Quantifying the relative information in noisy epidemic time series

Kris V Parag

University of Bristol and Imperial College London

k.parag@imperial.ac.uk, @krisparag1

- Joint work with: **Christl A Donnelly** & **Alexander E Zarebski**

- KV Parag, CA Donnelly & AE Zarebski. ***Quantifying the information in noisy epidemic curves***. medRxiv 10.1101/2022.05.16.22275147; 2022.

Tracking infectious disease spread

- How can we ***meaningfully*** infer changes in pathogen ***transmissibility***?
  - Crucial insights into ecological or biological factors (new variants)
  - Test hypotheses about transmission mechanisms (super-spreading)
  - Evaluate effectiveness of interventions (lockdowns, quarantines)
  - Forecast or communicate likely disease burden (lives, livelihoods)
- ***Meaningful***: summarises key dynamics, reliable given practical data

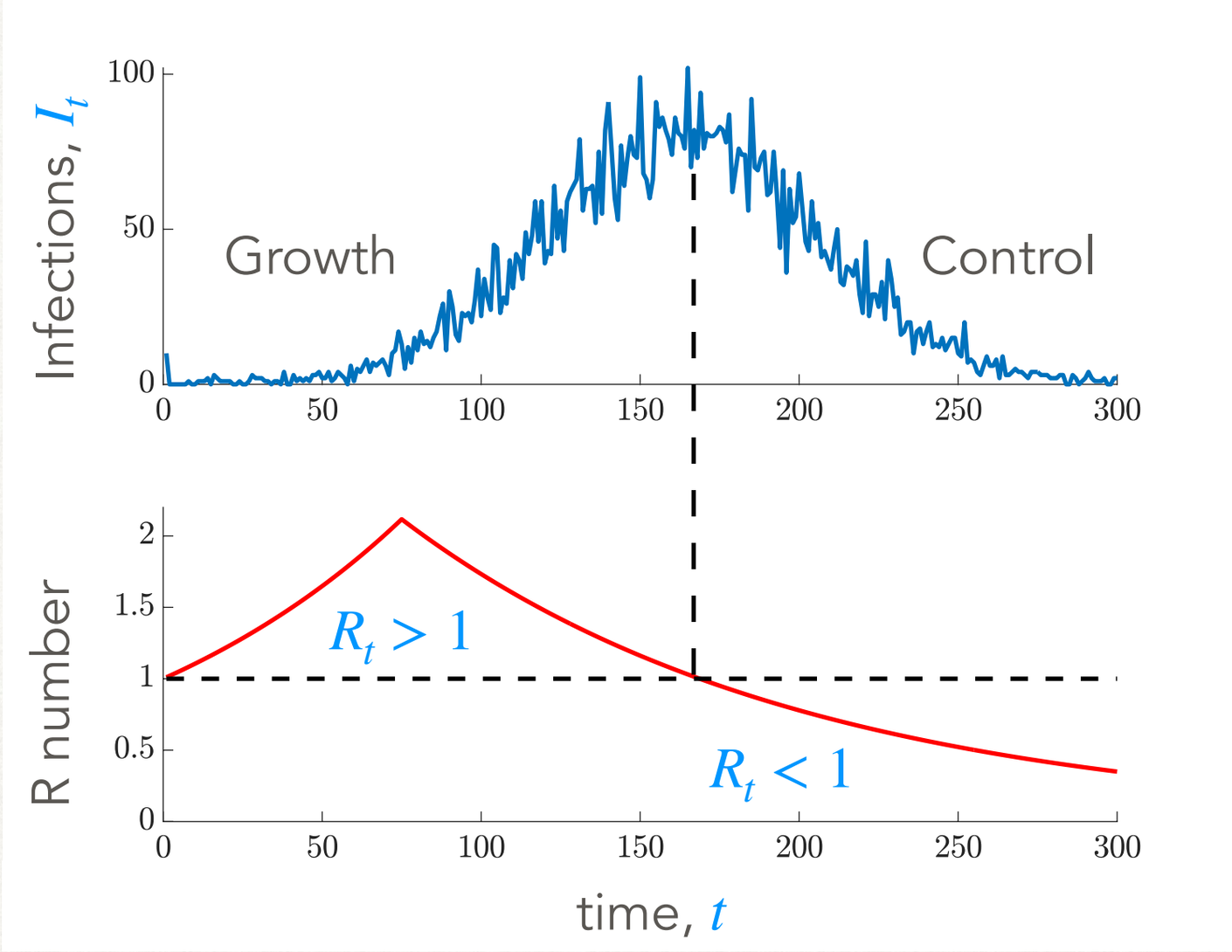
Effective reproduction number, R

- Measures temporal spread, threshold parameter or dominant pole

- R_t — average new infections at t per effective infection

- Are interventions or controls working?

- Are second waves imminent or likely?



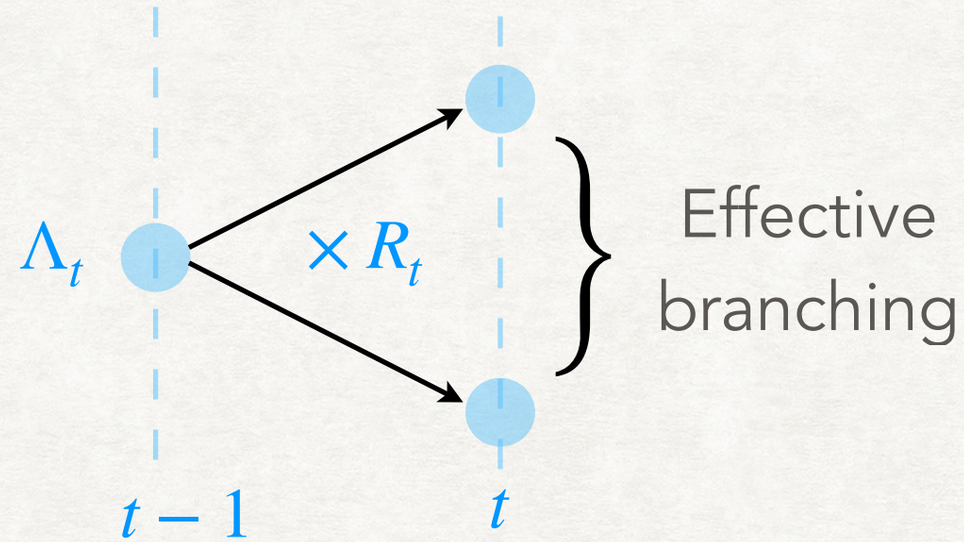
- R Anderson, C Donnelly et al. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK. The Royal Society (2020).

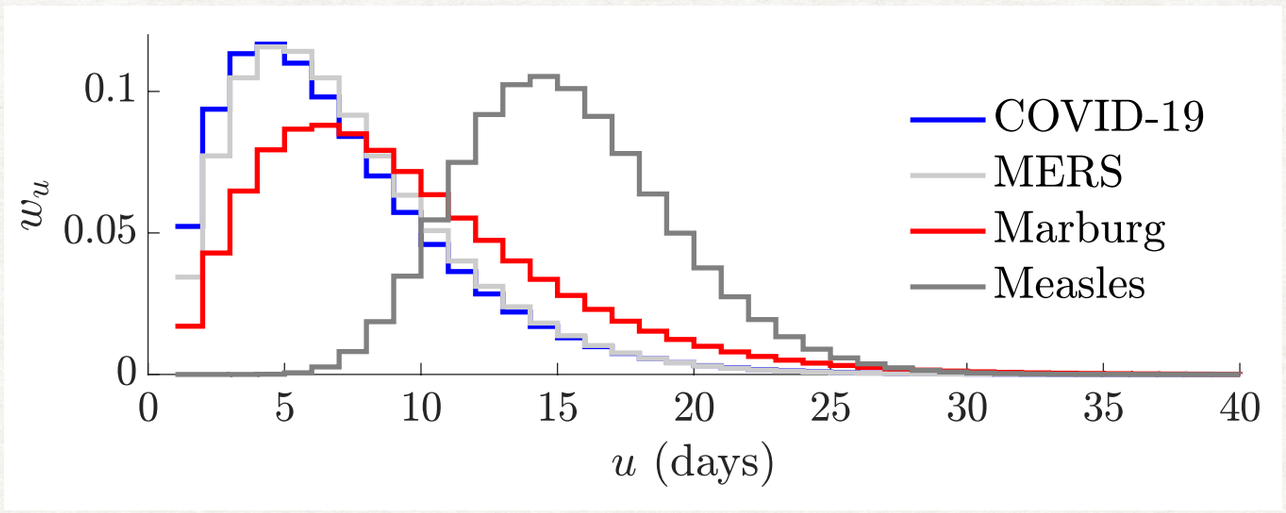
Renewal transmission models

- General epidemic model — links to Bellman-Harris; autoregressive

$$I_t \sim \text{Pois}\left(R_t \Lambda_t\right)$$ New infections

$$\Lambda_t = \sum_{u=1}^{t-1} w_u I_{t-u}$$ Active no. past infections





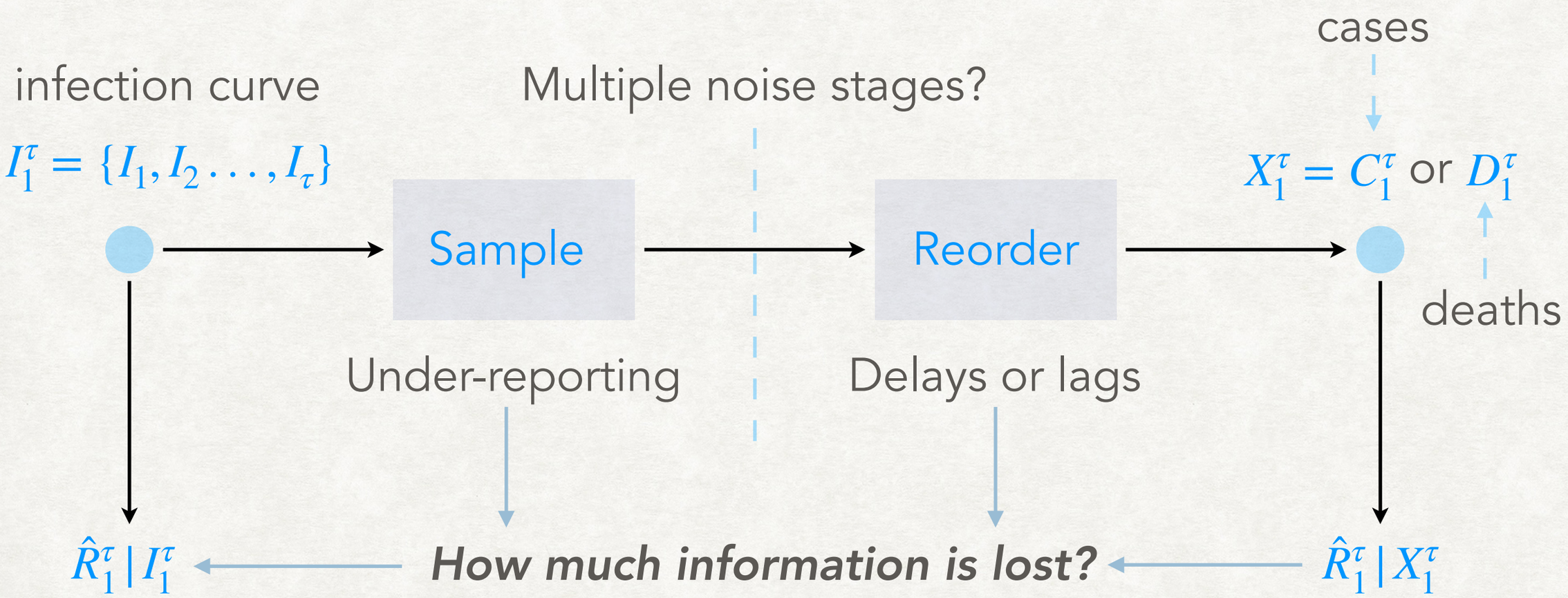
Distribution of times between infection

Shape encodes different transmission dynamics

- C Fraser, D Cummings, et al. Influenza transmission in households during the 1918 pandemic. Am J Epidemiol 174:505–14 (2011).

Epidemic time series or curves

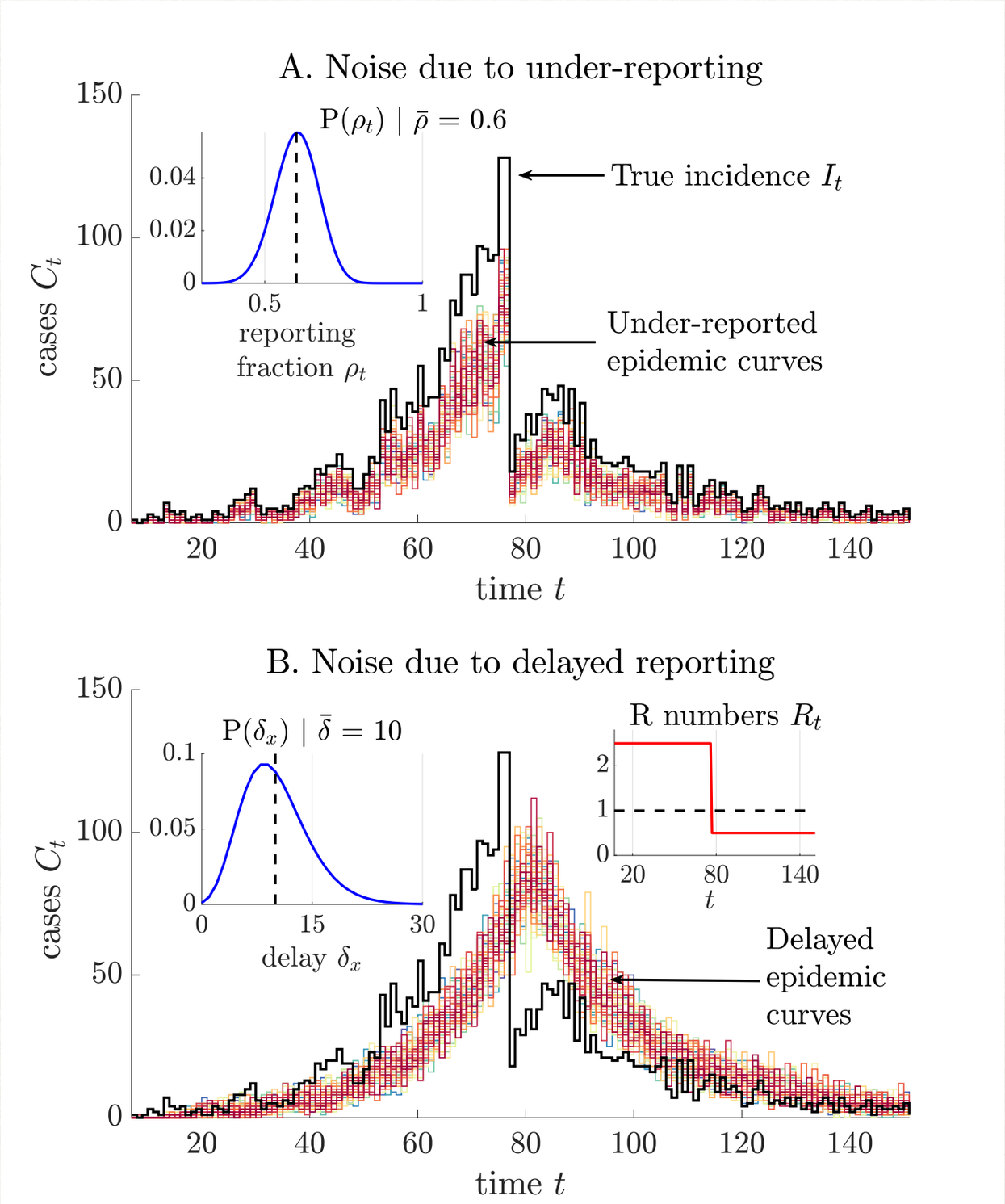
- Ideal: estimate R_t from I_t and Λ_t — but these are ***unobservable***
- Practical ***proxies*** — cases, deaths, hospitalisations, wastewater?



- Can we measure this loss — rank the ***reliability*** of proxy time series?

Noise — delays and under-reporting

- No framework exists for comparing noise sources
- ***Generalisable*** — avoid expensive simulations
- ***Interpretable*** — which noise source dominates
- Use Fisher information and experimental design?



The information in practical data

- Assume time units chosen such that every R_t is independent
- Reporting fraction is $\rho_t \leq 1$ and $\delta_x \leq 1$ prob. delay is x units

$$C_t \sim \text{Pois}\left(\sum_{s=1}^{t} \delta_{t-s}\rho_s\Lambda_s R_s\right) \xrightarrow[\delta_x = 1_{x=0}]{\rho_t = 1} I_t \sim \text{Pois}\left(R_t\Lambda_t\right)$$

- Fisher information gives best precision of unbiased estimators

$$\mathbb{F}_C(R_t) = \rho_t F_{\tau-t}\Lambda_t R_t^{-1} \longrightarrow \mathbb{F}_I(R_t) = \Lambda_t R_t^{-1}$$

with

$$F_{\tau-t} = \sum_{x \leq \tau-t} \delta_x$$

- K Parag & C Donnelly. Adaptive Estimation for Epidemic Renewal and Phylogenetic Skyline Models. Syst. Biol 69:1163–1179 (2020).

Total Fisher information ratios

- Reliability — measure total uncertainty when inferring R_1^τ from C_1^τ
- ***D optimal*** design — maximise the determinant of Fisher matrix
- Independence implies determinant$^{\frac{1}{2}}$:

$$\mathbb{T}(C_1^\tau) = \prod_{t=1}^{\tau} \sqrt{\rho_t F_{\tau-t} \Lambda_t R_t^{-1}}$$

- $\dfrac{1}{\mathbb{T}(C_1^\tau)} \propto$ volume of the (asymptotic) uncertainty ellipsoid around $\hat{R}_1^\tau$
- Propose novel ratio:

$$0 \leq \eta(C_1^\tau) = \frac{\mathbb{T}(C_1^\tau)}{\mathbb{T}(I_1^\tau)} = \prod_{t=1}^{\tau} \sqrt{\rho_t F_{\tau-t}} \leq 1$$

Generalisable, simulation agnostic, independent of unknown R_t, I_t, Λ_t

Interpretable information diagnostics

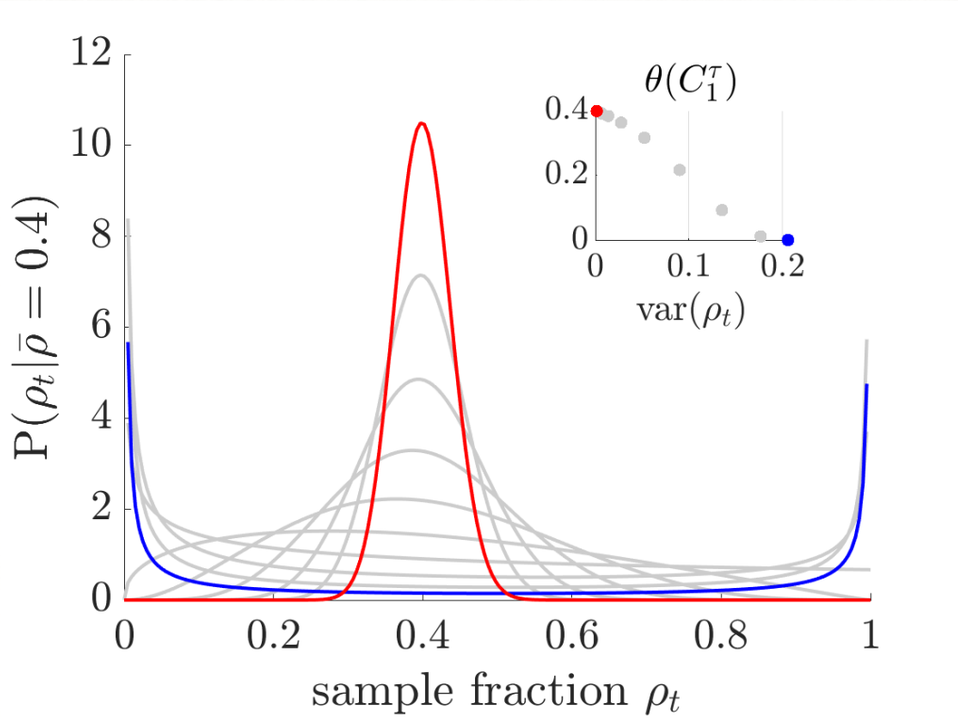
- A model with stable reporting $C_t \sim \mathrm{Pois}(\theta R_t \Lambda_t)$ has $\eta(C_1^\tau) = \sqrt{\theta^\tau}$
- Effective (fixed) reporting ratio: $\theta(C_1^\tau) = \mathbb{G}(\rho_t F_{\tau-t}) = \mathbb{G}(\rho_t)\mathbb{G}(F_{\tau-t})$

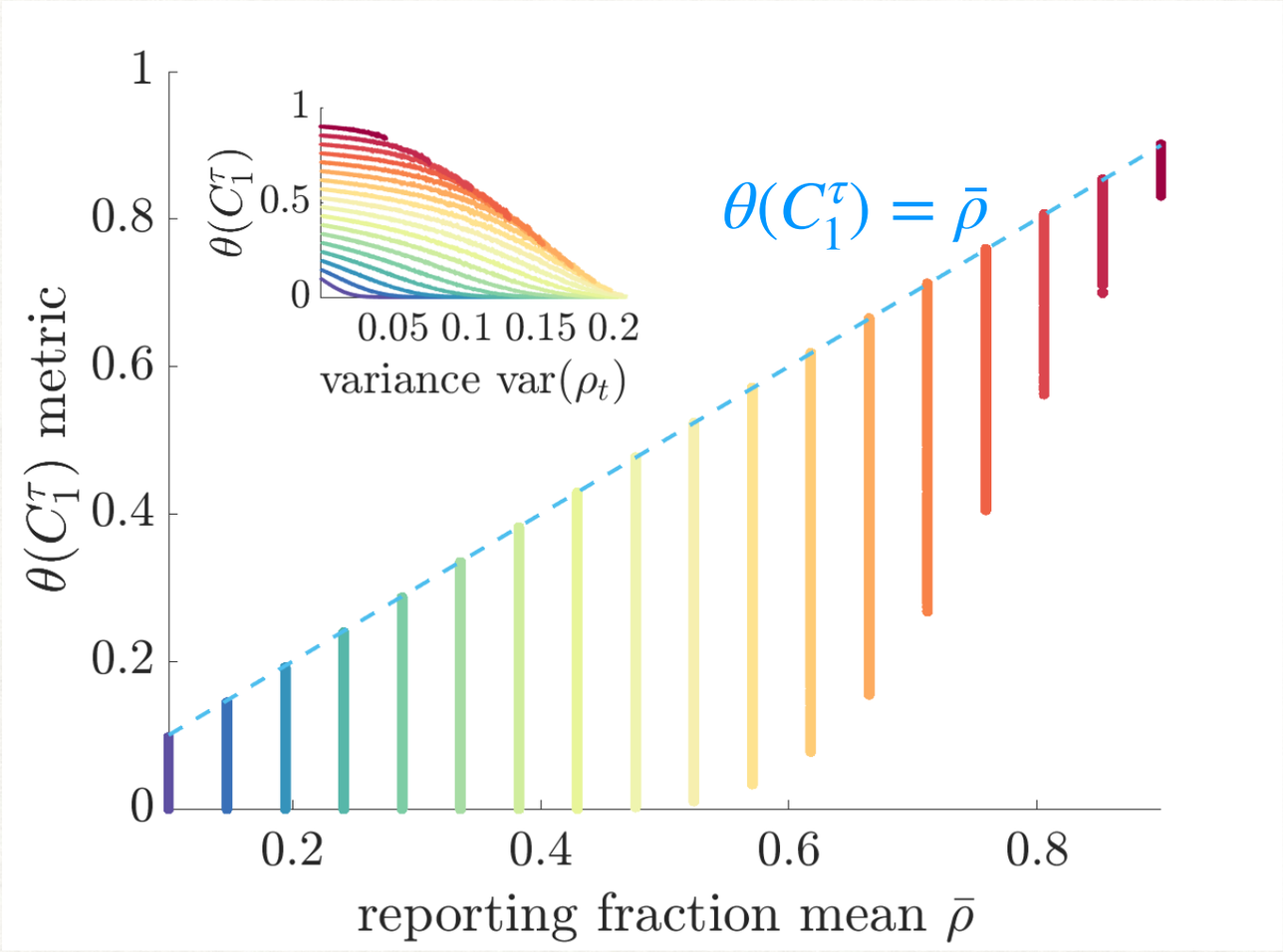
$$0 \leq \min_t \alpha_t \leq \mathbb{G}(\alpha_t) \leq \max_t \alpha_t \leq 1$$

Each noise source degrades information by its ***geometric mean***

- Under mean constraint — maximised by the most ***uniform*** α_t input
- Noise types add terms — can ***compare*** noise sources, data streams

Under-reporting

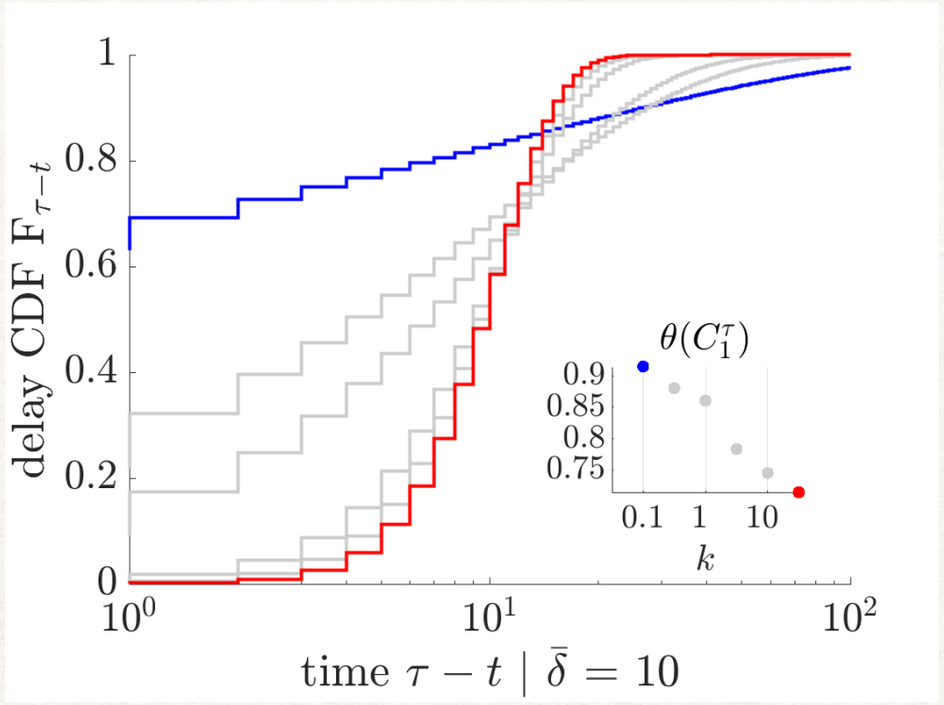


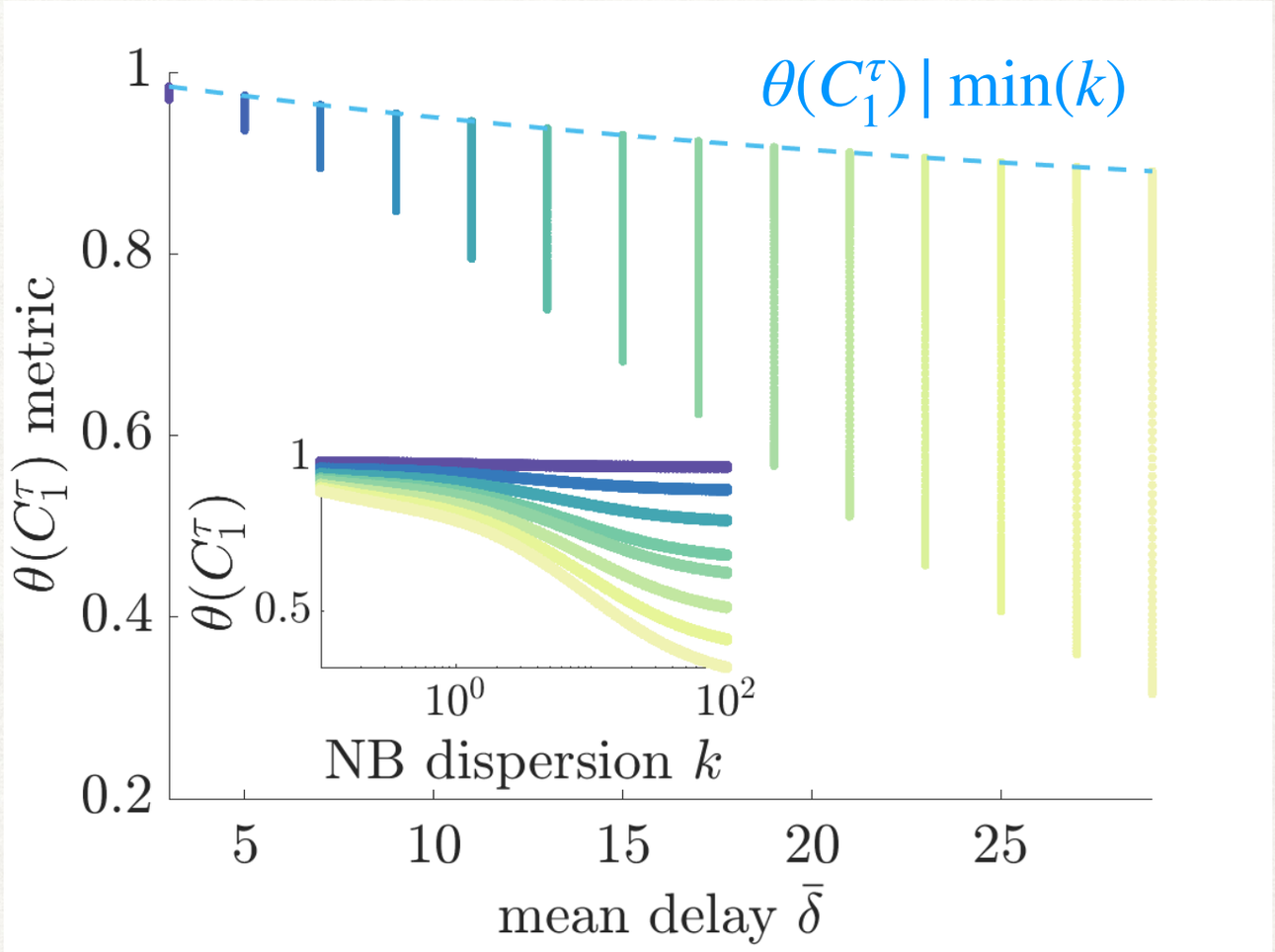


- Constant reporting is optimal — supported

- But variability $\mathrm{var}(\rho_t)$ and mean $\bar{\rho}$ reporting rates affect information
  - Higher $\mathrm{var}(\rho_t)$ on its own does not guarantee worse reliability

Delays in reporting





- Less variable delays are worse — contradiction

- Dispersion k and mean $\bar{\delta}$ reporting delays modulate information
  - Majority of cases with small delay, minority with very large lag

Are deaths more reliable than cases?

- ***Untested***, common assumption for COVID-19 — deaths better
- Wildly fluctuating testing rates and high fraction of cases missed
- Reporting of deaths is more stable, less behaviour-dependent

$$D_t \sim \text{Pois}\left(\sum_{x=1}^{t} \text{ifr}_x \gamma_{t-x} \sigma_x \Lambda_x R_x\right) \longrightarrow \theta(D_1^\tau) = \mathbb{G}(\text{ifr}_t)\mathbb{G}(\sigma_t)\mathbb{G}(H_{\tau-t})$$

- Infection fatality rate at t — ifr_t, prob. of delay from infection to death is x units — γ_x, death reporting rate — σ_t, $H_{\tau-t} = \sum_{x \leq \tau-t} \gamma_x$

- P Nouvellet, S Bhatia et al. Reduction in mobility and COVID-19 transmission. Nat. Commun 12:1090 (2020).

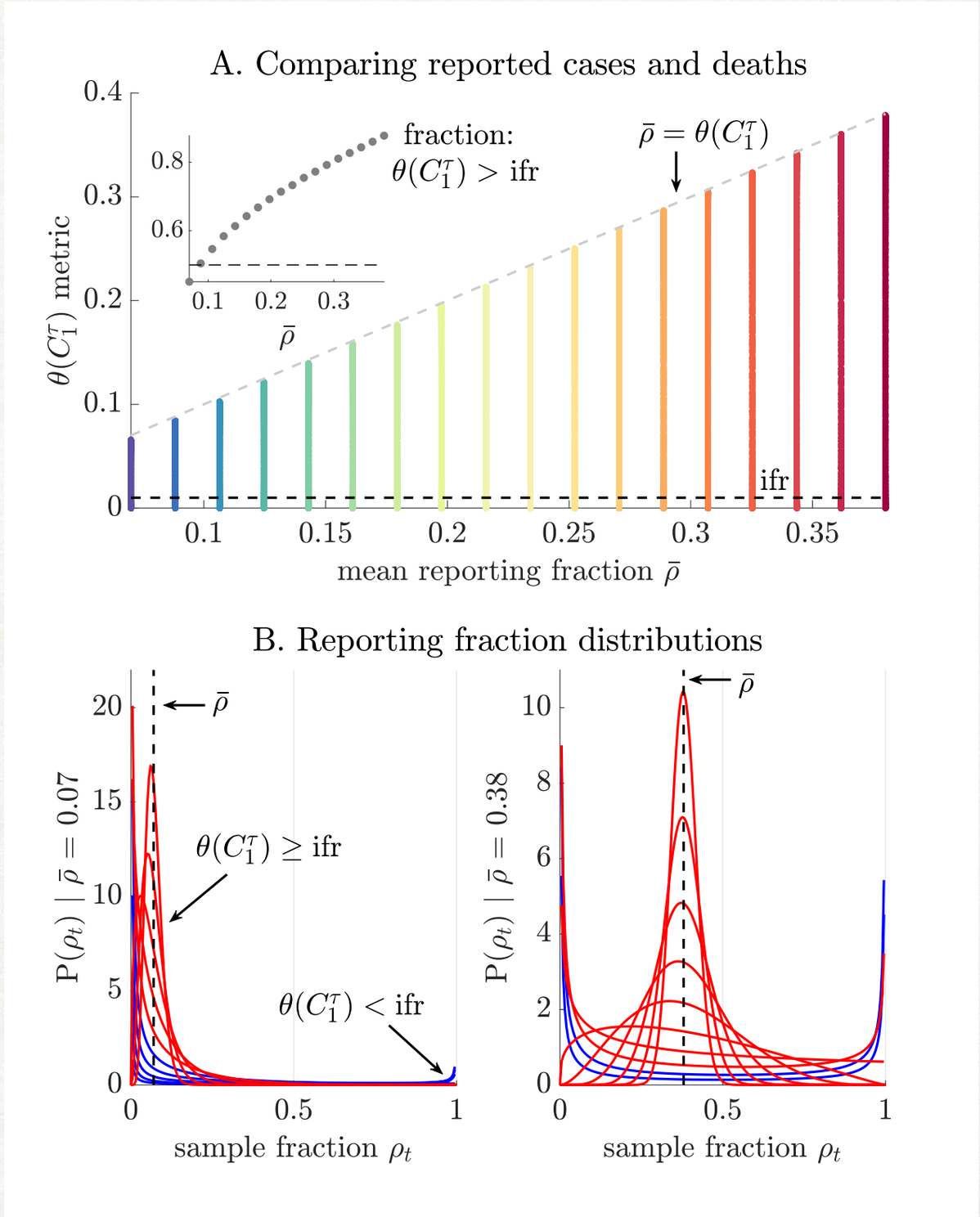
Cases C_1^τ vs deaths D_1^τ?

Cases better:

$$\mathbb{G}\left(\frac{\rho_t}{\sigma_t \text{ifr}_t}\right) \geq \mathbb{G}\left(\frac{H_{\tau-t}}{F_{\tau-t}}\right)$$

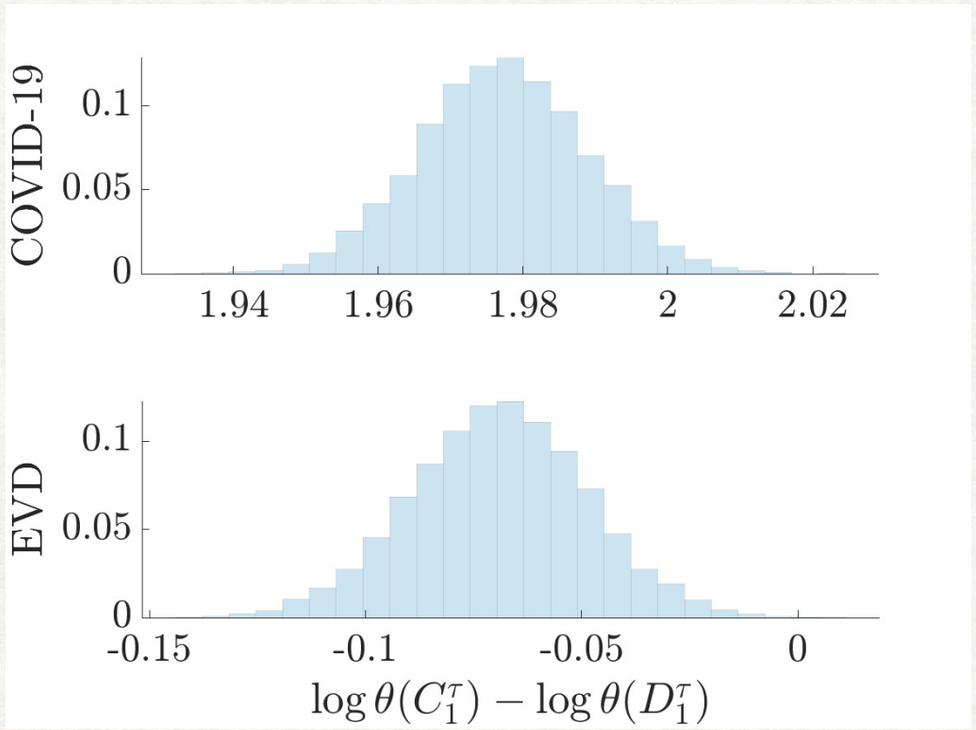
- Let $\sigma_t = 1$, $\text{ifr}_t = 0.01$ — max the information in deaths
- $\mathbb{G}(H_{\tau-t})$ upper bounds delay information, $\mathbb{G}\left(\frac{H_{\tau-t}}{F_{\tau-t}}\right) \leq 1$
- Estimated $0.07 \leq \rho_t \leq 0.38$

$$\mathbb{G}\left(\rho_t | \bar{\rho} \in [0.07, 0.38]\right) \geq 0.01$$



Benefits of information ratios

- Rigorously assess or interrogate the relative reliability of time series
  - Very unlikely COVID-19 deaths better — ifr means 99% I_1^τ lost



- Apply practically — ***independent*** of simulations or R-inference methods
- Interpretable — ***analytical insights***
- Pinpoint ***bottlenecks*** in surveillance

- ***Hospitalisations***, infection ***prevalence*** and virus ***wastewater*** surveys (novel) all conform to this framework — growing applicability

Limitations and outlook

- Renewal models — ***simplistic*** (Poisson) and ***well-mixed*** assumptions
  - Some studies find ***network*** models do not improve R-inference
- Fisher information is asymptotic and assumed independence of R_t
  - Variance stabilising transforms, non-diagonal Fisher matrix terms
  - ***Description length*** (in bits) strongly depends on $\log \mathbb{T}(X_1^\tau)$

We must think more carefully about the ***quality of data*** as pandemic response becomes more ***data driven*** — e.g., consensus weights

- Q Liu, M Ajelli et al. Measurability of the epidemic reproduction number in data-driven contact networks. PNAS 115:12680–12685 (2018).