Estimating parameters from data
Part I: epidemic data

Niel Hens

www.simid.be
www.socialcontactdata.org

Trento, 2019

Background

- Niel Hens, MMath, MSc BioStat, PhD BioStat
- Scientific chair in evidence-based vaccinology @UAntwerpen
- Professor of Biostatistics @UHasselt
- co-director of SIMID: www.simid.be

- Disclaimer: the course material presented here gives a flavour of methods and does not constitute a comprehensive overview. Comments can be made about the different methods and there exists improved methodology elsewhere.

Overview

Mathematical modelling of infectious diseases

- Purposes:
  - prediction: requires the inclusion of known complexities and population-level heterogeneity
  - understanding: investigating the factors that drive dynamics
- Building a model presents a trade-off:
  - accuracy: reproduce what is observed and predict future dynamics
  - transparency: ability to understand how model components influence the dynamics and interact
  - flexibility: ease of adapting the model to new situations

Mathematical modelling of infectious diseases

- Limitations:
  - models present a simplification of reality
  - chance events of infectious disease transmission hinder perfect prediction
- A good model:
  - suited to its purpose: simple as possible, but no simpler
  - balance accuracy, transparency, flexibility
  - parametrisable from available data

Mathematical modelling of infectious diseases

- Daniel Bernoulli was the first to present a mathematical model for smallpox in 1760, published in 1766
- Since then many people have developed models to describe infectious disease dynamics, see e.g. Bailey [1975], Anderson and May [1992], Grenfell and Dobson [1995], Daley and Gani [1999], Hethcote [2000]
- Several textbooks are available; amongst which Vynnycky and White [2010] provides an excellent introduction to mathematical modelling

Contact between individuals

- Predicting the number of infections at time $t+1$ based on the circumstance at time t
- The force of infection λ
  - the per capita rate at which a susceptible individual contracts infection
  - it is assumed proportional to the number of infectious persons at time t and that it is given by

$$\lambda_t = \beta I_t$$

- The per capita rate at which two specific individuals come into effective contact per unit time β
  - Denote c_e the number of effective contacts made by each person per unit time; β is then given by

$$\beta = c_e/N$$

Contact between individuals

- The number of new infections at time $t+1$ is given by $\lambda_t S_t$ and thus:

  $$I_{t+1} = \beta S_t I_t = c_e S_t I_t / N_t$$

  This is referred to as the mass action principle
- density-dependent transmission: $\beta = c_e / N$ remains constant:
  - as the population size increases, so does the number of effective contacts
  - mostly applicable to plant and animal diseases (homogeneity)
- frequency-dependent transmission: $\beta = c_e / N_t$ changes with time:
  - the number of individuals effectively contacted is assumed constant regardless of a change in population size
  - mostly applicable to human and vectorborne diseases (heterogeneity)

Contact between individuals

- Note that when in a constant population, both density- and frequency-dependent are equivalent
- One can write
$$I_{t+1} = c_e S_t I_t / N_t^\gamma,$$
with $0 \leq \gamma \leq 1$.
- In what follows, I will use
$$I_{t+1} = \beta S_t I_t,$$
without loss of generality.

Key questions

- What are the factors influencing epidemic dynamics?
- What can we learn from the early stages of an epidemic?
- What is likely to be the size of an epidemic?

The basic reproduction number

- The discrete time SIR model:

$$\begin{aligned} S_{t+1} &= S_t - \beta I_t S_t \\ I_{t+1} &= I_t + \beta I_t S_t - \nu I_t \\ R_{t+1} &= R_t + \nu I_t \end{aligned}$$

- Assumptions:
  - time spent in classes follows exp. distribution, the mean infectious period:

$$D = 1/\nu$$

  - closed population:

$$N_{t+1} = S_{t+1} + I_{t+1} + R_{t+1} = S_t + I_t + R_t = N_t$$

- Dynamics are determined by

initial condition: $(N - 1, 1, 0)$, parameters: (β, ν)

The basic reproduction number

- The continuous time SIR model:

$$\begin{aligned} S'(t) &= -\beta I(t)S(t) \\ I'(t) &= \beta I(t)S(t) - \nu I(t) \\ R'(t) &= \nu I(t) \end{aligned}$$

- Assumptions:
  - time spent in classes follows exp. distribution, the mean infectious period:

$$D = 1/\nu$$

  - closed population:

$$N'(t) = 0$$

- Dynamics are determined by

initial condition: $(N-1, 1, 0)$ parameters: (β, ν)

The basic reproduction number

- The basic reproduction number, R_0:

$$R_0 = \beta N D.$$

- The effective reproduction number, R_e:

$$R_e(t) = s(t)R_0 = \beta S(t) D.$$

- The herd immunity threshold:

$$1 - 1/R_0.$$

Estimation from data

- Methods to estimate R_0 depend on the available information
  - time series data
  - epidemic tree data
  - final size data
  - ...
- The formulas used here assume homogeneous mixing and are therefore approximate
- However they are useful!

Overview

Growth rate models

- Assume we are in the early stage of an epidemic:

$$I(t) \approx I(0)e^{\Lambda t}.$$

where Λ is called the growth rate of the epidemic, $I(t)$ is the number of infectious individuals at time t

- If we take the logarithm of both sides:

$$\log(I(t)) = \log(I(0)) + \Lambda t.$$

- So how can we estimate Λ?

Class Exercise

- Show using $I'(t) = \beta I(t)S(t) - \gamma I(t)$ that $I(t) \approx I(0)e^{\Lambda t}$. What is Λ?
- Use the 1976 Ebola outbreak data
- Plot the number of cases by time
- Can you identify a reasonable time range for the initial epidemic phase?
- Use the data in this time range to estimate Λ

R-code growth rate estimation

```
# Zaire Ebola 1976 data
data=read.table("ZaireEbola1976.txt")
names(data)=c("time","It")
plot(data$time,data$It)
abline(v=19-0.5)
lm(log(data$It)~data$time,subset=data$time<19)
```

Some issues

- Number of cases or cum. number of cases?
- The statistical model used here is approximate.
- What about underreporting, delays in reporting, reporting biases?
- What if we don't observe exponential growth?
- How do we relate the growth rate to R_0?

The generalized growth model

- Slower-than-exponential (sub-exponential) growth patterns can occur due to spatial heterogeneity, behavioural changes, etc.
- The generalized growth model (GGM) is given by:

$$\frac{dC(t)}{dt} = C'(t) = r * C(t)^p,$$

- where
  - $r * C(t)^p$: incidence curve over time t
  - $C(t)$: cumulative number of cases at time t
  - r: growth rate
  - $p \in [0, 1]$: growth scaling parameter
- $0 < p < 1$ describes sub-exponential growth (Viboud *et al.*, 2016)

The generalized growth model

- Least Squares typically used $\Rightarrow$ implicitly assumes error terms are independent and identically distributed with constant variance
- use GLM framework: Poisson model is a natural choice

$$y_t|C(t) \sim \text{Poisson}\big(\mu_t\big)$$

  - where y_t is incidence observed at time t; $\mu_t = r * C(t)^p$
- overdispersion common in practice due to unobserved heterogeneity, consider:

$$y_t|C(t), \xi_t \sim \text{Poisson}\big(\xi_t * \mu_t\big)$$

  - where ξ_t is a random error term uncorrelated with $C(t)$
  - assume ξ_t is Gamma white noise (Bretó *et al.*, 2009)

The generalized growth model

- Assuming $\xi_t \sim \text{Gamma}\left(\frac{1}{\theta}, \theta\right)$, it can be shown that integrating out ξ_t gives:

$$y_t | C(t) \sim \text{Negative binomial}\left(\frac{\theta^{-1}}{\mu_t + \theta^{-1}}, \theta^{-1}\right)$$

  - where θ is a dispersion parameter
- the conditional mean and variance are given by

$$E\Big(y_t | C(t)\Big) = \mu_t \text{ and,}$$

$$Var\Big(y_t | C(t)\Big) = \mu_t + \theta \mu_t^2,$$

- the Poisson distribution is obtained as $\theta \to 0$
- for each model, $\Theta_{Poisson} = \{r, p\}$ and $\Theta_{NB} = \{r, p, \theta\}$ can be estimated together within the framework of classical maximum likelihood theory

The generalized growth model

- Extensive simulation study comparing Poisson and Neg Bin (Ganyani et al. in rev.):
  - little evidence for bias for both models
  - lower coverage for the Poisson model due to underestimation of variance $\Rightarrow$ narrower confidence intervals
  - for the Poisson model coverage goes down with more data (the more the data the "greater" the precision) $\Rightarrow$ even narrower confidence intervals
  - Type I error increases with decrease in coverage $\Rightarrow$ more chance to incorrectly identify growth pattern as sub-exponential
- Conclusion: In practice Poisson model to be avoided even if overdispersion is unsuspected - inference using Neg Bin model is practically indistinguishable under assumption of equidispersion

R_0 and growth rates

- In case the pre-infectious period is short in comparison with the infectious period and the infectious period is assumed to follow an exponential distribution:

$$R_0 = 1 + \Lambda D,$$

  where D is the average infectious period.

- In case the pre-infectious period and the infectious period both follow an exponential distribution:

$$R_0 = (1 + \Lambda D)(1 + \Lambda D'),$$

  where D' and D are the average pre-infectious and infectious period, respectively.

R_0 and growth rates

- In case the pre-infectious and infectious periods are unknown but assumed to follow an exponential distribution but the serial interval is known:

$$R_0 = 1 + \Lambda T_s$$

- Given $R_0 = 1 + \Lambda D$, and assumptions, one can show:

$$R_0 = 1 + \frac{\log(2)}{T_d} D,$$

where T_d is the doubling time

Class exercises

- What is R_0 knowing that the pre-infectious period and infectious period are both exponentially distributed and on average 2 days long?
- If $R_0 = 2$ and our pre-infectious and infectious period are exponential and both, on average, 2 days long. What is the growth rate of the epidemic?
- What happens in the previous situation when the pre-infectious period is very short? What is the doubling time in this situation?

Some illustrations: modelling Ebola

PLOS | ONE

RESEARCH ARTICLE

Spatiotemporal Evolution of Ebola Virus Disease at Sub-National Level during the 2014 West Africa Epidemic: Model Scrutiny and Data Meagreness

Eva Santermans[1]*, Emmanuel Robesyn[2], Tapiwa Ganyani[1], Bertrand Sudre[2], Christel Faes[1], Chantal Quinten[2], Wim Van Bortel[2], Tom Haber[3], Thomas Kovac[1,3], Frank Van Reeth[3], Marco Testa[2,4], Niel Hens[1,5], Diamantis Plachouras[2]

1 Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium, **2** European Centre for Disease Prevention and Control, Stockholm, Sweden, **3** Expertise centre for Digital Media, iMinds, tUL, Diepenbeek, Belgium, **4** Department of Public Health, University of Turin, Turin, Italy, **5** Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

CrossMark
click for updates

Santermans et al. [2016]

Some illustrations: nonlinear mass action

DOI: 10.1002/sim.7935

RESEARCH ARTICLE

WILEY Statistics in Medicine

Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter

Tapiwa Ganyani[1] | Christel Faes[1] | Gerardo Chowell[2,3] | Niel Hens[1,4]

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, UHasselt (Hasselt University), Diepenbeek, Belgium

[2]School of Public Health, Georgia State University, Atlanta, Georgia

[3]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institute of Health, Bethesda, Maryland

[4]Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

Correspondence
Tapiwa Ganyani, Interuniversity Institute for Biostatistics and statistical Bioinformatics, UHasselt (Hasselt University), Diepenbeek, Belgium.
Email: tapiwa.ganyani@uhasselt.be

The standard mass action, which assumes that infectious disease transmission occurs in well-mixed populations, is popular for formulating compartmental epidemic models. Compartmental epidemic models often follow standard mass action for simplicity and to gain insight into transmission dynamics as it often performs well at reproducing disease dynamics in large populations. In this work, we formulate discrete time stochastic susceptible-infected-removed models with linear (standard) and nonlinear mass action structures to mimic varying mixing levels. Using simulations and real epidemic data, we demonstrate the sensitivity of the basic reproduction number to these mathematical structures of the force of infection. Our results suggest the need to consider nonlinear mass action in order to generate more accurate estimates of the basic reproduction number although its uncertainty increases due to the addition of one growth scaling parameter.

KEYWORDS
basic reproduction number, discrete time stochastic SIR model, early epidemic growth phase, epidemic modeling, mass action principle

Some considerations

- Growth rate models are naturally embedded into the well-known Generalized Linear Modeling framework
- Extensions exist and allow for taking into account e.g. spatial effects (not shown here)
- Linking growth models with mechanistic approaches relies on making assumptions and are thus approximate in that sense
- Because of their non-mechanistic nature predicting the size of an epidemic using growth models is not always meaningful (not shown here)

Incidence data

- Fitting more general incidence data patterns is hard
- Choice of model is much more important
  - Becker and Britton [1999]: maximum likelihood and martingale theory
  - Finkenstädt and Grenfell [2000]: dynamical systems approach applied to measles
  - many developments since
- Many parameters to be fitted:
  - Latin Hypercube Sampling from reasonable parameter ranges
  - study correlations between parameters
- More later on . . .

Overview

The Mukden 1946 Epidemic

Tieh, T.H., Landauer, E., Miyagawa, F., Kobayashi, G., Okayasu, G. (1948). Primary pneumonic plague in Mukden, 1946, and report of 39 cases with 3 recoveries. Journal of Infectious Diseases, 82; 52-58.

- The index case traveled to Mukden from Russia
- He was infected prior to arrival to Mukden
- This man infected the family he was visiting
- All together there were 39 cases of pneumonic plague with only 3 recoveries

The Mukden 1946 Epidemic

- The control measures began on the 12th day of the epidemic
    - Isolation and quarantine of all patients and contacts
    - Other citizens wore masks
    - Besides, there was vaccination against bubonic plague
- Lasted 19 days after implementing control measures
- 28 positives out of 39 cases (laboratory test)
- 4 persons were asymptomatic

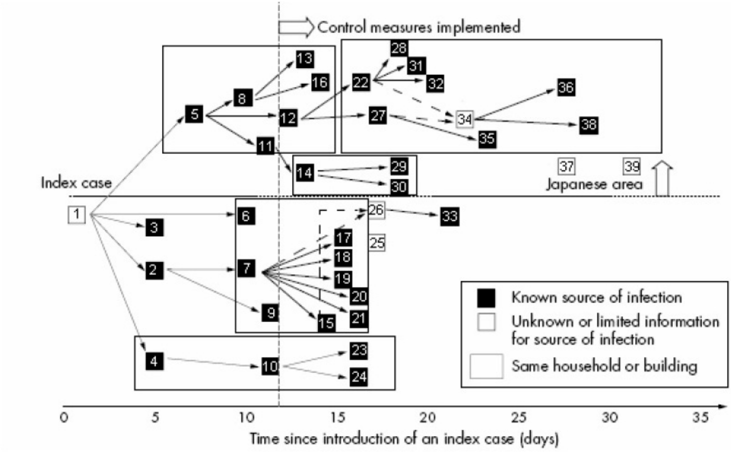
Pneumonic Plague

- Pneumonic plague
- Airborne infection
- 95% mortality rate
- Bioterrorism
- Epidemic tree

The Epidemic Tree

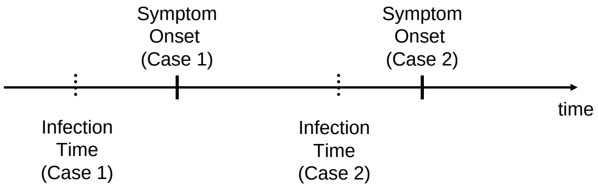
- Directed network or transmission tree
- The nodes or boxes represent disease cases
- The edges represent the transmission of the disease
- Each case has exactly one primary case
- Thus each node can have at maximum only one incoming edge
- There cannot be any edges from a node to itself

The Epidemic Tree



The Generation Interval

Proxy: Serial Interval



Notation

- Denote
  - t_i: the time of symptom onset for individual $i = 1, \ldots, n$
  - $t_{v(i)}$: the time of symptom onset for the source case for individual i
  - $X_i = t_i - t_{v(i)}$: the generation interval for individual i
- Assume $X_2, \ldots, X_n \sim g(\boldsymbol{x}|\boldsymbol{\theta})$
- We define the loglikelihood

$$\ell(\boldsymbol{\theta}|\mathbf{t}) = \sum_{\mathbf{i=2}}^{\mathbf{n}} \log \mathbf{g}(\mathbf{t_i} - \mathbf{t_{v(i)}}|\boldsymbol{\theta})$$

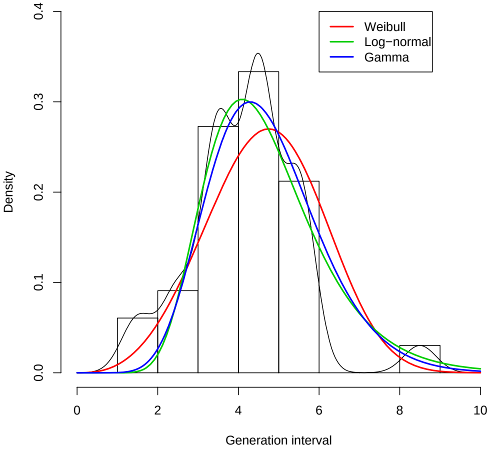
Estimating the Generation Interval Distribution

$$\boxed{g(t_i - t_{v(i)}|\boldsymbol{\theta})}$$

- Leads to the basic reproduction number R_0,
- and the effective reproduction number R_{eff}
- Can be used to find missing links!

- Gani & Leach (2004): Lognormal distribution
- Nishuira et al. (2006): Gamma distribution
- Generation interval of length 0: Weibull distribution

Result


Weibull
Log−normal
Gamma
Density
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
Generation interval

Incomplete Data Problem

- Literature:
  - Haydon et al. (2003): bootstrap approach
  - Wallinga and Teunis (2004): likelihood-based approach
  - Nishiura et al. (2006): likelihood-based, bootstrap approach
- Incomplete data modelling approach:
  - ignorable missingness - EM-algorithm
  - non-ignorable missingness - auxiliary information/sens. analysis

Likelihood-based approach

- Probability model
- Identify the most likely case
- $g(t_i - t_j|\hat{\boldsymbol{\theta}}_{ML})$
- Probability that the ith case is infected by the jth case

$$p_{ij} = \frac{g(t_i - t_j|\hat{\boldsymbol{\theta}}_{ML})}{\sum_{k \neq i} g(t_i - t_k|\hat{\boldsymbol{\theta}}_{ML})}$$

- Note: $g(t_i - t_j|\hat{\boldsymbol{\theta}}_{ML}) = 0$ if $t_i < t_j$ and $p_{ii} = 0$.
- Missing links and unlikely links

EM-algorithm

- Denote $\boldsymbol{X} \sim g(x|\boldsymbol{\theta})$ the observed generation intervals
- Denote $\mathbf{Z} \sim g(x|\boldsymbol{\theta})$ the unobserved generation intervals
- Let k $(n-k-1)$ be the number of missing (observed) links
- The complete data likelihood:

$$L^{\mathsf{C}}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}) = \prod_{i=2}^{n-k} g(x_i|\boldsymbol{\theta}) \prod_{i=1}^{k} g(z_i|\boldsymbol{\theta})$$

- Since z is unknown, we use the expected complete data loglikelihood

$$\mathrm{E}(\ell^{\mathsf{C}}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z})) = \sum_{\boldsymbol{z}} \ell^{\mathsf{C}}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}) h(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$$

EM-algorithm

- Assume given $\boldsymbol{\theta}$, the unobserved pairs are independent of the observed pairs
- We then calculate (E-step)

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k, \boldsymbol{x}) = \mathrm{E}(\ell^{\mathsf{C}}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})|\boldsymbol{\theta}_k, \boldsymbol{x}) = \sum_{\boldsymbol{z}} \ell^{\mathsf{C}}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) h(\boldsymbol{z}|\boldsymbol{\theta}_k)$$

- Here $h(\boldsymbol{z}|\boldsymbol{\theta}_k)$ is obtained using the expected transmission probabilities $g(t_i - t_j|\boldsymbol{\theta}_k)/\sum_{k \neq i} g(t_i - t_k|\boldsymbol{\theta}_k)$
- The M-step is then given by

$$\boldsymbol{\theta}_{k+1} = \mathrm{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k, \boldsymbol{x})$$

- We iterate the E- and M-step until convergence
- Use different starting values for $\boldsymbol{\theta}_0$

Incorporating Prior Knowledge

- Prior knowledge: same household, ...
- Until now

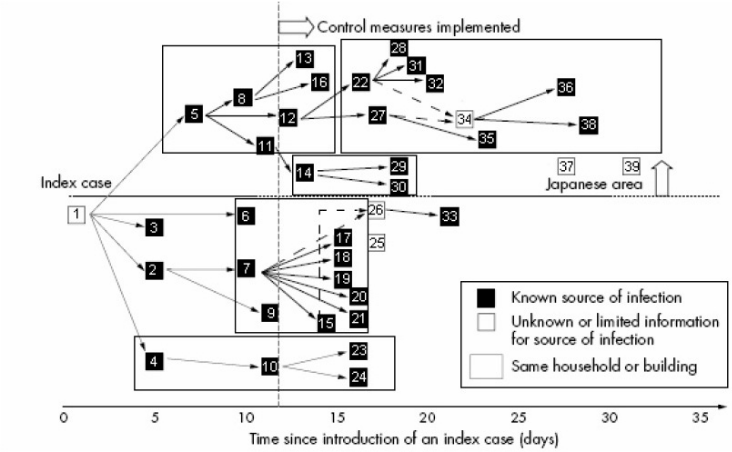
$$X \sim_{iid} g(x|\boldsymbol{\theta})$$

- Let us assume prior probabilities π_{ij} and call this the PEM (Prior-based EM)

$$p_{ij} = \frac{g(t_i - t_j|\boldsymbol{\theta}) \cdot \pi_{ij}}{\sum_{k \neq i} g(t_i - t_k|\boldsymbol{\theta}) \cdot \pi_{ik}}$$

- The original EM corresponds to an uninformative prior

Results


Control measures implemented
Index case
Japanese area
Known source of infection
Unknown or limited information for source of infection
Same household or building
0 5 10 15 20 25 30 35
Time since introduction of an index case (days)

Prior Knowledge

- Five unknown links: Cases 25, 26, 34, 37 and 39
- Case 25: 0.143 weight for cases 7, 15, 17-21 (Lee Family)
- Case 26: 0.5 weight for cases 2 or 7
- Case 34: 0.5 weight for case 22 or 27
- Case 37 and 39: weight 0.0625 and 0.0526 (Japanese area, infected before)

Data Analysis

- Gamma distribution
- 3 most likely sources:

Missing Link	Likelihood	EM-algorithm	PEM-algorithm
Case 25	14 (0.152)	14 (0.153)	15 (0.505)
	13 (0.152)	13 (0.153)	7 (0.228)
	12 (0.137)	12 (0.138)	
Case 26	14 (0.152)	14 (0.153)	7 (0.998)
	13 (0.152)	13 (0.153)	2 (0.002)
	12 (0.137)	12 (0.138)	
Case 34	30 (0.101)	30 (0.101)	27 (0.621)
	29 (0.101)	29 (0.101)	22 (0.379)
	28 (0.101)	28 (0.101)	
...			
Gamma Shape	11.57	11.72	11.02
Gamma Scale	0.41	0.40	0.43
Mean	4.70	4.70	4.75
Variance	1.91	1.88	2.04

Identifying Unlikely Links

- One option: posterior probabilities p_{ij}
  - p_{ij} is calculated conditional on the estimated $\hat{\boldsymbol{\theta}}$
  - influenced by excessively short (long) reported serial intervals
- Another option:

  global influence measures (Cook, 1982; Zhu et al. 2001)

Identifying Unlikely Links

- Define $\hat{\boldsymbol{\theta}}_{[-i]}$ the estimate of $\boldsymbol{\theta}$ with the i-th serial interval assumed unobserved for all analyses

- The global influence measure is then defined as

$$\mathsf{GI}_i^{\mathcal{F}}(\hat{\boldsymbol{\theta}}) = \mathcal{F}(\hat{\boldsymbol{\theta}}) - \mathcal{F}(\hat{\boldsymbol{\theta}}_{[-i]}),$$

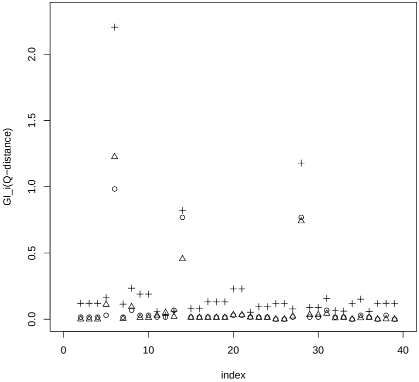
  where $\mathcal{F}$ could be any function of $\hat{\boldsymbol{\theta}}$: shape, scale, mean, ...

- "Q-distance"

$$\mathcal{F}(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{x}),$$

  $\rightarrow$ Zhu et al. (2001) with observed interval reclassified as non-observed

Data Analysis: Q-distance



Q-distance influence measures for the Mukden dataset using the likelihood method ($\circ$), the EM-algorithm ($\triangle$) and the PEM-algorithm ($+$), respectively. On the horizontal axis, the index value of the individuals are given while on the vertical axis the Q-distance is depicted.

Discussion

- Robust reconstruction of the epidemic tree: Hens et al. [2012]

  More flexible approach: MCMC [see e.g. te Beest et al., 2013]

- Concerns about contraction of the generation interval due to
  - depletion of susceptibles [Scalia Tomba et al., 2010]
  - competition of infectors [Kenah et al., 2008]

  $\rightarrow$ Hazard-based approach

Overview

Final size and R_0

- Assume we have the final size of our epidemic. If $R_0 < 3$ we can estimate it using the following methods:
  - Denote z_f the proportion of the population which has been infected by the end of the epidemic and assume that all individuals are susceptible at the start of the epidemic:

$$R_0 = -\frac{\log(1 - z_f)}{z_f}$$

  - Assume that s_0 and s_f are the proportions of the population susceptible at the start and at the end of the epidemic, respectively:

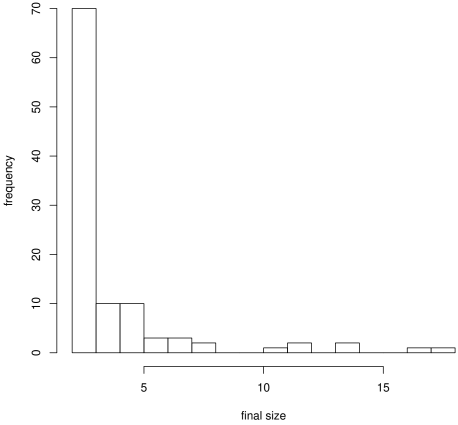
$$R_0 = \frac{\log(s_f) - \log(s_0)}{s_f - s_0}$$

  - ...
- Though these equations are deterministic and don't take stochasticity into account.

Distribution of final sizes: Hepatitis A outbreak data

- 113 outbreaks
- 5 foodborne outbreaks: excluded from the analysis
- Available information:
  - Final sizes
  - 5 provinces: Antwerp, East-Flanders, Flemish Brabant, Limburg, West-Flanders
  - Auxiliary information on outbreak nature: school, family, children
- Number of source cases unknown

Distribution of final sizes: Hepatitis A outbreak data



Issue: underreporting due to long incubation period (15-45 days)

Estimating the Reproduction Number

- Assume we know the number of source cases
- Final size: total number of infected persons
- Becker 1974: final size distribution

$$P(X = x; s) = b(x, s)\frac{\theta^{x-s}}{A(\theta)^x},$$

  where

  - $S = s$ initial cases
  - $X = x$ is the final size
  - θ is the Reproduction Number
  - $b(x, s)$ is constant.
- Power series family

Estimating the Reproduction Number

- Poisson offspring distribution: Borel-Tanner distribution (Haight and Breuer, 1960)

$$P(X = x; s) = \frac{s x^{x-s-1} \theta^{x-s} e^{-x\theta}}{(x-s)!}, \quad x = s, s+1, \ldots.$$

- Geometric offspring distribution:

$$P(X = x; s) = \frac{s}{2x - s} \binom{2x - s}{x - s} \frac{\theta^{x-s}}{(1+\theta)^{2x-s}}, \quad x = s, s+1, \ldots.$$

- Defined for $X < \infty$ and $\theta \leq 1$

Estimating the Reproduction Number

- Farrington, Kanaan and Gay (2003)
- $X = \infty$:

$$P(X = \infty) = 1 - q(\theta)^s$$

- Censored likelihood:

$$L(\theta; s, x, c) = \prod_{i=1}^{n} \left\{ P(X = x_i; s_i)^{c_i} \left(1 - \sum_{j=s_i}^{x_i - 1} P(X = j, s_i) \right)^{1-c_i} \right\}$$

- Profile likelihood confidence interval

Unknown Number of Source Cases

- Crucial assumption: conditioning $S = s$
  - assume initial case distribution
  - derive the joint and marginal likelihood $\{x \geq s\}$
- Crucial assumption: random mixing
  - include covariates: e.g. provinces
- Computational problem: large outbreaks are impossible

Initial Case Distribution

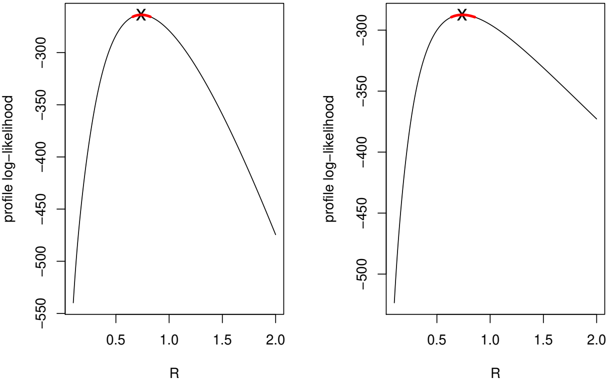
- Degenerate: $S = 1$
- Discrete cases: $S = 1, 2$
- Truncated Poisson
- Truncated Negative Binomial
- ...
- Comparison by goodness-of-fit

Underreporting

- Sensitivity Analysis
  - Censoring for school outbreaks
  - Uniform underreporting factors
  - Varying underreporting factors: school, family, w/o children

Fixed number of source cases

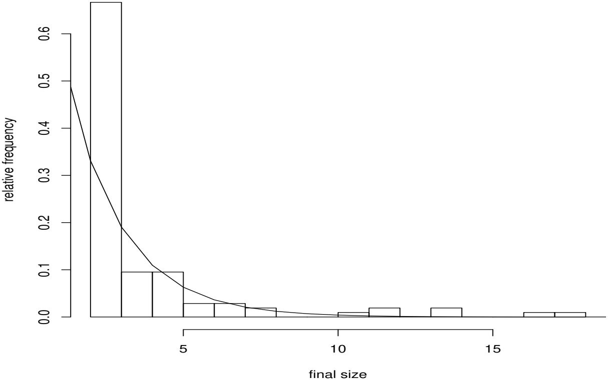
- Assume 1 source case for each outbreak



- Borrel-Tanner: $R = 0.74\ (0.66, 0.83)$
- Geometric Offspring: $R = 0.74\ (0.64; 0.86)$

Fixed number of source cases

- Goodness-of-fit (Geometric Offspring)



Unknown number of source cases

- Unknown number of source cases (Geometric Offspring):
  - Degenerate: 527.58
  - Poisson: 450.90
  - Discrete: 369.07
- Discrete: $R = 0.48$, $P(S = 2) \approx 1$
- Empirical Bayes estimates can give you the most likely value per outbreak

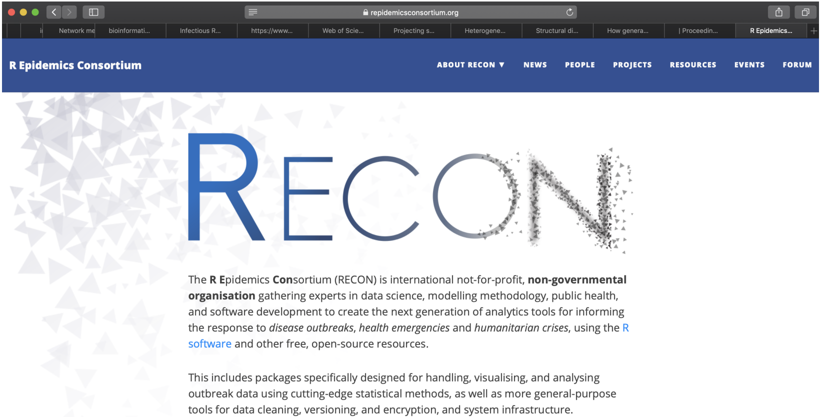
Discussion

- Closed form likelihood solution available for simple cases
- Conditioning on $R < 1$ can be undesirable
- Alternative approach: use stochastic SIR model and a simulation-based approach: computationally expensive - efficient algorithm by Black and Ross (2015)

General Discussion

- Growth rate models, Epidemic tree data, Final size data
- Many items haven't been covered: renewal equations, household models, . . .
- Take home message:
  - tailored methods are needed for inference
  - crucial to list assumptions made

The RECON initiative towards outbreak analysis



References

R. Anderson and R. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1992.

N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Charles Griffin and Company, London, 1975.

N. G. Becker and T. Britton. Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):287–307, 1999. doi: 10.1111/1467-9868.00177. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00177`.

D. Daley and J. Gani. *Epidemic Modelling: An Introduction*. Cambridge University Press, 1999.

B. F. Finkenstädt and B. T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(2):187–205, 2000. doi: 10.1111/1467-9876.00187. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00187`.

B. Grenfell and A. Dobson. *Ecology of Infectious Disease in Natural Populations*. Cambridge, UK: Cambridge University Press, 1995.

N. Hens, L. Calatayud, S. Kurkela, T. Tamme, and J. Wallinga. Robust reconstruction and analysis of outbreak data: influenza a(h1n1)v transmission in a school-based population. *American Journal of Epidemiology*, 2012.

H. Hethcote. The mathematics of infectious diseases. *SIAM REview*, 42(4):599–653, 2000.

E. Kenah, M. Lipsitch, and J. Robins. Generation interval contraction and epidemic data analysis. *Mathematical Biosciences*, 213:71–79, 2008.

E. Santermans, E. Robesyn, T. Ganyani, B. Sudré, C. Faes, C. Quinten, W. Van Bortel, T. Haber, T. Kovac, F. Van Reeth, M. Testa, N. Hens, and D. Plachouras. Spatiotemporal evolution of ebola virus disease at sub-national level during the 2014 west africa epidemic: Model scrutiny and data meagreness. *PLoS ONE*, 11(1):e0147172, 2016.

G. Scalia Tomba, A. Svensson, T. Asikainen, and J. Giesecke. Some model based considerations on observing generation times for communicable diseases. *Mathematical Biosciences*, 223:24–31, 2010.

D. E. te Beest, J. Wallinga, T. Donker, and M. van Boven. Estimating the generation interval of influenza a (h1n1) in a range of social settings. *Epidemiology (Cambridge, Mass.)*, 24:244–250, Mar. 2013. ISSN 1531-5487. doi: 10.1097/EDE.0b013e31827f50e8.

E. Vynnycky and R. White. *An Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.