Efficient Testing for Viral Kinetics Modeling

Project Update

Violet Ross

Github

Check out code, figures, and a written project update at:

https://github.com/Violet-Ross/efficient-sampling

Be warned: rough edges

Fitting VK Models is Expensive

Problem: To fit a model of within-host viral kinetics, you need to measure viral load regularly from infection time to clearance time in many individuals.

1. Finding infections: Finding an infected individual is expensive. (It is even more expensive to find such an individual before they get infected, but this is necessary for the assumed modeling approach.)
2. Learning from infections: Current models take in regular (e.g. daily) measurements of viral RNA concentration and output a fit model. However, taking daily measurements of viral load from an infected individual is expensive.

Goal: Reduce cost!

Sampling and Testing

- **sampling**: spit in tube
- **testing**: measure the concentration of viral RNA in the tube of spit
- they both cost time, money, etc
- right now: we assume the sampling has been done, and we try to *reduce the total testing cost*
- future goal: inform sampling efforts

Finding Infections

Problem Setup and Brute Force

Our Approach

Given: matrix of samples coming from m people over n days.

Assume: Each infection i has length ℓ_i drawn from some continuous positive distribution where $\ell_i > \ell_0$ with probability 0.95.

For each person:

1. Combine all n samples into a single pool and test that pool.
2. If the test in step 1 is positive, then individually test that person's samples in increments of ℓ_0. That is, test time 0, time ℓ_0, time $2\ell_0$, etc.

Cost Reduction

Applying our approach to a matrix of m individuals over n days with cumulative incidence c incurs cost

$$\begin{aligned}
\text{cost} &= \begin{pmatrix} \text{num} \\ \text{people} \end{pmatrix} \begin{pmatrix} \text{cost per} \\ \text{person} \end{pmatrix} + \begin{pmatrix} \text{num positive} \\ \text{people} \end{pmatrix} \begin{pmatrix} \text{cost per positive} \\ \text{person} \end{pmatrix} \\
&= m(1) + cm\frac{n}{\ell_0} \\
&= mn\left(\frac{1}{n} + \frac{c}{\ell_0}\right)
\end{aligned}$$

Accuracy Guarantees

Example: Exponential Distribution

Suppose $\ell_i \sim \text{Expo}(\lambda)$.

$$\begin{aligned} P(\text{miss traj } i) &= 0.05 - \frac{1}{\ell_0}\int_0^{\ell_0} P(\ell_i)\ell_i d\ell_i \\ &= 0.05 - \frac{1}{\ell_0}\int_0^{\ell_0} \lambda e^{-\lambda \ell_i}\ell_i d\ell_i \\ &\vdots \\ &= 0.05 + e^{-\lambda\ell_0} + \frac{1}{\ell_0\lambda}e^{-\lambda\ell_0} - \frac{1}{\ell_0\lambda} \end{aligned}$$

Recall our guarantee that $\ell_i > \ell_0$ with probability 0.95. In order to achieve this, we set

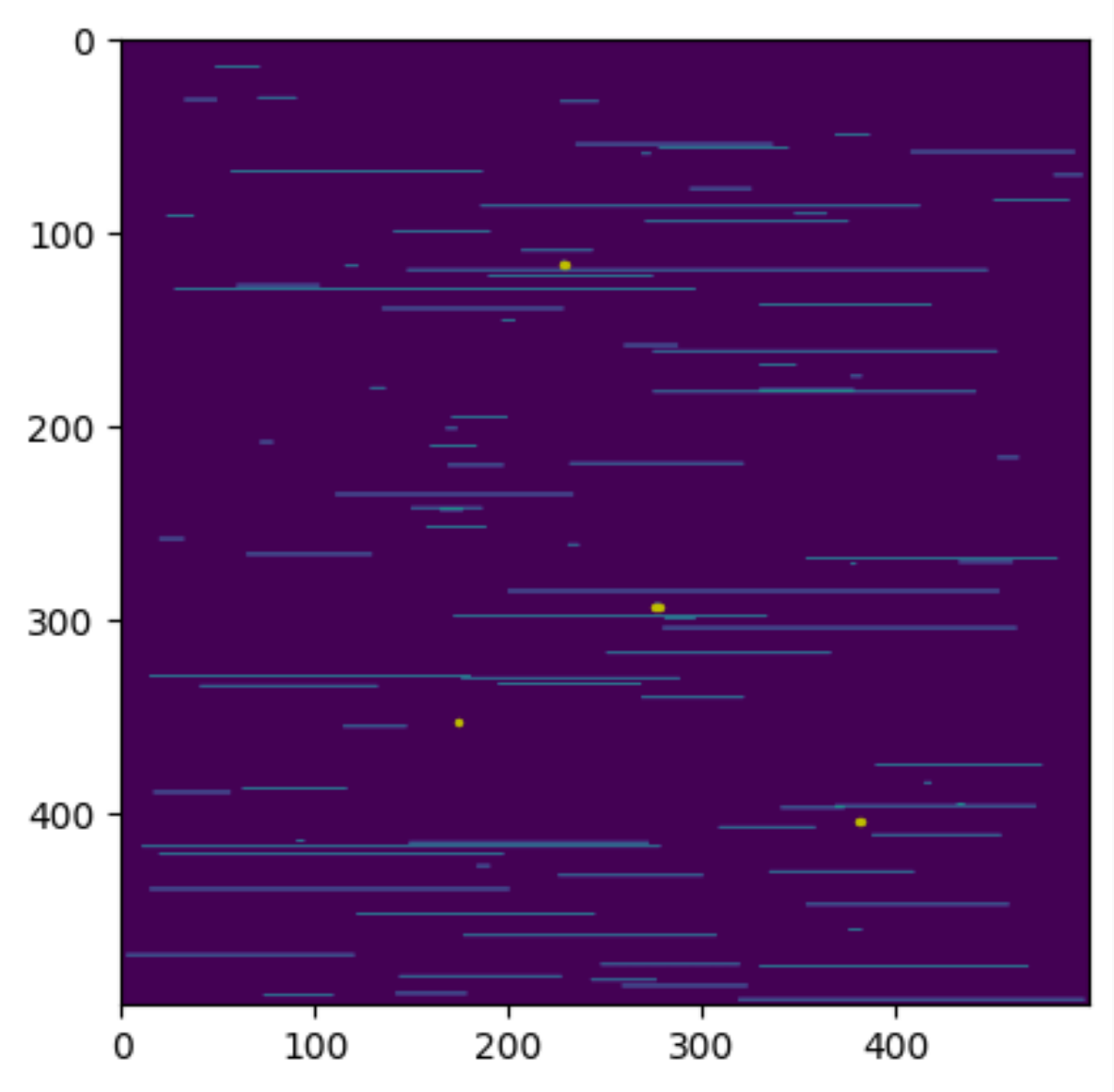
$$\lambda = -\ln(0.95)/\ell_0.$$

Substituting this into our equation for $P(\text{miss traj } i)$, we get

$$P(\text{miss traj } i) \approx 0.0252^{1}.$$

[1] What kinds of trajectories do we miss?

Simulation: Exponential Distribution



Exponentially distributed duration of infection with $\lambda = -\ln(0.95)/4$ and a cumulative incidence of 0.2.

Learning from Infections

Problem Setup

Segmented Regression

We'll model VK using a two-phase segmented regression model ("tent function")

$$y = \alpha + \beta_1 x + \beta_2 x \mathbb{1}_{\{x>\psi\}} + \varepsilon$$

Brute Force
Uniform Random Sampling

Variance Minimization

Think back to simple linear regression.

$$y = \alpha + \beta_1 x + \varepsilon$$

The variance of the slope estimate is

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.$$

So the more spread out our x_i, the less variable our β_1 estimator.

Variance Minimization

Segmented regression model:

$$y = \alpha + \beta_1 x + \beta_2 x \mathbb{1}_{\{x>\psi\}} + \varepsilon$$

The variances of the β estimates are:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i:x_i<\psi}(x_i - \bar{x}_-)^2}$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i:x_i<\psi}(x_i - \bar{x}_-)^2} + \frac{\sigma^2}{\sum_{i:x_i>\psi}(x_i - \bar{x}_+)^2}$$

Let's choose the combination of m points that minimizes

$$\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_1) = 2\frac{\sigma^2}{\sum_{i:x_i<\psi}(x_i - \bar{x}_-)^2} + \frac{\sigma^2}{\sum_{i:x_i>\psi}(x_i - \bar{x}_+)^2}.$$

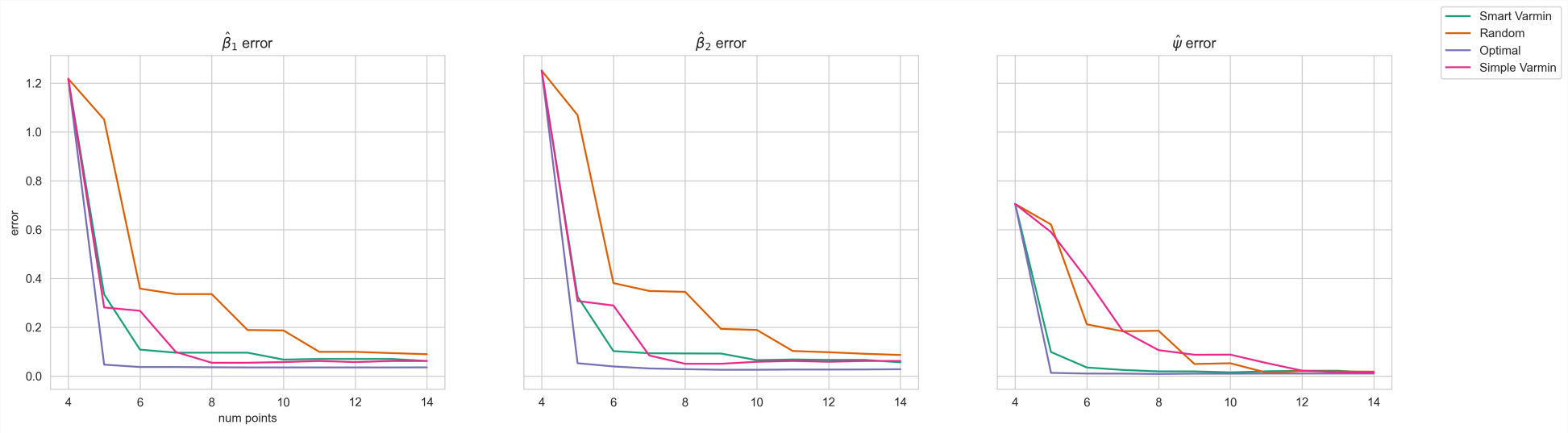
Variance Minimization

Algorithm 1 Smart Varmin Sampling

Input:
array of sample timestamps S where $S[i] = [\text{time of } i\text{th sample}]$,
number of samples to measure $m \geq 4$,
stepsize for breakpoint estimation c
Output: estimates $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\psi})$

$P \leftarrow (S[0],\ S[\text{len}(S)//3],\ S[(\text{len}(S)//3) * 2],\ S[-1])$
$M \leftarrow \text{MEASURE\_ALL}(P)$
$\hat{\alpha}^{(0)}, \hat{\beta}_1^{(0)}, \hat{\beta}_2^{(0)}, b^{(0)} \leftarrow \text{ESTIMATE}(M, c)$
$b \leftarrow b^{(0)}$
min_var $\leftarrow \infty$
for i in range(0, m - 4) **do**
    for s in $S \setminus P$ **do**
        combo $\leftarrow$ append s to P
        slope_var $\leftarrow \text{COMPUTE\_SLOPE\_VARIANCE}(b, \text{combo})$
        if slope_var $<$ min_var **then**
            min_var $\leftarrow$ slope_var
            optimal_combo $\leftarrow$ combo
    $P \leftarrow$ optimal_combo
    $M \leftarrow \text{MEASURE\_ALL}(P)$
    $\hat{\alpha}^{(i+1)}, \hat{\beta}_1^{(i+1)}, \hat{\beta}_2^{(i+1)}, b^{(i+1)} \leftarrow \text{ESTIMATE}(M, c)$
    $b \leftarrow b^{(i)}$
return $\text{ESTIMATE}(P, c)$

Simulation


$\hat{\beta}_1$ error
$\hat{\beta}_2$ error
$\hat{\psi}$ error
error
num points
Smart Varmin
Random
Optimal
Simple Varmin

Looking Forward

Progress

- Developed an efficient and accurate approaches for Phases 1 (finding an infection) and 2 (fitting a model to an infection).
- Submitted a GRFP on this topic.

Next Steps

- In Phase 1, we locate an infection. In Phase 2, we skip a step and assume that the start and end times of the infection are known. Can we **fit a segmented regression model starting only with a single positive point?**
- Right now, our approaches are designed to learn a model of a single trajectory. Can we **design a (Bayesian Hierarchical) approach which learns about the distributions from which the trajectories are generated?** That is, can we say something about the disease dynamics in general, rather than just discretely describing dynamics within individuals?
    - This also helps us move from descriptive modeling land to predictive modeling land
- Can we **translate these mathematical results into an explicit approach** to collecting and processing data for within-host viral kinetics research?
    - Open-source software?

Feedback?

Questions?