{I} + \lambda^{-1}\mathbf{X}_R\mathbf{D}^{-1}\mathbf{X}_R^T$ as detailed by [12].

In a Bayesian context, prior beliefs about parameters before observing the data are specified through a prior distribution. For smoothing, the penalty term $\beta^T\mathbf{S}\beta$ suggests a prior on β:

$$\beta \sim \mathcal{N}(0, \mathbf{S}^-/\lambda),$$

where $\mathbf{S}^-$ is the pseudoinverse of $\mathbf{S}$, accounting for its rank deficiency. This pseudoinverse $\mathbf{S}^-$ corresponds to $(\mathbf{D}^+)^{-1}$. This prior implies that the values of β are normally distributed with a mean of zero and a covariance matrix $\mathbf{S}^-$, tightening around zero as λ increases and becoming flatter as λ decreases. This Bayesian interpretation of smoothing is discussed in [4].

Consequently, the maximum posteriori (MAP) estimate of β is given by:

$$\beta \mid y \sim \mathcal{N}(\hat{\beta}, (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2), \tag{2.10}$$

which aligns with the solution of Equation 2.6 derived from Equation 2.5.

This reveals that methodologies designed for smoothing problems are applicable for estimating Gaussian random effects.

2.1.11 Gaussian processes regression smoothers

A *random field* is a function f that assigns a random value $f(x_i)$ at each point x_i within its domain $\mathcal{X}$. If we assume that the collection of values $f(x)$ for any finite selection of points $\{x_i\}_{i=1}^n$ follows a multivariate Gaussian distribution, then the subset $\{f(x_i)\}_{i=1}^n$ is jointly normally distributed. This distribution is characterized by a mean function $\mu(x)$ and a covariance function $C(x, x')$, which measures how correlations decay with distance between any two points, thereby defining a *Gaussian Random Field* (GRF).

Unlike being restricted to fixed or discrete locations, a GRF can be generalized by defining a distribution over functions, making the model continuous with respect to its

domain. How is this achieved? The covariance function C of the GRF, initially defined on a set of discrete points such as a lattice, can be generalized through a *kernel* $k(x, x_i)$. This kernel extends the covariance function to a continuous domain, ensuring that its realization over any finite subset of this domain yields a positive semi-definite matrix. This requirement extends the univariate requirement of a positive variance parameter σ^2 to a multivariate scenario.

A *Gaussian process* is thus defined as the model where any finite collection of realizations (i.e., n observations) is treated as having a multivariate normal distribution. The characteristics of these realizations are determined by the mean function $\mu(x)$ and the kernel $k(x, x_i)$, with the latter's realization forming a positive semi-definite symmetric matrix $\mathbf{K}$.

The function f can be represented as:

$$f(x) = (1, \mathbf{x}^T)\beta + \sum_i b_i C(x, x_i),$$

where $C(x, x_i)$ is a non-negative function measuring the distance between two points. The value of $C(x, x_i)$ should equal one when points are identical (indicating maximum correlation) and approach zero as the distance between points increases to infinity. Given the vector b, its prior distribution is $b \sim \mathcal{N}(0, \mathbf{S}^{-1}/\lambda)$. Here, β represents a vector of fixed effects parameters, and the model depicts f as a linear combination of these fixed effects and a random effect weighted by the kernel function $C(x, x_i)$.

In matrix form, f is represented as:

$$f = \mathbf{B}\beta + \mathbf{C}b.$$

To find the covariance matrix of f, compute:

$$\text{Cov}(f) = \text{Cov}(\mathbf{C}b) = \mathbf{C}\text{Cov}(b)\mathbf{C}^T = \mathbf{C}(\lambda\mathbf{C})^{-1}\mathbf{C}^T = \mathbf{C}/\lambda,$$

leveraging the fact that $\mathbf{C}$ is symmetric.

Minimizing the objective function:

$$\|y - \mathbf{B}\beta - \mathbf{C}b\|^2/\sigma^2 + \lambda b^T\mathbf{C}b$$

is equivalent to maximizing the posterior probability of the parameters given the data, usually incorporating σ^2 into the smoothing parameter λ. The values of β and b that minimize this function are the MAP estimates given in equation 2.10.

This methodology is known as *Gaussian process regression*. Originally referred to as kriging in the geostatistics literature of the 1960s, Gaussian process regression's complexity primarily resides in the choice of the kernel. The kernel encodes assumptions

about the function f by defining the concept of proximity or similarity used in the estimation. For an introduction to Gaussian Processes, including many covariance functions and further details, see [13].

2.1.12 Ornstein–Uhlenbeck process

The Ornstein-Uhlenbeck (OU) process emerges as a special case of the Matérn covariance function with $\nu = \frac{1}{2}$. This formulation leads to a stationary first-order Gaussian Markov process.

The *Matérn class* of covariance function is a n in Gaussian processes is defined as:

$$k(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} r}{\ell} \right),$$

where r is the distance between points, σ^2 is the variance, ℓ is the length scale, ν is a smoothness parameter and K_ν is a modified Bessel function of the second kind.

When $\nu = \frac{1}{2}$, the Matérn function simplifies significantly because the modified Bessel function of the second kind, $K_{1/2}(z)$, has a known simple form that relates to the exponential function. The formula for $K_{1/2}(z)$ is:

$$K_{1/2}(z) = \sqrt{\frac{\pi}{2z}} e^{-z}.$$

Substituting $\nu = \frac{1}{2}$ into the Matérn formula, using $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we get:

$$k(r) = \sigma^2 \frac{2^{1-1/2}}{\sqrt{\pi}} \left(\frac{\sqrt{1} r}{\ell} \right)^{1/2} \sqrt{\frac{\pi}{2\frac{\sqrt{1}r}{\ell}}} e^{-\frac{\sqrt{1}r}{\ell}}.$$

This simplifies to:

$$k(r) = \sigma^2 e^{-r/\ell}.$$

Here, ℓ acts as a scale parameter, and the resulting covariance function is the exponential covariance function.

The *Ornstein-Uhlenbeck process* is a continuous-time stochastic process that is both Gaussian and Markov, characterized by its mean-reverting property. The covariance function of the OU process over time t with mean reversion rate θ is given by:

$$k(t) = \sigma^2 e^{-\theta|t|},$$

where $\theta > 0$ is the rate at which the process reverts to its mean.

Comparing this with the exponential covariance function derived from the Matérn function, we see that they are essentially the same form when interpreted over time rather than space, with $\theta = 1/\ell$. Thus, the OU process, which has this exponential form of the covariance function, is a Gaussian Markov process.

Setting $\nu = \frac{1}{2}$ in the Matérn covariance function yields an exponential covariance function, which corresponds to the covariance structure of the OU process. The first-order Markov property in the OU process is given from the memoryless feature of the exponential decay in its covariance function. This demonstrates how the OU process, as a stationary first-order Gaussian Markov process, is a special case of the Gaussian processes modeled by the Matérn covariance function with $\nu = \frac{1}{2}$.

2.2 Compartmental models

Compartmental models are a technique for the mathematical modelling of infectious diseases. These models can be formulated as directed graphs using the language of graph theory [14].

Consider the stratification of the total total population of individuals N. Then suppose there are n possible states at which an individual can be in. Individuals transition from one state to another, and these transitions can be depicted as flows. Within the framework of graph theory, these flows are visualized as directed edges linking nodes, where each node represents a state.

A flow between compartments is defined by a function that may depend on the conditions of any of the n compartments and any of the m parameters. The collection of all possible flow functions constitutes the model's *state space*, and the set of all potential parameters forms the model's *parameter space*. This approach to compartmental models through graph theory allows for a rigorous definition, as discussed by Flynn-Primrose et al. The primary purpose of this method is to facilitate the construction of complex models from a modified Cartesian product of simpler directed graphs.

However, for our purposes, we employed two very simple compartmental models, so a deeper development of this theory is not necessary for the current discussion.

2.2.1 The SIR and SIRS models

The *SIR model* (Figure 2.1) categorizes the population into three distinct compartments: $S(t)$, $I(t)$, and $R(t)$, which represent the number of susceptible, infected, and recovered individuals, respectively. As the disease progresses, the number of individuals in each compartment changes over time, thus these compartments are represented as functions of time. The transitions between these compartments are guided by the processes depicted in the schematic shown in Figure 2.1.

The model incorporates two primary parameters: the transmission rate (β) and the recovery rate (γ). The *transmission rate*, β, measures the probability of disease transmission per contact per unit time and reflects the likelihood that an interaction between a susceptible and an infected individual results in transmission. This rate is used to calculate βSI, where SI is the total number of interactions between susceptible and infected individuals; βSI thus represents the expected number of new infections per unit time.

On the other hand, the *recovery rate*, γ, indicates how quickly infected individuals recover and gain immunity. If an individual is infectious for a period D, then γ is defined as $\gamma = \frac{1}{D}$. Therefore, γI calculates what proportion of the infected population will recover during any given time interval, moving from the infected compartment to the recovered compartment, based on the infectious period D.

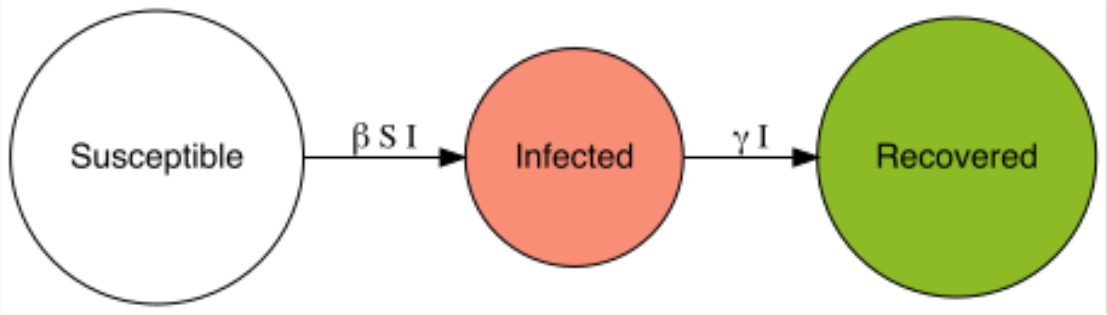

FIGURE 2.1: **A Susceptible-Infected-Recovered (SIR) model.** The edges are the flows from one compartment to another. The nodes are the compartments that represent an element in the stratification of the total population.

For any given time t, the rates of transition between compartments in the SIR model can be derived from Figure 2.1 and expressed as a nonlinear system of ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS, \\ \frac{dI}{dt} &= \beta IS - \gamma I, \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{2.11}$$

This set of equations assumes a closed population—meaning there are no births or deaths—such that the total rate of change across all compartments sums to zero:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0.$$

This conservation of the total population implies that the sum of susceptible, infected, and recovered individuals remains constant over time:

$$N = S + I + R.$$

The basic SIR model can be modified to include a process where recovered individuals become susceptible again after losing immunity. This adaptation introduces a waning

immunity parameter, ϕ, which quantifies the rate at which recovered individuals lose their immunity and return to the susceptible compartment. This extended model, illustrated in Figure 2.2, is known as the SIRS model or SIR model with waning immunity.

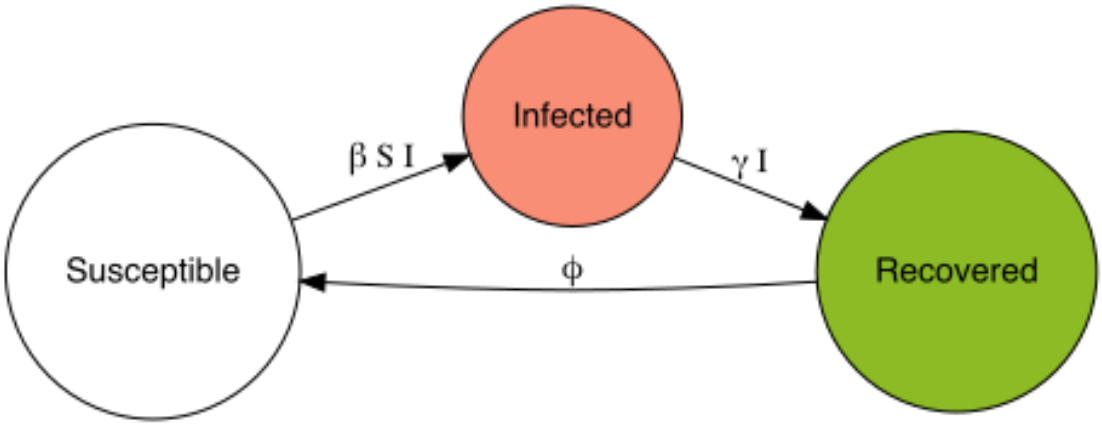

FIGURE 2.2: **A SIR model with waning immunity (SIRS).** The waning parameter represents the flow of individuals who have lost their natural immunity from the recovered back to the susceptible compartment.

2.2.2 Force of Infection (FOI) and the Basic Reproduction Number (R_0)

An important quantity is called the *force of infection* (FOI). In equation 2.11, for $\frac{dS}{dt}$, the term $\beta\frac{I}{N}S$ represents the rate at which susceptible are becoming infected. Here, $\lambda = \beta\frac{I}{N}$ is the force of infection, which means that each susceptible individual has a probability λ of becoming infected per unit time. The FOI is defined as as the per capita rate at which susceptible individuals contract the disease. Essentially, it quantifies the risk that a susceptible individual faces of becoming infected at a given time, depending on the current prevalence and contagiousness of the disease in the population. The force of infection is directly proportional to the number of infectious individuals in the population. As I increases, so does λ, increasing the risk of infection among susceptibles. A higher β increases λ, indicating that more contacts (or more effective transmission per contact) raise the likelihood of infection.

There is an important quantity $R_0 = \frac{\beta}{\gamma}$, called the basic reproduction number. An infectious individual contacts β individuals per unit time, and the proportion of susceptibles in the population is $\frac{S}{N}$. Therefore, the effective contacts that can result in a new infection are $\beta\frac{S}{N}$. An infectious individual remains infectious for $\frac{1}{\gamma}$ units of time, on average (since γ is the rate at which individuals recover and cease being infectious). The *basic reproduction number* R_0 can be calculated as the product of the infection rate per

contact, the number of contacts per unit time, and the duration of infectiousness:

$$R_0 = \beta \frac{S}{N} \frac{1}{\gamma} = \frac{\beta}{\gamma}.$$

Therefore R_0 is the average number of secondary cases of disease caused by a single infected individual over his or her infectious period. [15] discusses the subtleties in defining and estimating the reproductive number.

At the start of an epidemic, assuming almost the whole population is susceptible ($S \approx N$), this simplifies to $R_0 = \frac{\beta}{\gamma}$. If $R_0 > 1$, each infectious individual, on average, infects more than one other person, leading to the potential for an epidemic. Conversely, if $R_0 < 1$, the disease will likely die out in the population over time. Understanding R_0 helps in predicting disease behavior and controlling outbreaks. For instance if we can reduce β (e.g., through vaccination, social distancing, or wearing masks), or increase γ (e.g., through faster diagnosis and treatment), R_0 can be brought below 1, aiming to control the spread of the disease. The proportion of the population that needs to be immune (via recovery or vaccination) to stop disease spread is estimated by $1 - \frac{1}{R_0}$.

2.2.3 Numerical solutions of discrete time compartmental models

Compartmental models for infectious diseases are usually formulated as a system of differential equations. They can run with ordinary differential equations, whose trajectory is deterministic in the sense of being entirely determined by the model parameters. Given a model M, state space X, and parameters $P = (p_1, \ldots, p_n)$, the system's evolution is tracked by solving the differential equations numerically. This process involves choosing an initial state X_0 from the state space X, and then applying numerical integration techniques, such as the Euler method or Runge-Kutta methods, to compute the state $X(t)$ at future times t. The trajectory $X(t)$ for $t = 0, 1, 2, \ldots, T$ represents the evolution of the compartments in the model.

For the SIR model $X(t) = (S(t), I(t), R(t))$. For any discrete time step t, the state $X(t+1)$ at time $t+1$ can be derived from the state $X(t)$ at time t using the model's parameters $P = (\beta, \gamma)$. In the context of the SIR model, these updates can be described by the following set of *difference equations*, which approximate the continuous model's dynamics:

$$\begin{aligned} S(t+1) &= S(t) - \Delta t \cdot \beta \cdot I(t) \cdot S(t), \\ I(t+1) &= I(t) + \Delta t \cdot (\beta \cdot I(t) \cdot S(t) - \gamma \cdot I(t)), \\ R(t+1) &= R(t) + \Delta t \cdot \gamma \cdot I(t), \end{aligned}$$

Difference equations are discrete models that directly approximate changes in a system's state at discrete time steps. These do not require the computation of derivatives.

The Euler method is a numerical technique used to solve ordinary differential equations (ODEs) by approximating the solution at successive time steps. The Euler method

takes these derivatives and uses them to estimate the states at the next time step. The general form of the Euler update for a variable x is:

$$x(t + \Delta t) = x(t) + \Delta t \cdot f(t, x(t))$$

where $f(t, x(t))$ is the derivative $\frac{dx}{dt}$ at time t.

In the context of an SIR model described by ODEs, the *Euler method* would use the current state to estimate the derivative and then step forward in small increments Δt to update the state:

$$\begin{aligned} S(t + \Delta t) &= S(t) - \Delta t \cdot \beta \cdot I(t) \cdot S(t), \\ I(t + \Delta t) &= I(t) + \Delta t \cdot (\beta \cdot I(t) \cdot S(t) - \gamma \cdot I(t)), \\ R(t + \Delta t) &= R(t) + \Delta t \cdot \gamma \cdot I(t). \end{aligned} \tag{2.12}$$

See [16] for more information regarding numerical analysis in mathematical epidemiology.

Chapter 3

Materials and methods

This chapter covers the technical details of using semi-mechanistic models. There are two main aspects. The first aspect explains how to construct linear smoothers using the `mgcv` package. The second aspect describes how to implement these linear smoothers within a compartmental modeling framework using `macpan2`. If your goal is to implement your own model in `macpan2` using the smoothing parameter estimation methodology to infer a latent variable by fitting the model to data, this chapter provides the general methodology to do so. It also addresses some of the unique technical issues we encountered and their solutions.

3.1 Software

Mcmaster Pandemic 2 (`macpan2`) is an `R` modelling package designed as a compartmental modelling tool that is agnostic about its underlying computational but currently uses template mode builder (`TMB`). It allows the user to write complex bespoke compartmental models in a user friendly way.

Template Model Builder (`TMB`) is an R package specifically designed to fit latent variable models efficiently to data. With `macpan2`, the user is able to write the negative log-likelihood of the their objective function with respect to the parameters to be fit to the data, in R code. `macpan2` then converts this objective function to `C++` code. It then implements maximum likelihood estimation and uncertainty calculations by maximizing the Laplace approximation of the marginal likelihood. See appendix A.2 for a brief introduction to the Laplace approximation.

The use of the Laplace approximation to estimate model parameters and their uncertainties involves the computation of complex and high-dimensional second-order derivatives. This challenge arises because the likelihood function often exhibits non-linear behaviors characterized by multiple local maxima, steep regions, and flat plateaus. Computing the Hessian, which reflects the curvature of the likelihood surface at a point, can be numerically unstable if the surface is irregular or flat. The direct computation of the Hessian involves calculating second-order partial derivatives for every pair of parameters, significantly increasing computational load and the risk of numerical inaccuracies due to rounding and approximation errors, particularly when using discrete numerical methods

like Euler's method. Optimization algorithms such as BFGS, which utilize the Hessian, depend on accurate estimates of this matrix for efficient parameter updates. Inaccuracies in the Hessian can lead to suboptimal parameter updates, slow convergence, or convergence to non-optimal points.

`TMB` harnesses the capabilities of *automatic differentiation* (`AD`), a computational technique for accurately calculating derivatives of functions. Unlike numerical or symbolic differentiation, `AD` operates by exploiting the fact that all computationally implemented functions decompose into a finite sequence of elementary arithmetic operations and functions. Using the chain rule, `AD` breaks down these complex functions into simpler operations, computing derivatives in a sequence that parallels the function's evaluation. This method enables the precise calculation of derivatives up to machine precision. For its implementation of `AD`, `TMB` utilizes the `C++` libraries `CppAD` for automatic differentiation and `Eigen` for handling both sparse and dense matrix computations.

In `macpan2`, specifying that the optimizer include uncertainty estimates for parameters is straightforward. This functionality enables the computation of Wald confidence intervals with specified uncertainty levels. Confidence intervals, as discussed in the chapter 4, are computed using this method. See section 3.5 for details on how the confidence intervals are constructed in the case of aggregated data.

For further reading on `TMB`, refer to [17]. For more information on the Laplace Approximation, see [18]. Refer to the section 3.4 for details on deriving the objective function for the semi-mechanistic model and its implementation in `macpan2`.

3.2 Time varying transmission rate

We specify the *time-varying transmission rate* β in our model using a linear smoother defined as

$$\beta = \exp(b_0 + \mathbf{X}b), \tag{3.1}$$

where b_0 is the intercept, $\mathbf{X}$ is the basis matrix of dimensions $n \times (k-1)$, and b is a vector of basis coefficients of length $k-1$. The basis matrix $\mathbf{X}$ is constructed using the `smoothCon` function from the `mgcv` package in R. The structure of $\mathbf{X}$ depends on the selected type of smoother, as indicated by the `bs` parameter, and the number of knots k.

The `R` package *Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (`mgcv`), developed by Simon Wood, implements a variety of smoothers that can be used for penalized General Linear Models. In our approach, we utilize `smoothCon` to create the model matrix $\mathbf{X}$ and its corresponding penalty matrix $\mathbf{P}$ for β. This function facilitates the capture of the nonlinear relationships of the latent variable β from the data by constructing a univariate Gaussian regression smoother. While typically used internally by `mgcv` in calls to the `gam` function for fitting generalized additive models,

`smoothCon` serves as a critical low-level function for constructing smooth terms in our model.

This function is configured via the `bs` argument to select the type of smoother and the `k` argument to determine the number of basis functions, which we refer to as `num_variables`. The number of observations in the data is represented by `n`. The initial step involves constructing a simple data frame `dd = seq(from = 0, to = n, by = 1)`, which discretizes our time variable into intervals that map directly onto the domain over which the smoother operates. This setup allows `smoothCon` to accurately interpret the temporal structure of our data.

The command to execute this configuration in R is as follows:

```
s <- smoothCon(object = s(time, bs = smooth, k = num_variables, ...),
               absorb.cons = TRUE, data = dd,
               knots = num_variables)
```

This function call yields two components for the model: the basis matrix $\mathbf{X}$ and the penalty matrix $\mathbf{P}$.

For instance, specifying `bs = cr` configures the basis and penalty matrices for a cubic regression spline, which is detailed further in the subsection 2.1.3. Alternatively, when using `bs = gp` for a Gaussian process regression smoother, it becomes necessary to define the kernel type within the additional arguments (`...`). Guidelines on kernel specification and available smooths can be found in Simon Wood's `mgcv` package documentation [19].

Details on the range of smoothers implemented in our models and their respective kernel choices are discussed in the chapter 4.

The argument `absorb.cons = TRUE` absorbs the identifiability constraints into the basis matrix rather than being applied as external conditions or through additional penalty terms. The default *identifiability constraint* in `mgcv` ensures the smooth sums to zero over the observed values of x_j, i.e,

$$1^T \mathbf{X} \beta = 0.$$

This implies $1^T \mathbf{X} = 0$.

`mgcv` implements this constraint by constructing the following QR decomposition.

Let

$$\mathbf{C}^T = \mathbf{U} \begin{bmatrix} \mathbf{P} & 0 \end{bmatrix},$$

where $\mathbf{U}$ is a $p \times p$ orthogonal matrix and $\mathbf{P}$ is an $m \times m$ upper triangular matrix. The zero matrix appended to $\mathbf{P}$ is $m \times (p - m)$ to match the dimensions of $\mathbf{U}$. Now, $\mathbf{U}$ is partitioned as $\mathbf{U} \equiv (\mathbf{D} : \mathbf{Z})$, where $\mathbf{D}$ is a $p \times m$ matrix and $\mathbf{Z}$ is a $p \times (p - m)$ matrix.

Given that $\beta = \mathbf{Z}\beta_z$ and β_z is a $(p-m)$-dimensional vector, we compute:

$$\mathbf{C}\beta = \begin{bmatrix} \mathbf{P}^T \\ 0 \end{bmatrix} \begin{bmatrix} \mathbf{D}^T \\ \mathbf{Z}^T \end{bmatrix} \mathbf{Z}\beta_z = \begin{bmatrix} \mathbf{P}^T\mathbf{D}^T \\ 0 \end{bmatrix} \mathbf{Z}\beta_z = \begin{bmatrix} \mathbf{P}^T \\ 0 \end{bmatrix} \beta_z = 0,$$

where we utilized that $\mathbf{D}^T\mathbf{Z} = 0$, and $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}_{p-m}$ because $\mathbf{U}$ is orthogonal, hence $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$.

To minimize equation 2.5 such that $1^T\mathbf{X}\beta = 0$, find the $k \times (k-1)$ matrix Z and reparameterize the basis matrix to $\mathbf{XZ}$ and the penalty matrix to $\mathbf{Z^TPZ}$.

A computationally more expensive equivalent method to implement is to zero center $\mathbf{X}$ but it worth illustrating because it is more intuitive. By default, the basis matrix $\mathbf{X}$ produced by `mgcv::smoothCon` doesn't include an intercept. Zero-centering the spline basis functions ensures that the spline components represent deviations or variations around a central tendency, rather than absolute values. This allows the intercept, b_0, in the model to uniquely capture the central tendency of the response variable. Consequently, the intercept and the spline coefficients are identifiable as distinct contributors to the model: the intercept as the average response and the spline coefficients as the adjustments from this average. Each spline coefficient can be interpreted as the effect of that basis function relative to the central tendency captured by the intercept

Zero centering is implemented by subtracting the column mean from each column of $\mathbf{X}$. This reduces the rank of $\mathbf{X}$ to $k-1$. The solution then is to drop the row and column of $\mathbf{X}$ corresponding to the zero eigenvalue and delete the corresponding element of β.

$\mathbf{X}$ and b now have dimension one less than the number of knots owing to dimension of the null space of the penalty matrix. The null space of the smoothing penalty matrix being of dimension 1 means that there's essentially one direction (in the parameter space) along which the function can vary without incurring any penalty. This is associated with the smooth function being able to revert to a simple linear trend without penalty because a linear function in the space spanned by the spline basis functions wouldn't be penalized. If a heavy penalty is applied to the smooth terms the model output reverts to a linear trend. This occurs because, under heavy penalization, the model minimizes the penalized complexity by adopting the simplest form that incurs the least penalty, which is a linear function.

There are difficulties encountered when working with the penalty matrix that has been transformed into the constraint space of the sum to zero constraint. Computing the eigendecomposition of the penalty matrix, returned by `mgcv::smoothCon`, one of the eigenvalues is essentially zero, in terms of numerical precision, being on the order of $\leq 10^{-15}$. This implies that the penalty matrix is singular. It is not possible to take the logarithmic determinant of a singular matrix when taking numerical precision limitations into account. Computing the log determinant is part of the objective equation (3.2). To overcome this we can take the regularized determinant by adding a small value ($\approx 10^{-5}$)) to the diagonal of the penalty matrix. Another option (which we did not

implement) is to take the singular value decomposition $\mathbf{P} = \mathbf{U\Sigma V}^T$ and use the fact that $\text{logdet}\mathbf{P} = \sum_i \log\Sigma_{ii}$.

The Gaussian process smooth, `bs = gp`, has some unique issues. The null space of the basis were dominating the smooths. This was diagnosed by plotting the basis functions and noting that there was a large difference in magnitude of the linear constraints of the null space and the rest of the basis functions. The solution is to scale $\mathbf{X}$ to make it compatible with a choice of the standard deviation used to simulate the starting values for β. Each column of the basis matrix $\mathbf{X}$ is normalized by dividing it by its Euclidean norm, resulting in each column having a unit norm. Then $\mathbf{X}\beta$ is on the range of about of plus or minus $log(2)$ (i.e $\mathbf{X}\beta \pm 1$), since

$$\beta \sim \mathcal{MVN}(0, b_{sd}).$$

To ensure that the penalization term $\mathbf{XPX}^T$ in the likelihood equation reflects the scaling applied to $\mathbf{X}$, the penalty matrix $\mathbf{P}$ must be adjusted to align with the scaling of the basis matrix $\mathbf{X}$. Since the columns of $\mathbf{X}$ are normalized, any transformations applied to $\mathbf{X}$ necessitate corresponding adjustments to $\mathbf{P}$. When each column of $\mathbf{X}$ is divided by its norm, the scaling effect on $\mathbf{P}$ must square the scaling factors used on $\mathbf{X}$. Given that the scaling of $\mathbf{X}$ involves dividing each column by its Euclidean norm, and denoting these norms as $\|x_i\|$ for each column x_i, the appropriate scaling for $\mathbf{P}$ would involve multiplying it by the square of these norms on both sides, i.e., $\mathbf{P}_{\text{scaled}} = \mathbf{DPD}$, where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are $\frac{1}{\|x_i\|^2}$.

When the sum-to-zero constraints are absorbed into the basis matrix, this also sets the penalty matrix for the Gaussian process to have the last row and column equal to zero, effectively absorbing the null space constraints into the penalty matrix. This makes $\mathbf{P}$ singular. If we remove the final row and column of $\mathbf{P}$, then we get a non-singular matrix.

3.3 Time-varying effective reproduction number

The effective reproductive number, denoted as R_t, dynamically reflects the average number of secondary infections that an infectious individual can cause at a specific time t in a population where not all members are susceptible. Unlike R_0, which assumes that the entire population is susceptible, R_e adjusts for changes in susceptibility due to factors such as immunity from previous infections or vaccinations.

As the epidemic progresses, the proportion of susceptible individuals decreases either through infection—which can lead to immunity or death—or through vaccination. This reduction in the susceptible population is quantified by $S(t)/N$, where $S(t)$ represents the number of susceptible individuals at time t, and N is the total population.

The *effective reproduction number* R_e is then calculated by adjusting R_0 for the fraction of the population that remains susceptible:

$$R_e = R_0 \times \frac{S(t)}{N}$$

This equation implies that R_e will decrease over time as $S(t)$ reduces, either due to increasing immunity within the population or through interventions that effectively reduce the contact rate β.

Instead of treating β as a constant, it is estimated at every time step as a smooth function. This allows β, the time-varying transmission parameter, to adapt and change over time, resulting in the sequence $\{\beta(t_i)\}_{i=1}^{n}$. Meanwhile, the recovery rate γ remains fixed at its initial value. Consequently, this framework enables the calculation of the *time varying effective reproductive number* based on the dynamically adjusting β. It is defined as

$$R_t = \frac{\beta(t)}{\gamma} \times \frac{S(t)}{N}.$$

Note that in some definitions of the time varying effective reproduction number, both γ and β are estimated as time-varying parameters. However, in this context, γ is treated as a constant. In Chapter 4, the starting value of γ for each disease is derived from existing literature. Meanwhile, in the simulation study, γ is assigned a reasonable fixed value.

In section 3.2 we described how to estimate a time varying transmission parameter to compute estimates for the force of infection. This estimate is then used to compute the time varying effective reproduction number.

3.4 Inititial conditions and parameters

The simulation study employs the SIRS model, while the examples utilize the SIR model. The simulation study initializes the starting values of S, I, and R to the following endemic equilibrium solution states:

$$\begin{aligned} S &= \frac{\gamma N}{\beta}, \\ I &= \frac{\phi N(\beta - \gamma)}{\beta(\phi + \gamma)}, \\ R &= \frac{\gamma N(\beta - \gamma)}{\beta(\phi + \gamma)}. \end{aligned}$$

These expressions are derived by setting the derivatives of the flow equations to zero and solving for S, I, and R in terms of the model parameters. This procedure allows us to determine the equilibrium states where the number of individuals in each compartment remains constant over time.

By initializing the compartment values deterministically as functions of the total population N, the transmission rate β, the recovery rate γ, and the waning rate ϕ, we can start the simulation close to the endemic equilibrium. This approach stabilizes the influence of the starting parameters, particularly β, on the model dynamics. Consequently, the system begins in a balanced state, reducing the transient effects that might otherwise occur due to arbitrary initial conditions. This transformation was not necessary for the real-world data examples.

For these examples, the state vectors are initialized as follows:

$$\begin{aligned} S &= N - I_0, \\ I &= I_0, \\ R &= 0. \end{aligned}$$

The *initial number of infected individuals*, I_0, at time $t = 0$ is modeled as a fixed parameter. To incorporate flexibility and account for uncertainty about I_0, we employ a log-normal prior distribution. We use a similar approach for the recovery rate γ, also modeled with a log-normal prior to ensure positivity and reflect our uncertainty regarding its value.

In our computing environment (`macpan2`), lacking a dedicated log-normal density function akin to `stats::dlnorm`, we calculate on the log-scale. Specifically, we set the mean of the priors to the logarithmic values of I_0 and γ, and the standard deviation for each parameter is set to a small reasonable value to instill a sharp prior as in Figure 3.1

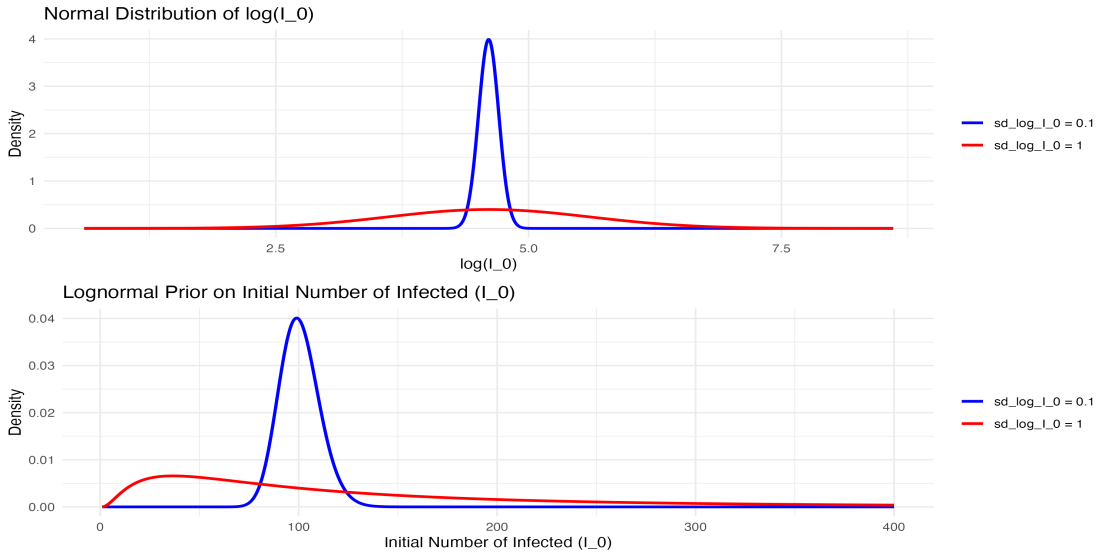

FIGURE 3.1: **Comparison of large and small values of standard deviation on the lognormal prior for the initial number of infected individuals.** The blue line displays a sharp prior corresponding to a small standard deviation, while the red line displays a weakly informative prior for a large standard deviation. The upper figure shows the distribution on the log scale and the lower figure shows the distribution on the exponentiated scale.

To compute the log-normal densities, we utilize the normal density function available within `TMB` as an engine function (functions that can be written in R code that are available to the `C++` compiler). This ensures that non-negativity constraints for both I_0 and γ are maintained by transforming and computing their distributions on the log-scale.

The smoothing coefficients vector $b = (b_1, \ldots, b_{k-2})$ is initialized using $k-1$ random draws from a standard normal distribution.

The intercept b_0 of the linear smoother (3.1), for the time varying transmission, is estimated in the model. It starting value is set to the logarithm of the starting value of β at time $t = 0$.

We can compute the likelihood of the penalty term by making the assumption that the spline basis coefficients β follow a multivariate Gaussian distribution, i.e., $\beta \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. The likelihood function for β is then

$$l(\beta) = (2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right).$$

The log-likelihood becomes

$$L(\beta) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{\Sigma})) - \frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu).$$

Now let $\beta = \mathbf{x} - \mu$ and $\Sigma^{-1} = a\mathbf{S}$, where $a = \frac{1}{\sigma^2}$, $\sigma^2 \in \mathbb{R}$ and $\mathbf{S}$ is the penalty matrix. This implies $\mathbf{\Sigma} = a^{-1}\mathbf{S}^{-1} = \sigma^2\mathbf{S}^{-1}$.

Then,

$$\begin{aligned}
L(\beta) &= -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{\Sigma})) - \frac{1}{2}\beta^T\mathbf{S}\beta \\
&= -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\det(\sigma^2\mathbf{S}^{-1})) + \frac{1}{\sigma^2}\beta^T\mathbf{S}\beta \\
&= -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log(\det(\mathbf{S}^{-1})) + \frac{1}{2\sigma^2}\beta^T\mathbf{S}\beta \\
&= -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \log(\det(\mathbf{S})) + \frac{1}{2\sigma^2}\beta^T\mathbf{S}\beta.
\end{aligned}$$

The term σ^2 represents a variance component that scales the penalty matrix $\mathbf{S}$. It acts as a global variance parameter that moderates the extent to which the penalty is applied. By scaling $\mathbf{S}$ with σ^2, you effectively adjust the strength of the regularization relative to the variance of the data. Consequently, $\lambda = \frac{1}{2\sigma^2}$ functions as a regularization parameter, controlling the "wiggliness" of the fit by influencing the variance of the distribution of the smoothing coefficients. This setup can be viewed as placing a prior distribution on β, with $\mathbf{S}$ acting as the precision matrix of the prior. This approach is analogous to the Bayesian perspective of smoothing, as discussed in equation 2.1.10. Here, σ^2 scales the precision matrix of the prior, influencing how strongly the prior beliefs (e.g., smoothness) affect the posterior estimates.

Therefore,

$$\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\lambda) - \log(\det(\mathbf{S})) + \lambda\beta^T\mathbf{S}\beta \tag{3.2}$$

is the derived form of the penalty functional $J(f)$ in equation 2.7 when the smoothing coefficients are assumed to be Gaussian.

Considering the data model $Y_i = f(x_i) + \epsilon_i$, where $i = 1, \ldots, n$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, the likelihood functional $L(f|\text{data})$ in equation 2.7 simplifies to a least squares functional. This is proportional to $\sum_{i=1}^{n}(Y_i - f(x_i))^2$, aligning with the Gaussian likelihood. Therefore, the likelihood can be expressed using a Gaussian density function evaluated at the vector of observed values $\mathbf{Y}$, with mean $f(\mathbf{x})$ and variance σ_Y^2.

Thus, to fit the combined objective function $L(f|\text{data}) + J(f)$, it is necessary to estimate both the smoothing parameter λ and the variance σ_Y^2.

The number of basis functions or knots to use in the model is not algorithmically optimized. The basis dimension k was chosen to be just large enough that the plotted fits appeared to converge to a stable fit. In [4], Simon Wood outlines the methodology to compute a quantitative measure of whether a particular choice of basis dimension is appropriate. However, in our case, the smoothing parameter does most of the work in avoiding overfitting. Our goal was to simply show that semi-mechanistic models can be used to give very reasonable looking fits in order to estimate latent variables in compartmental models.

3.5 Model formulation

Recall the following assumptions in the model:

$$\begin{aligned} I_0 &\sim \text{Lognormal}(\mu_{I_0}, \sigma^2_{I_0}) \\ \gamma &\sim \text{Lognormal}(\mu_\gamma, \sigma^2_\gamma) \\ Y &\sim \mathcal{N}(f(x), \sigma^2_Y) \\ \beta &\sim \mathcal{N}(0, \frac{\mathbf{S}^{-1}}{\sigma^2}), \end{aligned}$$

where $f(x)$ is the fitted values (incidence). Note that we are not estimating the observed values Y but estimating the variance σ^2_Y corresponding to its likelihood equation. In this way the fitted values $f(x)$ behave as a sort of Gaussian process. As each iteration of the simulation of the trajectory proceeds, the fitted values will be updated and the likelihood function will respond accordingly by adjusting the covariance function.

The model assumptions and starting conditions are specified and passed to a simulator object in `macpan2`. `TMB` simulates the trajectory using the Euler method as explained in subsection 2.2.3.

At $t = 0$ the smoothing basis $\mathbf{X}$, the vector of smoothing coefficients b and its intercept b_0 are used to construct the transmission rate β, a vector equal to the number of observations, of size n. At each time step, $1 \leq t \leq n$, β is used to compute the number of new infections (`incidence`), which in turn is used to compute the total number of infected (`I`) and susceptible (`S`) at that time point. Additionally, β is used to compute the instantaneous effective reproduction number R_t.

At time $t = n + 1$, the negative log likelihood is minimized subject to finding the optimal values of the starting values of the initial number of infected I_0, the recovery rate γ, the variance σ^2_Y of likelihood of the observed data and the regularization/smoothing parameter σ^2. Note that the priors on the recovery rate and the initial number of infected are not "fully Bayesian" in the sense that there are not priors placed on the mean and variance of the prior distribution, i.e no there are no hyper-priors. Parameter estimates for the intercept b_0 of the linear smoother are also obtained.

For each iteration, the simulated trajectory is matched to the observed values and the likelihood is calculated using the Laplace approximation. New parameter estimates are updated using quasi-newton methods via `nlminb`. This process is then iterated until the parameter estimates converge.

Here is an example of what the simulator object for an SIR model in `macpan2` looks like using the above formulation:

```
----------------------
Before the simulation loop (t = 0):
----------------------
1: I_0 ~ exp(log_I_0)
2: gamma ~ exp(log_gamma)
3: lambda ~ exp(log_lambda)
4: I_sd ~ exp(log_I_sd)
5: S ~ N - I_0
6: R ~ 0
7: I ~ I_0
8: S ~ N - I - R
9: eta ~ b_0 + (X %*% b)

----------------------
At every iteration of the simulation loop (t = 1 to n):
----------------------
1: theta ~ eta[time_step(1)]
2: beta ~ exp(eta[time_step(1)])
3: R_t ~ (log(beta) - log(gamma) + log(S) - log(N))
4: incidence ~ S * I * beta/N
5: recovery ~ gamma * I
6: S ~ S - incidence
7: I ~ I + incidence - recovery
8: R ~ R + recovery

----------------------
After the simulation loop (t = n+1):
----------------------
1: log_lik ~ -sum(dnorm(incidence_obs, incidence_fitted, incidence_sd)) -
              dnorm(log_gamma, mean_log_gamma, sd_log_gamma) -
              log(det(P)) -
              dnorm(log_I_0, mean_log_I_0, sd_log_I_0) +
              log(sigma^2) +
              ((t(b) %*% P %*% b) / sigma^2)
```

Sometimes it is useful to simulate the trajectory for n time steps and then calibrate the model over a smaller time series by aggregating the data into $\frac{n}{k}$ time steps by

averaging the trajectory and data over a period of k steps. For example, in the case of the Ireland Covid-19 dataset (4.3), the reported incidence was inconsistent on a daily scale. By averaging the observations over a weekly scale, several statistical improvements are achieved. First, variance reduction occurs as the averaging process diminishes the day-to-day fluctuations caused by sporadic reporting. Secondly, this reduces noise by smoothing out the random variations. Additionally, stabilization of the data set is achieved; this makes the trends more reliable and the overall dataset less susceptible to the anomalies of daily reporting. However, upon calibrating the model to the aggregated data, the estimated transmission rate was double the value then what was estimated when fitting to the unaggregated (daily) data. The effective reproduction number was around four times smaller than expected as well. We were unable to fix this issue but it is suspected that since the trajectory uses Euler steps, the step size of a week is too large and causing the inaccuracy in magnitude. We plan to address this issue so that these models can be calibrated to aggregated data accurately.

Extra care is needed to handle the uncertainty estimates of aggregated trajectory simulations. For unaggreggated data, `macpan2`, which has `TMB` for its optimization engine, uses the Laplace Approximation (see appendix A.2) to compute uncertainty estimates with the delta method. By averaging the trajectory over a time period of size k, we are in effect making a transformation of a random variable. The uncertainty estimates are required to take this into account. The variance of a function $h(\beta) = \mathbf{H}\beta$ of the random variable β is computed as

$$Var(\mathbf{H}\beta) = \mathbf{H}^T Cov(\beta)\mathbf{H},$$

where $\mathbf{H}$ is a $n \times n_k$ indicator matrix, where n_k is the integer ceiling of $\frac{n}{k}$ such that $\frac{n}{n_k} \in \mathbb{Z}$. The transformed variance is then used to compute the Wald confidence intervals as in the unaggreggated case.

Sometimes the estimates for the transmission rate can be close to zero and then the associated standard error produces negative uncertainty estimates which is a non nonsensical value for a transmission rate. A negative transmission rate does not have any meaning. To produce non negative uncertainty estimates we compute the confidence intervals on the logarithmic scale and then exponentiate the upper and lower bounds.

3.6 Model comparison and selection

How can we account for overfitting in statistical modeling when comparing the fit of two different models, which incorporate penalization, to data? The answer is quite complex and nuanced but the resulting measure is quite elegant.

See appendix A.3 for the background on the using *Akaike information criterion* (AIC) for comparison between models with unpenalized parameters. Given equation A.11, we now have a measure for model performance that takes into account complexity by penalizing for the number of parameters in the model. However, the notion of degrees of

freedom for penalized smoothing coefficients is more complicated than in the unpenalized case. To address this complexity, it is helpful to introduce the concept of natural parameterization and the effective degrees of freedom (EDF). These concepts explain the impact of penalties on model coefficients and for defining a correction notion for what p should be in expression A.11.

In the context of penalized smoothers, Simon Wood [4] describes a "natural" parameterization, also known as Demmler and Reinsch parameterization, that simplifies the understanding of how penalties affect model degrees of freedom. The *natural parameterization* transforms the parameter estimators such that they are independent with unit variance in the absence of a penalty, and the penalty matrix becomes diagonal.

Consider a model with a design matrix $\mathbf{X}$, parameter vector $\boldsymbol{\beta}$, wiggliness penalty matrix $\mathbf{S}$, and smoothing parameter λ. Using the QR decomposition, $\mathbf{X}$ is factorized as $\mathbf{X} = \mathbf{QR}$. Re-parameterizing in terms of $\boldsymbol{\beta''} = \mathbf{R}\boldsymbol{\beta}$ transforms the model matrix to $\mathbf{Q}$ and the penalty matrix to $\mathbf{R}^{-T}\mathbf{S}\mathbf{R}^{-1}$.

The penalty matrix is then eigen-decomposed as $\mathbf{R}^{-T}\mathbf{S}\mathbf{R}^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{U}$ is orthogonal and $\mathbf{D}$ is diagonal. Further re-parameterization via a rotation/reflection of the parameter space yields parameters $\boldsymbol{\beta'} = \mathbf{U}^T\boldsymbol{\beta''}$, resulting in a model matrix $\mathbf{QU}$ and penalty matrix $\mathbf{D}$. This "natural" parameterization allows for a clear understanding of the penalty's role in limiting parameter variance.

Each unpenalized coefficient has one degree of freedom. The penalized estimates are shrunken versions of the unpenalized estimates: $\hat{\beta'}_i = (1 + \lambda D_{ii})^{-1}\tilde{\beta'}_i$, where D_{ii} are the eigenvalues of the penalty matrix. The shrinkage factor $(1 + \lambda D_{ii})^{-1}$ represents the *effective degrees of freedom* for each penalized coefficient.

The total EDF for the smooth is the sum of the individual shrinkage factors:

$$\sum_i (1 + \lambda D_{ii})^{-1} = \mathrm{tr}(\tau) \quad \text{where} \quad \tau = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T\mathbf{X}$$

τ can be interpreted as the matrix that maps the un-penalized coefficient estimates to the penalized coefficient estimates. This means that the trace of τ can be understood as having the effect of being the averge shrinkage of the coefficients, multipled by the number of coefficients. This measure is bounded between the number of zero eigenvalues of the penalty (as $\lambda \to \infty$) and the total number of coefficients (when $\lambda = 0$).

The unpenalized estimators are unbiased, leading to the expected value of the penalized estimates: $E(\hat{\beta'}_i) = (1 + \lambda D_{ii})^{-1}\beta_i$. The shrinkage factors determine the relative smoothing bias.

The penalty suppresses variability in parameters corresponding to high eigenvalues D_{ii}, effectively reducing model complexity.

Therefore, the AIC formula corrected to incorporate the effective degrees of freedom is

$$\text{AIC} = -2\ell(\hat{\beta}) + 2\tau. \tag{3.3}$$

This is the formula used to compare models using different smoothing basis fitted to a given data set in chapter 4.

Chapter 4

Results

This chapter presents examples demonstrating the application of semi-mechanistic models in infectious disease modeling. These examples illustrate the inferential capabilities of semi-mechanistic models. We begin with a simulated data example to showcase the effectiveness of fitting semi-mechanistic models with a single unknown function. These structure of the compartmental models are intentionally kept simple to demonstrate that estimating unknown functions using penalized smoothers is relatively straightforward when integrated into the syntax and optimization engine of `macpan2`. The objective is to illustrate this approach so that modelers can utilize it without extensive knowledge of spline and nonlinear optimization literature. Following this, we provide three real-world examples: an epidemic of Scarlet Fever in Ontario from 1929 to 1931, the initial SARS-CoV-19 outbreak in Ireland at the start of the pandemic and 5 decades of measles in London UK.

4.1 SIRS model with simulated data

The challenge is that for a given epidemiological dataset containing incidence or prevalence observations, the transmission rate is not observed. This means the true shape of the unknown function used to estimate the transmission rate is also unknown. We aim to develop a compartmental model, formulated as a deterministic system of ordinary differential equations with a linear smoother component. If this model can predict the population dynamics from knowledge of the time-varying transmission rate, then it should be possible to infer the transmission rate by fitting the model to data. This allows us to prove the efficacy of the smoothing parameter estimation methodology and the use of these methods with models formulated as discrete systems of ordinary differential equations.

Consider the time-varying transmission rate β as a latent variable. 'Time-varying' means that at each observation x_i in the dataset, the model—comprising a system of non-stochastic ordinary differential equations—contains an estimated value for the transmission rate for each $1 \leq i \leq n$, where n is the number of observations.

β is constructed as the linear smoother $b_0 + \mathbf{X}b$, where b_0 is the intercept, $\mathbf{X}$ is the model matrix associated with the particular smoother, and b is the vector of smoothing coefficients. Both b_0 and b are parameters that need to be estimated. This linear smoother, which estimates β, is a non-parametric component inside a deterministic system of differential equations, whose trajectory is determined entirely by the starting parameters. This is why the model is called semi-mechanistic or partially specified. Some components of the model contain unknown functions, while the rest of the model comprises conventional elements with only unknown parameters.

To prove that semi-mechanistic models, within the `macpan2` modeling framework, are a capable methodology for this problem, we conduct a simulation study. For a chosen smooth, we simulate data and add Gaussian noise to the incidence. The simulated data is used to calibrate the model using a variety of the smoothers available in the `mgcv` package.

Consider a SIRS compartmental model with a fixed waning immunity parameter ϕ. The total number of individuals in the population is given by N. The initial values of S, I, and R are fixed at the endemic equilibrium solutions.

We construct the data generating model as follows. The smoother type and the number of knots k are specified using a particular `mgcv` smooth. This determines the form of $\mathbf{X}$ and the penalty matrix $\mathbf{P}$. The smoothing coefficients b, of dimension $k-2$, are assumed to be multivariate Gaussian. Thus, b is initialized as random normal deviates at the $k-2$ evenly spaced quantiles with a mean of 0 and a standard deviation specified as b_{sd}. As the variation of this distribution increases, the true function describing the time-varying transmission rate for the data model will become more complex. An initial value for b_0 is chosen as the log of the initial value of β. The recovery rate γ is fixed. Initial values for the remaining parameter estimates are set to 1. See subsection 3.4 for more details on initializing compartment values.

The model trajectory is simulated from these initial conditions using Euler steps. Gaussian noise ($sd = 800$) is added to the simulated incidence vector. This vector is used to test the efficacy of the semi-mechanistic models. The models are calibrated using `macpan2`, which utilizes the Laplace approximation via `TMB` to optimize the objective function and update parameter estimates using quasi-Newton methods.

The calibration model is constructed it the same way, except that we can choose to vary the smoother and the model granularity by specifying the smoother type and the number of knots. In figure 4.1 we illustrate the shape of the basis functions for each of the univariate, non-cyclic smoothing basis used in this methodology.

It is worth mentioning that we have kept γ fixed in the simulation study, but in the real-life data examples, we have used a log-normal prior to allow flexible deviation from the initialized starting value of γ. The reason for this is that we were not able to get this to work owing to a possible error in the use of `macpan2` syntax.

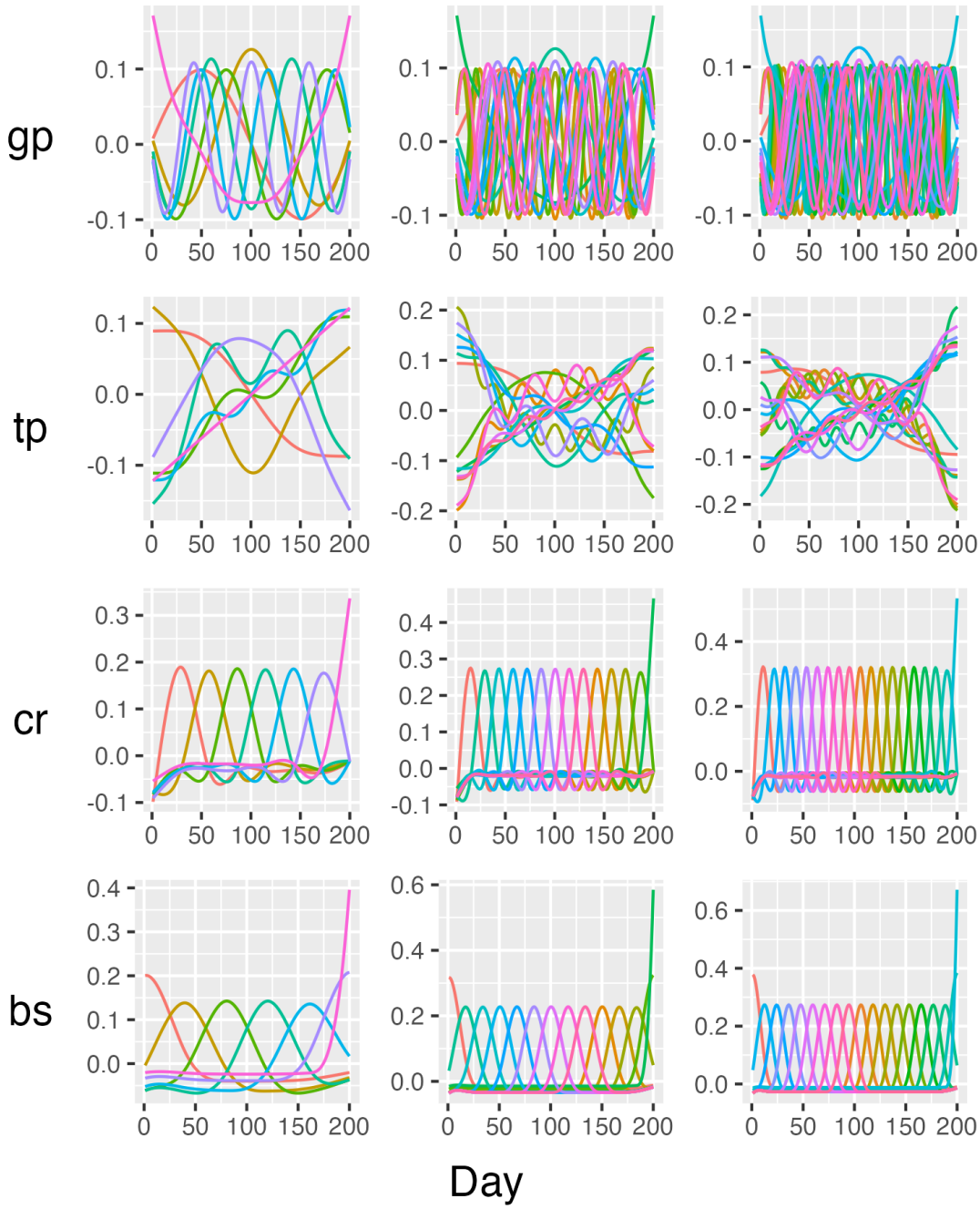

FIGURE 4.1: **Basis functions for calibrating smoothers**. Basis matrices are obtained via mgcv::smoothCon() with number of knots $k = 8$ and domain of $n = 200$ days. Each basis matrix is determined only by the input domain and the number of knots. Each basis function is a single column of the basis matrix. The acronyms for the smoothers are defined as follows: "gp" = Gaussian process, "tp" = thin plate regression spline, "cr" = cubic regression spline and "bs" = B-spline. Not included in this figure are the acronyms "ps" = P-spline, "ts" = cyclic thin plate regression spline, "cc" = cyclic cubic regression spline, and "cp" = cyclic P-spline.

The calibration models vary the number of knots ($k = 8, 15, 20$) while keeping the variance $b_{sd} = 2$ fixed for starting values. $\gamma = 1/14$, which represents a two-week period of an infectious individual being capable of transmitting the disease to a susceptible individual. $\phi = \frac{1}{300}$, which means that at each time point, that proportion of the recovered individuals lose immunity. An equivalent interpretation of the waning immunity parameter is that it represents the period of immunity for an individual, which in this case represents 300 days.

In Figures 4.2, 4.3, and 4.4, we present the predicted incidence, transmission rate, and effective reproduction number for data simulated using a Gaussian process (GP) smoother with $k = 20$ knots. We have used the univariate non-cyclic smoothers in the calibration model: Gaussian process regression smoothers, thin plate regression splines, B-splines, and cubic regression splines. We have chosen to omit the results of the calibration models utilizing a P-spline basis as they show poor ability to both predict incidence and infer the underlying true transmission rate when the number of knots used is less than the number used to generate the data.

Figure 4.2 shows the predicted incidence for the best models, for each smoothing type, fitted to simulated data with added noise. As we move from left to right across each figure, we can observe the effect of increasing model granularity by increasing the number of basis functions, or knots, used by each smoother in constructing the unknown function, β. In the leftmost and center columns, we see two degrees of underfitting, where the model smooths over the peaks and valleys. In the rightmost column, the model fully matches the incidence trajectory.

We observe the same phenomenon in the estimated transmission rate and effective reproduction number in Figures 4.3 and 4.4. In the rightmost column of these figures, for all bases, the model is able to recover the true shape of the unknown functions. However, it is only the Gaussian process basis that is capable of reasonably approximating the true shape of the unknown functions with low model granularity.

In Table 4.1, we present the comparison of the conditional AIC scores for the models corresponding to Figures 4.2, 4.3, and 4.4. We observe that the Gaussian process (GP) model consistently has the lowest AIC score across all evaluated models and levels of model granularity.

The AIC score, which balances model fit and complexity, suggests that the GP model provides the best trade-off between these two aspects among the models compared. A lower AIC score indicates that the GP model is relatively the best fit for the data, taking into account the effect of penalization on the number of knots used, i.e., effective degrees of freedom. Thus, the GP model is the most likely to minimize information loss when approximating the true data-generating process, compared to the other models tested.

We also observed that when the data-generating model used a thin plate regression spline as the basis for the linear smoother, the GP model continued to have the lowest AIC score, even compared to the model using a thin plate regression spline basis. Notably, the AIC score heavily penalizes model complexity for the B-spline and cubic

regression spline bases. For the GP and thin plate regression spline bases, the difference in AIC scores between models with lower and higher complexity increases by a negligible amount.

In the appendix, Figures B.1, B.2, and B.3, we simulated data using the cyclic cubic regression (CC) basis and fitted the data using models with cyclic bases. This demonstrated that the cyclic bases were capable of fitting cyclic data. Notably, the CC basis, unlike the TS or CP bases, was effective in fitting non-cyclic data. Not shown, the GP basis is able to also fit to the cyclic data well, while the other non-cyclic basis do not.

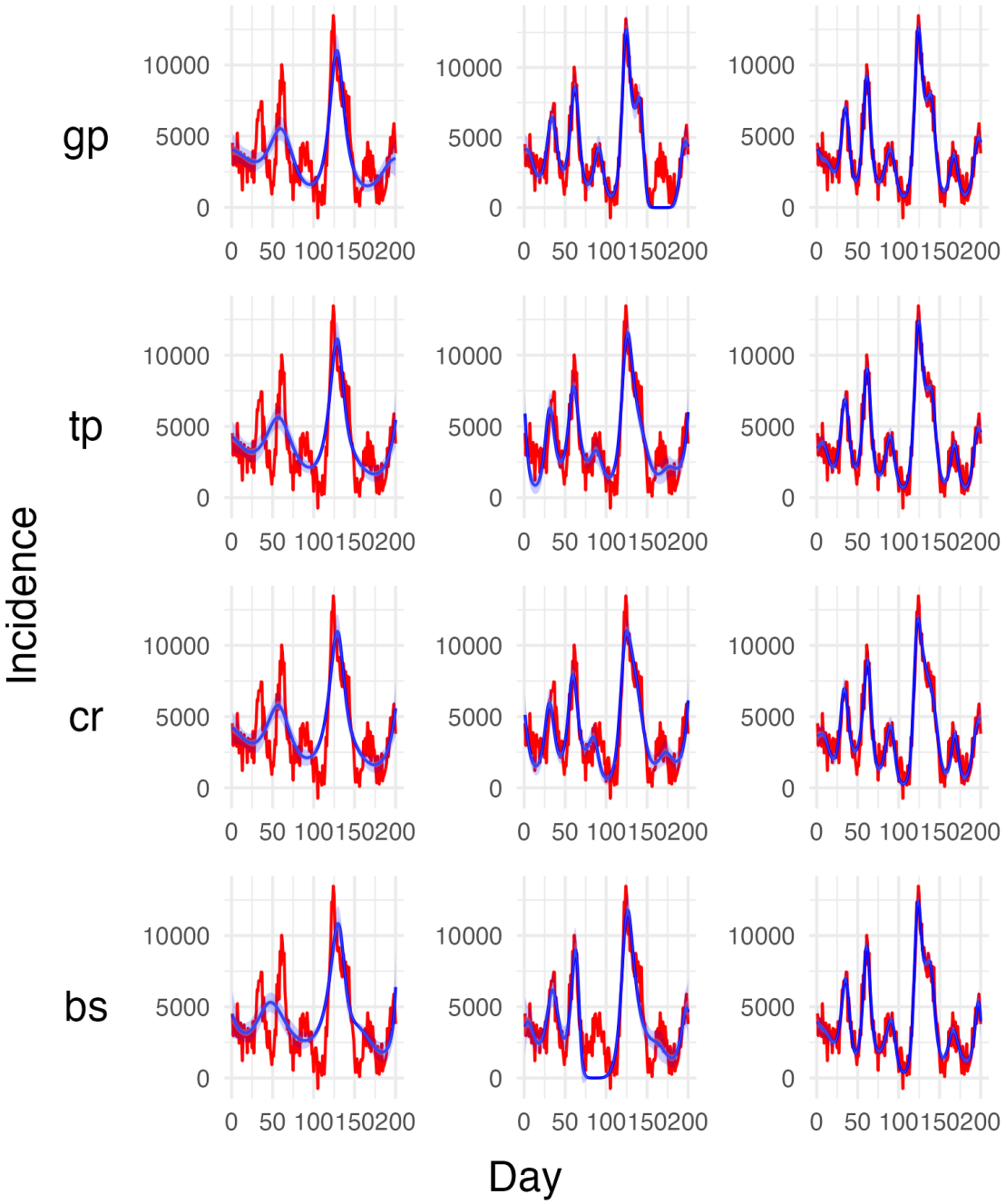

FIGURE 4.2: **Estimated Number of New Infections per day (Incidence) for data simulated using a Gaussian process (GP) regression smoother.** The first column is calibrated using $k = 8$ knots, the second with $k = 15$ knots, and the third with $k = 20$ knots. The red line represents the simulated trajectory with Gaussian noise. The blue line indicates the predicted incidence, with light and dark blue bands representing the 95% and 50% confidence intervals.

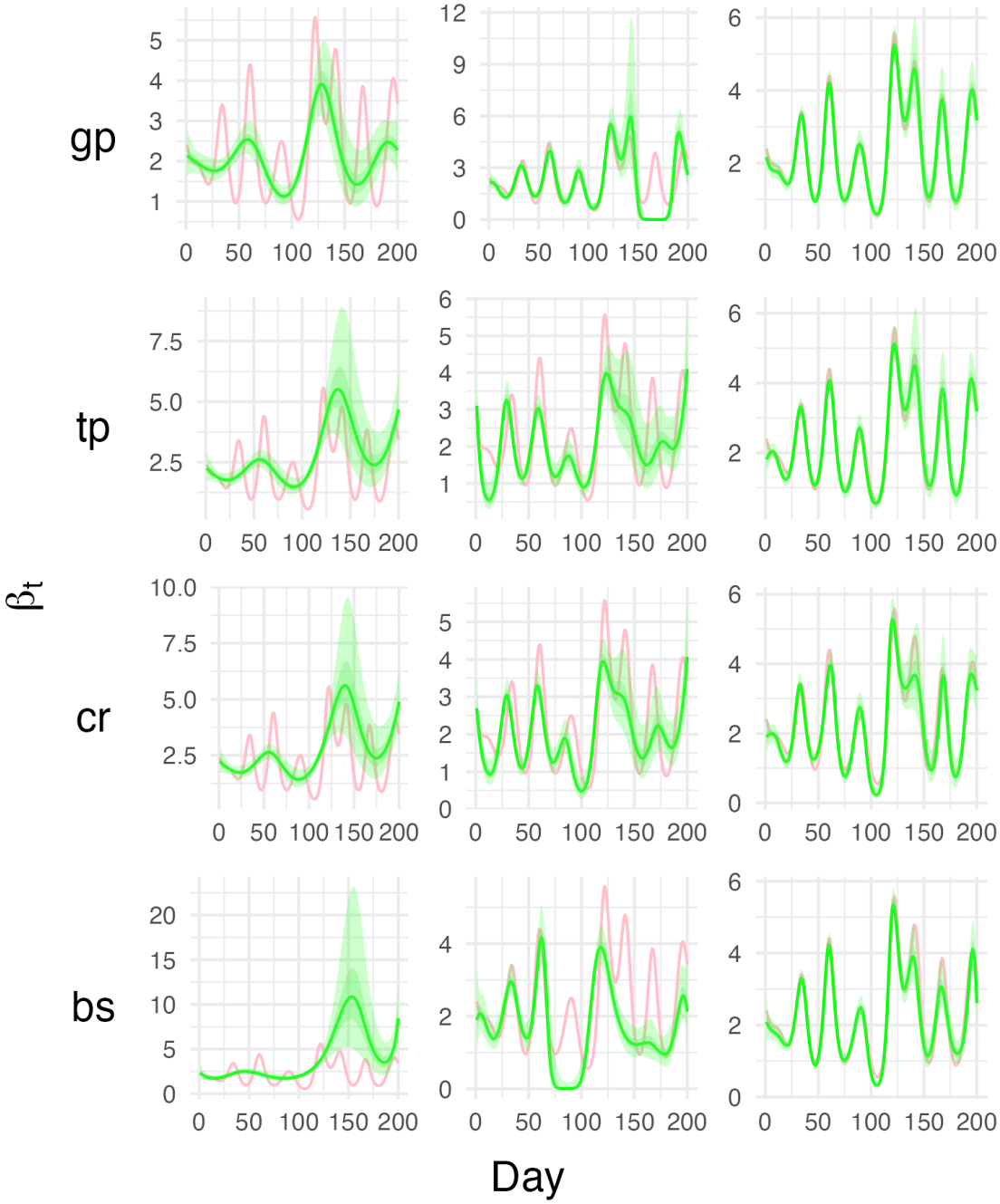

FIGURE 4.3: **Estimated transmission rate β per day for data simulated using a Gaussian process (GP) regression smoother.** The first column is calibrated using $k = 8$ knots, the second with $k = 15$ knots, and the third with $k = 20$ knots. The red line represents the true transmission rate function from the simulated data. The green line indicates the estimated transmission rate, with light and dark green bands representing the 95% and 50% confidence intervals.

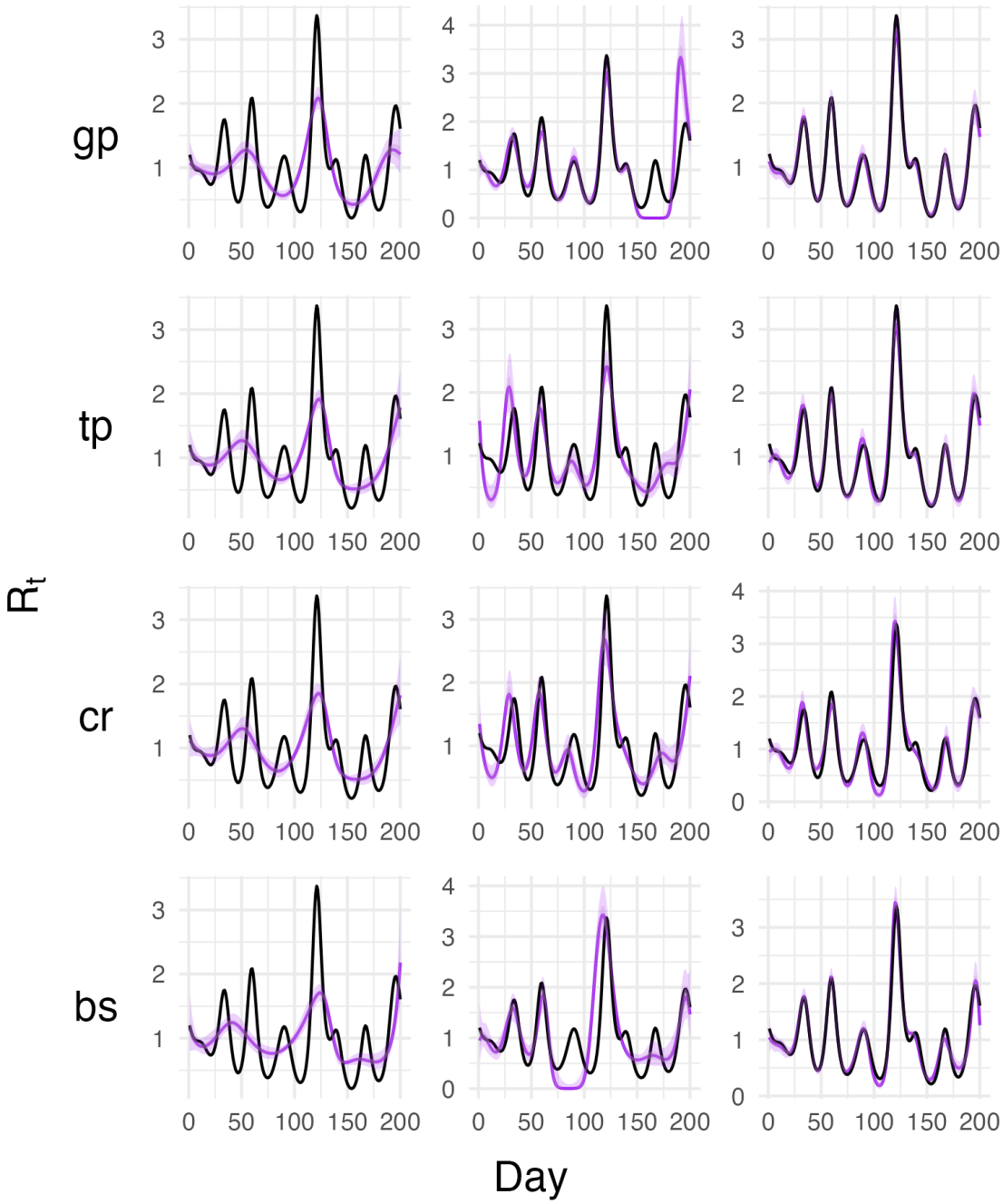

FIGURE 4.4: **Estimated effective reproduction R_t per day for data simulated using a Gaussian process (GP) regression smoother.** The first column is calibrated using $k = 8$ knots, the second with $k = 15$ knots, and the third with $k = 20$ knots. The black line represents the reproduction number from the simulated data. The purple line indicates the estimated reproduction number, with light and dark purple bands representing the 95% and 50% confidence intervals.

TABLE 4.1: **Conditional AIC Scores of calibrating models with varying model granularity, fitted to data simulated using a GP smoother with $k = 20$ knots.** The degrees of freedom are defined as the model degrees of freedom, which is computed as the trace of penalized smoothing matrix.

	Knots		
Smooth Type	k = 8	k = 15	k = 20
gp	-14.96674	-14.86925	-14.74068
tp	-12.95091	-12.81223	-12.60562
cr	-12.25615	-9.445058	-7.025796
bs	-12.28378	-10.16051	-6.892706

4.2 Scarlet Fever in Ontario 1929-1931

This example showcases the varying efficacy of the smoother bases from `mgcv`. We consider an outbreak of scarlet fever in Ontario from 1929 to 1930. The dataset comes from the International Infectious Disease Data Archives (iidda) [20].

The calibration model is constructed using the basic SIR model, with a starting recovery rate of $\gamma = 1/7$.

We present the results of the best models using the Ornstein-Uhlenbeck (OU), thin plate regression spline (TPRS), cubic regression spline (CR), and cyclic cubic regression spline (CC) bases for the linear smoother component. The other bases, as in Figure 4.1, did not perform well enough to be included; their incidence predictions were inaccurate, leading to inflated uncertainty estimates of the transmission rate, or the optimization did not converge.

In Figures 4.5, 4.6, and 4.7, we show the predicted incidence, transmission rate, and effective reproduction number for the best models. We varied the model granularity by increasing the number of knots. The third column of all three figures, where the number of knots $k = 10$, shows overfitting, introducing unnecessary inflection points into the estimates. The leftmost column, calibrated with $k = 5$ knots, shows a smooth, simple fit to the data and estimates.

As model granularity increases, the TPRS and CR bases perform poorly, indicated by the apparent width of the uncertainty estimates compared to the GP and CC bases. This is reflected in the conditional AIC scores of the nine models in Table 4.2. Notice that the AIC score for the GP basis increases very little as model granularity increases. In contrast, the CC, TPRS, and CR bases show a significant increase in AIC score with higher model granularity. This behavior, observed in the simulation study, leads to the hypothesis that the GP prior, of which the OU basis is a special case, is generally the most robust of the univariate smoothing bases available in the `mgcv` package for this type of model.

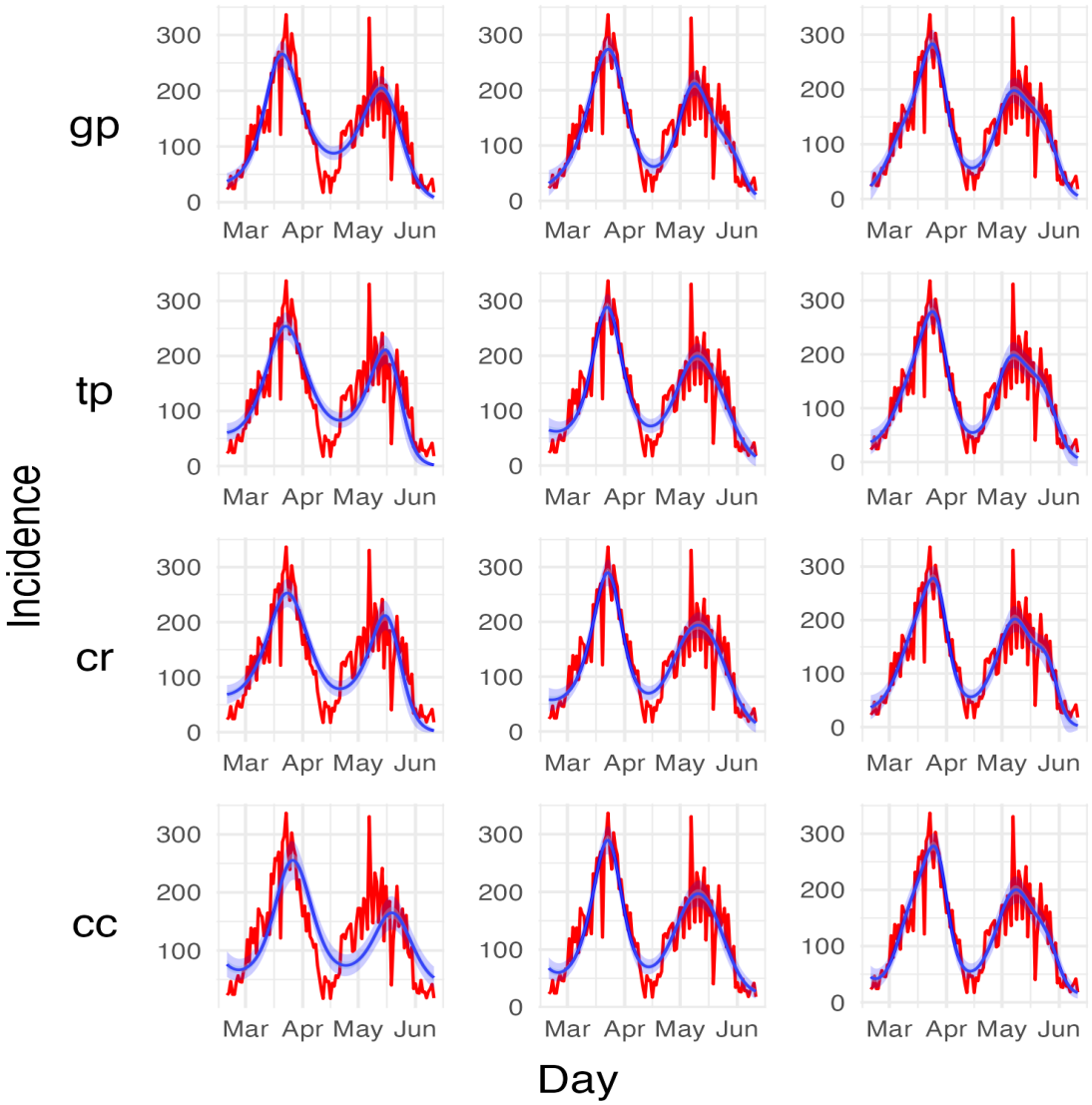

FIGURE 4.5: **Predicted incidence for weekly observed scarlet fever cases in Ontario, from 1929 to 1930.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 8, 10$. The covariance function of the GP basis uses a power exponential kernel, with power parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

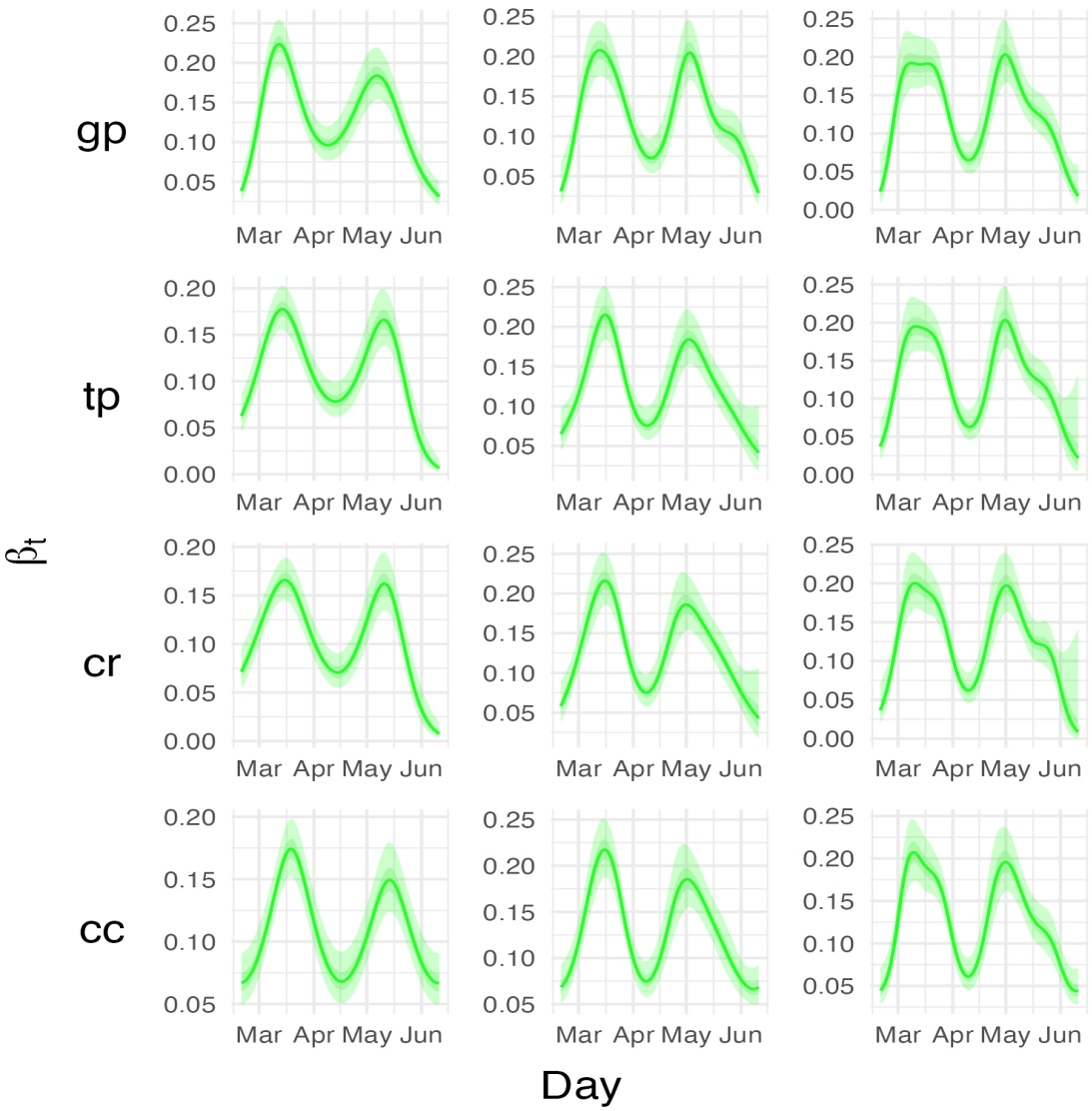

FIGURE 4.6: **Estimated transmission rate for weekly observed scarlet fever cases in Ontario, from 1929 to 1930.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 8, 10$. The covariance function of the GP basis uses a power exponential kernel, with power parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

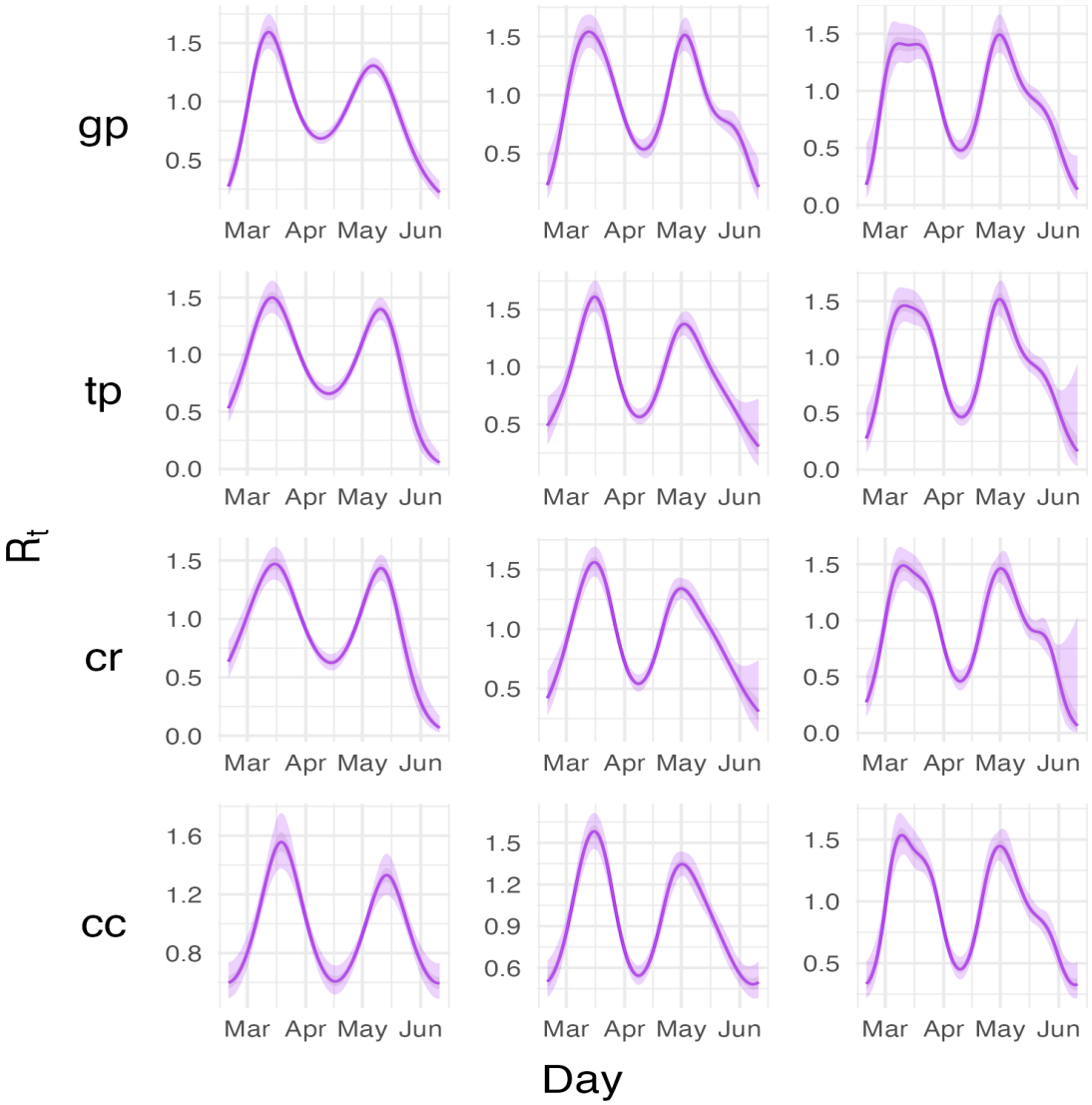

FIGURE 4.7: **Estimated effective reproduction number for weekly observed scarlet cases in Ontario, from 1929 to 1930.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 8, 10$. The covariance function of the GP basis uses a power exponential kernel, with power parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

4.3 Covid-19 Ireland 2020

Next, we fit to observations of the daily number of COVID-19 cases in Ireland from the onset of the outbreak, spanning February 20, 2020, to May 9, 2020. This data is sourced

TABLE 4.2: **Conditional AIC Scores of calibrating models with varying model granularity, calibrated to Ontario scarlet fever (1929-1930.** The degrees of freedom are defined as the model degrees of freedom which is computed as the trace of penalized smoothing matrix.

	Knots		
Smooth Type	k = 5	k = 8	k = 10
gp	-12.75613	-12.69357	-12.66127
tp	-10.8087	-10.65266	-10.56183
cr	-10.74505	-9.142717	-7.634613
cc	-12.84522	-11.9573	-10.65248

from the publication by Andrade and Duggan [21].

The calibration model is constructed using the basic SIR model, with an initial recovery rate of $\gamma = 1/14$.

Figures 4.8, 4.9, and 4.10 present the results of the best models, which utilize the GP, CC, TPRS and CR basis functions. The BS and PS basis functions exhibited very large uncertainty estimates, while the CP and TS basis functions showed both larger uncertainty estimates and lower conditional AIC scores than the CC basis. We chose not to include the latter basis. To illustrate again the concept of overfitting, model granularity is varied.

All basis displayed excellent fit to the observed data with appropriately sized uncertainty estimates.

Table 4.3 provides the conditional AIC scores. The CC and GP smoothers perform the best, although the AIC score for the GP basis does not increase as dramatically as for the CC basis. The TP basis demonstrates resilience to overfitting as model granularity increases but still maintains a higher AIC score compared to the GP and CC bases. The CR basis performs the worst.

Notably, there is a difference in the functional shape of the transmission rate for the CC and GP bases versus the CR and TP bases. The former display a near-zero transmission rate at time zero with a sharp peak and large amplitude. In contrast, the latter have an estimated transmission rate of approximately 0.2 at time zero and a more rounded peak with a smaller amplitude. Additionally, the uncertainty bounds are much larger at the beginning of the curve for the CR and TP bases, even with appropriate model granularity ($k = 5$). The uncertainty estimates for the CR and TP bases increase significantly near the origin as model granularity rises. Although the GP basis also shows this trend, it is not as pronounced as for the CP and TP bases. Surprisingly, despite the CC basis displaying overfitting (as indicated by the AIC score) when model granularity increases, it visually retains the tightest uncertainty estimates among all the smoothers. This behavior is also observed in the estimates for the effective reproduction number.

The model's sensitivity to changes in the starting value of γ (ranging from 7 to 14 days) does not alter the shape of the functional form of β, but it changes the amplitude of the peak, leading to an increase in R_t. Increasing the variance of the log-normal prior on γ causes the model to tend to select a larger γ. However, this also results in a dramatic increase in uncertainty estimates. We observed that when the starting value of γ is lowered and the variance of the log-normal prior is increased simultaneously, the model still predicts larger values of γ.

Andrade and Duggan (2022) used COVID-19 data to infer the effective reproduction number by employing an SEIR model with three different data-generating processes for the transmission rate. One of the processes they used was Geometric Brownian Motion (GBM). GBM is similar to the Ornstein-Uhlenbeck (OU) process that we used, but it is non-stationary, non-mean-reverting, and the response is log-normally distributed. In our thesis, we implemented the OU process as a special case of a Gaussian Process (GP) with an exponential covariance function, which differs from defining an OU process as a stochastic differential equation (SDE). By viewing the OU process as a GP, we focus on the joint distribution of values at different times, characterized by the mean and covariance functions. These functions describe the decay rate of the covariance and the process variance, respectively, emphasizing the correlation structure. Computationally, this approach allows us to use linear algebra, while the SDE formulation involves solving differential equations. Andrade and Duggan define the transmission rate using GBM formulated as an SDE. They also assume that the response, the incidence data, follows Poisson and Negative Binomial distributions. Additionally, they use Apple mobility data to adjust the transmission rate by assuming that the effect of social distancing is correlated with the transmission rate.

We compared our results, using the OU process basis for our linear smoother, with those of Andrade and Duggan. They aggregated their incidence data to a weekly scale to account for irregularities in daily reporting. We implemented this by computing the trajectory on a daily scale, aggregating the predicted incidence to a weekly scale, and then fitting the model to the weekly aggregated observed data. The shape of the inferred transmission rate and effective reproduction number were as expected, but the magnitude was lower than anticipated. We hypothesize that this is due to the use of too large Euler steps, on the order of one week. Therefore, we can only compare the results of our fit using daily observations to their fit using the aggregated weekly data.

In Figure 4.8, the predicted incidence fits the data well, with very reasonable uncertainty bounds, and the shape is very similar to Andrade and Duggan's results. In Figure 4.9, we present the estimated transmission rates. Andrade and Duggan did not include the rates at time zero in their plots, but we have. Ignoring the first week in our plots, both results show approximately the same shape, an exponential decay. The magnitude of the transmission rate is about twice as large for Andrade and Duggan. This could be attributed to their use of an SEIR model, which partitions the susceptible population by those exposed to the disease, informed by the mobility data. The difference could be explained by the fact that the disease would need to be more infectious to infect the same

number of people in a portion of the total susceptible population. Our SIR model does not make any such assumptions. The estimated effective reproduction number, which is the transmission rate scaled by the recovery rate and the proportion of susceptibles at any time t, also shows a similar pattern. Although Andrade and Duggan compute an analytical expression for the basic reproduction number and we compute ours as the transmission rate scaled by the recovery rate, we both scale the basic reproduction number by the ratio $\frac{S_t}{N_t}$ to obtain the effective reproduction number. Most notably, in Figure 4.10, we observe that the estimated effective reproduction number has not only the same shape as Andrade and Duggan's but also a similar shape and magnitude. The uncertainty estimates are tighter in our model, but our model is simpler.

TABLE 4.3: **Conditional AIC Scores of calibrating models with varying model granularity, calibrated to Ireland Covid-19 (2020).** The degrees of freedom are defined as the model degrees of freedom which is computed as the trace of penalized smoothing matrix.

	Knots		
Smooth Type	k = 5	k =6	k = 7
gp	-12.38633	-12.3596	-12.36127
cc	-12.39882	-12.27432	-11.83989
cr	-10.10928	-9.864844	-8.977689
tp	-10.39501	-10.34895	-10.31064

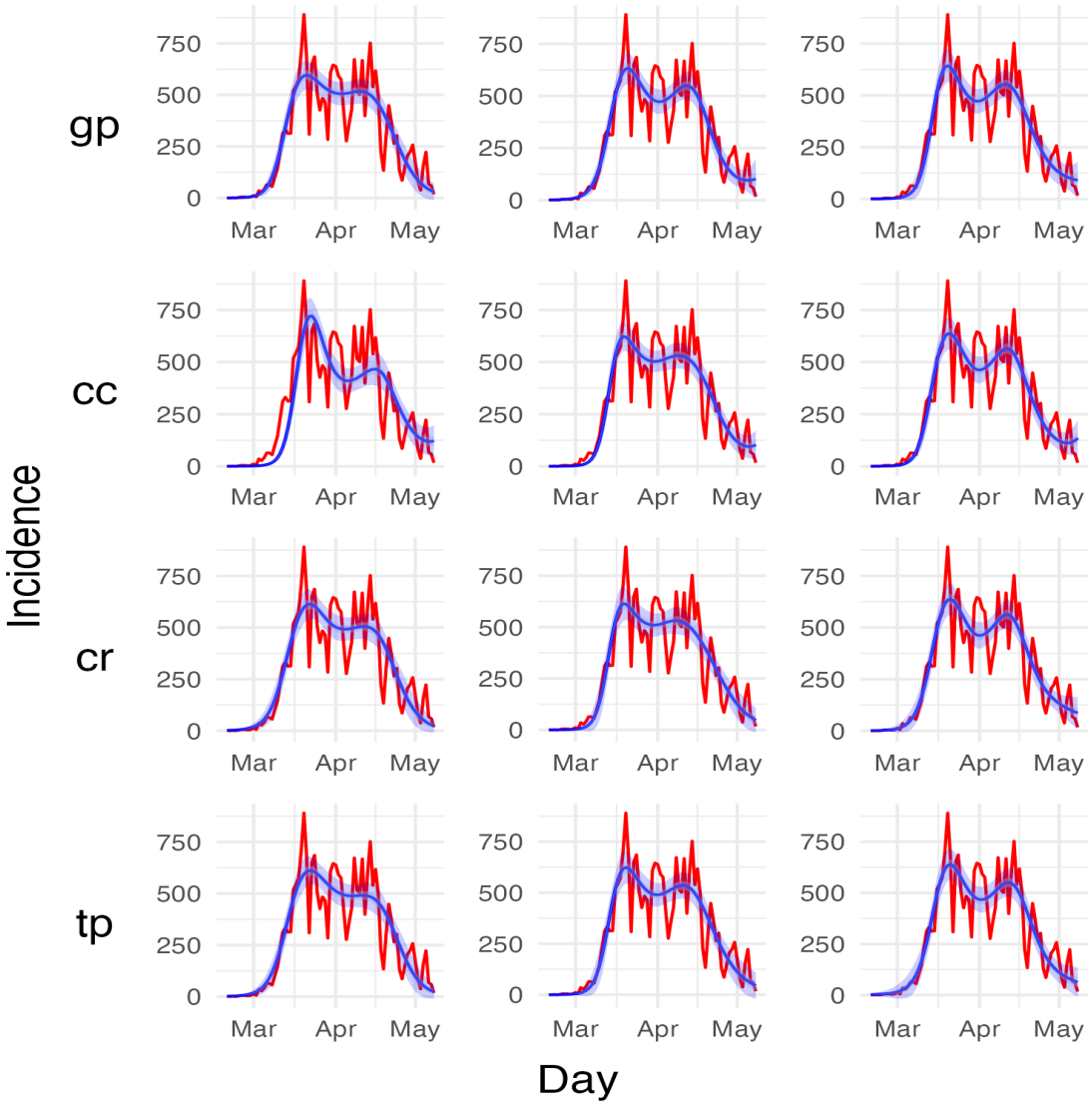

FIGURE 4.8: **Predicted incidence for weekly observed measles cases in Ireland 2020.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 6, 7$. The covariance function of the GP basis uses an exponential kernel function, with exponent parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

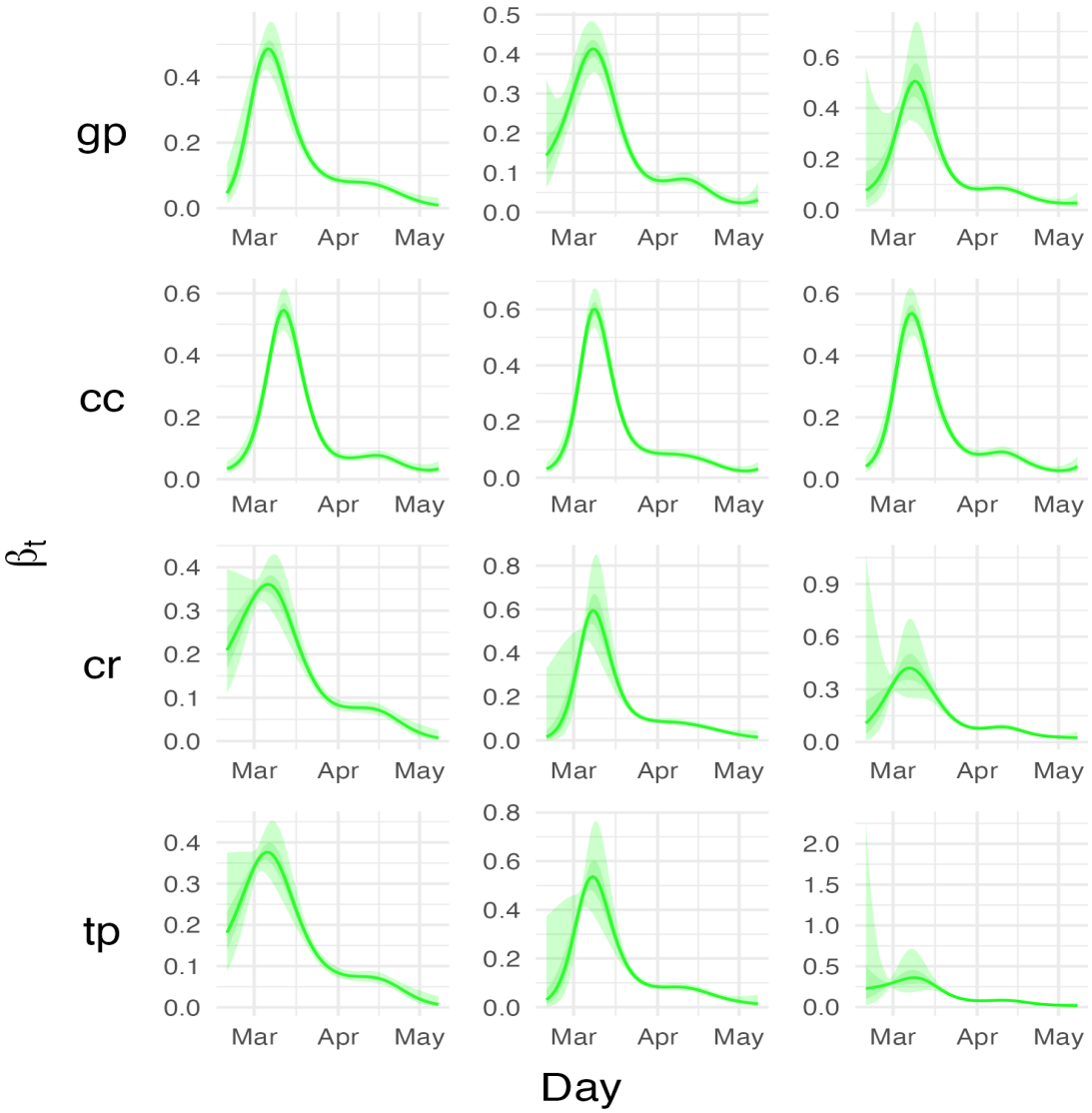

FIGURE 4.9: **Estimated transmission rate for weekly observed measles cases in Ireland 2020.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 6, 7$. The covariance function of the GP basis uses an exponential kernel function, with exponent parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

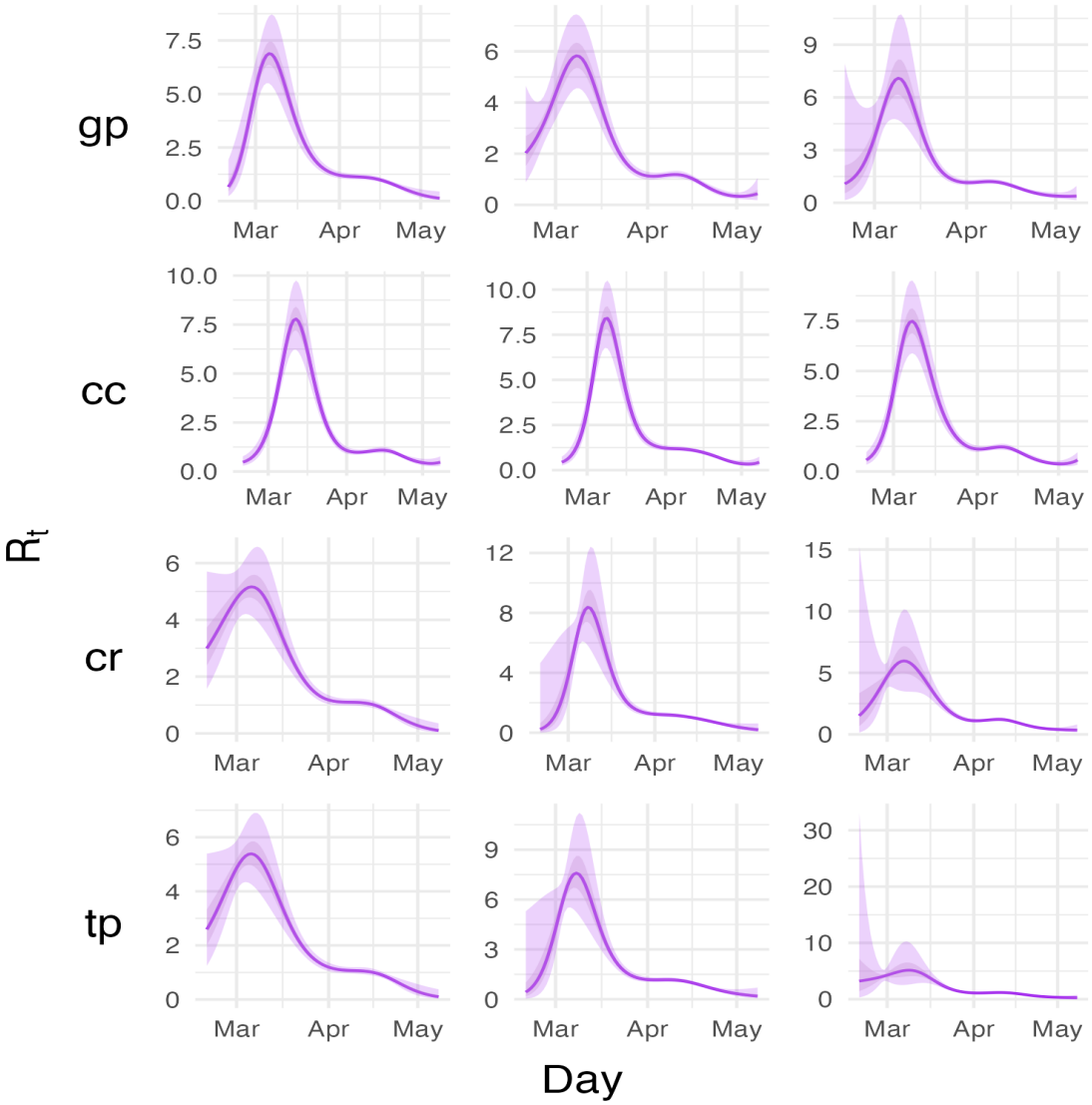

FIGURE 4.10: **Estimated effective reproduction number for weekly observed measles cases in Ireland 2020.** The linear smoother component of the semi-mechanistic compartmental model has model granularity increasing from right to left $k = 5, 6, 7$. The covariance function of the GP basis uses an exponential kernel function, with exponent parameter $\kappa = 1$ and range parameter $\ell = 2$. The figures display 95% and 50% confidence intervals.

4.4 Measles London UK 1944-1984

We now present a more challenging problem than the previous examples. We consider weekly observed measles cases in London, UK, from 1944 to 1984. This dataset was first utilized in a publication by David Earn et al. [22].

The calibration model is constructed using the basic SIR model. The starting value for the recovery rate is $\gamma = 1/8$, aligning with the initial value used in [22]. A key assumption we make is that the total population remains constant over time. This is due to the nature of the SIR model, which does not account for a time-varying total population component. We fixed the total population to $N = 8,000,000$, which approximates the population of London at the start of this dataset, and then rounded down. This assumption is reasonable, as the population of London has fluctuated between six and ten million from then until the present day. However, since the disease predominantly affects children, a more sophisticated model could include a time-varying total population component, partitioned according to age and weighted by the incidence rate per age demographic. We initialize the number of infected individuals at $I_0 = 250$.

The full dataset extends to 1984, but we had difficulty tuning the model to fit the last decade. During this period, the observed incidence was relatively flat compared to the previous years. We observed that the calibrated model inflated the predicted transmission rate to unreasonable levels. By truncating the last decade, reducing the observations from 2,660 to 2,140, we were able to calibrate the model more effectively.

We exclusively present the results for a Gaussian process basis for the linear smoother component of the model. Other smoothing bases were tested, but the optimizer failed to converge for more than 100 to 200 observations. In Figure 4.11, we show the predicted incidence, transmission rate, and effective reproduction number for the SIR model with a Gaussian process basis for the linear smoother component. The covariance function is the exponential kernel with a range parameter $\ell = 20$. Other kernels, such as the Matérn function, were tested by iterating over different range parameters $\ell = 30, 40, 50$. The resulting models showed little difference in optimized parameter values and conditional AIC scores. For simplicity, we present the parsimonious model with the simplest kernel function and the smallest range parameter. Each model took about 20 minutes to fit.

Figure 4.11 displays the optimized values of the parameters in the calibrated model and their uncertainty measurements. These are the log transformed values. Exponentiating, the partial prior on γ produced an optimal value of $\frac{1}{11.84}$.

During the period from 1950 to 1968, London experienced biennial cycles of measles epidemics. In Figure 4.11, these epidemic cycles are evident as peaks in the incidence observations. There are nine cycles of epidemics, corresponding to nine cycles of the transmission rate, indicating that our model effectively captures the seasonality of the transmission rate.

Earn et al. (2000) employed an SEIR model with seasonal forcing to explain the transitions in measles dynamics. They estimated the mean transmission rate (β), which is the product of the recovery rate and the empirically measured basic reproduction number (R_0). This approach provides a dimensionless parameter that incorporates both the contact rate and the probability of transmission per contact.

In contrast, our model estimates a time-varying β, representing the probability of transmission per contact per unit time. This more intuitive representation allows for

direct fitting to the observed data. However, it complicates direct comparisons with the high β values used by Earn et al., which are scaled differently.

Despite this difference, our model successfully fits the large dataset and captures the seasonal dynamics of measles transmission. This is notable given the lower magnitude of the effective reproduction number compared to empirically calculated results. One possible explanation for this discrepancy is the use of a simpler SIR model in our study, which does not account for factors like vaccination coverage and population changes over time.

To improve the alignment of our results with empirical findings, future work should consider using a more sophisticated model akin to the SEIR framework used by Earn et al.

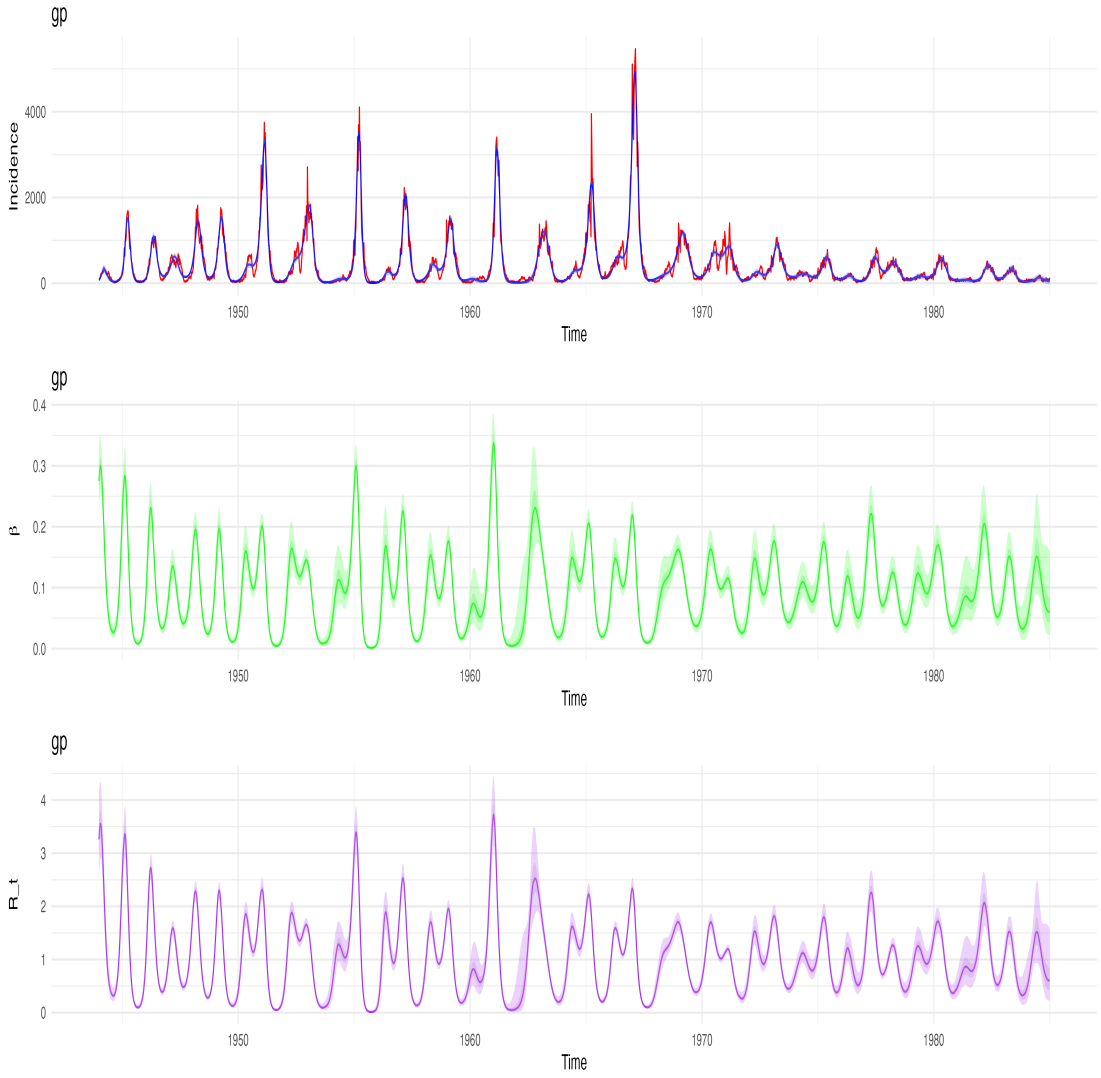

FIGURE 4.11: **Predicted incidence and estimated transmission rate and effective reproduction number for weekly observed measles cases in London, UK, from 1944 to 1984 using a Gaussian Process smoother.** The linear smoother component of the semi-mechanistic compartmental model employs an Ornstein-Uhlenbeck (OU) process with a range parameter $\ell = 20$ and $k = 100$ knots. The model estimates a recovery rate of approximately $\gamma = 11.84$ days and an initial number of infected individuals of approximately $I_0 = 250$. The figures display 95% and 50% confidence intervals.

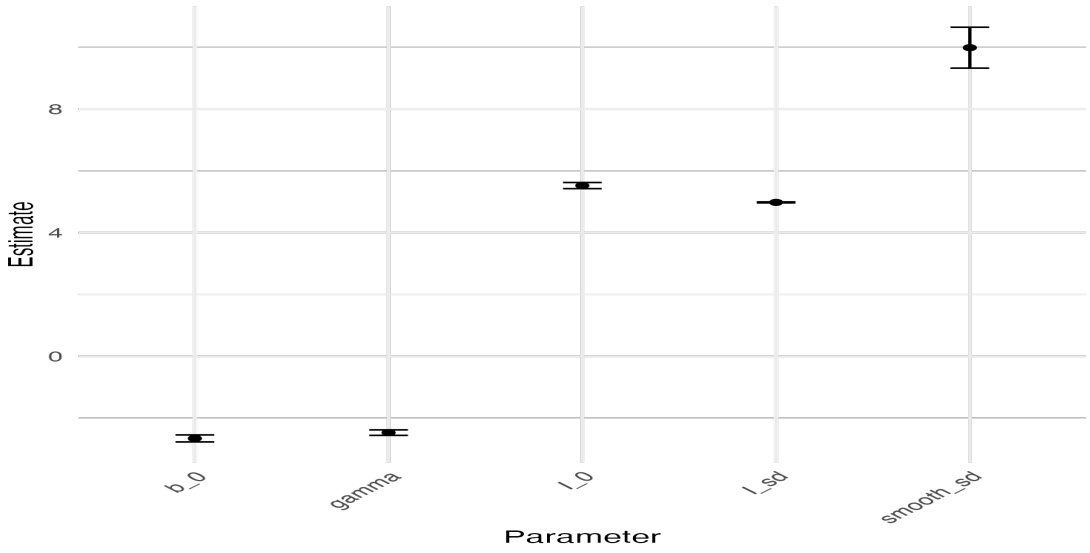

FIGURE 4.12: **Coefficients plot of the optimized parameters (log-transformed) for the model with a GP basis calibrated to the UK London Measles (1944-1984) dataset.** Uncertainty estimates are obtained as the wald confidence intervals computed using the Delta method.

Chapter 5

Discussion

The approach in this thesis outlines a way to easily formulate infectious disease compartmental models without having to make unjustified biological assumptions about the underlying disease transmission process in the way that using fixed or parametric models of the transmission rate do. Integrating this into the `macpan2` and `TMB` framework creates for the modeler a user friendly way to fit the model, select the best model and infer estimates of the latent variable at each point in time in the domain. We accomplished this by adapting the general methodology of Simon Wood [2] and utilizing it in conjunction with the latter mentioned model formulation tool and optimization engine.

What we have done differently then the work of Simon Wood is to have made this methodology accessible for a quantitatively minded modeler who may not be an expert in non-linear optimization and smoothing theory. This was made possible because `macpan2` is a driven from a software-engineering perspective. This means that `macpan2` wraps all of the `C++` code needed to interact with `TMB` in an `R` wrapper. It negates the requirement for the modeler to write bespoke optimization code, which can be the biggest hurdle to using such models.

Through simulation studies, we demonstrated the efficacy of penalized smoothing parameter estimation. The `mgcv` package greatly facilitated the construction of low-rank smoothing bases and penalty matrices for a given domain and model granularity. We compared different smoothing bases, evaluating their performance based on uncertainty estimates, AIC scores, and the shapes of the estimated functions.

The performance of the semi-mechanistic models to infer the time varying parameters was quite good. We were able to fit the models to the incidence data from the real world examples with the goodness of fit controlled by specifying model granularity and the smoothness by inferring the optimization parameter. We believe these results show that this methodology can be beneficial to any modeler that wishes to estimate a time varying latent variable by inferring the shape of the unknown function. Although we focused on modelling infectious disease with unknown transmission rate and reproduction number, it seems reasonable that it can be applied to an arbitrary compartmental model with some unknown time varying function that one wishes to estimate. This can even be applied to the more general case of dynamic ecological models formulated using compartmental models.

The models were effective in regards to inference of the fixed parameters. The recovery rate and the initial number of infected individuals were not fixed to to the initial values as is commonly done in most epidemiological studies. Instead, we introduced some flexibility by applying a sharp partial Bayesian log-normal prior. The quantile of this prior was treated as a parameter and calibrated using the data. For future research, a fully Bayesian approach could be adopted by specifying prior distributions for both the mean and variance of the log-normal distribution. This would allow a larger range for the parameter space while possibly not introducing a ton of variance, which occurs when we specify a uninformative prior, i.e, we allow the variance to become too large.

The relative performance of the different smoothing basis across all examples can be analyzed. In all examples (subsections 4.1, 4.2, 4.3, and 4.4), the Gaussian process (GP) regression smoothing basis proved to be the best in several respects. We observed that the GP basis was the most robust to increases in the effective degrees of freedom when the number of knots was increased. Tables 4.1, 4.2, and 4.3 show that during the model fitting procedure, selecting model complexity by adjusting the smoothing parameter λ resulted in a non-significant increase in the conditional AIC score. Checking the resultant effective degrees of freedom for the fitted models across different bases, the GP basis consistently showed the lowest values. This indicates that the GP basis, in conjunction with the `nlminb` optimizer used by `TMB`, is able to find the best smoothing parameter value compared to other bases. For the simulated data, this phenomenon was also observed with the TP, but the TP basis consistently had a conditional AIC score approximately two points lower than the GP basis. It is not clear to us why the difference between the AIC score for the GP and TP basis is constant. However, the uncertainty estimates for the TP basis increased dramatically with model granularity, far more than those for the GP basis. Additionally, across all datasets, the GP basis performed best or near-best. Notably, it was the only model capable of fitting the large Measles dataset. Our conclusion is that, using this methodology, a GP basis is generally an effective choice.

The cyclic basis assumes that the smoothing function takes the same values at the first and last knot for the zeroth and second derivatives. This implies that the shape and height of the function at the beginning and end of a cycle is equivalent. Figures 4.5 and 4.8 show that the shape of the curve for the observations and the fitted incidence at the beginning and end of the domain are about the same in terms of height and shape. Therefore, it is not surprising that the CC basis fits so well. What is not understood is why the CP and TS bases do not fit well for non-cyclic data, or if using a cyclic basis for these datasets is appropriate. Either way, it is interesting to observe and note this phenomenon regarding cyclic bases.

The effective degrees of the calibrated models are computed as the model degrees of freedom (subsection 3.6). Although the model degrees of freedom take into account the value of the smoothing parameter, it does not take into account the uncertainty estimates of the fitted smoothing parameter. We hypothesize that this may be a part of the reason why there is a constant AIC difference between the GP and TP basis, even

though the TP basis tends to have larger uncertainty estimates. From Appendix A.3, the effective degrees of freedom is equal to:

$$\tau = \text{tr}2\mathbb{E}\left[\frac{1}{2}(\hat{\beta} - \beta_K)^T \mathcal{I}_K(\hat{\beta} - \beta_K)\right] = \text{tr}\mathbb{E}[\chi^2_p] = p,$$

where β_K is the coefficient vector minimizing the K-L divergence and $\mathcal{I}_K$ is the expected negative Hessian of the log likelihood. In [23], Wood et al defines the corrected AIC as

$$\tau_2 = \text{tr}(\mathbf{V}'_\beta \hat{\mathcal{I}}),$$

where $\mathbf{V}'$ is an approximation of the covariance matrix of the Bayesian large sample approximation

$$\beta \mid y, \lambda \sim \mathcal{N}(\hat{\beta}_\lambda, \mathbf{V}_\beta)$$

and $\mathbf{V}_\beta = (\hat{\mathcal{I}} + \mathbf{P})^{-1}$. $\hat{\mathcal{I}}$ is the is the Hessian of the negative log likelihood at $\mathcal{I}_K$. The goal is to calculate a first-order adjustment to the posterior distribution of the model coefficients, taking into account the uncertainty in the smoothing parameter. After that, the penalty term in the AIC is represented using the Bayesian covariance matrix of the coefficients.

It would be informative, in future work, to compute the corrected AIC for the best fitting models and reevaluate the performance of the smoothing basis. For more complex compartmental models this might be essential as the conditional AIC with the model degrees of freedom is too likely to select a model which includes a random effect that is not present in the true model [24].

The SIR and SIRS compartmental models we used in this work are very basic compartmental models. Theoretically, a compartmental model can be as complex as the modeler wishes, but in practice, certain assumptions are made to make the fitting process tractable. The models we used are essentially toy examples that allow for a proof of concept of the efficacy of the methodology presented in this thesis.

In contrast, more realistic compartmental models can be highly complex. They may include numerous compartments (or nodes) and connections (or edges) between them, each with associated parameters or unknown functions that need to be estimated. These models can account for various factors such as different stages of infection, varying rates of transmission, recovery, and immunity, as well as heterogeneity in the population.

Despite their simplicity, our models effectively demonstrate the potential of the methodology. However, they do not capture the full complexity of real-world scenarios, where the intricate dynamics of disease spread necessitate more sophisticated models. By starting with these simpler models, we can establish a solid foundation for understanding

and validating the methodology before potentially extending it to more complicated and realistic compartmental models in future research.

In compartmental models for infectious disease, the response is often assumed to follow a Poisson distribution. This is because the Poisson distribution is well-suited for modeling count data, such as the number of new infection cases within a given time period. However, if the data exhibit overdispersion (i.e., the variance exceeds the mean), a negative binomial distribution might be used instead.

The objective function, as presented in Equation 3.2, assumes that both the observations and the smoothing coefficients are normally distributed. Consequently, when fitting the model to the data, we are essentially performing univariate Gaussian regression with respect to the underlying linear smoother. This assumption simplifies the model for the purpose of demonstrating methodological efficacy. An extension of this approach would be to assume that the observations and the smoothing coefficients follow distributions from the exponential family, such as Poisson or negative binomial distributions. Gu [25] describes how to construct the likelihood and penalty functionals for implementing penalized likelihood regression with non-Gaussian responses.

Assuming the unknown function is linear, as done in this thesis (when the smoother parameter is fixed for each iteration), is a step towards constructing more complex forms of the unknown function. If the modeler wishes to incorporate biological information about the transmission process into the model, they might impose qualitative conditions on the unknown function. For instance, if the unknown function is assumed to be logistic, the modeler can impose specific conditions of non-linearity and boundedness. Extending this methodology to include unknown functional forms other than linear allows the literature to guide the shape of the fitted unknown function by applying qualitative constraints to the smoothing functional.

In summary, what we have accomplished here is the proof of concept for the ability to estimate time varying unknown functions in deterministic compartmental models. There are many possible avenues of extension, some which we have discussed above. More generally this thesis suggests that it should be possible to be able to adapt this methodology to be used within any optimization framework that fits mixed models by using the theory of the duality of smooths and random effects, to rewrite the smoothing basis as random effects matrices, as we discussed in subsection 2.1.9. Another avenue of research is to estimate more than one unknown function. All of the methods here are easily extensible to fitting models to data with more than one unknown function, with their own smoothing parameter. Wood [26] and Gu [25] describe how to formulate models with more than one smoothing parameter.

Appendix A

Proofs, sketches and derivations

A.1 Matrix formulations and basis functions for cubic smoothing splines

The following sketch of a proof is taken from [7], where the reader can find the full proof.

To determine the function f that minimizes $J_2(f)$, we apply the Euler-Lagrange equation to a more general functional form $J(x) = \int L(x, x', x'', t)\, dt$. The Euler-Lagrange equation for this functional becomes:

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''} = 0. \tag{A.1}$$

When the function $f(x)$ extends beyond the range of the data points, or knots, we avoid imposing fixed boundary values for x and x' at the domain boundaries (i.e., $x(t_a)$, $x(t_b)$, $x'(t_a)$, and $x'(t_b)$). Instead, we use *natural boundary conditions*.

Natural boundary conditions are designed to ensure that the contributions from the boundary conditions to the first-order variation $\delta J = J(x + \delta x) - J(x)$ vanish, thus optimizing the solution. In standard scenarios, fixed values might be set for δx and $\delta x'$, where δ represents an infinitesimal change. However, this could potentially lead to undesirable behavior at the boundary points, especially outside the region defined by the knots.

From the application of the Euler-Lagrange equation A.1 and the principle that the first-order variation δJ should vanish at the optimal solution, we derive two critical *natural boundary conditions*:

$$\frac{\partial L}{\partial \dot{x}} - \frac{d}{dt}\frac{\partial L}{\partial \ddot{x}} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \ddot{x}} = 0,$$

where these conditions are each evaluated at t_a and t_b. These conditions help ensure that the function $f(x)$ not only fits the data within the knot range but also behaves optimally at the boundaries without artificial constraints.

Now, we can express Equation 2.2 in a variational form as follows:

$$J = \sum_{n=0}^{N-1} w_n(y_n - x(t_n))^2 + \lambda \int_{t_a}^{t_b} x''(t)^2 \, dt,$$

where the Lagrangian L is defined by:

$$L = \sum_{n=0}^{N-1} w_n(y_n - x(t))^2 \delta(t - t_n) + \lambda x''(t)^2.$$

Applying the Euler-Lagrange equation to this formulation yields:

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} + \frac{d^2}{dt^2}\frac{\partial L}{\partial x''} = -2 \sum_{n=0}^{N-1} w_n(y_n - x(t))\delta(t - t_n) + 2\lambda x^{(4)}(t) = 0. \tag{A.2}$$

The natural boundary conditions for this setup are:

$$x^{(3)}(t_a) = 0, \quad x''(t_a) = 0, \quad x^{(3)}(t_b) = 0, \quad x''(t_b) = 0.$$

By rearranging Equation A.2 to solve for $x^{(4)}(t)$, we derive:

$$x^{(4)}(t) = \lambda^{-1} \sum_{n=0}^{N-1} w_n(y_n - x(t_n))\delta(t - t_n), \tag{A.3}$$

This equation indicates that the third derivative of the spline function, $x(t)$, is zero except at the designated knot points t_n. Consequently, within each interval between knots, $x(t)$ must be represented as a cubic polynomial. The coefficients of these cubic polynomials may vary between intervals. The spline function transitions to first-degree polynomials in the endpoint intervals $[t_a, t_0]$ and $[t_{N-1}, t_b]$, defining what is meant by 'natural' in the context of natural cubic splines.

The explicit form of $x(t)$ is given by:

$$x(t) = \begin{cases} p_{-1}(t) = a_{-1} + b_{-1}(t - t_a), & t_a \le t \le t_0 \\ p_n(t) = a_n + b_n(t - t_n) + \frac{1}{2}c_n(t - t_n)^2 + \frac{1}{6}d_n(t - t_n)^3, & t_n \le t \le t_{n+1} \\ p_{N-1}(t) = a_{N-1} + b_{N-1}(t - t_{N-1}), & t_{N-1} \le t \le t_b \end{cases} \tag{A.4}$$

The coefficients are determined as follows:

$$\begin{aligned} a_n &= x(t_n) = p_n(t_n), \\ b_n &= p'_n(t_n), \\ c_n &= p''_n(t_n), \\ d_n &= p'''_n(t_n), \qquad\qquad \text{for } n = 0, 1, \ldots, N-1. \end{aligned}$$

From Equation A.3, we can establish the continuity and discontinuity conditions at the knots in terms of Equation A.4:

$$\begin{aligned} p_n(t_n) &= p_{n-1}(t_n), \qquad\qquad \text{for } n = 0, 1, \ldots, N-1, \\ p'_n(t_n) &= p'_{n-1}(t_n), \\ p''_n(t_n) &= p''_{n-1}(t_n), \\ p'''_n(t_n) - p'''_{n-1}(t_n) &= \lambda^{-1} w_n (y_n - a_n). \end{aligned} \tag{A.5}$$

These conditions ensure that each spline segment smoothly transitions into the next, preserving the continuity of the first, second, and third derivatives, except at the knots, where the third derivative may be discontinuous.

For the cubic spline model, there are $N-1$ cubic polynomials—one for each interval between knots—and two linear polynomials for the intervals at the domain boundaries, leading to a total of $4(N-1)+4 = 4N$ coefficients to solve in equations A.4. The equations derived using the constraints in equations A.5 form the basis functions for a cubic spline between knots x_j and x_{j+1}, with each interval defined by $h_j = x_{j+1} - x_j$. These basis functions are defined as:

$$\begin{aligned} a_j(x) &= \frac{x_{j+1} - x}{h_j}, \\ b_j(x) &= \frac{(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)}{6}, \\ c_j(x) &= \frac{x - x_j}{h_j}, \\ d_j(x) &= \frac{(x - x_j)^3/h_j - h_j(x - x_j)}{6}. \end{aligned}$$

The matrix elements for the non-cyclic spline are defined as follows:

$$\mathbf{B} = \frac{1}{6} \begin{bmatrix} 2(h_0 + h_1) & h_1 & 0 & 0 \\ h_1 & 2(h_1 + h_2) & h_2 & 0 \\ 0 & h_2 & 2(h_2 + h_3) & h_3 \\ 0 & 0 & h_3 & 2(h_3 + h_4) \end{bmatrix}$$

and

$$\mathbf{D} = \begin{bmatrix} h_0^{-1} & 0 & 0 & 0 & 0 \\ -(h_0^{-1} + h_1^{-1}) & h_1^{-1} & 0 & 0 & 0 \\ 0 & h_1^{-1} & -(h_1^{-1} + h_2^{-1}) & h_2^{-1} & 0 \\ 0 & 0 & h_2^{-1} & -(h_2^{-1} + h_3^{-1}) & h_3^{-1} \\ 0 & 0 & 0 & h_3^{-1} & -(h_3^{-1} + h_4^{-1}) \\ 0 & 0 & 0 & 0 & h_4^{-1} \end{bmatrix}.$$

Thus, these matrix formulations and the associated spline basis functions emerge from the process of optimizing the objective function outlined in Equation 2.2.

A.2 Laplace approximation

The following proof sketch is adapted from [17].

Let $f(u, \theta)$ denote the negative joint log-likelihood of the data and the random effects, where $u \in \mathbb{R}^n$ represents the unknown random effects and $\theta \in \mathbb{R}^n$ represents the model parameters. The MLE, in terms of the model parameters θ, is the marginal likelihood expressed as:

$$L(\theta) = \int_{\mathbb{R}^n} \exp(-f(u, \theta))\, du.$$

In this expression, the random effects are integrated out. Define $\hat{u}$ as the value that minimizes $f(u, \theta)$, leading to the Hessian $\mathbf{H}(\theta)$, which is the second partial derivative of $f(u, \theta)$ with respect to u, evaluated at $\hat{u}(\theta)$:

$$\mathbf{H}(\theta) = \frac{\partial^2 f}{\partial u^2}(\hat{u}(\theta), \theta).$$

Since f is approximated by a second-order Taylor expansion centered around $\hat{u}$, the first-order term disappears, resulting in the approximation:

$$f(u, \theta) \approx f(\hat{u}, \theta) - \frac{1}{2}(u - \hat{u})^T \mathbf{H}(\hat{u})(u - \hat{u}).$$

The Laplace approximation of $L(\theta)$ is then given by:

$$L^*(\theta) = (\sqrt{2\pi})^n \det(\mathbf{H}(\theta))^{-\frac{1}{2}} \exp(-f(\hat{u}, \theta)).$$

Taking the negative log of this approximation yields the objective function form used by `TMB` to estimate θ:

$$-\log L^*(\theta) = -n\log\sqrt{2\pi} + \frac{1}{2}\log\det(\mathbf{H}(\theta)) + f(\hat{u}, \theta). \tag{A.6}$$

This formulation of the objective function and its derivatives allows `TMB` to employ standard nonlinear optimization algorithms such as BFGS.

Uncertainty estimates for $\hat{\theta}$ or any differential function $\phi(\hat{\theta})$ are obtained through the delta method:

$$\mathrm{VAR}(\phi(\hat{\theta})) = -\left(\frac{\partial\phi}{\partial\theta}(\hat{\theta})\right)\left(\frac{\partial^2 \log L^*}{\partial\theta^2}(\hat{\theta})\right)^{-1}\left(\frac{\partial\phi}{\partial\theta}(\hat{\theta})\right)^T. \tag{A.7}$$

A.3 Akaike information criterion (AIC)

The following derivation of AIC is adapted from the proof sketch from [4].

Suppose we have two possible models P and Q for a data vector $\mathbf{X}$. We can think of these models as being a null hypothesis H and an alternative hypothesis A. Let $f_H(\mathbf{x})$ be the probability density of $\mathbf{X}$ under H and $f_A(\mathbf{x})$ under A. Define the log-likelihood ratio as

$$\eta(x) = \log\frac{f_A(x)}{f_H(x)}$$

Computing the expected value of $\eta(x)$ with respect to A is

$$\mathbb{E}_A[\eta(x)] = \int f_A(x)\log\frac{f_A(x)}{f_H(x)}\,dx.$$

This expected log-likelihood ratio can be interpreted as having the same form as the Kullback-Leibler (KL) divergence defined from the density f_A to the density f_H. When the alternative model f_A fits the data better than the wrong model f_H, i.e., A is true, the two models are well separated and the log-likelihood ratio will be positive. The ratio will be negative when f_H fits the data better than f_A, i.e., when H is true.

Suppose we misspecify the alternative hypothesis for some other model Q with density $f_Q(\mathbf{x})$. This leads to another interpretation of the KL divergence from f_A to f_Q as measuring how much power we lose with the likelihood ratio test if we misspecify the alternative hypothesis A as Q. Dualistically we can also make the mistake of taking the null hypothesis $f_H(\mathbf{x})$ to be $f_Q(\mathbf{x})$. The dual of this interpretation now says that the KL divergence from f_H to f_Q is the loss of power if we misspecify the null hypothesis.

Thus, the interpretation of the expected log-likelihood ratio statistic of two statistical models as the loss of power for specifying the model in terms of type one and type two

error can be insightful for finding the solution of the problem of accounting for model complexity in model selection.

If we were to judge between nested models on the basis of their fit to new data, not used in estimation, using the Likelihood Ratio Test (LRT), the model with the higher number of parameters will always have the higher likelihood. This is because the more complex model can better capture the nuances in the data. However, the Neyman-Pearson (NP) Lemma tells us that while the LRT is the most powerful test for simple hypotheses, in the context of model selection with multiple parameters (composite hypotheses), we need to balance model fit with complexity to avoid overfitting.

The Akaike Information Criterion (AIC) addresses this by incorporating a penalty for the number of parameters. This penalty helps control overfitting by favoring models that generalize better to new data, not just those that fit the training data well. Therefore, while the LRT tends to favor more complex models due to higher likelihoods, AIC provides a more balanced approach by considering both fit and parsimony. It accomplishes this in the following way.

Consider a scenario where our data are actually generated from a true density $f_{\theta_0}(y)$, while our model assumes a density $f_\theta(y)$, where θ represents the model parameters. Both y and θ are typically vectors, with θ having p dimensions. The Kullback-Leibler (KL) divergence between these densities is given by:

$$K(f_\theta, f_{\theta_0}) = \int [\log f_{\theta_0}(y) - \log f_\theta(y)] f_{\theta_0}(y)\, dy \tag{A.8}$$

This divergence quantifies how much the model f_θ deviates from the true density f_{θ_0}. When $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ, the KL divergence $K(f_{\hat{\theta}}, f_{\theta_0})$ serves as an indicator of the model's expected performance on new data, distinct from the data used to estimate $\hat{\theta}$. It's important to note that, for the purpose of evaluating this divergence, $\hat{\theta}$ is treated as a fixed value, independent of y.

We don't know what the density of the true model is. This can be overcome by constructing a truncated Taylor expansion of $\log(f_{\theta_0})$ about the unknown parameters θ_K, as the minimizer to equation A.8.

$$\log f_{\hat{\theta}}(y) \approx \log f_{\theta_K}(y) + (\hat{\theta} - \theta_K)^T g + \frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{H} (\hat{\theta} - \theta_K) \tag{A.9}$$

where g and $\mathbf{H}$ are the gradient vector and Hessian matrix of the first and second derivatives of $\log f_\theta(y)$ with respect to θ, evaluated at θ_K.

Substitute the Taylor expansion of $\log f_{\hat{\theta}}(y)$ into the KL divergence expression A.8:

$$K(f_{\hat{\theta}}, f_{\theta_0}) = \int \left[\log f_{\theta_0}(y) - \left(\log f_{\theta_K}(y) + (\hat{\theta} - \theta_K)^T g + \frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{H} (\hat{\theta} - \theta_K) \right) \right] f_{\theta_0}(y)\, dy$$

Separate the terms in the integral:

$$K(f_{\hat{\theta}}, f_{\theta_0}) = \int \left[\log f_{\theta_0}(y) - \log f_{\theta_K}(y) \right] f_{\theta_0}(y)\, dy - \int (\hat{\theta} - \theta_K)^T g f_{\theta_0}(y)\, dy - \int \frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{H} (\hat{\theta} - \theta_K) f_{\theta_0}(y)\, d$$

The first term is the KL divergence between f_{θ_K} and f_{θ_0}:

$$K(f_{\theta_K}, f_{\theta_0}) = \int \left[\log f_{\theta_0}(y) - \log f_{\theta_K}(y) \right] f_{\theta_0}(y)\, dy$$

Since θ_K minimizes the KL divergence $K(f_\theta, f_{\theta_0})$, the gradient vector g at θ_K will integrate to zero:

$$\int g f_{\theta_0}(y)\, dy = 0$$

The remaining term involves the Hessian $\mathbf{H}$:

$$\int \frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{H} (\hat{\theta} - \theta_K) f_{\theta_0}(y)\, dy$$

This term represents the second-order approximation of the KL divergence around θ_K.

Combining these results, we obtain:

$$K(f_{\hat{\theta}}, f_{\theta_0}) \approx K(f_{\theta_K}, f_{\theta_0}) + \frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{I}_K (\hat{\theta} - \theta_K)$$

Here, $\mathbf{I}_K$ is the Fisher information matrix evaluated at θ_K, which is equivalent to the negative expected value of the Hessian matrix $\mathbf{H}$.

Since we don't know θ_K, we take the expectation of the KL divergence approximation over the distribution of $\hat{\theta}$. This yields:

$$\mathbb{E}[K(f_{\hat{\theta}}, f_{\theta_0})] \approx K(f_{\theta_K}, f_{\theta_0}) + \mathbb{E}\left[\frac{1}{2}(\hat{\theta} - \theta_K)^T \mathbf{I}_K (\hat{\theta} - \theta_K) \right]$$

Under the assumption that the model is correct or nearly correct, $\hat{\theta}$ is approximately normally distributed around θ_K with covariance matrix $\mathbf{I}_K^{-1}$. Therefore, $(\hat{\theta} - \theta_K)^T \mathbf{I}_K (\hat{\theta} -$

θ_K) follows a chi-squared distribution with p degrees of freedom (p being the number of parameters).

The expected value of a chi-squared distribution with p degrees of freedom is p. Thus:

$$\mathbb{E}\left[\frac{1}{2}(\hat{\theta}-\theta_K)^T\mathbf{I}_K(\hat{\theta}-\theta_K)\right]=\frac{1}{2}\mathbb{E}[\chi_p^2]=\frac{p}{2}$$

Substituting this back into our expression, we get:

$$\mathbb{E}[K(f_{\hat{\theta}},f_{\theta_0})]\approx K(f_{\theta_K},f_{\theta_0})+\frac{p}{2} \tag{A.10}$$

The goal is to find an approximately unbiased estimator for $K(f_{\theta_K},f_{\theta_0})$, which is the Kullback-Leibler (KL) divergence between the true distribution f_{θ_0} and the model f_{θ_K}.

Given the log-likelihood function $l(\theta)=\log[f_\theta(y)]$, we start with

$$E[-l(\hat{\theta})]$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ. We decompose $E[-l(\hat{\theta})]$ as

$$E[-l(\hat{\theta})]=E[-l(\theta_K)]-E[l(\hat{\theta})-l(\theta_K)].$$

Next, we use the linearity of expectation:

$$E[-l(\hat{\theta})]=E[-l(\theta_K)]-E[l(\hat{\theta})-l(\theta_K)].$$

The term $E[-l(\theta_K)]$ corresponds to the expected log-likelihood under the true model, which can be linked to the KL divergence. Specifically,

$$E[-l(\theta_K)]=-\int\log[f_{\theta_K}(y)]f_{\theta_0}(y)\,dy.$$

The second term, $E[l(\hat{\theta})-l(\theta_K)]$, needs a bias correction. Considering the large sample result that $2(\log f_{\hat{\theta}}-\log f_{\theta_K})$ is approximately chi-squared distributed with p degrees of freedom, we use:

$$2(\log f_{\hat{\theta}}-\log f_{\theta_K})\sim\chi_p^2.$$

Thus, the bias correction term is $p/2$, leading to:

$$E[l(\hat{\theta}) - l(\theta_K)] \approx \frac{p}{2}.$$

So, we get:

$$E[-l(\hat{\theta})] = E[-l(\theta_K)] - \frac{p}{2}.$$

Recall the definition of the Kullback-Leibler (KL) divergence:

$$K(f_{\theta_K}, f_{\theta_0}) = -\int \log[f_{\theta_K}(y)] f_{\theta_0}(y)\, dy - \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy.$$

Notice that the first term on the right-hand side of our expectation equation is the negative expected log-likelihood of the model evaluated at θ_K:

$$E[-l(\theta_K)] = -\int \log[f_{\theta_K}(y)] f_{\theta_0}(y)\, dy.$$

So, we can express the KL divergence as:

$$K(f_{\theta_K}, f_{\theta_0}) = E[-l(\theta_K)] + \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy.$$

We have:

$$E[-l(\hat{\theta})] \approx K(f_{\theta_K}, f_{\theta_0}) - \frac{p}{2} - \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy.$$

Rearranging this for $K(f_{\theta_K}, f_{\theta_0})$:

$$K(f_{\theta_K}, f_{\theta_0}) \approx E[-l(\hat{\theta})] + \frac{p}{2} + \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy.$$

The log-likelihood evaluated at the MLE $\hat{\theta}$ is a random variable that converges in probability to the log-likelihood evaluated at the true parameter value θ because in this case the MLE is a consistent estimator. This means that The probability that the estimator is within an ϵ-neighborhood of the true parameter value approaches 1 as the sample size increases.

The expectation $E[-l(\hat{\theta})]$ involves the distribution of $\hat{\theta}$, which, for large samples, is concentrated around θ. Therefore we can approximate $E[-l(\hat{\theta})]$ with $-l(\hat{\theta})$ for large sample sizes. This approximation is justified because $\hat{\theta}$ is close to θ, and the observed log-likelihood $-l(\hat{\theta})$ will be close to its expected value.

Since $E[-l(\hat{\theta})]$ is the expectation of the negative log-likelihood, we approximate it with the observed value $-l(\hat{\theta})$. We obtain the unbiased estimator for the KL divergence between the true distribution f_{θ_0} and the model f_{θ_K}.

$$\widehat{K(f_{\theta_K}, f_{\theta_0})} \approx -l(\hat{\theta}) + \frac{p}{2} + \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy.$$

Substituting this into A.10 we obtain:

$$\mathbb{E}[K(f_{\hat{\theta}}, f_{\theta_0})] \approx l(\hat{\theta}) + p + \int \log[f_{\theta_0}(y)] f_{\theta_0}(y)\, dy..$$

Since we don't have $f_{\theta_0}(y)$, we drop the last term, as it is a constant across any set of models compared using the same data set:

$$\mathbb{E}[K(f_{\hat{\theta}}, f_{\theta_0})] \approx -\log f_{\hat{\theta}} + p = -l(\hat{\theta}) + p$$

Scaling the above equation by a factor of 2, the Akaike Information Criterion (AIC) is:

$$\text{AIC} = -2 \log \hat{L} + 2p \tag{A.11}$$

Appendix B

SIRS model with simualted data (cylic basis)

Figures B.1, B.2 and B.3 show the results of the predicted incidence, transmission rate, and effective reproduction number for data simulated using a cyclic cubic (CC) regression smoother with varying numbers of knots ($k = 8, 15, 20$), from left to right. The variance b_{sd} in the Gaussian coefficients used to generate the simulation model and calibrating models are fixed at $b\_sd = 2$. γ and ϕ are fixed at $\frac{1}{14}$ and 10^{-3} respectively.

The SIRS model is used for the simulation, with compartment values initialized at the endemic equalibrium solutions. The figures use the following visual elements:

- The red line represents the simulated trajectory with Gaussian noise.
- The blue line (for incidence) or green line (for transmission rate and reproduction number) indicates the predicted values.
- The light and dark colored bands represent the 95% and 50% confidence intervals, respectively.

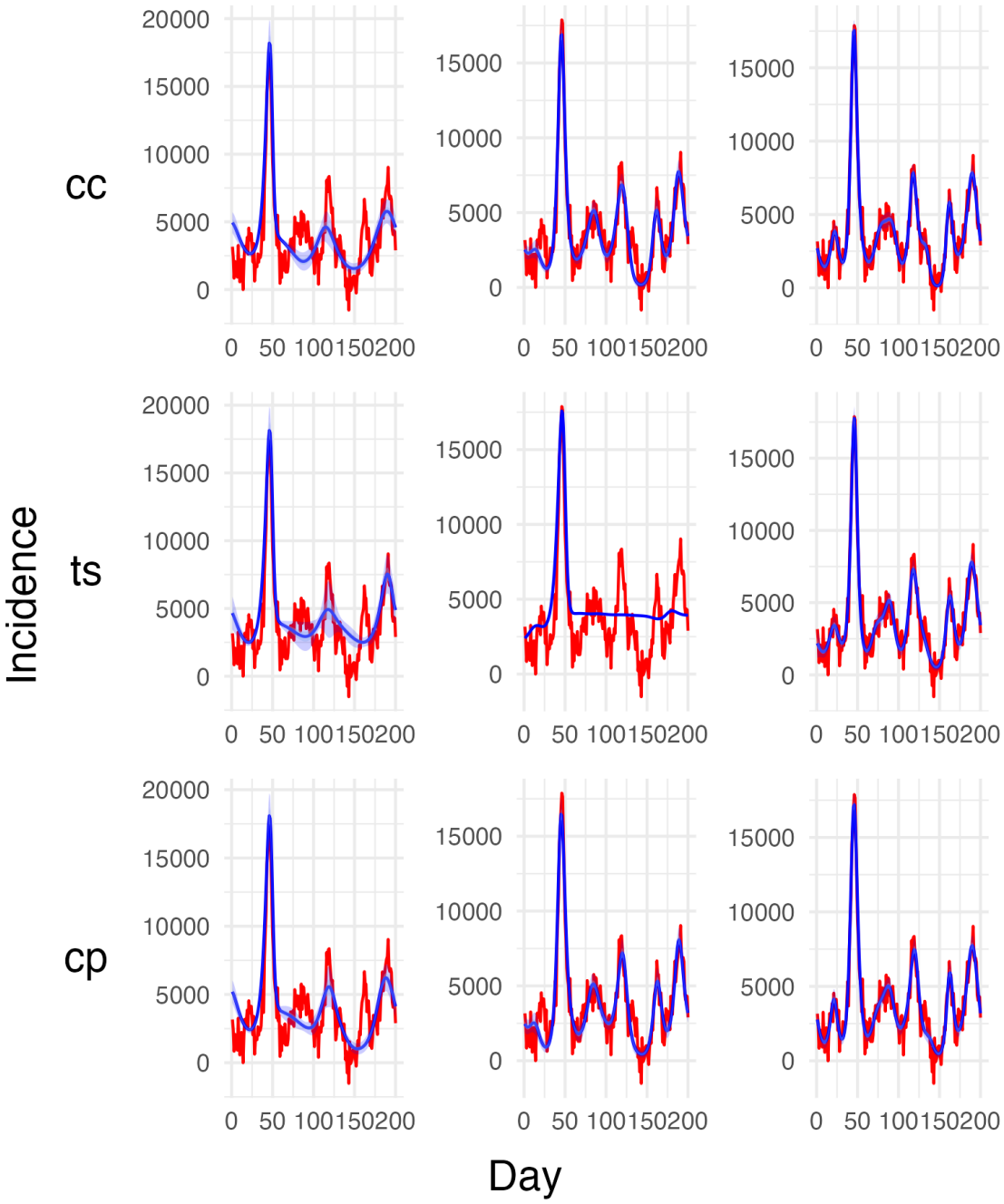

FIGURE B.1: **Predicted incidence for data simulated using a cyclic cubic (CC) regression smoother with varying numbers of knots ($k = 8, 15, 20$).** The plots show the predicted incidence fitted to data simulated from a SIRS model.

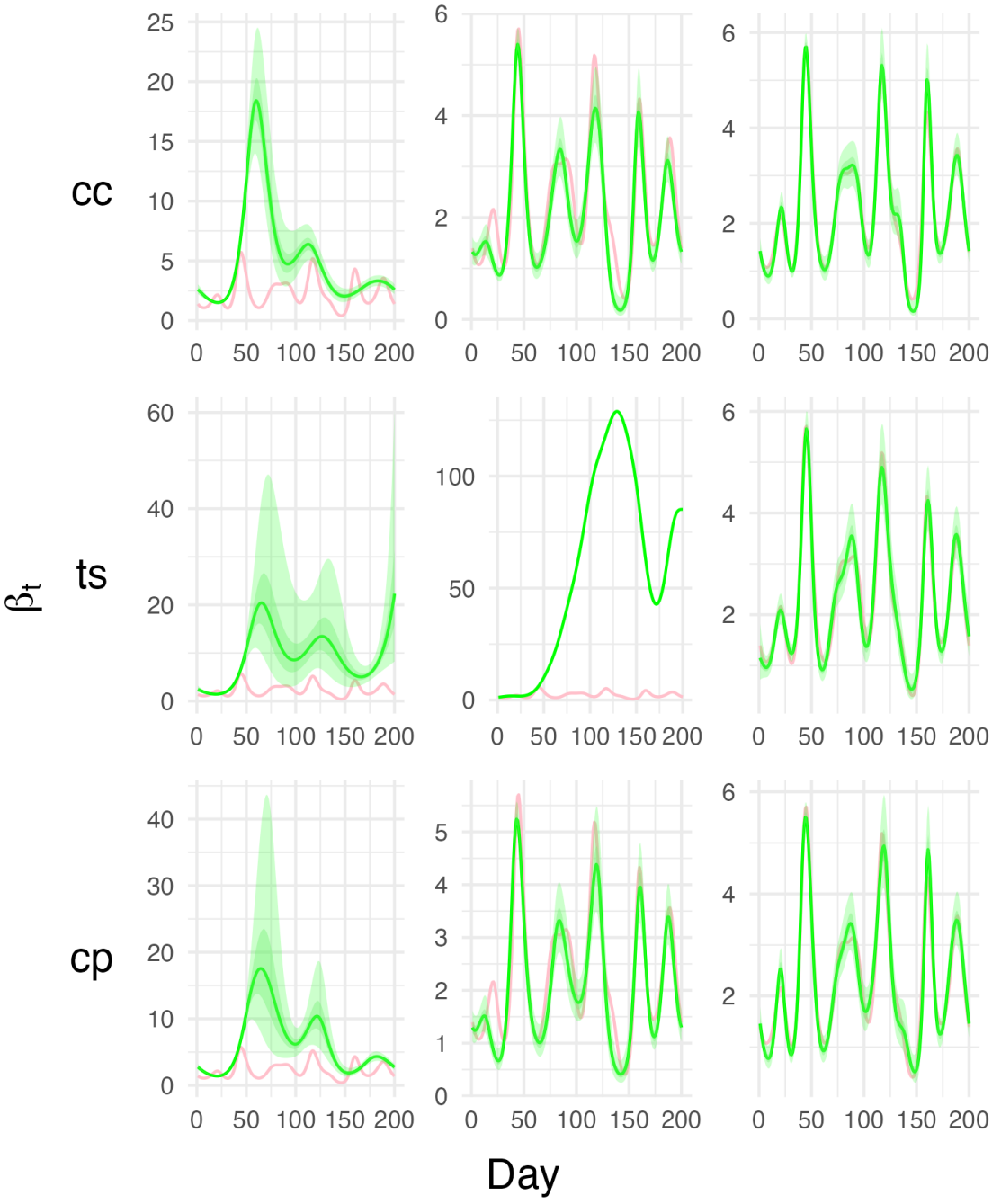

FIGURE B.2: **Estimated transmission rate for data simulated using a cyclic cubic (CC) regression smoother with varying numbers of knots ($k = 8, 15, 20$).** The plots show the predicted transmission rate fitted to data simulated from a SIRS model.

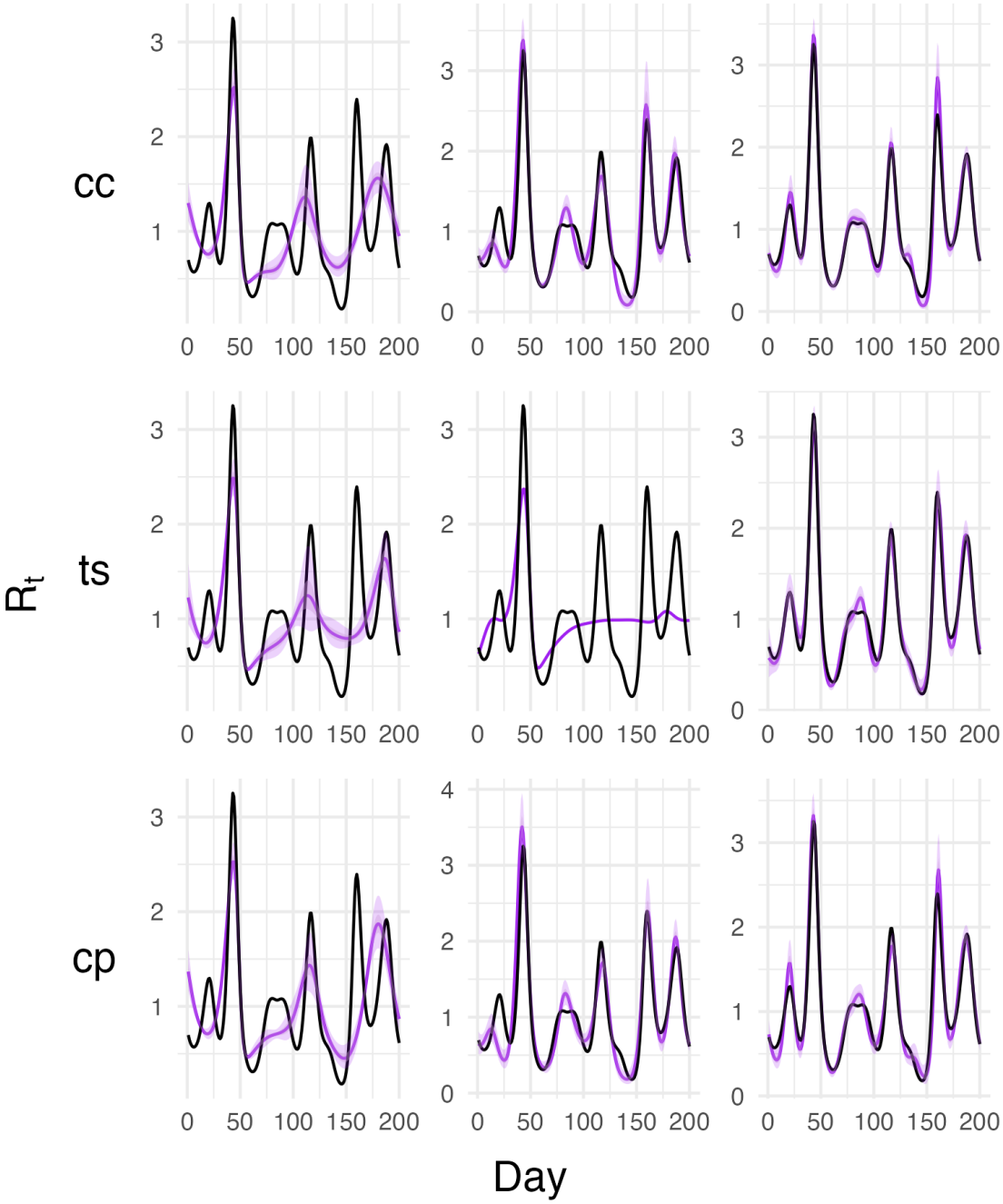

FIGURE B.3: **Estimated effective reproduction number for data simulated using a cyclic cubic (CC) regression smoother with varying numbers of knots ($k = 8, 15, 20$).** The plots show the predicted effective reproduction number fitted to data simulated from a SIRS model.

References

[1] Bolker BM. Ecological Models and Data in R. Princeton University Press; 2008. https://doi.org/10.2307/j.ctvcm4g37.

[2] Wood SN. Partially Specified Ecological Models. Ecological Monographs 2001;71.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Springer New York, NY; n.d.

[4] Wood SN. Generalized additive models: An introduction with R. Second edition. Boca Raton: CRC Press/Taylor & Francis Group; 2017.

[5] Carl de Boor. A Practical Guide to Splines. vol. Volume 27. Applied Mathematical Sciences, New York: Springer; 1978.

[6] Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models: A roughness penalty approach. New York: Chapman and Hall/CRC; 1993. https://doi.org/10.1201/b15710.

[7] Orfanidis SJ. Optimum signal processing. Alpha Books; 1989.

[8] Lancaster P, Salkauskas K. Curve and surface fitting - an introduction, 1986.

[9] Wood SN. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. Journal of the American Statistical Association 2004;99:673–86.

[10] Wood SN. Thin Plate Regression Splines. Journal of the Royal Statistical Society Series B: Statistical Methodology 2003;65:95–114. https://doi.org/10.1111/1467-9868.00374.

[11] David Ruppert, M. P. Wand, R. J. Carroll. Semiparametric Regression. Cambridge University Press; 2003.

[12] Goran Kauermann. Penalized Splines, Mixed Models and Bayesian Ideas. Statistical Modelling and Regression Structures, Springer-Verlag Berlin Heidelberg; 2010, p. 45–57.

[13] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. The MIT Press; 2005. https://doi.org/10.7551/mitpress/3206.001.0001.

[14] Flynn-Primrose D, Walker SC, Li M, Bolker BM, Earn DJD, Dushoff J. Toward a comprehensive system for constructing compartmental epidemic models 2023. http://arxiv.org/abs/2307.10308 (accessed March 31, 2024).

[15] Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. Am J Epidemiol 2013;178:1505–12. https://doi.org/10.1093/aje/kwt133.

[16] de Vries G, Hillen T, Lewis M, Müller J, chönfisch, Birgit. A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational Methods. SIAM; 2006.

[17] Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM. TMB: Automatic Differentiation and Laplace Approximation. Journal of Statistical Software 2016;70:1–21. https://doi.org/10.18637/jss.v070.i05.

[18] Madsen H, Thyregod P. Introduction to General and Generalized Linear Models. CRC Press; 2010.

[19] Wood S. Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation 2023.

[20] Walker S, Earn D. Canmod/iidda: International Infectious Disease Data Archive 2024. https://github.com/canmod/iidda/tree/main (accessed June 5, 2024).

[21] Andrade J, Duggan J. Inferring the effective reproductive number from deterministic and semi-deterministic compartmental models using incidence and mobility data. PLOS Computational Biology 2022;18:e1010206. https://doi.org/10.1371/journal.pcbi.1010206.

[22] Earn DJD, Rohani P, Bolker BM, Grenfell BT. A Simple Model for Complex Dynamical Transitions in Epidemics. Science 2000;287:667–7.

[23] Wood SN, Pya N, Säfken B. Smoothing Parameter and Model Selection for General Smooth Models. Journal of the American Statistical Association 2016;111:1548–63. https://doi.org/10.1080/01621459.2016.1180986.

[24] Greven S, Kneib T. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 2010;97:773–89.

[25] Chong Gu. Smoothing Spline ANOVA Models. 1st ed. New York, NY: Springer; 2013.

[26] Wood SN. Generalized additive models: An introduction with R. Second edition. Boca Raton: CRC Press/Taylor & Francis Group; 2017.