Politecnico di Milano
M.Sc. in Mathematical Engineering
Numerical Analysis for Partial Differential Equations
Professor A. Quarteroni
AY 2020/2021

AR Multinomial-Dirichlet model for time series of counts of COVID-19 data

Marchionni Edoardo*

Abstract

This is the report of the project of *Numerical Analysis for Partial Differential Equations* course held by Professor A. Quarteroni at Politecnico di Milano during academic year 2020/2021. The tutors of this work are Professor N. Parolini and Professor L. Dedè.

We adapt a Bayesian AR Dirichlet-Multinomial model for COVID-19 data to the scenario of a differential model for modelling epidemic waves (SUITHER model). From one side, this goes in the direction to find a new way to calibrate it. From the other side, it represents in general a good tool for validating results and getting insights about the dynamics of the epidemics, not otherwise inferable through data.

*M.Sc. student in Mathematical Engineering student at Politecnico di Milano.

Contents

1 Introduction

1.1 SUITHER model

In the framework of the COVID-19 epidemics that has been affecting humankind since two years up to now, several different models have been proposed in literature to address the problem of forcasting the epidemiological curves. These models are either statistical or dynamical, in particular we focus our attention on a new compartmental model named SUITHER proposed in Parolini et al. 2021. SUITHER stands for the seven compartments it includes, which are

1. S: ***susceptible*** uninfected and never before infected individuals
2. U: ***undetected*** infected individuals
3. I: home ***isolated*** infected individuals
4. H: ***hospitalized*** infected individuals
5. T: life-***threatened*** individuals i.e. people in intensive care units
6. E: ***extinct*** individuals
7. R: ***recovered*** individuals.

The model is described by the following system of ordinary differential equations

$$\begin{aligned}
\dot{S}(t) &= -S(t)\frac{\beta_U U(t) + \beta_I I(t) + \beta_H H(t)}{N} \\
\dot{U}(t) &= S(t)\frac{\beta_U U(t) + \beta_I I(t) + \beta_H H(t)}{N} - (\delta + \rho_U)U(t) \\
\dot{I}(t) &= \delta U(t) - (\rho_I + \omega_I + \gamma_I)I(t) + \theta_H H(t) \\
\dot{H}(t) &= \omega_I I(t) - (\rho_H + \omega_H + \theta_H + \gamma_H)H(t) + \theta_T T(t) \\
\dot{T}(t) &= \omega_H H(t) - (\theta_T + \gamma_T)T(t) \\
\dot{R}(t) &= \rho_U U(t) + \rho_I I(t) + \rho_H H(t)
\end{aligned} \tag{1}$$

We observe that the above system is characterized by 14 parameters that have to be calibrated. The time horizon we take into consideration from now on is the second epidemic outbreak in Italy, in particular from August, 20 2020 until December, 31 2020. We proceed by splitting up this period into ten different phases, which are the followings:

- Phase 1: August 20 - September 28
- Phase 2: September 29 - October 11
- Phase 3: October 12 - October 29
- Phase 4: October 30 - November 9

- Phase 5: November 10 - November 18
- Phase 6: November 19 - November 23
- Phase 7: November 24 - December 3
- Phase 8: December 4 - December 10
- Phase 9: December 11 - December 22
- Phase 10: December 20 - December 31.

For more details about this choice, please refer to Parolini et al. 2021.
In the most general assumptions, 14 time-dependant parameters have to be calibrated. First of all, we add the further hypothesis that the parameters are piece-wise constant on each of the above phases. Also in this framework, the optimization of the entire set of 14×10 parameters is problematic, hence the following additional assumptions are made:

- β_I and β_H are set equal to 0 on each phase, this means that the infection is presumed to occur only among a susceptible individual and an undetected one
- θ_H is set equal to 0 on each phase, assuming that hospitalized people are released only once recovered
- γ_H is set equal to 0 on each phase, assuming that before passing away, critical hospitalized patients are always moved to ICU
- $\delta, \rho_U, \rho_I, \rho_H, \gamma_I, \theta_T$ are constant on the entire time interval and not only piece-wise on the different phases.

With these constraints, the number of parameter to be calibrated is reduced to $4\times10+6$, where $(\beta_U, \omega_I, \omega_H, \gamma_T)$ are the parameters that vary over the different phases and $(\delta, \rho_U, \rho_I, \rho_H, \gamma_I, \theta_T)$ are the constant ones among all the different phases.

1.1.1 Calibration[1]

In ibid. the calibration of the model is led through two different steps . First a LS optimization problem is solved numerically, then to take into consideration the uncertainty about this estimate, a Bayesian model is employed. In particular, as parameters to be inferred from the model we set the parameters of our interest, choosing as priors uniform distributions centered in the LS estimate. The likelihood is on the reported data at some previously fixed time instants, which are assumed independently normally distributed with homogeneous variance and centered in the outcome of the system. The data considered are the official daily counts related to COVID-19 epidemics by the Italian Civil Defense (Dipartimento della Protezione Civile - Presidenza del Consiglio dei Ministri) (see reference). It is noteworthy that those data do not include both the counts for undetected individuals and for the actual number of recovered cases. Indeed, the available counts of recovered

[1]Please note that the explanation that follows is not intended to be exhaustive of all the details of the calibration procedure used for SUITHER model, refer to the paper for a more extensive explanation

people are actually the counts of people recovered after tested positive to COVID-19 that had been counted among the infected individuals. This lack has two consequences. First the time series used in the calibration procedure, both in LS optimization problem and as data on which we compute the likelihood in the Bayesian model, are the one of the categories infected, hospitalized, threatened, extinct and recovered from detected. We underline that susceptible counts are not considered, since they are actually computed as a difference between the total fixed population and the counts in other categories, but in this case, as above explained, we do not dispose of all the needed time series to make this calculation meaningful. Moreover, we can notice that even in the case those time series were available, it would be at this time redundant to include susceptible counts among the available data, since the number of total individuals is assumed constant over the time. On the other hand, two more values are added to the parameters vector of the calibration procedure in the Bayesian model, which are the initial datum for both recovered and undetected people. The prior distribution of these two other parameters of the model are centered in the estimate obtained reconstructing the underlying time series through a procedure described in 2.3. Their value is inferred through the Bayesian model, since we need to account for uncertainty also for this estimate.

1.2 Research question

The purpose of the project is to find a new way to calibrate the SUITHER model. In particular, this is intended to be done adapting the Bayesian autoregressive Multinomial-Dirichlet model for time series presented in Bartolucci, Pennoni, and Mira 2021 to our case in exam.
Unfortunately, it has not been possible to carry out the whole calibration procedure, due to the fact that the adaptation of that model to SUITHER scenario revealed to be more problematic than expected.
Nevertheless, we end up with a good tool that goes towards this goal and, at the current stage, is a valid supportive instrument both to validate results and to accomplish calibration. More specifically, through this model we estimate the fluxes of individuals from the different compartments to the others at each (discretized) time instant, in particular on daily basis, complementing available data with useful insights.

1.3 Outline

In the second section first the Bayesian Dirichlet-Multinomial hierarchical model is presented in a general framework, then the AR formulation for COVID-19 is exposed. In the very last part of the section, the reconstruction procedure for the undetected and susceptible time series is explained. Last, in the third section, sampling algorithm is illustrated, this is followed by a posterior inference analysis.

2 The model

2.1 Dirichlet-Multinomial model

We now introduce in a general framework the Dirichlet-Multinomial model and the related compound distribution, which will play a key role in our model.

Primarly, consider K categories and N independent trials, each of which selects one of the categories as a success. Being $\boldsymbol{X} = (X_1, \dots, X_K)^T$ the random variable such that X_i counts the number of times over N that category i was selected, we have

$$\boldsymbol{X} \sim \text{Multinomial}(N, \boldsymbol{p}) \tag{2}$$

where $\boldsymbol{p} = (p_1, \dots, p_K)^T$ is the fixed vector of probabilities associated to each category. From this definition, it follows immediately that $\sum_{j=1}^K X_j = N$ and $\sum_{j=1}^K p_j = 1$. Furthermore, it is worth noticing that the Multinomial distribution is a generalization of both the Binomial (and hence the Bernouilli) and the categorical distribution that can be recovered assigning particular values to K and N.

On the other hand, in a Bayesian framework parameters to be inferred are considered random variables, instead of quantities that vary in a given parametric space. We hence look for an suitable distribution for the vector of probabilities $\boldsymbol{p}$. The most used one in this context is the Dirichlet distribution. Heuristically, this distribution is a multivariate generalization of the beta distribution. Each components has support $[0, 1]$ and the whole vector has consequently as support the $K-1$-simplex, since the last component is deterministic, due to the constrain $\sum_{j=1}^K p_j = 1$. We hence have

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{p} &\sim \text{Multinomial}(N, \boldsymbol{p}) \\ \boldsymbol{p} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T$ is the hyperparameter vector of the Dirichlet distribution. In particular, each component α_i belongs to $\mathbb{R}_+$.

The model we will rely on is slightly more complicated than the above one. A hyperprior π is set on the vector of hyperparameters, finally getting the following hierarchical Bayesian model

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{p}, \cancel{\boldsymbol{\alpha}} &\sim \text{Multinomial}(N, \boldsymbol{p}) \\ \boldsymbol{p}|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{\alpha} &\sim \pi(\boldsymbol{\alpha}). \end{aligned} \tag{4}$$

Concerning the above model, there is one aspect worth to be underlined. Marginalizing out $\boldsymbol{p}|\boldsymbol{\alpha}$, i.e.

$$\mathcal{L}(\boldsymbol{X}|\boldsymbol{\alpha}) = \int_{\mathbf{P}} \mathcal{L}(\boldsymbol{X}|\boldsymbol{p})\mathcal{L}(\boldsymbol{p}|\boldsymbol{\alpha})d\boldsymbol{p},$$

we have that the distribution of $\boldsymbol{X}|\boldsymbol{\alpha}$ is known in closed form, and it is called Dirichlet-Multinomial distribution. Thus, we can rewrite the model (4) in the following more compact form

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{\alpha} &\sim \text{Dirichlet-Multinomial}(N, \boldsymbol{\alpha}) \\ \boldsymbol{\alpha} &\sim \pi(\boldsymbol{\alpha}). \end{aligned} \tag{5}$$

2.2 Model specification

As above said, our model will rely on the Dirichlet-Multinomial distribution. Data consists of counts over some time grid $\mathcal{T}$ of success of K different categories. We introduce the random vector $\boldsymbol{Y_t} = (Y_{t,1}, ..., Y_{t,K})$ that represents those counts at each time occasion, with the constraint $\forall t \in \mathcal{T} \quad \sum_{j=1}^K Y_{t,j} = N$.

In out specific case, namely in the framework of SUITHER model, we set as categories the compartments of SUITHER $\mathcal{K} = \{S, R, I, H, T, U, E\} = \{1, \dots, 7\}$, where $\mathcal{K}$ represents grouping those categories. Publicly available data plus the reconstructed time series (refer to 2.3), which are the counts of each category on a daily basis, are the realization of the above random vector for all the time instants. Moreover, the time grid $\mathcal{T}$ are the days from August, 20 2020 to December, 31 2020, labelled with progressive integers starting from 1. From now on, as usual in statistical frameworks, we will denote with lowercase letters observed data and with capital letters the associated random variables.

2.2.1 AR model

The counts for the first time occasion are considered given. To build an autoregressive model for our vector, we consider at each time instant a contingency table $\mathbb{X}_t = (X_{tjk})_{j,k} \quad t > 1$ such that each row j sums up to $Y_{t-1,j}$ and each column k sums up to $Y_{t,k}$. Each element X_{tjk} of this table is the random variable representing the counts of individuals that at $t-1$ are in category j and at time t move to category k.

	S	R	I	H	T	U	E	$\boldsymbol{Y_{t-1}}$
S								$Y_{t-1,1}$
R								.
I								.
H								.
T								.
U								.
E								$Y_{t-1,7}$
$\boldsymbol{Y_t}$	$Y_{t,1}$	.	.	.	.	.	$Y_{t,7}$	N

Table 1: ...

2.2.2 Likelihood and prior distribution specification

Denote now with $\boldsymbol{X_{tj}}. \quad t > 1$ each row of the contingency table. These vectors, given $\boldsymbol{Y_{t-1}}$, will be distributed as Multinomial distributions, as follows

$$\boldsymbol{X_{tj}}.|\boldsymbol{y_{t-1}}, \boldsymbol{p_{tj}} \sim \text{Multinomial}(y_{t-1,j}, \boldsymbol{p_{tj}}) \tag{6}$$

where the number of trials are the number of individuals in category j at time instant $t-1$ that have to be reassigned to the different categories at time t and $\boldsymbol{p_{tj}}$ is a transition probability vector from category j to other categories at the same time instant. We underline that these probability vectors are the parameters of

our interest. As prior structure on them we rely on the hierarchical model (4), getting in the end for $t > 1$

$$\begin{aligned} \boldsymbol{X_{tj}}.|\boldsymbol{y_{t-1}}, \boldsymbol{p_{tj}} &\sim \text{Multinomial}(y_{t-1,j}, \boldsymbol{p_{tj}}) \\ \boldsymbol{p_{tj}}|\boldsymbol{\alpha_{tj}} &\sim \text{Dirichlet}(\boldsymbol{\alpha_{tj}}) \\ \boldsymbol{\alpha_{tj}} &\sim \pi(\boldsymbol{\alpha_{tj}}) \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha_{tj}}$ are for all category j the hyperparameter vectors of $\boldsymbol{p_{tj}}$.
As discussed in 2.1, the above model can be reformulated as follows

$$\begin{aligned} \boldsymbol{X_{tj}}.|\boldsymbol{y_{t-1}}, \boldsymbol{\alpha}_{tj} &\sim \text{Dirichlet-Multinomial}(y_{t-1,j}, \boldsymbol{\alpha_{tj}}) \\ \boldsymbol{\alpha_{tj}} &\sim \pi(\boldsymbol{\alpha_{tj}}) \end{aligned} \tag{8}$$

2.2.3 Constraints

To fit the above Bayesian autoregressive model to the case of SUITHER model, we have to impose the same constraints on admissible transitions between categories that were exposed in 1.1. In particular, such constraints are expressed setting to 0 the value corresponding to the inadmissible transitions between categories in each contingency table. In Table 2 the final contingency table for all time instants is reported.

	S	R	I	H	T	U	E	$\boldsymbol{Y_{t-1}}$
S	X_{t11}	0	0	0	0	X_{t12}	0	$Y_{t-1,1}$
R	0	X_{t22}	0	0	0	0	0	.
I	0	X_{t32}	X_{t33}	X_{t34}	0	0	X_{t37}	.
H	0	X_{t42}	0	X_{t44}	X_{t45}	0	0	.
T	0	0	0	X_{t54}	X_{t55}	0	X_{t57}	.
U	0	X_{t62}	X_{t63}	0	0	X_{t66}	0	.
E	0	0	0	0	0	0	X_{t77}	$Y_{t-1,7}$
$\boldsymbol{Y_t}$	$Y_{t,1}$	.	.	.	.	.	$Y_{t,7}$	N

Table 2: ...

2.2.4 Parametrization and hyperprior specification

In order to reduce the dimensionality of the problem, we parametrize the hyperparameters of our model. First of all, let denote with $\mathcal{D}_j \subset \mathcal{K}$ the set of categories for whom the transition is possible starting from category j. We assume that for $t > 1$, $j \in \mathcal{K}$, and $k \in \mathcal{D}_j$ each $\alpha_{tjk} \in \mathbb{R}^+$ can be decomposed in two terms, one depending on time and the other one depending on the transition we are considering, in form of a generalized linear model, namely

$$\alpha_{tjk} = \exp(\boldsymbol{f}_{tjk}^T \boldsymbol{\beta}_{jk}) \tag{9}$$

where $\boldsymbol{f}_{tjk}^T$ is a vector of polynomial regressors in time and $\boldsymbol{\beta}_{jk}$ is a vector of coefficients, that will play the role of the actual parameters of our model. We underline that among those polynomial regressors in time, we may

add other variables, such as dummy variables, in order to take into account different phases of the epidemics, due for instance to containment policies or other nonnegligible events. This will be done in our case (see 3.2). The hyperprior $\pi(\boldsymbol{\alpha_{tj}})$ is set on those vectors of coefficients. We opt for a diffuse prior that makes independent and identically distributed each of the components of a single vector. The natural choice for this is assuming that for $k \in \mathcal{D}_j$ and $j \in \mathcal{K}$

$$\boldsymbol{\beta}_{jk} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}). \tag{10}$$

The final model we get is the following one

$$\begin{aligned} \boldsymbol{X_{tj}}.|\boldsymbol{y_{t-1}}, (\boldsymbol{\beta}_{jk})_{k\in\mathcal{D}_j} &\sim \text{Dirichlet-Multinomial}(y_{t-1,j}, \boldsymbol{\alpha_{tj}}) \\ \boldsymbol{\beta}_{jk} &\overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}) \end{aligned} \tag{11}$$

for $t > 1$, $j \in \mathcal{K}$, and $k \in \mathcal{D}_j$.

2.3 Reconstruction missing input data

One of the most critical issues is the lack of counts for undetected individuals and, as a consequence, we miss also the real number of recovered individuals and susceptible ones at each time instant. We remind that, as discussed in 1.1, the number of recovered individuals reported by public data is in actual fact the number of individuals recovered after tested positive to COVID-19 and that had been counted among the infected individuals. The devised strategy to overcome this problem is to reconstruct the underlining time series, as done in Parolini et al. 2021.

First, we define a number IFR, i.e. Infection Fatality Ratio, as the ratio between the number of passed away individuals and the number of resolved case ideally at the end of the epidemic

$$\mathrm{IFR} = \frac{\mathrm{E}}{\mathrm{R} + \mathrm{E}}. \tag{12}$$

This number is assumed constant over time and estimated to be approximately 1.2%. The number of recovered individuals at a given time can be straightforward computed knowing the number of extinct individuals and the IFR, in the following manner

$$\mathrm{R}(t) = \left(\frac{1}{\mathrm{IFR}} - 1\right) \mathrm{E}(t). \tag{13}$$

On the other hand, to estimate undetected counts, we proceed introducing a time-dependent ratio CFR, i.e. Case Fatality Ratio, defined as follows

$$\mathrm{CFR}(t) = \frac{\Delta \mathrm{E}(t)}{\Delta \mathrm{R}_D(t) + \Delta \mathrm{E}(t)} \tag{14}$$

where R_D indicates the number of recovered individuals form infected people (recovered from detected) and $\Delta\cdot$ stands for an increment calculated in a time window of Δt days around time t. More explicitly, $\Delta \mathrm{E}(t) = \mathrm{E}(t + \frac{\Delta t}{2}) - \mathrm{E}(t - \frac{\Delta t}{2})$ and $\Delta \mathrm{R}_D(t) = \mathrm{R}_D(t + \frac{\Delta t}{2}) - \mathrm{R}_D(t - \frac{\Delta t}{2})$. In our case, $\Delta t = 28$.

We now assume that the variation of positive individuals in the time horizon $[\![t - \Delta t, t + \Delta t]\!]$ can be well

approximated by the variation of recovered plus extinct (i.e. resolved cases) shifted by a confirmation-to-death delay of d, that is

$$\Delta \mathrm{U}(t) + \Delta \mathrm{I}(t) + \Delta \mathrm{H}(t) + \Delta \mathrm{T}(t) \approx \Delta \mathrm{R}(t+d) + \Delta \mathrm{E}(t+d) = \frac{\Delta \mathrm{E}(t+d)}{\mathrm{IFR}}. \tag{15}$$

In a similar way, also the variation of detected positive individuals are assumed in the time window $[\![t-\Delta t, t+\Delta t]\!]$ to be approximated by the variation of recovered from detected plus extinct (i.e. detected resolved cases) shifted by the same delay d

$$\Delta \mathrm{I}(t) + \Delta \mathrm{H}(t) + \Delta \mathrm{T}(t) \approx \Delta \mathrm{R}_D(t+d) + \Delta \mathrm{E}(t+d) = \frac{\Delta \mathrm{E}(t+d)}{\mathrm{CFR}(t+d)}. \tag{16}$$

In our case, we take d equal to 13.

Now, we can obtain a relation that allows us to reconstruct the undetected time series arguing that the detecting ratio at a certain time instant t, which is the ratio between the detected cases among the total positive cases at time instant t, can be approximated by the ratio of the increment of those quantities over Δt

$$\frac{\mathrm{I}(t) + \mathrm{H}(t) + \mathrm{T}(t)}{\mathrm{U}(t) + \mathrm{I}(t) + \mathrm{H}(t) + \mathrm{T}(t)} \approx \frac{\Delta \mathrm{I}(t) + \Delta \mathrm{H}(t) + \Delta \mathrm{T}(t)}{\Delta \mathrm{U}(t) + \Delta \mathrm{I}(t) + \Delta \mathrm{H}(t) + \Delta \mathrm{T}(t)} \approx \frac{\mathrm{IFR}}{\mathrm{CFR}(t+d)}. \tag{17}$$

Making explicit from the above relation $\mathrm{U}(t)$, we obtain the following

$$\mathrm{U}(t) = \left(\frac{\mathrm{CRF}(t+d)}{\mathrm{IFR}} - 1 \right) (\mathrm{I}(t) + \mathrm{H}(t) + \mathrm{T}(t)) \tag{18}$$

being finally able to reconstruct the time series of undetected individuals. Once undetected and recovered counts are computed through (13) and (18), we can compute by difference susceptible individuals, having assumed that the total population is constant over the time.

3 Posterior inference

Our purpose is to estimate transition probabilities vectors $\boldsymbol{p_{tj}}$, namely each p_{tjk} for $t > 1$, $j \in \mathcal{K}$, and $k \in \mathcal{D}_j$. This is performed finding an estimate of coefficients $\boldsymbol{\beta}_{jk}$, in particular, being in a Bayesian framework, we will sample from their posterior distribution. We straightforwardly face two critical aspects. From one side the posterior distribution cannot be expressed in closed form, but this is common in Bayesian models and can be easily overcome using Markov Chain Monte Carlo methods. From the other side, we note that we do not observe contingency tables, despite having set the "likelihood" on them and, as a consequence, we are considering them our data from a Bayesian perspective. At our disposal, as already stated, we have daily observation, respectively daily reconstructed values for undetected and recovered, of the counts of individuals in all the categories, which represent the margins of contingency tables (see 2.2.1). This will lead to a two-steps sampling algorithm, one to sample contingency tables for each time instant $t > 1$, given their margins and the current value of the parameters, and the subsequent one to update the coefficients $\boldsymbol{\beta}_{jk}$ for $j \in \mathcal{K}$ and $k \in \mathcal{D}_j$, given the updated contingency tables.

3.1 Algorithm

In this section we will present extensively the algorithm used. As explained in the above section, we will perform two MCMC steps, in particular two Metropolis-Hastings steps.
First, contingency tables are sampled for $t > 1$, using the technique formalized in Diaconis and Sturmfels 1996. This consists in three steps:

1. we randomly select a 2x2 subtable that allows to perform the moves explained in the next point without getting a noncoherent table with respect to margin constraints,

2. perform a so called "move", which involves adding, respectively subtracting, a random integer to the cells that lie on the main diagonal of the subtable and at the same time, subtracting, respectively adding, to the two remaining cells the same integer, ending up with a new proposed table,

3. after checking that the proposed table has all the elements positive and being sure the move was performed in a way the new table is coherent with the margins, accept it according to the following acceptance rate

$$\min\left(1, \prod_{j\in\mathcal{K}} \frac{\mathbb{P}(\boldsymbol{X_{tj}}. = \boldsymbol{x^*_{tj}}.|\boldsymbol{y_{t-1}}, (\boldsymbol{\beta}_{jk})_{k\in\mathcal{D}_j})}{\mathbb{P}(\boldsymbol{X_{tj}}. = \boldsymbol{x_{tj}}.|\boldsymbol{y_{t-1}}, (\boldsymbol{\beta}_{jk})_{k\in\mathcal{D}_j})}\right) \tag{19}$$

   where $\boldsymbol{x^*_{tj}}.$ and $\boldsymbol{x_{tj}}.$ are respectively the j-th row of the proposed table and the the one of the current table.

Please note that the issue of the admissible moves is a delicate one, in particular in our case due to the null-values constraints, the moves allowed in order to get a table coherent with given margins are not many. For further details on this aspect, please refer to the literature on Markov basis.
Once the table is update, either accepting the new proposed value or maintain the old one, we update the parameters performing another Metropolis Hastings step. The proposal distribution for each $\boldsymbol{\beta}_{jk}$ for $j \in \mathcal{K}$ and $k \in \mathcal{D}_j$ is a multivariate Normal distribution centered on the current value of the vector with a suitable covariance matrix. This proposed value is accepted according to the following acceptance rate

$$\min\left(1, \prod_{t>1} \frac{\mathbb{P}(\boldsymbol{X_{tj}}. = \boldsymbol{x_{tj}}.|\boldsymbol{y_{t-1}}, (\boldsymbol{\beta}^*_{jk})_{k\in\mathcal{D}_j})}{\mathbb{P}(\boldsymbol{X_{tj}}. = \boldsymbol{x_{tj}}.|\boldsymbol{y_{t-1}}, (\boldsymbol{\beta}_{jk})_{k\in\mathcal{D}_j})} \frac{\pi(\boldsymbol{\beta}^*_{jk})}{\pi(\boldsymbol{\beta}_{jk})}\right) \tag{20}$$

where $\boldsymbol{\beta}^*_{jk}$ is the new proposed vector.
Summing up:

Algorithm 1 Metropolis-Hastings sampler for AR COVID-19

for $s = 1, \dots, S$ **do**

1. Sample the contingency tables

for $t = 2, \dots, T$ **do**

sample $\mathbb{X}_t^{(s)}$ using *Diaconis (1998)* technique through a M-H step

2. Update parameters

for $j \in \mathcal{K} \wedge k \in \mathcal{D}_j$ **do**

sample $\boldsymbol{\beta}_{\boldsymbol{jk}}^{(\boldsymbol{s})}$ from the posterior distribution through a M-H step

3.2 Sampling

In setting algorithm running parameters, we rely on the values used in Bartolucci, Pennoni, and Mira 2021, except for the number of iterations, the tinning step and the number of burning in iterations. The regression model for parameters is assumed of the third order, as the most performative one in ibid. Moreover, the variance matrix of the proposal distribution is set with all elements equal to 0.1. As far as hyperprior elicitation is concerned, we set $\sigma^2 = 100$.

On the other hand, regarding the remaining running parameters, the number of iterations were set equal to $1e+6$, with $2e+5$ burning in iterations and a tinning step of 20 iterations. The number of total iterations as well as the number of burning in ones is higher than those set in ibid. This is because the number of admissible moves in sampling contingency tables, due to constraints on transitions, is much lower than in the less complicated case simulated in the paper. This may lead to a low-converging chain, needing a higher number of iterations to reach stationarity. Furthermore, at each step, we change the value only to four entries of the considered table. This may result in high autocorrelation, i.e. the correlation between consecutive draws, and this situation is worsen in our model due to the fact that out admissible moves are not many. To overcome this problem, we set a tinning step equal to 20, hence keeping one iteration out of 20 and ending up with a sample of cardinality equal to $5e + 4$.

Last, dummies variables were introduced to take into account the different phases as in the SUITHER calibration (see 1.1).

Running time was around sixteen hours and thirty-six minutes.

3.3 Parameter estimation

First of all, we notice that the output of the simulation is a posterior sample of beta coefficients, from which we can recover a posterior estimate of alpha hyperparameters. This has no pratical interest, since our aim is to get an estimate of transition probabilities. Having a posterior estimate of alpha hyperparameters, we may infer those probabilities in many different ways. The most natural choice is the one that follows

$$p_{tjk} = \frac{\exp(\alpha_{tjk})}{\sum_{l \in \mathcal{K}} \exp(\alpha_{tjl})} = \frac{\exp(\boldsymbol{f}_{tjk}^T \boldsymbol{\beta}_{jk})}{\sum_{l \in \mathcal{K}} \exp(\boldsymbol{f}_{tjl}^T \boldsymbol{\beta}_{jk})} \tag{21}$$

for $t > 1$, $j \in \mathcal{K}$, and $k \in \mathcal{D}_j$. We underline that this is the expected value of the full conditional $\boldsymbol{p}|\boldsymbol{\alpha}$. Computing this quantity for every time instant $t > 1$, we get a value of transition probabilities for each iteration. Then, fixing a time instant and a couple of categories (i, j) for which the transition is allowed, we get an estimate of transition probability p_{tjk}. Moreover, if we compute the mean of those estimates over the different time phases, we get an estimate of the transition probabilities over the different phases and computing their variance a dispersion estimate.

In Table 3 estimate of transition probabilities of day two are reported and in Table 4 the ones for the seventieth day. We note that some changes in this table are coherent with what a priori expected. For instance, the probability to move from being a susceptible individual to being an undetected one increases from the first to the second table, this is due to the boost of infection during the second wave. On the other hand, we observe that the probability of moving from being in ICU to being just hospitalized decreases of two order of magnitude from the first to the second table. This substantial change may be caused to the fact that during late summer we observed major contagion among young people who were on vacation and attended aggregation places (e.g. nightclubs, festivals), whereas approaching the peak a more consistent number of vulnerable people to severe illness were infected. This is confirmed from the increase also in the probability of moving from being an infected isolated individual to being hospitalized. On the other hand, the probability of dying when in ICU decreases from the first to the second scenario. This has no straightforward interpretation and might be due either to some day variability, in particular in counts of death people that sometimes were grouped the same day, even if occurred during different days, or we may argue this is due to a less mortality from severe illness as time goes on, thanks to a better comprehension of the illness mechanism and the consequent treatments.

	S	R	I	H	T	U	E
S	0.9999513	0	0	0	0	0.0000486	0
R	0	1.0000000	0	0	0	0	0
I	0	0.004362	0.9934381	0.0017327	0	0	0.0004663
H	0	0.014016	0	0.9805806	0.0054025	0	0
T	0	0	0	0.0518571	0.9002312	0	0.0479116
U	0	0.0260656	0.0530876	0	0	0.92084670	0
E	0	0	0	0	0	0	1.0000000

Table 3: Posterior estimate of probability rates at time 2

	S	R	I	H	T	U	E
S	0.9988434	0	0	0	0	0.0011565	0
R	0	1.0000000	0	0	0	0	0
I	0	0.0182375	0.9764264	0.0045394	0	0	0.0007965
H	0	0.0013721	0	0.9913784	0.0072493	0	0
T	0	0	0	0.0008668	0.9989810	0	0.0001521
U	0	0.0202165	0.0440959	0	0	0.935687	0
E	0	0	0	0	0	0	1.0000000

Table 4: Posterior estimate of probability rates at time 70

In Figure 1 we observe the evolution over time of the estimate of the probability of moving from being a susceptible individual to an undetected one, in particular that is the probability of being infected, since in our model we assumed this one as the only admissible transition from susceptible compartment. In accordance to what expected, we observe a peak in this curve just before the peak of the second wave, which is around the ninetieth day. Then we observe a rapid decrease of the curve, due to the social distancing and other restrictions imposed by the government to contain the spread of the epidemics. In the last part, an increasing tail is present, this is coherent with the relaxation of the restriction in the late phases.
On the other hand, in Figure 2 the evolution of the estimate of the probability of being detected once infected is reported. We note a first increase, this is due to people being tested after summer vacation, either under obligation coming back from other countries or by personal initiative. A decrease is observed since the contagion starts to spread consequently, where a reversal of the trend is remarkable until the peak of the second wave. After the peak, the curve goes down until an increasing final tail. This is not surprising, indeed we know the phenomenon of many people deciding to test themselves in view of Christmas season. We hence witness also in this case a behaviour consistent with the phenomenon.

3.4 Diagnostics

Markov chain Monte Carlo diagnostics is based mostly on graphical tools. Indeed, convergence is theoretically guaranteed, but the needed iterations to reach stationarity are not a priori known. One of the most used tools are trace plots, this is to check whether the chain has reached stationarity, and hence if we are sampling from the target distribution. Furthermore, at the same time it can help detecting whether the chain is autocorrelated, meaning that two subsequent draws are not independent. Many expedient are put in place during sampling to reach stationarity and try to reduce autocorrelation, please refer to 3.2 for more details. In our case, the complete graphical diagnostics cannot be performed, since we have too many parameters, namely 13 betas coefficients - considering third order polynomial plus 9 dummies for 10 phases - for each of the admissible transitions, which are 17. Nevertheless, a partial convergence analysis has been performed.
In general, the chain presents satisfactory convergence properties, despite observing in part of the parameters some autocorrelation. This is not surprising for the reasons already explained in 3.2.

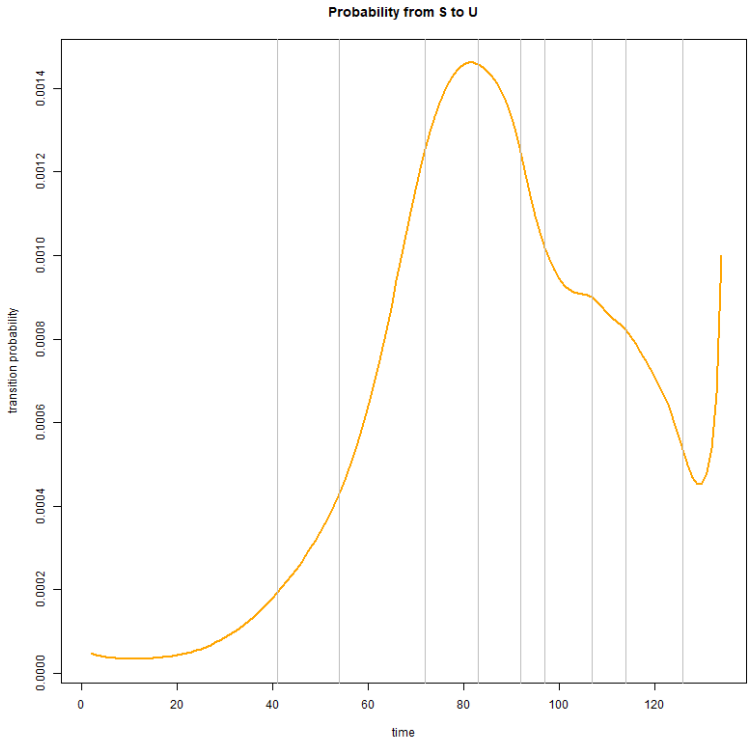

Figure 1: Evolution over time of the probability of moving from S to U, grey lines correspond to the different phases identified in the model

In Figure 3 we observe the trace plot of some selected coefficient, in particular the fourth and the fifth component of beta coefficients related to susceptible-susceptible transition and susceptible-undetected transition. We observe a light trend in subsequent iteration that is an indication of presence of autocorrelation, but in general the chain is able to explore the sampling space.

Our target is to estimate transition probabilities, hence we may analyze also their trace plots. We underline that those probability rates are not actually sampled, since we estimate them as a posterior expected value (see 21). Anyway, since their value is directly correlated to the sampling chain, their trace plots may give interesting insights to evaluate the goodness of the model. In Figure 4 we report the trace plots of transition probabilities at time 2 related to transitions from susceptible to susceptible and undetected, from hospitalized to threatened and from threatened to hospitalized. In general the chain explores well the space of possible values, however high peaks are observable and may hide some autocorrelation trends due to a nonoptimal scaling of the figure.

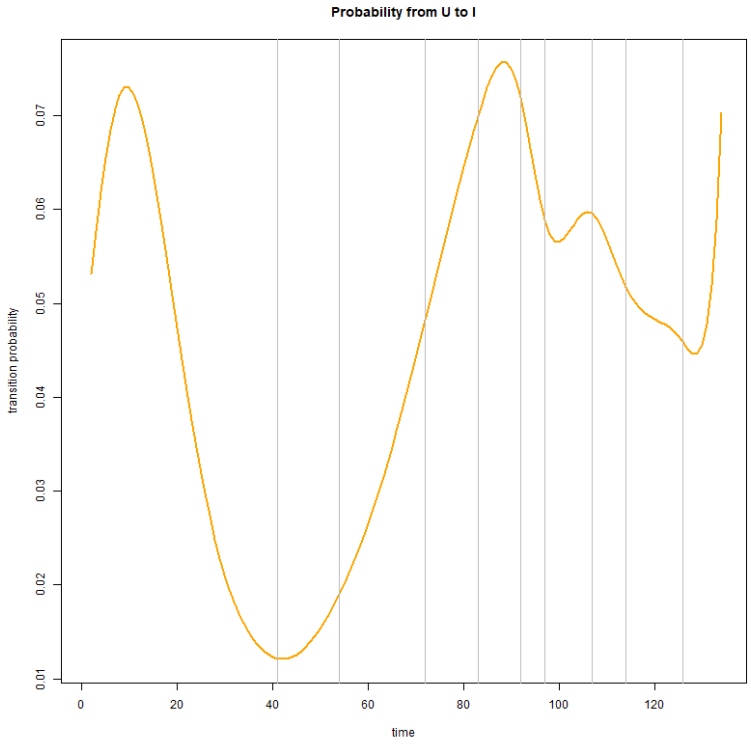

Figure 2: Evolution over time of the probability of moving from U to I, grey lines correspond to the different phases identified in the model

3.5 Contingency tables sample

During sampling, we get also a sample of contingency tables at each time instant. This may be useful to get a further insight in investigating transition trends. However, since the chain that samples contingency tables is not able to explore fast the sample space, this sample depends consistently on the initial values. In this regard, the chain was not completely randomly initialized, whereas the most of the transitions from one day to the following were believed to be for a major part among the same categories, namely the individuals in the main diagonal of each table were the most part of the minimum among the corresponding column sum counts and row sum counts. More specifically, tables were initialized first fixing the constrained values that are for each table the counts of susceptible-susceptible, susceptible-undetected, undetected-undetected, recovered-recovered, and extinct-extinct transitions. Then the minimum of the remaining not assigned individuals was computed for each row and column and then assigned for the major part to the main diagonal (90%), last the remaining not already attributed counts were assigned as randomly as possible.

Contingency tables can be estimated for instance computing the mean of each cell over the different iterations, an estimate of upper and lower bound may be given through 0.025 and 0.975 quantiles. In Table 5 we report an example of contingency table, in particular we report the estimated upper and lower bound of counts at time 2.

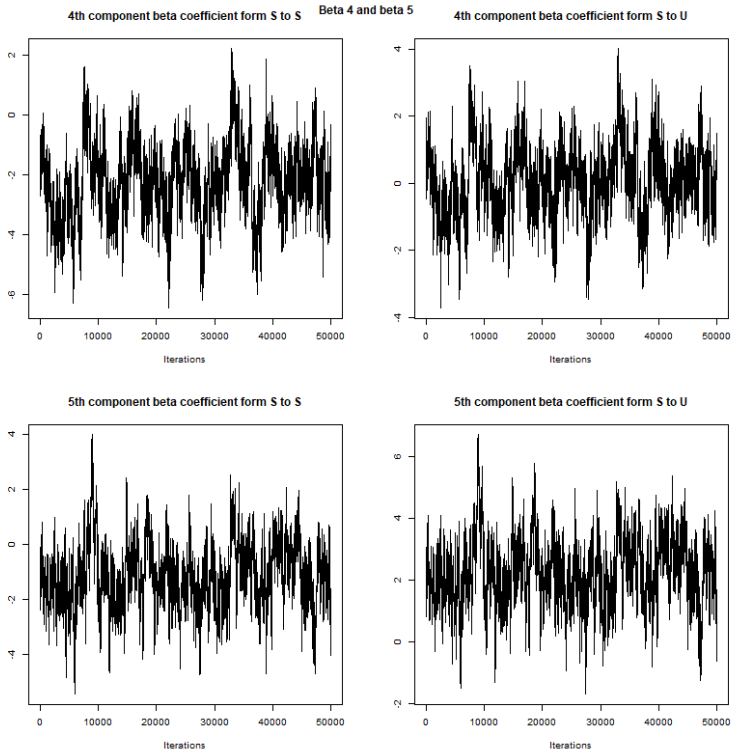

Figure 3: Trace plots of the 4th and the 5th components of the beta coefficients referring to the transition from S to S and S to U

4 Conclusions and further developments

The proposed model behaves in general as expected and is able to model efficiently the phenomenon. From one side, it might be a valid support tool to learn insight about the evolution of the epidemics, not otherwise inferable through publicly available data. This may help in structuring dynamical models, to help their calibration, or even as validation tool. From the other side, it might represent the starting point for calibration procedure for SUITHER model. In particular, relation between SUITHER parameters and transition probabilities may be derived and hence we are able to proceed to the calibration through the output of this AR model.

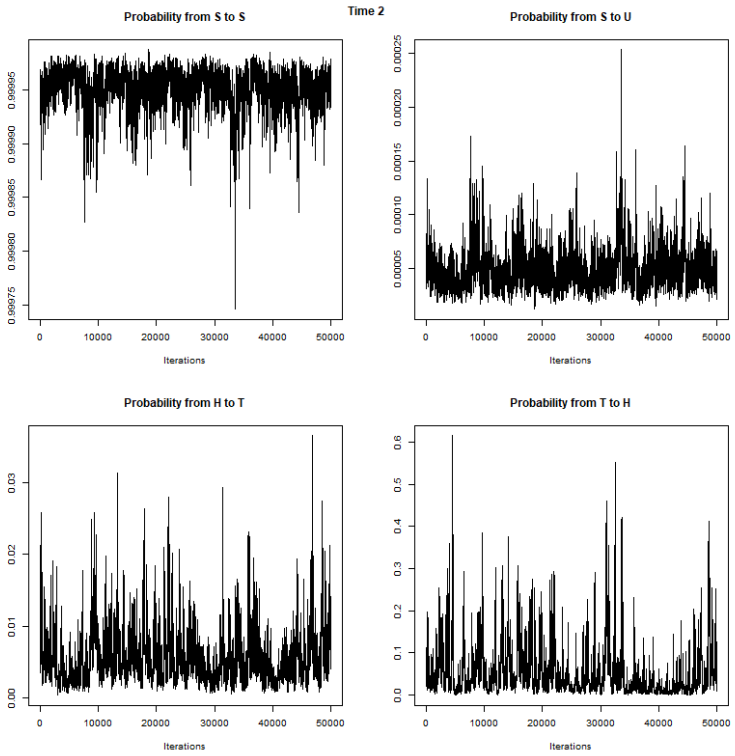

Figure 4: Trace plots of transition probabilities at time 2 related to transitions from S to S, S to U, H to T and T to H

	S	R	I	H	T	U	E
S	57335765	0	0	0	0	1448	0
R	0	2916082	0	0	0	0	0
I	0	(0, 371)	(14632, 15017)	(37, 97)	0	0	(0, 9)
H	0	(0, 55)	0	(817, 882)	(1, 19)	0	0
T	0	0	0	(0, 13)	(50, 68)	0	(0, 9)
U	0	(356, 741)	(673, 1058)	0	0	10859	0
E	0	0	0	0	0	0	35418

Table 5: Estimated upper and lower bound for counts at time 2, please note that for constrained values no interval is reported

Code

All the analysis were implemented in
`R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.`

Codes are publicly available in this Github repository. The output of the model is too heavy to be uploaded, hence please ask the author for it.
The code is structured as follows:

- `Script.R`: R-script containing all the steps to build the data set from raw data, estimating the missing time series, and to run the sampler
- `MCMC`: folder containing the sampling function and a folder `Initialization tables` containing all the code needed to initialize table plus the file `initial_tables.RData` with the used initial tables
- `Posterior inference`: folder contanining a script for performing posterior inference and a script for MCMC diagnostics plus the estimated values and the pics used in this report.

R-packages

1. `coda`: Martyn Plummer, Nicky Best, Kate Cowles and Karen Vines (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, vol 6, 7-11
2. `extraDist`: Tymoteusz Wolodzko (2020). extraDistr: Additional Univariate and Multivariate Distributions. R package version 1.9.1. https://CRAN.R-project.org/package=extraDistr
3. `pcmbapply`: Kevin Kuang, Quyu Kong and Francesco Napolitano (2019). pbmcapply: Tracking the Progress of Mc*pply with Progress Bar. R package version 1.5.0. https://CRAN.R-project.org/package=pbmcapply
4. `splines`, `stats` as part of R Core Team

Data

Data that are used in this analysis are publicy availabe at https://github.com/pcm-dpc/COVID-19.

References

Bartolucci, F., F. Pennoni, and A. Mira (2021). "Statistical approach to predict COVID-19 count data with epidemiological interpretation and uncertainty quantification". In: *Statistics in Medicine* 30;40.24, pp. 5351–5372.

Diaconis, P. and B . Sturmfels (1996). "Algebraic algorithms for sampling from conditional distributions". In: *The Annals of Statistics* 26, pp. 363–397.

Dobra, A., C. Tebaldi, and M. Westa (2004). "Data augmentation in multi-way contingency tables with fixed marginal totals". In: *Journal of Statistical Planning and Inference* 136.2, pp. 355–372.

Gelman, A. et al. (2020). *Bayesian Data Analysis*. 3rd ed. electronic edition.

Parolini, N. et al. (2021). "SUIHTER: a new mathematical model for COVID-19. Application to the analysis of the second epidemic outbreak in Italy". In: *Proceedings Royal Society A* 477.20210027.