Variant modeling proposal

Evan

October 1, 2024

1 Modeling motivations/ideas

1. For each variant, hierarchy across locations indicating that we expect trends to be similar across locations for the same variant.
2. hierarchy across variants, with the anchor point being the "start time" or introduction time for the variant. note, variants will generally be introduced to different locations at different times, but i think in practice we can probably ignore that and use the same start time across all locations, using different intercepts for different locations to handle this? i think modeling deltas in start times might be a pain and maybe not very identifiable. additionally, there is the issue that Maryclare brought up: if I haven't seen any sequences reported for a variant in a particular location yet, does that mean it's not there or just that reported counts are low?
3. a flexible trend in the relative growth advantage of novel variants over time. We saw consistent patterns of growth rates relative to a reference variant, with an initial phase of steady or possibly increasing values, then gradually decreasing. (We're not asserting why this is, could be due to model misspecification and/or actual biology...)
4. multinomial or dirichlet-multinomial observation process. to investigate whether this matters

2 Model

2.1 Notation

- Locations indexed by $l = 1, \ldots, L$
- Time indexed by t, with first data observed at time $t = 1$, current time is T, we will generate predictions in a window around T, $[T - h_{hind}, T + h_{fore}]$
- Variants indexed by $v = 1, \ldots, V$
- Observed data are counts of variants in each location and time point, $c_{l,t} := (c_{l,t,1}, \ldots, c_{l,t,V})$
  - Total sample size in location l and time t is $n_{l,t} := \sum_v c_{l,t,v}$. This is a nuisance parameter.

2.2 Observation Model

Given a parameter $\theta_{l,t} := (\theta_{l,t,1}, \ldots, \theta_{l,t,V})$ with each $\theta_{l,t,v} \geq 0$ and $\sum_v \theta_{l,t,v} = 1$ with (estimated) variant proportions at the population level, the observed data are modeled as a draw from a multinomial distribution. We condition on the observed sample size $n_{l,t}$:

$$C_{l,t} \mid \theta_{l,t}, n_{l,t} \sim \text{Multinomial}(\theta_{l,t}, n_{l,t})$$

As part of model development, we might check on calibration of predictions generated from this model. If predictive distributions are too narrow, we could update to use a Dirichlet-multinomial instead of multinomial.

2.3 Process Model

The proposed model is an exponential growth rate model with non-constant (relative) growth rates over time. Note that this means that we are accounting for any divergence from exponential growth via the time-varying growth rate. We could also shift to, e.g., a logistic growth model but that may not be necessary?

Let s_v denote the "start time" for variant v, defined to be the first time that it is observed in any location. We define the index $u_v := t - s_v$ to be the time since the introduction of variant v.

Let's use the notation $I_{l,t,v}$ to denote prevalent infections with variant v in location l at time t. To start, consider a single location l. For that location, we have an initial value of prevalent infections $I_{l,s_v,v}$ for variant v and $r_{l,v}(u)$ will be a function describing the exponential growth rate of variant v that is in effect for location l for the day u time units after its introduction. To simplify the calculations below, we'll work in discrete time. In other words, we will allow the value of $r_{l,v}(u)$ to change at integer values of u, holding the growth rate fixed throughout the course of each day but allowing it to be updated each day.

Let $\mathcal{A}_t$ be the collection of active variants at time t, i.e., those variants with $s_v \leq t$. For each variant in this set, we can find its total cumulative infections at time t by starting at the time of its introduction and working forward with exponential growth in each subsequent day:

$$I_{l,t,v} = \begin{cases} I_{l,s_v,v} \cdot \prod_{u=1}^{t-s_v} e^{r_{l,v}(u)} & \text{if } v \in \mathcal{A}_t \\ 0 & \text{otherwise} \end{cases},$$

where $r_{l,v}(u) \geq 0$. Another option could be

$$I_{l,t,v} = \begin{cases} I_{l,s_v,v} \cdot \prod_{u=1}^{t-s_v} [1 + e^{r_{l,v}(u)}] & \text{if } v \in \mathcal{A}_t \\ 0 & \text{otherwise} \end{cases},$$

allowing for $r_{l,v}(u) \in \mathbb{R}$, and in particular $r_{l,v}(u) < 0$. Alternatively, we can express this as follows

$$I_{l,t,v} = \begin{cases} 0 & t < s_v \\ I_{l,s_v,v} & t = s_v \\ I_{l,t-1,v} \cdot e^{r_{l,v}(t-s_v)} & t > s_v \end{cases}$$

We're in a world of unrealistic mechanistic models here and I'm not eager to go down mechanistic modeling rabbit holes, but:

- I think an MLR-style model would get to variant proportions by normalizing these cumulative incidence values. This says that anyone who was ever infected with the variant still has a chance of showing up in the sequenced cases for day t. Maybe that wasn't too implausible when fitting to a short window, but it seems broken when fitting to all data...
- More realistic would be to normalize the new incidence, right? $i_{l,t,v} = I_{l,t,v} - I_{l,t-1,v}$
- A next level of adding realism to the model would involve something saying that people remain infectious for some finite time and so the number of new infections at time t is not proportional to the cumulative count of all past infections. Here we'd be heading into a renewal equations setup, more of a full compartmental model setup. Probably this is another thing that a time-varying growth rate (that gets smaller as time since variant introduction rolls on) helps account for though...?

For now, I'm going to go with the intermediate route of realism and suggest the use of $i_{l,t,v}$. My sense is that mucking around with details of mechanistic setup is not a high-payoff activity; no matter what we do it will be wrong, and adding in statistical flexibility along the lines of time-varying growth will likely necessary for any mechanistic core...

Following the boxed discussion above, we arrive at the modeled variant proportions

$$\begin{aligned}
\theta_{l,t,v} &= \frac{i_{l,t,v}}{\sum_{v'=1}^{V} i_{l,t,v'}} \\
&= \frac{I_{l,t,v} - I_{l,t-1,v}}{\sum_{v'=1}^{V} (I_{l,t,v'} - I_{l,t-1,v'})} \\
&= \frac{I_{l,t-1,v}\left(e^{r_{l,v}(t-s_v)} - 1\right)}{\sum_{v'=1}^{V} I_{l,t-1,v'}\left(e^{r_{l,v'}(t-s_{v'})} - 1\right)},
\end{aligned}$$

where whenever $t = s_v$ or $t = s_{v'}$, we replace terms in the numerator and/or denominator with $I_{l,s_v,v}$ or $I_{l,s_{v'},v'}$ as necessary.

We can see how the variant proportions simplify when $r_{l,v}\,(t - s_v) = \beta_{l,v}$. Starting with the variant proportions at time $t = s_v$, we have:

$$\theta_{l,s_v,v} = \frac{I_{l,s_v,v}}{\sum_{v'=1}^{V} I_{l,s_v,v'}}.$$

Then at time $t = s_v + 1$, we have

$$\theta_{l,s_v,v} = \frac{(\exp\{\beta_{l,v}\} - 1)\, I_{l,s_v,v}}{\sum_{v'=1}^{V} (\exp\{\beta_{l,v'}\} - 1)\, I_{l,s_v,v'}}.$$

And at time $t = s_v + 2$, we have

$$\theta_{l,s_v,v} = \frac{(\exp\{\beta_{l,v}\} - 1)\exp\{\beta_{l,v}\}\, I_{l,s_v,v}}{\sum_{v'=1}^{V} (\exp\{\beta_{l,v'}\} - 1)\exp\{\beta_{l,v'}\}\, I_{l,s_v,v'}}.$$

Finally at time $t = s_v + k$, we have

$$\theta_{l,s_v,v} = \frac{(\exp\{\beta_{l,v}\} - 1)\exp\{\beta_{l,v}\,(k-1)\}\, I_{l,s_v,v}}{\sum_{v'=1}^{V} (\exp\{\beta_{l,v'}\} - 1)\exp\{\beta_{l,v'}\,(k-1)\}\, I_{l,s_v,v'}}.$$

Rearranging terms, we can see that this is equivalent to a multinomial linear regression model

$$\theta_{l,s_v,v} = \frac{\left(\frac{\exp\{\beta_{l,v}\}-1}{\exp\{\beta_{lv}\}}\right) I_{l,s_v,v}\exp\{\beta_{l,v}k\}}{\sum_{v'=1}^{V}\left(\frac{\exp\{\beta_{l,v'}\}-1}{\exp\{\beta_{l,v'}\}}\right) I_{l,s_v,v'}\exp\{\beta_{l,v'}k\}}.$$

with slope $\beta_{l,v}$ and intercept $\log\left(\frac{\exp\{\beta_{l,v}\}-1}{\exp\{\beta_{l,v}\}}\right) + \log\left(I_{l,s_v,v}\right)$.

> As an implementation detail, note that it may be more numerically stable to work on the logarithmic scale. In that case, a variety of intermediate representations are available and I'm not sure which would be most convenient. As an example, one option is to work with
>
> $$\log(I_{t,v}) = \begin{cases} \log(I_{s_v,v}) + \sum_{u=1}^{t-s_v} r_v(u) & \text{if } v \in \mathcal{A}_t \\ -\infty & \text{otherwise.} \end{cases}$$
>
> The value of $\log(i_{l,t,v})$ can be calculated stably in Stan using the log_diff_exp function.

The model is completed by giving the setup for the initial values $I_{l,s_v,v}$ and the growth rate functions $r_{l,v}(u)$. For the initial values, I propose something like the following hierarchical model on the log scale, where:

- On the log scale, for each variant v the initial values $I_{l,s_v,v}$ in different locations are distributed around a shared variant-specific initial level μ_v
- Variant-specific initial levels are distributed around a "grand mean initial level", μ.

$$\begin{aligned} \log(I_{l,s_v,v}) \mid \mu_v, \sigma_l &\sim \text{Normal}(\mu_v, \sigma_l) \text{ for each } l = 1, \ldots, L \\ \mu_v \mid \mu, \sigma_v &\sim \text{Normal}(\mu, \sigma_v) \text{ for each } v = 1, \ldots, V \\ \mu &\sim \text{Normal}(0, 3) \\ \sigma_l, \sigma_v &\sim \text{Half Cauchy} \end{aligned}$$

Notes:

- If we're early on in emergence of a new variant, it might be good to add more structure here? e.g., for locations that haven't yet reported anything for the new variant, presumably the value of $I_{l,s_v,v}$ will be lower than for locations that have reported presence of that variant?
- The use of two levels of hierarchy here may be unnecessary; we could start with just one level, i.e. $\log(I_{l,s_v,v}) \mid \mu, \sigma \sim \text{Normal}(\mu, \sigma)$.
- I have no justification for choosing a prior standard deviation of 3 for μ
- As I was thinking this through, I had a question about how population size would interact with this. But you could imagine multiplying by population size and then it would cancel out when getting to θ, right?

For the growth rate functions I propose the use of some hierarchical splines [but I have some concern that there may be issues with identifiability in what I suggest here]. Specifically, we could set

$$r_{l,v}(u) = X\beta_{l,v}$$

where X is a B-spline basis over the time since variant introduction, and the coefficients $\beta_{l,v}$ are given a hierarchical prior with similar structure to the above. Note that $\beta_{l,v} := (\beta_{l,v,1}, \ldots, \beta_{l,v,J})$ is a vector of length equal to the number of spline basis functions, J. We could adopt the following priors (independently of j, i.e. allowing the model to learn temporal structure over the spline coefficients from the data without imposing time structure via the priors??):

$$\begin{aligned} \log(\beta_{l,v,j}) \mid \beta_{v,j}, \xi_l &\sim \text{Normal}(\log(\beta_{v,j}), \xi_l) \\ \log(\beta_{v,j}) \mid \beta_j, \xi_v &\sim \text{Normal}(\log(\beta_j), \xi_v) \\ \log(\beta_j) \mid \beta, \xi_j &\sim \text{Normal}(\log(\beta), \xi_j) \end{aligned}$$

I have some question about whether the model could learn to correctly attribute variability in the coefficients $\beta_{l,v,j}$ across the hierarchy levels for variants v and locations l? Maybe it doesn't matter?

Or maybe some empirical Bayes thing could be done to get some priors that encode information about the relative magnitudes of ξ_l, ξ_v, and ξ_j? As with the initial level parameters, we might start by dropping the two levels of hierarchy across v and l to get to something simpler and potentially easier to fit...

3 Other Models

3.1 Linear Hierarchical Multinomial Logistic Regression (MLR)

3.1.1 Notation

- Let the locations be indexed by $l = 1 \dots L$
- Let time be indexed by t, normalized such that the first day of the dataset is $t = 1$. If we let the current time be T, we are interested in predicting over $[T - 31, T + 10]$
- Let the clades be indexed by $v = 1 \dots V$
- Let $c_{l,t} = (c_{l,t,1}, \dots c_{l,t,V})$ be the observed virus counts for each clade at the lth location and the tth time. Then $n_t = \sum_v c_{l,t,v}$ is the total observed counts for time t and location v, as noted above n_t will not be known for our predictions and it is a nuisance parameter.

3.1.2 Current Model

We define intercepts α and slopes β

$$\alpha = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1V} \\ \vdots & \ddots & \vdots \\ \alpha_{L1} & \dots & \alpha_{LV} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_{11} & \dots & \beta_{1V} \\ \vdots & \ddots & \vdots \\ \beta_{L1} & \dots & \beta_{LV} \end{pmatrix},$$

define the vectors of intercepts and slopes associated with location l as $\boldsymbol{\alpha}_{l+} = (\alpha_{l1}, \dots, \alpha_{lV})$ and $\boldsymbol{\beta}_{l+} = (\beta_{l1}, \dots, \beta_{lV})$, and define the vectors of intercepts and slopes associated with variant v as $\boldsymbol{\alpha}_{+v} = (\alpha_{1v}, \dots, \alpha_{Lv})$, $\boldsymbol{\beta}_{+v} = (\beta_{1v}, \dots, \beta_{Lv})$.

For this model, we assume that the case counts for the lth location and tth time conditional on n_t are linear in the logit with respect to time with

$$C_{1t1}, \dots C_{ltV} \mid \alpha_l, \beta_l, n_{lt} \sim \text{Multinomial}(\frac{\exp(\alpha_{lv} + \beta_{lv} t)}{\sum_v \exp(\alpha_{lv} + \beta_{lv} t)}, n_{lt})$$

We also assume that the $\boldsymbol{\alpha}_{+v}$ and $\boldsymbol{\beta}_{+v}$ are draws from distributions, i.e we think that the same variant in different locations will have similar trajectories. Based on fitting MLR models to each location individually, this seems like a reasonably true assumption, at least for the current time. At present, we have the following prior structure:

$$\begin{aligned} \boldsymbol{\alpha}_{l+} | \boldsymbol{\mu}_\alpha, \tau_\alpha^2 &\sim \text{Normal}(\boldsymbol{\mu}_\alpha, \tau_\alpha^2 I_V) \text{ for each } l = 1, \dots, L \\ \boldsymbol{\beta}_{l+} \mid \boldsymbol{\mu}_\beta, \tau_\beta^2 &\sim \text{Normal}(\boldsymbol{\mu}_\beta, \tau_\beta^2 I_V) \text{ for } l = 1, \dots, L \\ \boldsymbol{\mu}_\alpha &\sim \text{Normal}(\mathbf{0}, 0.2 \times I_V) \\ \boldsymbol{\mu}_\beta &\sim \text{Normal}(\mathbf{0}, 0.2 \times I_V) \\ \tau_\alpha^2 &\sim \text{TruncatedNormal}(1, 0.1, 0, \infty) \\ \tau_\beta^2 &\sim \text{TruncatedNormal}(1, 0.1, 0, \infty) \end{aligned}$$

Note here $bloc = (bloc_1, \dots bloc_V)$ and bsd is a scalar that is greater than 0. This prior structure is arbitrary; mostly based off the prior structure that the Seattle Flu alliance used. The prior for α_v feels a little wide to me, (most α are negative numbers between -1 and -10). Conversely, the priors for $bloc$ and bsd seem very strong, stronger then I would probably pick. If we want to use this model for actual nowcasting and forecasting, we will probably want to use different priors.

3.1.3 Other Thoughts

- Generally, when fitting this model we choose one variant to be the reference variant to allow the parameters to be identifiable. When fitting this model, I use an other clade made up of all the clades with low case counts in the recent days. I don't know how this would work for model submissions, probably doesn't matter because we are submitting samples which should be the same regardless of reference variant.
- This model generally preforms better in terms of energy score compared to just fitting a MLR model to each location individually. This may be because the way this model is fit allows for information sharing between locations, allowing for more accurate predictions in locations with little data.
- That being said the information sharing is quite weak, because it does not make use of geographical information.

3.2 Hierarchical MLR with Splines

3.2.1 Notation

We use all the same notation from the linear hierarchical model with the following additions:

- Let $f_{l,v}(t)$ be a spline for the lth location and vth clade evaluated at time t.
- let $b = 1, \dots B$ be the basis functions for the splines.

3.2.2 Current Model

This model has the same form as the linear form as the linear model but with the linear time term replaced with a spline. Note in the way the model is currently written the spline does not have have an intercept. The model form for the lth location at the t time is as follows:

$$C_{l,t,1}, \dots C_{l,t,V} \mid \alpha_l, f_l, n_{l,t} \sim \text{Multinomial}(\frac{\exp\left(\alpha_{l,v} + f_{l,v}(t)\right)}{\sum_v \exp\left(\alpha_{l,v} + f_{l,v}(t)\right)}, n_{l,t})$$

This model uses the same prior structure as the linear Heir MLR model, if we let $\beta_{v,b} = c(\beta_{v,b,1} \dots \beta_{v,b,L})$ be the vector of the bth basis coefficient for the vth variant, then we have

$$\begin{aligned}
\beta_{v,b} \mid bsd, bloc \sim & \text{Normal}(bloc_v, bsd) \text{ for each } v = 1, \dots, V \text{ and } b = 1, \dots, B \\
& \alpha_v \sim \text{Normal}(0, 6) \text{ for each } v = 1, \dots, V \\
& bloc \sim \text{Normal}(0, 0.2) \\
& bsd \sim \text{Normal}(1, 0.1)
\end{aligned}$$

These priors have the same idea as in the linear model that each variant should be similar across locations, but what priors to use in this case are even less clear than in the linear case. These priors should be thought of more as placeholders than true values, and even the hierarchical structure should not be thought of as fixed.

3.2.3 Other Thoughts

- The default number of basis functions for this model is 3, using a higher number of basis functions may produce a better fit, but we do need to be wary of the lack of data and over fitting.
- For the few models that I have tried, the cubic spline Heir MLR model and the linear Heir MLR model have had very similar probabilities and energy scores. This could suggest that fitting a linear model is reasonable in this case or perhaps that the number of basis functions is not high enough to show a difference.