Alma Mater Studiorum · Università di Bologna

Department of computer science and engineering - DISI

Bayesian melding

Luca Salvatore Lorello (919584) Giorgio Tsiotas (934519)

Compartmental models

Introduction

Epidemiological models are a class of models which describe and predict the outcome of infectious diseases. Compartmental models exploit the assumption of a large population by treating the pool of individuals as different compartments which change in size over time. Many compartmental models are deterministically described by a set of differential equations and can be complicated arbitrarily by designing how compartments behave, for example:

- compartments can be split by age,
- compartments can be split geographically,
- rates can be adjusted to consider symptomatic and asymptomatic cases,
- natural death causes,
- births,
- etc.

Differential equations can be solved naively with the Euler method:

1. Set each compartment $X \leftarrow X_0$
2. For t times: $X_i \leftarrow X_{i-1} + dX$

Compartmental models

Susceptible-infected-removed (SIR)

Only three compartments, governed by the equations:

$$dS = -\frac{\beta SI}{n}$$

$$dI = \frac{\beta SI}{n} - \gamma I$$

$$dR = \gamma I$$

This model doesn't take into account reinfections, incubation times or deceased.

- β: infection rate,
- γ: recovery rate,
- $n = S + I + R$: total population,
- $R_0 = \frac{\beta}{\gamma}$: basic reproduction number,
- $R_t = R_0 \frac{S}{n}$: effective reproduction number.

Compartmental models

Susceptible-infected-recovered-deceased (SIRD)

Four compartments, governed by the equations:

$$dS = -\frac{\beta SI}{n}$$
$$dI = \frac{\beta SI}{n} - \gamma I - fI$$
$$dR = \gamma I$$
$$dD = fI$$

This model doesn't take into account reinfections or incubation times.

- β: infection rate,
- γ: recovery rate,
- f: fatality rate,
- $n = S + I + R$: total population **still alive**,
- $R_0 = \frac{\beta}{\gamma}$: basic reproduction number,
- $R_t = R_0 \frac{S}{n}$: effective reproduction number.

Compartmental models

Susceptible-exposed-infected-recovered-deceased (SEIRD)

Five compartments, governed by the equations:

$$
\begin{aligned}
dS &= -\frac{\beta S I}{n} \\
dE &= \frac{\beta S I}{n} - \sigma E \\
dI &= \sigma E - \gamma I + c\frac{RI}{n} - fI \\
dR &= \gamma I - c\frac{RI}{n} \\
dD &= fI
\end{aligned}
$$

This model also takes into account reinfections and incubation times.

- β: infection rate,
- σ: incubation rate,
- γ: recovery rate,
- c: reinfection rate,
- f: fatality rate,
- $n = S + E + I + R$: total population **still alive**,
- $R_0 = \frac{\beta}{\gamma}$: basic reproduction number,
- $R_t = R_0 \frac{S}{n}$: effective reproduction number.

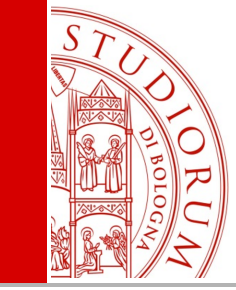
Compartmental models

Problems with SEIRD

The SEIRD model is very expressive, but it has problems:

- the E compartment is difficult (or impossible) to measure,
- the reinfection rate can be hard to estimate (especially in new diseases like COVID-19),
- too many degrees of freedom (eg. the I compartment can be tuned by changing 4 different parameters), combined with the uncertainty of measures (eg. is the estimate on the size of I compartment good or there are asymptomatic cases which make this number wrong?).

Hiding the E compartment (ie. taking E_0 as an initial parameter, but not generating E_t as output) can partially mitigate the first problem during fitting.



The dataset

National trend of COVID-19 cases in Italy published by Protezione Civile:

- multivariate time-series with daily values,
- contains missing values, but not on the columns we are interested in,
- the S compartment cannot be inferred (we assume an initial value of 610^7 individuals at t_0 and decrease it depending on the other compartments),
- split into **disjoint windows** for quick exploration of methods (not the best approach, but a good trade-off between speed and accuracy).

Deterministic seeding

Scipy's `curve_fit`

Scipy allows to fit an arbitrary function via non-linear least square optimization using the `curve_fit` method. Since it requires arrays of dependent and independent variables, the compartmental models need to be implemented in order to work vectorially:

- `__init__` method: sets the model's parameters,
- `eval_series(t)` method: returns the predictions at times $0..t$,
- `eval_last(t)` method: returns only the last prediction (at time t), used for melding,
- $f(t, \beta, \gamma, \dots)$ method: wraps `eval_last` in order to work with the parameters passed as arrays and sets the parameters at each call, used by `curve_fit`.

Fitting is performed on an ensemble of estimators, each optimized to predict a single window given the parameters at the first day. Values are bounded in $[0, \infty)$ for compartments and in $[0, 1]$ for rates. The initial guess for compartments is the value on the first day of the window, while for rates is 0.5.

Deterministic seeding

Results (3-days windows)

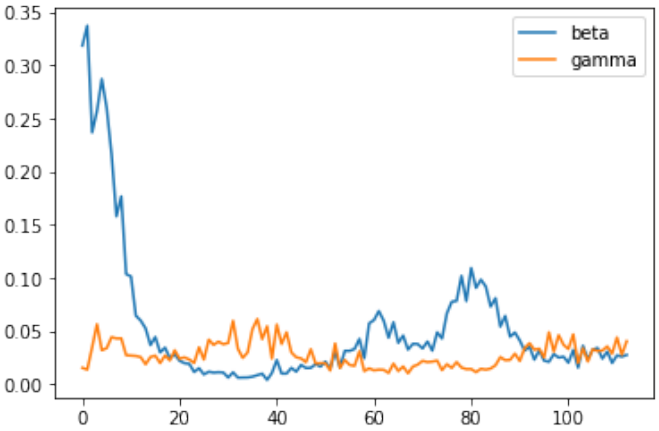


Figure: SIR

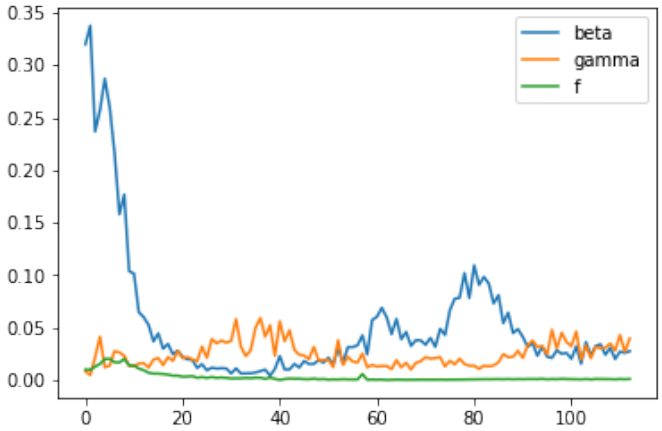


Figure: SIRD

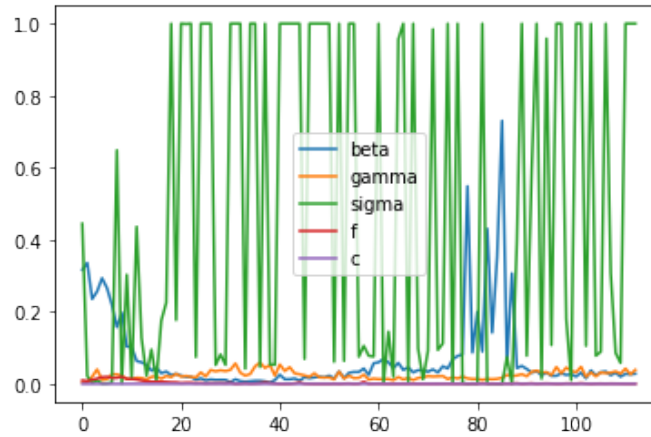


Figure: SEIRD

SEIRD has too many degrees of freedom (also because of hidden E compartment) and in order to fit the observations, the optimizer gives arbitrary values to the parameters (especially σ).

Deterministic seeding

Results (7-days windows)

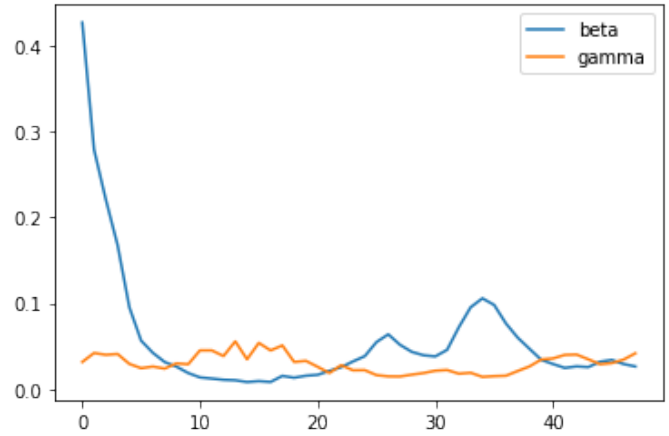


Figure: SIR

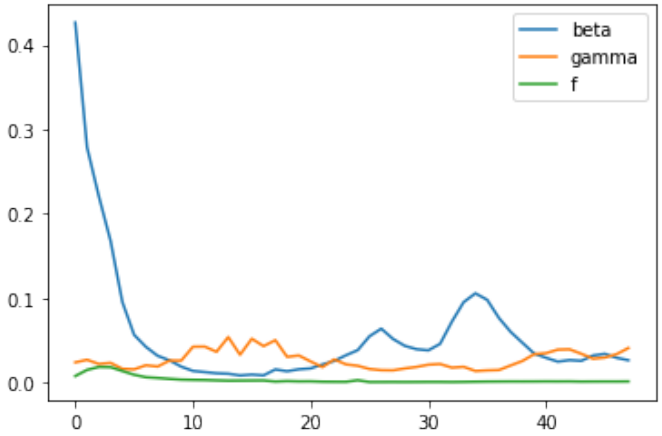


Figure: SIRD

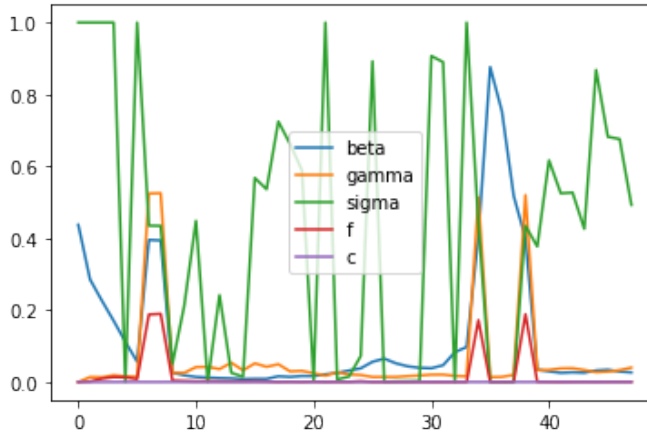


Figure: SEIRD

Deterministic seeding

Results (15-days windows)

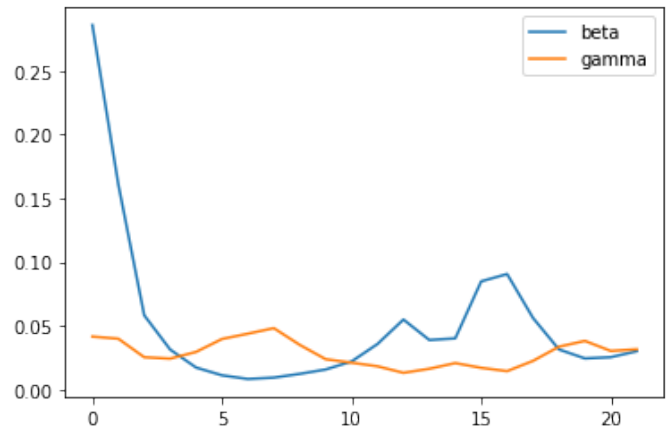


Figure: SIR

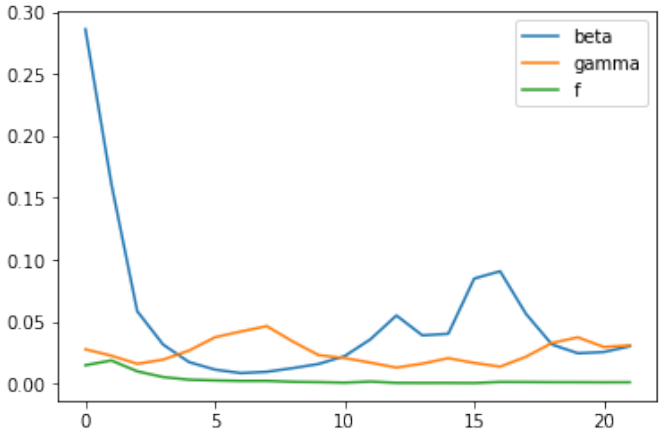


Figure: SIRD

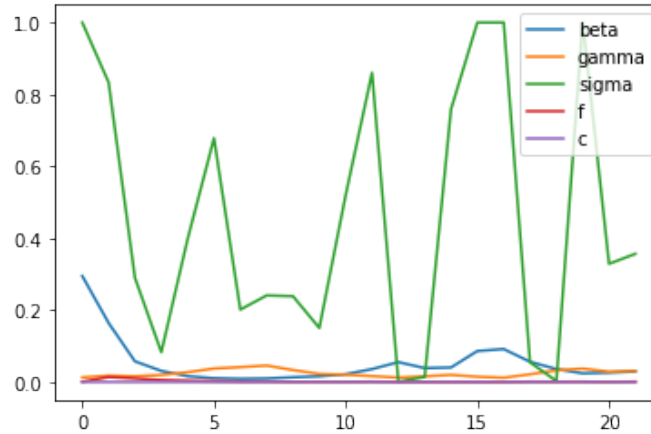


Figure: SEIRD

Deterministic seeding

Improving the Hidden SEIRD fit

In order to constrain the SEIRD model, we bounded σ to values from literature[1], but still got spurious spikes. In another test we further narrowed the interval to 0.093 ± 0.01 (the average between 8.2 and 15.6 days), obtaining better results.

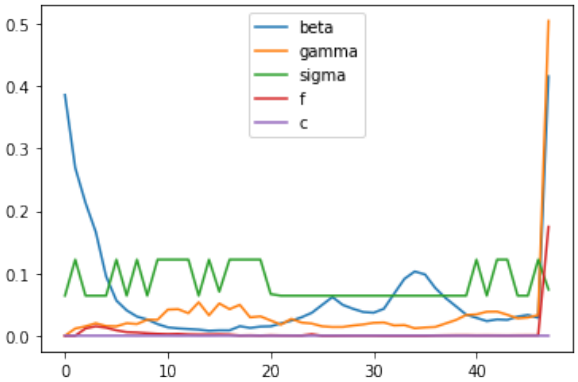


Figure: SEIRD with $\sigma \in [\frac{1}{15.6}, \frac{1}{8.2}]$ (7-days window)

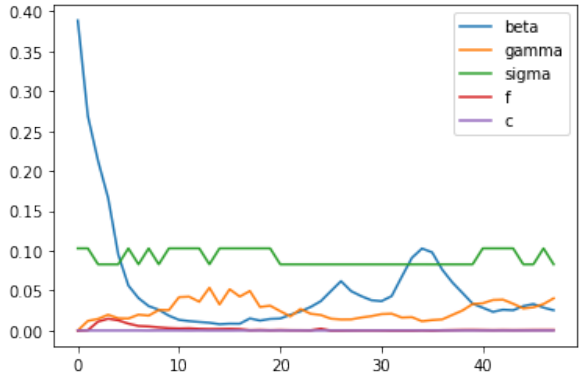


Figure: SEIRD with $\sigma \in [0.083, 0.103]$ (7-days window)

[1] Stephen A Lauer et al. "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application". In: *Annals of internal medicine* 172.9 (2020), pp. 577–582.

Bayesian melding

Idea

A deterministic model can be seen as a cause-effect relationship between inputs and outputs, however the probabilistic informations available are lost. The most simple approach would be to create a random variable $Y = f(X)$ as the application of the model to the random input, but this would discard all the available informations about the output distribution (ie. the likelihood of each possible outcome).
Bayesian melding is a statistical technique which puts together (meld) all the available informations about inputs and output distributions, without suffering from Borel's paradox (unlike older techniques as Bayesian synthesis).

Borel's paradox

Conditional probabilities on events with null probability causes the entire pdf not to be reparametrization-invariant (eg. simply changing the scale of variables causes the shape of the pdf to change).

Bayesian melding as consensus among many pdfs

A "coherent" pdf can be derived by pooling different pdfs. Logarithmic pooling is the only method which preserves **external Bayesianity** (on multivariate pdfs, applying Bayes theorem before or after creating the joint probability yields the same results).

$$T(q_1, q_2, \ldots, q_n) \propto \prod_{i=1}^{n} q_i^{\alpha_i}, \ with \ \sum_{i=1}^{n} \alpha_i = 1$$

Bayesian melding

Which pdfs to pool?

When fitting a deterministic model, we have at our disposal:

- hypotheses on inputs and outputs (priors),
- observations from the phenomenon we are trying to model (likelihoods).

We can meld the output prior q_2 and the induced output distribution q_1^* (computed by applying the input prior to the model) to get a "coherent" pdf on outputs by a factor α, then, applying Bayes theorem, we can condition that pdf to the actual observations:

$$P(\texttt{input} = \theta \mid \texttt{output} = M(\theta)) = \left(\frac{q_2(M(\theta))}{q_1^*(M(\theta))}\right)^{1-\alpha} L_1(\theta) L_2(M(\theta))$$

Bayesian melding

Sample importance-resampling algorithm

In order to compute the induced density, the model needs to be inverted ($q_1^*(M(\Theta)) = q_1(M^{-1}(\Phi))|J(\Theta)|$, where J is the Jacobian of the model in functional form), but this is usually not possible. The sample importance-resampling algorithm allows to bypass this problem, by approximating the pooled distribution as follows:

1. Sample phase: extract a large number of samples from the input prior,
2. Importance computation phase: weight each sample:
   - 2.1 run the model on each sample to get the posteriors,
   - 2.2 estimate q_1^* from the posteriors with a non-parametric method (eg. Gaussian KDE),
   - 2.3 associate each sample Θ_i to a weight $(\frac{q_2(M(\Theta_i))}{q_1^*(M(\Theta_i))})^{1-\alpha} L_1(\Theta_i) L_2(M(\Theta_i))$,
3. Resampling phase: take a small subset of the samples, following the weights computed,
4. The new samples (approximately) follow the input pdf which best explains the observed outputs (and therefore its mean can be used to fit the model, with the variance as a measure of confidence).

Execution can be very slow (5 days on SEIRD for the entire dataset split into 7-days windows) due to the large number ($\sim 10^5$) of samples and the posterior estimation.

Bayesian melding

Our assumptions on distributions

We are dealing with multivariate pdfs, both for inputs and outputs. The joint probability for a given model (eg. SEIRD) can be computed as the product of the marginals, under an independence assumption, or can itself be pooled. We needed to make further assumptions on the probabilities:

- input likelihoods for rates (eg. β, σ, etc.) are unknown, therefore they are set as uniform in $[0, 1]$,
- input priors for rates were seeded with the results of `curve_fit` and set as Gaussians with the given value as mean,
- every density related to compartments (both prior and likelihood) set to Gaussians with the known value as mean,
- for every marginal density function (except the uniform ones), deciding a variance is very problematic: an initial heuristic for compartments is to take the variance of the values inside the window we want to fit, when this doesn't work, a different value is set.

What happens when no consensus is reached?

When the melded pdfs don't have sufficient overlap, underflows and divisions by zero may occur. This makes choosing the right model and priors very important for the outcome.

Bayesian melding

Results (7-days windows, independent marginals assumption)

The SEIRD model behaved badly also during least square optimization, therefore its values were averaged (EMA, $\alpha = 0.05$) in order to try to reduce the chance of poor melding.

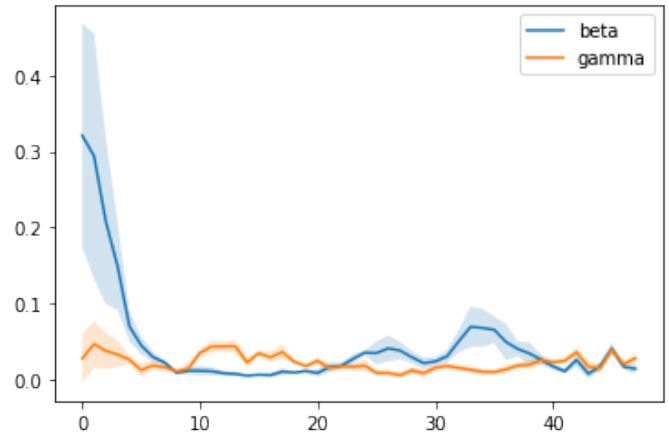


Figure: SIR (10^4 samples, 500 resamples)

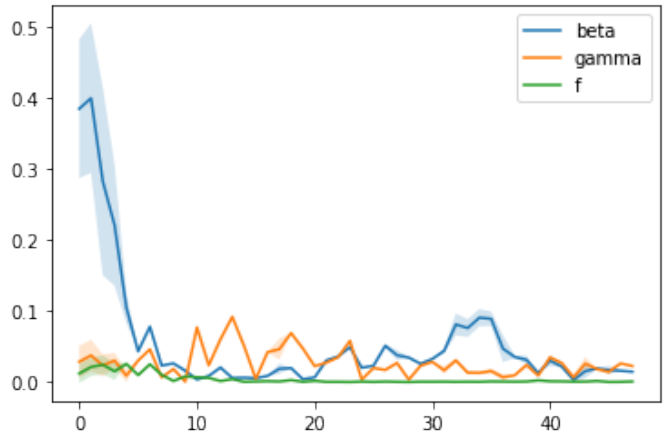


Figure: SIRD ($5 \cdot 10^4$ samples, 2000 resamples)

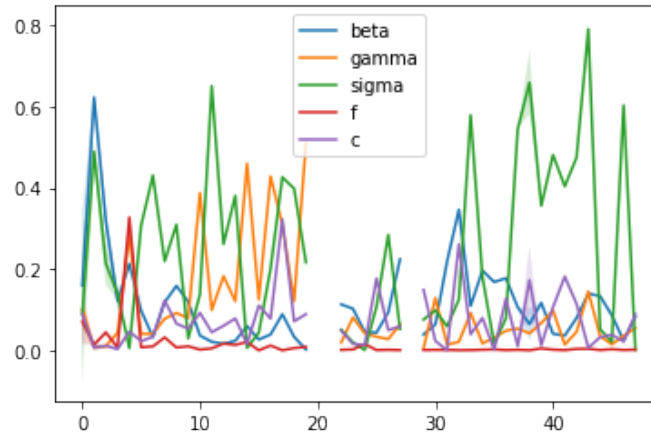


Figure: SEIRD (10^5 samples, 5000 resamples)

Bayesian melding

Results (7-days windows, marginals pooled geometrically)

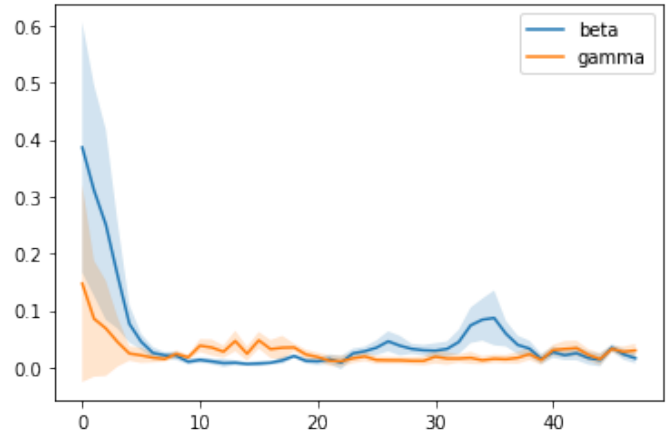


Figure: SIR (10^4 samples, 500 resamples)

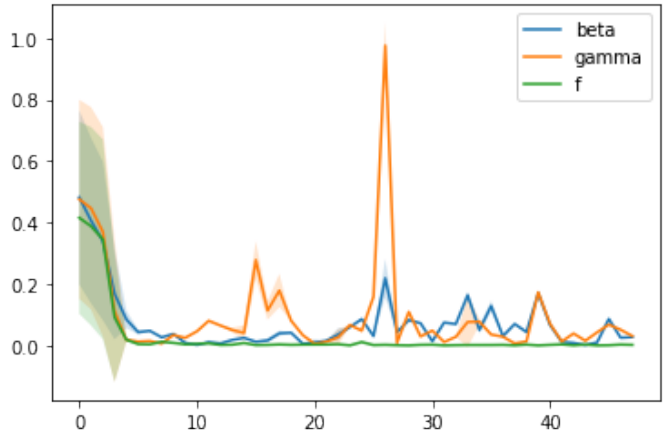


Figure: SIRD ($5 \cdot 10^4$ samples, 2000 resamples)

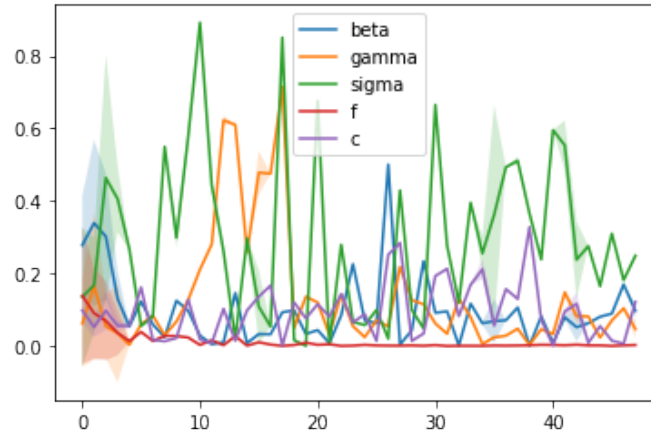


Figure: SEIRD (10^5 samples, 5000 resamples)

Bayesian melding

Results (15-days windows, independent marginals assumption)

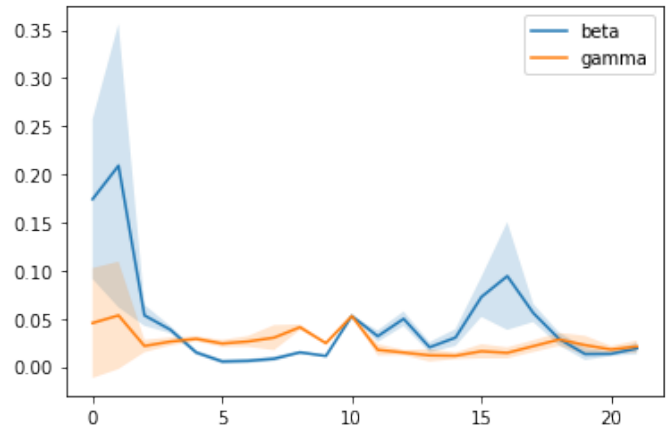


Figure: SIR (10^4 samples, 500 resamples)

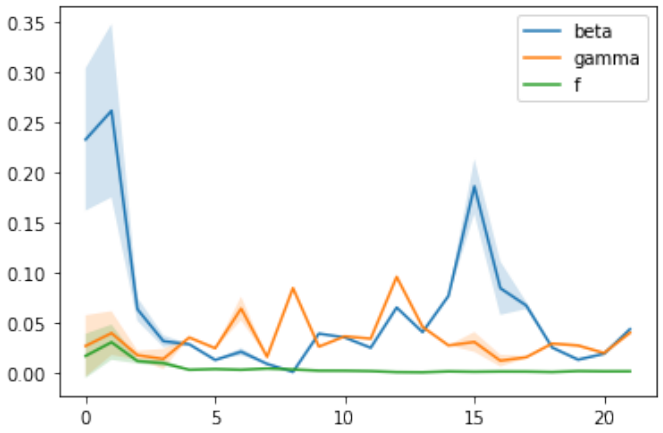


Figure: SIRD ($5 \cdot 10^4$ samples, 2000 resamples)

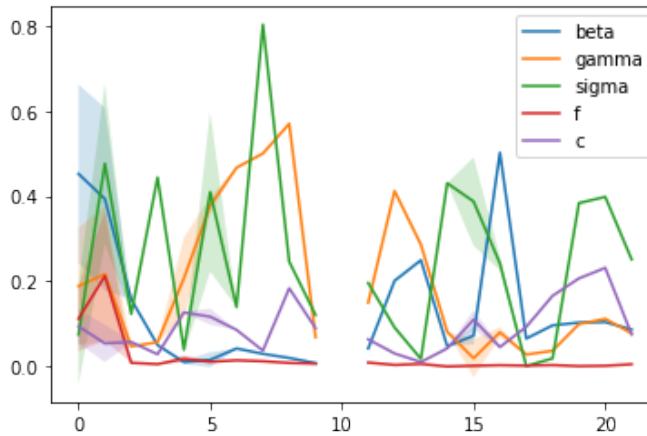


Figure: SEIRD (10^5 samples, 5000 resamples)

Bayesian melding

Possible sources of failure? (1/2)

Although melding has higher potential than deterministic fitting when handling stochastic informations, its performance was worse than `curve_fit` baseline. Possible causes of these results may be:

- observations are not really stochastic (the dataset on COVID-19 contains hard numbers and we created the pdfs synthetically),
- sampling was too shallow (other works on Bayesian melding applied to compartmental models suggest to use at least $10^5 - 10^6$ samples and $\sim 10^3$ resamples),
- consensus between priors and likelihoods was not reached (as evidenced by plenty of underflows during melding),

Bayesian melding

Possible sources of failure? (2/2)

- the model's domain and the pdfs support don't match (eg. $\beta \in [0, 1]$ for the model, but $S \in (-\infty, \infty)$ for a Gaussian pdf centered on the seed) and this adds noise during weighting (the model "saturates" invalid inputs, so an invalid value has the same weight as the boundary),
- bandwidth selection in Scipy's KDE is not optimized for multivariate case (the authors of melding propose to use Terrell's Maximum Smoothing Principle[2] to select automatically the best bandwidth in a scale independent and high-dimension tolerant fashion, but Scipy offers only Silverman's rule, which fails when the true distribution is very different from a Gaussian, or Scott's rule, which has similar problems[3]).

---

[2] George R Terrell. "The maximal smoothing principle in density estimation". In: *Journal of the American Statistical Association* 85.410 (1990), pp. 470–477.

[3] *Kernel Bandwidth: Scott's vs. Silverman's rules.* URL: https://stats.stackexchange.com/questions/90656/kernel-bandwidth-scotts-vs-silvermans-rules.

Appendix - Original project proposal slides

The following slides summarize some articles about Bayesian melding and were used to give an initial informal presentation about the method. They are not formatted, nor proof-read, and contain redundant content.

Introduction[4]

[4] *Bayesian melding - University of Washington.* URL: https://sites.stat.washington.edu/hana/talk/bm_slides.pdf.

Bayesian melding

What?

Statistic technique which is used to quantify a model output's uncertainty (ie. a probability distribution over the model predictions).
A deterministic model applied to a random variable produces another random variable, however posterior informations are discarded.
Posterior informations may not be the direct output of the model (eg. mortality rates in COVID-19 cases are known, but a model could output the number of infected people instead).

Bayesian melding

Ingredients

Let's denote the inputs as Θ and a model which produces the outputs Φ as $M(\Theta) = \Phi$.
We know from the real world:

- Prior on the inputs: $q(\Theta)$
- Data related to the outputs, which yield a posterior on the outputs $L(\Phi) = P(Data|\Phi)$
- Posterior on the inputs: $p(\Theta) \propto q(\Theta)L(\Phi)$ (Bayes theorem)

Typically we are interested not directly on the model outputs, but in quantities ψ which are a function of them (with their own distributions which can give additional insight).

Bayesian melding

What can be done

The model can be used to estimate the output, producing another posterior on the inputs which can be different than the one achieved according to Bayes theorem.

- Better prediction: a traditional approach tends to ignore the posterior on the outputs (or to delegate it as a validation for the model), but a better approach would be to consider all the available information, similar to majority voting in an ensemble model
- Better coverage: traditional simulations underestimate uncertainty, taking the posteriors into account the observed values fall more often inside an interval centered at the predictions
- Better calibration: the output space is sampled more efficiently.

Typically the model is not invertible, therefore the likelihoods can be estimated only via Monte Carlo approaches.

Deterministic models

1. Draw N samples $\{\Theta_1, \Theta_2, \ldots, \Theta_N\}$ from the prior $q(\Theta)$
2. For each Θ_i compute the model output $\Phi_i = M(\Theta_i)$
3. Compute weights $w_i = L(\Phi_i)$
4. The (approximate) posterior on the inputs can be approximated by weighting $\{\Theta_1, \Theta_2, \ldots, \Theta_N\}$ by $\{w_1, w_2, \ldots, w_N\}$
5. Likewise, the posterior on the quantities of interest uses the same weights on $\{\psi_1 = \Psi(\Phi_1), \psi_2 = \Psi(\Phi_2), \ldots, \psi_N = \Psi(\Phi_N)\}$

Stochastic models

1. Draw N samples $\{\Theta_1, \Theta_2, \ldots, \Theta_N\}$ from the prior $q(\Theta)$
2. For each Θ_i compute the model output J times obtaining $\Phi_{ij} = M(\Theta_i)$ $(j \in [1, J])$
3. Compute weights $w_i = L(\overline{\Phi}_i)$, where $\overline{\Phi}_i = \frac{1}{J} \sum_{j=1}^{J} \Phi_{ij}$
4. The (approximate) posterior on the inputs can be approximated by weighting $\{\Theta_1, \Theta_2, \ldots, \Theta_N\}$ by $\{w_1, w_2, \ldots, w_N\}$
5. This time, the posterior on the quantities of interest depends directly on the output distribution, without considering the weights.

Original paper[5]

[5] David Poole and Adrian E Raftery. "Inference for deterministic simulation models: the Bayesian melding approach". In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1244–1255.

Bayesian synthesis

Simpler method, but suffers from the Borel paradox. Given $p(\Phi, \Theta)$ (joint pre-model distribution, ie. everything known about inputs and outputs from the real world, except what is modeled deterministically) and the model $M : \Theta \mapsto \Phi$, the post-model distribution is:

$$\begin{cases} p(\Theta, M(\Theta)) & if\ \Phi = M(\Theta) \\ 0 & otherwise \end{cases}$$

Applying the Bayes theorem, $p(\Theta, \Phi) \propto q_1(\Theta) q_2(\Phi) L_1(\Theta) L_2(\Phi)$, where the distribution has been decomposed in terms of priors (q) and likelihoods (L) on inputs and outputs.
Sample importance resampling is used to get a prediction.

Borel paradox

The way the post-model distribution is defined in Bayesian synthesis is ill posed and its values depend on how the model is parametrized (a counter-example was given by simply modeling an exponential growth model and computing input distributions in linear or log scale and observing different results).
More generally, the Borel paradox manifests itself if a conditional distribution is defined on an event with probability 0.
The paradox in Bayesian synthesis is not due to likelihoods (L_1 and L_2, which are invariant wrt reparametrization), but due to prior distributions. Intuitively this is caused by the fact that from the input prior $q_1(\Theta)$ and the model $M(\Theta)$ another prior on the outputs ($q_1^*(\Phi)$) can be derived and may be different from $q_2(\Phi)$.

Bayesian melding as consensus pooling

Assuming a "coherent" prior $\tilde{q}$ can be derived, applying Bayes theorem yields $\tilde{q}(\Theta)L_1(\Theta)L_2(M(\Theta))$ which is a traditional Bayes posterior and only marginal distributions (instead of pre-model and post-model distributions) are needed.
Sample importance resampling is still required due to the impossibility of deriving the inverse of the model analytically.
Finding $\tilde{q}$ can be seen as determining the consensus in an ensemble. Pooling different distributions is a statistical way of determining consensus.
External Bayesianity is the property of a pooling method which guarantees posterior invariance with respect to forming the multivariate prior and then applying Bayes theorem or, viceversa, applying Bayes theorem on the univariate priors and then forming the multivariate posterior.

Logarithmic pooling

$$T(q_1, q_2, \ldots, q_n) \propto \prod_{i=1}^{n} q_i^{\alpha_i}, \ with \ \sum_{i=1}^{n} \alpha_i = 1$$

Logarithmic pooling is the only mechanism which satisfies external Bayesianity, it's also invariant wrt rescaling of priors and naturally sets to zero the pooled distribution if one of the priors is zero for a given value (zero preservation property, important in melding because allows to easily eliminate "impossible" values according to model or evidences).
For two distributions (the model output and the evidence), only α and $1 - \alpha$ are used and the value of α is arbitrary and tunes "reliability" of the distributions (not precision of melding).
Geometric pooling is logarithmic pooling with $\alpha = \frac{1}{2}$ (ie. same importance to model and evidence).

Sample importance resampling algorithm

On a non-invertible model pooling still produces unambiguous results, however inference on the inputs does not. SIR works both for continuous and discrete models, even when they are non-invertible or hard to compute.

1. Draw k samples of Θ from the prior $q_1(\Theta)$
2. For each Θ_i run the model to get Φ_i
3. Estimate $q_i^*(\Phi)$ with nonparametric density estimation (eg. KDE with Gaussian kernel)
4. Create weights $w_i = (\frac{q_2(M(\Theta_i))}{q_1^*(M(\Theta_i))})^{1-\alpha} L_1(\Theta_i) L_2(M(\Theta_i))$
5. Sample l values with probabilities proportional to w_i, which will be an approximation of the posterior.

SIR may be slow with complex models and large numbers of samples. Improved algorithms are available in literature.

Model validation

"Goodness" can be defined as the property of a model to produce for any input, output or quantity of interest, a "substantial overlap" with the other sources of informations (priors on inputs, priors on outputs, likelihoods on inputs and likelihoods on outputs, where some of them may be unavailable). Given two distributions, the overlap can be estimated qualitatively by sampling random values and observing the boxplots (which should be more or less of the same size and around the same mean).

- Prior on the outputs: $q_2(\Phi)$ is compared with $q_2^*(\Phi)$ (which is induced by the prior on the inputs)
- Prior on the inputs: $q_1^*(\Theta)$ is computed as the pooled prior with $\alpha = 0$ (ie. pooling considering only the model) and then compared with $q_1(\Theta)$
- Likelihoods need an heuristic to determine $L_1^*(\Theta)$ and $L_2^*(\Phi)$ distributions comparable with their counterparts.

Hypothesis testing

Given two models M_0 (null) and M_1 (alternative), a Bayes factor B_{10} can be computed and if it's less than 1, the null model explains better the evidence, while above the alternative is better:

- Between 1 and 3 the evidence on the alternative being better is weak
- Between 3 and 20 the evidence on the alternative being better is positive
- Between 20 and 100 the evidence on the alternative being better is strong
- Above 100 the evidence on the alternative being better is very strong.

The Bayesian factor can be computed by modifying slightly the SIR implementation for Bayesian melding.

Relabeling

In some cases the model may be reformulated by relabeling some inputs as outputs (or viceversa), eg. instead of knowing the values at time 0 and regressing at time t, it may be desirable to regress time 0 knowing the values at t.
If the variants of the model M are $\tau = M_1(\lambda, \gamma)$ and $\lambda = M_2(\tau, \gamma)$ (where γ are the shared inputs and the other two are variables which form the input for one model but the output of the other one) and the transformation $(\lambda, \gamma) \mapsto (\tau, \gamma)$ is one-to-one, then Bayesian melding can be applied on the (independent) priors $q_1(\lambda)$, $q_2(\gamma)$ and $q_3(\tau)$ (taking care on how pooling is done to consistently determine weights for both variants).

Melding on deterministic models[6]

[6] Leontine Alkema, Adrian E Raftery, and Tim Brown. "Bayesian melding for estimating uncertainty in national HIV prevalence estimates". In: *Sexually transmitted infections* 84.Suppl 1 (2008), pp. i11–i16.

Melding on time series

Plausibility bounds for HIV prevalence were given by experts, but lacked a formal statistical assessment (ie. they were not confidence intervals).

Bayesian melding was used to calibrate the prediction curves of an EPP model (deterministic).

Sampling prevalence curves: from the posterior on HIV prevalence SIR (sampling importance resample) algorithm was used:

1. Sample a large number of input parameters (according to their prior joint distribution)
2. Run the model to obtain a time series for each input
3. Compute weights based on how well the prediction fits the actual data (ie. $w_i = p(M(\Theta_i)|\Phi)$)
4. Resample input parameters with a probability proportional to weights and rerun the model

The input samples should be large enough (eg. 200000) to cover the input space and to produce a reasonable number of unique predictions. Resampling can be done with a much lower number of runs (eg. 3000)

Curve calibration based on known posterior

ANCs (antenatal curves) are used to estimate HIV prevalence, however these data differ from population prevalence.
Population-based surveys are used to meld the ANC curves and get a better estimate.

1. Rescale the time series based on the surveys (eg. shift by the mean and multiply by the variance)
2. Compute weights (based on the surveys' data, NOT the ANC curves)
3. Resample

Curve calibration based on similar data

ANCs (antenatal curves) are used to estimate HIV prevalence, however these data differ from population prevalence. If population-based surveys are unavailable, melding is performed against different countries which have those data.
Assume that the difference between ANC and surveys is given by a bias (which is unknown since surveys are missing in this case), the mean and variance of which can be estimated by data on other countries.

1. Compute mean and variance of bias (difference between ANC and survey data) in other countries
2. Rescale the time series based on the bias
3. Resample

Best trajectory

The best trajectory is given by the most likely prediction given prior and data (maximum a posteriori trajectory, MAP).
Due to Bayes theorem, the posterior probability is given by the product of prior on input, prior on output and likelihood of data.
The model is run for subpopulations (eg. rural vs urban areas), the posteriors are weighted by population size and the MAP is selected for the national trajectory. Confidence intervals can be evaluated on the MAP.

Weights for little data[7]

[7] Luiz K Hotta. "Bayesian melding estimation of a stochastic SEIR model". In: *Mathematical Population Studies* 17.2 (2010), pp. 101–111.

Best trajectory using limited data

Sample importance resampling algorithm was applied to a SEIR model on co-infection of HIV and tuberculosis on prison inmates (the removal rate included deaths, recoveries and releases from prison), with a lot of missing data (eg. if the inmates had HIV before entering prison). A priori distributions of HIV and TB cases are assumed independent.
The simulations were too few to get a good estimate of weights for resampling, so they were approximated by a multivariate normal, using means and variances from the posterior (symptomatic cases).
This procedure still achieved better results than least square error estimation on a Monte Carlo Markov Chain simulation on the same model applied to the same priors.

Exploit posterior knowledge to improve deterministic models[8]

[8] Miguel Sánchez-Romero et al. "How many lives can be saved? A global view on the impact of testing, herd immunity and demographics on COVID-19 fatality rates (supplementary material)". In: *medRxiv* (2020).

SEIR model

- 95 age groups for age-specific morality rates
- Isolation explicitly modeled
- Infectious period after testing
- E and I (one for each age group) components are split into E^u, E^d, I^u and I^d (undetected and detected fractions)
- E^d is assumed to be quarantined
- R (recoveries) and D deaths are modeled separately
- M (death by other causes) and $M^c(\varepsilon)$ (death by COVID-19) diagonal matrix of death rates for each of the 95 age groups. ε is the symptomatic fraction (the lower it is, the higher the divergence between the true fatality rate and the estimated one)
- Time steps are daily.

Calibration

Deaths are more reliable than the positive tests, even when the country decides to count as COVID-19 deaths only those of people who were tested positive.
Ordinary least square regression to model the age gradient in mortality (in log scale), very good fit after 30 years of age.
Prior on ε was set to be uniform in $]0; 1[$ and the mortality matrices were built by adjusting from annual to daily rates and by unpacking age groups with a penalized composite link model (the official statistics in the human mortality database are grouped in 5-years bins) and smoothed with a Kannisto model after 80 years of age.

Priors

Epidemiological characteristics of COVID-19 are still unknown, so all priors are uninformative (ie. uniform distributions between a minimum and a maximum value for each variable $S, E^u, E^d, I^u, I^d, R, D$).
Parameters are taken from the literature, except for ε which is sampled between 0 and 1.
Time s of the SEIR model is the number of days elapsed from the first death by COVID-19 in the country. An additional variable T models the time instant in which quarantine measures start to slow down spreading. Before T deaths reported can have an high error, so the joint prior on the outputs is uninformative as well (1 if the error at time T is less than the maximum error up to time T, 0 otherwise).

Melding

Weights were computed with geometric pooling on the priors $q_1^{\alpha} \cdot q_2^{1-\alpha}$ (q_1 prior on the inputs, q_2 prior on the outputs, $\alpha = 0.5$ to give the same importance to model and data). Posteriors on the input were estimated using a standard gaussian KDE.