Imprecise Compositional Data Analysis: Alternative Statistical Methods

Michael Smithson

The Australian National University

2-6 July 2019 /SIPTA 2019

Statistical methods for analyzing imprecise compositional data

Compositional data must sum to a constant value, e.g., probabilities that must sum to 1.
Statistical methods for analyzing imprecise compositional data are relatively under-developed.

Two alternative approaches are considered here:

- Log-ratio transforms (well-established, starting with Aitchison, 1982)
- Dirichlet regression (also well-established, including the IDM)
- Probability-ratio transforms (under development by the author)

Compositional data

Given a composition consisting of K parts, suppose that we have N collections of points in the K-simplex, $0 \leq \pi_{ki}^{(j_i)} \leq 1$, for $k = 1, \ldots, K$ and $i = 1, \ldots, N$, such that for each i they sum to 1 across the k. For the i^{th} collection, there are J_i points, indexed by the bracketed j_i superscript.

Our main topic is how to connect these collections with regression or generalized linear models (GLMs) that treat them as dependent variables.

Basics

The log-ratio transform method maps data from the simplex to an unrestricted vector space, via the logit transform of odds. Suppose the K^{th} composition part is the part of the composition against which we would like to compare the other parts. Then Aitchison's "additive log-ratio" transform would yield

$$\eta_{ki}^{(j_i)} = \log\left(\left(\frac{\pi_{ki}^{(j_i)}}{1-\pi_{ki}^{(j_i)}}\right) \Big/ \left(\frac{\pi_{Ki}^{(j_i)}}{1-\pi_{Ki}^{(j_i)}}\right)\right), \tag{1}$$

for $k = 1, \ldots, K-1$. The $\eta_{ki}^{(j_i)}$ are considered as continuous random variables on the real line, and therefore may be analysed with appropriate statistical methods for such variables.

Advantages

The log-ratio framework enjoys several attractive properties that account for its popularity.

- Subcompositional coherence means that the inferential outcomes of an analysis of any subcomposition should remain the same for that analysis in the entire composition.
- Permutation invariance guarantees that outcomes remain the same regardless of the ordering of the components in a composition.
- It is straightforward to use because the log-ratios can be analyzed with conventional methods such as linear regression with Gaussian errors.

Disadvantages

The log-ratio framework also has some limitations:

- It cannot deal with zeros in the data.
- It is unable to extend to non-Gaussian distributions without adding more parameters.
- Dispersion is routinely ignored in the log-ratio framework.

Basics

Dirichlet regression models are a natural and popular choice for modeling compositional data. These models have two main limitations.

- The marginal distributions are beta distributions sharing the same precision parameter, so all parts of the composition must have the same submodel for their precisions. This limits its ability to model multivariate heteroskedasticity.
- A single Dirichlet distribution can model only negative associations among the variables, although this restriction may be relaxed when covariates are modeled or other kinds of mixture models are employed.

Basics

Rather than taking logs of relative odds, we take the corresponding relative probabilities and model them.

Turning once again to our example with the K^{th} category as the base, the relevant probability ratios are

$$\nu_{ki}^{(j_i)} = \pi_{ki}^{(j_i)} \Big/ \left(\pi_{ki}^{(j_i)} + \pi_{Ki}^{(j_i)} \right), \tag{2}$$

for $k = 1, \ldots, K-1$. The $\nu_{ki}^{(j_i)}$ are random variables in the unit hypercube, and the marginals may be modeled by any distribution whose support is the unit interval (0,1). The dependency structure may be modeled using copulas.

Advantages

The advantages of the probability-ratio method are:

- The probablity-ratio approach includes the logistic-normal distribution but also other more flexible two-parameter distributions such as the beta and cdf-quantile family.
- Unlike the Dirichlet model, each marginal distribution can have a unique precision parameter, thereby able to model multivariate heteroskedasticity.
- Modeling dispersion is naturally done in the probability-ratio framework.
- It possesses both permutation invariance and subcompositional coherence.
- Zeros can be dealt with via hurdle models.
- The use of copulas enables flexible modeling of dependency structures, separately from the marginal structures.

Conclusions

- A new “probability ratio” approach to modeling compositional data has been proposed that can complement the well-established log-ratio approach.
- Both of these provide an alternative to Dirichlet models for imprecise compositional data.
- Much remains to be done in evaluating their merits, for instance their relative sensitivities to noise or other sources of imprecision.
- The probability ratio approach shows promise in overcoming some of the limitations of the other two approaches.

The End

Thanks!