The Best of Both Worlds: Combining MOCAR and MCDISP

Arjuna Tuzzi & Fiona J. Tweedie

Dipartimento di Scienze Statistiche, Università di Padova, via S. Francesco, 33, 35121 PADOVA, ITALY. & Department of Statistics, University of Glasgow, Mathematics Building, University Gardens, GLASGOW, G12 8QW, UK.

Abstract

In the analysis of a corpus of open-ended questions, one of the most important goals is to identify words which distinguish between groups of respondents. The MOCAR procedure within SpadT does this using hypergeometric probabilities (Lebart et al., 1998). However, while the words obtained may only occur within a particular group, the researcher has no indication of their distribution within that group. A word may be chosen which is specific to one or two responses, rather than being representative of the group as a whole. We address this problem using the MCDISP procedure developed by Baayen (1996). The words identified by MOCAR can then be checked for significant under-dispersion, which would indicate that they are confined to a subset of the texts. We illustrate this with data from a corpus of open interviews of graduates of the University of Padua.

Keywords: analysis of open questions, SpadT, dispersion

1. Introduction

Open interviews are special questionnaires in which questions are open, i.e. requiring a free response. They are frequently used in social surveys and in marketing applications. Since the responses to open questions are special text, the problem of processing textual information is not a new question. In the past manual post coding procedures were used to extract the apparent content (Bradburn and Sudman, 1979; Schuman and Presser, 1981), nowadays the modern textual analysis techniques are based on an automatic numerical labelling of word tokens and are latent content search oriented (Bolasco, 1999; Lebart et al, 1998). Open answers can be processed in their original form, but can also be matched with additional information such as respondents' demographic characteristics as well as their responses to other closed questions. The results of a survey based on open interviews constitute a corpus in which the single interviews are the natural partition but the available information on these texts can be exploited to identify grouping criteria useful for the analysis. In order to describe groups of texts it is possible to use the so-called "characteristic elements" (Lebart et al, 1998), i.e. the textual units (words, lemmas, segments, etc.) that are present in a given group either a great deal more or a great deal less than in the rest of the corpus. To detect characteristic elements the procedure MOCAR of the package SpadT (one of the most common programs for textual analysis) implements a method based on the hypergeometric model (Lebart et al, 1989).

1.1. The Hypergeometric Model

Given a corpus of dimension N in word tokens and $V(N)$ in word types, we suppose that all the texts of the corpus are classified into p different groups. Given the grouping criteria, each word type can be classified in the p groups (g_1, $ldots$, g_p) by means of a lexical table, a matrix ($V(N)$

	g_1	g_2	$\dots$	g_j	$\dots$	g_p	
ω_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1.}$
ω_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
ω_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\omega_{V(N)}$	$n_{V(N)1}$	$n_{V(N)2}$	$\dots$	$n_{V(N)j}$	$\dots$	$n_{V(N)p}$	$n_{V(N).}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.j}$	$\dots$	$n_{.p}$	N

Table 1: A generic lexical table (word types * groups)

x p) in which the p columns represent the groups and the $V(N)$ rows the word types (Table 1). The generic element n_{ij} of the lexical table is the number of times the word type on row i occurs in all the texts assigned to the group on column j. The marginal row totals $n_{i.}$ are the total occurences of the word type on row i in the whole corpus, and the marginal column totals $n_{.j}$ the dimension in word tokens of the group on column j. We select the word type on row i and the group on column j to see if the word type is characteristic for the group. In order to use the hypergeometric probability distribution we have to imagine a population of N objects. Only $n_{i.}$ of these objects are marked (the marked objects are the occurences of the word type), the other $N - n_{i.}$ are unmarked and indistinguishable (the unmarked objects are the occurences of all the other $V - 1$ word types). Using random sampling without replacement, we draw a sample of $n_{.j}$ objects (the group) and evaluate the number of marked objects in the sample. The number of marked objects in the sample is a realization of a hypergeometric random variable with parameters N, $n_{i.}$ and $n_{.j}$.

$$X \sim \text{Hyp}(N, n_{i.}, n_{.j})$$

$$\Pr(X = x|N, n_{i.}, n_{.j}) = \frac{\left(\begin{array}{c} n_{i.} \\ x \end{array}\right)\left(\begin{array}{c} N - n_{i.} \\ n_{.j} - x \end{array}\right)}{\left(\begin{array}{c} N \\ n_{.j} \end{array}\right)} \quad (1)$$

1.2. The MOCAR *Procedure*

For each word and for each group the observed value n_{ij} (occurences of word type i in group j) is compared with the mode of the hypergeometric distribution with parameters N, $n_{i.}$ and $n_{.j}$ in order to decide if the word type is a positive characteristic word ($n_{ij} > \textit{mode}$) or a a negative characteristic one ($n_{ij} < \textit{mode}$) for the group (Lafon, 1980). The MOCAR procedure computes the probability on the right tail if $n_{ij} >$ mode and the probability on the left tail if $n_{ij} <$ mode. The smaller this probability, the more characteristic the word is deemed to be of this group (in fact values closed to the mode represent banal words). We consider a word to be characteristic of a group if it is largely over-represented,

$$\sum_{n_{ij} \leq x \leq n_{i.}} \Pr(X = x|N, n_{i.}, n_{.j}) < \alpha \quad (2)$$

or under-represented,

$$\sum_{0 \leq x \leq n_{ij}} \Pr(X = x|N, n_{i.}, n_{.j}) < \alpha \quad (3)$$

in the group, according to a threshold that should be stricter than $\alpha = 5\%$ because of the problem of multiple comparisons (Lebart et al, 1998).

1.3. The under-dispersion problem in MOCAR *analysis*

The occurrence of a word type is not a simple attribute for a text because the word type can occur either many times, or never, inside it. A group of texts is formed by pooling a set of texts, so the occurences of the word type in the group is then the sum of the occurences for each text assigned to the group. If a word type occurs a great deal more in a given text than in the rest of a group of texts, or indeed the corpus as a whole, can it be considered to be a characteristic word for the group, or only for the single text? To answer this question we have to evaluate the dispersion of the word type inside the group. If the word type is restricted to one or a few texts, it is described as *under-dispersed* and this implies that the word is not characteristic for the group, but only for few of its components. If the word type is spread out over all the texts of the group the word can be considered to be important for the group.

Baayen (1996) defines *dispersion* as follows: "The dispersion d_i of a word ω_i is the number of text slices in which ω_i occurs. (462)". In the context of his investigation into the randomness assumption, that words occur randomly through text, Baayen uses a permutation test to identify words which are significantly under-dispersed using equal-sized slices of text. We have modified his MCDISP program to allow for variable-size slices, i.e. each text, or in this case, each respondent's answer, in a corpus can be treated as a unit. The permutation test permutes the text, randomising the order of the words, and then calculates the number of slices in which each word appears. This is carried out a great number of times, here we use 1000 permutations. The number of slices in which a word appears in the original text can then be compared with the empirical distribution of numbers of slices from the permutations and a p-value given. Thus, a word which occurs frequently in only a single text slice within the corpus ought to appear in many slices when the text is permuted, and its occurrence in the original single slice will hence be deemed significant.

1.4. Method

We will illustrate the method using textual data taken from the results of a wide survey of the professional profiles of the graduates of the Faculty of Political Sciences of the University of Padua. The corpus is composed of 256 open interviews made by a group of student interviewers submitting a set of 17 questions to a sample of graduates. Interviews are grouped by external criteria such as gender, date of graduation, etc, and also by internal criteria such as answers to closed questions within the questionnaire.

MOCAR was used to identify words that were considered characteristic of each group. Next, MCDISP was used to discover if these words were under-dispersed. In this context, under-dispersed words are likely to have been used by a single, or a few, respondents, rather than by the group as a whole and should therefore not be considered representative. We describe the results in the section below.

2. Results

We shall illustrate the method by considering two questions that could be asked about this corpus. The first is "Are the male graduates only interested in a job and a career?" where the grouping used is gender, and interest is in words relating to jobs and careers. The second, "Do

Male			Female		
ω_i	MOCAR	MCDISP	ω_i	MOCAR	MCDISP
laureato	< 0.001	0.943	*laureata*	< 0.001	0.675
soddisfatto	0.009	0.609	*soddisfatta*	< 0.001	0.157
segretario	0.003	< 0.001	*segretaria*	0.004	> 0.999
banca	< 0.001	< 0.001	*tesi (TDL)*	< 0.001	< 0.001
azienda	0.001	< 0.001	*Lingue*	< 0.001	< 0.001
consulente	< 0.001	< 0.001	*ricerca*	< 0.001	< 0.001
competenze	< 0.001	0.991			

Table 2: Some positive characteristic words in the analysis of gender. MOCAR between groups and MCDISP within groups

graduates from different times have different worries?", looks at differences in the responses of graduates from 1972–1980, 1981–1989, and 1990–1995.

2.1. Are the male graduates only interested in a job and a career?

In analysing the MOCAR results for gender we find trivially that male inflections for verbs, nouns and adjectives are positive characteristic words for male respondents and the female inflections are positive characteristic words for female respondents. Examples (Table 2) here include *laureato/laureata* (*graduated*) and *soddisfatto/soddisfatta* (*satisfied*) with MOCAR results $p < 0.01$ and $p < 0.001$ respectively for both males and females. The MCDISP test confirms these results because most of these words are not under-dispersed, that is, male and female inflections are used throughout the replies from men and women respectively and are not restricted to a subset of men or women (MCDISP with 1000 permutations: *laureato/a* $p = 0.943$ for men and $p = 0.675$ for women; *soddisfatto/a* $p = 0.609$ for men and $p = 0.157$ for women). Interestingly, *segretario/a* (*secretary*) is found to be significant by MOCAR for both male and female respondents. According to MCDISP it is dispersed throughout the female respondents' texts ($p > 0.999$) but is significantly under-dispersed in the male respondents' texts ($p < 0.001$), again with 1000 permutations.

The MOCAR results suggest other words which distinguish between genders; these words concern the attitude in preferring job and career subjects (for men) rather than university and post degree training (for women). However, this interpretation of the data is not supported by MCDISP. All the positive characteristic words for men concerning jobs and careers (*banca — bank, azienda — enterprise, consulente — consultant,* ...) are significantly under-dispersed and only *competenze — skills* can be considered to be representative of the group ($p = 0.991$). In the same way all the positive characteristic words for women concerning training and study (*tesi — thesis, lingue — foreign languages, ricerca — research* ...) are under-dispersed.

2.2. Do graduates from different times have different worries?

According to the MOCAR results, the group composed by the oldest graduates (who received their degree in the period 1972–80) speak about time using words that concern things which seemed far away at the moment of the interview (Table 3). They speak about their times and use words such as *anni — years, tempi — times, passato — past* and *Settanta — Sixties* (MOCAR $p < 0.02$). In contrast the youngest graduates (who received their degree in the period 1990–95)

1972–1980			1981–1989			1990–1995		
ω_i	MOCAR	MCDISP	ω_i	MOCAR	MCDISP	ω_i	MOCAR	MCDISP
anni	< 0.001	0.730	*flessibilità*	< 0.001	0.107	*mesi*	< 0.001	0.472
tempi	0.001	0.023	*autonomia*	< 0.001	0.169	*giorni*	< 0.001	> 0.999
passato	0.013	0.297	*capacità*	0.002	0.250	*presto*	0.019	0.038
Settanta	0.014	0.165	*preguidizi*	0.005	0.044	*tesi (TDL)*	< 0.001	< 0.001
insegnante	< 0.001	< 0.001	*preparati*	0.019	< 0.001	*estero*	< 0.001	< 0.001
insegnare	< 0.001	< 0.001	*serie B*	0.020	> 0.999	*Lingue*	< 0.001	< 0.001
insegnamento	< 0.001	0.013	*peggio*	0.020	> 0.999	*percorsi*	0.009	0.002
insegnanti	< 0.001	0.036	*carriera*	0.015	> 0.634	*percorso*	0.010	< 0.001
scuola	< 0.001	0.057	*prestigio*	0.016	0.237	*colloqui*	0.001	0.535
communale	< 0.001	< 0.001				*contratto*	0.001	0.380
giuridico	< 0.001	0.029				*occasioni*	0.003	0.075
pubblico	0.001	0.002				*umiltà*	0.005	> 0.999
Comune	0.003	< 0.001				*trovare*	0.007	0.175
						apprendere	0.009	0.494
						assumere	0.013	0.164
						adattarsi	0.009	0.518

Table 3: Some positive characteristic words in the analysis of date of graduation. MOCAR between groups and MCDISP within groups

are speaking about recent things so use words such as *mesi — months*, *giorni — days* and *presto — early*. These results are supported by MCDISP as most of these words are not significantly under-dispersed (only *presto — early* and *tempi — times* are under-dispersed, $p = 0.038$ and $p = 0.023$ for the youngest and oldest graduates respectively).

The interviews are about professional profiles, hence the job subject is pre-eminent and it is important to examine what the graduates of different ages are speaking about. According to MOCAR results, the oldest (1972–80) seem to be associated with a job in the public sector and particularly with teaching and schools (and we have to underline that in Italy the vast majority of schools are public i.e. state schools). This result seems very reliable because in the past graduates of the Political Sciences faculties very often became teachers at high schools. In fact we can find within the positive characteristic words for this group *insegnante — teacher*, *insegnare — to teach*, *insegnamento — teaching*, *insegnanti — teachers*, *scuola — school*, *comunale — municipal*, *giuridico — legal*, *pubblico — public* and *Comune — town council* (MOCAR $p \leq 0.003$). But MCDISP results suggest that these data interpretations are all incorrect because all these words are significantly under-dispersed (p-values from < 0.001 to 0.057 for *scuola*).

According to MOCAR results the youngest graduates (1990–95) are worried about their training; they have recently graduated and are still speaking about their study experiences using words such as *tesi — thesis*, *estero — abroad*, *Lingue — foreign languages*, *percorsi — forming lines*, *percorso — forming line* (MOCAR $p < 0.01$) and about their first employment; they are looking for jobs and so are worried about *colloqui — interviews*, *contratto — contract*, *occasioni — chances*, *umiltá — humility*, *trovare — to find*, *apprendere — to learn*, *assumere — to hire*,

adattarsi — to adapt, MOCAR $p < 0.02$). These appealing data interpretations are only partially reliable, in fact the MCDISP results show us that only the second group of words, those concerning the job search, are useful in understanding the peculiarities of the group of youngest graduates.

The graduates of the Eighties (1981–89) began looking for a job in a period of great depression for Italy and the unemployment nightmare clearly appears in their interviews (these results are supported by both MOCAR and MCDISP). But they graduated in a difficult period for the Faculty too (a group of teachers were arrested because of contrasts of ideology with the government) so in speaking about the problem of showing the right skills to find a job in a depressed period (*flessibilitá — flexibility*, *autonomia — autonomy*, *capacitá — ability*) they also seem very worried about discrimination (*pregiudizi — prejudices*, *preparati — prepared*, *serie B — second category*, *peggio — worse*). In Italy the Eighties were the years of Yuppies too, so graduates are very interested in *carriera — career* and *prestigio — prestige* too.

3. Conclusions

It is clear from the results obtained in the section above that use of the MOCAR procedure alone is not sufficient to draw general conclusions about differences between groups of texts, or in this case, respondents. The first set of conclusions described above are found to be strongly influenced by the views of a subset of men and women and cannot be held to be representative of male and female graduates of the Faculty of Political Science of Padua University. Turning to the second question, we find that some hypotheses, for example the oldest graduates being associated with public sector jobs, cannot be confirmed as being representative, while others, such as the Eighties graduates being associated with difficulty in finding a job, can be confirmed as being relevant to the group as a whole.

We must therefore conclude by advising those using MOCAR to supplement their analysis with a procedure such as MCDISP to ensure that their conclusions are representative of their groups and not unduly influenced by a smaller subset of texts.

References

Baayen R. H. (1996). The effect of lexical specialisation on the growth curve of the vocabulary. *Computational Linguistics*, 22:455–480.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bradburn N., Sudman S., and Associates (1979). *Improving Interview Method and Questionnaire Design*. Jossey-Bass publishers, San Francisco.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1(127–165).

Lebart L., Morineau A., and Bécue M. (1989). *SPAD.T Système Portable pour l'Analyse des Données Textualles*. CASIA, Paris.

Lebart L., Salem A., and Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht.

Schuman H. and Presser F. (1981). *Questions and Answers in Attitude Surveys*. Academic Press, New York.