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

Research Questions

- **PRQ1:** How can we use univariate Gaussian regression smoothers in statistical modeling?
- **RQ1:** How can we estimate the time-varying transmission rate by fitting a differential equation model to a single time series of incidence data?
- **RQ2:** How can we formulate and fit semi-mechanistic models within the `macpan2` compartmental modeling framework?

Literature Review

- Wahba (1990) and Hastie and Tibshirani (1990) developed methodologies for general non-parametric statistical modeling, which have been applied to ecological modeling.
- Ellner et al. (1998) introduced "semi-mechanistic" models that combine deterministic and parametric components with non-parametric methods for flexible function estimation.
- Simon N. Wood (2001) presented a methodology using penalized smoothing to estimate time-varying latent variables in dynamic ecological models, employing quasi-Newton methods and generalized cross-validation.

Workflow

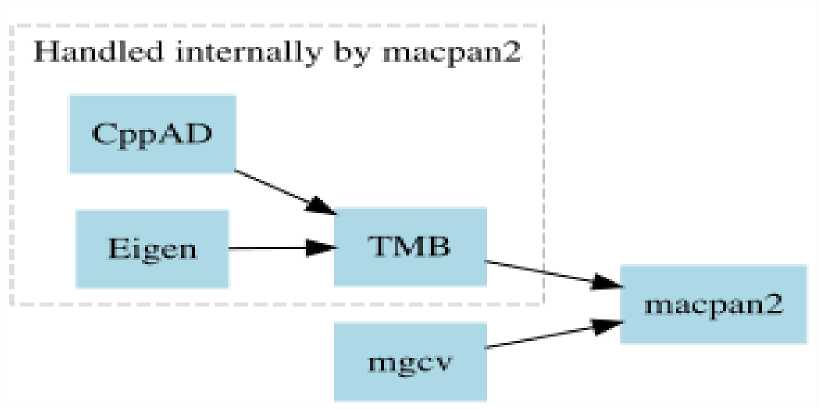

Figure 1: Workflow of software package relationships

Epidemeological Basics - What is an SIR Model?

The rates of transition between compartments in the SIR model can be expressed as a system of nonlinear ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

where:

- $\frac{dS}{dt}$ is the rate of change of the susceptible population.
- $\frac{dI}{dt}$ is the rate of change of the infectious population.
- $\frac{dR}{dt}$ is the rate of change of the recovered population.
- γ is the recovery rate.

Epidemiological Basics - What role does β play in the SIR model?

- β is the transmission rate, which represents the rate at which an infectious individual can spread the disease to susceptible individuals.
- The term SI is proportional to the number of contacts between susceptible and infectious individuals.
- β can be decomposed into three factors: (susceptibility) $\times$ (contact rate) $\times$ (infectiousness).
- Multiplying SI by β gives the number of new infections per unit time, as it scales the contact rate by the probability of transmission per contact.

Epidemiological Basics - What is the effective reproduction number R_t?

- The *effective reproduction number* R_t can be calculated as the product of the infection rate per contact, the number of contacts per unit time, and the duration of infectiousness:

$$R_t = \frac{\beta(t)}{\gamma} \frac{S(t)}{N}, \tag{1}$$

where N is the total population size.

- Therefore R_t is the average number of secondary cases of disease caused by a single infected individual over their infectious period.

PRQ1: Smoothing Basics - What is a linear smoother?

- Consider a typical univariate Gaussian data model $y = f(x) + \epsilon$.
- We can define a linear smoother by choosing a *basis*, which means choosing some basis functions.
- The unknown function f then has representation

$$f(x) = \sum_{i=1}^{k} \delta_i(x) b_i, \tag{2}$$

where b_i are the unknown parameters and δ_i represents the i^{th} basis function.

PRQ1: Smoothing Basics - How can we prevent overfitting?

- Smoothing is achieved by minimizing the following objective function:

$$L(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2\, dx. \tag{3}$$

- A computationally efficient form of the quadratic penalty functional is:

$$\int_{x_1}^{x_k} f''(x)^2\, dx = b^T \mathbf{P} b, \tag{4}$$

where **P** is called the *penalty matrix* for this basis.

- The penalized regression problem, is formulated to minimize

$$\|y - \mathbf{X}b\|^2 + \lambda b^T \mathbf{P} b, \tag{5}$$

where **X** is the called the *basis matrix*.

PRQ1: Smoothing Basics - What does overfitting look like?

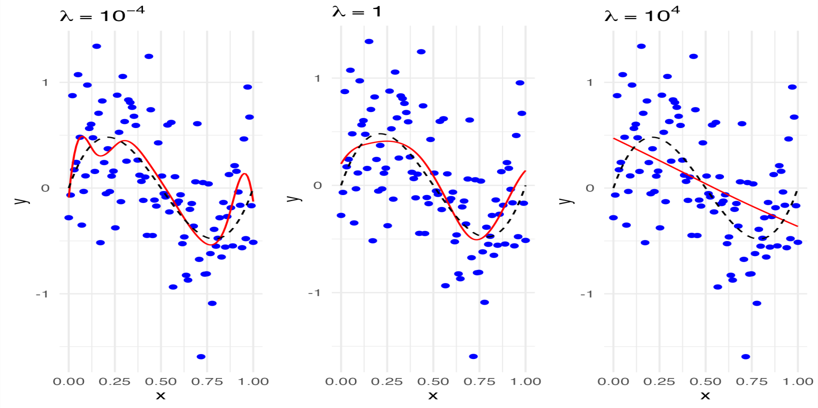

Figure 2: Penalized Regression Spline Fits with Different Smoothing Parameters. Data were generated using a polynomial function with added noise.

PRQ1: Smoothing Basics - What kind of basis functions can we use? (1)

- **Cubic Regression Splines (CR)**
  - Fits a cubic polynomial between each pair of data points, with continuity at the knots.
- **B-Splines (BS)**
  - Constructed from piecewise polynomials defined over a sequence of knots. Computationally efficient and each basis function is strictly local; i.e., only non-zero over the intervals between $m+3$ adjacent knots.

PRQ1: Smoothing Basics - What kind of basis functions can we use? (2)

- **Gaussian Process Regression Smoothers (GP)**
  - A Gaussian Process (GP) defines a distribution over functions with a mean function $\mu(x)$ and a covariance function $k(x, x')$. It provides a non-parametric way to interpolate and smooth data, where the kernel function controls the smoothness and correlation between points. Any finite collection of function values is jointly normally distributed.
- **Thin Plate Regression Splines (TP)**
  - Useful for smoothing in multiple dimensions, using derivative penalties of integer order and does not require knot placement. In the univariate case they reduce to radial basis functions.

PRQ1: Smoothing Basics - What kind of basis functions can we use? (3)

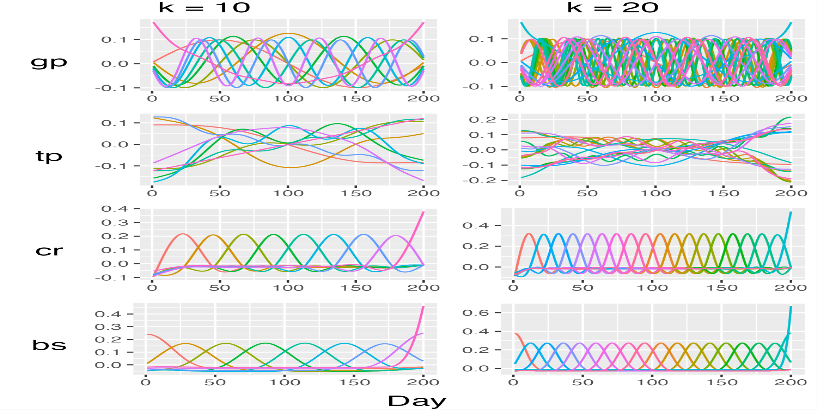

Figure 3: Basis functions for calibrating smoothers. Basis matrices are obtained via mgcv::smoothCon.

Summary of Progress So Far

To recap, our goal is to infer the unobserved transmission rate in a compartmental model formulated as a system of ordinary differential equations. We achieve this by:

- Representing the unknown function as a linear smoother.
- Avoiding overfitting by adding a penalization term.
- Using the `mgcv` package to obtain the basis and penalty matrices.

RQ1: Formulating the model - What are the model assumptions?

We have the following assumptions in the model:

$$
\begin{aligned}
&\text{Priors:} \\
I_0 &\sim \text{Lognormal}(\mu_{I_0}, \sigma^2_{I_0}) \\
\gamma &\sim \text{Lognormal}(\mu_\gamma, \sigma^2_\gamma) \\
&\text{Likelihood:} \\
Y &\sim \mathcal{N}(f(x), \sigma^2_Y) \\
\beta &\sim \mathcal{N}(0, \frac{\mathbf{P}^-}{\lambda}),
\end{aligned} \tag{6}
$$

where $f(x)$ are the fitted values (incidence) and $\mathbf{P}^-$ is the psuedoinverse of $\mathbf{P}$, which is rank deficient by the dimension of the penalty null space.

RQ1: Formulating the model - How can we specify a time varying transmission rate process using a linear smoother?

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

$$\beta = \exp(b_0 + \mathbf{X}b), \tag{7}$$

where b_0 is the intercept to be estimated, $\mathbf{X}$ is the basis matrix of dimensions $n \times (k-1)$, and b is a vector of basis coefficients of length $k-1$.

- The log likelihood function for b is

$$L(b) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\lambda) - \log(\det(\mathbf{P})) + \frac{1}{2\lambda}b^T\mathbf{P}b. \tag{8}$$

RQ2: Formulating the model - macpan2 Model Structure (1)

```
---------------------
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: beta ~ exp(eta[time_step(1)]) #extract current
# value of beta
2: R_t ~ (log(beta) - log(gamma) + log(S) - log(N))
3: incidence ~ S * I * beta/N
4: recovery ~ gamma * I
5: S ~ S - incidence
6: I ~ I + incidence - recovery
7: R ~ R + recovery
```

RQ2: Formulating the model - macpan2 Model Structure (3)

```
---------------------
After the simulation loop (t = n+1):
---------------------
1: log_lik ~ -sum(dnorm(I_obs,
                        I_fitted,
                        I_sd))
               - dnorm(log_gamma,
                       mean_log_gamma,
                       sd_log_gamma)
               - dnorm(log_I_0, mean_log_I_0, sd_log_I_0)
               + log(det(P))
               - 1/2 * log(log_lambda)
               + 1/2 * ((t(b) %*% P %*% b) / log_lambda)
```

Model Comparison: Conditional AIC i

- The 'natural' parameterization means parameter estimators are independent with unit variance in the absence of the penalty, and the penalty matrix is diagonal.
- Each unpenalized coefficient b_i counts for one degree of freedom.
- The penalized parameter estimates are shrunken versions of the unpenalized coefficients: $\hat{\beta}_i = (1 + \lambda D_{ii})^{-1}\beta_i$.
- The shrinkage factors, $(1 + \lambda D_{ii})^{-1}$, range from 0 to 1.
- Since unpenalized coefficients have one degree of freedom each, the shrinkage factor can be interpreted as the 'effective degrees of freedom' of $\hat{\beta}_i$.

Model Comparison: Conditional AIC ii

- The total effective degrees of freedom for the smooth is the sum of the individual shrinkage factors:

$$\sum_i (1 + \lambda \mathbf{D}_{ii})^{-1} = \mathrm{tr}(\tau) \quad \text{where} \quad \tau = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{P})^{-1}\mathbf{X}^T\mathbf{X}. \tag{9}$$

  where $\mathbf{D}$ is a diagonal matrix of the eigenvalues of the 'naturally' parameterized penalty matrix, arranged in decreasing order.

- Therefore, the AIC formula corrected to incorporate the effective degrees of freedom is the *conditional AIC*:

$$\text{cAIC} = -2\ell(\hat{b}) + 2\tau, \tag{10}$$

where $\ell(\hat{b})$ is the maximum likelihood estimate of the model (Simon N. Wood 2017).

Results: Simulation Study - How is the simulated data constructed? (1)

We construct the data-generating model as follows:

- **1.** The smoother type and the number of knots k are specified using a particular `mgcv` smooth.
- **2.** The smoothing coefficients b, of dimension $k-2$, are assumed to be multivariate Gaussian. Thus, b is defined as random normal deviates at the $k-2$ knots with a mean of 0 and a standard deviation specified as b_{sd}.
- **3.** An initial value for b_0 is chosen as the log of the initial value of β. The recovery rate γ is fixed.
- **4.** The model trajectory is simulated from these initial conditions using Euler steps. Gaussian noise (sd $= 0.2$) is added to the simulated incidence vector.

Results: Simulation Study (2)

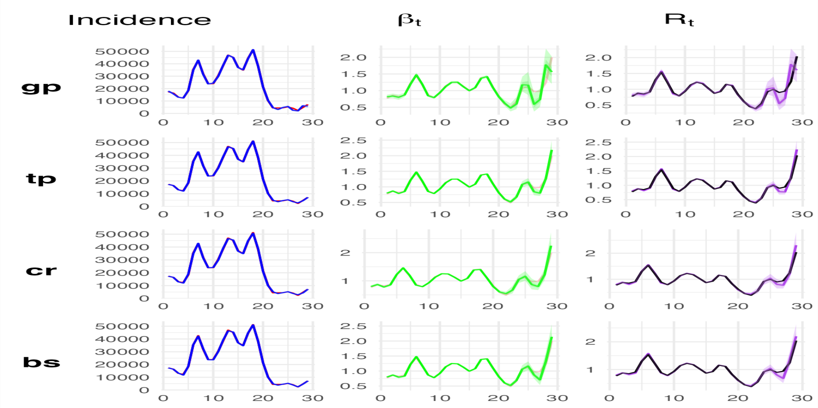

Figure 4: SIR Model with data simulated using a B-Spline basis. The data is aggregated to a weekly scale before fitting but after simulating the trajectory.

Table 1: Conditional AIC Scores for SIR Model with Simulated Data Aggregated to a Weekly Scale. The columns represent the smoothing basis used to fit the model, while the rows indicate the basis used to generate the simulated data. The trajectories were simulated on a daily scale and then aggregated to a weekly scale for model calibration. ΔAIC values are calculated relative to the best score within each row.

BasisType	gp	tp	cr	bs
gp	1.94	2.02	0	1.01
tp	0.90	0.69	0	NA
cr	1.78	1.98	0	0.93
bs	1.90	2.27	0	1.09

Results: Covid-19 in Ireland 2020 (1)

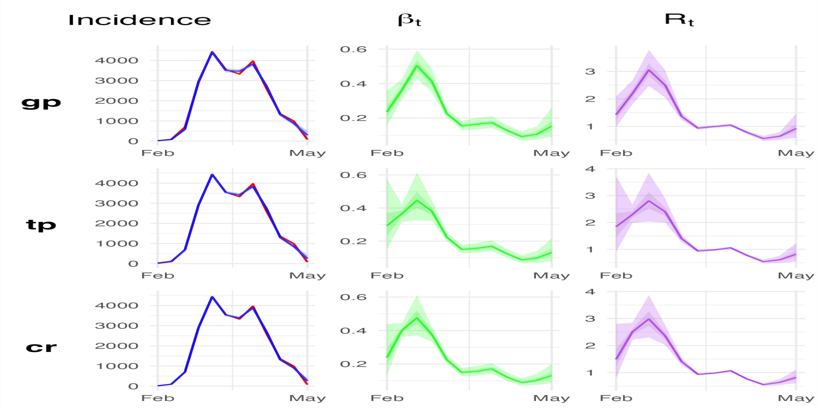

Figure 5: Combined analysis of predicted incidence, estimated transmission rate, and effective reproduction number for observed Covid-19 cases in Ireland, 2020.

Table 2: Conditional AIC Scores of calibrating models with varying smoothing basis, calibrated to Ireland Covid-19 (2020). The $\triangle$AIC values are calculated relative to the best score.

Smooth Type	Delta AIC
gp	1.24
tp	0.14
cr	0.00

Conclusion

- This thesis demonstrated a method for estimating time-varying functions in deterministic compartmental models.
- We formulated infectious disease models without relying on fixed or parametric assumptions about disease transmission.
- By integrating this approach into the macpan2 and TMB frameworks, we offer a user-friendly tool for model fitting, model selection, and inference of time-varying latent processes.
- Simulation studies confirmed the efficacy of penalized smoothing parameter estimation.
- The models were successfully applied to real-world incidence data.

RQ1: Formulating the model - How can we estimate the unknown parameters?

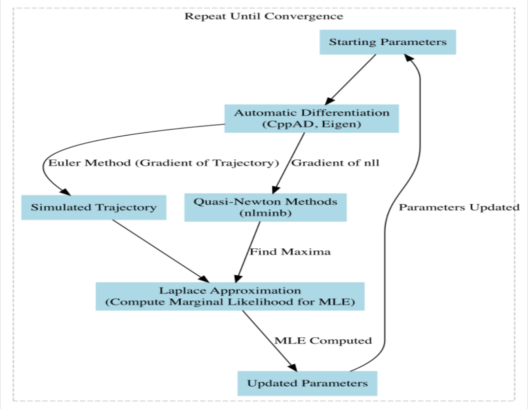

Figure 6: Flowchart illustrating the optimization process.

Ellner, S. P., B. A. Bailey, G. V. Bobashev, A. R. Gallant, B. T. Grenfell, and D. W. Nychka. 1998. "Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling." *The American Naturalist* 151 (5): 425–40. https://doi.org/10.1086/286130.

Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized Additive Models*. CRC Press. https://books.google.com?id=qa29r1Ze1coC.

Wahba, Grace. 1990. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970128.

Wood, Simon N. 2001. "Partially Specified Ecological Models." *Ecological Monographs* 71 (1).

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. Second edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press/Taylor & Francis Group.