To appear in *Statistics: A Journal of Theoretical and Applied Statistics*
Vol. 00, No. 00, Month 201X, 1–18

RESEARCH ARTICLE

Classical and robust orthogonal regression between parts of compositional data

K. Hrůzová[a*], V. Todorov[b], K. Hron[a] and P. Filzmoser[c]

[a] *Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17. listopadu 12, Olomouc, Czech Republic*;
[b] *United Nations Industrial Development Organization (UNIDO), Vienna International Centre, Vienna, Austria*;
[c] *Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria*

(*v4.0 released March 2013*)

The different parts (variables) of a compositional data set cannot be considered independent from each other, since only the ratios between the parts constitute the relevant information to be analyzed. Practically, this information can be included in a system of orthonormal coordinates. For the task of regression of one part on other parts, a specific choice of orthonormal coordinates is proposed which allows for an interpretation of the regression parameters in terms of the original parts. In this context, orthogonal regression is appropriate since all compositional parts–also the explanatory variables–are measured with errors. Besides classical (least-squares based) parameter estimation, also robust estimation based on robust principal component analysis is employed. Statistical inference for the regression parameters is obtained by bootstrap; in the robust version the fast and robust bootstrap procedure is used. The methodology is illustrated with a data set from macroeconomics.

Keywords: compositional data, orthogonal regression, isometric logratio coordinates, MM-estimates, bootstrap inference

AMS Subject Classification: 62H99; 62J05

1. Introduction

Compositional data are nowadays widely known as multivariate observations that carry only relative information [1, 27, 28], particularly also through their frequent representation in proportions or percentages. Consequently, the standard Euclidean geometry, based on the absolute (Lebesgue) measure, does not reflect the natural requirements for compositional data analysis that can be summarized into principles of scale invariance (multiplying by a positive constant does not alter the information conveyed by the composition), subcompositional coherence (analysis concerning a subset of parts must not depend on the other non-involved parts) and permutation invariance (the conclusions of a compositional analysis should not depend on the order of the parts) [11]. These principles are followed formally by the Aitchison geometry, named after the author of the seminal book on statistical analysis of compositional data [1]. The scale invariance

*Corresponding author. Email: klara.hruzova@gmail.com

property of compositional data allows for a representation of the observations with an arbitrary constant sum constraint, like the mentioned proportions or percentages, and thus we refer to the simplex as the sample space of (representations of) compositions. As most standard statistical methods rely on the Euclidean geometry [6], the aim is to express compositional data in the real space, i.e. to find an appropriate coordinate representation with respect to the Aitchison geometry.

Since the dimension of the simplex is one less than the actual number of D parts in the composition, there is no canonical basis that would allow to assign a coordinate to each compositional part simultaneously. Recently, the main focus is devoted to find interpretable orthonormal coordinates, guaranteeing isometry between the Aitchison geometry and the Euclidean real space, that would reflect the needs of a particular statistical problem of interest [9, 27]. One possibility are log-contrasts, each one within a different set of orthonormal coordinates, that make possible to assign one of the coordinates to a chosen compositional part [15, 16, 19, 21]. Nevertheless, a further challenge arises when not one part, but two (or even more) compositional parts are simultaneously of interest within one coordinate system. This is exactly the case in regression analysis, when both the response and explanatory variables come from one and the same composition. This particular problem will be thoroughly discussed.

The next section introduces one possible choice of orthonormal coordinates, following the idea of [2], extended for the purpose of regression analysis among compositional parts. In particular, it turns out that $D-1$ regression models are necessary to analyze comprehensively the relations of one compositional part to the rest. Moreover, as both the response and explanatory variables are coordinates of a random composition, they are random by nature. Consequently, standard least-squares regression would lead to biased results, so orthogonal regression (as a special case of errors-in-variable models) needs to be applied instead. This is treated in Section 3, and the estimation of the regression parameters via singular value decomposition of the input data matrix is described in detail together with the corresponding geometric motivation. Moreover, as real-world compositional data are usually contaminated by outlying observations, a robust counterpart to classical orthogonal regression (MM-estimates) is presented as well. In order to perform statistical inference, like deriving confidence intervals or testing hypotheses, bootstrap techniques for classical and robust orthogonal regression are described in Section 4. In Section 5, these procedures are applied to a problem from macroeconomics. Section 6 introduces the R package **oreg** which was used for computation, and the final Section 7 concludes.

2. Coordinate representation of compositional data

As already mentioned in the introduction, specific features of compositional data [7] are characterized by the Aitchison geometry on the simplex with Euclidean vector space structure [10]. Nevertheless, most multivariate statistical methods rely on the Euclidean geometry in real space [6, 26]. It is convenient to find orthonormal coordinates–in the ideal case interpretable coordinates–that allow for a statistical analysis using standard methods. This is possible through isometric logratio (ilr) coordinates [8] that assign a $(D-1)$-dimensional real vector to the D-part composition $\mathbf{x} = (x_1, \ldots, x_D)'$. Unfortunately, due to the dimension of the Aitchison geometry (one less than the number of parts of the composition), it is not possible to assign canonical coordinates to compositional data. Thus, a proper choice of interpretable coordinates is of particular interest. One possibility that seems to be advantageous from many different methodological aspects

[15, 19, 25] results in a set of D real vectors $\mathbf{z}^{(lk)} = (z_1^{(lk)}, \ldots, z_{D-1}^{(lk)})'$, $l = 1, \ldots, D$, where

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

Here $(x_1^{(lk)}, \ldots, x_D^{(lk)})'$ stands for such a permutation of the parts $(x_1, \ldots, x_D)'$, that always the l-th compositional part fills the first position and the k-th part the second one, $(x_l, x_k, x_1, \ldots, x_i, \ldots, x_D)'$, $i \notin \{l, k\}$. In such a configuration, the first ilr variable $z_1^{(lk)}$ explains all the relative information (logratios) about the original compositional part x_l, while the coordinates $z_2^{(lk)}, \ldots, z_{D-1}^{(lk)}$ explain the remaining logratios in the composition [16]. Note that the only important position is that of $x_1^{(lk)}$ (which is interpretable through $z_1^{(lk)}$), the other parts can be chosen arbitrarily from the perspective of x_l, because different ilr transformations are orthogonal rotations of each other [8]. Of course, $z_1^{(lk)}$ cannot be identified with the compositional part x_l, as the other parts are also naturally involved through the corresponding logratios. Therefore, due to the specific structure of the Aitchison geometry, its interpretation is limited. We can also see that this coordinate is formed by a logratio between the part x_l and an "average part", resulting from the geometric mean of the remaining parts in the composition. Therefore, the values of $z_1^{(lk)}$ represent a measure of dominance of the part x_l with respect to the other parts.

Regression analysis among compositional parts leads to a further complication, which comes from the fact that at least two parts in the composition are of simultaneous interest, the response part and the covariate part(s). Assume that x_l stands for the response and the remaining parts in the actual composition form the explanatory variables. We show that the coordinates (1) can be used for this purpose. In order to proceed, we follow methodologically the paper [2], where a similar problem was studied in a hydrological context, and the paper [19] which treats the case of regression of a real response on compositional explanatory variables. Let us start with the response variable. As it was mentioned above, if all relative information concerning part x_l in a given composition should be merged into one coordinate, then all pairwise logratios $\ln(x_l/x_1), \ldots, \ln(x_l/x_{l-1}), \ln(x_l/x_{l+1}), \ldots, \ln(x_l/x_D)$ need to be aggregated. So we arrive at

$$\ln(x_l/x_1) + \cdots + \ln(x_l/x_{l-1}) + \ln(x_l/x_{l+1}) + \cdots + \ln(x_l/x_D) = (D-1) \ln \frac{x_1^{(lk)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(lk)}}}; \tag{2}$$

up to a scaling constant, and this is nothing else than the coordinate $z_1^{(lk)}$ from (1). Indeed, since the main task is to analyze the influence of the other parts on x_l, it seems reasonable that also the corresponding coordinate will contain information on the relation of x_l to all remaining parts in the composition. From a mathematical perspective, the coordinates $z_1^{(lk)}$, $l = 1, \ldots, D$, are nothing else than logcontrasts, i.e. terms of the form $c_1 \ln x_1 + \cdots + c_D \ln x_D$ for $\sum_{i=1}^{D} c_i = 0$, proportional to the well-known centered logratio coefficients [1].

Now we can proceed with the coordinate representation of the explanatory subcomposition $(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$. In the above notation, e.g., $z_2^{(lk)}, \ldots, z_{D-1}^{(lk)}$ can serve for this purpose. Similarly, as we have done it for the response, we have to find an appropriate coordinate representation for the explanatory variables. An additional problem arises whether it is possible to treat $D-1$ covariates simultaneously, represented by the

respective coordinates. Unfortunately, this is not the case. The reason is that there would be an overlap of information, conveyed by pairwise logratios, used to construct the resulting coordinates. To see that, let us consider a pair of covariates x_k and x_m and aggregate the respective pairwise logratios from the explanatory subcomposition. Obviously, up to the sign, both of them contain $\ln(x_k/x_m)$, so it would not be possible to use both of them to construct orthonormal coordinates, required for a meaningful and interpretable statistical processing. For a particular single part x_k, however, we can continue to use the coordinates (1), and the constructed $z_2^{(lk)}$ would fit exactly to our needs. It is just not possible to consider both x_k and x_m (or even all covariates) simultaneously in one regression model.

Consequently, in order to analyze the influence of single explanatory parts (or, more precisely, their respective logratios) to the response, $D-1$ multiple regression models following the coordinate representations (1) need to be constructed. In each of these models, the response is represented by the coordinate $z_1^{(lk)}$ to capture the relative information on x_l. Note that this coordinate is the same for any $k \in \{1,\dots,D\}$, $k \neq l$. To each of the explanatory parts x_k, $k \neq l$, the coordinates $z_2^{(lk)},\dots,z_{D-1}^{(lk)}$ according to the reordered subcomposition $(x_k, x_2, \dots, x_i, \dots, x_D)'$, $i \notin \{k,l\}$, $k = 2,\dots,D$, are assigned. Similarly as before, the coordinate $z_2^{(lk)}$ explains all the relative information about part x_k in the resulting subcomposition. Considering the range of k, we arrive finally at $D-1$ regression models

$$z_1^{(lk)} = b_1^{(lk)} + b_2^{(lk)} z_2^{(lk)} + \ldots + b_{D-1}^{(lk)} z_{D-1}^{(lk)} + \varepsilon \tag{3}$$

(ε stands for an error term), assigned to single explanatory compositional parts. The interpretation of these models results from the interpretability of the coordinates, i.e., in each model just the absolute term parameter and the parameter corresponding to the coordinate $z_2^{(lk)}$ are used for further interpretation and for statistical inference (confidence intervals, hypotheses testing).

Since both the response and the explanatory variables originate from one composition, it cannot be assumed that the covariates represent errorless variables like in the case of a real valued response [19]. Consequently, the use of an ordinary multiple regression model is inappropriate and can even lead to biased results. Therefore, we apply an orthogonal regression model (or, equivalently, a total least squares model) for this purpose, which is a specific type of errors-in-variable (EIV) model [18].

3. Classical and robust orthogonal regression for compositional data

3.1 *Orthogonal regression*

For simplification of the notation, we denote the matrix of n realizations of the vector $(z_2^{(lk)},\dots,z_{D-1}^{(lk)})$, for a chosen $k \in \{1,\dots,D\}$, $k \neq l$, as $\mathbf{X} \in \mathbb{R}^{n\times D-2}$, and by $\mathbf{y} \in \mathbb{R}^n$ the observation vector of the response coordinate $z_1^{(lk)}$. For a further simplification of the notation, we assume without loss of generality that the response and the covariates are mean-centered, i.e. the sample mean is subtracted from the compositional data in coordinates. The total least-squares (TLS) method was originally introduced to solve overdetermined systems of equations $\mathbf{Xb} \approx \mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are given data (here compositions expressed in orthonormal coordinates), and $\mathbf{b} \in \mathbb{R}^{D-2}$ is the vector of unknown parameters. There is no exact solution; particularly in the case of $n > D-2$, we are seeking for an approximation.

In the classical TLS problem [22] we minimize the errors $\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y$, given the centered

data $\mathbf{X}, \mathbf{y}$ that make the system of equations $\widehat{\mathbf{X}}\mathbf{b} = \widehat{\mathbf{y}},\ \widehat{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}_X,\ \widehat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}_y$ solvable, i.e.

$$\{\widehat{\mathbf{X}}, \widehat{\mathbf{y}},\ \boldsymbol{\varepsilon}_X,\ \boldsymbol{\varepsilon}_y\} := \operatorname{argmin}_{\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y} \|\operatorname{vec}[\boldsymbol{\varepsilon}_X,\ \boldsymbol{\varepsilon}_y]\|\,, \tag{4}$$

subject to $(\mathbf{X} + \boldsymbol{\varepsilon}_X)\mathbf{b} = \mathbf{y} + \boldsymbol{\varepsilon}_y$ ("vec" forms one vector, composed of the columns of the matrix in the argument). The solution is a maximum likelihood estimator $\widehat{\mathbf{b}}$ in the optimally corrected EIV model $\widehat{\mathbf{X}}\widehat{\mathbf{b}} = \widehat{\mathbf{y}}$, $\widehat{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}_X$, $\widehat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}_y$, if the usual assumptions are fulfilled, namely that $\operatorname{vec}[\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y]$ has zero mean, and is a normally distributed random vector with a covariance matrix that is a multiple of the identity.

From the methodological point of view, singular value decomposition is applied to $\mathbf{Z} = [\mathbf{X}, \mathbf{y}] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$, where $\boldsymbol{\Lambda} = \operatorname{Diag}(\lambda_1, \ldots, \lambda_{D-1})$ and $\lambda_1 \geq \cdots \geq \lambda_{D-1} \geq 0$ are the singular values of $\mathbf{Z}$, and $\mathbf{U}$ and $\mathbf{V}$ are the corresponding orthonormal matrices. Let us define the partitions

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{21} & v_{22} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \lambda_D \end{bmatrix},$$

where the matrices $\mathbf{V}_{11}$ and $\boldsymbol{\Lambda}_1$ are of dimension $(D-2) \times (D-2)$. Then a TLS solution exists iff v_{22} is non-zero; moreover, it is unique iff $\lambda_{D-2} \neq \lambda_{D-1}$. In this case it is given by

$$\widehat{\mathbf{b}} = -\mathbf{v}_{12}/v_{22} \tag{5}$$

and the corresponding TLS error matrix equals $\boldsymbol{\varepsilon}_Z = [\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y] = -\mathbf{U}\operatorname{Diag}(\mathbf{0}, \lambda_{D-1})\mathbf{V}'$ [22], with $\mathbf{0}$ being a vector with $D-2$ zeros. Thus, when a unique solution $\widehat{\mathbf{b}}$ exists, it is computed from the scaled right singular vector corresponding to the smallest singular value. It is important to note that since the different ilr coordinate systems are just orthogonal rotations of each other [8], the TLS estimates will transform accordingly.

It is well known that the matrices $\boldsymbol{\Lambda}$ and $\mathbf{V}$ from SVD applied on the centered explanatory and response variables correspond to outputs of an eigenvalue decomposition on the (estimated) covariance matrix $\boldsymbol{\Sigma}$, as it is done in principal component analysis (PCA). Thus, except for the intercept term in the orthogonal regression model (that is discussed in the next section), the same results as above in (5) can be obtained using the smallest eigenvalue and the corresponding eigenvector (loading vector) of the covariance matrix. We will follow this approach further in the paper.

Regression estimators which are based on classical SVD or PCA are sensitive to outliers that occur in most real-world data sets. Therefore, we consider also a robust version of orthogonal regression. In [36], M- and S-estimators for robust orthogonal regression are presented. However, S-estimators are computed using inefficient algorithms and M-estimators have low breakdown point. Another possibility can be found in [3], where the projection-pursuit approach is used, which is also suitable for more than one response variable. In order to keep the structure of the paper consistent, but also to benefit from the better statistical properties of the MM-estimates, which will be combined with fast and robust bootstrap, we decided to follow the above (classical) approach in order to develop robust orthogonal regression in orthonormal coordinates.

Although robust versions of SVD are available (e.g. [4]), it is simpler and computationally more attractive to use robust PCA, which is obtained through a robust estimation of the covariance matrix (e.g. [13]). Among other possibilities like [3, 12, 24], in the following the MM-estimators [29] are employed for this purpose. The reason for choosing

MM-estimators is that they are highly efficient when the errors have a normal distribution, their breakdown point is 0.5 and they have bounded influence function.

Multivariate MM-estimators are extensions of S-estimators [23]. They are based on two loss functions ρ_0 and ρ_1. Generally, a ρ function has to satisfy the conditions:

(a) ρ is symmetric and twice continuously differentiable, with $\rho(0) = 0$;
(b) ρ is strictly increasing on an interval $[0, k]$ and constant on $[k, +\infty]$ for some finite constant k.

A standard choice of such loss function is Tukey's biweight function, defined as

$$\rho^c(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c, \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{6}$$

where $c > 0$ is a user-chosen tuning constant. Given the matrix with the observations in any chosen orthonormal coordinates (like those from the beginning of this section), $\mathbf{Z} = [\mathbf{X}, \mathbf{y}] = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' \in \mathbb{R}^{D-1}$, the MM-estimators for location and covariance are defined in two steps:

(1) Let $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ be S-estimators of location and covariance, respectively, that is $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ minimize $|\mathbf{C}|$, where $|\cdot|$ denotes the determinant of a matrix, subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left([(\mathbf{z}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{t})]^{1/2}\right) = b, \tag{7}$$

for $b > 0$, among all $(\mathbf{t}, \mathbf{C}) \in \mathbb{R}^{D-1}$. Denote $\widehat{s} = |\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]}$.

(2) The MM-estimator for location and shape $(\widehat{\boldsymbol{\mu}}_n,\ \widehat{\boldsymbol{\Gamma}}_n)$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\left[(\mathbf{z}_i - \mathbf{t})'\mathbf{S}^{-1}(\mathbf{z}_i - \mathbf{t})\right]^{1/2}/\widehat{s}\right)$$

among all $\mathbf{t}$ and all symmetric positive definite $\mathbf{S}$ with $|\mathbf{S}| = 1$. The MM-estimator of the covariance matrix is then $\widehat{\boldsymbol{\Sigma}}_n = \widehat{s}^2\widehat{\boldsymbol{\Gamma}}_n$.

Note that the values b in (7) and k_0 corresponding to condition (b) in the loss function ρ_0 (first step) are chosen in order to reach the maximal breakdown point, and the value k_1 corresponding to condition (b) in ρ_1 (second step) to achieve better efficiency of the estimator.

The idea is to estimate the scale by means of a very robust S-estimator and then to estimate the location and shape using different ρ functions to reach a better efficiency. Once location and covariance are obtained using the MM-estimator, they can be used to compute the robust orthogonal regression estimates as described above. Finally, it is important to stress that MM-estimators, applied to compositional data in orthonormal coordinates, i.e., to standard observations in real space, preserve all their important properties (like high breakdown point and good efficiency).

3.2 *Geometrical motivation*

As mentioned in the previous section, the TLS (orthogonal regression) estimates of the parameters can be obtained by means of principal component analysis. We apply the proposed procedure directly to the case of four-part compositional data where a geometrical illustration of the problem is still possible. For this purpose, we assume that we

have a random vector $\mathbf{z} = (z_1, z_2, z_3)'$ (an orthonormal coordinate representation of the composition). The task is to find a relationship between the response variable z_1 and the covariates z_2, z_3, expressed in the form $z_1 = b_1 + b_2 z_2 + b_3 z_3 + \varepsilon$, with the regression parameters b_1, b_2, b_3.

From a geometrical point of view, the basic idea is to fit a plane to the data using PCA. The loadings of the first two principal components define a basis of the plane. As the third principal component is orthogonal to the previous ones, its loadings define a unit normal vector to the plane, $\mathbf{n} = (n_1, n_2, n_3)'$, forming the last column of the matrix $\mathbf{V}$ in terms of the previous section. The plane passes through the point $\mathbf{t}$, representing the location estimate of the $n \times 3$ data matrix $\mathbf{Z}$ (the arithmetic mean in the classical case, equal to the zero vector for centered data), and its perpendicular distance from the origin is $\mathbf{t}'\mathbf{n}$. The perpendicular distance from each point in $\mathbf{Z}$ to the plane (the norm of the residuals) is the inner product of each centered point and the normal vector to the plane. The fitted plane minimizes the sum of squared errors.

Consequently, the estimated regression parameters are obtained using the elements of the normal vector, namely

$$\widehat{b}_1 = \frac{\mathbf{t}'\mathbf{n}}{n_3},\ \widehat{b}_2 = -\frac{n_1}{n_3},\ \widehat{b}_3 = -\frac{n_2}{n_3}.$$

4. Bootstrap sampling

For supporting the interpretation of the outcome of orthogonal regression, it is desirable to obtain confidence intervals for the regression parameters, and p-values for tests on these parameters. This statistical inference is only possible with strict distributional assumptions, but even then it would be challenging to derive the exact distribution of the parameters in the robust case. A better strategy is to derive the inference by resampling methods. In order to relax the assumptions about the distribution of the input data, the nonparametric bootstrap [17] was chosen for this purpose.

4.1 *Classical nonparametric bootstrap*

Generally, bootstrapping is based on building a sampling distribution for a statistic by resampling from the data at hand. Consequently, the nonparametric bootstrap allows us to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly. The basic idea is that, after drawing a sample of size n from $\mathsf{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ with replacement (without loss of generality, we fix the special choice of orthonormal coordinates again), we are treating the sample S as an estimate of the whole population. This means that each element $\mathbf{z}_i$ of S is selected with probability $1/n$ to mimic the original selection of the sample S. This procedure is repeated R times, where R is a large number, to obtain a sufficient number of bootstrap samples.

The r-th bootstrap sample is denoted as $\mathsf{S}^r = \{\mathbf{z}_{r_1}, \ldots, \mathbf{z}_{r_n}\}$, $r = 1, \ldots, R$. In the next step we compute the regression estimates $\widehat{b}_i$ for each bootstrap sample to get $\widehat{b}_i^{r*}$, $i = 1, \ldots, D-1$. Then the distribution of $\widehat{b}_i^{r*}$ around the original estimate $\widehat{b}_i$ is analogous to the sampling distribution of the estimator $\widehat{b}_i$ around the population parameter b_i. In the context of orthogonal regression, the bootstrap distribution of $\widehat{b}_i$ can

be directly used to derive sample p-values for significance testing of the regression parameters. For this purpose, the p-value p_i for the regression parameter b_i for a two-sided alternative is derived by comparing the values of the bootstrap parameter estimates to zero. By denoting l_i and h_i as the number of estimated values lower and higher than zero, respectively, we get $p_i = 2 \cdot \min\{l_i, h_i\}/R$ [5].

Furthermore, we can proceed also to construct bootstrap confidence intervals. For this purpose, several approaches are available. A natural choice is to take bootstrap percentile intervals that are free of any distributional assumptions [5, 17] and $\mathrm{BC_a}$ intervals (bias-corrected, accelerated percentile interval). The bootstrap percentile interval uses the empirical quantiles of $\widehat{b}_i^{r*}$ (computed from S^r, $r = 1, \ldots, R$) to form a confidence interval for b_i, $(\widehat{b}^*_{i(l)}, \widehat{b}^*_{i(u)})$, $i = 1, \ldots, D-1$. Hence, from the ordered bootstrap replicates of the statistic $\widehat{b}_i$, i.e. $\widehat{b}^*_{i(1)}, \widehat{b}^*_{i(2)}, \ldots, \widehat{b}^*_{i(R)}$, and for a given $\alpha \in (0,1)$ we set $l = [(R+1)\alpha/2], u = [(R+1)(1-\alpha/2)]$ (rounded to the nearest integer).

Due to lower accuracy of percentile intervals in case of small samples, the $\mathrm{BC_a}$ intervals are considered as well. To find the $\mathrm{BC_a}$ interval for b_i we need to compute correction factors z and a:

$$z = \Phi^{-1}\left[\frac{\#_{r=1}^{R}(\widehat{b}^*_{i(r)} \leq \widehat{b}_i)}{R+1}\right], \tag{8}$$

where $\Phi^{-1}(\cdot)$ is the standard-normal quantile function and $\#(\widehat{b}^*_{i(r)} \leq \widehat{b}_i)/(R+1)$ is the proportion of bootstrap replicates at or below the original-sample estimate $\widehat{b}_i$ of b_i;

$$a = \frac{\sum_{j=1}^{n}(\widehat{b}_{i(-j)} - \bar{b}_i)^3}{6[\sum_{j=1}^{n}(\widehat{b}_{i(-j)} - \bar{b}_i^2)]^{3/2}}, \tag{9}$$

where $\widehat{b}_{i(-j)}$ is defined as the value of $\widehat{b}_i$ produced when the jth observation is deleted from the sample and $\bar{b}_i = \sum_{j=1}^{n}\widehat{b}_{i(-j)}/n$. Then we need to compute values a_1 and a_2 that are used to locate the endpoints of the corrected percentile confidence interval, $\widehat{b}^*_{i(lower^*)} < b_i < \widehat{b}^*_{i(upper^*)}$, as $lower^* = [Ra_1]$ and $upper^* = [Ra_2]$:

$$a_1 = \Phi\left[z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})}\right];\ a_2 = \Phi\left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})}\right]. \tag{10}$$

Both percentile and $\mathrm{BC_a}$ confidence intervals can be computed with the software package described in Section 6.

4.2 *Fast and robust bootstrap*

The available theory for robust estimators is limited to asymptotic results. Although bootstrap is a very useful tool, in case of robust estimators there are two problems: computational complexity of robust estimators and the instability of the bootstrap in case of outliers. Therefore we used fast and robust bootstrap [33, 35] which is based on the fact that the robust estimators (namely S- and MM-estimators) can be represented by smooth fixed point equations which allow to calculate a fast approximation of the estimates in each bootstrap sample. For the case of MM-estimators, the fixed point

equations are as follows,

$$\widehat{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} \mathbf{z}_i \right); \tag{11}$$

$$\widehat{\boldsymbol{\Gamma}}_n = G\left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} (\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)' \right); \tag{12}$$

$$\widetilde{\boldsymbol{\Sigma}}_n = \frac{1}{nb}\left(\sum_{i=1}^{n} (D-1) \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} (\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)' + \left(\sum_{i=1}^{n} \widetilde{w}_i \right) \widetilde{\boldsymbol{\Sigma}}_n \right); \tag{13}$$

$$\widetilde{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} \mathbf{z}_i \right); \tag{14}$$

where we denote $G(\mathbf{A}) = |\mathbf{A}|^{-1/(D-1)}\mathbf{A}$ for a $(D-1)\times(D-1)$ matrix $\mathbf{A}$, $d_i = [(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Gamma}}_n^{-1}(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)]^{1/2}$, $\widetilde{d}_i = [(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)'\widetilde{\boldsymbol{\Sigma}}_n^{-1}(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)]^{1/2}$ and $\widetilde{w}_i = \rho_0(\widetilde{d}_i) - \rho_0'(\widetilde{d}_i)\widetilde{d}_i$. Here $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Gamma}}_n$ are the MM-estimators of location and shape, respectively, and $\widetilde{\boldsymbol{\mu}}_n$ and $\widetilde{\boldsymbol{\Sigma}}_n$ are the S-estimators of location and covariance, respectively (for more details, see Section 3.1). Generally, denote the equations (11) - (14) by means of a function $\mathbf{f} : \mathbb{R}^{2[(D-1)+(D-1)^2]} \to \mathbb{R}^{2[(D-1)+(D-1)^2]}$ such that $\mathbf{f}(\widehat{\Theta}_n) = \widehat{\Theta}_n$, where $\widehat{\Theta}_n$ contains all estimates in the vectorized form and can be represented as a solution of fixed-point equations. For example, for MM-estimators, we have $\widehat{\Theta}_n := \left((\widehat{\boldsymbol{\mu}}_n)', \mathrm{vec}(\widehat{\boldsymbol{\Gamma}}_n)', \mathrm{vec}(\widetilde{\boldsymbol{\Sigma}}_n)', (\widetilde{\boldsymbol{\mu}}_n)' \right)'$. Instead of recalculating the estimates $\widehat{\Theta}_n^*$ for each bootstrap sample we can calculate its one-step approximation starting from the initial value $\widehat{\Theta}_n$,

$$\widehat{\Theta}_n^{1*} = \mathbf{f}(\widehat{\Theta}_n). \tag{15}$$

Unfortunately, this approximation underestimates the variability of $\widehat{\Theta}_n$ because the initial value in the approximation remains the same. To remedy this we can apply a linear correction [32] as follows. Given the smoothness of $\mathbf{f}$ we can calculate a Taylor expansion about the limiting value of $\widehat{\Theta}_n$

$$\widehat{\Theta}_n = \mathbf{f}(\Theta) + \nabla\mathbf{f}(\Theta)(\widehat{\Theta}_n - \Theta) + R_n, \tag{16}$$

where $\Theta = (\boldsymbol{\mu}', \mathrm{vec}(\boldsymbol{\Gamma})', \mathrm{vec}(\boldsymbol{\Sigma})', \boldsymbol{\mu}')'$, R_n is the remainder term and $\nabla\mathbf{f}(\cdot)$ is the matrix of partial derivatives. If the remainder term is sufficiently small, we can rewrite (16) as

$$\sqrt{n}(\widehat{\Theta} - \Theta) \approx [I - \nabla\mathbf{f}(\Theta)]^{-1}\sqrt{n}(\mathbf{f}(\Theta) - \Theta). \tag{17}$$

Since both sides of this equation are asymptotically equivalent, the distribution of the bootstrapped statistics will also converge to the same limit. Finally, we can define the linearly corrected version of the one-step approximation (15) as

$$\widehat{\Theta}_n^{R*} := \widehat{\Theta}_n + [I - \nabla\mathbf{f}(\widehat{\Theta}_n)]^{-1}(\widehat{\Theta}_n^{1*} - \widehat{\Theta}_n). \tag{18}$$

Note that the estimating equations involve weighted least squares estimates and covariances, which are a generalization of the classical least squares method. Then the weights will be small or even zero for observations identified as outliers. This guarantees that $\widehat{\Theta}_n^{R*}$ is as robust as $\widehat{\Theta}_n$.

5. Example: Structure of gross value added

The procedures described in the previous sections are applied to a data set from macroeconomics representing the structure of gross value added and the relation between its components. The data set comes from the World Bank database (http://data.worldbank.org) and includes observations for 131 countries in 2010 at constant 2005 USD.

Gross value added (GVA) is the most important measure of productivity of the economy of a country or region, representing the difference between production output and intermediate consumption, i.e. the monetary value of the amount of goods and services that have been produced, less the cost of all inputs and raw materials that are directly attributable to that production. Gross value added is less than GDP because it excludes value-added tax (VAT) and other product taxes.

GVA can be decomposed into the following economic activities:

- agriculture (consisting of agriculture, forestry, hunting and fishing);
- manufacturing[1];
- other industry (consisting of mining and quarrying; electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities; construction);
- services (consisting of education, health and other personal services; public administration and defense).

Thus, GVA can be expressed as the sum of these four activities. The goal of the study is to analyze the relation between manufacturing and the rest of the activities by considering relative contributions of the mentioned activities to the overall GVA.

Although the original data are expressed in monetary units (USD), and no constant sum constraint is present (like it is the case of proportions or percentages), from the relative structure of GVA we can conclude that these four economic activities form a composition $\mathbf{x} = (x_1, x_2, x_3, x_4)'$, where x_1 corresponds to manufacturing, x_2 to agriculture, x_3 to other industry and x_4 to services. In such case, using an arbitrary regression technique either for the original observations or any constrained form of them, would lead to biased results [19]. Figure 1 displays a ternary diagram of the explanatory variables x_2, x_3, x_4. The ternary diagram is an equilateral triangle $X_2X_3X_4$ such that the three-part subcomposition $(x_2, x_3, x_4)'$ is plotted at a distance x_2 from the opposite side of vertex X_2, at a distance x_3 from the opposite side of vertex X_3, and at a distance x_4 from the opposite side of vertex X_4 (see, e.g., [1]). Accordingly, it can be observed that the part `srv` (services) contains the largest relative contribution and `agr` (agriculture) the smallest one in this subcomposition. This corresponds to the fact that the points are concentrated mainly along the segment between `srv` and `ind` (other industry), rather closer to the vertex `srv`.

For the further statistical processing, the compositional response and the explanatory variables are expressed in *ilr* coordinates (1). Following the previous considerations, the response coordinate is defined as $z_1^{(1k)} = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}$ for any $k \in \{2, 3, 4\}$, i.e., it ex-

[1]Manufacturing is defined as the physical or chemical transformation of materials of components into new products.

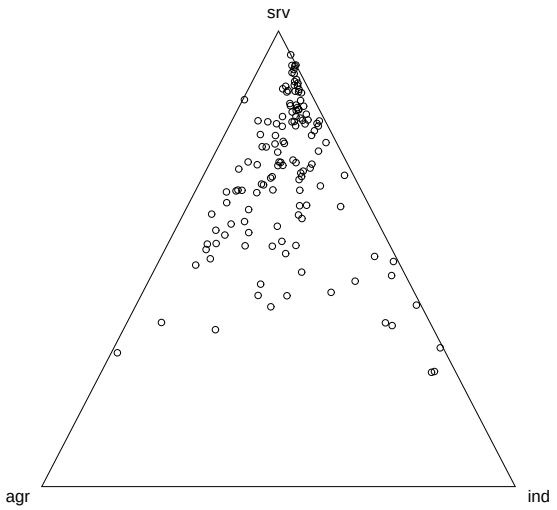

Figure 1. Ternary diagram of the explanatory variables `agr` (agriculture), `ind` (other industry), and `srv` (services).

plains all the relative information about manufacturing with respect to the other three parts in the composition through an aggregation of the corresponding logratios. Permutation of the remaining three activities results in three regression models, where always the respective coordinate $z_2^{(1k)}$ for $k = 2, 3, 4$ includes the most interesting information - (scaled) aggregation of logratios of x_i with the remaining explanatory parts. The resulting regression models that favor one of the explanatory compositional parts x_2, x_3, x_4 thus contains the following coordinates (in addition to $z_1^{(1k)}$),

$$\begin{aligned} z_2^{(12)} &= \sqrt{\tfrac{2}{3}} \ln \tfrac{x_2}{\sqrt{x_3 x_4}}, \quad z_3^{(12)} = \sqrt{\tfrac{1}{2}} \ln \tfrac{x_3}{x_4}; \\ z_2^{(13)} &= \sqrt{\tfrac{2}{3}} \ln \tfrac{x_3}{\sqrt{x_2 x_4}}, \quad z_3^{(13)} = \sqrt{\tfrac{1}{2}} \ln \tfrac{x_2}{x_4}; \\ z_2^{(14)} &= \sqrt{\tfrac{2}{3}} \ln \tfrac{x_4}{\sqrt{x_2 x_3}}, \quad z_3^{(14)} = \sqrt{\tfrac{1}{2}} \ln \tfrac{x_2}{x_3}, \end{aligned}$$

respectively.

In Figure 2, scatterplots of the explanatory coordinates are displayed, where the part of interest corresponds to x_2 (upper left), x_3 (upper right) and x_4 (lower left). Particularly, it can be seen that the x-coordinates of the upper left and upper right plots, $z_2^{(12)}$ and $z_2^{(13)}$, are mainly negative which means that the relative contributions of agriculture and other industry are lower than the mean contribution of the other parts. On the other hand, the coordinate $z_2^{(14)}$ clearly shows the relative dominance of services. The 3D scatterplot in Figure 2 (lower right) contains all three coordinates $z_2^{(12)}, z_3^{(12)}, z_1^{(12)}$ (in this order) to see the relation between the covariates and the response variable. Although a certain linear relationship can be observed from this scatterplot, orthogonal regression modeling needs to be performed in order to specify the possible influence of covariates.

The results of classical orthogonal regression in coordinates (following Sections 3 and 4) are summarized in Table 1. Note that the intercept for all regression models is identical (similar as for LS regression [19]), which is a consequence of the orthogonal relation between the different *ilr* coordinate systems. Therefore, although the basic model consists of three regression parameters (corresponding to intercept and two orthonormal coordinates), for the interpretation purposes it is enough to summarize just the intercept and

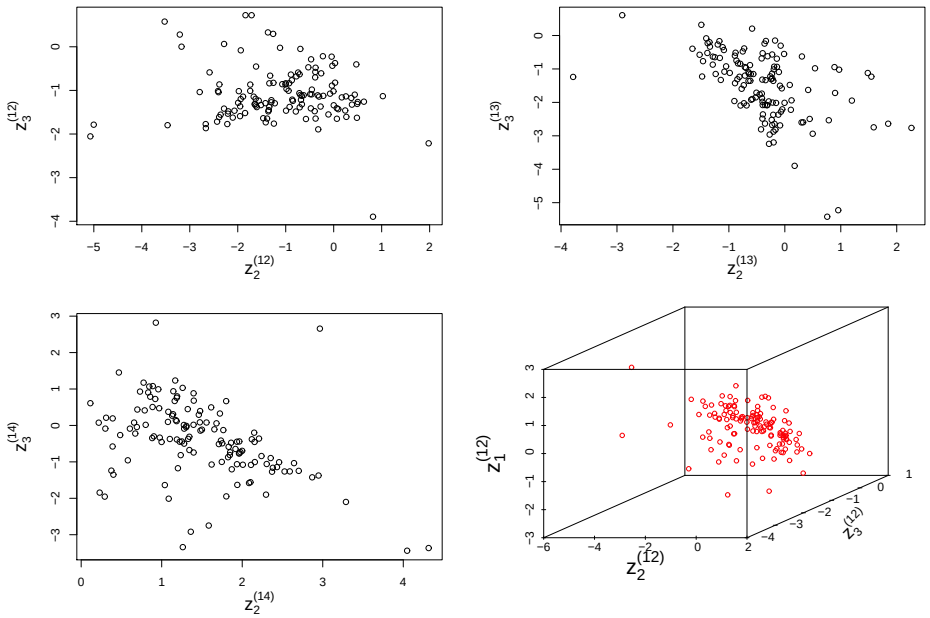

Figure 2. The plots of coordinates of explanatory variables and 3D scatterplot of the explanatory coordinates.

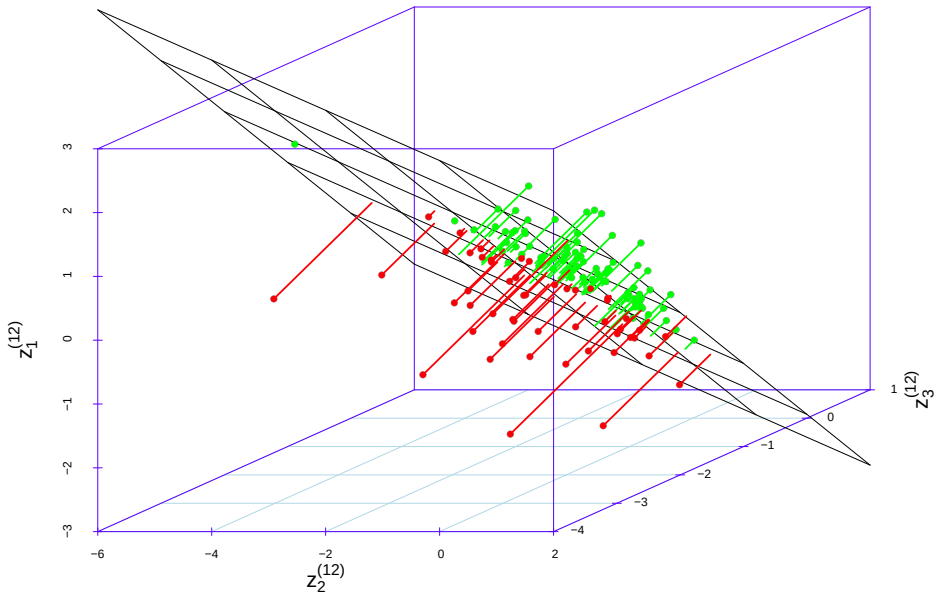

Figure 3. The 3D plot of estimated regression plane for coordinates $z_2^{(12)}, z_3^{(12)}, z_1^{(12)}$.

Table 1. Summary of regression outputs using classical orthogonal regression for all defined models.

	par. estimate	perc. CI	BC_a CI	p-value
intercept	-2.151	(-4.464, -1.559)	(-4.571, -1.562)	0.002
$b_2^{(12)}$ (agriculture)	-0.394	(-0.584, -0.115)	(-0.603, -0.011)	0.020
$b_2^{(13)}$ (other industry)	-0.878	(-2.745, -0.498)	(-3.390, -0.490)	0.000
$b_2^{(14)}$ (services)	1.272	(0.858, 2.978)	(0.832, 2.777)	0.002

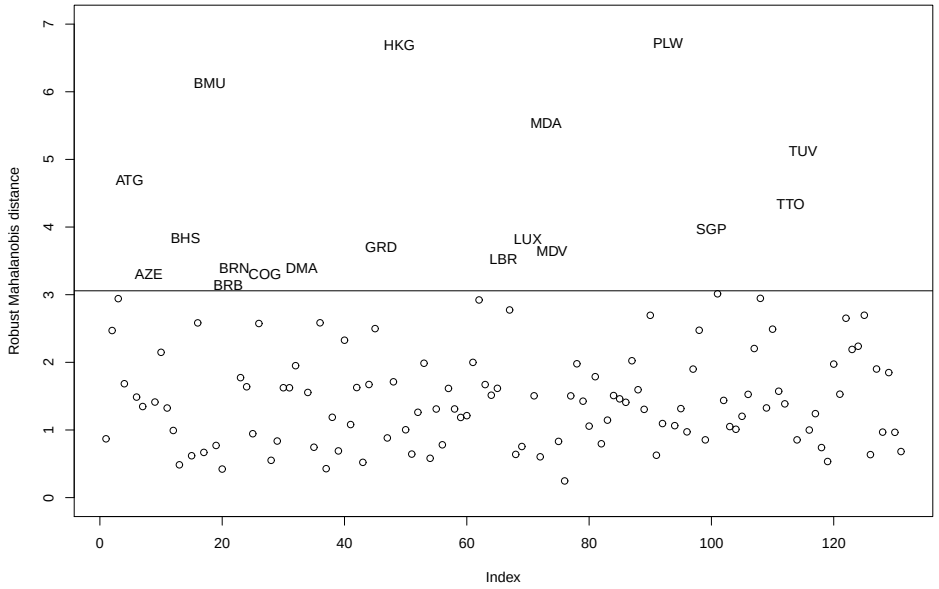

Figure 4. Scatterplot of outlying observations.

the parameters corresponding to the coordinates $z_2^{(12)}, z_2^{(13)}, z_2^{(14)}$ from all three models. Nonparametric bootstrap (with $R = 1000$) was used to derive the corresponding statistical inference (confidence intervals, p-values for significance testing). Note that it would be possible to compute the regression estimates for all remaining models just from the estimates in one concrete model using a orthogonal transformations, similar as in the standard case of LS (PLS) regression [20].

According to Table 1, all regression parameters are significant on the usual level $\alpha = 0.05$, although $\widehat{b}_2^{(12)}$ is closer to zero. Moreover, the estimated parameters $b_2^{(12)}$ and $b_2^{(13)}$ are negative which means that "agriculture" and "other industry" have small negative relative influence on manufacturing. On the other hand, "services" (resulting from the estimation in the last model) has strong positive relative influence on "manufacturing". This is explained by the fact that the growth of the manufacturing sector is inevitably induced by the growth of the services sector, necessary to support it. Transportation, communication, financial and business services are required by the manufacturing and thus there is no increase in manufacturing without (relative) growth of these services. To illustrate the regression results geometrically (see Section 3.2 to recall the geometric motivation), Figure 3 displays the 3D plot of the estimated regression plane for coordinates $z_1^{(12)}$ (response), and $z_2^{(12)}, z_3^{(12)}$ (covariates) with all the points projected on the plane.

Figure 4 shows the results of outlier detection for compositional data using Mahalanobis distances [14]. There are 18 outlying observations in our dataset, affecting both the response and the covariates. To restrict their possible influence on the estimates, a robust version of orthogonal regression using MM-estimators was applied as well. The summary of the regression outputs (including confidence intervals and p-values computed by fast and robust bootstrap) are displayed in Table 2. The results are similar to those from Table 1. In contrast to the classical analysis, here the regression parameter $\widehat{b}_2^{(12)}$ is not significant. Consequently, the difference for the inference on $\widehat{b}_2^{(12)}$ can be attributed to the outliers, underlining the need for a robust analysis. Of course, there are differences among countries, and the relation between agriculture and industry is a long debated

topic, see for example [30] and [31] for a detailed analysis of these linkages in India using the input-output framework.

Table 2. Summary of regression outputs using robust orthogonal regression for all defined models.

	par. estimate	perc. CI	BC_a	p-value
intercept	-2.311	(-6.391, -1.666)	(-6.040, -1.640)	0.006
$b_2^{(12)}$ (agriculture)	-0.389	(-0.605, 0.180)	(-0.590, 0.120)	0.116
$b_2^{(13)}$ (other industry)	-1.075	(-4.994, -0.556)	(-4.145, -0.549)	0.002
$b_2^{(14)}$ (services)	1.464	(0.996, 5.184)	(0.031, 2.640)	0.002

6. Software: the R package oreg

The computations in this paper were carried out using the R package *oreg* which provides functions for classical and robust orthogonal regression - `oregClassic()` and `oregMM()`. These functions can be applied on both compositional and non-compositional data. In case of compositional data, all $D-1$ regression models are estimated, one for each orthonormal basis, as described in Section 2. The regression parameters are estimated using (classical or robust) principal components. The MM-estimates are computed by a call to the implementation in the *rrcov* package [34].

The results can be viewed by standard `print()` and `plot()` functions, while a `summary()` function presents the parameter estimates and also the corresponding statistical inference (confidence intervals and p-values for significance testing) obtained through bootstrap. In the robust version, fast and robust bootstrap from the package *FRB* [35] is used.

The presented methodology is computationally intensive since we are dealing with $D-1$ regressions and each of them requires bootstrap and estimation of the robust MM regressions. While the computational effort is mitigated by the application of the fast and robust bootstrap approach, the question still remains how feasible the methodology is in case of larger data sets. The example presented in the previous section is in dimension 4 only and does not shed light upon this issue. To study the computational performance of the method, an experiment with simulated data with $n = 400$ observations in increasing dimensions, up to $D = 40$ was carried out on an average, modern PC. For each $D = 2, \ldots, 40$, all $D-1$ classical and robust regressions with the bootstrap inference were computed and the results (computation time in seconds) are presented in Figure 6. For less than 15 variables, the robust MM-regression with fast and robust bootstrap is even faster than the classical orthogonal regression with standard bootstrap (less than 1.5 minutes), but in higher dimensions the timings change. The classical orthogonal regression remains below 7 minutes up to $D = 40$, while the time consumed by the robust regression increases and becomes almost 50 minutes for $D = 40$.

A future extension of the package will include the weighted robust orthogonal regression [12] and the projection pursuit algorithm [3].

7. Summary and discussion

Within a composition, the measured variables are usually accompanied by a measurement error. Thus, if regression of one compositional part on other parts is to be done, the

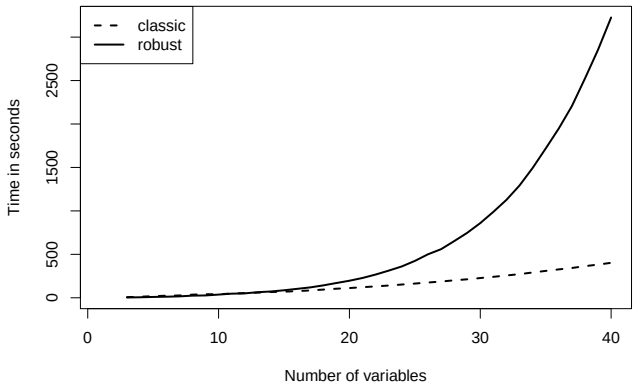

Figure 5. Computational performance of the robust and classical orthogonal regression for compositional data.

appropriate methodology is an errors-in-variable model, practically carried out through orthogonal regression. Since one is interested also in the interpretation of the model and in inference statistics, the task is to choose coordinates of the compositions which can be assigned to the measured parts. We have presented such a choice based on ilr coordinates, which have to be specified for each explanatory variable. All the different ilr coordinates are orthogonal transformations of each other. Due to the invariance of orthogonal regression to orthogonal rotations, the fit of the model does not change, and the intercept term remains the same.

The parameters of the orthogonal regression model can be estimated either via SVD (singular value decomposition) or by PCA (principal component analysis), since the outcomes of both methods are the same for the centered variables. The computational burden of both in the classical case is comparable up to moderate size of the data. Our interest was also in robust parameter estimation, since data outliers can have a strong effect on the classical estimates. Although procedures for robust SVD are available (e.g. [4]), it is more straightforward to employ a robust PCA procedure, since one simply can plug-in a robustly estimated covariance into the PCA procedure. Here we used the highly robust and efficient MM-estimator of location and scatter to robustify PCA, and thus estimate the parameters of the orthogonal regression model.

Statistical inference is particularly challenging in the robust case. Even if strict distributional assumptions are taken, like joint multivariate normal distribution of the response and the explanatory variables, it is not straightforward to derive the exact distribution of the robustly estimated regression parameters. In order to avoid these difficulties and in particular these strict assumptions, we used nonparametric bootstrap to estimate confidence intervals for the regression parameters and p-values of the tests for the parameters. This has been done for classical (least-squares based) and robust orthogonal regression. In the latter case, a much faster procedure for robust bootstrap has been employed, which is based on an approximation using fixed point equations.

Finally, the example has clearly shown that the proposed approach leads to meaningful results. Although there are no massive outliers in the data set, so that the classical and the robust analysis give about the same answer, they still differ in the significance of a regression parameter, important for interpretation purposes. From the economic perspective, it would also be interesting to investigate how the obtained relations change over time or how they change across different country groups, or even to employ more covariates (including non-compositional ones) for the regression purposes. We leave these topics for further research.

Acknowledgments

This research was supported by the grant COST Action CRoNoS IC1408 and the grant IGA_PrF_2015_013 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

References

[1] Aitchison J. The statistical analysis of compositional data. London: Chapman and Hall; 1986.
[2] Buccianti A, Pawlowsky-Glahn V, Egozcue JJ. Variation diagrams to statistically model the behavior of geochemical variables: Theory and applications. J Hydrol. 2014;519:988–998.
[3] Croux C, Fekri M, Ruiz-Gazen A. Fast and robust estimation of the multivariate errors in variables model. Test 2010;19:286–303.
[4] Croux C, Filzmoser P, Pison G, Rousseeuw P. Fitting multiplicative models by robust alternating regressions. Statistics and Computing 2003;13(1):23–36.
[5] Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997.
[6] Eaton ML. Multivariate Statistics. A Vector Space Approach. New York: Wiley; 1983.
[7] Egozcue JJ. Reply to "On the Harker Variation Diagrams" by J.A. Cortés. Math Geosci. 2009;41:829–834.
[8] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Math Geol. 2003;35:279–300.
[9] Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. Math Geol. 2005;37:795–828.
[10] Egozcue JJ, Pawlowsky-Glahn V. Simplicial geometry for compositional data. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, editors. Compositional data analysis in the geosciences: From theory to practice. Special Publications 264. London: Geological Society; 145–160.
[11] Egozcue JJ, Pawlowsky-Glahn V. Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A, editors. Compositional data analysis: Theory and applications. Chichester: Wiley; 2011. p. 12–28.
[12] Fekri M, Ruiz-Gazen A. Robust weighted orthogonal regression in the errors-invariables model. Journal of Multivariate Analysis, 2004;88:89–108.
[13] Filzmoser P. Robust principal components and factor analysis in the geostatistical treatment of environmental data. Environmetrics 1999;10:363–375.
[14] Filzmoser P. Hron K. Outlier detection for compositional data using classical and robust methods. Mathematical Geosciences, 2008;40;233–248.
[15] Filzmoser P, Hron K, Reimann C. Interpretation of multivariate outliers for compositional data. Comput Geosci. 2012;39:77–85.
[16] Fišerová E, Hron K. On interpretation of orthonormal coordinates for compositional data. Math Geosci. 2011;43:455–468.
[17] Fox, J. Bootstrapping Regression Models. Appendix to an R and S-PLUS Companion to Applied Regression. 2002
[18] Fuller, WA. Measurement Error Models. New York: Wiley; 1987.
[19] Hron K, Filzmoser P, Thompson K. Linear regression with compositional explanatory variables. J Appl Stat. 2012;39:1115–1128.
[20] Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T. PLS-DA for compositional data with application to metabolomics. Journal of Chemometrics 2015;29:21–28.
[21] Kynčlová P, Filzmoser P, Hron K. Compositional biplots including external noncompositional variables. Statistics 2016; DOI: 10.1080/02331888.2015.1135155.

[22] Markovsky, I., Van Huffel, S. Overview of total least-squares methods. Signal Processing 2007;87:2283–2302.
[23] Maronna RA, Martin RD, Yohai VJ. Robust statistics: Theory and methods. Chichester: Wiley; 2006.
[24] Maronna RA. Principal Components and Orthogonal Regression Based on Robust Scales. Technometrics, 2005;47:264–273.
[25] Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. Comput Stat Data An. 2012;56:2688–2704.
[26] Pawlowsky-Glahn V, Egozcue JJ. Geometric approach to statistical analysis on the simplex. SERRA 2001;15;384–398.
[27] Pawlowsky-Glahn V, Buccianti A, editors. Compositional data analysis: Theory and applications. Chichester: Wiley; 2011.
[28] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modeling and analysis of compositional data. Chichester: Wiley; 2015.
[29] Rousseeuw P, Hubert M. High-breakdown estimators of multivariate location and scatter. In Robustness and complex data structures. Springer, Heidelberg, 2013.
[30] Saikia D. Agriculture-industry interlinkages: Some theoretical and methodological issues in the Indian context, 2009, available at: http://mpra.ub.uni-muenchen.de/27820, last accessed on 1 December 2014.
[31] Saikia D. Analyzing inter-sectoral linkages in India, African Journal of Agricultural Research 2011;6:6766–6775.
[32] Salibian-Barrera M, Zamar RH. Bootstrapping robust estimates of regression. Ann Stat 2002;30:556–582.
[33] Salibian-Barrera M, Van Aelst S, Willems G. PCA based on multivariate MM-estimators with fast and robust bootstrap. J Am Stat Assoc 2006;101:1198–1211.
[34] Todorov V, Filzmoser P. An object oriented framework for robust multivariate analysis. Journal of Statistical Software 2009; 32/3.
[35] Van Aelst S, Willems G. Fast and robust bootstrap for multivariate inference: The R package FRB. Journal of Statistical Software 2013; 53/3.
[36] Zamar, RH. Robust estimation in the errors-in-variables model. Biometrika 1989;76(1):149–160.