Modelling structural zeros in compositional data

Dr John Bacon-Shone
Director, Social Sciences Research Centre
The University of Hong Kong
Pokfulam Road
Hong Kong
Email: johnbs@hku.hk
Tel: (852) 2859-2412

Abstract

This analysis was stimulated by the real data analysis problem of household expenditure data. The full dataset contains expenditure data for a sample of 1224 households. The expenditure is broken down at 2 hierarchical levels: 9 major levels (e.g. housing, food, utilities etc.) and 92 minor levels. There are also 5 factors and 5 covariates at the household level. Not surprisingly, there are a small number of zeros at the major level, but many zeros at the minor level. The question is how best to model the zeros. Clearly, models that try to add a small amount to the zero terms are not appropriate in general as at least some of the zeros are clearly structural, e.g. alcohol/tobacco for households that are teetotal. The key question then is how to build suitable conditional models. For example, is the sub-composition of spending excluding alcohol/tobacco similar for teetotal and non-teetotal households? In other words, we are looking for sub-compositional independence. Also, what determines whether a household is teetotal? Can we assume that it is independent of the composition? In general, whether teetotal will clearly depend on the household level variables, so we need to be able to model this dependence. The other tricky question is that with zeros on more than one component, we need to be able to model dependence and independence of zeros on the different components. Lastly, while some zeros are structural, others may not be, for example, for expenditure on durables, it may be chance as to whether a particular household spends money on durables within the sample period. This would clearly be distinguishable if we had longitudinal data, but may still be distinguishable by looking at the distribution, on the assumption that random zeros will usually be for situations where any non-zero expenditure is not small.

While this analysis is based on around economic data, the ideas carry over to many other situations, including geological data, where minerals may be missing for structural reasons (similar to alcohol), or missing because they

occur only in random regions which may be missed in a sample (similar to the durables).

Acknowledgements:

The ideas in this paper owe a considerable amount to discussions with John Aitchison last year.

Background

Early attempts to "solve" the problem of zeros in log-ratio compositional data analysis (Aitchison, Bacon-Shone) took the approach of trying to modify the log mappings at the extremes by adding small amounts to zeros or by using rank transformations. This can be seen in geological terms as an implicit assumption that there are no true zero elements, but just that some elements are below the measurable limits of the instrument used. This approach is limiting in that there may be good reasons why a particular component is wholly absent. This paper focuses on household expenditure data, but similar arguments apply in the case of geological data, and doubtless in many other application areas as well. In the expenditure case, zeros can arise in at least three contexts. Firstly, the expenditure may be too low to be measured, for example, expenditure is measured to the nearest dollar, and the expenditure is less than 50 cents. Secondly, expenditure may be a rare event, such as the purchase of a cooker or fridge, so that even if the expenditure covers a longer period for electrical appliances, it is quite feasible that no expenditure takes place for many households. Thirdly, there may be types of expenditure which some households never make, while others make them frequently. An example would be alcohol or tobacco, where the entire household may abstain completely. If we prefer geological analogies, the second and third situations could reflect that some minerals tend to be very unevenly distributed, while other minerals cannot occur together for reasons linked to geochemistry. The second and third situations can both be considered to be structural zeros, but the reasons can be linked either to the randomness of the sampling process, or to the nature of the objects being sampled. While the distinction may seem arbitrary, I would argue that there is an important difference. If the zeros are linked to the sampling process, this would suggest that there may be no clear difference between the samples with zeros and those without, either in terms of the subcomposition excluding the zeros or in terms of covariates. On the other hand, if there is some real difference between the nature of samples with zeros and those without, we should be able to distinguish between them either in terms of subcompositional differences or in terms of covariates. In the expenditure case, we can express the idea in the following way: we might expect teetotal households to be fundamentally different from households that drink, but we would not expect to be able to distinguish clearly between those who bought a fridge recently and those who did not. This paper reflects an early stage of research into trying to tease out these distinctions with models that are meaningful both in the statistical sense and in the context of the data in economics, geology etc. I apologize for concentrating on an economic dataset at a conference attended by geologists, but I believe that the core statistical problem is the same.

Conditional modeling

If we initially consider a single level of hierarchy, there are two key questions:

Firstly, how do we model the pattern of zeros for multiple components and secondly, how do we model the composition, conditional on a particular pattern of zeros.

Structural modeling of zeros

Clearly, this can be modeled using multivariate logistic or probit models. However, as we have observed with compositional data, logistic models may have too strong a pattern of independence to be practical. On the other hand, they are simple to fit and parsimonious, which is valuable when we may have relatively few occurrences of zeros for some components. Probit models allow many more parameters, but they require much more data to provide useful estimates. We will illustrate these tradeoffs using the household expenditure data modeled at the two different levels of hierarchy and with differing numbers of explanatory covariates and factors.

Conditional subcompositions

If there is no linkage between the zeros and the remaining subcompositions, then the model takes a relatively simple form, at least when there is only a single zero so we can use a product of a bivariate composition and the remaining composition. One interesting question is whether the bivariate composition, after excluding zeros, shows much chance of small values.

This may provide a diagnostic check on whether the zeros could indeed be reflecting measurement limits rather than structural zeros. It is necessary to include the covariates when trying to detect a difference between the subcompositions, as any apparent difference may disappear after controlling for the covariates. In the expenditure case, housing type (public vs private) and income decile both have large impacts on the compositions in general, so it is important to control for their effect. If there is a linkage between the zeros and the subcompositions, then the interesting question is to find a model that captures that dependence in a concise way. We will illustrate situations that show both independence and dependence.

Conclusions

While there is still much more careful analysis and modeling to be done, it is clear that the separation of the modeling into the two parts allows much more useful information to be extracted from compositions, in situations where the zeros may contain potentially important information. This seems to remove the major remaining barrier to the widespread use of log-ratio models for compositional data, particularly in geological and expenditure models.