EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Eberhard Karls University of Tübingen
Faculty of Mathematics and Natural Sciences
Department of Computer Science
Chair for the Methods of Machine Learning

Bachelor's Thesis Cognitive Science

Probabilistic Modelling of Infectious Disease Dynamics

A Case Study of COVID-19 in Germany

Hanna Dettki

November 20, 2020

Supervisor
Prof. Dr. Philipp Hennig
University of Tübingen
Department of Computer Science

Advisor
Jonathan Wenger
University of Tübingen
Department of Computer Science

Dettki, Hanna:
Probabilistic Modelling of Infectious Disease Dynamics
Bachelor's Thesis Cognitive Science
Eberhard Karls University of Tübingen
Student ID: 4106641
`hanna.dettki@student.uni-tuebingen.de`
Start: June 23, 2020
End: November 20, 2020

Abstract

In order to contain the pandemic caused by the novel coronavirus (SARS-CoV-2), first reported in China in December 2019, many countries imposed unprecedented lockdowns that have resulted in extraordinary consequences, far reaching into all facettes of life. There is great interest to understand the current infection process and development in a timely manner in order to allow for mitigation policies to be implemented aptly and promptly. A pivotal figure, often required for deriving further epidemiological parameters, is the number of infections occuring on a given day, which naturally, nobody can exactly determine. We provide a probabilistic model to approximate this number via the disease onset date using Bayesian inference. Focusing on case numbers in Germany, we find reporting-specific effects, which we "model-away" with a Gaussian process defined by a sum kernel. Finally, we derive a metric, which quantifies the growth rate of new infections. The uncertainty in the data, which is especially large at the beginning of the pandemic when test capacities were low, poses a major challenge. We compensate for this data scarcity by integrating prior knowledge into the model and provide the resulting estimates with uncertainty in order to keep them faithful.

Acknowledgements

Powered by Coffee! Over the course of this thesis, I had many cups of coffee. Perhaps too many. I even made it to the top of the Hall of Fame list with $\sim 84 - 104$ single shots over the past 30 days until I was passed by Yannick and Michael who were both working towards exams and AISTATS at the time. As my deadline is approaching and theirs have passed, I expect to see my name slowly but surely working its way back to the top of the list. From my love for both foam and coffee naturally followed the intent to add the skill of mastering latte art as yet another task to be achieved by submission date. Clearly, Figure 1 shows that I still have a long way ahead of me until my ”foam-pictures” can be considered true latte-art! This is why usually, I categorize my latte ”art” as *abstract art* and let people decide what it is supposed to depict which sometimes yielded quite funny interpretations. So, maybe I should extend this thesis solely for the purpose of getting some more latte art practice and possibly the opportunity to attend the planned barista workshop?

Figure 1: *Powered by Coffee.*

Acronyms

GP	Gaussian process
GPR	Gaussian process regression
LR	Linear Regression
rv	random variable
SE	Squared exponential
RKI	Robert Koch Institute
pdf	Probability density function
MLE	Maximum likelihood estimate / estimation
MAP	Maximum a posteriori

Notation

Scalars, Vectors, Matrices and Symbols

θ	Scalar or (probability distribution) parameter
$\boldsymbol{x}$	(Column) vector
I	Unit matrix
$\boldsymbol{y}^{\boldsymbol{T}}$	The transpose of vector $\boldsymbol{y}$
$\triangleq$	An equality which acts as a definition
$x \propto y$	x is directly proportional to y

Probability Theory

$p(x)$	Probability density function or probability mass function
$p(y\|x)$	Conditional density function
i.i.d.	Independent and identically distributed
$X \sim D$	Random variable X is distributed according to distribution D
Σ	Covariance
$\mathcal{N}(\mu, \Sigma)$	(Multivariate) normal distribution with mean μ and covariance Σ
$\mathcal{N}(x\|\mu, \Sigma)$	Density of the (multivariate) normal distribution
$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{x'})$	Covariance between vectors $\boldsymbol{x}$ and $\boldsymbol{x'}$
$\mathcal{GP}(\mu, k)$	Gaussian process with mean function $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$

Regression

$\boldsymbol{x}$	Input vector
x	Scalar input
$x_\star$	Scalar test point, (or vector if bold)
y	Observed target value which is assumed to be corrupted by noise, i.e., $y = f(\boldsymbol{x}) + \epsilon$
$\boldsymbol{y}$	Vector of targets
$y_\star$	Predicted target or output value corresponding to novel test input $\boldsymbol{x}_\star$
$\boldsymbol{w}$	Vector of weights (parameters)
$\phi(\boldsymbol{x})$	Function which maps a $D-$dimensional input vector $\boldsymbol{x}$ into an N dimensional feature space
ϕ_x	abbreviation for $\phi(\boldsymbol{x})$
$\phi_\star$	abbreviation for $\phi(\boldsymbol{x}_\star)$
$f(\boldsymbol{x})$	A real process where $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})))$. f evaluated at x is a random variable.
f_x	abbreviation for $f(\boldsymbol{x})$
$f(\boldsymbol{x}_\star)$	Gaussian process (posterior) prediction (random variable)
$\boldsymbol{f}_\star$	abbreviation for $f(\boldsymbol{x}_\star)$
ϵ	Noise by which observed values y differ from function values $f(\boldsymbol{x})$, assumed to be i.i.d. and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$
σ_n^2	Variance of noise
X	$N \times D$ Design matrix
D	Dimension of input space
$\mathcal{D}$	Training set, $\mathcal{D} = (X, \boldsymbol{y})$
$k(\boldsymbol{x}, \boldsymbol{x'})$	Kernel or covariance function evaluated at $\boldsymbol{x}$ and $\boldsymbol{x'}$
$\boldsymbol{k(x_\star)}$	Vector, short for $K(X, x_\star)$ when there is only a single test case $\boldsymbol{x}_\star$
$\boldsymbol{k}_\star$	Abbreviation for $\boldsymbol{k(x_\star)}$
K	$N \times N$ Covariance matrix $K(X, X)$ for the training points
$K_\star$	$N \times N_\star$ Covariance matrix $K(X, X_\star)$ between the training and test cases
K_f	Covariance matrix for the noise free $\boldsymbol{f}$ values

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The novel coronavirus disease (COVID-19), first reported in Wuhan, China, in December 2019, has rapidly grown into a pandemic as cases have been reported worldwide [1]. With more than 53 million confirmed cases in 191 countries and about the 1.3 million deaths [2] as of November 2020, the virus continues to spread across the globe.

The ongoing pandemic poses immense threats not only to public health, but also to the stability of society and economies across the world. In order to contain the outbreak, many countries have imposed unprecedented lockdowns and other policies with far reaching consequences for family life, work, sports and travel. In Germany, such measures eventually proved effective in reducing the number of fatalities and daily confirmed cases (incidence). This steady state w.r.t. the incidence rate over the summer was succeeded by a second wave of infections.

To be able to impose targeted, time-limited, strict mitigation policies, reliable estimates on how easily the virus is spreading under different scenarios are key. Hence, for pandemic mitigation to be effective, reliable forecasts are crucial. These are provided by data-driven models, which make predictions about the future by learning from the past.

However, at the beginning of the pandemic there is only a small amount of data available and additionally, little is known about epidemiological parameters such as the basic reproduction number. This is further complicated by changing protocols of data collection and low data quality in general. This makes generating reliable estimates challenging and can result in systematic and statistical errors in these initial stages of the pandemic. This is further complicated by the delay with which effects of policy interventions can be evaluated and the time-lag inherent in the reporting process yielding the daily reported case numbers. All of these factors result in the fact that decision making has to take place under large uncertainty.

This can only be done near-optimally if the resulting uncertainty can be quantified and possibly reduced by incorporating any prior knowledge, e.g. from similar diseases. This also includes information on factors such as the reporting delay and lag, the serial interval[1] (needed to estimate turnover of case generations and transmissibility),[2] the incubation period,[3] virulent window and proportion of transmissions occuring during the incubation period, as well as factors such as the latent number of infected, dead and recovered, etc. The need to obtain faithful estimates, which account for uncertainty, incorporate prior knowledge and integrate newly available data continuously, naturally leaves us with a Bayesian approach to model the dynamics of the coronavirus.

1.1 Contribution

In this work, we design a probabilistic model estimating the daily number of newly infected individuals with the novel coronavirus SARS-CoV-2 in Germany. In particular, we remove reporting-specific effects present in the data. We improve upon this model by also estimating the disease onset date with uncertainty, which serves as a proxy for the time of infection of a case, yielding a more accurate reflection of the pandemic's state for a given point in time. Finally, we propose a new metric, which reflects the trend of disease progression and may thus serve as a policy guiding tool.

[1] Serial interval: time from when one infected person starts showing symptoms to illness onset in a secondary case.

[2] The serial interval is estimated to have a median of 4.0 days which indicates rapid cycles from one generation to the next [3].

[3] Incubation period: time elapsed between exposure to virus and when symptoms start showing.

Chapter 2

Background

This chapter concerns itself with regression. We begin by introducing a basic linear model, point out its limitations and subsequently outline a way to circumvent these by projecting the inputs into a higher-dimensional feature space to instead perform inference there. As it turns out, performing linear regression in feature space naturally leads to a concept called **Gaussian Process** (**GP**) regression as discussed in Section 2.4.

For later reference, it may be useful to note that the following presentation of regression sometimes is referred to as the *weight space* or *parameter space* view, while the subsequent one using GPs in Section 2.4 is considered the *function space* view. This introduction mostly relies on standard textbooks on this topic, most notably Rasmussen and Williams [4], Bishop [5] and Murphy [6].

2.1 Introduction

We aim to find a relationship between the time and observed incidence, which would enable us to predict the future trend of incidence with uncertainty. To put it in a more formal way, observing some inputs X_i and outputs y_i, we assume $y_i = f\left(X_i\right)$ for some unknown function f. Inferring this function from the input-output pairs amounts to the regression problem that supervised learning is concerned with (next to classification problems which will not be considered here).

2.2 Standard Linear Model

A basic linear regression model, where the scalar output y_i is a linear combination of the input vector $\boldsymbol{x_i}$ and some weights $\boldsymbol{w}$, is easy to interpret but also limited in its expressiveness if the relationship between input and output can not be linearly approximated anymore.

From a Bayesian viewpoint on the standard linear regression model, we assume for some unknown function f, corrupted by Gaussian noise ϵ,

$$f\left(X\right) = X^T \boldsymbol{w}, \qquad y_i = f\left(X\right) + \epsilon, \tag{2.1}$$

where X is the input vector and $\boldsymbol{w}$ a vector of weights (parameters). If the observed values y were noise-free, our linear model should perfectly interpolate and predict $f\left(X\right)$ with zero uncertainty for a value of X that was already seen by the model, i.e., X_i is an element of the training set. Since we have assumed that the observed values y are a noisy version of the underlying function, however, the model is not expected to perfectly interpolate the data but only to come "close", i.e., with uncertainty. We assume the noise ϵ to be Gaussian distributed with zero mean and variance σ_n^2

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2). \tag{2.2}$$

Suppose we have a training set $\mathcal{D}$ of n observations with $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}$ where $\boldsymbol{x_i}$ denotes an input vector of dimension D and y_i corresponds to a scalar output or target (dependent variable). Then the column vector inputs for all n cases are aggregated in a $D \times n$ *design matrix* X, and we write $\mathcal{D} = (X, \boldsymbol{y})$.

In order to now perform Bayesian inference for a test case $\boldsymbol{x}_\star$, all possible parameter values are weighted by their posterior probability.

This is done by Bayes' rule

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}, \qquad p(\boldsymbol{w}|\boldsymbol{y}, X) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|X, \boldsymbol{w})}{p(\boldsymbol{y}|X)}, \tag{2.3}$$

where the evidence is independent of the weights $\boldsymbol{w}$. The evidence or normalization constant is also known as the marginal likelihood and given by

$$p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|X, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}. \tag{2.4}$$

Thus, the posterior in eq. (2.3) captures everything we know about the parameters by combining prior and likelihood. Conversely, in non-Bayesian settings, typically a single parameter is chosen by some criterion. The *likelihood* $p(\boldsymbol{y}|X, \boldsymbol{w})$ in eq. (2.3) is the probability density function (pdf) of y evaluated at $\boldsymbol{w}$.

When making predictions given new inputs $X_\star$ by applying Bayes' rule from eq. (2.3), we compute the predictive distribution $f_\star \triangleq f(X_\star)$, which amounts to averaging the output of all possible linear models w.r.t. the Gaussian posterior. The predictive distribution $f_\star$ is then given by

$$p(f_\star|X_\star, X, \boldsymbol{y}) = \int p(f_\star|X_\star, \boldsymbol{w})p(\boldsymbol{w}|X, \boldsymbol{y})d\boldsymbol{w}, \tag{2.5}$$

where again, $f_\star$ is the predictive distribution for some new inputs $\boldsymbol{x}_\star$ given the training data $(X, \boldsymbol{y})$. Since we assumed everything to be Gaussian distributed, the posterior distribution is Gaussian as well and is parametrized by a mean given by the posterior means of the weights multiplied by the test input and a predictive variance.

2.3 Feature Space

The Bayesian linear model lacks expressiveness in the case where input and output are non-linearly related. A way to circumvent this limitation is to first project the inputs into a high dimensional *feature space* by using a set of basis or feature functions $\phi(X)$. Subsequently, we can apply the linear model in the feature space rather than on the inputs themselves. A plethora of feature functions ϕ_x may be used, such as the Switch-, Pixel-, Fourier- Bell Curve, etc., resulting in different piecewise interpolants and therefore in various regression models. For instance, a polynomial feature function $\phi(X) = (1, X, X^2, X^3, ...)^T$ projects an input X into the space of powers resulting in polynomial regression.

For this section we rely on the lecture *Probabilistic Machine Learning* by Prof. Philipp Hennig [7]. The described model is still linear in the parameters, hence computationally tractable, as long as the projection functions independent of the parameters $\boldsymbol{w}$. By introducing the feature function $\phi(X)$ that projects a $D-$dimensional input vector X into an D'-dimensional feature space, the linear model $f\,(X) = X^T\boldsymbol{w}$ from eq. (2.1) now becomes

$$f\,(X) = \phi(X)^T\boldsymbol{w}, \tag{2.6}$$

where the parameter vector now has length D'. The analysis of this model is still analogous to the standard linear model described in eq. (2.1), except that all X are being substituted by $\phi(X)$ now.

If we assign a prior Gaussian distribution over the weights $\boldsymbol{w}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ, we can write the posterior predictive distribution as

$$\begin{aligned} p(f_\star|\boldsymbol{y}, \phi_X) = \mathcal{N}(f_\star;\, & \phi_\star^T\mu + \phi_\star^T\Sigma\phi_X(\phi_X^T\Sigma\phi_X + \sigma_n^2 I)^{-1}(y - \phi_X^T\mu), \\ & \phi_\star^T\Sigma\phi_\star - \phi_\star^T\Sigma\phi_X\,(\phi_X^T\Sigma\phi_X + \sigma_n^2 I)^{-1}\phi_X^T\Sigma\phi_\star), \end{aligned} \tag{2.7}$$

where $\phi_{\star,X} \triangleq \phi(X_\star, X)$.

Note that ϕ is always part of an inner product. Hence, we may define two abstract functions encapsulating the two operations in which ϕ is included respectively (color-coded by blue and red).

Now let

$$m_\star := \phi_\star^T\mu \tag{2.8}$$

be the *mean function*, where $\phi_\star^T$ denotes features of test data (ϕ_X^T respectively denotes features of training data), and μ is the mean of weights $\boldsymbol{w}$.

Let further

$$k_{ab} \coloneqq \phi_a^T \Sigma \phi_b \tag{2.9}$$

be the *covariance function* or **kernel.**[1] Then we can rewrite the posterior predictive distribution as

$$\begin{aligned} p(f_\star | \boldsymbol{y}, \phi_X) = \mathcal{N}(f_\star; m_\star + k_{\star X}(k_{XX} + \sigma_n^2 I)^{-1}(y - m_X), \\ k_{\star X} - k_{\star X}(k_{XX} + \sigma_n^2 I)^{-1} k_{X\star}). \end{aligned} \tag{2.10}$$

Note that the kernel k_{ab} in eq. (2.9) is replacing the inner product between two feature vectors ϕ_a and ϕ_b weighted by the covariance Σ of the weights, which also is referred to as the *kernel trick* [5]. Inner products are sums $\langle \phi_a, \phi_b \rangle \coloneqq \phi_a^T \phi_b = \sum_i^F (\phi_a)_i (\phi_b)_i$. Some feature functions $\phi_j(x_i)$ with $j = 1, ..., F$ and $i = 1, ..., N$, where F is the number of features and N the number of input data, can even be evaluated when $F \to \infty$. This is because in that case, these sums turn into integrals with analytical solutions. However, this requires some regularity assumptions about the features' shape, locations, etc.

When computing the kernel rather than the feature vectors themselves is cheaper, then the kernel trick is particularly valuable. This notion sets the stage for the concept of **Gaussian processes** (**GPs**), which we will shed more light on in the next Section 2.4.

2.4 Gaussian Process Regression

In the previous section we observed that sometimes it is possible to consider infinitely many features. We can then learn infinitely many features in parallel in closed form using the kernel operation. In that sense, we learn $p(\theta|\mathcal{D})$, as opposed to inferring $p(f|\mathcal{D})$, which we did previously when we focused on parametric representations of a function f_θ. Here, θ denotes some latent distribution parameters and $\mathcal{D}$ denotes the observed data. More generally, if the unknown variable or parameter is not a scalar or a fixed-length vector, but a function, we can perform Bayesian inference over functions themselves. The resulting (nonparametric) process is referred to as a **Gaussian process (GP)** if any finite projection is a Gaussian random variable. For this introduction on GPs, we again rely primarily on Bishop [5], Rasmussen and Williams [4] and Murphy [6].

[1] The terms *covariance function* and *kernel* will be used interchangeably throughout this thesis.

2.4.1 Introduction

In Gaussian process regression (GPR) we aim to find a function that is close to the latent, unknown function f that generated the data and that generalizes well. In this case, a GP is used as a prior probability distribution whose samples are continuous functions. Function values are then modeled as a draw from a multivariate normal distribution. GPs are a particularly convenient choice due to the analytic marginalization and conditioning properties innate to the multivariate normal distribution, which we will elaborate on in Sections 2.4.3 and 2.4.4 when discussing the inference and prediction step.

2.4.2 Definition

For our purposes a GP defines a prior over functions, which can be transformed into a posterior over functions after the model is presented with data. While one may think of a GP as describing a probability distribution over functions, the definition only refers to a *finite* number of evaluations. A Gaussian process is a random process, which, if evaluated at a number of input points $\boldsymbol{x_1}, ..., \boldsymbol{x_N}$, the function values $f(\boldsymbol{x_1}), ..., f(\boldsymbol{x_N})$ are jointly Gaussian distributed. We then write a GP as

$$f(X) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})), \tag{2.11}$$

solely parametrized by the mean function $m(\boldsymbol{x})$ and the covariance function $k(\boldsymbol{x}, \boldsymbol{x'})$, which is assumed to be positive semi-definite (see also eq. (2.8) and eq. (2.9)).

2.4.3 Inference

Inference refers to the process of fitting the underlying GP to the given training data. By definition, a GP induces a joint distribution, here for *noise-free* observations $\boldsymbol{y} = f(X)$ and $\boldsymbol{y_\star} = f(\boldsymbol{x_\star})$, which is a multivariate normal,

$$\begin{bmatrix} f \\ f_\star \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} m(x) \\ m(x_\star) \end{bmatrix}, \begin{bmatrix} k(X, X) & k(X, X_\star) \\ k(X_\star, X) & k(X_\star, X_\star) \end{bmatrix} \right), \tag{2.12}$$

where f are training and $f_\star$ test outputs, and $K(\cdot, \cdot)$ denotes the respective sub-matrix of the covariance given by the kernel. That is, $K(X, X_\star)$ defines the $N \times N_\star$ covariance matrix evaluated at all pairs of training points N and test points $N_\star$. The same notion applies to the remaining entries $K(X_\star, X)$, $K(X_\star, X_\star)$ and $K(X, X)$. Typically, the kernel function is chosen to express the property that, for any arbitrary points $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ which are similar, the corresponding output values $y(\boldsymbol{x_a})$ and $y(\boldsymbol{x_b})$ will be more correlated as well (see Figure Figure 2.1). Note, however, that the notion of *similarity* mainly depends on the context that GPR takes place in and may hence differ.

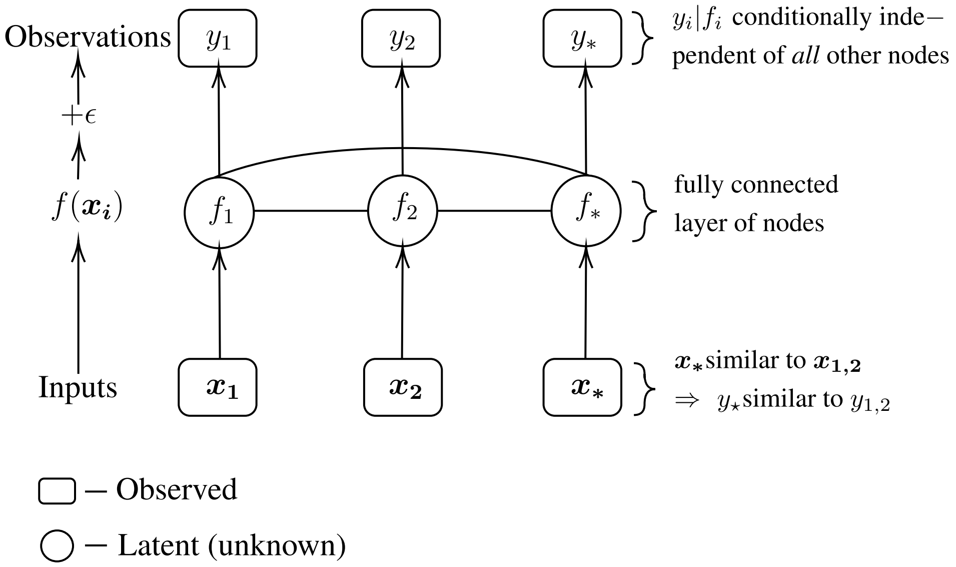

Observations
y_1
y_2
y_*
$y_i|f_i$ conditionally inde– pendent of *all* other nodes
$+\epsilon$
$f(\boldsymbol{x_i})$
f_1
f_2
f_*
fully connected layer of nodes
Inputs
$\boldsymbol{x_1}$
$\boldsymbol{x_2}$
$\boldsymbol{x_*}$
$\boldsymbol{x_*}$similar to $\boldsymbol{x_{1,2}}$
$\Rightarrow$ $y_\star$similar to $y_{1,2}$
— Observed
— Latent (unknown)

Figure 2.1: *Graphical model for GPR.* Graphical model (chain graph) for two training points $\boldsymbol{x_{1,2}}$ and one test point $\boldsymbol{x_\star}$. Adapted from [4], p. 17.

In order to predict $f_\star$ at new locations $\boldsymbol{x_\star}$, we need to first learn the latent function f, the data generating process for the observed data $\boldsymbol{y}$. We obtain the marginal likelihood (or evidence) $p(\boldsymbol{y}|X)$ by integrating over the unknown latent function f. The marginal likelihood then is likelihood times the prior

$$p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|X, \boldsymbol{f})p(\boldsymbol{f})\ d\boldsymbol{f}. \tag{2.13}$$

Conveniently, the corresponding marginal distribution can directly be read off of the joint distribution in eq. (2.12). Therefore

$$\boldsymbol{y}|X \sim \mathcal{N}(m(X), K). \tag{2.14}$$

We typically only have access to *noisy* observations $y = f(X) + \epsilon$. Supposing additive i.i.d. Gaussian noise ϵ with variance σ_n^2, the prior covariance for noisy observations then becomes

$$K = k(X, X) + \sigma_n^2 I, \tag{2.15}$$

where I is the unit matrix.

The joint distribution of the observed noisy target values $\boldsymbol{y}$ and the function values $f_\star$ at new input locations $X_\star$ then becomes

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{f}_\star\end{bmatrix} = \mathcal{N}\left(\begin{bmatrix}m(X)\\ m(\boldsymbol{x}_\star)\end{bmatrix}, \begin{bmatrix}k(X,X)+\sigma_n^2 I & k(X,X_\star)\\ k(X_\star,X) & k(X_\star,X_\star)\end{bmatrix}\right), \tag{2.16}$$

An example of a latent function f for a synthetic data set is shown in Figure 2.2.

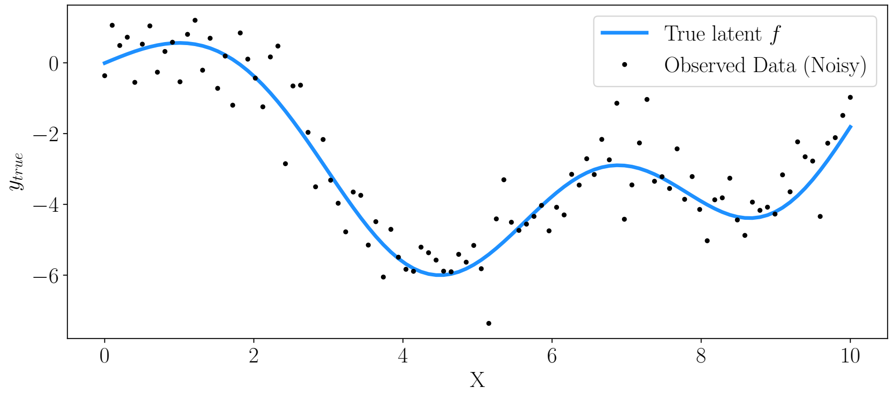


Figure 2.2: *Inference using a GP.* Observations y are the unknown function f plus some i.i.d. Gaussian distributed additive noise. The latent function was generated by a Gaussian Process with an RBF kernel.

Here, the observed data y are the sum of a GP with a radial basis function kernel (RBF) as covariance function and Gaussian noise. Observe that the RBF covariance function k_y explicitly refers to the noisy targets y (and not, say, k_f, referring to the signal) and is then given by

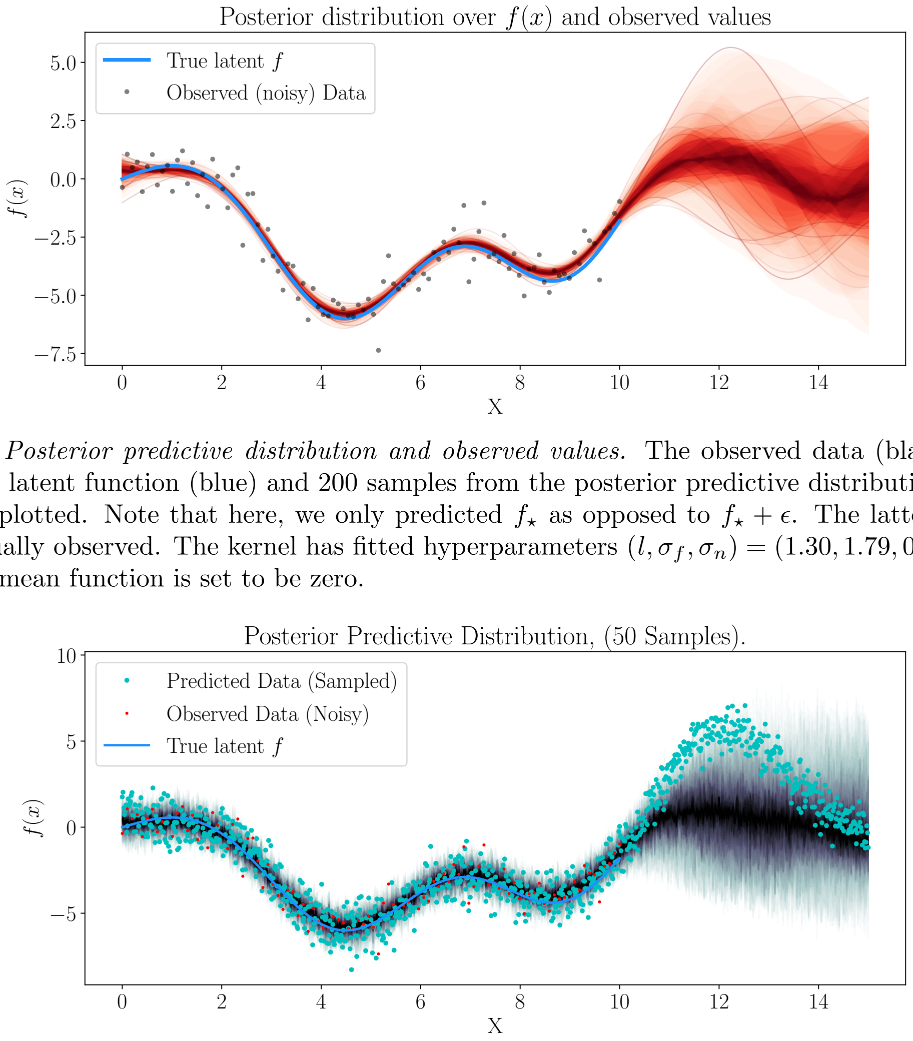
$$k_y(x_a, x_b) = \underbrace{\overbrace{\sigma_f^2}^{\text{signal variance}} \exp(-\frac{\|x_a - x_b\|^2}{2l^2})}_{\text{Radial Basis Function kernel } k_{RBF}(x_a,x_b)} + \underbrace{\sigma_n^2 \delta(x_a, x_b)}_{\text{Gaussian noise kernel } k_\epsilon(x_a,x_b)} . \tag{2.17}$$

Note that the lengthscale l, the signal variance σ_f and the noise variance σ_n can be varied and are called *hyperparameters*. We will expand on this notion in Section 2.4.5.

It turns out that inference in the regression setting of GPs has polynomial cost of $\mathcal{O}(N^3)$ in the number of data points N, which is due to inversion of an $N \times N$ kernel matrix (see eq. (2.7)). While inference is computationally tractable, the polynomial running time is still one principal drawback of GPs. While this is a valid concern for some applications (e.g. autonomous driving, online trading, etc.), this will not affect our use case of GPs we do not require real-time decision, which would necessitate fast computation. Rather, we simply exploit a GP's ability to approximate *any* function arbitrarily well via its posterior mean, given enough data [7].Since a GP's predictive performance depends on the suitability of the chosen kernel, we will take some time to explore different kernels and combinations thereof in Chapter 3. One can create new kernels from old ones since we can take the product and sum of kernels as well as scale the kernel's input and output [5]. This property gives rise to a powerful modelling language which allows for very expressive models while keeping computations tractable.

2.4.4 Prediction

Now that our model is fit to the data, we can predict new values $y_\star$ at new input locations $\boldsymbol{x}_\star$. Using the conditional distribution, we obtain the predictive distribution for the underlying function represented by the GP, inducing the conditional mean and variance. Figure 2.3 depicts the posterior predictive distribution using the GP defined above. Figure 2.3a shows the posterior predictive distribution $f_\star$ without noise, while Figure 2.3b depicts what we actually observe, namely $f_\star + \epsilon$ and additionally new data points.



(a) *Posterior predictive distribution and observed values.* The observed data (black), the true latent function (blue) and 200 samples from the posterior predictive distribution (red) are plotted. Note that here, we only predicted $f_\star$ as opposed to $f_\star + \epsilon$. The latter one is actually observed. The kernel has fitted hyperparameters $(l, \sigma_f, \sigma_n) = (1.30, 1.79, 0.77)$ and the mean function is set to be zero.

(b) *Posterior distribution of the data generating process.* In addition, we now predict new data points (shown as light colored dots) drawn from the posterior predictive distribution. Note that the posterior predictive density is wider than the conditional distribution of the noiseless function depicted in Figure 2.3a. This reflects the predictive distribution of the noisy, observed data (red). Further note, that the new points do not follow the spread of the predictive density exactly since they are a single draw from the GP's posterior plus noise.

Figure 2.3: *Posterior Predictive Distribution.*

Finally, we can plot the posterior mean and 2σ uncertainty region which is shown in Figure 2.4. Observe that the further we move away from observed data, the wider the uncertainty. It may be worth highlighting, that the predictive uncertainty is independent of the data (since we assumed i.i.d. additive Gaussian noise ϵ on each individual observation y_i) setting the stage to two, somewhat orthogonal properties. On the one hand, this allows for measurements $\boldsymbol{y}$ to be placed in a way such that the aggregated predictive uncertainty becomes fairly small for a lot of corresponding locations X which might be desired in an experimental context. On the other hand, since the predictive uncertainty does not adapt to the data, the model does not reliably estimate how well it fits to the test data.

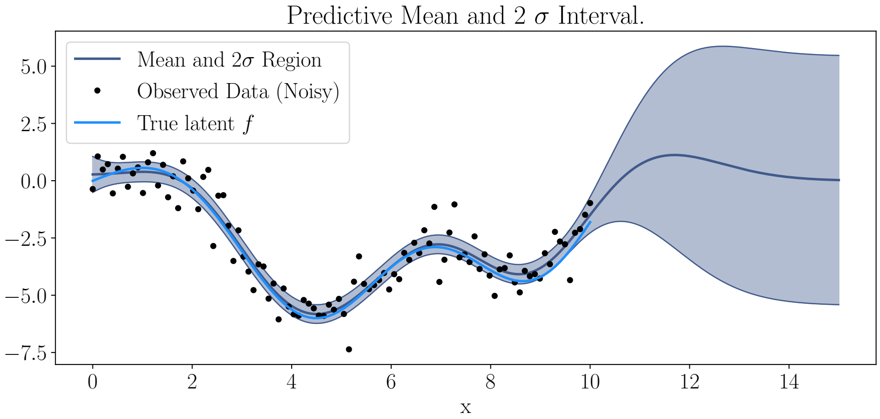


Figure 2.4: *Posterior mean and 2σ posterior credible interval.*

2.4.5 Kernel Parameters

In Section 2.4.4, we encountered parameters, that govern things such as the length scale or amplitude of a kernel which are also referred to as *hyperparameters*. Since the predictive performance of GPs depends on the choice of the kernel, it is hence crucial to have well chosen hyperparameters.

In the following, we discuss approaches to estimate the parameters by *maximum likelihood* (MLE type I), *maximum a posteriori* (MAP) and MLE type II.

MLE refers to maximizing the likelihood $p(\boldsymbol{y}|f, X, \boldsymbol{\theta})$ as a function of the given parameters θ. By minimizing the negative log of the likelihood using any standard gradient-based optimizer, we obtain estimates for optimal $\boldsymbol{\theta}$. One problem with this approach is, however, that the optimizer may only find local minima, thus possibly not translating into globally optimal $\boldsymbol{\theta}$. The MLE for $\boldsymbol{\theta}$ is prone to overfitting, that is, the function is fit too closely too the training data and consequently, will not generalize well to new input points. This can manifest itself in extreme parameter values found by the optimizer.

In order to mitigate this effect, a prior distribution $p(\boldsymbol{\theta})$ can be placed over the hyperparameters. This distribution then explicitly constrains the range of values that the hyperparameters can take. Instead of maximizing the likelihood, we maximize the posterior $p(\boldsymbol{\theta}|X, \boldsymbol{y})$. This procedure is called *maximum a posteriori* (MAP) estimation.

Alternatively, one can also use MLE type II. This amounts to maximizing the evidence in Bayes' theorem eq. (2.3), also referred to as the model's marginal likelihood

$$p(\boldsymbol{y}|\theta, X) = \int p(\boldsymbol{y}|\theta, \boldsymbol{f}, X)p(\boldsymbol{f}|\theta, X)d\boldsymbol{f}. \tag{2.18}$$

2.4.6 Source Separation

In the following section, we briefly digress and introduce the concept known as *source separation*, also referred to as *additive decomposition*. Here we rely on Duvenaud [8]. We will make use of this concept later on in this thesis, particularly in Section 3.2.2.

One convenient property of constructing an additive kernel, as in eq. (2.17), for instance, is that we can decompose our posterior into additive parts. That is, if our kernel is a sum $k_{sum}(\boldsymbol{x}, \boldsymbol{x'}) = k_1(\boldsymbol{x}, \boldsymbol{x'}) + k_2(\boldsymbol{x}, \boldsymbol{x'}) + \cdots + k_D(\boldsymbol{x}, \boldsymbol{x'})$, then our posterior can equally be decomposed into a sum of GPs, each with mean

$$m(f_d(\boldsymbol{x}_\star)) = k_d(\boldsymbol{x}_\star, X)K_{sum}(X, X)^{-1}f(X) \tag{2.19}$$

and variance

$$\begin{aligned} \text{cov}(f_d(\boldsymbol{x}_\star), f_d(\boldsymbol{x}_\star)) &= \text{var}(f_d(\boldsymbol{x}_\star)) \\ &= k_d(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - k_d(\boldsymbol{x}_\star, X)K_{sum}(X, X)^{-1}k_d(X, \boldsymbol{x}_\star), \end{aligned} \tag{2.20}$$

where k_d and f_d denote the corresponding decomposed entities.

Note, however, that this concept is constrained to *additivity* which is lost under any non-linear transformation of the output.

Chapter 3

Probabilistic Model

In order to assess the status of the pandemic and to derive the best possible informed decisions on mitigation policies, an estimate on the number of infected persons on a given day is an important figure. However, these numbers can only be estimated by counting backwards from what we actually can measure, that is the daily number of people tested positive for the virus, also known as incidence. Consequently, for the estimate to be as accurate as possible, several assumptions based on what we know about the data generating process must be incorporated into the model.

By its nature, one can never be 100% certain about these assumptions. This is why the models presented in this work consider many possible options while assigning different probabilities to these scenarios based on their likelihood. In a nutshell, we aim to infer the most likely curve depicting the latent incidence given the data we observe.

In what follows, we will present different models increasing in complexity in an attempt to estimate the latent number of infected people on a given day. This leaves us with an estimate with uncertainty about the incidence and hence provides some information about the current status of the pandemic. Based on this information, it is of great importance to act accordingly as a society. For this reason, we subsequently derive a metric called D-value, which assesses the current trend in transmission rates from which potential mitigation strategies can be derived by policy makers.

3.1 Data

We use daily counts of confirmed cases published by the Robert Koch Institute (RKI) [9], Germany's central scientific body, in charge of federal health reporting and of epidemiological evaluation of contagious diseases. An excerpt of the data published by the RKI is depicted in Table 3.1.

3.1.1 COVID-19-Reporting in Germany

In Germany, in accordance with the obligation to notify under the Infection-Protection Act (IfSG)[1], COVID-19 infections are reported to the local public health department in the respective districts. Subsequently, the data are transmitted to the responsible federal state health authorities. Each day by midnight the total number of infections per federal district reported to the respective health departments is transmitted to the RKI. Thereby the RKI evaluates all laboratory diagnostic evidence of SARS-CoV-2 as COVID-19 cases, regardless of presence and severity of clinical symptoms, which is in accordance with the international standards of the World Health Organization (WHO) and the European Center for Disease Control (ECDC). Hence, the number of COVID-19 cases summarizes both acute SARS-CoV-2 infections and COVID-19 diseases. In what follows, we provide a brief overview on important dates involved in the reporting process and the progression of the disease caused by the virus as can be seen in Figure 3.2.

Reporting Date. For the daily published number of COVID-19 cases, the *reporting date* is used, which refers to the date on which the local health department became aware of the case and recorded it electronically. The RKI uses the reporting date to display newly submitted cases per day on its dashboard [10] a screenshot of which can be seen in Figure 3.1.

Notification Delay. A few days may elapse between the notification by doctors and laboratories to the local health departments and the transmission of the cases to the responsible state authorities and the RKI. The number of new cases received by the RKI each day may have been reported to the health departments on the same day or on earlier days. Thus, the difference in reported cases w.r.t. the previous day may be distributed over multiple days. This is illustrated on the COVID-19 dashboard [10] *COVID-19 cases by date of report* by the orange bars, which represent the newly reported cases. (see Figure 3.1).

Time of Infection. The exact time of infection of the reported cases can usually not be be determined. The RKI states that the reporting date therefore best reflects the time the infection was detected and hence the current infection rate in a given region [10]. This is why we choose the daily counts by reporting date as an initial proxy to model the daily infections of COVID-19 cases. Additionally, we will consider the reporting delay in order to approximate the date of infection to ultimately model the latent ongoing transmission of infections more accurately.

[1] `https://www.rki.de/DE/Content/Infekt/IfSG/ifsg_node.html`

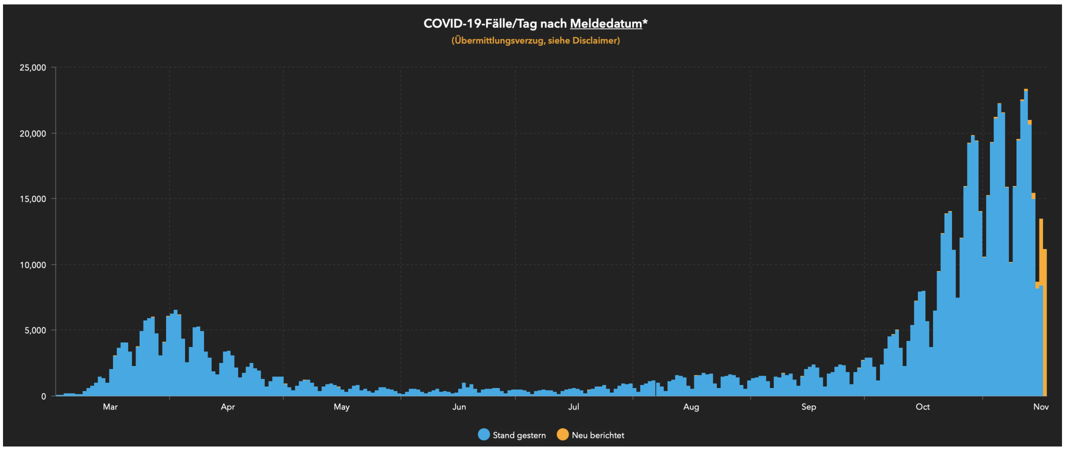


Figure 3.1: *Screenshot of the RKI dashboard.* Depicted are daily counts of COVID-19 cases by reporting date. As a first step, we use these numbers to approximate the daily number of infections of COVID-19 in Germany. Blue represents the status of the previous day's numbers, while orange depicts the newly submitted cases for today. Note that the newly reported cases (orange) can be distributed over multiple days.

Time of Disease Onset. The date of disease onset is the date on which the patient claims to have become ill with clinical symptoms according to their own information or an estimate by the treating physician. According to an der Heiden and Hamouda [11], the disease onset date may be the best way to approximate the time of infection since comparably reliable estimates on the incubation period are available, on average five to six days, which can be as high as 14 days, however.[2] According to the Centers for Disease Control and Prevention (CDC) in the U.S. the infectious period lasts for up to ten days following symptom onset.[3]

Estimate of Recovery Date. Based on the disease onset date (or if not known, based on the reporting date), an estimated recovery date results for each case. However, since disease progression may vary significantly, and the disease onset date is only known for about $\sim 60 - 70\%$ of the cases, the recovery dates are to be considered only rough estimates with the respective limitations taken into account. This date is especially of interest when trying to assess the degree of immunity present in a population (assuming recovery from disease prevents a reinfection), which will ultimately give rise to more accurate forecasts (see also compartmental models, e.g., *SEIR*).

It is important to note that the case numbers by reporting date do not

[2] https://www.rki.de/SharedDocs/FAQ/NCOV2019/gesamt.html

[3] https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html, last accessed: November 16, 2020

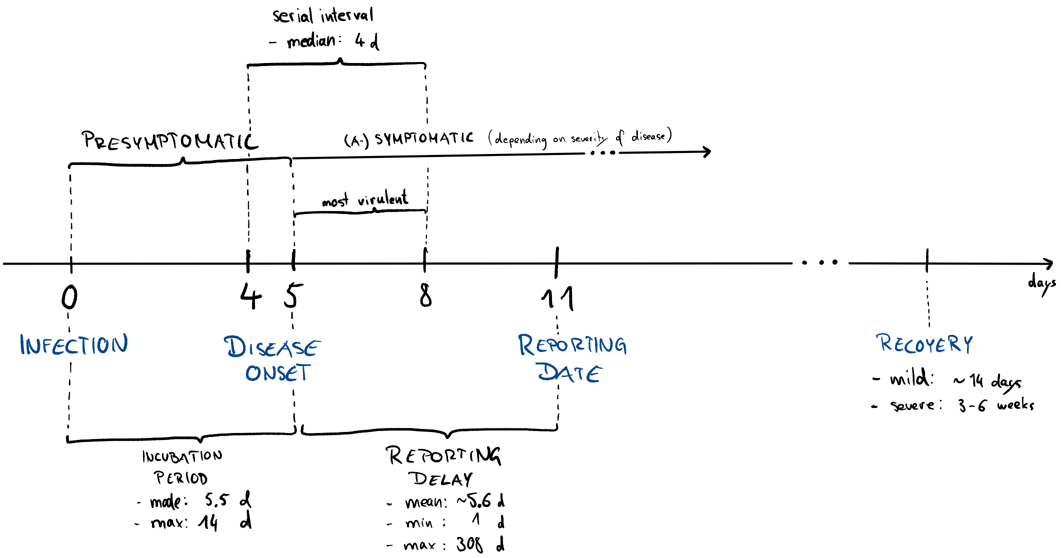


Figure 3.2: *Timeline of COVID-19.* Depicted are critical time ranges (in days) of COVID-19 regarding time of infection, disease/symptom onset, serial interval, virulent window, reporting and recovery date. Numbers come from our own analysis, the RKI, WHO and CDC.[4]

reflect the actual temporal progression of COVID-19-transmissions since the time intervals between symptom onset and disease diagnosis, reporting, as well as data transmission to the RKI varies greatly.

The issues of reporting delay and delays between symptom onset and infection are going to be accounted for in a second step in which we approximate the actual number of infected persons for a given day. Generally, it is important to bear in mind, that results presented here are sensitive to changes in testing practices and the degree of effort put into detecting cases, e.g. through contact tracing.

3.1.2 Data Processing

The data in its raw form is given as a csv file,[5] of which an extract is depicted in Table 3.1.

In a first step, we process the data such that we obtain a time-series for Germany as a whole (national level) by *date of report*. For sub-national analysis, we equally create a time-series for each of Germany's 16 federal states and 412 counties and districts.

Next, we create a time-series by *symptom-onset-date*[6] for each geographic

[5] `https://www.arcgis.com/sharing/rest/content/items/f10774f1c63e40168479a1feb6c7ca74/data`

[6] Date by symptom onset and disease onset are used interchangeably throughout this

Bundesland	Landkreis	AnzahlFall	AnzahlTodesfall	Meldedatum	NeuerFall	Refdatum	NeuGenesen	IstErkrankungsbeginn
Schleswig-Holstein	SK Flensburg	1	0	2020-09-30	0	2020-09-30	0	0
Schleswig-Holstein	SK Flensburg	1	0	2020-10-29	0	2020-10-29	-9	0
Schleswig-Holstein	SK Flensburg	1	0	2020-08-24	0	2020-08-24	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Baden-Württemberg	SK Stuttgart	3	0	2020-10-30	1	2020-10-30	-9	0
Baden-Württemberg	SK Stuttgart	4	0	2020-10-31	1	2020-10-31	-9	0
Baden-Württemberg	SK Stuttgart	1	0	2020-10-22	0	2020-10-22	-9	0
Baden-Württemberg	LK Böblingen	1	0	2020-03-24	0	2020-03-24	0	0
Baden-Württemberg	LK Böblingen	1	0	2020-04-05	0	2020-04-05	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Thüringen	LK Altenburger Land	2	0	2020-10-29	0	2020-10-29	-9	0
Thüringen	LK Altenburger Land	1	0	2020-10-30	0	2020-10-26	-9	1
Thüringen	LK Altenburger Land	1	0	2020-10-30	0	2020-10-30	-9	0

Table 3.1: *Excerpt of the raw data provided by the RKI.* Depicted are the columns used. Also provided by the RKI are the columns *FID, IdBundesland, Altersgruppe, Geschlecht, IdLandkreis, Datenstand, NeuerTodesfall, AnzahlGenesen* and *Altersgruppe2*, which are omitted here.

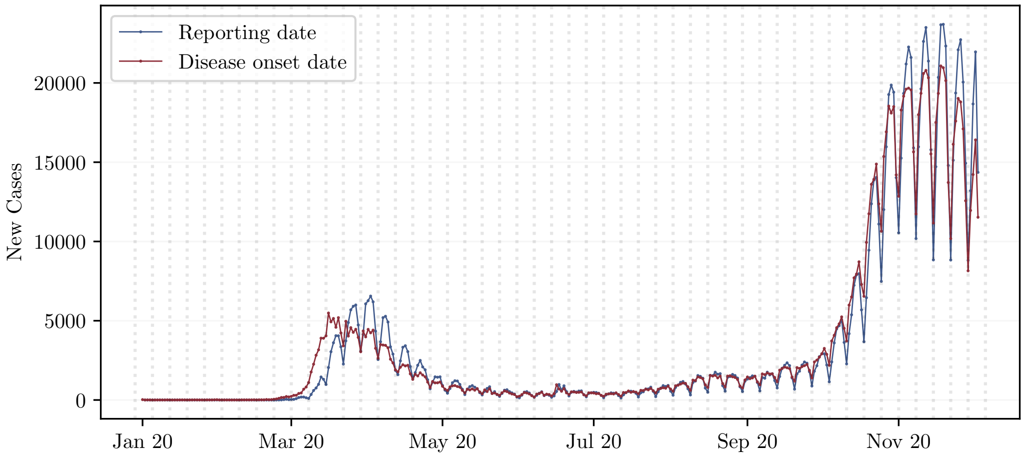


Figure 3.3: *Cases by reporting and disease onset date.* The disease onset date is not known for all notified cases. On November 4, 2020, the disease onset date was known for $\sim 67.09\%$. Additionally, the mean number of days between disease onset date and date of report for a given day varies over the course of the pandemic. While during the first wave in March and April a significant shift can be observed, this shift seems to have decreased as the pandemic has progressed. Moreover, for the latest seven to ten days, significantly fewer cases by disease onset date are known. This is due to the time-delay resulting from the point in time a test-positive case is notified to the RKI in the course of a case starting from infection to disease onset date to getting tested to being reported.

entity, which will serve as a better proxy for the *date of infection*. However, this date is only being reported for $\sim 67.09\%$[7] of the notified cases. The analysis of how the missing dates are estimated and all further analyses regarding data imputation are described in Section 3.2.3.

3.1.2.1 Data Transformation

We transform the data to a latent space in which we can make approximate normality and linearity assumptions via

$$y(t) = T(x) = \log(x + 1) \tag{3.1}$$

with inverse $T^{-1}(y) = \exp(y) - 1$. This also allows for modelling the heteroscedastic variance present in the data. For instance, Figure 3.1 illustrates that the variance in reported cases clearly is not constant over time.

thesis.

[7] As of November 4, 2020

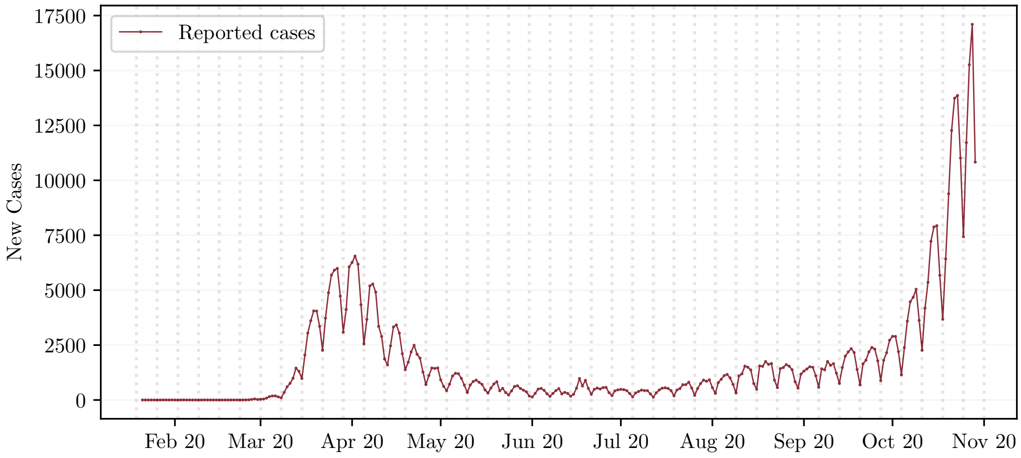


Figure 3.4: *Day-of-the-week-effect.* Daily confirmed cases by date of report. A day of the week-effect becomes apparent by Sundays, marked by dashed vertical lines, matching the local minima observed in the reported case numbers.

We standardize each time-series in time

$$y'(t) = l(y(t)) = \frac{y(t) - \text{mean}_t(y(t))}{\text{std}_t(y(t))}$$

resulting in the following data transformation

$$z(t) = (l \circ T)(x(t)). \tag{3.2}$$

Figure 3.5 depicts $z(t)$ from eq. (3.2) applied to the the daily confirmed cases by reporting date in Germany.

3.2 Time-series Modelling using Gaussian Processes

In this section, we perform GP regression for time-series modelling of COVID-19 cases in Germany. We are interested in the incidence, that is, the number of newly infected people per day. However, the disease onset date is only known for $\sim 60 - 70\%$ of the cases reported to the RKI, which is why we first base our model on cases by date of report. Consequently, when interpreting the model's predictions, the time delays innate to the reporting process must be considered. In particular this means that the resulting estimate does not represent a timely reflection of the incidence but rather a shifted one. This shift can be approximated by the number of days that on average lie between reporting date and disease onset date, which can be derived from the cases

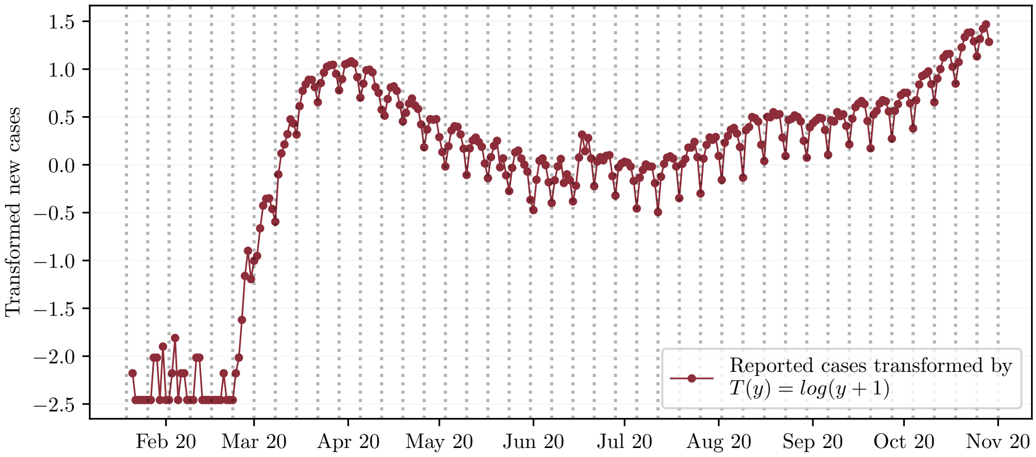


Figure 3.5: *Data transformation* $z(t)$ from eq. (3.2) applied to daily cases by reporting date in Germany. Vertical dashed lines represent Sundays, illustrating the day-of-the-week effect present in the data.

for which both dates are known. In Section 3.2.3.1 and Section 3.2.4 two different approaches to estimating the disease onset date are outlined. Finally, in Section 3.2.5.3 we derive an alternative metric to the R-value, called D-value.

3.2.1 Modelling Confirmed Cases by Date of Report

When considering the number of confirmed cases per day by reporting date (see Figure 3.4), two distinct patterns can be derived:

- a **latent trend**, approximately reflecting the progression over time of COVID-19 cases and
- a **periodic trend**, suggesting a day-of-week effect.

It turns out that the local minima to be observed in the periodic trend match to Sundays, which can be seen in Figure 3.4. This suggests that the cases to be considered by reporting date depict a poor reflection of the true underlying transmission process of the virus but rather reflect the way the reporting system is set up (e.g., fewer laboratories, local health departments etc. are actively reporting on Sundays). Derived from this observation, we expect the latent incidence to follow the trend of the curve of reported numbers (see Figure 3.4), which is based on the assumption that the number of people getting infected by the virus is distributed uniformly across weekdays. Further, due to the reporting delay, the most current data available only reflect the transmission rate from approximately ten days ago [11, 3].

One may conclude that a 10-day forecast represents the *current* transmission situation more accurately. As a first step, however, we aim to "model-away" the day-of-the week effect w.r.t. date of report which is why it is important to consider the limitations when interpreting the results.

We do this by assuming that in the transformed space, the reported cases constitute the sum of a latent, smooth trend, and a periodic trend, reflecting the day-of-the week effect. Re-transformed, the latent-underlying component should serve as a first approximation for the actual incidence.

3.2.1.1 Model specification

As a first approach, we assume the following Gaussian process model for the data.

$$z(t) \sim \mathcal{GP}(0, k(t,t')) \tag{3.3}$$

Our prior knowledge about the periodic reporting behavior is encoded in a sum kernel of the form

$$k(t,t') = k_{\text{latent}}(t,t') + k_{\text{weekly}}(t,t') + \varepsilon\delta(t,t')$$

where k_{latent} aims to model the latent trend in reported cases, while k_{weekly} aims to account for the day-of-week-effect, which we choose a locally periodic kernel for:

$$k(t,t') = k_{\text{latent}}(t,t') + \underbrace{k_{\text{latent}}(t,t')k_{\text{periodic}}(t,t')}_{\text{locally periodic kernel}} + \varepsilon\delta(t,t') \tag{3.4}$$

We choose a non-stationary kernel in the form of a locally periodic kernel because clearly the mean and covariance of the data are non-constant over time. Rather, we can observe the variance shifting over time, i.e. the data is heteroscedastic. In particular, the variance might be bigger when test capacities are being pushed to their limits. This might be especially true either at the beginning of the pandemic, when test facilities have not been properly set up yet, or in the event of another surge in infections happening, which would exceed available test capacities.

Our model is trained on daily counts of confirmed cases in Germany starting with the first reported case on January 21, 2020 and ending on September 30, 2020. Data from October 1, 2020 onwards serves as test data to evaluate our model's predictive performance.

3.2.1.1.1 Kernel Hyperparameters. We assume the following prior distributions for the kernel hyperparameters (see Appendix B) based on our prior knowledge:

- Radial Basis Function (RBF) kernel k_{RBF} modelling the latent trend:

$$k_{\mathrm{RBF}}(t,t') = \underbrace{\overbrace{\sigma_f^2}^{\text{signal variance}} \exp(-\frac{\|t-t'\|^2}{2l^2})}_{\text{RBF}} + \underbrace{\sigma_n^2\delta(t,t')}_{\text{Gaussian noise}} . \tag{3.5}$$

  - Length-scale l: We expect the number of cases to vary over a period of two to four weeks, which is why we assume a prior on the lengthscale $l \sim \mathcal{N}(20, 4)$.
  - Scale factor: output variance σ^2, which determines the average squared distance of our function from its mean (in latent space, i.e. transformed reported cases).

- Locally Periodic Kernel k_{localPer} modelling the day-of-the week effect:

  - To account for the periodic pattern in the data, we choose a locally periodic kernel, constituted by the product of of a periodic and an *RBF*-kernel. As opposed to a standard periodic kernel, a locally periodic kernel allows to account for the heteroscedasticity.
  - We place a very informative prior on the periodicity by fixing the period p to 7 since we assume the periodic pattern to be repeated on a weekly basis.
  - The RBF kernel parameters are identical to the ones simulating the latent trend for the same reason except for σ being set to 2 since we assume the latent trend to dominate the pattern in the data.

$$k_{\mathrm{localPer}}(t,t') = k_{periodic} \cdot k_{\mathrm{latent}} = \sigma_f^2 \underbrace{\exp(-\frac{2sin^2(\pi\|t-t'\|/p)}{l^2})}_{\text{periodic}} \underbrace{\exp(\frac{-\|t-t'\|^2}{2l^2})}_{\text{RBF}} \tag{3.6}$$

.

3.2.1.2 Predicted Case Numbers:

After having trained the model by fitting the hyperparameters via MAP, for which the exact values can be found in Table A.1, we obtain the following predictions for the case numbers after applying the inverse transformation T^{-1} as shown in Figure 3.6.

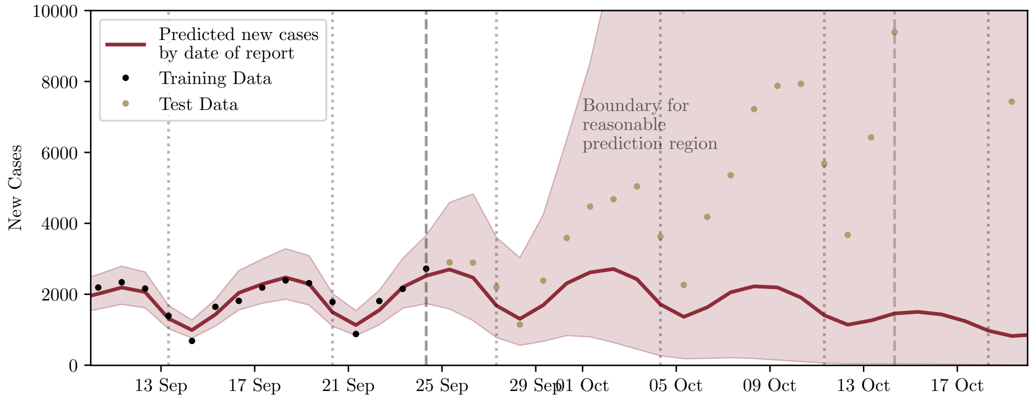


Figure 3.6: *Predicted reported case numbers.* Posterior mean of the model and 95%-CI are shown. Extrapolation (i.e. predicting into the future) is only significantly influenced by the data approximately l days away from the data, which, in this case amounts to a reasonable prediction range of $\sim$ 16 days, marked by a dashed line.

3.2.2 Source Separation / Additive Decomposition

For our model we chose a sum of GPs, where the corresponding kernels define the additive model given by

$$z(t) = l \circ T(x(t)) = f_{\text{latent}}(t) + f_{\text{locPer}}(t) + f_{\text{noise}}(t),$$

where $f_{\text{latent}}(t) \sim \mathcal{GP}(0, k_{\text{RBF}}(t,t'))$, $f_{\text{localPer}}(t) \sim \mathcal{GP}(0, k_{\text{RBF}}(t,t')k_{\text{periodic}}(t,t'))$ and $f_{\text{latent}}(t) \sim \mathcal{GP}(0, \varepsilon\delta(t,t'))$.

As a first step we aimed to "model-away" the day-of-the-week-effect in order to obtain an estimate of the latent underlying trend $f_{\text{latent}}(t)$ of reported cases. Due to the additive nature of our model we can perform source separation / additive decomposition as described in Section 2.4.6 to compute the individual posterior distributions of the components (see Figure 3.1). Since T is not linear, the following holds:

$$T^{-1} \circ l^{-1}(z(t)) \neq T^{-1} \circ l^{-1}(f_{\text{latent}}(t)) + T^{-1} \circ l^{-1}(f_{\text{locPer}}(t)) + T^{-1} \circ l^{-1}(f_{\text{noise}}(t))$$

Therefore, while terms like $(T^{-1} \circ L^{-1})(f_{\text{latent}}(t))$ are informative (and one can compute their means and covariances via Monte-Carlo), they have to be carefully interpreted. Only in the transformed space do their means and covariances actually add up to the mean and covariance of the entire model $z(t)$, which is shown in Figure 3.7.

3.2.2.1 Model Evaluation and Improvements

Generally speaking, the model seems to fit the data reasonably well, that is the number of daily reported COVID-19 cases in Germany as a whole.

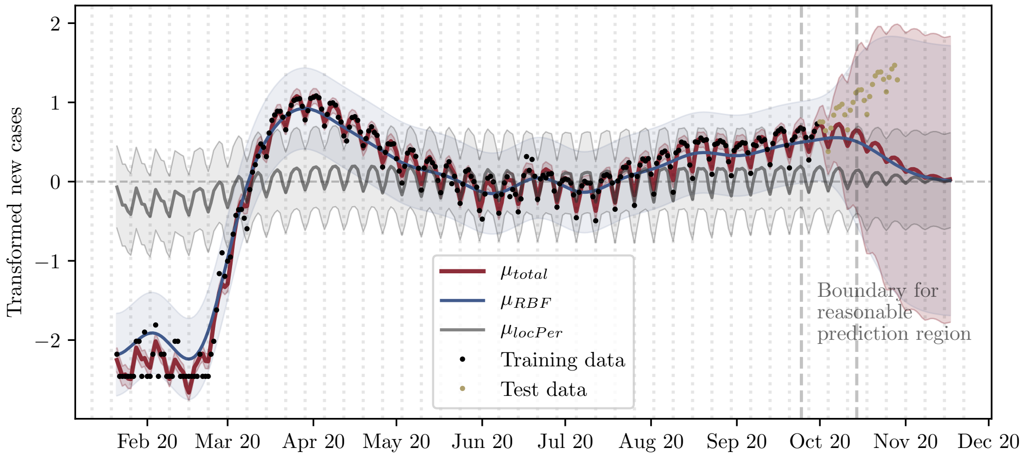


Figure 3.7: *Source separation / additive decomposition.* Latent Space: the reported numbers are the sum of local, periodic and a latent trend.

When it comes to estimating future reported case numbers, it should be noted that a reasonable prediction region only comprises the number of days given by the lengthscale parameter of the RBF-kernel. Considering this region, one can observe two things in Figure 3.8.

First, the periodic reporting behavior is indeed reflected by our model (red curve depicting the posterior predictive μ_{total}). Second, the latent number of infections (blue curve in Figure 3.8), which is supposed to "model away" the day-of-the-week-effect, equally reflects our expectation of reflecting a smooth trend, somewhat averaging the observed data across weekdays.

However, when considering the model's behavior when moving further away from the data that the model was trained on, one can assert two things: First, the "sausage of uncertainty" becomes wider. Second, both posterior predictive means μ_{total} and μ_{RBF} converge to zero in latent space, which is to be expected, given that we initialized our GPs with zero-mean functions (see Figure 3.8). Depending on the phase of the pandemic, this prior mean may or may not be a good reflection of the reality. Alternatively, a compartmental prior mean model, such as *SEIR* (Susceptible, Exposed, Infected, Recovered) might be more apt to reflect the prior underlying dynamics of the disease progression within a society.

Weekly Fluctuations in Reported Numbers and Testing. There are two quantities to be observed, neither one uniformly distributed across weekdays. One being the number of reported cases as depicted in Figure 3.4. The other being the number of daily tests carried out, which not only influences the number of reported cases with some time-lag but also varies over the course of

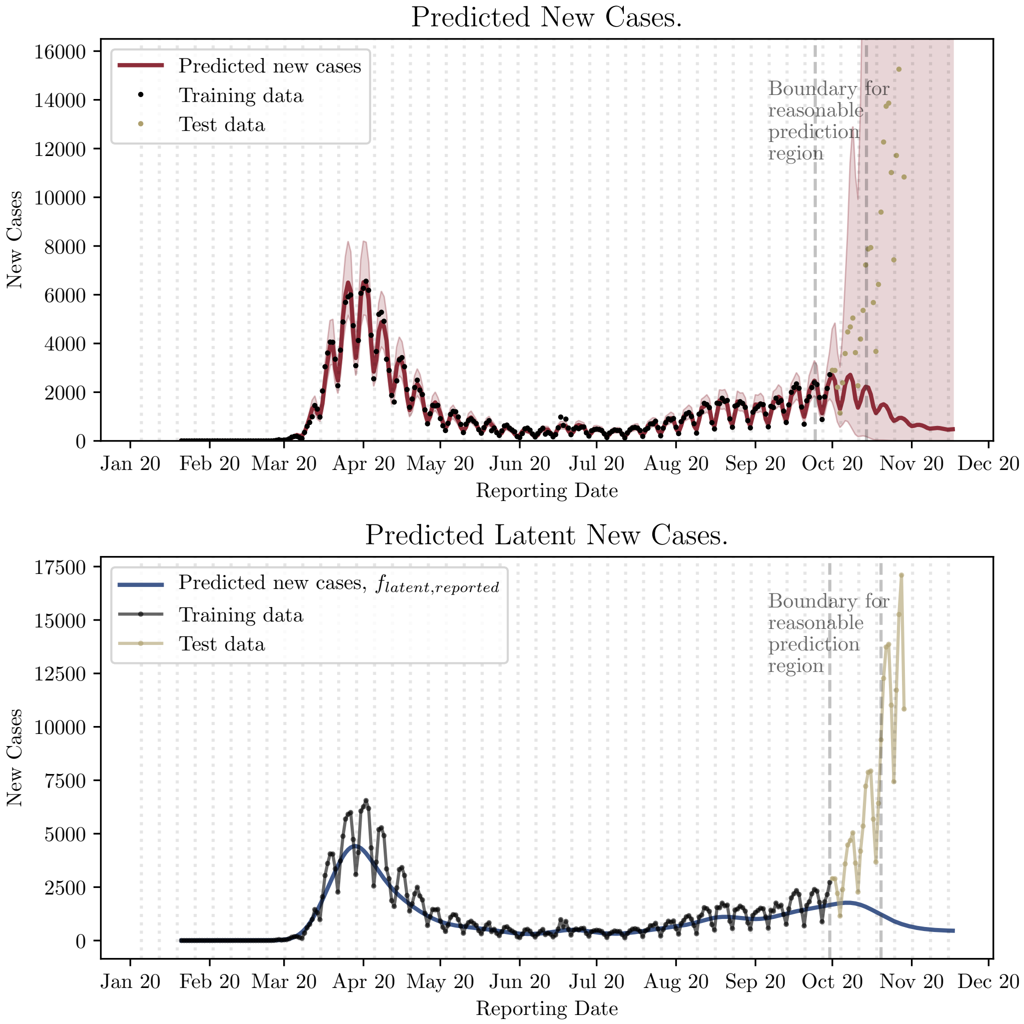


Figure 3.8: *Fit and prediction of reported and latent case numbers.* The upper curve represents fit and prediction to the actually reported data, while $f_{latent,reported}$ (blue curve) in the second graph represents the underlying reported incidence, that is, day-of-the-week effect instigated by the reporting dynamics is ”modelled away”. As we move further away from the data that the model has seen during training, the ”sausage of uncertainty” becomes wider. Additionally, the posterior predictive $f_{reported}$ describing the predicted newly reported case numbers converges to zero reflecting the fact that any GP converges to its prior mean function (which we set to zero) when moving sufficiently far away from the data.

the pandemic. At the beginning of the pandemic, for instance, test facilities were low, while by now test capacities allow for 219092 of daily tests (as of calender week 38 in 2020) [8] The number of tests, however, cannot be equated with the number of people being tested since the data may contain multiple tests per patient. If more tests are carried out, the probability of discovering an infected individual increases. By taking this probability into account, a model may compensate for the weekly fluctuations.

3.2.3 Modelling Confirmed Cases by Date of Disease Onset

In order to obtain a more timely reflection of the pandemic's progression, we now consider the cases by disease onset date. We do this first by shifting the reported cases by the mean number of days that pass from disease onset date to a case being reported (Section 3.2.3.1) and in a second step by applying a coregionalization approach described in Section 3.2.4.

3.2.3.1 Estimating Disease Onset Date via Shift

From the $\sim 60 - 70\%$ of cases for which both, disease onset and reporting date are known, we calculate the mean number of days that pass from a case's disease onset date to its date of report for every day. Next, we calculate the mean over all average shifts per day. This number is then subtracted from the reporting date yielding the an approximation of the disease onset date. The mean number of days between a cases's disease onset date and date of report is ≈ 5.67 days, while for a small number of cases this shift can be as large as 308 days. From Figure 3.9 can be inferred that for most cases three days (mode) pass from disease onset date to reporting date.[9] Running our model on data by disease onset date yields the prediction depicted in Figure 3.10.

Evidently, this is a poor fix for estimating the date of disease onset. In particular, *all* cases are shifted and not only the ones for which the disease onset date is not known. This means that we bias the data to a certain degree in the sense that we are agnostic to the $60 - 70\%$ of cases for which the disease onset date was reported to the RKI. Moreover, this approach does not account for the variation in time lag between a case's disease onset and reporting date as can be seen in Figure 3.11. The time-lag depends, for instance, on factors

[8] `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Sept_2020/2020-09-16-de.pdf?__blob=publicationFile`

[9] Refers to the date a test-positive case is reported to the local health department. Note that still a few days may elapse until a case is notified to the RKI. Nonetheless, the date of report in the RKI database refers to the date a case has been reported to the local health authorities.

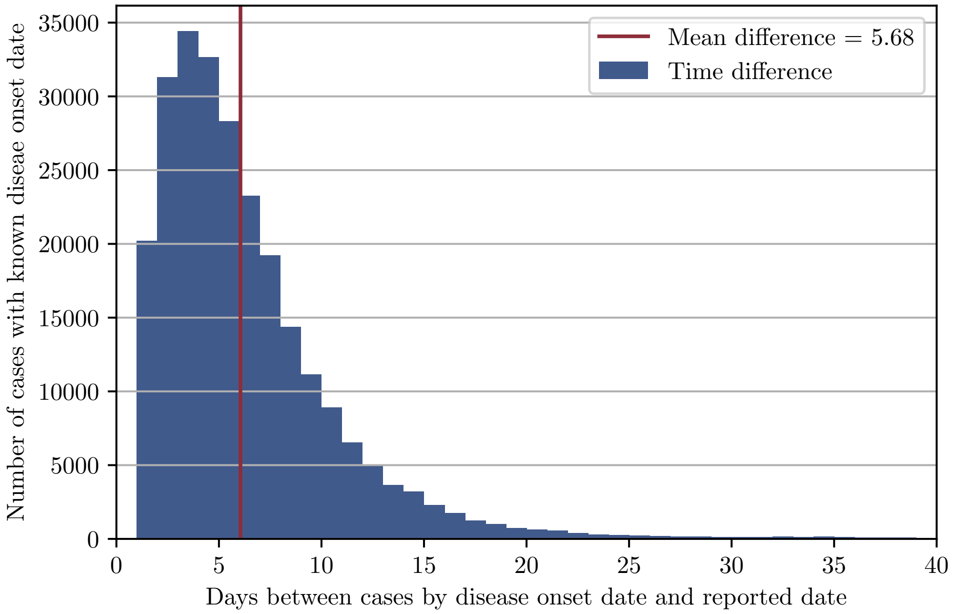


Figure 3.9: *Mean shift between cases by disease onset date and date of report.* For most cases, 3 days pass from disease onset date to being reported ($mode = 3$). On average, 5.67 days elapse with a minimum of 1 and maximum of 308 days and the median being at 4 days. For cases with unknown disease onset date, we estimate this date by shifting their date of report by this number back in time.

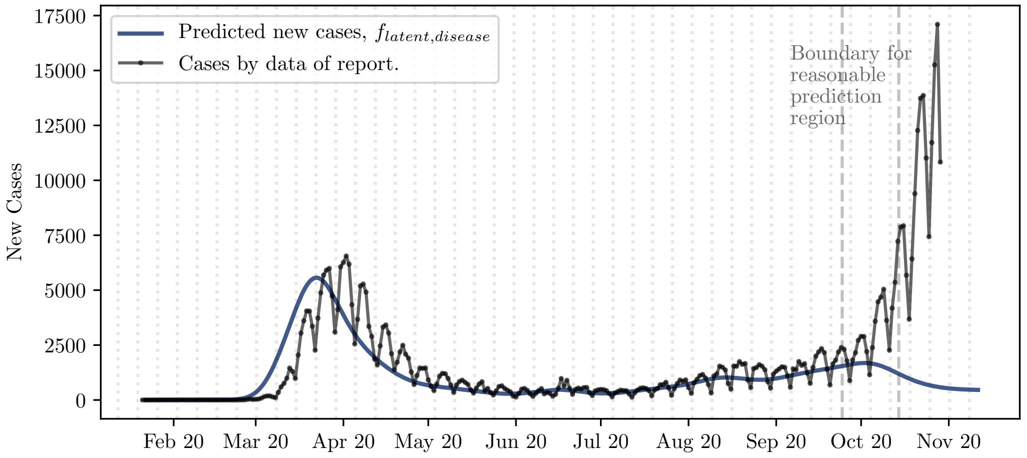


Figure 3.10: *Predicted case numbers by estimated disease date via mean shift.* Our model's prediction for the latent incidence by disease onset date is depicted. This date is estimated by shifting the date of report by 5.67 days back in time. This is the average number of days that has elapsed between a case's date of illness onset until its date of report inferred from cases notified to the RKI for which both dates are known.

such as test-capacities and absolute number of incidence, but also the day-of-the-week effect not only present in the reporting behavior to local health departments but also induced by the point in time people may or may not decide to go to tested. To account for this weekly fluctuation, a more sophisticated modelling approach is necessary. In particular, outliers represented by cases that are reported with a large number of days having elapsed since their date of disease onset date result in the mean to be a rather conservative shift. This way, the disease onset date is estimated to have happened earlier rather than later w.r.t. their date of report. On the positive side, we believe this conservative nature of shifting by the mean to be a rather desirable property resulting in the estimated date of disease onset to be closer to its respective date of infection. To that end, this results in forecasts reflecting the actual transmission dynamics more accurately, which is a crucial factor for mitigation strategies to be decided upon such that they can be effective in a timely manner. Alternatively, one could consider the mode ($\sim$ three days) or median (four days) in place of the mean, which would both be less sensitive to outliers.

3.2.4 Coregionalization

In Section 3.2.1 we learned the latent trend of reported cases, which resulted in a curve with no day-of-the-week effect (i.e. periodic pattern) present in the posterior predictive anymore. Aiming to "model-away" the periodic pattern

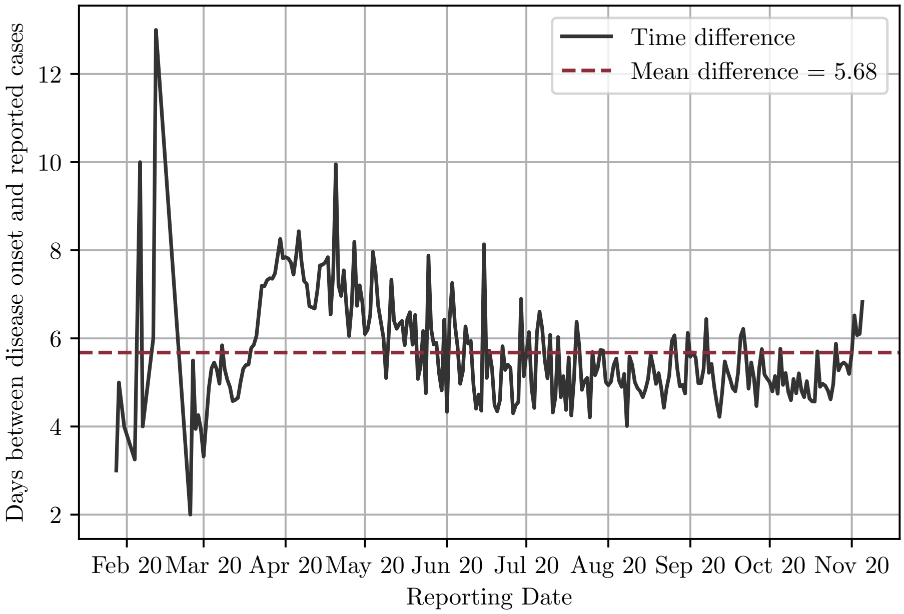


Figure 3.11: *Time-series of the average number of days elapsed from a case's disease onset date until its date of report.* For cases with both, disease onset and reporting date are known, the average number of days that has elapsed from their time of disease onset until they were reported is depicted. For instance, cases that were reported in mid-February have a disease onset date, which precedes their date of report on average by ~ 13 days.

was based on the assumption that the true, latent transmission rate is approximately constant across weekdays. Consequently, the same reasoning applies to cases by disease onset date, which, as previously stated, is only known for about $\sim 60 - 70\%$ of the notified cases.[10] From the latent pattern learned in Section 3.2.1 we now want to derive the cases by disease onset by taking a coregionalization approach. This allows us to use the pattern learned from the reported data to inform the pattern (with missing data) of the date of disease onset. In particular, this reduces the uncertainty in the prediction. To this end, we consider a two-dimensional Gaussian process model

$$\begin{pmatrix} f_{\text{reported}}(t) \\ f_{\text{disease}}(t) \end{pmatrix} \sim \mathcal{GP}(0, k) \tag{3.7}$$

where $f_{\text{reported}}(t)$ denotes the model for the structured noisy process and $f_{\text{disease}}(t)$ the second informative process for the latent function that we would like to retrieve. The covariance function (i.e. kernel k) is given by

$$k(f_i(t), f_j(t')) = k_{\text{reported}}(t, t') \cdot B(i, j)$$

where k_{reported} is an arbitrary kernel describing the way the function varies over time. B describes the degree of similarity between f_{reported} and f_{disease}. B has to be a symmetric positive (semi-)definite matrix in order for k to be a kernel. Thus, we parametrize B as

$$B = WW^\top + \text{diag}(\kappa)$$

where $W \in \mathbb{R}^{2\times p}$ and $\kappa \in \mathbb{R}^2$. While the off-diagonal entries of $WW^\top$ describe how the two processes vary together, the diagonal entries and κ describe how strongly the processes vary. This is analogous to the scaling parameter of an RBF kernel, for instance. This approach is also known as the *intrinsic model of coregionalization* [12] and in particular does *not* require the same data points for f_{reported} and f_{disease}.

Figure 3.3 not only illustrates that for a lot of cases the disease onset date is missing but also that the cases by disease onset roughly follow the trend of the reported cases. Yet, depending on the point in time considered in the pandemic, varying sizes in shift between disease onset and reporting date can be observed (see also Figure 3.11). In order to uncover the unobserved latent trend in cases by disease onset date, we design our prediction model as a combination of the coregionalization approach and an additive Gaussian

[10] Note that the RKI's daily situation report of 05/11/2020, page 4 states that the disease onset date is only known for 295,205 cases (49%) for the period 01/03/2020-05/22/2020. We however, consider the entire time range starting from when the first cases was reported on 28/01/2020. See `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Nov_2020/2020-11-05-en.pdf?__blob=publicationFile`

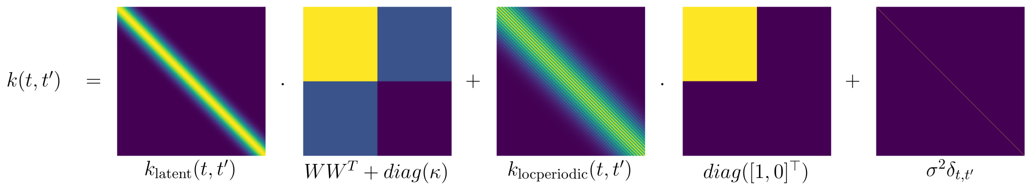


Figure 3.12: *Coregionalization-kernel.* In eq. (3.8), the sum kernel is rewritten as products of a coregionalization kernel and the corresponding kernel modelling either the latent trend or the periodic component

process model. The coregionalization part of the new model then accounts for the similarity in shapes of the curves depicting cases by disease onset date and reported date and the additive Gaussian process model allows to distill the latent trend in reported cases without the weekly pattern.

The sum kernel in eq. (3.8) can be rewritten as products of a coregionalization kernel and the corresponding kernel modelling either the latent trend or the periodic component:

$$k(t,t') = k_{\text{latent}}(t,t') \begin{bmatrix} B_{11} & B_{21} \\ B_{21} & B_{22} \end{bmatrix} + \begin{bmatrix} k_{\text{locPeriodic}}(t,t') & 0 \\ 0 & 0 \end{bmatrix} + \sigma^2\delta_{t,t'} \tag{3.8}$$

$$= k_{\text{latent}}(t,t')(WW^\top + \text{diag}(\kappa)) + k_{\text{locPeriodic}}(t,t')(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \end{bmatrix} + \text{diag}(\begin{bmatrix} 1 \\ 0 \end{bmatrix})) + \sigma^2\delta_{t,t'} \tag{3.9}$$

where $W \in \mathbb{R}^{2\times p}$ and $\kappa \in \mathbb{R}^2$. Graphically this is illustrated in Figure 3.12.

3.2.4.1 Prediction of Missing Disease Onset Dates

Since our model is the sum of individual *GP*s, we can again apply additive decomposition in the transformed space. This yields the individual posterior predictive distributions depicted in Figure 3.13.

Overall, the upper graphs in Figure 3.13 illustrate that our model fits well to the data for both dates, that is f_{reported} and f_{disease}. Further, it can be seen that the uncertainty is generally low except in regions where there is no data. This concerns the first few weeks in January for f_{reported} and the prediction region where uncertainty becomes higher the further we move away from data. The lower part of Figure 3.13 depicts the posterior predictive distributions of the periodic component and $f_{\text{latent, reported}}$ resulting from additive decomposition.

The latter one is then used to inform f_{disease} in the coregionalization approach. Several aspects can be inferred from the shape of the periodic component. First, it can be noted, that overall, the posterior predictive distribution including its uncertainty oscillates approximately constantly around zero,

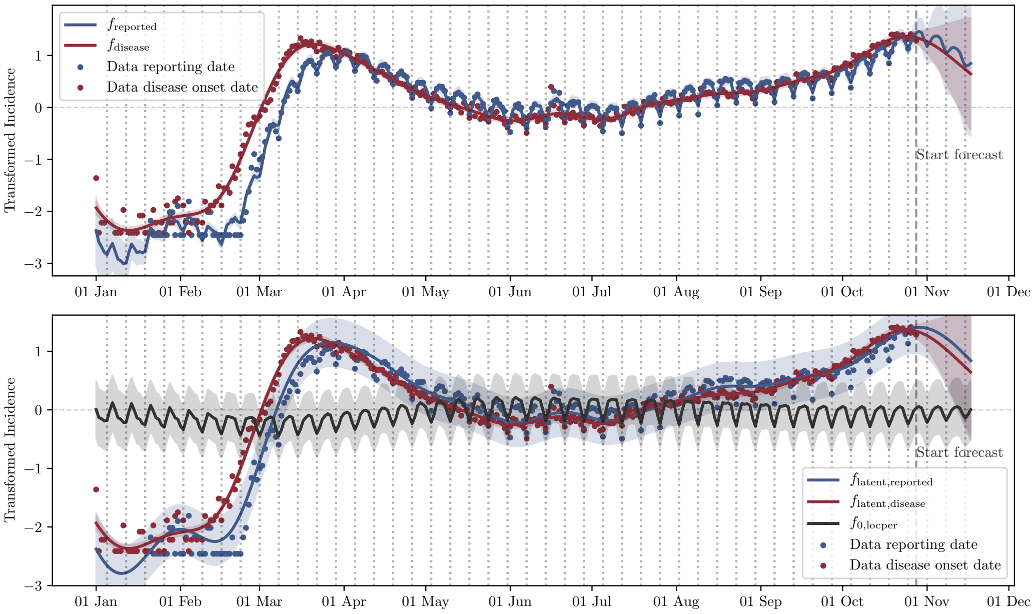


Figure 3.13: *Additive decomposition and coregionalization.* The upper graph shows f_{disease} and f_{reported}. The uncertainty is generally low except for in regions with no data. This concerns the prediction region where uncertainty increases as we move away from the data and the first weeks of January for f_{reported}. The lower graph depicts the additive decomposition of f_{reported} resulting in the latent trend $f_{\text{latent,reported}}$ and its periodic component f_{locper}. $f_{\text{latent,disease}}$ is what we are interested in and results from coregionalizing with $f_{\text{latent,reported}}$.

which suggests that it captures all structure hidden in f_{reported}. Conversely, if there were for instance a linear trend within the periodic posterior predictive apparent, this would suggest that f_{reported} is additionally made-up of a linear trend not yet accounted for in the model. Since this is not the case, however, we have reason to believe that our model constituted by a sum of GPs and the kernels chosen have been reasonable. Moreover, we can see that the periodic pattern has a slightly different shape at the beginning of the pandemic. This is because, back then, only few data was available, which in particular yielded no obvious day-of-the-week effect. However, we encoded this periodic behavior by fixing the period hyperparameter to 7 in the locally periodic kernel.

Transformed back to the incidence space, case numbers now become interpretable again and can be seen in Figure 3.14. The upper plot in Figure 3.14 indicates that the model is fitted well to the raw data for both dates while the lower plot shows particularly high uncertainty for the first and second waves for $f_{\text{latent,reported}}$. Moreover, the model's forecast for f_{reported} reflects the periodic pattern well. Given what we provided the model with, a rise in reported cases followed by declining incidence seems plausible.

Concerning the first wave, the disease onset peak exceeds the reported peak by ~ 2000 cases (when comparing the two latent trends in the lower plot) and a shift of approximately one week. Interestingly, the predicted cases by disease onset date fitted to either the known data (upper plot) or estimated by following the trend provided by $f_{\text{latent,reported}}$ do not differ significantly. When taking a closer look at the prediction region in Figure 3.15, $f_{\text{latent, disease}}$ declines immediately as we move away from the date while uncertainty increases. Some samples, however, illustrate a more probable scenario in which cases by disease onset further rise. Given that people experiencing symptoms today will only be notified as test-positive in a couple of days from today, this would be the most likely trajectory and consequently a desired property of the model. However, the declining posterior predictive of $f_{\text{latent, disease}}$ suggests that our model is strongly influenced by the declining case numbers by date of disease onset present in the most recent days of our data set which is due to the missing cases yet to be reported who have been infected within this period. This brings us to the limitations of the coregionalization approach, namely, that f_{reported} and f_{disease} inform each other *mutually.* In our case, this is an undesired property since we know that f_{disease} is to follow a rising trajectory at least for the period of time delay imposed by the reporting process. The bidirectional flow of information consequently also results in f_{disease} informing f_{reported}. Since we know that for the most recent days of the data set f_{disease} constitutes a lower bound to f_{reported}, this bidirectional flow hinders the prediction for f_{reported} to be as accurately as possible. To mitigate this effect a little (that is, information flowing from to f_{disease} to f_{reported}) we excluded the most recent days of the datasets such that declining case numbers for most recent days especially present for f_{disease} (since yet to be reported) do not affect prediction of f_{reported} too strongly. In particular, we excluded six days from the $\text{data}_{\text{reported}}$ and $12 \approx 6 + mean(\text{disease onset date} - \text{report date})$.

To conclude, this approach allows for a more accurate estimate of disease onset dates as opposed to simply shifting by the mean average of days between the two dates as in Section 3.2.3.1. This is because the coregionalization approach takes into account the varying numbers of days present between reported and disease onset date of cases over the course of the pandemic.

3.2.5 Rating the Spread of the Pandemic – Metrics, which Inform Mitigation Strategies

In order to assess how fast the pandemic evolves, the reproduction number R in conjunction with the number of daily newly infected persons (incidence) constitutes an important epidemiological metric.

The time-dependent effective reproduction ratio R_t describes the expected number of disease transmissions caused by a single infectious individual. Thus,

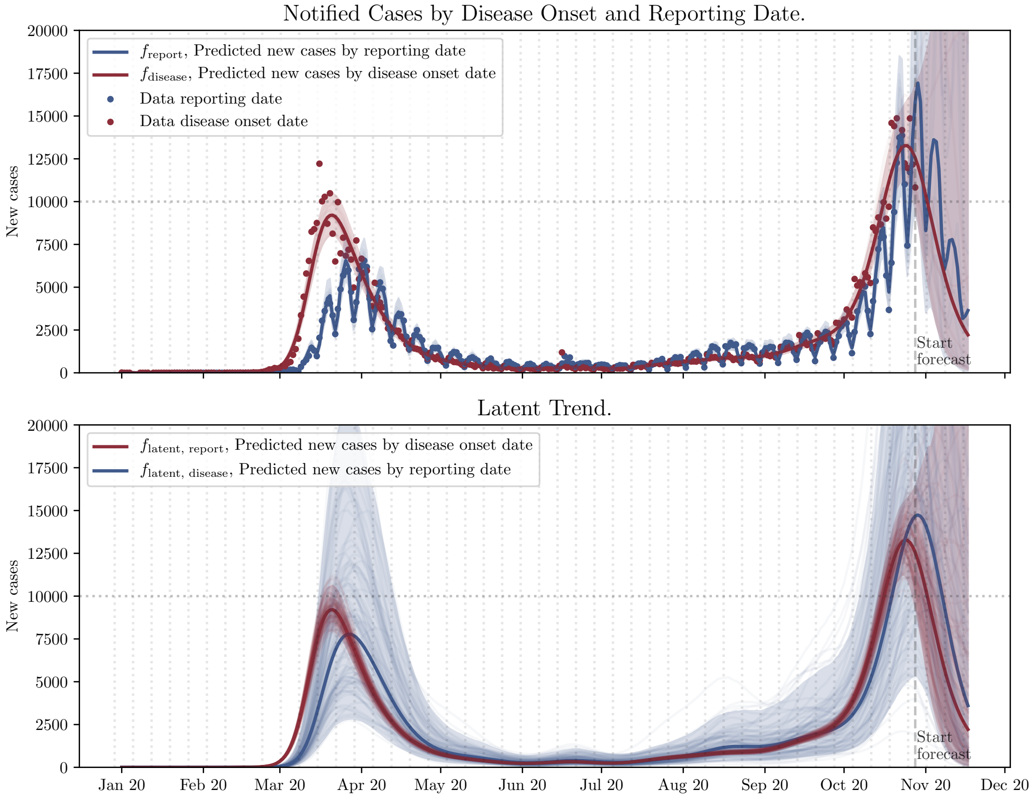


Figure 3.14: *Disease onset date estimated by coregionalizing approach.* The upper graph shows f_{disease} and f_{reported}. The uncertainty is generally low except for in regions with no data. This concerns the prediction region where uncertainty increases as we move away from the data and the first weeks of January for f_{reported}. The lower graph depicts the additive decomposition of f_{reported} resulting in the latent trend $f_{\text{latent,reported}}$. $f_{\text{latent,disease}}$ is what we are interested in and results from coregionalizing with $f_{\text{latent,reported}}$.

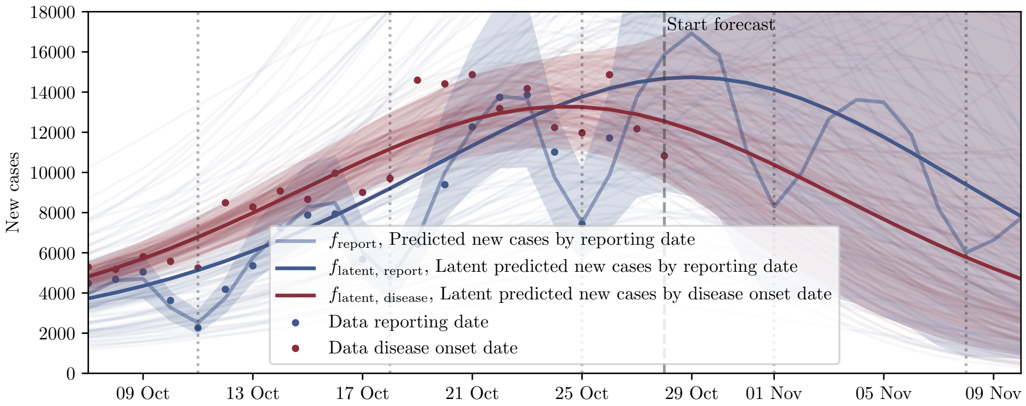


Figure 3.15: *Forecast with coregionalization approach.* Excerpts of the last five weeks of Figure 3.14 depicted. The forecast region is marked by the end of data points on October 28, 2020.

R_t essentially is a multiplier, which describes how effectively a virus is spreading and is a commonly used guidance instrument for policy decisions.

There are three cases to be distinguished:

- $R_t > 1$: number of new infections increases
- $R_t = 1$: number of new infections stagnates
- $R_t < 1$: number of new infections decreases

The RKI provides two estimates for the R_t values, which will be described below. Subsequently, we propose two alternative metrics, $R_{4,GP}$ and the D-value, which may equally inform mitigation strategies.

3.2.5.1 R-values published by the RKI

The RKI defines the R-value as the mean number of people being infected by one infected person[11] and critically notes that its value cannot be derived from the notification system but that it can only be estimated which an der Heiden and Hamouda [11] do via a statistical approach which they name Nowcasting. The RKI further notes that fluctuations regarding R-value are especially high when overall numbers are low. In order to account for these fluctuations, they apply a 4- or 7-day moving average, resulting in a 4− or 7−day-R-value which we will briefly elaborate on below. Both R-values are based on Nowcasting which predicts the number of cases with illness onset up to the date of four

[11] https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/2020-05-21-en.pdf?__blob=publicationFile

days ago, since no reliable prediction can be made about the number of new cases in the last three days. The 4− and 7−day-R-values as calculated by the RKI are depicted in Figure 3.16.

4-day R-value. The 4-day R-value is estimated by using a moving 4-day average of the number of cases with illness onset up to the date of four days ago. The 4-day mean of incident cases on one day is compared with the respective mean four days before. This takes into account that infections occur four to six days before symptom onset and reflects the course of infection from approximately one to two weeks ago. However, according to the RKI, the 4-day-R-value is still very sensitive to fluctuations in case numbers which is why they additionally provide a 7-day R-value.

7-day R-value. Aiming to account for the day-of-the-week effect to be observed in reported case numbers, the RKI provides a 7-day R-value based on data from a longer time period which is less subject to short-term fluctuations. It is calculated the same way as the 4-day R-value except that a moving 7-day average from the Nowcasting [11] curve is used. Hence, this reflects trends more reliably but is based on infections that occured on average earlier than those on which the more sensitive 4-day-R-value is based on. The 7-day R-value consequently represents a slightly later course of infection of about one to a little over two weeks ago.

3.2.5.2 Comparison to R-value Based on Coregionalization Model

We compute the 4-day incidence rate based on the case numbers by disease onset date described by $f_{\text{latent, disease}}$, which resulted from the coregionalization model in Section 3.2.4 the same way as the RKI calculates its R_4 depicted in Figure 3.16 which we denote as $R_{4,GP}$. According to the RKI, R_4 is more sensitive to fluctuating case numbers resulting from the day-of-the-week effect. Thus they also provide the R_7-value, which is supposed to mitigate said effect and is therefore more apt to serve as a policy guiding tool.

Supposedly, while R_7 is smoother than R_4, Figure 3.16 shows that R_7 still reflects the periodic pattern, which is also not covered by the "sausage of uncertainty". This implies that R_7 is quite certain about infections to follow a periodic pattern.

Conversely, this effect has already been "modeled-away" in $f_{\text{latent, disease}}$. Computing the more sensitive R_4-value based on this data ($R_{4,GP}$) depicted in Figure 3.16, results in no periodic pattern present anymore.

When comparing $R_{4,GP}$ to R_7, one can generally observe that $R_{4,GP}$ is much smoother than R_7 but mostly follows the same trajectory, approximately averaging the periodic pattern present in R_7. In particular, uncertainty for

$R_{4,GP}$ is very low in these regions. Combined with the observation that $R_{4,GP}$ has a similar shape to $R_{4,7}$, this adds to the plausibility of $R_{4,GP}$.

Furthermore, uncertainty for $R_{4,GP}$ increases, in regions where there is only little data available. This concerns the beginning of the pandemic (when $R_{4,7}$ following the Nowcasting approach were not yet computed) and the forecast. This is a desired property since the level of uncertainty suggests a degree of trust adequate to put into the model's estimate.

Moreover, there are a few points in time, where $R_{4,GP}$ and R_7 diverge. We observe this behavior first during the initial wave in March where $R_{4,GP}$ not only estimates a slightly higher R-value but also suggests a wider range in which the majority of infections were occuring, which, overall, are additionally predicted to have happened approximately a week earlier than suggested by R_7. Further, the peak in June is again predicted to have happened a few days earlier by $R_{4,GP}$ and with an estimated value of $R_{4,GP} \approx 1.25$ significantly lower than $R_4 \approx 2.1$. This shift in peaks may be explained by the moving average, which the R-values are based on. Since $R_{4,GP}$ is computed already on the latent trend, sudden fluctuations and delays are already smoothed and further washed out by the moving average. Lastly, $R_{4,GP}$ decreases and hence seems to underestimate the R-value starting ca. on October 18, which corresponds to ten days before the coregionalization model was not provided with data anymore. This decreasing trend of $R_{4,GP}$ is to be explained by the coregionalization model's reversion to the prior mean. In particular, we already pointed out that its ability to predict new cases by disease onset is limited due to the time delay with which they become notified to the RKI and the bidirectional flow of information present in the model which compromises its forecasting ability such that a reversion to the zero-mean function is eventually to be observed.

All things considered, the approach yielding $R_{4,GP}$ seems to be a worthwhile alternative to the Nowcasting approach by [11], which R_4 and R_7 are based on.

3.2.5.3 D-value

In the following, we propose an alternative metric to R named D-value, that estimates the spread of the disease via the normalized change in case numbers.

D is calculated by taking the derivative of the underlying trend $f_{\text{latent, disease}}$ (see eq. (3.10)), which resulted from the coregionalization model in Section 3.2.4 (see Figure 3.14). This is possible since $f_{\text{latent, disease}}$ is given by a GP itself, differentiation is a linear operation, and, applied to a GP, therefore results in a GP again. The result is normalized by dividing by the posterior mean of $f_{\text{latent, disease}}$. Uncertainty is represented as a 95%-CI, which is calculated based on samples drawn from the GP that were transformed back to the

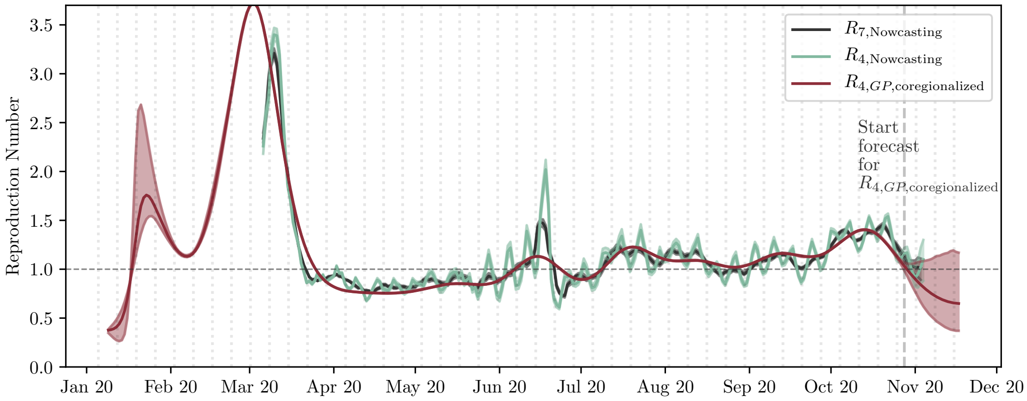


Figure 3.16: *R-values.* The 7-day R-value compares a 7-day moving average of new cases on one day with the corresponding 7-day average of new cases on the day a week before, while the more sensitive R_4-value is estimated by using a moving 4-day average of the number of cases with illness onset up to the date of four days ago. $R_{4,GP}$ is calculated the same way as R_4, on the case numbers described by $f_{\text{latent, disease}}$. While both, R_4 and R_7, still reflect some periodic pattern arising from the day-of-the-week effect, our estimate $R_{4,GP}$ does not show this undesirable periodicity. For each R-value, the upper and lower bound of the 95% prediction interval is depicted. Data for $R_{4,7}$ comes from a .csv-file provided by the RKI found under `https://tinyurl.com/y5e93kmf`. The Nowcasting data in this plot is based on Nov. 7, 2020.

incidence space.

$$D(t) = \frac{\frac{d}{dt} f_{\text{latent, disease}}(t)}{\mu_{\text{latent, disease}}(t)} \tag{3.10}$$

In principle, D can be computed for any GP-based model of similar form to the one we propose here. Given our model, estimating the change of the incidence by calculating the derivative is the intuitive metric to consider if one is interested in the dynamics of the disease. We therefore believe this to be a worthwhile approach, not least because it additionally provides uncertainty and also allows for improved predictions by incorporating prior knowledge.

D on any given day can now be interpreted as follows and is depicted in Figure 3.17:

- $D < 0$: case numbers are decreasing
- $D \approx 0$: case numbers stagnate
- $D > 0$: case numbers are increasing

Figure 3.17 shows that D generally follows a similar trend to $R_{4,GP}$ and $R_{4,7}$. In contrast to $R_{4,7}$ but similar to $R_{4,GP}$, no day-of-the-week effects are present in D. During much of the so called "steady-state"-phase, when case numbers were comparatively low (from May to August), D stays approximately constant. Analogous to $R < 1$, D is smaller than zero suggesting declining case numbers for that period. One may interpret, that during the steady-state phase, both, D and the R-values predicted a slightly decreasing trend in case numbers. Uncertainty in D being higher than in the R-values suggests that D deems a (slight) surge in case numbers more probable than $R_{4,7}$ and $R_{4,GP}$ do. Further, it can be observed that in October, D reflects the start of the second wave with high probability. Following this rise in D a downturn along with an increase in uncertainty can be observed. This effect can be explained by fewer cases by disease onset date being available in our dataset and by the the model's reversion to the prior mean as it moves away from the data.

Additionally, while both R_4 and R_7 come with uncertainty, no significant benefit is added since the periodicity imposed by the day-of-the-week effect is not covered by the "uncertainty sausage ". In contrast, uncertainty is generally wider in D compared to the R-values. In fact, $R_{4,GP}$ is particularly certain in the regions with a lot of data available, a result, which is to be questioned. We believe this to be caused by a "double-smoothing"-effect underlying the calculations yielding $R_{4,GP}$. Recall that we calculate the four-day incidence based on the coregionalized latent disease onset date in which the day-of-the-week effect has been "modeled-away" yielding a smooth curve. Further smoothing is induced by the 4-day moving average yielding an $R_{4,GP}$ with uncertainty that arguably does not faithfully represent uncertainty in the estimate.

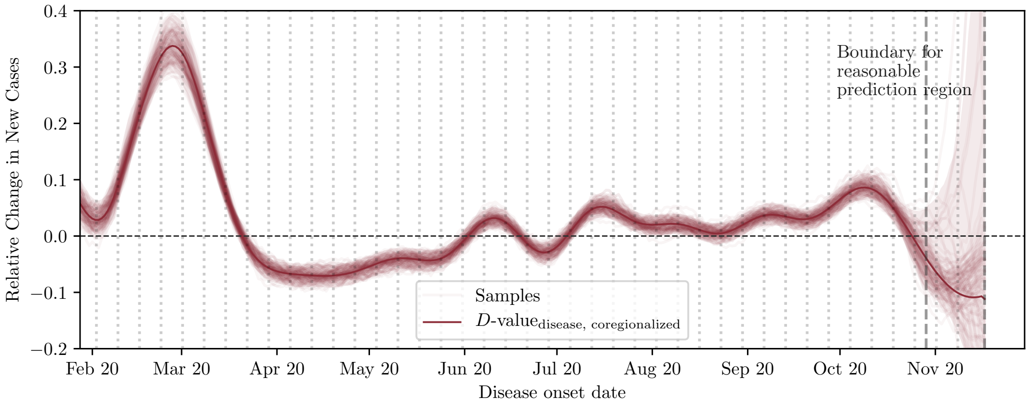


Figure 3.17: *D-value.* D is calculated via the normalized change in case numbers (see eq. (3.10)) based on the coregionalized disease onset date. D assesses the rate at which the virus is transmitted and therefore provides an alternative metric to R. The upper and lower bound of the 95% prediction interval is depicted.

With that in mind, it is important to consider the limitations of such metrics. Altogether, we have seen that even R_7, which intends to account for the day-of-the-week effect, still reflects the weekly periodicity. Conversely, both D and $R_{4,GP}$ based on the coregionalized disease onset dates do not reflect such periodicity. Both the D- and $R_{4,GP}$-value seem to provide a reasonable alternative to R_4 and R_7 as a metric to evaluate the rate the disease spreads and even seem to offer some advantages over R_4 and R_7 regarding fluctuations in reporting-behavior and in terms of possibly providing a more accurate estimate when it comes to the latent trend. All things considered, we believe that even though $R_{4,GP}$ improves upon $R_{4,7}$, D is a more informative metric given this model with the additional benefit of faithful uncertainty.

The previously described GP models were implemented in the probabilistic programming language PyMC3 [13], a Python framework, which includes predefined probability distributions and covariance functions serving as building blocks for our models.

3.3 Related Work

Since the ongoing pandemic affects countries across the globe, the scientific community has made an immense effort to understand the dynamics of the pandemic, the virus and treatments for the disease. Meanwhile, this research has manifested itself in a plethora of studies. Here, we only refer to a handful of studies, which have influenced this work.

Do we need to act? In order to decide whether to impose new lockdown policies, it is crucial to have faithful estimates reflecting the *current* stage of the pandemic. A principal epidemiological parameter is the reproduction number R, which quantifies the average number of people that one infectious individual will pass on a virus to [14].

R is derived from the daily number of reported cases and provides a metric for the relative growth or decline of the virus. The RKI therefore computes its R-values not directly based on reported case numbers but data that has been nowcasted to correct for the delay between disease onset and day of report. Nowcasting is a statistical correction by an der Heiden and Hamouda [11], which accounts for the diagnosis, reporting and notification delay in the reporting process of test-positives. They simulate the date of illness onset by computing an empirical distribution for the number of days elapsed from a case's disease onset to its reported date. Based on this, the date of infection can be estimated by assuming a serial interval of four days.

An alternative approach to estimating R is proposed by Systrom et al. [15] who provide a real-time view of R_t. In essence, amongst the infinite number of curves potentially describing the true trajectory of R_t, their model[12] searches for the one with the highest probability of explaining the observed data. Their model integrates data on test-adjusted positives, which is a relative measure of how many true positives there are, which they admit to not be an ideal measure. However, more informative ones, such as the hospitalization and death rate are drastically time-shifted from the infection date.

How to contain the spread of the virus? Having reliable estimates on the current stage of the pandemic w.r.t. incidence is key to decide on mitigation strategies such as (lockdown) policies in order to curb case numbers. Counterfactual analyses answering "What if ..."-questions regarding hypothetical scenarios and illustrating different intervention strategies can be an informative tool for policy makers.

Qian et al. [16] provide a machine-learning based decision-making tool named "Policy Impact Predictor (PIP) for COVID-19"[13] aimed at guiding governments in their decisions on measures that prevent the virus from spreading. PIP uses a two-layer Gaussian process model where the lower layer has country and policy specific parameters as a prior mean described by a variant of the *SEIR*-model, which, in their case, comprises six compartments (Susceptible, Exposed, Infected, Recovered, Critically-ill, Recovered, Dead) – *SEICRD*. This layer captures fatality curves under counterfactual policies within each country whereas the upper layer is shared across all countries and learns lower-layer *SEICRD*-parameters as a function of a country's features and policy parame-

[12] R_t.live is a Bayesian model equally implemented in PyMC3 as our model is.

[13] `https://www.vanderschaar-lab.com/policy-impact-predictor-for-covid-19/`

ters.

Evaluating the effects of interventions in a timely manner is possible with the Bayesian framework for the spread of COVID-19 by Dehning et al. [17]. Their model infers principal epidemiological parameters and the timing and magnitude of policy effects, which also allows for short-term forecasts of future interventions' effects (or effects from lifting restrictions).

Chapter 4

Conclusion and Future Work

The pandemic known as the coronavirus (SARS-CoV-2) has significantly disrupted public life world-wide and was directly responsible for approximately 1.3 million excess fatalities as of November 2020 [2] . In order to contain the disease's spread, mitigation strategies need to be decided upon in a timely manner. These should be guided in part by metrics reflecting the current stage of the pandemic. Such important epidemiological parameters are the incidence in conjunction with the basic reproduction value R.

This thesis concerned itself with modelling the current state of the pandemic in Germany by estimating the daily number of newly infected people. Based on our model, we also provide a metric assessing the trend of the incidence.

modelling the number of COVID-19 cases should ideally be based on the date of infection since this reflects the current stage of the transmission rate as timely as possible. The infection date is not known, however, and can only be inferred by the disease onset date of a case. Unfortunately, even this date is not known for all cases notified to the RKI. Ultimately, only the date of report is provided for all notified cases which is why this date served as a first proxy in our models.

We observed a weekly periodicity in the case numbers by date of report, which we attributed to the time lags innate in the reporting process of a test-positive case to the RKI. As an initial model we assumed a transformed Gaussian process with a sum kernel. Via source separation, we were able to disentangle the periodic effect from the latent trend which we presume to be a more accurate description of the true transmission dynamics.

However, the date of report only reflects a "back-cast" of the status of the pandemic due to its inherent time lag. For this reason, we next modeled the incidence by disease onset date. As of today this date is only known for $\sim 60-70\%$.[1] Therefore, we computed the empirical distribution of the average

[1] This range stems from considering the cases numbers provided by the RKI from the

number of days that elapsed from a case's date of illness onset until its date of report. We observed this time lag between disease onset and reporting date to vary over time. This variation can be attributed to several factors, such as the phase of the pandemic (beginning vs. first or second wave), a surge or decline in case numbers, available test capacities and other resources (health-care staff, test-kits, etc.), etc.

As an initial improvement over the model based on the date of report we imputed the missing disease onset dates with the mean difference between disease onset and reporting date and ran the same model again on this data.

In order to account for the varying differences between disease onset and reporting date, we next applied a coregionalization approach with the intent of informing the curve modelling the disease onset date via the shape of the latent time-series describing cases by date of report. We again accounted for the reporting specific pattern in the data and were left with a reasonable fit with uncertainty to the raw data (both dates). The model's prediction ability, however, was impaired by the fact that the two time-series were able to inform each other mutually, which yielded too optimistic forecasts. The ideal property would have been an unidirectional flow of information, such that the reporting date curve (which is "most complete") only informs the disease onset curve and not vice versa. Nonetheless, this approach improved upon our initial model.

Finally, we proposed a metric that we named D-value, which quantifies the change in daily confirmed cases by calculating the normalized derivative of the incidence on coregionalized disease onset data. Importantly, it provides a confidence estimate for the current dynamic of the pandemic.

4.1 Shortcomings and Improvements

Our models provide reasonable estimates on the latent trend of the incidence by removing reporting-specific patterns in the data, specifically a pronounced periodicity. However, forecasts eventually always revert to zero and come with rapidly increasing uncertainty as we move away from the data. The former effect can be explained by the GP reverting to its prior mean away from the data which we initialized to be zero. Depending on the phase of the pandemic, this may or may not be a reasonable choice. An improvement upon this behavior would be a compartmental prior model, for instance the *SEIR*-model.

The latent trend of the incidence by disease onset date is equally estimated reasonably well but also falls short with predictions where the trajectory of the curve is eventually decreasing as well. Again, this can be partly explained by reversion to the prior mean. Additionally, while the coregionalization ap-

first case up until early November.

proach generally accounts for the varying shift over time, it results in a mutual exchange of information between the two time-series results in predictions of incidence by reported date to be underestimated. Moreover, estimates generally suffer from the curves informing each other mutually.

An improvement upon this would be to e.g. sample shifts from an empirical distribution. This could be even extended to constraining sampling a shift to an empirical distribution estimated from the same time period as the case's date of reporting.

4.1.1 Geographic Knowledge as a Predictor

Oftentimes, surging case numbers are restricted to a local region often resulting from superspreader events (e.g. slaughter houses such as Tönnies, ...). To contain such outbreaks, temporary mitigation policies might need to be imposed locally but not be required on a national scale. Hence, making predictions for different counties would be informative given varying causal factors influencing the progression of the COVID-19 case numbers. Our model in Section 3.2.1.1 does not take geographic information available from the RKI into account. Using the same assumptions for the national model also for an individual county does not result in an ideal prediction as Figure 4.1 shows.

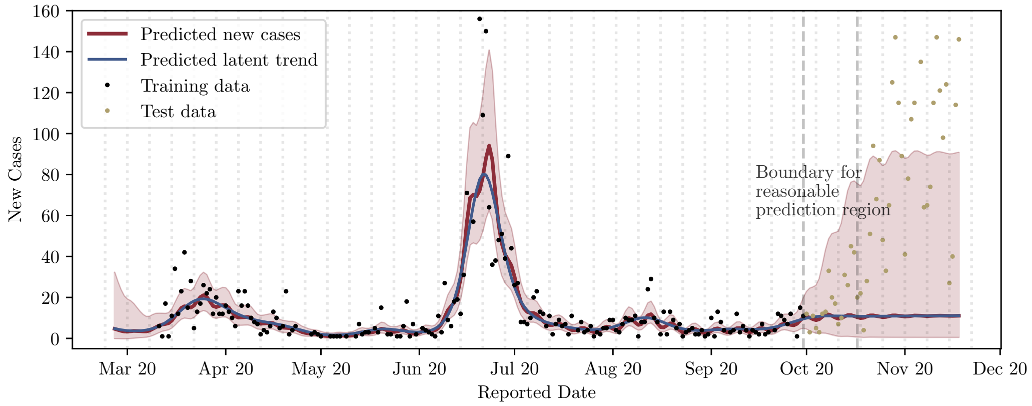


Figure 4.1: *Naive model on the federal county of Gütersloh.* The resulting posterior means show a somewhat poor fit when reporting behavior does not reflect the national trend. Gütersloh is a county in the federal state of North Rhine-Westphalia where. In June 2020, there was a surge in cases due to an outbreak of COVID-19 in a meat processing plant.

A more finely grained approach could be to model each county in Germany separately and taylor the assumptions of the model to the individual geographic regions. In particular this also considers spatial proximity. In our framework a *hierarchical* Gaussian process model would be an appropriate choice, where we partially pool the data per geographic region and thus allow for these levels to mutually inform each other. In a recent episode of the Coronavirus-Update Podcast, Professor Christian Drosten[2] mentioned that it may be of importance to take a closer look at the incidence in individual districts lending credence to this extension of our approach. Figure B.4 illustrates that the national shape of reported COVID-19 cases is mainly influenced by the reporting shape found in Bayern and Baden-Wuerttemberg, while several other states and districts do not reflect the pattern found in the national trend. This further illustrates that a "one-size-fits-all"-approach may not be suitable here and motivates a model of the following form:

$$\begin{aligned} f_{\text{national}}(t) &\sim \mathcal{GP}(0, k) \\ f_{\text{local, i}}(t) &\sim \mathcal{GP}(f_{\text{national}}(t), k_{\text{local, i}}), \end{aligned}$$

where $f_{\text{national}}(t)$ describes the national trend, defined by a GP with zero-mean function and a kernel accounting for the day-of-the-week-effect in the reporting behavior, amongst others. $f_{\text{local, i}}(t)$ defines a respective GP for a local geographic region (state- or county level), defined by a mean-function which now corresponds to some pooled geographic posterior predictive, e.g. $f_{national}$, and a respective kernel. This additional granularity would allow for a more detailed incorporation of structural knowledge (demographic data, spatial proximity of counties, population density, urban - rural, vacation returnees, school openings, ...) to enter the model.

4.2 Outlook

With the recent announcement of the primary efficacy analysis of a vaccine by Biontech / Pfizer, which suggests to be 95% effective against COVID-19,[3] followed by similar results of Moderna's vaccine candidate,[4] an end to the pandemic is a possibility. However, it will take a considerable amount of time until sufficient doses of a vaccine for the entire world population can be produced. Hence, it is important to distribute vaccines in such a way that transmission of the virus is curbed and a level of herd immunity is achieved to ensure the resumption of public life and economic activity. At the same time, however, we would like to keep the hospitalization rate and lethality as low as possible. These factors necessitate an informed vaccine distribution strategy. One may intuitively think that vaccines should be brought first to the most vulnerable groups in the population, namely elderly and those suffering from

[2] Das Coronavirus-Update von NDR Info Folge 56

[3] `https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine`, last accessed on November 19, 2020

[4] `https://www.modernatx.com/cove-study`, last accessed on November 19, 2020

pre-existing medical conditions. However, it is important to consider, that the vaccine could cause side-effects that these groups may be suffering from more severely than healthy people. Other factors to consider are differences in mobility between population groups, occupational differences in number of contacts, probability of a surge in cases in certain areas and the severity of mitigation policies (e.g. lockdowns). Devising a vaccine distribution strategy therefore has to make use of the current state of knowledge about the pandemic and its effects, derived from models and data similar to what we consider in this work. The more that is known about the spread of the virus, the faster immunity across the population can be achieved.

Appendix A

Further Tables and Figures

Parameter	Value at MAP
l_{RBF}	15.352809
θ_{RBF}	0.850290
l_{pRBF}	23.434164
θ_{pRBF}	0.547721
θ_{per}	0.547721
σ_{noise}	0.104464

Table A.1: *Hyperparameter values after fitting naive model to incidence by reported date.*

Parameter	Value at MAP
l_{RBF}	18.378534
W	$\begin{bmatrix} 0.812192 - 2.208457 \\ 0.617260 - 1.820872 \end{bmatrix}$
κ	$\begin{bmatrix} 0.812192 \\ 0.617260 \end{bmatrix}$
θ_{RBF}	0.418021
σ	0.110604
σ_{locper}	0.302919
l_{locper}	38.787971

Table A.2: *Hyperparameter values after fitting corgionalization model.*

Appendix B

Figures

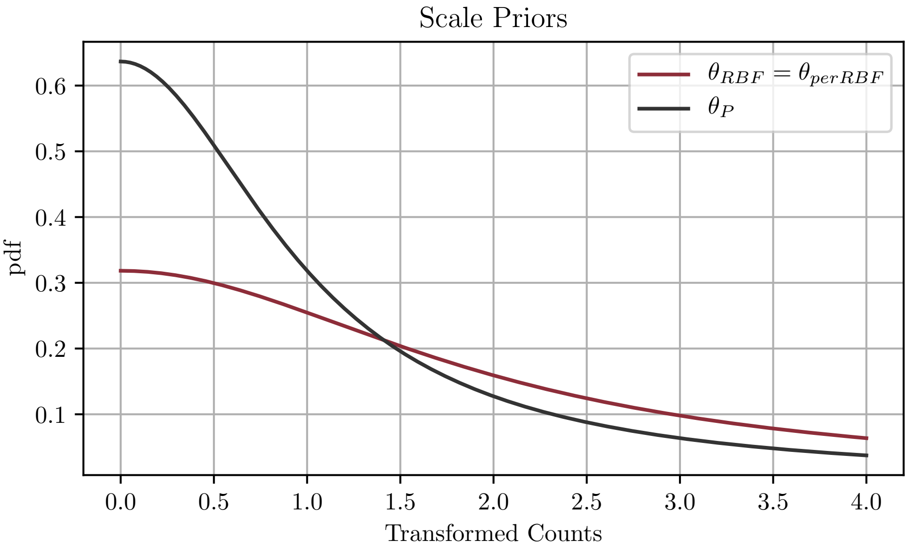


Figure B.1: PDFs of scaling hyperparameters for naive model on reported case numbers in Germany.

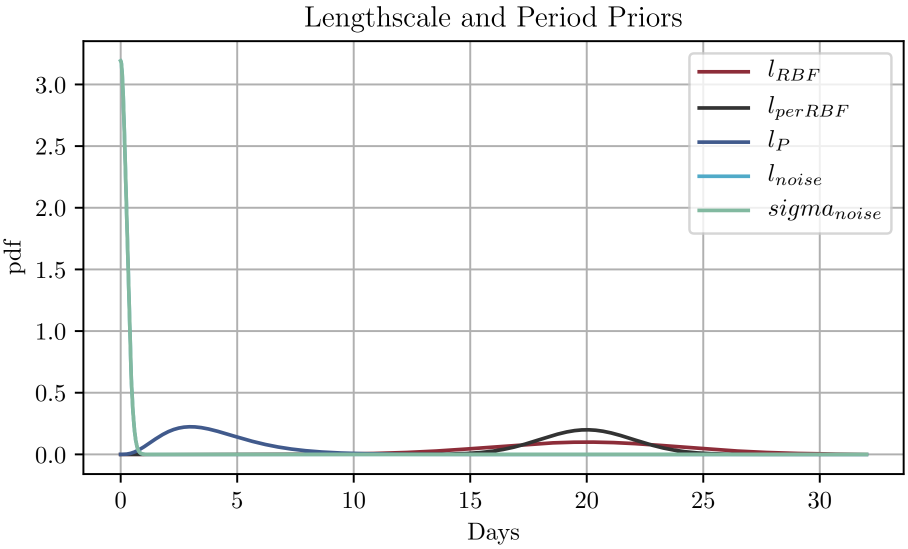


Figure B.2: PDFs of lengthscale hyperparameters for naive model on reported case numbers in Germany.

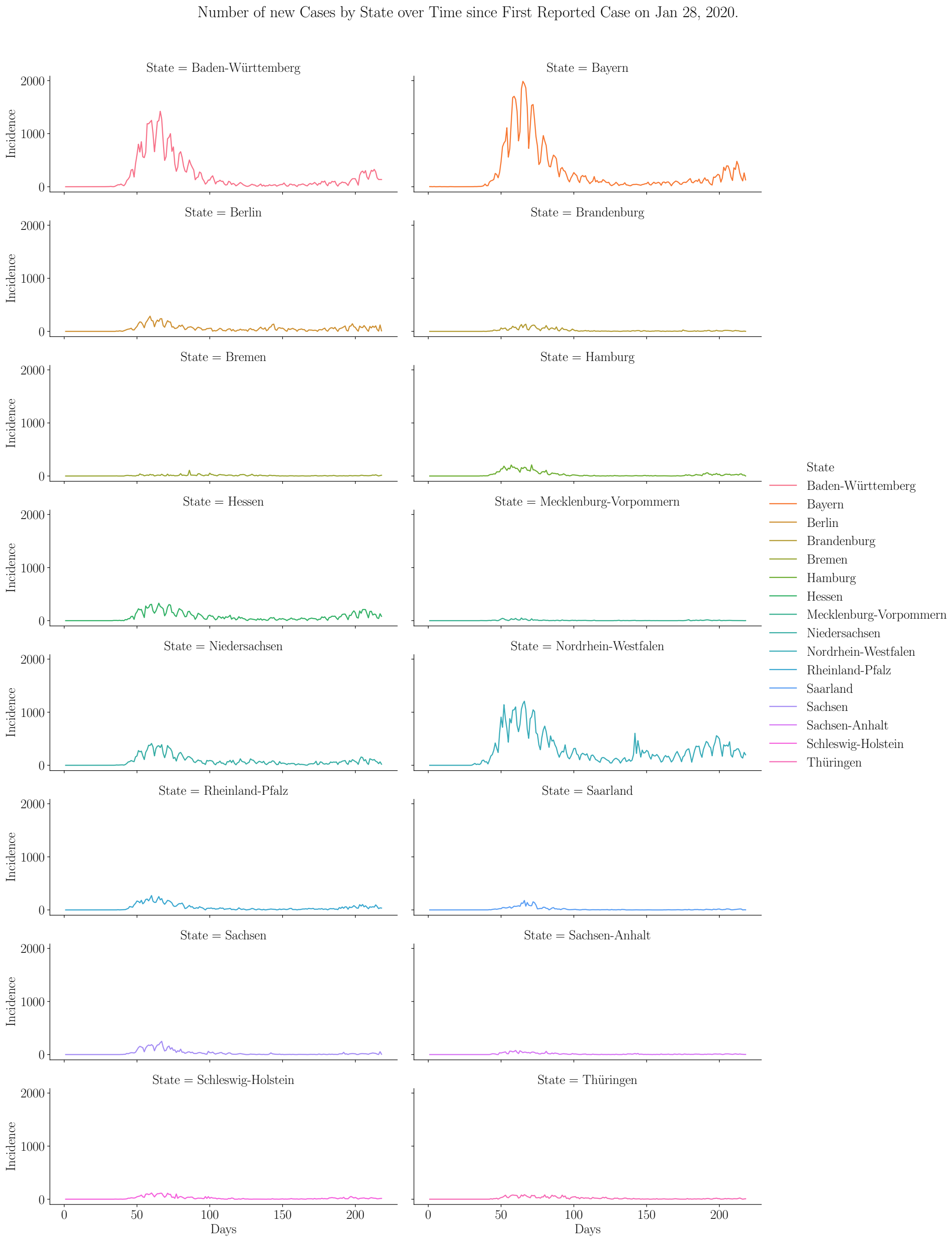


Figure B.3: Number of new cases by reporting date per federal state in Germany.

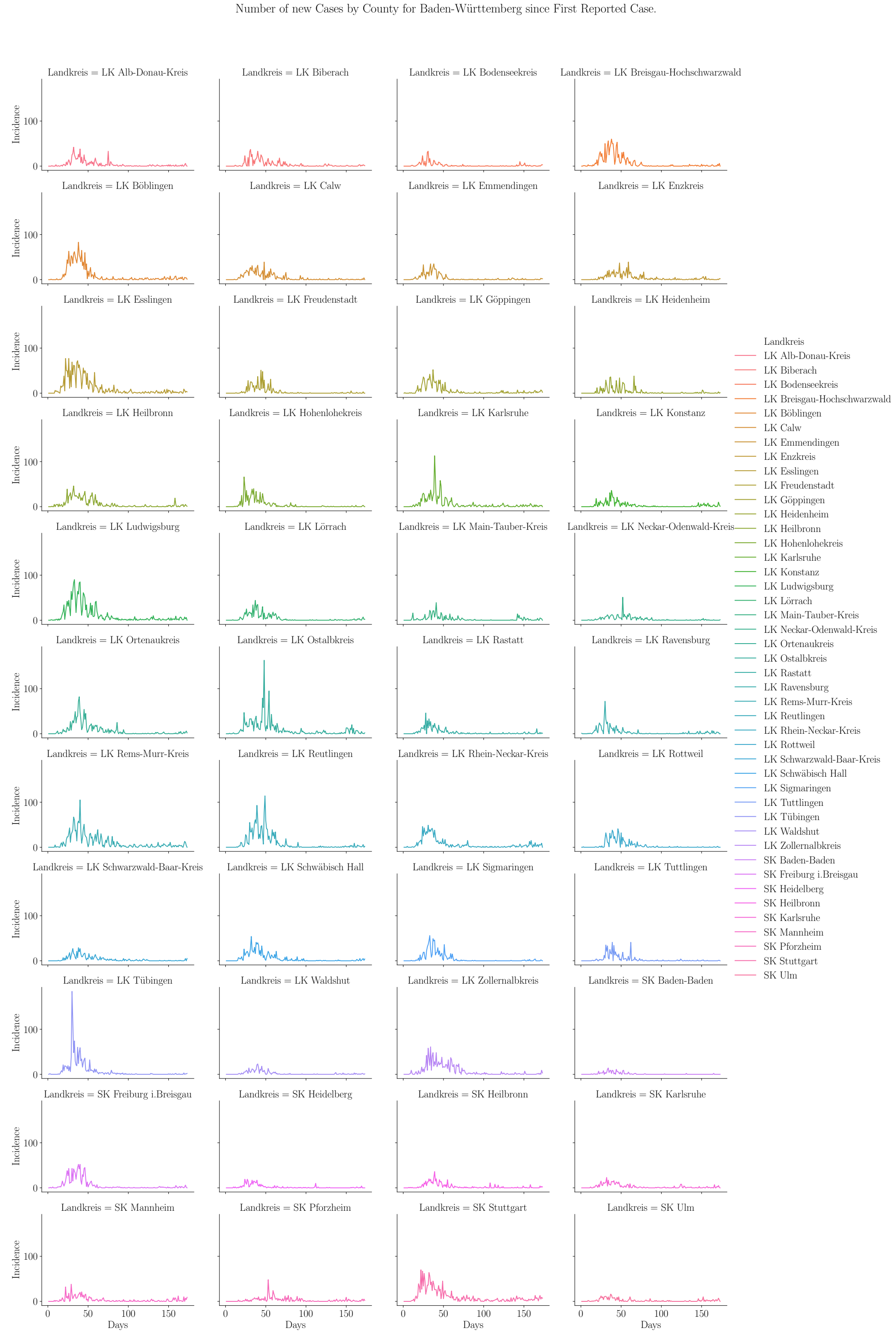


Figure B.4: Number of new cases by reporting date per county in Baden-Württemberg.

Bibliography

[1] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 2020.

[2] Johns Hopkins University, November 2020. URL `https://coronavirus.jhu.edu`.

[3] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov. Serial interval of novel coronavirus (COVID-19) infections. *International journal of infectious diseases*, 2020.

[4] CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[6] Kevin Patrick Murphy. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.

[7] Philipp Hennig. Probabilistic machine learning. lecture course, University of Tuebingen, 2020. https://uni-tuebingen.de/en/180804.

[8] David Duvenaud. *Automatic model construction with Gaussian processes.* PhD thesis, University of Cambridge, 2014.

[9] Robert Koch Institute, November 2020. URL `https://tinyurl.com/yys4vbpk`.

[10] Robert Koch Institute: COVID-19-Dashboard, November 2020. URL `https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4/page/page_1/`.

[11] Matthias an der Heiden and Osamah Hamouda. Schätzung der aktuellen entwicklung der SARS-CoV-2- epidemie in deutschland – nowcasting. *Epidemiologisches Bulletin*, 2020(17):10–15, 2020. doi: http://dx.doi.org/10.25646/6692.2.

[12] Edwin V Bonilla, Kian Chai, and Christopher Williams. Multitask Gaussian process prediction. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 153–160. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf`.

[13] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

[14] David Adam. A guide to R — the pandemic's misunderstood metric, November 2020. URL `https://www.nature.com/articles/d41586-020-02009-w`.

[15] Kevin Systrom, Thomas Vladek, and Mike Krieger. Model powering rt.live. `https://github.com/rtcovidlive/covid-model`, 2020.

[16] Zhaozhi Qian, Ahmed M Alaa, and Mihaela van der Schaar. When and how to lift the lockdown? Global COVID-19 scenario analysis and policy assessment using compartmental Gaussian processes. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] Jonas Dehning, Johannes Zierenberg, F. Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, 2020. ISSN 0036-8075. doi: 10.1126/science.abb9789. URL `https://science.sciencemag.org/content/early/2020/05/14/science.abb9789`.

Versicherung

Hiermit versichere ich, dass sich der Ausdruck dieser Arbeit und die per Email eingereichte Version nicht unterscheiden.
(Grundlage: Beschluss des Prüfungsausschuss-Vorsitzenden Kognitionswissenschaft am 18.03.2020.)

Tübingen, November 20, 2020 Unterschrift

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

I assure the single handed composition of this bachelors's thesis only supported by the declared resources.

Tübingen, November 20, 2020

Unterschrift

Hanna Dettki