AN EMPIRICAL BAYES APPROACH FOR CONSTRUCTING CONFIDENCE INTERVALS FOR CLONALITY AND ENTROPY

By Zhongren Chen\(^{1,a}\), Lu Tian\(^{2,b}\) and Richard A. Olshen\(^{3,c}\)

\(^{1}\)Department of Statistics, Stanford University, \(^{2}\)zhongren.chen@stanford.edu
\(^{2}\)Department of Biomedical Data Science, Stanford University, \(^{b}\)lutian@stanford.edu
\(^{3}\)Department of Biomedical Data Science, Stanford University, \(^{c}\)olshen@stanford.edu

This paper is motivated by the need to quantify human immune responses to environmental challenges. Specifically, the genome of the selected cell population from a blood sample is amplified by the well-known PCR process of successive heating and cooling, producing a large number of reads. They number roughly 30,000 to 300,000 in our applications. Each read corresponds to a particular rearrangement of so-called V(D)J sequences. In the end, the observed data consist of a set of integers, representing numbers of reads corresponding to different V(D)J sequences. The underlying relative frequencies of distinct V(D)J sequences can be summarized by a probability vector, with the cardinality being the number of distinct V(D)J rearrangements present in the blood. The statistical question is to make inferences on a summary parameter of the probability vector based on a single multinomial-type observation of large dimension. Popular summaries of the diversity of a cell population include clonality and entropy, or more generally, a suitable function of the probability vector. A point estimator of the clonality based on multiple replicates from the same blood sample has been proposed previously. After obtaining a point estimator of a particular function, the remaining challenge is to construct a confidence interval of the parameter to appropriately reflect its uncertainty. In this paper, we have proposed to couple the empirical Bayes method with a resampling-based calibration procedure to construct a robust confidence interval for different population diversity parameters. The method is illustrated via extensive numerical studies and real data examples.

1. Introduction. This paper is motivated by the need to construct confidence intervals for parameters summarizing the diversity of a cell population consisting of cells of different types. We first introduce the sources of data and give some details here. More biological and statistical background in the case of point estimation can be found in a previous paper (Tian et al., 2019). The problem has its biomedical origin in attempting to quantify human immune responses to an environmental challenge; more specifically, to quantify the adaptive immunologic response to any antigen, e.g., vaccination against Covid 19 virus. Briefly, blood is sampled from a patient. This blood sample may be divided, as equally as possible, into several parts, i.e., replicates. The genome of the selected cell subpopulation in each replicate is amplified by the well-known PCR process of successive heating and cooling. One resultant from each replicate is then chosen without prejudice for sequencing, producing a large number of reads. They number roughly 30,000 to 300,000 per replicate. Each read corresponds to a particular rearrangement of the so-called V(D)J sequence. In the end, the observation from a particular replicate consists of a set of numbers of reads for different V(D)J sequences. Mathematically, each observation can then be thought of as a finite dimensional random vector \(Z = (Z_1, Z_2, \ldots, Z_C)\)' reflecting the underlying relative frequency of different cell subpopulations with particular V(D)J rearrangements in the entire circulation system from
which the blood was sampled originally. According to the Rao-Blackwell theorem, the underlying relative frequencies can be summarized by a probability vector \( p = (p_1, p_2, \ldots, p_{C_0})' \), where \( C_0 \), the cardinality of the vector \( p \), represents the total number of different V(D)J rearrangement in the circulation. While \( C_0 \) is finite, its value is unknown, since some of the V(D)J rearrangements may not be observed in the replicate due to their rarity. In other words, \( C \), the number of observed V(D)J rearrangements could be substantially smaller than the number of different V(D)J rearrangements in the blood. While we know from criticisms that the PCR process may favor some rearrangements over others, that distinction is ignored here.

Statistical concerns are making inferences on a summary parameter based on this single multinomial-type observation, \( Z \). One popular summary of the diversity of a cell population is its clonality, \( G_C(p) = \|p\|_2^2 \), the squared \( l_2 \) norm of vector \( p \in \mathbb{R}^{C_0} \), which obviously varies between \( C_0^{-1} \) and 1. We have reported a study of point estimation for clonality based on multiple replicates for each blood sample in a previous paper (Tian et al., 2019). After obtaining a point estimator of a particular function of \( p \) such as clonality characterizing the composition of the cell population, the remaining challenge is to construct a confidence interval for the relevant parameter to appropriately reflect the uncertainty in this point estimator. Consequently, in what follows, attention is devoted to interval estimation. Specifically, we propose to study point and interval estimations for clonality without requiring multiple replicates. Another extremely popular measure for population diversity is entropy, which is defined as

\[
G_E(p) = \sum_{j=1}^{C_0} (-p_j \log p_j)
\]

(see, for example, Glanville et al. (2017)); the logarithm is taken to be to the natural base as the convention. Many authors prefer to summarize the variability of \( p \) by its entropy. Of course, the notion of entropy has a storied history in communications, indeed in many aspects of engineering (Zurek, 2018). The estimation of entropy itself is very difficult (Chao and Shen, 2003). Especially, we are unaware of careful attempts to form confidence intervals for entropy. The method proposed in this paper will cover the point and interval estimations for parameters such as entropy or other functions of the probability vector \( p \).

We begin with an explanation of matters that bear upon parametric approaches in Section 2. All material applies to a fixed \( p \). However, the key is that we assume that these \( C_0 \) components of the probability vector \( p \) are realizations from a parametric “prior” distribution, which can be estimated based on observed data, although there is oftentimes no sufficient information in estimating the values of all individual components. With the estimated “prior” distribution generating the individual probability components, point and interval estimates for the function of interest can be obtained based on the “posterior distribution” derived from the Bayesian theorem. This is essentially an Empirical Bayesian approach, where the “prior” distribution generating the probability vector is empirically estimated. However, such a naive interval may fail to cover the true value of the function with the desired coverage level due to the simple fact that the uncertainty of the estimated “prior” distribution is not considered in this Empirical Bayesian approach (Casella, 1985; Carlin and Louis, 2000). Therefore, we have proposed one additional calibration step to correct the under-coverage of this naive empirical Bayesian confidence interval.

We have realized that a good part of understanding the performance of the constructed confidence interval should be summarized by extensive computations on simulated clinical or other suitable data. Some computations are given in the Numerical Study Section (Section 3) and real data example Section (Section 4), that follow the Method Section (Section 2). Additional suggestions for further research are topics of our Discussion Section(Section 5).

2. Method.
2.1. The General Framework. The complete data consist of \( n \) pairs of observations \( \{(Z_i, Y_i), i = 1, \cdots, n\} \). Let 
\[
(Z_i, Y_i) \overset{i.i.d.}{\sim} p(z, y \mid \theta_0), i = 1, \cdots, n,
\]
where \( p(z, y \mid \theta_0) \) is the density function for the joint distribution with respect to \((Z_i, Y_i), i = 1, \cdots, n\). Suppose that we only observe \( Z = \{Z_1, \cdots, Z_n\}, \ Y = \{Y_1, \cdots, Y_n\} \) is missing but correlated with \( Z = \{Z_1, \cdots, Z_n\} \), and \( \theta_0 \) is a fixed but unknown parameter. Our aim is to construct a confidence interval covering \( G(Y) = G(Y_1, \cdots, Y_n) \) with a specified probability based on observed data \( Z \) only, where \( G(\cdot) \) is a given function. Note that different from the conventional setting, where the parameter of interest is a deterministic population parameter, \( G(Y) \) is a random variable that varies from data to data. For this reason, we use credible interval and confidence interval interchangeably in the current paper with a slight abuse of notations.

Assuming that a point estimator \( \hat{\theta}(Z) \) for \( \theta_0 \) based on \( Z \) is available, we can derive the conditional distribution of \( Y_i \mid Z_i, \theta_0 = \hat{\theta}(Z), i = 1, \cdots, n \), and simulate multiple copies of \( G^* = G(Y) \) directly from the conditional distribution of \( G(Y) \mid Z, \theta_0 = \hat{\theta}(Z) \). A credible interval for \( G(Y) \) can then be constructed based on sampled \( G^* \)s. See Algorithm 1 for this naive approach.

**Algorithm 1** A Naive Algorithm to Construct a 95% credible Interval for \( G(Y_1, Y_2, \cdots, Y_n) \)

1: Find a point estimator of \( \theta_0, \hat{\theta}(Z) \).
2: Simulate \( Y^*_i \) from the conditional distribution \( Y_i \mid Z_i, \theta_0 = \hat{\theta}(Z) \), for \( i = 1, \cdots, n \).
3: Compute \( G^* = G(Y^*_1, \cdots, Y^*_n) \).
4: Repeat steps 2 and 3 a large number of times, obtain \( G^*_1, \cdots, G^*_B \), and return the 2.5th and 97.5th percentiles of the empirical distribution of \( \{G^*_1, \cdots, G^*_B\} \) as a 95% credible interval for \( G(Y) \).

In practice, however, the behavior of this interval is largely dependent on how well \( \hat{\theta}(Z) \) is in estimating \( \theta_0 \). A significant discrepancy may be observed between \( \theta_0 \) and \( \hat{\theta}(Z) \) resulting severe under-coverage. To make the constructed intervals robust to the variability of \( \hat{\theta}(Z) \), we propose to couple a method incorporating the variance of \( \hat{\theta}(Z) \) and a calibration step using the parametric bootstrap method (Efron and Tibshirani, 1994).

Specifically, we first obtain an estimator of \( \theta_0, \hat{\theta}(Z) \), and an estimator of the variance of \( \hat{\theta}(Z), \hat{J}(Z) \). Then, we attempt to construct the credible interval based on a “posterior” distribution:

\[
G(Y) \mid Z, \theta_0 = \theta^*; \theta^* \sim N \left\{ \hat{\theta}(Z), \hat{J}(Z) \right\},
\]
which, in general, has a larger variability than the naive “posterior” distribution
\[
G(Y) \mid Z, \theta_0 = \hat{\theta}(Z)
\]
and likely results in a wider credible interval. In addition, we also consider a calibration step based on parametric bootstrap (Efron and Tibshirani, 1994) similar to the correction introduced in Carlin and Gelfand (1990). To be specific, we simulate “observed” data from an assumed model with \( \theta_0 = \theta(Z) \), simulate realizations of \( G(Y) \) from the “posterior” distribution given above, and construct the \( 100(1 - \alpha)\% \) credible intervals for different \( \alpha \) based on the quantiles of samples drawn from the posterior distribution. After repeating this simulation a large number of times, we will examine the empirical coverage level of the constructed credible interval with respect to the true \( G(Y) \) in the simulated data, anticipating that the real coverage level of \( 100(1 - \alpha)\% \) credible intervals may differ from \( (1 - \alpha) \). We will find the \( \alpha \) level
such that the corresponding $100(1 - \alpha)\%$ credible interval has a coverage level of 95%. This $\alpha$ level will then be used to construct the credible interval based on the “posterior” distribution (1) from the original data. The mathematical rationale is that

$$P \{ G(Y) < q_\alpha(Z) \mid \theta_0 \} \approx P \{ G(Y) < q_\alpha(Z) \mid \hat{G}(Z) \},$$

where $q_\alpha(Z)$ is the $\alpha$th quantile of the “posterior” distribution (1), and the parametric bootstrap was used to estimate $P \{ G(Y) < q_\alpha(Z) \mid \hat{G}(Z) \}$. This step can be viewed as a calibration step based on a bootstrap method (Efron and Tibshirani, 1994). In the end, if denoting the resulting credible interval as $[\hat{L}(Z), \hat{U}(Z)]$, we expect that

$$P \{ \hat{L}(Z) \leq G(Y) \leq \hat{U}(Z) \} = 0.95,$$

where the probability is with respect to the joint distribution of $(Y, Z)$. The detailed steps are outlined in algorithm 2.

**Algorithm 2** A General Algorithm to Construct a 95% Confidence Interval for $G(\hat{\lambda})$

1: Obtain an point estimator $\hat{\theta}(Z)$ of $\theta_0$.
2: Compute a consistent variance estimator of $\hat{\theta}(Z)$, $J(Z)$.
3: for $b = 1, 2, \ldots, B$, do
4: Simulate $\tilde{\theta}_1(b), \ldots, \tilde{\theta}_n(b) \sim i.i.d. N(\hat{\theta}(Z), J(Z))$.
5: Simulate $Y_{11}(b) \sim p_Y(y \mid Z_1, \tilde{\theta}_1(b)), \ldots, Y_{n}(b) \sim p_Y(y \mid Z_n, \tilde{\theta}_n(b))$, where $p_Y(y \mid z, \theta)$ is the density function of $Y$ conditional on $Z = z$ and $\theta_0 = \theta$.
6: Compute $G^*_b(Z) = G(Y_{11}, \ldots, Y_{n})$.
7: end for
8: for $i = 1, 2, \ldots, R$, do
9: Simulate a new dataset $\{Y^*_i, Z^*_i\}_{j=1}^{\frac{n}{2}} \sim i.i.d. p(y, z \mid \hat{\theta}(Z))$. Let $Z^*_i = (Z^*_{ij})_{j=1}^{n}$ and $Y^*_i = (Y^*_{ij})_{j=1}^{n}$.
10: Calculate the random variable to be estimated: $G^*_i = G(Y^*_{11}, \ldots, Y^*_{n}) = G(Y^*_i)$.
11: Obtain the point estimator $\hat{\theta}(Z^*_i)$ of $\theta$ based on the generated data $Z^*_i$.
12: Obtain a variance estimator of $\hat{\theta}(Z^*_i)$, $J(Z^*_i)$.
13: Using Steps 6-9 to obtain $\{G_1(Z^*_i), \ldots, G_B(Z^*_i)\}$.
14: Construct the $100(1 - \alpha)\%$ confidence interval as the $100\alpha / 2$th percentile and $100(1 - \alpha / 2)$th percentile of $\{G_1(Z^*_i), \ldots, G_B(Z^*_i)\}$, denoted by $\hat{CI}_i(1 - \alpha)$.
15: end for
16: Calculate the empirical coverage level of $\hat{CI}_i(1 - \alpha)$ as

$$\frac{1}{R} \sum_{i=1}^{R} I \left( G^*_i \in \hat{CI}_i(1 - \alpha) \right),$$

where $I(\cdot)$ is an indicator function.
17: Determine the value of $\alpha_0$ such that the empirical coverage level of $\hat{CI}_i(1 - \alpha_0)$ is 95%, i.e.,

$$\frac{1}{R} \sum_{i=1}^{R} I \left( G^*_i \in \hat{CI}_i(1 - \alpha_0) \right) = 0.95$$

18: Return the 95% credible interval for $G(Y)$ based on observed data as the interval between the $100\alpha_0 / 2$th and $100(1 - \alpha_0 / 2)$th quantiles of $\{G_1(Z), \ldots, G_B(Z)\}$.

### 2.2. Interval Estimates of Clonality and Entropy

We now apply the general algorithm introduced in the previous section to constructing confidence intervals for clonality and entropy.
Consider a parametric model for the number of cells from a clone, i.e., cells with a particular V(D)J rearrangement:

\[ \lambda_i \sim \text{Gamma}(a_0, b_0), \quad i = 1, \ldots, C_0, \]

\[ Z_i \sim \text{Poisson}(\lambda_i), \quad i = 1, \ldots, C_0, \]

where \( C_0 \) is the total number of clones. \( \{\lambda_i\}_{i=1}^{C_0} \) and \( (a_0, b_0) \) are unknown to us, but we observe \( Z_i \) if \( Z_i > 0 \), which is the number of reads corresponding to the \( i \)th V(D)J rearrangement. Therefore, the observed data consist of the truncated independent Poissons with parameters \( \lambda_i, \ldots, \lambda_{C_0} \):

\[ \{Z_i | Z_i > 0, \ 1 \leq i \leq C_0\} \]

Here \( \{Z_i | Z_i = 0, \ 1 \leq i \leq C_0\} \) are missing data. Now, we define clonality as \( G_C \) and entropy as \( G_E \):

\[ G_C := \sum_{i=1}^{C_0} \left( \frac{\lambda_i}{\sum_{j=1}^{C_0} \lambda_j} \right)^2, \]

and

\[ G_E := -\sum_{i=1}^{C_0} \left\{ \frac{\lambda_i}{\sum_{j=1}^{C_0} \lambda_j} \log \left( \frac{\lambda_i}{\sum_{j=1}^{C_0} \lambda_j} \right) \right\}. \]

Both are functions of \( \{\lambda_1, \cdots, \lambda_{C_0}\} \). We aim to obtain point estimations as well as confidence intervals for \( G_C \) and \( G_E \).

To fit into the general framework described above, we use an estimator of \( C_0 \), denoted by \( \hat{C} \), to replace \( C_0 \). Then \( \mathcal{Z} \) corresponds to data \( \{Z_1, \cdots, Z_{\hat{C}}\} \), which are partially observed, \( \mathcal{Y} \) corresponds to \( \{\lambda_1, \cdots, \lambda_{\hat{C}}\} \) and \( \theta_0 \) corresponds to \( (a_0, b_0) \), the shape and rate parameters in the gamma distribution. Below, we outline the algorithm step by step.

### 2.2.1. Estimation of \( (a_0, b_0) \) and \( C_0 \).

Marginally, \( Z_i \) follows a negative binomial distribution, i.e.,

\[
\mathbb{P}(Z_i = z) = p_Z(z | a_0, b_0) = \int_0^\infty \frac{e^{-\lambda} \lambda^z}{z!} \frac{\lambda^{a_0-1} e^{-b_0 \lambda} b_0^{a_0}}{\Gamma(a_0)} d\lambda
\]

\[ = \left( \frac{z + a_0 - 1}{z} \right) \left( \frac{b_0}{b_0 + 1} \right)^{a_0} \left( \frac{1}{b_0 + 1} \right)^z, \]

where \( p_Z(\cdot | a, b) \) represents the density function of the distribution of \( Z_i \). We aim to first estimate the parameters \( a_0 \) and \( b_0 \) in the gamma distribution based on observed data consisting of positive reads

\[ \mathcal{Z}_O = \{Z_i | Z_i > 0, i = 1, \cdots, C_0\}. \]

The resulting log-likelihood function in terms of \( a_0, b_0 \) is

\[
l(a, b) = \sum_{z_i \in \mathcal{Z}_O} \log \left\{ \frac{p_Z(z_i | a, b)}{1 - p_Z(0 | a, b)} \right\}, \tag{2}
\]

which can be maximized via an EM algorithm (Dempster, Laird and Rubin, 1977; McLachlan and Jones, 1988).
If $C_0$ is known, i.e., we observed clones with $Z_i = 0$, the full log-likelihood function becomes

$$
\sum_{z_i \in Z_0} \log\{p(z_i|a_0, b_0)\} + \sum_{z_i=0} \log\{p(0|a_0, b_0)\}.
$$

Therefore, the EM algorithm consists of the following E and M steps.

- **E-step:** For a given estimator $\hat{a}$ and $\hat{b}$, estimate $E[C_0 - C | \hat{a}, \hat{b}]$ as

$$
E[C_0 - C | \hat{a}, \hat{b}] = \frac{p_Z(0 | \hat{a}, \hat{b})}{1 - p_Z(0 | \hat{a}, \hat{b})} C
$$

based on the relationship

$$
p_Z(0 | \hat{a}, \hat{b})(E[C_0 - C | \hat{a}, \hat{b}] + C) = E[C_0 - C | \hat{a}, \hat{b}],
$$

where $C$ is the number of observed nonzero $Z_i$s. Since

$$
p_Z(0 | a_0, b_0) = \left(\frac{b_0}{b_0 + 1}\right)^{a_0},
$$

we have

$$
\hat{n}_0 = E[C_0 - C | \hat{a}, \hat{b}] = \frac{\hat{b}^\hat{a} C}{(\hat{b} + 1)^\hat{a} - \hat{b}^\hat{a}}
$$

- **M-step:** Update the MLE of $a_0$, $b_0$ by maximizing

$$
\sum_{z_i \in Z_o} \log\{p_Z(z_i | a_0, b_0)\} + \hat{n}_0 \log\{p_Z(0 | a_0, b_0)\}.
$$

This optimization can be achieved via an **Inner EM algorithm** treating $\{\lambda_i\}$ as missing variables:

1. **E-step:** For given $(\hat{a}, \hat{b})$

$$
E[\lambda_i | \hat{a}, \hat{b}] = \frac{Z_i + \hat{a}}{1 + \hat{b}},
$$

$$
E[\log(\lambda_i) | \hat{a}, \hat{b}] = \Psi(Z_i + \hat{a}) - \log(1 + \hat{b}), \quad i = 1, \ldots, C,
$$

$$
E[\lambda_0 | \hat{a}, \hat{b}] = \frac{\hat{a}}{1 + \hat{b}},
$$

$$
E[\log(\lambda_0) | \hat{a}, \hat{b}] = \Psi(\hat{a}) - \log(1 + \hat{b}),
$$

where $\lambda_0$ corresponds to the Poisson rate of unobserved clones with $Z_i = 0$ and $\Psi(\cdot)$ is the dgamma function, the derivative of the log-transformed Gamma function, i.e.,

$$
\Psi(x) = \left[\log \left\{ \int_0^\infty t^{x-1}e^{-t}dt \right\} \right]'.
$$

2. **M-step:** Maximize the log-likelihood function

$$
l(a_0, b_0)
$$

$$
= (a_0 - 1) \left[ \sum_{i=1}^C E[\log(\lambda_i) | \hat{a}, \hat{b}] + \hat{n}_0 E[\log(\lambda_0) | \hat{a}, \hat{b}] \right]
$$
\[-(C + \hat{n}_0)[\log \Gamma(a_0) - a_0 \log(b_0)]\]
\[-b_0 \left[ \sum_{i=1}^{C} \mathbb{E}[\lambda_i | \hat{a}, \hat{b}] + \hat{n}_0 \mathbb{E}[\lambda_0 | \hat{a}, \hat{b}] \right]\]
\[= (a_0 - 1) \left[ \sum_{i=1}^{C} \{ \Psi(z_i + \hat{a}) - \log(1 + \hat{b}) \} + \hat{n}_0 \{ \Psi(\hat{a}) - \log(1 + \hat{b}) \} \right]\]
\[-(C + \hat{n}_0)[\log \Gamma(a_0) - a_0 \log(b_0)] - b_0 \sum_{i=1}^{C} (z_i + \hat{a}) + \hat{n}_0 \hat{a} \frac{1}{1 + \hat{b}}\]

In practice, we iterate the inner EM algorithm to maximize
\[\sum_{z_i \in Z_O} \log \{ p_Z(z_i | a_0, b_0) \} + \hat{n}_0 \log \{ p_Z(0 | a_0, b_0) \}\]
for a given \(\hat{n}_0\) and iterate the outer EM algorithm until \(\hat{n}_0\) converges. The final convergence of \((\hat{a}, \hat{b}, \hat{n}_0)\) can be assessed by the relative change of the log-likelihood:
\[\sum_{z_i \in bZ_O} \log \left\{ \frac{p_Z(z_i | \hat{a}, \hat{b})}{1 - p_Z(0 | \hat{a}, \hat{b})} \right\}.
whose value should increase with each iteration of the outer EM algorithm.

2.2.2 Estimation of the variance of \((\hat{a}, \hat{b})\). After obtaining the MLE \((\hat{a}, \hat{b})\) for \((a_0, b_0)\) using the EM algorithm, we can estimate the variance of \((\hat{a}, \hat{b})\) by the inverse of the observed information matrix, which can be calculated by taking the second derivative of the log-likelihood function w.r.t. \(a\) and \(b\), i.e.,
\[
\hat{J}(Z_O) = \begin{pmatrix}
\frac{\partial^2 l(a, b)}{\partial a^2} & \frac{\partial^2 l(a, b)}{\partial a \partial b} \\
\frac{\partial^2 l(a, b)}{\partial b \partial a} & \frac{\partial^2 l(a, b)}{\partial b^2}
\end{pmatrix}
\bigg|_{(a, b) = (\hat{a}, \hat{b})}^{-1},
\]
where \(l(a, b)\) is given in (2).

2.2.3 Sampling from the posterior distribution. In constructing the 95% confidence interval, we need to generate
\[
\begin{pmatrix} a^*_i \\ b^*_i \end{pmatrix} \sim \exp \left\{ N \left( \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}, \hat{A}(Z_O) \hat{J}(Z_O) \hat{A}(Z_O) \right) \right\}
\]
and
\[\lambda^*_i \sim \lambda_i | Z_i, a^*_i, b^*_i;\]
which is a gamma distribution with the shape and rate parameters being \(a^*_i + Z_i\) and \(b^*_i + 1\), respectively, where \(\hat{A}(Z_O) = \text{diag}(\hat{a}^{-1}, \hat{b}^{-1})\). This approach ensures that sampled \(a^*_i\) and \(b^*_i\) are always positive. Operationally, we pretend that \(C_0 = \hat{C}\) and let
\[Z = \{ Z_1, Z_2, \ldots, Z_C, Z_{C+1} = 0, \ldots, Z_{\hat{C}} = 0 \} = Z_O \cup \{ 0, \ldots, 0 \}.
\]
and
\[Y = \{ \lambda_1, \lambda_2, \ldots, \lambda_C \}.\]
The credible intervals $\hat{C}I_C(Z_O)$ and $\hat{C}I_E(Z_O)$ are calibrated such that

$$P\left\{ \sum_{i=1}^{\hat{C}} \left( \frac{\lambda_i}{\sum_{j=1}^{C} \lambda_j} \right)^2 \in \hat{C}I_C(Z_O) \mid \hat{a}, \hat{b} \right\} = 0.95$$

and

$$P\left\{ \sum_{i=1}^{\hat{C}} - \frac{\lambda_i}{\sum_{j=1}^{C} \lambda_j} \log \left( \frac{\lambda_i}{\sum_{j=1}^{C} \lambda_j} \right)^2 \in \hat{C}I_E(Z_O) \mid \hat{a}, \hat{b} \right\} = 0.95$$

The full algorithm is provided in Algorithm 3. Note the algorithm for constructing confidence intervals for entropy is the same as that for clonality except that we replace all $G$ by $E$.

### Algorithm 3: The Algorithm to Construct a 95% Confidence Interval for Clonality

1. Use the proposed EM algorithm to obtain the MLE for $a_0$, $b_0$, and $C_0$, denoted by $\hat{a}$, $\hat{b}$, and $\hat{C}$, respectively.

2. Compute $\hat{J}(Z_O)$ as the inverse of the observed information matrix for $a_0$ and $b_0$ and $\hat{A}(Z_O)$.

3. Let $Z = \{Z_1, \cdots, Z_C, Z_{C+1} = 0, \cdots, Z_C = 0 \}$.

4. for $b = 1, \cdots, B$, do

5. Simulate $(a_{i(b)}^*, b_{i(b)}^*) \sim \text{Gamma}(a(i), b(i))$ for $i = 1, \cdots, \hat{C}$.

6. Simulate $\lambda_{i(b)}^* \sim \text{Gamma} \left( \lambda_{i}^* + z_i b_{i(b)}^* + 1 \right)$ for $i = 1, \cdots, \hat{C}$.

7. Compute $G_C^*(b)(Z_O) = \sum_{i=1}^{\hat{C}} \left( \frac{\lambda_{i(b)}^*}{\sum_{j=1}^{C} \lambda_{j(b)}^*} \right)^2$.

8. end

9. for $i = 1, \cdots, R$, do

10. Simulate $(\lambda_{i1}^*, \cdots, \lambda_{i\hat{C}}^*) \sim \text{Gamma}(\hat{a}, \hat{b})$ and let $A_i^* := \left\{ \lambda_{ij}^* \mid j = 1, \cdots, \hat{C} \right\}$.

11. Compute $G_C(A_i^*) = \sum_{j=1}^{\hat{C}} \left( \frac{\lambda_{ij}^*}{\sum_{j=1}^{C} \lambda_{j}^*} \right)^2$.

12. Simulate a new set of data $Z_{iO}^* = \{Z_{ij}^* \mid j = 1, \cdots, \hat{C} \}$ where $Z_{ij}^* \sim \text{Poisson}(\lambda_{ij}^*)$.

13. Induce the observed data $Z_{iO} = \{Z_{ij} \mid Z_{ij} > 0, j = 1, \cdots, \hat{C} \}$.

14. Repeat steps 4-5 for $Z_{iO}^*$ and obtain $\left\{G_C^*(1)(Z_{iO}^*), \cdots, G_C^*(B)(Z_{iO}^*) \right\}$.

15. Let $\hat{C}I_{\alpha}(Z_{iO}^*)$ be the interval between the $100\alpha/2$th and $100(1 - \alpha/2)$th percentiles of $\left\{G_C^*(1)(Z_{iO}^*), \cdots, G_C^*(B)(Z_{iO}^*) \right\}$ for various $\alpha$.

16. end

17. Determine the value of $\hat{\alpha}_0$ such that the proportion of $G_C(A_i^*)$ falls in $\hat{C}I_{\hat{\alpha}_0}(Z_{iO}^*)$ is closest to 95%, i.e.,

$$\frac{1}{R} \sum_{i=1}^{n} I \left\{ G_C(A_i^*) \in \hat{C}I_{\hat{\alpha}_0}(Z_{iO}^*) \right\} \approx 0.95.$$ 

18. Return the $100\alpha/2$th and $100(1 - \alpha/2)$th percentiles of $\left\{ G_C^*(b)(Z_O), b = 1, \cdots, B \right\}$, which is our 95% confidence interval for the clonality.

### 3. Simulation Study

We have conducted a comprehensive simulation study to examine the empirical performance of the constructed confidence intervals for both clonality and entropy. For each given set of $(a_0, b_0, C_0)$, we have repeated the experiments 500 times to...
compute the empirical coverage level of the constructed confidence intervals. In constructing the confidence interval, we set the number of resampling for the calibration to be $R = 200$. The number of posterior sampling was $B = 500$. $C_0$, the number of clones was set at 10,000. A higher $C_0$ is expected to yield better confidence intervals. The values of $a_0$ and $b_0$ were fixed at their maximum likelihood estimators based on data examples in Section 4 to mimic real practice. For comparison purpose, we have also constructed the confidence interval based on the naive Bayesian procedure given in Algorithm 1 and the confidence interval based on the posterior distribution without the calibration step.

The simulation results were summarized in Table 1. The empirical coverage level of our proposed confidence intervals was close to its nominal level. On the other hand, the confidence intervals based on the Naive EB approach undercovered the true parameter, which was expected considering the fact that the variability of $(\hat{a}, \hat{b})$ was ignored in constructing the confidence interval. The coverage level of confidence intervals based on the posterior distribution directly was higher than the nominal level, suggesting that those intervals were too conservative. It could be due to the fact that this method did not consider the fact that $(\hat{a}, \hat{b})$ was also depended on $Z_O$ and thus was correlated with quantiles of the posterior distribution. This observation confirmed the essential role played by the calibration step, which was also the most computationally intensive part of the proposed algorithm. We have also plotted confidence intervals for the clonality and their corresponding true clonalities from 500 data simulated with $(a_0, b_0, C_0) = (0.086, 0.111, 10,000)$ in Figure 1. The confidence intervals are sorted according to the size of true clonalities. It is clear that, unlike the conventional case, the true clonalities have quite substantial variation relative to the width of the 95% confidence intervals, highlighting the importance of treating $G(Y)$ as a random quantity.

Next, we examined the empirical performance of the proposed confidence intervals when the underlying rates $\lambda_i, i = 1, \cdots, C_0$ do not follow a gamma distribution, i.e., the assumed Poisson-Gamma model was misspecified in characterizing the data generation process. Specifi-
The simulation results on empirical coverage level of constructed confidence intervals

| (a₀, b₀)         | Method       | Clonality Coverage | Entropy Coverage |
|------------------|--------------|--------------------|------------------|
| (0.732, 0.882)   | EB w Calibration | 94.0%              | 92.2%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 63.2%              | 45.6%            |
| (0.414, 0.335)   | EB w Calibration | 91.2%              | 95.8%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 78.0%              | 69.4%            |
| (0.596, 0.960)   | EB w Calibration | 98.4%              | 96.6%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 67.2%              | 40.6%            |
| (0.551, 0.775)   | EB w Calibration | 97.0%              | 96.0%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 68.4%              | 44.8%            |
| (0.171, 0.301)   | EB w Calibration | 98.4%              | 99.2%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 88.6%              | 70.6%            |
| (0.126, 0.132)   | EB w Calibration | 95.8%              | 98.2%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 94.0%              | 83.8%            |
| (0.0860, 0.111)  | EB w Calibration | 94.4%              | 99.6%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 94.2%              | 84.8%            |
| (0.113, 0.142)   | EB w Calibration | 95.6%              | 98.6%            |
|                  | EB w/o Calibration | 100%              | 100%             |
|                  | Naive EB      | 91.2%              | 83.2%            |

The simulation results when the underlying model for rate is log-normal rather than Gamma

| (μ₀, σ₀²)         | Coverage for Clonality | Coverage for Entropy |
|-------------------|------------------------|----------------------|
| (−1.38, 1.64²)    | 80.4%                  | 98.2%                |
| (−1.27, 1.72²)    | 86.8%                  | 97.8%                |
| (−1.22, 1.50²)    | 80.2%                  | 98.4%                |
| (−1.02, 1.62²)    | 82.7%                  | 95.8%                |

Gally, we simulated λᵢ from a log-normal distribution $\exp \{ N(μ₀, σ₀²) \}$ with chosen $μ₀$ and $σ₀$ and generated the observed reads $Z_O$. Using the same steps described above, we constructed confidence intervals based on 500 simulated datasets and calculated the empirical coverage level of these confidence intervals. The results are summarized in Table 2. The empirical coverage level of entropy was quite close to 95% even though the Poisson-Gamma parametric model was misspecified, implying the robustness of the proposed confidence intervals. On the other hand, the constructed confidence interval for clonality severely underestimate the true parameter suggesting the importance of assuming an appropriate distribution of the Poisson rates.
4. Real Data Studies. In this section, we illustrate our approach by applying it to analyze a recent study conducted by Qi et al. (2014). The objective was to investigate human T cell receptor (TCR) diversity. Specifically, we were interested in measuring TCR diversity by clonality and entropy. In the study conducted by Qi et al. (2014), five replicate TCR libraries of CD4 naive T cells and CD4 memory T cells are sequenced from each of seven participants. The total number of reads varied from $8.9 \times 10^4$ to $7.4 \times 10^5$. First, we counted the total number of reads in each clone across five replicates, and Figure 2 shows the observed cumulative proportions of clones sorted from the largest to the smallest for both CD4 naive and memory T cells. From the Figure, it is clear that large clones contain a bigger proportion of CD4 memory T cells than naive CD4 T cells, reflecting the relative evenness of the distribution of clone sizes of naive T cells.

Next, we apply the proposed method to construct confidence intervals of log-transformed clonality and entropy based on data from five replicates per patient, which are plotted in Figures 3 and 4. It is clear that other than a few outliers the confidence intervals based on different replicates of the same participant are fairly consistent, suggesting a low within-person variation relative to between-person variation supporting the validity of the experiment result. The clonality of CD4 memory T cells is substantially greater than that of the CD4 naive T
Fig 3: 95% confidence intervals for clonality based on five replicates per patient (in the log scale for better visualization)

cells, again confirming observations on the evenness of the distribution of the naive T cell clones in Figure 2.

Lastly, we estimate the “average” (log-transformed) clonality and entropy of naive CD4 T cells for three participants younger than 40 and four participants older than 70, separately, based on a random effects model used in meta analysis. We have then compared the clonality and entropy between younger and older participants. The average log-transformed clonality for memory T cells was -9.45 [-9.66 to -9.24] for young participants and -7.79 [-9.09 to -6.50] for old participants with a two-sided p-value of 0.014 for two group comparison. The average entropy is -10.97 [-11.17 to -10.76] for young participants and -10.37 [-10.87 to 9.88] for old participants with a two-sided p-value of 0.029 for two group comparison. These results suggest that immune diversity in old participants is lower than that in young participants, as anticipated, implying the aging effect on human immune system.

5. Discussion. In this paper, we have discussed a method for constructing confidence intervals for entropy and clonality: both are functions of a high dimensional probability vector. The method was developed under a general Empirical Bayesian framework and thus is parametric in nature. As a consequence, the performance of the confidence interval depends on relevant parametric assumptions and the gamma distribution for the intensity rate in particular, even though the simulation study demonstrates certain robustness in the performance of constructed confidence intervals. One generalization is to replace this gamma distribution by a more flexible distribution. For example, it is appealing to consider distributions from a nonparametric exponential family: \( p(\lambda | \eta) \propto p_0(\lambda) \exp \{B(\lambda)'\eta\} \), where \( p(\cdot | \eta) \) is the density function of the intensity rate, and \( B(\lambda) \) is a set of flexible basis functions given a priori such as \( B(\lambda) = (\lambda, \lambda^2, \lambda^3)' \) (Schwartzman, 2008). The extension in this direction warrants further study. Lastly, in the proposed approach, the actual number of distinct clones was replaced by its estimator, which may affect the performance of the subsequent point and
interval estimation. It is conceivable that the impact is greater for some functions such as entropy which is more sensitive to small-size clones than other functions such as clonality which is robust to small-size clones. However, estimating the number of distinct clones is analogous to estimating the number of unseen species, which is a difficult problem and in the current case depends on the parametric assumption for the intensity rate (Efron and Thisted, 1976). Therefore, it is important to study the impact of this estimator on the construction of the confidence interval for different diversity parameters.

REFERENCES

Carlin, B. P. and Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. Journal of the American Statistical Association 85 105–114.
| Patient | $\hat{a}$ | $\hat{b}$ | $\hat{C}$ | Clonality (95% CI)* | Entropy (95% CI) |
|---------|---------|---------|---------|---------------------|------------------|
| Patient 1 | 0.171   | 0.301   | 242994  | 6.06 (2.61, 8.73)   | -10.79 (-11.01, -10.52) |
| Patient 2 | 0.126   | 0.132   | 415556  | 9.21 (8.90, 9.50)   | -10.84 (-10.88, -10.79) |
| Patient 3 | 0.086   | 0.111   | 399632  | 239 (235, 244)      | -9.66 (-9.67, -9.65)   |
| Patient 4 | 0.113   | 0.142   | 443870  | 37.0 (35.3, 38.5)   | -10.59 (-10.65, -10.52) |
| Patient 5 | 0.134   | 0.184   | 391341  | 9.70 (9.10, 10.3)   | -10.83 (-10.88, -10.77) |
| Patient 6 | 0.150   | 0.204   | 439771  | 7.14 (6.74, 7.48)   | -11.11 (-11.16, -11.07) |
| Patient 7 | 0.163   | 0.293   | 204378  | 33.5 (31.6, 35.5)   | -10.448 (-10.452, -10.445) |

* $\times 10^{-5}$

CARLIN, B. P. and LOUIS, T. A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association* 95 1286–1289.
CASSELLA, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician* 39 83–87.
CHAO, A. and SHEN, T.-J. (2003). Nonparametric estimation of Shannon’s index of diversity when there are unseen species in sample. *Environmental and ecological statistics* 10 429–443.
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39 1–22.
EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biomетrika* 63 435–447.
EFRON, B. and TIBSHIRANI, R. J. (1994). *An introduction to the bootstrap*. CRC press.
GLANVILLE, J., HUANG, H., NAU, A., HATTON, O., WAGAR, L. E., RUBElt, F., JI, X., HAN, A., KRAMS, S. M., PETTUS, C. et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547 94–98.
MCLACHLAN, G. and JONES, P. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics* 571–578.
QI, Q., LIU, Y., CHENG, Y., GLANVILLE, J., ZHANG, D., LEE, J.-Y., OLSHEN, R. A., WEYAND, C. M., BOYD, S. D. and GORONZY, J. J. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences* 111 13139–13144.
SCHWARTZMAN, A. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics* 2 1332–1359.
TIAN, L., LIU, Y., FIRE, A. Z., BOYD, S. D. and OLSHEN, R. A. (2019). Clonality: point estimation. *The Annals of Applied Statistics* 13 113–131.
ZUREK, W. H. (2018). *Complexity, entropy and the physics of information*. CRC Press.