Backfitting Optimization

Let y_t be the observed cases at time t, and x_t be the time index. We have a set of convolution matrices $C_1, \dots, C_K$, $k = 1, \dots, K$ such that each corresponds to a different variant. These have the property that $C_1 + \cdots + C_K$. Each row of C_k is a convolution of the t-specific reporting delay (common across variants), with a variant-specific infection-symptom delay (common across time). Finally, we multiply each row by the variant-specific circulation proportion (smoothed) associated to that variant.

The problem is to estimate the variant-specific deconvolved cases simultaneously using trend filtering. We write this as:

$$\min_{\theta_1,\dots,\theta_K} \frac{1}{2}\|y - \sum_{k=1}^{K} C_k\theta_k\|_2^2 + \lambda \sum_{k=1}^{K} \|D^{(4)}\theta_k\|_1.$$

There is also a constraint that $\theta_{tk} \geq 0$ for all k.

Coupled with a small ridge penalty, the current backfitting code gives Figure 1 with $\lambda = 10^4$ which is somewhat smooth. This was 25 backfitting iterates with 200 ADMM iterates per-variant per backfitting pass.

The problems here seem to be:

1. The ordering of the variants is alphabetical, but this bungles the backfitting.
2. Total cases is 1.21×10^7 compared to 1.71×10^7 for deconvolved cases.
3. The nonnegativity constraint forces some weird behavior, though might disappear with a fix for the ordering.

The current ADMM steps (single variant) are:

1. Solve the Least Squares problem for θ_k using a QR (+small ridge penalty).
2. Hard threshold the solution to nonnegativity.
3. Use the DP.
4. Update the dual variable.

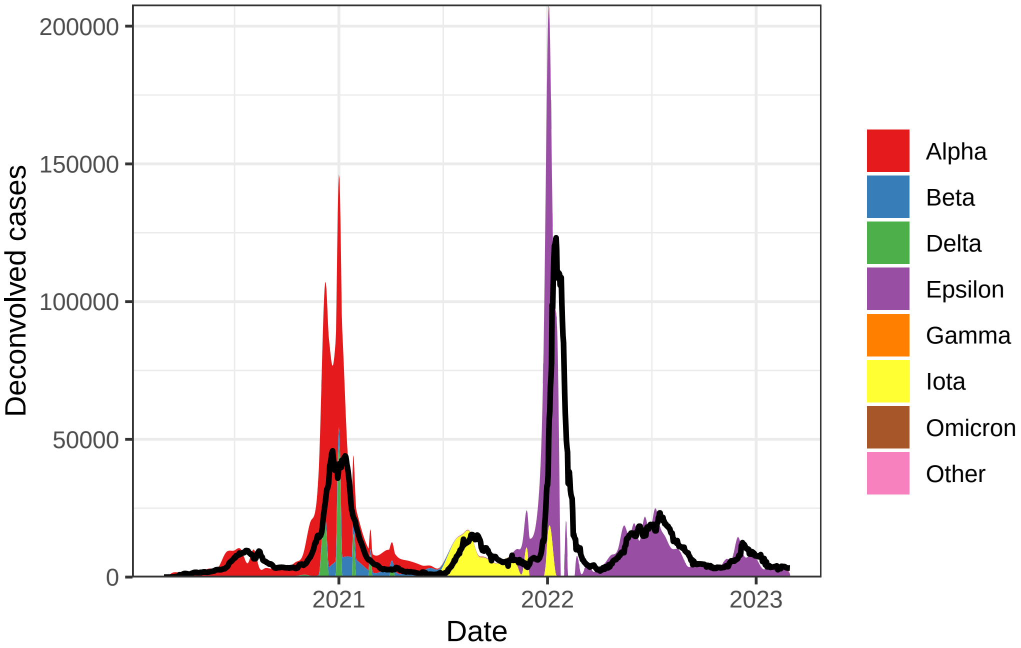


Figure 1: Deconvolved cases for California (backfitting)

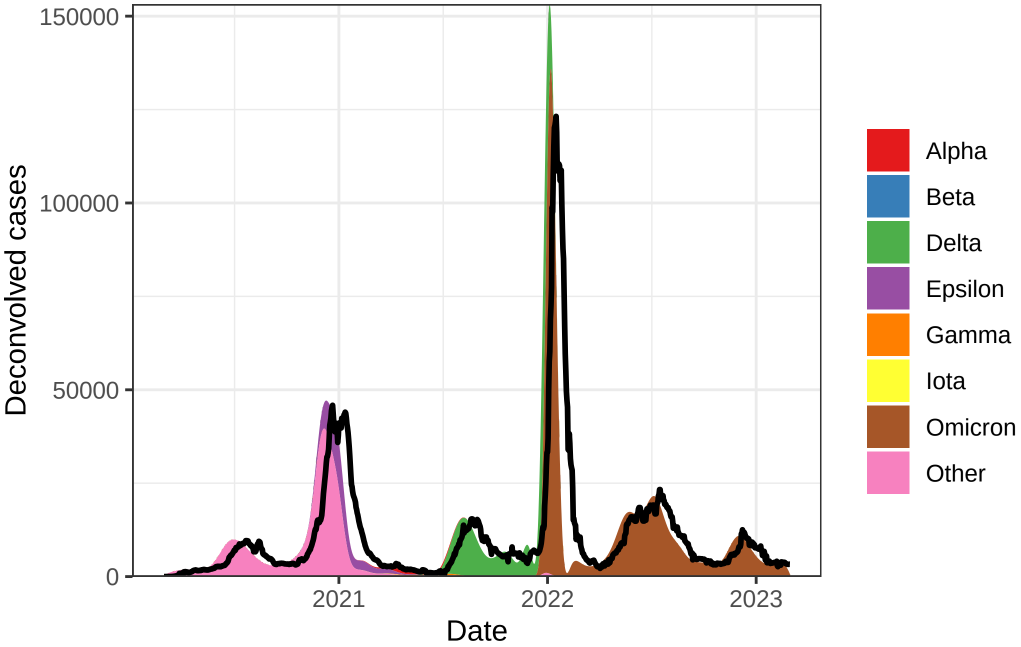
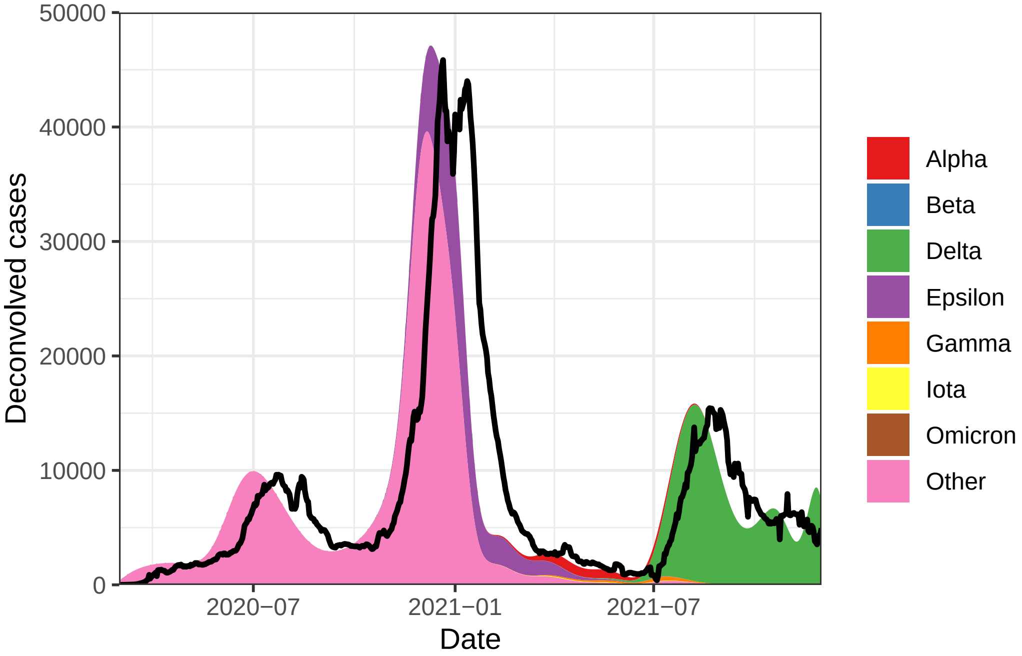


Figure 2: Deconvolved cases for California (easy way)

Redoing the figure simply multiplying deconvolved cases by the proportion of the variant in circulation gives Figure 2.

And Figure 3 is the same as Figure 2, but zoomed in to our period


Deconvolved cases
Date
Alpha
Beta
Delta
Epsilon
Gamma
Iota
Omicron
Other