**Noname manuscript No.**
(will be inserted by the editor)

# Discriminant analysis for compositional data and robust parameter estimation

**Peter Filzmoser · Karel Hron · Matthias Templ**

Received: date / Accepted: date

**Abstract** Compositional data, i.e. data including only relative information, need to be transformed prior to applying the standard discriminant analysis methods that are designed for the Euclidean space. Here it is investigated for linear, quadratic, and Fisher discriminant analysis, which of the transformations lead to invariance of the resulting discriminant rules. Moreover, it is shown that for robust parameter estimation not only an appropriate transformation, but also affine equivariant estimators of location and covariance are needed. An example and simulated data demonstrate the effects of working in an inappropriate space for discriminant analysis.

**Keywords** Compositional data · Logratio transformations · Discriminant analysis · Robustness

## 1 Introduction

Discriminant analysis is a widely used statistical method for supervised classification. While for the observations of a training data set the group memberships are known, discriminant analysis aims for reliable group assignments of new observations forming the test data set. The group assignment is based on a *discriminant rule*. Two well

P. Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology
Tel.: +43-1-58801-10733
Fax: +43-1-58801-10799
E-mail: P.Filzmoser@tuwien.ac.at

K. Hron
Department of Mathematical Analysis and Applications of Mathematics
Palacký University Olomouc, Faculty of Science
E-mail: hronk@seznam.cz

M. Templ
Department of Statistics and Probability Theory
Vienna University of Technology and Statistics Austria
E-mail: templ@statistik.tuwien.ac.at

established rules are the *Bayesian* and the *Fisher* discriminant rules (see, e.g., Johnson and Wichern, 2007). Generally speaking, the Bayesian rule results in linear or quadratic group separation, while the Fisher rule leads only to linear boundaries. However, the latter method allows for simple dimension reduction and thus for a better visualization of the separation boundaries.

Suppose that $n$ observation of a training data set are given which have been measured at $p$ characteristics. The $n$ observations are originating from $g$ different populations $\pi_1, \ldots, \pi_g$, with sample sizes $n_1, \ldots, n_g$, where $\sum_{j=1}^{g} n_j = n$. The $p$-variate observations will be denoted as column vectors $\mathbf{x}_{ij}$ with index $j = 1, \ldots, g$ representing the groups and index $i = 1, \ldots, n_j$ numbering the samples within a group.

We will first briefly describe the Bayesian discriminant rule. It is assumed that the observations from group $j = 1, \ldots, g$ have been sampled from a population $\pi_j$ with an underlying density function $\mathbf{f}_j$. Usually, $\mathbf{f}_j$ is assumed to be the $p$-variate normal density with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Moreover, $p_j$ denotes the prior probability of this group, with $\sum_{j=1}^{g} p_j = 1$. A new observation $\mathbf{x}$ from the test set is then assigned to that population $\pi_k$, $k \in \{1, \ldots, g\}$, for which the expression $\ln(p_j \mathbf{f}_j(\mathbf{x}))$ is maximal over all groups $j = 1, \ldots, g$. Using the assumption of normal densities for the groups, this rule is equivalent to assigning $\mathbf{x}$ to that group $k$ for which

$$d_j^Q(\mathbf{x}) = -\frac{1}{2}\ln[\det(\boldsymbol{\Sigma}_j)] - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) + \ln(p_j) \tag{1}$$

is the largest for $j = 1, \ldots, g$. The values $d_j^Q(\mathbf{x})$ are called *quadratic discriminant scores*, and the discussed method is called *quadratic discriminant analysis* (QDA). The assumption that all the group covariance matrices are equal, i.e. $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, leads to a simplification of equation (1), namely to the *linear discriminant scores*

$$d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln(p_j), \tag{2}$$

and accordingly to *linear discriminant analysis* (LDA). The assignment of $\mathbf{x}$ to a group is done in an analogous way.

For the Fisher discriminant rule (Fisher, 1938; Rao, 1948) the assumption of normal distribution of the groups is not explicitly required, although the method loses its optimality in case of deviations from normality. The rule is based on the matrix $\mathbf{B}$ describing the variation *between the groups*, and the matrix $\mathbf{W}$ denoting the *within groups covariance matrix*. Using the notation $\boldsymbol{\mu} = \sum_{j=1}^{g} p_j \boldsymbol{\mu}_j$ for the overall weighted mean of all populations, the two matrices are defined as

$$\mathbf{B} = \sum_{j=1}^{g} p_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^t$$

and

$$\mathbf{W} = \sum_{j=1}^{g} \boldsymbol{\Sigma}_j.$$

Under the assumption of equal group covariance matrices it can be shown that the best separation of the group means can be achieved by maximizing

$$\frac{\mathbf{a}^t \mathbf{B} \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} \qquad \text{for } \mathbf{a} \in \mathbf{R}^p, \mathbf{a} \neq \mathbf{0}. \tag{3}$$

The solution of this maximization problem is given by the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_l$ of the matrix $\mathbf{W}^{-1}\mathbf{B}$, scaled so that $\mathbf{v}_i^t \mathbf{W} \mathbf{v}_i = 1$ for $i = 1, \ldots, l$. The number $l$ of strictly positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ turns out to be $l \leq \min(g-1, p)$. Arranging these eigenvectors in the matrix $\mathbf{V}$ allows for a definition of the *Fisher discriminant scores*

$$d_j^F(\mathbf{x}) = \left[(\mathbf{x} - \boldsymbol{\mu}_j)^t \mathbf{V}\mathbf{V}^t(\mathbf{x} - \boldsymbol{\mu}_j) - 2\ln(p_j)\right]^{\frac{1}{2}} \tag{4}$$

for $j = 1, \ldots, g$. A new observation $\mathbf{x}$ is assigned to group $k$ if $d_k^F(\mathbf{x})$ is the smallest among the scores for all groups. The Fisher rule is equivalent to the linear discriminant rule (2) if the populations are normally distributed with equal covariance matrices and if $l$ is the number of strictly positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. However, as mentioned above the main advantage of the Fisher discriminant rule is its ability for dimension reduction. For example, by projecting the data in the space of the first two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, one obtains a data presentation in the plane that best captures the differences among the groups.

For the above discriminant rules several parameters have to be estimated. If the number of observations $n_j$ in the data is representative for the $j$-th group in the population, $n_j/n$ can be used to estimate $p_j$. However, if the data have not been sampled completely at random from the mixture, this estimate can be quite unrealistic, and sampling weights need to be included. For the quadratic discriminant scores (1) the group means $\boldsymbol{\mu}_j$ and covariance matrices $\boldsymbol{\Sigma}_j$ need to be estimated. This can be done in the classical way, by taking the arithmetic means $\bar{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{x}_{ij}$ and the sample covariance matrices $\mathbf{S}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^t$ for $j = 1, \ldots, g$. Alternatively, one can use robust estimators of location and covariance for the observations of each group. The MCD and the S estimator have been successfully used to robustify discriminant analysis (He and Fung, 2000; Hubert and Van Driessen, 2004). Both estimators have good robustness properties, and both are affine equivariant (see Maronna et al, 2006). A location estimator $T$ and a covariance estimator $C$ are called affine equivariant, if, for a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, for any nonsingular $p \times p$ matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathbf{R}^p$ the conditions

$$\begin{aligned} T(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) &= \mathbf{A}T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{b}, \\ C(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) &= \mathbf{A}C(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}^t \end{aligned}$$

are fulfilled. This property is important in the context of discriminant analysis for compositional data, as we will show later on.

For linear discriminant analysis a joint covariance matrix $\boldsymbol{\Sigma}$ is required. There are several possibilities for estimation; one way is to use a pooled estimate of the group covariance matrices. For the classical estimation this is

$$\mathbf{S}_{pooled} = \frac{\sum_{j=1}^{g}(n_j - 1)\mathbf{S}_j}{\sum_{j=1}^{g} n_j - g},$$

and in the robust case the robust group covariance matrices have to be averaged in an analogous way. Another method for deriving an estimate of the joint covariance matrix is to center the observations with the estimated group means, and to estimate the covariance matrix of the centered observations. In the robust case an iterative scheme can be used, by updating the estimated group means with the location estimate of the centered observations (for details, see He and Fung, 2000). Finally, for the Fisher

discriminant rule (4), in addition the overall weighted mean $\boldsymbol{\mu}$ needs to be estimated, which can be done either based on the arithmetic means of the groups, or using robust location estimates, for instance from the MCD or S estimator.

Besides appropriate parameter estimation, another aspect needs to be considered for discriminant analysis. Many practical data sets, like environmental data or data from official statistics, are compositional data. This means that the ratios between variables (usually called components or parts), rather than the reported values of the variables, contain the relevant information. Thus, scaling the observations to a constant sum, e.g. 1, or 100, when the data are expressed in percentages, has no effect on the information contained in the data. These properties have consequences for the statistical analysis, not only for correlation analysis (Pearson, 1897; Filzmoser and Hron, 2009), but also for discriminant analysis. The problem is not only a computational (singularity that arises from the constant sum constraint) as is sometimes wrongly stated (Bohling et al, 1998), but also a conceptual one (Barceló-Vidal et al, 1999). Namely, the standard statistical methods were developed for the case when absolute information is contained in the data, i.e. when the sample space follows the laws of the standard Euclidean geometry. However, compositional data with positive entries naturally induce another sample space, a simplex, with its own geometry. It is often called Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001; Egozcue and Pawlowsky-Glahn, 2006) with special operations of perturbation and power transformation, and the Aitchison inner product that enables us to construct the Euclidean geometry also on the simplex (for a more detailed discussion, see Pawlowsky-Glahn et al, 2008; Hron et al, 2010). Rather than working directly in the simplex, compositional data are usually transformed to the Euclidean space where standard statistical methods can be used. This is equivalent to finding a proper basis (or generating system) on the simplex and to expressing the compositions in coordinates. This approach induces a family of one-to-one logratio transformations, namely the additive, centered, and isometric logratio transformations (Aitchison, 1986; Egozcue et al, 2003) (see Section 2 for details).

Discriminant analysis is also very popular in the context of compositional data, and in the last decade a number of case studies using the logratio approach have been carried out (see, e.g., Gorelikova et al, 2006; Kovács et al, 2006; Martín-Fernández et al, 2005; Thomas and Aitchison, 2005, 2006; Von Eynatten et al, 2003). The major developments for discriminant analysis in the context of compositional data are summarized in Kovács et al (2006); Von Eynatten et al (2003). The identified groups were visualized by biplots for centered logratio transformed compositions, so called *compositional biplots* (Aitchison and Greenacre, 2002), being appropriate for visualizing the structure of compositional data. An important numerical result was that the misclassification rate turned out to be independent of the choice of the logratio transformation in the case of linear discriminant analysis (Kovács et al, 2006).

This paper combines both issues, robust parameter estimation and the compositional nature of the data for discriminant analysis. In Section 2 we will be interested in answering the question, which logratio transformations will be most suitable for discriminant analysis with compositional data, and whether the type of data transformation changes the results. These issues are also investigated for robust parameter estimation. Section 3 shows a numerical example and focuses on the geometry of discriminant rules for compositional data. A simulation study in Section 4 compares the

results for discriminant analysis with and without data transformation. The final Section 5 concludes.

## 2 Logratio transformations and invariance properties for discriminant analysis rules

For a $D$-part composition, $\mathbf{x} = (x_1, \ldots, x_D)^t$, the *additive logratio (alr) transformations* (Aitchison, 1986) from the simplex $\mathcal{S}^D$, the set of all $D$-part compositions, to the $(D-1)$-dimensional real space $\mathbf{R}^{D-1}$ are defined as

$$\mathbf{y}^{(j)} = (y_1^{(j)}, \ldots, y_{D-1}^{(j)})^t = \left( \ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right)^t. \qquad (5)$$

The index $j \in \{1, \ldots, D\}$ refers to the variable that is chosen as the ratioing variable in the transformation. The choice of the ratioing variable usually depends on the context of the data under consideration. Without loss of generality, we will choose the last part $x_D$ as the ratioing part, and we simplify the notation from $\mathbf{y}^{(D)}$ to $\mathbf{y}$. Results from the alr approach are usually quite easy to interpret, and therefore this approach is frequently found in applications. However, the alr transformation is not isometric and thus it should be rather avoided (Pawlowsky-Glahn et al, 2008). The main reason is that the corresponding basis on the simplex is not orthonormal with respect to the Aitchison geometry. The desired orthonormality is fulfilled for the *isometric logratio (ilr) transformations* from $\mathcal{S}^D$ to $\mathbf{R}^{D-1}$ (Egozcue et al, 2003), defined for a chosen orthonormal basis as

$$\mathbf{z} = (z_1, \ldots, z_{D-1})^t, \quad z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \quad \text{for } i = 1, \ldots, D-1. \qquad (6)$$

The ilr transformation can be viewed as expressing the compositions in orthonormal coordinates. The inverse ilr transformation is then obtained as

$$x_i = \exp\left( \sum_{j=i}^{D} \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} z_{i-1} \right) \quad \text{with } z_0 = z_D = 0 \quad \text{for } i = 1, \ldots, D. \qquad (7)$$

As the name of the transformation suggests, ilr already moves the whole Aitchison geometry on the simplex to the standard Euclidean geometry, and thus it seems to be the best choice for performing statistical analysis of compositions. However, the interpretation of the ilr variables is not easy and it is connected with a rather complicated procedure (Egozcue and Pawlowsky-Glahn, 2005). On the other hand, for discriminant analysis one is more interested in the structure of the observations than in single compositional parts, and thus their concrete interpretation is not of major importance.

Let us remark that the ilr transformations are strongly connected with the *centered logratio (clr) transformation* from $\mathcal{S}^D$ to $\mathbf{R}^D$,

$$\mathbf{w} = (w_1, \ldots, w_D)^t = \left( \ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)^t, \qquad (8)$$

introduced by Aitchison (1986). Although this transformation has a very useful interpretation via compositional biplots (Aitchison and Greenacre, 2002; Filzmoser et al, 2009), it produces singular observations of rank $D-1$ and thus it does not allow application of robust statistical methods. However, by choosing an orthonormal basis on the hyperplane $\mathcal{H} : w_1 + \ldots + w_D = 0$, formed by the clr transformation, we obtain exactly the ilr transformation as defined above.

All the logratio transformations are connected with linear relationships; following Aitchison (1986) and Egozcue et al (2003) we can summarize that

$$\mathbf{z} = \mathbf{U}\mathbf{w}, \quad \mathbf{y} = \mathbf{C}\mathbf{z}, \quad \mathbf{y} = \mathbf{F}\mathbf{w}. \tag{9}$$

The $(D-1) \times D$ matrices $\mathbf{U}$ and $\mathbf{F}$ are of full rank in rows and $\mathbf{U}\mathbf{U}^t = \mathbf{I}_{D-1}$, the identity matrix of order $D-1$; consequently $\mathbf{C} = \mathbf{F}\mathbf{U}^t$ is a regular matrix of order $D-1$. In more detail, the rows of the matrix $\mathbf{U}$ are the orthonormal basis vectors of the hyperplane $\mathcal{H}$ (Egozcue et al, 2003) and $\mathbf{F} = [\mathbf{I}_{D-1}, -\mathbf{1}_{D-1}]$, where $\mathbf{1}_{D-1}$ stands for the $(D-1)$ vector of ones; see Aitchison (1986) for details. Consequently,

$$\mathbf{w} = \mathbf{U}^t\mathbf{z}, \quad \mathbf{z} = \mathbf{C}^{-1}\mathbf{y}, \quad \mathbf{w} = \mathbf{F}^{+}\mathbf{y}, \tag{10}$$

where $\mathbf{F}^{+}$ denotes the Moore-Penrose inverse of the matrix $\mathbf{F}$. Note that the relations with the alr transformations could also be generalized to different choices of the ratioing part in the alr transformation due to an existing linear relationship between any two alr versions (see, e.g., Filzmoser and Hron, 2008). In addition, ilr transformations among each other are joined with an orthogonal relation, i.e. for $\mathbf{z}$ and $\mathbf{z}^*$ as results of the ilr transformation by different choices of the orthonormal basis on the simplex, there exists an orthonormal matrix $\mathbf{P}$ of order $D-1$ ($\mathbf{P}\mathbf{P}^t = \mathbf{P}^t\mathbf{P} = \mathbf{I}_{D-1}$) such that $\mathbf{z}^* = \mathbf{P}\mathbf{z}$.

With these relations we can investigate under which transformations the invariance of the discriminant rules is given:

**Theorem 1 (LDA):** The classical linear discriminant analysis rule is invariant to alr, ilr and clr transformations, i.e., the different versions of alr and ilr transformations, as well as the clr transformation lead to the same discriminant rules (see Kovács et al, 2006). For robust LDA, the same holds for alr and ilr transformations if affine equivariant robust estimators of location and covariance are taken.

This follows directly from the fact that (robust) LDA is based on (robust) Mahalanobis distances, being invariant to the logratio transformation chosen (see Filzmoser and Hron, 2008). Details are provided in the Appendix. Note that the clr transformation produces singular data, which cannot be used for robust affine equivariant location and covariance estimators.

**Theorem 2 (QDA):** The classical and robust quadratic discriminant analysis rules are invariant to alr transformations, i.e., decisions based on different versions of alr transformations are the same (also the discriminant scores are the same when changing from one alr transformation to another). The clr transformation leads to an ill-defined problem. The rules are invariant to ilr transformations (and the rules yield exactly the same values of the discriminant scores). All these properties are valid in the robust case if affine equivariant estimators of location and covariance are taken.

The proof is given in the Appendix.

**Theorem 3 (Fisher):** The classical Fisher discriminant rule is invariant to all alr, ilr and clr transformations. The robust rule is invariant to alr and ilr transformations if affine equivariant robust estimators of location and covariance are taken.

The proof is given in the Appendix.

Summarizing, any alr or ilr transformation is suitable for discriminant analysis. In the robust case it is essential to use affine equivariant estimators of location and covariance. In the following we will use an ilr transformation for both the data representation and the discrimination. In particular in the first case a representation in orthogonal coordinates has advantages over an oblique basis (Pawlowsky-Glahn et al, 2008).

## 3 Example

As an example data set we consider the distribution of labor force by status in employment, given for most countries of the world. The data set has been published by the United Nations Statistics Devision, and it contains the percentages of female and male employees, employers, own-account workers, and contributing family workers. For more information, see http://unstats.un.org/unsd/demographic/products/indwm/tab5c.htm. We combined the last two categories because of their rather low values, and denote them by "Self-employed". The resulting three compositional parts do not sum to 100% for females and males, because members of producers' cooperatives and workers not classifiable by status are not covered in the data. This, however, allows us to apply robust discriminant analysis to the original data because the robust covariance matrices are not singular. Figure 1 shows the results from LDA, applied to the ilr-transformed (left panels) and original (right panels) data, and for the classical (upper panels) and robust (lower panels) version. The original data are visualized in ternary diagrams (after centering) (Aitchison, 1986; Pawlowsky-Glahn et al, 2008). The symbols in the plots represent the data from females (triangles) and males (circles). It can be seen that the two groups are overlapping to some extent which may lead to misclassifications in discriminant analysis.

The ilr space consists of two variables $z_1$ and $z_2$ which can be easily visualized (Figure 1, left column). The ellipses in the plots indicate the structure of both groups. These are the 90% tolerance ellipses, meaning that in case of bivariate normal distribution they would include 90% of the observations of each group. For the upper left picture of Figure 1 the ellipses are constructed with classical estimates of group means and covariances, while for the lower left picture robust (MCD) estimates were used. Accordingly, classical and robust LDA leads to the separating lines in the plots. The misclassified observations are shown with black filled symbols, Since we use the same data set for deriving the classification rule and evaluation, the resulting error rate, called apparent error rate (AER), is usually too optimistic. Nevertheless, it gives an impression of the situation: the AER is 0.25 for classical LDA and 0.20 for robust LDA.

Since there exists a one-to-one relation between the ilr space and the original data space, the ellipses and the separating LDA lines can be back-transformed to the simplex in Figure 1 (right column). Due to the relative scale of compositions, the shapes of the ellipses and lines look rather unusual, especially at the boundary of the simplex (see, e.g., Pawlowsky-Glahn et al, 2008; Hron et al, 2010). LDA applied in the original space of the three compositions will lead to different results. Since the separating LDA

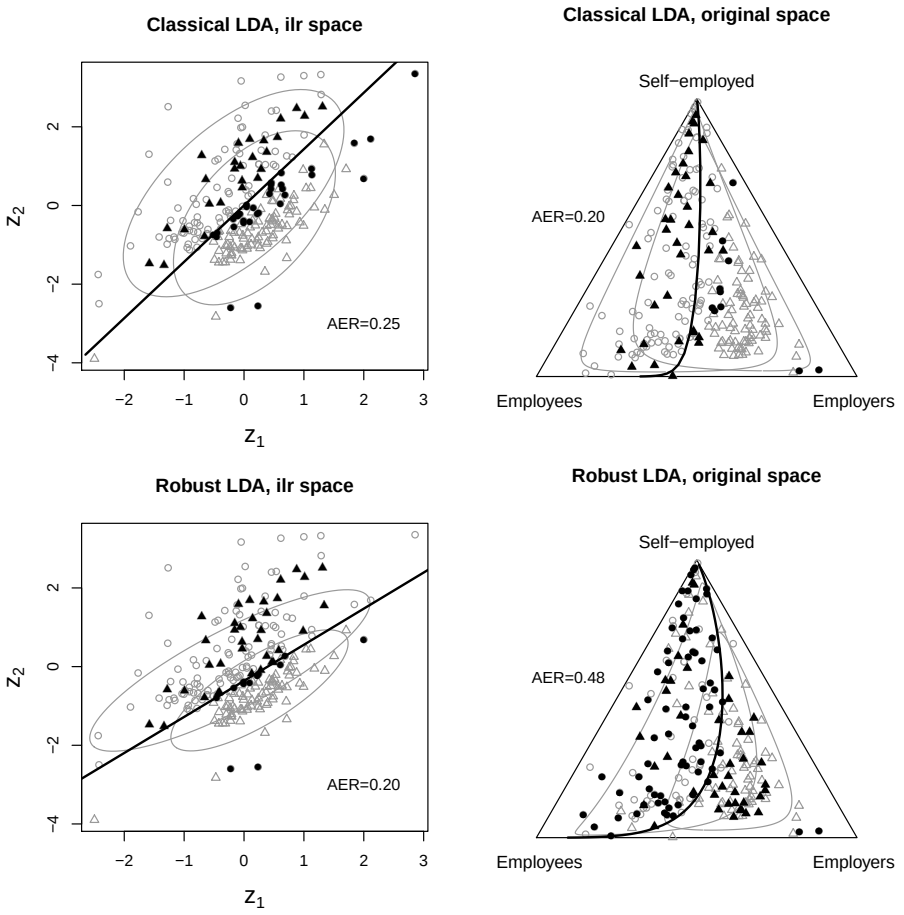


**Fig. 1** Employment data for females (triangles) and males (circles), presented in orthonormal coordinates (left column) and in the simplex (right column). The separation lines from classical (upper left) and robust (lower left) LDA applied in the ilr space as well as the 90% tolerance ellipses of the groups are shown in the ilr space and back-transformed to the simplex. The dark filled symbols represent the misclassified objects when classical and robust LDA is applied to the ilr space and to the original data space, respectively; see also the resulting apparent error rates (AER).

plane is difficult to visualize, we only show the resulting misclassified objects by black symbols. The AER for classical LDA is 0.20 (upper right), while the AER for robust (MCD-based) LDA is 0.48 (lower right).

This simple example already demonstrates the difficulties for discriminant analysis applied to multivariate observations, here specially to compositional data. Namely, depending on the dataset and the true distributions, the classification errors of different discriminant methods are not strictly ordered. As a consequence, working in the ilr space does not necessarily mean that the error rates are optimized; for classical LDA the AER was smaller in the simplex. Moreover, using robust rather than classical discriminant analysis in case of data containing outliers will also not necessarily lower

the misclassification rate: it could be reduced in the ilr space while in the simplex it was increased significantly. The outcome depends very much on the data structure and on the position of the outliers. In general, outliers will affect the covariance estimation, but it could happen that non-robustly estimated covariance matrices even appear more separated than robustly estimated ones, leading to smaller error rates (see, e.g., Croux et al, 2008). Similarly, points on the boundary of the simplex can have very different effects on discriminant analysis applied in the ilr space or in the original data space. In the next section we will use a simulation study to investigate in more detail how the position of the groups and the outliers in space affect the misclassification rates.

## 4 Simulation study

### 4.1 Two groups in low dimension

In the simulation study we will now discuss the patterns of the Aitchison geometry in more detail. To make the design simple, the simulations are based on using the normal distribution on the simplex, i.e. the ilr-transformed compositions follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (see Mateu-Figueras and Pawlowsky-Glahn, 2008, for details). According to practical experiences, the normal distribution on the simplex is often a reasonable assumption for the theoretical underlying distribution, followed by the majority of the data (Von Eynatten et al, 2003). To make the data yet more realistic, the sum of the compositional parts varies uniformly between 0.75 and 1, as this often occurs in geochemical data sets (Reimann et al, 2008).

In the *first design*, we have formed two groups with the same covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.4 & 0.3 \\ 0.3 & 0.4 \end{pmatrix}.$$

The mean vectors of the groups are chosen as $\boldsymbol{\mu}_1 = (-1+\delta, 0)^t$ and $\boldsymbol{\mu}_2 = (1+\delta, 0)^t$, respectively, where $\delta$ is varied from 0 to 3 within the simulation, using 25 equidistant steps. In each of these steps we generate 500 data sets according to the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$, with $n_1 = n_2 = 100$ observations. As in the example in the previous section, the generated data sets are treated as training data and test data at the same time, and thus the apparent error rate (AER) will be reported for discriminant analysis. Figure 2 shows the shapes of both simulated groups by 90% tolerance ellipses, where the solid ellipses correspond to $\delta = 0$ and the dashed ones to $\delta = 3$. The left picture is for the ilr space, while the right picture shows the simplex sample space. In the simplex we can see the same phenomenon as in Figure 1, namely that on the boundary of the simplex the geometrical structure of both groups appears much more different than in the center if the Euclidean geometry were applied to this space.

The results of the simulation are visualized in Figure 3 (left). Both linear and quadratic discriminations were performed together with their classical and robust versions. These methods were applied to the data in the original space and in the ilr space. So, in total 8 lines are shown, corresponding to the legends in Figure 3 left and right. For each of 25 steps of the simulation (horizontal axis) the discrimination procedures were applied to the 500 generated data sets; the average misclassification rate is shown (vertical axis). The results of discriminant analysis for the ilr-transformed data (triangles) are very similar, and there is practically no change when varying the

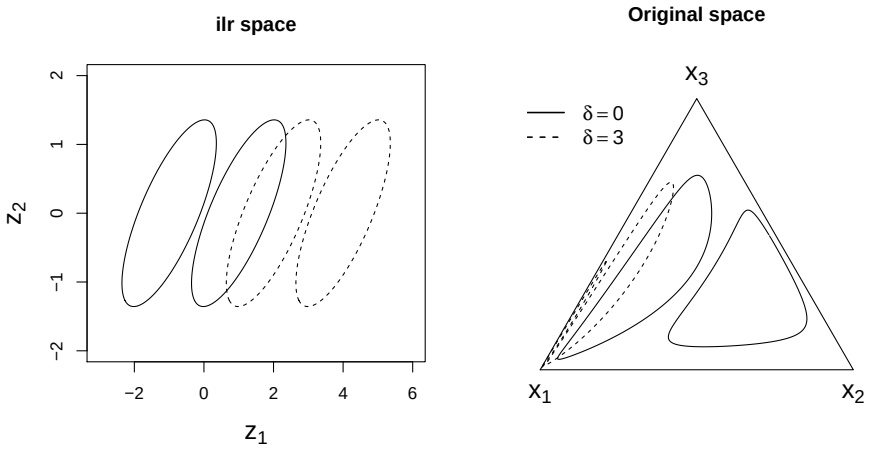


**Fig. 2** Simulated compositions according to the first design, displayed in the ilr space (left) and in the original space with the ternary diagram (right). The groups are shown by 90% tolerance ellipses, and their centers are shifted simultaneously.

shift parameter $\delta$. This is what can be expected, because the group covariance matrices are the same, there are no outliers, and the separation between the groups does not change with $\delta$. The results are quite different if the methods are applied in the original data space (circles). Robust LDA and QDA lead to higher misclassification rates than their classical counterparts, and QDA seems to be preferable to LDA. The latter result is logical because in the original space the two group covariance matrices are very different. Also the difference between classical and robust estimation is reasonable because the rules require normally distributed groups. The data points that deviate from this elliptical shape in the wrong geometry are treated as outliers by the robustified methods, and this leads to an increase of the error rate.

The first simulation design clearly shows how misleading it could be to analyze the data directly on the simplex. Although both groups are simulated with the same covariance structure, they differ a lot in the simplex sample space.

An analogous conclusion can be derived also from the *second design*, although this provides a different situation. Now the group centers are fixed at $\boldsymbol{\mu}_1 = (0,0)^t$ and $\boldsymbol{\mu}_2 = (2,1)^t$, but the group covariance matrices are varied. Both covariance matrices are taken as $\delta\boldsymbol{\Sigma}$, with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{pmatrix},$$

and $\delta$ varied from 0.3 to 1.2 in 25 equidistant steps, see Figure 4 (left). Thus, for increasing values of $\delta$ the groups will have more and more overlap, leading to an increase of the misclassification rate. Also in the original data space the groups tend to overlap more with increasing covariance. However, in this geometry the group sizes appear more and more unequal with increasing $\delta$, see Figure 4 (right).

As before, we generate 500 data sets in each of the 25 steps for $\delta$, with 100 observations in each group. We compute the average misclassification rate (average of AER) for classical and robust LDA and QDA, applied to the original and to the ilr-transformed data. The results are displayed in Figure 3 (right), and they show a simultaneous in-

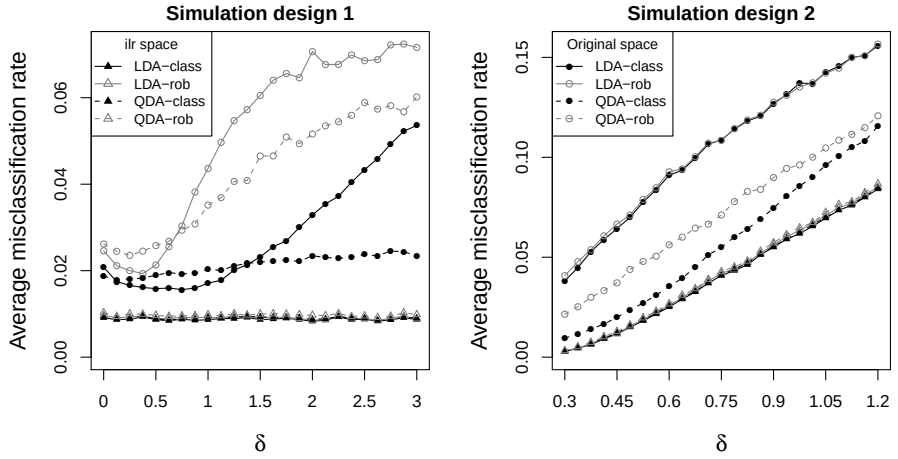


**Fig. 3** Results of the simulation study for the first and second design. The symbols, circles and triangles, refer to the discrimination of the original and ilr-transformed compositions, respectively. For both simulation designs, all methods yield almost the same results when working in the ilr space (four lines with symbols triangles on top of each other).

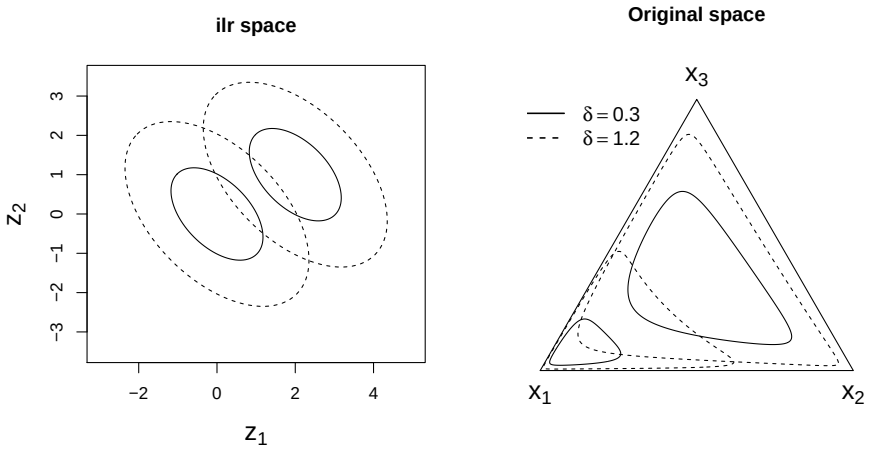


**Fig. 4** Simulated compositions according to the second design, displayed in the ilr space (left) and in the original space with the ternary diagram (right). The groups are shown by 90% tolerance ellipses, and their covariance matrices are increased simultaneously.

crease of all misclassification rates due to the increasing overlap. The best results are achieved by applying the methods in the ilr space. When working in the simplex, QDA is preferable over LDA due to the seemingly different group covariances. Robust LDA is worse than classical LDA for the same reasons as in the first simulation design.

### 4.2 Two groups in low dimension with outliers

The effects of outliers in the context LDA and QDA have been thoroughly studied in the literature (see, e.g., Hawkins and McLachlan, 1997; He and Fung, 2000; Croux and Dehon, 2001; Hubert and Van Driessen, 2004). We can expect the same effects when working in the corresponding geometry. However, it is unclear how outliers will affect discriminant analysis if this method is applied directly to the simplex sample space.

For this purpose we use the previous two simulation designs. The training sets are simulated with $\delta = 0$, and for the outliers we vary $\delta$ in the same way as before. In more detail, two data groups of 100 observations each are generated from normal distributions on the simplex, using the same means and covariances as before. In addition, each group includes 20 outlying observations, and the outliers are varied in 25 steps according to the varying parameter $\delta$. In each step we perform 500 replications. The error rate is computed for test data sets with 100 observations in each group, which are generated like the outlier-free training data. Thus, the outliers are only important for the computation of the discriminant rules, but they are not used for computing the misclassification rates.

Figure 5 shows the results. For the first simulation design (left picture) we can see how the outliers affect the classical discriminant methods in the ilr space. With increasing values of $\delta$ corresponding to increasing remoteness of the outliers the misclassification rates increase for the classical methods, but they remain stable for robust LDA and QDA. Interestingly, a small shift of the outliers (up to $delta = 0.5$) has no effect on the classical methods. When applying the methods in the simplex, the robust methods also lead to better results than their classical counterparts. This is in contrast to the results from the study without outliers (Figure 3, left).

The outliers in the second simulation design have a quite different effect, see Figure 5 (right). Since they are generated in both data groups, and since they are placed symmetrically around the centers of the distributions, they have practically no effect on classical and robust LDA and QDA in the ilr space. The behavior in the original simplex sample space is different. QDA leads to better results than LDA, because the covariances look different in this space. The contamination is asymmetric in the Euclidean geometry, and here this leads to a preference for the classical methods, especially in the case of QDA. The reason is that the wrong geometry distorts the normal distribution of the groups, and the robust method declares good data points as outliers, which results in erroneous discriminant rules.

### 4.3 Two groups in higher dimension

In this section we extend the simulation examples from Section 4.1 to higher dimension. We use the structure of the previous simulation designs, but now we simulate 10-dimensional normally distributed data on the simplex: For the first design we use the group means $\boldsymbol{\mu}_1 = (-1 + \delta, 0, \ldots, 0)^t$ and $\boldsymbol{\mu}_2 = (1 + \delta, 0, \ldots, 0)^t$, and a $(10 \times 10)$-dimensional covariance matrix $\boldsymbol{\Sigma}$ with values 0.4 in the main diagonal and values 0.3 as off-diagonal elements. For the second design we also use 10-dimensional data, with $\boldsymbol{\mu}_1 = (0, \ldots, 0)^t$, $\boldsymbol{\mu}_2 = (2, 1, \ldots, 1)^t$, and covariance matrix $\delta\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ has values -0.5 in the off-diagonal, and values 5 in the main diagonal (in order to be positive definite). All other parameters, including the values of $\delta$, are kept as in the original simulation designs.

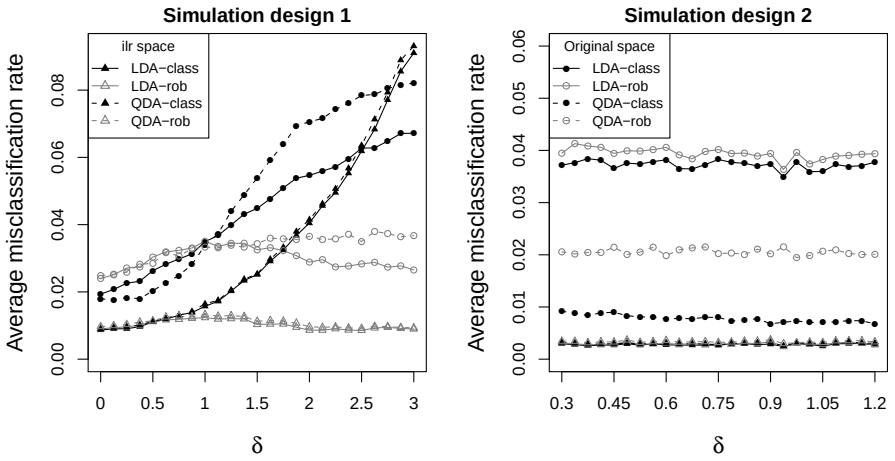


**Fig. 5** Results of the simulation study with outliers generated according to the first (left) and second (right) simulation design. For both simulation designs, the methods work best in the ilr space; in the first design, robust LDA and QDA yield stable results.

The results are presented in Figure 6, and they can be compared with the results for the low-dimensional data (Figure 3). For the methods applied in the ilr space (triangles) we can see the same behavior, except that the misclassification rates are generally lower due to better group separation in the higher-dimensional space. When working in the original data space (circles) we observe some major differences in the results. For simulation design 1 the robust discriminant analyses improve considerably with increasing $\delta$, until their misclassification rate becomes about constant for $\delta > 1$. The non-robust methods show a decreasing misclassification rate with increasing values of $\delta$. Thus, the closer we get to the boundary of the simplex (higher values of $\delta$), the more artifacts we can expect in the Euclidean geometry–and these artifacts depend on the dimensionality.

For simulation design 2 we get essentially the same results for the ilr-transformed data in lower and higher dimension (Figure 3 and 6, right). The results based on the original sample space show the same increasing trend of the misclassification rate as in lower dimension, but now robust QDA is worst, closely followed by classical LDA. This simulation example shows once more that discriminant analysis applied to the original data space may give unexpected or even unpredictable results.

It should be noted that the Fisher rule would give the same results as the LDA rule in all simulations, as long as no dimensionality reduction is performed, because both data groups had the same covariance structure and the same number of observations.

Rather than presenting further results with outliers, we apply Fisher discriminant analysis to the considered 10-dimensional data. We construct four groups, each with 100 data points: The first two groups consist of randomly generated data according to the first simulation design with parameter $\delta = 2$, and the remaining two groups are simulated according to the second design with $\delta = 0.3$. Figure 7 presents the results. We applied classical (left column) and robust (right column) Fisher discriminant analysis to the original (upper row) and ilr-transformed (lower row) data. Note that the

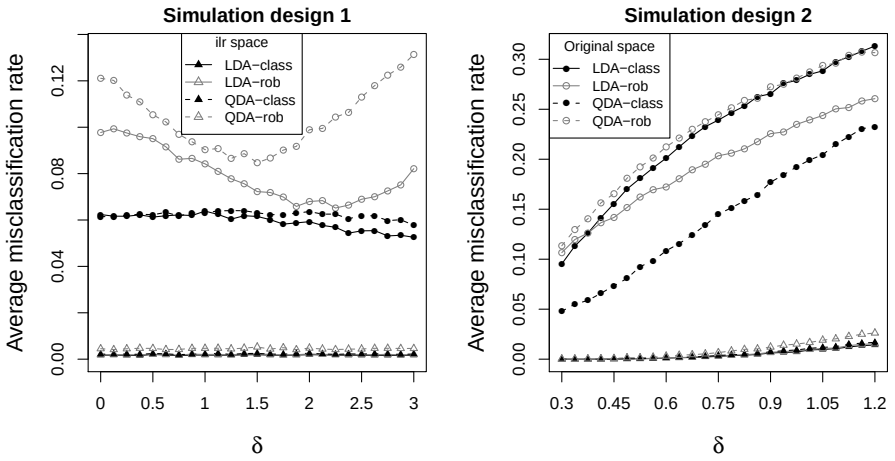


**Fig. 6** Results of the simulation study for the first (left) and second (right) design extended to 10 dimensions. The circles present the results for discriminant analysis applied to the original data, while the triangles are the results for the ilr-transformed data. For both simulation designs, all methods yield almost the same results when working in the ilr space (four lines with symbols triangles on top of each other).

Fisher rule is not optimal in this case, because the group covariances are different. The results are presented by the first and second Fisher direction, i.e. by the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, see Section 1. This is the two-dimensional projection that reveals the group separation in the best possible way, and the results confirm the findings in the previous simulations. If the Fisher method is applied to the ilr-transformed data, the plot shows the elliptical structure of the individual groups which are, however, overlapping due to their construction (Figure 7, lower row). An application to the original data makes the geometrical artifact visible. Figure 7 (upper row) suggests that the groups are not originating from an elliptical distribution, and that they are not linearly separable. Robustness is not changing much; a robust treatment is not able to repair an inappropriate geometry.

## 5 Conclusions

Many theoretical and practical results have been derived for compositional data, including investigations about their geometrical structure and appropriate transformations. We have shown that with respect to alr and ilr transformations, all considered discriminant rules are invariant, and that their robust versions are invariant as long as affine equivariant estimators of location and covariance are used.

The example and the simulation studies have shown that in an appropriately transformed space (alr or ilr) we can expect what we are used to expect from discriminant analysis: For example, differences in the covariance structure will usually give preference to QDA over LDA, or in presence of outliers in the data the robust versions of discriminant analysis are preferable to their classical counterparts.

It should also be mentioned that robust versions of discriminant analysis will not necessarily improve the misclassification rates, which depend very much on the position

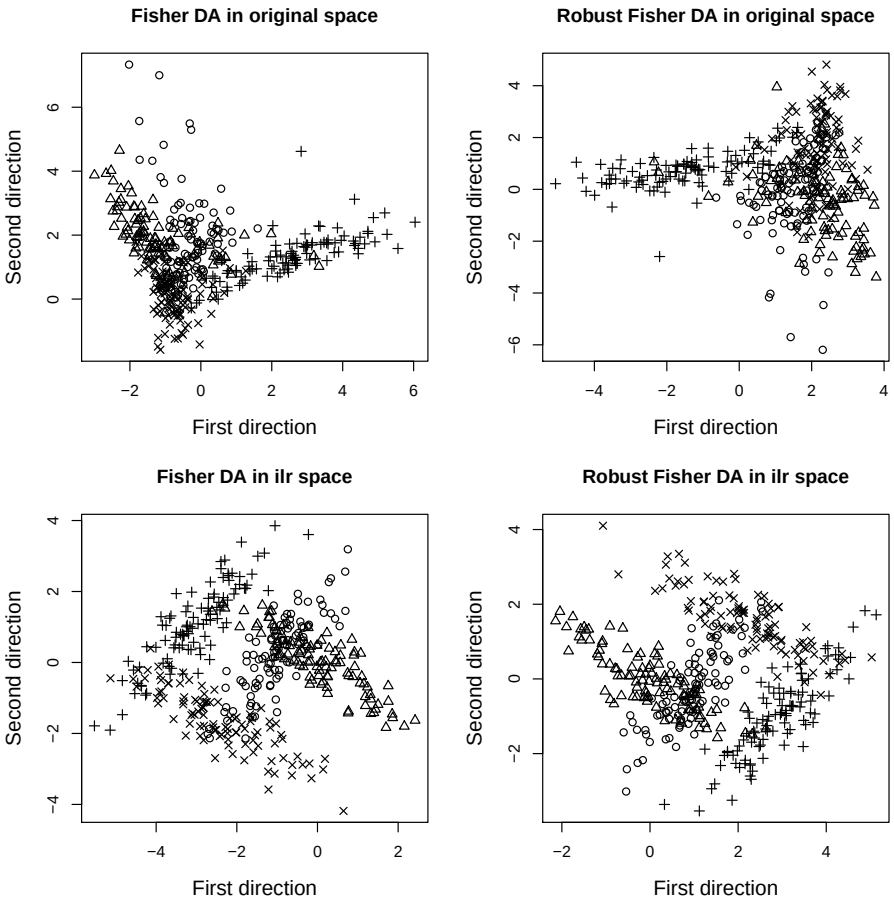


**Fig. 7** Presentation of four groups of 10-dimensional data by the first two Fisher directions. Left column: classical Fisher method; right column: robust Fisher method; upper row: Fisher rule applied to the original space; lower row: Fisher rule applied to the ilr-transformed space.

of the outliers in space: It can happen that classical parameter estimates where the group covariances are inflated by the outliers even cause the groups to appear more separated.

Not so well known are the properties of discriminant analysis in the simplex sample space. We have shown with simulated data that even if the data structure is very clear in Euclidean space, it can appear very distorted in the simplex. Especially on the boundary of the simplex the geometrical artifacts can become severe. This has consequences for discriminant analysis applied to the original data: Robust discriminant analysis methods try to cope with these artifacts by declaring extreme observations–thus typically observations that are on the boundary of the simplex–as outliers, although they are only appearing as outlying because of the wrong geometry. Thus, the concept of robustness is not really the solution to the problem–the solution would be an appropriate data transformation. These "outliers" will then affect the misclassification rates from

classical and robust discriminant analysis, and again the performance of the methods depends on the position of the outliers.

One could argue that for discriminant analysis we are only interested in a small misclassification rate (for the test set), independent of the geometry used for presenting the data. In other words, for a given data (test) set one could apply different variants of discriminant analysis to the original data, as well as to other transformations of the data. The optimal discriminant rule would then be based on that method and for that geometrical space where the misclassification rate (or, even better, a function of all individual group misclassifications) is minimized. This approach might have practical appeal, but it can be very misleading if plots are shown and interpreted: regular observations may appear as outliers, group covariances may appear very different, etc. Moreover, due to an incorrect geometrical presentation the results of discriminant analysis are "unpredictable". They do not depend only on the group structure, but in addition on the location in the simplex. Especially the latter influence is difficult or even impossible to visualize or assess.

## Acknowledgements

The authors are very grateful to Prof. Gerald van den Boogaart for his valuable comments and suggestions. We would like to thank the editor, the associate editor, and the referees for their constructive comments that greatly improved the presentation of these results. This work was supported by the Council of the Czech Government MSM 6198959214.

## References

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman and Hall, London

Aitchison J, Greenacre M (2002) Biplots of compositional data. Applied Statistics 51:375–392

Barceló-Vidal C, Martín-Fernández J, Pawlowsky-Glahn V (1999) Comment on 'Singularity and nonnormality in the classification of compositional data' by G.C. Bohling, J.C. Davis, R.A. Olea, and J. Harff. Mathematical Geology 31(5):581–585

Bohling G, Davis J, Olea R, J H (1998) Singularity and nonnormality in the classification of compositional data. Mathematical Geology 30(1):5–20

Croux C, Dehon C (2001) Robust linear discriminant analysis using $S$-estimators. The Canadian Journal of Statistics 29:473–492

Croux C, Filzmoser P, Joossens K (2008) Classification efficiencies for robust linear discriminant analysis. Statistica Sinica (18):581–599

Egozcue J, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7):795–828

Egozcue J, Pawlowsky-Glahn V (2006) Simplicial geometry for compositional data. In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional Data Analysis in the geosciences: From Theory to Practice, Geological Society, London, pp 145–160

Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Mathematical Geosciences 40(3):233–248

Filzmoser P, Hron K (2009) Correlation analysis for compositional data. Mathematical Geosciences 41:905–919

Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics 20:621–632

Fisher RA (1938) The statistical utilization of multiple measurements. Annals of Eugenics 8:376–386

Gorelikova N, Tolosana-Delgado R, Pawlowsky-Glahn V, Khanchuk A, Gonevchuk V (2006) Discriminating geodynamical regimes of tin ore formation using trace element composition of cassiterite: the Sikhote'Alin case (Far Eastern Russia). In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional Data Analysis in the geosciences: From Theory to Practice, Geological Society, London, pp 43–57

Hawkins D, McLachlan G (1997) High-breakdown linear discriminant analysis. Journal of the Americal Statistical Association 92:136–143

He X, Fung W (2000) High breakdown estimation for multiple populations with applications to discriminant analysis. Journal of Multivariate Statistics 72:151–162

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

Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Computational Statistics and Data Analysis (45):301–320

Johnson RA, Wichern DW (2007) Applied Multivariate Statistical Analysis, sixth edn. Prentice Hall, New York

Kovács L, Kovács G, Martín-Fernández J, Barceló-Vidal C (2006) Major-oxide compositional discrimination in Cenozoic volcanites of Hungary. In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional Data Analysis in the geosciences: From Theory to Practice, Geological Society, London, pp 11–23

Maronna R, Martin R, Yohai V (2006) Robust Statistics: Theory and Methods. John Wiley & Sons, New York

Martín-Fernández J, Barceló-Vidal C, Pawlowsky-Glahn V, Kovács L, Kovács G (2005) Subcompositional patterns in Cenozoic volcanic rocks of Hungary. Mathematical Geology 37(7):729–752

Mateu-Figueras G, Pawlowsky-Glahn V (2008) A critical approach to probability laws in geochemistry. Mathematical Geosciences 40:489–502

Pawlowsky-Glahn V, Egozcue J (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environ Res Risk Assess 15(5):384–398

Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2008) Lecture Notes on Compositional Data Analysis. Universitat de Girona, URL http://hdl.handle.net/10256/297

Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60:489–502

Rao CR (1948) The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society, Series B 10:159–203

Reimann C, Filzmoser P, Garrett R, Dutter R (2008) Statistical Data Analysis Explained. Applied Environmental Statistics with R. John Wiley, Chichester

Thomas C, Aitchison J (2005) Compositional data analysis of geological variability and process: a case study. Mathematical Geology 37(7):753–772

Thomas C, Aitchison J (2006) Log-ratios and geochemical discrimination of Scottish Dalradian limestones: a case study. In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional Data Analysis in the geosciences: From Theory to Practice, Geological Society, London, pp 25–41

Von Eynatten H, Barceló-Vidal C, Pawlowsky-Glahn V (2003) Composition and discrimination of sandstones: a statistical evaluation of different analytical methods. Journal of Sedimentary Research 73(1):47–57

## Appendix

**Proof of Theorem 1:**

From Barceló-Vidal et al (1999) and in particular from Filzmoser and Hron (2008), where the invariance of the Mahalanobis distance under alr and ilr transformations for affine equivariant robust (and classical) estimators of location and covariance was proven, it directly follows that the values of the expressions $\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\mathbf{x}$ and $\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$ in Equation (2) remain unchanged. Moreover, it is easy to see that the Moore-Penrose inverse of the clr covariance matrix $\boldsymbol{\Sigma}_\mathbf{w}$ can be expressed as a linear transformation of the inverse ilr covariance matrix, $\boldsymbol{\Sigma}_\mathbf{w}^+ = \mathbf{U}^t\boldsymbol{\Sigma}^{-1}\mathbf{U}$, where $\mathbf{U}\mathbf{U}^t = \mathbf{I}_{D-1}$. Thus, as in the case of clr-transformed compositions, the inverse of the covariance matrix is replaced exactly by its pseudo-inverse. In the classical case this invariance can be extended also to the clr transformation.

**Proof of Theorem 2:**

Let $\mathbf{u}$ and $\mathbf{v}$ any different alr and/or ilr transformations of a composition $\mathbf{x}$. Then there exists a non-singular matrix $\mathbf{A}$ of order $D-1$ such that $\mathbf{v} = \mathbf{A}\mathbf{u}$. Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ denote the mean and covariance matrix of the $j$-th group of $\mathbf{u}$, respectively. Then we can immediately see from the QDA rule in Equation (1) that

$$d_j^Q(\mathbf{v}) = -\frac{1}{2}\ln(\det(\mathbf{A}\boldsymbol{\Sigma}_j\mathbf{A}^t)) - \frac{1}{2}(\mathbf{A}\mathbf{u}-\mathbf{A}\boldsymbol{\mu}_j)^t(\mathbf{A}\boldsymbol{\Sigma}_j\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{u}-\mathbf{A}\boldsymbol{\mu}_j) + \ln(p_j) =$$

$$= -\frac{1}{2}\ln(\det(\mathbf{A})\det(\boldsymbol{\Sigma}_j)\det(\mathbf{A}^t)) - \frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_j)^t\mathbf{A}^t(\mathbf{A}^t)^{-1}\boldsymbol{\Sigma}_j^{-1}\mathbf{A}^{-1}\mathbf{A}(\mathbf{u}-\boldsymbol{\mu}_j) + \ln(p_j) =$$

$$= d_j^Q(\mathbf{u}) - \frac{1}{2}\ln(\det(\mathbf{A}^2)),$$

where the property $\det(\mathbf{A}) = \det(\mathbf{A}^t)$ was utilized. Now we can distinguish four cases for the transformation from $\mathbf{u}$ to $\mathbf{v}$:

- $\mathbf{u}$ and $\mathbf{v}$ are from two different ilr transformations: Then $\mathbf{A}$ is orthogonal, $\det(\mathbf{A}^2) = 1$, and the scores are unchanged.
- $\mathbf{u}$ and $\mathbf{v}$ are from two different alr transformations: The simplest form of a transformation matrix $\mathbf{A}$ between two different alr transformations is an upper diagonal matrix. The determinant of such a matrix is the product of the diagonal elements, which is $-1$. Any other alr transformation can be built by a matrix where columns and/or rows of the simple transformation matrix are permuted. The determinant of a permuted matrix is only multiplied by $+1$ or $-1$. This means that $\det(\mathbf{A}^2) = 1$, and the scores remain unchanged.

- $\mathbf{u}$ is represented in an ilr space, and $\mathbf{v}$ in an alr space: Then the matrix $\mathbf{A}$ linking $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{A} = \mathbf{F}\mathbf{U}^t$, see Section 2. In general, the determinant of $\mathbf{F}\mathbf{U}^t$ is different from $\pm 1$, and thus the ilr scores differ from the alr scores by a dimension depending constant. The result of the discrimination remains the same.
- $\mathbf{u}$ is represented in an alr space, and $\mathbf{v}$ in an ilr space: This is analoguous to the previous point.

The clr transformation gives a determinant of zero, and the logarithm of zero is not defined. All the above properties are valid for location and covariance estimators that are affine equivariant.

**Proof of Theorem 3:**

As in the proof of Theorem 2 we use the results of two different alr and/or ilr transformations $\mathbf{u}$ and $\mathbf{v}$ with the relation $\mathbf{v} = \mathbf{A}\mathbf{u}$. Then the within and between groups covariance matrices are related by

$$\mathbf{W_v} = \mathbf{A}\mathbf{W_u}\mathbf{A}^t, \quad \mathbf{B_v} = \mathbf{A}\mathbf{B_u}\mathbf{A}^t$$

and thus

$$(\mathbf{W_v})^{-1}\mathbf{B_v} = (\mathbf{A}^t)^{-1}\mathbf{W_u}^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{B_u}\mathbf{A}^t = (\mathbf{A}^t)^{-1}\mathbf{W_u}^{-1}\mathbf{B_u}\mathbf{A}^t.$$

The $l \times (D-1)$ matrix $\mathbf{V}$ from the spectral decomposition of $\mathbf{W_u}^{-1}\mathbf{B_u}$ becomes $(\mathbf{A}^t)^{-1}\mathbf{V}$ for $\mathbf{W_v}^{-1}\mathbf{B_v}$ (belonging to the same strictly positive eigenvalues). The Fisher discriminant scores are then

$$d_j^F(\mathbf{v}) = \left((\mathbf{A}\mathbf{u} - \mathbf{A}\boldsymbol{\mu}_j)^t(\mathbf{A}^t)^{-1}\mathbf{V}\mathbf{V}^t\mathbf{A}^{-1}(\mathbf{A}\mathbf{u} - \mathbf{A}\boldsymbol{\mu}_j) - 2\ln(p_j)\right)^{\frac{1}{2}} =$$

$$= \left((\mathbf{u} - \boldsymbol{\mu}_j)^t\mathbf{A}^t(\mathbf{A}^t)^{-1}\mathbf{V}\mathbf{V}^t\mathbf{A}^{-1}\mathbf{A}(\mathbf{u} - \boldsymbol{\mu}_j) - 2\ln(p_j)\right)^{\frac{1}{2}} = d_j^F(\mathbf{u})$$

for $j = 1, \ldots, g$. Thus, different alr and ilr transformations are invariant for the Fisher rule as long as location and covariance estimators are affine equivariant.

For classical estimators, the invariance can be extended also to clr transformation (using the relation to ilr transformations). Here analogously the relation $\mathbf{U}\mathbf{U}^t = \mathbf{I}_{D-1}$ of the $(D-1) \times D$ matrix $\mathbf{U}$ from Equation (10) will be utilized. For

$$\mathbf{W_w} = \mathbf{U}^t\mathbf{W_z}\mathbf{U}, \quad \mathbf{B_w} = \mathbf{U}^t\mathbf{B_z}\mathbf{U}$$

we obtain

$$\mathbf{W_w^+}\mathbf{B_w} = \mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U}\mathbf{U}^t\mathbf{B_z}\mathbf{U} = \mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{B_z}\mathbf{U}.$$

where the properties of $\mathbf{W_w^+}$ as the Moore-Penrose inverse of $\mathbf{W_w}$ are fulfilled. Namely

$$\begin{aligned}
&\mathbf{W_w}\mathbf{W_w^+}\mathbf{W_w} = \mathbf{U}^t\mathbf{W_z}\mathbf{U}\mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U}\mathbf{U}^t\mathbf{W_z}\mathbf{U} = \mathbf{U}^t\mathbf{W_z}\mathbf{U} = \mathbf{W_w},\\
&\mathbf{W_w^+}\mathbf{W_w}\mathbf{W_w^+} = \mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U}\mathbf{U}^t\mathbf{W_z}\mathbf{U}\mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U} = \mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U} = \mathbf{W_w^+},\\
&\mathbf{W_w}\mathbf{W_w^+} = \mathbf{U}^t\mathbf{W_z}\mathbf{U}\mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U} = \mathbf{U}^t\mathbf{U},\\
&\mathbf{W_w^+}\mathbf{W_w} = \mathbf{U}^t\mathbf{W_z}^{-1}\mathbf{U}\mathbf{U}^t\mathbf{W_z}\mathbf{U} = \mathbf{U}^t\mathbf{U},
\end{aligned}$$

where the matrix $\mathbf{U}^t\mathbf{U}$ is obviously a symmetric matrix. Finally, for $j = 1, \ldots, g$ we get

$$d_j^F(\mathbf{w}) = \left((\mathbf{U}^t\mathbf{z} - \mathbf{U}^t\boldsymbol{\mu}_j)^t\mathbf{U}^t\mathbf{V}\mathbf{V}^t\mathbf{U}(\mathbf{U}^t\mathbf{z} - \mathbf{U}^t\boldsymbol{\mu}_j) - 2\ln(p_j)\right)^{\frac{1}{2}} =$$

$$= \left((\mathbf{z} - \boldsymbol{\mu}_j)^t\mathbf{U}\mathbf{U}^t\mathbf{V}\mathbf{V}^t\mathbf{U}\mathbf{U}^t(\mathbf{z} - \boldsymbol{\mu}_j) - 2\ln(p_j)\right)^{\frac{1}{2}} = d_j^F(\mathbf{z})$$

which shows that the Fisher rule for clr-transformed data gives the same results as for ilr-transformed data.