# Bias-corrected Maximum-likelihood Estimates of Multiplicity of Infection

Meraj Hashemi and Kristan A. Schneider

Department of Applied Computer and Biosciences,
University of Applied Sciences Mittweida,
Technikumplatz 17, 09648
Mittweida, Germany.

June 18, 2018

**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.

## 1 Introduction

In epidemiology, metrics capable to monitor changes in exposure and transmission intensity are of particular interest. While the entomological inoculation rate (EIR) and the basic reproduction number R0 are still the gold standards to measure transmission [3][10], molecular metrics such as multiplicity of infection (MOI) and molecular force of infection (mFOI) are recognized as being more appropriate [10]. In some diseases, e.g., malaria this is well recognized [6][7]. Multiplicity of infection refers to the number of super-infections with the same or different pathogen variants in the course of an infection (Figure 1). MOI is of clinical importance as it might correlate with disease severity (e.g., in the case of malaria) and the interaction of different pathogen variants within an infection might affect the course of the disease. Moreover, the distribution of MOI in a population correlates with transmission intensities, underlying its epidemiological importance. Moreover, MOI is an important genetic quantity as it mediates recombination between pathogen variants. Although it is possible to control or measure the number of distinctive pathogen lineages in models and experimental settings (e.g. [3]), a totally different scenario is the one faced by those studying naturally occurring infections in the context of ecological and epidemiological investigations [4]. Under such circumstances, MOI is usually measured by ad hoc metrics that rely on a set of genetic markers or the observed polymorphism in one or several genes [2]. The need for an experimental definition of MOI has generated approaches based on phylogenetic frameworks (e.g. many viruses) or some form of multi-locus genotyping [1] Whereas such approximations have been useful, there is still need for a formal statistical framework that allows the estimation of the actual number of

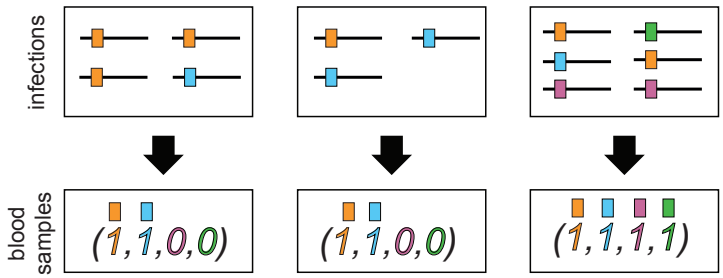


***Figure 1:*** *Information contained in blood samples. The top shows three (co-)infections, while the bottom shows the respective information about the infection that can be reconstructed from a blood sample. The first individual was infected by $m = 4$ lineages, three times with the orange and once with the blue lineage. Hence the orange and blue lineages are detectable in the blood sample, while the pink and green lineages were not detected. In the middle a co-infection with $m = 3$ lineage is illustrated which differs from the first infection but results in the same 0-1-vector. All four lineages were infecting in the third example, however $m = 6$ co-infections occurred.*

lineages and other approximations to MOI that felicitates and/or considers confounding factors. In the context of malaria (and related diseases) such a framework was introduced and developed by [6][8][9]. More precisely, a maximum-likelihood framework was employed to estimate MOI and the frequency of pathogen lineages from molecular data obtained from a collection of blood samples of disease-positive patients. Although, the method is promising, it leads to biased results, especially for small sample sizes. Here, we report how a bias correction can improve the estimate in terms of bias and variance.

# 2 Methods

In the following we briefly describe the model and show how the maximum likelihood is derived.

## 2.1 Model Background

We refer to lineages as pathogen variants that can be identified by allelic variants at a considered locus, therefore we can use the terms *"lineage"* and *"allele"* synonymously. Lineages can also be interpreted as haplotypes in a non-recombining region. Suppose $n$ lineages of a pathogen, $A_1, \ldots, A_n$, circulating in a population. Their frequencies are denoted by the vector $\boldsymbol{p} = (p_1, \ldots, p_n)$. It is assumed that a host is infected by one lineage at each infective event. Importantly, hosts can be super-infected multiple times by the same lineage or different ones. Let $m_k$ be the number of times a host is infected by lineage $A_k$, so that $m = m_1 + \ldots + m_n$ is the number of times a host was (super-)infected during the course of an infection. Consequently, conditioned on being super-infected $m$-times,the probability that the host is infected $m_k$ times with lineage $A_k$ $(k = 1, \ldots, n)$ follows a multinomial distribution with parameters $(m, p_1, \ldots, p_n)$. The quantity $m$ is called multiplicity of infection (MOI).

If infections with the pathogen are rare and independent, a natural assumption is that the number of pathogens infecting a host is Poisson distributed, or more precisely follows a conditional

Poisson distribution (CPD) i.e.

$$\kappa_m = \frac{1}{e^\lambda - 1}\frac{\lambda^m}{m!} \qquad \text{for } m \geq 1. \tag{1}$$

Under this assumption, the distribution of MOI is identified by the parameter $\lambda$. The average MOI is $\dfrac{\lambda}{1-e^{-\lambda}}$.

The objective is to estimate the distribution of $m$ or equivalently $\lambda$ from $N$ disease-positive blood samples by a maximum-likelihood approach.

In practice $m$ is unknown for a given host (see Figure 1) and it is impossible to reconstruct $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)$, however, it is possible to detect the absence and presence of lineages in a blood sample. Let $i_k \in \{0, 1\}$ denote the absence and presence of lineage $A_k$. Therefore, a clinical sample is represented by the configuration $\boldsymbol{i} = (i_1, \ldots, i_n) \in \{0, 1\}^n \backslash \{\mathbf{0}\}$. Notably, the configuration $\mathbf{0}$ represents an uninfected host($m = 0$). The probability that a clinical sample has configuration $\boldsymbol{i}$ is given by

$$Q_{\boldsymbol{i}} = \frac{1}{e^\lambda - 1}\prod_{k=1}^{n}(e^{\lambda p_k} - 1)^{i_k} \tag{2}$$

according to [8]. This model is identifiable, i.e., different sets of parameters lead to different distributions of $\boldsymbol{i}$.

Data is obtained by collecting clinical samples (e.g. blood samples) from $N$ disease-positive hosts. Let $N_k$ be the number of samples in which lineage $A_k$ is found, i.e., $N_k = \sum_{j=1}^{N} i_k^{(j)}$ where $\boldsymbol{i}^{(j)}$ corresponds to configuration of $j$-th sample. Under the model (2), the log-likelihood function is given by

$$L = L(\lambda, \boldsymbol{p}) = -N\log(e^\lambda - 1) + \sum_{k=1}^{n} N_k \log(e^{\lambda p_k} - 1), \tag{3}$$

cf. [8].

## 2.2 Maximum Likelihood

The maximum-likelihood estimate (MLE) for the model parameters $\theta = (\lambda, \boldsymbol{p})$ exists and is uniquely defined except in two pathologic situations. In the first, only one lineage is found in each blood sample, i.e., $\sum_{k=1}^{n} N_k = N$ so there is no indication of super-infections. In the second, at least one lineages is found in every blood sample, i.e., $N_k = N$ for at least one $k$. Except these cases the maximum likelihood can be properly evaluated. For a regular case the maximum likelihood estimate of $\boldsymbol{\theta}$ is

$$p_k = -\frac{1}{\hat{\lambda}}\log\big(1 - \frac{N_k}{N}(1 - e^{-\hat{\lambda}})\big), \tag{4a}$$

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

$$\lambda_{t+1} = \lambda_t - \frac{\lambda_t + \sum_{k=1}^{n} \log\left(1 - \frac{N_k}{N}(1 - e^{-\lambda_t})\right)}{1 - \sum_{k=1}^{n} \frac{N_k}{Ne_t^{\lambda} - N_k(e^{\lambda}-1)}}. \tag{4b}$$

The sequence (4b) converges monotonically and at quadratic rate from any initial value $\lambda_1 \geq \lambda$. Hence, it is guaranteed to find the MLE as long as the initial value $\lambda_1$ is chosen sufficiently large.

## 2.3 Bias corrections

Maximum-likelihood estimates have many desirable asymptotic properties. In particular, they are typically asymptotically unbiased, where the bias is of order $\mathcal{O}(\frac{1}{N})$. Indeed, in our case, bias might be too large for samples of size $N < 100$, especially if $\lambda$ is small, i.e., if transmission intensities are small. Since transmission intensities correlate not only with MOI but also with disease prevalence, it will be difficult to collect a large number of clinical samples in low-transmission settings, rendering $N \sim 90$ a realistic sample size. Therefore, it is important to apply bias corrections to (4), which reduce bias to the order $\mathcal{O}(\frac{1}{N^2})$.

We adopt the bias-correction outlined in [5], which requires the likelihood function to be well-behaved, as in the present case. The advantage of this method is that it can be explicitly derived, although the formulas are complicated and hence omitted. We conducted a simulation study, in which for each parameter choice $\boldsymbol{\theta} = (\lambda, p_1, \ldots, p_n)$ 10 000 samples of size $N$ (=50, 70, 100, 150, 200, 250) were randomly drawn from the conditional-Poisson model and the MLE and its bias-corrected version were derived. Bias and variance of the estimates were derived based on these samples. Since the MLE cannot be derived in pathologic situations, especially in the case $\sum_{k=1}^{n} N_k = N$ , in which formally $\hat{\lambda} = 0$ and (4) is undefined, pathologic samples were replaced until a total number of 10 000 was reached. Hence, bias was calculated conditionally on non-pathologic data. The case $\sum_{k=1}^{n} N_k = N$ occurs mainly for very small $\lambda$.

## 2.4 Results and conclusions

The bias-corrected estimate (BCE) shows clear improvement over the original estimate, as shown in Figure 2. The parameters chosen in Figure 2 are representative choices as far as results change only quantitatively but not qualitatively. (The parameters used lead to high coefficients of variation, hence worst case scenarios are shown). The original MLE tends to overestimate the true parameter, even substantially for small values of $\lambda$, but bias vanishes as sample size increases. The BCE can be regarded unbiased irrespectively of the sample size, except for small $\lambda$. The reason for the overestimation of small true $\lambda$, is that occasional data sets with high MOI yield a huge overestimation. (Remember also that we report bias conditioned on non-pathologic data. Since $\hat{\lambda}$ if $\sum_{k=1}^{n} N_k = N$, actual bias is smaller than the one reported.)

The BCE shows small improvement in standard deviation compared with the MLE (cf. Figure 2c, g), especially for small sample size. Together with the reduction in bias, this leads to a small decrease in the coefficient of variation of the BCE (cf. Figure 2d, h). Summarizing, the BCE has reduced bias and smaller variance as the original estimate and is therefore preferable, especially for small samples.

# References


[1] *Reconstructing the dynamics of hiv evolution within hosts from serial deep sequence data.*, PLoS Comput Biol, 8 (2012), p. e1002753.

[2] O. Balmer and M. Tanner, *Prevalence and implications of multiple-strain infections*, The Lancet Infectious Diseases, 11 (2011), pp. 868 – 878.

[3] F. Ben-Ami, L. Mouton, and D. Ebert, *The effects of multiple infections on the expression and evolution of virulence in a daphnia-endoparasite system.*, Evolution, 62 (2008), pp. 1700–1711.

[4] T. Cohen, P. D. van Helden, D. Wilson, C. Colijn, M. M. McLaughlin, I. Abubakar, and R. M. Warren, *Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control*, Clinical Microbiology Reviews, 25 (2012), pp. 708–719.

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

[6] W. G. Hill and H. A. Babiker, *Estimation of numbers of malaria clones in blood samples*, Proceedings of the Royal Society of London. Series B: Biological Sciences, 262 (1995), pp. 249–257.

[7] E. Y. Klein, D. L. Smith, R. Laxminarayan, and S. Levin, *Superinfection and the evolution of resistance to antimalarial drugs.*, Proc Biol Sci, 279 (2012), pp. 3834–3842.

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

[9] K. A. Schneider and Y. Kim, *An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance.*, Theor Popul Biol, 78 (2010), pp. 93–108.

[10] L. S. Tusting, T. Bousema, D. L. Smith, and C. Drakeley, *Chapter three - measuring changes in plasmodium falciparum transmission: Precision, accuracy and costs of metrics*, vol. 84 of Advances in Parasitology, Academic Press, 2014, pp. 151 – 208.

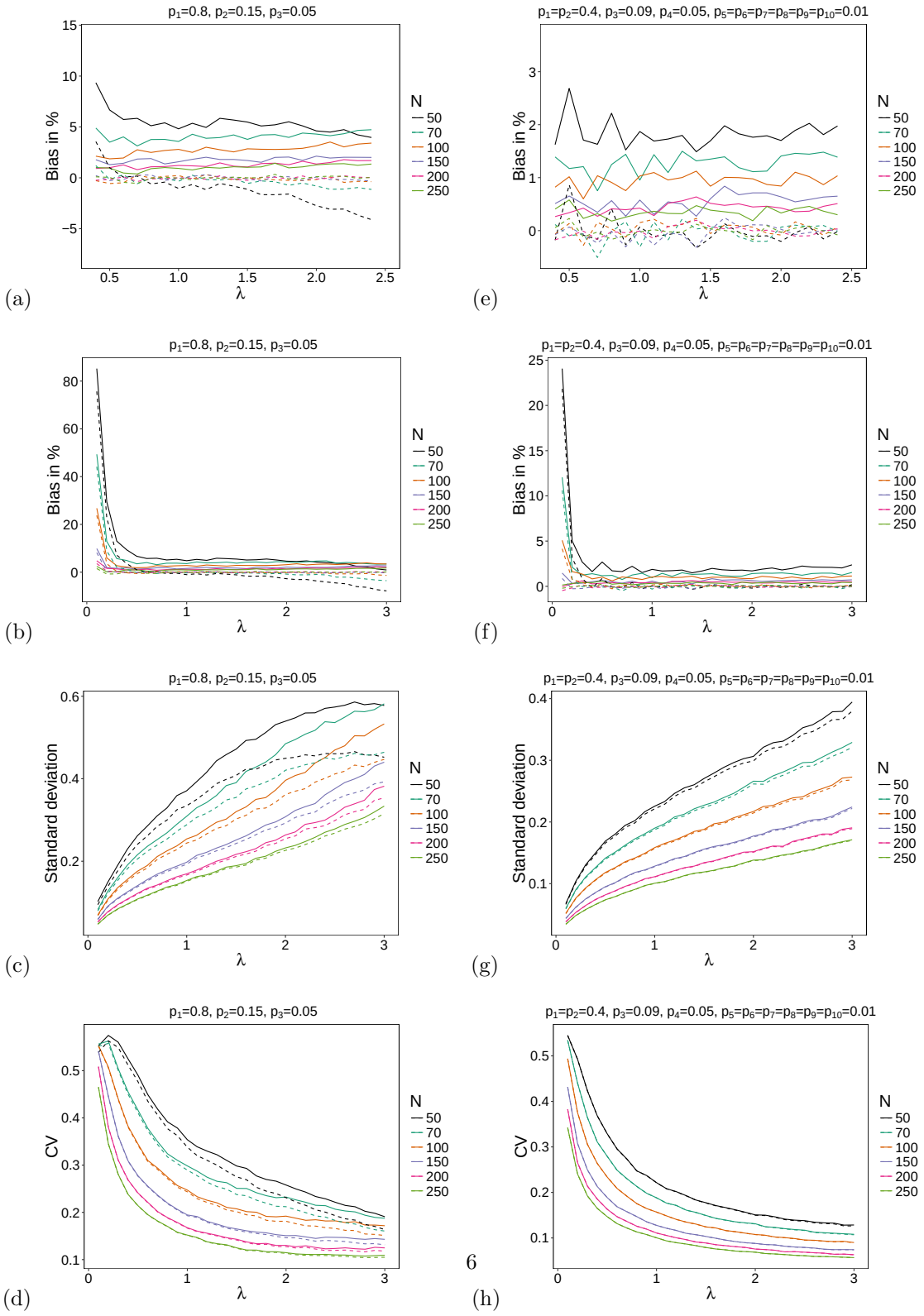


***Figure 2:*** *Through different inferences The plots illustrate the improvements that has been obtained by bias correcting the MLE. Different colors are representative of different sample sizes, and the bias-corrected MLE(BCE) is shown by dotted lines.*