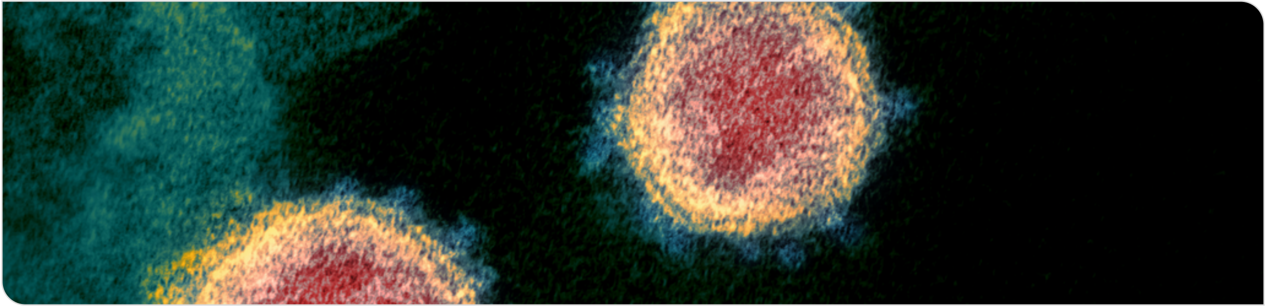
Karlsruher Institut für Technologie

A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts

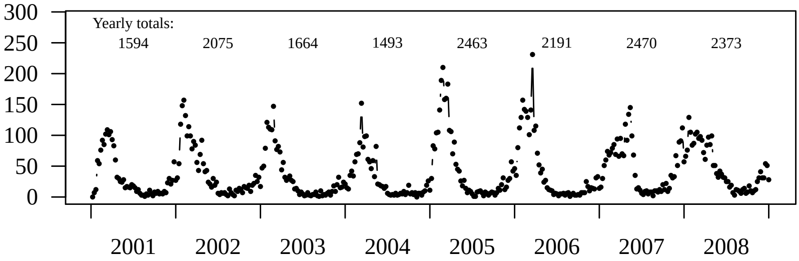
Workshop on the Endemic-Epidemic framework for infectious disease modelling, LSHTM, 23 March 2022

Johannes Bracher | Karlsruhe Institute of Technology / Heidelberg Institute for Theoretical Studies
joint work with **Leonhard Held**, University of Zurich



Motivation: Rotavirus in Berlin

- Gastrointestinal disease affecting mainly young children
- A vaccine has existed since 2006 (used widely since 2009), but we consider the pre-vaccination period
- Mean serial interval estimated as ~ 5 days (Grimwood 1983)
- **Known to be severely underreported: estimated reporting probability in Western Germany (incl Berlin): 4.3% (Weidemann et al 2014)**



Data: survstat/RKI

Goals

- **Simple model** distinguishing within-region spread and import of cases, **accounting for underreporting.**
- Computationally efficient **inference scheme.**
- Apply to **estimate time varying local reproductive numbers**: *how many new cases does one case cause on average within a given geographic unit?* (See C. Bauer's talk later)
- Analyse **biases** resulting when ignoring underreporting

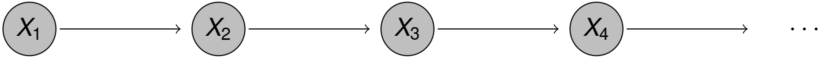
The endemic-epidemic model (Held et al 2005)

Simplest form: model the number X_t of new cases at discrete time t as

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \lambda_t, \text{disp} = \psi)$$

$$\lambda_t = \underbrace{\nu}_{\text{"endemic component"}} + \underbrace{\phi X_{t-1}}_{\text{"epidemic component"}}$$

Simple Markov structure:



The endemic-epidemic model (Held et al 2005)

Simplest form: model the number X_t of new cases at discrete time t as

$$X_t \mid \text{past} \sim \text{NegBin}(\text{mean} = \lambda_t, \text{disp} = \psi)$$
$$\lambda_t = \underbrace{\nu}_{\text{"endemic component"}} + \underbrace{\phi X_{t-1}}_{\text{"epidemic component"}}$$

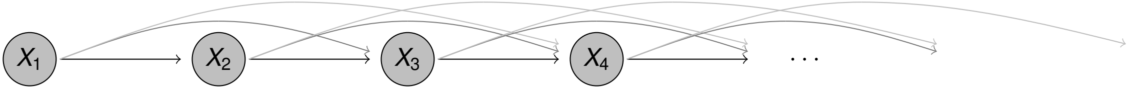
Intuition:

- Some cases are due to cases outside the study region / infection from reservoirs / environmental sources.
- Some cases are due to cases from the time $t - 1$ (serial interval = 1 time period).
- ϕ is local effective reproductive number R (assuming serial interval = 1 time step; Bauer/Wakefield 2018).
- With average number ν of importations and local reproduction number $\phi < 1$ there will on average be $\nu/(1 - \phi)$ new cases per time point.

Relaxing the AR(1) assumption

Extend the conditional mean structure to

$$
\begin{aligned}
X_t \mid \text{past} &\sim \text{NegBin}(\lambda_t, \psi) \\
\lambda_t &= \nu + \phi X_{t-1} + \kappa \lambda_{t-1}
\end{aligned}
$$



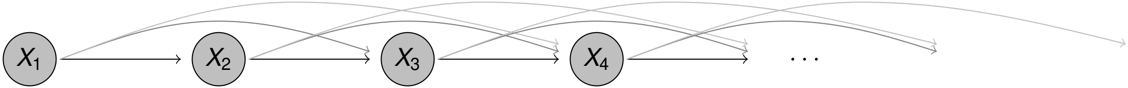
- Equivalent to NB-INGARCH(1, 1) model (Zhu 2011)

Relaxing the AR(1) assumption

Extend the conditional mean structure to

$$X_t \mid \text{past} \sim \text{NegBin}(\lambda_t, \psi)$$

$$\lambda_t = \frac{\nu}{1-\kappa} + \frac{\phi}{1-\kappa} \cdot \sum_{d=1}^{t-1} (1-\kappa)\kappa^{d-1} X_{t-d}$$

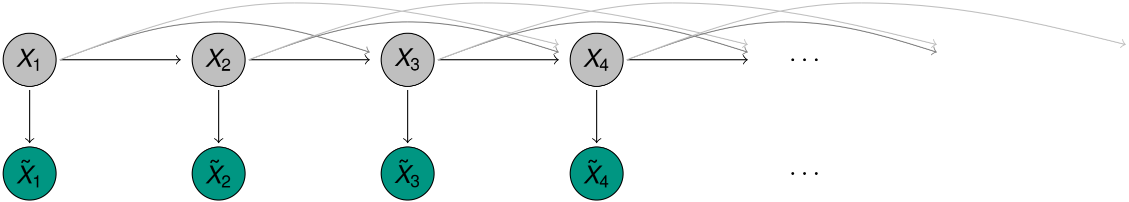


- Equivalent to NB-INGARCH(1, 1) model (Zhu 2011)
- Effective reproductive number: $R_{\text{eff}} = \frac{\phi}{1-\kappa}$
- Geometric serial interval distribution (mean $= 1/(1-\kappa)$)

Adding a reporting process to the model

We denote the **reported** cases by $\tilde{X}_t$ and link them to X_t via a time-constant reporting probability π:

$$\begin{aligned} X_t \mid \text{past} &\sim \text{NegBin}(\lambda_t, \psi) \\ \lambda_t &= \nu + \phi X_{t-1} + \kappa\lambda_{t-1} \\ \tilde{X}_t &\sim \text{Bin}(X_t, \pi) \end{aligned}$$



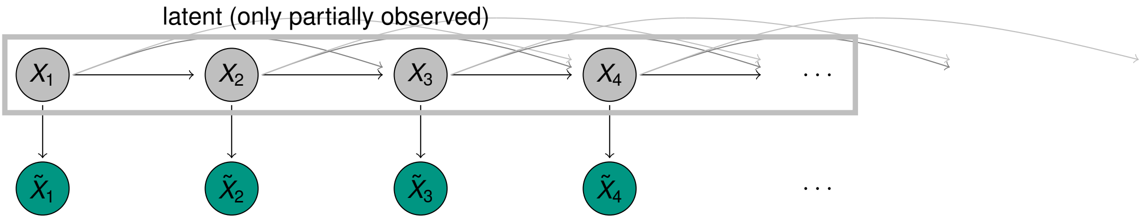
Adding a reporting process to the model

We denote the **reported** cases by $\tilde{X}_t$ and link them to X_t via a time-constant reporting probability π:

$$X_t \mid \text{past} \sim \text{NegBin}(\lambda_t, \psi)$$

$$\lambda_t = \nu + \phi X_{t-1} + \kappa \lambda_{t-1}$$

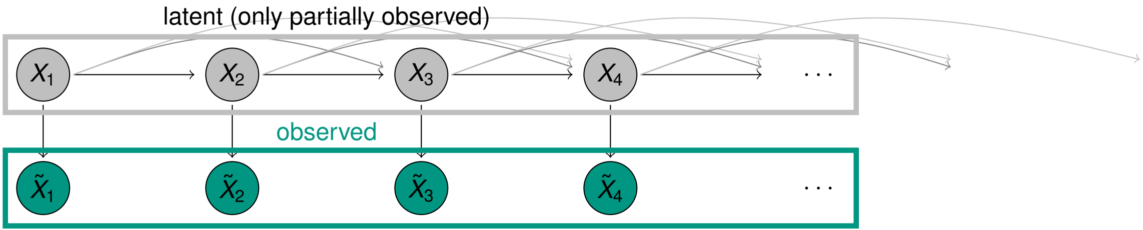
$$\tilde{X}_t \sim \text{Bin}(X_t, \pi)$$



Adding a reporting process to the model

We denote the **reported** cases by $\tilde{X}_t$ and link them to X_t via a time-constant reporting probability π:

$$X_t \mid \text{ past } \sim \text{NegBin}(\lambda_t, \psi)$$
$$\lambda_t = \nu + \phi X_{t-1} + \kappa \lambda_{t-1}$$
$$\tilde{X}_t \sim \text{Bin}(X_t, \pi)$$

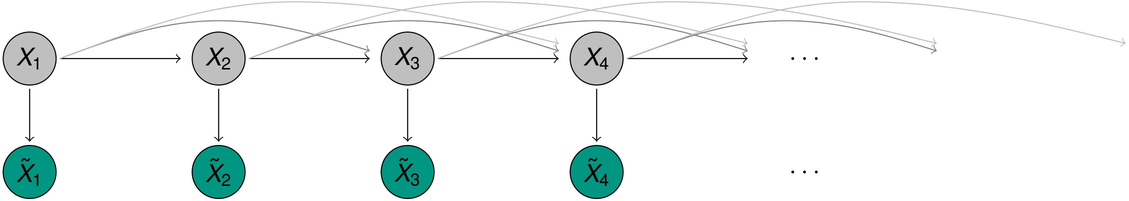


Some theory without formulas

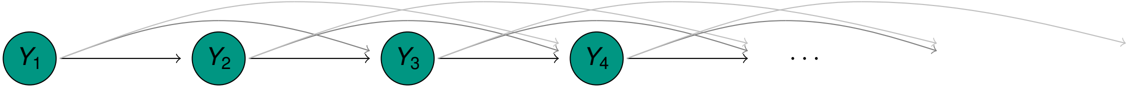
- Stationary means, variances and autocorrelations can be obtained for the latent and observable processes (provided $\phi + \kappa < 1$).
- The latent $\{X_t\}$ and observed $\{\tilde{X}_t\}$ both have geometrically decaying autocorrelation functions.
- Different combinations of parameters ν, ϕ, κ, ψ and reporting probability π can lead to **second-order equivalent** observable processes.

Some theory without formulas (contd)

For each model of type



there is a second-order equivalent model (from the same class) with the simpler struture structure



Simulation studies show that second-order equivlent models behave *very* similarly.

Practical implications

- The model is **not identifiable** – we cannot estimate π along with the other parameters

Practical implications

- The model is **not identifiable** – we cannot estimate π along with the other parameters
- It can be shown that if we ignore underreporting, we will
  - **underestimate** ν, ϕ and the reproductive number R_{eff}
  - **overestimate** the share of imported cases
  - **overestimate** κ (and thus the mean serial interval $1/(1-\kappa)$), ψ

Practical implications

- The model is **not identifiable** – we cannot estimate π along with the other parameters
- It can be shown that if we ignore underreporting, we will
  - **underestimate** ν, ϕ and the reproductive number R_{eff}
  - **overestimate** the share of imported cases
  - **overestimate** κ (and thus the mean serial interval $1/(1-\kappa)$), ψ
- If π is **known** we can set up an **approximate maximum likelihood** inference scheme

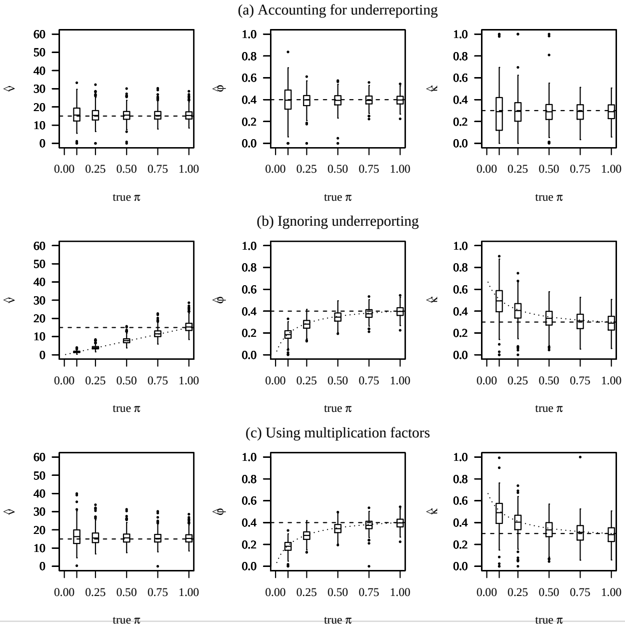
Sketch of the approximate maximum likelihood estimation

- In order to evaluate the likelihood for parameters ν, ϕ, κ, ψ and reporting probability π:
  - Approximate the process with underreporting by a second-order equivalent one without underreporting.
  - Use (conditional) likelihood of this fully observed process as an approximation.

- Optimize the approximate likelihood function numerically.

- This generalizes well to time-varying parameters and even temporally aggregated observations.

Simulation studies



(a) Accounting for underreporting

(b) Ignoring underreporting

(c) Using multiplication factors

Full model for rotavirus data

We assume the following model:

$$X_t \mid X_{t-1}, \dots, X_1, \lambda_1 \sim \text{NegBin}(\lambda_t, \psi)$$

$$\lambda_t = \nu_t + \phi_t X_{t-1} + \kappa \lambda_{t-1}$$

$$\log(\nu_t) = \alpha^{(\nu)} + \gamma^{(\nu)} \sin(2\pi t/52) + \delta^{(\nu)} \cos(2\pi t/52)$$

$$\log(\phi_t) = \alpha^{(\phi)} + \gamma^{(\phi)} \sin(2\pi t/52) + \delta^{(\phi)} \cos(2\pi t/52)$$

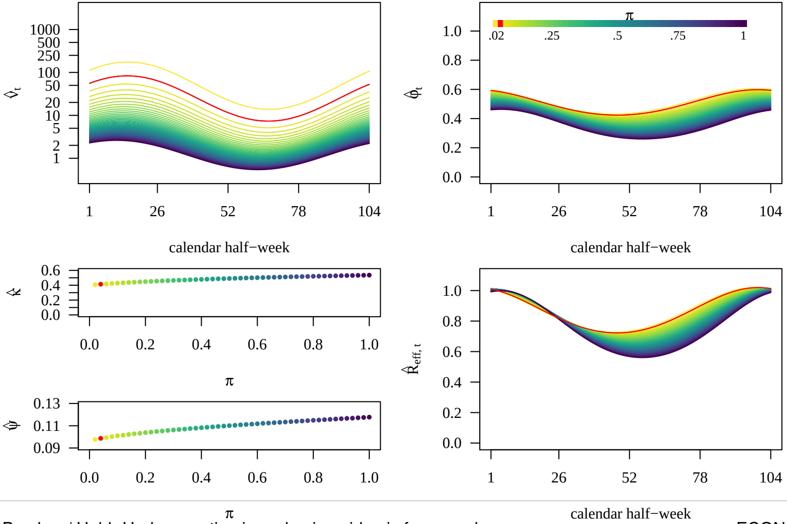
$$\tilde{X}_t \mid X_t \sim \text{Bin}(X_t, \pi),$$

$$\tilde{X}_t^* = \tilde{X}_{2t-1} + \tilde{X}_{2t},$$

where we vary $\pi \in \{0.02, 0.04, \dots, 0.98, 1\}$

Results

Estimated parameters as a function of the assumed reporting probability π:



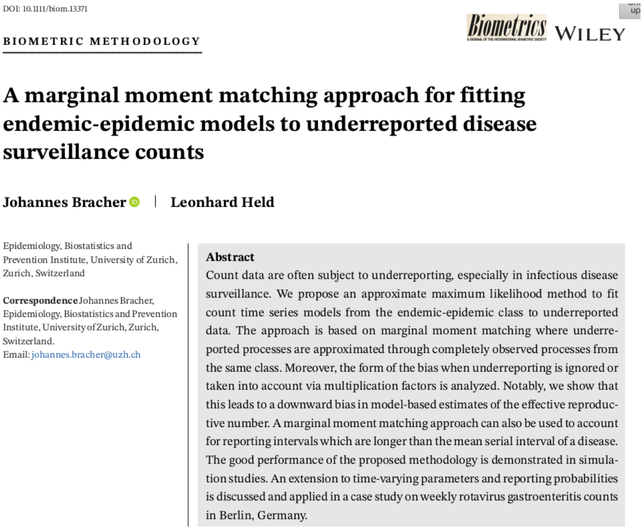
Interpretation

- Assuming $\pi = 0.043$ (Weidemann et al 2014) we obtain
  - a mean serial interval of 6.0 days
  - a local effective reproductive number between 0.72 and 1.02
  - an estimated one in ten cases imported from outside Berlin

- Assuming complete reporting ($\pi = 1$) we would get
  - a mean serial interval of 7.5 days
  - a local effective reproductive number between 0.56 and 1.00
  - an estimated one in seven cases imported from outside Berlin

→ Considerable dependence of results on assumptions on reporting.

- Caution: Even given strong commuter streams to Berlin, one imported case in ten seems high – other biases may be at work.



DOI: 10.1111/biom.13371

BIOMETRIC METHODOLOGY

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts

Johannes Bracher | Leonhard Held

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

Correspondence Johannes Bracher, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland.
Email: johannes.bracher@uzh.ch

Abstract

Count data are often subject to underreporting, especially in infectious disease surveillance. We propose an approximate maximum likelihood method to fit count time series models from the endemic-epidemic class to underreported data. The approach is based on marginal moment matching where underreported processes are approximated through completely observed processes from the same class. Moreover, the form of the bias when underreporting is ignored or taken into account via multiplication factors is analyzed. Notably, we show that this leads to a downward bias in model-based estimates of the effective reproductive number. A marginal moment matching approach can also be used to account for reporting intervals which are longer than the mean serial interval of a disease. The good performance of the proposed methodology is demonstrated in simulation studies. An extension to time-varying parameters and reporting probabilities is discussed and applied in a case study on weekly rotavirus gastroenteritis counts in Berlin, Germany.

References

- Bauer, C. and Wakefield, J. (2018). Stratified spacetime infectious disease modelling, with an application to hand, foot and mouth disease in China. Journal of the Royal Statistical Society: Series C (Applied Statistics), (available online first, DOI 10.1111/rssc.12284).
- Bracher, J. and Held, L. (2020). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. Biometrics, in press.
- Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fevre, E. M., and Kretzschmar, M. E. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health, 14(1):1–17.
- Grimwood, K., Abbott, G.D., Fergusson, D.M., Jennings, L.C. and Allan, J.M. (1983). Spread of rotavirus within families: a community based study. British Medical Journal, 287(6392), 575–577.
- Held, L., Höhle, M., and Hofmann, M. (2005). *A statistical framework for the analysis of multivariate infectious disease surveillance counts.* Statistical Modelling, 5(3):187–199.

References (contd)

- King, A., Nguyen, D. and Ionides, E. (2016) Statistical inference for partially observed Markov processes via the R package pomp. Journal of Statistical Software, 69(12), 1–43.
- Weidemann, F., Dehnert, M., Koch, J., Wichmann, O., and Höhle, M. (2014). Bayesian parameter inference for dynamic infectious disease modelling: Rotavirus in Germany. Statistics in Medicine, 33(9):1580–1599.
- Zhu, F. (2011) A negative binomial integer-valued GARCH model. Journal of Time Series Analysis, 32(1), 54–67.

The same in a few increasingly unwieldy formulas (1)

Stationary second-order properties of $\{X_t\}$ (Zhu 2011):

$$\mu = \frac{\nu}{1 - \phi - \kappa}, \tag{1}$$

$$\sigma^2 = \frac{1 - (\phi + \kappa)^2 + \phi^2}{1 - (\phi + \kappa)^2 - \psi\phi^2} \cdot (\mu + \psi\mu^2), \tag{2}$$

$$\rho(d) = \eta\xi^{d-1}, \quad d = 1, 2, \ldots \tag{3}$$

where

$$\eta = \phi \cdot \frac{1 - \kappa(\phi + \kappa)}{1 - (\phi + \kappa)^2 + \phi^2} \tag{4}$$

$$\xi = \phi + \kappa. \tag{5}$$

→ **has an ARMA(1, 1) covariance structure.**

The same in a few increasingly unwieldy formulas (2)

Stationary second-order properties of $\{\tilde{X}_t\}$ with $\tilde{X} \mid X_t \sim \text{Bin}(X_t, \pi)$:

$$\begin{aligned} \tilde{\mu} &= \pi\mu, && (6)\\ \tilde{\sigma}^2 &= \pi^2\sigma^2 + \pi(1-\pi)\mu, && (7)\\ \tilde{\rho}(d) &= \tau\eta\xi^{d-1}, \quad d = 1, 2, \dots, && (8) \end{aligned}$$

where

$$\tau = \frac{\tilde{\sigma}^2 - (1-\pi)\tilde{\mu}}{\tilde{\sigma}^2}$$

$\rightarrow$ **Still an ARMA(1, 1) covariance structure!**

The same argument in a few unwieldy formulas (3)

Consider a process $\{X_t\}$ with parameters ν, ϕ, κ, ψ and associated underreported process $\{\tilde{X}_t\}$ with reporting probability π.

For most π_Y (and any $\pi_Y > \pi$) there is a pair of latent process $\{Y_t\}$ and underreported process $\{\tilde{Y}_t\}$ such that $\{\tilde{X}_t\}$ and $\{\tilde{X}_t\}$ are second-order equivalent. The parameters of $\{Y_t\}$ are given by

$$\nu_Y(\pi_Y) = \frac{\pi}{\pi_Y} \cdot \nu \tag{9}$$

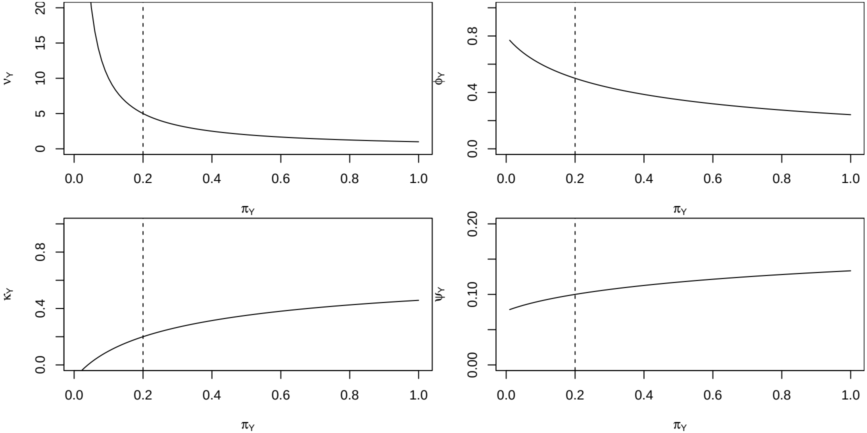
$$\phi_Y(\pi_Y) = \frac{\sqrt{\tau_Y^2(1-\xi^2)^2 + 4(\tau_Y\xi - \tau\eta)\tau\eta(1-\xi^2)} - \tau_Y(1-\xi^2)}{2(\tau_Y\xi - \tau\eta)} \tag{10}$$

$$\text{with } \tau_Y = \frac{\tilde{\sigma}^2 - (1-\pi_Y)\tilde{\mu}}{\tilde{\sigma}^2}$$

$$\kappa_Y(\pi_Y) = \xi - \phi_Y \tag{11}$$

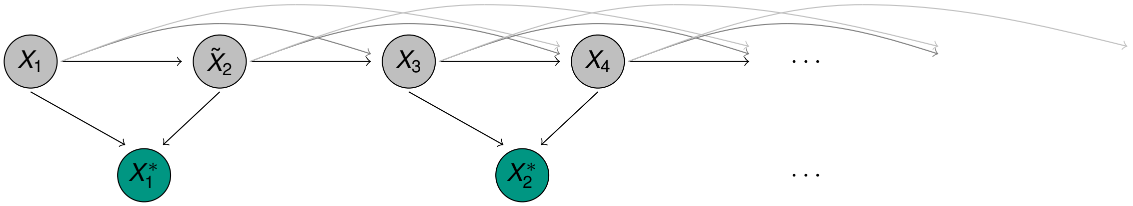
$$\psi_Y(\pi_Y) = \frac{\{\tilde{\sigma}^2 - (1-\pi_Y)\tilde{\mu}\}(1-\xi^2) - \pi_Y\tilde{\mu}\{1-\xi^2+\phi_Y^2\}}{\phi_Y^2\{\tilde{\sigma}^2 - (1-\pi_Y)\tilde{\mu}\} + \tilde{\mu}^2\{1-\xi^2+\phi_Y^2\}}. \tag{12}$$

Illustration



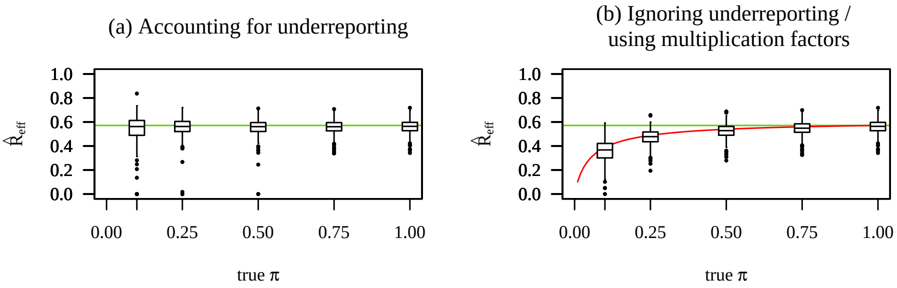
Side note: accounting for temporal aggregation

The same general idea can be used to approximate the likelihood of models defined at a “half-weekly” scale given weekly data:



This is useful for diseases with a mean serial interval below one week, such as rotavirus.

Simulation results


(a) Accounting for underreporting
(b) Ignoring underreporting / using multiplication factors
$\hat{R}_{eff}$
true π