საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 9, №1, 2015
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 9, no. 1, 2015

Bioinformatics

On the Estimators for some Frequency Distributions Arising in Bioinformatics

Davood Farbod

Department of Mathematics, Quchan University of Advanced Technology, Quchan, Iran

(Presented by Academy Member Mindia Salukvadze)

ABSTRACT. We consider two well-known frequency distributions which can be used for modeling phenomena arising in bioinformatics. These are: the two-parameter Waring frequency distribution and the two-parameter Pareto-like frequency distribution, both introduced by V. A. Kuznetsov. We propose the system of finding the *maximum likelihood estimator*s (MLE) for the unknown parameters of such frequency distributions. The proposed MLE are coincided with some moment estimators. Moreover, a method of approximate computation of the MLE for the models parameters is obtained. Simulation studies are given to support the theoretical results. *© 2015 Bull. Georg. Natl. Acad. Sci.*

Key words: *waring distribution, Pareto-like distribution, Markov Chain Monte Carlo (MCMC), MLE.*

1. Introduction and Preliminaries

Several parametric frequency distributions have been proposed for the needs of biomolecular applications. Among them, there are two well-known classes of frequency distributions introduced by V. A. Kuznetsov [1, 2]: Waring and Pareto-like frequency distributions. These distributions may be described for modeling phenomena in biomolecular systems such as for the number of expressed genes in the transcriptome, the number of protein domain occurrences in the proteomes. For details about the applications and properties of such distributions see [1-4].

The Waring frequency distribution is a two-parameter distribution with the following probability function (see [1, 3]):

$$\begin{cases} f_x(\theta) = \left(f_0(\theta)\right)^{-1} \cdot \prod_{k=0}^{x-1} \dfrac{p+k}{q+k}, & x = 1, 2, \ldots, \\ f_0(\theta) = \sum_{y=1}^{\infty} \prod_{k=0}^{y-1} \dfrac{p+k}{q+k} \end{cases} \tag{1}$$

where $\theta = (p, q)$ and $0 < p < q < \infty$. Note that p and q are the model's parameters.

Another important family is the two-parameter Pareto-like frequency distribution with the following prob-

ability function (see [2], [3]):

$$\begin{cases} f_x(\alpha) = \left(f_0(\alpha)\right)^{-1} \cdot (x+b)^{-\rho}, & x = 1, 2, \ldots, \\ f_0(\alpha) = \sum_{y=1}^{\infty} (y+b)^{-\rho} \end{cases} \tag{2}$$

where $\alpha = (\rho, b)$, $1 < \rho < \infty$ is the shape parameter and $-1 < b < \infty$ is the location parameter.

It is of interest to investigate the statistical properties of the parameters for the models (1) and (2). But, the lack of simple closed formulas for the probability mass and cumulative distribution functions were major drawbacks to the use of such frequency distributions. In this paper we are going to study some statistical inferences for the models (1) and (2).

Assuming $X^n = (X_1, \ldots, X_n)$, with observation $x^n = (x_1, \ldots, x_n)$, is sample corresponding to a random variable ξ with the probability distribution given by (1) or (2). For $x \in \mathbb{N}$, we fix the following notations to be used in this paper:

$$h(x;\theta) = \sum_{k=0}^{x-1} \frac{1}{p+k}, \qquad l(x;\theta) = \sum_{k=0}^{x-1} \frac{-1}{q+k}, \qquad k(x;\theta) = \sum_{k=0}^{x-1} \frac{-1}{(p+k)^2},$$

$$t(x;\theta) = \sum_{k=0}^{x-1} \frac{1}{(q+k)^2}, \qquad m(x;\alpha) = -\ln(x+b), \qquad w(x;\alpha) = \frac{-\rho}{(x+b)},$$

$$\overline{f^n(\theta)} = \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta), \qquad \overline{f^n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} f(x_i;\alpha)$$

Also, the symbols $E[\cdot]$, $Var(\cdot)$ and $Cov(\cdot,\cdot)$ refer to the mathematical expectation, variance and covariance, respectively.

2. MLE

Supposing, as above, ξ is a discrete random variable with the probability distribution given by formula (1) or (2). We first establish a lemma, which will be used in the proof of the main theorem. Here, we consider the model (1).

Lemma 1. *For the model (1), we have*

$$E_\theta\left[h(\xi;\theta)\right] < \infty, \qquad E_\theta\left[l(\xi;\theta)\right] < \infty.$$

Proof. By using definition of mathematical expectation the proof is met obviously.

Now, from Lemma 1 the following theorem is presented.

Theorem 1. *The MLE of the parameter θ is obtained from the following moment equations:*

$$\begin{cases} E_\theta\left[h(\xi;\theta)\right] = \overline{h^n(\theta)}, \\ E_\theta\left[l(\xi;\theta)\right] = \overline{l^n(\theta)}. \end{cases} \tag{3}$$

Proof. We derive the logarithm of likelihood function as follows

$$l(X^n;\theta) = \ln \prod_{i=1}^{n} \left(f_0(\theta) \cdot \prod_{k=0}^{x_i-1} \frac{p+k}{q+k} \right) = -n \ln f_0(\theta) + \sum_{i=1}^{n}\sum_{k=0}^{x_i-1} \ln\left(\frac{p+k}{q+k}\right). \tag{4}$$

In order to existence of the MLE for the parameter θ, the necessary conditions are expressed as

$$\frac{\partial l\left(X^n;\theta\right)}{\partial \theta_i}=0,\quad i=1,2,$$

where $\theta_1 = p$, $\theta_2 = q$.

Let us establish the derivatives with respect to the parameters p and q. We obtain:

$$\frac{\partial l\left(X^n;\theta\right)}{\partial p}=-n\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial p}+\sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{1}{p+k},$$

where $\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial p}=E_\theta\left[h(\xi;\theta)\right]$. From $\frac{\partial l\left(X^n;\theta\right)}{\partial p}=0$ it follows that

$$E_\theta\left[h\left(\xi;\theta\right)\right]=\overline{h^n(\theta)}.$$

Meanwhile, we get

$$\frac{\partial l\left(X^n;\theta\right)}{\partial q}=-n\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial q}+\sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{-1}{q+k}$$

Taking into account that $\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial q}=E_\theta\left[l(\xi;\theta)\right]$, then from the condition $\frac{\partial l\left(X^n;\theta\right)}{\partial q}=0$ we conclude

$$E_\theta\left[l\left(\xi;\theta\right)\right]=\overline{l^n(\theta)}.$$

The proof of Theorem 1 is complete.

Now, we are going to show that the solution $\hat{\theta}=\hat{\theta}_i^n=\left(\hat{\theta}_i^n\right)_{i=1}^2$ of the system (3) (if it exists) is the MLE of the parameter θ. It suffices to establish that the matrix

$\widehat{M}_n=\left(\widehat{M}_{i,j}^n\right)_{i,j=1}^2$ with $\widehat{M}_{i,j}^n=\widehat{M}_{i,j}^n\left(\hat{\theta}\right)$, $\widehat{M}_{i,j}^n\left(\hat{\theta}\right)=\left.\frac{\partial l(X^n,\theta)}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}}$ is *negative definite*. We record the following assertion.

Lemma 2. *Assuming the solution $\hat{\theta}$ of the system (3) (if it exists) holds in the following conditions*

$$\begin{cases}E_\theta\left[k\left(\xi;\theta\right)\right]=\overline{k^n(\theta)}\\ E_\theta\left[t\left(\xi;\theta\right)\right]=\overline{t^n(\theta)}\end{cases}\tag{5}$$

then the elements of the matrix $\widehat{M}_n$ are

$$\widehat{M}_{11}^n=-nVar_{\hat{\theta}}\left(h(\xi;\theta)\right),$$
$$\widehat{M}_{12}^n=\widehat{M}_{21}^n=-nCov_{\hat{\theta}}\left(h\left(\xi;\theta\right),l(\xi;\theta)\right),$$
$$\widehat{M}_{22}^n=-nVar_{\hat{\theta}}\left(l(\xi;\theta)\right).$$

Proof. It is easy to see that

$$M_{11}^n=\frac{\partial^2 l\left(X^n;\theta\right)}{\partial p^2}=-n\left(\frac{1}{f_0(\theta)}\cdot\frac{\partial^2 f_0(\theta)}{\partial p^2}-\left(\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial p}\right)^2\right)+n\overline{k^n(\theta)},$$

$$M_{12}^n = M_{21}^n = \frac{\partial^2 l\left(X^n;\theta\right)}{\partial p\,\partial q} = \frac{\partial^2 l\left(X^n;\theta\right)}{\partial q\,\partial p} = -n\left(\frac{1}{f_0(\theta)}\cdot\frac{\partial^2 f_0(\theta)}{\partial p\,\partial q} - \frac{1}{\left(f_0(\theta)\right)^2}\cdot\frac{\partial f_0(\theta)}{\partial p}\cdot\frac{\partial f_0(\theta)}{\partial q}\right),$$

$$M_{22}^n = \frac{\partial^2 l\left(X^n;\theta\right)}{\partial q^2} = -n\left(\frac{1}{f_0(\theta)}\cdot\frac{\partial^2 f_0(\theta)}{\partial q^2} - \left(\frac{1}{f_0(\theta)}\cdot\frac{\partial f_0(\theta)}{\partial q}\right)^2\right) + n\overline{t^n(\theta)}$$

After some calculations and simplifications we obtain

$$M_{11}^n = -nVar_\theta\left(h(\xi;\theta)\right) - \left(n\,E_\theta\left[k(\xi;\theta)\right] - n\overline{k^n(\theta)}\right)$$

$$M_{12}^n = M_{21}^n = -n\,Cov_\theta\left(h\left(\xi;\theta\right), l(\xi;\theta)\right),$$

$$M_{22}^n = -nVar_\theta\left(l(\xi;\theta)\right) - \left(n\,E_\theta\left[t\left(\xi;\theta\right)\right] - n\overline{t^n\left(\theta\right)}\right).$$

From the condition (5) the proof is completed.

Lemma 3. *Under satisfying the conditions (5), the matrix* $\widehat{M}_n$ *is negative definite.*

Proof. In order to due that it is enough to demonstrate $\widehat{M}_{11}^n < 0$ and $\det(\widehat{M}_n) > 0$.

From Lemma 2 it is obvious $\widehat{M}_{11}^n < 0$. To prove that $\det(\widehat{M}_n) > 0$ we have

$$\det(\widehat{M}_n) = \widehat{M}_{11}^n\,\widehat{M}_{22}^n - \left(\widehat{M}_{12}^n\right)^2.$$

According to the value of $\widehat{M}_{11}^n$, $\widehat{M}_{12}^n$, $\widehat{M}_{22}^n$ and based on Cauchy-Bunyakovski-Schwartz inequality (see, for example, [5]) the proof is finished.

As an immediate consequence of Lemmas 2 and 3 the following result is received.

Corollary 1. *If the solution of the system (3) satisfies the conditions (5), then it coincides with the MLE of the parameter* θ.

Remark 1. We notice that the above mentioned results can also be proved for the model (2).

Thus, compare to Theorem 1, let us propose the following assertion for the model (2).

Theorem 2. *The MLE of the parameter* α *is given from the following moment equations:*

$$\begin{cases} E_\alpha\left[m\left(\xi;\alpha\right)\right] = \overline{m^n\left(\alpha\right)}, \\ E_\alpha\left[w\left(\xi;\alpha\right)\right] = \overline{w^n\left(\alpha\right)}. \end{cases} \tag{6}$$

Proof. Similar to the proof of Theorem 1.

Corollary 2. *Similarly, the Lemmas 2 and 3 may be considered for the two-parameter Pareto-like frequency distribution given by formula (2). Because of similarity, we leave the results and proofs for the readers.*

3. Approximate Computation of the MLE

Based on the systems (3) and (6), it is not simple to derive closed forms for the solutions. So, we need to propose a method of the approximate computation of the MLE for the models parameters. In order to do that (compare to [6], p. 228), the Fisher's accumulation method is suggested. For more details about the Fisher's accumulation method we refer the readers to, for example, [7], p. 88.

3.1 Waring Distribution

Supposing $\theta_0 = (p_0, q_0)$ is an initial value for the parameter $\theta = (p, q)$. Recurrently, we find out the $(s+1)_{th}$ approximation by using the formula (compare to [6], p. 228):

$$\theta_j\left(s+1\right)=\theta_j\left(s\right)+\frac{\Lambda_{\mathrm{j}}(\theta(\mathrm{s}))}{n\cdot\det I(\theta(s))}, \qquad j=1,2, \tag{7}$$

where $I(\cdot)$ is the Fisher's information measure. In details, the formula (7) is considered as

$$\begin{cases} p\left(s+1\right)=p\left(s\right)+\dfrac{\Lambda_{\mathrm{p}}(\theta(\mathrm{s}))}{n\cdot\det I(\theta(s))}, \\ q\left(s+1\right)=q\left(s\right)+\dfrac{\Lambda_{\mathrm{q}}(\theta(\mathrm{s}))}{n\cdot\det I(\theta(s))}, \end{cases} \tag{8}$$

where

$$\Lambda_{\mathrm{p}}\left(\theta(\mathrm{s})\right)=\begin{pmatrix}\mathrm{U}_1(\theta) & I_{12}(\theta)\\ \mathrm{U}_2(\theta) & I_{22}(\theta)\end{pmatrix}, \qquad \Lambda_{\mathrm{q}}\left(\theta(\mathrm{s})\right)=\begin{pmatrix}I_{11}(\theta) & \mathrm{U}_1(\theta)\\ I_{21}(\theta) & \mathrm{U}_2(\theta)\end{pmatrix},$$

and $\mathrm{U}_1\left(\theta\right)=\dfrac{\partial l\left(X^n;\theta\right)}{\partial p}$ and $\mathrm{U}_2\left(\theta\right)=\dfrac{\partial l\left(X^n;\theta\right)}{\partial q}$ are the *contribution functions*. We obtain

$$\mathrm{U}_1\left(\theta\right)=-n\,E_\theta\left[h(\xi;\theta)\right]+n\,\overline{h^n(\theta)},$$

$$\mathrm{U}_2\left(\theta\right)=-n\,E_\theta\left[l(\xi;\theta)\right]+n\,\overline{l^n(\theta)}.$$

Comparing to Farbod and Gasparian ([6], p. 229), let us give the following algorithm for the model (1).

Algorithm I.

1. Generate random numbers from the model (1) using MCMC method;
2. Use (7) (or (8)) in order to calculate $\theta_j\left(s\right)$, $j=1,2$, $s=0,1,2,\ldots$;
3. If $\left|\theta_j\left(s+1\right)-\theta_j\left(s\right)\right|<\varepsilon$ (ε is some small positive constant), then $\theta_j\left(s+1\right)=\hat{\theta}$ is the MLE, otherwise go to the step 2.

3.2 Pareto-Like Distribution

In a similar way, for the model (2) we establish

$$\alpha_j\left(\mathrm{s}+1\right)=\alpha_j\left(s\right)+\frac{\Lambda_{\mathrm{j}}(\alpha(\mathrm{s}))}{n\cdot\det I(\alpha(s))}, \qquad j=1,2. \tag{9}$$

In details, we get

$$\begin{cases} \rho\left(s+1\right)=\rho\left(s\right)+\dfrac{\Lambda_{\rho}(\alpha(\mathrm{s}))}{n\cdot\det I(\alpha(s))} \\ b\left(s+1\right)=b\left(s\right)+\dfrac{\Lambda_{\mathrm{b}}(\alpha(\mathrm{s}))}{n\cdot\det I(\alpha(s))} \end{cases} \tag{10}$$

where

$$\Lambda_{\rho}\left(\alpha(\mathrm{s})\right)=\begin{pmatrix}\mathrm{U}_1(\alpha) & I_{12}(\alpha)\\ \mathrm{U}_2(\alpha) & I_{22}(\alpha)\end{pmatrix}, \qquad \Lambda_{\mathrm{b}}\left(\alpha(\mathrm{s})\right)=\begin{pmatrix}I_{11}(\alpha) & \mathrm{U}_1(\alpha)\\ I_{21}(\alpha) & \mathrm{U}_2(\alpha)\end{pmatrix},$$

and $\mathrm{U}_1\left(\alpha\right)=\dfrac{\partial l\left(X^n;\alpha\right)}{\partial \rho}$ and $\mathrm{U}_2\left(\alpha\right)=\dfrac{\partial l\left(X^n;\alpha\right)}{\partial b}$. We have

$$U_1(\alpha) = -n E_\alpha\left[m(\xi;\alpha)\right] + n\overline{\left(m^n(\theta)\right)}, \qquad U_2(\alpha) = -n E_\alpha\left[w(\xi;\alpha)\right] + n\overline{\left(w^n(\theta)\right)}.$$

Again, by comparing Farbod and Gasparian ([6], p. 229), we record the following algorithm for the model (2).

Algorithm II.

1. Generate random numbers from the model (2) using MCMC method;
2. Use (9) (or (10)) in order to calculate $\alpha_j(s)$, $j = 1,2$; $s = 0,1,2,\ldots$;
3. If $\left|\alpha_j(s+1) - \alpha_j(s)\right| < \varepsilon$ (ε is some small positive constant), then $\alpha_j(s+1) = \widehat{\alpha}$ is the MLE, other wise go to the step 2.

3.3 Simulation Studies

As we refereed in Introduction, to the author's knowledge, by now there have not been presented any closed expressions for the *cumulative distribution functions* of the models (1) and (2). Therefore, it is not possible to generate sample numbers at random based on the *cumulative distribution functions*. To overcome this difficulty, the MCMC method (see [8]) is used. Because of numerical calculations, let us consider the random variable ξ as *truncated*. Here, random variable ξ is restricted to 100. From Algorithms I and II we propose the Tables 1 and 2 for the models (1) and (2), respectively.

Note 1. The value $\theta = (p = 0.5,\ q = 1.9)$ is given as true value of the parameters of the model (1). Simulation studies are done for 1000 times to illustrate the behavior of the MLE. Namely, we do simulation for *M =1000* (*M* is the number of iteration) and *N=20, 50, 100, 200* (*N* is the sample size). Consider $\varepsilon = 0.0005$.

Table 1. The mean of estimations and the mean square errors (MSE) for the model (1)

	N=20		*N=50*	
Case	Mean	MSE	Mean	MSE
p	2.7226527	4.940185	2.7222685	4.938477
q	3.5981751	3.029295	3.4578110	2.962381
Iteration	274	---	208	---
	N=100		*N=200*	
Case	Mean	MSE	Mean	MSE
p	2.1383215	4.620950	1.8123654	3.818908
q	3.1511743	2.254472	2.5301244	1.743261
Iteration	139	---	187	---

Table 2. The mean of estimations and the MSE for the model (2)

	N=20		*N=50*	
Case	Mean	MSE	Mean	MSE
ρ	4.865513	4.266343	4.105983	1.705592
b	5.559738	18.145364	3.777491	6.137962
Iteration	1451	---	715	---
	N=100		*N=200*	
Case	Mean	MSE	Mean	MSE
ρ	2.081051	0.5168874	2.766523	0.0011207
b	0.477143	0.6770937	1.165229	0.0181633
Iteration	208	---	367	---

The mean of estimations and the MSE are given. As we see from Table 1, with increasing sample size the MSE decreases.

Note 2. The value $\alpha = (\rho = 2.8,\ b = 1.3)$ is considered as true value of the parameters of the model (2). We present the mean of estimations and the MSE. As we expected, from Table 2 it is readily seen that with increasing sample size the MSE decreases.

Note 3. All computations and simulations have been performed by using "R" statistical software.

Acknowledgement. This work was supported by a grant from Quchan University of Advanced Technology, Quchan, Iran.

ბიოინფორმატიკა

ბიოინფორმატიკაში წარმოშობილი სიხშირის განაწილების შეფასების ფორმულები

დ. ფარბოდი

ქუჩანის მოწინავე ტექნოლოგიების უნივერსიტეტი, მათემატიკის განყოფილება, ქუჩანი, ირანი

(წარმოდგენილია აკადემიის წევრის მ. სალუქვაძის მიერ)

განვიხილავთ სიხშირის განაწილების ორ ცნობილ სახეს, რომელიც გამოიყენება ინფორმატიკაში წარმოშობილი ფენომენების მოდელირებისათვის. ესენია, გარინგის ორპარამეტრიანი სიხშირის განაწილება და პარეტოს ტიპის ორპარამეტრიანი სიხშირის განაწილება, რომელიც ვ. ა. კუზნეცოვმა გაგვაცნო. გთავაზობთ *მაქსიმალური ალბათობის შეფასების ფორმულებს* (მაშფ) ასეთი სიხშირის განაწილების უცნობი პარამეტრებისათვის. შემოთავაზებული მაშფ-ი მყისიერი შეფასების ზოგიერთ ფორმულას ემთხვევა. ამასთან ერთად, მიღებულია მაშფ-ის მიახლოებითი გამოთვლის მეთოდი მოდელირების პარამეტრებისთვის. თეორიული შედეგების დასაბუთების მიზნით მოცემულია სიმულაციური კვლევები.

REFERENCES

1. *Kuznetsov V. A.* (2003) Signal Processing, **33**(4): 889-910.
2. *Kuznetsov V. A.* (2001) EURASIP J. Appl. Signal Processing, **4**: 285-296.
3. *Astola J. and Danielian E.* (2007) Frequency Distributions in Biomolecular Systems and Growing Networks, TICSP Series no. 31, Tampere, Finland
4. *Danielian E. and Astola J.* (2007) Proceedings of International TICSP Workshop on Spectral Methods and Multirate Signal Processing, Moscow, Russia, 235-237.
5. *Shiryaev A. N.* (1995) Probability (Graduate Texts in Mathematics), **95**, Springer. Translated from Russian.
6. *Farbod D. and Gasparian K.* (2013) J. Iranian Statist. Soc. (JIRSS), **12**(2): 211-234.
7. *Ivchenko G. and Medvedev Yu.* (1990) Mathematical Statistics. Translated from Russian.
8. *Rizzo M. L.* (2008) Statistical Computing with R, Chapman and Hall/CRC.

Received February, 2015