Orthogonal regression among parts of compositional data

Klára Hrůzová, Karel Hron, Peter Filzmoser, Valentin Todorov

25th August 2014

Contents

Compositional data

- strictly positive real numbers carrying only relative information, $\mathbf{x} = (x_1, \ldots, x_D)'$;
- sample space is simplex, $\mathcal{S}^D$;
- Aitchison geometry with Euclidean vector space structure (perturbation, power transformation, distance, norm and inner product).

Motivation-example

- GVA: difference between gross output and intermediate consumption
- GVA can be decomposed by these activities:
  1. Agriculture;
  2. Manufacturing: the physical or chemical transformation of materials of components into new products;
  3. Other industry;
  4. Services;

  $\rightarrow$ GVA can be expressed as the sum of these four activities
- $\mathbf{Y}_{\text{manufacturing}} \sim \mathbf{X}_{\text{agriculture}} + \mathbf{X}_{\text{other industry}} + \mathbf{X}_{\text{services}}$;
- dataset comes from the World Bank database (http://data.worldbank.org), includes the data from 131 countries of the world from the year 2010 in constant 2005 USD.

Isometric logratio transformation

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1, \qquad (1)$$

- $\mathbf{z}_l = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$, $l = 1, \ldots, D$ is a real vector;
- $(x_1^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ stands for a permutation of the parts of $\mathbf{x}$ where always the l-th compositional part occupies the first position;
- the first ilr variable $z_1^{(l)}$ explains all the relative information about the original part x_l;
- the coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ explain the remaining logratios in the composition.

Analysis of the relation between parts of a composition

- $\mathbf{Y}_{\text{manufacturing}} \sim \mathbf{X}_{\text{agriculture}} + \mathbf{X}_{\text{other industry}} + \mathbf{X}_{\text{services}}$;
- $\mathbf{x} = (x_1, x_2, x_3, x_4)' \rightarrow x_1 \sim x_2 + x_3 + x_4$

$\rightarrow$ at least two compositional parts are of simultaneous interest, although their positions in the regression model are different

- the response variable: $z_1^{(l)}$ from (1);
- the explanatory variables: $z_2^{(k)}, \ldots, z_{D-1}^{(k)}$ corresponding to reordered subcomposition $(x_k, x_1, \ldots, x_i, \ldots, x_D)', \ i \neq \{k, l\}, \ k = 1, \ldots, D, \ k \neq l$;

$\rightarrow$ we obtain $D - 1$ regression models assigned to single explanatory compositional parts

Orthogonal regression

- both the response and explanatory variables come from one composition $\rightarrow$ all the variables are burdened by errors;
- belongs to so called errors-in-variable models (EIV), forms a special case of total least squares regression;
- notation: $\mathbf{X} \in \mathbb{R}^{n \times D-2}$ is the matrix of n replicates of the vector $(z_2^{(k)}, \ldots, z_{D-1}^{(k)})$, for a chosen $k = 1, \ldots, D$, $k \neq l,$, $\mathbf{y} \in \mathbb{R}^n$ the observation vector of the response coordinate $z_1^{(l)}$

Total least-squares method

- originally introduced to solve overdetermined systems of equations $\mathbf{Xb} \approx \mathbf{y}$;
- in the case of $n > D - 2$, there is no exact solution $\rightarrow$ we are seeking for an approximation;
- classical TLS problem is looking for the minimal errors $\varepsilon_X, \varepsilon_y$ on the given data $\mathbf{X}, \mathbf{y}$ that make the system of equations $\widehat{\mathbf{X}}\mathbf{b} = \widehat{\mathbf{y}},\ \widehat{\mathbf{X}} = \mathbf{X} + \varepsilon_X,\ \widehat{\mathbf{y}} = \mathbf{y} + \varepsilon_y$ solvable;
- singular value decomposition is applied to $\mathbf{Z} = [\mathbf{X}, \mathbf{y}] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$, where $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_{D-1})$ and $\sigma_1 \geq \cdots \geq \sigma_{D-1} \geq 0$ are the singular values of $\mathbf{Z}$;
- partitionings:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & v_{22} \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma_1} & \mathbf{0} \\ \mathbf{0} & \sigma_D \end{bmatrix},$$

- TLS solution exists iff v_{22} is non-singular; moreover, it is unique iff $\sigma_{D-2} \neq \sigma_{D-1}$, then

$$\widehat{\mathbf{b}} = -\mathbf{V}_{12}/v_{22}. \tag{2}$$

Use of principal component analysis

- matrices $\mathbf{\Sigma}$ and $\mathbf{V}$ from SVD of the centered explanatory and response variables correspond to outputs of eigenvalue decomposition of the covariance matrix, performed within principal component analysis (PCA);
- except of the intercept term in the orthogonal regression model, the same results as above in (2) can be obtained also using the smallest eigenvalue and the corresponding eigenvector (loading vector);
- $\mathbf{z} = (z_1, z_2, z_3)'$
- estimated parameters $\boldsymbol{\beta}$ are obtained using the values of the normal vector, $\mathbf{n} = (n_1, n_2, n_3)$, (i.e. the loading vector corresponding to the smallest eigenvalue), namely

$$\widehat{\beta}_0 = \frac{\mathbf{t}'\mathbf{n}}{n_3}, \ \widehat{\beta}_1 = -\frac{n_1}{n_3}, \ \widehat{\beta}_2 = -\frac{n_2}{n_3}.$$

Bootstrap sampling

- statistical inferences are not defined for orthogonal regression $\rightarrow$ use of resampling methods;
- we draw a sample $\mathbf{S} = (X_1, \ldots, X_n)'$ from a population $\mathbf{P} = (x_1, \ldots, x_N)'$, where $N >> n$ and we are interested in some statistic $T = t(\mathbf{S})$ which is estimate of the corresponding population parameter $\theta = t(\mathbf{P})$;
- nonparametric bootstrap allows us to estimate the sampling distribution of a statistic T empirically without making assumptions about the form of the population P;

Nonparametric bootstrap

- we draw a sample of size n from $\mathbf{S}$ with replacement;
- the sample $\mathbf{S}$ is treated as an estimate of the population $\mathbf{P}$ which means that each element of X_i of $\mathbf{S}$ is selected with probability $1/n$ to mimick the original selection of the sample $\mathbf{S}$ from the population $\mathbf{P}$;
- procedure is repeated R-times to obtain many bootstrap samples;
- next step is to compute the statistic T for each bootstrap sample.

Notation - ilr transformation in the example

- dependent variable: $z_1^{(1)} = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}$;
- the other two coordinates are formed by permutation of the remaining three parts,
  $z_2^{(2)} = \sqrt{\frac{2}{3}} \ln \frac{x_2}{\sqrt{x_3 x_4}}$; $z_3^{(2)} = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}$,
  $z_2^{(3)} = \sqrt{\frac{2}{3}} \ln \frac{x_3}{\sqrt{x_2 x_4}}$; $z_3^{(3)} = \sqrt{\frac{1}{2}} \ln \frac{x_2}{x_4}$,
  $z_2^{(4)} = \sqrt{\frac{2}{3}} \ln \frac{x_4}{\sqrt{x_2 x_3}}$; $z_3^{(4)} = \sqrt{\frac{1}{2}} \ln \frac{x_2}{x_3}$
  $\rightarrow$ three regression models are formed;
- $\mathbf{z}_2 = (z_2^{(2)}, z_3^{(2)}, z_1^{(1)})'$, $\mathbf{z}_3 = (z_2^{(3)}, z_3^{(3)}, z_1^{(1)})'$ and $\mathbf{z}_4 = (z_2^{(4)}, z_3^{(4)}, z_1^{(1)})'$.

Results

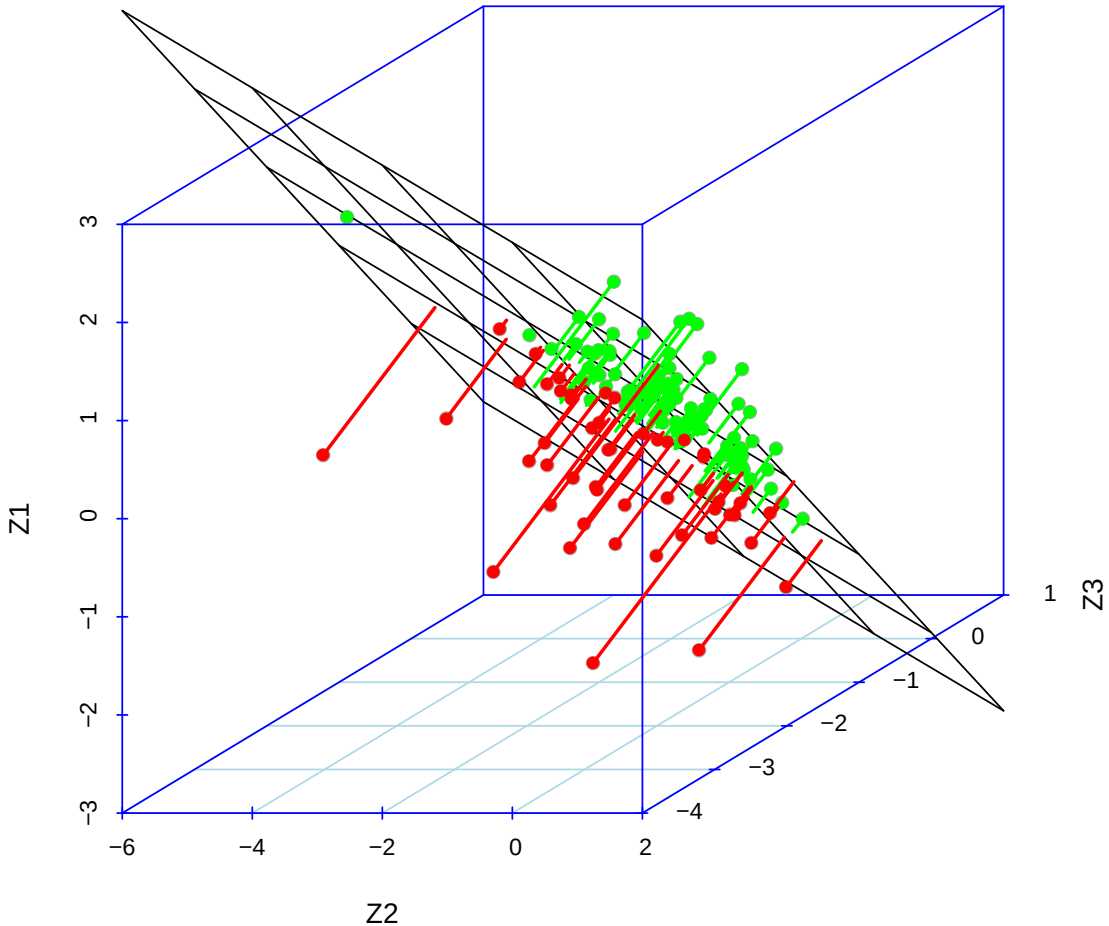
Table: Estimated parameters for the classical orthogonal regression.

parameter	$\mathbf{z}_2$	std. error	$\mathbf{z}_3$	std. error	$\mathbf{z}_4$	std. error
$\widehat{\beta}_0$	-2.151	1.018	-2.151	0.859	-2.151	0.853
$\widehat{\beta}_1$	-0.394	0.147	-0.878	0.720	1.272	0.653
$\widehat{\beta}_2$	-1.242	0.940	-0.962	0.370	0.280	0.475

Bootstrap confidence intervals

Table: Normal bootstrap confidence intervals for the model $\mathbf{z}_2$.

parameter	$\mathbf{z}_2$	confidence interval
$\widehat{\beta_0}$	-2.151	(-3.938, 0.078)
$\widehat{\beta_1}$	-0.394	(-0.6919, -0.1204)
$\widehat{\beta_2}$	-1.242	(-2.893, 0.833)


Z1
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2
Z2
1
0
−1
−2
−3
−4
Z3

Conclusions

- when we are dealing with compositional data, it is necessary to firstly transform the data in to the Euclidean space
- the choice of coordinates depends on the methodology used and on the interpretability of the results
- when both response and explanatory variables contain errors, it is necessary to use error-in-variable models
- orthogonal regression which is generally solved by SVD of $[\mathbf{X}, \mathbf{y}]$, our approach is based on PCA (eigenvalue decomposition) for interpretation purposes
- it is not possible to obtain statistical inference in the standard way because of the different estimation of parameters

References


Aitchison, J. (1986).
The Statistical Analysis of Compositional Data.
Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).


Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barceló-Vidal (2003).
Isometric logratio transformations for compositional data analysis.
Math Geol. 35, 279–300.


Fuller, W. A. (1987).
Measurement Error Models.
John Wiley & Sons, New York.


Hron, K., P. Filzmoser and K. Thompson (2012).
Linear regression with compositional explanatory variables.
Journal of Applied Statistics 39(5), 1115–1128.