A Statistical Approach to Find the Multiplicity of Infection

Meraj Hashemi, Kristan Schneider

Faculty of Applied Computer- and Biosciences,
University of Applied Sciences Mittweida

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Abstract

Multiplicity of infection (MOI) refers to the presence of multiple pathogen variant within an infection due to multiple infective contacts. MOI is an important clinical, genetic and epidemiological parameter, hence accurate estimates are highly desirable. Here, we show how a maximum-likelihood estimate of MOI can be improved by applying bias correction.

Background

In epidemiology, metrics capable to monitor exposure and transmission intensity are of particular interest. Hosts in areas of moderate/high transmission typically infected by several genetically distinct pathogen variants (lineages). Classical metrics for monitoring the transmission intensity:

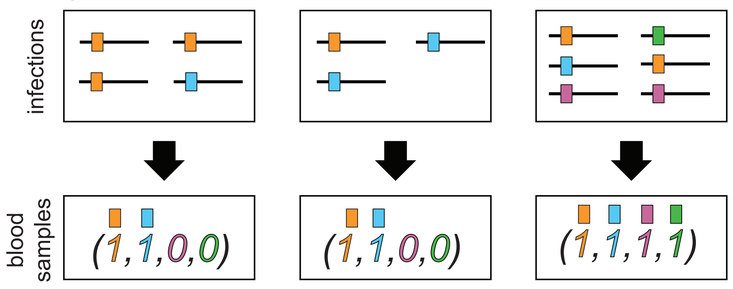
i entomological inoculation rate (EIR),

ii basic reproduction rate (R_0).

EIR and R_0 are difficult to measure. More appropriate: molecular metrics, e.g. multiplicity of infection (MOI), defined as the number of super infections.
properties of MOI:

- informative of transmission intensities,
- mediating the recombination rate,
- correlates with disease severity.

MOI is usually measured by ad hoc metrics that rely on a set of genetic markers. In practice MOI is unknown for a given host, however, it is possible to detect absence/presence of lineages in a blood sample (cf. [2] and figure below).
Figure: colored squares=variants of genetic marker, infections with, respectively, MOI=4, 3, 6 with 2, 2 and 4 different lineages.



Bias-corrected MLE (BC-MLE)

The properties of MLE:

- asymptotically unbiased,
- strongly consistent,
- efficient.

Nevertheless, could perform poorly for small samples ($N < 100$), bias is of order $\mathcal{O}(N^{-1})$.

Solution: bias correction to reduce bias to order $\mathcal{O}(N^{-2})$. We adopt the bias-correction outlined in [1]. The method gives an explicit formula for the bias, but it is omitted because of its complicated expressions.

Bias-corrected MLE's performance is checked by Monte Carlo simulations.

Simulation

For each set of parameters $\boldsymbol{\theta} = (\lambda, \boldsymbol{p})$, 10 000 samples of various sizes of N are generated randomly from the underlying statistical model. The MLE and its bias-corrected version were derived, where the bias was calculated conditionally on non-degenerate data.

Likelihood Estimate

Assume n different lineages $A_1, A_2, \ldots, A_n$, with the frequencies $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ in a pathogen population.

Assumptions:

1. infections are rare and independent,
2. only one lineage per infective event,
3. lineages mutate rarely during the course of infection,
4. infections are not chronical.

$m \ldots$ number of super infections. Assumptions imply:

$$m \sim \text{Poiss}(\lambda), \text{i.e., } P(m) = \frac{1}{e^{\lambda}} \frac{\lambda^m}{m!}, \quad m \geq 0.$$

When considering only disease-positive patients:

$$m \sim \text{conditional Poiss}(\lambda), \text{i.e.,}$$
$$P(m) = \frac{1}{e^{\lambda} - 1} \frac{\lambda^m}{m!}, \quad m \geq 1.$$

Average MOI is:

$$\psi = \frac{\lambda}{1 - e^{\lambda}}.$$

Log-likelihood based on N random samples is:

$$L = L(\lambda, \boldsymbol{p}) = -N \log(e^{\lambda} - 1) + \sum_{k=1}^{n} N_k \log(e^{\lambda p_k} - 1),$$

where

$N_k \ldots$ number of samples containing lineage A_k.

The maximum-likelihood estimate (MLE) for the model parameters is ([2]):

$$\hat{p}_k = \frac{1}{\hat{\lambda}} \log\left(1 - \frac{N_k}{N}(1 - e^{-\hat{\lambda}})\right),$$

where $\hat{\lambda}$ is derived by iterating

$$\lambda_{t+1} = \lambda_t - \frac{\lambda_t + \sum_{k=1}^{n} \log\left(1 - \frac{N_k}{N}(1 - e^{-\lambda_t})\right)}{1 - \sum_{k=1}^{n} \frac{N_k}{N e_t^{\lambda} - N_k (e^{\lambda} - 1)}}.$$

Recursion converges monotonically at quadratic rate from initial value $\lambda_1 \geq \lambda$.

Existence of MLE is ensured if

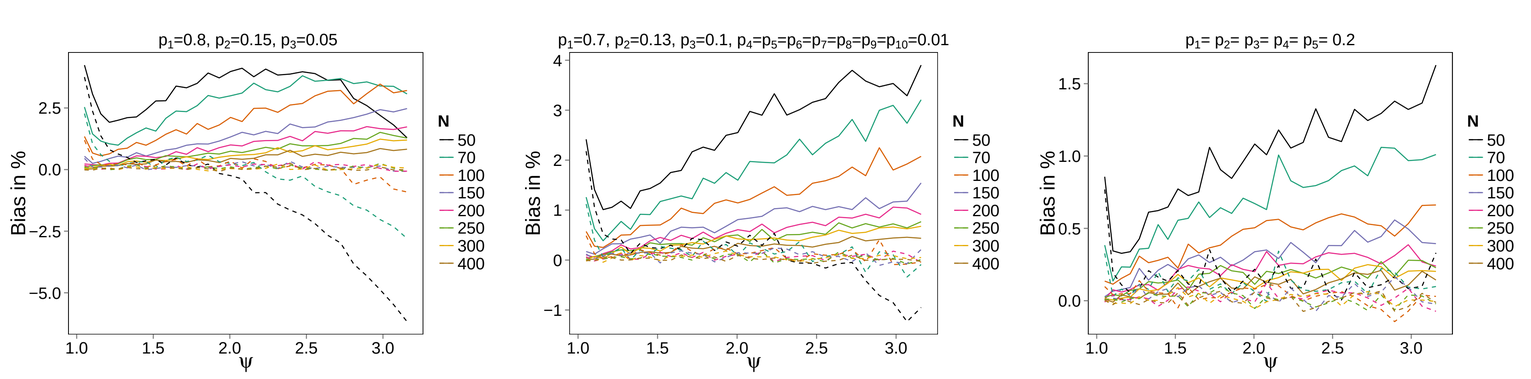
- $N_k \neq N \; \forall k$ i.e., no lineage in all samples.,
- $\sum_{k=1}^{n} N_k > N$ i.e., at least one polycolonal sample.

Data violating these properties is called degenerate.

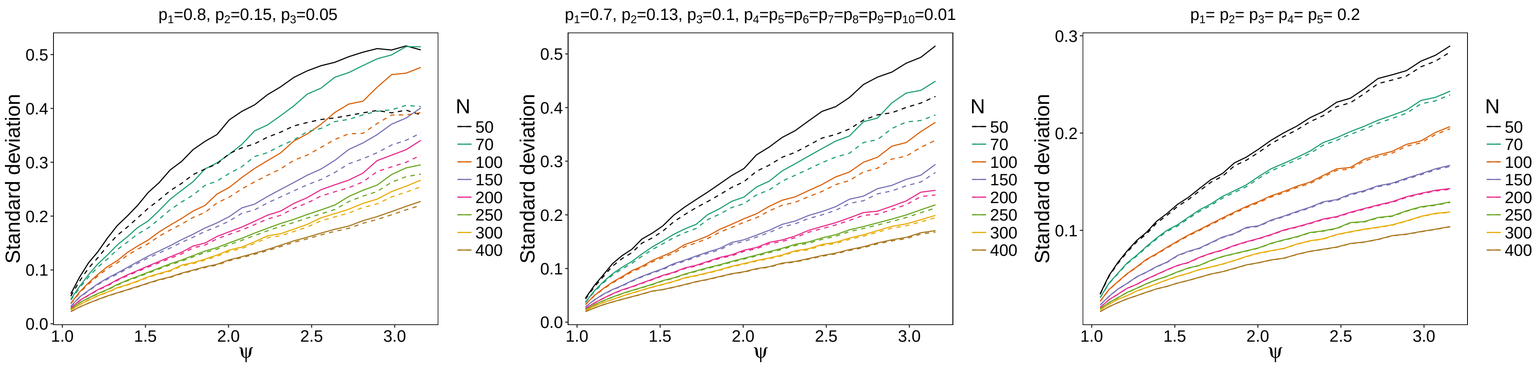
Results

Figures illustrate the performance of MLE (solid) vs. BC-MLE (dashed).

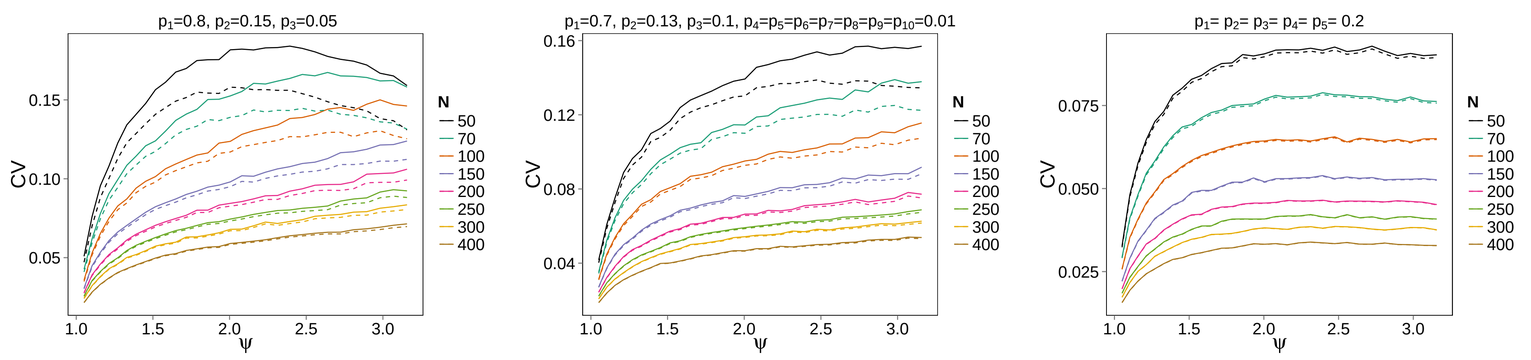
a) The BC-MLE shows a clear improvement over the original estimate. For typical values of ψ i.e., $1.1 < \psi < 2.7$, the BC-MLE is almost unbiased. For skewed lineage frequencies and especially when n is small, the BC-MLE is still unbiased. However, it outperforms the MLE.



b) The BC-MLE has a slightly smaller standard deviation (sd) than the MLE, and the sd is closer to Creme-Rao lower bound (not shown).



c) The dimension-free coefficient of variation is a better way to comparethe estimate's variances across ψ-values.



Summarizing, the BC-MLE can be readily calculated and has notably smaller bias (it is almost unbiased) and smaller variance than the original MLE, and it is therefore preferable.

Acknowledgements

This research is supported by a scholarship from Hans-Seidel-Stiftung(HSS) and the SMWK-SAB project "Innovationsvorhaben yur Profilschäferung an Hochschulen für Angewandte Wissenschaften" (Project number 100257255).

References

[1] G. M. Cordeiro and F. Cribari-Neto, *An introduction to bartlett correction and bias reduction*, 2014.

[2] K. A. Schneider and A. A. Escalante, *A likelihood approach to estimate the number of co-infections*, PLoS ONE, 9 (2014), p. e97899.