Authors

Jonas Dehning, Johannes Zierenberg, F. Paul Spitzner, Joao Pinheiro Neto, Michael Wilczek, Viola Priesemann

MPI for Dynamics and Self-Organization, Göttingen

Acknowledgements.
We thank Tim Friede, Michael Wibral and Vladimir Zykov for carefully and promptly reviewing our work. We thank the Priesemann group - Matthias Loidolt, Daniel Gonzalez Marx, Fabian Mikulasch, Lucas Rudelt & Andreas Schneider - for exciting discussions and for their valuable comments. We thank the colleagues of the Göttingen Campus, with whom we were discussing the project and the case forecast in the past weeks very intensively, for their valuable input: Heike Bickeböller, Eberhard Bodenschatz, Wolfgang Brück, Alexander Ecker, Theo Geisel, Ramin Golestanian, Helmut Grubüller, Reinhard Jahn, Norbert Lossau & Simone Scheithauer.

Model and Methods

Introduction/Overview

The aim of our modeling is to forecast different scenarios on the spread of COVID-19 in Germany. Apart from picking a suitable model, the main challenge is to estimate model parameters. As a basis, we use the differential equations of the well-established SIR (Susceptible-Infected-Recovered) model. Simply put, we first infer the parameters that best describe the observed situation, and then we use those parameters to forecast future developments. For the parameter estimation, Monte Carlo importance sampling is applied on the model parameters to infer a distribution of parameters that well describes the observed data. For the forcast, we evolve the model equations with parameter samples from this distribution.

You find the results on the webpage of the [Campus Göttingen (https://goettingen-campus.de/research/szenarien-covid-19)], and the code with figures on github [github.com/Priesemann-Group (https://github.com/Priesemann-Group/covid_bayesian_mcmc/blob/master/Corona_germany_SIR.ipynb)].

Model Equations

We consider a discrete SIR (Susceptible-Infected-Recovered) model. In short, we assume that the disease spreads at rate λ from an infected person (I) to a susceptible person (S) and that an infected person becomes a recovered person (R) at rate μ, i.e.

$$S + I \xrightarrow{\lambda} I + I$$
$$I \xrightarrow{\mu} R\,.$$

This well-established model for disease spreading can be described by the following set of (deterministic) ordinary differential equations [see, e.g., Wikipedia (https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology) or recent works on the spread of covid-19 (https://arxiv.org/abs/2002.07572)]. Within a population of size N,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\lambda \frac{SI}{N}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \lambda \frac{SI}{N} - \mu I$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I\,.$$

Because our data set is discrete in time ($\Delta t = 1\ \mathrm{day}$), we solve the above differential equations with a discrete time step ($\mathrm{d}I/\mathrm{d}t \approx \Delta I/\Delta t$), such that

$$\begin{aligned} S_t - S_{t-1} &= -\lambda \Delta t \frac{S_{t-1}}{N} I_{t-1} =: -I_t^{\mathrm{new}} \\ R_t - R_{t-1} &= \qquad \mu \Delta t I_{t-1} =: \quad R_t^{\mathrm{new}} \\ I_t - I_{t-1} &= \left(\lambda \frac{S_{t-1}}{N} - \mu\right) \Delta t I_{t-1} = I_t^{\mathrm{new}} - R_t^{\mathrm{new}}\,. \end{aligned}$$

Importantly, I_t models the number of all active, (currently) infected people, while I_t^{new} is the number of new infections that is reported according to standard WHO convention. Furthermore, we explicitely include a reporting delay D between new infections I_t^{new} and reported cases when generating the forecast.

Exponential growth during outbreak onset

Note that in the onset phase, only a tiny fraction of the population is infected (I) or recovered (R), and thus $S \approx N \gg I$ such that $S/N \approx 1$. Therefore, the differential equation for the infected reduces to a simple linear equation, exhibiting an exponential growth

$$\frac{dI}{dt} = (\lambda - \mu) I \quad \text{solved by} \quad I(t) = I(0)\, e^{(\lambda-\mu)t}\,.$$

Estimating model parameters

We estimate the set of model parameters $\theta = \{\lambda, \mu, \sigma, I_0\}$ using Bayesian inference with Markov-chain Monte-Carlo (MCMC). Our implementation relies on the python package pymc3 with NUTS (No-U-Turn Sampling) (https://docs.pymc.io/api/inference.html).

The structure of our approach is the following:

- **Choose random initial parameters and evolve according to model equations.** Initially, we choose paramters θ from prior distributions that we explicitly specify below. Then, time integration of the model equations generates a (fully deterministic) time series of new infected cases $I^{\text{new}}(\theta) = \{I_t^{\text{new}}(\theta)\}$ of the same length as the observed real-world data $\hat{I}^{\text{new}} = \left\{\hat{I}_t^{\text{new}}\right\}$.
- **Recursively update the parameters using MCMC.** The drawing of new candidate parameters and the time integration is repeated in every MCMC step. The idea is to propose new parameters and to accept them in a way that overall reduces the deviation between the model outcome and the real-world data. We quantify the deviation between the model outcome $I_t^{\text{new}}(\theta)$ and the real-world data $\hat{I}_t^{\text{new}}$ for each step t of the time series with the local likelihood

  $$p\left(\hat{I}_t^{\text{new}} | \theta\right) \sim \text{StudentT}_{\nu=4}\left(\text{mean} = I_t^{\text{new}}(\theta), \text{ width} = \sigma\sqrt{I_t^{\text{new}}(\theta)}\right).$$

  We chose the Student’s t-distribution because it approaches a Gaussian distribution but features heavy tails, which make the MCMC more robust with respect to outliers [Lange et al, J. Am. Stat. Assoc, 1989] (https://doi.org/10.2307/2290063). The square-root width models the demographic noise of typical mean-field solutions for epidemic spreading [see, e.g., di Santo et al. (2017) (https://link.aps.org/doi/10.1103/PhysRevE.95.032115)].

  For each MCMC step, the new parameters are drawn so that a set of parameters that minimizes the previous deviation is more likely to be chosen. In our case, this is done with an advanced gradient-based method (NUTS (https://arxiv.org/abs/1111.4246)). Every time integration that is performed (with its own set of parameters) yields one Monte Carlo sample, and the MCMC step is repeated to create the full set of samples. Eventually, the majority of sampled parameters will describe the real-world data well, so that consistent forecasts are possible in the forecast phase.
- **Forecast using Monte Carlo samples.** For the forecast, we take all samples from the MCMC step and continue time integration according to different forecast scenarios explained below. Note that the overall procedure yields an ensemble of predictions — as opposed to a single prediction that would be solely based on one set of (previously optimized) parameters.

Constraining parameters

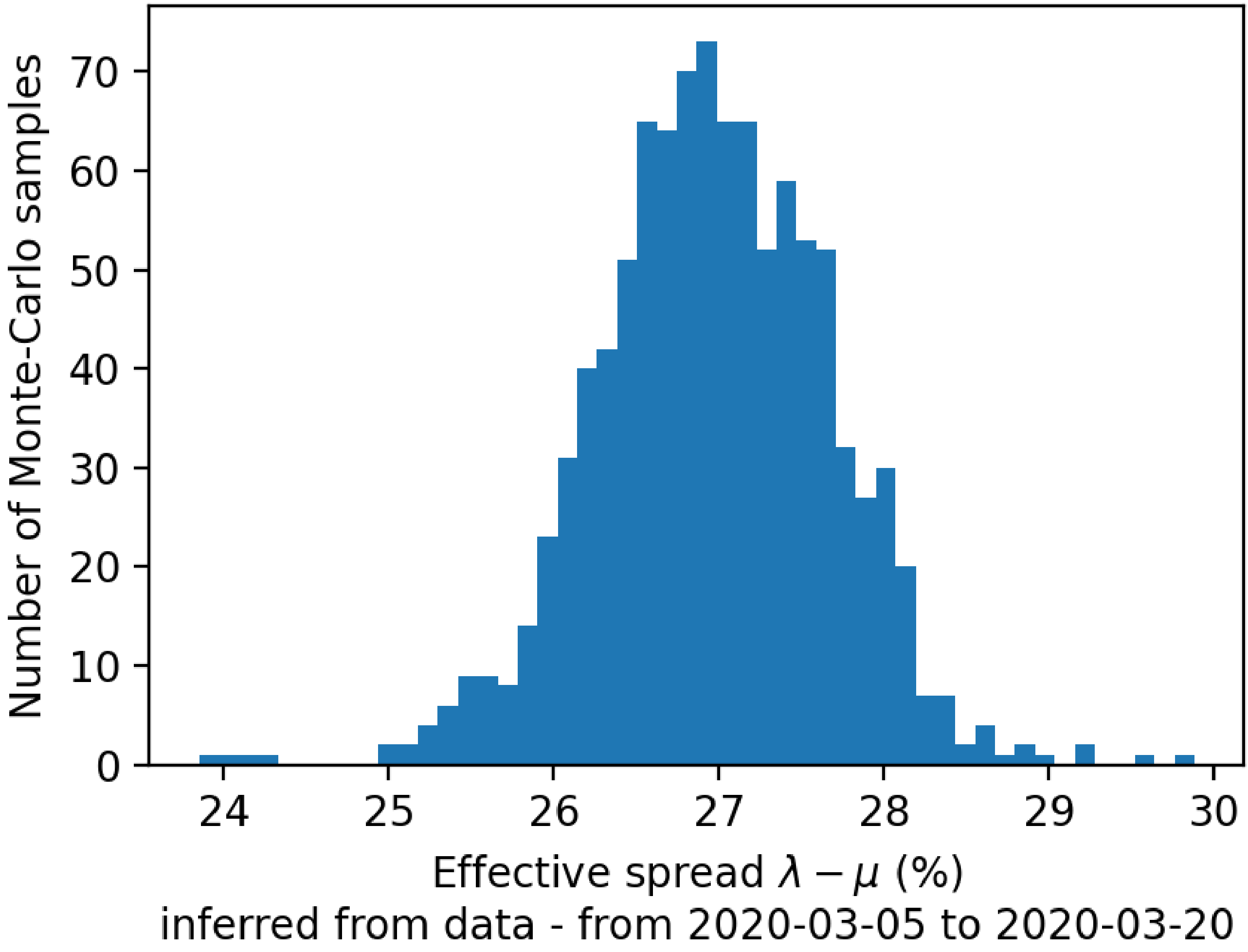
The available real-world data is not informative enough to fit all free parameters, or to empirically find the underlying distributions. We set the following informative priors on the initial model rates:

- The spreading rate is set to $\lambda \sim \mathrm{LogNormal}(\log(0.4), 0.5)$, where 0.4 is an initial guess that corresponds to 40% new infections day over day.
- The recovery rate is set to $\mu \sim \mathrm{LogNormal}(\log(1/8), 0.2)$, which corresponds to median recovery time of 8 days.

The remaining model parameters are constrained by uninformative priors, in practice the Half-Cauchy distribution [Gelman, Bayesian Anal., 2006 (https://doi.org/10.1214/06-BA117A)]:

- $\sigma \sim \mathrm{HalfCauchy}(1)$ — The scale factor of the width of the Student's t-distribution of new cases.
- $I_0 \sim \mathrm{LogNormal}(\log(\hat{I}_0), 0.9)$ — The number of infected people in the model (I_0) is constrained to be distributed around the recorded number of infected people ($\hat{I}_0$) on the day before the first data point we fit.

The MCMC sampler finds the posterior distribution $p\left(\theta \middle| \hat{I}^{\mathrm{new}}\right)$ of model parameters θ such that the model equations match the real-world data. As an example, below we plot the effective spread ($\lambda - \mu$, which corresponds to the daily rate of case increase) inferred from the data of the past.



Scenarios

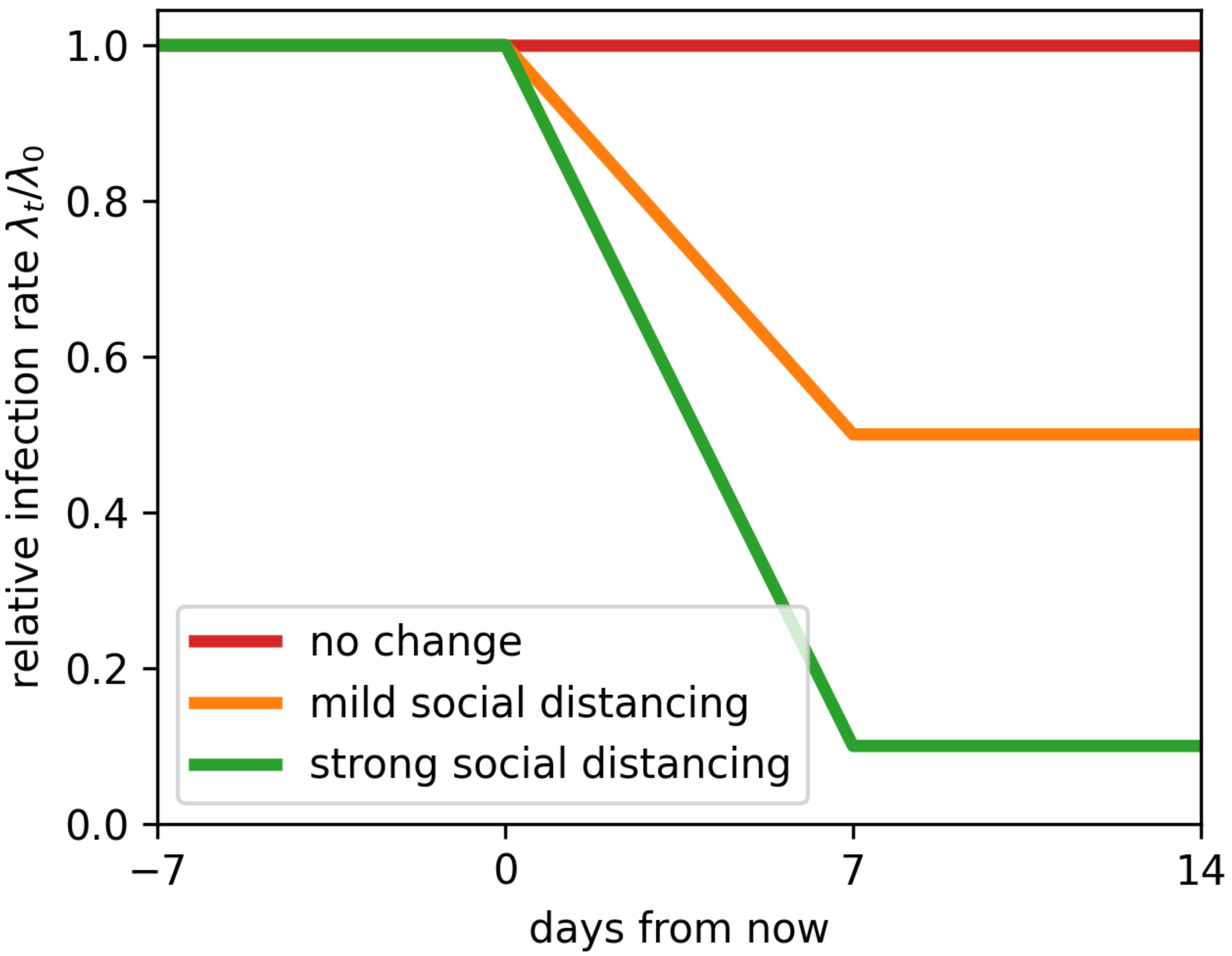
In addition to the above constraints, we add explicit prior distributions for additional parameters that are (so far) only considered for the extrapolation beyond the observed measurement. This includes

- a delay between infection and report $D \sim \mathrm{LogNormal}(\log(8), 0.1)$ assuming a median delay of 8 days.

With these priors, we now extrapolate the fitted model into three potential future scenarios. The scenarios are implemented by introducing a time-dependent spreading rate λ_t with

1. Everything stays the same and the spread continues with the inferred rate $\lambda_t = \lambda$.
2. Mild social distancing: We assume λ_t gets reduced down to 50%, linearly within 7 days.
3. Strong social distancing: We assume that λ_t gets reduced down to 10%, linearly within 7 days.

Note that, for each realization of the model, a different set of parameters (including λ_0) is drawn from the posterior distribtion $p(\theta|\hat{a})$ that was determined by the MCMC approach.



Overview of model parameters

Variable	**Parameter**
$\theta = \{\lambda, \mu, \sigma, I_0\}$	Set of model parameters that are optimized
λ	Spreading rate
μ	Recovery rate
σ	Scale factor of the width of Student’s t-distribution
N	Population size (86.700.000)
S_t	Susceptible at time t
I_t	Infected at time t
R_t	Recovered at time t
Δt	Time step
$I_t^{\text{new}} = \lambda \Delta t \frac{S_{t-1}}{N} I_{t-1}$	New infections at time t
$R_t^{\text{new}} = \mu \Delta t I_{t-1}$	New recoveries at time t
$A_t = \sum_{t'=0}^{t} I_{t'}^{\text{new}}$	Cumulative active cases until time t
$a_t = \alpha A_t$	Subsampled cum. active cases until time t
α	Subsampling fraction
D	Delay of case detection

Discussion

In the following we discuss our choice of model paramters and priors.

- **Choice of future scenarios**
  We chose to model three future scenarios. We consider two of them as quite extreme scenarios: The spread either continues as in the past (i.e. no change in social distancing), or the spread is reduced by a factor 10 (‘strong social distancing’). We display them because they set the bounds: any future scenario is likely to lie between the two.

- **Parameterizing future scenarios with changes in λ**
  We model the effect of social distancing as a reduction in λ, assuming that an infected person infects less people because e.g. of larger distance or less contacts.
  In the model, we kept μ constant. However, we expect that also μ, the rate of recovery, might increase with social distancing: People become more aware of the risk they pose to others in case of flu symptoms, and thus they “self-isolate” earlier; in this model, self-isolation is equivalent to recovery, because an isolated person does not infect other

susceptible persons anymore. However, we did not model a change in μ explicitely, but instead assume that it can be incorporated into a change of λ, as the main control parameter of the dyamics is the differnce $\lambda - \mu$ (see below).

- **Parametrization of the model with λ and μ**
  During the intial phase of exponential growth, but also further on, the dynamics is mainly determined by the difference $\lambda - \mu$. This difference is the main determinant of the daily growth rate, especially in the initial growth phase.

- **Stationarity of model parameters**
  The model parameters λ and μ are probably not stationary, but change over time, because people change their behavior over time. However, we assumed for the model inference that these two parametrs are constant over the past weeks. In the curren scenario (as of 2020/03/23), the stationarity assumption is quite reasonable, because the growth rate hardly changed. In the future, we expect that this assumption might be problematic, and that an explicit modeling of the time-dependence of change of λ and μ, e.g. via hierachical Bayesian approaches, will be necessary to make full use of the past data for prediction.