CSE 397: Validation & UQ Robert Moser Course Project

Shreyas Gaikwad

We will develop a plan for validation and uncertainty quantification for a model problem of epidemic modeling.

A validation plan consists of the following elements, which should provide answers to the indicated questions as concisely and precisely as possible.

1. Goals and Objective

   (a) What system is to be simulated and why?

   We simulate the spread of an epidemic in Travis County. Mathematical models can be used to project how highly contagious diseases spread in order to show likely outcomes of epidemics and thus help inform public policy.

   (b) What scenarios are to be simulated (the scenarios of interest)?

   The scenario of interest is the reopening of Travis County in order to resume economic activities. We will assume there is no migration and thus, the scenario we study is one where the total population (living and dead) is constant.

   (c) What quantities are to be computed as outputs of the simulation (the quantities of interest)?

   We wish to predict how the number of infected people changes with time. We are also interested in predicting the number of critically infected people as well as the number of deaths.

   (d) What will the simulation outputs (QoIs) be used for?

   The projection of the total number of infected people will be used to determine if we can fully reopen. The projected number of critically infected people will be helpful for the authorities to stock up on the medical equipment such as testing kits, ventilators, masks, etc. The aim is to avoid straining the local healthcare system. And obviously, we do not want "too many" deaths happening, so an accurate forecast of the number of deaths is pretty important too.

   (e) What are the requirements for uncertainty characterization?

   The uncertainty characterization is supposed to give us a confidence interval for forecasting, since even if we have a really optimistic forecast, it could be extremely sensitive to the uncertainties in our system and the on-the-ground situation might be much worse

than what we forecast. It is important to thus get an estimate of how far off we could be in our estimate. This characterization is supposed to be an assessment of the risk in reopening the Travis County.

2. Model Specification

   (a) What mathematical model is to be used in the simulations?

   We use the modified *Susceptible-Exposed-Infected-Recovered-Model (SEIR)* for these simulations. This choice is made because we know as metadata/prior information ("word on the street") that the virus has a significant incubation period and thus normal *SIR* models cannot capture the aspect that accounts for the population that is exposed to the virus and is asymptomatic. Based on our QoI, the *SEIR* Model has been enriched to have four extra classes in this case *Total Population(N), Deaths (D), Severe cases (d), Turnover(C).* N is the total population (living and dead) which under suitable assumptions discussed later, will not change since there is no migration. R accounts for both recovered and dead people. C represents the total number of people infected thus far. The model is mathematically expressed as (the models for the QoI are an addition to the basic model and will be discussed later),

$$N = S + E + I + R$$
$$N_t = 0 \implies S_t + E_t + I_t + R_t = 0$$

   (b) What are the reliable aspects of the model; that is, what parts of the model need not be questioned in the context of the system and scenarios to be simulated?

   There is no migration and we can ignore the dynamics of birth and death (here death refers to deaths not caused by the coronavirus). Hence, one reliable aspect of the *SEIR* model is that the total population (living or dead) is constant.

   (c) What reliable theory or principle underlies the model and what are the limits on the reliability of these aspects of the model.

   The total number of people, living or dead, in an isolated area, will not change with time. The limits on the reliability of this principle would be that the county (county is just a placeholder for any region since our discussion is pretty general) is not actually isolated and there is a significant influx and outflux of people. The reliability is most questionable if the influx brings in a greater fraction of exposed/infected people than the fraction of exposed/infected people in the county and there is no outflux.

   The dynamics of an epidemic, for example, the coronavirus, are often much faster than the dynamics of birth and death, therefore, birth and death are often omitted in simple compartmental models (like the *SEIR* models). If the spread of the epidemic is sufficiently curtailed, the birth-death dynamics are likely to be more important. We would then have

to suitably amend the model.

(d) How do we know that the simulation scenarios are within these limits?

We know that that there is no immigration or emigration due to the quarantine and social-distancing norms.

Let us look at the time scales of birth-death dynamics and the coronavirus. We need to compare the birth-death rates with the coronavirus infection rates in an unrestricted scenario (since our scenario of interest is reopening the county). We do some simple back of the envelope calculations. Consider the case of the US. A child is born every 8 seconds, meaning there are a total 324,000 births every month. A person dies every 12 seconds, meaning around 216,000 deaths per month. Therefore the population increases by 108,000 every month. In comparison, the number of cases (based only on the limited testing done so far) in the month of April alone rose by about 850,000. This growth is following an exponential trend and we have not even begun to flatten it (meaning that there will even more cases/month in the next few months), which further strengthens our case that the simulation scenarios are within the limits of applicability of our model.

3. Embedded Models

(a) For what quantities are embedded models needed?

Embedded models are needed for $S_t, E_t, I_t, R_t, D_t, d_t$ while C_t does not have an independent embedded model, but is simply dependent on other embedded models.

(b) What are the simplest plausible embedded models for each unclosed quantity?

$$
\begin{aligned}
&Model\ Outputs \\
&N_t = 0 \\
&S_t = -\beta SI \\
&E_t = \beta SI - \sigma E \\
&I_t = \sigma E - \gamma I \\
&R_t = \gamma I \\
&Quantities\ of\ interest\ (QoI) \\
&D_t = f_D \gamma I, f_D < 1 \\
&d_t = f_d \gamma I, f_d < 1 \\
&C_t = \sigma E
\end{aligned}
$$

where *S,E,I,R,D,N,C* are as defined above in 2(a).

Exposure E intuitively increases in proportion to both S & I, hence we have the term βSI, where β is the transmission rate. σ, γ, λ are just constants of proportionality for

the various "source" and "sink" terms. S has a sink term which is a source term for E. E has a sink term too, which is a source term for I and so on.

In the *QoI*, d denotes the fraction of severe cases. D denotes the number of deaths. C is the cumulative number of cases. f_D, f_d are just constants of proportionality.

(c) What are the simplifications, assumptions and approximations that go into each embedded model?

We have assumed that the rates are only linearly proportional to the various values while in reality any such dependence is likely to be a complex non-linear function. We have assumed that the model parameters we have are just unknown constants and not stochastic variables. We have also only taken first order differential equations where there are likely higher order dynamics at work too.

One of the possible structural flaws in our model is that the recovered people may again be susceptible and our model does not account for that.

We are also looking at a very coarse picture. There are no small-scale dynamics, no inclusion for the possibility of local hotspots. One assumption which we implicitly make is that a deterministic set of equations can be used to model an inherently stochastic process which is not homogeneous/isotropic. For example, the pandemic is more likely to spread along routes with more traffic. Our model has no structures to account for this.

The solutions to our system are non-oscillatory (numerical solutions of these equations are non-oscillatory), meaning we are ignoring the inherent oscillatory trends that can be common in epidemics. In fact, our model might include them as part of the noise, which might not be true.

We are also assuming that there is no immunity via maternal anti-bodies (the M component of *MSEIR* models). There is no accounting for births and natural deaths. We have assumed no migration. There is no dependence on parameters reflecting the strength of the local healthcare i.e. number of healthcare workers, the quality of equipment, etc.

(d) Which of the assumptions and approximations are most questionable for the target model use cases?

One of the most questionable approximations is that the transmission rate β is a constant. Since most of the data we have is not from our county, we must take into account the effect of population density on the transmission rate. It does not take into account any measures the government might introduce even after the reopening (like compulsory use of gloves, masks) and also does not take the effect of public perception into account. A more fearful public is likely correlated to a lower transmission rate. A better model for β would be $\beta = f(\alpha, P, \rho)$, where α is a measure of government regulation, P is public

perception and ρ is population density. f is monotonically decreasing function with α, P but is a monotonically increasing function with ρ.

Another very questionable approximation is that the death rate is just directly proportional to the infection rate I. If the healthcare system gets overwhelmed, the death rate will definitely go up and hence a more reasonable function would be $D_t = g(I)$ where g is monotonically increasing when a certain value of I is breached. Similarly, the rate of increase of severe cases d_t would also be a function of I. Hence, $d_t = h(I)$ where h is monotonically increasing.

For higher fidelity, we might need constitutive models for these parameters to better capture the small scale dynamics at work here.

(e) What model parameters need to be set in each embedded model, and what is known about their values?

We need to set (infer) parameters such as $\beta, \gamma, \sigma, \lambda, f_d, f_D$. We know that all these parameters are positive. Not much is known about their values though. Since f_D, f_d are fractions they should be in the interval $(0, 1)$. We have some estimate about these values though from observations by health authorities. For example, we may know that f_D is 2%-3% as a general estimate. Something similar might be known about f_d as well.

(f) How will their a priori uncertainty be modeled?

For f_D, f_d we have a range of values from various sources. We can have some kind of distribution in this range. If we are really uncertain, it could be uniform.

For $\beta, \gamma, \sigma, \lambda$ we may choose a $\mathcal{U}(0, \infty)$ distribution since that would be the maximum entropy distribution in this case and not much else is known about these values. These values must vary wildly from disease to disease and the prior information from other diseases is certainly irrelevant here.

(g) Are the uncertainties aleatoric or epistemic for the target use of the model?

Our model inadequacies result in both epistemic and aleatoric uncertainties. Since we are not considering small-scale dynamics, our description of the phenomenon is incomplete, resulting in epistemic uncertainties. For example, our model will not model the oscillation in epidemiology trends. We are trying to model an inherently stochastic process with a deterministic system of equations, resulting in more epistemic uncertainties. Also, any stochastic process will mostly have aleatoric uncertainties built-in.

4. Model Inputs

(a) What inputs are required to complete the specification of the model (in addition to model parameters above)?

The model needs initial conditions of S,E,I,R,D,d,C to close the system of equations for a given set of parameters.

(b) How will they be determined for the scenarios of interest, and how will their uncertainty be modeled?

We will have to depend on our data on the last day of quarantine for our initial values. It is difficult to get accurate data for exposure. Any data we have for exposure is probably highly noisy. We might have to filter this data. The method of data collection matters and will be described in detail later on.

We will also have to assume that the data we have has tracked the number of infected people pretty accurately. This might not always be possible if the number of asymptomatic people is significant. If that is the case, it could make sense to split I into symptomatic and asymptomatic groups. We will have reliable data for the former, but the latter will have to be modeled based on more research about the fraction of people that stay asymptomatic.

We should have much more accurate data for D,d since these people come to the hospitals. Their values will have minimal uncertainties.

(c) Are the uncertainties aleatoric or epistemic for the target use of the model?]

It depends on how we have collected our data during the quarantine period. (Remember that the data collected during quarantine gives us the initial conditions for the model.)

Almost any method of data collection used will have aleatoric uncertainties, since no testing is 100% accurate, and it is very difficult to track all of the population for such a fast-changing situation.

Epistemic uncertainties might arise if you only test people with symptoms. This means there is no input data for asymptomatic people. On the other hand, if we collected many random samples all over the county, we might reduce epistemic uncertainties since you have a better representation of the population and also a very good estimate of the number of asymptomatic infected people.

5. Calibration

(a) For each embedded model that has parameters that need to be calibrated, what reliable data can be used to determine them?

You can calibrate your parameters using the data for S, I, R, D, d and the not-so-reliable data (generally) for E, which may have to be filtered.

(b) Under what scenarios would that data be obtained?

You cannot really experiment to collect your data. We want the data where the virus

spreads without being kept in check since we are modeling that. This can be done with the data from Wuhan, Italy, Iran, Spain and New York City and many more such areas. We also have the cruise ship data from various stranded cruises all over the world. We also want have data from Travis county and other counties before the quarantine was enforced.

(c) What is expected to dominate uncertainties in the data?

The method of data accumulation is crucial to calibration. If we only have the data of people who were symptomatic, it is a systematic flaw in our data collection and hence should increase how uncertain we are. Instead we could collect random samples of the population for better estimates.

The *Exposure* E is expected to dominate uncertainties in our data because it is very difficult to estimate/measure. It is very difficult to find how many people were exposed. You would have to independently model that and many factors such as range of infection, susceptibility, etc. would play a significant role. We might be able to get some noisy data on exposure though, that we could filter. For example, you know through contact tracing that a person X got the infection from person Y at a party where 100 people were present. This means it is possible that all 100 people were exposed. Or not even 20 people were exposed.

(d) If new measurements could be made, what opportunities might there be for optimizing scenarios?

We could maximize the Expected Information Gain with respect to the scenarios. Apriori, the results are difficult to estimate. We know, however, that the scenarios that lead to maximum gain are the same as the scenarios that our parameters can provide the maximum information on (Consequence of writing EIG in two different ways). Therefore, the new measurements should ideally be made on scenarios for which no data is easily available since they will inform our parameters the best. In our case, we will likely already have data from a variety of scenarios.

If we could make any kind of measurements, we should collect data in counties demographically similar to our county all over the US (and no restrictions enforced), where we should look to take a completely random sample of the population as a whole and not only those people who are symptomatic. This sampling would give us an indication of how widespread the disease actually is and give us some indication about the exposure.

6. Validation

(a) What reliable data can be used for validation, besides the calibration data above?

You can validate your parameters using the data for S, I, R, D, d and the not-so-reliable data (generally) for E, which may have to be filtered. We can find data from various

counties before the quarantine was enforced. We also have data from other countries and the cruise ships. Some of the data could be used for calibration and the rest for validation (these partitions could be randomly or strategically selected). Or if we are desperate to calibrate on as much data as possible, we may want to use a "leave-one-out" strategy, where one datum is excluded from the calibration, and used for validation. Each datum is then left out in turn.

(b) Under what scenarios would that data be obtained?

This question is not relevant to what we are doing in part 6(a) since it would be a repeat of what was mentioned in calibration.

But let us assume for a moment that any kind of data could be gathered for validation. Ideally, you would want data that questions the model assumptions and approximations.

We would like to validate on data with varying population densities, varying healthcare system quality, varying initial conditions, varying levels of migration as well as data from places where significant number of people get reinfected. We would like to validate on data which contains local hot-spots, or significant inhomogeneities, etc.

(c) What is expected to dominate uncertainties in the data?

The method of data accumulation is crucial to calibration. If we only have the data of people who were symptomatic, it is a systematic flaw in our data collection and hence should increase how uncertain we are. Instead we could collect random samples of the population for better estimates.

The *Exposure* E is expected to dominate uncertainties in our data because it is very difficult to estimate/measure. It is very difficult to find how many people were exposed. You would have to independently model that and many factors such as range of infection, susceptibility, etc. would play a significant role.

(d) How well would the validation cases test the questionable model assumptions described in (3)?

We can arrange this data obtained in 6(b), testing all our assumptions one at a time, and then in combinations through the "validation pyramid".

The lowest level would check one assumption/approximation at a time. For example, "For everything else held constant does the population density affect transmission rate?" This level should include checks for transmission rate being constant with respect to public perception, level of preventive measures (masks, etc.), population density. Similarly we should check if the healthcare quality affects d, D significantly. We should also do a sensitivity analysis at this level for checking if our QoI are sensitive to noisy data on exposure. We should do a sensitivity check of the QoI with respect to uncertainties in in the initial conditions. We should validate our assumption that very less people get

reinfected and it is fair to completely neglect this aspect. We should do a validation check for insensitivity to model inadequacies. It could be done by modeling the inadequacy using stochastic parameters and propagating it to the QoI. We should validate that small scale dynamics are not essential to trend of various parameters. The data for scenarios of local hotspots, or just otherwise inhomogeneous in some other way helps us validate this assumption. While doing these tests, we are checking various dependencies and have the opportunity to modify the functional forms based on what we find.

Then on the middle level of the pyramid, we validate cases with combinations of the things mentioned above and finally on the top level we validate for all of the assumptions and approximations together, i.e. pretty close to our actual prediction scenario.

Not sure if this is relevant here, but it is worth mentioning that we should do a comprehensive sensitivity analysis test to see if there are parameters that the calibration and validation observables were insensitive to (and it is thus informed only by our prior) but our QoI are sensitive to. These parameters have not been sufficiently informed by the data and need further inspection.

Overall, the diversity and amount of data we have should be able to adequately question all model assumptions.

(e) What are appropriate validation metrics?

We need to ensure that our model is entitled to make predictions. We have a hypothesis about the cause of our inadequacy. We can use our numerical solutions and make some mathematical arguments to see if the inadequacy is expected to be lower for our prediction case. (i.e. if our prediction case is a conservative case, the inadequacy for validation cases should also be applicable to the prediction case). Another thing to see is if the validation data is a credible outcome of the posterior distribution. Since we have both aleatoric and epistemic uncertainties in the system, we can use posterior predictive check for validation (where we marginalize epistemic uncertainties and deal with aleatoric uncertainties).

7. Concluding Remarks

I was slightly unsure (uncertain haha) about certain (many) sections since I have not explicitly though about how the Bayesian inverse problem's formulation will actually be done. Even without actually doing any inference though, the model is likely going to be inadequate for the task at hand given the enormous simplifying assumptions we have made. Reference (v) contains an example of a more sophisticated model that I derived some ideas from about what the basic SEIR model lacks.

8. References

i. Lecture Notes on VUQ by Dr. Robert D. Moser, UT Austin

ii. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

iii. https://www.indexmundi.com/blog/index.php/2018/03/05/how-many-babies-are-born-each-second-in-the-us/

iv. https://www.indexmundi.com/blog/index.php/2018/03/05/how-many-people-die-a-day-in-the-us/

v. https://doi.org/10.1016/J.IJID.2020.02.058

vi.