Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1575–1583
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Maximum likelihood estimation for constrained parameters of multinomial distributions—Application to Zipf–Mandelbrot models

F. Izsák[a, b, *]

[a] *ELTE, Institute of Mathematics, P.O. Box 120, 1518 Budapest, Hungary*
[b] *University of Twente, EWI, P.O. Box 217, 7500 AE Enschede, Netherlands*

Received 3 June 2005; received in revised form 10 May 2006; accepted 11 May 2006
Available online 12 June 2006

Abstract

A numerical maximum likelihood (ML) estimation procedure is developed for the constrained parameters of multinomial distributions. The main difficulty involved in computing the likelihood function is the precise and fast determination of the multinomial coefficients. For this the coefficients are rewritten into a telescopic product. The presented method is applied to the ML estimation of the Zipf–Mandelbrot (ZM) distribution, which provides a true model in many real-life cases. The examples discussed arise from ecological and medical observations. Based on the estimates, the hypothesis that the data is ZM distributed is tested using a chi-square test. The computer code of the presented procedure is available on request by the author.

Keywords: Maximum likelihood estimation; Multinomial distribution; Zipf–Mandelbrot model

1. Introduction

Analysis of data sets arising from a multinomial distribution is a frequently studied topic in a wide range of applications. In many cases, one has to test the validity of several models which are proposed to describe the investigated structure. Practically, based on an observation, one has to fit the parameters arising from the proposed model and, afterwards, a goodness-of-fit test should be performed.

A natural way to use this procedure is to apply a maximum likelihood (ML) estimation for the unknown parameter, and then, based on Fisher's classical theorem, the goodness of fit can be tested using the *chi-square* statistics (Lehmann, 1997). In spite of this simple draft, in the course of implementing the scheme we are confronted with some difficulties: the essential problem is the numerical treatment of the ML method. Big samples and multidimensional parameter sets lead us to considerable maximization problems. In our case, the likelihood function (l_θ) should be computed with a high precision since it is very small and, therefore, tiny errors can result in considerably unsharp estimates after the maximization procedure. As an application, we investigate the ML estimation for the parameters of the Zipf–Mandelbrot (ZM) distributions in detail and describe the fitting procedure to data sets arising from some ecological observations.

* Corresponding author at: ELTE, Institute of Mathematics, P.O. Box 120, 1518 Budapest, Hungary. Tel.: +36 13812157; fax: +36 13812158.
E-mail address: izsakf@cs.elte.hu (F. Izsák).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.05.008

In an early work (Zipf, 1949), Zipf proposed a probability distribution to describe the frequencies of word occurrences in linguistics. This model has been modified by Mandelbrot, who applied it to observations in economical studies (Mandelbrot, 1977a,b) focusing on the description of the income distribution; a further generalization can be found in Zornig and Altmann (1995). For other applications of the ZM model to linguistics we refer to Egghe (1999) and Meadow et al. (1993).

Nowadays, the main field of the applications of the ZM model is related to biology. Scientists observed that, in many ecological communities, the frequencies of each species can also be related to these kinds of distributions (see Frontier, 1985, and the references therein). A great deal of medical observations (for example, the spreading of diseases (Sabatier et al., 1998) and mortality statistics (Li and Yang, 2002)) can also be connected to this model.

Recently, analysis of complex systems like the Internet (Huberman et al., 1998) or DNS molecules (Kuznetsov, 2002) have produced data sets which seem to be well characterized in the framework of Zipf's law or the ZM distribution. In light of the broad occurrence of data structures related to the ZM distribution in real-life situations (see also http://www.nslij-genetics.org/wli/zipf/), an effective fitting procedure and an appropriate test for the goodness of fit may be of great importance.

There are several attempts and suggestions in the literature about how to estimate the parameters of the ZM distribution, but none of the authors presented a procedure for the ML estimation. For a detailed review on the existing results, we refer to Rousseau (2002) and Wilson (1991).

The outline of the paper is as follows. After stating our problem mathematically, we provide the proposed numerical treatment of the ML estimation for the parameters of a constrained multinomial distribution. Then, we apply it to parameter estimation in the ZM model using concrete ecological/medical observations. The goodness of fit will also be checked by performing a chi-square test.

We also demonstrate the power of our method compared with some conventional techniques.

2. Mathematical formalization

For ease of the presentation, we use terms such as *population* or *species* in this section.

The proposed model for the distribution of the population is given as a family of probability distributions

$$\left\{\mathbf{p}_\theta = \left(p_{\theta,1}, p_{\theta,2}, \ldots, p_{\theta,r}\right), \theta \in \Theta \subset \mathbf{R}^k\right\},$$

where $p_{\theta,j}$ gives the probability of the occurrence of the jth species in the population. Taking a random sample $x = (n_1, n_2, \ldots, n_r)$ of size n from this population, with n_j being the frequency of the jth species, we obtain that x is multinomially distributed with the parameters $(n, \mathbf{p}_\theta)$. Our aim is to estimate θ. Moreover, we intend to control the goodness of fit for the obtained estimate $\hat{\theta}$ by a hypothesis testing procedure.

Proposition 1. *Using the notations above we obtain the likelihood equation for the sample x*:

$$l_\theta(x) = \frac{n!}{n_1!n_2!\ldots n_r!} \prod_{i=1}^{r} p_{\theta,i}^{n_i}, \tag{1}$$

which gives the probability of the observed sample supposed that it arises from a $\mathbf{p}_\theta$-distributed population.

The ML estimate $\hat{\theta}$ of θ is the maximum site of the function $l : \theta \to l_\theta(x)$. Several methods have been proposed to numerically deal with the related maximization problem whenever there is no known explicit formula for $\hat{\theta}$.

A conventional method to obtain $\hat{\theta}$ involves the application of the EM algorithm (McLachlan and Krishnan, 1997). This algorithm is rather general and special procedures have been developed for treating maximization problems related to ML estimates (also with constraints). For a broad overview and further literature, we refer the reader to Jamshidian (2004).

Here, we propose an alternative method which is easy to implement: the numerical treatment of the maximization problem can be facilitated by expanding (1) into a telescopic product and calculating the multinomial coefficients in this way.

Lemma 1. *Using the notations*

$$n^j = n - \sum_{i=1}^{j-1} n_i \quad \text{and} \quad p_\theta^j = 1 - \sum_{i=1}^{j-1} p_{\theta,i}$$

the likelihood equation (1) *can be rewritten as follows*:

$$l_\theta(x) = \prod_{j=1}^{r-1} B_{n^j}^{p_{\theta,j}/p_\theta^j}\left(n_j\right),$$

where $B_N^p(k) = \binom{N}{k} p^k (1-p)^{N-k}$ *denotes the kth member of the binomial distribution with the parameters* (N, p).

Proof. We proceed by induction. For $r = 2$ the product consists of only one component, namely

$$B_{n_1+n_2}^{p_{\theta,1}}(n_1) = \frac{n!}{n_1! n_2!} p_{\theta,1}^{n_1} p_{\theta,2}^{n_2}$$

with $p_{\theta,1} + p_{\theta,2} = 1$ and $n_1 + n_2 = n$, which proves the lemma in this case.

Assuming that the lemma is proved for r, we apply it in the second step for the r-tuple $(n_1, n_2, \ldots, n_{r-1}, n_r + n_{r+1})$ and the related (discrete) probability distribution $\left(p_{\theta,1}, p_{\theta,2}, \ldots, p_{\theta,r-1}, p_{\theta,r} + p_{\theta,r+1}\right)$ as follows:

$$\begin{aligned}
l_\theta(x) &= \frac{n!}{n_1! n_2! \ldots n_{r+1}!} \prod_{i=1}^{r+1} P_{\theta,i}^{n_i} \\
&= \frac{(n_r + n_{r+1})!}{n_r! n_{r+1}!} \cdot \frac{p_{\theta,r}^{n_r} p_{\theta,r+1}^{n_{r+1}}}{\left(p_{\theta,r} + p_{\theta,r+1}\right)^{n_r+n_{r+1}}} \\
&\quad \times \frac{n!}{n_1! n_2! \ldots n_{r-1}! (n_r + n_{r+1})!} \cdot \left(p_{\theta,r} + p_{\theta,r+1}\right)^{n_r+n_{r+1}} \prod_{i=1}^{r-1} p_{\theta,i}^{n_i} \\
&= \frac{(n_r + n_{r+1})!}{n_r! n_{r+1}!} \left(\frac{p_{\theta,r}}{p_{\theta,r} + p_{\theta,r+1}}\right)^{n_r} \left(\frac{p_{\theta,r+1}}{p_{\theta,r} + p_{\theta,r+1}}\right)^{n_{r+1}} \prod_{j=1}^{r-1} B_{n^j}^{p_{\theta,j}/p_\theta^j}\left(n_j\right) \\
&= B_{n^r}^{p_{\theta,r}/p_\theta^r}(n_r)\, B_{n^{r+1}}^{p_{\theta,r+1}/p_\theta^j}(n_{r+1}) \prod_{j=1}^{r-1} B_{n^j}^{p_{\theta,j}/p_\theta^j}\left(n_j\right) = \prod_{j=1}^{r+1} B_{n^j}^{p_{\theta,j}/p_\theta^j}\left(n_j\right). \qquad (2)
\end{aligned}$$

This proves the lemma for $r + 1$ and we are done. □

In order to apply the representation in Lemma 1 for developing a fast and precise ML estimation algorithm, we need to have a subroutine which provides the terms $B_{n^j}^{p_{\theta,j}/p_\theta^j}\left(n_j\right)$ effectively. Then a (numerical) maximization process serves the (approximation of the desired) ML estimate $\hat{\theta}$.

3. Application

We demonstrate the effectiveness of the computational procedure (based on Lemma 1) by testing the hypothesis that some data sets arise from the ZM distribution defined as follows.

Definition 1. A random variable X is ZM distributed ($X \sim ZM_{b,c}$) if for some $b, c \in \mathbf{R}$

$$P(X = i) = \frac{a}{(b+i)^c} \quad \text{for } i = 1, 2, \ldots, r, \qquad (3)$$

where $a = \left(\sum_{i=1}^{r} (b+i)^{-c}\right)^{-1}$.

Qualitatively, data sets arising from ZM distributed populations exhibit fast decreasing profile; sometimes the last observations are just ignored. Formally, (3) makes also sense for non-integer i, in this context the function $f : \mathbf{R}^+ \to \mathbf{R}$, $f(x) = a/(b+x)^c$ is called ZM curve.

Many efforts have been made to provide an estimation for the parameters of the ZM distribution. One can use momentum methods based on some of the first observations (Zornig and Altmann, 1995), or in an other study (Piqueira et al., 1999), the authors apply only a least-square fitting procedure and propose the parameters of the fitted ZM-like curve to be $\hat{b}$ and $\hat{c}$. One realizes immediately the weakness of this latter approach by the exhibited instability and horrible growing of the "fitted parameters" $\hat{b}$ and $\hat{c}$. In the fitting problem alone, one frequently considers "log" or "log–log" scales which translates the problem into that of a linear fitting.

We formalize first the problem in the framework introduced in Section 2: we investigate the case when the family of probability distributions is given by

$$\mathbf{p}_{(b,c)} = \left\{ \left(p_{(b,c),1}, p_{(b,c),2}, \ldots, p_{(b,c),r} \right) : (b,c) \in \mathbf{R}^2 \right\}$$

with

$$p_{(b,c),i} = \frac{a}{(b+i)^c}, \quad i \in \{1, 2, \ldots, r\}.$$

An observation is identified with the vector $x = (n_1, n_2, \ldots, n_r)$, where $n_1 + n_2 + \cdots + n_r = n$ which consists of the observed frequencies of the species. For the likelihood function one should know the number of the species in the whole population whenever they have not been observed. "How many species there are?", as it is usually asked. A scale of methods has been elaborated to give a meaningful estimate (Chao, 1984; Colwell and Coddington, 1994; Palmer, 1990, 1991); for a comparison of its performance, we refer to Papp et al. (1997). In practice, one should use a particular estimate according to the origin and the structure of the data: in the present situation, some non-parametric methods can be proposed. In this way, the observation x should be enlarged with some zeros, such that its length is the estimated number of species.

Although we do not estimate the number of species, we point out that the estimate in our second example is robust in the sense that, through assuming more species in the population (i.e., more than observed), the estimate $(\hat{b}, \hat{c})$ does not change substantially.

A further problem arises, when we try to find a correspondence between the observed frequencies $\{n_i\}_{i=1}^r$ and the probabilities $\left\{p_{\theta,i}\right\}_{i=1}^r$ in the model. To solve this, we state the following:

Lemma 2. *Assume that for some indices $1 \leqslant i < j \leqslant r$ the inequalities $p_{\theta,i} \geqslant p_{\theta,j}$ and $n_i \geqslant n_j$ hold. Then*

$$\frac{n!}{n_1!n_2!\ldots n_r!} p_{\theta,1}^{n_1} \cdots p_{\theta,i}^{n_i} \cdots p_{\theta,j}^{n_j} \cdots p_{\theta,r}^{n_r} \geqslant \frac{n!}{n_1!n_2!\ldots n_r!} p_{\theta,1}^{n_1} \cdots p_{\theta,i}^{n_j} \cdots p_{\theta,j}^{n_i} \cdots p_{\theta,r}^{n_r}.$$

Proof. Since $p_i \geqslant p_j$ we obtain that

$$p_{\theta,i}^{n_i} p_{\theta,j}^{n_j} = p_{\theta,i}^{n_i - n_j} p_{\theta,i}^{n_j} p_{\theta,j}^{n_j - n_i} p_{\theta,j}^{n_i} = \left(\frac{p_{\theta,i}}{p_{\theta,j}} \right)^{n_i - n_j} p_{\theta,i}^{n_j} p_{\theta,j}^{n_i} \geqslant p_{\theta,i}^{n_j} p_{\theta,j}^{n_i}$$

which proves the lemma. □

In the ML procedure, one has to find the probabilities $\left\{p_{\theta,i}\right\}_{i=1}^r$ such that the likelihood function (1) is maximal. According to Lemma 2, for any parameters b, c and probabilities $p_{\theta,1} \geqslant p_{\theta,2} \geqslant \cdots \geqslant p_{\theta,r}$ the likelihood function is the largest possible if the frequencies in the observation are ordered also decreasingly: $n_1 \geqslant n_2 \geqslant \cdots \geqslant n_r$. Using Proposition 1, the explicit form of the likelihood function is

$$l_{b,c}(x) = \frac{n!}{n_1!n_2!\ldots n_r!} \prod_{i=1}^{r} \left(\frac{a}{(b+i)^c} \right)^{n_i}, \tag{4}$$

which can be maximized numerically using the formula in Lemma 1.

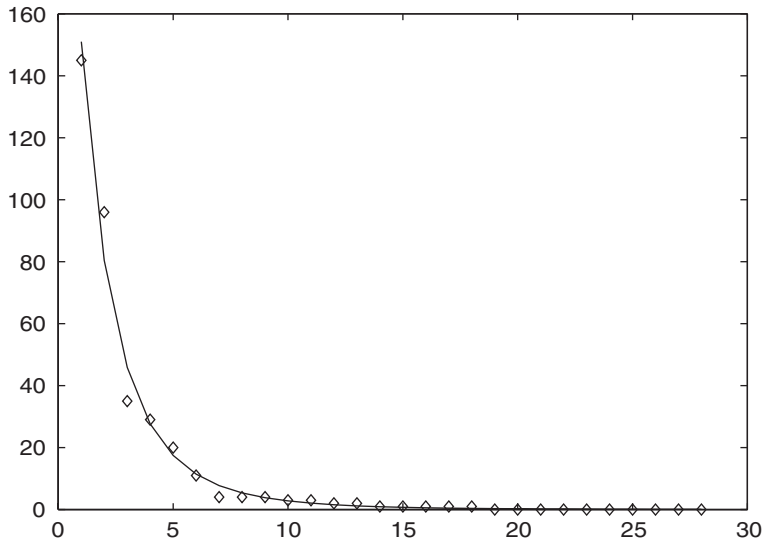

Fig. 1. ML estimate for the data set (5): we plotted the piecewise linear interpolation of the fitted ZM curve based on the estimate $\hat{b} = 4.743$ and $\hat{c} = 4.124$. The observed data set is depicted with the $\diamond$ symbols.

4. Numerical results

4.1. Parameter estimation

In the first example, we refer to an ecological observation (Papp, 1992) on the diversity of fly species in a certain territory. The observation is represented by the vector

$$x = \texttt{[145 96 35 29 20 11 4 4 4 3 3 2 2 1 1 1 1 1].} \tag{5}$$

The ML estimate of b and c has been executed with MATLAB based on the telescopic expansion in Lemma 1. In course of the numerical computation of the likelihood function, we made use of the built-in binomial distribution function `binopdf.m` and the maximization subroutine `fmins.m` which needs also an initial value (as an input parameter). For the details of these subroutines we refer to Loader (2002) and http://www.mathworks.com/access/helpdesk_r13/help/techdoc/ref/fmins.html.

The code is available on request by the author.

Fig. 1 shows the result of the parameter estimation together with the observed data set: we presented a linear interpolation (between some integers) of the ZM curve based on the ML estimate: $\hat{b} = 4.743$ and $\hat{c} = 4.124$. In the maximization procedure, we chose initially $b_0 = 1$ and $c_0 = 3$. For a better visual outcome we did not add zeros to the observation (which could represent the non-observed species).

The second example arises from medical statistics resulting from a study on cancer diseases of rats (Lang, 1992), where

$$\begin{aligned} x = \texttt{[902 393 221 131 91 76 53 51 49 48 42 32 28 18} \\ \texttt{16 14 13 13 10 9 8 7 6 6 6 6 6 5 5 4 4 4 4 4} \\ \texttt{3*ones(1,11), 2*ones(1,14), ones(1,50)]} \end{aligned} \tag{6}$$

using the short notation of MATLAB. In this case we obtained the estimate: $\hat{b} = 0.619$, $\hat{c} = 1.747$. The related ZM curve along with the observation is shown in Fig. 2. In the maximization procedure we chose initially $b_0 = 0.1$ and $c_0 = 3$.

Based on the size of the observation, we expect that some of the species in the population are not represented in this study. Therefore, we also computed $\left(\hat{b}, \hat{c}\right)$ assuming 1, 2, 5, 10 and 20 non-observed species, respectively. The results are shown in Table 1.

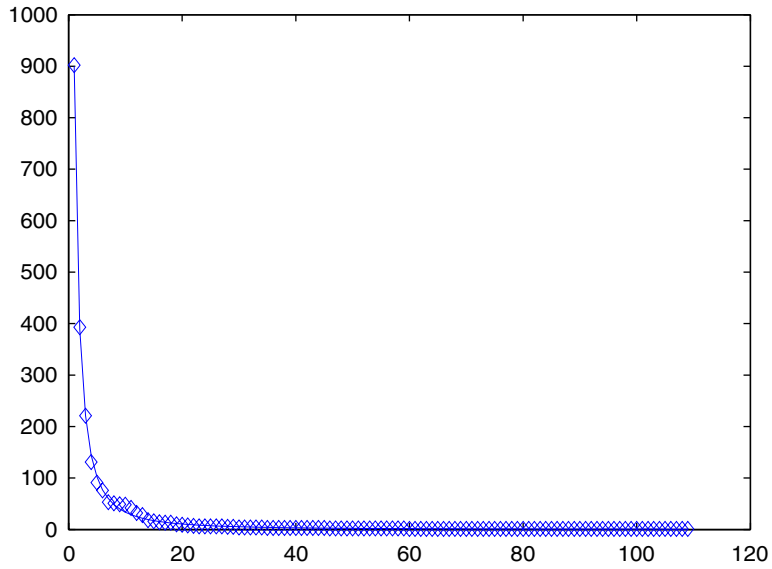

Fig. 2. ML estimate for the data set (6): we plotted the piecewise linear interpolation of the fitted ZM curve based on the estimate $\hat{b} = 0.619$ and $\hat{c} = 1.747$. The observed data set is depicted with the $\diamond$ symbols.

Table 1
ML estimate for the data set (6) involving non-observed species represented by zero in the sample

$\hat{b}$	$\hat{c}$	$l_{\hat{b},\hat{c}}(x)$	Number of zeros
0.6193	1.7469	3.6×10^{-77}	0
0.6246	1.7496	2.1×10^{-77}	1
0.6300	1.7522	1.19×10^{-77}	2
0.6455	1.7599	2.47×10^{-78}	5
0.6699	1.7719	2.19×10^{-79}	10
0.7137	1.7933	3.35×10^{-81}	20

4.2. Goodness of fit

The goodness of fit is investigated using the chi-square test (Lehmann, 1997). This is justified since $\left(\hat{b}, \hat{c}\right)$ is the ML estimate. For a discussion of this topic, we refer to Berkson (1980). In case of observation (5), we performed the chi-square test such that the last 5 observations have been contracted into a new one and we performed in the same way with the preceding 3, 2 and again 2 observations such that we obtained

$x = [145\ 96\ 35\ 29\ 20\ 11\ 8\ 7\ 7\ 5].$

Here, the resulting chi-square statistic is $\chi = 9.557$ with the degrees of freedom $10 - 3 = 7$. Therefore, we should accept the hypothesis that the data set (4.2) is ZM distributed with the parameters $\hat{b} = 4.743$ and $\hat{c} = 4.124$.

In the second case, we modify (6) in the way that we omit the observations with the frequencies less than 5. Then we obtain $\chi = 30.57$ for the chi-square test statistic, with the degrees of freedom $29 - 3 = 26$. In this way, we accept the hypothesis that the observation is arising from a ZM distributed population with the parameters $\hat{b} = 0.619$, $\hat{c} = 1.747$.

To summarize, the present investigation (using parameter estimation and hypothesis testing) confirms the validity of the ZM model.

4.3. Comparison with an approximation

In this subsection, we point out the power of the method developed in this paper. We compare it with a quite straightforward approximation for the likelihood function.

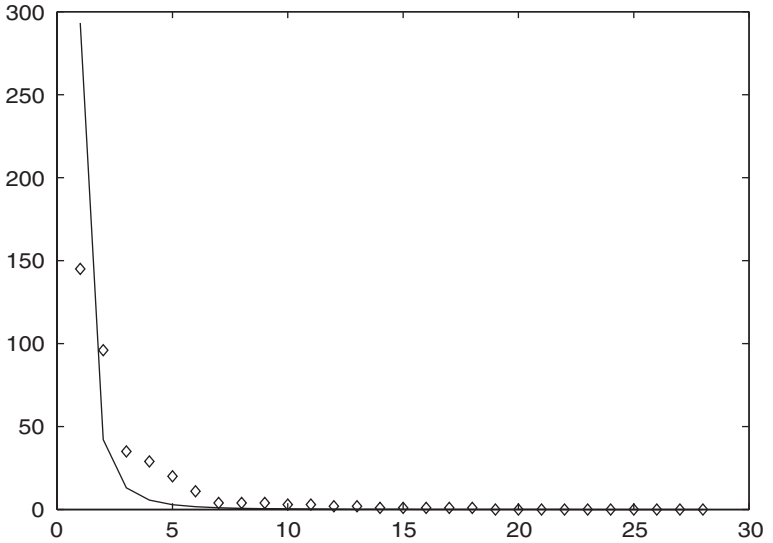

Fig. 3. Result of the parameter estimation using (8): the data set (5) and a piecewise linear interpolation of the fitted ZM curve obtained by ML estimation using the approximation in (8).

Since the values of a multinomial distribution function consist of many factorials (with quite big numbers, depending on the observation), it seems to be handy to apply the Stirling formula as an approximation. We use the following simple version:

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}. \tag{7}$$

For improving the sharpness of the approximation, one could use further terms on the right-hand side of (7). This, however, does not have any effect on the maximization procedure since the values n_i are fixed in the computations. For the same reason, we may also omit the constant $\sqrt{2\pi n}$ in (7) and obtain the following approximation of the log-likelihood function:

$$\begin{aligned}\log l_{b,c}(x) &= \log\left[\frac{n!}{n_1!n_2!\ldots n_r!}\prod_{i=1}^{r}\left(\frac{a}{(b+i)^c}\right)^{n_i}\right]\\ &= \log n! - \sum_{i=1}^{r}\log\, n_i! + \sum_{i=1}^{r} n_i(\log a - c\log(b+i)).\end{aligned} \tag{8}$$

We also implemented this approximation in course of the numerical computation of the ML estimate. Results for the first and second observation are shown in Figs. 3 and 4, respectively.

In the first case (using the same input parameters in the maximization), the chi-square statistics was $\chi = 101.4$. In the second case, we started the maximization with the input parameters (0.5, 2) (nearly to the exact one) and obtained $\chi = 189.7$. These show clearly the inconvenience of the classical approximation.

5. Discussion

Based on the numerical experiments, we propose that the direct method developed here provides an effective procedure for computing multinomial coefficients. This makes it possible to obtain the ML estimates that we pointed out in the case of the ZM model.

The method should be improved in some ways. First, during the estimation, we can proceed only in the case of $n \lessapprox 3000$. In order to force this condition, we could reduce the order of the data, e.g. by taking the annual income per thousand dollars or measure the ten thousands of inhabitants in an economical and demographical analysis, respectively.

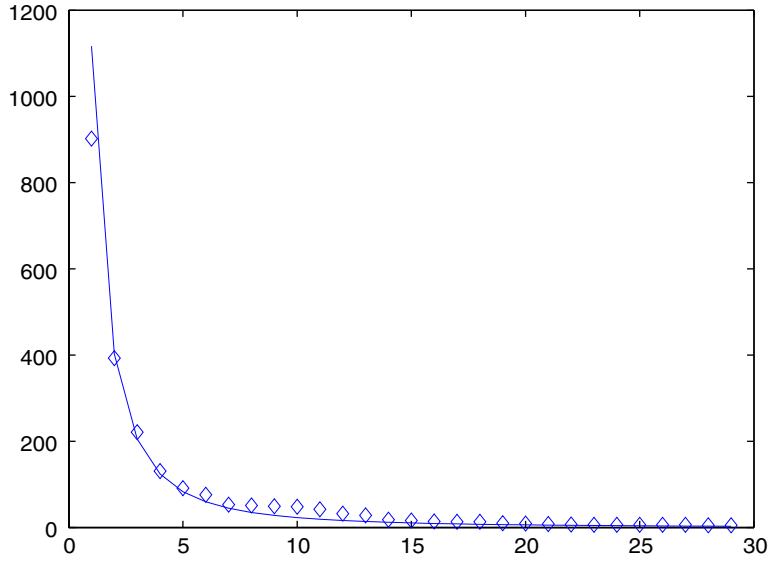

Fig. 4. Result of the parameter estimation using (8): the data set (6) and a piecewise linear interpolation of the fitted ZM curve obtained by ML estimation using the approximation in (8).

This method is a remedy for the problem from a practical point of view (Marsili and Zhang, 1998), but, mathematically, it is hardly acceptable.

Another crucial question emerges while testing the hypothesis in Section 4.2. While using the chi-square test, it is generally suggested that observations should be grouped in such a way that each of the groups contains at least five elements. Since it was not the case in the presented examples, we omitted the low-frequency observations during the test. Indeed, these low-frequency observations do not really reflect the theoretical frequencies of the species.

References

Berkson, J., 1980. Minimum chi-square, not maximum likelihood! Ann. Statist. 8 (3), 457–487.

Chao, A., 1984. Non-parametric estimation of the number of the classes in a population. Scand. J. Statist. 11, 265–270.

Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. In: Hawksworth, D.L. (Ed.), Biodiversity. Measurement and Estimation. The Royal Society and Chapman & Hall, London, pp. 101–118.

Egghe, L., 1999. On the law of Zipf–Mandelbrot for multi-word phrases. J. Amer. Soc. Inform. Sci. 50, 233–242.

Frontier, S., 1985. Diversity and structure in aquatic ecosystems. Oceanogr. Mar. Biol. (Ann. rev.) 23, 253–312.

Huberman, B.A., Pirollo, P.L.T., Pitkow, J.E., Lukose, R.M., 1998. Strong regularities in world wide web surfing. Science 280, 95–97.

Jamshidian, M., 2004. On algorithms for restricted maximum likelihood estimation. Comput. Statist. Data Anal. 45, 137–157.

Kuznetsov, V.A., 2002. Statistics of the numbers of transcripts and protein sequences encoded in the genome. In: Computational and Statistical Methods to Genomics, Kluwer, Boston, pp. 125–171.

Lang, P.L., 1992. Spontaneous neoplastic lesions and selected non-neoplastic lesions in the Crl: CD^R BR Rat. Charles River Laboratories.

Lehmann, E.L., 1997. Testing Statistical Hypotheses. second ed. Springer, New York. pp. 480–494.

Li, W., Yang, Y., 2002. Zipf's law in importance of genes for cancer classification using microarray data. J. Theoret. Biol. 219, 539–551.

Loader, C., 2002. Fast and accurate computation of binomial probabilities. Available at ⟨http://www.herine.net/stat/papers/dbinom.pdf⟩.

Mandelbrot, B.B., 1977a. Fractals: Form, Chance and Dimension. W.H. Freeman and Co., San Francisco, pp. 365–370.

Mandelbrot, B.B., 1977b. The Fractal Geometry of Nature. W.H. Freeman and Co., San Francisco, pp. 468–480.

Marsili, M., Zhang, Y.-C., 1998. Interacting individuals leading to Zipf's Law. Phys. Rev. Lett. 80 (12), 2741–2744.

⟨http://www.mathworks.com/access/helpdesk_r13/help/techdoc/ref/fmins.html⟩.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.

Meadow, C., Wang, J., Stamboulie, M., 1993. An analysis of Zipf–Mandelbrot language measures and their application to artificial languages. J. Inform. Sci. 19, 247–258.

⟨http://www.nslij-genetics.org/wli/zipf/⟩.

Palmer, M.W., 1990. The estimation of species richness by extrapolation. Ecology 71 (3), 1195–1198.

Palmer, M.W., 1991. The estimation of species richness: the second-order Jacknife reconsidered. Ecology 72 (4), 1512–1513.

Papp, L., 1992. Drosophilid assemblages in mountain creek valleys in Hungary (Diptera: Drosophilidae) I. Folia Entomol. Hung. 53, 139–153.

Papp, L., Izsák, J., Ádám, L., 1997. Dipterious assemblages of sheep-run droppings: number of species observed, estimated, and generated by simulation. Acta Zool. Acad. Sci. H. 43 (3), 191–205.

Piqueira, J.R.C., Monteiro, L.H.A., deMagalhães, T.M.C., Ramos, R.T., Sassi, R.B., Cruz, E.G., 1999. Zipf's Law organizes a psychiatric ward. J. Theoret. Biol. 198, 439–443.

Rousseau, R., 2002. Lack of standardization of informetric research. Comments on "Power laws of research output. Evidence for journals of economics" by Matthias Sutter and Martin G. Kocher. Scientometrics 55, 317–327.

Sabatier, P., Guigal, P.-M., Dubois, D.M., 1998. Fractals and epidemic process. Internat. J. Comput. Anticipatory Systems (Publ. by CHAOS) 1, 135–145.

Wilson, J.B., 1991. Methods for fitting dominance/diversity curves. J. Vegetation Sci. 2, 35–46.

Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Hafner, New York.

Zornig, P., Altmann, G., 1995. Unified representation of Zipf distributions. Comput. Statist. Data Anal. 4, 461–473.