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

GOURIEROUX, JASIAK

Forhcoming Journal of Econometrics

1. INTRODUCTION

The aim of the paper :

address the problems of partial observability encountered in epidemiological research on COVID-19.

- First, only cross-sectionally aggregated data are easily available, not the individual medical histories of the individuals.

- Second, some individuals are infected and asymptomatic. They are undetected in the early phase of the epidemics and the number of recovered, immunized people is not available.

We develop a model based approach to solve this issue.

- We extend the standard SIRD (Susceptible, Infected, Recovered, Diceased) model, introduced in Kermack, Mc Kendrick (1927), PRSS, by disentangling among the infected, the detected and the undetected.

and look for additional information in order to identify the number of infected undetected : the total daily number of deceased people (not only from coronavirus).

An analysis done rather early.

• on French data (See Brown et al. (2020) for North Carolina).

• from March, 16, to April 4, 2020, (22 days)

• in a stable environment of social distancing measures (total lock-down on the week end of March 16, just after the first round of municipal elections)

• based on data available on April 06, the first publication of the daily total number of deceased.

• used with other aggregate counts :

detected, hospitalized, recovered (detected), deceased (total and from COVID) (more or less reliable).

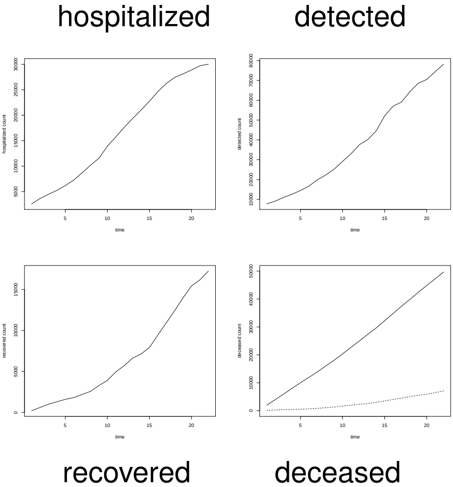


Figure 3 : Evolution of Observed Counts, 03/16 to 04/06, France

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

2. THE MODEL

A model with 5 states :

1 = S (Susceptible), 2 = IU (Infected, Undetected), 3 = ID (Infected, Detected,) 4= R (Recovered), 5 = D (Deceased)

$p_{1t}, p_{2t}, \ldots, p_{5t}$ the marginal probabilities of the states at date t.

The dynamics is defined by the structure of the transition matrix.

The model is written in discrete time.

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

$$\text{where : } \begin{array}{rcl} \pi_{11,t} & \simeq & 1, \pi_{12t} \simeq \exp[a_1 + b_1 p_2(t-1) + c_1 p_3(t-1)] \\ \pi_{13t} & \simeq & \exp[a_2 + b_2 p_2(t-1) + c_2 p_3(t-1)] \end{array}$$

row 2 : 0 ; p_{22} ; p_{23} ; p_{24} ; p_{25}
row 3 : 0 ; 0 ; p_{33} ; p_{34} ; p_{35}
row 4 : 0 ; 0 ;0 ; p_{44} ; p_{45}
row 5 : 0 ; 0 ;0 ;0 ; 1

A specific structure

• multinomial logit for the transmission effects
$a_1, b_1, c_1, a_2, b_2, c_2$: transmission parameters

• triangular form, in particular recovered are permanently immunized (or at least for a long period)

• constant recovery rate, with duration of infection period (hospitalization) with exponential distribution

D as an absorbing state

A model with 13 parameters :
6 for the transmission, 7 time independent transition probabilities

This differs from the basic SIR in two respects :

• the introduction of two compartments IU, ID instead of a single one I.

• the form of the transmission function, that in a standard version would be :

$$\begin{array}{lcl} \pi^*_{12t} & \simeq & b_1 p_2(t-1) + c_1 p_3(t-1) \\ \pi^*_{13t} & \simeq & b_2 p_1(t-1) + c_2 p_3(t-1) \end{array}$$

Standard SIR : No infected people in the country, no infection [closed economy implying herd immunity]

Modified SIR :
if $p_2(t-1) = p_3(t-1) = 0, \pi_{12t} \simeq \exp a_1, \pi_{13t} \simeq \exp a_3$
[open economy implying no herd immunity]

Simulation for given values

$p_{15} = p_{45}$ fixed at the long term (daily) mortality rate
p_{25} the average mortality rate for people detected in hospital
p_{35} between both to account for less fragile asymptomatic people.
The transmission parameters have been set to provide about 60 new daily detected infection at the beginning (for a French population of 60 millions of inhabitants)
1 500 new daily detected infections 30 days after the beginning.

Note : no effect of an increase of tests during the epidemics [due to shortage of test components for PCR and no validated serological tests]

$$p(0) = (1, 0, 0, 0, 0,).$$

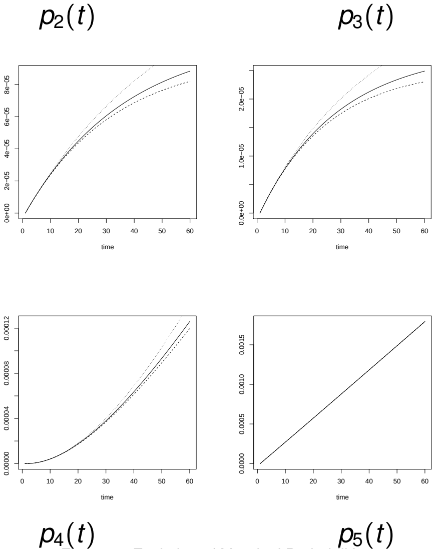


Figure 1 : Evolution of Marginal Probabilities
Solid line-baseline, dotted line - doubled transmission parameters, dashed line - halved transmission parameters

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

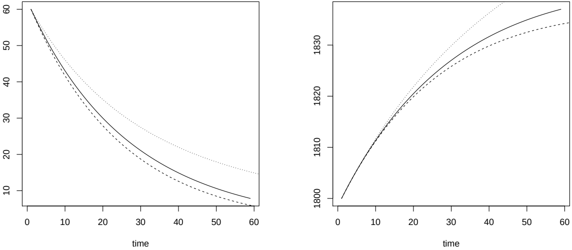


Figure 2 : Evolution of New Counts
Solid line-baseline, dotted line - doubled transmission parameters, dashed line - halved transmission parameters

3. ESTIMATION

Latent model
A large panel of individual histories

$$Y_{i,t}, i = 1, \ldots, n, t = 1, \ldots, T.$$

Assumption A.1 :
i) At t fixed, the variables $Y_{i,t}, i = 1, \ldots, n$ have the same marginal distribution $p(t)$ with components :

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

ii) The processes $(Y_{i,t}, t = 1, \ldots, T), i = 1, \ldots, n$ are independent (heterogenous) Markov processes with transition $P[p(t-1), \theta]$, where θ are the parameters.

Under Assumption A.1, the cross-sectional frequencies $f(t)$ are consistent of $p(t)$, asymptotically normal, with a structure of variance-covariance given in the paper.

Assumption A.2 :
The observations are $\hat{A}_t = Af_t$, where A is an aggregation matrix with zeros and ones.

• In the literature : $A = Id$, all $f_t's$ are observed Mc Rae (1977), Econometrica, Miller, Judge (2015), Econometrics.

• In our application : $f_3(t), f_5(t)$ only are observed

• In a $(SI)^2$ model with 2 countries

S1 : susceptible in 1, I1 : Infected in 1
S2 : susceptible in 2, I2 : Infected in 2

We may observe aggregates corresponding to S1US2 and I1UI2.

The estimation is based on estimating equations [Godambe, Thompson (1974), AMS]

$$p(t) = P[p(t-1);\theta]' p(t-1), t = 1, \dots, T.$$

More precisely :

$$\begin{aligned} [\hat{p}(1), \dots, \hat{p}(T), \hat{\theta}] &= \arg\min_{p(t),\theta} \|p(t) - P[p(t-1);\theta]' p(t-1)\|^2, \\ \text{s.t. } Ap(t) &= Af(t) = \hat{A}_t, t = 1, \dots, T, \end{aligned}$$

where $\|.\|$ is an Euclidean norm.

The estimators are consistent, asymptotically normal, for $n \rightarrow$, T fixed

A Kalman filter can be applied to a (pseudo) state space model to find numerically the solution :

$$\begin{aligned}
\hat{A}_t &= Af(t), \\
f(t) &= p(t) + u(t), \\
p(t) &= P[p(t-1); \theta)p(t-1),
\end{aligned}$$

with three layers, including two layers of state variables for $p(t)$ and $f(t)$, respectively, and some deterministic equations.

Remark : Different forms of $V[u((1), \dots, u(T)]$ can be used, providing always consistent estimators.

Identification

Partial observability may create identification issues for θ. This is not the case in our framework, due to :

- the triangular form of P.
- the nonlinear dynamics in the transmission function.

Intuitively, the identification arises, since

- we observe a nonstationary evolution
- the unobserved IU imply a mixture of "logistic" evolutions at different speeds.

Estimated parameters

Focus on the transmission parameters and recovery parameters

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]

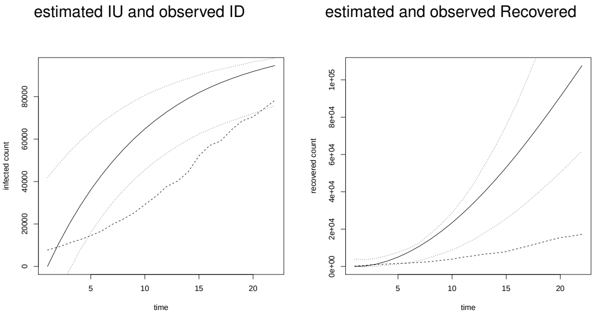
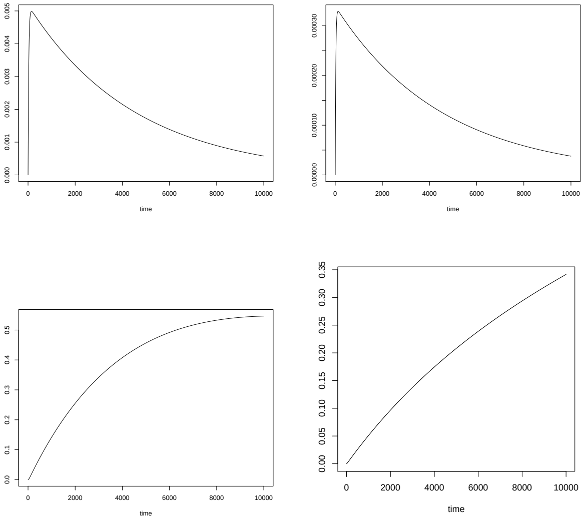


Figure 4 : Estimated and Observed Counts
The estimated counts - solid line, observed counts - dashed line. The figure compares the estimated counts of Infected and Undetected with the observed Infected Detected (top panel), and Recovered estimated and reported as hospitalizations (bottom panel). The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



projected $p_2(t)$

projected $p_3(t)$

projected $p_4(t)$

projected $p_5(t)$

Figure 5 : Projected Evolution of Marginal Probabilities.

Ex-post it is possible to compare the predictions of the model with the ex-post realizations.

- the database on total deceased has been updated even for the initial period
- strong week end effects (but not at the beginning)

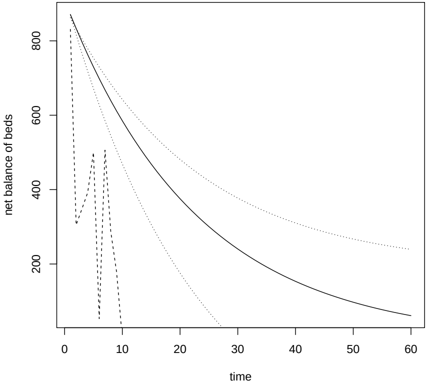


Figure 6 : Projected Evolution of Net Balance of Hospitalization
The dashed line shows the projected daily net changes in hospitalization $\Delta p_3(t) * pop$ over 60 days following the end of sample on April 6. The dashed line depicts the true net changes in hospitalization observed ex-post. On April 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.

4. CONCLUDING REMARKS

We have introduced a model based methodology to solve the problems of partial observability, in particular to reconstitute the number of infected, undetected individuals.

message on data :
important data are often missing

total daily number of deceased (important for identification)(see also Li et al. (2020) with a multicities model and travel data)

the aggregate flows, not only the cross-sectional aggregates (for improving the accuracy) [Breto et al. (2009)].

Other data are not reliable, due to unprecise definitions, i.e. the number of deceased due to COVID,

or too sentitive to some health (test) policies as the number of confirmed cases.

Message for test policies

There exist several types of tests :

- The PCR [Polymerase Chain Reaction] tests
- The serological tests

It is expected to diminish the number of infected undetected (spot and ex-post) by increasing the number of tests, but

- At the beginning of the epidemics, serological tests were not validated ;

- There are a significant rates of false negative and of false positive ;

- This is costly 80 € per PCR test

- The test assignment is endogenous (tracing) (not at random), implying a selectivity bias.

Message for epidemiological modelling

- The pseudo state space representation is valid is n is large, with also the $np_j(t)$ large enough. This is not the case at the very beginning of the epidemics, or if the analysis is performed on small regions, or subgroups.

The asymptotic normality is replaced by asymptotic Poisson, with consequences on consistency.

- The basic model has "time" independent transmission parameters : this does not account for evoluting health policies as well as for frailty effects (i.e. the mover-stayer phenomenon).

5. COVARIANCE OPERATOR

(supplementary material)

The frequencies $\hat{f}(t)$ are such that :

$$Cov[\hat{f}(t), \hat{f}(\tau)] = n\ Cov[Z(t), Z(\tau)],$$

where $Z(t)$ is the $J = 5$ dimensional vector, whose components are the indicators of $Y(t) = 1, 2, 3, 4, 5$.

We have : $E(Z_t | Z_{t-1}) = P(t-1)\ Z_{t-1}$.

By iterated expectation :

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

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

Therefore,

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

6. REFERENCES

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

Brown, Ghysels and Yi (2020) : "Estimating Undetected COVID-19 Infections : the Case of North Carolina", DP Univ. North Carolina.

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

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

Li, Pei, Chen, Song, Zhang, Yang and Shaman (2020) : "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus", Science, 368, 489-493.

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

Miller, and Judge (2015) : "Information Recovery in a Dynamic Statistical Markov Model", Econometrics, 312, 187-198.