Mechanistic-statistical SIR modelling for early estimation of the actual number of cases and mortality rate from COVID-19

Lionel Roques, Etienne Klein, Julien Papaïx and Samuel Soubeyrand
INRAE, BioSP, 84914, Avignon, France
Contact : lionel.roques@inrae.fr

Abstract

The first cases of COVID-19 in France were detected on January 24, 2020. The number of screening tests carried out and the methodology used to target the patients tested do not allow for a direct computation of the actual number of cases and the mortality rate. In this note, we develop a 'mechanistic-statistical' approach coupling a SIR ODE model describing the unobserved epidemiological dynamics, a probabilistic model describing the data acquisition process and a statistical inference method. The objective of this model is not to make forecasts but to estimate the actual number of people infected with COVID-19 during the observation window in France and to deduce the mortality rate associated with the epidemic.
Main results. The actual number of infected cases in France is probably much higher than the observations: we find here a factor $\times 15$ (95%-CI: $4 - 33$), which leads to a 5.2/1000 mortality rate (95%-CI: $1.5/1000 - 11.7/1000$) at the end of the observation period. We find a R_0 of 4.8, a high value which may be linked to the long viral shedding period of 20 days.

Introduction. The COVID-19 epidemic started in December 2019 in Hubei province, China. Since then, the disease has spread around the world reaching the pandemic stage, according to the WHO, on March 11. The first cases were detected in South Korea on January 19, 2020, and in France on January 24. The number of screening tests carried out varies widely from country to country ($36,747$ in France *vs* $268,212$ in South Korea up to March 15, 2020; Sources: Santé Publique France and Korean Center for Disease Control) and does not make possible to know with certainty the actual number of infected people in the population. Thus, even if the number of deaths linked to COVID-19 is known with more certainty, ignoring the total number of patients does not allow a direct calculation of the mortality rate. Using the data available in France and South Korea, our objectives are:

- to estimate the number of people infected with COVID-19 in France;
- to deduce from this number the associated mortality rate;
- to calculate the parameters of a SIR-type model representing the early phase of the outbreak in France;
- to compare the results obtained for France and South Korea.

In this aim, we carried out an analysis grounded on a mechanistic-statistical formalism. This formalism allows the analyst to couple a mechanistic model, here an ordinary differential equation model (ODE) of the SIR type, and uncertain, non-exhaustive and not necessarily commensurable data with the solutions of the ODE. This formalism, which we have popularized by applying it to biological invasions (Roques et al., 2011; Roques and Bonnefon, 2016; Abboud et al., 2019), combines (1) the mechanistic model, (2) a probabilistic model describing the data collection process conditional on the solution of the mechanistic model and (3) a statistical method for estimating the parameters of the mechanistic model.

Data. The COVID-19 screening data in France and South Korea used for preparing this note are those available from January 22, 2020 to March 17, 2020. These data describe the number of positive cases and deaths, day by day (source: Johns Hopkins University Center for Systems Science and Engineering, https://github.com/ CSSEGISandData/COVID-19). The number of tests carried out is only available from February 22 (Sources: Santé publique France and Korean Center for Disease Control). As some data (positive cases, deaths) are not fully reliable (example: 0 new cases detected in France on March 12, 2020), we smoothed the data with a moving average over 5 days.

Mechanistic model. SIR models are the most standard ODE-based epidemiological models. These models are compartmental: they divide the population into subclasses: the susceptibles (S), the infected (I) and the recovered (R, immune individuals in our case). The simplest SIR compartmental model does not take into account the demography of the S compartment:

$$\begin{cases} S' = -\dfrac{\alpha}{N}\, S\, I, \\ I' = \dfrac{\alpha}{N}\, S\, I - \beta\, I, \\ R' = \beta\, I, \end{cases} \tag{1}$$

with $N = S + I + R$ the total population, which is constant. The impact of the compartment D (number of dead people) on the dynamics of the SIR system is therefore neglected here. The compartment D satisfies:

$$D'(t) = \gamma(t)\, I. \tag{2}$$

We use this equation later to compute the death rate due to the disease. The initial conditions are $S(t_0) = N - 1$, $I(t_0) = 1$ and $R(t_0) = 0$, where N corresponds to the population size in France or South Korea ($67{\cdot}10^6$ and $52{\cdot}10^6$ people). The SIR model is started at $t = t_0$, which should approach the date of introduction of the virus in the considered country (this point is shortly discussed at the end of this note).

Note that $I'(t) = \beta\, I\, (R_0\, S/N - 1)$, with $R_0 = \alpha/\beta$ the basic reproduction rate (Murray, 2002). When $R_0 < 1$, we observe that $I' < 0$, and so the epidemic cannot spread in the population. When $R_0 > 1$, the infected compartment I increases as long as $R_0\, S > N = S + I + R$.

The system (1) can be analytically solved, using a change of time variable which requires a numerical integration. Here, we rather solve the ODE system thanks to a standard numerical algorithm, using Matlab® *ode23s* solver.

Observation model. We denote by $\hat{\delta}_t$ the number of cases tested positive on day t. We suppose that these increments follow independent binomial laws, conditionally on the number of tests, on $I(t)$ and on $S(t)$:

$$\hat{\delta}_t \sim Bi(n_t, p_t), \tag{3}$$

where n_t corresponds to the number of tests carried out on day t and p_t the probability of being tested positive in this sample. The tested population consists of a fraction of the infecteds and a fraction of the susceptibles: $n_t = \tau_1(t)\, I(t) + \tau_2(t)\, S(t)$. Thus,

$$p_t = \frac{\tau_1(t)\, I(t)}{\tau_1(t)\, I(t) + \tau_2(t)\, S(t)} = \frac{I(t)}{I(t) + \kappa_t\, S(t)},$$

with $\kappa := \tau_2(t)/\tau_1(t)$, the relative probability of undergoing a screening test for an individual of type S *vs* an individual of type I (probability of being tested conditionally on being S / probability of being tested conditionally on being I). We assume that the ratio κ does not depend on t at the beginning of the epidemic (i.e., over the period that we use to estimate the parameters of the model). The daily number of deaths caused by the virus is assumed to be known exactly.

In order to compute the maximum likelihood estimator (the MLE, i.e., the parameters that maximize $\mathcal{L}$), we use the BFGS constrained minimization method, applied to $-\ln(\mathcal{L})$, via the Matlab ® function *fmincon*. In order to find a global maximum of $\mathcal{L}$, we apply this method starting from random initial values for α, t_0, κ drawn uniformly in the following intervals:

$$\begin{cases} \alpha \in (0,1), \\ t_0 \in (1,30), \ \text{(January 1st - January 30th)} \\ \kappa \in (0,1). \end{cases} \tag{4}$$

For each country, the minimization algorithm is applied to 2000 random initial values of the parameters.

The posterior distribution of the parameters (α, t_0, κ) is computed with a Bayesian method, using uniform prior distributions in the intervals given by (4). This posterior distribution corresponds to the distribution of the parameters conditionally on the observations:

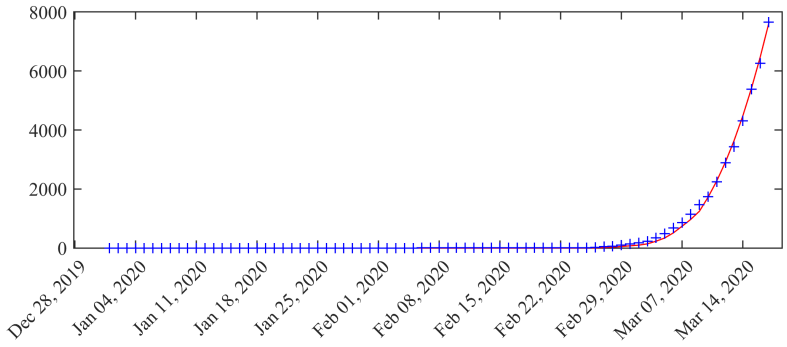
$$P(\alpha, t_0, \kappa | \{\hat{\delta}_t\}) = \frac{\mathcal{L}(\alpha, t_0, \kappa)\, \pi(\alpha, t_0, \kappa)}{C},$$

where $\pi(\alpha, t_0, \kappa)$ corresponds to the prior distribution of the parameters (therefore uniform) and C is a normalization constant independent of the parameters. The numerical computation of the posterior distribution (which is only carried out for French data) is performed with a Metropolis-Hastings (MCMC) algorithm, using 5 independent chains, each of which with 10^6 iterations, starting from random values close to the MLE.

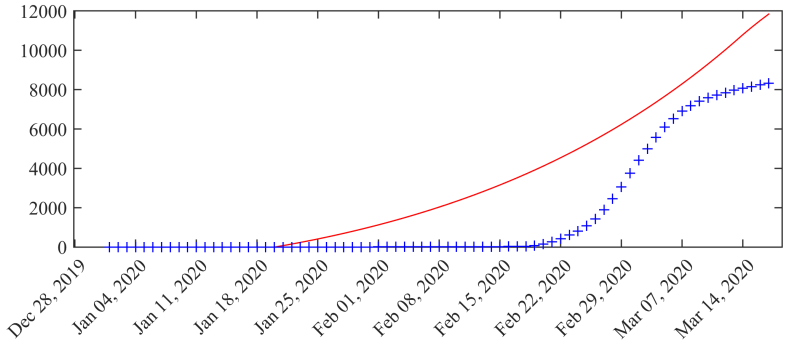
Unless otherwise mentioned, the data $\hat{\delta}_t$ used to compute the MLE and the posterior distribution are those corresponding to the period from February 29 to March 17.

Results. *Model fit.* We denote $(\alpha^*, t_0^*, \kappa^*)$ the maximum likelihood estimator (MLE), and $I^*(t)$, $S^*(t)$ the solutions of the system (1) associated with these values. In France, we get $(\alpha^*, t_0^*, \kappa^*) = (0.24, 26, 2 \cdot 10^{-4})$. The expectation of the observations associated with this MLE is $n_t\, p_t^*$ (expectation of a binomial) with

$$p_t^* = \frac{I^*(t)}{I^*(t) + \kappa^*\, S^*(t)}.$$



(a) France

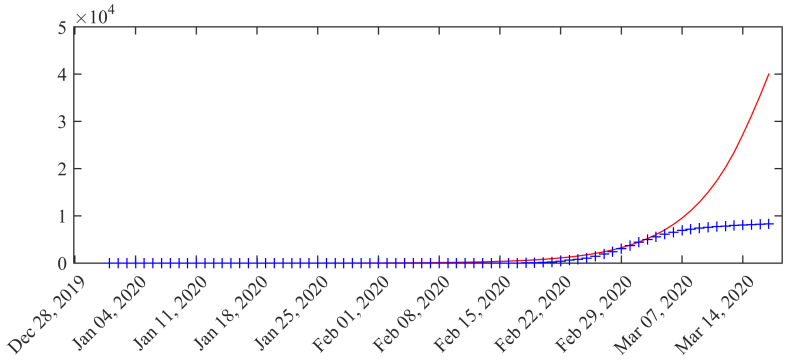


(b) South Korea

Figure 1: **Expected number of observed cases associated with the MLE** *vs* **number of cases actually detected (total cases).** The red curve corresponds to the expectation $n_t\, p_t^*$, the blue crosses to the data (cumulated values of $\hat{\delta}_t$). The computation of the MLE is based on data from February 29 to March 17.

Fig. 1 compares this expectation with observations. In France, we get a good match between $n_t\, p_t^*$ and the data. In South Korea, on the other hand, the gap between the data and the model prediction is significant: the SIR model, which leads to an exponential trajectory of I at the beginning of the epidemic, cannot properly render the dynamics. Using data obtained at an earlier stage in South Korea, the model fit is better for the initial outbreak phase (Fig. 2). The corresponding MLE is: $(\alpha^*, t_0^*, \kappa^*) = (0.13, 3, 3 \cdot 10^{-5})$.

Parameter distribution. The joint posterior distributions of the three pairs of parameters (α, κ), (t_0, α) and (t_0, κ) for France are presented in Appendix A (Fig. 6). Note that the distributions are very different from the uniform prior distribution. However, the distributions of t_0 and κ are quite dispersed. The joint distribution of (t_0, κ), presented in Appendix A shows a correlation between t_0 and κ. Based on this distribution, in order to reduce the uncertainty on κ, we assume that t_0 is between January 13 and 30.

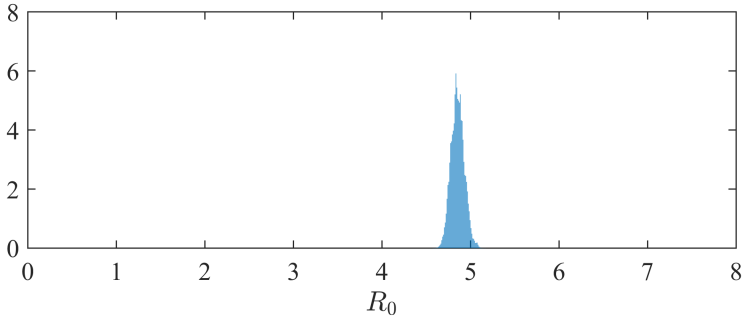


(a) South Korea

Figure 2: **Expected number of observed cases associated with the MLE** *vs* **number of cases actually detected (total cases) in South Korea.** The red curve corresponds to the expectation $n_t\, p_t^*$, the blue crosses to the data (cumulated values of $\hat{\delta}_t$). The computation of the MLE is based on data from February 18 to March 4.

Fig. 3 depicts the marginal posterior distribution of the basic reproduction rate R_0. The value of R_0 corresponding to the MLE in France is $R_0^* = \alpha^*/\beta = 4.8$. A similar computation in South Korea, based on the data used in Fig. 2 gives $R_0^* = 2.6$.

Actual number of infected cases. Using the posterior distribution of the model parameters, with the constraint that t_0 is between January 13 and 30, we can compute the daily distribution of the actual number of infected peoples. This distribution is represented across time in Fig. 4. We deduce the following ratios between the actual number of infected people and the observations, $I(t)/\Sigma\hat{\delta}_t$ (with $\Sigma\hat{\delta}_t$ the sum of the observed infected cases at time t). Thus, in France, the estimated ratio between the actual number of infected and observed cases is 15 (95%-CI: 4-33).



(a) France

Figure 3: **Posterior distribution of the basic reproduction rate R_0 in France.**

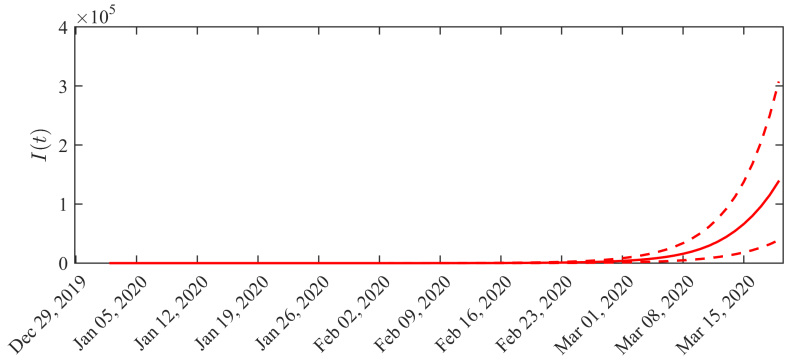


Figure 4: **Distribution of the actual number of infected cases in France across time.** Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise posterior quantiles.

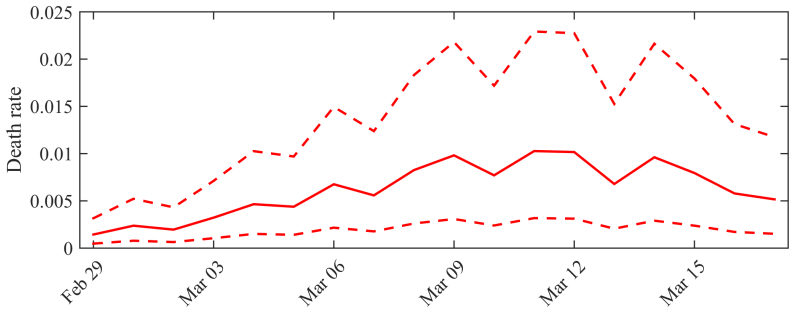


Figure 5: **Dynamics of the mortality rate in France.** Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise quantiles.

Actual mortality rate. The mortality rate corresponds to the fraction of the infected who die, that is $\gamma(t)/(\gamma(t)+\beta)$. The term $\gamma(t)$ is computed using the formula (2) and the mortality data. We thus obtain, on March 17, a mortality rate in France of 5.2/1000 (95%-CI: $1.5/1000 - 11.7/1000$). The temporal dynamics of the mortality rate are depicted in Fig. 5.

Discussion. *On the number of infecteds and the mortality rate.* The actual number of infected individuals in France is probably much higher than the observations (we find here a factor $\times 15$), which leads at a lower mortality rate than that calculated on the basis of the observed cases. However, if the virus led to contaminate 80% of the French population (Ferguson et al., 2020), the total number of deaths to deplore in the absence of variation in the mortality rate (increase induced for example by the saturation of hospital structures, or decrease linked to better patient care) would be $277,000$ (95%-CI : $81,000 - 629,000$)). This estimate could be corroborated or

invalidated when 80% of the population will be infected, eventually over several years, assuming that an infected individual is definitively immunized. It has to be noted that measures of confinement or social distancing can decrease both the percentage of infected individuals in the population and the degree of saturation of hospital structures.

On the differences between France and South Korea. The mechanistic-statistical SIR model achieves a satisfactory goodness-of-fit for French data, but does not capture the decline in the number of cases observed in South Korea. Grounding the analysis of the South Korean situation on the first phase of the epidemic improves the goodness-of-fit and yields an estimated R_0 twice as low as the French R_0, resulting on a slower epidemic dynamic. The difference between the dynamics predicted by the SIR model and the South Korean data is probably linked to a different management of the epidemic in Korea, having a strong impact on the epidemic dynamics (more important screening, tracing, social distancing in South Korea).

On the value of R_0. The value of R_0 obtained in South Korea is consistent with recent estimates for COVID-19 (2.0,2.6); see Ferguson et al. (2020). The estimated distribution in France is therefore surprisingly high. This difference could be due to a different definition of R_0 depending on the type of model used to calculate it. A direct estimate, by a non-mechanistic method, of the parameters (ρ, t_0) of a model of the form $\hat{\delta}_t = e^{\rho\,(t-t_0)}$ gives $t_0 = 30$ (January 30) and $\rho = 0.19$. With the SIR model, $I'(t) \approx I\,(\alpha - \beta)$ for small times ($S \approx N$), which leads to a growth rate equal to $\rho \approx \alpha - \beta$, and a value of $\alpha \approx 0.24$, that is to say $R_0 = 4.8$, which is consistent with the distribution presented in Fig. 3. Note that $\beta = 1/20$ corresponds to the median period of viral shedding of 20 days described by Zhou et al. (2020). A shorter period would lead to a lower value of R_0.

On the uncertainty linked to the data. The uncertainty on the actual number of infected and therefore the mortality rate are very high. We must therefore interpret with caution the inferences that can be made based on the data we currently have in France. In addition, we do not draw forecasts here: the future dynamics will be strongly influenced by the containment measures that will be taken and should be modeled accordingly.

On the hypotheses underlying the model. The data used here contain a limited amount of information, especially since the observation period considered is short and corresponds to the initial phase of the epidemic dynamics, which can be strongly influenced by discrete events. This limit led us to use a particularly parsimonious model in order to avoid problems of identifiability for the parameters. The assumptions underlying the model are therefore relatively simple and the results must be interpreted with regard to these assumptions. For instance, the date of the introduction t_0 must be seen as an *efficient* date of introduction for a dynamics where a single introduction would be decisive for the outbreak and the other (anterior and posterior) introductions would have an insignificant effect on the dynamics.

Appendix A: joint posterior distributions. The joint posterior distributions of the three pairs of parameters (α, κ), (t_0, α) and (t_0, κ) are depicted in Fig. 6.

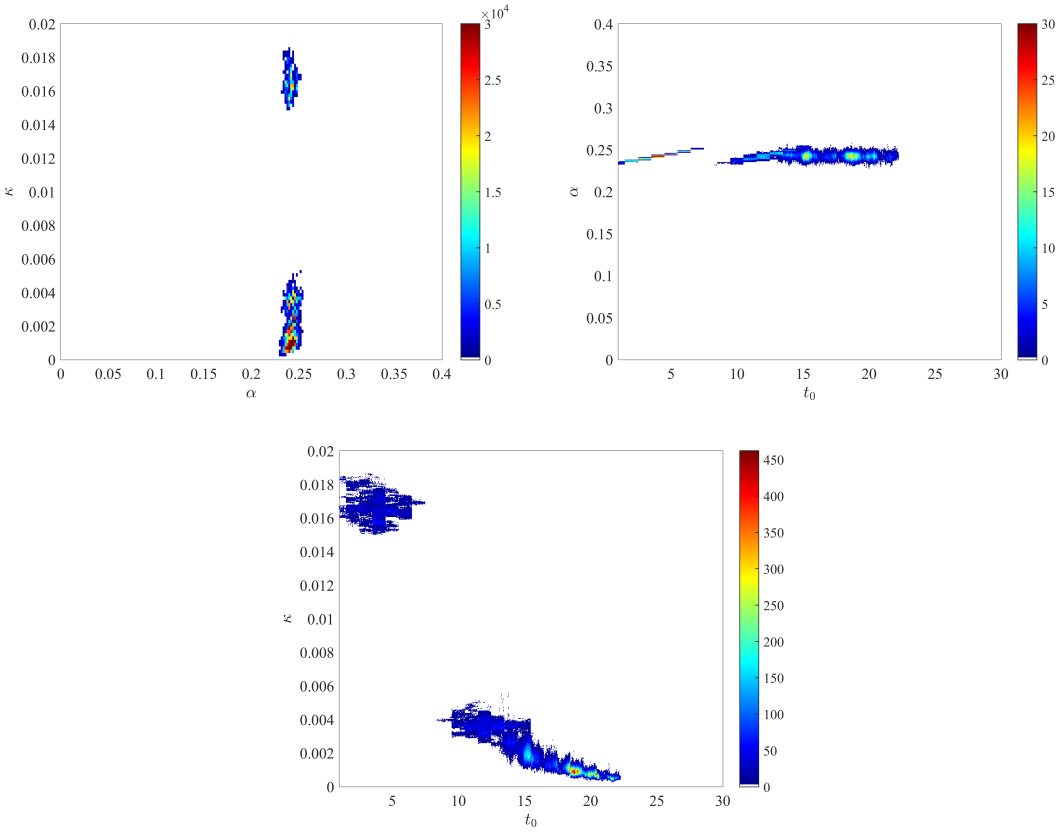


Figure 6: **Joint posterior distributions of** (α, κ)**,** (t_0, α) **and** (t_0, κ) **in France.**

Appendix B: maximum likelihood estimators. The computation of the MLEs were based on a BFGS minimization algorithm, applied to $-\ln(\mathcal{L})$, using Matlab ® *fmincon* function, starting from 2000 randomly drawn initial guesses for the parameters. This led to 2000 values of $(\alpha^*, t_0^*, \kappa^*)$. We only retained the value leading to the highest likelihood. The other values are depicted in Fig. 7.

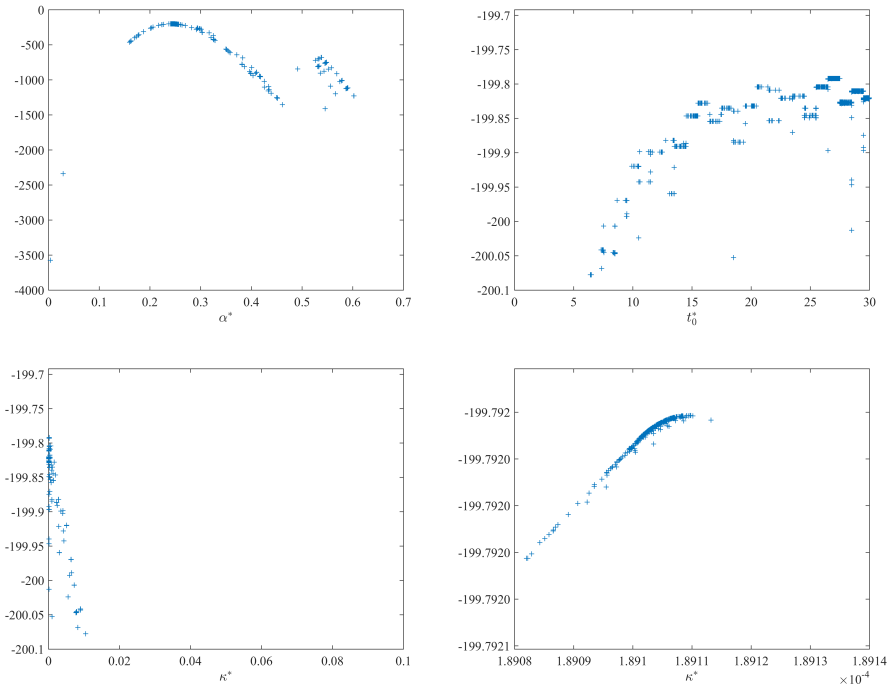


Figure 7: **Maximum likelihood estimators** computed from 2000 randomly drawn initial guesses in the intervals (4), and using the French data.

References

Abboud, C., O. Bonnefon, E. Parent, and S. Soubeyrand (2019). Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model. *Journal of mathematical biology 79*(2), 765–789.

Ferguson, N. M., D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Imperial College, London. DOI: https://doi.org/10.25561/77482*.

Murray, J. D. (2002). *Mathematical Biology.* Third edition, Interdisciplinary Applied Mathematics 17, Springer-Verlag, New York.

Roques, L. and O. Bonnefon (2016). Modelling population dynamics in realistic landscapes with linear elements: A mechanistic-statistical reaction-diffusion approach. *PloS one 11*(3), e0151217.

Roques, L., S. Soubeyrand, and J. Rousselet (2011). A statistical-reaction-diffusion approach for analyzing expansion processes. *J Theor Biol 274*, 43–51.

Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*.