Time Varying Markov Process with Partially Observed Aggregate Data: An Application to Coronavirus

C. Gourieroux *, J. Jasiak ‡

first version: March 31, 2020
revised: September 18, 2020

A major difficulty in the analysis of Covid-19 transmission is that many infected individuals are asymptomatic. For this reason, the total counts of infected individuals and of recovered immunized individuals are unknown, especially during the early phase of the epidemic. In this paper, we consider a parametric time varying Markov process of Coronavirus transmission and show how to estimate the model parameters and approximate the unobserved counts from daily data on infected and detected individuals and the total daily death counts. This model-based approach is illustrated in an application to French data, performed on April 6, 2020.

Keywords: Markov Process, Partial Observability, Information Recovery, Estimating Equations, SIR Model, Coronavirus, Infection Rate.

*University of Toronto, Toulouse School of Economics and CREST, *e-mail*: gouriero@ensae.fr
‡York University, Canada, *e-mail*: jasiakj@yorku.ca,
The authors gratefully acknowledge financial support of the chair ACPR: Regulation and Systemic Risks, the ERC DYSMOIA, the Agence Nationale de la Recherche: (ANR-COVID) grant ANR-17-EUR-0010 and Natural Sciences and Engineering Research Council of Canada (NSERC).
The authors thank A. Djogbenou, C. Dobronyi, Y. Lu, A. Monfort, P. Rilstone and J. Wu for helpful comments.

1 Introduction

The aim of this paper is to address the problem of partial observability, encountered recently in epidemiological research on Covid-19. More specifically, some individuals are infected and asymptomatic. Therefore, they remain undetected and unrecorded, especially during the early phase of the epidemic [1]. As a consequence, the total count of recovered and immunized individuals is unknown, as only the number of recovered detected individuals is available. This problem of partial observability of counts renders difficult the estimation of an epidemiological SIRD (Susceptible, Infected, Recovered, Deceased) model, extended to disentangle the infected and undetected from the infected and detected individuals. Moreover, such substantial undocumented infection can facilitate fast transmission of the virus (Li et al.(2020)).

The unknown total counts of infected individuals can be approximated by sampling the population daily and performing serological tests on the sampled individuals to estimate the rates of infected undetected and recovered individuals. However, it takes time to validate and produce reliable serological tests for Covid-19. Moreover, regularly performed sampling can be costly, especially in terms of time of health care providers. The alternative method, proposed in this paper, is purely model-based. Loosely speaking, under the standard extended SIRD model, the evolution of death rates might be different, depending on whether all infected individuals are detected or not. This implied difference will allow us for a model-based estimation of the proportions of infected undetected individuals (resp. recovered immunized) [see, Verity et al.(2019) for pure model based estimation of coronavirus infection, Manski, Molinari (2020) for set estimation of the infection rate].

This paper discusses the general case of time varying Markov processes when aggregate counts are partially observed. It is organized as follows. Section 2 describes the latent model of qualitative individual histories. These histories follow a time varying Markov process with transition probabilities that can depend on latent counts and unknown parameters. The observations are functions of the frequencies of individual states (called compartments in epidemiology), although not all of those frequencies are observed, in general. More specifically, only some states can be observed and/or a sum of frequencies

[1] Even though some data on asymptomatic ratios are available [see e.g. Nishiura et al.(2020)], some individuals may remain undetected for other reasons, e.g. an individual may refuse to be tested, or get a false negative tests result.

over subsets of states can be observed. Section 3 introduces the estimation method, which jointly estimates the unknown parameters and the unknown state probabilities. We derive the asymptotic properties of the estimators under identification. Identification, which is the main challenge of the proposed approach, is the topic of Section 4. First, we discuss the identification in a homogeneous Markov process, when the transition matrix is not time varying. Without additional restrictions on the transition probabilities, that model is not identifiable and the proposed approach cannot be used. However, it is not the case for a time varying Markov process that includes contagion effects and, in particular, for the SIR-type models used in epidemiology. The estimation approach is illustrated in Section 5 with a SIR type model for French data. Section 6 concludes. Some technical problems are discussed in the Appendices.

2 Latent Model and Observations

2.1 Latent Model

We consider a large panel of individual histories $Y_{i,t}, i = 1, ..., N$, $t = 1, .., T$,where the latent variable is qualitative polytomous with J alternatives denoted by $j = 1, .., J$.

Assumption A1: The individual histories are such that:
i) The variables $Y_{i,t}$, $i = 1, .., N$, at t fixed, have the same marginal distributions. This common marginal distribution is discrete and summarized by the J-dimensional vector $p(t)$, with components:

$$p_j(t) = P(Y_{i,t} = j).$$

ii) The processes $\{Y_{i,t}, t = 1, .., T\}$, $i = 1, .., N$, are independent (heterogeneous) Markov processes with transitions between times $t-1$ and t summarized by a $J \times J$ transition matrix $P[p(t-1); \theta]$ parametrized by θ. This matrix is such that each row sums up to 1.

Thus, we consider a discrete time model applicable to data on a homogeneous population of risks. The time dependent transition matrix is written in terms of marginal distributions for compatibility with the SIR-type epidemiological models.

Let $f(t)$ denote the cross sectional frequency, i.e. the sample counterpart of $p(t)$. It follows from the standard limit theorem that:

Proposition 1: Under Assumption A1, the frequencies $f(t)$ are consistent of $p(t)$ and asymptotically normal for large N. Their variance-covariance matrix is given in Appendix 1.

This specification of the transition matrix includes the homogeneous Markov chain, when there is no effect of lagged $p(t-1)$. It also includes the standard contagion SIR-type models used in epidemiology [see, McKendrick (1926), Kermack, McKendrick (1927) for early articles on SI and SIR models in the literature, Hethcote (2000), Brauer et al.(2001), Vinnicky, White (2010) for general presentations of epidemiological models, Allen (1994) for their discrete time counterparts, Gourieroux, Jasiak (2020) for an overview, and also examples given below].
As vectors $p(t)$ change over time, stationarity is not assumed.

2.2 Observations

In practice, the individual histories, or the counts of flows between the states [2] may not be observed, while cross-sectional frequencies are generally available. These can be the frequencies $f(t), t = 1, ..., T$, or aggregates of such frequencies.

Assumption A2: The observations are: $\hat{A}_t = Af(t), t = 1, .., T$, where A is a $K \times J$ state aggregation matrix, that is a matrix with rows containing zeros and ones. The aggregation matrix is known and of full rank K.

Example 1: When $A = Id$, all $f(t)$'s are observed. This is the case considered in McRae (1977), Miller, Judge (2015).

Example 2: In a model of the coronavirus transmission, the following 5 individual states can be distinguished: $1 = S$, for Susceptible, $2 = IU$, for Infected and Undetected, $3 = ID$ for Infected and Detected, $4 = R$ for Recovered, and $5 = D$ for Deceased. Frequencies $f_3(t)$ and $f_5(t)$ are observed, and the other frequencies are unobserved. We have a 2×5 matrix A given by:

[2] See Breto et al. (2009) for the treatment of flow information.

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

which characterizes the selection of the frequencies.

Example 3: In other applications, matrix A truly aggregates the frequencies, as for instance, in applications of cascade processes and percolation theory to an epidemiological model [3]. Let us consider a country with two regions and a SI model distinguishing these regions. We get a 4 state model: 1=S1, susceptible in region 1, 2=S2, susceptible in region 2, 3=I1, infected in region 1, 4=I2, infected in region 2. A transition model can be written at a disaggregate level to account for both disease transmissions within and between the regions. Thus, there is a competition between regions 1 and 2 as the sources of contagion. However, only aggregate data for the entire country may be available. Then, the aggregating matrix A is equal to:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Although, in general, the process of aggregate counts: $f_1(t) + f_2(t), f_3(t) + f_4(t)$ may not be Markov, it is important to consider the special case when it is, and then explore the possibility of identifying the parameters of the regional, i.e. disaggregated dynamics. This is the objective of the percolation theory [see, Garet, Marchand (2006) for a detailed analysis of competing contagion sources].

3 Estimation

Under Assumption A1, we can use the Bayes' theorem to link the marginal theoretical probabilities $p(t)$ to the transition probabilities as follows:

$$p(t) = P[p(t-1);\ \theta]'p(t-1),\ t = 2, ..., T. \tag{3.1}$$

The nonlinear implicit recursive equation (3.1) is the discrete time counterpart of the deterministic differential system, called the mechanistic model, which is commonly used in epidemiology [see, Gourieroux, Jasiak (2020)]. It defines the "dynamic equilibrium"

[3] See Good (1949) and Hammersley (1957) for the introductory articles on cascade processes and percolation, respectively.

for the sequence of cross-sectional distributions. These equations will be used as the estimating equations in the asymptotic least squares estimation method outlined below [4]. In our framework, the parameter of interest includes θ as well as the (equilibrium) sequence of vectors $p(t)$. They can be jointly estimated from the following optimization:

$$(\hat{p}(1),..,\hat{p}(T),\hat{\theta}) = ArgMin_{p(t),\theta} \sum_{t=2}^{T} ||p(t) - P[p(t-1);\theta]'p(t-1)||^2 \tag{3.2}$$

$$\text{s.t. } Ap(t) = Af(t) = \hat{A}_t, t = 1,..,T,$$

where $||.||$ denotes an Euclidean norm. This estimation is constrained to account for the positivity and unit mass restrictions on the $p(t)$'s, and for potentially other restrictions on parameter θ (see, Section 5.3).

The estimation method depends on the selected norm, such as $||p||^2 = p'p$, or $||p||^2 = p'\Omega^{-1}p$, where Ω is a symmetric positive definite matrix, or a norm, which varies during the disease transmission depending on the precision of frequencies $f(t)$ [see, Gourieroux, Jasiak (2020)].

Proposition 2: If the constrained optimization given above has a unique solution, which is continuously differentiable with respect to $Af(t) = \hat{A}_t, t = 1, ..., T$, then the estimator is asymptotically consistent, converges at rate $1/\sqrt{N}$, and is asymptotically normally distributed.
Proof: see Appendix 2.

The expression of the asymptotic variance-covariance matrix is derived by a delta method from the asymptotic variance-covariance matrix of $f(t)$ given in Appendix 1.

If $A = Id$, that is, if all frequencies are observed, we obtain the case analysed in McRae (1977). In the general framework, this optimization is not only used to estimate parameter θ, but also to approximate the unobserved marginal probabilities.

For ease of exposition, let us consider $A = Id$. The constrained optimization (3.2) can be interpreted in a (pseudo) state space framework with the measurement equation:

$$f(t) = p(t) + u(t),\ t = 1, ..., T, \tag{3.3}$$

[4] see, Godambe, Thompson (1974), Hardin, Hilbe (2003).

the deterministic state equation:

$$p(t) = P[p(t-1;\theta);\theta]'\, p(t-1),\ t = 1, ..., T, \tag{3.4}$$

and an assumption on the variance of $u(t)$, depending on the selected Euclidean norm. This is a pseudo-state space representation, rather than an exact state space representation, as the errors $u(t)$ are serially dependent [see, Appendix 1]. A Kalman filter [5] can be applied to the above pseudo-state space [for example, under the assumption of independent errors $u(t) \sim N(0, \Sigma)$] to estimate numerically equation (3.2). However, the estimated elements of the variance-covariance matrix of $\hat{\theta}, \hat{p}_t, t = 1, ..., T$ provided by a Kalman filter are incorrect due to misspecified serial dependence. The estimated standard errors can be adjusted either by applying the "sandwich" variance estimator, or by using the bootstrap. The bootstrap can additionally adjusts for the non-normality of errors $u(t)$, at the beginning of the epidemic, when the distribution may by closer to a multivariate Poisson distribution than to a normal distribution.

The condition for the uniqueness of the solution given in Proposition 2 is an identification condition, which is discussed in detail in the next section.

4 Identification Condition

In this section we discuss the (asymptotic) identification corresponding to the objective function with $||p||^2 = p'p$ given in Section 3. For a homogeneous Markov process with $\theta = P$, this objective function has a simple form, as under linear constraints it is quadratic with respect to the sequence $p(t)$. This allows us for an optimisation in two steps: first with respect to the $p(t)$'s, and next, with respect to θ after concentrating. This is the approach used below for identification [6]. Next, the analysis is extended to the SIR model to observe the outcomes of a path dependent transmission effect.

[5] a standard Kalman filter, or an extended Kalman filter [Song, Grizzle (1995), Julien, Uhlman (1997), Einicke, White (1999), Krener (2003)], or unscented Kalman filter [Wan, Van der Merwe (2000)].

[6] This numerical simplification will not arise for other measures of distance in the Cressie-Read family between probability distributions [Cressie,Read (1984), Miller,Judge (2015)].

4.1 Order Condition

By taking into account the fact that probabilities sum up to one, we can compare the number of moment conditions equal to $(J-1)(T-1)$ with the number of parameters of interest $(J-K-1)T + \dim\theta$. Therefore, the order condition is $KT-(J-1) \geq \dim\theta$. It is satisfied iff the number of days T is sufficiently large. However, in a non-linear framework the order condition is insufficient for identification, in general. Let us now consider the rank condition, which is a condition of local identification.

4.2 Rank condition for Homogenous Markov

For ease of exposition, we first consider the example of a homogenous Markov model with 3 states: $J = 3$, which is the number of states in a SIR model [see, Section 4.3]. The parameter θ includes the elements of the transition matrix P, which has 6 independent components, given that each row of P sums up to 1. We assume that the observed marginal probabilities are $p_3(t), t = 1, .., T$. Thus, we have partial observability. From the Bayes' theorem, it follows that

$$p(t) = P'p(t-1),\ t = 2, ..., T, \tag{4.5}$$

leading to $2(T-1)$ independent moment restrictions that are the estimating equations:

$$\begin{aligned} p_2(t) &= p_{12}p_1(t-1) + p_{22}p_2(t-1) + p_{32}p_3(t-1), \\ p_3(t) &= p_{13}p_1(t-1) + p_{23}p_2(t-1) + p_{33}p_3(t-1), \end{aligned}$$

or equivalently,

$$\begin{aligned} p_2(t) &= p_{12}[1 - p_2(t-1) - p_3(t-1)] + p_{22}p_2(t-1) + p_{32}p_3(t-1), \\ p_3(t) &= p_{13}[1 - p_2(t-1) - p_3(t-1)] + p_{23}p_2(t-1) + p_{33}p_3(t-1). \end{aligned} \tag{4.6}$$

To discuss identification, we search for the solutions in $\theta = P$ and $p(t), t = 1, ..., T$ of system (4.3) written for $t = 2, ..., T$. We have the following result:

Proposition 3: For a homogeneous Markov model with $J = 3$ and observed $p_3(t),\ t = 1, ..., T$, generically, i.e. up to a (Lebesgue) negligible set of parameter values, and if $T \geq 6$, we have that:

i) Parameter P is not identifiable, with an under-identification order equal to 3.

ii) There exist 3 functions of P that are identifiable. These functions are independent of T.

iii) These functions are over-identified with an over-identification order equal to $T-5$.

Proof: The proof is based on a concentration with respect to the values of $p_2(t)$. From the second equation of system (4.3), we see that $p_2(t-1)$ is a linear affine function of $p_3(t), p_3(t-1)$, with coefficients that depend on P. These linear affine expressions can be substituted into the first equation of system (4.3) to show that the observed sequence $p_3(t)$ satisfies a linear affine recursion of order 2:

$$p_3(t) = a(P) + b(P)p_3(t-1) + c(P)p_3(t-2),\ t = 3, ..., T,$$

with coefficients that depend on P. The results follow since:

i) the functions $a(P), b(P), c(P)$ are identifiable;

ii) the degree of under-identification of P is: 6-3=3;

iii) the degree of over-identification of the identifiable parameters is: $T-2-3 = T-5$.

Q.E.D.

Appendix 3 provides the expressions of functions $a(P), b(P), c(P)$ and points out that Proposition 3 holds, except for conditions that are (Lebesgue) negligible. In particular, identification requires that observations $p_3(t)$ correspond to a nonstationary episode as shown in the remark below.

Remark 1: Let π denote the stationary probability solution of the Markov chain, defined by:

$$\pi = P'\pi.$$

If the observed $p_3(t) = \pi_3$ were associated to a stationary episode, the sole identifiable function of parameters would be $\pi_3(P)$ and the under-identification degree would be equal to 6-1=5. Therefore, by observing the process during a nonstationary episode, we gain 2 identification degrees.

Remark 2: If the Markov structure is recursive, that is, if matrix P is upper triangular, the under-identification degree becomes 3-3=0, and the parameter is generically identifiable.

Proposition 3 shows that we can expect to identify the parameter of interest if we either consider a) a homogeneous Markov and constrain the parameters, as illustrated in Remark 2 by an example of the recursive system, or b) a non-homogeneous Markov discussed in the next subsection.

Remark 3: The rank condition can be derived in the general case of any number of states J and any type of partial observability of A. The relation between the observations A_t (for N large) and the parameters of interest $P,\ p(t),\ t = 1, ..., T$ is given by:

$$\begin{aligned} A_t &= Ap(t),\ t = 1, ..., T, \\ p(t) &= P'p(t-1),\ t = 2, ..., T. \end{aligned} \tag{4.7}$$

The second equation can be solved for $p(t)$ as a function of P and $p(1)$, as $p(t) = (P')^{t-1}p(1)$. Next, this expression of $p(t)$ can be substituted into the measurement equation to get:

$$A_t = A(P')^{t-1}p(1),\ t = 1, ..., T. \tag{4.8}$$

Next, we need to find the Jacobian of the transformation associating $A_1, ..., A_T$ to $P, p(1)$. This Jacobian can be obtained by considering the impact of small shocks δP and $\delta p(1)$ to P and $p(1)$ on A_t. By differentiating equation (4.8), we get a linear system in δP and $\delta p(1)$:

$$\delta A_t = A \sum_{k=0}^{t-2} \left[(P')^k (\delta P)' (P')^{t-k-2}\right] p(1) + A(P')^{t-1}\delta p(1),\ t = 1, ..., T. \tag{4.9}$$

System (4.9) can be rewritten in terms of $[vec'(\delta P'), vec'\delta p(1)]'$ as:

$$\begin{bmatrix} \delta A_1 \\ \vdots \\ \delta A_T \end{bmatrix} = \mathcal{J}[vec'(\delta P'), vec'\delta p(1)]',$$

and the rank of Jacobian $\mathcal{J}$ can be compared with the parameter dimension (taking into account the unit mass restrictions). In applications, the rank condition has to be checked for each specific model of interest, as shown above for $J = 3$.

4.3 Rank Condition in a Disease Transmission Model

Let us now consider an epidemiological model with $J = 3$ states to facilitate the comparison with the example in Section 4.2. The states of the SID model are: 1=S for susceptible, 2=I for infectious (individuals stay infectious, even if they recover), 3=D for deceased. The rows of the transition matrix are the following:

row $1 = S$: $(1-p_{13})[1-logist(a_1+a_2p_2(t-1))]$; $(1-p_{13})logist(a_1+a_2p_2(t-1))$; p_{13}

row $2 = I$: 0; $1-p_{23}$; p_{23}

row $3 = D$: 0, 0, 1

where $logist(x) = 1/[1 + exp(-x)]$ is the logistic function, i.e. the inverse of the logit function. We obtain a triangular transition matrix with state D as an absorbing state. The contagion effect is characterized by parameter a_2 and follows a nonlinear logistic function. We also expect that mortality rate p_{23} is strictly larger than mortality rate p_{13}. There are 4 independent parameters in $\theta = [a_1, a_2, p_{13}, p_{23}]'$.

Proposition 4: The SID model with observed $p_3(t)$ given above is generically identifiable. Parameter θ is over-identified with an over-identification order equal to 5.

Proof: The proof is similar to the proof of Proposition 3. The two independent moment conditions are:

$$\begin{aligned} p_2(t) &= (1-p_{13})logist[a_1 + a_2p_2(t-1)][1 - p_2(t-1) - p_3(t-1)] + (1-p_{23})p_2(t-1), \\ p_3(t) &= p_{13}[1 - p_2(t-1) - p_3(t-1)] + p_{23}p_2(t-1). \end{aligned}$$

From the second equation, it follows that $p_2(t-1)$ is a linear affine function of $p_3(t)$ and $p_3(t-1)$. Next by substituting into the first equation, we find that the observed $p_3(t)$ satisfies a nonlinear recursive equation of order 2 of the type:

$p_3(t) = a_1(\theta) + b_1(\theta)p_3(t-1) + c_1(\theta)p_3(t-2) + [a_2(\theta) + b_2(\theta)p_3(t-1) + c_2(\theta)p_3(t-2)]\, logist[a_3(\theta) + b_3(\theta)p_3(t-1) + c_3(\theta)p_3(t-2)]$.

If T is sufficiently large, this nonlinear observed dynamics allows us to identify 9 nonlinear functions of parameter θ. Thus, parameter θ is identifiable with an over-identification order equal to 5.

Q.E.D.

Remark 2 suggested earlier that the triangular form of the transition matrix alone would facilitate the identification. However, the order of over-identification reveals the additional role of the contagion effect. The nonlinear dynamics induced by the logistic transformation also facilitates identification.

Remark 4: As in the case of a homogeneous Markov process, it is theoretically possible to compute the Jacobian associating the observed aggregates A_t to the underlying parameters $\theta, p(t),\ t = 1, ..., T$. The condition on the rank of the Jacobian is difficult to interpret in epidemiological terms, except for specific models, such as the SID model given above.

5 An Illustration

This section illustrates the estimation approach and its performance in an epidemiological model. It is intended to recover the rate of infected undetected individuals, who are often asymptomatic.

5.1 The model and observations

We consider a model with 5 states: 1=S, 2=IU, 3=ID, 4=R ,5=D, and the following rows of the transition matrix:

row 1: $(1-p_{15})\pi_{11t};\ (1-p_{15})\pi_{12t};\ (1-p_{15})\pi_{13t};\ 0;\ p_{15}$,

where the $\pi_{1jt}, j = 1, 2, 3$ sum up to 1, and are proportional to:

$\pi_{11t} \approx 1;\ \pi_{12t} \approx exp[a_1 + b_1 p_2(t-1) + c_1 p_3(t-1)];\ \pi_{13t} \approx exp[a_2 + b_2 p_2(t-1) + c_2 p_3(t-1)]$

row2: $0; p_{22}; p_{23}; p_{24}; p_{25}$

row3: $0; 0; p_{33}; p_{34}; p_{35}$

row 4: $0; 0; 0; p_{44}; p_{45}$

row 5: $0; 0; 0; 0; 1$

Conditional on staying alive, the first row includes a multinomial logit model for the competing disease transmission driven by either lagged IU, or lagged ID [see, e.g.

McFadden (1984)]. The transmission parameters b_1, c_1, b_2, c_2 are non-negative and allow for different impacts of $p_2(t-1)$ and $p_3(t-1)$, as the detected individuals are expected to be self-isolated more often. There is no contagion effect from the recovered R, who are assumed no longer infectious [7]. The structure of zeros in the transition matrix indicates that one cannot recover without being infected, one cannot be infected twice [8] and death is considered as an absorbing state.

This is a parametric model with 6+7=13 parameters, i.e. the 6 parameters $a_l, b_l, c_l, l = 1, 2$ and 7 independent transition probabilities.

Among the 5 series of frequencies $f_j(t),\ j = 1, .., 5$ that sum up to 1 at each date, $f_3(t)$ and $f_5(t)$ of infected detected and of deceased, respectively, are assumed to be observed. The frequencies $f_2(t)$ and $f_4(t)$ are unobserved and will be considered as additional quantities of interest to be estimated jointly. They are crucial for a model-based inference on counts of infected undetected and of recovered immunized individuals.

As illustrated in Section 4.3, the triangular form of the transition matrix and the nonlinear doubly logistic contagion dynamic will provide generic identification.

5.2 Simulations

The above model can be used for simulation of the Covid-19 transmission for given values of parameter θ and initial value $p(1)$. These values are set as follows:

The daily mortality rates are: $p_{15} = p_{45} = 3e-05,\ p_{25} = 0.004, p_{35} = 0.013$. The mortality rates $p_{15} = p_{45}$ correspond to the long term mortality rates in France; p_{35} is an average mortality rate of individuals detected with Covid-19 in hospitals [see, Verity et al.(2020), Table 1 for a comparison], p_{35} has been fixed between those numbers to account for a lower rate due to the presence of asymptomatic individuals [see e.g. Nishiura et al.(2020) for the asymptomatic ratio].

We assume that there are about 3 times more transitions to IU than to ID,i.e.

$$\exp(a_1) = 3\exp(a_2),\ b_1 = b_2,\ c_1 = c_2,$$

[7] For viruses other then Covid-19, the recovered, immunized individuals can stay infectious.

[8] This was initially anticipated for Covid-19. Recently, it has been documented that some recovered individuals have not become immune. The number of repeated infections is too low for reliable statistical analysis.

and the transmission effects due to IU and ID, are equal, i.e. $b_2 = c_2$. Then a_2, b_2 are set such that:

$\exp(a_2) = 1e-06$ and $\exp(2b_2/1000) = 25$. These parameters have been set to provide about 60 new daily detected infections at the beginning of the epidemic for a population of 60 millions of inhabitants, and 1500 new daily infections later on, about 30 days after the beginning.

The parameters p_{23}, p_{24}, p_{34} are as follows:

$p_{24} = p_{34} = 0.03$, representing an average recovery time of about 33 days before being immunized. This average time is fixed equal for the IU and ID states in the simulation.

Rate p_{23} is fixed equal to $p_{12} = 1e - 06$.

Coefficient a_2 is strictly positive. This means that there can exist exogenous sources of infections for the population of interest, either from animals to humans, or more importantly from humans of another population to humans in the population of interest, due to either tourism, or migration. Thus, we consider an open economy from the epidemiological point of view [9]. We do not account for the increase of daily tests for Covid-19 performed during the epidemic (its effect in France during the early phase of the epidemic was negligible due to shortages of test components [10]).

Next, the parameters of the diagonal transition probabilities are computed from the unit mass restrictions on each row.

All probabilities of transitions out of the diagonal are very small as a consequence of the daily frequency of our data. The initial marginal probabilities are set equal to: $p(0) = (1, 0, 0, 0, 0)$, which corresponds to an initial population with no prior infection from the coronavirus in this population. Thus, the first cluster of infections has to be linked to travellers arriving to the country.

Two types of dynamic analysis can be performed, depending whether the sequence of $p(t)$, or the sequence of $f(t)$ are considered. The dynamics of $p(t)$'s are deterministic, and driven by the deterministic system (3.1). They provide us the dynamics of the expected values of $f(t)$'s. The dynamics of $f(t)$'s are stochastic with trajectories obtained

[9] The idea of collective immunity, which implies that the infection disappears if more than 60% of people are immune, implicitly assumes a closed economy. It is valid for the world in its entity, but not for each open country separately.

[10] Our model does not take into account the reliability of the tests for Covid-19, i.e. the proportion of false negative outcomes.

by simulating the time varying Markov process. As an additional outcome, the difference between the $p(t)$'s and $f(t)$'s provides a measure of uncertainty on any predictions obtained from the deterministic model of $p(t)$'s [see, Appendix 1 for the autocovariance function of $u(t) = f(t) - p(t)$].

Figure 1 shows the evolutions of $p_2(t), p_3(t), p_4(t), p_5(t)$ in separate panels as their ranges and evolutions differ, due to the selected parameter values. In addition, Figure 1 illustrates the effect of an increase (decrease) of transmission parameters b_1, b_2, c_1 and c_2 on the marginal probabilities.

[Insert Figure 1: Evolutions of Marginal Probabilities]

The solid lines represent the trajectories of $p(t)'s$ computed from the baseline parameter values given above. The dotted and dashed lines, respectively, depict the trajectories obtained when parameters b_1, b_2, c_1 and c_2 increase and decrease by a factor of 2, respectively.

The change of transmission parameters has an impact on the shape of curves, resulting in faster (slower) rates of increase in all panels, except for the bottom right one. The dynamic of $p_5(t)$ does not seem affected, as the trajectories computed from the baseline and increased (decreased) parameter values overlap one another.

Figure 2 displays the evolutions of $p_3(t) - p_3(t-1)$ and $p_5(t) - p_5(t-1)$ multiplied by the total size of the population, i.e. 60 millions. These are the new counts of ID, to be compared with the health system capacity, and the numbers of new deaths D, including, but not limited to the confirmed deaths from Covid-19.

[Insert Figure 2: Evolution of New Counts]

As before, the solid lines represent the trajectories computed from the baseline parameter values and the dotted and dashed lines show the trajectories obtained by increasing and decreasing the parameter values, respectively. A change in transmission parameters affects the shape of the curves of new counts, resulting in higher (lower) growth rates of new counts.

The evolutions are computed over a period of 60 days, i.e. 2 months. During this episode, the total number of infected individuals remains rather small, as compared to the size of the population and so does the total count of deaths. The above figures have to be interpreted in terms of stocks and flows as the numbers associated with R and D (resp.

IU, ID) are cumulated and are interpretable as stocks (resp. flows). This cumulation effect explains the increasing patterns in Figure 1, with higher rates for higher values of transmission parameters.

The counts of individuals in the two Infected states IU and ID are flows, as they are observed between the times of entry in, and exit from the state of infection. Moreover, the probability of exiting after 20 days is very close to 1. We usually expect a "phase" transition effect: For small t, these counts increase quickly as new infected individuals are cumulated without a sufficiently high number of exits to compensate for the arrivals. This explains an increase of the curves at the beginning of the period. After that initial period, the counts of exits tend to grow and offset the new arrivals so that the curves tend to flatten. More precisely, they continue to increase, due to the disease transmission effect, but at a very low rate. This is the so-called flattening of the curve. This theoretical evolution depends on the choice of parameter values, especially the transmission parameters. Given the selected parameter values that allow for exogenous sources of infection, the initial convex pattern in the counts of infected is not visible. Only the concave part of the curve, up to its flat part, is observed. One can perform similar dynamic sensitivity analysis for other credible scenarios.

The Figures given above have been simulated with time independent propagation parameters. A self-isolation measure introduced at some point would have changed subsequent evolution. There is first a tendency to reach a flat part on the curve without self-isolation, and then to reach a lower flat part on the curve with self-isolation measures. Therefore, over a longer period, the first flat part can appear as a smoothed peak. If self-isolation measures are lifted afterwards, a second peak of infections is expected, and so on, resulting in a sequence of stop and go [Ferguson et al.(2020), Gourieroux, Jasiak (2020)].

5.3 Estimation

This section presents the estimation of the extended SIRD model from data on Covid 19 transmission in France over the period of 22 days between 03/16 to 04/06, 2020.

The model introduced in Section 5.1 assumes a stable environment of constant social distancing measures, which was the case in France during the observation period. A total

lock-down was implemented on the weekend of March 16 (after the first round of municipal elections), with the closure of shops, schools, universities and strict social distancing rules. This self-isolation measure had an impact on the spread of the disease, especially on the transmission parameters and some mortality parameters [11]. To detect that effect, it would be necessary to estimate separately the model over the periods of March 1 to 15, and March 16 until April 7, which would be possible as these periods are sufficiently long for identification (see Proposition 3) [12]. Then, we could compare the results to measure the efficiency of the lock-down and perform predictions including the effects of different stages of reopening.

We focus on the second period which is sufficiently stable for the estimation purpose. The fully observed states are the states ID and D. State ID is assumed equivalent to hospitalization, as the counts of ("confirmed") detected, which are publicly available, are measured with error and are not reliable. This is due to the counts of detected individuals being derived from the PCR test results, while not all tests results may have been recorded, some people could have been tested multiple times, inflating the counts, or people might have not been tested at random, or without an adequate exogenous stratification, which creates a selectivity bias [13] [14]. In contrast, the hospitalization data are more reliable and regularly updated. State D is assumed observed through total death counts. These include deaths from Covid-19, which are reported on-line as D/H, i.e. death after hospitalization, and are known to underestimate the true number of deaths due to the coronavirus, as they do not include all deaths from Covid-19 at home, or in the long-term health care institutions.

The series to be estimated are the theoretical proportions of infected undetected IU and recovered R. We use the available series of ("confirmed") detected and of recovered

[11] The effect of Covid-19 on the total mortality rate is unclear. There is a negative effect of the virus. However, there also are some positive effects due to better protection against other viruses, such as the influenza, and a reduced number of car accidents.

[12] It is not the case for countries where the outbreak is very recent, or self-isolation implemented too late, or data are unreliable at the beginning of the outbreak (Wuhan), or the isolation period is too short (Denmark), or introduced in successive steps, or self-isolation measures are different across the regions (Germany and the US).

[13] During this period, the PCR tests were processed only in hospital laboratories, as private laboratories were not sanctioned. Moreover, the serological tests were not publicly available or officially authorized.

[14] Similar data are used for estimation in Manski, Molinari(2020), but not adjusted for the significant selectivity bias.

after hospitalization, for comparison with the estimates.

More specifically, we use the French data on the total daily number of deaths from the French National Statistical Institute INSEE (2020) and the daily data on coronavirus pandemic from Sante Publique France (2020) reported at https://dashboard.covid19.data.gouv.fr/ and https://www.linternaute.com/actualite/guide-vie-quotidienne/2489651-covid-19-en-france-les-dernieres-statistiques-au-06-avril-2020/, available on April 06 [15]. The daily evolutions of total counts of hospitalized, detected, recovered and deceased individuals reported by these sources on April 6 are displayed in Figure 3. Note that the data used in this study can differ from the data currently reported, due to updating. In particular, the daily data on overall death counts in France have been since updated and adjusted for individuals deceased at home or in long term health care facilities. For example, the new records report 2713 deaths on April 6, 2020, as compared to the initially reported number of 2401 used in this study.

[Insert Figure 3: Evolution of Observed Counts, 03/16 to 04/06, France]

The panels display the series of "hospitalized", "confirmed" (i.e. detected), "returned from hospital" (i.e. recovered after hospitalization) in the top row and left bottom panels, respectively. In the bottom right panel, the dynamics of counts of total deceased (solid line) and deceased due to Covid-19 (dashed line) are distinguished.

The model introduced in Section 5.1 has been estimated by optimizing objective function (3.2) under the constraints of positivity, unit mass of the rates and non-negativity of the transmission parameters b_1, c_1, b_2, c_2. The results are as follows: The estimated coefficients are $a_1 = -8.6517, a_2 = -11.1481$, $b_1 = 0.0034$, $b_2 = 2.499e-05$, $c_1 = 8.482e-05$ $c_2 = 0.00028$. The estimated coefficient of mortality rate p_{15} is 3.1575e-05, which is close to the mortality rate in France of 3e-05 = 0.03/1000, used in the simulation study in Section 5.2. The remaining estimated parameters of the transition matrix are given below in Table 1:

Table 1. Estimated Transition Matrix

[15] The size of the French population is 66,9 millions of inhabitants.

	1= S	2=IU	3=ID	4=R	5=D
2=IU	0	0.9022	0.0386	0.0571	0.00207
3=ID	0	0	0.7926	0.1032	0.0158
4=R	0	0	0	0.9999	1.514e-5
5=D	0	0	0	0	1

As pointed out in the simulation, some parameters, such as transmission parameters and transition probabilities are very small, and difficult to estimate. These parameter values are determined by their epidemiological interpretation and the selected time unit. The transition parameters take positive values, even when estimated under the non-negativity constraint.

Table 2 provides the confidence intervals (CI) for selected transmission parameters and transition probabilities. They have been computed by bootstrap in order to accommodate the finite sample properties of estimators, especially those with small positive values, whose finite sample distributions are asymmetric. For that reason, some confidence intervals are not centered at the estimated values. Yet, the focus is on the transmission parameters, regardless of their small values. The epidemiological models are nonlinear dynamic models with chaotic features, in the sense that small changes in some parameters can have a substantial impact in the long run. Note that the traditional representation of the confidence intervals (CI) can be misleading, especially for the parameters that sum up to one.

Table 2. Confidence Intervals

parameter	CI	parameter	CI
b_1	[0.0031, 0.0052]	p_{23}	[0.0099, 0.0560]
b_2	[0.252e-05, 4.032e-05]	p_{24}	[0.0273, 0.0942]
c_1	[4.497e-05, 17.203e-05]	p_{25}	[0.00098, 0.00356]
c_2	[0.00023, 0.00047]	p_{34}	[0.068, 0.1057]
		p_{35}	[0.0092, 0.0214]

The evolutions of estimated counts of IU, i.e. infected and undetected and of R, i.e. recovered are shown in Figure 4 (solid line). The estimates are compared with the available counts of ("confirmed") detected individuals and of recovered after being hospitalized (R—H).

[Insert Figure 4: Estimated and Observed Counts]

The estimated counts exceed those reported by the media in April 2020. In particular, the observed and estimated counts on April 06, 2020, which is the last day of sample are as follows: The final observed count of ("confirmed") detected is equal to 78167 and is 1.2 times smaller that the estimated final count of infected and undetected (IU) equal to 94461. The observed final count of Recovered (after being hospitalized) equal to 17250 is 6.24 times smaller than the estimated final count of Recovered equal to 107640.

Let us now present a scenario of a projected evolution, based on the estimated coefficients values and probabilities. These projections were performed on April 06, without taking into account future social distancing measures, increase of PCR tests, mandatory personal protective equipment (PPE), or the retrospective updates of databases. Figure 5 below shows the projected evolution of the marginal probabilities of IU, ID, R and D over the period of 25 years. This long horizon gives insights into the long run properties of the estimated dynamic model. It corresponds to the duration of the measles epidemic in London, prior to the vaccine, with infections documented over the period 1948 to 1964.

[Insert Figure 5: Projected Evolution of Marginal Probabilities]

Figure 5 displays peaks in marginal probabilities of states 2 and 3 that occur after about 98 days. At the peak, the projected count of infected and undetected (IU) individuals is over 300,000. In addition, we observe that the estimated model reveals no collective immunity. After 25 years, 35 % of the population-at-risk from March 16 die (not necessarily from Covid) and about 65 % are immunized. The existence of collective immunity depends on the selected model. In the standard SIR model, the collective immunity exists if the reproductive number R_0 is larger than 1, and it does not, otherwise. The specification outlined in Section 5.1 differs from the standard SIR in terms of the expressions of transmission functions $\pi_{12,t}, \pi_{13,t}$. They are equal to $\exp(a_1)$, $\exp(a_2)$, respectively, if $p_2(t-1) = p_3(t-1) = 0$, whereas in the standard SIR, they are equal to 0. The estimated non-zero values of $\exp(\hat{a}_1)$, $\exp(\hat{a}_2)$ reflect the transmission due to individual travelling between countries and regions. The projected results need to be interpreted with caution, due to the uncertainty on parameter estimates [see, Table 2].

Another pessimistic outcome is that without any social distancing measures, medical treatment for Covid-19, or a vaccine, it takes about 25 years for the marginal probabilities of IU and ID to decline to 0.

Figure 6 below shows the projected daily new counts of ID, as approximated by the net balance of hospitalizations over an initial period of 60 days, which can be used for the assessment of the capacity of the health sector.

[Insert Figure 6: Projected Evolution of Net Balance of Hospitalization]

The dashed line shows the projected daily net changes in hospitalization, computed as $\Delta p_3(t) * pop$, over 60 days following the end of sample on April 6. The dashed line depicts the true net changes in hospitalization observed ex-post. On April 15 (i.e. after 10 days) the net changes in hospitalization become negative (-513) and remain negative with high variation between -792 on 06/05 and 0 on 04/26. Nevertheless, on April 15, there are 2415 new hospitalizations and 275 new admissions to the ICU. From April 15 on, the number of patients released from the hospital exceeds the number of new admissions, resulting in negative net changes. The dotted lines represent the CI of the projection.

We observe that the projection detects the flattening of the curve of infections, although it overestimates the timing of the peak, i.e. the timing of the first value 0 on April 15, known ex-post. The predicted curve lies above the realized curve, revealing a prediction of the number of beds required for Covid-19 hospitalizations [16] However, the projection performed on April 6 has not been updated at any future date, as it is done in practice. The prediction can be updated daily, without re-estimating the model. In particular, the Kalman filter algorithm applied to the pseudo state space representation (see, Section 3) accommodates easily daily prediction updating.

6 Concluding Remarks

This paper is intended to provide a solution for incomplete counts of infected and undetected individuals and of recovered individuals. These unknown quantities can be estimated jointly with the parameters of a compartmental epidemiological model. This approach is illustrated in an estimation involving French count data on Covid-19 infections [see also Brown et al. (2020) for an application to North Carolina]. Our methodology

[16] The estimation performed on April 06 could not take into consideration the retrospectively updated total death counts. This could explain, at least to some extent, the observed bias. According to the updated sources, the evolution of deaths was more explosive at the beginning, i.e. close to March 16, and its inflection changed earlier too.

required daily data on the total counts of deaths, comprising the deaths due to Covid-19. These data are available in France and other European countries [see, the website Euromomo], but may be publicly unavailable in other countries, such as Canada. The results derived for one country (France) cannot be extrapolated directly to another country or state, because of differences in age structure and comorbidity.

More specifically, our results cannot be directly compared with other studies of undocumented infections in the US [see, Hortacsu, Liu, Schwieg (2020)] and China [Li et al. (2020)]. The comparisons are difficult, as each study employs different models, aggregate data and estimation methods. For example, Li et. al. (2020) use a (multicities) four state model with only 6 parameters, including 2 transmission parameters. They do not include the states D of Death and R of Recovered and they do not use the observations on the total number of deaths. Their estimation method is also different. More specifically, Li et al. (2020) use Bayesian methods, which are sensitive to the selected priors (Section 1 of the on-line "Supplementary Material"). As another example, Hortacsu et al. (2020), (Section 4), use conterfactual analysis, with fixed values of relevant parameters, such as the rate of asymptomatic, which is set equal to 0.6 and 0.1.

We consider a discrete time model, although the epidemiological literature relies mostly on the continuous time mechanistic model. The discrete time model provides consistent parameter estimates of the pseudo state-space representation and better accommodates daily data. This is because the trajectory of a Euler discretized continuous time model, even with a very short timestep, can be significantly different from the continuous time trajectory. Moreover, the conditions of collective immunity inferred from the discrete and continuous time models can differ [see, Boalto et al. (2018) and Allen (1994)]. This difficulty, due to the sensitivity of nonlinear dynamics with respect to the size of timestep, is out of the scope of this paper.

Various extensions of the model examined in this paper can be considered:

i) As mentioned earlier, the model is a special case of a nonlinear pseudo state space model, with states $p(t)$, deterministic state equations (3.1), and measurement equations: $\hat{A}_t = Ap(t) + Au(t)$, where $u(t)$ denotes the difference between the observed frequencies $f(t)$ and $p(t)$. Additional state space variables could also be introduced to account for individual compliance with self-isolation measures and their dynamic [see e.g. Alvarez et al.(2020), Chudik et al.(2020),Ferguson et al.(2020), Tang et al. (2020)].

ii) The individual efforts (moral hazard phenomenon) have impact on the transmission parameters. These can be captured by introducing transmission parameters with stochastic heterogeneity over time. In particular, some specific heterogeneity dynamics would allow for reproducing the stop and go phenomenon [see e.g. Ferguson et al.(2020), Figure 4]. More generally, the model can be extended by introducing time dependent or stochastic time dependent transmission parameters [see e.g. Dureau (2013), Boato et al (2018), Gourieroux, Lu (2020) for extensions of the SIR model]. It may be important to account of the mover-stayer phenomenon, as over time, the remaining Susceptibles are those who are more resistant to the infection.

iii) Other specifications of the propagation functions π_t can also be considered and compared [see Wu et al. (2020)]. The treatment of missing data can likely be improved by introducing additional explanatory variables that are expected to impact the virus trnsmission. This approach is followed in Hortacsu et al. (2020) who use hospitalization data from various regions and interregional transportation data to forecast infection rates.

REFERENCES

Allen, L. (1994): Some Discrete-Time SI, SIR and SIS Epidemic Models, Mathematical Biosciences, 124, 83-105.

Alvarez, F., Argente, D. and F. Lippi (2020) :A Simple Planning Problem for COVID-19 Lockdown,DP University of Chicago.

Boatto, S., Bonnet, C., Cazelles, B. and F. Mazenc (2018): SIR Model with Time Dependent Infectivity Parameter: Approximating the Epidemic Attractor and the Importance of the Initial Phase, HAL-01677886.

Brauer, F. and C. Castillo-Chavez (2001): Mathematical Models in Population Biology and Epidemiology,Springer,New York.

Breto, C., He, D., Ionides, E. and A. King (2009): Time Series Analysis via Mechanistic Models, Annals of Applied Statistics, 3, 319-348.

Brown, G., Ghysels, E. and L. Yi (2020): Estimating Undetected COVID-19 Infections. The Case of North Carolina, DP, University of North Carolina.

Chudik, A., Pesaran, H. and A. Rebucci (2020): Voluntary and Mandatory Social Distancy: Evidence on COVID 19 Exposure Rates from Chinese and Selected Countries,NBER 27034.

Cressie, N. and T. Read (1984): Multinomial Goodness-of-Fit Tests, JRSS,B,46,440-464.

Dureau, J., Kalegeropoulos, K. and M. Buguelin (2013): Capturing the Time Varying Drivers of an Epidemic Using Stochastic Dynamical Systems, Biostatistics, 14, 541-555.

Einicke, G. and L., White (1999): Robust Extended Kalman Filtering,IEEE Trans. Signal. Process.,47,2596-2599.

Ferguson N. et al. (2020): Estimating the Number of Infections and the Impact of Non-Pharmaceutical Interventions on Covid-19 in 11 European Countries,Imperial College, London.

Garet, O. and I. Marchand (2006): Competition Between Growths Governed by Bernoulli Percolation,Polymath,12,695-734.

Godambe, V. and M. Thompson (1974): Estimating Equations in the Presence of a Nuisance Parameter,Annals of Statistics,3,568-571.

Good, I. (1949): The Number of Individuals in a Cascade Process, Proc. Camb. Phil. Soc.,45,360-363.

Gourieroux, C. and J. Jasiak (2020): Analysis of Virus Transmission: A Transition Model Representaton of Stochastic Epidemiological Models, ARXiv 2006.10265

Gourieroux, C. and Y. Lu (2020): SIR Model with Stochastic Transmission, CREST DP.

Hammersley,J. (1957): Percolation Processes: Lower Bounds for the Critical Probability, Annals Math. Stat.,28,790-795.

Hardin, J. and J. Hilbe (2003): Generalized Estimating Equations, Chapman & Hall

Hethcote, H. (2000): The Mathematics of Infectious Diseases,SIAM Review, 42, 599-653.

Hortacsu, A., Liu, J. and T. Schwieg (2020): Estimating the Fraction of Unreported Infections in Epidemics with a Known Epicenter: An Application to COVID-19, University of Chicago DP.

INSEE (2020): Nombre de deces quotidiens par departement, April 10.

Julien, S. and J. Uhlman (1997): A New Extension of the Kalman Filter to Nonlinear Systems,11th Int.Symp. on Aerospace/Defence,Sensing,Simulation and Controls.

Kermack, W. and A. McKendrick (1927): A Contribution to the Mathematical Theory of Epidemics,Proceedings of the Royal Statistical Society, A, 115,700-721.

Krener, A. (2003): The Convergence of the EKF, in: Directions in Mathematical Systems:Theory and Optimization,173-182, Springer

Li, R., Pei, S., Chen, B., Song, Y., Zhang, J., Yang, W. and J. Shaman (2020): Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus(SARs-CoV2), Science, 368, 489-493.

Manski,C. and F. Molinari (2020): Estimating the COVID-19 Infection Rate:Anatomy of an Inference Problem,Northwestern Univ. DP.

McFadden, D. (1984): Econometric Analysis of Qualitative Response Models,in Handbook of Econometrics,Vol 2,1395-1457, Elsevier.

McRae, E. (1977): Estimation of Time Varying Markov Processes with Aggregate Data,Econometrica, 45,183-198.

McKendrick, A. (1926): Applications of Mathematics to Medical Problems, Proceedings of the Edinburgh Mathematical Society, 14, 9-130.

Miller,J. and G. Judge (2015): Information Recovery in a Dynamic Statistical Markov Model, Econometrics, Vol 3/2, 187-198.

Nishiura, H. et al. (2020): Estimation of the Asymptomatic Ratio of Novel Coronavirus Infection (COVID-19),Int.J. Infect.Dis.

Sante Publique France (2020): Donnees Hospitalieres Relatives a l'Epidemie Covid-19.

Song, Y. and J. Grizzle (1995): The Extended Kalman Filter as a Local Asymptotic Observer,Estimation and Control,5,59-78.

Tang, B. et al.(2020) : Estimation of the Transmission Risk of the 2019-nCoV and its Implication for Public Health Interventions,Journal of Clinical Medicine,9.

Verity, R. et al. (2019): Estimates of the Severity of Coronavirus Disease 2019;A Model Based Analysis,Lancet Infect.Dis.

Vynnicky, E. and R. White (eds)(2010): An Introduction to Infectious Disease Modelling, Oxford Univ Press.

Wan, E. and R. Van der Merwe (2000) : The Unscented Kalman Filter for Nonlinear Estimation,in Adaptive Systems for Signal Processing,Communication and Control Symposium,IEEE 2000,153-158

Wu, K., Darcet, D., Wang, Q. and D. Sornette (2020): Generalized Logistic Growth Modelling of the Covid 19 Outbreak in 29 Provinces in China and the Rest of the World,DP Univ.Zurich.

Appendix 1

Expression of the Autocovariance Operator

Instead of characterizing the individual histories by the qualitative sequences Y_{it}, a sequence of J-dimensional vectors Z_{it} can alternatively be considered, where component j is the 0-1 indicator of $Y_{it} = j$. Then we have:

$$E(Z_t|Z_{t-1}) = P(t-1)Z_{t-1},$$

where $P(t-1)$ denotes the transition matrix from date t-1 to date t. By the iterated expectations theorem, we get:

$$E(Z_t|Z_{t-h}) = \Pi(t-1;h)Z_{t-h},$$

where $\Pi(t-1;h) = P(t-1)...P(t-h)$.

Let us now consider the covariance:

$$\begin{aligned} \Omega_{t,t-h} &= Cov(Z_t, Z_{t-h}) = E(Z_t Z'_{t-h}) - E(Z_t)E(Z_{t-h})' \\ &= E(\Pi(t-1;h)Z_{t-h}Z'_{t-h}) - p(t)p(t-h)' \end{aligned}$$

(by the iterated expectation and using $E(Z_t) = p(t)$)

$$= \Pi(t-1;h)E[diag(Z_{t-h})] - p(t)p(t-h)'$$

(by taking into account the 0-1 components of Z)

$$= \Pi(t-1;h)diag[p(t-h)] - p(t)p(t-h)'.$$

This is the expression of the autocovariance as a function of the $p(t)$'s and model parameters. Under Assumptions A.1. and after a normalization by $1/N$ we obtain the autocovariance of the frequencies $f(t), t = 1, ..., T$ and of the measurement equation error $u(t), t = 1, ..., T$ in the pseudo state space representation.

Appendix 2
Asymptotic Expansions

The asymptotic expansions are easily derived, given that the optimization in Proposition 2 is deterministic. Therefore, estimators $\hat{p}(1), \hat{p}(2), ..., \hat{p}(T), \hat{\theta}$ are deterministic functions of observations $\hat{A}_t = Af(t), t = 1, .., T$. If the transition matrix is twice continuously differentiable with respect to $p(t-1)$ and θ in a neighbourhood of the true values, these deterministic functions are continuously differentiable. Then, by using the asymptotic normality of $f(t)$'s (Proposition 1), we can apply the delta method to deduce the $1/\sqrt{N}$ rate of convergence of the estimators and their asymptotic variance-covariance matrix from the one of the $f(t)$'s (see Appendix 1).

When the number of observation dates and of missing counts is too large, the use of the delta method can be numerically cumbersome. It can be replaced by a bootstrap method (for which the regularity conditions of validity are satisfied in our framework), or by the approximated standard errors provided by an EKF, or UKF algorithm, after adjusting for the misspecification of the autocovariances of the measurement equation errors $u(t)$.

Appendix 3
Nongeneric Cases in Proposition 3

This Appendix derives the equations used in the proof of Proposition 3. It provides the closed form expressions of functions $a(P), b(P), c(P)$, and outlines conditions 1 to 4 for the validity of Proposition 3.

i) Let us first solve the second equation of system (4.3). We get:

$$(p_{23} - p_{13})p_2(t-1) = p_3(t) + (p_{13} - p_{33})p_3(t-1) - p_{13},$$

or,

$$p_2(t-1) = [p_3(t) + (p_{13} - p_{33})p_3(t-1) - p_{13}]/(p_{23} - p_{13}),$$

if the following condition is satisfied:

condition 1: p_{23} is different of p_{13}.

ii) Next, let us consider the first equation of system (4.3):

$$p_2(t) = p_{12} + (p_{22} - p_{12})p_2(t-1) + (p_{32} - p_{12})p_3(t-1)$$

and substitute into this equation the expression of $p_2(t)$ derived in part i). We get:

$p_3(t+1) + (p_{13} - p_{33})p_3(t) - p_{13} = p_{12}(p_{23} - p_{13}) + (p_{22} - p_{12})[p_3(t) + (p_{13} - p_{33})p_3(t-1) - p_{13}] + (p_{23} - p_{13})(p_{32} - p_{12})p_3(t-1)$.

It follows that:

$$\begin{aligned} a(P) &= p_{12}(p_{23} - p_{13}) + p_{13}(1 - p_{22} + p_{12}), \\ b(P) &= p_{22} - p_{12} + p_{33} - p_{13}, \\ c(P) &= (p_{22} - p_{12})(p_{13} - p_{33}) + (p_{23} - p_{13})(p_{32} - p_{12}). \end{aligned}$$

To get a recursive equation of order 2, we need the second condition:

condition 2: $c(P) \neq 0$

To identify functions a, b, c from the observed $p_3(t)$, we need:

condition 3: The matrix $3 \times (T-2)$ with columns $(1, ..., 1)'$, $(p_3(T-1), p_3(T-2), .., p_3(2))'$ and $(p_3(T-2), p_3(T-3), ..., p_3(1))'$ is of full column rank.

This implies, in particular, the order condition: $T \geq 5$ in Proposition 3.

The following condition 4 is needed for computing the exact under-identification order of P from functions a, b, c.

condition 4: By taking into account the unit mass restrictions on the rows of P,the Jacobian of (a, b, c) has rank 3.

Note that condition 4 implies condition 2.

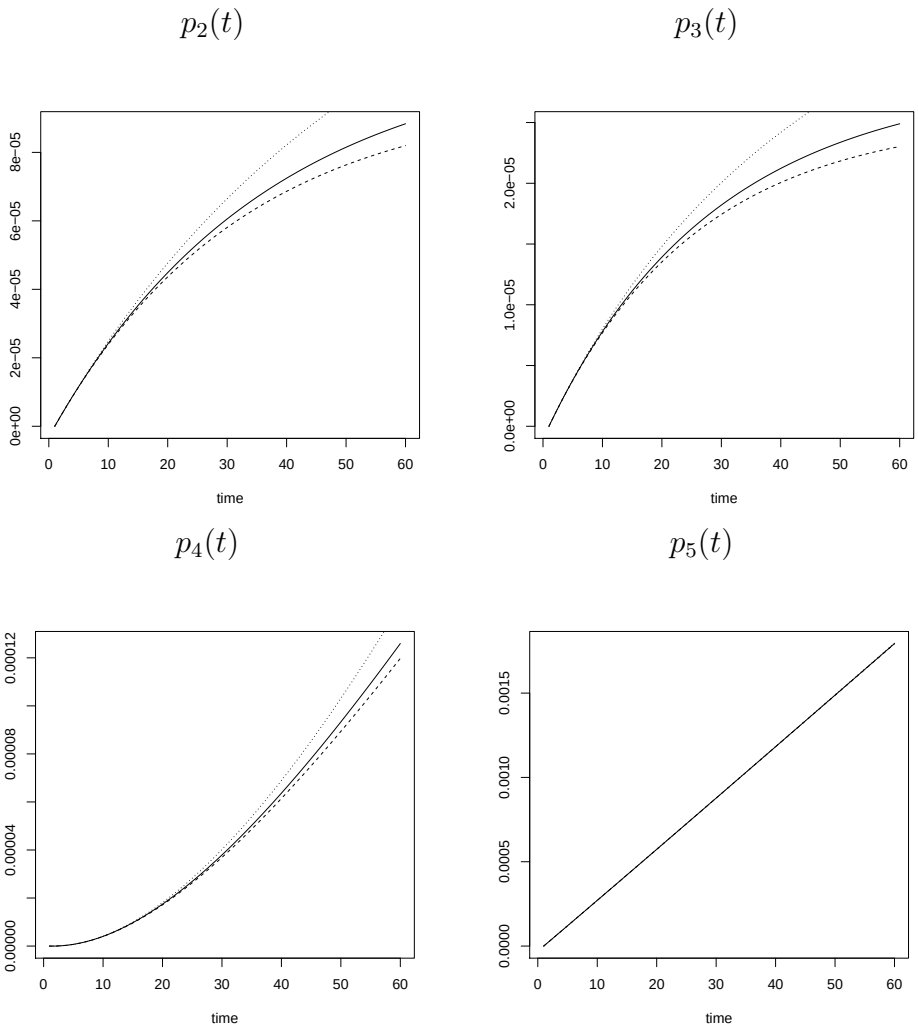


Figure 1: Evolution of Marginal Probabilities
Solid line-baseline, dotted line - doubled transmission parameters, dashed line - halved transmission parameters

$\Delta p_3(t) * pop$ $\Delta p_5(t) * pop$

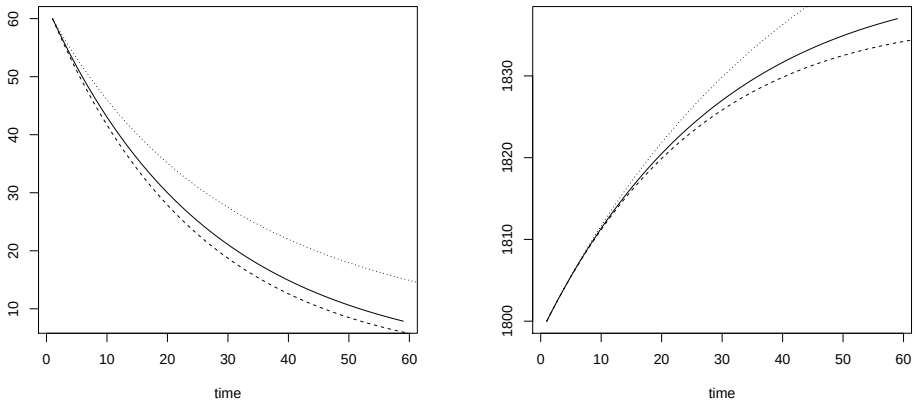


Figure 2: Evolution of New Counts
Solid line-baseline, dotted line - doubled transmission parameters, dashed line - halved transmission parameters

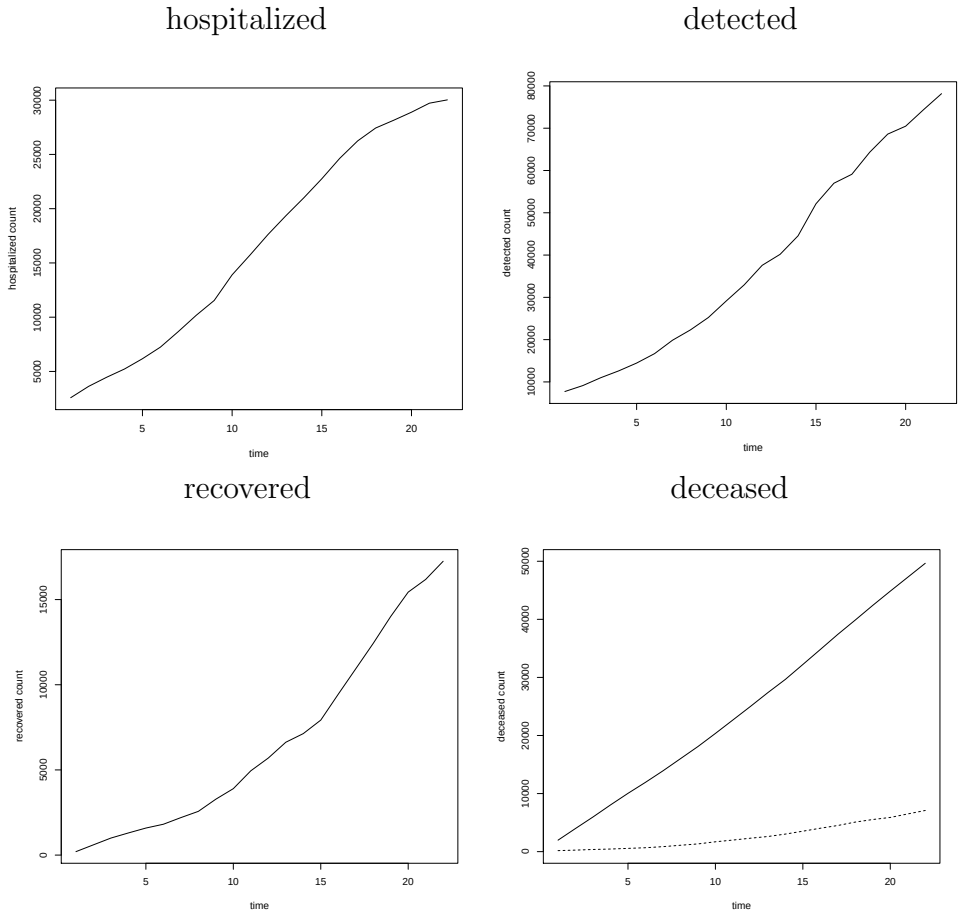
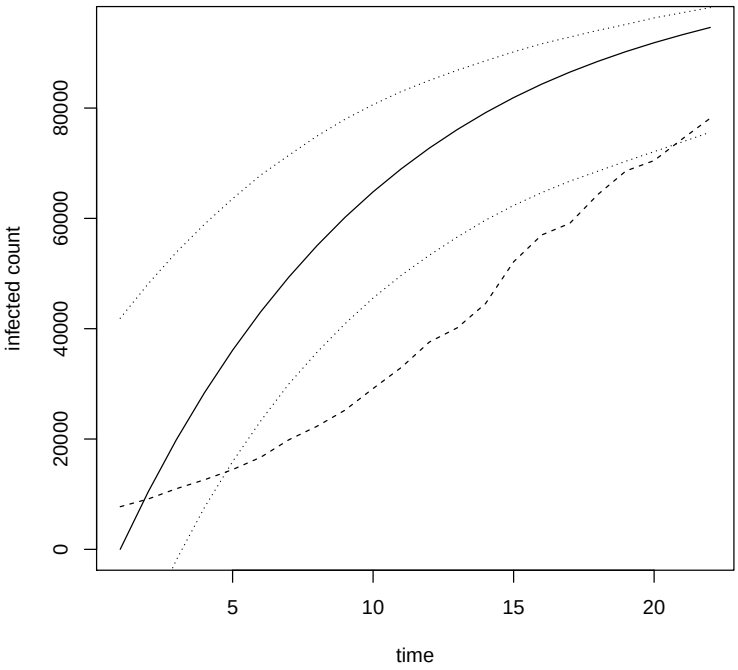


Figure 3: Evolution of Observed Counts, 03/16 to 04/06, France
The figure shows the evolution of observed daily counts. In the panel of deceased (bottom, right), the solid line shows the total deceased in France and the dashed line the (reported) deceased due to Covid-19

estimated IU and observed ID



estimated and observed Recovered

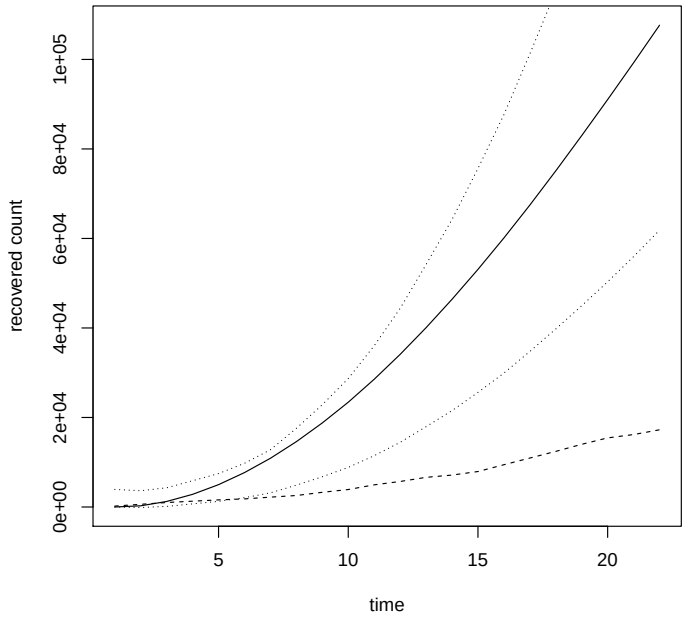


Figure 4: Estimated and Observed Counts
The estimated counts - solid line, observed counts - dashed line. The figure compares the estimated counts of Infected and Undetected with the observed Infected Detected (top panel), and Recovered estimated and reported as hospitalizations (bottom panel). The dotted lines depict the confidence intervals.

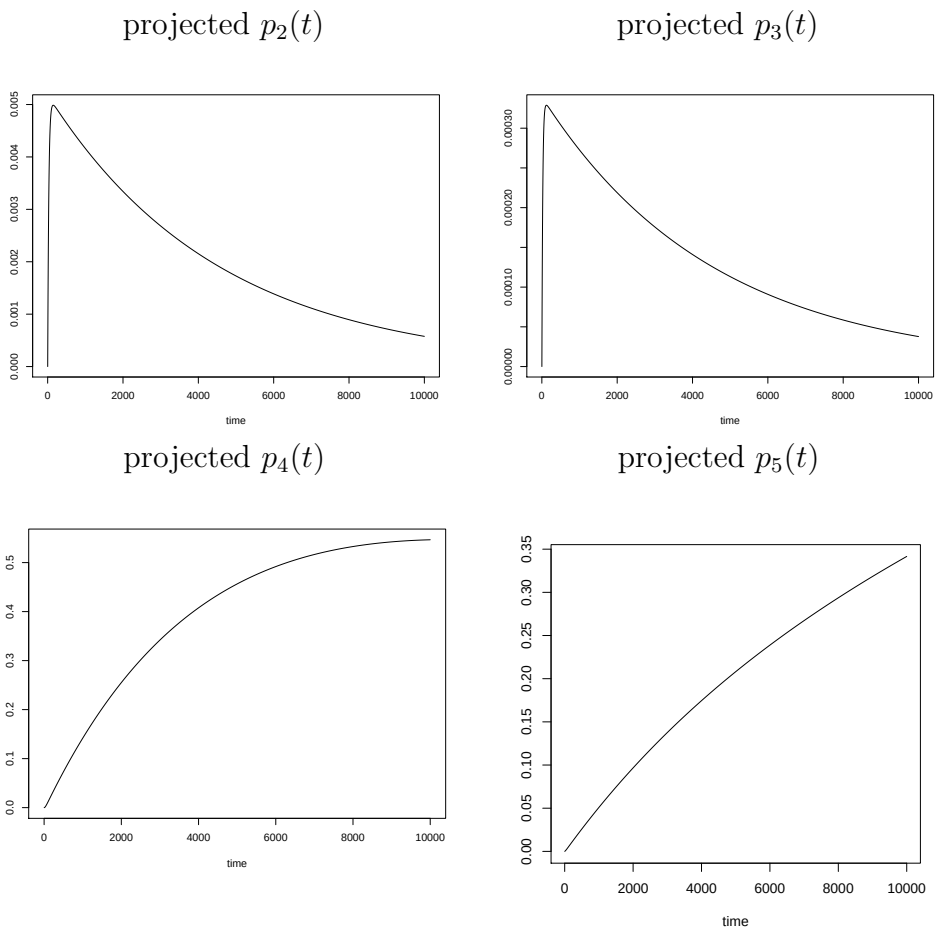


Figure 5: Projected Evolution of Marginal Probabilities.
The figure displays projected daily marginal probabilities of all states over 25 years.

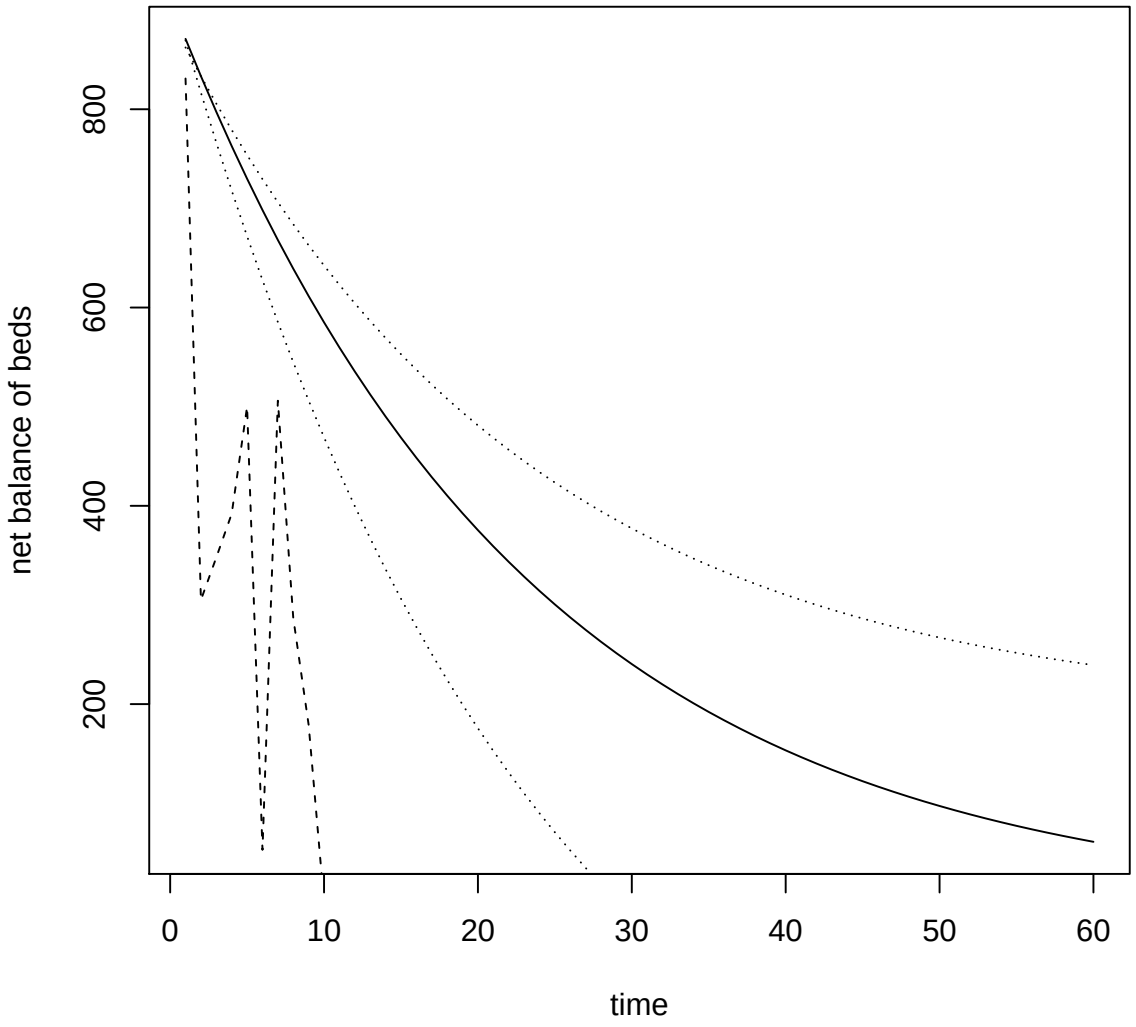


Figure 6: Projected Evolution of Net Balance of Hospitalization

The dashed line shows the projected daily net changes in hospitalization $\Delta p_3(t) * pop$ over 60 days following the end of sample on April 6. The dashed line depicts the true net changes in hospitalization observed ex-post. On April 15 (i.e. after 10 days) the net changes in hospitalization become negative (-513) and remain negative with high variation between -792 on 06/05 and 0 on 04/26. The dotted lines represent the CI of the projection.