Unveiling and Unraveling Aggregation and Dispersion Fallacies in Group MCDM

Mohammadi, Majid; Tamburri, Damian A.; Rezaei, Jafar

DOI
10.1007/s10726-023-09822-4

Publication date
2023

Document Version
Final published version

Published in
Group Decision and Negotiation

Group Decision and Negotiation
https://doi.org/10.1007/s10726-023-09822-4

Unveiling and Unraveling Aggregation and Dispersion Fallacies in Group MCDM

Majid Mohammadi[1] · **Damian A. Tamburri**[2] · **Jafar Rezaei**[3]

Accepted: 4 March 2023

Abstract

Priorities in multi-criteria decision-making (MCDM) convey the relevance preference of one criterion over another, which is usually reflected by imposing the non-negativity and unit-sum constraints. The processing of such priorities is different than other unconstrained data, but this point is often neglected by researchers, which results in fallacious statistical analysis. This article studies three prevalent fallacies in group MCDM along with solutions based on compositional data analysis to avoid misusing statistical operations. First, we use a compositional approach to aggregate the priorities of a group of DMs and show that the outcome of the compositional analysis is identical to the normalized geometric mean, meaning that the arithmetic mean should be avoided. Furthermore, a new aggregation method is developed, which is a robust surrogate for the geometric mean. We also discuss the errors in computing measures of dispersion, including standard deviation and distance functions. Discussing the fallacies in computing the standard deviation, we provide a probabilistic criteria ranking by developing proper Bayesian tests, where we calculate the extent to which a criterion is more important than another. Finally, we explain the errors in computing the distance between priorities, and a clustering algorithm is specially tailored based on proper distance metrics.

Keywords Group decisions and negotiations · Multi-criteria decision-making (MCDM) · Priorities aggregation · Clustering · Compositional data

✉ Majid Mohammadi
m.mohammadi@vu.nl

Damian A. Tamburri
d.tamburi@tue.nl

Jafar Rezaei
j.rezaei@tudelft.nl

[1] Computer Science Department, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] Jheronimus Academy of Data Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[3] Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

Published online: 17 April 2023

1 Introduction

Multi-criteria decision-making (MCDM) problems typically involve evaluating a set of alternatives with respect to a handful of criteria based on the preferences of one or a group of decision-makers (DMs), with the ultimate goal of selecting, sorting, or ranking available alternatives. For such evaluation, the performance of alternatives for each criterion is acquired by employing a crucial data collection approach, whose results are stored in a so-called *performance matrix*. There are several methods to elicit the preferences of DMs, including but not limited to Tradeoff (Keeney et al. 1976), SMART (simple multi-attribute rating technique) (Edwards 1977), Swing (Mustajoki et al. 2005), AHP (analytic hierarchy process) (Saaty 1977), ANP (analytic network process) (Saaty 1990), and BWM (best-worst method) (Rezaei 2015). For more information about popular MCDM methods, see (Triantaphyllou 2000).

Conventional MCDM methods typically analyze a decision-making problem that entails one DM only. Members usually have distinct preferences when there is a group of DMs for decision-making. Three approaches exist to deal with the differences in preferences: sharing, comparing, and aggregating (Belton and Pictet 1997). In sharing, the whole group will arrive at a unified preference structure so that the group is treated as a single DM problem. For instance, if one wants to find the weights of a set of criteria in sharing, they will end up with a single set of weights through negotiation among members and by using a weighting method (e.g., AHP). In contrast, individual preferences are considered in comparing and aggregating, where we have different sets of weights from individual members. In aggregating, for instance, we try to aggregate the weights from all members to come up with a single set of weights (Amenta et al. 2020).

Aggregating is usually performed in two different ways. First, the methods for one DM are extended to encompass multiple DMs' preferences. The most popular technique known so far is arguably the geometric mean method (Ramanathan and Ganesh 1994; Forman and Peniwati 1998) that is applied, for instance, to the pairwise comparison matrices (PCMs) computed within the AHP, the result of which is an aggregated PCM representing the preferences of the whole group. Other methods use more complicated techniques, such as evolutionary algorithms (Blagojevic et al. 2016; Abel et al. 2015) and Bayesian statistics (Altuzarra et al. 2007; Mohammadi and Rezaei 2020), to aggregate the preferences of multiple DMs. This is also called input-based aggregation (Dias and Clımaco 2005). The second approach for aggregation is to compute all DMs' priorities individually and then conduct the consequent processes based on such priorities, such as aggregation and clustering of DMs. This approach is called output-based aggregation (Dias and Clımaco 2005). The arithmetic mean is typically used for output-based aggregation and is seemingly appropriate since it satisfies the non-negativity and constant-sum (e.g., unit-sum) constraints required in most MCDM methods. Other processes, such as computing the standard deviation of priorities of the group (Tomashevskii and Tomashevskii 2019; Tomashevskii 2015), clustering of DMs (Abel et al. 2014; Meixner et al. 2016), and statistical significance

tests (Chiclana et al. 2013; Blagojevic et al. 2016), are directly applied to the priorities computed based on the preferences of DMs.

While the methods for aggregating multiple DMs' preferences are often statistically sound, applying the same statistical operations to the priorities is incorrect because the priorities are ratios lying on a simplex and not on the real space. Like the priorities of a DM, a vector that satisfies the non-negativity and constant-sum constraints is called a *composition*[1] (Aitchison 1982). From a statistical point of view, the statistical operations should be adjusted to apply to such compositions; otherwise, the follow-up methods or outcomes are unreliable and statistically incorrect. While there is a tremendous effort, such as several books (Aitchison 1982; Pawlowsky-Glahn and Buccianti 2011; Buccianti et al. 2006), in providing proper statistical tools and methods for analyzing compositional data, MCDM researchers often neglect them, resulting in improper statistical analysis with disastrous consequences. This article aims to study three of such essential and prevalent fallacies in analyzing the priorities of a group of DMs.

First, we analyze the aggregation of different DMs' priorities using the notions of compositional data since computing the arithmetic or geometric mean directly from the priorities is not in line with their compositional nature. We show that the aggregated priorities based on our analysis are equivalent to the normalized geometric mean of the priorities, called the geometric mean method (GMM), in the MCDM literature (Forman and Peniwati 1998; Ramanathan and Ganesh 1994). The byproduct of such an analysis is that using the arithmetic mean for aggregating the priorities should be avoided. This analysis ends a lengthy discussion in MCDM regarding using the arithmetic or geometric mean for aggregating priorities. In addition, the adaptive weighted geometric mean method (AWGMM) is developed, which is a more robust surrogate to the GMM. The robustness of the method is against DMs whose preferences significantly deviate from the majority of other DMs. The AWGMM uses the robust Welsch estimator developed in robust statistics (Huber 2004), which adaptively assigns a weight for each DM based on their priorities. A DM with a small weight is deemed *deviant* and thus contributes less to the final aggregated priorities computed in the AWGMM. Aside from the aggregation, the identification of DMs with deviating preferences could also be used in the negotiation process to converge the preferences of the whole group.

Second, the errors in estimating priorities' standard deviation (as a measure of dispersion) are explained. Since statistical tests are typically based on the standard deviation of samples, the flaw in computing the standard deviation directly impacts the tests. Thus, the change in calculating the standard deviation for priorities (as compositional data) means that statistical tests cannot be directly applied to priorities to verify the difference between the weights of the two criteria. For example, statistical tests, such as paired t-test and Wilcoxon Signed-rank test, are used to check the importance difference of criteria based

[1] Even if they do not lie on a standard simplex, the priorities in MCDM convey the relative importance of one criterion over another, meaning that they are still compositional and statistical methodologies for processing them should be different than other standard multivariate data.

on the priorities of a group by *subtracting* their weights (Chiclana et al. 2013), which is in not statistically correct given the compositional nature of the priorities. Instead, we develop a Bayesian Wilcoxon-type test to verify if the difference between the weights of two criteria is significant based on a group of DMs' priorities. The Bayesian tests do not suffer from the pitfalls of the frequentist tests and enable us to compute the extent to which one criterion is preferred over another based on the priorities of a group of DMs. This approach provides us with a probabilistic ranking of criteria. Aside from the Wilcoxon-type test, the ways to use the Bayesian *t*-test and Bayesian Sign test (i.e., beta-binomial conjugate) are also explained.

Third, we put forward the proper ways to measure the distance (as another measure of dispersion) between two priorities. One of the most popular distance functions is the Euclidean distance which is directly applied to the weights in the priorities (Abel et al. 2014). The distance functions are used for different aims, but one of the essential uses of a distance function is clustering the DMs based on their priorities. Typically, the Euclidean distance and mean absolute deviation (MAD) are used for clustering the DMs using clustering algorithms such as K-means and fuzzy C-means. Instead, we introduce the compositional extension of the distance metrics, according to which we extend the K-means algorithm that is statistically appropriate and correct for grouping the DMs based on their priorities.

In summary, the contributions of this article are as follows:

- We show that the proper way to aggregate the priorities is to use the geometric mean method (GMM) and that the arithmetic mean should be avoided. Further, the adaptive weighted geometric mean method (AWGMM) is also developed by using the robust Welsch estimator.
- The errors of computing the standard deviation of priorities are discussed, and proper Bayesian statistical tests for verifying the difference between the weights of two criteria are developed, the result of which is a probabilistic ranking of criteria.
- A proper distance function between two priorities is reviewed, according to which a new clustering method is developed for grouping multiple DMs based on their priorities.

The paper is structured as follows. Section 2 presents the preliminary concepts of MCDM and compositional data, which will be extensively used in the subsequent sections. Section 3 is dedicated to the aggregation and the fallacies therein and then puts forward a new robust aggregation method. In Sect. 4, we discuss the shortcoming in computing standard deviation and statistical tests in MCDM and develop mathematically sound tests that can be used in such situations. Section 5 is devoted to the applications of distance-related measures in MCDM, where we also develop some new correct ways which can be used in the context of MCDM. Section 6 studies a real-world MCDM example, where we show the use of the proposed methods. Finally, the paper is concluded in Sect. 7.

Table 1 An example of pairwise comparison matrix (PCM) for three criteria, i.e., $C = \{c_1, c_2, c_3\}$

	c_1	c_2	c_3
c_1	1	2	8
c_2	1/2	1	4
c_3	1/8	1/4	1

2 Preliminary

This section presents the preliminary concepts from MCDM and compositional data, which will be used in the following sections. We first review some rudimentary concepts from pairwise comparison matrices (PCMs) and discuss the consistency of such an MCDM problem. We then look into the compositional data and provide necessary introductory notions.

2.1 Pairwise Comparison Matrix

Pairwise comparison matrices (PCMs) were first introduced and used in AHP (analytic hierarchy process) (Saaty 1977; Ishizaka and Labib 2011), where they were used to identify the importance of a set of criteria. For a set of criteria $C = \{c_1, c_2, \dots, c_n\}$, the following definition provides the notion of PCMs.

Definition 2.1 For a set of n criteria, a PCM $M = \{m_{ij}\}_{i,j=1}^{n}$ is a square matrix of order n, whose element m_{ij} expresses the relative importance of criterion c_i over c_j.

A particular example of a PCM is shown in Table 1. The preferences of criteria over each other are shown on a scale of 1–9. A number like 8 in this table indicates that c_1 in eight times more important than c_3.

In MCDM, the importance of a set of criteria like C is denoted by a priority vector $w \in \mathcal{R}^n$, where each w_j is non-negative for all $j = 1, \dots, n$, and the sum of all weights come to one. Therefore, each element of a PCM should satisfy the following property:

$$m_{ij} = \frac{w_i}{w_j}, \quad \forall i, j = 1, \dots, n. \tag{1}$$

There does not need to exist a unique vector w such that Eq. (1) satisfies all the elements in a PCM. But if such a unique vector exists, the PCM is said to be fully consistent. The following two definitions provide the notion of consistency for a PCM differently.

Definition 2.2 (Saaty 1977) A PCM is said to be fully consistent if and only if it satisfies the multiplicative-transitivity property, defined as:

$$m_{ij} = m_{ih} \times m_{hj}, \quad i,j,h = 1,\ldots,n. \tag{2}$$

Definition 2.3 (Saaty 2000) A PCM is said to be fully-consistent if and only if there exists a unique $w \in \mathcal{R}^n$, for which Eq. (1) holds true.

When a PCM is not fully consistent, there are different ways to compute the priorities w. The popular methods are the maximal eigenvector method (EVM) and the geometric mean method. The EVM is based on the eigenvalue decomposition, where the criteria priorities are set to be the maximal eigenvector. On the other hand, according to the geometric mean method for PCM, the optimal priorities of criteria are computed by taking geometric means of columns, defined as:

$$w_i = \sqrt[n]{\prod_{j=1}^{n} \hat{m}_{ij}}, \quad i = 1,\ldots,n, \tag{3}$$

where $\hat{M} = \{\hat{m}\}_{i,j=1}^{n}$ is the column-wise normalized version of a PCM. In both methods, the ratio of weights in Eq. (1) approximately holds. In the case that the PCM is fully consistent, the column-wise normalization of all columns is the same, making any normalized column the desired criteria priorities.

2.2 Compositional Data

The difficulty of processing ratios with common elements in their nominators or denominators is recognized in statistics a long time ago. In his famous paper on spurious correlation in Pearson (1897) pointed out that the correlation of ratios with common parts in their nominators/denominators may not be precisely correct. However, Pearson's precautions went unheeded until the 1980s, when some studies highlighted the methods and tools for analyzing ratios with common elements (Aitchison 1982). These studies were based on data whose sum is a fixed number, which were called *compositions*. The following definition provides a concise explanation of this notion.

Definition 2.4 (Aitchison 1982) A composition w of n parts is an $n \times 1$ vector with positive components $w_1,\ldots,w_n$ whose sum is 1.

Be aware that the sum of components of a composition in Definition 2.4 is not necessarily 1 but can be any other number, e.g., 100. Moreover, even the sum of the parts can be unknown and not equivalent for different compositions in the data set. An example of such data is the household budget, where the expenses of families are classified different categories four categories (Aitchison 1982). Usually, different families have different costs, typically commensurate to their income, but for analyzing such data, the ratio between different expense classes is critical, and the data is thus compositional. In this circumstance, we can divide the value of an expense class for each family by the total expenses of the family to form a compositional vector from the raw expense values. As a result, a different (unknown) constant-sum

does not change the nature of the compositional data, where the ratio between different parts matters, not the magnitude of its parts.

Given that compositions have a limiting constraint, the statistical analysis of compositional data must be different. As a concrete example, we cannot assume that a composition follows a normal distribution since it does not lie in the real space $\mathcal{R}^n$. Instead, it lies in lower-dimensional simplices, making the available statistical tools inapplicable to such data.

The priorities of criteria in MCDM are indeed compositional. As a result, statistical analysis—even simple arithmetic or geometric mean—should be investigated before applying directly to such data. In the following sections, several statistical methods are analyzed, and proper methodologies for processing priorities in MCDM are developed and put forward, which align with the priorities' compositional nature.

3 Priorities Aggregation

Group members involved in a decision-making problem can be expected to have different preferences, which might be due to different causes, including (1) uncertainty (due to lack of relevant information and proper structuring of the problem); (2) conflict (due to different values or priorities); or (3) misunderstanding (due to different perspective and partial information) (Belton and Pictet 1997). There are generally three approaches to handling the differences among group members: sharing, comparing, and aggregating (Belton and Pictet 1997). By sharing, the analyst/facilitator tries to moderate a discussion, for instance, via the decision conferencing (Phillips 1990; McCartt and Rohrbaugh 1989), among the members to address the differences by discussing the causes of different preferences with the ultimate aim of finding common ground and agreement. In this approach, the group is treated as a single DM. In comparing, on the other hand, members are considered individually, and the preferences obtained from members are used for further negotiation and comparison to reach a consensus. Finally, in aggregating, the individual preferences are not negotiated, nonetheless, the analyst/facilitator tries to find common ground where the differences among the preferences are reduced.

While reaching a consensus among the group members by following sharing or comparing approaches is desirable, a unanimous agreement is not always guaranteed. This is why the aggregating approach could also be seen as complementary to the comparing approach. A responsible analyst/facilitator should try to moderate a discussion among the members and make their views close. However, if reaching a consensus seems impossible, aggregating the preferences (which have already become closer) comes to the picture in the end. Also, in some situations where sharing and comparing is not easy, e.g., when we need to collect the preferences of a relatively large number of participants, aggregating seems the only possible option. Examples could be arriving at a decision in a municipality by collecting the preferences of citizens of a town or formulating a policy by collecting the preferences of electric vehicle users in a region. We refer to Belton and Stewart (2002) for detailed explanations of the three approaches mentioned here. Regarding aggregating the

Table 2 Notation used throughout the paper

Variable	Description
K	Number of DMs
n	Number of criteria
$W \in R^{K \times n}$	matrix containing all the priorities of K DMs for n criteria
$\hat{W} \in R^{K \times n(n-1)/2}$	Log-ratio transformation of priorities
P	Performance matrix
$w^g \in \mathcal{R}^n$	Aggregated priority
$\hat{w} \in R^{n(n-1)/2}$	Log-ratio transformation of the aggregated priority

priorities, as we discussed earlier, we need to use operations suitable for priorities, i.e., operations on compositions.

In this section, we first review the compositional approach to aggregating the priorities of multiple DMs and show that the geometric mean of priorities should be utilized for aggregation, and the use of the arithmetic mean should be avoided in this case. Further, a method based on robust statistics is developed for aggregating the priorities that are robust to deviant DMs and identifies the DMs whose priorities are different from the majority of other DMs, making them have a lesser impact on the final aggregated priorities. Identifying deviant DMs can also be used in the negotiation process to converge the preferences of the DMs in the group. The notations used in this article are also shown in Table 2.

3.1 Arithmetic Versus Geometric Mean

There are two widely-used approaches for aggregating multiple DM priorities: arithmetic and geometric mean. Initially, the arithmetic mean method (AMM) was reported to be the most appropriate method because the geometric mean violates the Pareto optimality (Ramanathan and Ganesh 1994). Later on, Forman and Peniwati showed that the geometric mean preserved the Pareto optimality as well (Forman and Peniwati 1998) and recommended using the geometric mean for aggregating priorities without a solid justification. However, this recommendation is taken for granted, and the arithmetic mean has been used mainly for aggregating the priorities. Even in a more recent review article (Ishizaka and Labib 2011), only the arithmetic mean is mentioned as the only means for aggregating the priorities in the AHP.

These two methods have a fundamental problem of averaging DM priorities, each element showing only the relative importance of one criterion concerning other criteria. Since only the ratio between the weights of different criteria, and not the magnitude of weights, matters, we cannot directly apply the arithmetic or geometric mean over the priorities of multiple DMs. Instead, we first need to compute all the possible ratios between the elements in a priority or weight vector and then take an average over the ratios. For n criteria, all the possible ratios are $n(n-1)/2$ placed in a *compositional average array*.

Definition 3.1 (Compositional average array (Aitchison 1982)) For an aggregated n-part compositions like w, the compositional average array is given by

$$E = \begin{bmatrix} 0 & \xi_{12} & \xi_{13} & \cdots & \xi_{1n} \\ \xi_{21} & 0 & \xi_{23} & \cdots & \xi_{2n} \\ \vdots & & & & \vdots \\ \xi_{n1} & \xi_{n2} & \xi_{n3} & \cdots & 0 \end{bmatrix},$$

where $\xi_{ij} = \mathbb{E}\{\ln \frac{w_i}{w_j}\}$, and $\mathbb{E}$ is the mathematical expectation.

The mathematical expectation in Definition 3.1 can be replaced by the average for the empirical analysis since the log-ratios lie within the real space $\mathcal{R}$. Thus, one can compute each element of the compositional average array based on the priorities of K DMs as:

$$\xi_{ij} = \mathbb{E}\left\{\ln \frac{w_i}{w_j}\right\} = \frac{1}{K}\sum_{k=1}^{K} \ln \frac{W_{ki}}{W_{kj}}, \quad \forall i, j = 1, \ldots, n. \tag{4}$$

In addition, it is evident that:

$$\mathbb{E}\{\xi_{ij}\} = \mathbb{E}\{\xi_{ik}\} + \mathbb{E}\{\xi_{kj}\}, \quad \forall i, j = 1, \ldots, n. \tag{5}$$

On the other hand, a PCM, like M, is said to be fully-consistent if $m_{ij} = m_{ih} \times m_{hj}$ for all i, j, k in the range of the matrix (see Definition 2.2). Now, if we define the matrix $\hat{E}$ by taking element-wise exponential from E, i.e., $\hat{E} = exp(E)$, this matrix can be viewed as a fully-consistent PCM. For fully consistent PCMs, a normalized column would yield the final aggregated priorities. The following lemma shows that a normalized column of matrix $\hat{E}$ is equivalent to the normalized geometric mean of all priorities or the GMM.

Lemma 3.2 *A normalized column of the exponential-transformed compositional average array is tantamount to the geometric mean method* (*GMM*).

Proof Without loss of generality, we take the first column of the compositional average array and show that the normalization of the first column is equivalent to the GMM. Let $c \in \mathcal{R}^n$ be the exponential-transformed of the first column, then,

$$\begin{aligned} c_i &= exp(\xi_{i1}), \quad i = 1, \ldots, n, \\ &= exp\left(\frac{1}{K}\sum_{k=1}^{K} \ln \frac{W_{ki}}{W_{k1}}\right) \\ &= exp\left(\ln \prod_{k=1}^{K} \left(\frac{W_{ki}}{W_{k1}}\right)^{\frac{1}{k}}\right) \\ &= \prod_{k=1}^{K} \left(\frac{W_{ki}}{w_{k1}}\right)^{\frac{1}{K}}. \end{aligned} \tag{6}$$

Now, the normalization of c yields

$$\begin{aligned}\frac{c_i}{\sum_{j=1}^{n} c_j} &= \frac{1}{\sum_{j=1}^{n} \prod_{k=1}^{K} \frac{W_{kj}^{\frac{1}{K}}}{W_{k1}^{\frac{1}{K}}}} \prod_{k=1}^{K} \left(\frac{W_{ki}}{W_{k1}}\right)^{\frac{1}{K}} \\ &= \frac{\prod_{k=1}^{K} W_{ki}^{\frac{1}{K}}}{\sum_{j=1}^{n} \prod_{k=1}^{K} W_{kj}^{\frac{1}{K}}},\end{aligned} \tag{7}$$

which is the normalized geometric mean of priorities or the GMM, and that completes the proof.

Corollary 3.3 *Taking any average over the raw priorities is not statistically correct. Nevertheless, the analysis based on compositional data showed that the GMM is the proper average of all the priorities. This implies that using the arithmetic mean for aggregating DM priorities should be avoided.*

3.2 Robust Aggregation Based on the Welsch Estimator

In Eq. (4), the mathematical expectation is replaced by the averaging over the available samples, i.e., the priorities of K DMs, and the final aggregated outcome becomes equivalent to the GMM. Another option is to use more robust statistics, such as the median, to produce more robust priorities for deviant DMs. In aggregating priorities, a deviant is a DM whose preferences deviate significantly from most other DMs. Note also that using the median on the original data is erroneous. The final aggregated results using the median in Eq. (4) are not tantamount to the GMM but are a more robust surrogate. However, the median would not lead to a fully consistent PCM based on the associated compositional average array.

Another choice, instead of averaging, is to use different estimators. In robust statistics, there are a number of estimators that can provide more robust estimations. Such estimators are called M-estimators that can replace the mean or median in Eq. (4). For example, the Welsch M-estimator is one of the well-known estimators that has shown promising performance in noisy environments (He et al. 2010; Mohammadi et al. 2016). To use this estimator for computing the compositional average array, we consider the log-ratio transformed data $\hat{W}$ and estimate the aggregated log-ratio transformed priority $\hat{w}^g$ by solving the following optimization problem:

$$\min_{\hat{w}^g} \sum_{k=1}^{K} \phi(\|\hat{W}_{k\cdot} - \hat{w}^g\|), \tag{8}$$

where $\phi(x) = exp(-x^2/\sigma^2)$ is the Welsch estimator. For $\phi(x) = x$, minimization (8) yields the same solution as the arithmetic mean of log-ratio transformed data, so the

result of the compositional average array would be identical to the GMM. Using the Welsch, a more robust estimator, we expect that the aggregation outcome will be more robust to deviants, i.e., the DMs whose preferences are significantly different from the majority of other DMs. Problem (8) is not convex (Geman and Reynolds 1992), but the following lemma paves the way for an efficient solution.

Lemma 3.4 (Geman and Reynolds 1992) *For a fixed x, there is a potential dual function ψ such that*:

$$\phi(x_i) = \inf_{\alpha_i} \alpha_i x_i^2 + \psi(\alpha_i), \tag{9}$$

where $\psi(.)$ is the convex conjugate of $\phi(.)$, and $\alpha_i > 0$ is an auxiliary variable, which is determined by the so-called minimizer function $\delta(.)$ defined as:

$$\delta(x) = exp\left(-\frac{x^2}{\sigma^2}\right). \tag{10}$$

Using Lemma 3.4, problem (8) can be written as:

$$\min_{\hat{w}^g,\alpha} \sum_{k=1}^{K} \alpha_k \|\hat{W}_{k\cdot} - \hat{w}^g\|^2 + \psi(\alpha_k). \tag{11}$$

Thus, the following steps must be iterated until convergence is reached:

$$\begin{aligned} &\alpha_k = \delta\left(\|\hat{W}_k - \hat{w}^g\|\right) = exp\left(-\frac{\|\hat{W}_{k\cdot} - \hat{w}^g\|}{\sigma^2}\right), \quad \forall k = 1, \ldots, K, \\ &\hat{w}^g = arg\min_{\hat{w}^g} \sum_{k=1}^{K} \alpha_k \|\hat{W}_{k\cdot} - \hat{w}^g\|^2. \end{aligned} \tag{12}$$

For the second step in (12), we need to take the derivative and find the optimal solution as:

$$\begin{aligned} &\frac{\partial}{\partial \hat{w}^g} \sum_{k=1}^{K} \alpha_k \|\hat{W}_{k\cdot} - \hat{w}^g\|_2^2 = 0 \\ &\Rightarrow \sum_{k=1}^{K} \alpha_k \hat{W}_{k\cdot} = \sum_{k=1}^{K} \alpha_k \hat{w}^g \\ &\Rightarrow \hat{w}^g = \sum \lambda_k \hat{W}_{k\cdot}, \quad \lambda_k = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}. \end{aligned} \tag{13}$$

In addition, the performance of the Welsch estimator is heavily reliant on selecting its parameter σ. As the recent studies suggest (He et al. 2010; Mohammadi et al. 2015), this parameter can be recursively updated in each iteration as:

$$\sigma = \frac{\sum_{k=1}^{K} \|\hat{W}_{k\cdot} - \hat{w}^g\|_2^2}{n^2}. \tag{14}$$

Algorithm 1 summarizes the overall procedure for aggregating the priorities of multiple DMs by using the Welsch estimator.

Algorithm 1 Adaptive weighted geometric mean method (AWGMM)

Input: Priorities $W \in R^{K\times n}$.
while *NotConverged* **do**
$\alpha_k = exp(-\frac{\|\hat{W}_{k\cdot} - \hat{w}^g\|}{\sigma^2}), \quad \forall k = 1, ..., K$
$\lambda_k = \alpha_k / \sum_j \alpha_j, \quad k = 1, 2, ..., K$
$\hat{w}^g = \sum_k \lambda_m \hat{W}_k$.
end while
Output Final weight ratios $\hat{w}^g$, λ

The value of λ_k in Algorithm 1 shows the contribution of each DM to the final aggregated priorities, and the DMs with deviating preferences from the majority of other DMs are assigned a lower λ_k. As a result, the DMs with deviating preferences from the vast majority of other DMs can be detected by considering their associated λ_k. Note also that $\lambda \in R^K$ satisfies the non-negativity and unit-sum constraints, allowing us to view it as a weight for DMs based on their opinion proximity to other DMs.

Remark 3.5 The values of λ from Algorithm 1 identifies the deviant DMs. Instead of aggregation, the decision facilitator/analyst can use these values to understand what DMs have different preferences and use such information in the negotiation process.

Remark 3.6 We do not claim the preference of w^g_{AWGMM} and w^g_{GMM} over each other. While both are mathematically sound approaches for aggregating priorities, their main difference is handling the deviants. While w^g_{GMM} assigns equal weights to all the DMs, w^g_{AWGMM} assigns lower weights to the DMs whose priorities are far from the majority of the DMs. We think the suitability of the two approaches depends on a particular decision-making situation; if all the DMs involved in the decision-making process must have equal contributions, then the GMM must be used. However, if the majority opinion is of more interest, then AWGMM seems more appropriate. In any case, using the arithmetic mean for aggregating priorities should be avoided.

After obtaining $\hat{w}^g$ from Algorithm 1, it can be placed in a compositional average array, which happens to be fully consistent, making the aggregated priorities be obtained by normalizing a column of such a matrix. The following lemma proves the consistency of the compositional average array built based on $\hat{w}^g$.

Lemma 3.7 *The compositional average array computed based on Algorithm* 1 *is fully consistent*, *thereby providing unique aggregated priorities*.

Proof Algorithm 1 outputs the log-ratio transformed $\hat{w}^g$ and $\lambda \in R^K$. Accordingly, the compositional average array can be constructed as in Definition 3.1, where

$$\begin{aligned}\xi_{ij} &= \mathbb{E}\left\{ \ln \frac{W_{ki}}{W_{kj}} \right\} \\ &= \sum_{k=1}^{K} \lambda_k \ln \frac{W_{ki}}{W_{kj}}, \quad \forall i,j = 1,\ldots,n,\end{aligned}$$

where λ is obtained from Algorithm 1. It is now simple to show that:

$$\xi_{ij} = \xi_{ik} + \xi_{kj},$$

which means that the exponential-transformed of such a matrix is fully consistent and provides unique aggregated priorities, and that completes the proof.

The proposed aggregation method has several features, some of which are listed in the following:

- *Pareto optimality* entails that if all DMs in a group prefer A over B, then the group decision should also favor A (Forman and Peniwati 1998). In the proposed aggregation method, if all DMs favor criterion i over j, then:

$$\ln \frac{W_{ki}}{W_{kj}} > 0, \quad \forall k = 1,\ldots,K, \tag{15}$$

  and since λ_k is non-negative and is summed to one in Algorithm 1, it follows

$$\sum_{k=1}^{K} \lambda_k \ln \frac{W_{ki}}{W_{kj}} > 0, \tag{16}$$

  which means that the proposed aggregation method satisfies Pareto optimality.
- *Non-dictatorship* refers to the fact that no individual priorities become the priorities of the group automatically, regardless of the preferences of other members in the group. The proposed method assigns a weight to each DM, the magnitude representing the corresponding DM's contribution to final aggregated weights. Although the procedure might give some DMs a lower weight (even nearly zero), the weights are only assigned after considering the priorities of all DMs in a group. In addition, if a DM is added or removed from a group, the weights of DMs and the final aggregated priorities will change, confirming that the aggregated priorities are not automatically biased toward one of the DMs.
- *Recognition* means that the group decision is arrived at after considering all the members' priorities. In the proposed aggregation method, each DM is assigned a weight after considering the priorities of all group members: If the priorities of a DM are different from those of other group members, then it is assigned a lower weight. It means that the final aggregated priorities are computed based on the priorities of *all* decision-makers. Further, adding or removing a DM will change

the weights assigned to each DM and the final aggregated priorities, which also corroborates that AWGMM considers all DMs' priorities before arriving at the aggregation priorities.

3.3 An Illustrative Example

This section illustrates the aggregation procedure through an example. In this regard, assume that five DMs have expressed their preferences on four criteria $C = \{c_1, c_2, c_3, c_4\}$, and the matrix W containing the priorities of five DMs is as follows:

$$W = \begin{array}{c} \\ DM1 \\ DM2 \\ DM3 \\ DM4 \\ DM5 \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{pmatrix} 0.220 & 0.435 & 0.295 & 0.050 \\ 0.210 & 0.434 & 0.312 & 0.044 \\ 0.363 & 0.312 & 0.107 & 0.218 \\ 0.243 & 0.386 & 0.332 & 0.039 \\ 0.227 & 0.381 & 0.339 & 0.053 \end{pmatrix} \end{array}. \tag{17}$$

The result of the arithmetic mean is:

$$w^g_{AMM} = \begin{bmatrix} 0.253 & 0.389 & 0.277 & 0.081 \end{bmatrix}. \tag{18}$$

We first create the compositional average array, as defined in Definition 3.1. In this regard, we compute element ξ_{12} as:

$$\begin{aligned} \xi_{12} &= \frac{1}{5}\left(\ln \frac{0.220}{0.435} + \ln \frac{0.210}{0.434} + \ln \frac{0.363}{0.312} + \ln \frac{0.243}{0.386} + \ln \frac{0.227}{0.381} \right) \\ &= -0.446. \end{aligned}$$

Similarly, other elements in the compositional average array are computed, the result of which is as follows:

$$E = \begin{pmatrix} 0 & -0.446 & -0.036 & 1.371 \\ 0.446 & 0 & 0.411 & 1.817 \\ 0.036 & -0.411 & 0 & 1.407 \\ -1.371 & -1.817 & -1.407 & 0 \end{pmatrix}. \tag{19}$$

By taking exponential and normalizing a column of this matrix, we arrive at the aggregated priorities as:

$$w^g_{GMM} = \begin{bmatrix} 0.260 & 0.405 & 0.269 & 0.066 \end{bmatrix}, \tag{20}$$

which is identical to the GMM and is considerably different from (18). We further apply Algorithm 1 to find the aggregated priorities. This algorithm works with the log-ratio transformed data $\hat{W}$ and has, as one of its outputs, $\lambda \in R^K$ that works as a weight for different DMs. The $\hat{W}$ and λ for this example are:

$$\hat{W} = \begin{array}{c} \lambda \\ 0.31 \\ 0.33 \\ 0.00 \\ 0.24 \\ 0.12 \end{array} \begin{array}{c} \begin{array}{ccccccc} \frac{c_1}{c_2} & \frac{c_1}{c_3} & \frac{c_1}{c_4} & \frac{c_2}{c_3} & \frac{c_2}{c_4} & \frac{c_3}{c_4} \end{array} \\ \begin{pmatrix} -0.681 & -0.294 & 1.480 & 0.387 & 2.161 & 1.774 \\ -0.722 & -0.391 & 1.560 & 0.331 & 2.282 & 1.952 \\ 0.151 & 0.221 & 0.511 & 1.070 & 0.360 & -0.710 \\ -0.519 & -0.403 & 1.461 & 0.116 & 1.980 & 1.864 \\ -0.462 & -0.311 & 1.841 & 0.151 & 2.303 & 2.152 \end{pmatrix} \end{array}.$$

As is evident from this matrix, the third DM is assigned a weight of approximately zero,[2] mainly because their preferences are significantly different from others. Thus, it is treated like a deviant and does not influence the final aggregated log-ratio $\hat{w}^g$ that is computed as:

$$\hat{w}^g = \begin{bmatrix} -0.628 & -0.354 & 1.546 & 0.274 & 2.174 & 1.90 \end{bmatrix},$$

which can be placed in an array as:

$$E = \begin{pmatrix} 0 & -0.628 & -0.354 & 1.546 \\ 0.628 & 0 & 0.274 & 2.174 \\ 0.354 & -0.274 & 0 & 1.90 \\ -1.546 & 2.174 & -1.90 & 0 \end{pmatrix}. \tag{21}$$

It can be verified that the exponential-transformed of this matrix is a fully consistent PCM, which was proved in Lemma 3.7. Now, if we take the exponential and then normalize a column, the final aggregated priorities are obtained as:

$$w^g_{AWGMM} = \begin{bmatrix} 0.225 & 0.410 & 0.319 & 0.046 \end{bmatrix}. \tag{22}$$

As expected, the average computed by the AWGMM in (22) is different from those in (20) and (18) since one of the DMs is assigned a small weight (nearly zero) and has therefore minimal impact on the final aggregated priorities. Note also that the values of λ can be used in the negotiation, as the third DM has significantly different preferences than others.

4 Standard Deviation and Statistical Tests

Similar to the discussion regarding the central tendency of priorities, the standard deviation for the priorities is defined differently. The change of central tendency and standard deviation influences statistical tests, such as paired *t*-test and Wilcoxon Signed-rank test, which are often used in processing the priorities. The use of such statistical tests is to verify if the difference between the weights of the two criteria is significantly different (Chiclana et al. 2013). In this section, we first study the estimation of standard deviation from the perspective of compositional data analysis and show that such a definition would provide correct and more meaningful results than computing the standard deviation from raw priorities. We then combine compositional data analysis with Bayesian statistics to calculate the probability that one

[2] Note that it is not absolute zero, but an infinitesimal number.

criterion is more important than another based on the priorities of multiple DMs and provide a probabilistic ranking of the criteria.

4.1 Standard Deviation

The standard deviation of elements in priorities has also been used without considering the inherent constraints in the compositional data (Tomashevskii and Tomashevskii 2019; Tomashevskii 2015). Like the arithmetic mean, the standard deviation calculation based on raw priorities is also incorrect and invalidates the consequent decisions and methods. In addition, the standard deviation estimated based on the priorities typically has a magnitude that is as large as or even larger than the mean of the priorities (see Tomashevskii and Tomashevskii 2019; Tomashevskii 2015), thereby having narrow applicability in practice.

For compositional data, we need to compute the deviation of all possible ratios in a composition. Therefore, we define the *compositional deviation array* that includes the standard deviation of all possible ratios.

Definition 4.1 (Compositional deviation array (Aitchison 1982)) For an n-part composition like w, the compositional deviation array is given by

$$\mathcal{T}=\begin{bmatrix} - & \tau_{12} & \tau_{13} & \dots & \tau_{1n} \\ \tau_{21} & - & \tau_{23} & \dots & \tau_{2n} \\ \vdots & & & & \vdots \\ \tau_{n1} & \tau_{n2} & \tau_{n3} & \dots & - \end{bmatrix},$$

where $\tau_{ij}=\sqrt{var\left(\ln\frac{w_i}{w_j}\right)}$, and var is the variance operator.

The variance in Definition 4.1 can be replaced by the empirical variance defined as:

$$\tau_{ij}^2=var\left(\ln\frac{w_i}{w_j}\right)=\frac{1}{K-1}\sum_{k=1}^{K}\left(\ln\frac{W_{ki}}{W_{kj}}-\xi_{ij}\right)^2,\quad \forall i,j=1,\dots,n. \tag{23}$$

Definition 4.1 shows that for n criteria, we indeed have $n(n-1)/2$ unique standard deviations, equivalent to the number of possible ratios for an n-part composition. This means that, based on such a definition, the analysis based on the average and standard deviation of each criterion computed according to raw priorities makes no sense. Besides, by using the deviation and average arrays, it is possible to provide more insights regarding the priorities of different DMs over a set of criteria, which provides more description of DMs' priorities and the relative importance of different criteria rather than a sole ratio.

Before illustrating the usefulness of the compositional deviation array through an example, it is worth noting that the definition of variance for the deviation array can also be replaced by the median absolute deviation (MAD), which is based on the median and is a more robust surrogate for standard deviation. In addition, based on the proposed aggregation method AWGMM, we can define a robust variance as:

$$var\left(\ln\frac{w_i}{w_j}\right)^2 = \sum_{k=1}^{K}\lambda_k\left(\ln\frac{W_{ki}}{W_{kj}} - \xi_{ij}^{AWGMM}\right)^2, \tag{24}$$

where λ and ξ^{AWGMM} are computed based on the proposed aggregated method in Algorithm 1. Given these definitions, we illustrate the use of standard deviation by an example. In addition, since the magnitudes of values in average and deviation arrays are identical, we can place the numbers in an array to summarize the priorities of multiple DMs in one array only. The next definition presents such an array.

Definition 4.2 (Compositional average-deviation array) For an n-part composition like w, the compositional average-deviation array (AD) is defined as:

$$\begin{pmatrix} 0 & \xi_{12} & \xi_{13} & \cdots & \xi_{1n} \\ \tau_{21} & 0 & \xi_{23} & \cdots & \xi_{2n} \\ \vdots & & & & \vdots \\ \tau_{n1} & \tau_{n2} & \tau_{n,3} & \cdots & 0 \end{pmatrix},$$

where ξ_{ij} is the average, i.e., mean, median, or a robust estimation based on Algorithm 1, and τ_{ji} is the associated standard deviation to ξ_{ij}.

Example 4.3 We compute the compositional average-deviation array for the example in Sect. 3.3 by using the mean, median, and the method proposed in Algorithm 1. We first begin with the mean and standard deviation, that is computed as:

$$AD_{mean} = \begin{array}{c} \\ c1 \\ c2 \\ c3 \\ c4 \end{array}\begin{array}{c} \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{pmatrix} 0 & -0.446 & -0.035 & 1.370 \\ 0.351 & 0 & 0.410 & 1.817 \\ 0.704 & 0.386 & 0 & 1.406 \\ 0.504 & 0.824 & 1.191 & 0 \end{pmatrix} \end{array}. \tag{25}$$

The magnitude of the above-diagonal values suggests the difference between the two criteria, and its sign indicates which criterion is more important. For instance, the value -0.446 indicates that criterion c_2 is more important than criterion c_1, since the value is negative. The biggest difference is between c_2 and c_4 whose magnitude of average differences is 1.817. The lower-diagonal entries represent the standard deviation, which can help realize how reliable the difference between the two criteria is based on a group of DMs' priorities. For instance, the average difference between c_1 and c_3 has the infinitesimal value of -0.035, while the standard deviation is 0.704. Therefore, one can readily understand that these criteria have similar priorities or importance to the group of DMs.

The same average-variation array can be computed by using the median and median absolute deviation, as well as the proposed robust aggregation of priorities, i.e.,

$$AD_{median} = \begin{pmatrix} 0 & -0.518 & -0.311 & 1.479 \\ 0.162 & 0 & 0.330 & 2.160 \\ 0.080 & 0.179 & 0 & 1.864 \\ 0.081 & 0.142 & 0.090 & 0 \end{pmatrix}, \tag{26}$$

$$AD_{AWGMM} = \begin{pmatrix} 0 & -0.628 & -0.354 & 1.546 \\ 0.100 & 0 & 0.274 & 2.175 \\ 0.048 & 0.113 & 0 & 1.901 \\ 0.117 & 0.122 & 0.118 & 0 \end{pmatrix}. \tag{27}$$

The matrices computed based on the median and the proposed aggregated method are similar but significantly different from those computed based on the mean. As an instance, the difference between criteria c_1 and c_3 are more significant than that computed by mean (-0.311 and -0.354 for median and AWGMM, respectively, but -0.035 for mean), while the standard deviation is also tiny (0.080 and 0.048 for median and AWGMM, respectively, but 0.704 for mean), making the difference between the two criteria significantly different. By looking at the priorities provided by DMs, we can readily realize that c_3 is consistently better than c_1, except for the third DM, and the difference between the weights of these two criteria for four DMs is also significant with the average ratio of 1.42. At the same time, the third DM favors c_1 to c_3 with the factor of 3.39. This influences the average differences between the two criteria and skews the mean statistics. However, in more robust approaches, such as median and AWGMM, such priorities have less of an impact on the aggregated statistics, allowing us to capture robustly the statistical description of criteria importance. The same considerable difference exists for comparing c_1 and c_4 as well.

4.2 Statistical Comparison of Two Criteria: A Bayesian Approach

In the previous section, we show how notions of compositional data can help describe the importance of criteria in MCDM, given a number of priorities from multiple DMs. Along the same line, this section is devoted to studying the statistical tests for comparing the significance of the difference between the criteria based on various DMs' priorities.

We have three statistical tests for comparing two criteria: paired t-test (or one-sample t-test), Wilcoxon Signed-rank test, and Sign test. There is a comprehensive comparison of these tests in practical problems (Demšar 2006; Mohammadi et al. 2018), concluding that each test is appropriate in a given circumstance. However, the statistical tests based on p value have many drawbacks (Benavoli et al. 2017), making the outcome of the tests unreliable and of little practical importance. Therefore, it is highly recommended to use Bayesian tests instead of using p value inferences. The use of Bayesian statistics also allows us to make a more meaningful comparison: We can compute the extent to which one criterion is more important than another based on the priorities of a group of DMs. Such a meaningful comparison was the primary driver of some recent studies in group MCDM, which tried to compute the extent to which one criterion is more important than another by using the standard deviation of priorities (Tomashevskii and Tomashevskii 2019; Tomashevskii 2015). While these studies are fallacious due to an incorrect and improper calculation of standard deviation, we here provide a statistically-sound method based

on compositional data and Bayesian statistics. We first review two basic definitions introduced in Mohammadi and Rezaei (2020).

Definition 4.4 (Credal ordering (Mohammadi and Rezaei 2020)) For a pair of criteria c_i and c_j, the credal ordering O is defined as:

$$O = (c_i, c_j, R, d), \tag{28}$$

where

- R is the relation between the criteria c_i and c_j, i.e., $<$, $>$, or $=$;
- $d \in [0, 1]$ represents the confidence of the relation.

Definition 4.5 (Credal ranking (Mohammadi and Rezaei 2020)) For a set of criteria $C = (c_1, c_2, \ldots, c_n)$, the credal ranking is a set of credal orderings which includes all pairs (c_i, c_j), for all $c_i, c_j \in C$.

Using Bayesian statistics to compare every pair of criteria in a given problem will finally result in the credal ranking of all criteria. What is required to be computed is the confidence d for each credal ordering, that could be computed by the Bayesian counterpart of three tests: paired t-test, Sign test, and Wilcoxon Signed-rank test.

The paired t-test requires the average and the standard deviation of differences, which can be simply supplied by using the average-deviation array. One of the three arrays, i.e., mean, median, and AWGMM, can be used to conduct the paired t-test. There are a number of Bayesian counterparts for the sample paired t-test (Rouder et al. 2009; Kruschke 2013; Fox and Dimmic 2006). For the case of comparing criteria importance, the Bayesian test proposed in Rouder et al. (2009) is recommended since it takes the average, standard deviation, as well as the sample size (i.e., the number of DMs in our case) and provides the extent to which one criterion is more important than another, allowing us to experiment with the average-deviation arrays.

For the Sign test, we need to count the DMs that favor one criterion over another and then use the beta-binomial conjugate as a Bayesian test, according to which we can compute the confidence in the credal orderings.

However, developing a Wilcoxon Signed-rank test for compositional data is a bit tricky. Suppose we apply the Wilcoxon Signed-rank test directly to the priorities of multiple DMs. In that case, it accounts for the difference between the weights of two criteria for all the DMs, assigns a rank based on the magnitude of the difference, and computes the statistics. However, for compositional data like the priorities of DMs, the ratio between the weights for each DM should be considered. Therefore, instead of computing the difference between the weights of the two criteria, we should calculate the *ratio* and assign a rank based on the magnitude of the ratios. So, if two criteria are deemed the same, the ratio between their weights should be one. Instead of taking the ratios and comparing them against one, we can take the logarithm of the weights and then apply the conventional Wilcoxon Signed-rank test: If two criteria have the exact

Table 3 An example of applying the proposed Wilcoxon-type test for verifying the significance of the difference between two criteria based on the priorities of multiple DMs

	c_1	c_2	log-ratio	Rank
DM1	0.125	0.243	– 0.6650	13
DM2	0.143	0.224	– 0.4490	9
DM3	0.147	0.231	– 0.4520	10
DM4	0.164	0.209	– 0.2420	6
DM5	0.197	0.151	0.2660	7
DM6	0.157	0.256	– 0.4890	12
DM7	0.153	0.232	– 0.4160	8
DM8	0.115	0.249	– 0.7730	14
DM9	0.178	0.167	0.0640	1
DM10	0.164	0.183	– 0.1100	2
DM11	0.175	0.211	– 0.1870	5
DM12	0.168	0.192	– 0.1340	3
DM13	0.155	0.251	– 0.4820	11
DM14	0.126	0.273	– 0.7730	15
DM15	0.199	0.17	0.1580	4

weights, then the log-ratio will be zero, and the deviation of the log-ratio from zero indicates that one criterion is significantly more important than another.

For K DMs, the proposed Wilcoxon-type test contains the following steps:

Step 1. Compute the ratio r_k between the weights of the two criteria for all the DMs.
Step 2. Compute $\hat{r}_k$ by taking the logarithm of ratios r_k. Then, rank $\hat{r}_k$ according to their absolute magnitude. In the case of ties, the average rank is assigned.
Step 3. Compute R^+ as the sum of ranks of the DMs whose corresponding $\hat{r}_k$ is positive.
Step 4. Similarly, R^- is computed as the sum of ranks of DMs whose corresponding $\hat{r}_k$ is negative.
Step 6. For p value statistics, define $T = min(R^+, R^-)$. Most statistical books contain a table of exact critical values for T and K. For the Bayesian test, the ranks are given to the Bayesian Wilcoxon test (Benavoli et al. 2014) and compute the extent to which one criterion is more important than another.

We explain the procedure of the Wilcoxon-type problem using an example. Table 3 shows the weights of two criteria obtained based on the preferences of 15 DMs. Each row in this table corresponds to each DM. In the last two columns, we show the log ratios and the associated rank of each DM based on the Wilcoxon Signed-rank test. Accordingly, R^+ and R^- are calculated as:

$$\begin{aligned} R^+ &= 7 + 1 + 4 = 12 \\ R^- &= 108. \end{aligned} \tag{29}$$

Given R^+ and R^-, we can then apply the Bayesian Wilcoxon Signed-rank test to compute the extent to which one criterion is more important than another based on the priorities of a group of DMs.

5 Distance Metrics and Clustering Methods

In processing multiple DM priorities, the Euclidean distance is widely used, where it directly computes the distance based on the original priorities. In compositional data analysis, however, the distance between two compositions is defined differently, taking the nature of the compositions into account. In this section, the Aitchison distance (Aitchison et al. 2000), which is a proper and arguably the most popular distance metric for compositional data, is reviewed, according to which a clustering method is developed to cluster multiple DMs based on their priorities.

5.1 Distance Metrics for Priorities in Group MCDM

The Euclidean distance is typically used as the distance measure between two priorities. Let $w, w' \in \mathcal{R}^n$ be the priorities of two DMs, the Euclidean distance is defined as:

$$d_e(w, w') = \sqrt{\sum_{i=1}^{n} \left(w_i - w'_i\right)^2}. \tag{30}$$

The Euclidean distance is based on the Euclidean space and computes the distance based on the original priorities that lie on a simplex (not Euclidean space). Therefore, a proper distance metric should be used to measure the distance more in line with the compositional nature of the priorities. The Aitchison distance is arguably the most popular distance metric in compositional data analysis, which also correlates with the Euclidean distance. The Aitchison distance is defined as the Euclidean distance of the log-ratio transformed data (Aitchison et al. 2000), i.e.,

$$d_a(w, w') = \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\ln \frac{w_i}{w_j} - \ln \frac{w'_i}{w'_j} \right)^2}. \tag{31}$$

Similarly, the mean absolute deviation (MAD) distance for the compositional data, shown by *MADC*, can be defined as:

$$d_{MADC}(w, w') = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| \ln \frac{w_i}{w_j} - \ln \frac{w'_i}{w'_j} \right|. \tag{32}$$

As a result, the distance metrics in (31) or (32) should be used for clustering the DMs based on their priorities, which respect the compositional nature of the priorities.

5.2 K-Means Clustering for Priorities

Grouping DMs into a number of clusters is also used in group MCDM (Abel et al. 2014; Meixner et al. 2016). One way is to group them based on their priorities using clustering methods, the core building block of which is a distance function. Using an improper distance function would group the DMs on a wrong basis. On top of that, the centroids of clusters would not necessarily satisfy the unit-sum constraint, thereby failing to represent the priorities properly. To prevent these pitfalls, we now extend the K-means clustering algorithm that uses a compositional distance metric, e.g., distance metrics in (31) or (32), to group the DMs based on their priorities.

The K-means needs to know the number of clusters, o, and identifies o centroids of clusters $l_1, \ldots, l_o$. The standard K-means uses the Euclidean distance and the arithmetic mean for clustering. However, a compositional distance metric and normalized geometric mean should be utilized for clustering compositional data.

The steps required to cluster compositional data $W_i, i = 1, \ldots, K$ by using K-means are as follows:

Step 1 Place centroids $l_1, \ldots, l_o$ at random locations.
Step 2 Until convergence, repeat the following steps:

Step 2.1 For each W_i, find the nearest centred l_j as:

$$j = arg \min_j d_a(w_i, l_j), \tag{33}$$

where $d_e(.,.)$ is a compositional distance. Then, assign W_i to cluster j.

Step 2.2 For each cluster $j = 1, \ldots, o$, update the centroids as the mean of the points within the cluster, i.e.,

$$l_j = \frac{\prod_{k=1}^{K_j} W_{k.}^{\frac{1}{K_j}}}{e^T \left(\prod_{k=1}^{K_j} W_{k.}^{\frac{1}{K_j}} \right)}, \tag{34}$$

where K_j is the number of priorities in cluster j, $\prod$ and power are element-wise operations, and e is a vector with elements of one.

Since the distance function and the average are different from those of the standard K-means, the outcome of clustering will be distinct as well. In addition, the centroids identified by the proposed clustering method would be compositional, while the centroids of standard K-means for clustering compositional data are not necessarily compositional, e.g., the centroids are not compositional for the compositional mean absolute deviation.

6 Numerical Results

In this section, we show the correct ways of analyzing the priorities of multiple DMs through a real case study in airline baggage handling. The baggage handling system is essential to ground handling operations and impacts passengers' satisfaction. A recent study develops a model, namely SERVQUAL, for assessing the quality of service for baggage handling (Rezaei et al. 2018). Grounded on the literature review and the interviews with the passengers, the SERVQUAL includes five main criteria for evaluating the quality of the airline baggage handling systems: *tangibles*, *reliability*, *responsiveness*, *assurance*, and *empathy*. The estimate the importance of the criteria, the preferences of 148 passengers from several nationalities are elicited according to the best–worst method (Mohammadi and Rezaei 2020). The analyses conducted in this section are based on the priorities of the 148 participants. The data and MATLAB implementation of the corresponding analyses are publicly available.[3]

6.1 Aggregating Priorities

We first look into the aggregation of priorities by different methods. We first compute by the AMM (which is not correct but is typically used in the literature), and the result is:

$$w^g_{AMM} = \begin{bmatrix} 0.1397 & 0.3459 & 0.2289 & 0.1519 & 0.1336 \end{bmatrix}. \tag{35}$$

Similarly, the results of the GMM and AWGMM are as follows:

$$\begin{aligned} w^g_{GMM} &= \begin{bmatrix} 0.1376 & 0.3502 & 0.2347 & 0.1527 & 0.1248 \end{bmatrix}, \\ w^g_{AWGMM} &= \begin{bmatrix} 0.1234 & 0.4462 & 0.1932 & 0.1490 & 0.0883 \end{bmatrix}. \end{aligned} \tag{36}$$

The AMM has a different aggregation than the GMM, so the result of such aggregation can distort the follow-up decisions based on the aggregated weights. Also, the AWGMM has a different aggregation than GMM, especially with respect to the weights of *reliability*, *responsiveness*, and *empathy*. In particular, the weights of *empathy* and *responsiveness* are much less in the AWGMM, while the weight of *reliability* is more significant. To inspect this difference, we looked into the participants' priorities and realized that 7 out of 140 participants had been assigned an infinitesimal weight. By looking into the priorities of these participants, we realize that most of them have either assigned significantly higher weights to *empathy* (mainly more than 0.50) or *tangibles* and instead a much lower weight to *reliability*. Hence, since these participants are deemed deviants and assigned a lower weight, they have a lesser impact on the final aggregated priorities. As a result, the AWGMM aggregated priorities have a higher weight for *reliability* and lower weights for *tangibles* and *empathy* in comparison to the GMM.

[3] https://github.com/Majeed7/MCDMfallacies_compositional.

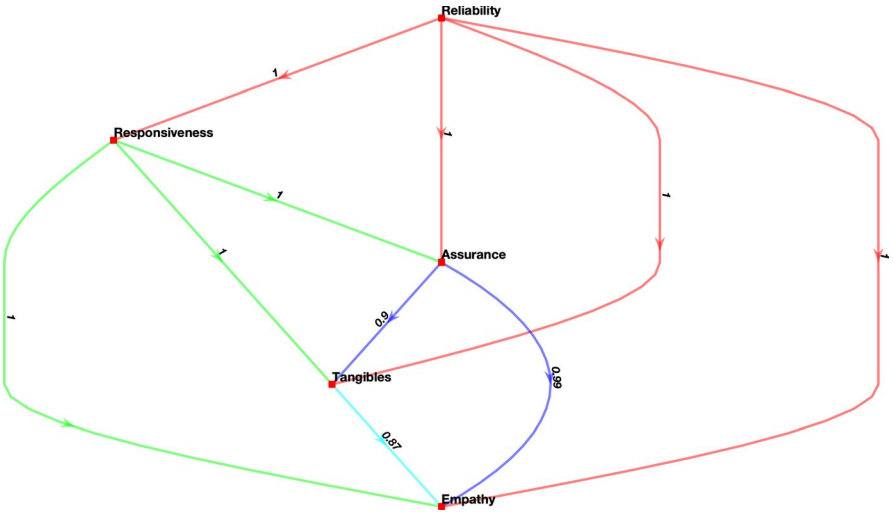

Fig. 1 The credal ranking of criteria for assessing the service quality of airline baggage handling system

Table 4 The center of three clusters identified by K-means and K-means for compositional data

	Tangibles	Reliability	Responsiveness	Assurance	Empathy	Sum
K-means	0.1211	0.5019	0.1402	0.1302	0.0651	0.9584
	0.1085	0.1434	0.4957	0.1185	0.1142	0.9803
	0.1092	0.1545	0.1549	0.1806	0.1765	0.7757
Compositional K-means	0.1524	0.1524	0.4653	0.0773	0.1524	1
	0.1375	0.5325	0.1375	0.1375	0.0550	1
	0.0726	0.3224	0.1921	0.2005	0.2123	1

6.2 Credal Ranking

We now compute the credal ranking of criteria based on the Bayesian Wilcoxon Signed-rank test. We specifically use the Bayesian Wilcoxon Signed-rank test because it entails fewer assumptions on the input data and is typically used in MCDM. To better summarize the credal ranking of criteria, we visualize it using a weighted, directed graph. The nodes in the graph are the criteria, and each directed arc shows that the criterion in origin is much more important than that at the other end by a confidence level specified by the weight of the associated arc. Each arc in the graph visualizes a credal ordering of two criteria, and the whole graph visualizes the credal ranking of all criteria.

Figure 1 shows the credal ranking of five criteria for the baggage handling systems. According to this graph, *reliability* is by far the most important criterion, followed by *responsiveness*. *Assurance* is the third important criterion, and it is more important than *tangible* and *empathy* with confidence levels of 0.90 and 0.99,

respectively. Also, *tangible* is the fourth criterion and is more important than *empathy* with a confidence of 0.87.

6.3 Clustering the Participants

We now show how clustering can group the participants into multiple mutually exclusive groups. For doing so, we apply K-means for compositional data discussed in Sect. 5, and compare with conventional K-means. Also, the number of clusters is set to three (Rezaei et al. 2018). We use the mean absolute deviation and its compositional version for clustering to highlight the advantages of the compositional distance metrics for grouping the DMs based on their priorities.

We first compare the outcome of clustering methods based on their centroids. Table 4 shows the centroids of clusters identified by the clustering methods. An essential difference between the methods is the difference between the sum of centroids: The centroids of K-means do not sum up to one, while those of the compositional K-means will add up to one, satisfying the unit-sum constraint required for the priorities of criteria. As a result, the centers of clusters provided by the proposed method are better representatives of the different groups of participants.

We also compare the clusters of the participants assigned by different clustering methods. The two clustering methods group the participants differently: By repeating the clustering methods multiple times, the two methods assign around 50 participants into different clusters. Thus, without considering the compositional nature, clustering would group the participants differently. While we cannot compare the participants in the clusters, we should note that the clustering based on the conventional distance metric (without considering the compositional nature) is theoretically incorrect, so we should favor the clusters provided by the proposed clustering algorithm that is specially tailored for the compositional data.

7 Conclusion and Discussion

In this paper, we studied three different errors in processing the priorities of multiple decision-makers (DMs) in group decision-making problems, and correct ways for processing the priorities were introduced. The first error discussed in this article was the aggregation of priorities and showed that the compositional analysis for aggregation would result in the normalized geometric mean of priorities. An essential by-product is that the use of the arithmetic mean of priorities should be avoided.

We also discussed the error regarding the computation of the standard deviation of weights and proposed using Bayesian statistics to provide a probabilistic ranking of criteria, called credal ranking. The credal ranking gives the extent to which a group of DMs prefers one criterion over another, computed based on the Bayesian Wilcoxon Signed-rank test. We finally explained a proper distance metric for gauging the distance between two priorities, according to which we modified K-means clustering for grouping the DMs based on their priorities.

The findings of this study also have implications for other statistical tools when we use them for priorities [e.g., analysis of variance (ANOVA)]. Generally speaking, wherever we need to do computations with priorities, we need to consider their compositional nature.

In the future, more errors in MCDM should be studied from a compositional data perspective. A crucial case is when the performance matrix of alternatives is also created by some MCDM methods, each column resulting in a composition. Then, the compositions from different criteria should be merged into global priorities by considering the importance of criteria shown by another composition. Considering the compositional nature of the data might allow us to extend those methods to be normalization-agnostic and obviate the rank reversal phenomenon.

Funding The authors have no relevant financial or non-financial interests to disclose.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

Abel E, Mikhailov L, Keane J (2014) Clustering decision makers with respect to similarity of views. In: 2014 IEEE symposium on computational intelligence in multi-criteria decision-making (MCDM). IEEE, pp 40–47

Abel E, Mikhailov L, Keane J (2015) Group aggregation of pairwise comparisons using multi-objective optimization. Inf Sci 322:257–275

Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc Ser B (Methodol) 44:139–160

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32:271–275

Altuzarra A, Moreno-Jiménez JM, Salvador M (2007) A Bayesian priorization procedure for AHP-group decision making. Eur J Oper Res 182:367–382

Amenta P, Ishizaka A, Lucadamo A, Marcarelli G, Vyas V (2020) Computing a common preference vector in a complex multi-actor and multi-group decision system in analytic hierarchy process context. Ann Oper Res 284:33–62

Belton V, Pictet J (1997) A framework for group decision using a MCDA model: Sharing, aggregating or comparing individual information? J Decis Syst 6:283–303

Belton V, Stewart T (2002) Multiple criteria decision analysis: an integrated approach. Springer, Berlin

Benavoli A, Corani G, Mangili F, Zaffalon M, Ruggeri F (2014) A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In: International conference on machine learning, pp 1026–1034

Benavoli A, Corani G, Demšar J, Zaffalon M (2017) Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J Mach Learn Res 18:2653–2688

Blagojevic B, Srdjevic B, Srdjevic Z, Zoranovic T (2016) Heuristic aggregation of individual judgments in AHP group decision making using simulated annealing algorithm. Inf Sci 330:260–273
Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (2006) Compositional data analysis in the geosciences: from theory to practice. Geological Society of London, London
Chiclana F, GarcíA JT, del Moral MJ, Herrera-Viedma E (2013) A statistical comparative study of different similarity measures of consensus in group decision making. Inf Sci 221:110–123
Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
Dias LC, Clımaco JN (2005) Dealing with imprecise information in group multicriteria decisions: a methodology and a GDSS architecture. Eur J Oper Res 160:291–307
Edwards W (1977) 12 use of multiattribute utility measurement for social decision making. Conflicting 247:326–340
Forman E, Peniwati K (1998) Aggregating individual judgments and priorities with the analytic hierarchy process. Eur J Oper Res 108:165–169
Fox RJ, Dimmic MW (2006) A two-sample Bayesian t-test for microarray data. BMC Bioinform 7:126
Geman D, Reynolds G (1992) Constrained restoration and the recovery of discontinuities. IEEE Trans Pattern Anal Mach Intell 14:367–383
He R, Zheng WS, Hu BG (2010) Maximum correntropy criterion for robust face recognition. IEEE Trans Pattern Anal Mach Intell 33:1561–1576
Huber PJ (2004) Robust statistics, vol 523. Wiley, Hoboken
Ishizaka A, Labib A (2011) Review of the main developments in the analytic hierarchy process. Expert Syst Appl 38:14336–14345
Keeney RL, Raiffa H et al (1976) Decisions with multiple objectives: preferences and value trade-offs. Wiley, Hoboken
Kruschke JK (2013) Bayesian estimation supersedes the t test. J Exp Psychol Gen 142:573
McCartt AT, Rohrbaugh J (1989) Evaluating group decision support system effectiveness: a performance study of decision conferencing. Decis Support Syst 5:243–253
Meixner O, Haas R, Pöchtrager S (2016) AHP group decision making and clustering. In: International symposium on the analytic hierarchy process (ISAHP). http://www.isahp.org/uploads/paper_mo_hr_isahp2016rev_001.pdf. Accessed Sept 2016
Mohammadi M, Hodtani GA, Yassi M (2015) A robust correntropy-based method for analyzing multisample ACGH data. Genomics 106:257–264
Mohammadi M, Noghabi HS, Hodtani GA, Mashhadi HR (2016) Robust and stable gene selection via maximum–minimum correntropy criterion. Genomics 107:83–87
Mohammadi M, Hofman W, Tan YH (2018) A comparative study of ontology matching systems via inferential statistics. IEEE Trans Knowl Data Eng 31:615–628
Mohammadi M, Rezaei J (2020) Bayesian best–worst method: a probabilistic group decision making model. Omega 96:102075
Mustajoki J, Hämäläinen RP, Salo A (2005) Decision support by interval smart/swing-incorporating imprecision in the smart and swing methods. Decis Sci 36:317–339
Pawlowsky-Glahn V, Buccianti A (2011) Compositional data analysis. Wiley, Hoboken
Pearson K (1897) Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 60:489–498
Phillips LD (1990) Decision analysis for group decision support. In: Eden C, Radford J (eds) Tackling strategic problems: the role of group decision support. Sage Publications, London (**ISBN 9780803982604**)
Ramanathan R, Ganesh L (1994) Group preference aggregation methods employed in AHP: an evaluation and an intrinsic process for deriving members' weightages. Eur J Oper Res 79:249–265
Rezaei J (2015) Best–worst multi-criteria decision-making method. Omega 53:49–57
Rezaei J, Kothadiya O, Tavasszy L, Kroesen M (2018) Quality assessment of airline baggage handling systems using SERVQUAL and BWM. Tour Manag 66:85–93
Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 16:225–237
Saaty TL (1977) A scaling method for priorities in hierarchical structures. J Math Psychol 15:234–281
Saaty TL (1990) Decision making for leaders: the analytic hierarchy process for decisions in a complex world. RWS Publications, Pittsburgh
Saaty TL (2000) Fundamentals of decision making and priority theory with the analytic hierarchy process, vol 6. RWS Publications, Pittsburgh
Tomashevskii I (2015) Eigenvector ranking method as a measuring tool: formulas for errors. Eur J Oper Res 240:774–780

Tomashevskii I, Tomashevskii D (2019) A non-heuristic multicriteria decision-making method with verifiable accuracy and reliability. J Oper Res Soc 72:1–15
Triantaphyllou E (2000) Multi-criteria decision making methods: a comparative study. Kluwer Academic Publishers (now Springer), Dordrecht