Prioritization of Issues and Requirements by Cumulative Voting: A Compositional Data Analysis Framework

Panagiota Chatzipetrou
Lefteris Angelis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
{pchatzip, lef@csd.auth.gr}

Per Rovegård
Ericsson AB,
Karlskrona, Sweden
per.rovegard@ericsson.com

Claes Wohlin
Blekinge Institute of Technology
Karlskrona, Sweden
Claes.Wohlin@bth.se

Abstract

Cumulative Voting (CV), also known as Hundred-Point Method, is a simple and straightforward technique, used in various prioritization studies in software engineering. Multiple stakeholders (users, developers, consultants, marketing representatives or customers) are asked to prioritize issues concerning requirements, process improvements or change management in a ratio scale. The data obtained from such studies contain useful information regarding correlations of issues and trends of the respondents towards them. However, the multivariate and constrained nature of data requires particular statistical analysis. In this paper we propose a statistical framework; the multivariate Compositional Data Analysis (CoDA) for analyzing data obtained from CV prioritization studies. Certain methodologies for studying the correlation structure of variables are applied to a dataset concerning impact analysis issues prioritized by software professionals under different perspectives. These involve filling of zeros, transformation using the geometric mean, principle component analysis on the transformed variables and graphical representation by biplots and ternary plots.

1. Introduction

Prioritization is a procedure of principal importance in decision making. It is encountered in cases where multiple choices have to be considered in order to take a decision regarding a product or a managerial strategy. Software engineering is a wide field where prioritizations of process improvement issues, stakeholders, software requirements, etc play significant role. For a comprehensive account of prioritization and its importance in software product management we refer to [1]. There are two main problems in the application of prioritization in practice: First, the large number of choices makes the procedure of prioritization complicated. Second, the human subjectivity is a source of variation when different people try to prioritize independently a certain amount of issues. These aspects led to the adoption of voting schemes where stakeholders express their relative preferences to certain issues or requirements in a systematic and controlled manner.

The Cumulative Voting (CV) or 100-Point Method or Hundred-Dollar ($100) test, described by Leffingwell and Widrig [2], is a simple, straightforward and intuitively appealing voting scheme where each stakeholder is given a constant amount (e.g. 100, 1000 or 10000) of imaginary units (for example monetary) that he or she can use for voting in favor of the most important issues. In this way, the amount of money assigned to an issue represents the respondent's relative preference (and therefore prioritization) in relation to the other issues. The points can be distributed in any way that the stakeholder desires. Each stakeholder is free to put the whole amount given to him or her on only one issue of dominating importance. It is also possible for a stakeholder to distribute equally the amount to many of, or even to all of the issues.

CV is sometimes known as "proportional voting" since the amount of units assigned to an issue represents the relative priority of the specific issue in relation to the other issues. The term "proportional" in this case also reflects the fact that if the amount of units assigned to an issue is divided by the constant number of units available to each stakeholder, the result becomes a proportion between zero and one. The stakeholder's ratings for a set of issues can be therefore considered as the “composition” or “mixture” of a person’s opinion towards the issues, in the abstract sense that each issue occupies a certain proportion (or percentage) of preference inside the person’s belief or judgment.

The procedure may result to issues that are assigned zero units showing that the specific stakeholder considers these issues completely unimportant. The zeros are generally a problem in this kind of data, because they make the notion of relative preference or importance completely meaningless and the computation of ratios impossible. Of course, a questionnaire where zeros are not allowed could be designed, but in general, the principle of CV is to allow stakeholders to spread freely their total amount without further restrictions.

In empirical studies aiming to investigate the different views of stakeholders towards prioritization of issues or requirements, the hundred dollar test is given to a sample of people, the results are coded as variables and they are statistically analyzed in order to find differences or agreements in views and maybe correlations with other variables coming from the questionnaire (attitudes, opinions etc). The data gathered in such studies are affected by various sources of variation and are therefore subject to large variability. The statistical analysis of such data can reveal significant differences, trends, disagreements and groupings between the respondents and can constitute a valuable aid for understanding the attitudes and opinions of the interviewed persons and therefore a tool for decision-making.

In statistical terms, each person in the sample produces a vector of numerical values with fixed sum, i.e. the total amount of imaginary units. As we already mentioned, by a simple division each data vector is transformed to a compositional vector of proportions summing up to one. It is obvious that the random variables corresponding to such proportion vectors are inherently correlated and bounded since their sum is one. For this reason, the classic parametric statistical tests assuming that the sample comes from a multivariate normal distribution are not valid for the original data. It is obvious therefore that special transformation of the data and statistical techniques should be used, specifically designed for proportions.

In this paper we suggest the use of a statistical framework, suitable for the analysis of proportions, known as Compositional Data Analysis (CoDA). This methodology has been widely used in the analysis of materials composition in various scientific fields like chemistry, geology and archaeology but its principles fit to the data obtained by CV. We present first the general principles of CoDA and then we apply certain methods to CV prioritization data from a previous empirical study on impact analysis. Specifically, we discuss the problem of zeros and imputation methods for handling them as missing values, the transformation of the original values dividing them by their geometric mean, the application of principal components analysis on the transformed data and graphical methods such as the biplot and the ternary plot for representing the variability and the correlation of data.

The paper is structured as follows: Section II provides an outline of the related work. Section III presents the basic principles of CoDA and discusses various problems related to its application. Section IV presents the results from the application of CoDA on the dataset, which are discussed in Section V. Finally, in Section VI we summarize providing conclusions and directions for future work.

2. Related work

The method of cumulative voting is described in [3] as the procedure of voting for a company's directors and boards (see also [4] and [5]). Specifically, CV has been proposed as an alternative voting mechanism, suitable for enhancing the opinion and voice of minorities in majority-dominated companies [6]. In politics, the method has been applied for election of councils since 1854 [7].

In Software Engineering, CV is known as a prioritization technique, used in decision making in various areas, such as requirements engineering, impact analysis or process improvement ([8], [9], [10], [1]). Prioritization is performed by stakeholders (users, developers, consultants, marketing representatives or customers), under different perspectives or positions, who respond in questionnaires appropriately designed. CV has been proposed as an alternative to the Analytical Hierarchy Process (AHP) and its use is continuously expanding to areas such as requirements prioritization and prioritization of process improvements [2], [11].

In [8], CV is used in an industrial case study where a distributed prioritization process is proposed, observed and evaluated. The stakeholders prioritized 58 requirements with $100,000 to distribute among the requirements (the large amount of “money” was chosen to cope with the large number of requirements). In [12] the hundred dollar test is discussed in the framework of requirements triage. In [13] the CV is considered as one of the prioritization approaches among others in a research framework for studies about requirements prioritization. In [14] the $100-test was used in the design of the New Jersey Homeland Security Technology Systems Center (NJHSTSC) web site. Specifically, it was used to prioritize the aspects of the webpage development. In [15] the CV was used for an industrial case study on the choice between language customization mechanisms. In [16] CV is one of the four prioritization methods examined, evaluated and recommended for certain stages of a software project.

Although CV is a simple method for taking aggregated results, it has not been thoroughly studied as a method for collecting data. Indeed, only few papers have considered a statistical analysis of these data. In [8] novel approaches to visualize the priority distribution among stakeholders are presented, together with measures on disagreement and satisfaction. In [17] the data used in the present paper, collected through a CV procedure for impact analysis, were analyzed using the nonparametric method Kruskal-Wallis, alternative to ANOVA. In [18] the same data were further analyzed by multivariate methods, especially using distances and measures of agreement, taking into account the proportional nature of data.

In the present paper we use the same data as in [17 and 18]. However, our goal is not just to analyze the same data with different technique, but to introduce a comprehensive framework of statistical tools that can deal with all kinds of data collected through a CV questionnaire. The methods we apply focus on the multidimensional nature of data and aim to study their correlation structure. We note that the same techniques can be applied to all quantitative data from other variations of CV, for example the Hierarchical Cumulative Voting [19].

3. The statistical framework

3.1. Compositional data analysis

Compositional Data Analysis (CoDA) is a multivariate statistical analysis framework for vectors of variables having a certain dependence structure: The values in each vector have sum equal to a constant. Usually, for easy reference to the same problem, after division by that constant, the sum of the values of each vector becomes one. Thus from now on, we can assume that our data set consists of vectors of proportions or percentages in the form:

$$(p_1,...,p_k) \text{ where } p_i \geq 0 \text{ , } \sum_{i=1}^{k} p_i = 1 \qquad (1)$$

The important point here is to understand that the data are constrained and every sample comes from a sample space which is the unit simplex. Therefore, the techniques applied to samples from the real Euclidean space are not applied in a straightforward manner.

In fact, there are various problems associated with the analysis of those vectors: First, there is a problem of interdependence of the proportions (since their sum is 1) and therefore they cannot be treated as independent variables (the usual assumption of the multivariate methods). Second, their values are restricted in the [0,1] interval, so the normality assumptions are invalid. Third, we are not really interested in absolute values here, but rather for relative values (that is actually the meaning of a proportion). So the whole problem is transferred to the analysis and the interpretation of the ratios of the proportions, i.e. values of the form p_i / p_j. The statistical analysis of these data, using methods based on ratios, tries to provide answers to some research problems which we encounter in any multivariate statistical analysis.

Concerning now the prioritization questionnaires using the 100$ (or the 1000$) test, the data are essentially representing proportions of the overall importance allocated to each of the issues examined in a study. The relative importance of the issues is represented by their ratios, so CoDA seems the appropriate framework for their study. Historically, Karl Pearson in 1897 [20] posed the problem of interpreting correlations of proportions while the milestone for this type of statistical analysis is the pioneer work of John Aitchison [21], [22]. A freeware package for compositional data analysis is the CoDaPack3D [23] which we use in our analysis.

3.2. The problem of zeros

The variables which form the constrained vectors are the issues or the requirements in our context while their values satisfying (1) are the proportions of priorities or simply the priorities. The data from the CV questionnaires have some special characteristics which cause problems in the analysis.

The problem of zeros is of principal importance. When the number of issues is large and the individuals are only few, the data matrix is usually sparse with a large number of zeros. This structure causes problems of interpretation when we consider the relative importance. Another problem associated with zeros, is to determine their actual meaning. According to the general theory of CoDA applied to composition of materials, the zeros can be *essential* (i.e. complete absence of a component) or *rounded* (i.e. the instrument used for the measurements cannot detect the component). However, in our context, the importance of an issue is a completely abstract notion. The instrument of its measurement is the human judgment so we can assume that a low prioritization is actually measured by a very low value in the 100$ scale, which is rounded to zero. Besides, it is known that humans tend to allocate "rounded" numbers in such tests. Otherwise, we should assume the absence of importance, which is more or less irrational since at least one of the respondents considered it as important and allocated a nonzero account. Of course, when all respondents allocate zero to an issue, this can be excluded from the analysis and considered as essential zero.

Due to the problems of zeros, the various ratios needed for the analysis are impossible to be computed. It is therefore essential and necessary to find first a way of dealing with the zeros. In [24] the method of zero replacement in economic data is considered the most appropriate among other techniques for handling it. The problem of rounded zeros was addressed by [22] as an imputation problem applied to missing data. A simple solution introduced in [22], was the *additive replacement strategy*. In [25] a new simple method is proposed that is most stable regarding the choice of the imputed values. This is called *multiplicative replacement strategy* and according to it, every vector $\mathbf{p} = (p_1, ..., p_k)$ with the properties in (1), having c zeros, can be replaced by a vector $\mathbf{r} = (r_1, ..., r_k)$ where

$$r_j = \delta_j \ \ (\text{if } p_j = 0) \text{ or}$$
$$r_j = p_j \left(1 - \sum_{l:p_l=0} \delta_l \right) \ \ (\text{if } p_j > 0) \tag{2}$$

where δ_j is a (small) imputed value for p_j. The advantages of multiplicative replacement are discussed extensively in [25], [26] and [27]. Other, more complicated methods, such as a modified EM algorithm, have been also proposed [28]. However, in our context the simplicity of the multiplicative replacement, the very good results and the lack of any theoretical assumptions make the method most preferable than any other. Note that this method is implemented in CoDaPack3D.

3.3. The clr transformation

Aitchison [22] proposed the *centered logratio* (*clr*) *transformation* for transforming the raw proportional dataset to the real space and at the same time for retaining their correlation structure. The transformation is simple and achieved after dividing each component of a vector of proportions by their geometric mean. Since in our context the data usually contain zeros, the transformation can be applied only after the replacement of zeros by a method like (2). The formula for the clr transformation of the compositional vector $\mathbf{r}$ is:

$$\mathbf{y} = clr(\mathbf{r}) = \left[\log \frac{\mathbf{r}}{g(\mathbf{r})} \right] \tag{3}$$

where the geometric mean is given by:

$$g(\mathbf{r}) = \left(\prod_{i=1}^{k} r_i \right)^{1/k} \tag{4}$$

The clr transformation has been proposed ([22] and [29]) for producing data that can be subsequently used in multivariate methods like *Principal Components Analysis* (*PCA*) [30]. PCA produces uncorrelated linear combinations of the original variables. The new variables (or *components*) account for decreasing amounts of the total variation (i.e. the first component explains the maximum variance, and so on) and their estimations can be used for variable reduction and representation of the data points in lower dimensions.

3.4. The biplot

The *biplot* [31], [32] is a graphical tool that has been used in various applications. Its compositional data version is a straightforward and useful tool for exploring trends and peculiarities in data. A biplot is a graphical display of the rows and columns of a rectangular $n \times k$ data matrix $\mathbf{X}$, where the rows represent individuals and the columns stand for variables. The biplot analysis is applied after performing a transformation on the elements of $\mathbf{X}$, depending on the nature of the data, in order to obtain a transformed matrix $\mathbf{Z}$ which is the matrix that is actually displayed. For compositional data, biplots are derived from the covariance matrix of centered log-ratios of proportions (clr transformation). More specifically, for an $n \times k$ compositional data matrix $\mathbf{X}$ the biplot is based on a singular value decomposition of the doubly centered logratio matrix $\mathbf{Z} = [z_{ij}]$, where:

$$z_{ij} = \log\left\{ \frac{x_{ij}}{g(\mathbf{x}_i)} \right\} - \frac{1}{n} \sum_{i=1}^{n} \log\left\{ \frac{x_{ij}}{g(\mathbf{x}_i)} \right\} \tag{5}$$

An example of a biplot is given in Figure 1. The basic characteristics of a biplot are *lines* (or *rays*) and *dots*. Rays are used to represent the variables of the dataset (in the example there are 3 variables labelled *A*, *B*, and *C*), and dots are used to represent the individuals (in the example there are 4 observations numbered from 1 to 4). Every ray starts from an origin O which represents the centre of the compositional data set. Another characteristic is the angle between the rays.

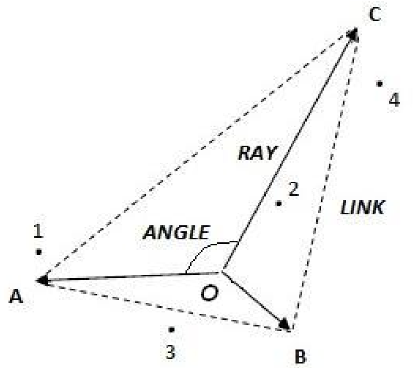

Figure 1. The basic characteristics of a biplot

The characteristics of a biplot have certain interpretation [33]: the length of a ray is an expression of the variance of the corresponding variable. Specifically, the square of the length of a ray (distance from *O*) is proportional to the variance of the clr transformation of the corresponding variable. Longer rays depict high variance. In our example, we can infer from Figure 1 that the variable corresponding to ray *C* has by far the highest variance among the variables in the biplot, while the variable corresponding to *B* has the lowest.

A *link* is a line connecting the ends of two rays. For two variables $\mathbf{x}_i$ and $\mathbf{x}_j$ the squared length of the link is an estimation of $\mathrm{var}\{\log(\mathbf{x}_i / \mathbf{x}_j)\}$, showing the difference between the two variables. Large links show large proportional variation. In the example of Figure 1, the lengths of the links show that the log ratio having the largest variation is *A/C*.

It is important to emphasize that in the context of compositional data, the most important characteristics of a biplot are considered the links and not the rays since the variables are examined in a proportional and relative manner.

The cosine of the angle between the rays is an expression of the correlation between the variables they represent (more precisely between their clr transformations). The closer the angle is to 90, or 270 degrees, the smaller the correlation. An angle of 0 or 180 degrees reflects a correlation of 1 or −1, respectively. The position of the dots with respect to the links indicates the relative values of variables. In the example of Figure 1, the dot representing individual 1 has far less *B* and *C* compared to *A*.

3.5. The Ternary Plot

Another way to illustrate compositional data is by using ternary plots [22], [34]. Ternary plots are used to represent the distribution of data with respect to only three different variables. An example of a ternary plot is presented in Figure 2. The three vertices of the equilateral triangle with unit altitude represent the variables *A*, *B* and *C*. The dots represent the individuals while their placement inside the triangle depends on the values of *A*, *B* and *C* for each individual. Note that these values are rescaled to sum up to 1. Specifically, the distance from a dot to the side opposite a vertex represents the value of the individual for the specific variable. So when a dot is near *A*, the distance from the opposite side is large and therefore has a large value for *A*. When the number of variables is more than three, the ternary plot allows us to focus on certain triples of variables which we consider important. We can also combine original variables, by taking their sum for example, in order to form sub-compositions and present aggregated percentages.

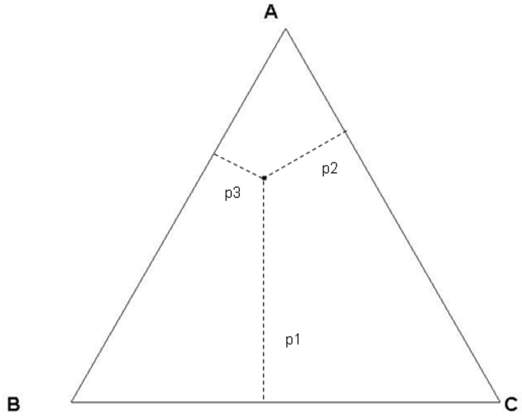

Figure 2. The basic characteristics of a ternary plot

4. Application to data

4.1. Description of the Dataset

The data used in this paper were collected during an empirical study on the role of Impact Analysis (IA) in the change management process at Ericsson AB in Sweden [18]. In the study, impact analysis issues were prioritized with respect to criticality by software professionals from an *organizational perspective* and an *individual* (or *self*) *perspective*. The software professionals belonged to three organizational levels: *operative*, *tactical* and *strategic*. In order to understand how issues associated with IA are seen at the different levels and under the different perspectives, 18 employees were interviewed at the company in their roles as industrial experts.

Table 1. The issues used in prioritization

id	Issue
i1	It is difficult to find resources for performing impact analysis.
i2	There is not enough time to perform impact analysis.
i3	System impact is underestimated or overlooked.
i4	Change requests are unclear.
i5	Responsibility and product/project balance are difficult to handle for analyses that span several systems.
i6	Analyses are incomplete or delayed.
i7	Analyses require much expertise and experience.
i8	Analyses are too coarse or uncertain.
i9	It is difficult to handle conflicting and synergetic change requests.
i10	Analyses are not prevented from being disregarded.
i11	Existing traceability is manual and cumbersome.
i12	It is difficult to see trends and statistics for collective impact.
i13	Tools to support the analysis are missing.
i14	Affected parties are overlooked.
i15	Analyses are performed by the wrong persons.
i16	Change request decisions are based on interest.
i17	Requirements and baseline are missing for early change requests.
i18	Analyses and change implementation evoke stress.
i19	It is not possible to see the outcome of a change request.
i20	It is difficult to see status and updates for a change request.
i21	Different change request have different levels of complexity, and there is no single method for handling all levels.
i22	Cheap, short-term solutions win over good, long-term solutions.
i23	Solutions are specified with too much detail by high-level analysis.
i24	Hardware and protocol dependencies are difficult to handle for late change requests.
i25	Relevant structure and documentation to support the analysis are missing.

The resulting list of issues was subsequently subjected to prioritization by the interviewees. The interviewees were asked to prioritize 25 issues (Table 1) using CV by distributing 1000 imaginary points to the issues. Each interviewee prioritized the 25 issues twice: Under the organizational perspective and under the self-perspective. There were 8 interviewees belonging in the operative, 5 in the strategic and 5 in the tactical level. The goal was to investigate whether people see IA differently depending on their level and perspective, and whether these groupings are meaningful in process improvement efforts. In the present study we consider only the different perspective and not the level. Specifically, we are interested in studying the data in a multivariate framework in order to investigate their correlation structure under each perspective.

Obviously, as we already mentioned, the data obtained from the interviews after division by 1000 consist of vectors with proportions as elements. For each participant $i = \{1, \ldots, 18\}$ there are two corresponding vectors:

$$\mathbf{p}_i^{(j)} = \left(p_{i,1}^{(j)} \quad p_{i,2}^{(j)} \quad \cdots \quad p_{i,25}^{(j)}\right), \; j = O, I \qquad (6)$$

where the superscript (*j*) is either (*O*) for the organizational perspective and (*I*) for the individual. So, we have essentially two datasets with proportional data.

4.2. Transformations and principal component analysis

The two datasets of proportions (6) were analyzed separately with the methods of CoDA described in the previous section. First the zeros were replaced by the multiplicative replacement technique and then the clr transformation was applied to each dataset. The variables obtained were analyzed by principle component analysis (PCA) with varimax rotation of the axes [30].

Table 2. Principal components for the organizational data

number of component	Variance explained (%)	Issues correlated with the component
1	12.94	i2(-), i1(-), i14(+), i20(+)
2	12.04	i8(-), i4(-), i16(+), i13 (+), i22 (-)
3	11.50	i21(+), i17(+), i11(-)
4	10.83	i23(-), i9(-), i5(+)
5	9.60	i24(+), i18(+)
6	9.00	i15(-),i6(+), i10(+)
7	8.10	i25(+), i3(-)
8	7.35	i7(+), i12 (-)
9	7.00	i19(+)

The application of PCA to organizational data revealed the existence of nine principle components explaining 88.35% of the total variation. Each principal component extracted is highly correlated with a number of issues either positively or negatively. The nine components in descending order of importance (% of the variance they explain) together with the issues correlated with them are given in Table 2. The sign (+) or (-) following each issue shows a positive or negative correlation. Similarly, the PCA applied to the individual data produced nine principal components explaining 87.28% of the total variance. The components and the issues correlated with them are shown in Table 3.

Table 3. Principal components for the individual data

number of component	Variance explained (%)	Issues correlated with the component
1	12.94	i5(+), i9(+), i20(+)
2	11.28	i17(+), i1(-), i2(-), i4(+)
3	10.97	i14(-), i10(-), i16(+), i25(+)
4	10.32	i23(+), i24(+), i3(-), i6(-)
5	10.08	i22(-), i13(+)
6	9.88	i19(+), i18(+)
7	8.12	i7(-), i21(+)
8	7.46	i12(+), i11(-), i8(+)
9	6.24	i15(-)

4.3. Biplots

The biplots for organizational and individual perspectives with the respondents as dots are shown in Figures 3 and 4 respectively. The rays representing the 25 issues have been labelled as i1,….., i25.

As we can see from Figure 3, issues i10, i13, i14, and i15 have the longest rays in the biplot, which indicates that they have by far the highest variance among the issues. Moreover, the longest link in Figure 3 is the one between i15 and i13. This indicates that those log ratios have the greatest variation. In contract, the short link between the ends of i7 and i17 or between i19 and i21, indicates that the ratios of those two pairs vary slightly across the set of the samples.

Furthermore, the rays corresponding i15 and i14, are nearly at right angles. This indicates that the correlation between these log ratios is close to zero. A strong positive correlation is shown between i18 and i24, while the correlation between i6 and i22 is negative. Finally, the proximity of dots to links helps to interpret the characteristics of outliers. The respondent #2 represented by the dot at the top right of the plot has high proportions of i15 relative to i6 and i10, respondent #18, at the left bottom of the plot, has high proportions of i14 relative to i22 and i13 while respondent #5, at the bottom of the plot has high proportions of i10 and i6 relative to i15 and i22.

Similarly, the biplot for individual perspective with the respondents is shown in Figure 4. As we can see, issues i1, i2, i4, i14 and i24 have the longest rays, indicating that they have the highest variance among the issues. The longest link is the one between the pairs of rays i14 and i24. This indicates that those log ratios have the greatest variation. In contrast, the short link between the ends of i1 and i2 indicates that those ratios vary slightly across the set of the samples.

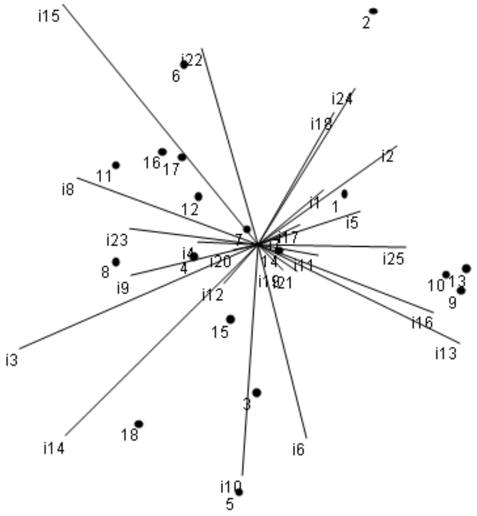

Figure 3. The biplot with the respondents for the organizational data

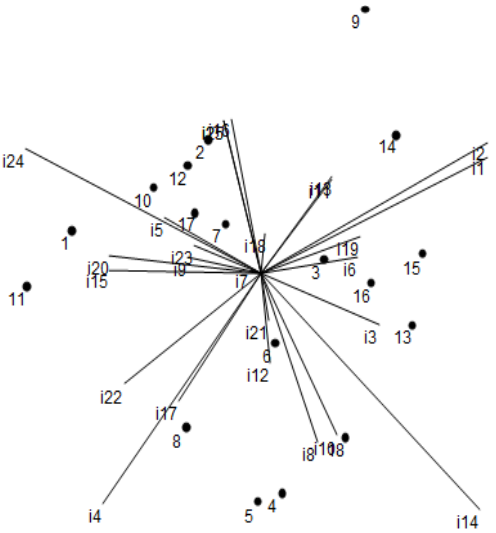

Figure 4. The biplot with the respondents for the individual data

Furthermore, the rays linking i1 and i25, are almost at right angles. This indicates that the correlation between these log ratios is close to zero. A strong positive correlation is represented between i1 and i2, while the correlation between i19 and i22 is negative. From the proximity of the dots to the links we can see the outlier respondent #9 at the top right of the plot having high proportions of i25 relative to i1 and i2, respondent #8, at the left bottom of the plot, has high proportions of i4 relative to i2 and i24 while respondent #16, at the right bottom of the plot has high proportions of i3 and i14 relative to i4 and i24.

4.4. Ternary plots

The ternary plots are used to show the distribution of the data with respect to a triple of issues. The decision on the importance of the triple is based on various criteria. A possible criterion is the high correlation between variables. Another possibility is to add the percentages of some variables in order to see how the data points are distributed with respect to those aggregations which remain percentages.

For our data sets, in Figures 5 and 6 (made by MATLAB functions, available at [35] and [36]) we give the ternary plots for organizational and individual data set respectively with respect to three variables "LOW", "MEDIUM" and "HIGH". Each of these variables was obtained by adding certain issues. Specifically, "LOW" in each dataset was derived by adding the issues that received on average less than 25/1000 points. Also, "MEDIUM" is the sum of issues that received on average 25/1000 to 40/1000 points, while 'HIGH" is the sum of issues that received on average more than 40/1000 points. These thresholds were defined arbitrarily.

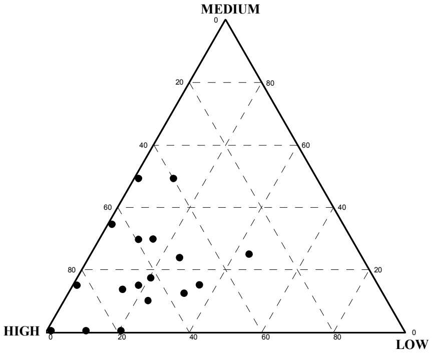

Figure 5. The ternary plot for the organizational data

In Figure 5 we can see the points corresponding to the organizational data, while in Figure 6 to the individual data. It is clear that for both datasets, the issues that were prioritized on average highly (i1, i2, i3, i6, i14, i15, i16, i22, i24 for the organizational and i1, i2, i3, i4, i6, i14, i15, i22, i24, i25 for the individual data) sum up to large values. This is depicted by the concentration of points (respondents) near the "HIGH" vertex, which is denser for the individual data. This particular distribution shows that the issues that received on average either low or medium prioritization do not manage to accumulate considerable amounts.

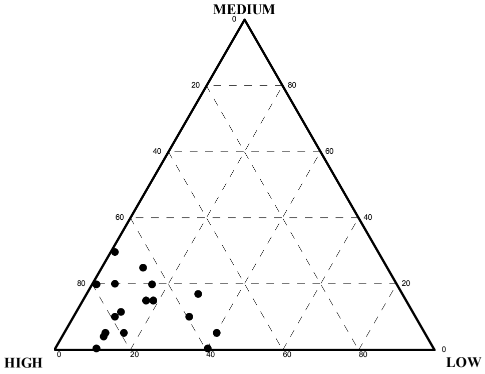

Figure 6. The ternary plot for the individual data

5. Discussion

The application of certain CoDA methods to the specific dataset in the context of software engineering showed that CoDA can form the basis for a complete analysis of prioritization data. Data from CV can come from various studies on requirements or any other issues in software engineering needing prioritization. The CoDA methods of zero replacement and clr transformation can be used as a preprocess stage in order to prepare the data for further statistical analysis.

The subsequent analysis of the transformed data by multivariate methods as the PCA and the biplots gives useful insight of the correlation structure of data. Specifically, it is interesting to find groupings of the prioritized issues which present significant correlation and examine how these groupings differ under different views. For example, in our data, issues i1 and i2 are highly correlated and although they are grouped together under both perspectives, the other issues of the groupings are different (i14, i20 under the organizational and i4, i17 under the individual perspective).

The correlations and groupings can be very useful for managerial decisions regarding improvements that address the various issues. In general, it is desirable to select improvements that are able to address as many issues as possible simultaneously [18]. However, since the cost of multi-issue improvements is large, the groupings resulting from statistical methods can form the basis for mapping targeted improvements to groups of highly correlated issues. Furthermore, the inclusion of risk and effort as measured variables in a dataset can help the subsequent analysis and the decision making process to take into account these characteristics too.

Interesting groupings of the respondents and various outliers can also be discovered by the CoDA plots. The relative placement of respondents either in the biplots or in the ternary plots shows their attitudes towards the issues examined. The biplots give the general picture while the ternary plots can help to focus on triples of combinations or aggregations of issues that are of special importance. It is important to realize that both plots offer visualization means of datasets with high dimensionality and help us to explore and understand the data. These plots can be also useful for managerial decision since they depict the trend of the respondents to "concentrate" on certain groups of issues or requirements.

In general, we believe that the proposed analysis can be useful in software engineering when prioritization data from CV are used for managerial decisions regarding multi issue improvements. The main advantage of CoDA framework is the analysis of the prioritized issues or requirements as a whole, focusing on the interdependencies between the issues and the trends of the respondents. It should be noted that the methods we applied here are only a part of the CoDA theory which involves advanced and complicated methods. The adaptation of other CoDA methods to prioritization in software engineering is a direction for future research.

Regarding the limitations and the drawbacks of the method, it is clear that any statistical method is appropriate for specific types of data. CoDA is suitable for data collected from the CV procedure, wherever this can be applied. Obviously, in some software development processes (for example the agile methodology) CV collection and CoDA analysis are meaningless. Another arguable point is that CoDA is too complicated for such a simple tool of prioritization like CV. However, if we go a step beyond the fact that CV offers a straightforward way of aggregating the prioritization of several stakeholders, CV produces a multidimensional numerical dataset of special type. The analysis aiming to discover significant information on correlation, agreement, trends, groupings and outliers from such data cannot be simplistic and requires advanced statistical methods with strong mathematical background.

6. Conclusions

Compositional data analysis (CoDA), as introduced by Aitchison [21] and evolved by several researchers from various disciplines, is a statistically sound and comprehensive set of methodologies which can set a framework for analyzing prioritization data from cumulative voting (CV) in software engineering. The methodologies discussed and applied in this paper can be applied in any dataset resulting from CV or related procedures in order to study in a multidimensional manner the correlation structure of the data, groupings of variables and respondents and also to detect outliers. The knowledge obtained from such a study can be used for managerial decisions in software product management.

Future research is necessary for the exploration of potentials and the full adaptation of CoDA methods in software engineering prioritization studies and involves application of other transformations, graphical and statistical methods from the general CoDA methodology, experimentation with more datasets and finally meaningful interpretation of the results for the enhancement of the statistical inference and decision making.

7. References

[1] Berander P., *Evolving prioritization for software product management*, Phd Thesis Department of Systems and Software Engineering School of Engineering Blekinge Institute of Technology, Sweden, 2007.

[2] Leffingwell, D. and D. Widrig, *Managing software requirements: A Use Case Approach*, 2nd ed. Addison-Wesley, Boston, 2003.

[3] http://www.investopedia.com/terms/c/cumulativevoting.asp

[4] G. J. Glasser., "Game theory and cumulative voting for corporate directors", *Management Science*, Vol. 5, No. 2, Jan., 1959, pp. 151-156.

[5] G. F. Davis, and E. Han Kim, "Business ties and proxy voting by mutual funds", *Journal of Financial Economics* 85, 2007, 552–570.

[6] C. Xi, "Institutional shareholder activism in China: Law and practice (Part 1)", 17 *International Company and Commercial Law Review*, 2006, 251-262.

[7] S. Bowler, T. Donovan, and D. M. Farrell, "Party strategy and voter organization under Cumulative Voting in Victorian England political studies", *Political studies*, 1999, 47 (5), 906–917.

[8] B. Regnell, M. Host, J. Natt och Dag, P. Beremark, and T. Hjelm, "An industrial case study on distributed prioritisation in market-driven requirements engineering for packaged software", *Requirements Eng*, 2001, 6:51–62.

[9] P. Berander, and C. Wohlin, "Difference in views between development roles in software process improvement – A quantitative comparison", *Empirical Assessment in Software Engineering* (EASE 2004).

[10] Karlsson L., *Requirements prioritisation and retrospective analysis for release planning process improvement*, PhD Thesis, Department of Communication Systems Lund Institute of Technology, 2006.

[11] D. Firesmith, "Prioritizing requirements", *Journal Of Object Technology*, Vol. 3, No.8, September-October 2004.

[12] A. M. Davis, "The art of requirements triage", *Computer, Published by the IEEE Computer Society,* vol 36, no3, p.42-49, March 2003.

[13] P. Berander, K. A. Khan, and L. Lehtola, “Towards a research framework on requirements prioritization”, *SERPS’06*, October 18–19, 2006, Umeå, Sweden.

[14] O. Eljabiri, R. Tawfik, S. Silva, M. Ahmed, S. Vivianni, R. Lanche and R. Chien, “NJHSTSC. NJIT-HOMELAND SECURITY, CIS 491/101”, 2005. http://eljabiri1.tripod.com/sitebuildercontent/sitebuilderfiles/ NewSample2.pdf

[15] M. Staron, and C. Wohlin, “An industrial case study on the choice between language customization mechanisms”, J. Münch, and M. Vierimaa (Eds.): *PROFES 2006, LNCS 4034*, pp. 177 – 191, 2006. Springer-Verlag Berlin Heidelberg, 2006.

[16] S. Hatton, “Choosing the “right” prioritisation method”, *19th Australian Conference on Software Engineering*.

[17] P. Jönsson, and C. Wohlin, “Understanding impact analysis: An empirical study to capture knowledge on different organisational levels”, *International Conference on Software Engineering and Knowledge Engineering (SEKE05)*, pp. 707-712, July 2005, Taipei, Taiwan.

[18] P. Rovegard, L. Angelis, and C. Wohlin, “An empirical study on views of importance of change impact analysis issues”, *IEEE Transactions On Software Engineering*, Volume 34, Issue 4, 2008, pp. 516-530.

[19] P. Berander, and P. Jonsson, “Hierarchical Cumulative Voting (HCV) – Proritization of requirements in hirarchies”, *International Journal of Software Engineering and Knowledge Engineering*, Vol. 16, No 6 ,2006, pp. 819-849.

[20] K. Pearson, “Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs”, *Proc. Roy. Soc. 60*, 1897, pp.489-98.

[21] J. Aitchison, “The statistical analysis of compositional data (with discussion)”, *J. R. Statist. Soc*. B, v.44,1982, pp. 139-177.

[22] Aitchison, J. *The statistical analysis of compositional data*, The Blackburn Press, London, 2003.

[23] S. Thió-Henestrosa, and V. Pawlowsky-Glahn, “Compositional data package user's guide”,2008, http://ima.udg.edu/CoDaPack

[24] J.M. Fry, T.R.L. Fry, and K.R. McLaren, “Compositional data analysis and zeros in micro data”, *Appl. Economics*, 32, 2000, pp. 953 – 959.

[25] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, “Zero replacement in compositional data sets”, In H. Kiers, J. Rasson, P. Groenen and M. Shader (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000)*, pp. 155–160. Berlin, Springer-Verlag.

[26] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, “Dealing with zeros and missing values in compositional data sets using non-parametric imputation”, *Mathematical Geology*, 35(3), 2003, pp.253–278.

[27] J.A. Martín-Fernández, J. Palarea-Albaladejo, and J. Gómez-García, “Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach”, In S. Thió-Henestrosa and J.A. Martín-Fernández (Eds.), *Proceedings of CODAWORK'03 - Compositional Data Analysis Workshop*,2003,ISBN 84-8458-111-X. Girona.

[28] J. Palarea-Albaladejo, J. A. Martín-Fernández, and J. Gómez-García, “A parametric approach for dealing with compositional rounded zeros”, *Mathematical Geology*, 2007, 39: 625–645.

[29] D. Howel,“Multivariate data analysis of pollutant profiles: PCB levels across Europe”, *Chemosphere 67*, 2007 1300-1307.

[30] Bartholomew, D.J., F., Steele , I., Moustaki, and J.I., Galbraith, *The analysis and interpretation of multivariate data for social scientists*, Chapman & Hall/CRC, 2002.

[31] K. R., Gabriel, “The biplot-graphic display of matrices with application to principal component analysis”, *Biometrika* 58,1971, pp. 453-467.

[32] K. R., Gabriel, “Biplot display of multivariate matrices for inspection of data and diagnosis”, In: V. Barnett, Ed., *Interpreting Multivariate Data,* Wiley, New York, 1981, 147-173.

[33] J. Aitchison, and K.W. Ng., “Conditional compositional biplots: theory and application”, *2nd Compositional Data Analysis Workshop CoDaWork'05*, 2005, http://ima.udg.edu/Activitats/CoDaWork05/CD/Session1/Aitchison-Ng.pdf

[34] S. Thió-Henestrosa, and J. A. Martín-Fernández, “Dealing with compositional data: The freeware CoDaPack”,*Mathematical Geology*, Vol. 37, No. 7, October 2005, DOI: 10.1007/s11004-005-7379-

[35] http://www.mathworks.com/matlabcentral/fileexchange/2299-ternplot

[36] http://www.mathworks.com/matlabcentral/fileexchange/7210