Three approaches to supervised learning for compositional data with pairwise logratios

Coenders, Germà
germa.coenders@udg.edu
Universitat de Girona
Girona, Spain

Greenacre, Michael
michael.greenacre@upf.edu
Universitat Pompeu Fabra
Barcelona, Spain

Statistično društvo Slovenije

17th Applied Statistics 2021

International Conference

Introduction

The usual approach to Compositional Data (CoDa) is to use existent statistical methods on transformed data. Logarithms of ratios are the standard transformation.

- The simplest case is a logratio between only two components.
- *D*-1 pairwise logratios contain the whole information for a *D*-part composition if each part participates in at least one logratio (Greenacre, 2019).

When the number of parts is large (sometimes even larger than *n*), some form of selection of fewer than *D*-1 logratios is convenient or even necessary.

Greenacre (2018; 2019) developed an unsupervised learning method based on a stepwise selection of the pairwise logratios that explain the largest possible variance of the composition itself.

In this paper we are interested in supervised learning, i.e. selecting logratios which best explain or predict a target variable.

We present three alternative stepwise supervised learning methods to select the best set of predive pairwise logratios in a generalized linear model. The dependent variable can be of any kind supported by generalized linear models, including binary (Bernouilli), continuous (normal) or count (Poisson).

We use the forward selection method as follows:

- In the first variant, any pairwise logratio is eligible to belong to the model.
- In the second variant only logratios whose pairs of parts do not overlap are eligible.
- The third variant aims at selecting a subset of parts (i.e. a subcomposition).

The methods will be available in the R package easyCODA (Greenacre, 2018).

This paper adds to the literature on variable selection in explanatory compositions.

- Regularized regression, including Lasso and related methods (e.g. Combettes & Müller, 2020; Lin et al., 2014; Louzada et al., 2019; Lu et al., 2019; Shi et al., 2016; Susin et al., 2020)
- The discriminative balance approach (Quinn and Erb 2020).
- The selbal approach (Rivera-Pinto et al., 2018; Susin et al., 2020).

We include an application predicting Crohn's disease from the microbiome composition.

Method

Compositions and their logratios

A *D*–part composition can be defined as an array of strictly positive numbers for which ratios between them are considered to be relevant:

$$\mathbf{x} = \left(x_1, x_2, ..., x_D \right) \text{ with } x_j > 0,\ j = 1, 2, ..., D.$$

The common approach to compositional data is to use standard analysis methods after transforming the data with logratios.

In the additive logratio transformation (alr), *D*-1 logratios are computed with a common denominator which can be any by permutation:

$$\log\left(\frac{x_j}{x_D} \right) = \log\left(x_j \right) - \log\left(x_D \right), \text{ with } \ j = 1, 2, ..., D-1.$$

This can easily be generalized to any of the possible $D(D\text{-}1)/2$ pairwise logratios between any two components:

$$\log\left(\frac{x_j}{x_k}\right) = \log\left(x_j\right) - \log\left(x_k\right), \qquad j = 2,3,\ldots,D;\ k = 1,2,\ldots,j-1,$$

although the inherent dimensionality of a composition is D-1. This means that the $D(D\text{-}1)/2$ pairwise logratios are mutually redundant.

Greenacre (2018; 2019) showed that D-1 pairwise logratios such that each part participates in at least one logratio, are always non-redundant.

Even D-1 logratios are too many when D is large, and the aim of this paper is to select a smaller optimal subset.

The most general form of a logratio is the log-contrast:

$$\begin{pmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_D\end{pmatrix}\begin{pmatrix}\log(x_1)\\ \log(x_2)\\ \vdots\\ \log(x_D)\end{pmatrix}, \quad \text{with} \sum_{j=1}^{D}\alpha_j = 0.$$

A pairwise logratio is a special case with one value in vector $\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_D)$ equal to 1, one equal to -1 and the remaining equal to zero.

Other ways of computing logratios have been suggested in the literature with the requirement of orthogonality of the α vectors of any two log-contrasts (e.g., the so-called isometric log ratios). This has implications for logratio interpretation.

It must be noted that the α vectors two pairwise logratios are mutually orthogonal if they do not overlap, i.e., if no part participates in both logratios. e.g. the logratios of x_2 over x_1 and x_4 over x_3 have the orthogonal α vectors (-1,1,0,0) and (0,0,-1,1).

Stepwise regression

Logratio selection in linear or generalized linear models belongs to the domain of statistical learning, and, more precisely, supervised statistical learning.

Stepwise regression is one of the earliest forms of supervised statistical learning.

The forward selection method of stepwise regression is especially interesting due to its ability to handle any number of logratios, even if D-1 is larger than n.

- In the first step, the algorithm selects the logratio leading to the lowest model deviance.
- In the second and subsequent steps, the algorithm adds the logratio leading to the largest reduction in the deviance, one logratio at a time.

Since adding a logratio always decreases the deviance, a stopping criterion is needed in order not to reach the trivial solution with D-1 logratios.

A penalty term is added to the deviance. For instance, the Bayesian Information Criterion (BIC), minimises (m is the number of model parameters):

$$\text{deviance} + \log(n)\, m.$$

Another possibility is to set such a penalty that logratios are added only if statistically significant.

Since D-1 non-redundant logratios are simultaneously being tested, α has to account for multiple testing. A popular criterion is the Bonferroni correction which uses the $\chi^2_{1\text{d.f.}}$ quantile with a tail area $\alpha/(D\text{-}1)$.

For instance, if D-1=10, and α=0.05, the tail area is 0.005 and we minimise:

$$\text{deviance} + 7.87944m.$$

Caveats

- It is well-known that estimates and t-values are biased upwards in stepwise regression, because the variables included are those with the highest values for the particular sample. This requires testing the final model with a fresh cross-validation sample.

- Even after cross-validation, many models may have a similar fit to the data while the procedure has only found one of them. This makes stepwise regression fit for predictive and exploratory purposes but not for theory testing.

Having said this, many learning methods in compositional data use stepwise algorithms (Greenacre, 2018; 2019; Hron et al., 2013; Rivera-Pinto et al., 2018; Susin et al, 2020).

Expert knowledge

Expert knowledge is a welcome complement to data-driven statistical learning and can help overcome the limitations of the stepwise method. The user should be able to:

- Force certain theoretically relevant logratios into the equation.
- Force certain theoretically relevant non-compositional control variables into the equation.
- Choose among logratios with about the same significance or BIC improvements.

The suggested algorithms 1. Unrestricted search

The final solution may be a combination of overlapping and non-overlapping pairs of parts. Imagine we have *D*=7 parts (A,B,C,D,E,F,G). The stepwise algorithm might select log(B/A), log(C/B) and log(G/F). The pairs B/A and C/B overlap.

The interpretation of models with overlapping logratios, whose α vectors are not orthogonal, is not intuitive (Coenders & Pawlowsky-Glahn, 2020; Hron et al., 2021).

The rule "keeping all other predictors constant" has to be applied with care. The effects of overlapping logratios do not correspond to trade-offs between the numerator and denominator parts.

- e.g. The coefficient associated to log(B/A) is interpreted as increasing B at the expense of decreasing A, while keeping the remaining ratios constant. Keeping the ratio of C over B constant means that C increases by the same factor as B. Thus, the coefficient is interpreted as increasing B and C together at the expense of decreasing A.

This algorithm selects the logratios which contribute most to predictive power. If the research purpose is only prediction, this method may still be the best choice.

The suggested algorithms 2.Search of non-overlapping logratios

Under this approach, at most $D/2$ non-overlapping logratios are selected.

For instance, if log(B/A) is chosen in the first step from composition (A,B,C,D,E,F,G), the only feasible choices for the second step are log(D/C), log(E/C), log(F/C), log(G/C), log(E/D), log(F/D), log(G/D), log(F/E), log(G/E) and log(G/F).

The limitation to at most $D/2$ logratios may yield a lower predictive power, but may be welcome for high dimensional compositions.

Non-overlapping logratios have orthogonal α vectors by construction. For this reason, their effects on the dependent variable can be interpreted in a straightforward manner in terms of trade-offs between only the numerator and the denominator parts.

The suggested algorithms 3.Search for completely overlapping pairwise logratios (search for a subcomposition with alr)

This algorithm draws from the fact that a subcomposition with *k* parts can be represented by *k*-1 pairwise logratios as long as each part participates in at least one log-ratio (Greenacre, 2019) and that any logratio selection fulfilling this criterion has identical predictions and goodness of fit (Pawlowsky-Glahn & Coenders, 2020).

This includes the additive logratios (alr) with any part of the subcomposition in the denominator and the remaining *k*-1 parts in the numerator This leads to a shorter search of candidate logratios and makes interpretation easier.

This algorithm searches for the *k*-part subcomposition with the highest explanatory power by means of the forward stepwise selection of alr, whose denominator part is determined by the logratio entered in the first step. Each step brings both an additional logratio in the equation and an additional part in the subcomposition.

In our (A,B,C,D,E,F,G) example, if log(B/A) is chosen in the first step [(A,B) subcomposition], the feasible choices for the second step are log(C/A), log(D/A), log(E/A), log(F/A), log(G/A). Selecting, for instance log(G/A) leads to subcomposition (A,B,G).

alr are do not have orthogonal α vectors. The effects are not interpretable as trade-offs between pairs of parts (Hron et al., 2021) but can be easily interpreted by reexpressing the equation as a log-contrast. From equation:

$$Y = b_0 + b_1 \log\left(\mathrm{B}/\mathrm{A}\right) + b_2 \log\left(\mathrm{G}/\mathrm{A}\right) + e,$$

the log-contrast is:

$$\cdots b_1 \log\left(\mathrm{B}\right) + b_2 \log\left(\mathrm{G}\right) + \left(-b_1 - b_2\right)\log\left(\mathrm{A}\right)\cdots$$

Interpretation: Increasing the parts with positive log-contrast coefficients at the expense of decreasing the parts with negative log-contrast coefficients leads to an increase in the dependent variable.

Illustration

We reanalyse one of the data sets used by Rivera-Pinto et al. (2018) to relate microbiome and Crohn's disease. Patients with Crohn's disease (n=662) are coded as 1, and those without any symptoms (n=313) as 0. The operational taxonomic unit (OTU) table was agglomerated to the genus level, resulting in a matrix with 48 genera.

Since the dependent variable is binary, an appropriate member of the generalized linear model family is a logit model.

The stopping criterion is set to ensure significance at α=5% with the Bonferroni correction.

Unrestricted stepwise search (BIC 932.5519)

	Estimate	p-value
g_Roseburia/g_Streptococcus	-0.3022	<0.0001
f_Peptostreptococcaceae_g/g_Dialister	-0.1618	<0.0001
g_Bacteroides/g_Dorea	-0.2393	<0.0001
g_Prevotella/g_Aggregatibacter	-0.1008	<0.0001
g_Adlercreutzia/g_Lachnospira	0.1158	<0.0001
o_Lactobacillales_g/g_Streptococcus	0.1482	<0.0001
g_Oscillospira/o_Clostridiales_g	0.1688	<0.0001
g_Sutterella/g_Bilophila	0.0873	0.0004

The unrestricted approach leads to the lowest BIC and is preferable for prediction, but is not interpretable in an intuitive manner. Bold-faced logratios overlap

Non-overlapping search (BIC 939.6419)

	Estimate	p-value
g_Roseburia/g_Streptococcus	-0.2444	<0.0001
f_Peptostreptococcaceae_g/g_Dialister	-0.1702	<0.0001
g_Bacteroides/g_Dorea	-0.2272	<0.0001
g_Prevotella/g_Aggregatibacter	-0.1087	<0.0001
g_Adlercreutzia/g_Lachnospira	0.1139	<0.0001
o_Clostridiales_g/f_Ruminococcaceae_g	-0.2553	<0.0001
g_Sutterella/g_Bilophila	0.0844	0.0006
g_Faecalibacterium/g_Oscillospira	-0.1088	0.0011

Interpretation: The incidence of Crohn's disease is significantly associated to:

- Increases in the relative importance of taxon g_Streptococcus at the expense of decreases of the relative importance of taxon g_Roseburia.
- Increases in the relative importance of g_Adlercreutzia at the expense of decreases of the relative importance of g_Lachnospira….

alr search (BIC 967.4244)

	Estimate	p-value
g_Roseburia/g_Streptococcus	-0.3375	<0.0001
g_Dialister/g_Streptococcus	0.1407	<0.0001
f_Peptostreptococcaceae_g/g_Streptococcus	-0.2065	<0.0001
o_Lactobacillales_g/g_Streptococcus	0.1420	0.0003
g_Bacteroides/g_Streptococcus	-0.2792	<0.0001
g_Dorea/g_Streptococcus	0.2021	<0.0001
g_Adlercreutzia/g_Streptococcus	0.1511	<0.0001
g_Aggregatibacter/g_Streptococcus	0.1378	<0.0001
g_Prevotella/g_Streptococcus	-0.0920	0.0004

Interpretation as a log contrast:

+0.2021log(g_Dorea)+0.1511log(g_Adlercreutzia)
+0.1420log(o_Lactobacillales_g)+0.1415log(g_Streptococcus)
+0.1407log(g_Dialister) +0.1378log(g_Aggregatibacter)
-0.0920log(g_Prevotella)-0.2065log(f_Peptostreptococcaceae_g)
-0.2792log(g_Bacteroides)-0.3375log(g_Roseburia)

Acknowledgements

We thank J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle for sharing their data with us.

This work was supported by the Spanish Ministry of Science, Innovation and Universities/FEDER [grant number RTI2018–095518–B–C21]; the Spanish Ministry of Health [grant number CIBERCB06/02/1002]; and the Government of Catalonia [grant number 2017SGR656].

Thanks very much for your attention!

germa.coenders@udg.edu