Cellwise robust regression on compositional variables

Nikola Štefelová,
Andreas Alfons, Javier Palarea-Albaladejo,
Peter Filzmoser, Karel Hron

25th January 2018

Compositional data (CoDa)

- Composition: D-part vector $\mathbf{x} = (x_1, \ldots, x_D)'$ of strictly **positive values** (compositional parts) carrying **relative information** [Pawlowsky-Glahn and others, 2015]
- Data representing parts of some whole, e.g. proportions, percentages
- All the relative information about $\mathbf{x}$ contained in the ratios between its parts
- Working with logratios
  $\Rightarrow$ moves range from positive numbers to real axis
  $\Rightarrow \ln \frac{x_i}{x_j} = -\ln \frac{x_j}{x_i}$
- Compositions follow the Aitchison geometry on simplex
- **Logratio methodology** $\Rightarrow$ mapping compositions from simplex into real Euclidean space

Pivot coordinates

- Composition expressed in orthonormal coordinate system that highlights the role of a single compositional part

$$\mathbf{x} = (x_1, \ldots, x_D)' \rightarrow \mathbf{z} = (z_1, \ldots, z_{D-1})',$$

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, \quad j = 1, \ldots, D-1$$

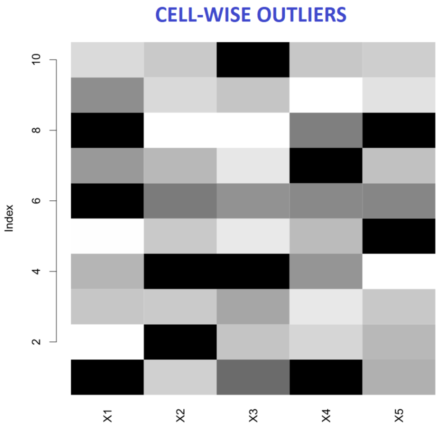
- z_1 explains all the relative information about part x_1

$$\mathbf{x}^{(l)} = (x_l, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)' = (x_1^{(l)}, \ldots, x_D^{(l)})' \rightarrow$$

$$z_j^{(l)} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^{(l)}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k^{(l)}}}, \quad \begin{array}{l} j = 1, \ldots, D-1, \\ l = 1, \ldots, D \end{array}$$

- D different orthonormal coordinate systems which are just rotations of each other

Cellwise outliers



- **Casewise** outlier - observation outlying as a whole
  Cellwise outlier - contamination only at a cell level
- Outlyingness of a cell in composition results from outlying pairwise logratio(s) with the respective part
- Ordinary robust estimators designed to deal with casewise outliers
- If contamination occurs only at the cell level $\Rightarrow$ unnecessary loss of information

Regression settings

- Real response Y
- Explanatory variables
  - $D-$part composition $\mathbf{x} = (x_1, \ldots, x_D)'$
  - p real variables $V_1, \ldots, V_p$
  - (Factor with k levels $\Rightarrow$ dummy variables $F_1, \ldots, F_{k-1}$)
- n observations available

Designed procedure

- Detection of cellwise outliers
  $\Rightarrow$ replacing them by missing values (NA's)
- Imputation of NA's
- Compositional MM-regression
- Multiple imputation

Detection of cellwise outliers

- Bivariate filter [Rousseeuw and Van den Bossche, 2017]
- Assumption - data follow multivariate normal distribution but after some cells were contaminated
- Detecting deviating cells in each column of standardized data
- Flagging cells that deviate from the correlation structure of data
  - Each cell predicted based on the unfiltered cells in the same row whose column correlate (robust $\rho > 0.5$) with the column in question
  - Observed value differs much from its predicted value $\Rightarrow$ cell detected as outlying

- Filter performed on the matrix with $p + 1 + D(D-1)/2$ columns
  - $p + 1$... real (explanatory and response) variables
  - $D(D-1)/2$... detecting deviating cells in CoDa via matrix of pairwise logratios
- For some observation at least half of the logratios with part x_i detected as outliers $\Rightarrow x_i$ flagged as outlying
- Flagged cells replaced by missing values (NA's)

Imputation of missing values (NA's)

- Adaptation of iterative model-based imputation for CoDa [Hron and others, 2010]
- Separate ordering of compositional parts and real variables based on amount of outliers
- Initialization - geometric/arithmetic mean
- First D steps in each iteration for updating $x_l, \quad l = 1, \dots, D$
    - $z_1^{(l)}$ set as a response, the rest of the variables as covariates
    - Observations with not outlying x_l used for the regression coefficients estimation
    - Obtained coefficients estimates taken to predict $z_1^{(l)}$ in observations with outlying x_l
    - Inverse mapping $\Rightarrow$ updated values of x_l

- Next $p + 1$ steps in each iteration for updating real variables - analogy (each time, different variable serves as response)
- Stop when the Frobenius norm of difference between the present and the previous empirical covariance matrix is smaller than a chosen boundary η ($\eta = 0.5$) - few iterations needed
- Robust MM-regression used in the iteration process

Compositional MM-regression

- Highly efficient robust MM-regression conducted on imputed data
- Compositional regression with interpretable regression coefficients [Hron and others, 2012]
- D different models

$$Y = \alpha + \beta_1^{(l)} z_1^{(l)} + \ldots + \beta_{D-1}^{(l)} z_{D-1}^{(l)} + \gamma_1 V_1 + \ldots + \gamma_p V_p + \varepsilon,$$
$$l = 1, \ldots, D$$

- Interest in $(\hat{\alpha}, \hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(D)}, \hat{\gamma}_1, \ldots, \hat{\gamma}_m)$

- Default standard errors and test statistics assume data to be complete
- Standard errors underestimated, significance inflated
  $\Rightarrow$ MI estimation of the regression

Multiple imputation

- Regression analysis carried out on m different datasets [Rubin and Schenker, 1986]
  - m set as number of observations containing outlying cells
- In each of the m datasets random error term is added to each imputed values (to the $z_1^{(l)}$ for CoDa)
- Noise - sample from $N(0, \sigma_j^2)$ multiplied by correction factor $\sqrt{1 + \frac{1}{n} m_j}$
  - m_j denotes the number of NA's in the jth response, $j = 1, \ldots, D + p + 1$
  - σ_j taken as a scale estimate of the reweighted residuals from jth step of the last iteration

- Final coefficient estimate taken as the average of the m estimates
- Estimation of variance of the estimator - sum of within-imputation variance and between-imputation variance multiplied by correction factor $\frac{m+1}{m}$.

Simulation study

- Simulation settings:

$S = 500$	# simulations
$n = 300$	# observations
$D = 6$	# compositional parts
$c = \frac{5n}{100}$	# outlying cells in each compositional part
$m = 3$	multiplicator for outlying parts

- Generating data for in each simulation run:

$$\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,D-1})' \sim \mathcal{N}_{D-1}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \text{ from VFA data}$$

$$\mathbf{z}_i \rightarrow \mathbf{x}_i = (x_{i,1}, \ldots, x_{i,D})'$$

$$y_i = \beta_0 + \beta_1 z_{i,1} + \ldots + \beta_{D-1} z_{i,D-1} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 0.5),$$

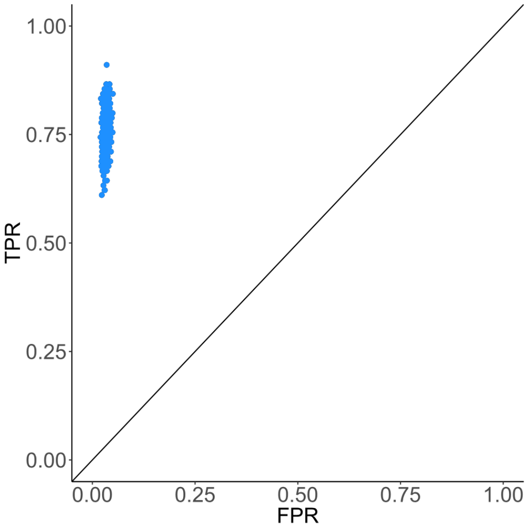
$$(\beta_0, \beta_1, \ldots, \beta_{D-1}) = (0, 1, \ldots, 1)$$

- Creating outlying cells:

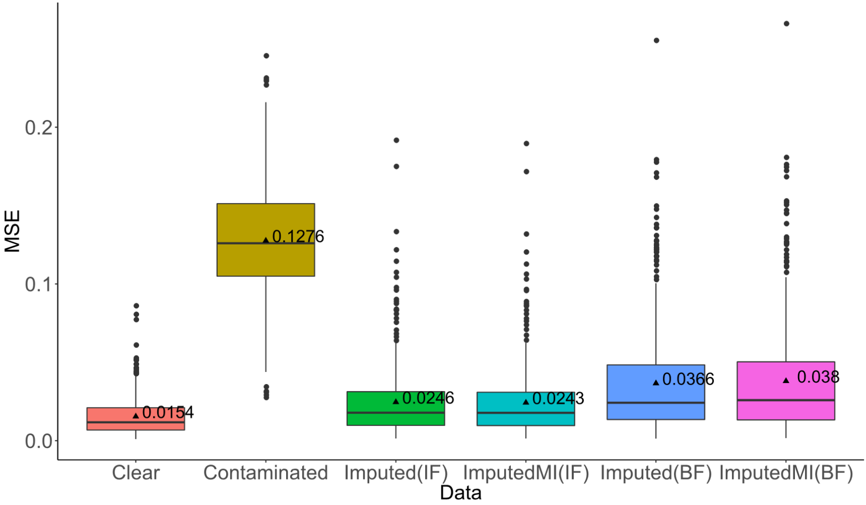
$$I_j = \{I_{j_1}, \ldots, I_{j_c}\} \subset \{1, 2, \ldots, n\}, \quad j = 1, \ldots, D$$

$$\hat{x}_{i,j} = \begin{cases} x_{i,j} m & \text{if} \quad i \in I_j \\ x_{i,j} & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, D$$

- The performance of the filter



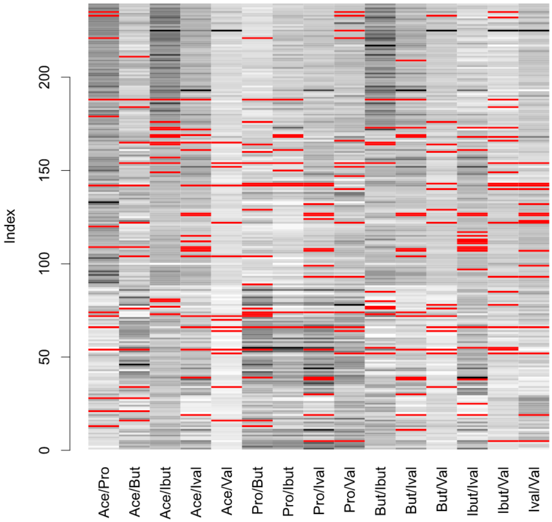
- The performance of the procedure



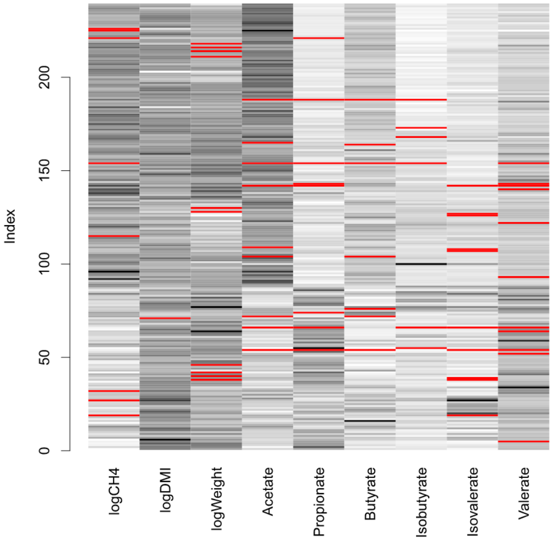
Example with VFA data

- 239 observations
- 3 real variables, 6-part composition and 1 factor
- Response
  - *CH4*: methane emissions [g/kgDMI]
- Explanatory variables
  - Volatile fatty acid (*VFA*) composition in mmol/mol (closed to 1000): *Acetate*, *Propionate*, *Butyrate*, *Isobutyrate*, *Isovalerate*, *Valerate*
  - *DMI*: actual dry matter intake [kg/day]
  - *Weight* [kg]
  - *Diet* factor with 2 levels: *Concentrate*/*Mixed*

- Logratios flagged as outlying



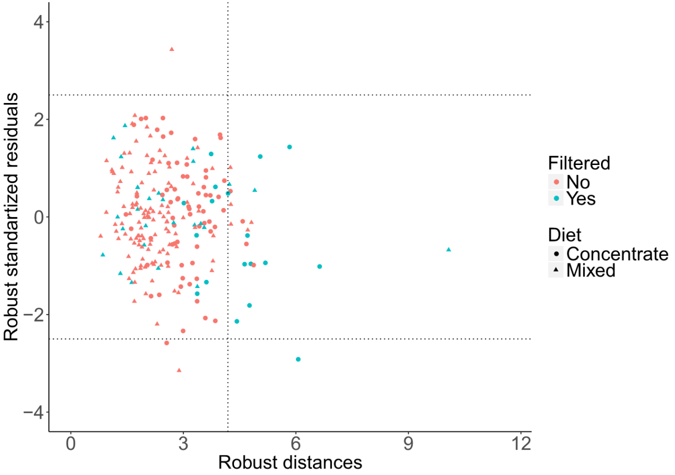
- Variables/parts flagged as outlying (3.25%)



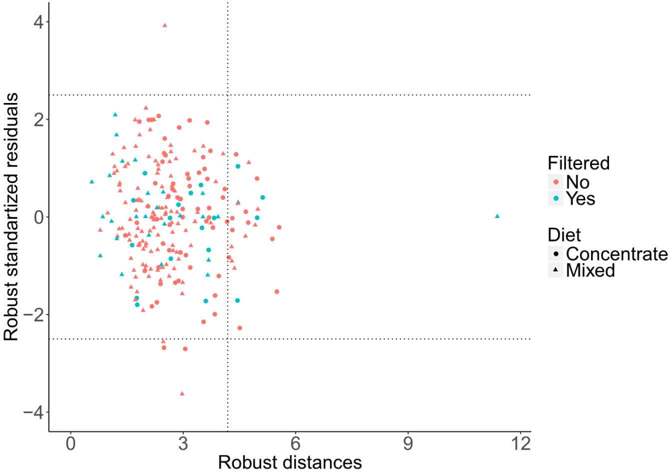
- Estimates of the regression coefficients, standard errors and $p-$values for ordinary vs. cellwise MM-regression

	Ordinary			Cellwise		
Variable	Coeff.	Std. Error	p-value	Coeff	Std. Error	p-value
Intercept	0.075	0.945	0.937	-0.760	0.922	0.410
$z_1^{(\text{Acetate})}$	0.147	0.089	0.102	0.191	0.088	0.030
$z_1^{(\text{Propionate})}$	-0.281	0.062	<0.001	-0.322	0.060	<0.001
$z_1^{(\text{Butyrate})}$	0.074	0.053	0.162	0.075	0.052	0.149
$z_1^{(\text{Isobutyrate})}$	0.011	0.047	0.816	0.012	0.043	0.783
$z_1^{(\text{Isovalerate})}$	0.013	0.034	0.715	0.041	0.034	0.228
$z_1^{(\text{Valerate})}$	0.037	0.036	0.305	0.003	0.030	0.921
log(DMI)	-0.379	0.061	<0.001	-0.388	0.051	<0.001
log(Weight)	0.554	0.156	<0.001	0.680	0.151	<0.001
F_{Mixed}	0.265	0.041	<0.001	0.228	0.038	<0.001

- Regression diagnostics for ordinary MM-regression



- Regression diagnostics for cellwise MM-regression



References

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

Hron, K., Templ, M., Filzmoser, P. (2010).
Imputation of missing values for CoDa using classical and robust methods.
Computational Statistics & Data Analysis 54(12), 3095 – 3107.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015).
Modeling and analysis of compositional data.
Chichester: Wiley.

Rousseeuw, P.J., Van den Bossche, W (2017).
Detecting deviating data cells.
Technometrics, DOI: 10.1080/00401706.2017.1340909.

Rubin, D.B, Schenker (1986).
Multiple imputation for interval estimation from simple random samples with ignorable nonresponse.
Journal of the American Statistical Association 81(394), 366 – 374.