Fitting a SEIR model to COVID-19 data from Lombardy, Italy

17 April 2020

Robert Perrotta

We use the pymc3 probabilistic programming library to fit a simplified SEIR model to the COVID-19 data recorded for Lombardy, Italy by the Protezione Civile and made available at https://github.com/pcm-dpc/COVID-19 (https://github.com/pcm-dpc/COVID-19). Model assumptions are discussed and the quality of the fit model is examined.

The model has the following compartments:

- s: susceptible
- e: exposed
- $i0$: infectious with mild symptoms
- i_{0d}: i_0 patients with confirmed cases
- i_1: infectious with severe symptoms (always detected)
- i_2: infectious in critical condition (always detected)
- f: fatalities that went undetected
- f_d: fatalities from detected cases
- r: recovered patients that went undetected
- r_d: recovered patients from detected cases

Transitions in the model happen according to the following:

- s->e: susceptible patients are exposed by coming into contact with infectious patients ($s/n * (\beta_0 * i_0 + \beta_{0d} * i_{0d} + \beta_1 * i_1 + \beta_2 * i_2)$)
- e->i_0: exposed individuals develop mild symptoms and become infectious ($e * \sigma_e$)
- i_0->i_{0d}: some individuals with mild symptoms are tested and their condition is detected ($i_0 * \theta$)
- i_0->i_1: symptoms can progress from mild to severe ($i_0 * \sigma_0$)
- i_{0d}->i_1: progression for detected cases can be different ($i_{0d} * \sigma_{0d}$)
- i_1->i_2: symptoms can progress from severe to critical ($i_1 * \sigma_1$)
- i_0->r: undetected patients can recover ($i_0 * \gamma_0$)
- $i\_$->rd: likewise detected cases ($i\_ * \gamma\_$)
- i_0->f: as with recovery, undetected cases can result in undetected fatalities (are they tested after death?) ($i_0 * \mu_0$)
- $i\_$->f_d: likewise detected cases ($i\_ * \mu\_$)

Parameters used above:

- n: the total population size, which is fixed in our model
- $\beta\_$: the rate of transmission from patients in the corresponding group
  $\beta\_ = R_0 * \lambda\_$: The transmissibility, R_0, times the interaction constant $\lambda\_$
- θ: the rate of detection. Incorporates rate of testing and rate of true positives.
- $\sigma\_$: the progression factors
- $\gamma\_$: the recovery factors
- $\mu\_$: the fatality factors

Our goal is to fit this model to the data from Lombardy, Italy.

- Confirmed cases (totale*casi) are the sum of $i*{0d},i_1,i_2,f_d, $and*r_d$.
- Fatalities (dimessi_guariti) are f_d.
- i_{0d} cases are put in home isolation (isolamento_domiciliare).
- i_1 cases are admitted into the hospital (ricoverati_con_sintomi).
- i_2 cases are sent to ICU (terapia_intensiva).
- r_d cases are marked recovered (dimessi_guariti).

We also have data for total tests administered (tamponi) and total patients hospitalized (totale_ospedalizzati).

For simplicity, we model the observation errors generically as Gaussian noise with constant plus linear scaling sigma.

In [2]:

```
from itertools import chain, islice
import pickle

import numpy as np
import pandas as pd
import pymc3 as pm
from scipy.interpolate import interp1d
from tqdm import tqdm

from data import lombardia
import seir
import util

import holoviews as hv
hv.notebook_extension('bokeh', logo=False)
```

Analyzing the model fit

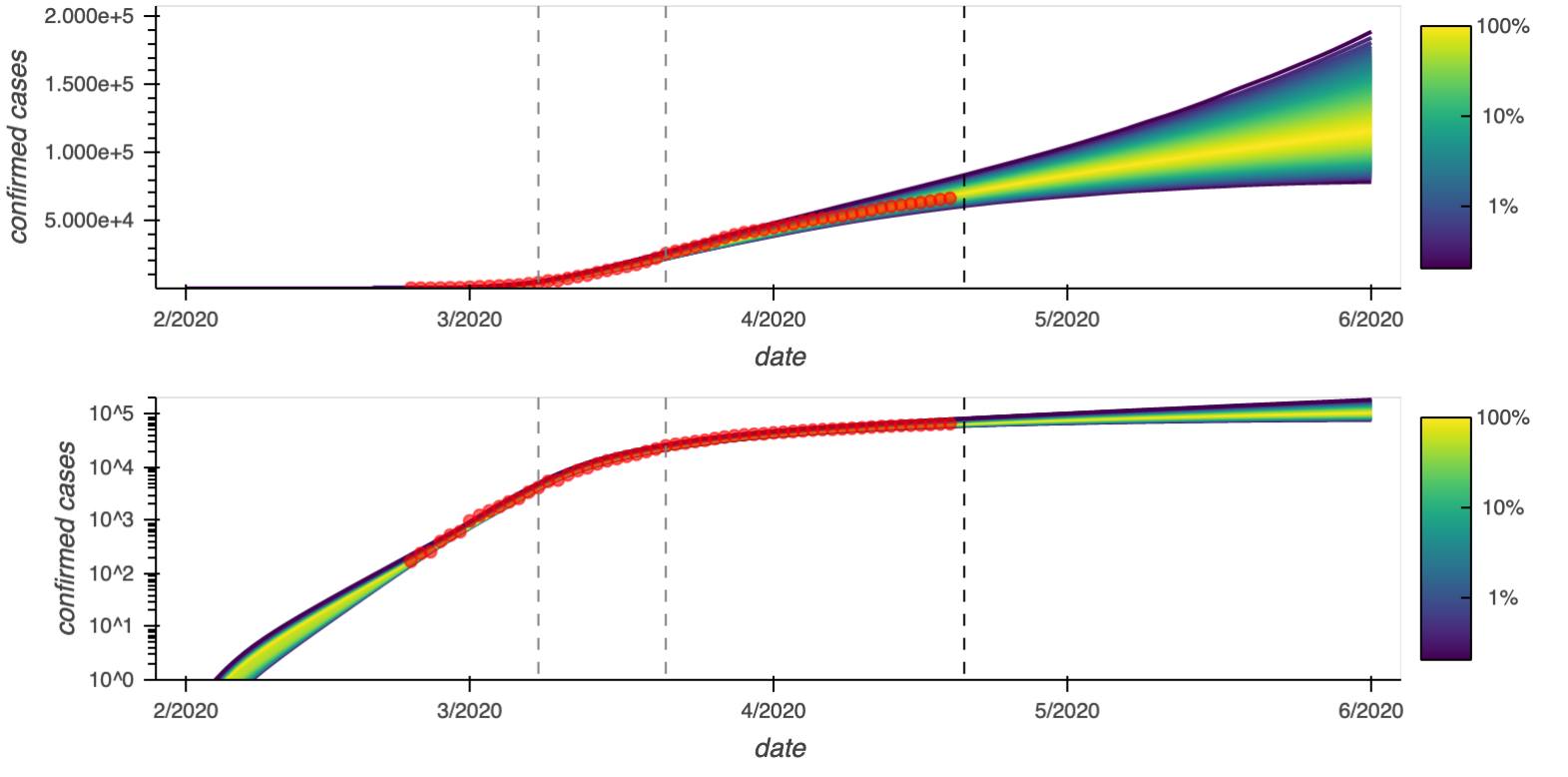
The model trace contains samples from the posterior for all our parameters. After discarding the burn-in period and sub-sampling to get greater statistical independence between samples, we can use these parameter sets to generate plausible model configurations. For each model state, instead of a single best-fit trace, we get a distribution of traces. Because probability density is not very intuitive, we instead map each trace to a probability on the cumulative distribution of our samples, then compute the tail probability, i.e. the probability of the true value being farther from the model median.

Modeled total confirmed cases

Our model makes the following distribution of predictions for total confirmed cases, which we observe to be well fit to the confirmed cases in the data. The plots below show the model predictions through the first of June assuming the current policies remain in effect. The bottom plot is identical to the top except that it's y-axis is log-scaled.

In [8]:
```
plot
```

Out[8]:

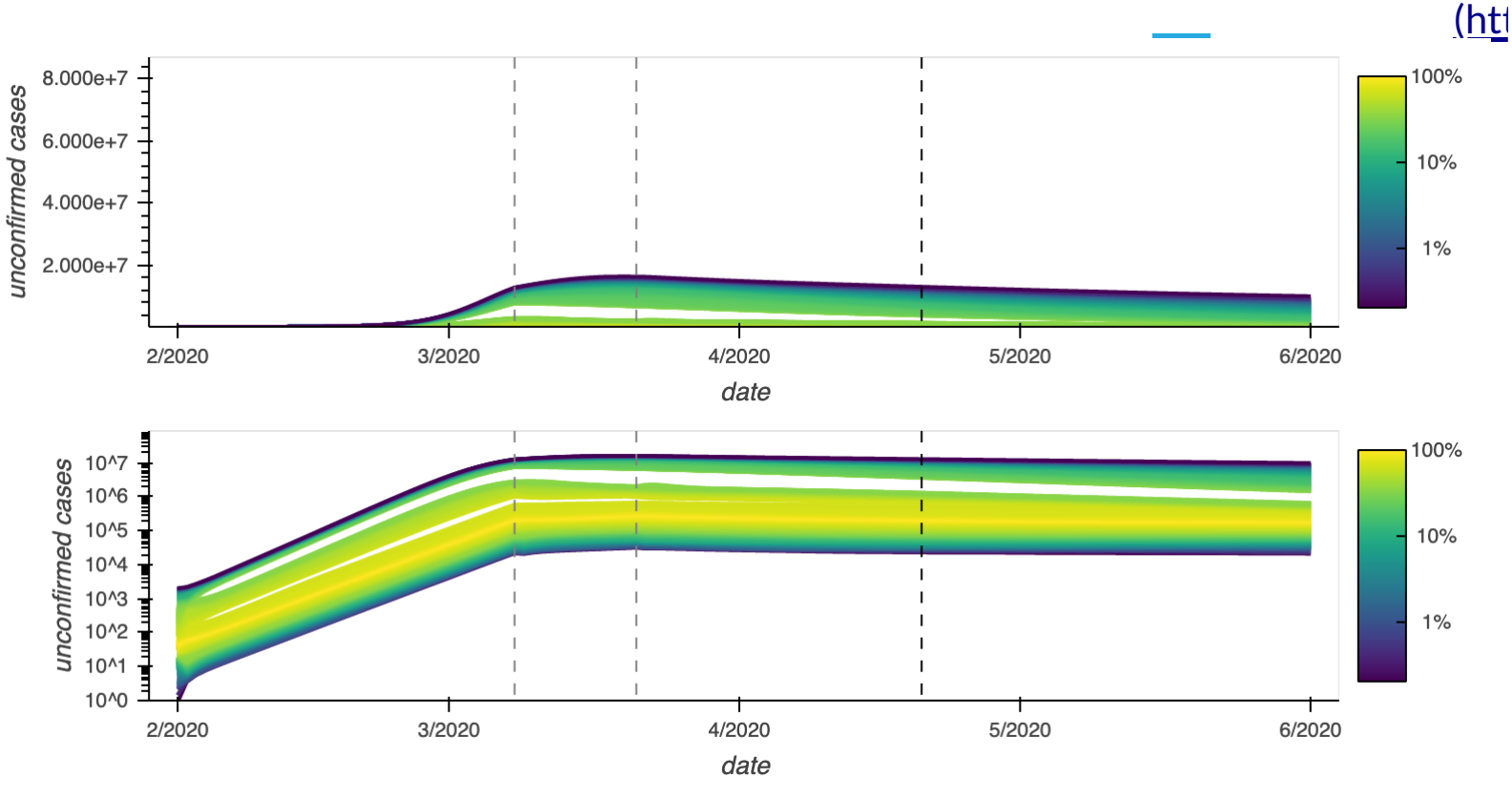


Modeled unknown cases

Our model predicts the following distribution of unconfirmed cases.

In [10]:
```
plot
```

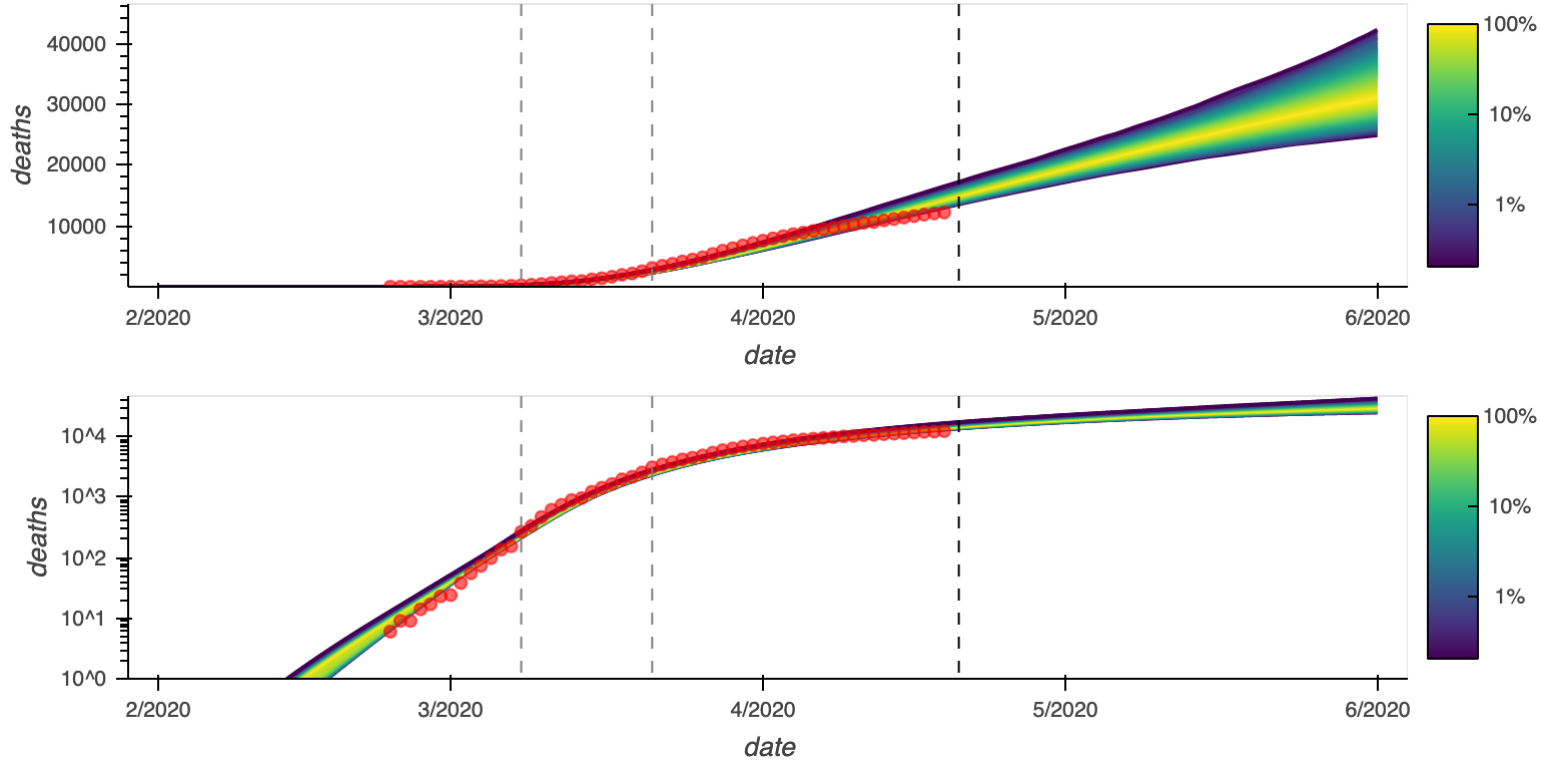
Out[10]:



Modeled number of deaths from known cases

In [12]:
```
plot
```

Out[12]:



Model assumptions and simplifications

- No resusceptibility
- No birth and no death except from COVID-19
- Model parameters are constant over time except transmission rate between unconfirmed cases, which change twice -- once on February 22nd when Lombardy was first put under lockdown and again on March 8th when Italy shut down all non-essential businesses nation-wide.
- No-one is tested until they are infectious
- Patients with mild symptoms recover at home
- Patients with severe symptoms are always admitted to the hospital
- Patients with critical symptoms are always treated in ICU
- Patients with severe or critical symptoms are always detected

Model limitations

- No attempt to compensate for reporting lag.
- Model fit diagnostics suggest poor convergence
- Slow ODE solution gradient calculation prohibits use of gradient-based samplers
- Treats Lombardy as independent, no attempt to model interactions with or to learn shared parameters from the rest of the world

Possible next steps

- Hold-out latest data to assess quality of predictions
- Develop more sophisticated models of reporting error
- Use fraction of mild/severe/critical cases, typical duration of cases, total mortality, etc. to further restrict results
- Use model to predict possible outcomes of lifting restrictions
- Explore probabilistic analysis of IHME model (healthdata.org)