Mixture models for word frequency distributions

Fiona J. Tweedie & R. Harald Baayen

Department of Statistics, University of Glasgow, Mathematics Building, University Gardens, Glasgow, G12 8QW, UK, & IWTS, University of Nijmegen, P.O.Box 310, 6500 AH, Nijmegen, The Netherlands.

Abstract

Word frequency distributions are generally extremely skewed and are described as having a Large Number of Rare Events (LNRE). LNRE distributions have been found to provide fits to many examples of word frequency distributions. However, Baayen and Tweedie (1998b) present a distribution of the Dutch suffix *-heid* which cannot be fitted by standard methods. In this case, the data come from a composite source and we introduce the idea of mixture distributions to deal with this. We present expressions for the expected word frequency distribution, the expected number of tokens, and the number of types in the population. An acceptable fit to the *-heid* data is presented.

Keywords: word frequency distributions, mixture models, LNRE models

1. Introduction

Word frequency distributions, the number of words that occur once, twice, and so on in a text are generally found to be extremely skewed, with often around half of the words in a text being found at a single location (see, e.g., Muller, 1977; 1979). These distributions are described as having a Large Number of Rare Events (LNRE) and a review of such distributions can be found in Chitashvili and Baayen (1993). Chitashvili and Baayen describe three models for LNRE distributions; Carroll's lognormal model, the Yule-Simon model and Sichel's Generalised Inverse Gauss-Poisson (GIGP) model. In this paper we shall concentrate on the last of these.

We shall consider the word frequency spectrum $V(m, N)$, the number of words occurring m times in a text of length N, $m = 1 \ldots, V$. Sichel's GIGP model gives the following expression for the spectrum elements,

$$\mathrm{E}[V(m, pN)|\{Z, b, \gamma]\} = \frac{2Z}{bK_{\gamma+1}(b)(1+N/Z)^{\gamma/2}} \frac{\left(\frac{bN}{2Z\sqrt{1+N/Z}}\right)^m}{m!} \cdot K_{m+\gamma}(b\sqrt{1+N/Z}), \quad (1)$$

where $K_a(b)$ is the modified Bessel function of the second kind of order a. The expression for $V(N)$, the number of word types in the text is

$$\mathrm{E}[V(N)|\{Z, b, \gamma]\} = \frac{2Z}{b}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}\left[1 - \frac{K_\gamma(b\sqrt{1+N/Z})}{(1+N/Z)^{\gamma/2}K_\gamma(b)}\right]. \quad (2)$$

The number of different types in the population S is given by the expression

$$S = \frac{2Z}{b}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}. \quad (3)$$

These expressions have been used to model successfully a variety of texts (Tweedie and Baayen, 1998; Baayen and Tweedie, 1998a), as well as in other application areas (e.g. Heller, 1997; Burrell and Fenton, 1993).

However, Baayen and Tweedie (1998b) present a word frequency distribution which cannot be fitted with the GIGP model. The Dutch suffix *-heid*, which, like *-ness* in English, coins abstract nouns from adjectives, has two distinct semantic functions. First, -HEID is used to create words for concepts. For instance, the Dutch word SNEL-HEID, literally QUICK-NESS, has SPEED as its English translation equivalent. Thus the concept of 'speed' is realized by a monomorphemic word in English, just as the concepts 'house' and 'tree' are realized by monomorphemic words. In Dutch, the concept for 'speed' happens to be realized by a morphologically complex word. We will henceforth refer to this use of -HEID as its conceptual use.

Second, -HEID is also used to refer to states of affairs that have been introduced previously in the discourse. For instance, if John has been described as grateful, this situation can later be referenced to by 'John's gratefulness'. We will refer to this use of -HEID as its anaphoric use. The two functions of -HEID, concept-formation versus anaphoric discourse referencing, are not mutually exclusive. One and the same word can realize both functions. Nevertheless, the two functions can be distinguished quantitatively. Baayen and Neijt (1997), using contextual clues for a sample of low and high-frequency nouns in -HEID in a corpus of newspaper Dutch, show that the conceptual function is more commonly realized for high-frequency words, while the anaphoric use is more typical for low-frequency words.

Mixture models (see for example Titterington, Smith, and Makov, 1985) describe distributions where the data come from one or more source. In the example described above, data are coming from both the anaphoric and conceptual uses of words, but we only observe the final word frequency distribution spectrum. The translation equivalent method described in Baayen and Tweedie (1998b) can indicate an approximate split between the distributions, but mathematical methods are necessary to find a more suitable fit.

When we model a word frequency spectrum we are interested in finding expected values of the elements $V(m, N)$. The parameters of such models as the GIGP are then chosen to make the expected value of the spectrum elements, $\mathbf{E}[V(m, N)]$ as close to the observed $V(m, N)$ as possible. In some cases, such as the one discussed above, a single distribution is not enough to deal with the observed data. In these cases we can consider the use of a mixture distribution, where the expected values are made up as follows:

$$\mathbf{E}[V(m, N)] = \mathbf{E}_1[V(m, pN)] + \mathbf{E}_2[V(m, (1-p)N)],$$

where p is the proportion of the data coming from the first distribution, usually called the mixing parameter, and $(1-p)$ the proportion of the remainder which comes from a second distribution. $\mathbf{E}_1$ and $\mathbf{E}_2$ indicate the expected values under the different distributions.

2. LNRE mixture models

We take the frequency spectrum as point of departure, and study the simple case in which $V(m, N)$ originates from just two distributions, one with parameters Z_1, a_1, and b_1, and one with parameters Z_2, a_2, and b_2. We assume that pN of our tokens have been sampled from the first distribution, and that $(1-p)N$ tokens come from the second distribution. For the

expectation of $V(m,N)$ we have

$$\begin{aligned}\mathrm{E}[V(m,N)] &= \mathrm{E}[V(m,pN)|\{Z_1,a_1,b_1\}+V(m,(1-p)N)|\{Z_2,a_2,b_2\}] \\ &= p\mathrm{E}[V(m,N)|\{\frac{Z_1}{p},a_1,b_1\}]+(1-p)\mathrm{E}[V(m,N)|\{\frac{Z_2}{1-p},a_2,b_2\}], \quad (4)\end{aligned}$$

where we make use of the fact that for any LNRE model,

$$\mathrm{E}[V(m,pN)|\{Z,\ldots]\} = p\mathrm{E}[V(m,N)|\{\frac{Z}{p},\ldots\}]. \quad (5)$$

Equivalently, writing $Z' = Z/p$, we can rephrase (5) as follows:

$$\mathrm{E}[V(m,pN)|\{pZ',\ldots]\} = p\mathrm{E}[V(m,N)|\{Z',\ldots\}],$$

in other words, the expected spectrum is linear in p with respect to N and Z. A proof for the case of the GIGP distribution proceeds as follows:

$$\begin{aligned}&\mathrm{E}[V(m,pN)|\{Z,b,\gamma]\} = \\ &= \frac{2Z}{bK_{\gamma+1}(b)(1+(pN)/Z)^{\gamma/2}}\frac{\left(\frac{b(pN)}{2Z\sqrt{1+(pN)/Z}}\right)^m}{m!}\cdot K_{m+\gamma}(b\sqrt{1+(pN)/Z}) \\ &= p\frac{2\frac{Z}{p}}{bK_{\gamma+1}(b)(1+N/\frac{Z}{p})^{\gamma/2}}\frac{\left(\frac{bN}{2\frac{Z}{p}\sqrt{1+N/\frac{Z}{p}}}\right)^m}{m!}\cdot K_{m+\gamma}\left(b\sqrt{1+N/\frac{Z}{p}}\right) \\ &= p\mathrm{E}[V(m,N)|\{\frac{Z}{p},b,\gamma\}]. \quad (6)\end{aligned}$$

Similar proofs for the lognormal and Yule-Simon models can be found in Baayen (forthcoming). For the expected vocabulary we have that

$$\begin{aligned}\mathrm{E}[V(N)] &= \mathrm{E}[V(pN)|\{Z_1,a_1,b_1\}+V((1-p)N)|\{Z_2,a_2,b_2\}] \\ &= p\mathrm{E}[V(N)|\{\frac{Z_1}{p},a_1,b_1\}]+(1-p)\mathrm{E}[V(N)|\{\frac{Z_2}{1-p},a_2,b_2\}], \quad (7)\end{aligned}$$

where we use the fact that the vocabulary size can be expressed as the sum of the spectrum elements:

$$\begin{aligned}\mathrm{E}[V(pN)|\{Z,a,b\}] &= \mathrm{E}[\sum_m V(m,pN)|\{Z,a,b\}] \\ &= \sum_m p\mathrm{E}[V(m,N)|\{\frac{Z}{p},a,b\}] \\ &= p\mathrm{E}[V(N)|\{\frac{Z}{p},a,b\}]. \quad (8)\end{aligned}$$

Variances and covariances of the mixture model can be expressed as the sums of variances and covariances of the components. For the variance of the vocabulary size we have:

$$\begin{aligned}\mathrm{VAR}[V(N)] &= \mathrm{E}[V(2N)]-\mathrm{E}[V(N)] \\ &= \mathrm{E}[V(2pN)|\{Z_1,a_1,b_1\}]+\mathrm{E}[V(2(1-p)N)|\{Z_2,a_2,b_2\}] \\ &\quad -[\mathrm{E}[V(pN)|\{Z_1,a_1,b_1\}]+\mathrm{E}[V((1-p)N)|\{Z_2,a_2,b_2\}]] \\ &= \mathrm{VAR}[V(pN)]+\mathrm{VAR}[V((1-p)N]. \quad (9)\end{aligned}$$

For the covariances of the spectrum elements and the vocabulary size we obtain

$$\begin{aligned}\mathrm{COV}[V(m,N),V(N)] &= \frac{1}{2^m}\mathrm{E}[V(m,2N)] \\ &= \frac{1}{2^m}\left[\mathrm{E}[V(m,2pN)] + \mathrm{E}[V(m,2(1-p)N)]\right] \\ &= \mathrm{COV}[V(m,2pN)] + \mathrm{COV}[V(m,2(1-p)N)], \qquad (10)\end{aligned}$$

and along similar lines it can be shown that

$$\begin{aligned}\mathrm{COV}[V(m,N),V(k,N)] &= \mathrm{COV}[V(m,pN),V(k,pN)] + \\ &\quad \mathrm{COV}[V(m,(1-p)N),V(k,(1-p)N)]. \qquad (11)\end{aligned}$$

The population number of types in a mixture with L components, each with a population number of types S_i, equals

$$\begin{aligned}\lim_{N\to\infty}\mathrm{E}[V(N)] &= \lim_{N\to\infty}\sum_{i=1}^{L}\mathrm{E}[V(p_iN)|\{Z_i,\ldots\}] \\ &= \sum_{i=1}^{L} p_i \lim_{N\to\infty}\mathrm{E}[V(N)|\{\frac{Z_i}{p_i},\ldots\}] \\ &= \sum_{i=1}^{L} p_i \frac{2Z_i}{bp_i}\frac{K_\gamma(b)}{K_{\gamma+1}(b)} \\ &= \sum_{i=1}^{L} S_i. \qquad (12)\end{aligned}$$

3. Computational issues

Parameter estimation for the GIGP model is non-trivial. Closed-form expressions for Maximum Likelihood estimators are not available and numerical methods must be employed to find the best-fitting parameters. For mixtures of two LNRE distributions the situation is further complicated by the need to estimate two sets of parameters as well as the mixing parameter, p. At present we estimate the parameters by a user-guided search for p and the first set of parameters. The second set of parameters is found by fitting what remains of the observed frequency spectrum once the first distribution has been subtracted. For a mixture of two GIGP distributions with parameters $\{\hat{Z}=42.0,\hat{b}=0.084,\hat{\gamma}=-0.5\}$ and $\{\hat{Z}=18.2,\hat{b}=0.000000000514,\hat{\gamma}=-0.5092\}$ respectively and mixing parameter $p=0.67$, we obtain an acceptable fit with the multivariate X^2 test ($X^2_{(9)}=19.29, p=0.023$). Figures 1 and 2 summarize graphically the improvements achievable by changing from a simple LNRE model to a mixture model.

For the case of two GIGP distributions, a simplex search can be implemented to range over the space of p and the first set of parameters. The parameters of the second GIGP distribution are then fitted to the remaining spectrum elements, again after subtracting those fitted by the first distribution. We are currently implementing the simplex search algorithm. At the conference, we will illustrate the software that we have developed for mixture distributions, which will become available under the GNU general public license.

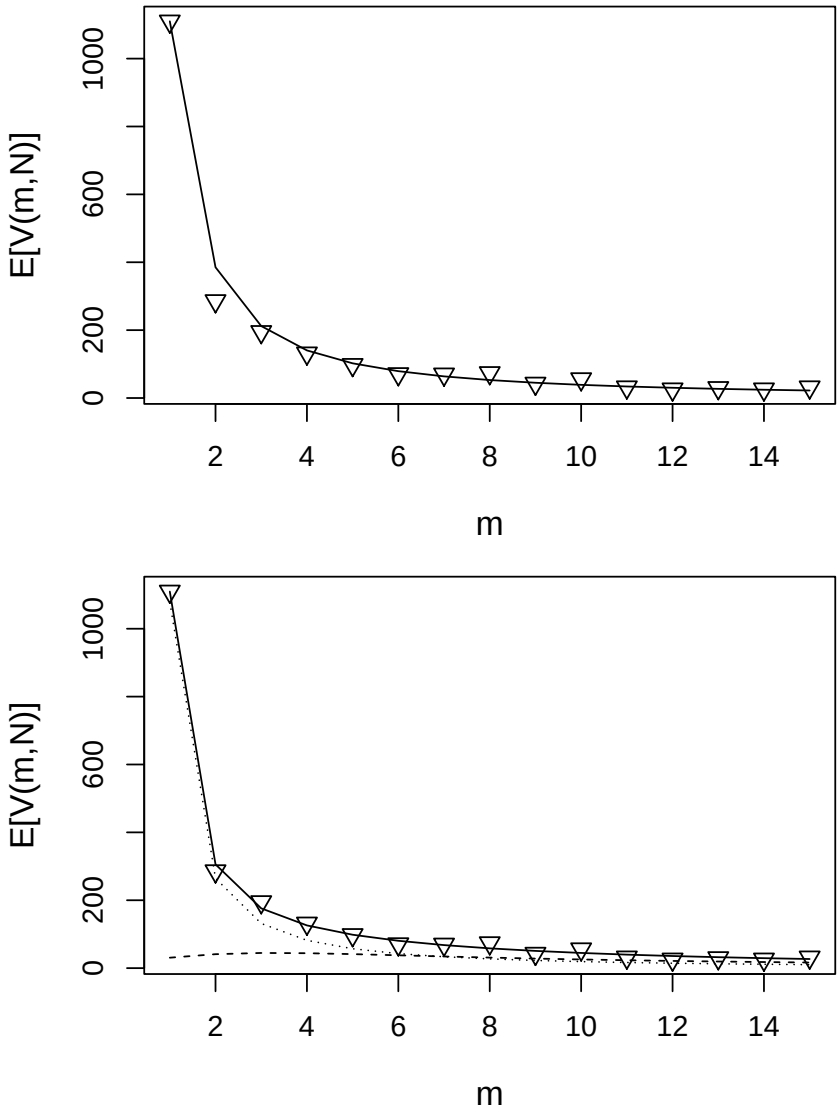

Figure 1: A simple GIGP fit to the first fifteen elements of the frequency spectrum of *-heid* (top panel), and the corresponding GIGP-GIGP mixture fit (bottom panel). In the bottom panel, the solid line represents the mixture model, the dotted line and the dashed line represent the component distributions.

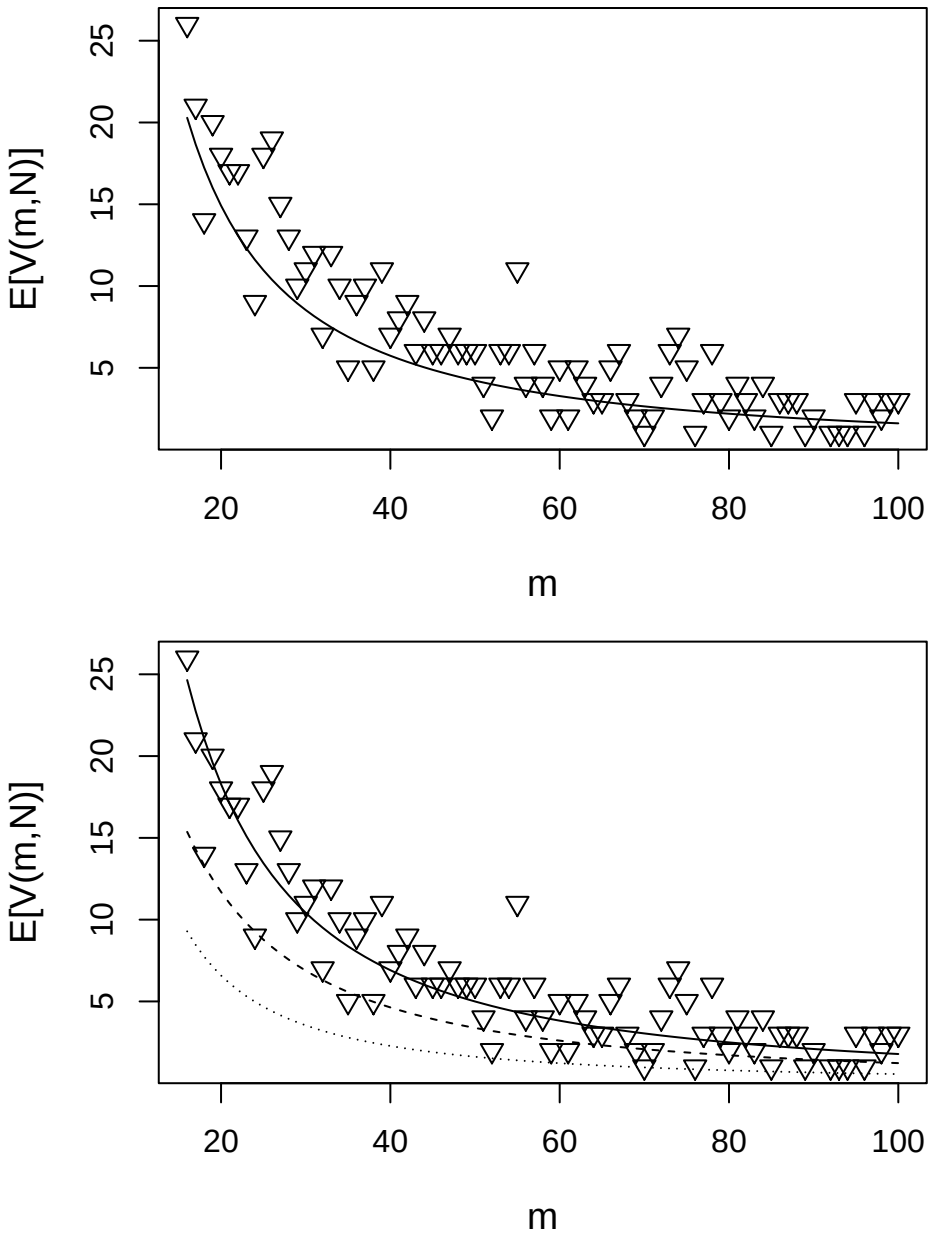

Figure 2: The underestimation bias of the simple GIGP model for the medium frequency spectrum elements (top panel), and the GIGP-GIGP fit (solid line) and its component distributions (dashed and dotted lines, bottom panel).

References

Baayen R. H. (1999). *Word Frequency Distributions*. (forthcoming), Nijmegen.

Baayen R. H. and Neijt A. (1997). Productivity in context: a case study of a Dutch suffix. *Linguistics*, 35:565–587.

Baayen R. H. and Tweedie F. (1998a). Enhancing LNRE models with partition-based adjustment. In *Proceedings of JADT 1998*, pages 29–37, Nice. Université Nice Sophia Antipolis.

Baayen R. H. and Tweedie F. J. (1998b). A mixture model for a uni-modal word frequency distribution. In *ALLC/ACH'98 conference abstracts*, pages 15–18, Debrecen.

Burrell Q. and Fenton M. (1993). Yes, the GIGP really does work — and is workable! *Journal of the American Society for Information Science*, 44:61–69.

Chitashvili R. J. and Baayen R. H. (1993). Word frequency distributions. In Altmann G. and Hřebíček L. editors, *Quantitative Text Analysis*, pages 54–135. Wissenschaftlicher Verlag Trier, Trier.

Heller G. Z. (1997). Estimation of the number of classes. *South African Statistical Journal*, 31:65–90.

Muller C. (1977). *Principes et méthodes de statistique lexicale*. Hachette, Paris.

Muller C. (1979). Peut-on estimer l'étendue d'un lexique? In *Langue Française et Linguistique Quantitative*, pages 399–425. Slatkine, Genève.

Titterington D. M., Smith A. F. M., and Makov U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Chichester.

Tweedie F. and Baayen R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.