Poisson mixture distribution analysis for North Carolina SIDS counts using information criteria

Tyler J. Massaro [1]

(1) Duke Clinical Research Institute

CORRESPONDING AUTHOR: *Tyler J. Massaro, Duke Clinical Research Institute - 2400 Pratt St., Box 4332 - Durham, NC 27705 - Office: (919) 668-8350 - Email: tyler.massaro@duke.edu*

DOI: ***10.2427/12550***
Accepted on June 17,2017

ABSTRACT

Aims: In this paper, we demonstrate the use of information criteria in a finite mixture of Poisson models framework to choose the optimal number of clusters in the North Carolina SIDS data (Symons, et al. 1983), a set of 100 overdispersed counts (mean = 6.67, var. = 60.55).
Methods: In addition to deriving information criteria with likelihood functions, we provide an empirical comparison between minimum Hellinger distance (MHD) estimation and EM estimation for finding parameters in a mixture of Poisson distributions with artificial data. This is further supplemented by an analysis of Bayes error in the context of classification problems with mixtures of 2, 3, 4, and 5 Poisson models.
Results: Our mixtures of Poisson distributions framework identified 4 naturally occurring clusters, each of which is nearly equidispersed. We determined that Robeson county has a suspiciously elevated number of counts, which we independently verified using 3 spatially explicit clustering methods from the literature. Through the examples with artificial data, we found that a combination of BIC with an EM parameter estimation procedure is best-suited for the proposed framework.
Conclusion: Using information criteria to select the optimal number of clusters from a finite mixture of Poisson distributions provides an effective, data-driven method for thresholding a set of counts into clusters in which the equivariance assumption of the Poisson model is upheld.

Key words: Finite mixture model; Poisson distribution; model selection; overdispersion; count data.

INTRODUCTION

The techniques described in this paper belong to the branch of machine learning that deals with unsupervised classification. The data have no group labels, although one may suspect that there is underlying group structure present due to heterogeneity within the population from which the data have been sampled. The primary goal is to develop tools, called mixture distribution models or finite mixture (FM) models, that provide group labels for these kinds of data. We are specifically interested in labeling count data since they are important for the fields of epidemiology and health care. Note, we will use the terms group, class, and cluster interchangeably.

Computational advances within the last 20 years have facilitated the emergence of FM modeling as one of the most popular ways to perform unsupervised classification tasks. The underlying principle behind FM modeling is that one treats data as having been sampled from a convex sum of distributions [2] [3] [4]. This idea reflects the

main assumption in FM modeling, which is that the data themselves come from a population that is segmented into homogeneous subpopulations [4].

The goal when we use FM modeling is to determine the true number of clusters that exist in a data set. We first estimate the maximum number of clusters that we suspect may exist, G_{max}, based on prior knowledge of the system from which the data are sampled. Alternatively, see [3] for heuristics on how to choose G_{max}. For each G = 1, …, G_{max}, we compute information criteria scores; the true number of clusters in the data set is the G for which the information criteria are minimized.

Statistical model selection via information criteria exists as an alternative approach to traditional p-value-driven statistics. The technique relies on the choice of an appropriate criterion that can be used to compare 2 or more candidate models. In this paper, we will be scoring AIC and BIC. Formal expressions of these criteria follow in Section 2.

Most algorithms that exist for parameter estimation of FM models are expectation-maximisation (EM) procedures. In addition to EM, we will use the minimum Hellinger distance (MHD) procedure to find parameter estimates. This technique was first described in [5], and extended for use on count data in [6]. The benefit of using a MHD procedure is that its parameter estimates are much more robust to contaminated data compared to parameter estimates from EM estimation [7] [8].

In this paper, we will utilize the class of minimum Hellinger distance algorithms popularized by Karlis and his collaborators [8] [9] [10] [11]. Although we did not use them in this paper, we also mention algorithms from Woo and Sriram (2007) who proposed an estimator based on minimizing their Hellinger Information Criterion [12]; and, from Umashanger and Sriram (2009) who described an "L_2E" estimator that minimized the L2-error between distributions [13].

The data we use to demonstrate our mixture distribution analysis framework come from a study that sought to identify counties at high risk of SIDS in North Carolina. The value assigned to each county represents the number of deaths in infants between 28 days and 1 year old attributed to SIDS in a 4-year span, from 1-Jul-1974 to 30-Jun-1978. The data were first introduced in their entirety in [1], and have often been used to demonstrate various clustering applications in the literature.

The remainder of this paper is structured as follows: in section 2, we derive formulas to compute the information criteria scores and describe an algorithm that can be used to select the optimal number of clusters; in section 3, we introduce the concept of Bayes error. In sections 4 and 5, we apply our algorithm to synthetic and real data using MATLAB 2016b. Sections 6 and 7 are for the discussion and conclusion.

METHODS

Observe that the joint pmf for n observations taken from a finite mixture of G Poisson distributions is given in [8] as

$$f(x;\, \pi_1, \ldots, \pi_G, \lambda_1, \ldots, \lambda_G) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \frac{\lambda_g^{x_i} e^{-\lambda_g}}{x_i!}. \quad (1)$$

The mean parameters within each component, λ_g, g = 1, …, G, contribute G parameters to the total overall number of parameters to be estimated, while the mixing proportions, π_g, g = 1, …, G, contribute G − 1 since $\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$. This gives a total of 2G − 1 estimated parameters.

Suppose we have estimated the set of parameters to find the model from equation (1) that best fits a given dataset. Let $\widehat{LL}$ be the maximized log-likelihood of equation (1), so that $\widehat{LL} = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} \left\{\hat{\pi}_g \hat{\lambda}_g^{x_i} e^{-\hat{\lambda}_g}/x_i!\right\}\right)$, where ^ denotes a parameter estimate. Then, the equations for expressing the information criteria are given below:

- AIC = $-2\,\widehat{LL} + 4G - 2$,
- BIC = $-2\widehat{LL} + (2G - 1)(\log n + 1)$.

For each G = 1, …, Gmax, where we have predetermined a suitable Gmax, we compute EM and/or MHD parameter estimates to score AIC and BIC. The G for which one, the other, or both of these information criteria is minimized is deemed the optimal number of groups for the given set of data.

When using FM modeling to sort data, observe that we assign observation i to group g = 1, …, G, which we denote by γ_i = g, if the following holds:

$$\gamma_i = \arg\max_{h \in 1,\ldots,G} \frac{\pi_h \frac{\lambda_h^{x_i} e^{-\lambda_h}}{x_i!}}{\sum_{g=1}^{G} \pi_g \frac{\lambda_g^{x_i} e^{-\lambda_g}}{x_i!}},$$

That is, observation i belongs to the group where its posterior probability is maximised. See [2] for more information on this concept.

Bayes error estimation

The Bayes error is the smallest error rate that a classifier, e.g. a finite mixture of Poisson distributions, may achieve [14]. Let C_g refer to the g-th component in equation (1), and define $H_g \subseteq \mathbb{Z}^+ \cup \{0\}$ as the region in which the FM model classifies observations to Cg. From [14], the Bayes error, EB, is precisely

$$E_B = \sum_{C_g \neq C_G} \sum_{x \in H_g} \pi_g \frac{\lambda_g^x e^{-\lambda_g}}{x!}.$$

It is much easier to visualize Bayes error. Suppose we have an artificial mixture of 2 Poisson distributions whose pmf is $f(x;\pi,\lambda) = 0.4f_1(x;\lambda_1 = 7) + 0.6f_2(x;\lambda_2 = 20)$, where f1 and f2 refer to the usual Poisson model for the given mean parameters. The weighted versions of f1 and f2 are shown in Figure 1, not f itself. In particular, the red and blue regions in Figure 1 correspond to C1 and C2, respectively, as they were defined in the preceding paragraph.

Where C1 and C2 overlap, in violet, is the region where observations from C1 are misclassified as belonging to C2 and vice versa. The area of this region is the Bayes error. Hence, if we used the model described by f to classify a data set for which we know the true group labels, the minimum possible misclassification rate we could achieve is 3.42%.

Artificial data example

We constructed separate examples using artificial data to separately (1) compare MHD and EM parameter estimates for Poisson mixture models; and, (2) assess the component selection accuracy of models fit using MHD and EM parameter estimation.

Experiments 1 and 2

In our first example, we were interested in estimating how closely parameter estimates, recovered from random samples taken from a mixture of 2 Poisson distributions, could estimate the true Bayes error of the distribution. For each δ from 1 to 15 we randomly sampled 1000 values from a Poisson distribution with mean parameter $\lambda_1^{(1)} = 1$, and another 1000 values from a Poisson distribution with mean parameter $\lambda_2^{(1)} = \lambda_1^{(1)} + \delta$ (note: the superscripts denote experiment). We recorded the true Bayes error for the pair $(\lambda_1^{(1)}, \lambda_2^{(1)})$, then recorded the Bayes error for the pair of parameters estimated from the artificial data using EM and MHD. We repeated this experiment 1,000 times for each δ.

The second example was set up in the same way as the first, except this time we used samples taken from a Poisson distribution with mean parameter $\lambda_1^{(2)} = 201$, and a separate Poisson distribution with parameter $\lambda_2^{(2)} = \lambda_1^{(2)} + \delta$, for $\delta = 1, \ldots, 100$. In doing so, we hoped to demonstrate the importance of data scale when considering Bayes error within our proposed framework.

Figure 2 shows the true and estimated Bayes error rates for experiments 1 and 2. Notice in the left-hand pane how the estimates are nearly identical. Further, the error rate is essentially negligible once the difference in mean parameters is 15. Contrast this with the results from the second experiment in the right-hand pane: we are still able to achieve negligible error rates but only once the distributions are much further apart, corresponding to $\lambda_1^{(2)} = 201$ and $\lambda_2^{(2)} = 301$.

FIGURE 1. Graphic showing the pmfs of the Poisson component models comprising the mixture model referred to as f in the text of section 3. In red, is the region corresponding to the first component, C1, and in blue is the region corresponding to the second component, C2. Where the two overlap, in violet, observations in C1 are misclassified as belonging to C2, and vice versa. The area of this region is the Bayes error, equal here to approximately 0.0342.

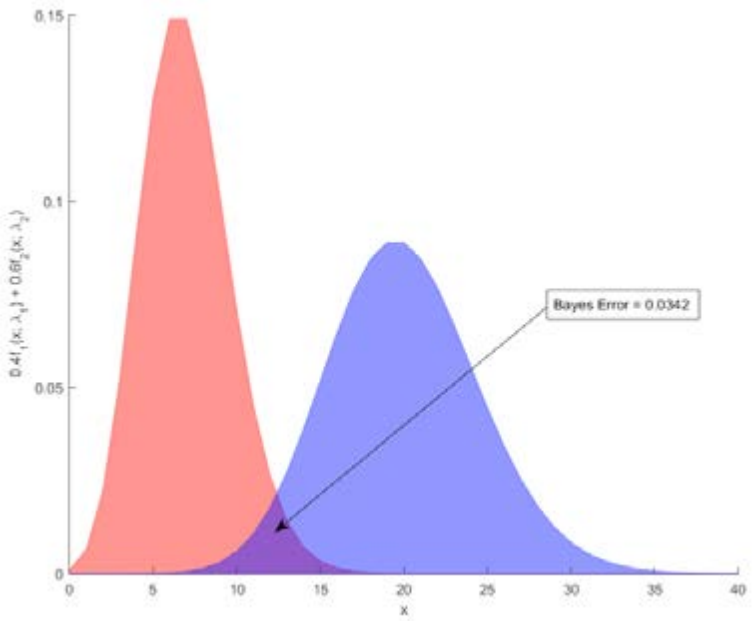
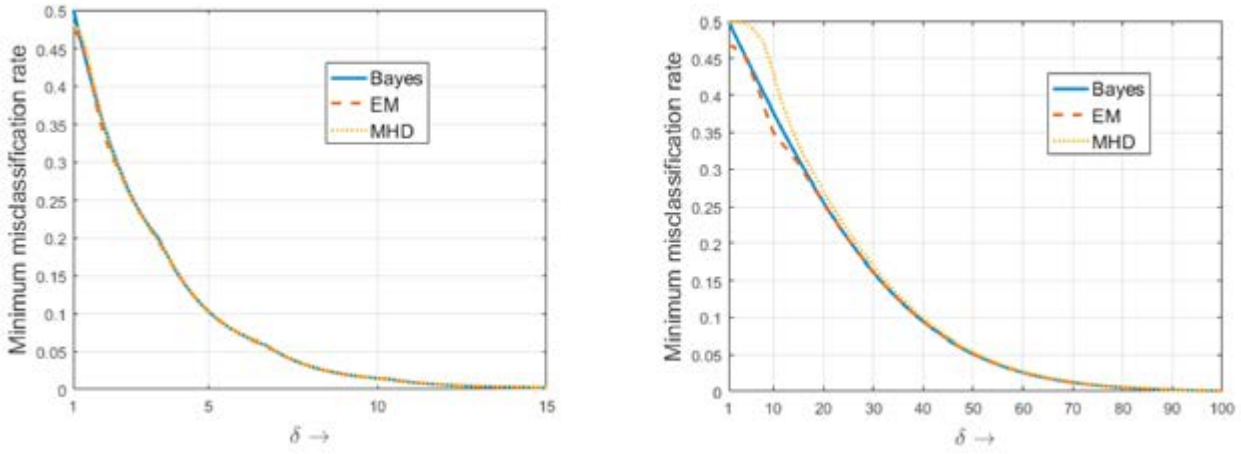


FIGURE 2. Minimum misclassification rates from experiments 1 and 2 as described above, where λ_1 =1 or 201, and $\lambda_2 = \lambda_1 + \delta$. In the left-hand pane, the estimated Bayes error rates from EM and MHD parameter estimation are nearly identical to the true error, this error is mostly negligible once samples are being drawn from distributions with mean parameters at 1 and 15. In the right-hand pane, the estimated errors from EM and MHD converge to the true error at different values of δ. Moreover, we achieve near-perfect classification again, but only once the mean parameters are located at 201 and 300.



Experiments 3, 4, and 5

The goal of experiments 3, 4, and 5 was to demonstrate using the information criterion framework to choose the optimal number of mixture components from artificial datasets where we knew the true distribution. A total of 1000 samples were drawn 1,000 times from the densities summarized below in Table 1. The performance of our method with respect to MHD and EM parameter estimation, and with respect to AIC and BIC, is summarized in Table 2. Additionally, Figure 3 shows the estimated Bayes errors for these experiments.

In every example, the average optimal number of groups chosen by BIC is closer to the truth than that chosen by AIC, suggesting that this score is less susceptible to model over- or under-fitting in the context of finite mixtures of Poisson distributions. With respect to parameter estimation, EM estimates lead to more accurate classifications. This tendency is most overt when examining the results from experiments 4 and 5: in particular, the parameter estimates from MHD seem to encourage model overfitting.

The estimated Bayes error rates shown in Figure 3 are consistent with the results in Table 2 in that we see better performance from EM estimates than from MHD. The Bayes errors computed using EM parameter estimates are, on average, closest to the true Bayes error (the solid black line) for each experiment.

Note that for each of the 1000 trials there is a corresponding Bayes error estimate, however we constructed these boxplots using only the trials where the true number of groups was identified; going from experiment 3 to 5, this means we used 989, 983, and 736 out of 1000 trials each for EM, and 991, 170, and 204 out of 1000 trials each for MHD. Despite this, the error rates for MHD estimates in experiment 5 are still much higher than the true error rate.

Figure 3: Boxplots of the estimated Bayes error rates for experiments 3, 4, and 5 as described above, with the true Bayes error of the model used in each experiment represented by the solid black line. We only used trials where BIC correctly identified the true number of groups to create the boxplots; from experiment 3 to 5, that is 989, 983, and 736 out of 1000 for EM, and 991, 170, and 204 out of 1000 for MHD.

NORTH CAROLINA SIDS DATA EXAMPLE

Classifying North Carolina counties by SIDS count

Our example with real data looks at sudden infant death syndrome (SIDS) cases in North Carolina counties between 1-Jul-1974 and 30-Jun-1978. These data were originally published by Symons et al [1]. They used mixtures of Poisson distributions to sort the rate data into 2 groups corresponding to normal- and high-risk counties. No consideration was given to fitting more than these 2 groups.

The data are briefly summarized in Table 3. The smallest count was 0, seen in 13 counties, up to a maximum count of 44. The variance is one order of magnitude higher than the mean, suggesting the presence of real overdispersion that may be caused by clustering. In support of this, Dean's test [15] for overdispersion is significant (P_B = 7.23, p-value << 0.0001). Zero-generation was restricted only to SIDS observations, so there was no need to consider a zero-inflated model.

To test for heterogeneity in the data, we fit mixture models of 1 up to G_{max} = 5 components using the EM estimates. As we did this, we compared the information criteria scores.

TABLE 1. True mixture model densities that were used for drawing artificial data samples in Experiments 3, 4, and 5.

Experiment	Density
3	$0.4f_1(x;\lambda_1 = 100) + 0.5f_2(x;\lambda_2 = 10) + 0.1f_3(x;\lambda_3 = 120)$
4	$0.2f_1(x;\lambda_1 = 5) + 0.6f_2(x;\lambda_2 = 15) + 0.15f_3(x;\lambda_3 = 50) + 0.05f_4(x;\lambda_4 = 75$
5	$0.1f_1(x;\lambda_1 = 2) + 0.15f_2(x;\lambda_2 = 12) + 0.3f_3(x;\lambda_3 = 25) + 0.05f_4(x;\lambda_4 = 40$ $0.40f_5(x;\lambda_5 = 80)$

TABLE 2. Summary of results from experiments 3, 4, and 5, carried out as described above. The frequency with which the number of components are chosen is shown for each information criterion score, and for parameters estimated using MHD and EM.

EXPERIMENT 3: TRUE G = 3

		EM			MHD	
G =		**AIC**	**BIC**		**AIC**	**BIC**
2		0	0		0	1
3		883	989		978	991
4		117	11		22	8
Avg. (s.d.)		3.12 (0.32)	3.01 (0.10)		3.02 (0.15)	3.01 (0.09)

EXPERIMENT 4: TRUE G = 4

		EM			MHD	
G =		**AIC**	**BIC**		**AIC**	**BIC**
3		0	0		217	236
4		880	983		180	170
5		120	17		603	594
Avg. (s.d.)		4.12 (0.33)	4.02 (0.13)		4.39 (0.82)	4.36 (0.84)

EXPERIMENT 5: TRUE G = 5

		EM			MHD	
G =		**AIC**	**BIC**		**AIC**	**BIC**
4		103	106		123	136
5		736	838		196	204
6		161	56		681	660
Avg. (s.d.)		5.06 (0.51)	4.95 (0.40)		5.56 (0.70)	5.52 (0.72)

RESULTS

Table 4 shows the information criteria scores for mixture models fitting up to 5 components to the NC SIDS data. Both AIC and BIC criteria select G = 4 as the optimal number of groups.

We show the summary statistics for each group in the G = 4 model in Table 5. Observe that groups 1, 2, and 3 show near equidispersion. In group 4, the variance is still higher than the mean suggesting some overdispersion is still present. However, the amount of excess variation within the component is substantially less than what was originally observed, and is not significant according to (Dean's test (; P_B = 1.44, p-value = 0.08).

FIGURE 3. Boxplots of the estimated Bayes error rates for experiments 3, 4, and 5 as described above, with the true Bayes error of the model used in each experiment represented by the solid black line. We only used trials where BIC correctly identified the true number of groups to create the boxplots; from experiment 3 to 5, that is 989, 983, and 736 out of 1000 for EM, and 991, 170, and 204 out of 1000 for MHD.

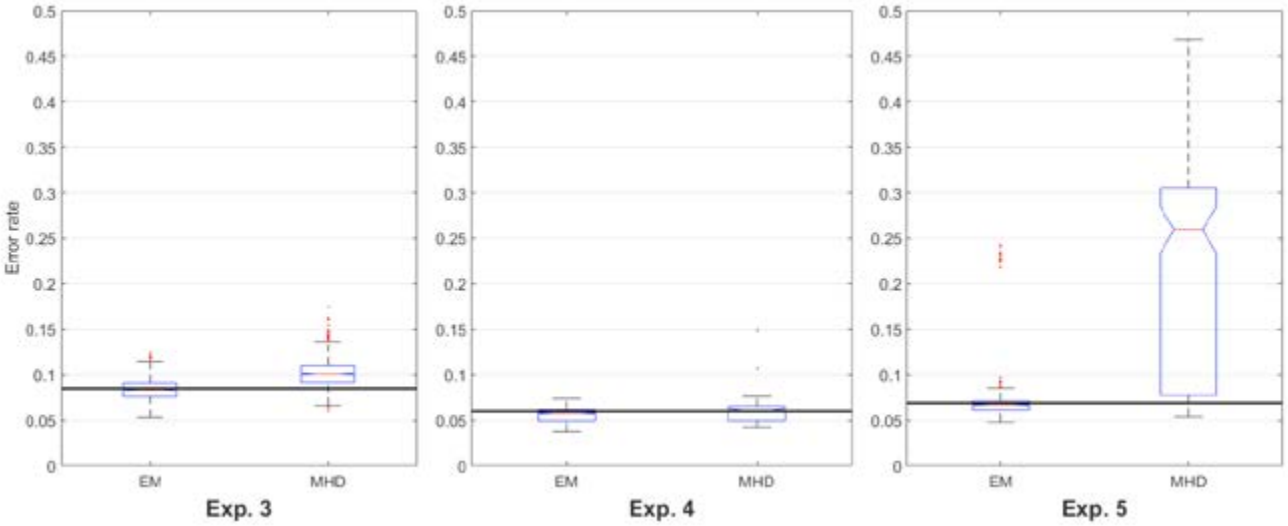


TABLE 3. Summary statistics for the North Carolina SIDS data, published by Symons et al. (1983).

No. obs.	100
Min.	0
Max.	44
Mean	6.67
Var.	60.55

TABLE 4. Information criteria scores for finite mixtures of up to 5 Poisson models fit to the NC SIDS data. Notice that both AIC and BIC are minimized for G = 4.

G	AIC	BIC
1	1014.29	1016.89
2	684.77	689.98
3	617.93	625.75
4	594.08	604.50
5	595.92	608.95

TABLE 5. Summary statistics for each component in the mixture of 4 Poisson models chosen by the information criteria to model the NC SIDS data.

	C_1	C_2	C_3	C_4
No. obs.	24	51	21	4
Min.	0	2	9	29
Max.	1	8	23	44
Mean	0.46	4.57	13.38	35.50
Var.	0.26	3.29	13.05	47.00

FIGURE 4. Map of North Carolina counties colorized by component ID from a finite mixture of G = 4 Poisson models. Notice that most of the higher SIDS counts are associated with counties where a populous city is located. The 2 counties without a city shown are Robeson (southwest of Fayetteville) and Onslow (southeast of Fayetteville).

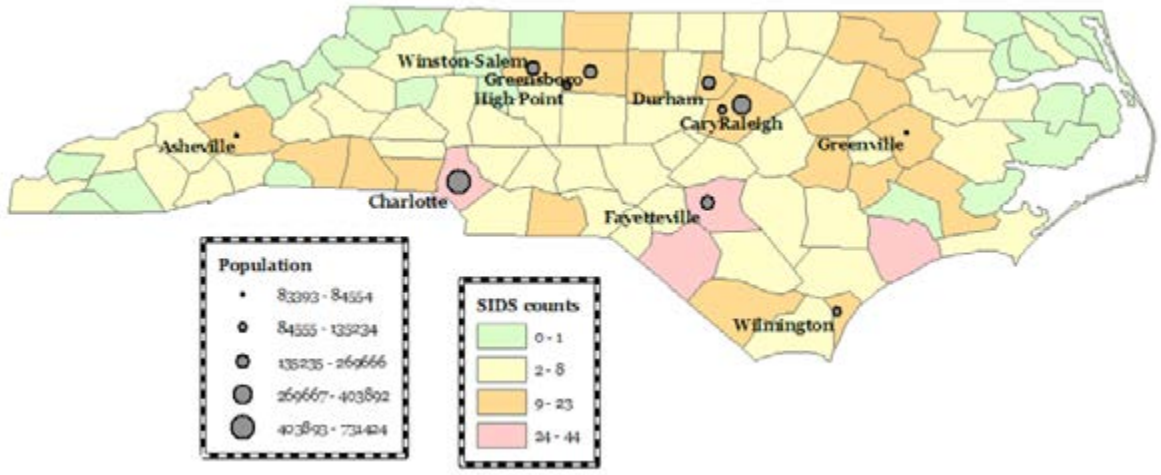
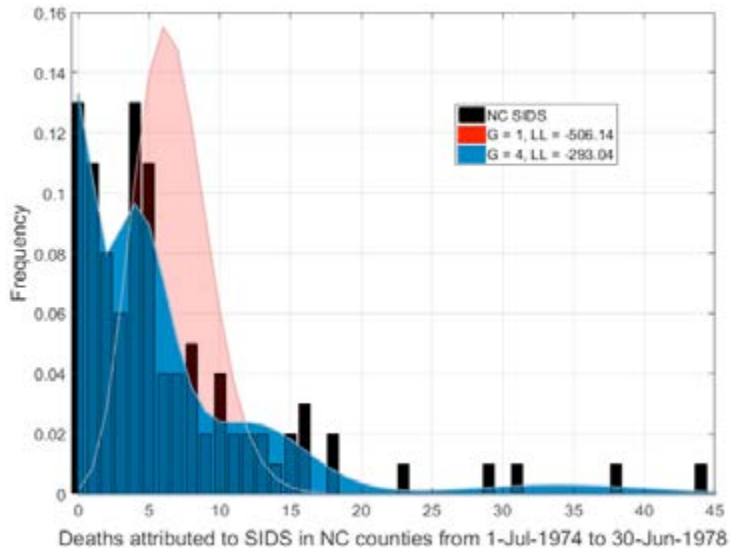


FIGURE 5. The black bars show the frequencies of observed counts in the North Carolina SIDS data set. We overlaid the estimated finite mixture model pmfs for G = 1 and 4 components, and show their respective log-likelihood values in the legend.



We provided a map of North Carolina counties with the counties colored by component ID in Figure 4 using ArcMap 10.4.1. Green counties contain the lowest counts (either 0 or 1), and these are traditionally rural regions of North Carolina. Meanwhile, orange and red counties are medium-high and high counts, and we see that these generally correspond to larger metropolitan areas. The top 11 cities in North Carolina by population as of the 2010 U.S. Census are also shown on the map [16], and each of these cities is in an orange or red county.

In Figure 5, we created a bar chart of the SIDS data with the finite mixture model pmf overlaid for the G = 1 and 4 models. The model that was chosen by the information criteria, G = 4, is in blue. We see that the first two peaks of the G = 4 model follow the peaks in the data quite closely; the tail behavior of this model is also much heavier than the G = 1 model, and the G = 2 and 3 models (not shown). The log-likelihood values, which we use to estimate the lack-of-fit portions of AIC and BIC, heavily favors the G = 4 model.

DISCUSSION

One general criticism of the Symons paper is with how they handled the mixtures of Poisson distributions problem. In brief, they found a p-value corresponding to a null hypothesis that the rates of SIDS deaths per 1000 live births were all sampled from the same distribution. Once

they determined that this p-value was sufficiently small, they chose to accept an alternative hypothesis that the rate data must have been sampled from 2 separate distributions instead of investigating other integer values.

Regardless of whether theirs was the correct decision, when we fit the same rate data to mixtures of up to 5 exponential distributions, G = 1 was chosen. Therefore, at least based on the rate data they were using, there is no reason to suspect underlying heterogeneity if we treat the rates as being sampled from exponentials. That they used Poisson distributions and then normalized based on number of live births, when perhaps modeling the rates as exponential data would have been more appropriate, is our other main criticism.

Another way of handling count data such as these would be to use a negative binomial distribution. However, we fit mixtures of up to 5 negative binomial distributions and discovered that G = 1 was chosen, despite the clear multimodal behavior in Figure 5Figure 4. The negative binomial distribution is undoubtedly useful for modeling overdispersed count data, howeverthough its tendency to oversmooth heterogeneous data is in fact a detriment in the context of this problem and the NC SIDS datahese NC SIDS data.

Out of 100 counties, Symons et al. identified 15 as high-risk, and the rest as normal. Of these 15 counties, we identified only 7 of these as either high or medium-high risk in the model: Anson, Northampton, Halifax, Columbus, Rutherford, Robeson, and Rockingham (listed in order of decreasing rate). In fact, these 7 counties found in common with the Symons approach were the 7 counties with a high rate and a SIDS count of at least 9 or higher in the 4-year window.

This reflects one of the major limitations of our approach, which is that we do not in any way normalize when using count data in a mixtures of Poisson distributions setting. It is vitally important to understand this beforehand. Essentially, using this tool allows one to perform data-driven thresholding in lieu of any subjective alternative.

Given this drawback, there is still much information to be gained in this and other similar types of problems involving count data. For example, of the 4 counties in red corresponding to the highest SIDS counts, 2 of these counties do not have a highly populous city highlighted. One of these is Robeson county (located to the southwest of Fayetteville, NC), which has been identified as a majority-minority county [17]. As of the 2010 U.S. Census, only 29.0% of respondents identified as white; 24.3% were black or African American, 38.4% were American Indian or Alaska Native, and 8.1% were Hispanic or Latino [16]. Each of these minority demographics has been shown to have significant barriers to health care coverage.

Another limitation to point out is that our approach does not incorporate any spatial dependence or adjacency like the class of spatially-explicit clustering methods. Moran's I [18], which tests for the presence of spatial autocorrelation, is significant (I = 0.24, p-value = 0.02). Thus, there is evidence which suggests spatial dependence in these data, though we did not consider it in our framework.

We ran Besag-Newell [19], Openshaw's Geographical Analysis Machine (GAM) [20], and Kuldorff-Nagarwalla [21] tests in R 3.3.1 using DCluster [22] to check for the presence of clustering. In general, each of these tests looks for a group of adjacent counties where the collective number of observed counts is higher than expected, taking into consideration different factors depending on which of the tests was chosen.

Openshaw's and Kuldorff and Nagarwalla's tests both identified a cluster of counties including and surrounding Robeson county; Openshaw's GAM identified 2 additional clusters, one near Anson county (the orange county due east of Charlotte in Figure 4) and the other near Northampton and Halifax counties (orange counties in northeast NC, along the border with Virginia in Figure 4). Besag-Newell identified a cluster consisting of only Columbus county, which is due south of Robeson county, plus an additional cluster of Northampton and Halifax counties. Finally, we determined that Stone's statistic [23] for Robeson county was significant (T = 2.08, p-value = 0.01); this tested the null hypothesis that relative risks are constant as distances to the centroid of Robeson county increase, with the alternative that relative risks decrease with increasing distance. Although we did not explicitly account for spatial dependence with our approach, we are able to obtain similar information.

It is important to point out that the MHD procedures, which have been previously advocated due to being less susceptible to outliers, seemed to struggle in our framework. In fairness, none of the experiments we designed was set up to validate claims related to this point, so we will not draw any conclusions beyond stating that an EM procedure for parameter estimation appears to be a better match for our modeling framework than MHD.

Keeping this in mind, one final observation comes from the results of experiment 2. Given that a Poisson distribution with sufficiently large mean parameter can be approximated by a normal, it is reasonable to assume that the data being generated from each Poisson distribution would significantly overlap for smaller mean parameter values. In these instances, estimates that are less susceptible to extreme values (i.e., MHD estimates) would tend to be closer together and therefore lead to increased Bayes error estimates, while we would expect the EM estimates to be further apart and lead to decreased Bayes error estimates. This kind of behavior is evident in the right-hand pane of Figure 2, as the estimated Bayes error appears lower for the EM estimates and higher for the MHD estimates.

CONCLUSION

In this paper, we demonstrated how to use model selection theory to choose the optimal number of

components in a finite mixture of univariate Poisson distributions. Using the North Carolina SIDS data set, we provided objective results based on the data rather than imposing any kind of subjective restraints. While our results demonstrated that higher counts will be associated with areas of higher population, we were nonetheless able to implicate 2 counties. One of these, Robeson county, was independently implicated in 3 separate spatially-explicit clustering tests that we ran.

Through a series of experiments with artificial data, we showed that an expectation-maximisation procedure for parameter estimation appears better-suited for integration within our modeling framework versus a minimum Hellinger distance procedure. Further, we showed the importance of data scale in achieving low misclassification rates.

Future research into this problem will investigate the role of data scale within the context of the finite mixtures of Poisson distributions problem. It will be interesting evaluating to what extent classification rates can be improved by shifting count data closer to 0.

References

1. Symons, M. J., R. C. Grimson, Y. C. Yuan, "Clustering of Rare Events," Biometrics, vol. 39, no. 1, 193-205, 1983.
2. Titterington, D. M., A. F. M. Smith, U. E. Makov. Statistical Analysis of Finite Mixture Distributions, Chichester, GBR: John Wiley & Sons Ltd., 1985.
3. Bozdogan, H., "Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity," in Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Information Approach (Vol. 2), H. Bozdogan, Ed., Dordrecht, NED, Kluwer Academic Publishers, 1994, 69-113.
4. Marin, J.-M., K. Mengersen, and C. Robert, "Bayesian Modelling and Inference on Mixtures of Distributions," in Handbook of Statistics (Vol. 25), C. Rao, Ed., NY, Springer-Verlag, 2005.
5. Beran, R., "Minimum Hellinger Distance Estimates For Parametric Models," The Annals of Statistics, vol. 5, no. 3, 445-463, 1977.
6. Simpson, D. G., "Minimum Hellinger Distance Estimation for the Analysis of Count Data," Journal of the American Statistical Association, vol. 82, no. 399, 802-807, 1987.
7. Lindsay, B. G., "Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods," The Annals of Statistics, vol. 22, no. 2, 1081-1114, 1994.
8. Karlis, D. and E. Xekalaki, "Minimum Hellinger distance estimation for Poisson mixtures," Computational Statistics & Data Analysis, vol. 29, 81-103, 1998.
9. Karlis, D. and E. Xekalaki, "Robust inference for finite poisson mixtures," Journal of Statistical Planning and Inference, vol. 93, 93-115, 2001.
10. Karlis, D., "An EM algorithm for multivariate Poisson distribution and related models," Journal of Applied Statistics, vol. 30, no. 1, 63-77, 2003.
11. Karlis, D. and L. Mligkotsidou, "Finite mixtures of multivariate Poisson distributions with application," Journal of Statistical Planning and Inference, vol. 137, 1942-1960, 2007.
12. Woo, M.-J., and T. N. Sriram, "Robust estimation of mixture complexity for count data," Computational Statistics & Data Analysis, vol. 51, 4379-4392, 2007.
13. Umashanger, T., and T. N. Sriram, "L2E estimation of mixture complexity for count data," Computational Statistics & Data Analysis, vol. 53, 4243-4254, 2009.
14. Tumer, K. and J. Ghosh, "Estimating the Bayes Error Rate through Classifier Combining," Proceedings of the International Conference on Pattern Recognition, 695-699, 1996.
15. Dean, C. B., "Testing for Overdispersion in Poisson and Binomial Regression Models," Journal of the American Statistical Association, vol. 87, no. 418, 451-457, 1992.
16. United States Census Bureau, "American FactFinder," [Online]. Available: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml#. [Accessed 7 May 2017].
17. Pollard, K. and M. Mather, "Population Reference Bureau," 2008. [Online]. Available: http://www.prb.org/Publications/Articles/2008/majority-minority.aspx. [Accessed 7 May 2017].
18. Moran, P. A. P., "The interpretation of statistical maps," Journal of the Royal Statistical Society, Series B, vol. 10, no. 2, 243-251, 1948.
19. Besag, J. and J. Newell, "The Detection of Clusters in Rare Diseases," Journal of the Royal Statistical Society, Series A, vol. 154, no. 1, 143-155, 1991.
20. Openshaw, S., M. Charlton, C. Wymer, and A. Craft, "A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets," International Journal of Geographical Information Systems, vol. 1, 335-358, 1987.
21. Kulldorff, M. and N. Nagarwalla, "Spatial disease clusters: detection and inference," Statistics in Medicine, vol. 14, 799-910, 1995.
22. Gomez-Rubio, V., J. Ferrandiz-Ferragud, and A. Lopez-Quilez, "Detecting clusters of disease with R," Journal of Geographic Systems, vol. 7, no. 2, 189-206, 2005.
23. Stone, R. A., "Investigations of excess environmental risks around putative sources: statistical problems and a proposed test," Statistics in Medicine, vol. 7, 649-660, 1988.