INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)

EMPIRICAL PREDICTION OF LIBRARY CIRCULATIONS BASED ON NEGATIVE BINOMIAL PROCESSES

Quentin L. BURRELL

Statistical Laboratory, Department of Mathematics,
University of Manchester,
Oxford Road, Manchester M13 9PL, United Kingdom

Abstract

The empirical method for prediction of library circulations in the presence of ageing recently proposed by Burrell is extended to cover mixtures of negative binomial processes. This makes the method more flexible but requires assumptions about an unknown parameter. Predictions using this method are compared with the earlier ones based on mixtures of Poisson processes.

§ 1. INTRODUCTION

Recent work has suggested, on the one hand, that the "mixture of Poisson processes" is invalid as a mathematical model for library circulations ([1,2]), but on the other ([3]) that it can provide estimates of future performance which are sufficiently accurate that they deserve consideration by library managers. In this note we show how to extend the "empirical Bayes method" described by Burrell [4] to mixtures of negative binomial processes which have been advocated in [2] as the most appropriate simple general model for library circulations. Once again we apply the method to the University of Saskatchewan data originally presented by Tague and her colleagues [5,6] in order to compare it with earlier work.

§ 2. THE RESULT

The context of interest is the usage, as measured by circulations or checked-out borrowings, of a library collection when ageing is present. A detailed description of mathematical modelling of this situation has been given by Burrell [7,8,9] and we shall not repeat the background here. The empirical approach in [4] is based on the assumption that individual items are borrowed according to a Poisson process, different items having different (unknown) rates. However, by considering the correlation structure of bivariate loan distributions, Burrell [2] has suggested that negative binomial (NB) processes might be more realistic. Parametric mixtures of such processes have previously been applied to bibliometric problems by the author [10,11] but here we make no assumptions regarding the form of the mixing distribution.

Suppose that we have a collection of N items available for loan and that the items of borrowings of the items may be modelled by a NB process of index ν. Thus if X_t denotes the number of times that an item is borrowed in a period of effective length t (see [4] for a discussion of the notion of effective time) then $X_t = NB(\nu t,p)$, i.e. :

$$P(X_t = r) = \binom{r+\nu t-1}{r} p^{\nu t}(1-p)^r \quad , \text{ for } r = 0,1,2,\ldots$$

(Note that this should not be confused with the gamma-poisson proces [3,10] which also has a negative binomial form but has a very different time-dependence). The index ν is assumed to be a characteristic of the collection while different items will in general have different, and unknown, values of p, say $p_1, p_2, \ldots, p_N$. The only restriction on the p's is that $0 < p \leq 1$, where the case $p = 1$ corresponds to a "dead" item, i.e. one which is never, or can never be, borrowed.

As in [4] we suppose that our data are just the observed circulation frequencies during the period [0,1] :

$$f_r = \text{number of items circulating r times, for } r = 0,1,2,\ldots,$$

and we wish to predict the number of items circulating r times during a future period of effective length $t < 1$. Letting X and Y denote the number of times an item is borrowed during [0,1] and the later period respectively, what we wish to estimate is therefore

$$\gamma_r = E[\#\text{ of items for which } Y = r] \ .$$

In the Appendix we prove the following, which is completely analogous to the Theorem based on Poisson processes proved in [4] :

Theorem : Unbiased estimates of γ_r, $r = 0,1,2,\ldots$, the predicted frequencies for the Y-distribution, are given by

$$g_r = \sum_{n=r}^{M} \frac{\binom{r+\nu t-1}{r}\binom{n-r+\nu(1-t)-1}{n-r}}{\binom{n+\nu-1}{n}} f_n, \text{ for } r = 0,1,2,\ldots,M \quad (*)$$

and

$$g_r = 0 \text{ for } r > M, \text{ where M is the greatest observed value of X.}$$

Although this is an empirical result, it is only so in that it makes no assumption regarding the distribution of the negative binomial parameter p; there is still the problem of the NB index ν, which is assumed to be a characteristic of the collection but whose value is, of course, unknown. Perhaps the best we can do is to note that, since necessarily $0 < \nu < \infty$, predictions based on these two extreme values are truly "empirical" and might in some sense be regarded as the limits of what can be termed empirical predictions based on mixtures of negative binomial processes (of constant index). In this case it is worthwhile to note the Corollary to the above Theorem as follows :

(i) as $\nu \to 0$,

$$g_r \to \begin{cases} f_0 + (1-t)(f_1 + f_2 + \ldots + f_M), \text{ for } r = 0 \ , \\ tf_r \ , \text{ for } r = 1,2,\ldots,M \ . \end{cases} \quad (**)$$

(ii) as $\nu \to \infty$,

$$g_r \to \left(\frac{t}{1-t}\right)^r \sum_{n=r}^{M} \binom{n}{r} (1-t)^n f_n \ , \text{ for } r = 0,1,2,\ldots,M \ . \quad (***)$$

§ 3. ILLUSTRATION

For much the same reasons as given in [4], we illustrate the Theorem by comparing its predictions with what actually happened in the University of Saskatchewan Library during 1968-1978. While this allows us to illustrate the performance of the model with the benefit of hindsight, it also means that we can put ourselves in the position of the library manager and investigate various possible "futures" as predicted by the model.

In the first place we imagine that the only information we have is the observed frequency-of-circulation distribution for 1968-69 (≈ year 1) and the total number of circulations for year 2. As in [7], this allows us to estimate the annual ageing factor as 0.8229 and then, further assuming that ageing occurs at an exponential rate, we would estimate the effective length of year n to be $(0.8229)^{n-1}$. (Note that the assumption of exponential ageing in fact proves to be a very crude approximation.) This is the situation covered by Table 1 for the years n = 4, 7 and 10, exactly as in Table 1 of [4] for Poisson mixtures. As the value of ν is unknown, we have given the predicted circulation distributions for various illustrative values together with what was in fact subsequently observed. (Note that we do not make use of the actually observed effective lengths of the years of interest; we are genuinely making predictions.)

These predictions seem to be most satisfactory in the middle of the range of r-values and, particularly for small values of ν, there is too much weight in the tails.

As we do not wish merely to illustrate the success (or otherwise) of the empirical method based on mixtures of NB processes as evinced by the application of (*), but also to compare it with results based on the empirical Poisson procedures derived in [4], we give in Table 2 values of the χ^2-statistic for the various prediction methods. As has been noted in the Appendix, the empirical mixed-Poisson case may be regarded as the limiting case (as $\nu \to \infty$) of a mixture of NB processes as given by (***). We also give the values based on the "deterministic" predictions based on (**) given by $\nu \to 0$. The general picture gained by consideration of Table 2 is that as the effective length of time becomes smaller, the process becomes more and more like a mixture of Poisson processes while for longer effective times smaller values of the NB index give better results.

§ 4. CONCLUDING REMARK

Clearly, more work is required before we can say with any certainty what is the "true" nature of the library loan process. For the moment it seems that mixtures of NB processes give the best predictions for collections with ageing, although Poisson mixtures give reasonable approximations and are somewhat simpler. In either case, in using the empirical approach, the very fact that we are able to dispose of explicit assumptions regarding the mixing distribution means that we have an extremely flexible tool to aid the library manager.

REFERENCES

[1] Gelman, E. & Sichel, H.S. (1987). Library book circulation and the beta-binomial distribution. Journal of the American Society for Information Science, 38, p.4-12.

[2] Burrell, Q.L. (1988). Correlation structure of library circulation data : a case study. Part 1 : The empirical view. Part 2 : Theoretical aspects. Research Report 117, Statistical Laboratory, University of Manchester.

[3] Burrell, Q.L. (1990). Using the gamma-Poisson model to predict library circulations. Journal of the American Society for Information Science, (to appear).

[4] Burrell, Q.L. (1988). A simple empirical method for predicting library circulations. Journal of Documentation, vol.44 (1988), pp.302-314.

[5] Beheshti, J. & Tague, J.M. (1984). Morse's Markov model of book use revisited. Journal of the American Society for Information Science, 35, 259-267.

[6] Tague, J. & Ajiferuke, I. (1987). The Markov and the mixed-Poisson models of library circulation compared. Journal of Documentation, 43, 212-231.

[7] Burrell, Q.L. (1985). A note on ageing in a library circulation model. Journal of Documentation, 41, 100-115.

[8] Burrell, Q.L. (1986). A second note on ageing in a library circulation model : the correlation structure. Journal of Documentation, 42, 114-128.

[9] Burrell, Q.L. (1987). A third note on ageing in a library circulation model : application to future use and relegation. Journal of Documentation, 43, 24-25.

[10] Burrell, Q.L. (1988). Modelling the Bradford phenomenon. Journal of Documentation, 44, 1-18.

[11] Burrell, Q.L. (1988). Predictive aspects of some bibliometric processes. In Egghe, L. & Rousseau, R. (eds.) : "Informetrics 87/88 : Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval", pp.43-63, Elsevier, Amsterdam.

APPENDIX

Proof of Theorem

As in [4] we might give a formal, analytic proof but this is not very enlightening so we prefer merely to give a heuristic proof analogous to the naive version given for Poisson mixtures in [4].

Let X and Y denote the number of loans of an item during [0,1] and the subsequent period of length $t < 1$ respectively. We wish to estimate

$$\gamma_r = E[\#\text{ of items for which } Y = r], \text{ for } r = 0,1,2,\ldots$$

on the basis of the known frequencies

$$f_n = \text{observed number of items for which } X = n, \text{ for } n = 0,1,2,\ldots .$$

By the stationarity of increments of mixtures of negative binomial processes, the distribution of Y is the same as the distribution of the increment during any other period length t and hence, in particular, the interval [0,t] . We seek to estimate the number of items circulating r times during this period.

Consider those items which circulate n times in [0,1], where $n \geq r$. Now for a negative binomial process, and any mixture of such processes, it can be shown that if there are n "events" in [0,1], then the conditional distribution of the number in [0,t] is a sort of hypergeometric form given by

$$p_t(r|n) = \frac{\binom{r+\nu t-1}{r}\binom{n-r+\nu(1-t)-1}{n-r}}{\binom{n+\nu-1}{n}}, \text{ for } r = 0,1,\ldots,n . \qquad (1)$$

But there are f_n such items and so the expected number of these producing r events (i.e. circulations) during [0,t] is just $p_t(r|n)f_n$. Now summing over all $n \geq r$ gives the total expected number of items circulating r times during [0,t] as

$$g_r = \sum_{n \geq r} \frac{\binom{r+\nu t-1}{r}\binom{n-r+\nu(1-t)-1}{n-r}}{\binom{n+\nu-1}{n}} f_n , \text{ for } r = 0,1,2,\ldots \qquad (*)$$

When we note that, by the definition of M as the greatest observed value of X, $f_n = 0$ for $n > M$ we have the required result.

Corollary

(i) In the limit as $\nu \to 0$,

$$g_r = \begin{cases} f_0 + (1-t)(f_1 + f_2 + \ldots + f_M), & \text{for } r = 0 \\ tf_r , & \text{for } r = 1,2,\ldots,M . \end{cases} \qquad (**)$$

(ii) In the limit as $\nu \to \infty$,

$$g_r = \left(\frac{t}{1-t}\right)^r \sum_{n=r}^{M} \binom{n}{r} (1-t)^n f_n \text{ , for } r = 0,1,2,\ldots,M \text{ .} \qquad (***)$$

The proofs of (i) and (ii) as corrolaries of the main result are quite straightforward. It is worth remarking that (i) is a sort of deterministic result while (ii) is the form given by a mixture of Poisson processes as in [4], i.e. the corresponding result for empirical mixtures of "purely random" processes.

Table 1 : Predicted frequency-of-circulation distributions for the University of Saskatchewan Library using empirical mixtures of negative binomial processes.

Predictions are based on the observed distribution in year 1 = 1968-1969 and the annual ageing factor is 0.823.

(a) Year 4 = 1971-1972

Number of circulations	Predicted number of items			Observed number
	(ν = 0.5)	(ν = 5.0)	(ν = 10.0)	
0	58235	56741	56548	58073
1	5411	6796	6944	5305
2	2199	2604	2679	2467
3	1162	1214	1235	1242
4	650	598	596	644
5	372	306	297	394
6	229	161	149	214
7	142	84	74	111
8	79	43	36	62
9	48	22	17	47
10	29	11	8	16
11	15	5	4	10
≥ 12	19	5	3	5

(b) Year 7 = 1974-1975

Number of circulations	Predicted number of items			Observed number
	(ν = 0.5)	(ν = 5.0)	(ν = 10.0)	
0	62430	60773	60496	60565
1	3444	5235	5532	5177
2	1277	1561	1621	1753
3	636	587	575	688
4	342	242	219	263
5	191	106	87	90
6	114	47	35	36
7	68	21	14	13
≥ 8	88	18	9	5

(c) Year 10 = 1977-1978

Number of circulations	Predicted number of items			Observed number
	(ν = 0.5)	(ν = 5.0)	(ν = 10.0)	
0	65027	63746	63496	63251
1	2068	3577	3900	3976
2	725	842	848	997
3	349	266	236	260
4	183	96	73	67
5	100	37	24	34
≥ 6	138	26	13	5

Table 2 : χ^2-values for various prediction methods applied to 1968-69 data from the University of Saskatchewan

Method	Year 4	Year 7	Year 10
(a) Parametric mixed Poisson			
(i) p = 0.298, ν = 0.243	1151.7	301.4	216.1
(ii) p = 0.270, ν = 0.211	764.1	170.4	147.2
(b) Empirical mixed Poisson	902.3	203.7	203.6
(c) Empirical mixed negative binomial			
(i) ν = 0.5	71.3	1375.2	2178.9
(ii) ν = 5.0	465.2	61.6	103.0
(iii) ν = 10.0	617.3	68.6	40.6
(d) Deterministic	494.2	4203.4	74603.6