Proceedings of Machine Learning Research 103:364–366, 2019 ISIPTA 2019

Imprecise Compositional Data Analysis: Alternative Statistical Methods

Michael Smithson MICHAEL.SMITHSON@ANU.EDU.AU
Research School of Psychology, The Australian National University, Canberra, Australia

Abstract

This paper briefly describes statistical methods for analyzing imprecise compositional data that might be elicited from approximate measurement or from expert judgments. Two alternative approaches are discussed: Log-ratio transforms and probability-ratio transforms. The first is well-established and the second is under development by the author. The primary focus in this paper is on generalized linear models for predicting imprecise compositional data.

Keywords: compositional data, imprecise data, beta distribution, cdf-quantile distribution, general linear model, copula

1. Introduction

Data that must sum to a constant value are known as "compositional", and coherent probability assignments are a typical example. 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. These collections may have been derived from credal sets, coherent lower and/or upper previsions, or sets of desirable gambles.

For the most part, this paper will set aside the details of these collections. Instead, our main topic is how to connect these collections with regression or generalized linear models (GLMs) that treat them as dependent variables. We further assume that the modeler has a set of covariates (whether continuous or categorical) to be used as predictors in such a GLM.

Although methods for statistical analysis of precise compositional data are well-established, their application to imprecise compositional data does not seem to have been developed. This short paper initiates investigations into this application. There are two well-established approaches to modeling compositional data: Dirichlet regression and log-ratio regression. A third, probability-ratio regression, is under development by the author. We will focus primarily on comparing the log-ratio transforms and probability-ratio transforms because these approaches are directly related. The primary focus in this paper is on generalized linear models for predicting imprecise compositional data.

2. Log-Ratio Transform Method

The log-ratio transform method [1] maps data from the simplex to an unrestricted vector space, via the logit transform of odds-ratios. For instance, 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

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

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, such as multi-level linear regression with multivariate Gaussian errors. Predictors in these regression models may be introduced at either the collection level, the compositional part level, or even at the j_i level (e.g., for distinguishing between lower and upper probabilities).

The log-ratio framework enjoys several attractive properties that account for its popularity. Chief among these are subcompositional coherence and permutation invariance. 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.

Although it is straightforward to use, the log-ratio framework has important limitations. First, it is unable to extend to non-Gaussian distributions without adding more parameters (e.g., via skew-normal distributions). Second, dispersion is routinely ignored in the log-ratio framework, despite its obvious relevance to doubly-bounded and sum-constrained random variables. Third, the logit transform is unsuited for dealing with zeroes or ones in the data. The popular approach of adding a small constant to the zeroes or subtracting it from the ones introduces difficulties regarding the arbitrary choice of its magnitude and sensitivity to that choice.

3. Probability-Ratio Transform Method

The limitations imposed by the assumption of multivariate normality, on the other hand, motivate us to seek alternatives to the log-ratio transform. We therefore now consider probability-ratios. 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 $k = 1, \ldots, K-1$. Any of the log-ratio transforms has a probability-ratio analogue, so this approach inherits both subcompositional coherence and permutation invariance. Moreover, the probability-ratio approach includes structures not found in the log-ratio literature, such as the stick-breaking construction (See [6] for a demonstration). It is less clear at the present stage of development whether distance-based techniques (e.g., certain types of cluster analysis) employed in the log-ratio framework translate into the probability-ratio framework. Smithson (2019) identifies conditions under which Euclidean distances in the unit hypercube between probability-ratios do not have the same rank-order as their log-ratio Euclidean distance counterparts.

The $\nu_{ki}^{(j_i)}$ may be modeled using any distribution whose support is the open unit interval (0,1). Many flexible two-parameter distributions are available, such as the beta distribution and the CDF-quantile family [7]. GLMs may be constructed following the same structure as presented in [7], although these are not GLMs in the formal sense because the distributions employed here are not members of the exponential family. In fact, if necessary, a different marginal distribution can be specified for each probability-ratio component. Note that all of this added flexibility in the marginal distribution models is achieved with the same number of parameters as in the log-ratio setup, i.e., two-parameter marginal distributions and the same regression models. Furthermore, the inclusion of the CDF-quantile distribution family enables flexible quantile regression models that are unavailable in the log-ratio framework.

Dispersion typically is not modeled with covariates in the log-ratio literature, although this could be done with a bit of extra work. However, explicit dispersion submodels are readily accommodated in both beta-distribution [8] and CDF-quantile distribution regression models [7]. Preliminary work using real data-sets has already indicated that modeling dispersion adds important insights to compositional data analysis and can correct location model misspecification [6].

The dependency structure may be modeled via random-effects models for the marginal parameters [9]. Alternatively, the dependency structure can be modeled with copulae or copula vines [5]. The conventional two-stage maximum-likelihood estimation procedure for this enables the marginal distributions to be modeled separately from the dependency structure.

Finally, the zeroes problem can be dealt with via hurdle models (in the log-ratio as well as probability-ratio setting). Shou and Smithson [5] have implemented hurdle models in their cdfquantreg package. Given that the beta regression GLM and many of the CDF-quantile distributions' GLMs use link functions that are defined only on (0,1), a hurdle model rather than a zero-inflated model is technically the relevant alternative.

4. Dirichlet Methods

Dirichlet regression models are a natural and popular choice for modeling compositional data [4]. These models have two main limitations. First, the Dirichlet distribution's 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. On the other hand, a probability-ratio model with beta marginals can incorporate a unique precision submodel for each marginal distribution. Second, 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 [3], the probability-ratio approach does not have this limitation even for the single-distribution case.

Promising imprecise Dirichlet process models [2] have been developed and applied to non-parametric Bayesian versions of hypothesis tests such as Wilcoxon tests. A systematic comparison between log-ratio, probability-ratio, and imprecise Dirichlet approaches to modeling imprecise compositional data has yet to be completed and so is beyond the scope of this paper.

In conclusion, we have briefly surveyed practical statistical methods for analyzing the kinds of imprecise compositional data that may arise from measurement or expert judgments, focusing on two related methods: The well-established log-ratio approach, and a new "probability ratio" approach. Much remains to be done in evaluating their respective 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 log-ratio method, and both methods can complement one another.

References

[1] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[2] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pages 1026–1034, 2014.

[3] Rafiq H Hijazi and Robert W Jernigan. Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91, 2009.

[4] Marco J Maier. Dirichletreg: Dirichlet regression for compositional data in r. *Research Report Series*, 125, 2014.

[5] Yiyun Shou and Michael Smithson. cdfquantreg: An r package for cdf-quantile regression. *Journal of Statistical Software*, 88:1–30, 2019.

[6] Michael Smithson. A new approach to compositional data analysis. The Australian National University, Canberra, Australia, 2019.

[7] Michael Smithson and Yiyun Shou. Cdf-quantile distributions for modelling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology*, 70(3):412–438, 2017.

[8] Michael Smithson and Jay Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, 2006.

[9] Jay Verkuilen and Michael Smithson. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, 37(1):82–113, 2012.