flan: An R Package for Inference on Mutation Models

Adrien Mazoyer, Rémy Drouilhet, Stéphane Despréaux, Bernard Ycart

flan: An R Package for Inference on Mutation Models.

by Adrien Mazoyer, Rémy Drouilhet, Stéphane Despréaux, and Bernard Ycart

Abstract

This paper describes **flan**, a package providing tools for fluctuation analysis of mutant cell counts. It includes functions dedicated to the distribution of final numbers of mutant cells. Parametric estimation and hypothesis testing are also implemented, enabling inference on different sorts of data with several possible methods. An overview of the subject is proposed. The general form of mutation models is described, including the classical models as particular cases. Estimating from a model, when the data have been generated by another, induces different possible biases, which are identified and discussed. The three estimation methods available in the package are described, and their mean squared errors are compared. Finally, implementation is discussed, and a few examples of usage on real data sets are given.

1 Introduction

Mutation models are probabilistic descriptions of the growth of a population of cells, where mutations occur randomly during the process. Data are samples of integers, interpreted as final numbers of mutant cells. These numbers may be coupled with final numbers of cells (mutant and non mutant). The frequent appearance in the data of very large mutant counts, usually called "jackpots", evidences heavy-tailed probability distributions. The parameter of interest is the mutation probability for a mutant cell to appear upon any given cell division, denoted by π. In practice, π is typically of order 10^{-9}–10^{-11}. Computing robust estimates for π is of crucial importance in medical applications, like cancer tumor relapse or multidrug resistance of *Mycobacterium Tuberculosis* for instance.

Any mutation model can be interpreted as the result of the three following ingredients:

- a random number of mutations occurring with small probability among a large number of cell divisions. Due to the law of small numbers, the number of mutations approximately follows a Poisson distribution. The expectation of that distribution, denoted by α, is the product of the mutation probability π with the total number of divisions.
- from each mutation, a clone of mutant cells growing for a random time. Due to exponential growth, most mutations occur close to the end of the experiment, and the developing time of a random clone has exponential distribution. The rate of that distribution, denoted by ρ, is the relative fitness, i.e. the ratio of the growth rate of normal cells to that of mutants.
- the number of mutant cells that any clone developing for a given time will produce. The distribution of this number depends on the distribution of division times of mutants.

Using the theory of continous time branching processes [Bellman and Harris, 1952, Athreya and Ney, 1972], and under specific modeling assumptions, it can be proved that the asymptotic distribution of the final number of mutants has an explicit form. A first mutation model with explicit distribution is the well known Luria-Delbrück model [Luria and Delbrück, 1943]. Other mathematical models were introduced by Lea and Coulson [1949], followed by Armitage [1952] and Bartlett [1978]. In these models, division times of mutant cells were supposed to be exponentially distributed. Thus a clone develops according to a Yule process, and its size at a given time follows a geometric distribution. The distribution of final mutant counts is also explicit when division times are supposed to be constant. This latter model is called Haldane model by Sarkar [1991]; an explicit form of the asymptotic

distribution is given in Ycart [2013]. General division times have been studied by Ycart [2013], but no explicit distribution is available appart from the exponential and constant division times.

The first estimation method was given by Luria and Delbrück [1943]. It is based on the simple relation between the probability of null counts in the sample, and the mutation probability, and it is called P0 method. Of course, if the sample does not contain null counts, the method cannot be applied. Appart from the P0 method, all other methods couple the estimation of π or α, with the estimation of ρ. When the distribution of final numbers has an explicit form, the Maximum Likelihood (ML) is an obvious optimal choice [Ma et al., 1992, Zheng, 2005]. However, because of the jackpots, likelihood computation can be numerically unstable. There are several ways to reduce tail effects [Wilcox, 2012, Sec. 2.2], among which "Winsorization" consists in truncating the sample beyond some maximal value. Another estimation method uses the probability generating function (GF) [Rémillard and Theodorescu, 2000, Hamon and Ycart, 2012]. The estimators of α and ρ obtained with the GF method proved to be close to optimal efficiency, with a broad range of calculability, a good numerical stability, and a negligible computing time. For the three methods, P0, ML, and GF, the estimators of α and ρ are asymptotically normal. Thus confidence intervals and p-values for hypothesis testing can be computed, for one sample and two sample tests.

The problem with classical mutation models, is that they are based on quite unrealistic assumptions: constant final number of cells [Angerer, 2001, Komarova et al., 2007, Ycart and Veziris, 2014], no cell deaths (Angerer [2001, Sec. 3.1]; Dewanji et al. [2005], Komarova et al. [2007], Ycart [2014], or, as mentioned above, exponential distribution of division times. Using a model for estimation, when the data have been generated by another one, necessarily induces a bias on estimates. For instance, if cell deaths are neglected, mutation probability will be underestimated.

The package **flan** described here, is dedicated to mutation models, and parameter estimation with the three methods P0, ML, and GF. It includes a set of functions for the distribution of mutant cell counts (`dflan`, `pflan`, `qflan`, `rflan`) and a graphic function (`draw.clone`). They treat general models, with fluctuating final numbers, cell deaths, and other division time distributions than exponential and contant. The general estimation function is `mutestim`. It returns estimates for the parameters α, π and ρ, with the three estimation methods, constant or exponential division times, and cell deaths. As a wrapper, a hypothesis testing function (`flan.test`) is provided. In order to make the package user-friendly, the functions have been designed to resemble classical R functions, like `t.test` or `rnorm`.

The paper is organized as follows. Section 2 is devoted to the probabilistic setting: the hypotheses of the different models are described, and the asymptotic results are explained. In section 3, the three estimation methods are exposed, and the biases described above are discussed. A comparison of the three methods in terms of mean squared errors is provided. The user interface and the **Rcpp** implementation is treated in section 4; examples of execution are shown in section 5.

2 Mutation models

In this section, probabilistic mutation models are described. The basic modeling hypotheses are the following:

- at time 0 a homogeneous culture of n_0 normal cells is given;
- the lifetime of any normal cell is a random variable with distribution function F;
- upon completion of the generation time of a normal cell:
  - with probability π one normal and one mutant cell are produced;
  - with probability $1-\pi$ two normal cells are produced;
- the lifetime of any mutant cell is a random variable with distribution function G;
- upon completion of the lifetime of a mutant cell:
  - with probability δ the cell dies out;
  - with probability $1-\delta$ two mutant cells are produced;
- all random variables and events (division times, mutations, and deaths) are mutually independent.

Consider that the initial number n_0 tends to infinity, the mutation probability $\pi = \pi_{n_0}$ tends to 0, and the time $t = t_{n_0}$ at which mutants are counted tends to infinity. The scale of time is supposed to be adjusted so that the exponential growth rate of mutants is 1; thus the exponential growth rate of normal cells is ρ. See Athreya and Ney [1972, Chap. IV Sec. 4] or Hamon and Ycart [2012] for the definition of the growth rate (also called "Malthusian parameter"). The expected number of mutations before t_{n_0} is proportional to $n_0\pi_{n_0}\mathrm{e}^{\rho t_{n_0}}$, and the asymptotics are assumed to be such that this number converges as n_0 tends to infinity to α, positive and finite.

Under the above hypotheses, as n_0 tends to $+\infty$, the final number of mutants converges in law to the distribution with PGF:

$$g(z) = \exp\left(-\alpha(1-h(z))\right) , \tag{1}$$

with

$$h(z) = \int_0^\infty \psi(z,t)\rho\mathrm{e}^{-\rho t}\mathrm{d}t \, , \tag{2}$$

where $\psi(z,t)$ is the PGF of the number of cells at time t in a mutant clone, starting from a single cell at time 0. Observe that it depends on the lifetime distribution of normal cells F only through ρ. The above result is deduced from the theory of continous time branching processes [Hamon and Ycart, 2012]). The expressions (1) and (2) translate the three ingredients described in the introduction:

1. the Poisson distribution with intensity α models the total number of mutations which occur during the process;
2. the exponential distribution with rate ρ is that of the time during which a random clone develops;
3. the distribution with PGF $\psi(\cdot,t)$ is that of the number of cells in a random clone developing during a time interval of length t. It is the solution of a Bellman-Harris equation [Bellman and Harris, 1952] in terms of δ and G.

Hence the expressions of h as an exponential mixture, and of g as a Poisson compound.

The PGF (1) defines a parameterized family of distributions, denoted herafter by $MM(\alpha,\rho,\delta,G)$ (Mutation Model). This is a family of heavy-tailed distributions, with tail exponent ρ: the higher the fitness, the heavier the tail. This directly influences the number and the amount of jackpots.

At this point, the PGF ψ can be given as an explicit expression only for two particular lifetime distributions G: exponential, and Dirac (constant lifetimes). The corresponding mutation models will be denoted respectively by $LD(\alpha,\rho,\delta)$ (Luria-Delbrück), and $H(\alpha,\rho,\delta)$ (Haldane). The functions `dflan`, `pflan`, and `qflan` compute densities, probabilities, quantiles of LD and H distributions.

Assuming that a consistent estimator of α has been defined, the problem in practice is to compute reliable estimates of the mutation probability π. The simplest approach assumes that the final number of cells, denoted by N, is constant. An estimate of π is then obtained by dividing the estimate of α by N. However, even under close experimental monitoring, assuming that the final number of cells is a constant is quite unrealistic. Thus, N must be viewed as a random variable with a certain probability distribution function K on $[0,+\infty)$. By analogy with (1), the conditional PGF of the number of mutants given $N = n$, can be given by the following expression:

$$g\left(z \,|\, N=n\right) = \exp\left(-\pi n(1-h(z))\right) .$$

Or else, the conditional distribution of the number of mutants given $N = n$ is the distribution $MM(\pi n,\rho,\delta,G)$. The distribution function K is supposed to be known and its Laplace transform is denoted by $\mathcal{L}$:

$$\mathcal{L}(z) = \mathbb{E}\left[\,\mathrm{e}^{-zN}\,\right] = \int_0^\infty \mathrm{e}^{-zn}\mathrm{d}K(n) \, ,$$

Thus the PGF of the final number of mutants is given by:

$$g(z) = \int_0^\infty g\left(z \,|\, N=n\right)\mathrm{d}K(n) = \mathcal{L}\left(\pi(1-h(z))\right) . \tag{3}$$

Remark that if N is constant, (3) reduces to (1) with $\alpha = \pi N$. In general, the PGF (3) defines a new parametrized family of mutation distributions, denoted herafter by $MMFN(\pi,\rho,\delta,G,K)$ (Mutation Models with Fluctuating Numbers of cells).

The two particular cases for the distribution G previously mentioned above (exponential and Dirac) will be denoted by $LDFN(\alpha, \rho, \delta, K)$ (Luria-Delbrück with Fluctuating Nmbers of cells) and $HFN(\alpha, \rho, \delta, K)$ (Haldane with Fluctuating Numbers of cells). As will be shown in section 3, estimating π by the ratio of an estimate of α by the expectation of N induces a negative bias.

The function `rflan` outputs samples of pairs (mutant counts–final counts) following $MMFN$ distributions where G is an exponential, Dirac, log-normal or gamma distribution, and K is a log-normal or Dirac distribution.

3 Statistical inference

Here the three estimation methods P0, ML and GF are described. The main features and the limitations of each method are discussed. The three methods compute estimates of α and ρ, under the LD and H models. When couples (mutant counts–final numbers) are given, estimates of π and ρ are calculated under the $LDFN$ or HFN models.

Even if the probabilities and their derivatives with respect to δ for LD and H distributions can be computed, the variations of the whole distribution as a function of δ are too small to enable estimation in practice (see Ycart [2014] for more details). Thus, the parameter δ is supposed to be known for the three methods.

In the rest of this section, the three estimators are described, their performances compared in terms of MSE, and the possible sources of biases discussed.

3.1 Estimators

P0 estimator: The first method was introduced by Luria and Delbrück [1943] when $\delta = 0$. In that case, the probability of null counts in the sample is $\mathrm{e}^{-\alpha}$. Hence α can be estimated taking the negative logarithm of the relative frequency of zeros among mutant counts. Hence the method cannot be applied if the sample does not contain null counts.

If $\delta > 0$, the probability of null counts in the sample depends also on δ. Assuming $\delta < 1/2$, a fixed point of the PGF $\psi(\cdot, t)$ is the extinction probability of a mutant clone [Athreya and Ney, 1972, Theorem 1, Chap.I]:

$$\delta_* = \frac{\delta}{1-\delta}\,.$$

By definition, δ_* is also a fixed point of the PGF (2). Then the probability of null counts in the sample is $\mathrm{e}^{-\alpha(1-\delta_*)}$. A consistent and asymptotically normal estimator of α is given by:

$$\hat{\alpha}_0 = \frac{-\log\left(\hat{g}(\delta_*)\right)}{1-\delta_*}\,,$$

where $\hat{g}$ denotes the empirical PGF of the final number of mutants. Remark that the P0 method does not directly yield an estimator of ρ. If an estimate is desired, the ML method can be used for ρ only, setting $\alpha = \hat{\alpha}_0$.

ML estimators: Since algorithms [Zheng, 2005, Hamon and Ycart, 2012, Ycart and Veziris, 2014] enable to compute the probabilities of the LD, and H models, the ML method seems to be an obvious choice. It can be used on two kinds of samples:

1. sample of mutant counts: In that case, the likelihood is computed with the probabilities of the model LD or H. The parameter of interest is α.
2. sample of pairs of (mutant counts–final numbers): In that case, the likelihood is computed with the probabilities of the model $LDFN$ or HFN. The parameter of interest is π.

In both cases, ρ can also be estimated.

However, when the sample maximum is large, sums of products of small terms have to been computed [Hamon and Ycart, 2012]. The procedure can be very long and numerically unstable. Thus, the ML estimators can fail for large α and small ρ. In practice, this instability problem is

avoided using Winsorization [Wilcox, 2012, Sec. 2.2], which consists in replacing any value of the sample that pass a certain bound by the bound itself. The bound is 512 by default, and it could be necessary to increase it. All information above the bound is lost, and in an extreme case where the sample minimum is greater than the bound, irrelevant results will be returned.

GF estimators: The GF method uses the PGF to estimate the parameter of a compound Poisson distribution [Rémillard and Theodorescu, 2000, Hamon and Ycart, 2012]. Let $0 < z_1 < z_2 < 1$ and z_3 in $(0\,;\,1)$. The estimators of α and ρ are the following:

$$\hat{\alpha}_{GF}(z_3) = \frac{\log\left(\hat{g}(z_3)\right)}{\hat{h}(z_3) - 1} \quad \text{and} \quad \hat{\rho}_{GF}(z_1, z_2) = f^{-1}_{z_1,z_2}(\hat{y})\,,$$

where $\hat{g}$ denotes the empirical PGF of the final number of mutants, $\hat{h}$ is the PGF (2) with $\rho = \hat{\rho}_{GF}(z_1, z_2)$, and:

$$f_{z_1,z_2}(\rho) = \frac{h(z_1) - 1}{h(z_2) - 1} \quad \text{and} \quad \hat{y} = \frac{\log\left(\hat{g}(z_1)\right)}{\log\left(\hat{g}(z_2)\right)}\,.$$

From Rémillard and Theodorescu [2000], it can be proved that the couple of estimators $(\hat{\alpha}_{GF}, \hat{\rho}_{GF})$ is strongly consistent and asymptotically normal, with explicit asymptotic variance [Hamon and Ycart, 2012].

The GF estimators depend on the three arbitrary values of z_1, z_2, z_3. Those tuning parameters are set to $z_1 = 0.1$, $z_2 = 0.9$, and $z_3 = 0.8$. For more details about the choice of those values, see Hamon and Ycart [2012].

In practice, the GF estimators are quite comparable in precision to ML estimators, with a much broader range of calculability, a better numerical stability, and a negligible computing time, even in case where the ML method fails. For that reason, we have chosen to initialize the ML optimization by GF estimates, to improve both numerical stability and computing time.

The only practical limitation of this method is the following. A zero of the monotonous function $f_{z_1,z_2}(\rho) - \hat{y}$ must be computed. An upper bound for the domain of research must be given, which can be a problem if the sample does not contain jackpots. However in that case, a mutation model is not adapted.

The function `mutestim` computes estimates and their respective standard deviations for α, π and ρ according to the type of input. Moreover, the estimators mentioned here are asymptotically normal. Thus, one and two sample tests can be performed, using the function `flan.test`. The null hypothesis will be either fixed theoretical values of α, π, ρ in the one sample case, or a difference of the same in the two sample case.

3.2 Comparison of the three estimators

The Figure 1 (drawn using **ggplot2**) shows a "map of usage" of the estimation methods. They are compared in terms of the relative MSE of $(\hat{\alpha}, \hat{\rho})$ defined as:

$$\sqrt{\left(1 - \frac{\hat{\alpha}}{\alpha}\right)^2 + \left(1 - \frac{\hat{\rho}}{\rho}\right)^2}\,. \tag{4}$$

The RGB code is used: red for GF, green for P0, blue for ML. Twenty values of α between 0.5 and 10, and as many values of ρ from 0.2 to 5, were chosen. Thus 400 couples were considered. For each of them, the following procedure was applied:

1. draw 10^4 samples of size 100 of the $LD(\alpha, \rho, 0)$;
2. for each sample, compute ML, GF and P0 estimates of $(\alpha\,,\,\rho)$;
3. from the 10^4 estimates, compute the relative MSEs of each method;
4. assign a RGB color according to the MSEs. For each method:
   - if the MSE is less than 0.05, assign 1 to the corresponding RGB component;

- if the MSE is greater than 1, assign 0 to the corresponding RGB component;
- else, assign 1 minus the MSE to the corresponding RGB component.

The map has been drawn with a $\log_5$-scale for ρ (y-axis). The map can be roughly divided into four

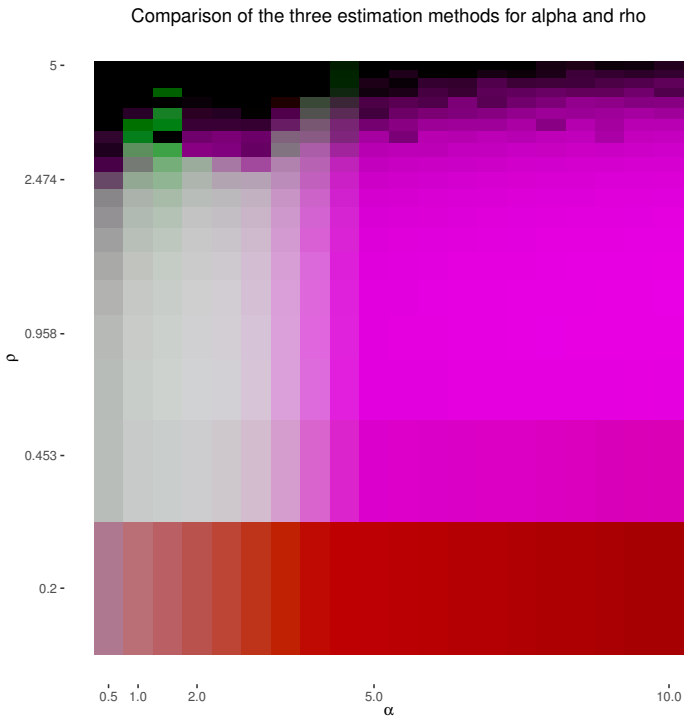

Figure 1: **Map of usage of the estimation methods.** The map compares the three methods according to their relative MSE (4). For each of 400 couples of parameters $\alpha = 0.5 \ldots 10$ (x-axis) and $\rho = 0.2 \ldots 5$ (y-axis, $\log_5$-scale), 10^4 samples of size 100 of the $LD(\alpha, \rho, 0)$ distribution were simulated. The estimates of $(\alpha\,,\,\rho)$ were calculated with the three methods GF (red), P0 (green) and ML (blue).

distinct parts:

- For $(\alpha\,,\,\rho) \in (0.5\,;\,3) \times (0.2\,;\,2.5)$, the color is essentially grey: the three methods are more or less equivalent.
- For $(\alpha\,,\,\rho) \in (3\,;\,10) \times (0.2\,;\,3.5)$, the color is magenta: the ML and GF methods are equivalent. The P0 method provides estimates with large MSEs or cannot be used because of the absence of null counts.
- For small values of ρ, the color is mainly red: The GF method is the only method with an acceptable MSE. Small values of ρ induce large jackpots. Moreover, the number of jackpots increases with α. Because of the winsorization, the ML and P0 method (which uses ML to estimate ρ) provide estimates with very large MSEs.
- For ρ large, the color is darker and tends to black: the three methods provide estimates with large MSEs, specially for $\rho \in (3.5\,;\,5)$, where jackpots are very small or absent. In those cases, estimating ρ with the GF method is not possible in practice (see previous sub-section). Consequently, the GF method will provide a biased estimate for α. The ML method, which uses the GF estimates to initialize the optimization of the log-likelihood, also provides biased estimates. The P0 method can provide good estimates of α whatever the value of ρ, which explains the presence of green areas at the top of the map. In a case where no jackpots are

present in the sample it should be considered that a (heavy tailed) mutation model is not adapted.

The three methods should also be compared in terms of computational time. An illustration on real data will be given in section 5. The slowest method is ML, for the reasons discussed in the previous section. It is even slower when the estimates are calculated under Haldane models H or HFN, when δ is positive, or if the initialization of ρ with the method GF fails. The method GF computes estimates of α and ρ (when possible) in negligible time. The method P0 outputs estimates of α in negligible time, but estimates of ρ are as slow as with ML.

3.3 Bias evidence

If the model used for the estimation does not correspond to the theoretical model, the estimates can be biased. Three different sources of bias are considered:

1. the final counts are random in the data, constant for the estimation model;
2. cell deaths occur in the data, not in the estimation model;
3. the lifetime distribution is different in the data and the estimation model.

In each case, simulation experiments have been made along the following lines:

1. draw 10000 samples of size 100, under one model;
2. for each sample, compute ML estimates of α and ρ, using another model;
3. observe the empirical distribution of $\hat{\theta}/\theta$, where $\hat{\theta}$ is the estimator and θ the true value.

Fluctuation of final counts: When N is constant, the estimate of π is derived by dividing the estimate of α by N. As mentioned in previous section, if N is a random variable, the relation between α and π can be explicit if the distribution K is known. However, this is not the case in practice: estimates of the expectation and variance of N are usually available at best. Assume that only the first two moments μ and σ^2 of N are known. Then a first order approximation of the Laplace transform $\mathcal{L}$ can be used to reduce the bias. This method is explained in Ycart and Veziris [2014] for the P0 method. It has been adapted to ML and GF estimates. Figure 2 shows the influence of the coefficient of variation $C = \sigma/\mu$ on the ML estimate of π. The estimates were calculated with three different approaches:

- divide ML estimates of α by the empirical mean of N and ignore fluctuations of N (left boxplots);
- directly compute ML with the sample of pairs (mutant counts–final counts) (center boxplots);
- derive from ML estimates of α, taking into account of the empirical fluctuations of N (right boxplots).

According to the visual observations, the efficiency of bias reduction decreases as α and C increase. It could be improved with a better appproximation of $\mathcal{L}$, that implies knowing or estimating higher moments of N. Another solution is to improve the estimation of C. Here C was estimated by the ratio of the empirical standard deviation of the empirical mean, which is known to be a bad method in terms of MSE [Breunig, 2001].

Cell deaths: The PGFs (1), (3) and (2) depend on δ. Ignoring cell deaths involves a bias on the estimate of α. Assuming the exact value is known, this bias is removed. Figure 3 shows the influence of the death parameter δ on the ML estimate of α. The estimates are calculated with two different approaches:

1. computing ML estimates of α with $\delta = 0$ (left boxplots);
2. computing ML estimates of α with theoretical value of δ (right boxplots).

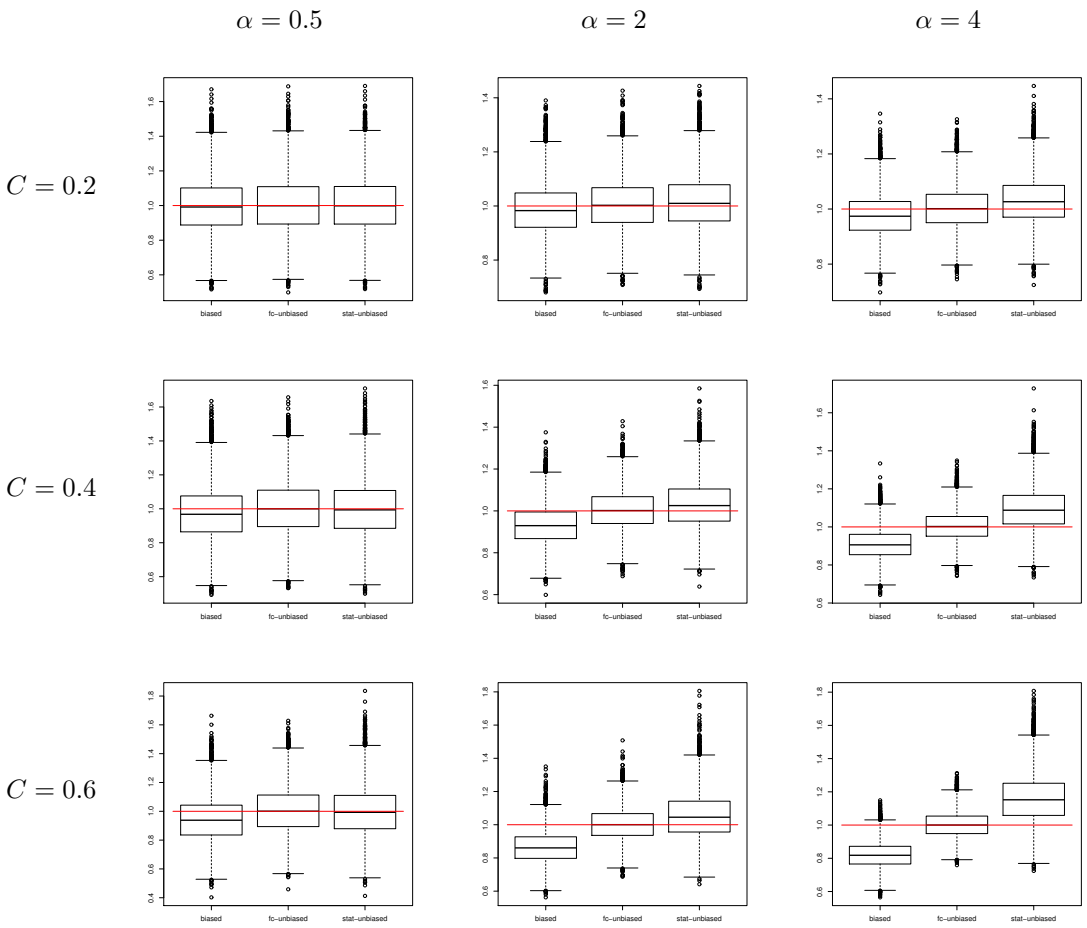

Figure 2: **Boxplots of ML estimates of π ignoring fluctuations of final numbers or not.** For each of the 9 sets of parameters $\pi = (0.5/\mu\,,\, 2/\mu\,,\, 4/\mu)$ (columns), and $C = (0.2\,,\, 0.4\,,\, 0.6)$ (rows), 10000 samples of size 100 of the $LDFN(\pi, \rho, 0, K)$ distribution were simulated, K being the Log-normal distribution adjusted to mean $\mu = 10^9$ and coefficient of variation C. The estimates of π were calculated with the model LD. Each boxplot represents the distribution of the 10000 ratio $\hat{\pi}/\pi$ obtained with $C = 0$ (left), $LDFN$ model (center), LD model with bias reduction (right).

The visual results show that the negative bias induced by ignoring cells death increases with the value of δ. From the theory of branching processes, the growth process of a mutant clone is supercritical and δ has to be smaller than 0.5. In practice δ is smaller than 0.3. According to the boxplots, the relative bias induced by ignoring cell deaths can reach 0.80. These experiments illustrate also the difficulty to estimate δ. For example, the boxplots at the top right of the figure seems to show that the value of the likelihood for $\alpha = 4$ and $\delta = 0$ is very close to its value for $\alpha = 4$ and $\delta = 0.05$.

Lifetime distribution: As mentioned earlier, the PGF h is explicit only for the LD and H distributions, i.e. when lifetimes are either exponential or constant. This is not the case in practice. If another lifetime distribution is used to simulate the data, and either LD or H are used to estimate the parameters, a bias will be induced on α and ρ. Figure 4 illustrates these observations. It shows

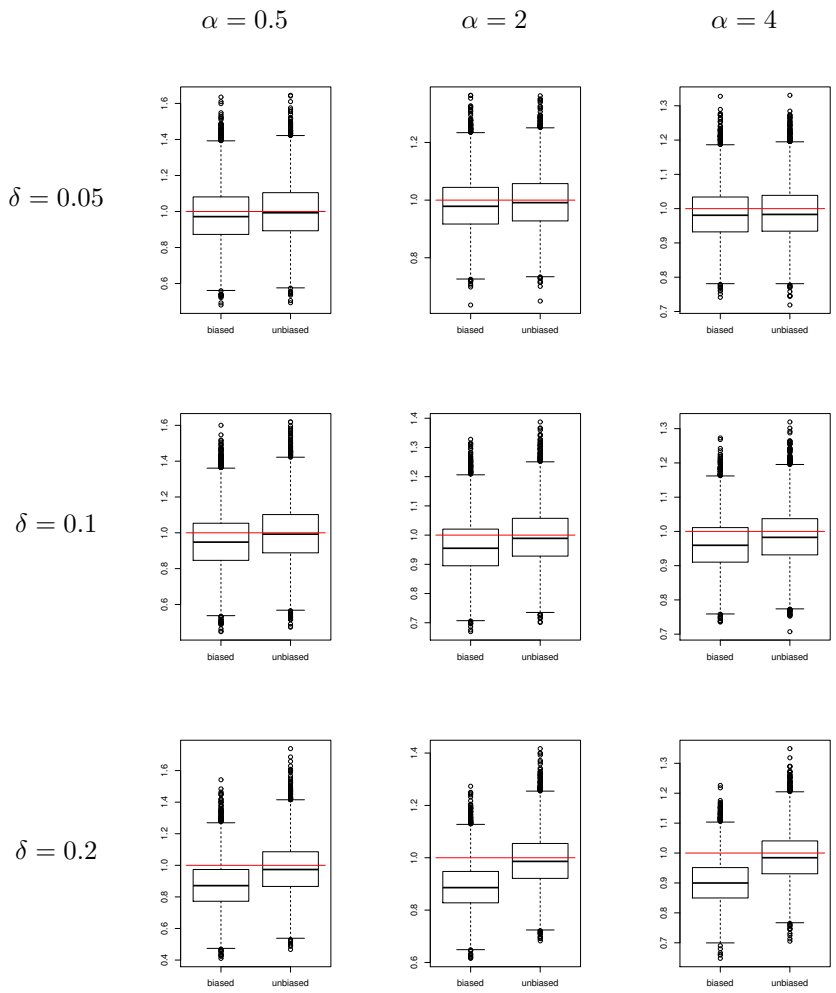

Figure 3: **Boxplots of α ML estimates taking account or not of cell deaths.** For each of the 9 sets of parameters $\alpha = (0.5\,,\, 2\,,\, 4)$ (columns), and $\delta = (0.05\,,\, 0.1\,,\, 0.2)$ (rows), 10000 samples of size 100 of the $LD(\alpha, 1, \delta)$ distribution were simulated. The estimates of α were calculated under LD model. Each boxplot represents the distribution of the 10000 ratio $\hat{\alpha}/\alpha$ of the estimates obtained without taking account of cells death (left) and with the theoretical value of δ (right).

the influence of the lifetime distribution on the ML estimates of α and ρ. The samples are drawn assuming the lifetimes are log-normally distributed. The estimates of α and ρ are calculated under LD (left boxplots) and H models (right boxplots).

From the visual observations, the LD and H models can be seen as extreme values for the lifetime distribution:

- the LD model correctly estimates α;
- the H model overestimates α. The bias seems to decrease as α increases;
- the LD model overestimates ρ and has a rather large dispersion of estimated values. The bias seems to increase as α increases;
- the H model correctly estimates ρ.

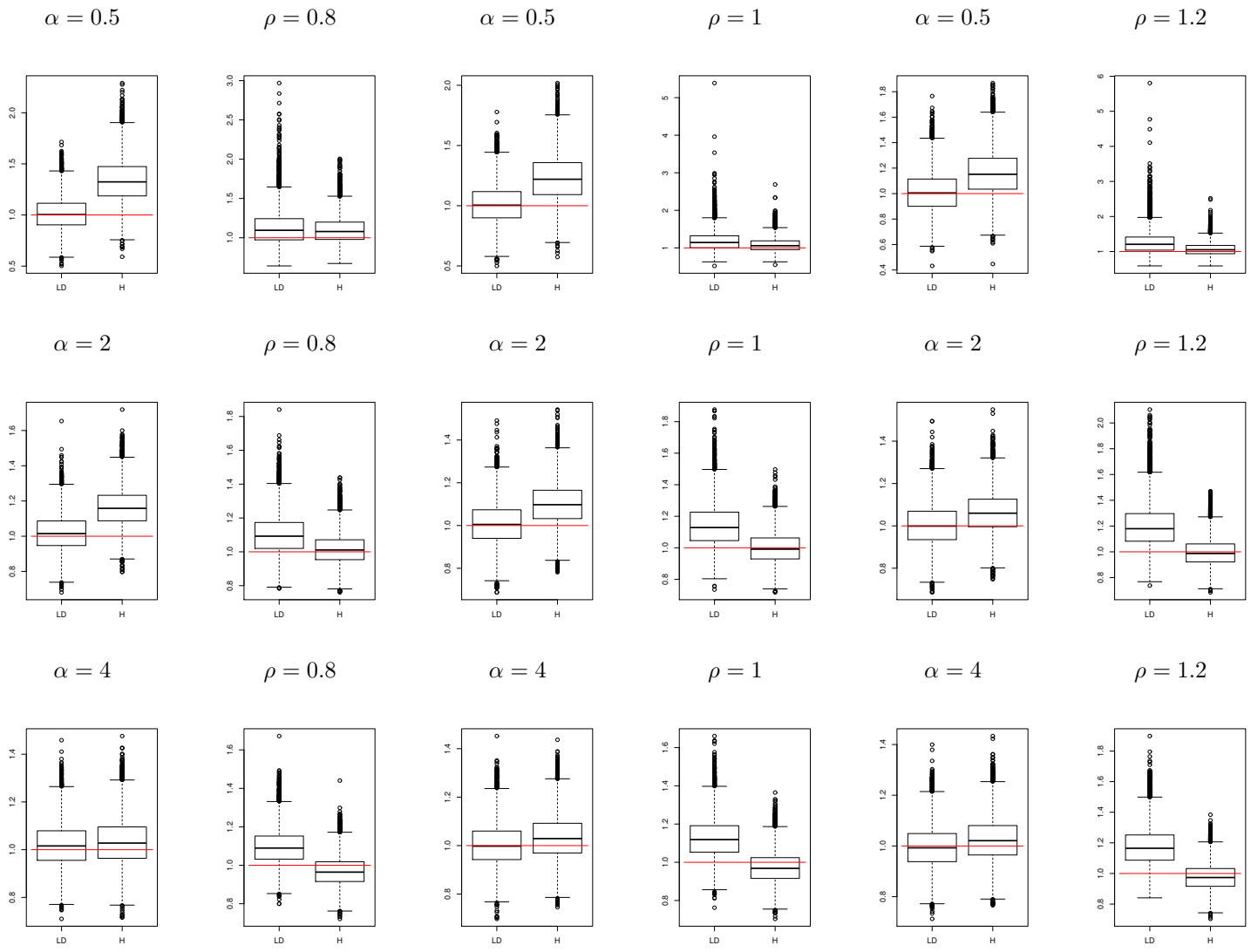

Figure 4: **Boxplots of α and ρ ML estimates under LD and H models.** Red horizontal lines mark unit. For each of the 9 sets of parameters $\alpha = (0.5\,,\ 2\,,\ 4)$ (rows), and $\rho = (0.8\,,\ 1\,,\ 1.2)$ (columns), 10000 samples of size 100 of the $MM(\alpha, \rho, 0, G)$ distribution were simulated, G being the Log-normal distribution adjusted on Kelly and Rahn's data [Kelly and Rahn, 1932]. The estimates of α and ρ were calculated with the two distributions $LD(\alpha, \rho, 0)$ and $H(\alpha, \rho, 0)$. Each boxplot represents the distribution of the 10000 ratio $\hat{\alpha}/\alpha$ and $\hat{\rho}/\rho$ obtained by the LD model (left) and the H model (right).

4 Implementation details

The available functions are described here; more details are given in the manual. The behavior of inference functions for inputs which are out of practical limitations is described. Some details about the **Rcpp** implementation are also provided.

4.1 User interface

flan can be split into two distinct parts: the distribution of the final number of mutants, and statistical inference. The functions `dflan`, `pflan`, `qflan` compute densities, probabilities and quantiles of LD and H distributions. The function `rflan` outputs samples of pairs (mutant counts–final counts) following $LDFN$, HFN, or $MMFN$ where G is log-normal or gamma distribution and K is log-normal or Dirac distribution. K is adjusted to the mean and coefficient of variation provided by the user. Those functions have been designed on the principle of the classical distribution functions of R. A graphic function `draw.clone` is also provided. It represents with a binary tree the growth of a clone starting from a single normal cell with mutation occurrences until a finite time. The function `mutestim` computes estimates of α or π and ρ, using LD or H models. The three estimation methods are available. Fluctuations of final numbers and cells death are included. It returns estimate(s) of the parameter(s) of interest and the standard deviations. The function `flan.test` uses asymptotic normality to perform one or two sample hypothesis testing. It has been designed on the principle of the classical hypothesis testing functions of R, such as `t.test`.

As mentioned in section 3, there are practical limitations for each estimation method. If the inputs of the `mutestim` function do not respect those limitations, it will output errors or warning messages:

- If $\delta = 0$, the P0 method can not be used if the sample does not contain any null counts. In that case, the `mutestim` function will return an error message.
- Issues of the Winzorization parameter w of ML method:
  1. If the minimum of the sample is larger than w, then the sample of mutant counts will be constant.
  2. If w is too large, then the optimization process can be very long.
  3. In the `mutestim` function w is set at 512 by default.
- The GF method does not have limitations of usage, even for extreme cases where the ML estimators fail, i.e. samples with theoretical large α and small ρ. However, estimating ρ requires to solve the zero equation discussed in section 3, which is theoretically solvable on $\mathbb{R}^+$. In practice the interval of research is bounded. Thus, if the sample does not contain any jackpots, which means ρ is very large, the zero equation may not have solution on the interval. In that case, the function will return a warning message, and set the estimate of ρ at 1, and the estimate of the standard deviation at 0. In the `mutestim` function, the domain of research is $[0.01\,;\,100]$.
- Moreover, the initialization of the ML method is done with GF method. Then the domain of optimization is $[0.1 \times \hat{\theta}_{GF}\,;\,10 \times \hat{\theta}_{GF}]$, where $\hat{\theta}_{GF}$ is the GF estimate(s) of the parameter(s) of interest. Then, if the GF method does not success to estimate ρ, there is no chance to estimate it with ML. A warning message is returned if the initialization of the estimate of ρ with GF fails.

The function `flan.test` is a wrapper function of `mutestim`. It will ouput the same errors or warning messages if its inputs do not respect the practical limitations.

4.2 Implementation

Since most functions involve loops that are more expensive in R than in C, **flan** has been implemented with the package **Rcpp** [D.Eddelbuettel, 2013]. **Rcpp** modules are used. This paradigm provides an easy way to expose C++ functions and classes to R. There are four main classes in the C++ implementation:

- `FLAN_Sim`: random generation for MM and $MMFN$ distributions. One of its members is a variable of following type;
- `FLAN_SimClone`: random generation for clone size distribution according to the lifetimes distribution;
- `FLAN_MutationModel`: computation of the descriptive functions (probabilities, PGF,...) for LD and H distributions. One of its members is a variable of following type;

- `FLAN_Clone`: computation of the descriptive functions for clone size distribution according to the lifetimes distribution.

The **Rcpp** interface enables also to import into the C++ code any R function. In particular, it is interesting to import the R functions which are already implemented in C. Thus no external C/C++ library is required. The installation remains basic, and the size of the installed package is reduced. For example, the computations of LD distributions involve numeric integrations. The C libraries **integration** and **alglib** compute integrals with an accuracy close to machine precision. We could use those libraries but the R function `integrate` is actually implemented in C. Then instead of importing the C libraries, the R function is directly called into the C++ code. Computing the probabilities for the H distribution with $\delta > 0$ involves squaring high degree polynomials. Such polynomials are easily treated by the package **polynomial**. However its implementation raises memory issues, because of the degree of the polynomials involved. A more efficient way is to use the Fast Fourier Transform. It is provided by the C library **fftw3**, which can raise some installation issues. Instead of importing it, the R function `fft` is directly called into the C++ code.

Finally, likelihood optimizations in the ML and P0 methods are done with a bounded BFGS optimizer. The package **lbfgsb3** provides the eponymous function which is implemented in Fortran. It is much faster than the basic R function `optim`.

5 Examples of usage

Some examples on the real data included in **flan** are provided. Practical limitations, influence of bias sources and comparison of the estimation methods in terms of computational time are illustrated.

Consider first the eleventh sample of mutant counts of the `werhoff` data [Werngren and Hoffner, 2003]:

```
werhoff$samples$W11$mc
[1] 0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 4 4 4 4 5 5
```

Denote it by W_{11}. This sample does not contain any jackpot, then the theoretical fitness in a mutation model, should be very large. If the GF method is used, it will output a warning message and set the fitness at $\rho = 1$, as customarily done in the litterature [Foster, 2006]. It is possible to find a numeric value for ρ, by enlarging the domain of research. The solution to the equation discussed in section 3 is $\rho' = 833.4272$. This value is quite unrealistic, indicating that the LD model is not adapted. However the estimate of α cannot be the same for $\rho = 1$ and $\rho = \rho'$. Using `mutestim` function setting the value of the input `fitness` enables to observe this difference. The GF estimate of α is $\hat{\alpha}_{GF} = 0.8792917$ if $\rho = 1$, and $\hat{\alpha}'_{GF} = 2.207935$ if $\rho = \rho'$. Using the P0 method is another way to realize that setting $\rho = 1$ by default can be misleading. Since this method does not depend on the lifetime distribution, the estimate of α will not depend on the value of ρ. The P0 estimate of α for W_{11} is $\hat{\alpha}_0 = 2.525729$, which is closer to $\hat{\alpha}'_{GF}$ than $\hat{\alpha}_{GF}$.

Consider now the only sample which includes final counts, the eleventh of the `david` data [David, 1970]:

```
david$D11
$mc
[1] 4 0 1 0 1 0 0 0 0 0
$fn
[1] 1.3e+09 9.2e+08 1.3e+09 2.5e+09 1.3e+09
[6] 1.6e+09 1.3e+09 2.5e+09 2.5e+09  2.0e+09
```

Remark firstly that the 4 value can be seen as a jackpot, and the GF method can be used to estimate α and ρ. Now let us compute the ML estimates of π and ρ taking into account or not of the final counts, under the LD model. The sample of final counts is denoted by $D_{11}^{(FN)}$. The empirical mean of the final counts is denoted by $\overline{\mu}$, the empirical coefficient of variation by $\overline{C}$. Table 1 displays these estimates, in the same way as for Figure 2. Comparing the first row to the third, one can see that

neglecting final number fluctuations induces a bias of order 5% on π, 10% on ρ. From the second row, it turns out that the correction taking into account $\overline{C}$, has not improved the estimate of π.

Using:	$\hat{\pi}_{ML}$	$\hat{\rho}_{ML}$
$\mu = \overline{\mu}$, $C = 0$	$2.067641 \cdot 10^{-10}$	2.214676
$\mu = \overline{\mu}$, $C = \overline{C}$	$2.094837 \cdot 10^{-10}$	2.214676
$D_{11}^{(FN)}$	$1.977135 \cdot 10^{-10}$	2.048984

Table 1: **Influence of fluctuations of final numbers on real data.** Each row shows the estimates of π and ρ, deducing from LD model with $C = 0$ (first row), $C = \overline{C}$ (second row), and directly with $LDFN$ model (third row).

Consider finally the data `boeal` [Boe et al., 1994]. Here the 23 samples are concatenated as one (`unlist(boeal)`), and used to compare the three estimation methods in terms of computational time. The package **microbenchmark** is used, evaluating 10^4 times each method on the sample. The estimates of α and ρ are computed under the model LD with $\delta = 0$. The results are shown on Figure 5, as boxplots of timing distributions. Times are in milliseconds and plotted on log-scale.

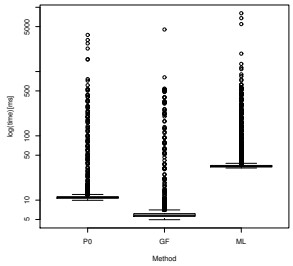

Figure 5: **Computational time of the three methods on real data.** Data consist in the 23 samples of `boeal` concatenated as one. For each method, the estimates of α and ρ have been computed under model LD. The timings have been returned with **microbenchmark**, evaluating 10^4 times each method. Times are in milliseconds and plotted on log-scale.

As mentioned earlier, the method ML is the slowest. The methods GF and P0 seem to be equivalent in terms of computational time. However, the estimates of ρ of the P0 method is calculated using the ML method. If only α has to be estimated, the P0 method is faster than the GF method.

References

W. Angerer. An explicit representation of the Luria-Delbrück distribution. *J. Math. Biol.*, 42(2): 145–174, 2001.

P. Armitage. The statistical theory of bacterial populations subject to mutation. *J. R. Statist. Soc. B*, 14:1–40, 1952.

K. Athreya and P. Ney. *Branching processes*. Springer Berlin Heidelberg, 1972.

M. S. Bartlett. *An introduction to stochastic processes, with special reference to methods and applications*. Cambridge University Press, 3rd edition, 1978.

R. Bellman and T. Harris. On age-dependent binary branching processes. *Ann. Math.*, 55(2):280–295, 1952.

L. Boe, T. Tolker-Nielsen, K. M. Eegholm, H. Spliid, and A. Vrang. Fluctuation analysis of mutations to nalidixic acid resistance in *Escherichia Coli*. *J. Bacteriol.*, 176(10):2781–2787, 1994.

R. Breunig. An almost unbiased estimator of the coefficient of variation. *Econ. Lett.*, 70(1):15–19, 2001.

H. L. David. Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis*. *Appl. Microbiol.*, 20(5):810–814, 1970.

D.Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer-Verlag New-York, 2013.

A. Dewanji, E. Luebeck, and S. Moolgavkar. A generalized Luria-Delbrück model. *Math. Biosci.*, 197(2):140–152, 2005.

P. Foster. Methods for determining spontaneous mutation rates. *Method. Enzymol.*, 409:195–213, 2006.

A. Hamon and B. Ycart. Statistics for the Luria-Delbrück distribution. *Elect. J. Statist.*, 6:1251–1272, 2012.

C. D. Kelly and O. Rahn. The growth rate of individual bacterial cells. *J. Bacteriol.*, 23(2):147–153, 1932.

N. L. Komarova, L. Wu, and P. Baldi. The fixed-size Luria-Delbrück model with a nonzero death rate. *Math. Biosci.*, 210(1):253–290, 2007.

D. Lea and C. Coulson. The distribution of the number of mutants in bacterial populations. *Journal of Genetics*, 49(3):264–285, 1949.

S. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

W. T. Ma, G. v. H. Sandri, and S. Sarkar. Analysis of the Luria-Delbrück distribution using discrete convolution powers. *J. Appl. Probab.*, 29(2):255–267, 1992.

B. Rémillard and R. Theodorescu. Inference based on the empirical probability generating function for mixtures of poisson distributions. *Statist. Decisions*, 18:349–366, 2000.

S. Sarkar. Haldane's solution of the Luria-Delbrück distribution. *Genetics*, 127:257–261, 1991.

J. Werngren and S. E. Hoffner. Drug susceptible *Mycobacterium tuberculosis* Beijing genotype does not develop motation-conferred resistance to Rifampin at an elevated rate. *J. Clin. Microbiol.*, 41 (4):1520–1524, 2003.

R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Elsevier, Amsterdam, 3$^{\text{rd}}$ edition, 2012.

B. Ycart. Fluctuation analysis: can estimates be trusted? *PLoS One*, 8(12), 2013.

B. Ycart. Fluctuation analysis with cell deaths. *J. Appl. Probab. Statist*, 9(1):13–29, 2014.

B. Ycart and N. Veziris. Unbiased estimates of mutation rates under fluctuating final counts. *PLoS One*, 9(7), 2014.

Q. Zheng. New algorithms for Luria-Delbrück fluctuation analysis. *Math. Biosci.*, 196(2):198–214, 2005.