Prediction of Covid-19 Cases and Deaths in California

Grant Hutchings

Abstract

This analysis, with the help of our previous results [1] aims to produce a model that can accurately predict case count and deaths in California due to Covid-19. Unlike our previous analysis, we introduce uncertainty in the number of cases, n. We use predictions of n, together with predictions of mortality rate from our previous analysis to predict death at the county level. We compare predictions from two models with the goal of understanding the variability between counties. Model 1, a standard hierarchical model, utilizes a poisson likelihood on case count, a gamma prior on the poisson rate, and gamma hyperpriors. The poisson likelihood includes an exposure parameter to control for population size. Model 2 modifies model 1 by considering the population-normalized case count to also be dependent on mortality, which we include as an addition to the poisson exposure parameter.

Our analysis will focus on predicting the number of cases and deaths per 1000 habitants of each county. We will show that population normalization is more informative on the variability between counties than if we left the data unnormalized.

1. Data

Results in our previous analysis [1] were heavily influenced by counties with high populations like LA. Figure 1 illustrates how LA is an extreme outlier in both case count and total deaths. As we will be modeling the number of cases by county with a poisson exposure proportional to population, we are interested in the variability in the number of cases. Figure 2 shows how variability is much more apparent when cases and deaths are normalized by population. LA is still on the extreme end, but is much less of an outlier.

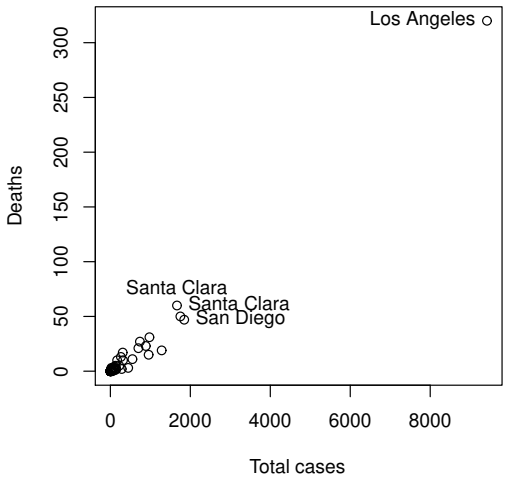

Figure 1: Total cases and total death by county. Most of the mass falls near the origin and variability is generally low.

Figure 3 shows only case data, and we see that nearly all the counties fall below .75 cases per 1000 habitants. The distribution tapers down smoothy and is much more informative than unnormalized data shown in figure 1. Mono county, by population, has been most impacted

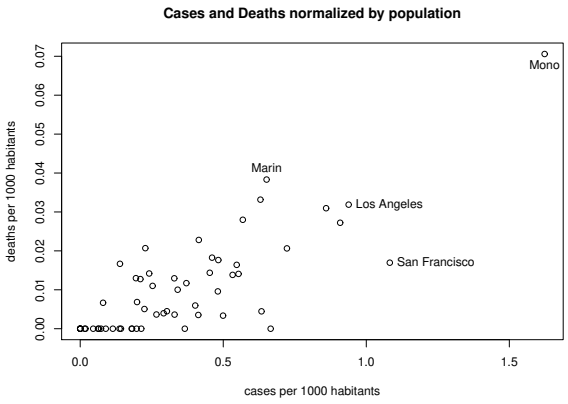

Figure 2: Cases and deaths normalized per 1000 habitants. Variability is much more visible, and outliers are less extreme.

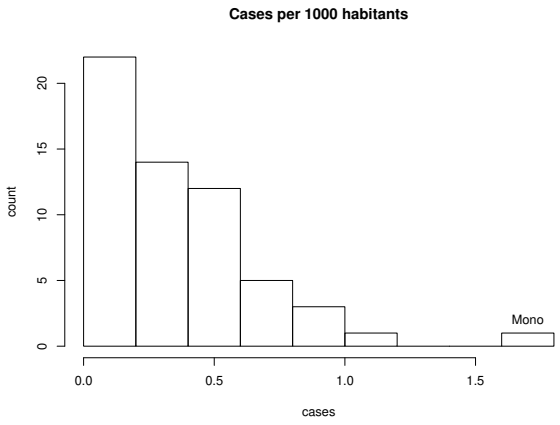

Figure 3: Cases per 1000 habitants by county. Mono county is labeled as having the most cases per person at 1.6, and LA is no longer an outlier with .94 cases per 1000.

by Covid-19 with 1.6 cases per 1000 habitants. We will continue our analysis based on this normalized case count by modeling the number of cases as an exposure weighted poisson with exposure of $population/1000$.

2. Hierarchical Poisson-Gamma Model

2.1 Methods

We fit a hierarchical model with an exposure weighted poisson likelihood, a gamma prior, and gamma hyperpriors.

$$n_i \sim Pois(\lambda_i c_i/10^3) \tag{1}$$

$$\lambda_i \sim Ga(\alpha, \beta) \tag{2}$$

$$p(\alpha, \beta) = Ga(\alpha|a, b)Ga(\beta|c, d) \tag{3}$$

where a, b, c, d are hyperparameters fixed so that, for each county, the mean number of cases is equal to 20% of the population. There are infinitely many choices for a, b, c, d that satisfy this condition. We choose values $a = 5 \times 10^{-4}, b = 1.6 \times 10^{-5}, c = 3.9 \times 10^{-4}, d = 5 \times 10^{-4}$. These values, while somewhat arbitrary were chosen to satisfy the above condition with the additional requirement that all values be less that 10^{-3}. This choice was made so that our prior distributions remain uninformative.

To obtain posterior samples, we factorize the joint posterior distribution.

$$p(\boldsymbol{\lambda}, \alpha, \beta|\boldsymbol{n}) = p(\boldsymbol{\lambda}|\alpha, \beta, \boldsymbol{n})p(\alpha, \beta|\boldsymbol{n}) \tag{4}$$

The full conditional of $\boldsymbol{\lambda}$ is

$$(\boldsymbol{\lambda}|\cdot) \sim Ga(\boldsymbol{\lambda}|\alpha + \boldsymbol{n}, \frac{\boldsymbol{c}}{10^3} + \beta) \tag{5}$$

where $\boldsymbol{n}$ is the vector of case counts, and $\boldsymbol{c}$ is the vector of populations. The full conditional for $\boldsymbol{\lambda}$ can therefore be sampled in Gibbs fashion using posterior samples of α, β.

The full conditional $(\alpha, \beta|\boldsymbol{n})$ is not available in closed form, and is sampled using a Sampling-Importance-Resampling algorithm with a multivariate-t proposal distribution centered at the posterior mode, with variance proportional to the curvature at the posterior

mode. Figure 4 shows the posterior contours of $p(\alpha, \beta|n)$ and samples obtained via SIR, which explore the distribution well.

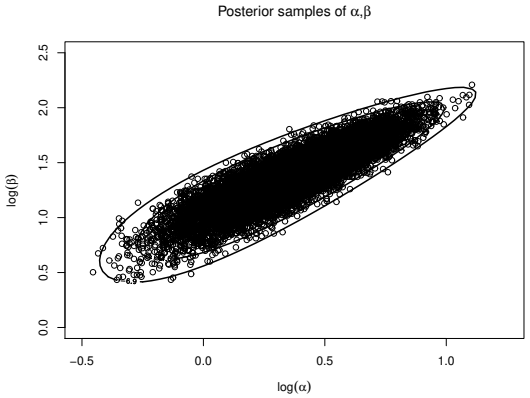

Figure 4: Posterior samples of (α, β) on log scale

2.2 Analysis

Having acquired S posterior samples of $\boldsymbol{\lambda}$, we can test the in-sample predictive accuracy of our model. The posterior predictive distribution for the number of cases in county i, is simply the likelihood defined in equation 1. For each county, we predict by sampling

$$n_i^{(s)} \sim Pois(\frac{c_i}{10^3}\lambda_i^{(s)}) \quad s = 1, ..., S \qquad (6)$$

Figure 5 shows in-sample predictions for the number of cases per 1000 habitants in each county, along with blue X's indicating the true value from the data. We can see that our predictions are very accurate. The true data value falls at or very near the median of our predictive distributions in almost all counties. As mentioned in [1], prediction for counties with zero cases and/or deaths is challenging. Uncertainty is higher, and prediction will usually be greater than zero, missing the true data value. This can be seen at the far right end of figure 5 where we consistently predict more than zero cases. Mono county is important to highlight here. We saw in figure 2 that Mono is an abnormally high in both cases and deaths per person. We see how this information makes it into the posterior distribution. Our model underestimates the cases in Mono county, and the uncertainty is very large.

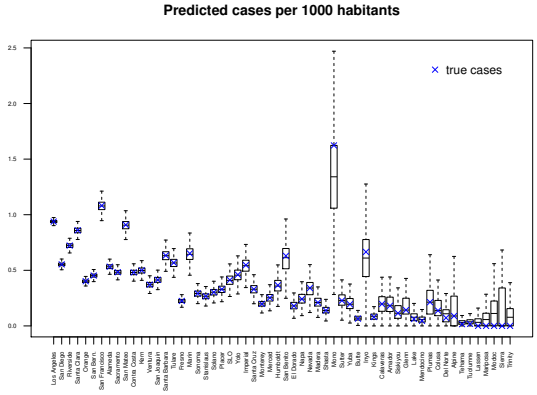

Figure 5: Predicted cases per 1000 habitants for all 58 counties in California. Hierarchical Poisson-Gamma model. Blue x indicates true value from data.

Given the in-sample accuracy of our predictions we feel confident using our posterior samples of n, along with our posterior samples of mortality rate $\boldsymbol{\theta}$ from [1] to predict the number of expected deaths per 1000 habitants. We assume a binomial likelihood for deaths

$$y_i \sim Bin(y_i|n_i, \theta_i). \qquad (7)$$

Figure 6 shows these distributions. We see that our predictive accuracy is still quite high, but our ability to predict death is not as good as our ability to predict cases. This is possibly due to the predictive distribution for cases depending on only one sampled parameter, seen

in equation 6, while the predictive distribution for deaths (equation 7) depends on two, introducing another level of uncertainty. The true data values are frequently close to the boundaries of the IQR rather than the median. Notice Mono county, which has an especially high number of deaths per person, is again underestimated in our model. This is not unexpected as the data are pulled towards something of a grand mean in the hierarchical model. Many counties had zero deaths or a very low case count, which is reflected in the right half of figure two. For counties with zero cases, we nearly always predict zero deaths because we are sampling from a binomial with a very small n and a very small probability of success. Variability is much higher for these counties. For counties with very low case counts, around the third quartile of the plot, variability is high because information in the data is limited. Counties like LA, with a lot of information in the data, have very low predictive variability. From this plot, it is not convincingly clear that there exists significant variation between the counties. An analysis of variance type procedure could possibly be implemented to test that hypothesis, but we will not pursue a quantitative answer to that question here.

This model is generally very good at in-sample prediction, and is not too complex in terms of structure and number of parameters.

3. Model 2

We now consider a modification of the previous model where the rate of infection depends on mortality.

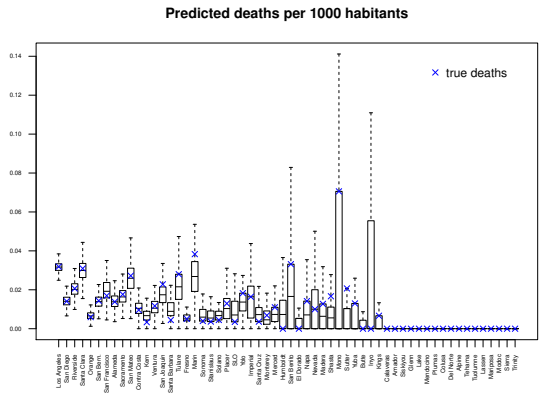

Figure 6: Predicted deaths per 1000 habitants for all 58 counties in California. Hierarchical Poisson-Gamma model. Blue X's indicates true value from data.

3.1 Methods

We define the likelihood in factorized form as

$$\begin{aligned} f(y_i, n_i|\theta_i, \lambda) &= f(y_i|n_i, \theta_i, \lambda) f(n_i|\theta_i, \lambda) \\ &= Bin(y_i|n_i, \theta_i) \times \\ &\quad Pois(n_i|\theta_i \lambda c_i/10^3) \\ &\quad \theta_i \in (0,1),\ \lambda > 0. \end{aligned} \tag{8}$$

We assume

$$\theta_i \sim Be(\theta_i|\mu\tau, (1-\mu)\tau) \tag{9}$$

$$\lambda \sim Ga(\lambda|a, b) \tag{10}$$

with prior

$$\begin{aligned} p(\mu, \tau) &= (\mu(1-\mu)(1+\tau)^2)^{-1}, \\ \mu &\in (0,1),\ \tau > 0 \end{aligned} \tag{11}$$

where a and b are fixed hyperparameters.

We again model the number of cases with an exposure weighted poisson, but this time our exposure depends on mortality as well as a common rate parameter λ, as opposed to county specific rates as in the previous model. We again take the likelihood of y_i to be a $Bin(n_i, \theta_i)$.

We choose the uninformative values $a = .001,\ b = .001$ for our hyperparameters. We will justify these hyperparameters in section 3.2.1 when we analyze predictive variability with respect to hyperparameter choice. We will also show that this model provides us with the ability to alter posterior estimates through prior knowledge via these parameters.

Using the formulation above, we can derive full conditionals for μ,τ,λ, and $\boldsymbol{\theta}$.

$$p(\mu|\cdot) \propto \prod_{i=1}^{N} B(\mu\tau,(1-\mu)\tau)^{-1}\theta_i^{\mu\tau-1} \times (1-\theta_i)^{(1-\mu)\tau-1}(\mu(1-\mu))^{-1} \tag{12}$$

$$p(\tau|\cdot) \propto \prod_{i=1}^{N} B(\mu\tau,(1-\mu)\tau)^{-1}\theta_i^{\mu\tau-1} \times (1-\theta_i)^{(1-\mu)\tau-1}(1+\tau)^{-2} \tag{13}$$

$$p(\boldsymbol{\theta}|\cdot) \propto \prod_{i=1}^{N} \theta_i^{y_i+n_i+\mu\tau-1} \times (1-\theta_i)^{n_i-y_i+(1-\mu)\tau-1} e^{\frac{-\theta_i\lambda c_i}{10^3}} \tag{14}$$

$$\lambda|\cdot \sim Ga\left(\lambda|\sum_{i=1}^{N} n_i + a, \sum_{i=1}^{N}\frac{\theta_i c_i}{10^3} + b\right) \tag{15}$$

We obtain posterior samples using a Metropolis within Gibbs algorithm on these full conditional distributions. The full conditional distribution of λ is in a convenient closed form, allowing Gibbs sampling, but the remaining parameters must be samples using a Metropolis step. Note that sampling from these conditionals requires samples of n_i and y_i. In the previous model, we obtained samples from the posterior predictive distributions of n_i, and y_i using samples of $\boldsymbol{\theta}$ from the hierarchical model in [1]. We use those samples here. Posterior chains for μ and τ exhibit quick convergence, but λ seemed to require a large number of iterations. To be safe we run our algorithm for 20,000 iterations, and burn the first 10,000 samples. Slow convergence of λ is shown in figure 7. We see that λ converges to 21.915+/-0.631 in standard error. This indicates that we expect the number of cases in county i to be about $22 \times c_i\theta_i/10^3$.

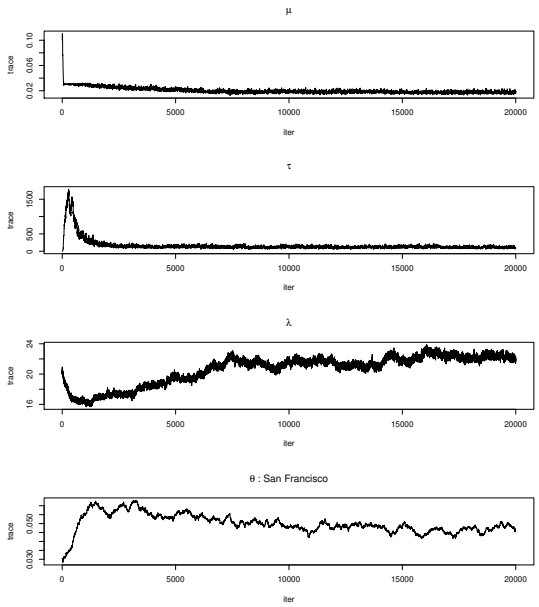

Figure 7: MCMC trace plots for $\mu, \tau, \lambda, \theta$ using San Francisco as an example of mortality rate convergence.

3.2 Analysis

With posterior samples of λ and $\boldsymbol{\theta}$ in hand, we turn our attention to the analysis of sampled mortality rate, and the posterior predictive distributions of n_i and y_i.

3.2.1 Mortality Rate

Figure 8 shows predicted mortality rate by county. We find that our model does relatively

poorly recovering the true data. This may be a related to the pooled rate parameter λ as opposed to a county specific rate λ_i. Figure 9 shows the residuals of these distributions against the true values. We do not see any clear over or underestimation trends. The only conclusion we make is that this model is not set up to make accurate predictions on mortality rate.

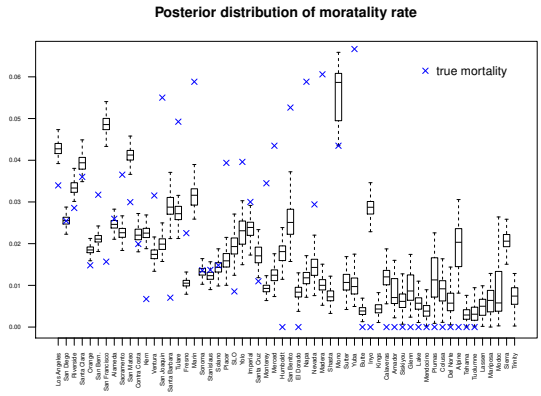

Figure 8: Posterior predictive distributions of mortality rate by county. True mortality rate labeled as blue X's. We see model 2 does poorly at in-sample prediction of mortality rate.

3.2.2 Predicting Cases and Deaths

Figures 10 and 11 show posterior predictive distributions for the number of cases and deaths per 1000 habitants by county. We are glad to see in-sample prediction much improved from that of mortality rate, however, we see that prediction is not as good as model 1 (this will be quantified in section 4 on model comparison). We again see very large uncertainty in the right half of the plots, indicating that counties with low information are hard to predict. Counties like Mono, which have abnormally high case and death counts are again underestimated with high

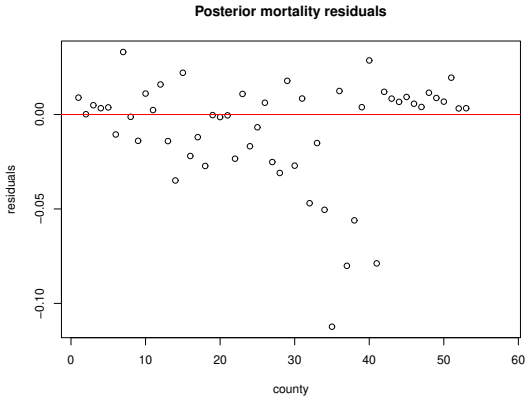

Figure 9: Residuals of posterior predicted mean mortality rate

uncertainty. We see a very similar uncertainty structure as in model 1, but the distributions tend to be slightly shifted, resulting in worse prediction.

A straightforward answer to why these predictions are worse than model 1 lies in figure 8. Predictions of n_i and y_i are based on samples of θ_i (see equation 8), so we are not surprised to see that the inaccuracies in mortality predictions have propagated to the predictive distributions of n_i and y_i.

In the following section we will show how these distributions change when hyperparameters are chosen differently.

3.2.3 Hyperparameter Choice

In equation 15 we saw the full conditional for λ is a gamma distribution with parameters $\sum_{i=1}^{N} n_i + a$ and $\sum_{i=1}^{N} \theta_i c_i/10^3 + b$ where $N = 58$ is the number of counties in California. We began by choosing $a = .001, b = .001$ as hyperpriors because they carry very little weight in the posterior distribution of λ as the summations are much larger. But clearly, as

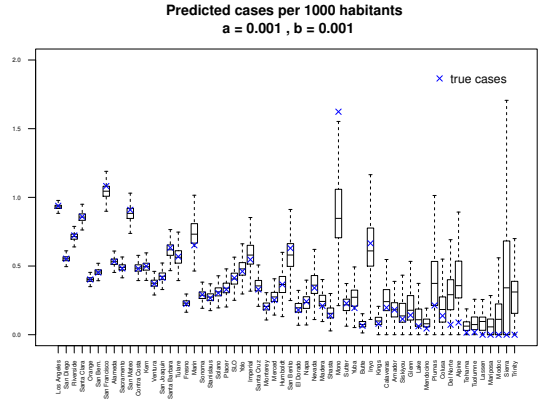

Figure 10: Predicted cases per 1000 habitants for all 58 counties in California. Non-informative hyperparameters $a = b = .001$. Blue x indicates true value from data.

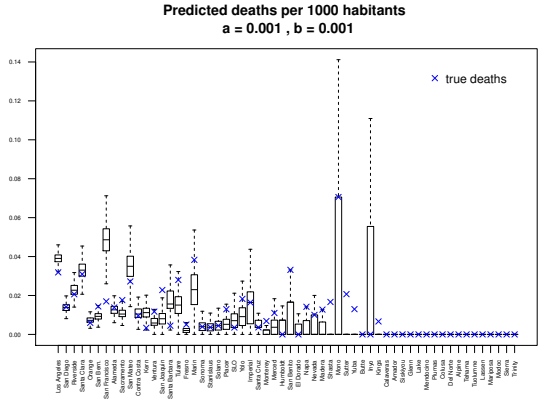

Figure 11: Predicted deaths per 1000 habitants for all 58 counties in California. Non-informative hyperparameters $a = b = .001$. Blue x indicates true value from data.

we increase a and b, they will begin to become more influential on our samples of λ, which will propagate to our other parameters in each metropolis step. We consider how our posterior predictions for deaths and cases change as we vary the hyperparameters. We consider three cases. In the first we vary a between .001 and 100, while fixing $b = .001$. In the second we similarly hold a fixed and vary b. and in the third, we consider varying a and b together, from 10^{-5} to 100.

· Case 1: vary a, fix b

In this case there is very little effect on posterior samples of λ. Increasing a occasionally causes slower convergence, but in the full range of $a \in (.001, 100)$, λ converges to the same posterior mean of about 22. Likewise, increasing a does not have a large effect on predicted cases or deaths except for some instabilities in counties with the lowest true case and death counts. These instabilities arise in the form of counties occasionally exhibiting large posterior uncertainty. Three counties in particular show this instability in predicted deaths; Inyo, Plumas, and Alpine, all of which are counties with zero recorded deaths. Sierra county, which has no cases and no deaths, shows this instability in predicted cases. We highlight the fact that these instabilities only arise in counties where the data provide very little information, i.e. very low or zero cases and/or deaths.

· Case 2: vary b, fix a

When we instead vary b, while holding a fixed at .001, we see catastrophic effects on posterior samples of λ. As we increase b, λ quickly tends to zero. This effect causes an increase in predicted deaths. This increase does not appear to be unbounded. Take San Francisco as an example. Non-informative hyperparameter choice leads to a prediction of .05 deaths per 1000 habitants. When we increase b to 20, this estimate increases to .17, and stays within the range of .14 to .18 for b up to 100. This effect

can be seen for many other counties as well. A similar, but opposite, and more gradual effect is seen in case count. Again, using San Francisco as an example of the trend, non-informative hyperparameters lead to an estimate of just over 1 case per 1000 habitants. Increasing b to 20 reduces that estimate to .8. As we increase b to 100, we steadily reduce our estimate to .5.

· Case 3: vary a, vary b

For the third case, our samples of λ appear unaffected while a, b are small, representing a lack of prior information. With uninformative hyperparameters, we get a posterior mean of about 22. When $a, b > 0.1$, posterior samples of λ begin to decrease steadily with a minimum posterior mean of about 1.5 when $a = b = 100$. Similarly, the posterior distributions for the number of deaths per 1000 habitants remains stable while a, b are in the uninformative range, say $a, b < 0.1$. As we increase a, b we start predicting many more deaths, often by as much as a factor 4. The opposite effect is seen in predicted case count per 1000 habitants. Stability is again seen in the uninformative range, but as we continue increasing a, b, we predict fewer and fewer cases. The implication here is that, if we increase our parameters outside of the uninformative range, but keep them equal, we predict a higher mortality rates.

Hyperparameter choice is clearly important for this model. We therefore feel justified in our choice of uninformative hyperparameters, but proceed with the knowledge that a, b can be adjusted to reflect prior information on mortality rate. We must however understand that inference on counties with low information may be subject to instabilities. The information presented in this section is challenging to visualize is a succinct manner. 12 plots showing these effects can be found on the project github page linked in the appendix.

4. Model Comparison

We have seen from figures 5,6 and 10,11 that model 1 seems to do better at in-sample prediction than model 2. We will now quantify this using the posterior predictive loss criterion developed by Gelfand and Ghosh in 1998 [2]. The criterion is defined as the sum of a "goodness of fit" term and a "penalty term". The goodness of fit term is the sum of squared deviations between replicated data means and observed data means, and the penalty term is the sum of the variance of replicated data.

$$\begin{aligned} D &= G + P; \\ G &= \sum_{s=1}^{S} (\mu_s - x_s)^2 \\ P &= \sum_{s=1}^{S} (\sigma_s^2) \end{aligned} \tag{16}$$

Table 1. shows values of D for 4 different situations. By row, the situations are; prediction of cases at the county level, prediction of the total number of cases in California, prediction of deaths at the county level, and prediction of the total number of deaths in California. We see that model 1 is better under this criterion in all but predicting total cases. This criterion generally favors model 1, but model 2 may still be useful. A metric such as total cases might be of interest to the state government of California or the federal government where high granularity prediction at the county level may not of interest. Local governments however, would be more interested in predictions from model 1, as they are superior at the county level. We have seen the most effective and proactive responses in managing covid cases from local governments, and

we would recommend they base predictions on model 1.

While it is beyond the scope of this analysis, there may be a way to dramatically improve model 2. In section 3.2.1, we showed that when we consider informative hyperpriors, we see substantial changes in the posterior predictive distributions. Therefore, there may exist a set of hyperparameter values that make model 2 consistently superior to model 1 in terms of posterior predictive loss. Searching for those hyperparameter values would be a good avenue for future work.

Table 1: Posterior predictive loss criteria for models 1 and 2. Calculated for cases and deaths both by county and total over all counties

	M1	M2
cases by county	0.756	3.122
total cases	875.322	756.520
deaths by county	0.013	0.027
total deaths	2.643	4.523

5. Conclusion

We have presented two fairly similar models for predicting Covid-19 cases and deaths in California. We have shown that model 1 is generally superior at in-sample prediction. Model 2 has trouble recovering true mortality rates, which in turn results in poor predictive performance. This is due to the dependence of the posterior predictive distributions on mortality rate. Model 1 bases predictions on better samples of mortality obtained in [1], and does not consider the case count to be dependent on mortality. We believe however that model 2 still has merit. With a more exhaustive modeling procedure, model 2 could be greatly improved through proper hyperparameter choice.

References

[1] Hutchings, G. (2020), "Analysis of Covid-19 cases in California Counties"

[2] ALAN E. GELFAND, SUJIT K. GHOSH, Model choice: A minimum posterior predictive loss approach, Biometrika, Volume 85, Issue 1, March 1998, Pages 1–11, https://doi.org/10.1093/biomet/85.1.1

6. Appendix

Images from section 3.2.1 on hyperparameter choice.
https://github.com/granthutchings/Covid19