

# Estimating Mean Viral Load Trajectory from Intermittent Longitudinal Data and Unknown Time Origins

Yonatan Woodbridge, Micha Mandel, Yair Goldberg and Amit Huppert

May 29, 2024

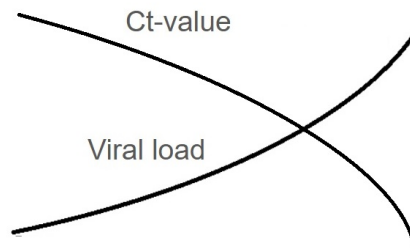
# Some background

## Viral load

- How many viruses does an infected person host?
- The amount of viral nucleic acid within the host, expressed as the number of viral particles in a given volume

## Ct-value

- How is the viral load measured?
- Cycle-threshold value: the number of viral nucleic acid replications required for detection
- The lower the Ct-value, the higher the viral load

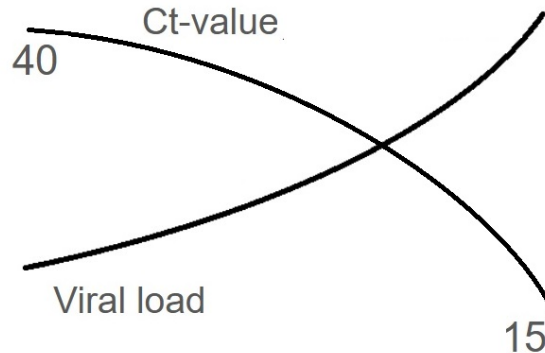


# Some background

## Ct-value in SARS-Cov-2

- Measured on nasal samples (as in PCR)
- Bounded by 15 (high viral load) and 40 (no infection)

Viral load estimation  $\iff$  Ct-value estimation



# Main motivation

## Viral load trajectory

- How does the viral load change over time from infection?
- Should increase (at early stage of infection), and then decrease (following host's immune response)

## Importance

- Viral load trajectory can help predict:
  - Rates of infectiousness (SIR model)
  - Generation time
  - Disease duration

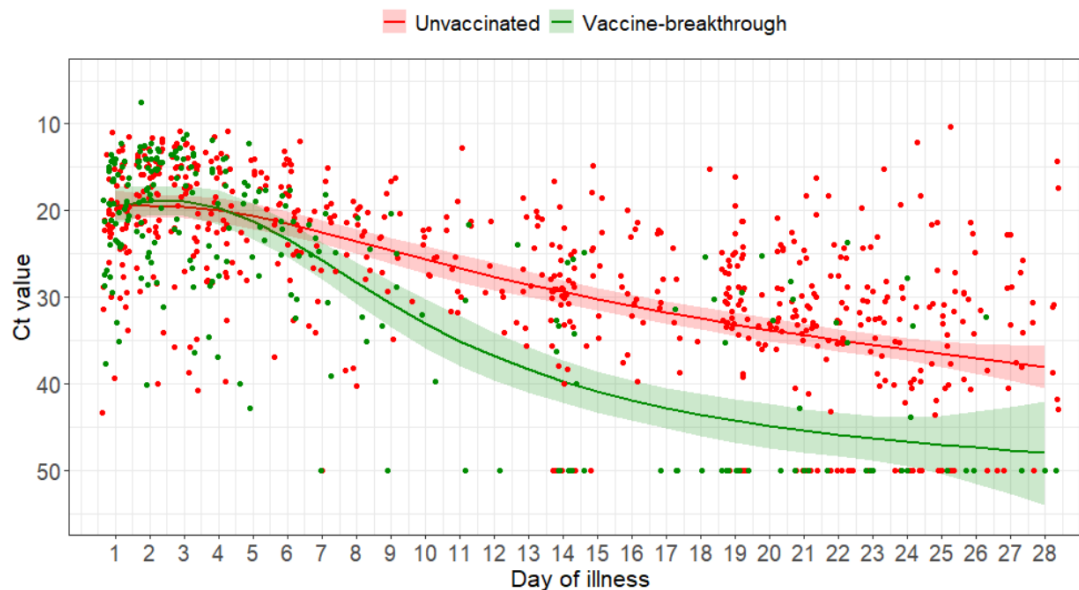
## Goal

How can we reconstruct the viral load trajectory?

# SARS-Cov-2 viral load trajectory

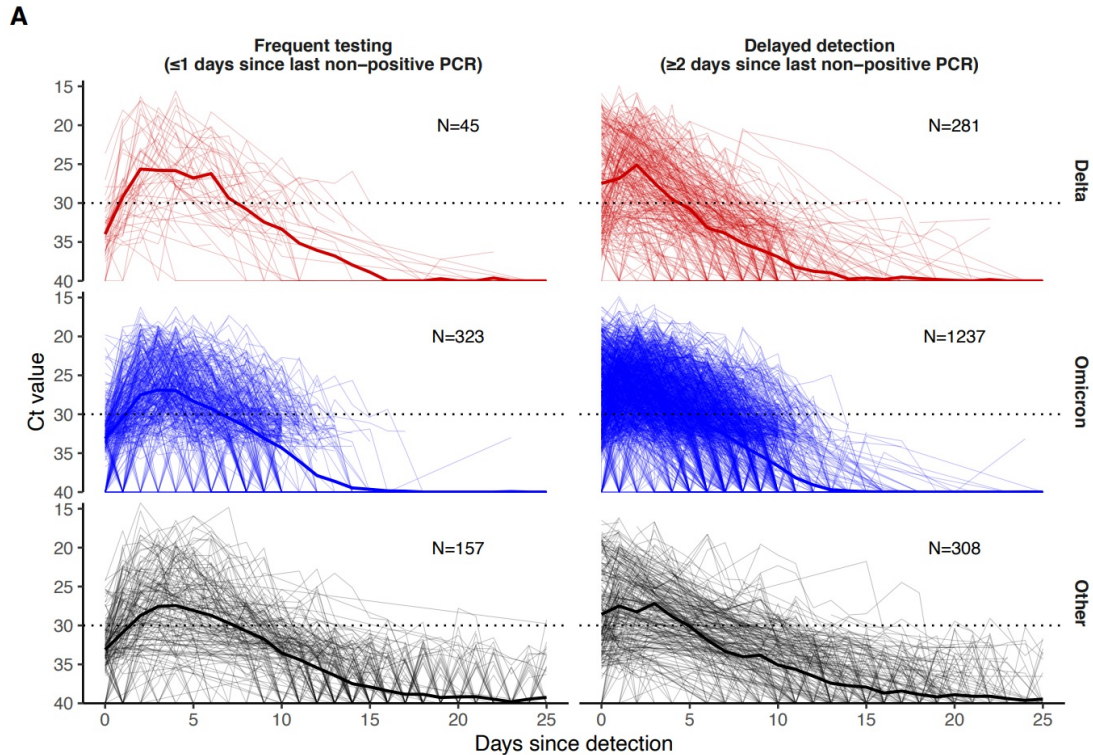
Chia, P. Y., et. al., 2021: "Virological and serological kinetics of SARS-CoV-2 Delta variant vaccine breakthrough infections: a multicentre cohort study" [conducted in a Singapore hospital](#)

- Major issue: "day of illness": symptomatic/positive PCR
- day of illness  $\neq$  day of infection



# SARS-Cov-2 viral load trajectory

Hay, J. A., et. al., 2022: "Quantifying the impact of immune history and variant on SARS-CoV-2 viral kinetics and infection rebound: A retrospective cohort study" Ct-value of NBA staff measured daily



# Viral load trajectory reconstruction

## Can we reconstruct the mean Ct-value trajectory over time?

- Such longitudinal studies are expensive to conduct, require constant monitoring, and are limited to small samples
- Is there a cheaper/more efficient way to collect data?
- How accurate is the estimated trajectory then?

## The data we had

- Ct-value measurements were performed by major labs in Israel
- $\sim 222,000$  Ct-value records of PCR-confirmed infection cases
- $\sim 6,500$  individuals whose Ct-value was measured twice or more on different days during an infection event
- Main problem: day of infection ("time origins") is unknown

## HIV biomarkers trajectory

- Unknown time of infection
- Previous studies incorporate dynamical/biological models and prior assumption
  - *Drylewicz J, et. al., Modeling the dynamics of biomarkers during primary HIV infection; 2010*
  - *Degruttola V, et. al., Modeling the progression of HIV infection; 1991 (slope estimation)*
  - *Berman SM. A stochastic model for the distribution of HIV latency time based on T4 counts; 1990*

## Biomarkers trajectory

- *Wang T, et. al., Time-to-Event Analysis with Unknown Time Origins via Longitudinal Biomarker Registration; 2022*
- Proposes a continuous-time likelihood model with Gaussian process

## Discrete-time likelihood model

- For individual  $i$ , let:
  - $x$ : the day after infection ( $1 \leq x \leq 2d - 1$ )
  - $y_{i,x}$ : the  $i$ -th individual's Ct-value on day  $x$  after infection
- Unconstrained model:

$$y_{i,x} = \theta_x + \epsilon_{i,x}, \quad (\epsilon_{i,1}, \dots, \epsilon_{i,2d-1}) \sim \mathcal{N}(0, \Sigma)$$

- Vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{2d-1})$ : Mean trajectory
- Covariance  $\Sigma$ : within-individual variability, e.g., high-risk individuals tend to have negative  $\epsilon_x$ 's ( $y_{i,x} < \theta_x$ )

## Discrete-time likelihood model

- Unconstrained model:

$$y_{i,x} = \theta_x + \epsilon_{i,x}, \quad (\epsilon_{i,1}, \dots, \epsilon_{i,2d-1}) \sim \mathcal{N}(0, \Sigma)$$

- Parameter  $\theta = (\theta_1, \dots, \theta_{2d-1})$ : Mean trajectory
- Parameter  $\Sigma$ : within-individual variability. E.g., high-risk individuals tend to have negative  $\epsilon_x$ 's ( $y_{i,x} < \theta_x$ )

## Additional assumptions

- Trajectories are statistically independent
- 1st Ct-value measurement  $\leq d$  days after infection
- Constant  $\theta_x$  and  $\Sigma_{x,x}$  for  $d \leq x \leq 2d - 1$

## Discrete-time likelihood model

$$y_{i,x} = \theta_x + \epsilon_{i,x}, \quad (\epsilon_{i,1}, \dots, \epsilon_{i,2d-1}) \sim \mathcal{N}(0, \Sigma)$$

- If  $x$ 's were known, we could compute the sample mean of  $y_{i,x}$ 's, grouped by  $x$ , to estimate  $\theta_x$
- But  $x$  ("time origin") is unknown
- Further, if only a single  $y_x$  was observed per individual (one measurement from a single day), the problem would become unidentifiable

## Key idea

The mean Ct-value trajectory can be recovered if multiple measurements are taken per individual

## Discrete-time likelihood model

For the  $i$ -th individual:

- $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,2d-1}) \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ : the Ct-value trajectory (vector)
- Only partial observation of  $\mathbf{y}_i$  is given:

$$(y_{i,x_1}, \dots, y_{i,x_{m_i}})$$

$$1 \leq x_1 \leq \dots \leq x_{m_i} \leq 2d - 1 \quad \text{and} \quad 1 \leq m_i \leq d$$

- Unknown indexing  $x_1, \dots, x_{m_i}$
- Instead, known differences between indices (time difference between consecutive measurements):

$$\Delta_{i,j} = x_{i,j+1} - x_{i,j}$$

- Given:  $\sum_j \Delta_{i,j} < d \implies$  implies:  $x_{m_i} \leq 2d - 1$

## Maximum-likelihood estimation

We derive an EM-algorithm, where the latent variables are:

- All unobserved entries of  $(y_{i,1}, \dots, y_{i,2d-1})$
- $x_{i,1} \sim \mathbf{q}$  (unknown discrete prior probability)
- Complete likelihood:

$$\prod_{i=1}^n \prod_{j=1}^d \left[ \frac{q_j}{(2\pi)^{(2d-1)/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\theta})^T \Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\theta})\right) \right]^{1_{x_{i,1}=j}}$$

## EM-algorithm main steps

- E-step: Partition each  $\mathbf{y}_i$  to  $\mathbf{y}_i^{observed}$  and  $\mathbf{y}_i^{unobserved}$  for each  $x_{i,1} = j$ , and compute conditional expectation
- M-step over  $\boldsymbol{\theta}$ : a quadratic program

## Model identifiability

When pairs of observations are given per individual, the model is identifiable in  $\boldsymbol{\theta}$ ,  $\boldsymbol{\Sigma}$ ,  $\mathbf{q}$ , under certain conditions:

- Uniqueness of pairs  $(\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{x,x})$
- Positive priors  $\mathbf{q}$

Proof establishes on identifiability of Gaussian mixture models

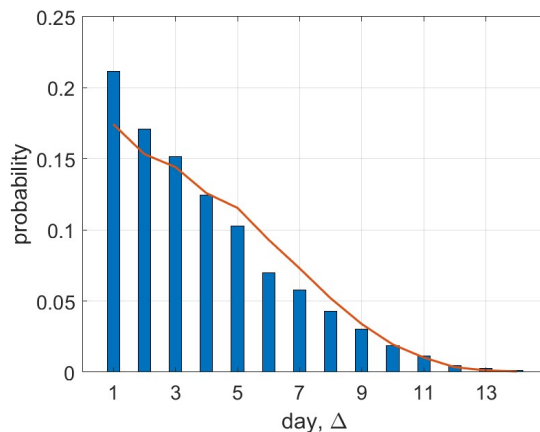
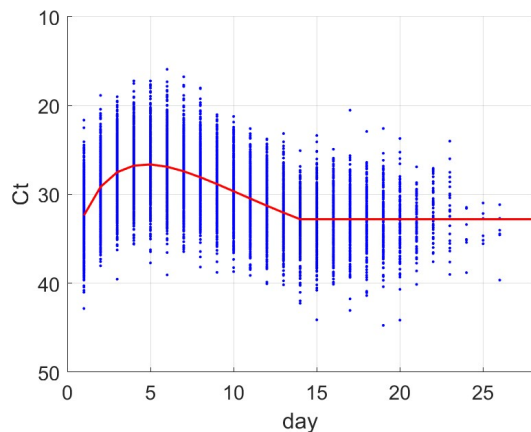
## Remarks

- Due to data's high variability, we use structural constraints:
  - Unimodal trajectory of  $\boldsymbol{\theta}$  with a peak
  - Parametric  $\boldsymbol{\theta}$ , e.g.,  $\boldsymbol{\theta}_x = \alpha_1 x^{\alpha_2 - 1} e^{-x/\alpha_3}$  ("Gamma")
  - $AR(1)$ /linear covariance  $\boldsymbol{\Sigma}$
- The likelihood function is not convex! try several random initializations

# Simulations

## Synthetic samples

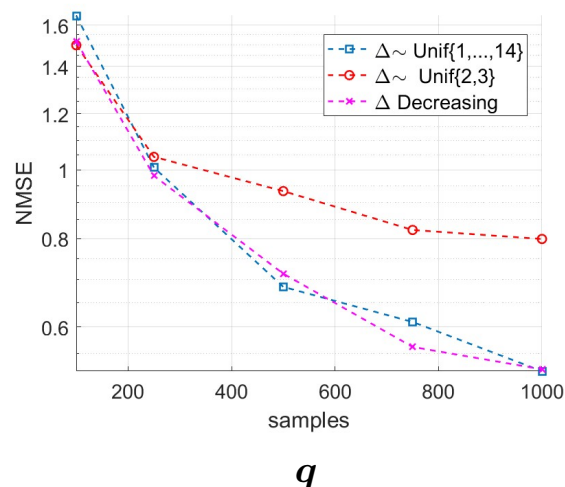
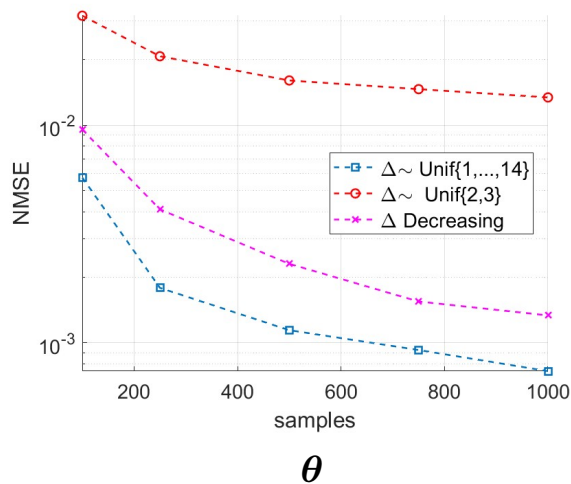
- $\theta$  was obtained from NBA dataset (Hay, J. A., et. al., 2022)
- Structure:  $\theta \implies \text{Gamma}$ ,  $\Sigma \implies \text{AR}(1)$
- $d = 14$ , only pairs of Ct-values per individual
- Samples were generated by  $\mathcal{N}(\theta, \Sigma)$
- 3 types of  $\Delta$  distribution:  $\mathcal{U}(1, \dots, 14)$ ,  $\mathcal{U}(2, 3)$ , decreasing



# Simulations

## Synthetic samples

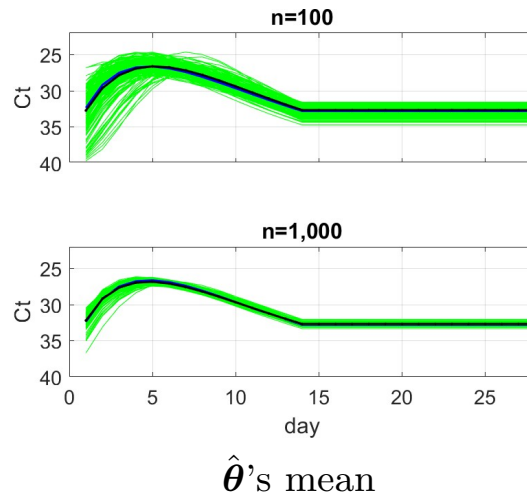
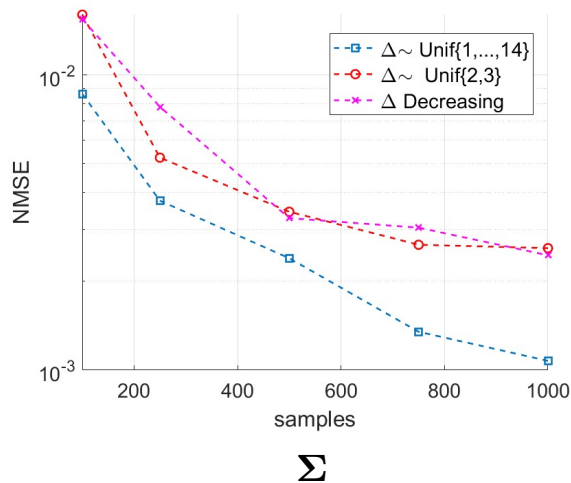
- Normalized mean-square error (NMSE):  
$$\text{NMSE}(\hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 / \|\boldsymbol{\theta}\|^2$$
- Consistent estimates of  $\boldsymbol{\theta}$ ,  $\boldsymbol{\Sigma}$ ,  $\mathbf{q}$
- Best performance when  $\Delta \sim \mathcal{U}(1, \dots, 14)$



# Simulations

## Synthetic samples

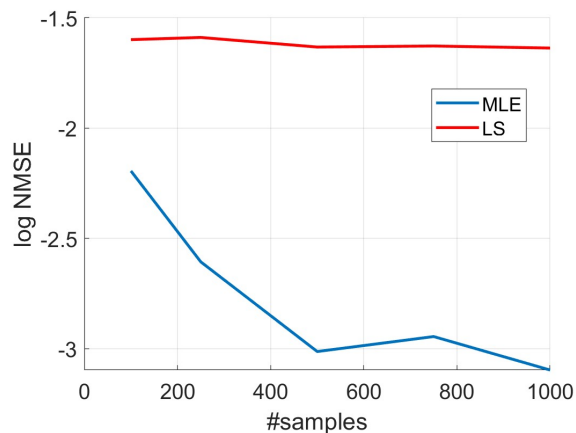
- Normalized mean-square error (NMSE):  
$$\text{NMSE}(\hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 / \|\boldsymbol{\theta}\|^2$$
- Consistent estimates of  $\boldsymbol{\theta}, \boldsymbol{\Sigma}, \mathbf{q}$
- Best performance when  $\Delta \sim \mathcal{U}(1, \dots, 14)$



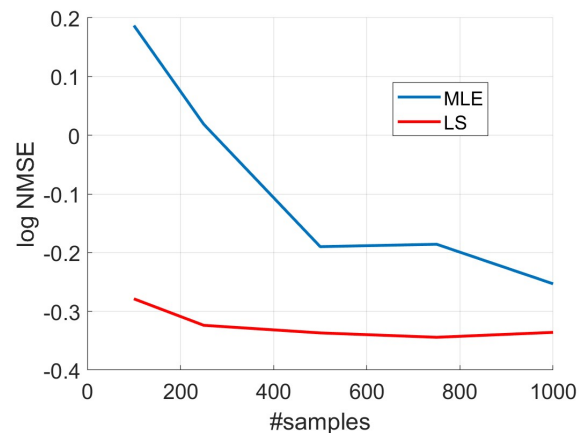
# Simulations

## Synthetic samples: comparison to least-squares method (LS)

- LS can be achieved by applying a grid search over three Gamma parameters
- LS does not account for within individual correlation



$\theta$

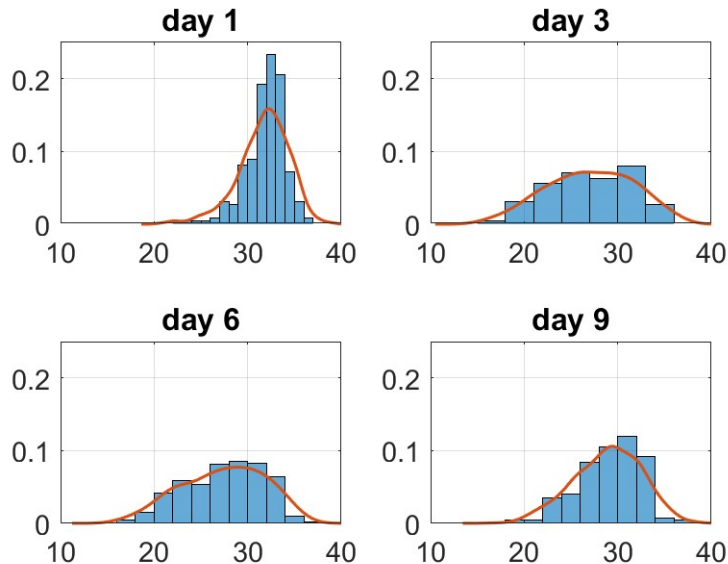


$q$

# Simulations

## Semi-synthetic samples

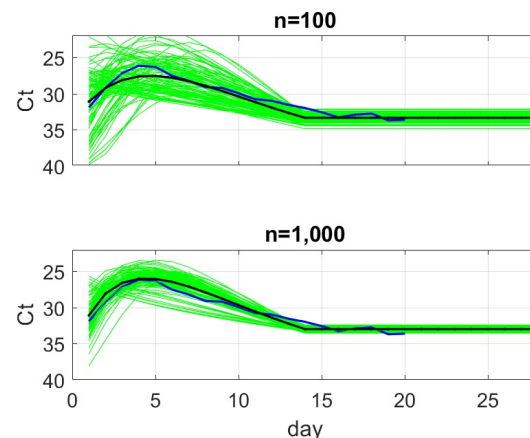
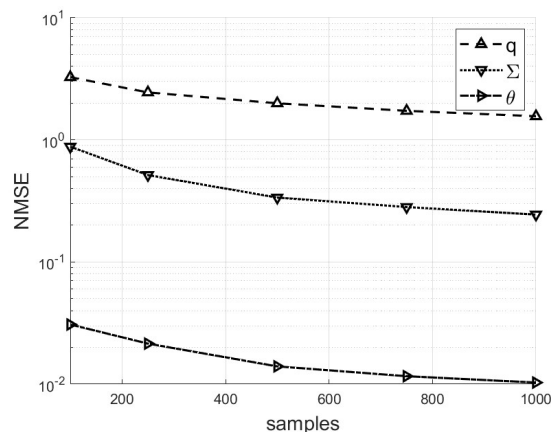
- NBA Ct-values are not normally distributed (histograms)
- We generated samples using the marginal empirical distribution
- We added  $\mathcal{N}(0, 1)$  random effect +  $\mathcal{N}(0, 1)$  noise (red lines)



# Simulations

## Semi-synthetic samples

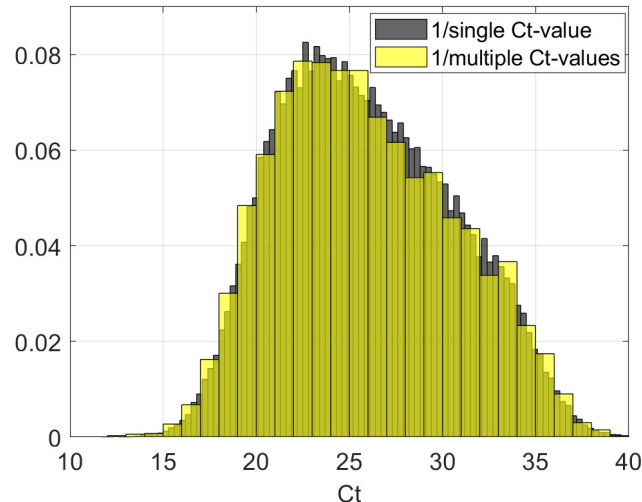
- Normalized mean-square error (NMSE):  
$$\text{NMSE}(\hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 / \|\boldsymbol{\theta}\|^2$$
- Consistent estimates of  $\boldsymbol{\theta}$ ,  $\boldsymbol{\Sigma}$ ,  $q$



$\hat{\boldsymbol{\theta}}$ 's mean

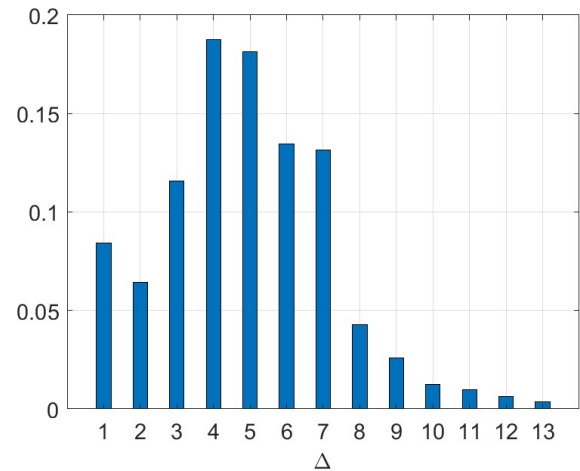
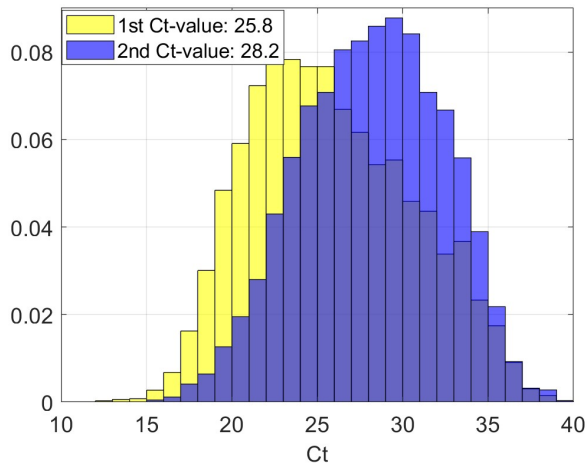
# Real data

- 222,668 records:
  - 97% include a single measurement per individual
  - 2.7% include pairs of measurements per individual
- Individuals who were measured twice vs once: a selection bias?
- Similar histograms: 1st among single vs. multiple Ct-values



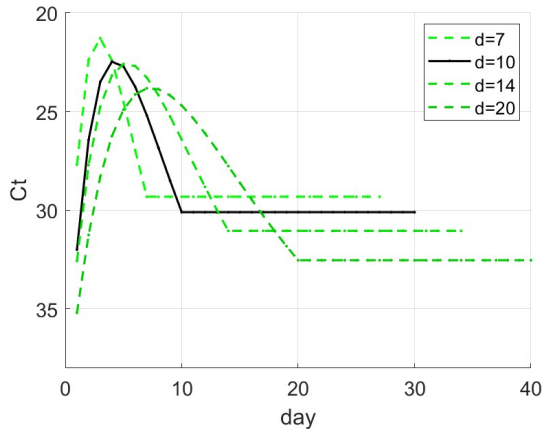
# Real data

- Among individuals with multiple Ct-values:  
1st Ct-value < 2nd Ct-value (on average)

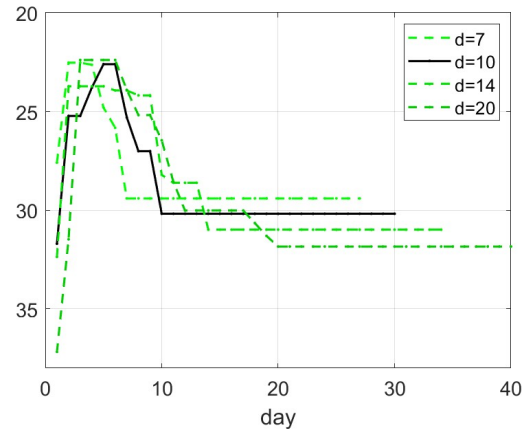


## Estimation results

- Estimates for  $d = 7, 10, 14, 20$
- Imposing Gamma (left) and unimodal (right) constraints



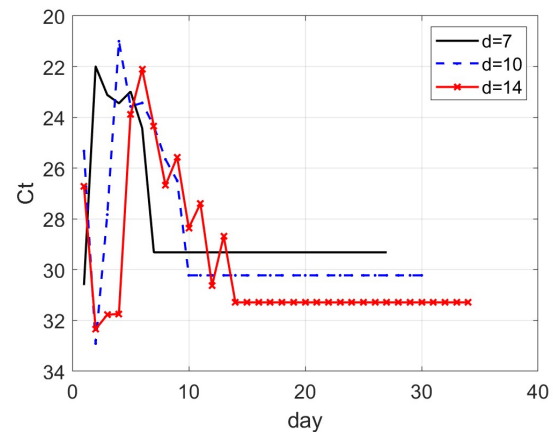
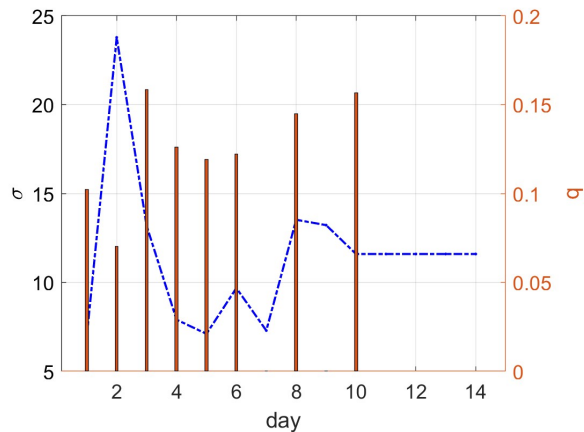
Gamma:  $\theta_x = \alpha_1 x^{\alpha_2 - 1} e^{-x/\alpha_3}$



Unimodal

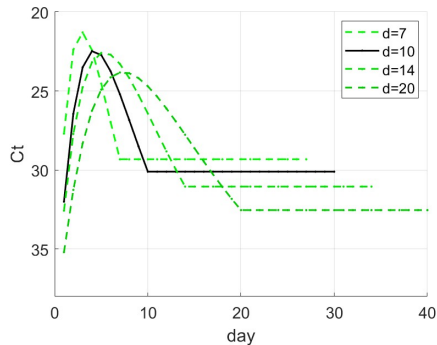
## Estimation results

- Left: Variance (blue) and  $\mathbf{q}$  (red) estimates for Gamma constraints on  $d = 10$
- Right:  $\theta$  estimate without constraints



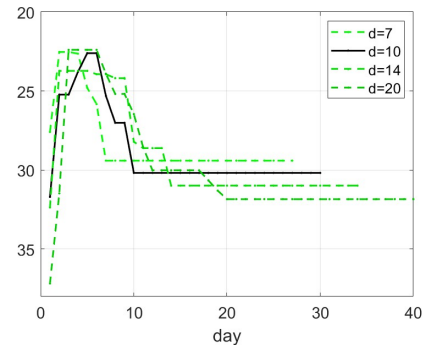
## Interpreting the estimate

- Estimate should be interpreted carefully
- Trajectory is conditioned on  $C_t < 40$  since data does not include recovered persons
- Other possible selection biases
- We focus on the EM-algorithm performance



Gamma:

$$\theta_x = \alpha_1 x^{\alpha_2 - 1} e^{-x/\alpha_3}$$



Unimodal

- Viral load trajectory is a key factor in understanding infectiousness
- Usually requires resource-intensive and complex longitudinal studies
- We developed an estimation method on readily available data
- Estimation method can be applied to data containing pairs/triplets of Ct-values per individual
- The model offers a practical guideline for effective data collection:
  - A pair of Ct-value measurements per individual suffices
  - Ct-value measurements on confirmed cases

## Possible future research directions

- Extension to non-Gaussian distribution: e.g., bimodal (to include recovered)
- Experimental design: e.g., analyzing optimal distribution of  $\Delta$  (recall best NMSE when  $\Delta \sim \mathcal{U}(1, \dots, 14)$ )
- Pairs or triplets of measurements?

Thank you!