Proceedings of the 2013 International Conference on Information, Operations Management and Statistics (ICIOMS2013), Kuala Lumpur, Malaysia, September 1-3, 2013

Fitting Com-Poisson Mixtures to Bimodal Count Data

Smarajit Bose[a,*], Galit Shmueli[b], Pragya Sur[a], Paromita Dubey[a]

[a]*Indian Staitsical Institute, Kolkata 700108, India*
[b]*Indian School of Business, Hyderabad 500032, India*

Abstract

Bi-modal truncated count distributions are frequently observed in aggregate surveys and ratings when respondents are mixed in their opinion. They also arise in censored count data, where the highest category might create an additional mode. The Poisson distribution is the most common distribution for fitting count data and can be modified to achieve mixtures of truncated Poisson distributions. However, it is suitable only for modeling equi-dispersed distributions and is limited in its ability to capture bi-modality. The Conway-Maxwell-Poisson (CMP) distribution is a two-parameter generalization of the Poisson distribution that allows for over- and under-dispersion. While the CMP is much more flexible, it still cannot capture bi-modality. In this work, we propose a mixture of CMPs for capturing a wide range of truncated count data, which can exhibit unimodal behavior (with equi-, under- or over-dispersion) as well as bimodal behaviour. We present methods for estimating the parameters of a mixture of two CMP distributions using an EM approach. Our approach introduces a special two-step optimization within the M-step to estimate multiple parameters. The methods are illustrated using simulated and real data.

Keywords: Mixture distributions; Conway-Maxwell-Poisson (CMP) distribution; EM algroithm.

1. Introduction and Motivation

Count data arise in many fields, including transportation, marketing, healthcare and biology among many others. The most commonly used distribution for modeling count data is the Poisson distribution. One of the major features of the Poisson distribution is that the mean and variance of the random variable are equal. However, in real life data often exhibit over- or under-dispersion. In such cases, the Poisson distribution often does not provide good approximations. For over-dispersed data, the negative Binomial model is a popular choice [2]. Other over-dispersion models include Poisson mixtures [3]. However, these models are not suitable for under-dispersion. A more flexible alternative that captures both over- and under-dispersion is the Conway-Maxwell-Poisson (CMP) distribution. The CMP is a two-parameter generalization of the Poisson distribution which also includes the Bernoulli and geometric distributions as special cases [4]. The CMP distribution has been used in a variety of count-data applications and has been extended methodologically in various directions (see a survey of CMP-based methods and applications in [5]).

* Corresponding author. Tel.: +91-33-2575-3412; fax: +91-33-2577-3104.
E-mail address: smarajit@isical.ac.in.

This paper is motivated by the need for a flexible distribution for modeling count data that arise in truncated environments, and in particular, where the empirical distributions exhibit bimodal behavior. One example is when only a censored version of the data is available. For example, when the data provider combines the highest values into a single "larger or equal to" bin, it often creates another mode at the last bin.

Real data in the above contexts can take a wide range of shapes, from symmetric to left- or right-skewed and from unimodal to bimodal. Count data arising from ratings or Likert-scale[†] questions exhibit bimodality when the respondents have mixed opinions. For example, respondents might have been asked to rate a certain product on a ten-point scale. If some respondents like the item considerably and others do not, we would find two modes in the resulting data.

In addition to bimodality, data from different groups of respondents might be under dispersed or over dispersed, due to various causes (e.g., dependence between responders' answers can cause over-dispersion). Under such setups, mixtures of two CMP distributions are likely to fit the data much better than a mixture of two Poisson distributions. While the CMP distribution has been the basis for various models, to the best of our knowledge, it was not extended to mixtures.

1.1. An Example

The Heritage Provider Network is a healthcare provider who launched a $3,000,000 contest with the following goal: "Identify patients who will be admitted to a hospital within the next year, using historical claims data."[‡] While the contest is much broader, for simplicity we look at one of the main outcome variables, i.e., the distribution of the number of days spent in the hospital (for claims received in a two-year period)[§]. The censoring at 15 days of hospitalization creates a second mode in the data, as can be seen in Figure 1.

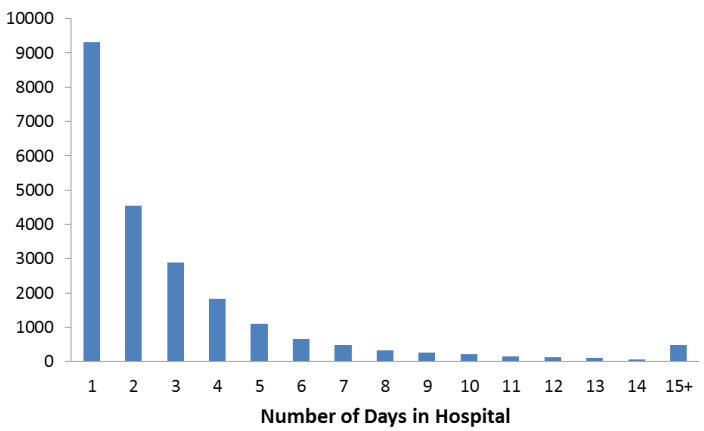

Figure 1: Distribution of numbers of days at the hospital. Data reported in censored form

[†]An example of a typical 5-point Likert scale is: strongly disagree, disagree, neutral (undecided), agree, strongly agree

[‡]http://www.heritagehealthprize.com

[§]We excluded zero counts which represent patients who were not admitted at all. The latter consist of nearly 125,000 records.

The remainder of the paper is organized as follows: In Section 2 we introduce a mixture of truncated CMP distributions for capturing bimodality, and describe the EM algorithm for estimating the five CMP mixture parameters. Section 3 illustrates our proposed methodology by applying it to simulated and real data examples. We conclude the paper with some discussions in Section 4.

2. A Mixture of Truncated CMP Distributions

2.1. The CMP Distribution

The Conway-Maxwell-Poisson distribution is a generalization of the Poisson distribution obtained by introducing an additional parameter ν which can take any non-negative real value and accounts for the cases of over and under dispersion in the data. The distribution was briefly introduced by Conway and Maxwell in 1962 for modeling queuing systems with state-dependent service rates. Non-Poisson data sets are commonly observed these days. Over-dispersion is often found in sales data, motor vehicle crashes counts, etc. Under-dispersion is often found in data on word length, airfreight breakages, etc. (see [5] for a survey of applications). The statistical properties of the CMP distribution, as well as methods for estimating its parameters were established by [4]. Various CMP-based models have since been published, including CMP regression models (classic and Bayesian approaches), cure-rate models, and more. The various methodological developments take advantage of the flexibility of the CMP distribution in capturing under- and over-dispersion, and applications have shown its usefulness in such cases. However, to the best of our knowledge, there has not been an attempt to fit bimodal count distributions using the CMP. The use of CMP mixtures is advantageous compared to Poisson mixtures, as it allows the combination of data with different dispersion levels with a resulting bimodal distribution.

If X is a random variable from a CMP distribution with parameters λ and ν, its pmf is given by

$$P(X = x) = \frac{\lambda^x}{x!^\nu}.\frac{1}{\sum_{j=0}^{\infty}\frac{\lambda^j}{j!^\nu}}, for\ \ x = 0,1,2,\ldots\ \ \lambda > 0\,, \nu \geq 0 \qquad (1)$$

Denote $(\lambda, \nu) = \sum_{j=0}^{\infty}\frac{\lambda^j}{j!^\nu}$. Some common features of this distribution are:

1. If ν < 1 successive ratios decrease at a slower rate compared to the Poisson distribution giving rise to a longer tail. This corresponds to the case of over-dispersion. The reverse occurs for the case of under-dispersion.
2. This distribution is a generalization of a number of discrete distributions:
    - For ν=0 and λ < 1, this is a geometric distribution with parameter (1-λ).
    - For ν=1, this is the Poisson distribution with parameter λ.
    - For ν → ∞, this is a Bernoulli distribution with parameter $\frac{\lambda}{1+\lambda}$.

We modify the CMP distribution to the truncated scenario considered in this paper. For data in the range t,t+1,t+2,…,T, we shift the distribution from 0 to t and then truncate values above T. For example, for data from a 10-point Likert scale, the truncated CMP pmf is given by:

$$P(X = x) = \frac{\lambda^x}{x!^\nu}.\frac{1}{\sum_{j=1}^{10}\frac{\lambda^j}{j!^\nu}},\ x = 1,2,\ldots,10\ ;\ \lambda > 0\,, \nu \geq 0 \qquad (2)$$

2.2. CMP Mixtures

The principal objective of this paper is to model bimodality in count data. Since both the Poisson and CMP can only capture unimodal distributions, for capturing bimodality we resort to mixtures. The standard technique for fitting a mixture distribution is to employ the Expectation-Maximization (EM) algorithm [1]. For example, in case of Poisson mixtures, one assumes that the underlying distribution is a mixture of two Poisson component distributions with unknown parameters while the mixing parameter p is also unknown. Further it is also assumed that there is a hidden variable with a Bernoulli(p) distribution, which determines from which component the data is coming from. Starting with some initial values of the unknown parameters, in the first step (E-step) of the algorithm, the conditional expectations of the missing hidden variables are calculated. Then in the second step (M-step), parameters are estimated by maximizing the full likelihood (where the values of the hidden variables are replaced with the expected values calculated in the E-step). Using these new estimates, the E-step is repeated, and iteratively both steps are continued until convergence.

Let the distribution of a random variable X be a mixture of CMP(λ_1, ν_1) and CMP(λ_2, ν_2) with probability p of being generated from the first CMP distribution. We also assume that each CMP is truncated to the interval [1,2,…, T]. If $f_1(x)$ and $f_2(x)$ are the pmfs of the two CMP distributions respectively, the pmf of X is

$$f(x) = pf_1(x) + (1-p)f_2(x) \;\; for\; x = 1,2,\dots,T \qquad (3)$$

If $X_1, X_2, \dots, X_n$ are iid random variables from the above mixture of two CMP distributions, their joint likelihood function is given by

$$L' = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \{pf_1(x_i) + (1-p)f_2(x_i)\}$$

$$\text{i.e.,} \quad \log L' = \sum_{i=1}^{n} \log \left\{ p.\frac{{\lambda_1}^{x_i}}{{x_i!}^{\nu_1}}.\frac{1}{\sum_{j=1}^{T}\frac{{\lambda_1}^{j}}{{j!}^{\nu_1}}} + (1-p).\frac{{\lambda_2}^{x_i}}{{x_i!}^{\nu_2}}.\frac{1}{\sum_{j=1}^{T}\frac{{\lambda_2}^{j}}{{j!}^{\nu_2}}} \right\} \qquad (4)$$

We would like to find the estimates $(\hat{p}, \widehat{\lambda_1}, \widehat{\nu_1}, \widehat{\lambda_2}, \widehat{\nu_2})$ by maximizing the likelihood function. However, due to the non-linear structure of the likelihood function, differentiating it with respect to each of the parameters and equating the partial derivatives to zero does not yield closed form solutions for any of the parameters. We therefore adapt an alternative procedure for representing the likelihood function.

Define a new set of random variables Y_i as follows:

$$Y_i = 1 \;\; if\; X_i \sim CMP(\lambda_1, \nu_1)$$
$$= 0 \; if\; X_i \sim CMP(\lambda_2, \nu_2)$$

Then the likelihood and log-likelihood functions can be written as

$$L = \prod_{i=1}^{n} \left\{ \left(pf_1(x_i)\right)^{y_i} \left((1-p)f_2(x_i)\right)^{(1-y_i)} \right\}$$

$$i.e., \;\; l = \log(L) = \sum_{i=1}^{n} y_i\{log(p) + logf_1(x_i)\} + \sum_{i=1}^{n}(1-y_i)\{log(1-p) + logf_2(x_i)\} \qquad (5)$$

From here we get a closed form solution for $\hat{p}$ by differentiation:

$$\frac{\delta l}{\delta p} = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The problem lies in the fact that the y_i's are unknown. We therefore use the EM algorithm technique.

E Step: Here we replace the y_i's with their conditional expected value

$$\widetilde{Y}_i := E(Y_i|X_i = x_i) = \frac{pf_1(x_i)}{pf_1(x_i)+ (1-p)f_2(x_i)}. \quad (6)$$

M Step: Thus, by replacing the unobserved y_i's in the E-step, we get

$$\hat{p} = \frac{\sum_{i=1}^{n} \widetilde{y}_i}{n}. \quad (7)$$

For the other parameters, none of the equations $\frac{\delta l}{\delta \lambda_1} = 0, \frac{\delta l}{\delta \nu_1} = 0, \frac{\delta l}{\delta \lambda_2} = 0, \frac{\delta l}{\delta \nu_2} = 0$ yield closed form solutions. We propose an iterative technique for obtaining the remaining estimates.

As an estimate of p is easy to obtain, we only maximize the likelihood over the remaining four parameters and then iterate. Plugging in the value of $\hat{p}$, the likelihood function L becomes a function of $\lambda_1, \nu_1, \lambda_2, \nu_2$.

The idea is to use the grid search technique to maximize L. In this technique, we divide the parameter space into a grid, evaluate the function at each grid point and find the grid point where the maximum is obtained. Then, a neighbourhood of this grid point is further divided into finer areas and the same procedure is repeated until convergence. We continue until the grid spacing is sufficiently small. Since we have four parameters to estimate, carrying out a grid search for all of them simultaneously is computationally infeasible. We therefore propose a two-step algorithm. First, we fix any two of the parameters at some initial value and carry out grid search for the remaining two. Then, fixing the values of these estimated parameters in the first step, we carry out grid search for the first two.

From simulation studies we observed that fixing the $\lambda's$ and obtaining $\hat{\nu}'s$ and then carrying out a grid search for estimating the λ's reduces the run time of the algorithm.

2.3. Model Estimation

If the empirical distribution exhibits a single peak, p is set to zero and a single CMP is estimated using ordinary maximum likelihood estimation [3] with adjustment for the truncation. Otherwise if the empirical distribution shows two peaks, we execute the following steps.

2.3.1. Initialization

Fit a Poisson mixture. If the resulting estimates of λ_1, λ_2 are sufficiently different, use these three estimates as the initial values for p, λ_1, and λ_2 and set the initial $\nu_1=\nu_2=1$.

If the estimated Poisson mixture fails to identify a mixture of different distributions, that is, when λ_1 and λ_2 are very close, then use the estimated p as the initial mixing probability, but initialize λ's by fixing them at the two peaks of the empirical distribution and set the initial $\nu_1=\nu_2=1$.

2.3.2. Iterations

After fixing the five parameters at initial values, the two-step optimization performs the following sequence:

For a given p,

1. Optimize the likelihood for ν's, fixing p, λ_1 and λ_2 using a grid search.
2. The optimal ν_1, ν_2 are then fixed (along with p). A grid search finds the optimal λ_1, λ_2.

3. Repeat steps 1 and 2 until some convergence stopping rule is reached.

Once the λ's and ν's are estimated, go back to estimate *p*.
Finally, the E step and M step are run until convergence.

3. Application to Simulated and Real Data

To illustrate and evaluate our CMP mixture approach and to compare it to simpler Poisson mixtures, we simulated bimodal discrete data over a truncated region, similar to the example of real data in Section 1.1.

3.1. Simulated Bimodal Data on 10-point Scale

This is an example of a mixture of two CMP distributions on a 10-point scale, one under-dispersed (λ_1=1, ν_1=3) and the other over-dispersed (λ_2=8, ν_2=0.7), with mixing parameter p=0.3. Figure 2 shows the empirical distribution for 100 observations simulated from this distribution. We see a mode at 1 and another at 10. We first fit a Poisson mixture, resulting in the fit shown in Table and Figure 2. As can be seen, the Poisson mixture properly captures the two modes, but their peak magnitudes are incorrectly flipped (thereby identifying the highest peak at 1). It also does not capture the single dip at 3, but rather estimates a longer dip throughout 3, 4 and 5. Finally, the estimated overall U-shape is also distorted.

We then fit a CMP mixture using the algorithm described in Section 4. The results are shown in Table and Figure 2. The fit appears satisfactory in terms of correctly capturing the two modes as well as the magnitudes of the peaks and dip. Note that the AIC statistic is very close to that from the Poisson mixture, yet the two models are visibly very different in terms of capturing modes, magnitudes and the overall shape.

Although the good fit of the CMP mixture might not be surprising (because the data were generated from a CMP mixture), it is reassuring that the algorithm converges to a solution with good fit. We also note that the estimated parameters are close to the generating parameters. Finally, we note that the runtime was about a minute.

Table :Simulated 10-point data (n=100) and expected counts from Poisson and CMP mixtures

Value	Simulated Data	Poisson Mixture	CMP Mixture
1	22	36	22
2	2	7	2
3	0	1	0
4	1	1	1
5	1	1	2
6	4	3	4
7	7	6	7
8	15	10	13
9	22	15	20
10	26	20	29
Estimates			
p	0.3	0.32	0.24
λ1, λ2	1,8	0.41, 13.58	1.13, 9.00
ν1, ν2	3, 0.7		3.75, 0.8
First Mode	1	1	1
Second Mode	10	10	10
AIC		370.6	370.0

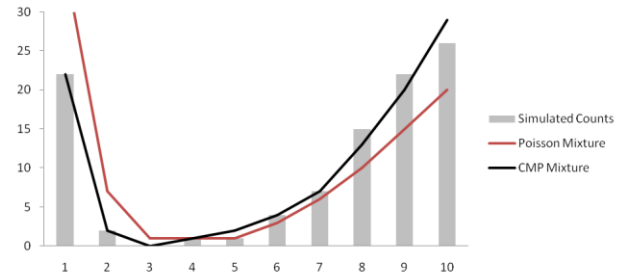

Figure 2: Fit of estimated Poisson mixture (p=0.3221, λ_1=0.4094, λ_2=13.5844) and CMP mixture (p=0.24, λ_1=1.1, ν_1=3.75, λ_2=9, ν_2=0.8)

3.2. Real Data Example

We return to the example from Section 1.1. The results of fitting a Poisson mixture and CMP mixture are shown in Table 2 and Figure 3.

Table 2: Observed and fitted counts for Heritage Insurance Competition data

# Days in hospital	Data	Poisson Mixture	CMP Mixture
1	9299	3284	7410
2	4548	2994	5567
3	2882	1860	3704
4	1819	976	2260
5	1093	641	1290
6	660	713	698
7	474	994	361
8	316	1327	183
9	263	1600	96
10	209	1742	62
11	145	1725	62
12	135	1566	89
13	111	1313	142
14	65	1021	227
15+	479	742	347
Estimates			
p		0.4132	0.96
λ_1, λ_2		1.8156, 10.8937	0.93, 13.4
ν_1, ν_2			0.3, 0.8
First Mode	1	1	1
Second Mode	15+	10	15+
Dip Location	14	5	10-11
Log likelihood		-5.6x10^4	-4.25x10^4
AIC		112006	85010

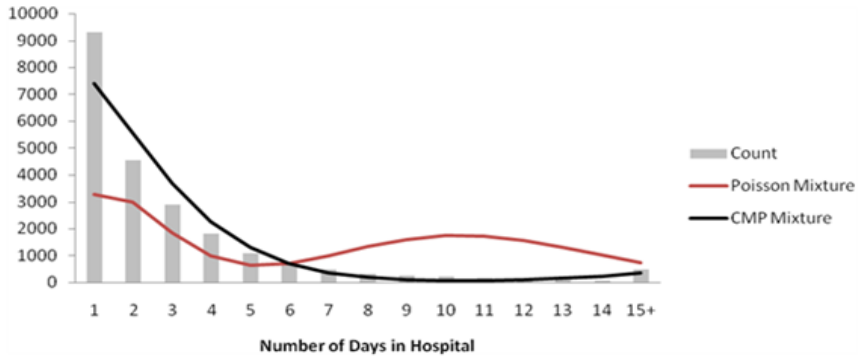

Figure 3: Observed and fitted Poisson and CMP mixture counts for Heritage Insurance Competition data

In this example, the two likelihood-based measures are very similar but the CMP fit is visibly much better. The CMP mixture correctly identifies the two modes and the magnitude of their frequencies. In contrast, the Poisson mixture not only misses the mode locations, but also the magnitude of inaccuracy for those frequencies is quite high.

4. Discussions

Discrete data often exhibit bimodality that is difficult to model with standard distributions. A natural choice would be a mixture of two (or more) Poisson distributions. However due to the presence of under- or over-dispersion often the Poisson mixture appears to be inadequate. The more general CMP distribution can capture under- or over-dispersion in the data. Therefore a mixture of CMP distributions (if necessary, properly truncated) may be appropriate to model such data.

The usual EM algorithm for fitting mixtures of distribution can be employed in this scenario. However, as the CMP distribution has an additional parameter (compared to the Poisson distribution), the maximization of the likelihood is nontrivial. In the absence of closed form solutions, iterative numerical algorithms are used for this purpose. An innovative two-step optimization with more than one possible initialization of the parameters has been suggested to ensure and speed up the convergence of the resulting algorithm. In our experiments, the proposed algorithm for fitting CMP mixture models take less than two minutes even for very large datasets (such as the Example 1.1: Heritage Competition dataset). Further reduction in runtime may be possible by invoking more efficient optimization techniques.

References

[1] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39, p. 1.
[2] Hilbe, J. M..2011. *Negative Binomial Regression*, 2nd edition, Cambridge University Press
[3] McLachlan, G. J., 1997. On the EM algorithm for overdispersed count data. Statistical Methods in Medical Research 6, p. 76.
[4] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., Boatwright, P., 2005 A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of The Royal Statistical Society, Series C 54, p. 127.
[5] Sellers, K. F., Borle, S., Shmueli, G., 2012 The CMP Model for Count Data: A Survey of Methods and Applications. Applied Stochastic Models in Business and Industry 28, p. 104.