An efficient EM algorithm for the mixture of negative binomial models

Chunmao Huang¹, Xingwang Liu¹, Tianyuan Yao² and Xiaoqiang Wang², ³

¹School of Science, Harbin Institute of Technology at Weihai, 2 Wenhua Xilu, Weihai, Shandong Province, P.R. China 264209
²School of Mathematics and Statistics, Shandong University, Weihai, 180 Wenhua Xilu, Weihai, Shandong Province, P.R. China 264209
³Email: xiaoqiang.wang@sdu.edu.cn

Abstract. Overdispersion is a widespread phenomenon in most count data sets. The negative binomial distribution is commonly adopted to fit over-dispersed count data. On the other hand, the mixture model always plays an important role in unsupervised classification. However, when estimating the parameters in the mixture of negative binomial models, the typical generalized Expectation Maximization (EM) algorithm which involves additional iterative procedures in M-step increases computational time. Hence, there remains a need for an efficient algorithm that can speed up the procedure of parameter estimation. For this purpose, here we develop a novel EM algorithm that successfully avoids the typical numerical solution in M-step for the mixture of negative binomial models. We extend further this EM algorithm to the zero-inflated negative binomial model. In the simulation studies, we focus on the runtimes and the classification performance of our proposed algorithm implemented in the mixture of negative binomial model. We found that our proposed EM algorithm can reduce the runtime of maximum likelihood estimation effectively, while achieving the similar classification performance in comparison with the typical EM algorithm. The mixture of negative binomial model and the proposed EM algorithm finally illustrates their good performance of fitting the real earthquake count data.

1. Introduction

In recent years, the negative binomial (NB) distribution has become increasingly popular because it can fit the over-dispersed count data effectively. Furthermore, these count data are widely distributed in various fields such as biology [1], insurance [2] and earthquake [3]. In particular, it is widely recognized that the NB distribution provides a good fit to read counts data produced by next generation sequencing (NGS) methods. In this paper, we focus on the maximum likelihood estimation under the framework of the heterogeneity and overdispersion of count data by considering a mixture of NB distributions.

The EM algorithm is usually adopted to estimate the maximum likelihood for mixture models. However, the digamma functions in M-step lead to no closed form when updating the dispersion parameters for the mixture of NB model. Consequently, some numerical solutions including Nelder-Mead simplex algorithm, the quasi Newton-Raphson algorithm must be carried out in M-step. It is worth noticing that the additional iterative processes highlight the time-consuming shortcoming of the generalized EM algorithms.
Quenouille (1949) [4] implied that the NB distribution is one of the compound Poisson distributions. Making use of this property, Adamidis (1999) [5] overcame the iterative scheme within the EM algorithm by introducing a truncated exponential distribution. Nevertheless, Adamidis’ EM algorithm was only realized for the samples drawn from an overall population. As to our knowledge, no similar algorithms have been proposed to estimate the maximum likelihood for the mixture of NB distributions.

Motivated by exploiting a more efficient EM algorithm for the mixture of NB model, in this article we fabricate some variables for estimating the maximum likelihood. This strategy is implemented in the mixture of NB distribution and the mixture of zero-inflated NB distribution respectively. We show how employing the proposed fictitious variables can improve the performance in terms of runtimes on simulated data. Simulation studies further demonstrate that our algorithm has the similar performance of classification in comparison with the generalized EM algorithm. It implies that the proposed EM algorithm is convergent.

The remainder of the article is organized as follows. In Section 2, we review firstly the model and the EM algorithm. Then, we present a novel EM algorithm with closed form in M-step on the mixture of model. This EM algorithm is developed on the mixture of zero-inflated NB model. In Section 3, we assess the performance of our proposed algorithm on simulated data and carry out the comparisons with the other generalized EM algorithm. In Section 4, we illustrate an application of the proposed algorithm for fitting the real earthquake data.

### 2. Mixture model and EM algorithm

#### 2.1. Review of negative binomial model and EM algorithm

The negative binomial distribution can be identified by any two of the four parameters, say size \( r \), probability \( \theta \), mean \( \mu \) and variance \( \sigma^2 \). Ordinarily, the probability mass function (PMF) is expressed as

\[
h(x|r, \theta) = \frac{\Gamma(r + x)}{x! \Gamma(r)} \theta^r (1 - \theta)^x, \quad (x = 0, 1, 2, \ldots). \tag{1}
\]

Quenouille (1949) presents firstly the relationship between the logarithmic, Poisson and negative binomial distribution. It results that the negative binomial distribution \( NB(r, \theta) \) can be considered as one of the compound Poisson distribution. Concretely, assuming the random variable \( M \) follows Poisson distribution with the mean \( \lambda \). Assuming further that \( Y_j, j = 1, \ldots, M \) is a sequence of i.i.d. random variables, each one having the logarithmic distribution with probability mass function

\[
p(y|\theta) = -\frac{1}{\ln \theta} \frac{(1 - \theta)^y}{y}, \quad (y = 1, 2, \ldots),
\]

then the random sum \( X = \sum_{j=1}^M Y_j \) follows \( NB \) distribution with parameters \( (r, \theta) \), where \( r = -\lambda / \ln \theta \).

In the context, the mixture model refers to the weighted average of a plurality of distributions

\[
p(x|\omega, r, \theta) = \sum_{k=1}^K \omega_k h(x|r_k, \theta_k), \quad \sum_{k=1}^K \omega_k = 1, \omega_k > 0, k = 1, \ldots, K,
\]

where \( h(x|r_k, \theta_k) \) is the PMF of NB distribution with parameters \( (r_k, \theta_k) \).

EM algorithm [6] is a standard methodology for maximum likelihood estimation with latent variables. This algorithm is performed by two alternative steps which consist of an expectation (E) step and a maximization (M) step. In M-step we need to update the parameters though maximizing the conditional expected log-likelihood found on the E-step. However, for the mixture of NB model with the PMF (1), the M-step has no closed form solution when solving the equation involving the digamma functions. This leads to the fact that the numerical solution must be employed for optimizing the objective function in M-step. Nevertheless, the iterative EM algorithm including another iterative procedure will inevitably increase the runtime of algorithms. Hence, the exact EM algorithm should be further investigated to boost up the speed. Adamidis (1999) firstly derived an EM algorithm that can
avoid the iterative process in M-step while updating the parameters in EM algorithm applied in the NB distribution. However, the derivation of such EM algorithm has not been proposed yet for the commonly used mixture of NB model.

2.2. Mixture of NB model and EM algorithm

In this section, we aim to present an efficient EM algorithm to estimate the parameters in the mixture of NB models. We omit the process of algorithm derivation, a similar deviation of EM algorithm can be found in Adamidis (1999).

Let \( x = (x_t)_{t=1,...,T} \) denote a sample of size \( T \) which sampling from the above mixture model. For each observation sampling from the \( k \)th subpopulation, inspired from Adamidis (1999), we fabricate a triple consisting of the independent random variables \((M_t, Y_t, Z_t)\) corresponding to \( x_t \), where \( M_t \) is a Poisson variable with the parameter \( \lambda_k \), \( Y_t = (Y_{t1},...,Y_{tM_t}) \) is a random vector comprised of \( M_t \) i.i.d. logarithmic variables with the parameter \( \theta_k \) and \( Z_t = (Z_{t1},...,Z_{tM_t}) \) comprised of \( M_t \) i.i.d. truncated exponential variables with the density

\[
p(z|\theta_k) = \frac{-\ln\theta_k}{1-\theta_k} \theta_k^z, \quad (0 < z < 1).
\]

Moreover, we introduce the hidden variable \( S = (S_t)_{t=1,...,T} \) associated with the allocation of subpopulation. Furthermore, the likelihood function of the imaginary complete data \((S, M, Y, Z)\) can be decomposed as

\[
p(S, M, Y, Z) = \prod_{t=1}^{T} \prod_{k=1}^{M} \omega_k p(M_t|\phi_k) \prod_{j=1}^{M_t} p(Y_{tj}|\phi_k) \prod_{j=1}^{M_t} p(Z_{tj}|\phi_k)
\]

where \( \omega_k = p(S = k) \), \( S^k_t = 1_{(S_t = k)} \) is an indicative function and \( \phi = (\phi_k) = (r_k, \theta_k) \).

Given the observations \( x \), we will maximize the loglikelihood \( \mathcal{L}(\phi; x) \) rather than maximizing the typical loglikelihood \( \mathcal{L}(\phi; x) \). Therefore, the Q function can be written as

\[
Q(\phi, \phi^{(h)}) = E[\log(p(S, M, Y, Z; \phi)) | x, \phi^{(h)}] = Q(\omega, \omega^{(h)}) + Q(\phi, \phi^{(h)})
\]

where

\[
Q(\omega, \omega^{(h)}) = \sum_t \sum_k \tau^{(h)}_{k} \log \omega_k,
\]

\[
Q(\phi, \phi^{(h)}) = \sum_t \sum_k \tau^{(h)}_{k} \left[ \delta^{(h)}_{k} \log \lambda_k - \lambda_k (x_t - \delta^{(h)}_{k}) \log (1 - \theta_k) + \theta^{(h)}_{k} \theta^{(h)}_{k} \log \theta_k \right]
\]

with \( \tau^{(h)}_{k} = P(S_t = k | x_t, \phi^{(h)}) \), \( \delta^{(h)}_{k} = E(M_t | S_t = k, x, \phi^{(h)}) \) and \( \theta^{(h)}_{k} = E(Z_{tj} | S_t = k, x, \phi^{(h)}) \).

In E-step, we need to calculate the conditional probabilities as

\[
\tau^{(h)}_{k} = P(S_t = k | x_t, \phi^{(h)}) \propto \omega_k^{(h)} h \left( x_t | r^{(h)}_{k}, \theta^{(h)}_{k} \right).
\]

\[
\delta^{(h)}_{k} = \frac{1}{1 - \theta^{(h)}_{k}} \left( \psi \left( \tau^{(h)}_{k} + x_t \right) - \psi \left( \tau^{(h)}_{k} \right) \right),
\]

where \( \psi(\cdot) \) is digamma function and \( r^{(h+1)}_{k} = -\lambda^{(h+1)}_{k} / \log \theta^{(h+1)}_{k} \).

In M-step, we need to update the parameters as

\[
\omega^{(h+1)}_{k} = \sum_t \tau^{(h)}_{k}, \quad \lambda^{(h+1)}_{k} = \frac{\sum_t \tau^{(h)}_{k} \delta^{(h)}_{k}}{\sum_t \tau^{(h)}_{k}}, \quad \theta^{(h+1)}_{k} = \frac{\theta^{(h)}_{k} \sum_t \tau^{(h)}_{k} \delta^{(h)}_{k}}{\sum_t \tau^{(h)}_{k} [x_t - (1 - \theta^{(h)}_{k}) \delta^{(h)}_{k}]}.
\]

2.3. Mixture of ZINB and EM algorithm

The zero-inflated negative binomial (ZINB) model is usually used for count data that exhibit both overdispersion and excess zero. Their PMF can be written as

\[
p(x_t|\phi) = \alpha 1_{(x_t=0)} + (1-\alpha) h(x_t|\phi),
\]
where \( h(x_t|\phi) \) is the PMF of NB distribution as expressed in (1).

The \( Q \) function consists of three parts

\[
Q(\varphi, \phi^{(h)}) = \sum_{t,k} \tau_k^{(t,h)} \log \omega_k + \sum_{t,h} \tau_k^{(t,h)} \left[ \sigma_k^{(t,h)} \log \alpha_k + \zeta_k^{(t,h)} \log (1 - \alpha_k) \right]
+ \sum_{t,k} \tau_k^{(t,h)} \zeta_k^{(t,h)} \log h(M_t, Y_t, Z_t|\phi_k)
\]

E-step:

\[
\tau_k^{(t,h)} \propto \omega_k^{(h)} p(x_t; \phi_k),
\]

\[
\sigma_k^{(t,h)} \propto \begin{cases} 
\alpha_k^{(h)} 1_{|x_t=0|}, & q = 1 \\
(1 - \alpha_k^{(h)}) h(x_t; \phi_k^{(h)}), & q = 2
\end{cases}
\]

\[
\delta_k^{(t,h)} = r_k^{(h)} \left[ \psi(r_k^{(h)}) + x_t - \psi(r_k^{(h)}) \right], \text{ with } r_k^{(h+1)} = -\frac{\lambda_k^{(h+1)}}{\log \theta_k^{(h+1)}},
\]

M-step:

\[
\omega_k^{(h+1)} = \sum_t \tau_k^{(t,h)}, \quad \sigma_k^{(h+1)} = \frac{\sum_t \tau_k^{(t,h)} \sigma_k^{(t,h)}}{\sum_t \tau_k^{(t,h)}}, \quad \theta_k^{(h+1)} = \frac{\beta_k^{(h)} \sum_t \tau_k^{(t,h)} \delta_k^{(t,h)}}{\sum_t \tau_k^{(t,h)} \zeta_k^{(t,h)} \left[ x_t - (1 - \beta_k^{(h)}) \delta_k^{(t,h)} \right]},
\]

3. Simulation

The process of drawing samples from the NB mixture model was achieved by the following two steps. The first step consists in determining the subpopulation of the drawn sample. We draw this subpopulation from the categorical distribution with the weight parameters \( \omega = (\omega_1, \ldots, \omega_k) \). In the second step, we draw the sample from the NB distribution with the parameters \((r, \theta)\) corresponding to the subpopulation selected in the first step.

Notice that NB distribution can be also parameterized by the mean and variance \((\mu, \sigma^2)\). For the purpose of having a more intuitive visualization, we set \((\mu, \sigma^2)\) for the different subpopulations. Correspondingly, the parameters size and rate \((r, \theta)\) can be obtained by \( r = \mu^2 / (\sigma^2 - \mu), \theta = \mu / \sigma^2 \).

The simulation studies were investigated to check the performance of our proposed algorithm in the terms of runtimes and classification performance. The words “closed form” represent our developed EM algorithm without the additional iterations in M-step. We performed the simulation schemes on a PC with CPU: Intel Core i5-8500 @ 3.00GHz × 6, Memory: 64 GB, System: Ubuntu 18.04.2 LTS, R: 3.6.0, RStudio: Version 1.2.1335.

3.1. Runtime

We measured the runtimes of closed form EM algorithm in comparison with a generalized EM algorithm in which the numerical solution in M-step is accomplished by Nelder-Mead simplex method. In order to make a fair platform of comparison, all the algorithms are programmed in R language. One thousand samples are generated with a set of fixed parameters: \( \omega = (0.2, 0.6, 0.2) \), \( \mu = (100, 200, 300) \), \( \sigma^2 = (1000, 1000, 1000) \). We repeat 100 times for this drawing samples scheme, the EM algorithms stop iterate till the precision reaches \(10^{-7}\). The results are shown in Figure 1:
It can be seen that, with the closed form approach, the running speed is obviously improved, about 3.5 times compared to Nelder-Mead method. We believe that is because the tedious iterations of numerical parameter optimizing steps of EM were simplified by the closed form EM algorithm.

3.2. Classification performance

The NB mixture model provides an important tool for the classification problems. Receiver Operating Characteristic (ROC) curve is commonly used to compare the performance of different classifiers. According to the ROC curves, we would check the classification performance of the closed form EM algorithm in comparison with the generalized EM algorithm with Nelder-Mead method. We repeat 100 times for drawing one thousand samples with a set of fixed parameters: \( \omega = (0.2,0.6,0.2), \mu = (100,200,300), \sigma^2 = (2000,2000,2000) \) and obtained the similar ROC curves for two approaches. Figure 2 shows one of results in respect to micro- and macro-average ROC curves.

![Figure 2. Micro- and macro-average ROC curves.](image)

From the perspective of micro- and macro-average ROC curves, the closed form EM method has the identical performance compared to Nelder-Mead EM method. This fact implies that the convergency of the proposed closed form EM algorithm is the same as the convergency of the other generalized EM algorithm for the mixture of NB model. The simulation results indicate that the closed form EM algorithm is an alternative method with higher efficiency.
4. Application
Earthquakes are a natural phenomenon which happen every day on our planet. Though predicting an earthquake is considered impossible, statistical number such as times of earthquakes happened in a year might have some pattern to be discovered. Here, the numbers of major earthquakes (M6.0+) in the world, 1900-2018, were downloaded from https://earthquake.usgs.gov/earthquakes/search/ on May, 2019.

Intuitively, the frequency of earthquake could be classified into three groups. Afterwards, we applied the closed form EM algorithm to estimate the parameters. With the precision of $10^{-7}$, our algorithm stops with the output: $\omega = (0.151,0.274,0.575)$, $\mu = (8.71,48.01,141.29)$, $\sigma^2 = (13.7,195.5,804.5)$. Clustering by maximum posterior probability, the frequency of each year was classified into one of three classes shown in Figure 3. Apparently, the major earthquakes happen more and more frequently, but there might be biases due to equipment level.

![Figure 3. Classification of years 1900-2018 by earthquake frequency level.](image)

Figure 4 shows the histogram of earthquake frequency and the fitted density estimated by closed form EM algorithm. The estimated probability density curve in Figure 4 illustrated a good fit by the mixture of the negative binomial model.

5. Discussion
The mixture model is widely used in the unsupervised classification problems. However, when estimating the parameters in the mixture of NB models, the additional iterative procedures are always embedded in the current EM algorithms. This fact aggravates the time-consuming feature of EM algorithm. For the purpose of reducing the runtimes of parameter estimations, we present an efficient EM algorithm for the mixture of NB models. We overcome the drawback of iterations which involves another iterative procedure by optimizing the fabricated likelihood instead of the classical likelihood. The simulation studies demonstrated that our developed EM algorithms effectively reduce the runtimes of parameter estimations in comparison with a typical generalized EM algorithm. Meanwhile, the simulation studies also show that our proposed algorithms have the same classification performance as the typical EM algorithm, it implies that our algorithm is convergence with the fabricated likelihood. Both simulated data and real data show that our proposed closed form EM algorithm could be applied as an alternative method with greater performance.

In addition to runtimes, we also studied the rate of convergence for the closed form EM algorithm. At each iterative EM step, we make a criterion to compare two EM algorithms by the following formula:
The Second International Conference on Physics, Mathematics and Statistics  
IOP Conf. Series: Journal of Physics: Conf. Series 1324 (2019) 012093  
doi:10.1088/1742-6596/1324/1/012093

\[ D^{(h)} = \sum_{k=1}^{K} \left( r_k^{(h)} - r_k^{(0)} \right)^2 + \left( \theta_k^{(h)} - \theta_k^{(0)} \right)^2 + \left( \omega_k^{(h)} - \omega_k^{(0)} \right)^2 \]

where \( r_k^{(0)}, \theta_k^{(0)}, \omega_k^{(0)} \) means the true value of parameters while drawing samples. Figure 5 shows the results related to the rate of convergence.

![Figure 5. Deviation of closed form and Nelder-Mead in each EM step.](image)

It is obvious that the closed form EM algorithm has a slower convergence rate. However, we recall that the runtimes of the closed form EM algorithm are much less than the generalized EM algorithm. This fact implies that the runtimes of EM algorithm for the mixture of NB model can also be reduced. For instance, the iterative schemes such as squared iterative methods [7] might be used to accelerate the closed form EM algorithm.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 11601286) and Beijing Municipal Natural Science Foundation (Grant No. 1182008).

References
[1] Li Q, Cassese A, Guindani M and Vannucci M 2018 Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data *Biometrics* 75(1) 183-192
[2] Wang Z 2011 One mixed negative binomial distribution with application *Journal of Statistical Planning and Inference* 141(3) 1153-1160
[3] Kagan Y 2017 Earthquake number forecasts testing *Geophysical Journal International* 211(1) 335-345
[4] Quenouille M H 1949 A relation between the logarithmic, Poisson, and negative binomial series *Biometrics* 5(2) 162-164
[5] Adamidis K 1999 Theory & Methods: An EM algorithm for estimating negative binomial parameters *Australian & New Zealand journal of statistics* 41(2) 213-221
[6] 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* 39(1) 1-38
[7] Varadhan R and Roland C 2008 Simple and globally convergent methods for accelerating the convergence of any EM algorithm *Scandinavian Journal of Statistics* 35(2) 335-353