CORRECTION

# Correction: A Likelihood Approach to Estimate the Number of Co-Infections

**Kristan A. Schneider, Ananias A. Escalante**

The authors wish to acknowledge an error that was overseen in the proof of Result 1 in [1]. By carefully inspecting the authors' proof, it becomes clear that an MLE cannot exist if $N_k = N$ for at least one $k$. This is easily seen from the authors' formula for $\hat{p}_k$ in Result 1, where $N_k = N$ implies $\hat{p}_k = 1$, a contradiction since then $L(\lambda, \hat{p}|x) = -\infty$. All results however remain valid as stated if their Assumption 1 is replaced by the following version.

**Assumption 1** *Assume that the sum over the lineages' prevalences is larger than one, but no alleles is* 100% *prevalent. In other words, more than one lineage is found in at least one infection, i.e.,* $\sum_{k=1}^{n} N_k > N$ *and all lineages are not found in every infection, i.e.,* $N_k \neq N$ *for all k.*

By replacing Assumption 1 with the version above in [1], the results hold without modifications. All other modifications that need to be made in the article are minor and obvious. However, the case $N_k = N$ for at least one $k$ was not properly addressed. This occurred because it was overseen that the proof of in Result 1 is not applicable then. What goes wrong in this case? The answer is somewhat subtle. Heuristically, this contradiction occurs because no point in the parameter space is a critical point, i.e., a point at which all derivatives of $L$ vanish. However, for any fixed $\lambda$, $L(\lambda, p|x)$ attains a maximum for some $\hat{p}^{(\lambda)}$, with $0 < \hat{p}_k^{(\lambda)} < 1$. The reason is that $L(\lambda, p|x) = -\infty$ for $p \in bd\mathcal{S}_n$ (where $\mathcal{S}_n$ denotes the $n-1$-dimensional simplex). For $\lambda \to 0$, $L(\lambda, \hat{p}^{(\lambda)}|x) \to -\infty$. Hence, $L(\lambda, \hat{p}^{(\lambda)}|x)$ is necessarily monotonically increasing in $\lambda$, implying that no MLE exists. In mathematical terms this can be formulated as follows:

**Remark 1** *Assume that at least one lineage is found in every sample, i.e.,* $N_k = N$ *for at least one k, but not all are found in every sample, i.e.,* $N_k \neq N$ *for at least one j. Then, the log-likelihood function does not attain a maximum. However, its smallest upper bound is*

$$\sup_{p \in \mathcal{S}_n, \lambda > 0} L(\lambda, p|x) = \sum_{\substack{k=1 \\ N_k \neq N}}^{n} (N - N_k)\log(1 - \frac{N_k}{N}).$$

*The supremum is reached in the limit of any sequence* $(\lambda_t, p_t)$ *with* $\lim_{t \to \infty} \lambda_t = \infty$, $\lim_{t \to \infty} p_k^{(t)} = -\log(1 - \frac{N_k}{N})$ *if* $N_k \neq N$ *and* $\lim_{t \to \infty} p_k^{(t)} \lambda_t = \infty$ *if* $N_k = N$.

**Proof**. Because $L(\lambda, p|x)$ is bounded by 0, the supremum exists. Furthermore, a sequence $(\lambda_t, p_t)$ exists with $L(\lambda_t, p_t|x) \to \sup_{p \in \mathcal{S}_n, \lambda > 0} L(\lambda, p|x)$.

Check for updates

Without loss of generality let $N_1, \ldots, N_m < N$ and $N_{m+1} = \ldots = N_n = N$. Hence,

$$
\begin{aligned}
L(\lambda, \boldsymbol{p} | \boldsymbol{x}) &= -N \log(e^\lambda - 1) + \sum_{k=1}^{n} N_k \log(e^{\lambda p_k} - 1) \\
&= N \log \frac{(e^{\lambda p_{m+1}} - 1) \cdot \ldots \cdot (e^{\lambda p_n} - 1)}{e^\lambda - 1} + \sum_{k=1}^{m} N_k \log(e^{\lambda p_k} - 1) \\
&= N \log \frac{(1 - e^{-\lambda p_{m+1}}) \cdot \ldots \cdot (1 - e^{-\lambda p_n})}{1 - e^{-\lambda}} e^{-\lambda(1 - p_{m+1} - \cdots - p_n)} + \sum_{k=1}^{m} N_k \log(e^{\lambda p_k} - 1) \\
&= N \log \frac{(1 - e^{-\lambda p_{m+1}}) \cdot \ldots \cdot (1 - e^{-\lambda p_n})}{1 - e^{-\lambda}} e^{-\lambda(p_1 + \cdots + p_m)} + \sum_{k=1}^{m} N_k \log(e^{\lambda p_k} - 1).
\end{aligned}
\tag{1}
$$

Let $(\lambda_t)$ be any monotone sequence with $\lim_{t \to \infty} \lambda_t = \infty$. Moreover, let $c_k > 0$ for $k = 1, \ldots, m$. Now let $\boldsymbol{p}_t$ be a sequence satisfying $\lim_{t \to \infty} p_k^{(t)} \lambda_t = c_k$ for $k = 1, \ldots, m$ and $\lim_{t \to \infty} p_k^{(t)} \lambda_t = \infty$ for $k = m+1, \ldots, n$. Without loss of generality let $p_k^{(t)} = \frac{c_k}{\lambda_t}$ for $k = 1, \ldots, m$ and $p_k^{(t)} = \frac{1}{n-m}(1 - \sum_{k=1}^{m} \frac{c_k}{\lambda_t})$ for $k = m+1, \ldots, n$. For sufficiently large $t$ this sequence is defined and $\boldsymbol{p}_t \in \mathcal{S}_n$. Hence,

$$
\begin{aligned}
\lim_{t \to \infty} L(\lambda_t, \boldsymbol{p}_t | \boldsymbol{x}) &= N \log 1 \cdot e^{-c_1 - \cdots - c_m} + \sum_{k=1}^{m} N_k \log(e^{c_k} - 1) \\
&= -N(c_1 + \ldots + c_m) + \sum_{k=1}^{m} N_k \log(e^{c_k} - 1).
\end{aligned}
$$

Next define $f(c_1, \ldots, c_m) := \lim_{t \to \infty} L(\lambda_t, \boldsymbol{p}_t | \boldsymbol{x})$. Note that this definition is independent of the sequence $(\lambda_t, \boldsymbol{p}_t)$, with $\lambda_t \boldsymbol{p}_t \to (c_1, \ldots, c_m, \infty, \ldots, \infty)$ for $t \to \infty$.

The next aim, is to identify potential maxima of $f$. Clearly, $\frac{\partial f}{\partial c_k} = -N + N_k \frac{e^{c_k}}{e^{c_k} - 1}$. Equating the partial derivatives to zero gives $\hat{c}_k = -\log(1 - \frac{N_k}{N})$. The Hessian matrix is given by $H = -\operatorname{diag}(N_k \frac{e^{c_k}}{(e^{c_k} - 1)^2})_{k=1,\ldots,n}$ and clearly negative definite. Thus, $f$ attains a global maximum at $\hat{c}_k = -\log(1 - \frac{N_k}{N})$. Therefore $f(\hat{c}_1, \ldots, \hat{c}_m) \leq \sup_{p \in \mathcal{S}_n, \lambda > 0} L(\lambda, \boldsymbol{p} | \boldsymbol{x})$.

If $(\lambda_t, \boldsymbol{p}_t)$ is any sequence with $\lambda_t p_k^{(t)} \to \infty$ for a $k$ with $1 \leq k \leq m$, it is easily seen from (1) that $\lim_{t \to \infty} L(\lambda_t, \boldsymbol{p}_t | \boldsymbol{x}) = -\infty$. Moreover, if $\lambda_t p_k^{(t)} \to c_k < \infty$ for $1 \leq k \leq m$ and at least one $k$ with $m+1 \leq k \leq n$, without loss of generality $\lambda_t p_k^{(t)} \to c_k < \infty$ for $m+1 \leq k \leq \quad$, (1) implies

$$
\begin{aligned}
\lim_{t \to \infty} L(\lambda_t, \boldsymbol{p}_t | \boldsymbol{x}) &= -N(c_1 + \cdots + c_m) + \sum_{k=1}^{m} N_k \log(e^{c_k} - 1) + N \sum_{k=m+1}^{l} \log(1 - e^{-c_k}) \\
&< -N(c_1 + \cdots + c_m) + \sum_{k=1}^{m} N_k \log(e^{c_k} - 1)
\end{aligned}
$$

implying that this limit is less than the maximum of $f$. The above considerations imply that the supremum of the log-likelihood function must be the maximum of $f$. Deriving $f(\hat{c}_1, \ldots, \hat{c}_m)$ finishes the proof.

The case that $N_k = N$ for all $k$ is treated in [1]. Moreover, obviously in Remark 1 of [1] a misprint occurred. The expression $\sum_{k=1}^{n} N_k \geq N$ needs to be replaced by $\sum_{k=1}^{n} N_k > N$, while the same expression needs to be replaced by $\sum_{k=1}^{n} N_k = N$ in the paragraph below Result 2.

## Reference

1. Schneider KA, Escalante AA (2014) A Likelihood Approach to Estimate the Number of Co-Infections. PLoS ONE 9(7): e97899. https://doi.org/10.1371/journal.pone.0097899 PMID: 24988302