POSSIBLE SOLUTIONS OF SOME ESSENTIAL ZERO PROBLEMS IN COMPOSITIONAL DATA ANALYSIS

J. Aitchison[1] and J. W. Kay[2]
Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland
[1]Email: john.aitchison@btinternet.com [2]Email: jim@stats.gla.ac.uk

Abstract

One of the tantalising remaining problems in compositional data analysis lies in how to deal with data sets in which there are components which are essential zeros. By an essential zero we mean a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring instrument has not been sufficiently sensitive to detect a trace of the part. Such essential zeros occur in many compositional situations, such as household budget patterns, time budgets, palaeontological zonation studies, ecological abundance studies. Devices such as non-zero replacement and amalgamation are almost invariably ad hoc and unsuccessful in such situations. From consideration of such examples it seems sensible to build up a model in two stages, the first determining where the zeros will occur and the second how the unit available is distributed among the non-zero parts. In this paper we suggest two such models, an independent binomial conditional logistic normal model and a hierarchical dependent binomial conditional logistic normal model. The compositional data in such modelling consist of an incidence matrix and a conditional compositional matrix. Interesting statistical problems arise, such as the question of estimability of parameters, the nature of the computational process for the estimation of both the incidence and compositional parameters caused by the complexity of the subcompositional structure, the formation of meaningful hypotheses, and the devising of suitable testing methodology within a lattice of such essential zero-compositional hypotheses. The methodology is illustrated by application to both simulated and real compositional data.

1. Introduction

One of the tantalising remaining problems in compositional data analysis lies in how to deal with data sets in which there are components which are essential or structural zeros. By an essential zero we mean a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring instrument has not been sufficiently sensitive to detect a trace of the part. Such essential zeros occur in many compositional situations.

1. In *household budget patterns*, where some households may spend nothing on such commodity groups as tobacco, alcohol, entertainment, over the period of observation.

2. In *time budgets*, where the individual subject may not take part in one or more of the assigned activities during the recording period.

3. In palaeontology, for example in *pollen zonation studies*, where at some levels a number of different varieties of the pollen fossils may be absent.

4. In *ecological abundance studies* where the abundances of different species are often expressed as percentages, and for some regions some species are absent..

Experience suggests that when faced with such compositional data the analyst, realising that the standard logratio transformation methodology cannot be applied (You can't take the logarithm of zero), may consider whether a solution is perhaps to replace the zeros by some small proportion. But how, and would such a replacement strategy not be rather arbitrary in the case of essential zeros? Another common way out may be to amalgamate parts in such a way that all zeros are eliminated. This may be a sensible solution if the parts amalgamated are similar in character and where the zeros may have arisen because of the definition of an unnecessarily fine division of parts. Often such steps to overcome the so-called zero problem are far from satisfactory and the analyst is left with a compositional data set that appears to defy proper analysis.

This view is not to detract from the use of replacement strategies for rounded or trace zeros. For such situations the subcompositionally coherent replacement strategy of Martin- Fernández, Barceló-Vidal and Pawlowsky-Glahn (2000) and Fry, Fry and McLaren. (2000) is appropriate, especially when these are subjected to a sensitivity analysis to detect how stable the inference is with respect to a reasonable range of replacement values.

One of the problems with compositional data sets with essential or structural zeros is that the owner of the data quite often has no well defined hypothesis to test or indeed any obvious inference purpose for the data. It is then the first task of the compositional data consultant to try to elicit the precise aim of the study. Sometimes when the appropriate question is formulated the zero problem disappears. For example. the compositional data set of Aitchison (1986, Data 35, Fig. 11.6) consists of three-part compositions (p, q, r) consisting of relative proportions of a predator species P and two prey species Q and R at 25 different sites. At 10 of these sites no predators are recorded, so there is an essential zero problem. But when we realise that the question being asked of the data is whether the presence of predators affect the relative abundance of the two types of prey we realise that we can answer this question by testing the hypothesis that the (Q, R) subcompositions of the 15 full compositions (with predator present) and the 10 two-part (Q, R) compositions with predator absent have identical distributions.

In this note we attempt to set a fairly general framework within which problems of essential zeros may be addressed. We make no claim that this framework will allow the resolution of all such problems for we are well aware of the diversity of the nature of essential zeros. But we hope that it may provide a stimulus for further modelling along similar lines.

In our view the stumbling block with essential zeros has been the conception that the data are compositional vectors and that somehow their analysis must fall within existing compositional data methods. If, however, we consider the above four

examples we may be directed to a different view and a different approach to modelling. In the household budget pattern example a household of non-smoking, non-drinking members is aware, even before it embarks on its monthly spending, that it will spend nothing on these commodity groups and will allocate its expenditure over the remaining commodity groups. In the time budget situation the individual may be aware at the start of the recording period that he has no skills in some of the activities and so allocate the whole time period over the other activities. In the pollen zonation study weather conditions over the period may have annihilated some species making way for other species to distribute themselves with less competition. In the ecological abundance studies some regions may not contain particular vegetation which is necessary to the life of some species.

2. Modelling

From the above argument it seems sensible to build up a model in two stages, the first determining where the zeros will occur and the second how the unit available is distributed among the non-zero parts. In terms of compositional data sets consisting of N D-part compositions we can visualise the data presented as two matrices. The first is an *incidence matrix I*, for example

$$I = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

where the first row with 1 in each position indicates a full 4-part composition, the 0's in the first and third positions in the second row indicate essential zero values for the components of the first and third parts of the second composition, and so on. The second matrix is a component matrix in which the component values associated with the 1's in the incidence matrix X are recorded, for example

$$X = \begin{bmatrix} 0.32 & 0.15 & 0.42 & 0.21 \\ 0 & 0.44 & 0 & 0.56 \\ 0.11 & 0.37 & 0.52 & 0 \\ 0 & 0..26 & 0.45 & 0.29 \\ 0.57 & 0 & 0 & 0.43 \end{bmatrix}$$

The independent binary model

As a first stage define a very simple model for the generation of such data in terms of a D-variate binary density function

$$b(u\,|q) = \prod_{i=1}^{D} q_i^{u_i}(1-q_i)^{1-u_1} \qquad \{u_i = 0,1 \quad (i = 1,\ldots,D)\} \qquad (2.1)$$

which is simply the density function of D independent binomial variables with different 'success' probabilities.

At the second stage nonzero components are generated by subcompositional density functions based on a logistic normal density function $y(x\,|x,\mathrm{T})$ associated with the full composition; in other words x follows a $L^D(x,\mathrm{T})$ distribution. For ease of exposition we have adopted the parametrisation based on the centre x and variation matrix T.

Let us now show how we combine these two concepts to provide a probabilistic mechanism for producing a composition with essential zeros. Suppose that the D-variate binary density function yields an observational vector u. Let

$$\begin{aligned} J(u) &= \{i : u_i = 1\}, \quad D_1 = \sum_{i=1}^{D} u_i, \\ K(u) &= \{i : u_i = 0\}, \quad D_0 = 1 - D_1. \end{aligned} \qquad (2.2)$$

Thus $J(u)$ is a D_1-vector containing the serial numbers of the non-zero parts, $K(u)$ is the D_0-vector of serial numbers of the essential-zero parts. Denote by $x_{J(u)}$ a D_1-part composition with parts labelled by the elements of $J(u)$. Also denote by $x_{J(u)}$ the D_1-part subcomposition formed from the $J(u)$ parts of x, and by $\mathrm{T}_{J(u)}$ the $D_1 \times D_1$ variation matrix formed from the $J(u)$ rows and columns of T. Then the non-zero components arise as a compositional vector from $L^{D_1}(x_{J(u)}, \mathrm{T}_{J(u)})$.

The dependent binary model

An obvious criticism of the above model is the independence of the binomial outcomes. This can easily be remedied by imposing a hierarchical prior on the binomial parameters. Probably the simplest way of doing this is through a simple reparametrisation of the binomial probability parameters, by

$$q_i = \exp(l_i) / \{\exp(l_i) + 1\} \quad (i = 1,\ldots,D) \qquad (2.3)$$

The hierarchical prior can then be taken to be a D-dimensional normal prior with density function $f^D(l\,|m,\Sigma)$.

Another criticism of such modelling may be that all the subcompositional distributions are based on one full compositional distribution. This might not be the case; for example, non-tobacco, non-alcohol spenders may act differently from non-travel spenders. But if this is thought to be a possibility, even a hypothesis, then the groups should be separated and the hypothesis of no difference tested within a lattice

with maximum model in which different full compositional distributions are allowed for each group.

3. Statistical aspects

Fry, Fry and McLaren (200) came close to considering such modelling but seemed to dismiss it as too complex. There are certainly questions of estimability of parameters and awkwardness of computation but there are no new statistical principles required.

Let us first consider the model where compositional aspects are based on the assumption of a common full $L^D(x,\mathrm{T})$ distribution. First we make an intuitive note on the estimability of parameters. The possibility of estimating the parameters x,T of the full compositional distribution clearly depends on the incidence matrix *I*. A simple test is to form $I^T I$. If all the non-diagonal elements are at least 2 then estimation is possible.

For most problems in parametric statistical inference a crucial step is the formation of the likelihood, given the data. A compositional data set here consists of a $N \times D$ incidence matrix I with *n*th row u_n $(n = 1, \ldots, N)$ together with the relevant $N \times D$ compositional data matrix X with *n*th row effectively the subcomposition $x_{J(u_n)}$ in the notation of Section 2. For the independent binomial model the likelihood can then be written as

$$L(q,x,\mathrm{T}|data) = \prod_{n=1}^{N} p(u_n|q)y\,(x_{J(u_n)}|\,x_{J(u_n)},\mathrm{T}_{J(u_n)}), \qquad (3.1)$$

where

$$p(u_n|q) = \prod_{i=1}^{D} q_i^{u_{ni}}(1-q)^{1-u_{ni}}\ . \qquad (3.2)$$

For the dependent binary modelling case the $p(u_n|q)$ of (3.1) is replaced by

$$.\ p(u_n|m,\Sigma) = \int_{R^{D-1}} \prod_{i=1}^{D} \frac{\exp(l_i u_{ni})}{1+\exp(l_i)}\ j\,(l|p,\Sigma)dl\ . \qquad (3.3)$$

It is worth noting that the dependent binomial model reduces to the independent binomial model when $\Sigma = 0$.

An important point to note in both forms of likelihood is that the binomial parameters, either q or (m,Σ), and the compositional parameters (x,T) are separable, so that inference questions about these parameters can be treated separately.

With explicit expressions for the likelihood the remaining problems are computational, how to compute maximum likelihood estimates and thereby generalised likelihood ratio test statistics for any hypothesis under test. Apart from the logistic problem of identifying the different subcompositions within the likelihood a

main problem is the integrals involved in (3.3). We believe that these are most easily tackled by an MCMC approach and are currently working on programs for this purpose.

This model-building and its implications for statistical inference for compositional data sets with essential zeros will be illustrated by a number of examples at CODAWORK03.

References

Aitchison, J.,1986., The Statistical Analysis of Compositional Data: Chapman and Hall, London, 416 p. Reprinted (2003) with extra material by The Blackburn Press.

Fry, J.M., Fry, T.R.L and McLaren, K.R., 2000, Compositional data analysis and zeros in micro data: Appl. Economics, v. 32, p. 953-959.

Martin- Fernández, J. A. Barceló-Vidal, C. and Pawlowsky-Glahn, V., 2000,.Zero replacement in compositional data sets, *in* Kiers, H., Rasson, J., Groenen, P and Shader, M., eds., Studies in Classifiacation, Data Analysis and Knowledge Organisation. Proceedings of 7th Conferencee of the International Federation of Classification Societies p. 155-160. Springer-Verlag, Berlin.