Discriminant Analysis for Compositional Data Incorporating Cell-Wise Uncertainties

Solveig Pospiech¹ ·
Raimon Tolosana-Delgado¹ ·
K. Gerald van den Boogaart^{1,2}

Received: 30 January 2020 / Accepted: 8 June 2020 / Published online: 20 July 2020
© The Author(s) 2020

Abstract In the geosciences it is still uncommon to include measurement uncertainties into statistical methods such as discriminant analysis, but, especially for trace elements, measurement uncertainties are frequently of relevant size. Uncertainties can be reported by each measured variable, by each observation or by individual cells (i.e., each observation has an individual uncertainty for each variable). Most methods incorporating uncertainties use the uncertainties as weights for the variables or observations of the data set. The method proposed in this contribution uses variance additivity properties and generalised least squares to calculate better estimates of group variances and group means, which then influence the decision rules of linear and quadratic discrimination algorithms. This methodological framework allows incorporation of cell-wise uncertainties, and would be largely valid if the information about co-dependency between variable errors within each observation were reported. The method is also appropriate for incorporating uncertainties into compositional data sets—for example, those formed by concentrations, proportions, percentages or any other form of information about the relative abundance of a set of components forming a whole—even if such uncertainties are nearly never reported considering this compositional nature. The methods are illustrated by means of case studies with simulated data.

✉ Solveig Pospiech
s.pospiech@hzdr.de

¹ Modelling and Evaluation, Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Chemnitz Str. 40, 09599 Freiberg, Germany

² Institute for Stochastics, Technische Universität Bergakademie Freiberg, Prüferstrasse 1, 09599 Freiberg, Germany

Graphical abstract

Keywords Weighted discriminant analysis · Linear discriminant analysis · Quadratic discriminant analysis · CoDa · Cell-wise uncertainty · Measurement uncertainty · Geochemical data

1 Introduction

All measurements come with measurement uncertainties. Many statistical procedures are not considering measurement uncertainties because they are often either I) small with respect to the overall variability or II) of the same variability for each measurement within a dataset and can thus be treated as part of the overall variability of the measurand.

In geochemistry, however, measurement uncertainties are quite frequently of relevant size, especially when looking at trace elements. They can also differ between observations, for example if measurements from different laboratories or measuring systems need to be combined or if the measurands had been analysed at a sensor range outside the linear sensor response.

Moreover, the implementation of uncertainty-aware methods is further complicated in geosciences because data sets often present a compositional nature, for example they are given as concentrations, proportions, percentages or any other form of information about the relative abundance of a set of components forming a whole (Aitchison 1986; Pawlowsky-Glahn et al. 2015). For these data the respective uncertainties are nearly never reported considering their compositional nature, that is, in terms of relative errors. Even for standard methods like discriminant analysis (DA, linear or quadratic) tools incorporating uncertainties of multivariate data, including compositional data, are not developed yet. This publication discusses a method to incorporate cell-wise uncertainties for compositional and non-compositional data into DA and how the structure of the uncertainties affects the discrimination function.

'Uncertainty' in the geochemical context is defined in Potts (2012) following the International Vocabulary of Metrology (BIPM et al. 2008) guidelines as 'non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used'. The uncertainties are often expressed as an interval around the (unbiased estimated) mean of the measurements. For example often a mean zero Gaussian distribution is assumed and reported by an n -fold of its standard deviation (typically $n = 1, 2$ or 3). Regarding the source of the uncertainties it has to be distinguished between systematic uncertainties and independent, random uncertainties. Systematic uncertainties result from systematic errors which are also referred to as bias (Potts 2012; BIPM et al. 2008). Biases can occur if the offset from the 'true' value of the measurand depend on the order of samples for analysis, on the sampling scheme, the person conducting the experiment, etc. Random dispersions occur for example due to the analytical uncertainty of the measurement device, the (in)homogeneity of samples, etc. This paper focuses on the case where all reported uncertainties are independent and random.

Uncertainties can be considered in a statistical analysis either by each measured variable, by each observation or by using the individual, cell-wise uncertainties. The term 'cell-wise uncertainties' will here be used for a data set of d analysed variables where each observation has an individual uncertainty for each of the d variables conforming it. Hence, a data set of $n \times d$ data values has associated a data set of $n \times d$ individual uncertainties. There are several methods for incorporating variable-wise or observation-wise uncertainties into DA. In this contribution the proposed method uses uncertainties to recalculate better estimates of the true group variances and true group means. This methodological framework does not only allow to incorporate cell-wise uncertainties, but also would largely be valid if the information about the co-dependency between uncertainties within each observation would be reported.

This paper is structured as follows: First comes a short review of the quadratic and linear discriminant analysis. This review comprises also the basic concepts of compositional data analysis and the log-transformations required for this publication. The main part describes how individual (cell-wise) uncertainties can be incorporated in discriminant analysis, with a special emphasis on this method applied for compositional data. In the last part case studies for simulated data and real data are presented to visualize the method.

2 Review of Discriminant Analysis and Compositional Data

2.1 Concepts and Notation

In the following, the i th observation (sample) $\mathbf{z}_i$ of a data set is written as a vector of d values (for example d elements)

$$\mathbf{z}_i = [z_{1i}, z_{2i}, \dots, z_{di}].$$

This observation is considered to be randomly distributed around a nominal value $\boldsymbol{\mu}_i = [\mu_{1i}, \mu_{2i}, \dots, \mu_{di}]$, with a variability

$$S_i = \text{var}[\mathbf{z}_i - \boldsymbol{\mu}_i] \in \mathbb{R}^{d \times d}.$$

One could for instance obtain an estimate of this variability from repeated measurements of the same sample.

The uncertainties of the observation $\mathbf{z}_i$ are almost always reported as n -fold standard deviations for each variable. This corresponds to using only the diagonal elements of these variance-covariance matrices S_i , id est the variances. Nevertheless, the methodology presented in this paper is also valid when the full matrices are available, for instance constructed from the individual variances plus the correlations between components.

A DA requires a data set $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\}$ of n observations available which is split into several groups. The group index is denoted by $g = 1, 2, \dots, K$, and the number of observations per group by $n_1, n_2, \dots, n_K$. The set G_g represents the set of indices of observations belonging to the group g .

The group means $\boldsymbol{\mu}_g$ and the group variance matrices S_g^z are estimated by the standard formulas for the method of moments. The estimator $\bar{\mathbf{z}}_g$ of the true mean $\boldsymbol{\mu}_g$ of each group mean is

$$\bar{\mathbf{z}}_g = \frac{1}{n_g} \sum_{i \in G_g} \mathbf{z}_i \quad (1)$$

and the unbiased estimate of the group variance S_g^z is

$$\hat{S}_g^z = \frac{1}{n_g - 1} \sum_{i \in G_g} (\mathbf{z}_i - \bar{\mathbf{z}}_g)(\mathbf{z}_i - \bar{\mathbf{z}}_g)^t. \quad (2)$$

In case all group variances can be assumed to be equal they are estimated as

$$\hat{S}^z = \frac{1}{n - K} \sum_g \sum_{i \in G_g} (\mathbf{z}_i - \bar{\mathbf{z}}_g)(\mathbf{z}_i - \bar{\mathbf{z}}_g)^t. \quad (3)$$

2.2 Quadratic and Linear Functions for Discriminant Analysis

Discrimination analysis (DA) is a classical tool in statistics which is used to train a classification rule assigning a new observation $\mathbf{z}_0$ to one group out of several given. This contribution focuses on linear and quadratic discriminant analyses (resp. LDA and QDA), both providing a probabilistic approach to the classification problem: initially each group is given a prior probability $\tilde{p}_g$, which is updated via a Bayes rule with evidence from the data to obtain the posterior probability. This discrimination function can be used either to calculate for each observation a posterior probability for each given group, or to determine which variables provide the evidence for the discrimination. QDA should be used to discriminate between groups which exhibit strong variability structure, while LDA can be used for discrimination if the group variances can be assumed to be similar. If the assumption of homogeneous group variances is given, the LDA can be valuable option, because it is simpler and less likely to produce overfit.

For normal distributed data the estimated density functions for a new observation $\mathbf{z}_0$ (ignoring the uncertainties of $\bar{\mathbf{z}}_g$ and $\hat{S}_g^z$) is

$$f_g(\mathbf{z}_0) = \frac{1}{\sqrt{(2\pi)^d |\hat{S}_g^z|}} \exp\left(-\frac{1}{2}(\mathbf{z}_0 - \bar{\mathbf{z}}_g)^t (\hat{S}_g^z)^{-1} (\mathbf{z}_0 - \bar{\mathbf{z}}_g)\right) \quad (4)$$

where $|\hat{S}_g^z|$ denotes the determinant of matrix $\hat{S}_g^z$. The posterior probability $P_g(\mathbf{z}_0)$ of observation $\mathbf{z}_0$ to belong to group g can be calculated via Bayesian updating as the product of the estimated density function $f_g(\mathbf{z}_0)$ and the prior probability $\tilde{p}_g$

$$P_g(\mathbf{z}_0) = \frac{f_g(\mathbf{z}_0) \cdot \tilde{p}_g}{\sum_{g'=1}^G f_{g'}(\mathbf{z}_0) \cdot \tilde{p}_{g'}}. \quad (5)$$

2.2.1 Quadratic Discriminant Analysis

The method of QDA uses estimated density functions for each group to estimate the probability of an unknown sample to belong to the respective groups. The classification function for QDA with different group variances can be derived from the Gaussian density function by taking logarithms and eliminating additive constants

$$\begin{aligned} D_g(\mathbf{z}_0) &= -\frac{1}{2} \log |\hat{S}_g^z| - \frac{1}{2}(\mathbf{z}_0 - \bar{\mathbf{z}}_g)^t (\hat{S}_g^z)^{-1} (\mathbf{z}_0 - \bar{\mathbf{z}}_g) + \log \tilde{p}_g \\ &= \underbrace{-\frac{1}{2} \log |\hat{S}_g^z|}_{\text{constant}} + \underbrace{\mathbf{z}_0^t (\hat{S}_g^z)^{-1} \bar{\mathbf{z}}_g}_{\text{linear term of } \mathbf{z}_0} - \underbrace{\frac{1}{2} \bar{\mathbf{z}}_g^t (\hat{S}_g^z)^{-1} \bar{\mathbf{z}}_g}_{\text{constant}} - \underbrace{\frac{1}{2} \mathbf{z}_0^t (\hat{S}_g^z)^{-1} \mathbf{z}_0}_{\text{quadratic term of } \mathbf{z}_0} + \underbrace{\log \tilde{p}_g}_{\text{constant}} \end{aligned} \quad (6)$$

For a certain observation $\mathbf{z}_0$ one computes $D_g(\mathbf{z}_0)$ for each group. Then the posterior probability of Eq. (5) can be recovered by normalizing the exponential of each $D_g(\mathbf{z}_0)$ by the sum of all exponentials of $D_g(\mathbf{z}_0)$, so that the vector $\kappa [\exp(D_1(\mathbf{z}_0)), \exp(D_2(\mathbf{z}_0)), \dots, \exp(D_K(\mathbf{z}_0))]$ sums to one, id est with

$$\kappa = \exp(D_1(\mathbf{z}_0)) + \exp(D_2(\mathbf{z}_0)) + \dots + \exp(D_K(\mathbf{z}_0)) \quad (7)$$

2.2.2 Linear Discriminant Analysis

The method of LDA assumes the same covariance matrix for all groups: $\hat{S}_g^z = \hat{S}^z, \forall g$. The classification function Eq. (6) simplifies then to

$$D_g(\mathbf{z}_0) = \mathbf{z}_0^t (\hat{S}^z)^{-1} \bar{\mathbf{z}}_g - \frac{1}{2} \bar{\mathbf{z}}_g^t (\hat{S}^z)^{-1} \bar{\mathbf{z}}_g + \log \tilde{p}_g \quad (8)$$

by eliminating all terms that are independent of the group index, for example $\log |\hat{S}_g^z|$ or $\mathbf{z}_0^t (\hat{S}_g^z)^{-1} \mathbf{z}_0$. Probabilities can be recovered from these discriminant functions equivalent to for QDA.

2.3 Compositional Data and Log-Ratio Transformations

Compositional data (CoDa) are vectors of positive components which inform of the relative contribution of the set of parts forming a total (Aitchison 1986; Pawłowsky-Glahn et al. 2015). Commonly they are identified with closed data, for example where these components sum up to a constant total sum c . A frequently used type of compositional data in geochemistry are concentrations given in percentages % (constant sum $c = 100$) or parts per million ppm (constant sum $c = 10^6$). Here we denote a composition as a vector of D components $\mathbf{z} = [z_1, z_2, \dots, z_D]$ such that

$$z_j > 0 \quad \text{and} \quad \sum_{j=1}^D z_j = c. \quad (9)$$

It is common that only $d, d \leq D$, components have been analyzed or are considered as important in the modeling task. In this case one speaks of a *subcomposition*. Because for calculating with compositional data the sum of components is irrelevant since only the relative information is needed, we can safely take any subcomposition for the following study.

One important principle for working with compositional data is the scale invariance Aitchison and Pawłowsky-Glahn (1997). This states that results have to be independent of any scaling applied to the data, such as, for instance, the units selected to represent the data. This can be achieved for example by working with log-ratios.

There are various log-ratio transformations. Two frequently used are the centered log-ratio (clr) and isometric log-ratio (ilr). For a detailed description of properties and fields of application of the various log-ratio transformations refer to Egozcue and Pawłowsky-Glahn (2011).

For a centered log-ratio transformation (Aitchison 1986) all parts of the composition are set into relation to the geometric mean (g_m) of the composition

$$\text{clr}(\mathbf{z}) = \left[\ln \left(\frac{z_j}{g_m(\mathbf{z})} \right) \right]_{j=1, \dots, d} \quad \text{with} \quad g_m(\mathbf{z}) = \left(\prod_{j=1}^d z_j \right)^{\frac{1}{d}}. \quad (10)$$

The centered log-ratio transformation results in a vector of d log-ratios that sum to zero. The isometric log-ratio transformation, introduced by Egozcue et al. (2003), is related to the clr by a set of $d - 1$ linear combinations of the clr-coefficients

$$\text{ilr}_i(\mathbf{z}) = \mathbf{v}_i^t \text{clr}(\mathbf{z}) = \sum_{j=1}^d v_{ij} \text{clr}_j(\mathbf{z}), \quad (11)$$

where the $(d - 1)$ vectors $\mathbf{v}_i = [v_{i1}, v_{i2}, \dots, v_{id}]$ with zero-sum constraint are pairwise orthogonal and they are all normalized to length 1. Alternatively, this can be written for all ilr scores at the same time in matrix form

$$\text{ilr}(\mathbf{z}) = V^t \text{clr}(\mathbf{z}), \quad (12)$$

with $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{d-1}] = [v_{ji}]$ containing by columns the vectors of coefficients of each ilr score. In this way, the variance-covariances of the log-transformed data and the ilr transformed data are then related by

$$\text{var}(\text{ilr}(\mathbf{z}_i)) = V^t \text{var}(\text{clr}(\mathbf{z}_i)) V = V^t \text{var}(\log(\mathbf{z}_i)) V, \quad (13)$$

relations readily invertible as

$$\text{clr}(\mathbf{z}) = V \text{ilr}(\mathbf{z}) \quad \text{and} \quad \text{var}(\text{clr}(\mathbf{z})) = V \text{var}(\text{ilr}(\mathbf{z})) V^t. \quad (14)$$

The advantage of an ilr transformation above the clr transformation resides on the fact that the ilr-covariance matrix is invertible. This is required by the methods discussed in this publication.

3 Incorporating Individual Uncertainties in Discriminant Analysis

3.1 Data and Uncertainty Pre-processing

The DA presented in this contribution can be used with data sets from a variety of scales, in particular multivariate interval scale, multivariate ratio scale and compositional scale (Stevens 1946; van den Boogaart and Tolosana-Delgado 2013).

The methods described from the next section on require all variables to be in multivariate interval scale. Consequently, for multivariate ratio and compositional scales it is necessary to express first the data (and their uncertainties) in interval scales. This step is straightforward for the data, as it only requires to transform its values through logarithms (ratio scale) or log-ratios (compositional scale). The appropriate expression of uncertainties is the focus of this section, because uncertainties are often reported as absolute values, for example by commercial laboratories.

Let the cell-wise uncertainties be given as relative uncertainties, $\Delta = \frac{sd(X)}{E[X]}$, a one-fold standard deviation around the estimated value. The amount z_{ij} of component j in sample i can be decomposed in its nominal value μ_{ij} and its residual ϵ_{ij}

$$z_{ij} = \mu_{ij} + \epsilon_{ij}. \quad (15)$$

Transferring z_{ij} into the log-space and applying the law of error propagation results into

$$\log(z_{ij}) = \log(\mu_{ij}) + \frac{1}{\mu_{ij}} \epsilon_{ij} + o(\epsilon_{ij}^2). \quad (16)$$

If the analytical uncertainty, for example the uncertainty from the machines, is small then $o(\epsilon_{ij}^2) \rightarrow o_g$ and the log-uncertainties of the cell-wise uncertainties can be simplified to

$$sd(\log(z_{ij})|\mu_i) \approx \frac{sd(\epsilon_{ij})}{\mu_{ij}} = \Delta_{ij}. \quad (17)$$

This approximates the relative uncertainties by considering it equal to log-standard deviations of the measurement error.

The individual variance induced by uncertainties can be expressed as a diagonal matrix, filled with the log-transformed uncertainties

$$\text{var}(\log(\mathbf{z}_i)|\boldsymbol{\mu}_i) \approx \begin{pmatrix} \Delta_{i1}^2 & & 0 \\ & \ddots & \\ 0 & & \Delta_{id}^2 \end{pmatrix} = S_i \quad (18)$$

Equation (18) provides thus an expression of the uncertainty of data in multivariate ratio scale. To obtain the uncertainty expressed in compositional scale we just need to project it into for example the ilr space, with Eq. (13). Equivalent expressions exist for projecting into the clr-space, but these are not appropriate in discriminant analysis due to the singularity of the clr-covariance matrix.

3.2 Discriminant Analysis Incorporating Uncertainties

3.2.1 Estimation of Group Variances Incorporating Uncertainties

As stated in Eq. (15) the observed compositions $\mathbf{z}_i$ deviate from their respective true value $\boldsymbol{\mu}_i$. If a group variance is derived from the observed values $\{\mathbf{z}_i\}$ it will be addressed in the following as the 'total' or 'observed' group variance S_g^z . The so called 'true' or 'natural' group variance S_g^μ is the theoretical value if the group variances would be calculated from the true values $\boldsymbol{\mu}$ of the observations $\mathbf{z}$. The observed groups variances S_g^z are composed of the true variability of the groups and the variability induced by various sources of uncertainties

$$S_g^\delta = \frac{1}{n_g} \sum_{i \in G_g} S_i, \quad (19)$$

which by additivity of the variance gives

$$S_g^z = S_g^\mu + S_g^\delta, \quad (20)$$

assuming that uncertainties are independent of nominal values.

If the presence of uncertainties is not accounted for, the decision rules given in Eq. (6) respectively in Eq. (8) are based on the observed group variances S_g^z . But this observed group variance S_g^z might deviate notably from the true group variance. In particular, if a variable presents a huge error in a few observations and a very small one in other observations, the resulting increased variability will obscure the ability of this variable to discriminate between groups even for the observations with low error.

A better estimation closer to the true group variance can be achieved by correcting the total variance by the individual uncertainties S_i : It can be shown by straight forward calculations that

$$E \left[\hat{S}_g^z \right] = S_g^\mu + \frac{1}{n_g} \sum_{i \in G_g} S_i. \quad (21)$$

Thus, S_g^μ can be estimated unbiasedly by

$$\hat{S}_g^\mu = \hat{S}_g^z - \frac{1}{n_g} \sum_{i \in G_g} S_i. \quad (22)$$

It should be noted that through the subtraction the estimate of the group variance could become non-positive definite. This extreme must be monitored, and eventually ensured that the variance matrix is positive definite, for example by forcing all eigenvalues of the matrix to be non-negative if a few eigenvalues are slightly to moderately negative. If the eigenvalues are notably below zero, this proposed method is not feasible: In that case, the variables showing negative eigenvalues should be discarded.

If the group variances can be assumed to be equal, for example if a LDA should be applied, the group variance for all groups can be estimated by

$$\hat{S}^\mu = \hat{S}^z - \frac{1}{n} \sum_g \sum_{i \in G_g} S_i. \quad (23)$$

3.2.2 Estimation of Group Means Incorporating Uncertainties

The individual uncertainties affect also the estimation of the group means. In correspondence with Eq. (2), an estimation of the group means incorporating the individual uncertainties is calculated by generalized least squares. This assumes weighting matrices for each observation inversely proportional to the uncertainty of that observation, id est $\Lambda_i \propto (S_g^\mu + S_i)^{-1}$, subject to $\sum \Lambda_i = I_{(d,d)}$, the identity matrix. Accordingly, the estimated general least squared (GLS) group mean $\bar{z}_g^\mu$ can be calculated by

$$\bar{z}_g^\mu = \left(\sum_{i \in G_g} (\hat{S}_g^\mu + S_i)^{-1} \right)^{-1} \sum_{i \in G_g} (\hat{S}_g^\mu + S_i)^{-1} \mathbf{z}_i. \quad (24)$$

Under the assumption that $\hat{S}_g^\mu$ is the correct variance this multivariate version of a weighted mean is the best linear unbiased estimator for this situation. Like every weighted mean it reacts unstable in case of negative or very large weights. For multivariate weights this corresponds to negative or large eigenvalues. It is thus important to ensure that the estimated variances are positive definite and do not contain unrealistically small eigenvalues. In our cases we achieved this through ensuring the positive definiteness of $\hat{S}_g^\mu$ by setting negative eigenvalues to 0. It is worth noting that this weighting does not ensure that the weighted mean is in the multivariate convex hull of the data. This is an effects of different weighting of the vector component in different datapoints in case of different errors and can occur in any estimator considering the different precisions, like component-wise weighted means, maximum likelihood and Bayes estimators, for the same situation.

The GLS estimation of the group means $\bar{z}_g^\mu$ has again an uncertainty, which can be described by the estimated variance of $\bar{z}_g^\mu$ to the true group mean μ_g

$$\hat{S}_{\mu_g}^\mu = \widehat{\text{var}}(\bar{z}_g^\mu - \mu_g) = \left(\sum_{i \in G_g} (\hat{S}_g^\mu + S_i) \right)^{-1} \quad (25)$$

3.2.3 Implementing the Corrected Estimations of Group Variances and Means

The probability of an observation $\mathbf{z}_0$ to belong to the respective groups is calculated by the classification functions for Eq. (6) for QDA and Eq. (8) for LDA. Normally, if uncertainties are not considered, the observed group variance(s) are used, S_g^z , as well as the observed group means $\bar{z}_g$. If the uncertainties are used to correct the estimations of group variance(s) and group means this results into using $\bar{z}_g = \bar{z}_g^\mu$ and $\hat{S}_g^z = \hat{S}_g^\mu + S_i$, estimation of the true group variance combined with the individual variance.

Calculating the classification functions for a new observation $\mathbf{z}_0$ to retrieve the group probabilities, the function for QDA from Eq. (6) becomes

$$D_g(\mathbf{z}_0) = -\frac{1}{2} \log |\hat{S}_g^\mu + S_0| - \frac{1}{2} (\mathbf{z}_0 - \bar{z}_g^\mu)^t (\hat{S}_g^\mu + S_0)^{-1} (\mathbf{z}_0 - \bar{z}_g^\mu) + \log \tilde{p}_g \quad (26)$$

and for LDA from Eq. (8)

$$D_g(\mathbf{z}_0) = \mathbf{z}_0^t (\hat{S}^\mu + S_0)^{-1} \bar{z}_g^\mu - \frac{1}{2} (\bar{z}_g^\mu)^t (\hat{S}^\mu + S_0)^{-1} \bar{z}_g^\mu + \log \tilde{p}_g. \quad (27)$$

As mentioned before, if the observations are multivariate ratio respectively compositional data, all data and associated uncertainties have to be transferred into an adequate log space respectively log-ratio space prior to calculating the decision rules. As a last step the posterior probabilities are calculated by the density functions and the prior probabilities with Eq. (5).

Finally, if the focus lies on determining the parameters of the decision rule, the density function must be computed by using the group variances $\hat{S}_g^\mu$ without extending them with the individual error. The resulting classification rule describes the influence of each variable in the ideal situation that analytical errors would be negligible, and represents the part of the classification rule that is attributable to natural phenomena, and thus interpretable.

4 Case Studies

4.1 Simulated Compositional and Non-compositional Data

For the two case studies with simulated data two data sets of 200 observations were simulated, 100 observations of 'Group 1' and 100 observations of 'Group 2'. For both data sets first 'true' data had been simulated. To calculate the 'observed data' errors had been simulated for each data point and added to the true data. The errors had been

simulated by firstly generating values from a chi-squared distribution with high number of degree of freedom, according to the numbers of observations. For each variable and each group these values had been multiplied with chosen variance of errors for each variable. Secondly, the measurement errors had been simulated by generating for each observation a value from the normal distribution with mean 0 and variance being the previously generated value from the chi-squared distribution. The relative difference of observed to true data are used as uncertainties. On both sets of observed data a LDA, a LDA incorporating the uncertainties, a QDA and a QDA incorporating the uncertainties had been applied to discriminate between the two groups of observations. The one data set contains two variables, 'Var1' and 'Var2', with normally distributed values which are not compositional ('non-compositional'/'normal' data). The other data set contains three variables, 'Var1', 'Var2' and 'Var3', forming a compositional data set. It is of compositional nature because for each observation the values of the three variables sum up to 1. The results for the second data set are displayed by ternary diagrams because the ternary diagrams represent better the constrained nature of the data.

The posterior likelihood for the LDA and the QDA without uncertainties had been calculated by the built-in functions of the statistic software R (R Core Team 2019), package 'MASS' (Venables and Ripley 2002). The posterior likelihood for the LDA and the QDA incorporating uncertainties had been calculated by own functions modified after the R-function 'qda.default'. The functions are published in the R-package 'vdar' (Pospiech and van den Boogaart 2020). The code is based on the Eq. (27) and Eq. (26). For calculating the log-ratio transformations of data and uncertainties the package 'compositions' (van den Boogaart and Tolosana-Delgado 2008) had been used. The graphics had been generated by the R-packages 'ggplot2' (Wickham 2016) and 'ggtern' (Hamilton and Ferry 2018).

The results of the DAs are shown in Fig. 1 for the 'non-compositional' data and in Fig. 2 for the example of compositional data. The observed data are coloured by the posterior probability for an observation to belong to Group 1.

The plots 'LDA' and 'QDA' in Fig. 1 and in Fig. 2 show that for the individual observations the values for the posterior likelihood for Group 1 are very similar for these two methods. Even for the observations with posterior likelihood values between 0.25 and 0.75 display only minor differences between LDA and QDA. These results are an effect of the structure of the group variances, which are approximately similar.

The plots 'LDA with uncertainty' and 'QDA with uncertainty' in Figs. 1 and 2 display the effect of incorporating the individual uncertainties on the posterior probability values as compared to the values resulting from the standard DAs. For these methods the likelihood values are not changing continuously between the two group means, but individual observations can have notably different likelihood values than neighbouring observations. Observations for which the observed values are considerably different than the true values have in this simulation also high uncertainties and the corresponding likelihoods can significantly differ from values derived by the standard DAs. For example, some observations of Group 1 have for the standard DAs posterior probability values of < 0.1 , but for the DAs incorporating the uncertainties the values > 0.2 . This shows, that the method of incorporating the individual uncer-

Fig. 1 Discrimination analysis of two groups non-compositional data set

Fig. 2 Discrimination analysis of two groups compositional data set

Table 1 Results of discrimination analysis for case study 1

	Group 1	Group 2
LDA Group 1	91	9
LDA Group 2	6	94
LDA w/u Group 1	92	8
LDA w/u Group 2	8	92
QDA Group 1	92	8
QDA Group 2	8	92
QDA w/u Group 1	92	8
QDA w/u Group 2	6	94

The table shows how many observations of which group had been classified into which group. If the posterior likelihood for Group 1 is above 0.5 the class of the observation is counted as Group 1, if it is below 0.5 the class is counted as Group 2. 'w/u' refers to 'with uncertainties'

Table 2 Results of discrimination analysis for case study 2

	Group 1	Group 2
LDA Group 1	95	5
LDA Group 2	7	93
LDA w/u Group 1	98	2
LDA w/u Group 2	5	95
QDA Group 1	96	4
QDA Group 2	7	93
QDA w/u Group 1	96	4
QDA w/u Group 2	6	94

The table shows how many observations of which group had been classified into which group. If the posterior likelihood for Group 1 is above 0.5 the class of the observation is counted as Group 1, if it is below 0.5 the class is counted as Group 2. 'w/u' refers to 'with uncertainties'

tainties changes the posterior probability of the individual observations with respect to the uncertainty.

Incorporating the uncertainties may result into either a better discrimination or poorer discrimination between the groups, depending on the relation between natural group variance S_g^μ and the variance induced by other sources, the S_g^δ . For these two simulations the discrimination between Group 1 and Group 2 are similarly good for all four DA methods (Tables 1 and 2). In Table 1 the LDA incorporating uncertainties is the best discrimination function and in Table 2 the QDA incorporating uncertainties is the best discrimination

4.2 Plant Geochemistry Data

The example data set originates from a study about trace element uptake of crop plants for biogas production (Fahlbusch 2018; Fahlbusch 2018). The selected plant data

Table 3 Overview of plant species per field

Location	Species	Number of samples
Reinshof	Amaranth	36
	Faba Bean	32
	Maize	33
	Ryegrass	24
	Triticale	30
	other species	63
Soemmerling	Amaranth	9
	Annual Flowering Mix	12
	Faba Bean	29
	Maize	11
	Ryegrass	24
	Triticale	30
	other species	57

comprise various crop plants mainly of the family of Poaceae which had been harvested on two fields in Germany, one close to the city Uslar, called 'Soemmerling' and one close to the city Göttingen, called Reinshof', between June 2015 and November 2016 (Table 3). Plant geochemistry is controlled by the geochemistry of the soil, but also many other parameters, such as fertilizer applications, dust contaminations, climate, etc. The discriminant analysis was applied to distinguish crop plants from the field 'Soemmerling' to plants from the field 'Reinshof' independent of annual or seasonal variations like field treatment and adhering particles on leaf material. This data set is used to demonstrate how the incorporation of individual errors affects the discrimination analysis.

All plant samples had been dissolved by a four acid total digestion (HNO_3 , HCl , HClO_4 , HF), and after removal of the acids and dissolution in 2% HNO_3 analysed by ICP-MS and ICP-AES. This analytical system allows to measure broad spectrum of elements, ranging from main elements to numerous trace and ultra-trace elements like the rare earth elements (REE). Still, several ultra-trace elements show very weak precision and accuracy. In this data set these are B, Be, Ga, Ge, Ir, Nb, Pt, Se, Sn, Te, V and W and they are discarded from the data set. The procedure of the four acid digestion dissolves not only organic components but also inorganic plant components and potential adhering dust or soil particles. The elements being mainly affected by contamination through soil particles are REE and Al, Bi, Cs, Fe, Hf, Li, Sb, Sc, Ta, Th, Ti, Tl, U, Y and Zr (Pospiech et al. 2017). These are also discarded to avoid an annual or seasonal effects, for example if plants in one harvesting period had more adhering soil particles than in an other year or harvesting period. The elements typically heavily affected by fertilizer application are in this data set K, S and P, and eventually also Ca. The elements which can be used for discrimination analysis between the two fields only based on element uptake without strong annual or seasonal effects are Ba, (Ca,) Cd, Cr, Cu, Mg, Mn, Mo, Na, Rb, Sr and Zn.

Table 4 Summary of relative uncertainties (in %) per element for the used plant data set

Ba	Ca	Cd	Cr	Cu	Mg
Min. : 3.1	Min. :8.4	Min. : 3.9	Min. : 3	Min. :3.1	Min. : 1.1
1st Qu.: 4.5	1st Qu.:8.4	1st Qu.: 8.5	1st Qu.: 61	1st Qu.:3.1	1st Qu.: 6.0
Median : 6.0	Median :8.4	Median : 18.8	Median : 109	Median :3.1	Median : 8.1
Mean : 8.4	Mean :8.4	Mean : 26.4	Mean : 42853	Mean :3.1	Mean : 9.5
3rd Qu.: 8.7	3rd Qu.:8.4	3rd Qu.: 30.3	3rd Qu.: 165	3rd Qu.:3.1	3rd Qu.:12.5
Max. :46.2	Max. :8.4	Max. :280.7	Max. :3416894	Max. :3.1	Max. :22.3
Mn	Mo	Na	Rb	Sr	Zn
Min. :10	Min. : 0.65	Min. : 7.9	Min. :18	Min. :3.7	Min. :2.9
1st Qu.:10	1st Qu.: 3.45	1st Qu.: 7.9	1st Qu.:18	1st Qu.:3.7	1st Qu.:2.9
Median :10	Median : 7.19	Median : 7.9	Median :18	Median :3.7	Median :2.9
Mean :10	Mean : 10.15	Mean : 10.3	Mean :18	Mean :3.7	Mean :2.9
3rd Qu.:10	3rd Qu.: 12.49	3rd Qu.: 7.9	3rd Qu.:18	3rd Qu.:3.7	3rd Qu.:2.9
Max. :10	Max. :229.16	Max. :204.1	Max. :18	Max. :3.7	Max. :2.9

Values rounded to two significant digits

The relative uncertainties for each sample and element had been derived by an uncertainty model which takes into account the measured concentration of the measurand. The parameter of the uncertainty model are derived from a separate data set of measurements of blank samples and replicate measurements (replicate including digestion) of four different plant certified reference materials (CRM): GBW 10052 (green tea), NCS DC 73349 (bush leaves), WEPAL IPE 126 (maize) and WEPAL IPE 168 (sunflower). The CRMs are digested and measured together with samples in the respective sample batches. Please note that ‘blank’ is an empty sample which runs the whole analytical process. The measurements had been conducted during three PhD-theses (Fahlbusch 2018; Pospiech 2018; Nguyen 2019) between 2011 and 2018 in the same laboratory with the same measuring system. Accordingly, the calculated relative uncertainties describe the relative uncertainties of plant samples analysed in the laboratory over the period from 2011 to 2018 for the named measuring system. The calculation of relative uncertainties is based on the uncertainty model described in Hawkins (2014). The summary of the values of the relative uncertainties is given in Table 4.

4.2.1 Example with Trace Elements and Ca

The element concentrations of Ba, Ca, Cd, Cr, Cu, Mg, Mn, Mo, Na, Rb, Sr and Zn had been used to calculate a LDA and a QDA with and without relative uncertainties. The values of the posterior probabilities are displayed in Fig. 3 and the classifications are in Table 5. The results show that for both, LDA and QDA, the discrimination is a little poorer if the uncertainties are included into the discrimination function. This can be explained by the included elements: With Ca, Cu, Mn, Na, Rb, Sr and Zn there

Fig. 3 Discrimination analysis for trace element data plus Ca concentration data of crop plants harvested in two different fields

Table 5 Results of discrimination analysis for plant data with selected REE element concentrations

	Reinshof	Soemmerling
LDA Reinshof	207	11
LDA Soemmerling	13	159
LDA w/u Reinshof	193	25
LDA w/u Soemmerling	20	152
QDA Reinshof	210	8
QDA Soemmerling	12	160
QDA w/u Reinshof	202	16
QDA w/u Soemmerling	18	154

The table shows how many observations of which sampling field had been classified into which sampling field. If the posterior likelihood for Group 'Reinshof' is above 0.5 the class of the observation is counted as Group 'Reinshof', if it is below 0.5 the class is counted as Group 'Soemmerling'. 'w/u' refers to 'with uncertainties'

are seven elements out of 12 which have very uniform errors and apart from Rb minor uncertainties (Table 4). This shows that the discrimination without errors is already supported by well measurable elements, and the incorporation of uncertainties induces less well defined likelihoods for the samples. This results into more mis-classifications than without including the uncertainty.

4.2.2 Example with Trace Elements without Ca

The element concentrations of Ba, Cd, Cr, Cu, Mg, Mn, Mo, Na, Rb, Sr and Zn had been used to calculate a LDA and a QDA with and without relative uncertainties. The values of the posterior probabilities are displayed in Fig. 4 and the classifications are in Table 6. The results show that for both, LDA and QDA, the discrimination is a

Fig. 4 Discrimination analysis for trace element data of crop plants harvested in two different fields

Table 6 Results of discrimination analysis for plant data with selected REE element concentrations

	Reinshof	Soemmerling
LDA Reinshof	185	33
LDA Soemmerling	24	148
LDA w/u Reinshof	188	30
LDA w/u Soemmerling	19	153
QDA Reinshof	204	14
QDA Soemmerling	16	156
QDA w/u Reinshof	201	17
QDA w/u Soemmerling	18	154

The table shows how many observations of which sampling field had been classified into which sampling field. If the posterior likelihood for Group 'Reinshof' is above 0.5 the class of the observation is counted as Group 'Reinshof', if it is below 0.5 the class is counted as Group 'Soemmerling'. 'w/u' refers to 'with uncertainties'

little better if the uncertainties are included into the discrimination function. This can be explained by the included elements: With Ba, Cd, Cr, Mg and Mo there are five elements out of 11 which have very variable uncertainties (Table 4). This shows that the discrimination without errors is partly based on measurands with poor analytical certainty, and the incorporation of uncertainties reveals better variance structure of the sample groups in the data set. This results into less mis-classifications than without including the uncertainty.

4.2.3 Example with Selected REE

The element concentrations of Dy, Er, Gd, Ho, La, Nd, Sm, Tb, Tm and Yb had been used to calculate a LDA and a QDA with and without relative uncertainties. The values

Fig. 5 Discrimination analysis for selected REE element data of crop plants harvested in two different fields

Table 7 Summary of relative uncertainties (in %) per element for the used plant data set

Dy	Er	Gd	Ho	La
Min. :14	Min. :10	Min. :14	Min. : 9.6	Min. :12
1st Qu.:14	1st Qu.:10	1st Qu.:14	1st Qu.: 9.7	1st Qu.:12
Median :14	Median :10	Median :14	Median : 9.8	Median :12
Mean :14	Mean :10	Mean :14	Mean :10.6	Mean :12
3rd Qu.:14	3rd Qu.:10	3rd Qu.:14	3rd Qu.:10.2	3rd Qu.:12
Max. :14	Max. :10	Max. :14	Max. :41.0	Max. :12
Nd	Sm	Tb	Tm	Yb
Min. : 12	Min. : 14	Min. : 12	Min. : 14	Min. :16
1st Qu.: 12	1st Qu.: 15	1st Qu.: 12	1st Qu.: 15	1st Qu.:16
Median : 13	Median : 17	Median : 13	Median : 17	Median :16
Mean : 19	Mean : 69	Mean : 17	Mean : 24	Mean :16
3rd Qu.: 17	3rd Qu.: 22	3rd Qu.: 16	3rd Qu.: 23	3rd Qu.:16
Max. :145	Max. :8922	Max. :107	Max. :156	Max. :16

Values rounded to two significant digits

of the posterior probabilities are displayed in Fig. 5 and the classifications are in Table 8. The results show that for both, LDA and QDA, the discrimination notably poorer if the uncertainties are included into the discrimination function. This can be explained by the high uncertainties of the included REE (Table 7): The discrimination without errors is based on measurands with generally poor analytical certainty. There might be a bias in the data resulting from sample handling in the lab, for example offset between sample batches because samples had not been completely randomized concerning the field origin, or a bias induced by the plant mix, which is slightly different between the

Table 8 Results of discrimination analysis for plant data with selected REE element concentrations

	Reinshof	Soemmerling
LDA Reinshof	183	34
LDA Soemmerling	64	108
LDA w/u Reinshof	140	77
LDA w/u Soemmerling	56	116
QDA Reinshof	186	31
QDA Soemmerling	48	124
QDA w/u Reinshof	132	85
QDA w/u Soemmerling	50	122

The table shows how many observations of which sampling field had been classified into which sampling field. If the posterior likelihood for Group 'Reinshof' is above 0.5 the class of the observation is counted as Group 'Reinshof', if it is below 0.5 the class is counted as Group 'Soemmerling'. 'w/u' refers to 'with uncertainties'

field, etc. Applying the long-term uncertainties into the discriminant function shows clearly that REE are not a save element suite for discriminating plants between the two fields: As Table 8 shows, the mis-classification rate is considerably higher if the relative uncertainty of the measurements is included.

5 Conclusions

This contribution proposes a method to incorporate individual, cell-wise uncertainties into a DA for compositional and non-compositional data. It allows to use the method in all sciences working with relative data for example element concentration data, election data, or other relative data. The method provides a more 'honest' approach to the results of a DA, because the uncertainties help to determine how much contributes each and every component of each and every observation to the classification between groups. For sciences working with element concentrations this method is a first step towards a change of paradigm allowing to work without detection limits, but with uncertainties instead.

Incorporating the uncertainties into building the decision rule of the discrimination function has two effects: (I) The impact of certain variables on the discrimination function may change. Formerly unimportant variables may become more important and vice versa. (II) The posterior probabilities of the observations change and the respective observations may be assigned to other groups, based on their individual uncertainties.

Acknowledgements This work is part of the NEXT project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.460776804 - H2020-SC5-2017. We thank Wiebke Fahlbusch for providing the data set of her PhD-thesis for the case study with plant data and for reading. We would like to thank also our colleagues from the IZNE of the Georg-August-University of Göttingen, Benedikt Sauer and Hans Ruppert for providing additional data for the CRMs. The additional data had been generated within a Project (<http://www.tfz.bayern.de/-rohstoffpflanzen/einjaherigekulturen/091177/index.php>) of the TFZ (Technologie- und Förderzentrum im

Kompetenzzentrum für Nachwachsende Rohstoffe). We also thank Thuy Phuong Nguyen to allow us to use unpublished CRM-data from her PhD-thesis.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall Ltd., London
- Aitchison J, Pawlowsky-Glahn V (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. *Proc IAMG* 97:3–35
- BIPM, IEC and IFCC, ILAC and ISO, IUPAC, IUPAP, OIML (2008) The international vocabulary of metrology—basic and general concepts and associated terms (VIM), JCGM 200: 2008. Online
- Egozcue JJ, Pawlowsky-Glahn V (2011) Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A (eds) *Compositional data analysis: theory and applications*, vol 378. Wiley, Chichester, pp 12–27
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300 ISSN 1573-8868
- Fahlbusch W (2018) Transfer of main and trace elements from soil to plant with an emphasis on trace element supply for biogas digestion plants. Ph.D. thesis, Georg-August-Universität Göttingen
- Fahlbusch W, Hey K, Sauer B, Ruppert H (2018) Trace element delivery for biogas production enhanced by alternative energy crops: results from two-year field trials. *Energy Sustain Soc* 8(1):1–11
- Hamilton NE, Ferry M (2018) ggtern: ternary diagrams using ggplot2. *J Stat Softw Code Snippets* 87(3):1–17
- Hawkins DM (2014) A Model for Assay Precision. *Stat Biopharm Res* 6(3):263–269. <https://doi.org/10.1080/19466315.2014.899511>
- Nguyen TP (2019) Transfer of Nutrient and Harmful Elements from Soil to Rice and Health Risk Assessments for the Vietnamese Population. Ph.D. thesis, Georg-August-Universität Göttingen
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data*. Wiley, Hoboken
- Pospiech SI (2018) Geochemical Characterization of Tea Leaves (*Camellia sinensis*) and Soils for Provenance Studies based on Compositional Data Analysis. Ph.D. thesis, Georg-August-Universität Göttingen
- Pospiech S, Fahlbusch W, Sauer B, Pasold T, Ruppert H (2017) Alteration of trace element concentrations in plants by adhering particles—methods of correction. *Chemosphere* 182:501–508 ISSN 0045-6535
- Pospiech S, van den Boogaart KG (2020) Discriminant analysis incorporating individual uncertainties. R package version 0.1.2-1. <https://CRAN.R-project.org/package=vdar>
- Potts PJ (2012) A proposal for the publication of geochemical data in the scientific literature. *Geostand Geoanal Res* 36(3):225–230
- R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Stevens SS (1946) On the theory of scales of measurement. *Science* 103(2684):677–680
- van den Boogaart KG, Tolosana-Delgado R (2008) “Compositions”: a unified R package to analyze compositional data. *Comput Geosci* 34(4):320–338 ISSN 0098-3004
- van den Boogaart KG, Tolosana-Delgado R (2013) *Analyzing compositional data with R*. Springer, Heidelberg
- Venables WN, Ripley BD (2002) *Modern applied statistics with S*, 4th edn. Springer, New York
- Wickham H (2016) *ggplot2: elegant graphics for data analysis*. Springer, New York