Inference for emerging epidemics among a community of households

Frank Ball

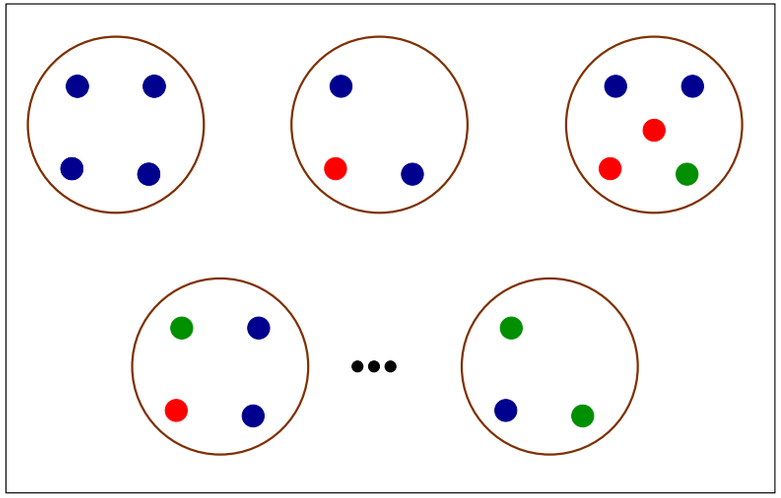
Frank.Ball@nottingham.ac.uk

University of Nottingham

III Workshop on Branching Processes and their Applications, Badajoz, Spain, 7–10 April 2015

Joint work with Laurence Shaw (University of Nottingham)

Households SIR epidemic model



m_n households of size n
$(n = 1, 2, \cdots, n_{\max})$

total no. of households $m = \sum_{n=1}^{n_{\max}} m_n$

total no. of individuals $N = \sum_{n=1}^{n_{\max}} n m_n < \infty$

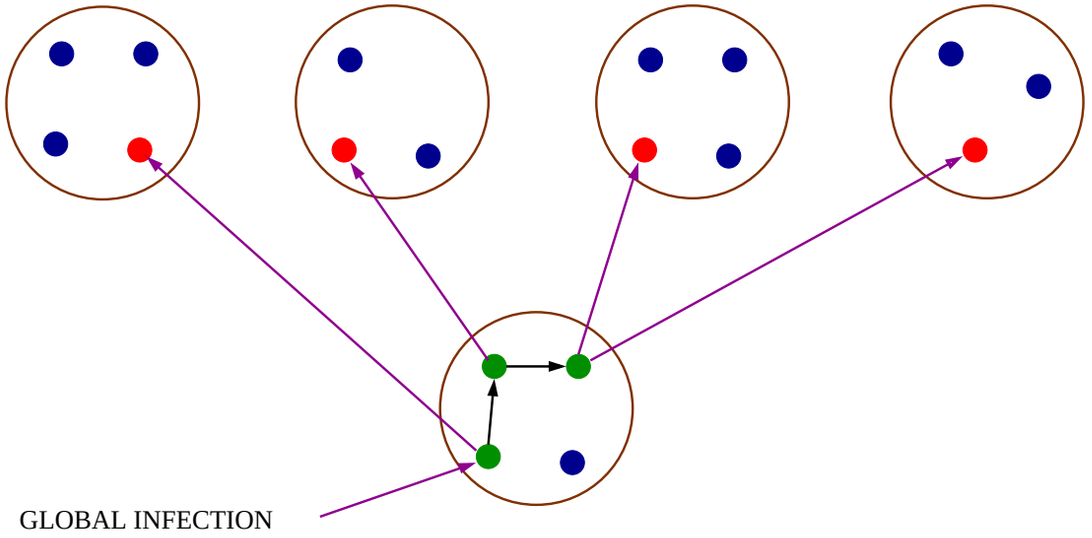
- SIR (susceptible $\rightarrow$ infective $\rightarrow$ recovered)
- Infectious period $\sim T_I$, having an arbitrary but specified distribution
- Infection rates (individual $\rightarrow$ individual)
  - (i) local (within-household) λ_L
  - (ii) global (between-household) λ_G/N
- Latent period/infectivity profiles

(Bartoszyński (1972), Becker and Dietz (1995), Ball, Mollison and Scalia-Tomba (1997))

Why study households models?

- Household structure is a key departure from homogeneous mixing for human populations and can have significant impact on disease dynamics
- There are outbreak control measures associated with households and similar structures (e.g. schools and workplaces)
- Epidemic data are often collected at the household level
- Households models are mathematically reasonably tractable and consequently are generally easier to interpret than complex simulation models

Threshold parameter R_*



- R_* = mean number of global contacts emanating from a typical single-household epidemic

$$R_* = \sum_{n=1}^{n_{\max}} \tilde{\alpha}_n \mu_n(\lambda_L) \lambda_G \mathrm{E}[I],$$

where

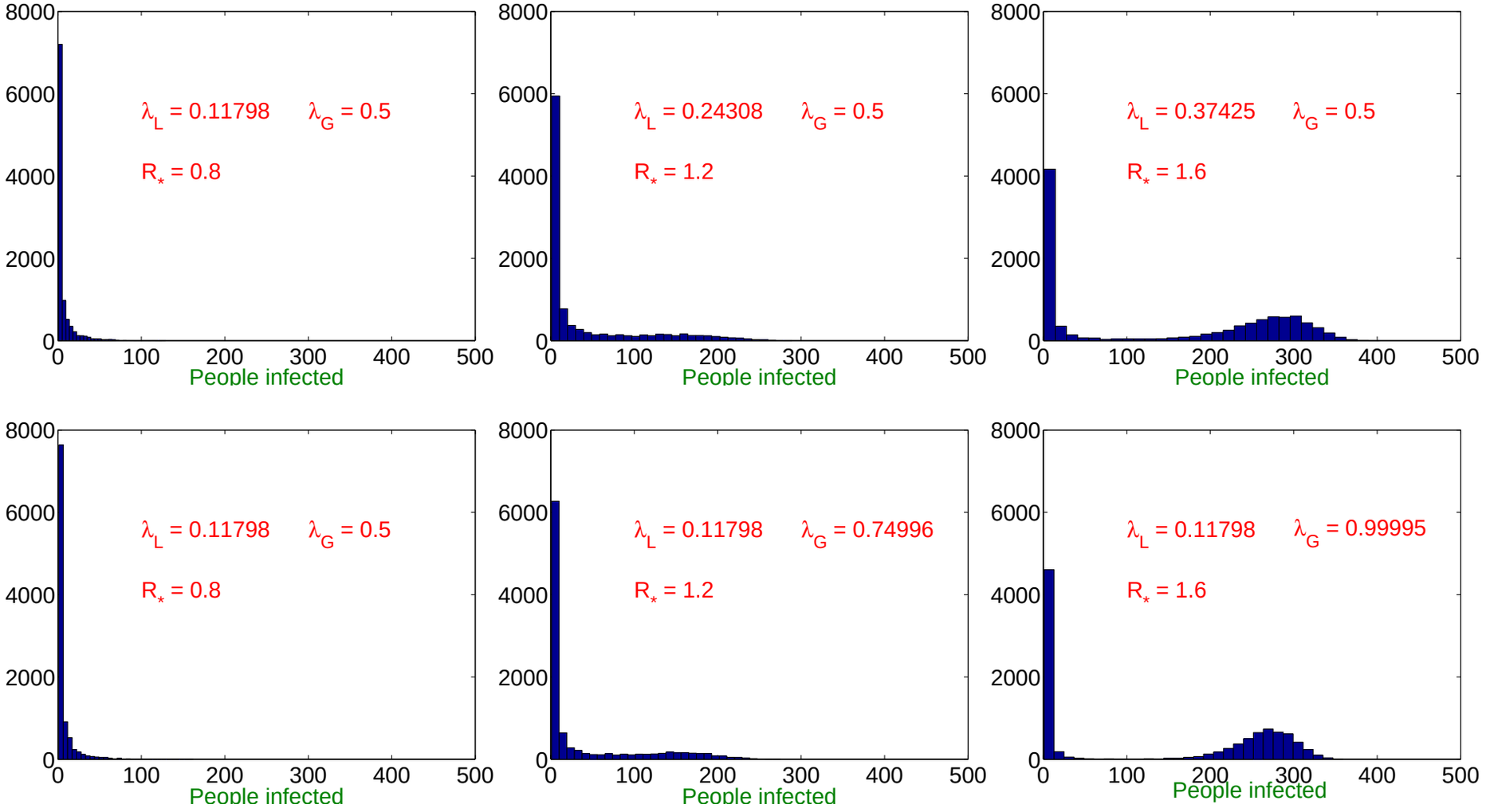
$\tilde{\alpha}_n = \frac{n m_n}{N}$ = P(randomly chosen person lives in a household of size n)

$\mu_n(\lambda_L)$ = mean size of single (size-n) household epidemic with 1 initial infective

- $\mathrm{P}(\text{global epidemic}) > 0 \iff R_* > 1$

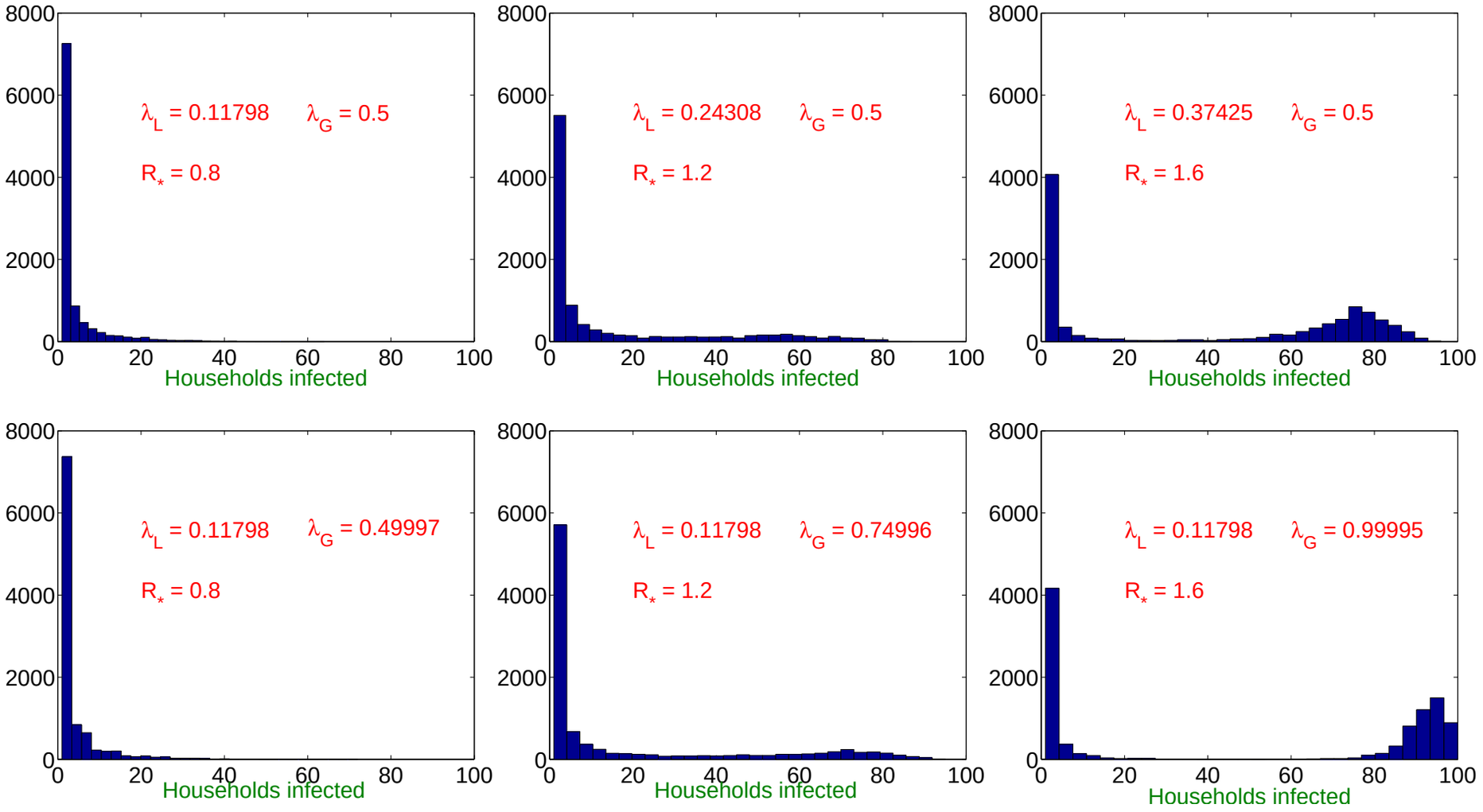
(Ball, Mollison and Scalia-Tomba (1997), Becker and Dietz (1995))

Number of people infected



Number of people infected in each set of 10,000 simulations, $T_I \equiv 1$, population consisting of 100 households of size 5

Number of households infected



Number of households infected in each set of 10,000 simulations, $T_I \equiv 1$, population consisting of 100 households of size 5

Estimation in an emerging epidemic

- Suppose that an epidemic is observed in its emerging phase and
  - population household-size distribution is known (e.g. from census data);
  - an estimate of the early exponential growth rate r of the epidemic is available;
  - more-detailed, household-level data are available in a sample of households.
- Goal is to estimate local infection rate λ_L.
- If the distribution of infectious period T_I is known, (λ_L, r) determines the global infection parameter λ_G.

Estimation in an emerging epidemic

- Consider an SIR epidemic among $1,000,000$ households, with

$$(\alpha_1, \alpha_2, \cdots, \alpha_6) = (0.29, 0.35, 0.16, 0.14, 0.04, 0.02),$$

where α_i is the fraction of households having size i. Suppose that

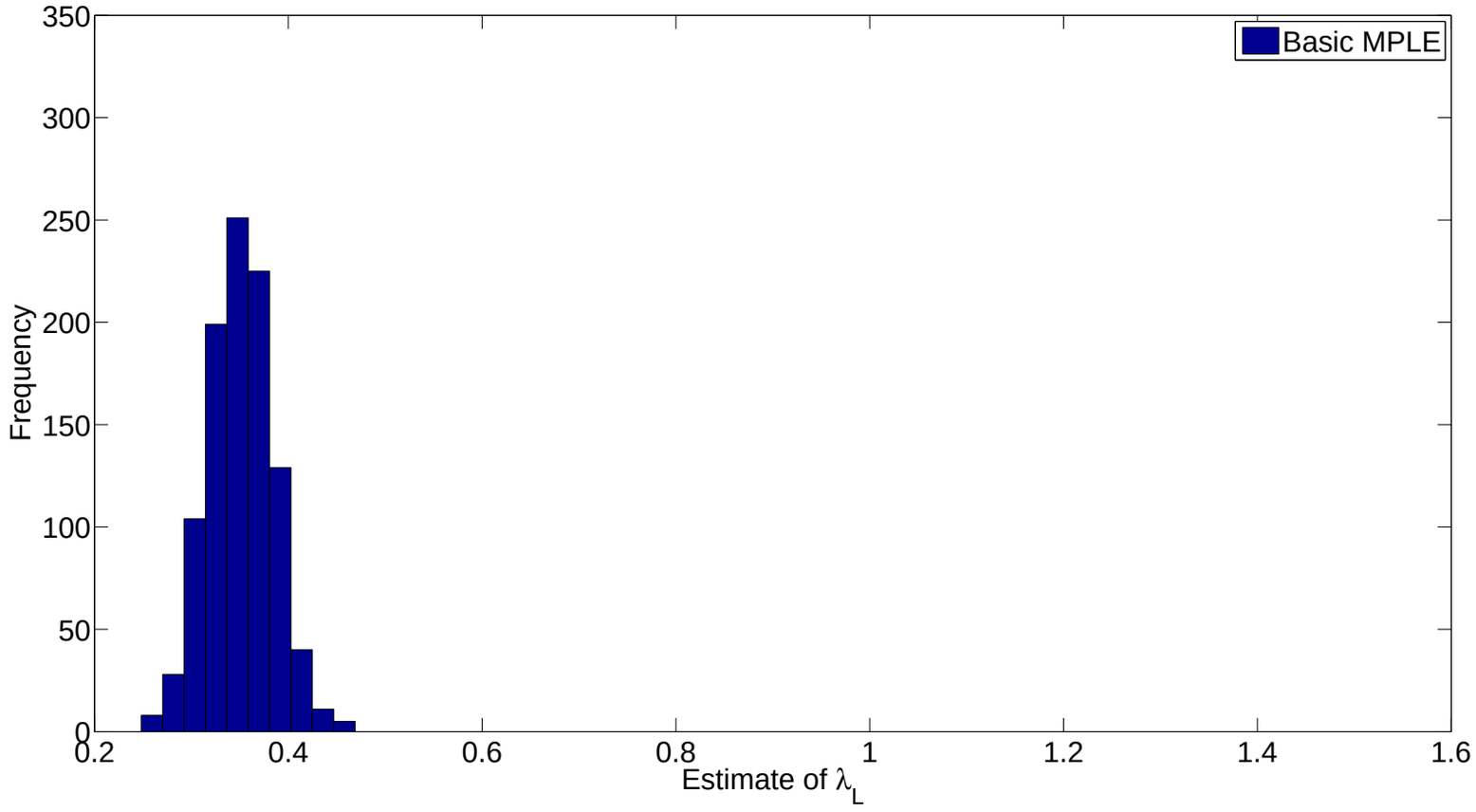
$$\lambda_L = 1, \lambda_G = 1 \text{ and } T_I \sim \text{Exp}(1).$$

Cf. the influenza example in Fraser (2007).

- After $1,000$ individuals have recovered, estimate λ_L by fitting final size distribution $p_n(\cdot|\lambda_L)$, to observed completed single-household outbreaks.

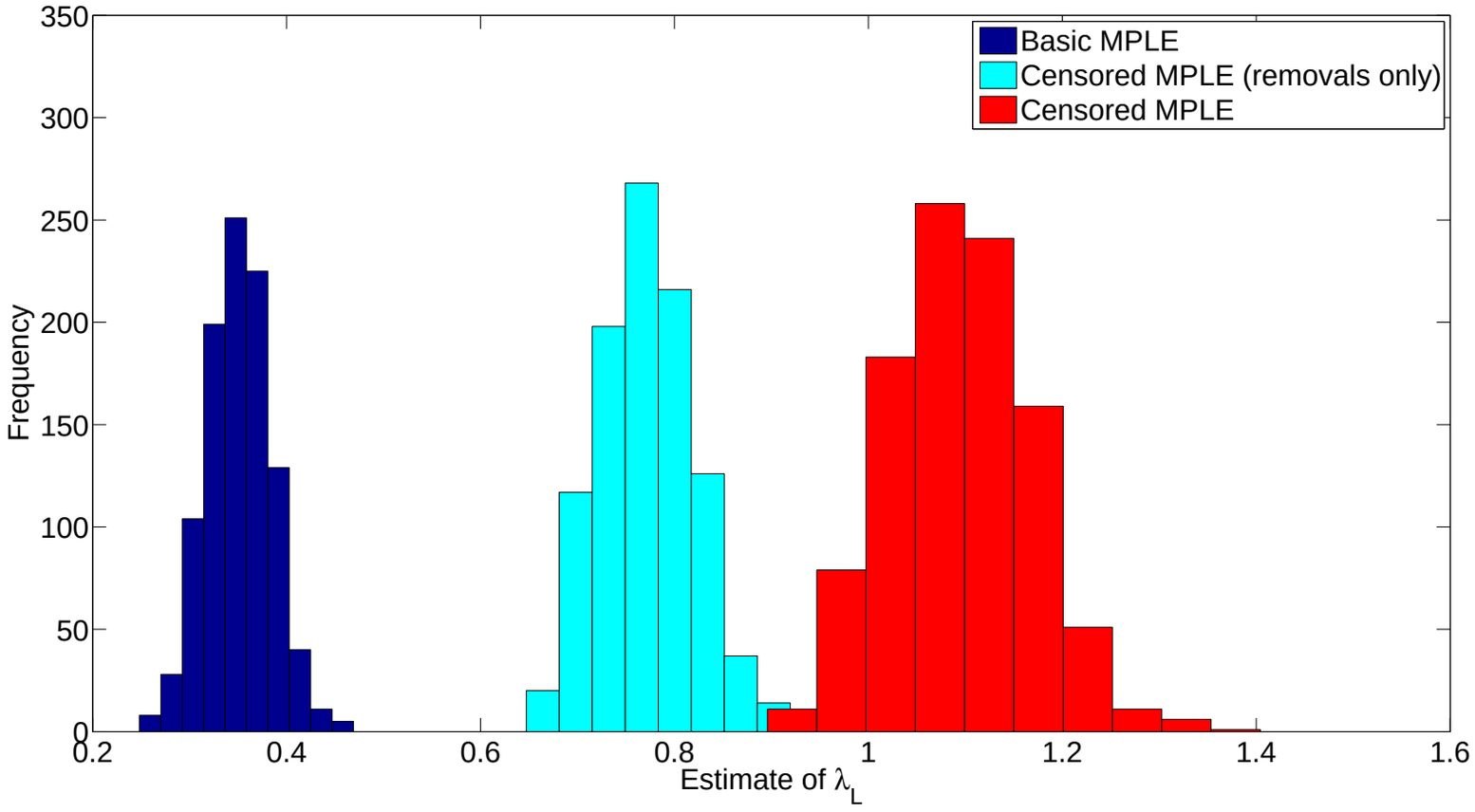
- $p_n(k|\lambda_L)$ $(k = 1, 2, \cdots, n)$ is the probability that a single-size-n-household epidemic, with 1 initial infective and no global infection, has k recovered cases in total.

Estimation in an emerging epidemic



Histogram of estimates of within-household infection rate λ_L based on $1,000$ simulated epidemics with $\lambda_L = 1$ and $\lambda_G = 1$ that took off.

Estimation in an emerging epidemic



Histograms of estimates of within-household infection rate λ_L based on $1,000$ simulated epidemics with $\lambda_L = 1$ and $\lambda_G = 1$ that took off.

Single-household epidemic

- Let $E_H^{(n)}$ denote a typical size-n single-household epidemic, started by one household member being infected at time $t = 0$.
- For $t \geq 0$, let $X_H^{(n)}(t)$ and $Y_H^{(n)}(t)$ be the numbers of susceptibles and infectives in $E_H^{(n)}$ at time t.
- Let $\mathcal{T}^{(n)} = \{(x, y) : \ x = 0, 1, ..., n-1; \ y = 0, 1, ..., n-x\}$ be the state space for $\left\{ \left(X_H^{(n)}(t), Y_H^{(n)}(t) \right) : \ t \geq 0 \right\}$.
- For $(x, y) \in \mathcal{T}^{(n)}$, let

$$p_{x,y}^{(n)}(t|\lambda_L) = \mathrm{P}(X_H^{(n)}(t) = x, \ Y_H^{(n)}(t) = y) \quad (t \geq 0)$$

and

$$\tilde{p}_{x,y}^{(n)}(r|\lambda_L) = \int_0^\infty e^{-rt} p_{x,y}^{(n)}(t|\lambda_L) \ \mathrm{d}t \quad (r \geq 0).$$

Approximating branching process

- Let E^∞ denote the general (Crump-Mode-Jagers) branching process which approximates the early stages of the epidemic in a community of households, in which individuals correspond to single-household epidemics and an individual reproduces in E^∞ whenever a global contact emanates from the corresponding single-household epidemic.
- For $n = 1, 2, \cdots, n_{\max}$, let $\xi^{(n)}$ be the point process of ages at which a typical size-n individual in E^∞ reproduces and let $\mu^{(n)}(t) = \mathrm{E}[\xi^{(n)}([0,t])]$ $(t \geq 0)$. Then

$$\mu^{(n)}(\mathrm{d}t) = \lambda_G \sum_{(x,y)\in\mathcal{T}^{(n)}} y p_{x,y}^{(n)}(t|\lambda_L)\ \mathrm{d}t.$$

- Let ξ be the point process of ages at which a typical individual in E^∞ reproduces and $\mu(t) = \mathrm{E}[\xi([0,t])]$ $(t \geq 0)$.

Approximating branching process

- A typical individual in E^{∞} has household size distributed according to the size-biased distribution $\tilde{\alpha}_n$ $(n = 1, 2, \cdots, n_{\max})$, so

$$\mu(\mathrm{d}t) = \sum_{n=1}^{n_{\max}} \tilde{\alpha}_n \mu^{(n)}(\mathrm{d}t) = \lambda_G \sum_{n=1}^{n_{\max}} \tilde{\alpha}_n \sum_{(x,y)\in\mathcal{T}^{(n)}} y p_{x,y}^{(n)}(t|\lambda_L)\ \mathrm{d}t.$$

- Suppose that $R_* > 1$. Then E^{∞} has a Malthusian parameter $r > 0$ given by

$$\int_0^{\infty} \mathrm{e}^{-rt} \mu(\mathrm{d}t) = 1.$$

- Note that r satisfies

$$\lambda_G \sum_{n=1}^{n_{\max}} \tilde{\alpha}_n \sum_{(x,y)\in\mathcal{T}^{(n)}} y \tilde{p}_{x,y}^{(n)}(r|\lambda_L) = 1.$$

Approximating branching process

- Assume that individuals live forever in E^{∞}.
- For $n = 1, 2, \cdots, n_{\max}$ and $(x, y) \in \mathcal{T}^{(n)}$, an individual is said to be in state (n, x, y) if it corresponds to a single size-n household epidemic and there are x susceptibles and y infectives in that epidemic.
- For $t \geq 0$ and $(n, x, y) \in \mathcal{T} = \{(n, x, y) : \ n = 1, 2, \cdots, n_{\max}$ and $(x, y) \in \mathcal{T}^{(n)}\}$, let $Y_{n,x,y}(t)$ be the number of individuals in state (n, x, y) at time t in E^{∞}.
- Suppose that $R_* > 1$. Then, using Nerman (1981), there exists a random variable $W \geq 0$, where $W = 0 \iff E^{\infty}$ goes extinct, such that for all $(n, x, y) \in \mathcal{T}$,

$$\mathrm{e}^{-rt} Y_{n,x,y}(t) \xrightarrow{a.s.} \tilde{\alpha}_n \tilde{p}^{(n)}_{x,y}(r|\lambda_L) W \quad \text{as} \quad t \to \infty.$$

Approximating branching process

- Recall that if $R_* > 1$ then, for all $(n, x, y) \in \mathcal{T}$,

$$\mathrm{e}^{-rt} Y_{n,x,y}(t) \xrightarrow{a.s.} \tilde{\alpha}_n \tilde{p}^{(n)}_{x,y}(r|\lambda_L) W \quad \text{as} \quad t \to \infty,$$

  where $W = 0 \iff E^\infty$ goes extinct.

- Note that, for $n = 1, 2, \cdots, n_{\max}$,

$$\sum_{(x,y)\in\mathcal{T}^{(n)}} p^{(n)}_{x,y}(t|\lambda_L) = 1 \implies \sum_{(x,y)\in\mathcal{T}^{(n)}} \tilde{p}^{(n)}_{x,y}(r|\lambda_L) = \frac{1}{r}.$$

- Thus, if E^∞ does not go extinct, as $t \to \infty$, the proportion of individuals that are in state (n, x, y) converges almost surely to $\tilde{\alpha}_n r \tilde{p}^{(n)}_{x,y}(r|\lambda_L)$.

- These yield the correct probabilities for an emerging epidemic.

Estimation in an emerging epidemic

- Suppose that
  - an estimate $\hat{r}$ of the growth rate r is available;
  - the epidemic has taken off, is still in its exponentially growing phase but has been running sufficiently long for the asymptotic composition of the branching process E^{∞} to be applicable.
- For $(n, x, y) \in \mathcal{T}$, let $a_{x,y}^{(n)}$ be the number of size-n households with x susceptibles and y infectives when estimation is made.
- Assuming complete knowledge of the current state of each single-household epidemic, λ_L may be estimated by maximising the normalised "pseudolikelihood" function

$$L_{\text{full}}(\lambda_L|\boldsymbol{a}, \hat{r}) = \prod_{n=2}^{n_{\max}} \prod_{(x,y)\in\mathcal{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(\hat{r}|\lambda_L)^{a_{x,y}^{(n)}}.$$

Estimation in an emerging epidemic

- Suppose that estimation is based only on completed single-household epidemics. Then λ_L may be estimated by maximising

$$L_{\text{final}}(\lambda_L|\boldsymbol{a},\hat{r}) = \prod_{n=2}^{n_{max}} \prod_{x=0}^{n-1} \tilde{p}_{x,0}^{(n)}(\hat{r}|\lambda_L)^{a_{x,0}^{(n)}}.$$

- Subject to mild conditions,

$$\lim_{t\to\infty} p_{x,0}^{(n)}(t|\lambda_L) = \lim_{r\to 0^+} r\tilde{p}_{x,0}^{(n)}(r|\lambda_L),$$

so using the usual single-household final size distribution yields "unbiased" estimates as the growth rate $r \downarrow 0$.

Estimation in an emerging epidemic

- Suppose that only recoveries are observed.
- For $n = 1, 2, \cdots, n_{\max}$ and $j = 1, 2, \cdots, n$, let $c_j^{(n)}$ be the observed number of size-n households with j recoveries, $\mathcal{A}_j^{(n)} = \{(x, y) \in \mathcal{T}^{(n)} : \ x + y = n - j\}$ and

$$\tilde{q}_j^{(n)}(r|\lambda_L) = \sum_{(x,y)\in\mathcal{A}_j^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\lambda_L)/(\frac{1}{r} - \tilde{q}_0^{(n)}(r|\lambda_L)),$$

  where

$$\tilde{q}_0^{(n)}(r|\lambda_L) = \sum_{y=1}^{n} \tilde{p}_{n-y,y}^{(n)}(r|\lambda_L).$$

- Then λ_L may be estimated by maximising

$$L_{\text{rec}}(\lambda_L|\boldsymbol{c}, \hat{r}) = \prod_{n=2}^{n_{max}} \prod_{j=1}^{n} \tilde{q}_j^{(n)}(\hat{r}|\lambda_L)^{c_j^{(n)}}.$$

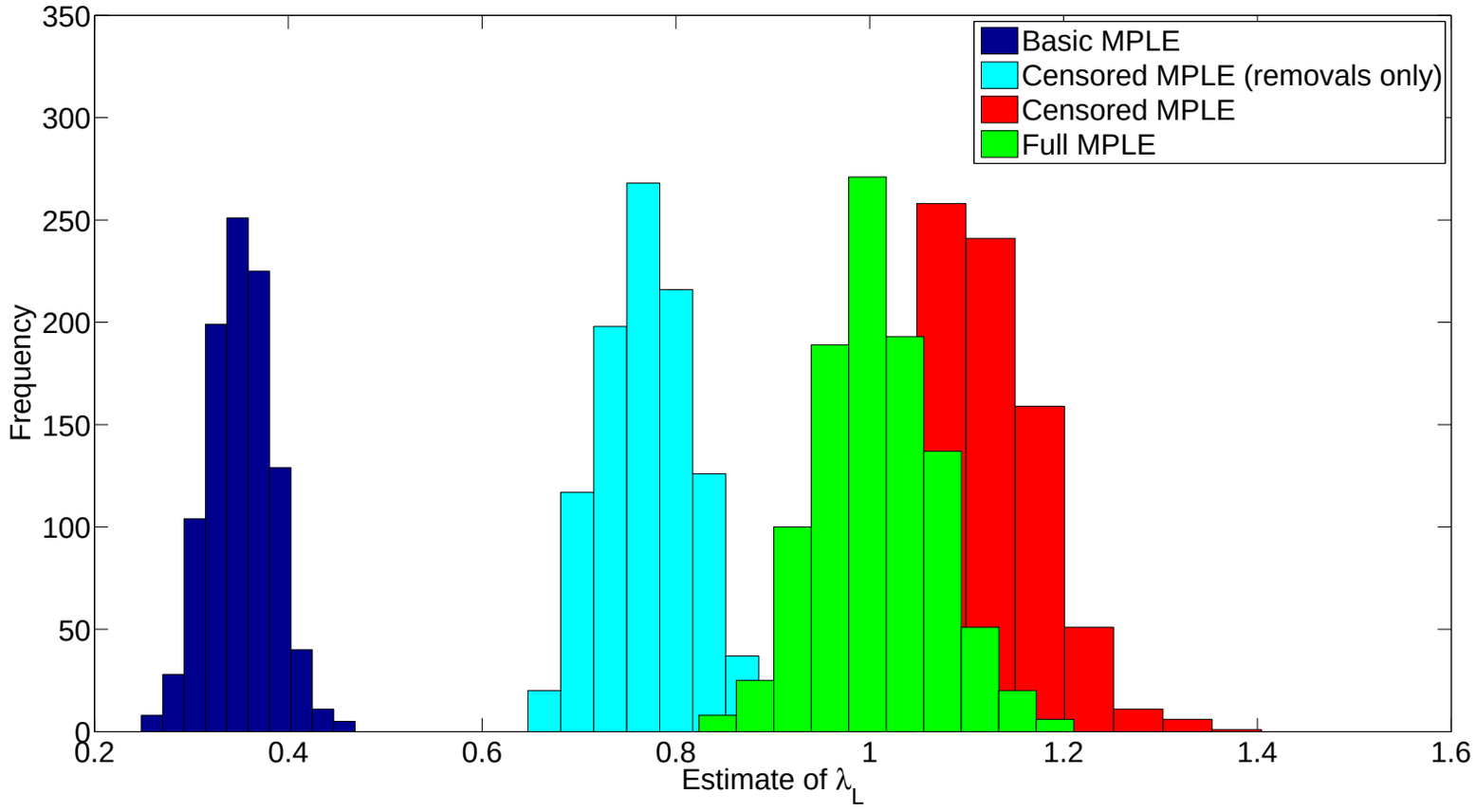
Calculation of $\tilde{p}^{(n)}_{x,y}(r|\lambda_L)$

- Generally difficult!
- If $T_I \sim \mathrm{Exp}(\gamma)$ then $\{(X_H^{(n)}(t), Y_H^{(n)}(t)) : t \geq 0\}$ is a continuous-time Markov chain with state space $\mathcal{T}^{(n)}$.
- Assign labels $1, 2, \cdots, s^{(n)}$ to states in $\mathcal{T}^{(n)}$, where $s^{(n)} = |\mathcal{T}^{(n)}|$ $(= n(n+3)/2)$.
- Let $P^{(n)}(t)$ and $Q^{(n)}$ be the transition-probability and transition-rate matrices of $\{(X_H^{(n)}(t), Y_H^{(n)}(t)) : t \geq 0\}$ using this labelling. Then

$$P^{(n)}(t) = \exp(tQ^{(n)}) \implies \int_0^\infty e^{-rt} P^{(n)}(t)\ \mathrm{d}t = (rI_{s^{(n)}} - Q^{(n)})^{-1} \quad (r > 0),$$

  and $\tilde{p}^{(n)}_{x,y}(r|\lambda_L)$ follows.
- Extends to case when T_I has a phase-type distribution.

Estimation in an emerging epidemic



Histograms of estimates of within-household infection rate λ_L based on $1,000$ simulated epidemics with $\lambda_L = 1$ and $\lambda_G = 1$ that took off.

Estimation in an emerging epidemic

- Consider an SIR epidemic among $10,000$ households, with

$$(\alpha_1, \alpha_2, \cdots, \alpha_6) = (0.13, 0.30, 0.23, 0.18, 0.09, 0.07),$$

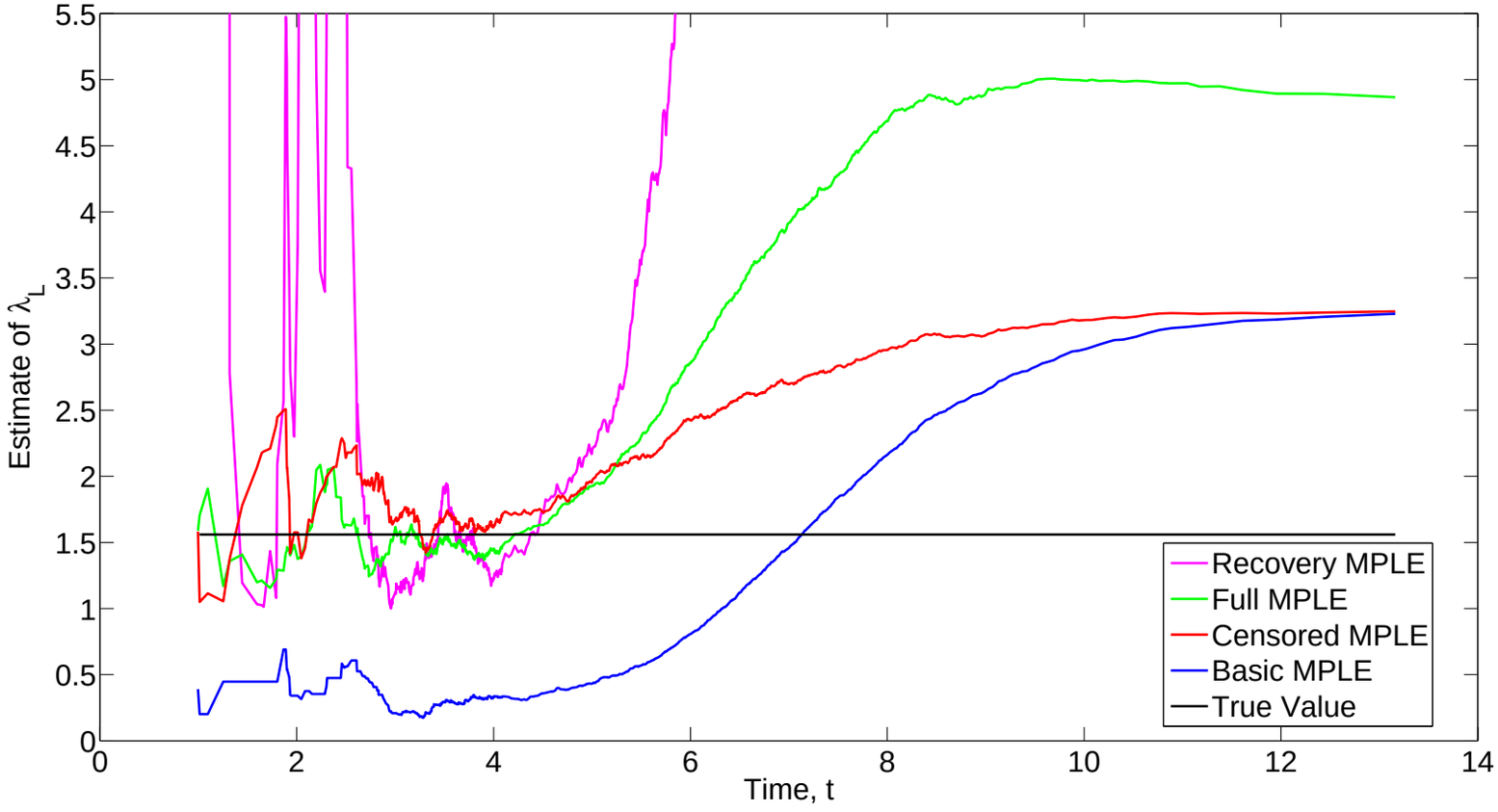
where α_i is the fraction of households having size i. Suppose that

$$\lambda_L = 1.56, \lambda_G = 1.21 \text{ and } T_I \sim \text{Exp}(1).$$

Cf. the varicella example in Fraser (2007).

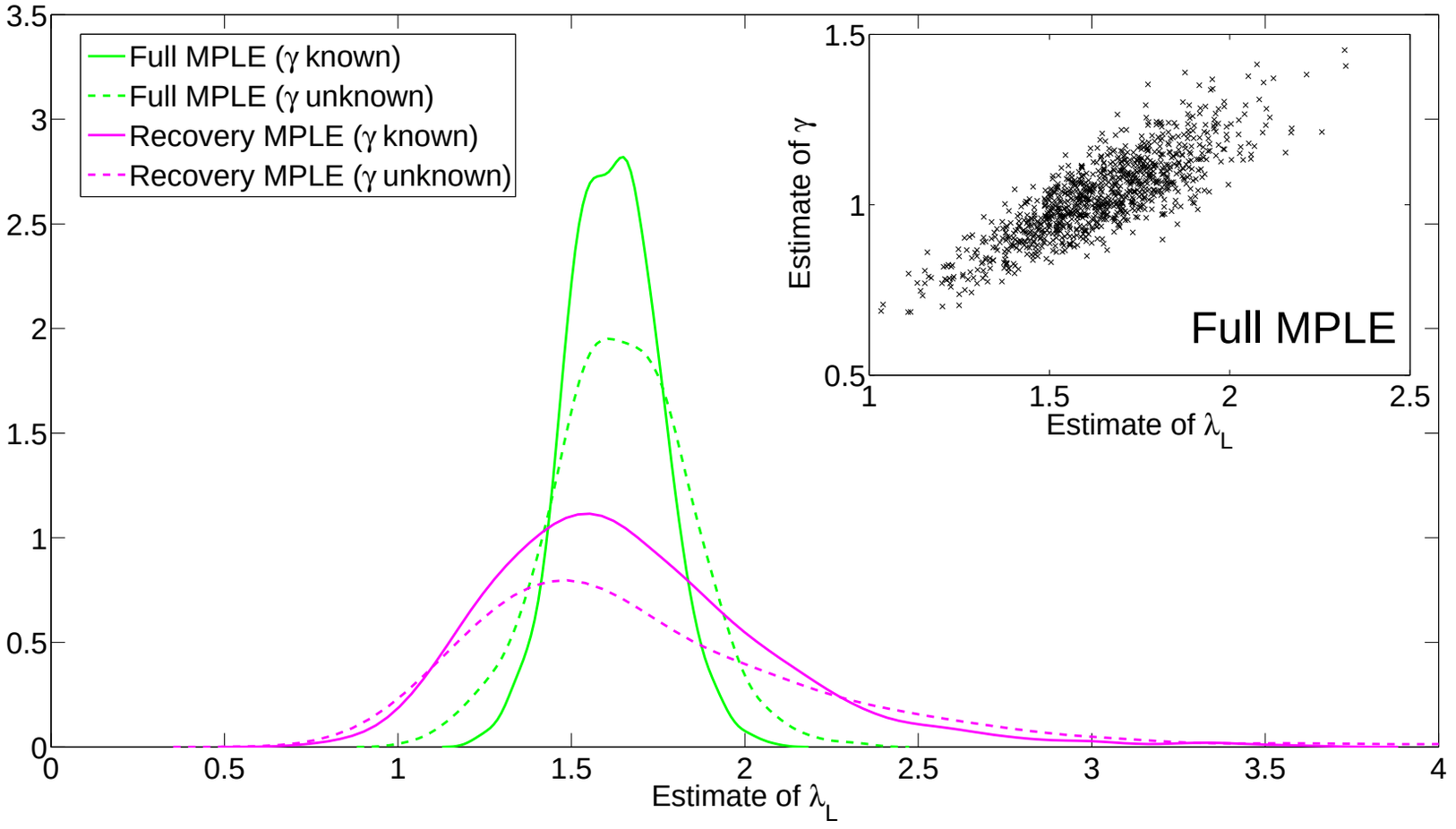
- After 500 individuals have recovered, estimate λ_L using the methods described above, with the growth rate estimate $\hat{r}$ being obtained by fitting a straight line to the logarithm of the number of recoveries, ignoring the first 20 recoveries.

Estimates of λ_L with time



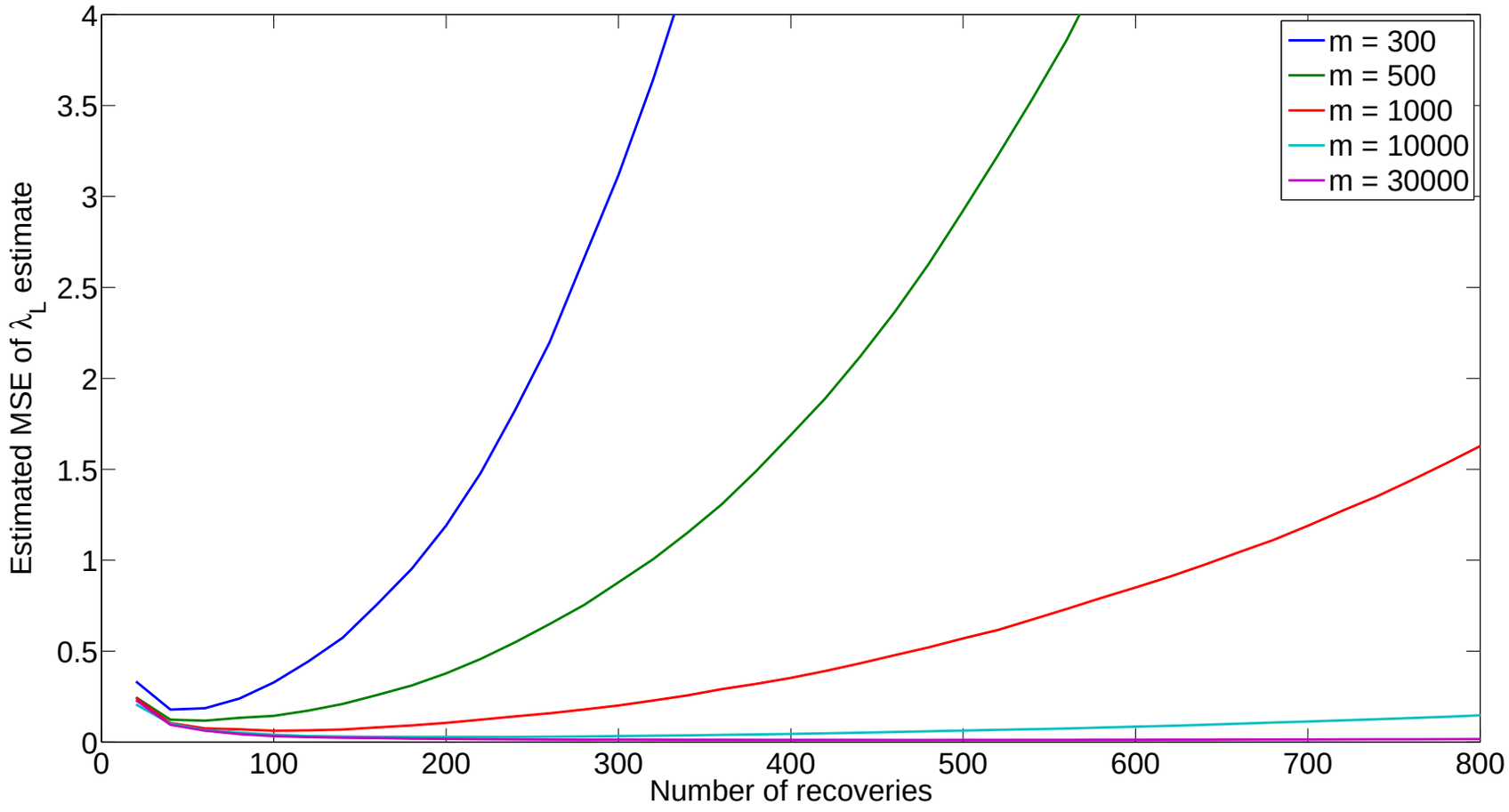
Estimates of within-household infection rate λ_L as a function of time for a single simulated epidemic with $\lambda_L = 1.56$ and $\lambda_G = 1.21$ that took off.

Estimation in an emerging epidemic



Kernel-density estimates of distributions of estimators of within-household infection rate λ_L based on $1,000$ simulated epidemics with $\lambda_L = 1.56$ and $\lambda_G = 1.21$ that took off.

Mean squared error (MSE)



MSE of estimates of λ_L using the full-pseudolikelihood method with known recovery rate $\gamma = 1$ during the emerging phase of $1,000$ simulated epidemics with different population sizes. Recall m is the number of households.

Concluding comments

- When fitting household and other models to data on an emerging epidemic, the data collected need to be modelled very carefully taking due account of the emerging nature of the epidemic.
- Asymptotic stable composition of supercritical branching processes provides a flexible framework for modelling such data.
- Areas for further research include
  - estimation of growth rate r
  - numerical implementation for non-Markovian models
  - variance of estimators
  - multitype epidemics — e.g. age-stratified populations, asymptomatic cases
  - temporal data within households.

Infinite data

- Suppose that the final sizes in m households are observed, each distributed according to $\tilde{p}_n(\cdot|\lambda_L)$ but the maximum-likelihood estimate $\hat{\lambda}_L^{(m)}$ is obtained using $p_n(\cdot|\lambda_L)$.
- Then

$$\hat{\lambda}_L^{(m)} \xrightarrow{\text{a.s.}} \hat{\lambda}_L^* \quad \text{as} \quad m \to \infty,$$

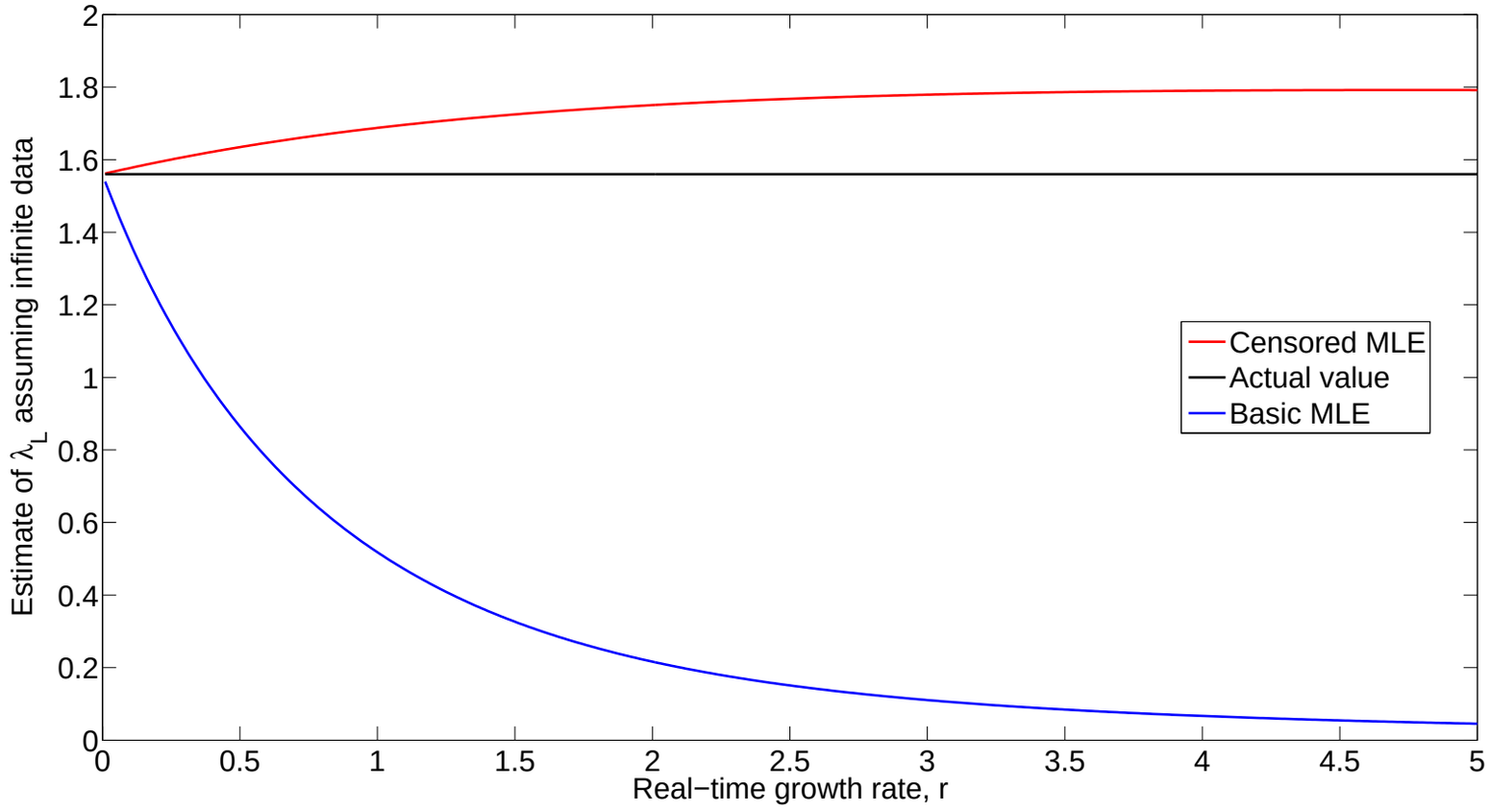
where $\hat{\lambda}_L^*$ maximises

$$\sum_{i=1}^{n} \tilde{p}_n(i|\lambda_L) \log p_n(i|\lambda_L).$$

Equivalently, $\hat{\lambda}_L^*$ minimises the Kullback-Leibler divergence of $p_n(\cdot|\lambda_L)$ from $\tilde{p}_n(\cdot|\lambda_L)$.

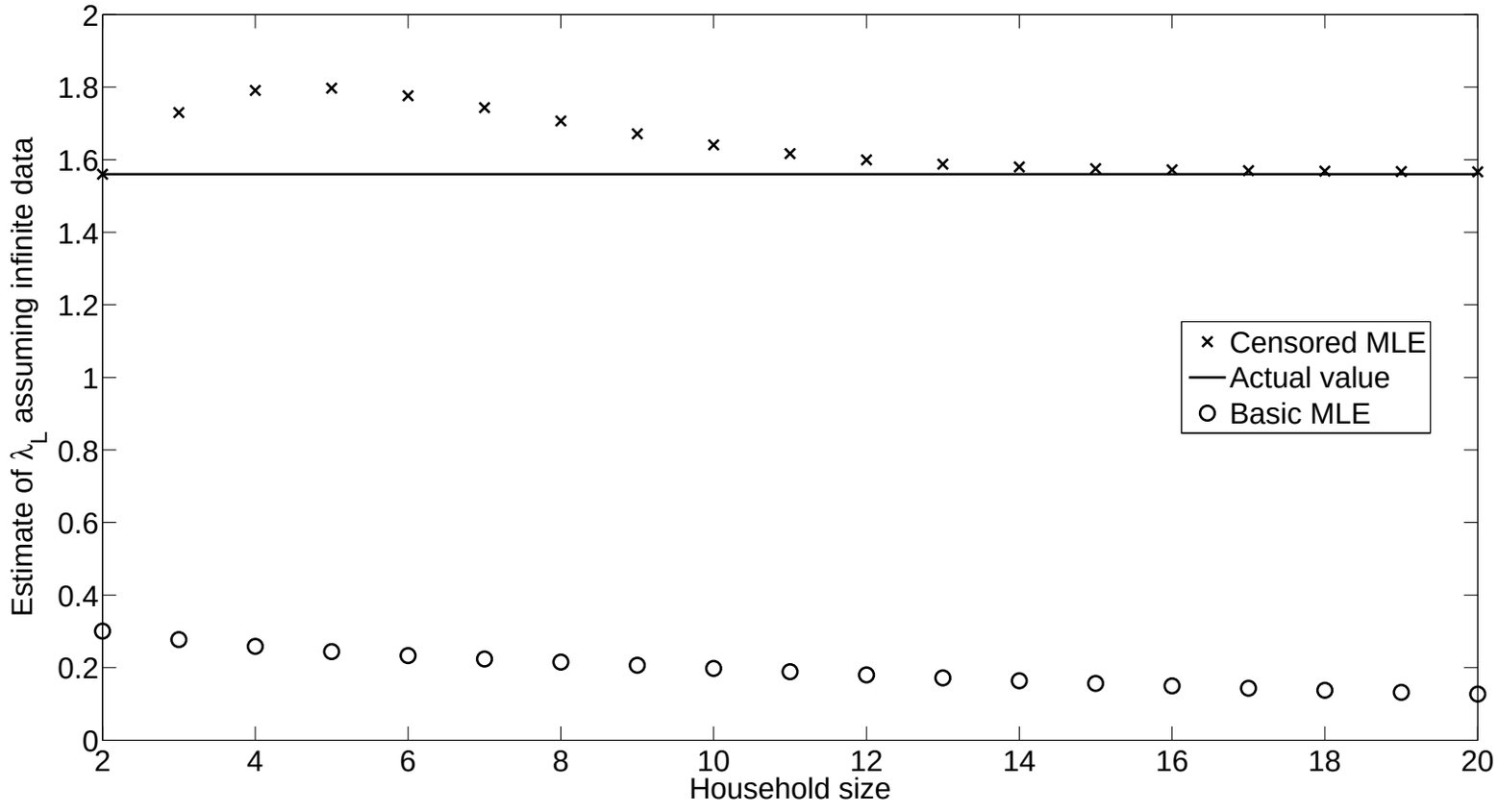
- Obvious extension to unequal household sizes.

Dependence of estimates on r



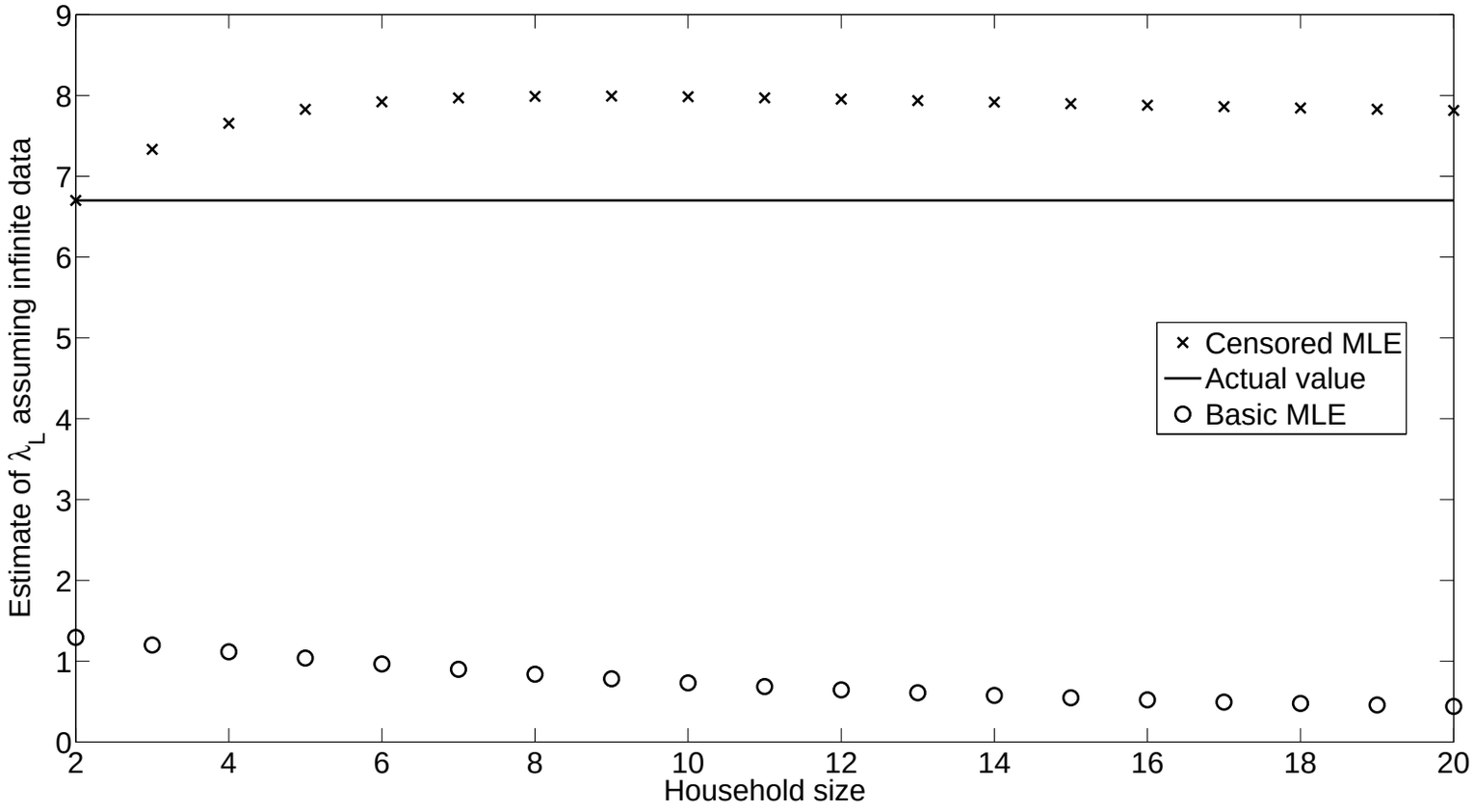
Estimates of λ_L using usual final size distributions $p_n(\cdot|\lambda_L)$ ($n = 1, 2, \cdots, n_{\max}$), assuming infinite data.

Effect of household size



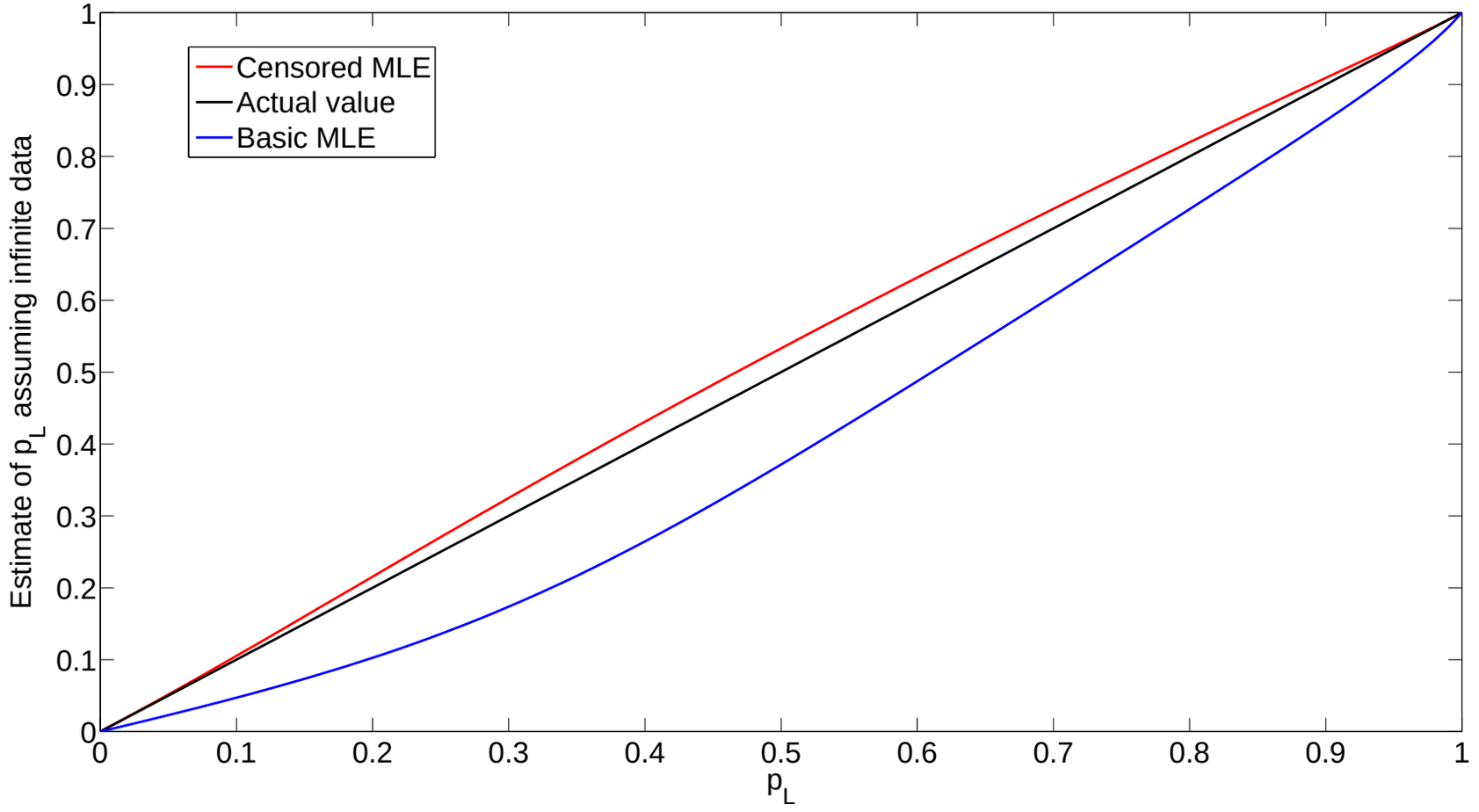
Estimates of λ_L assuming infinite data, for emerging epidemics with $r = 1.76$ and $\lambda_L = 1.56$ for populations with constant household size.

Effect of household size



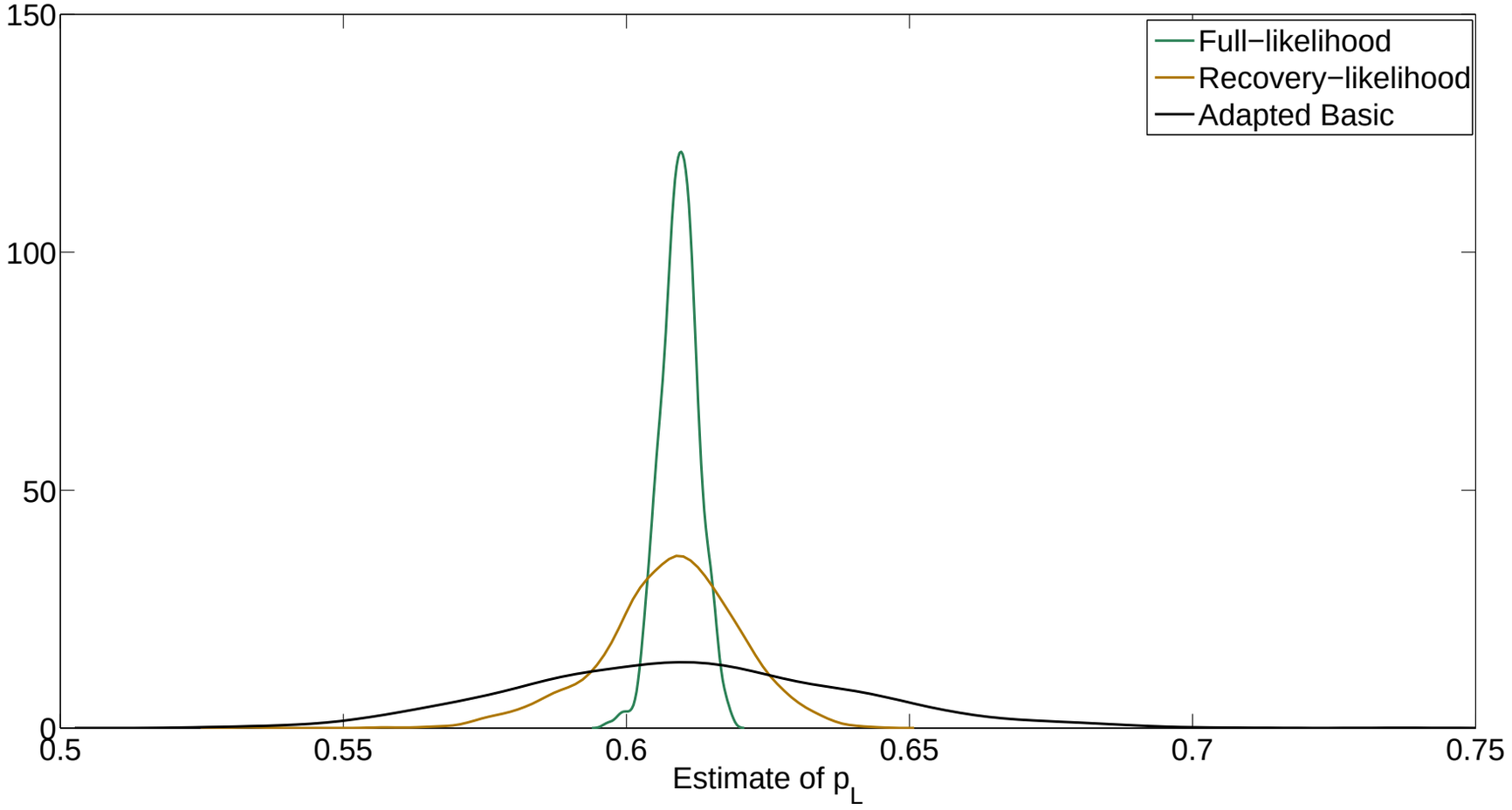
Estimates of λ_L (true value 6.7) assuming infinite data, for emerging epidemics with $r = 1.76$ and $\lambda_L^{(n)} = \lambda_L/n$ for populations with constant household size.

Reed-Frost model



Estimates of p_L using usual final size distribution for Reed-Frost single-household epidemic assuming infinite data with growth rate r fixed at 0.8109.

Reed-Frost model



Kernel-density estimate of distribution of "MLE" $\hat{p}_L$ based on $1,000$ simulated epidemics with $p_L = 0.61$ and $\mu_G = 1.21$ that took off.

Concluding comments

- When fitting household and other models to data on an emerging epidemic, the data collected need to be modelled very carefully taking due account of the emerging nature of the epidemic.
- Asymptotic stable composition of supercritical branching processes provides a flexible framework for modelling such data.
- Areas for further research include
  - estimation of growth rate r
  - numerical implementation for non-Markovian models
  - variance of estimators
  - multitype epidemics — e.g. age-stratified populations, asymptomatic cases
  - temporal data within households.