



# Estimating Latent Linear Correlations from Fuzzy Frequency Tables

Antonio Calcagni<sup>1,2</sup> 

Received: 6 November 2021 / Revised: 27 December 2021 / Accepted: 15 March 2022  
© The Author(s) 2022

## Abstract

This research concerns the estimation of latent linear or polychoric correlations from fuzzy frequency tables. Fuzzy counts are of particular interest to many disciplines including social and behavioral sciences and are especially relevant when observed data are classified using fuzzy categories—as for socioeconomic studies, clinical evaluations, content analysis, inter-rater reliability analysis—or when imprecise observations are classified into either precise or imprecise categories—as for the analysis of ratings data or fuzzy-coded variables. In these cases, the space of count matrices is no longer defined over naturals and, consequently, the polychoric estimator cannot be used to accurately estimate latent linear correlations. The aim of this contribution is twofold. First, we illustrate a computational procedure based on generalized natural numbers for computing fuzzy frequencies. Second, we reformulate the problem of estimating latent linear correlations from fuzzy counts in the context of expectation–maximization-based maximum likelihood estimation. A simulation study and two applications are used to investigate the characteristics of the proposed method. Overall, the results show that the fuzzy EM-based polychoric estimator is more efficient to deal with imprecise count data as opposed to standard polychoric estimators that may be used in this context.

**Keywords** Fuzzy frequency · Generalized natural numbers · Polychoric correlations · Fuzzy data analysis

**Mathematics Subject Classification** 60A86 · 62H17 · 62F86 · 62-08 · 62P25

---

✉ Antonio Calcagni  
antonio.calcagni@unipd.it

<sup>1</sup> DPSS, University of Padova, Via Venezia 8, 35031 Padua, Italy

<sup>2</sup> Member of the GNCS Research Group, National Institute of Advanced Mathematics (INdAM), P.le Aldo Moro 5, 00185 Roma, Italy

## 1 Introduction

The latent linear correlation (LLC) is a measure of bivariate association which is usually adopted when variables are measured at an ordinal level or when data are available in the form of frequency or contingency tables. Because LLCs are quite often used in analyzing categorical ordered variables, they are also known as polychoric correlations [59]. Latent linear or polychoric correlations differ from other measures of association such as Goodman and Kruskal's  $\gamma$  or Kendall's  $\tau$  in that they are based on a latent continuous parametric model according to which LLCs behave. Given a set of  $J$  variables, LLCs are computed pairwise for each pair  $j, k$  of variables by considering their joint frequencies  $\mathbf{N}_{R \times C}^{jk} = (n_{11}^{jk}, \dots, n_{rc}^{jk}, \dots, n_{RC}^{jk})$  over a  $R^{jk} \times C^{jk}$  partition space of the variables' domain. The general idea is to map the observed counts  $\mathbf{N}_{R \times C}^{jk}$  to the real domain of the bivariate latent density model via the Muthen's thresholds-based approach [57], under the constraint that the volumes of the rectangles of the latent density should be equal to the observed frequencies. In doing so, changing the covariance parameter of the latent model will change the probability distribution over the latent rectangles and hence the probability masses over the cells of  $\mathbf{N}_{R \times C}^{jk}$ . Although several parametric models are available for estimating LLCs (e.g., elliptical, skew-Gaussian, Copula-based models; see [47, 63, 67]), the standard formulation based on the Gaussian density with zero means and latent correlations  $\mathbf{R}_{R \times C}^{jk}$  is strong enough to be of practical use for many empirical applications. (For a recent study, see [39, 56].) Some of these include inter-rater agreement [60], reliability measurement [11, 70, 81], ordinal CFA and SEM [50, 58, 78], fuzzy cluster analysis [73], and polychoric PCA for dimensionality reduction in discrete data [45].

Fuzzy frequency or contingency tables are of particular concern across several disciplines including social, behavioral, and health sciences. Overall, there are two main situations which give rise to fuzzy frequencies, namely when precise data are classified into imprecise categories or, in the opposite case, when fuzzy data are classified into either precise or imprecise categories. Examples of the first case may be found in studies involving socioeconomic variables (e.g., income, labor flushes, employment) [19, 77], images or scenes classification [26, 38], content analysis [43], reliability analyses [23], evaluation of user-based experiences [46], multivariate analysis of qualitative data [3, 8], spatial distributional data [31], and human-based risk assessment [20]. By contrast, examples of the second case are most common in studies involving rating scales-based variables such as satisfaction, quality, attitudes, and motivation [12, 21]. What both of these situations have in common is that the  $R^{jk} \times C^{jk}$  space constitutes a fuzzy partition and, consequently, observed counts in the classification grid are no longer natural numbers. There have been a number of studies that have tried to deal with fuzzy contingency tables and fuzzy association measures. For instance, Kahraman et al. [42] proposed some nonparametric tests generalized to the case of fuzzy data, Grzegorzewski [32] studied fuzzy hypotheses testing based on fuzzy random variables, Denceux [24] proposed a rank-sum test based on fuzzy partial ordering and introduced a modelization of fuzzy statistical significance test, Hryniewicz [36] generalized the Goodman and Kruskal's  $\gamma$  measure to the case of fuzzy observations arranged into contingency tables, and Taheri et al. [69] presented the analysis

of contingency tables for both fuzzy observations/crisp categories and crisp observations/fuzzy categories cases along with a fuzzy generalization of association measures based on frequencies. Although they differ in some respects, all of them generalize the analysis of contingency tables to the fuzzy case either by the Zadeh’s extension principle or by the  $\alpha$ -cuts-based calculus [72]. Fuzzy statistics aside, a more recent strategy to incorporate imprecision and indeterminacy in count data is that of using a neutrosophic-based generalization of the standard chi-square and F-statistics [4, 5].

Based on this research stream, this article focuses on estimating latent linear correlations from fuzzy frequency tables, which include both the cases of crisp observations/fuzzy categories and fuzzy observations/fuzzy or crisp categories. Unlike the aforementioned studies, we develop our results by generalizing the standard LLC problem to cope with fuzzy frequencies under the general fuzzy maximum likelihood framework [24, 62]. In particular, we define the fuzzy frequency table  $\tilde{N}_{R \times C}^{jk}$  in terms of fuzzy cardinality and generalized natural numbers first, and then we extend the sample space of the LLC model to deal with fuzzy counts  $\tilde{n}_{11}^{jk}, \dots, \tilde{n}_{rc}^{jk}, \dots, \tilde{n}_{RC}^{jk}$ . In doing so, the fuzziness of the observations enters the model as a systematic and non-random component, while the model’s parameters are still crisp (i.e., the estimated latent correlation matrix  $\hat{R}_{R \times C}^{jk}$  is a non-fuzzy quantity). This offers an attractive solution to the problem of estimating LLCs with fuzzy information, with the additional benefit that statistical models that use the LLCs statistic as input data (e.g., CFA, PCA, SEM) do not need any further generalization to cope with fuzzy data.

The remainder of this article is structured as follows. Section 2 introduces the concept of fuzzy frequency through fuzzy cardinalities and generalized natural numbers. Section 3 describes the fuzzy LLCs model and its characteristics in terms of parameters estimation and interpretation. Section 4 reports the results of a simulation study performed to assess the finite sample properties of the fuzzy LLCs model as compared with standard defuzzification-based estimation methods. Section 5 describes the application of the proposed method to two real case studies, and Sect. 6 concludes the article by providing final remarks and suggestions for further extensions of the current findings. All the materials like algorithms and datasets used throughout the article are available to download at <https://github.com/antcalcagni/fuzzypolychoric/>.

## 2 Fuzzy Frequencies

### 2.1 Preliminaries

A fuzzy subset  $\tilde{A}$  of a universal set  $\mathcal{A} \subset \mathbb{R}$  can be defined by means of its characteristic function  $\xi_{\tilde{A}} : \mathcal{A} \rightarrow [0, 1]$ . It can also be expressed as a collection of crisp subsets called  $\alpha$ -sets, i.e.,  $\tilde{A}_\alpha = \{x \in \mathcal{A} : \xi_{\tilde{A}}(x) > \alpha\}$  with  $\alpha \in (0, 1]$ . If the  $\alpha$ -sets of  $\tilde{A}$  are all convex sets, then  $\tilde{A}$  is a convex fuzzy set. The support of  $\tilde{A}$  is  $A_0 = \{x \in \mathcal{A} : \xi_{\tilde{A}}(x) > 0\}$  and the core is the set of all its maximal points  $A_1 = \{x \in \mathcal{A} : \xi_{\tilde{A}}(x) = \max_{z \in \mathcal{A}} \xi_{\tilde{A}}(z)\}$ . In case  $\max_{x \in \mathcal{A}} \xi_{\tilde{A}}(x) = 1$ , then  $\tilde{A}$  is a normal fuzzy set. If  $\tilde{A}$  is a normal and convex subset of  $\mathbb{R}$ , then  $\tilde{A}$  is a fuzzy number (also called fuzzy interval). The quantity  $l(\tilde{A}) = \sup A_0 - \inf A_0$  is the length of the

support of the fuzzy set  $\tilde{A}$ . The simple cardinality of a fuzzy set  $\tilde{A}$  is defined as  $|\tilde{A}| = \int_{\mathcal{A}} \xi_{\tilde{A}}(x) dx$ . Given two fuzzy sets  $\tilde{A}, \tilde{B}$ , the degree of inclusion of  $\tilde{A}$  in  $\tilde{B}$  is  $\epsilon_{\tilde{A}\tilde{B}} = \left| \min_{x \in \mathcal{A}} (\xi_{\tilde{A}}(x), \xi_{\tilde{B}}(x)) \right| / \max(1, |\tilde{A}|)$ , with  $\epsilon_{\tilde{A}\tilde{B}} \in [0, 1]$ . The case  $\epsilon_{\tilde{A}\tilde{B}} = 1$  indicates that  $\tilde{A}$  is fully included in  $\tilde{B}$ . The class of all normal fuzzy numbers is denoted by  $\mathcal{F}(\mathbb{R})$ . Fuzzy numbers can conveniently be represented using parametric models that are indexed by some scalars. These include a number of shapes like triangular, trapezoidal, Gaussian, and exponential fuzzy sets [34]. A relevant class of parametric fuzzy numbers are the so-called LR-fuzzy numbers [27] and their generalizations [13, 70]. The trapezoidal fuzzy number is one of the most common fuzzy set used in many applications, and it is parameterized using four parameters as follows:

$$\xi_{\tilde{A}}(x) = \mathbb{1}_{(c_1, c_2)}(x) + \left( \frac{x - x_l}{c_1 - x_l} \right) \mathbb{1}_{(x_l, c_1)}(x) + \left( \frac{x_u - x}{x_u - c_2} \right) \mathbb{1}_{(c_2, x_u)}(x) \quad (2.1)$$

with  $x_l, x_u, c_1, c_2 \in \mathbb{R}$  being lower, upper bounds, and first and second modes, respectively. The symbol  $\mathbb{1}_{(a,b)}(x)$  denotes the indicator function in the interval  $(a, b)$ . Interestingly, the trapezoidal fuzzy set includes the triangular (if  $c_1 = c_2$ ) and rectangular (if  $x_l = c_1, c_2 = x_u$ ) fuzzy sets as special cases. A degenerated fuzzy number  $\tilde{A}$  is a particular fuzzy set with  $\xi_{\tilde{A}}(c) = 1$  and  $\xi_{\tilde{A}}(x) = 0$  for  $x \neq c, x \in \mathcal{A}$ . Note that rectangular and degenerated fuzzy numbers can be adopted to represent crisp categories and crisp observations, respectively. When a probability space is defined over  $\mathcal{A}$ , the probability of a fuzzy set  $\tilde{A}$  can be defined as  $\mathbb{P}(\tilde{A}) = \int_{\mathcal{A}} \xi_{\tilde{A}}(x) d\mathbb{P}$  (Zadeh’s probability). Similarly, the joint probability of two fuzzy sets is  $\mathbb{P}(\tilde{A}\tilde{B}) = \int_{\mathcal{A}} \xi_{\tilde{A}}(x)\xi_{\tilde{B}}(x) d\mathbb{P}$  under the rule  $\xi_{\tilde{A}\tilde{B}}(x) = \xi_{\tilde{A}}(x)\xi_{\tilde{B}}(x)$  (independent fuzzy sets) [52].

### 2.2 Fuzzy Granules

Let  $\mathcal{S} = \{\tilde{A}_1, \dots, \tilde{A}_i, \dots, \tilde{A}_I\}$  be a sample of  $I$  fuzzy or non-precise observations with  $\tilde{A}_i$  being a fuzzy number as defined by Eq. (2.1). Then, the interval  $\mathcal{R}(\mathcal{S}) = [r_0, r_1] \subset \mathbb{R}$  is the range of the fuzzy sample where  $r_0 = \min\{A_{0_1}^\dagger, \dots, A_{0_I}^\dagger\}$  and  $r_1 = \max\{A_{0_1}^\dagger, \dots, A_{0_I}^\dagger\}$ , with  $A_{0_i}^\dagger$  being the infimum of the support set  $A_{0_i}$  computed for the  $i$ th fuzzy observation. A collection  $\mathcal{G} = \{\tilde{G}_1, \dots, \tilde{G}_c, \dots, \tilde{G}_C\}$  of  $C$  fuzzy sets is a fuzzy partition of  $\mathcal{R}(\mathcal{S})$  if the following two properties hold (i)  $\max_{i=1, \dots, I} l(\tilde{A}_i) \leq \min_{c=1, \dots, C} l(\tilde{G}_c)$  and (ii)  $\sum_{c=1}^C \xi_{\tilde{G}_c}(x) = 1$  (Ruspini’s partition) [9, 29]. The fuzzy sets in  $\mathcal{G}$  are also called granules of  $\mathcal{R}(\mathcal{S})$ . The evaluation of the amount of fuzzy observations in a granule  $\tilde{G}_c$  is called cardinality (scalar or fuzzy) and can be used to compute fuzzy frequencies or counts for a partition  $\mathcal{G}$  given a sample  $\mathcal{S}$ . Figure 1 (left-side panels) shows an example of fuzzy granulation for both fuzzy and crisp observations.

### 2.3 Fuzzy Counts as Generalized Natural Numbers

Let  $\tilde{\mathbf{x}}_j = \{\tilde{x}_1^j, \dots, \tilde{x}_i^j, \dots, \tilde{x}_I^j\}$  and  $\tilde{\mathbf{x}}_k = \{\tilde{x}_1^k, \dots, \tilde{x}_i^k, \dots, \tilde{x}_I^k\}$  be two samples of fuzzy observations and  $\tilde{\mathbf{g}}_j = \{\tilde{g}_1^j, \dots, \tilde{g}_c^j, \dots, \tilde{g}_C^j\}$  and  $\tilde{\mathbf{g}}_k = \{\tilde{g}_1^k, \dots, \tilde{g}_r^k, \dots, \tilde{g}_R^k\}$

be two fuzzy partitions of the domains  $\mathcal{R}(\tilde{\mathbf{x}}_j)$  and  $\mathcal{R}(\tilde{\mathbf{x}}_k)$ . Given a pair of granule  $(\tilde{g}_r, \tilde{g}_c)$ , a fuzzy or imprecise count for the joint sample  $(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k)$  is a fuzzy set  $\tilde{n}_{rc}^{jk}$  with membership function  $\xi_{\tilde{n}_{rc}^{jk}} : \mathbb{N}_0 \rightarrow [0, 1]$ . As it is defined over natural numbers, a fuzzy count is a finite generalized natural number for which extended operations are available (e.g., addition, multiplication) [75]. Analogously to fuzzy intervals, the class of all fuzzy counts is denoted as  $\mathcal{F}(\mathbb{N}_0)$ . There are different choices for the computation of  $\xi_{\tilde{n}_{rc}^{jk}}$  (e.g., see [22, 36, 69, 71, 72]). In this contribution, we will follow the findings of [9, 10] which are based on Zadeh’s fuzzy counting functions [79] and fuzzy cardinalities [14]. More precisely, let

$$\epsilon_{rc}^{jk} = \left( \epsilon_{rc1}^{jk}, \dots, \epsilon_{rci}^{jk}, \dots, \epsilon_{rcI}^{jk} \right)$$

be the vector of joint degrees of inclusion for the  $r$ th granule where

$$\begin{aligned} \epsilon_{rci}^{jk} &= \min \left( \epsilon_{r_i}^j, \epsilon_{c_i}^k \right), \\ \epsilon_{r_i}^j &= \left| \min_{x \in \mathcal{R}(\tilde{\mathbf{x}}_j)} (\xi_{\tilde{x}_{ji}}(x), \xi_{\tilde{g}_r}(x)) \right| / \max(1, |\tilde{x}_i^j|), \\ \epsilon_{c_i}^k &= \left| \min_{x \in \mathcal{R}(\tilde{\mathbf{x}}_k)} (\xi_{\tilde{x}_{ki}}(x), \xi_{\tilde{g}_c}(x)) \right| / \max(1, |\tilde{x}_i^k|), \end{aligned}$$

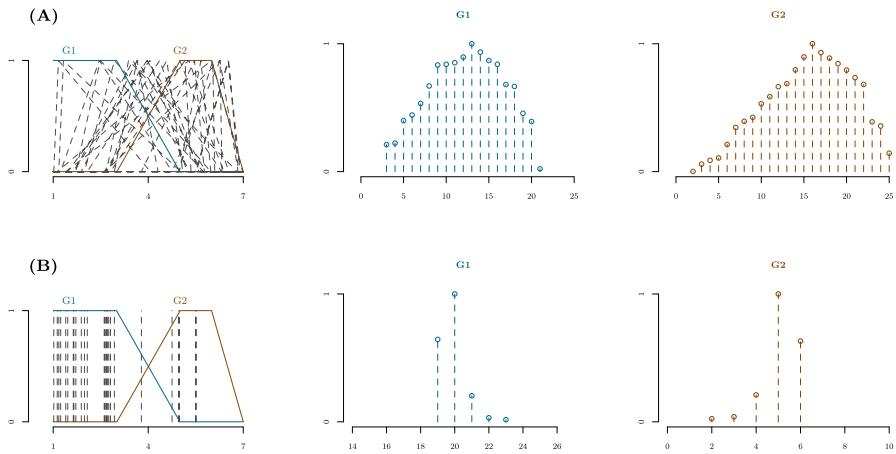
with  $|\cdot|$  being the simple cardinality according to the definition given in Sect. 2.1. For  $n \in \mathbb{N}_0$ , the fuzzy count is as follows:

$$\xi_{\tilde{n}_{rc}^{jk}}(n) = \min (\mu_{\text{FLC}}(n), \mu_{\text{FGC}}(n)) \tag{2.2}$$

with  $\mu_{\text{FLC}}(n)$  and  $\mu_{\text{FGC}}(n)$  being the output of the Zadeh’s fuzzy counting functions [79]. The following calculus can be used for  $\mu_{\text{FLC}}(n)$  and  $\mu_{\text{FGC}}(n)$ . First, compute the square matrix of differences  $\mathbf{Z}_{I \times I} = \left( \epsilon_{rc}^{jk} \mathbf{1}_I^T - \mathbf{1}_I (\epsilon_{rc}^{jk})^T \right)$ , with  $\mathbf{1}_I$  being an  $I \times 1$ , vector of all ones. Then, for each  $i = 1, \dots, I$  the vector  $\mathbf{z}_{I \times 1}$  is computed, with  $z_i = \mathbf{1}_I^T \mathcal{H}(\mathbf{Z},_i)$  and  $\mathcal{H}(x)$  being the Heaviside step function defined by  $\mathcal{H}(x) := \{0 \text{ if } x < 0, 1 \text{ if } x \geq 0\}$ . The vector  $\mathbf{z} = (z_1, \dots, z_i, \dots, z_I)$  contains the sums of the output of the Heaviside function applied column-wise on  $\mathbf{Z}$ . Finally, for  $n = 0, 1, 2, \dots, I$  the Zadeh’s counting functions are as follows:

$$\begin{aligned} \mu_{\text{FGC}}(n) &= \max \left( \mathcal{H}(\mathbf{z} - n) \odot \epsilon_{rc}^{jk} \right), \\ \mu_{\text{FLC}}(n) &= 1 - \max \left( \mathcal{H}(\mathbf{z} - n + 1) \odot \epsilon_{rc}^{jk} \right), \end{aligned} \tag{2.3}$$

where  $\odot$  is the element-wise product, whereas  $\mathcal{H}(x)$  is the Heaviside function defined as above. Thus, the membership function of  $\tilde{n}_{rc}^{jk}$  is defined as the minimum among the degree of possibility that at least  $n$  elements from  $(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k)$  are included in the  $r$ th granule (FGC count) and the degree of possibility that at most  $n$  elements are included in the  $r$ th granule (FLC count). By applying Eqs. (2.2) and (2.3) for each pair of granules  $(\tilde{g}_1, \tilde{g}_1), \dots, (\tilde{g}_R, \tilde{g}_C), \dots, (\tilde{g}_R, \tilde{g}_C)$ , one obtains the fuzzy frequency matrix



**Fig. 1** Examples of fuzzy granules and fuzzy counts for **A** fuzzy triangular observations and fuzzy trapezoidal categories and **B** crisp observations and fuzzy trapezoidal categories. Note that in both cases frequencies are represented as generalized natural numbers

$\tilde{N}_{R \times C}^{jk}$ . Note that the resulting fuzzy set  $\xi_{\tilde{n}_{rc}}$  may not be normal, i.e.,  $\max_n \xi_{\tilde{n}_{rc}}(n) \leq 1$ , and a post hoc normalization should be applied if normal fuzzy sets were needed.

Finally, it is relevant to point out that Eqs. (2.2) and (2.3) are quite general and can be applied for the cases of fuzzy observations/fuzzy categories, crisp observations/fuzzy categories, and fuzzy observations/crisp categories. In this context, crisp observations and crisp categories can be realized by means of degenerated fuzzy sets and rectangular fuzzy sets, respectively. For the special case of crisp observations/crisp categories, the resulting fuzzy set  $\xi_{\tilde{n}_{rc}}$  is degenerate. Figure 1 shows an exemplary case of fuzzy frequencies for fuzzy observations and fuzzy categories (Fig. 1A, middle and rightmost panels) and crisp observations and fuzzy categories as well (Fig. 1B, middle and rightmost panels).

### 3 LLCs for Fuzzy Frequency Tables

In this section, we describe the statistical procedure for computing latent linear correlations when observations are in the general form of fuzzy frequencies.

#### 3.1 Model

Let  $X = (X_i^j, X_i^k) i = 1, \dots, I$  be a collection of pairs of continuous random variables ( $j, k \in \{1, \dots, J\}, j \neq k$ ) following the bivariate Gaussian distribution centered at zero with correlation parameter  $\rho_{jk} \in [-1, 1]$  and density

**Algorithm 1** Computing fuzzy frequencies

```

procedure MAIN( $\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k, \tilde{\mathbf{g}}_j, \tilde{\mathbf{g}}_k$ )
  for  $r = 1, \dots, R$  and  $c = 1, \dots, C$  do
     $\epsilon_r^j \leftarrow \text{DoI}(\tilde{\mathbf{x}}_j, \tilde{\mathbf{g}}_r)$  ▷ Compute degrees of inclusion
     $\epsilon_c^k \leftarrow \text{DoI}(\tilde{\mathbf{x}}_k, \tilde{\mathbf{g}}_c)$ 
     $\epsilon_{rc}^{jk} \leftarrow \min(\epsilon_r^j, \epsilon_c^k)$  ▷ Compute joint degree of inclusion
     $\mathbf{Z} \leftarrow (\epsilon_{rc}^{jk} \mathbf{1}_I^T - \mathbf{1}_I^T \epsilon_{rc}^{jk})$ 
    for  $i = 1, \dots, I$  do
       $z[i] \leftarrow \mathbf{1}_I^T \mathcal{H}(\mathbf{Z}[:, i])$ 
    end for
    for  $n = 0, \dots, I$  do
       $\mu_{\text{FGC}}[n] \leftarrow \max(\mathcal{H}(\mathbf{z} - n) \odot \epsilon_{rc}^{jk})$  ▷ Fuzzy counting functions
       $\mu_{\text{FLC}}[n] \leftarrow 1 - \max(\mathcal{H}(\mathbf{z} - (n + 1)) \odot \epsilon_{rc}^{jk})$ 
       $\xi_{\tilde{n}_{rc}}^{jk}[n] \leftarrow \min(\mu_{\text{FLC}}[n], \mu_{\text{FGC}}[n])$  ▷ Compute fuzzy frequencies
    end for
  end for
  return  $\xi_{\tilde{n}}^{jk}$ 
end procedure

```

```

procedure DOI( $\tilde{\mathbf{x}}, \tilde{\mathbf{g}}$ )
  for  $i = 1, \dots, I$  do
     $\epsilon[i] \leftarrow \int_x \min(\xi_{\tilde{x}_{[i]}}(x), \xi_{\tilde{g}}(x)) dx / \max(1, \int_x \xi_{\tilde{x}_{[i]}}(x) dx)$  ▷
  end for
  return  $\epsilon$ 
end procedure

```

NOTE: The algorithm requires as input the  $I \times 1$  arrays of fuzzy observations  $\tilde{\mathbf{x}}_j$  and  $\tilde{\mathbf{x}}_k$  along with the fuzzy categories  $\tilde{\mathbf{g}}_j$  and  $\tilde{\mathbf{g}}_k$  for the  $j, k$ th pair of variables and returns as output the  $R \times C$  array of membership functions  $\xi_{\tilde{n}}^{jk} = (\xi_{\tilde{n}_{11}}, \dots, \xi_{\tilde{n}_{rc}}, \dots, \xi_{\tilde{n}_{RC}})$  associated to each fuzzy count  $\tilde{n}_{rc}^{jk}$ .

$$f_X(\mathbf{x}; \rho_{jk}) = \frac{1}{2\pi\sqrt{1 - \rho_{jk}^2}} \exp\left(-\frac{1}{2} \left[ \frac{(x^j)^2 + (x^k)^2 - 2x^j x^k \rho_{jk}}{1 - \rho_{jk}^2} \right]\right), \quad (3.1)$$

for  $-\infty < x^j < \infty$  and  $-\infty < x^k < \infty$ . Without loss of generality, consider the collection of fuzzy observations

$$\tilde{\mathbf{y}} = \{(\tilde{y}_1^j, \tilde{y}_1^k), \dots, (\tilde{y}_i^j, \tilde{y}_i^k), \dots, (\tilde{y}_I^j, \tilde{y}_I^k)\},$$

which relates to the (latent) bivariate Gaussian model in Eq. (3.1) via the constraint

$$(\tilde{y}_i^j \in \tilde{g}_r^j) \wedge (\tilde{y}_i^k \in \tilde{g}_c^k) \quad \text{iff} \quad (X_i^j, X_i^k) \in (\tau_{r-1}^j, \tau_r^j] \times (\tau_{c-1}^k, \tau_c^k] \subset \mathbb{R}^2, \quad (3.2)$$

where  $\in$  is intended as fuzzy membership,  $(\tilde{g}_r^j, \tilde{g}_c^k)$  are observed fuzzy categories or granules, and the arrays  $\tau_{X^j} = (\tau_0^{X^j}, \dots, \tau_r^{X^j}, \dots, \tau_R^{X^j})$  and  $\tau_{X^k} = (\tau_0^{X^k}, \dots, \tau_c^{X^k}, \dots, \tau_C^{X^k})$  are thresholds of the bivariate support  $\mathbb{R}^2$  under the conventions  $\tau_0^{X^j} = \tau_0^{X^k} = -\infty$  and  $\tau_R^{X^j} = \tau_C^{X^k} = \infty$ . Note that since fuzzy numbers encompass crisp observations and crisp categories as special cases (i.e., degenerated and rectangular fuzzy numbers, respectively), the expression (3.2) can be used for the non-fuzzy case as well. For instance, the simplest situation involving non-fuzzy observations and non-fuzzy categories can be obtained rewriting the left part of the constraint as  $(y_i^j = r) \wedge (y_i^k = c)$ , which indicates that crisp observations take the indices of the categories.

The parameter space for the LLCs model is

$$\theta = \{\rho_{jk}, \tau_{X^j}, \tau_{X^k}\} \in [-1, 1] \times \mathbb{R}^{R-1} \times \mathbb{R}^{C-1},$$

whereas the log-likelihood function takes the following form in the case of independent and identically distributed fuzzy observations [47, 59]:

$$\begin{aligned} \ln \mathcal{L}(\theta; \tilde{\mathbf{N}}) &= K - \sum_{r=1}^R \sum_{c=1}^C \sum_{n \in \mathbb{N}_0} n \xi_{n_{rc}}^{jk}(n) \ln \pi_{rc}^{jk}(\theta) \\ &= K - \sum_{r=1}^R \sum_{c=1}^C \sum_{n \in \mathbb{N}_0} n \xi_{n_{rc}}^{jk}(n) \ln \int_{\tau_{r-1}^{X^j}}^{\tau_r^{X^j}} \int_{\tau_{c-1}^{X^k}}^{\tau_c^{X^k}} f_X(\mathbf{x}; \rho_{jk}) \, dx^j dx^k, \end{aligned} \quad (3.3)$$

where  $f_X(\mathbf{x}; \rho_{jk})$  is the model's density in Eq. (3.1),  $\xi_{n_{rc}}^{jk}(n)$  is the  $r$ cth fuzzy count, and  $K$  is a constant term. Note that  $f_X(\mathbf{x}; \rho_{jk})$  is not fuzzy in this context and its realizations represent unobserved (latent) quantities. The evaluation of  $(\tilde{y}_i^j \in \tilde{g}_r^j) \wedge (\tilde{y}_i^k \in \tilde{g}_c^k)$  gives raise to a collection of fuzzy counts  $\tilde{n}_{11}^{jk}, \dots, \tilde{n}_{rc}^{jk}, \dots, \tilde{n}_{RC}^{jk}$  acting as possibilistic constraints on the unobserved non-fuzzy counts which would be observed if fuzziness was missed. As such, the expression  $\xi_{n_{rc}}^{jk}(n_{rc}) \in [0, 1]$  should be interpreted as the possibility that the crisp count  $n_{rc}$  has to occur, with  $\xi_{n_{rc}}^{jk}(n_{rc}) = 1$  indicating that  $n_{rc}$  is fully possible. According to the epistemic viewpoint on fuzzy statistics [37], the sampling process is thought as being the consequence of a two-stage generation mechanism, the first of which is a random experiment and the second is a non-random fuzzification of the outcome being realized. As an example of this schema, consider the simplest case of crisp observations (e.g., income and tobacco use) that are classified by a group of raters or an automatic classification system on the basis of fuzzy categories (e.g., income levels: low, medium, high; tobacco use: none, sporadic, habitual). Stated in this way, the fuzzy frequencies associated with income and tobacco use encapsulate two sources of uncertainty, namely the random component due to the sampling process and the non-random component due to the post-sampling fuzzy classification.



### 3.2 Parameter Estimation

To estimate  $\theta$ , we adopt the Olsson’s two-stage approach for latent linear correlations which iteratively alternates between approximating  $\hat{\tau}$  from the observed count data and maximizing Eq. (3.3) with respect to  $\hat{\rho}$  given the current thresholds [59]. In the case of fuzzy data, this procedure can be implemented using a variant of the expectation–maximization algorithm generalized to the case of fuzzy observations [24]. Likewise for the standard EM algorithm, the fuzzy EM version alternates between the E-step, which requires computing the expected complete log-likelihood given the candidate  $\theta' = \theta^{(q-1)}$ , and the M-step, which maximizes the expected complete log-likelihood w.r.t.  $\theta^{(q)}$ . More precisely, in the fuzzy EM algorithm the complete data log-likelihood is that obtained if the matrix of counts  $\mathbf{N}_{R \times C}^{jk}$  was precisely observed, namely:

$$\ln \mathcal{L}(\theta; \mathbf{N}) = \ln I! - \sum_{r=1}^R \sum_{c=1}^C n_{rc}^{jk} \ln \pi_{rc}^{jk}(\theta) - \sum_{r=1}^R \sum_{c=1}^C \ln n_{rc}^{jk}!. \tag{3.4}$$

Given the estimates  $\theta'$ , the E-step for the  $(q)$ th iteration consists of computing the Q-function via conditional expectation on the observed fuzzy counts:

$$\begin{aligned} Q(\theta, \theta') &= \mathbb{E}_{\theta'} \left[ \ln \mathcal{L}(\theta; \mathbf{N}) \mid \tilde{\mathbf{N}} \right] \\ &\propto \sum_{r=1}^R \sum_{c=1}^C \mathbb{E}_{\theta'} \left[ N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk} \right] \ln \pi_{rc}^{jk}(\theta) - \mathbb{E}_{\theta'} \left[ \ln N_{rc}^{jk}! \mid \tilde{n}_{rc}^{jk} \right]. \end{aligned} \tag{3.5}$$

The conditional expectations involve the density of a discrete random variable  $N_{rc}$  conditioned on a fuzzy event  $\tilde{n}_{rc}$  that, under the multinomial schema for random counts, can reasonably be modeled as Binomial [1]. Thus, using the definition of fuzzy probability,  $N_{rc} | \tilde{n}_{rc}$  is as follows:

$$\begin{aligned} p_{N_{rc}^{jk} | \tilde{n}_{rc}^{jk}}(n; \pi_{rc}^{jk}(\theta)) &= \frac{\mathbb{P}_{\theta} \left( N_{rc}^{jk}, \tilde{n}_{rc}^{jk} \right)}{\mathbb{P}_{\theta} \left( \tilde{n}_{rc}^{jk} \right)} = \frac{\xi_{\tilde{n}_{rc}^{jk}}^{jk}(n) p_{N_{rc}^{jk}}(n; \pi_{rc}^{jk}(\theta))}{\sum_{n \in \mathbb{N}_0} n \xi_{\tilde{n}_{rc}^{jk}}^{jk}(n) p_{N_{rc}^{jk}}(n; \pi_{rc}^{jk}(\theta))}, \tag{3.6} \\ \pi_{rc}^{jk}(\theta) &= \int_{\tau_{r-1}^{x_j}}^{\tau_r^{x_j}} \int_{\tau_{c-1}^{x_k}}^{\tau_c^{x_k}} f_X(\mathbf{x}; \rho_{jk}) \, dx^j dx^k, \end{aligned} \tag{3.7}$$

where  $p_{N_{rc}^{jk}} = \text{Bin}(n; \pi_{rc}^{jk}(\theta))$  and  $f_X(\mathbf{x}; \rho_{jk})$  is the latent model’s density in Eq. (3.1). Note that the quantity  $I \pi_{rc}^{jk}(\theta)$  is the reconstructed count from the bivariate latent model given the current parameters  $\theta'$  [66]. The linear form of the expectations in Eq. (3.5) is

$$\mathbb{E}_{\theta'} \left[ N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk} \right] = \sum_{n \in \mathbb{N}_0} n p_{N_{rc}^{jk} | \tilde{n}_{rc}^{jk}}(n; \pi_{rc}^{jk}(\theta')), \tag{3.8}$$

**Table 1** Expectation–maximization algorithm for estimating  $\theta = (\tau_{Xj}, \tau_{Xk}, \rho_{jk})$  in LLCs model with fuzzy frequency data

Algorithm 2	Olsson's two-stage approach via fuzzy EM algorithm	
	for $j \in (1, \dots, J)$ and $k \in (1, \dots, J)$ , $j \neq k$ , do:	
$\mathbf{q} = \mathbf{1}$ :	Set $\theta^{(q)} = (\rho_{jk}^0, \tau_{Xj}^0, \tau_{Xk}^0)$ , $l^{(q)} = l^0$ , $\epsilon = 1e^{-09}$	INITIALIZATION
$\mathbf{q} > \mathbf{1}$ :	Compute $\pi^{jk}(\theta^{(q-1)})$ from Eq. (3.7)	E- STEP
	Compute $\widehat{\mathbf{N}}^{jk}$ given $\theta^{(q-1)}$ from Eq. (3.8)	
	Compute $\ln \widehat{\mathbf{N}}^{jk}!$ given $\theta^{(q-1)}$ from Eq. (7.1)	
	Compute $\{\widehat{\tau}_{Xj}^{(q)}, \widehat{\tau}_{Xk}^{(q)}\}$ from Eqs. (3.9) and (3.10)	M- STEP
	Set $\theta^{(q)} = (\rho_{jk}^{(q-1)}, \tau_{Xj}^{(q)}, \tau_{Xk}^{(q)})$	
	Solve $\frac{\partial}{\partial \rho_{jk}} \mathcal{Q}(\theta, \theta^{(q)}) = 0$ w.r.t. $\rho_{jk}$ see Eq. (3.11)	
	Set $\theta^{(q)} = (\rho_{jk}^{(q)}, \tau_{Xj}^{(q)}, \tau_{Xk}^{(q)})$	
	Evaluate $l^{(q)} = \ln \mathcal{L}(\theta^{(q)}; \widehat{\mathbf{N}})$ see Eq. (3.4)	FINALIZATION
	Compute $l_\delta = (l^{(q)} - l^{(q-1)})$	
	If $l_\delta < \epsilon$ , set $\widehat{\theta} = \theta^{(q)}$ and <u>stop</u> the algorithm	
	$\mathbf{R}[j, k] = \rho_{jk}^{(q)}$	

whereas, since it is not involved in the M-step of the algorithm, the nonlinear expectation is provided in Appendix A for the sake of completeness.

Finally, the M-step for the  $(q)$ th iteration requires maximizing the functional  $\mathcal{Q}(\theta, \theta')$  with respect to  $\theta$ . Given the filtered counts at the current iteration  $\widehat{\mathbf{N}}_{R \times C}^{jk}$  (see Eq. 3.8), the Olsson's two-stage estimation approach requires the estimation of thresholds from the cumulative marginals of filtered counts first:

$$\widehat{\tau}_{Xj}^{(q)} = \Phi^{-1} \left( \mathbf{A}_{R \times R} \widehat{\mathbf{N}}^{jk} \mathbf{1}_C \right), \quad (3.9)$$

$$\widehat{\tau}_{Xk}^{(q)} = \Phi^{-1} \left( \mathbf{A}_{C \times C} (\widehat{\mathbf{N}}^{jk})^T \mathbf{1}_R \right), \quad (3.10)$$

where  $\mathbf{A}$  is a lower triangular matrix of ones,  $\mathbf{1}$  is a vector of appropriate order of all  $1/I$ , and  $\Phi$  is the Gaussian univariate distribution function with mean zero and unitary variance. Next, conditioned on  $\{\widehat{\tau}_{Xj}^{(q)}, \widehat{\tau}_{Xk}^{(q)}\}$ , the remaining parameter is found by solving the score equation of  $\mathcal{Q}(\theta, \theta^{(q)})$  numerically w.r.t.  $\rho_{jk}$ :

$$\mathcal{U}_{\rho_{jk}} = \frac{\partial \mathcal{Q} \left( \rho_{jk}, \{\widehat{\tau}_{Xj}^{(q)}, \widehat{\tau}_{Xk}^{(q)}\} \right)}{\partial \pi^{jk}} \frac{\partial \pi^{jk}}{\partial \rho_{jk}} = 0. \quad (3.11)$$

The algorithm proceeds iteratively until the log-likelihood does not increase significantly. Table 1 summarizes the fuzzy EM algorithm for the LLCs model.

### 3.3 Remarks

*About the Convergence of the Algorithm* Given a candidate  $\theta'$ , the fuzzy EM starts by constructing the surrogate  $\mathcal{Q}(\theta, \theta')$  that lower-bounds the observed data log-likelihood  $\ln \mathcal{L}(\theta; \tilde{\mathbf{N}})$  (E-step). Next, it is maximized to get the current estimates  $\theta^{(q)}$  (M-step), which is in turn used to construct a new lower bound  $\mathcal{Q}(\theta, \theta^{(q)})$  in the next iteration to get a new estimate  $\theta^{(q+1)}$ . The estimates in the M-step are chosen so that  $\mathcal{Q}(\theta, \theta^{(q)}) \geq \mathcal{Q}(\theta, \theta')$ , which forms the base of the monotonicity condition  $\ln \mathcal{L}(\theta^{(q+1)}; \tilde{\mathbf{N}}) \geq \ln \mathcal{L}(\theta^{(q)}; \tilde{\mathbf{N}})$  [54]. As for the standard case, the monotonicity of the sequence  $\{\ln \mathcal{L}(\theta^{(q)})\}_{q \in \mathbb{N}}$  implies the convergence to a stationary value, which can be global or local depending on the characteristics of the log-likelihood function and the starting point  $\theta^0$ . A sketch of the proof of the monotonicity of the fuzzy EM for the LLCs is provided in Appendix B, whereas the formal equivalence between EM and fuzzy EM is detailed in [24, 62].

*About the Starting Values of the Algorithm* Suitable starting values  $\theta^0$  can be obtained by first defuzzifying the observed fuzzy frequencies matrix  $\tilde{\mathbf{N}}^{jk}$  to obtain non-fuzzy counts and then applying the standard Olsson’s two-stage approach [59] on defuzzified data. In general, this yields convenient starting values. In the LLCs model, defuzzification can be performed via mean or max-based procedures as follows:  $\hat{n}_{rc}^{\text{mean}} \approx \sum_{n \in \mathbb{N}_0} n \xi_{\tilde{n}_{rc}}(n) / (\sum_{n \in \mathbb{N}_0} \xi_{\tilde{n}_{rc}}(n))$ ,  $\hat{n}_{rc}^{\text{max}} = \max\{n \in \mathbb{N}_0 : \xi_{\tilde{n}_{rc}}(n) = \max_{z \in \mathbb{N}_0} \xi_{\tilde{n}_{rc}}(z)\}$ ,  $r = 1, \dots, R$ ,  $c = 1, \dots, C$ .

*About the Term  $p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\theta))$ .* The term  $p_{N_{rc}|\tilde{n}_{rc}}$  represents the density of a non-fuzzy random variable conditioned on fuzzy numbers and can mathematically be interpreted as the combination of two independent components, namely the random mechanism underlying the sampling process and the observer’s partial knowledge (imprecision) about the sample realizations. In this sense, as it weights each fuzzy datum by the probability that it has to occur [52],  $p_{N_{rc}|\tilde{n}_{rc}}$  should not be confused with the mean-based defuzzification of fuzzy numbers. A nice property of this formulation is that fuzziness vanishes when precise observations are available. Indeed, the conditional density involving a degenerated fuzzy number  $\hat{n}_{rc}$  boils down to a degenerated discrete density  $p_{N_{rc}|\hat{n}_{rc}}$  with nonzero probability masses only for those  $n$  such that  $\xi_{\tilde{n}_{rc}}(n) = 1$ . As a consequence, the fuzzy EM procedure reduces to standard Olsson’s two-stage maximum likelihood estimation. In general, there are a number of ways for plugging-in non-stochastic components of uncertainty into  $p_{N_{rc}|\tilde{n}_{rc}}$ , such as those involving imprecise probability [7], conditional probability [18], belief measures [76], and random fuzzy variables [30].

*About the Computation of Standard Errors and Inference* Standard errors for  $\hat{\rho}_{jk}$  might be computed by following the general results of the EM algorithm [54]. In particular, there are a number of procedures which have been suggested to this purpose. A common way is that of calculating the square root of the inverse of the empirical information matrix [40], which approximates the expected information matrix by using the observed score statistic. Similarly, another strategy has been suggested by Louis [53] and requires the computation of the expected complete and missing information matrices. Alternatively, standard errors can also be obtained by means of

nonparametric or parametric bootstrap techniques, which have been demonstrated to be robust under several circumstances [55]. In the context of this article, we resorted to use the nonparametric bootstrap technique to compute standard errors for the fuzzy polychoric estimator (see [55], section 4.6). A particular advantage of this procedure is that  $(1 - \alpha)$  confidence intervals (CIs) can also be obtained as a by-product of the bootstrap technique, for instance the bias-corrected and accelerated (BCa) CIs. More precisely,  $Q$  bootstrap samples of the fuzzy matrix  $\{\tilde{\mathbf{N}}^{(jk)}\}_{q=1, \dots, Q}$  can be obtained by drawing from the  $\alpha$ -cuts of the  $r$ c $t$ h element of  $\tilde{\mathbf{N}}^{(jk)}$  (for each  $r = 1, \dots, R$  and  $c = 1, \dots, C$ ) and then fuzzifying back the bootstrapped sample of count data [33]. Finally, the sequence of estimates  $\{\hat{\rho}_{jk}\}_{q=1, \dots, Q}$  is used to compute the bootstrap covariance matrix  $\text{Cov}[R]_{jk} \approx \frac{1}{Q-1} \sum_{q=1}^Q (\hat{\rho}_{jk}^{(q)} - \frac{1}{B} \sum_{q=1}^Q \hat{\rho}_{jk}^{(q)})^2$ , which is in turn used for the computation of the standard errors  $\hat{\sigma}_{\rho_{jk}} = \sqrt{\text{Cov}[R]_{jk}}$ , and the  $(1 - \alpha)$  BCa confidence intervals [25].

*About the Polychoric Correlation Matrix  $\mathbf{R}_{J \times J}$*  As for the standard approach used in computing polychoric correlation matrices (e.g., see [41, 59]), also in the case of fuzzy data, the matrix of latent linear correlations is obtained by calculating each element  $\rho_{jk}$  of the correlation matrix pairwise. Although this approach offers a simple and effective alternative to more challenging methods (e.g., see [48, 68]), in some circumstances, it may lead to non-positive definite correlation matrices. This can be problematic, especially when such matrices are used as input of other statistical models such as factor analyses or SEMs [51]. In these cases, eigenvalue decomposition-based smoothing [44], least squares [44] or Dykstra's [35] corrections constitute workable solutions to solve this issue.

## 4 Simulation Study

The aim of this simulation study is twofold. First, we wish to evaluate the performances of fuzzy EM algorithm in estimating parameters of the LLCs model and, second, to investigate whether the standard Olsson's maximum likelihood procedure performs as good as the proposed method if applied on max-based and mean-based defuzzified data. The case  $J = 2$  has been considered for the sake of simplicity. The Monte Carlo study has been performed on a (remote) HPC machine based on 16 CPU Intel Xeon CPU E5-2630L v3 1.80 GHz, 16x4 GB RAM, whereas computations and analyses have been done in the R framework for statistical analyses.

*Design* The design of the study involved three factors, namely (i)  $I \in \{150, 250, 500, 1000\}$ , (ii)  $\rho^0 \in \{0.15, 0.50, 0.85\}$ , (iii)  $R = C \in \{4, 6\}$ , which were varied in a complete factorial design with  $4 \times 3 \times 2 = 24$  possible combinations. The threshold parameters were held fixed under the equidistance hypothesis [41], namely  $\tau_{X^j}^0 = \tau_{X^k}^0 = (-2.00, -0.66, 0.66, 2.00)$  for the conditions with  $R = C = 4$  and  $\tau_{X^j}^0 = \tau_{X^k}^0 = (-2.00, -1.20, -0.40, 0.40, 1.20, 2.00)$  for  $R = C = 6$ . For each combination,  $B = 5000$  samples were generated yielding to  $5000 \times 24 = 120000$  new data and an equivalent number of parameters.

*Data Generation and Procedure* Let  $I_a, \rho_b^0, R_d = C_d$  be distinct levels of the factors  $I, \rho^0, R,$  and  $C$ . Then, fuzzy frequency data have been generated according to the following procedure. For each  $r = 1, \dots, R_d$  and  $c = 1, \dots, C_d$ :

- (i) Set  $n_{rc} = I_a \pi_{rc}$  (see Eq. (3.7)) given  $\tau_{Xj}^0, \tau_{Xk}^0, \rho_b^0,$  and  $I_a$ ;
- (ii) the imprecision concerning  $n_{rc}$  was generated as follows:  $m_1 \sim \text{Gamma}_d(\alpha_{m_1}, \beta_{m_1})$  where  $\alpha_{m_1} = 1 + n_{rc} \beta_{m_1}, \beta_{m_1} = (n_{rc} + n_{rc}^2 + 4s_1^2)^{\frac{1}{2}} / 2s_1^2, s_1 \sim \text{Gamma}_d(\alpha_{s_1}, \beta_{s_1}), \alpha_{s_1} = 1 + m_0 \beta_{s_1}, \beta_{s_1} = (m_0 + m_0^2 + 4s_0^2)^{\frac{1}{2}} / 2s_0^2, m_0 = 1$  and  $s_0 = 0.25,$  with  $\text{Gamma}_d$  indicating the discrete Gamma random variable with shape and rate being reparameterized in terms of mean  $m$  and variance  $s$ ;
- (iii) the fuzzy set associated with  $\tilde{n}_{rc}$  was obtained via the following probability–possibility transformation:  $\xi_{\tilde{n}_{rc}} = f_{\tilde{G}_d}(\mathbf{n}; \alpha_{rc}, \beta_{rc}) / \max f_{\tilde{G}_d}(\mathbf{n}; \alpha_{rc}, \beta_{rc}),$  with  $\mathbf{n} = \{0, 1, \dots, I_a\}, \alpha_{rc} = 1 + m_1 \beta_{s_1}, \beta_{s_1} = 1 + (m_1 + m_1^2 + 4s_1^2)^{\frac{1}{2}} / 2s_1^2, \beta_{rc} = (m_1 + m_1^2 + 4s_1^2)^{\frac{1}{2}} / 2s_1^2,$  and  $f_{\tilde{G}_d}(n; \alpha_{rc}, \beta_{rc})$  being the discrete Gamma density normalized to one in order to mimic the behavior of a normal fuzzy set [27]. The discrete density  $f_{\tilde{G}_d}$  is computed as a difference of survival functions of the continuous Gamma density  $S_G(x) - S_G(x + 1)$  [15, 74].

Note that step (ii) is required in order to make crisp counts entirely imprecise so that  $\tilde{n}_{rc}$  is no longer centered on  $n_{rc}$ . Finally, parameters  $\theta = \{\rho, \tau_{Xj}, \tau_{Xk}\}$  were estimated from the fuzzy counts  $\tilde{N}_{R_d \times C_d}$  using the fuzzy EM algorithm (fEM) and the standard Olsson’s two-stage maximum likelihood on max-based (dML-max) and mean-based (dML-mean) defuzzified counts (see Sect. 3.3).

*Outcome Measures* For each condition of the simulation design, the three methods (i.e., fEM, ML-max, ML-mean) were evaluated in terms of bias and root mean square errors. In addition, for each method thresholds were aggregated to form a scalar statistic, namely  $\hat{\tau} = \mathbf{1}_{R_d}^T \hat{\tau}_{Xj}$  and  $\hat{\tau} = \mathbf{1}_{C_d}^T \hat{\tau}_{Xk}$ . (Note that  $\tau_{Xj}$  and  $\tau_{Xk}$  are equal by design.) For the sake of completeness, bootstrap standard errors and 95% BCa confidence intervals were computed for the three methods along with coverage probability and interval lengths.

*Results* Tables 2, 3, 4, and 5 report the results of the simulation study with regards to  $\hat{\rho}$  and  $\hat{\tau}$  for both  $R = C = 4$  and  $R = C = 6$  cases. We begin with the correlation parameter  $\rho$  for the case  $R = C = 4$  (see Table 2). Considering  $\rho^0 = 0.15,$  the methods showed negligible bias in estimating  $\rho$ . However, they differed in terms of RMSE, with fEM showing lower values with increasing sample size if compared to dML-max and dML-mean. With increasing correlation length ( $\rho^0 > 0.15$ ), bias of estimates as well as RMSE was more pronounced for dML-max and dML-mean. The same results were also observed for the case with  $R = C = 6$  (see Table 3). With regards to the overall statistic  $\hat{\tau}$  for the threshold parameters, all the methods achieved comparable results regardless of  $\rho^0$ . In particular, fEM showed slightly higher bias and RMSE and then dML-max and dML-mean methods across  $R = C = 4$  (see Table 4) and  $R = C = 6$  (see Table 5) conditions. To further investigate these results, we studied average bias and variance of estimates for  $\hat{\tau}_{Xj}$  (or  $\hat{\tau}_{Xk}$ ) as a function of sample size  $I$  and  $\rho^0$ . We found that the leftmost and rightmost thresholds tended to be slightly larger for fEM as opposed to the innermost thresholds for both  $R = C = 4$  (see

**Table 2** Simulation study: average bias and root-mean-square errors for  $\rho$  in the condition  $R = C = 4$ 

$R = C = 4$	fEM		dML-max		dML-mean	
	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>
$\rho = 0.15$						
$I = 150$	0.03401	0.08911	-0.01653	0.11826	-0.04354	0.08824
$I = 250$	0.00455	0.05062	-0.02821	0.08106	-0.04020	0.06766
$I = 500$	0.01047	0.02974	0.00311	0.04180	-0.00743	0.03339
$I = 1000$	-0.00321	0.01515	-0.01012	0.02441	-0.01421	0.02276
$\rho = 0.50$						
$I = 150$	0.01265	0.07236	-0.08807	0.15014	-0.17694	0.19253
$I = 250$	-0.03699	0.06349	-0.12376	0.15052	-0.17174	0.18119
$I = 500$	-0.00151	0.02688	-0.04673	0.06983	-0.08356	0.09120
$I = 1000$	-0.00050	0.01389	-0.02226	0.03582	-0.03921	0.04459
$\rho = 0.85$						
$I = 150$	0.00194	0.04504	-0.21865	0.25598	-0.32889	0.33729
$I = 250$	-0.00285	0.02903	-0.17042	0.19816	-0.25843	0.26540
$I = 500$	-0.00104	0.01586	-0.10519	0.12382	-0.16418	0.16884
$I = 1000$	-0.00056	0.00810	-0.06598	0.07880	-0.10451	0.10760

Note that fEM is the fuzzy EM algorithm, whereas dML-max and dML-mean denote the standard maximum likelihood based on max-based and mean-based defuzzified counts

**Table 3** Simulation study: average bias and root-mean-square errors for  $\rho$  in the condition  $R = C = 6$ 

$R = C = 6$	fEM		dML-max		dML-mean	
	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>
$\rho = 0.15$						
$I = 150$	0.02884	0.10022	-0.01490	0.10067	-0.04919	0.08355
$I = 250$	0.00860	0.05720	-0.02461	0.07169	-0.04289	0.06501
$I = 500$	-0.01619	0.03395	-0.02606	0.04555	-0.03853	0.04869
$I = 1000$	0.00064	0.01474	-0.00539	0.01947	-0.01021	0.01892
$\rho = 0.50$						
$I = 150$	-0.05241	0.10238	-0.17183	0.20107	-0.25228	0.26146
$I = 250$	-0.00259	0.04950	-0.09811	0.12241	-0.16604	0.17374
$I = 500$	-0.00644	0.02414	-0.05111	0.06359	-0.08845	0.09304
$I = 1000$	-0.00214	0.01222	-0.02415	0.03205	-0.04278	0.04594
$\rho = 0.85$						
$I = 150$	-0.00245	0.05111	-0.24268	0.26670	-0.38252	0.38855
$I = 250$	-0.01614	0.03414	-0.18437	0.19946	-0.28388	0.28831
$I = 500$	0.00078	0.01412	-0.09831	0.10892	-0.16062	0.16358
$I = 1000$	-0.00219	0.00694	-0.05869	0.06422	-0.09167	0.09336

Note that fEM is the fuzzy EM algorithm, whereas dML-max and dML-mean denote the standard maximum likelihood based on max-based and mean-based defuzzified counts

**Table 4** Simulation study: average bias and root-mean-square errors for the aggregated thresholds  $\hat{\tau} = \mathbf{1}_{R_d}^T \hat{\tau}_{Xj}$  in the condition  $R = C = 4$

$R = C = 4$	fEM		dML-max		dML-mean	
	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>
$\rho = 0.15$						
$I = 150$	0.11082	0.30335	0.03419	0.16958	0.11037	0.14607
$I = 250$	0.07881	0.12035	0.01038	0.10643	0.03761	0.08513
$I = 500$	0.02189	0.04664	0.01363	0.05270	0.01595	0.04706
$I = 1000$	0.02916	0.03685	0.01158	0.02967	0.01314	0.02781
$\rho = 0.50$						
$I = 150$	0.00597	0.12056	0.09273	0.16206	0.14804	0.17656
$I = 250$	0.04446	0.09646	0.02888	0.10059	0.05722	0.09249
$I = 500$	0.04445	0.06225	0.00718	0.05440	0.00284	0.04488
$I = 1000$	0.01771	0.02797	0.00017	0.02606	0.00288	0.02370
$\rho = 0.85$						
$I = 150$	0.06325	0.32381	0.08313	0.16289	0.13419	0.16534
$I = 250$	0.03843	0.09054	0.03940	0.10066	0.06298	0.09680
$I = 500$	0.01918	0.04347	0.02264	0.05314	0.02680	0.04995
$I = 1000$	0.03036	0.03668	0.00637	0.02696	0.00632	0.02414

Note that fEM is the fuzzy EM algorithm, whereas dML-max and dML-mean denote the standard maximum likelihood based on max-based and mean-based defuzzified counts

Supplementary Materials, Figure S1) and  $R = C = 6$  conditions (see Supplementary Materials, Figure S2). Moreover, the variance of estimates for the leftmost and rightmost thresholds was higher if compared to the innermost thresholds (see Supplementary Materials, Table S2) but, as expected, it reduced with increasing sample size regardless of  $\rho^0$ . This is not surprising given that we implemented a standard LLCs model in which no particular constraints were applied on threshold estimates, such as  $\mathbf{1}_{R_d}^T \hat{\tau}_{Xj} = 0$  (e.g., see [28]).<sup>1</sup> Most importantly, according to the Gaussianity assumption underlying the LLCs model, estimated thresholds were symmetric and equidistant with respect to the fixed point zero (see Supplementary Materials, Table S1). Overall, the results suggest that fEM should be preferred over defuzzified maximum likelihood when the interest is in estimating the latent linear association  $\rho$  among pairs of variables and fuzzy frequency statistics are available. On the contrary, for those particular cases where  $\rho$  is known and the interest is in estimating the true threshold parameters, standard Olsson’s maximum likelihood method can directly be applied after defuzzifying observed fuzzy frequency counts. With regards to the estimation of the standard errors, the three algorithms showed comparable results. As expected, the statistic  $\hat{\sigma}_{\rho_{jk}}$  decreased as a function of the sample size  $I$  for both  $R = C = 4$  and  $R = C = 6$  conditions (see Supplementary Materials, Tables S3 and S4). Instead, with regards to 95%

<sup>1</sup> It should be remarked that the unconstrained approach is most common in LLCs-based applications, especially when the primary interest lies in making inference about  $\rho$ . In this case, the threshold parameters play an auxiliary role as they only affect the scale of the latent variables underlying the LLCs model (e.g., see [49]).

**Table 5** Simulation study: average bias and root-mean-square errors for the aggregated thresholds  $\hat{\tau} = \mathbf{1}_{R_d}^T \hat{\tau}_{X_j}$  in the condition  $R = C = 6$ 

$R = C = 6$	fEM		dML-max		dML-mean	
	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>	<i>bias</i>	<i>rmse</i>
$\rho = 0.15$						
$I = 150$	0.07773	0.32426	0.06858	0.15506	0.14761	0.17434
$I = 250$	0.04557	0.10842	0.02624	0.09560	0.06524	0.09514
$I = 500$	0.04214	0.06111	0.00563	0.05203	0.00926	0.04363
$I = 1000$	0.01893	0.02878	0.00156	0.02522	0.00456	0.02381
$\rho = 0.50$						
$I = 150$	0.13719	0.43117	0.06764	0.15797	0.14535	0.17206
$I = 250$	0.06769	0.13028	0.02734	0.09601	0.06286	0.09325
$I = 500$	0.01777	0.04274	0.01712	0.04958	0.02562	0.04733
$I = 1000$	0.02963	0.03633	0.00881	0.02693	0.00994	0.02522
$\rho = 0.85$						
$I = 150$	0.02021	0.17906	0.10739	0.15707	0.16366	0.18348
$I = 250$	0.01492	0.06809	0.06039	0.10015	0.09371	0.11295
$I = 500$	0.02625	0.04338	0.01779	0.04829	0.03246	0.04939
$I = 1000$	0.02293	0.02893	0.00222	0.02407	0.00696	0.02142

Note that fEM is the fuzzy EM algorithm, whereas dML-max and dML-mean denote the standard maximum likelihood based on max-based and mean-based defuzzified counts

CIs, only the fEM method showed consistent results in terms of coverage probability and interval lengths over all the simulation conditions (see Supplementary Materials, Table S3 and S4). In particular, with the exception of the condition  $\rho^0 = 0.15$ , the dML-max and dML-mean algorithms did not reach the nominal coverage probability. By contrast, the empirical coverage probability for the fEM algorithm was close to (or higher then) the nominal value, with interval lengths decreasing as a function of the sample size  $I$ .

## 5 Applications

In this section, we describe the application of the proposed method to two case studies from health and natural sciences, involving the assessment of a psychotherapeutic intervention (application 1) and the evaluation of meteorological characteristics for forty Turkish cities (application 2). Note that both the applications are provided to merely illustrate the use of fuzzy LLCs model when dealing with imprecise data.

### 5.1 Application 1: Assessing the Outcome of a Therapy

Evaluating the quality of a psychotherapy session plays a central role in evidence-based medicine. A typical approach to understand the fundamentals of the therapeutic



**Table 6** Application 1: Fuzzy categories for the three variables of the assessment task

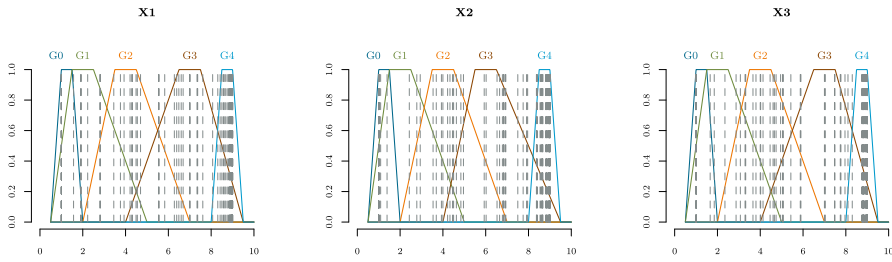
	$X_1$				$X_2$				$X_3$			
	$x_l$	$c_1$	$c_2$	$x_u$	$x_l$	$c_1$	$c_2$	$x_u$	$x_l$	$c_1$	$c_2$	$x_u$
$r = 1$	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00	0.50	1.00	1.50	2.00
$r = 2$	0.50	1.50	2.50	5.00	0.50	1.50	2.50	5.00	0.50	1.50	2.50	5.00
$r = 3$	2.00	3.50	4.50	7.00	2.00	3.50	4.50	7.00	2.00	3.50	4.50	7.00
$r = 4$	4.00	6.50	7.50	9.50	4.00	5.50	6.50	9.50	4.00	6.50	7.50	9.50
$r = 5$	8.00	8.50	9.00	9.50	8.00	8.50	9.00	9.50	8.00	8.50	9.00	9.50

Note that each category is represented by means of trapezoidal fuzzy numbers (see Eq. 2.1)

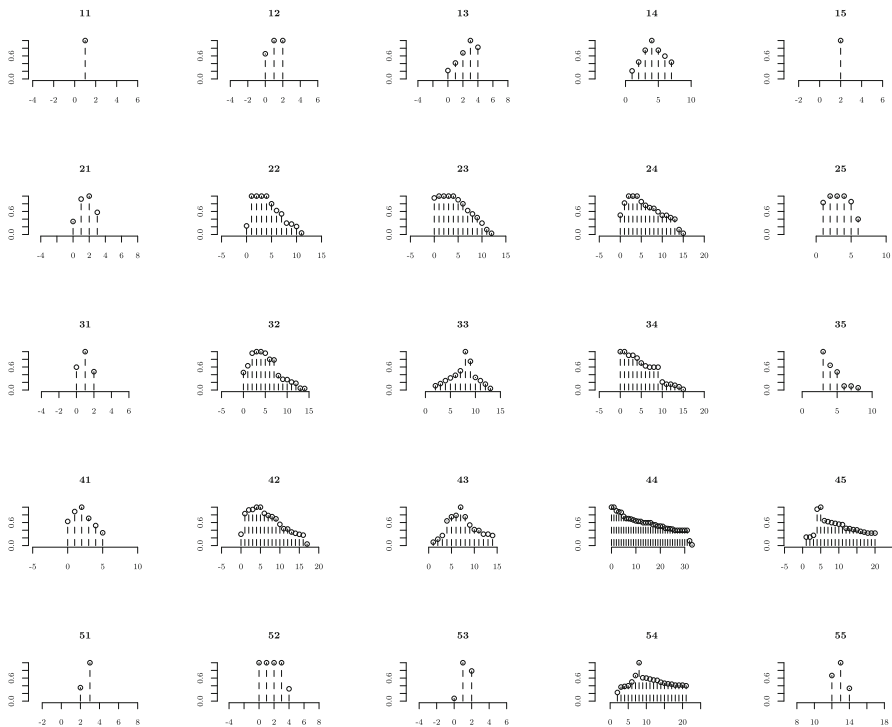
**Table 7** Application 1: Latent linear correlation matrix estimated via Olsson’s two-stage fuzzy EM algorithm (the bootstrap standard errors are reported in parentheses)

	$X_1$	$X_2$	$X_3$
$X_1$	1.00000		
$X_2$	0.06948 (0.11044)	1.00000	
$X_3$	0.00004 (0.09693)	0.21762 (0.10389)	1.00000

process consists in asking experts to assess the global quality and characteristics of the therapist–patient relationship through specialized instruments such as the PQS questionnaire [61]. The data thus collected generally consist either of ratings or of classification of attributes made through bounded and graded scales. Because of their characteristics, these tasks often involve imprecision and vagueness that can adequately be accounted for by the fuzzy statistical modeling. In this application, we consider the assessment of a psychotherapy session by means of the PQS questionnaire. Data were originally collected by [17] and refer to  $I = 60$  evaluations of psychotherapy on a 9-point scale over  $J = 3$  dimensions of assessment. Given the nature of the task, the three variables were originally considered to be fuzzy, each with three trapezoidal fuzzy categories. To account for the extremes of the classification scale, two more outer categories were added so that  $R = C = 5$  (see Table 6). Figure 2 shows the granulation based on five fuzzy categories ( $G_0, \dots, G_4$ ) for each dimension of assessment along with the corresponding crisp observations. The aim is to compute the correlation matrix for the three fuzzy variables, with the hypothesis that the higher degree of association is related to a good therapeutic outcome. The first step requires computing the fuzzy frequency matrix  $\tilde{N}_{5 \times 5}$  for each pair of  $J = 3$  fuzzy variables given the crisp observed data. Next, the matrix of fuzzy counts is used to estimate the latent linear correlation matrix  $\tilde{R}_{5 \times 5}$ . Figure 3 shows a graphical representation of the matrix of fuzzy counts  $\tilde{N}_{5 \times 5}$  for one pair of variables (i.e.,  $X_2, X_3$ ). It contains fuzzy numbers with various degrees of fuzziness and includes combinations with degenerated fuzzy counts as well (i.e.,  $G_0^{(2)}, G_4^{(3)}$  and  $G_0^{(2)}, G_0^{(3)}$ ). Table 7 reports the estimates of LLC coefficients. Overall, the results showed a low level of association among the three dimensions, which in turn indicated that the psychotherapy being assessed cannot be classified as having a good outcome.



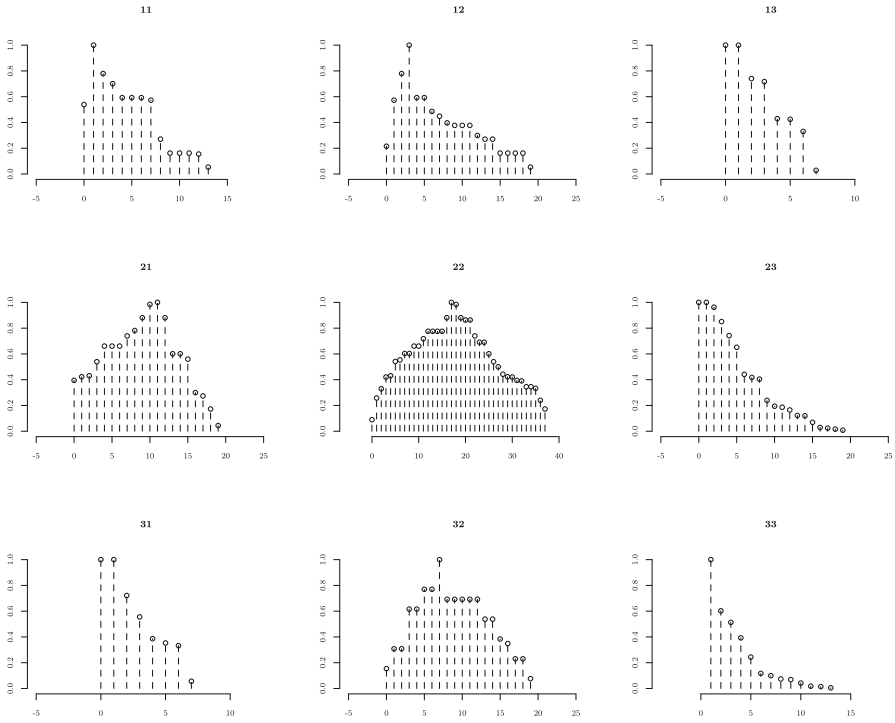
**Fig. 2** Application 1: Granulation for the three fuzzy variables along with crisp observations (dashed gray lines)



**Fig. 3** Application 1: Fuzzy frequency matrix for the pair  $X_2, X_3$ . Note that each cell contains a fuzzy natural number  $\tilde{n}_{rC}$  for a specific combination of the  $R \times C$  granulation space

### 5.2 Application 2: Effect of Climatic Variables on Rainfall

Meteorological variables are generally used to assess the impact of climatic characteristics in many phenomena including human as well as non-human activities. Although often regarded as discrete or continuous measurements, these variables can benefit from fuzzy coding in some circumstances. Examples include cases in which these variables are imprecisely coded (e.g., when data are available in terms of intervals or linguistic categories) or when they are derived from a variety of sources (e.g., sam-



**Fig. 4** Application 2: Fuzzy frequency matrix for the pair PRE, HUM. Note that each cell contains a fuzzy natural number  $\tilde{n}_{r_C}$  for a specific combination of the  $R \times C$  granulation space

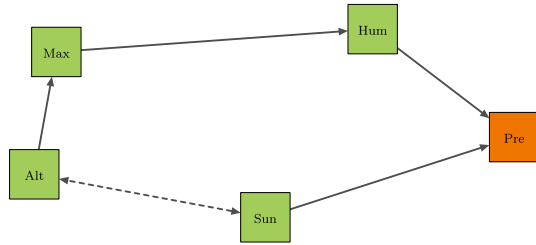
ples, historical databases, experts) that need to be integrated before being used for data analysis [8, 16]. In this application, we consider the analysis of  $J = 5$  meteorological variables (i.e., SUN: daily hours of sunshine; HUM: percentage of humidity; PRE: precipitations; ALT: altitude; MAX: maximum daily temperature) which were collected in 40 cities of Turkey during 2004 [2]. Data were originally coded using  $R = C = 3$  fuzzy triangular categories ( $G_0$ : minimum;  $G_1$ : medium;  $G_2$ : maximum) and membership grades  $\epsilon_1^{(j)}, \epsilon_2^{(j)}, \epsilon_3^{(j)}$   $j = 1, \dots, 5$  constitute the input data for the subsequent analysis. The aim is to explore the effects of climatic variables on rainfall (PRE) by means of a path analysis model. Likewise for the first application, the first step consisted in computing the fuzzy frequency matrix  $\widehat{N}_{3 \times 3}$  for each pair of the five climatic variables given the observed membership degrees. Then, the LLCs matrix was estimated using the fuzzy EM algorithm. Figure 4 shows an example of fuzzy counts for the pair of variables PRE-HUM, whereas Table 8 reports the estimated correlations for the variables involved in the study. As expected, the results showed a certain level of association among the five climatic variables.

Once the LLCs matrix has been estimated, we proceeded by modeling the effects of the climatic variables on PRE via path analysis (see Fig. 5). In particular, we expected that a higher humidity (HUM) increased rainfall (PRE) and that sunshine duration (SUN) decreased the levels of precipitation (PRE). Similarly, we also expected an

**Table 8** Application 2: Latent linear correlation matrix estimated via Olsson’s two-stage fuzzy EM algorithm (the bootstrap standard errors are reported in parentheses)

	SUN	HUM	PRE	ALT	MAX
SUN	1.00000				
HUM	-0.73125 (0.23112)	1.00000			
PRE	-0.58327 (0.2998)	0.31726 (0.29214)	1.00000		
ALT	-0.21412 (0.25751)	-0.47867 (0.26487)	-0.92587 (0.26166)	1.00000	
MAX	0.43941 (0.32198)	-0.23214 (0.28778)	0.17976 (0.26014)	-0.58675 (0.28806)	1.00000

**Fig. 5** Application 2: Path model for the effect of the climatic variables on the response variable PRE. Note that straight lines represent direct effects, whereas dotted lines indicate correlations



**Table 9** Application 2: Estimated coefficients  $\hat{\beta}$  and residual variances  $\hat{\sigma}_\epsilon^2$  for the path model depicted in Fig. 5 along with the standard errors  $\hat{\sigma}_\beta^2$  of the estimates

Path	$\hat{\beta}$	$\hat{\sigma}_\beta^2$	$\hat{\sigma}_\epsilon^2$
HUM-PRE	0.1844	0.1386	0.7488
SUN-PRE	-0.3406	0.1386	0.9295
MAX-HUM	-0.2161	0.1544	0.6717
ALT-MAX	-0.5577	0.1312	0.975

indirect effect of altitude (ALT) on humidity (HUM) through temperatures (TEMP). The path model has been estimated on the LLCs matrix via maximum likelihood as implemented in the R library lavaan [64]. Overall, the estimated model showed a moderate fit ( $R^2 = 0.20$ ). The results (Table 9) highlighted that PRE increased as a function of HUM ( $\hat{\beta} = 0.1844$ ,  $\hat{\sigma}_\beta^2 = 0.1386$ ) and decreased as sunshine duration increased ( $\hat{\beta} = -0.3406$ ,  $\hat{\sigma}_\beta^2 = 0.1386$ ). Humidity was negatively related to temperature ( $\hat{\beta} = -0.2161$ ,  $\hat{\sigma}_\beta^2 = 0.1544$ ), which was in turn negatively associated with altitude ( $\hat{\beta} = -0.5577$ ,  $\hat{\sigma}_\beta^2 = 0.1312$ ) as expected .

## 6 Conclusions

In this article, we described a novel approach to estimate latent linear correlations (LLCs) when data are in the form of fuzzy frequency tables. In particular, we represented fuzzy counts in terms of generalized natural numbers first, and then we generalized the sample space of the standard LLCs model to cope with fuzzy counts while retaining its parameter space as non-fuzzy. The resulting model encapsulated both random and non-random/imprecision components in a unified statistical representation. Since the inferential interest is on estimating the latent correlation matrix of the observed variables, parameter estimation was performed via fuzzy maximum likelihood using the expectation–maximization algorithm. A simulation study and two real applications were developed to highlight the characteristics of the fuzzy LLCs model. Overall, the simulation results revealed that the fuzzy LLCs model showed more accurate results in estimating the true correlation matrix as opposed to standard methods which can be applied on defuzzified data. The applications showed how the proposed method can be of particular value in situations involving fuzzy classification and fuzzy coding as well.

A particular advantage of the fuzzy LLCs model is its simplicity and ability to deal with situations involving imprecise classification problems. Moreover, the proposed method works with both fuzzy observations/crisp categories and crisp observations/fuzzy categories and, as such, it includes the standard crisp observations/crisp categories as a special case. Again, the fuzzy LLCs model does not require the extension of its parametric representation to account for fuzzy frequency data and consequently, parameter estimation and inference can be performed using the asymptotic properties of maximum likelihood theory. This is quite convenient and obviates the need of generalizing LLCs-based statistical modeling—such as structural equation models and factor analysis—to the fuzzy case. A limitation of the proposed approach is that it is based on the simplest, but still used, assumption of Gaussianity for LLCs. Although it has been proved that the assumption holds in several empirical contexts, there may be the need of LLCs based on more general probabilistic models (e.g., skew-Gaussian, elliptical,  $t$ , copula-based). As a result, the problems already identified by other researchers, for instance, bias in estimating the asymptotic covariance matrix of the LLCs matrix [28, 56], still persist in the fuzzy case. The fuzzy bootstrap technique used to approximate the covariance matrix of the fuzzy polychoric matrix might constitute an additional limitation of the current study. Indeed, although it provides a computational solution to calculate standard errors and CIs, it might suffer from the curse of dimensionality (e.g., in the case of a higher number of variables or response categories) as well as from a larger variance in the estimates. This is a well-known issue in the fuzzy statistics literature (e.g., see [33]), and it is particularly due to the fact that fuzzy bootstrap techniques handle with two sources of variability simultaneously, i.e., one related to the randomness of the estimator and the second related to the effect of the fuzziness in the data.

There are a number of further extensions to this project that can be undertaken in future research studies. For instance, the use of more general probabilistic model would extend the proposed method to handle with situations involving violations of Gaussianity assumption. In this line, further investigations should be undertaken to study the problem of deriving asymptotically efficient estimators for covariance matrix and standard error, for instance, by obtaining a fuzzy generalization of the Louis' method [53]. Similarly to the non-fuzzy case, this is still an open question. At the same time, building interval estimators for the polychoric fuzzy estimator—beyond the point-wise solution described in this article—might constitute a further generalization of the findings of the present study. Another aspect which might be interesting to investigate is the case where data need to be represented using more general fuzzy numbers (e.g., beta, exponential, Gaussian), which would allow the proposed method to cope with cases requiring more flexible models to represent non-random imprecision. Further, studying the properties of fuzzy LLCs-based statistical models like structural equation modeling or factor analysis would also constitute a research topic to be considered in a further study. Finally, neutrosophic-based generalizations of the proposed LLC statistic might also be a further research line to be investigated (e.g., see [6, 65]).

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40304-022-00295-6>.

**Acknowledgements** The author wishes to acknowledge and thank Dr. Andrea Spirito for his valuable comments on various issues concerning this research study.

**Funding** No funding was received for conducting this study.

## Declarations

**Data and Code Availability** The algorithms and datasets used during the current study are available in the GitHub repository <https://github.com/antcalcagni/fuzzypolychoric>.

**Conflict of interest** The author has no relevant financial or non-financial interests to disclose.

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

## Appendix A

To compute the nonlinear expectation  $\mathbb{E}_{\theta'}[\ln N_{rc}^{jk} | \tilde{n}_{rc}^{jk}]$ , we first approximate the factorial term via Stirling’s formula:

$$\begin{aligned} \mathbb{E}_{\theta'}[\ln N_{rc}^{jk} | \tilde{n}_{rc}^{jk}] &= \mathbb{E}_{\theta'}\left[N_{rc}^{jk} \ln N_{rc}^{jk} - N_{rc}^{jk} + \frac{1}{2} \ln 2\pi \mid \tilde{n}_{rc}^{jk}\right] \\ &= \mathbb{E}_{\theta'}\left[N_{rc}^{jk} \ln N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right] - \mathbb{E}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right] \\ &= \mathbb{E}_{\theta'}\left[g\left(N_{rc}^{jk}\right) \mid \tilde{n}_{rc}^{jk}\right] - \mathbb{E}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right] \end{aligned}$$

with  $g(x) := x \ln x$ . Next, since the nonlinear transformation  $g(\cdot)$  is smooth and twice-differentiable on  $(0, \infty)$  with  $g''(x) = 1/x$ , a second-order Taylor expansion around the first conditional moment  $\mathbb{E}_{\theta'}[N_{rc}^{jk} | \tilde{n}_{rc}^{jk}]$  can be developed to get the closed-form expression of the expectation term:

$$\mathbb{E}_{\theta'}\left[g\left(N_{rc}^{jk}\right) \mid \tilde{n}_{rc}^{jk}\right] \approx g\left(\mathbb{E}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right]\right) + \frac{\text{Var}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right]}{2\mathbb{E}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right]} \tag{7.1}$$

with the conditional variance being defined by

$$\text{Var}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right] = \sum_{n \in \mathbb{N}_0} \left(n - \mathbb{E}_{\theta'}\left[N_{rc}^{jk} \mid \tilde{n}_{rc}^{jk}\right]\right)^2 p_{N_{rc}^{jk} | \tilde{n}_{rc}^{jk}}(n; \pi_{rc}^{jk}(\theta')),$$

where  $\mathbb{E}_{\theta'}[N_{rc}^{jk} | \tilde{n}_{rc}^{jk}]$  is as in Eq. (3.8).

## Appendix B

To establish monotonicity for a sequence of log-likelihood evaluations  $\{\ln \mathcal{L}(\boldsymbol{\theta}^{(q)}; \tilde{\mathbf{N}})\}_{q \in \mathbb{N}}$  of the fuzzy expectation–maximization algorithm, we will follow the general results of [54], Sect. 3.2. A similar proof is also given by [80] for the case of rectangular fuzzy numbers (i.e., interval-valued data). In what follows, we will omit the indices  $j, k$  for the sake of simplicity. Given  $\boldsymbol{\theta}' = \boldsymbol{\theta}^{(q-1)}$  and by rearranging Eq. (3.6), we get by standard calculus:

$$\begin{aligned}
 & \ln \mathcal{L}(\boldsymbol{\theta}; \tilde{\mathbf{N}}) \\
 &= \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) - \sum_{r=1}^R \sum_{c=1}^C \ln p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta})) \\
 &= \mathbb{E}_{\boldsymbol{\theta}'} \left[ \sum_{r=1}^R \sum_{c=1}^C \ln \mathcal{L}(\boldsymbol{\theta}; N_{rc}) \Big| \tilde{n}_{rc} \right] - \mathbb{E}_{\boldsymbol{\theta}'} \left[ \sum_{r=1}^R \sum_{c=1}^C \ln p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta})) \Big| \tilde{n}_{rc} \right] \\
 &= \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}') - \mathcal{S}(\boldsymbol{\theta}; \boldsymbol{\theta}'). \tag{7.2}
 \end{aligned}$$

Then, an increasing of the observed log-likelihood can be written in terms of the result (7.2) as follows:

$$\begin{aligned}
 & \ln \mathcal{L}(\boldsymbol{\theta}^{(q)}; \tilde{\mathbf{N}}) - \ln \mathcal{L}(\boldsymbol{\theta}'; \tilde{\mathbf{N}}) \\
 & \geq \left( \mathcal{Q}(\boldsymbol{\theta}^{(q)}; \boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}'; \boldsymbol{\theta}') \right) - \left( \mathcal{S}(\boldsymbol{\theta}^{(q)}; \boldsymbol{\theta}') - \mathcal{S}(\boldsymbol{\theta}'; \boldsymbol{\theta}') \right).
 \end{aligned}$$

Note that because  $\boldsymbol{\theta}^{(q)}$  is chosen so that  $\mathcal{Q}(\boldsymbol{\theta}^{(q)}; \boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}'; \boldsymbol{\theta}') \geq 0$  [54], the condition  $\mathcal{S}(\boldsymbol{\theta}; \boldsymbol{\theta}') - \mathcal{S}(\boldsymbol{\theta}'; \boldsymbol{\theta}') \leq 0$  must hold for each  $\boldsymbol{\theta}$ . To do so, we proceed as follows:

$$\begin{aligned}
 & \mathcal{S}(\boldsymbol{\theta}; \boldsymbol{\theta}') - \mathcal{S}(\boldsymbol{\theta}'; \boldsymbol{\theta}') \\
 &= \sum_{r=1}^R \sum_{c=1}^C \mathbb{E}_{\boldsymbol{\theta}'} \left[ \ln p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta})) \Big| \tilde{n}_{rc} \right] - \mathbb{E}_{\boldsymbol{\theta}'} \left[ \ln p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}')) \Big| \tilde{n}_{rc} \right] \\
 &= \sum_{r=1}^R \sum_{c=1}^C \mathbb{E}_{\boldsymbol{\theta}'} \left[ \ln \left( \frac{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}'))} \right) \Big| \tilde{n}_{rc} \right] \\
 &\leq \sum_{r=1}^R \sum_{c=1}^C \ln \mathbb{E}_{\boldsymbol{\theta}'} \left[ \frac{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}'))} \Big| \tilde{n}_{rc} \right] \quad \text{using Jensen's inequality} \\
 &\leq \sum_{r=1}^R \sum_{c=1}^C \ln \sum_{n \in \mathbb{N}_0} \frac{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}{p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}'))} p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}')) n \\
 &\leq \sum_{r=1}^R \sum_{c=1}^C \ln \sum_{n \in \mathbb{N}_0} p_{N_{rc}|\tilde{n}_{rc}}(n; \pi_{rc}(\boldsymbol{\theta})) n
 \end{aligned}$$



$$\leq \sum_{r=1}^R \sum_{c=1}^C \ln 1 = 0 \quad \square$$

Hence, an increasing of  $\ln \mathcal{L}(\theta^{(q)}; \tilde{\mathbf{N}}) - \ln \mathcal{L}(\theta'; \tilde{\mathbf{N}}) \geq 0$  is guaranteed as soon as  $Q(\theta^{(q)}; \theta') - Q(\theta'; \theta') \geq 0$ .

## References

1. Agresti, A.: *Categorical Data Analysis*, vol. 482. Wiley, Hoboken (2003)
2. Asan, Z., Greenacre, M.: Measures of fit in multiple correspondence analysis of crisp and fuzzy coded data. Available at SSRN 1107815 (2008)
3. Aşan, Z., Greenacre, M.: Biplots of fuzzy coded data. *Fuzzy Sets Syst.* **183**(1), 57–71 (2011)
4. Aslam, M.: Chi-square test under indeterminacy: an application using pulse count data. *BMC Med. Res. Methodol.* **21**(1), 1–5 (2021)
5. Aslam, M.: Neutrosophic statistical test for counts in climatology. *Sci. Rep.* **11**(1), 1–5 (2021)
6. Aslam, M., Sherwani, R.A.K., Saleem, M.: Vague data analysis using neutrosophic jarque-bera test. *PLoS ONE* **16**(12), e0260689 (2021)
7. Augustin, T., Coolen, F.P., De Cooman, G.: *Introduction to Imprecise Probabilities*. Wiley, Hoboken (2014)
8. Blasius, J., Greenacre, M.: *Visualization and Verbalization of Data*. CRC Press, Boca Raton (2014)
9. Bodjanova, S., Kalina, M.: Cardinalities of granules of vague data. In: Magdalena, L., Ojeda-Aciego, M., Verdegay, J.L. (eds.) *Proceedings of IPMU2008, Torreliminos (Malaga), June 22–27 2008*, pp. 63–70 (2008)
10. Bodjanova, S.: A generalized histogram. *Fuzzy Sets Syst.* **116**(2), 155–166 (2000)
11. Bonanomi, A., Ruscone, M.N., Osmetti, S.A.: The polychoric ordinal alpha, measuring the reliability of a set of polytomous ordinal items. In: *SIS 2013 Conference: Advances in latent variables: Methods, models and applications*, Brescia, Italy, pp. 19–21. Citeseer (2013)
12. Calcagni, A., Lombardi, L.: Dynamic fuzzy rating tracker (dyfrat): a novel methodology for modeling real-time dynamic cognitive processes in rating scales. *Appl. Soft Comput.* **24**, 948–961 (2014)
13. Calcagni, A., Lombardi, L., Pascali, E.: Non-convex fuzzy data and fuzzy statistics: a first descriptive approach to data analysis. *Soft. Comput.* **18**(8), 1575–1588 (2014)
14. Casanovas, J., Torrens, J.: An axiomatic approach to fuzzy cardinalities of finite fuzzy sets. *Fuzzy Sets Syst.* **133**(2), 193–209 (2003)
15. Chakraborty, S., Chakravarty, D.: Discrete gamma distributions: properties and parameter estimations. *Commun. Stat. Theory Methods* **41**(18), 3301–3324 (2012)
16. Chevene, F., Doleadek, S., Chessel, D.: A fuzzy coding approach for the analysis of long-term ecological data. *Freshw. Biol.* **31**(3), 295–309 (1994)
17. Ciavolino, E., Salvatore, S., Calcagni, A.: A fuzzy set theory based computational model to represent the quality of inter-rater agreement. *Qual. Quant.* **48**(4), 2225–2240 (2014)
18. Coletti, G., Scozzafava, R.: Conditional probability, fuzzy sets, and possibility: a unifying view. *Fuzzy Sets Syst.* **144**(1), 227–249 (2004)
19. Da Roit, B., Weicht, B.: Migrant care work and care, migration and employment regimes: a fuzzy-set analysis. *J. Eur. Soc. Policy* **23**(5), 469–486 (2013)
20. Dan, J.R.G., Arnaldos, J., Darbra, R.M.: Introduction of the human factor in the estimation of accident frequencies through fuzzy logic. *Saf. Sci.* **97**, 134–143 (2017)
21. de Sáa, S.D.L.R., Gil, M.Á., González-Rodríguez, G., López, M.T., Lubiano, M.A.: Fuzzy rating scale-based questionnaires and their statistical analysis. *IEEE Trans. Fuzzy Syst.* **23**(1), 111–126 (2014)
22. Delgado, M., Gonzalez, A.: An inductive learning procedure to identify fuzzy systems. *Fuzzy Sets Syst.* **55**(2), 121–132 (1993)
23. Demertzis, K., Iliadis, L.S., Anezakis, V.-D.: An innovative soft computing system for smart energy grids cybersecurity. *Adv. Build. Energy Res.* **12**(1), 3–24 (2018)
24. Dencœux, T.: Maximum likelihood estimation from fuzzy data using the em algorithm. *Fuzzy Sets Syst.* **183**(1), 72–91 (2011)
25. Diccicio, T.J., Romano, J.P.: A review of bootstrap confidence intervals. *J. R. Stat. Soc. Ser. B (Methodological)* **50**(3), 338–354 (1988)

26. Dou, W., Ren, Y., Qian, W., Ruan, S., Chen, Y., Bloyet, D., Constans, J.-M.: Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing* **70**(4–6), 726–734 (2007)
27. Dubois, D., Prade, H.: *Fundamentals of Fuzzy Sets*, vol. 7. Springer, Berlin (2012)
28. Foldnes, N., Grønneberg, S.: Pernicious polychorics: the impact and detection of underlying non-normality. *Struct. Equ. Model.* **27**(4), 525–543 (2020)
29. Gil, M.A., López, M.T., Gil, P.: Comparison between fuzzy information systems. *Kybernetes* (1984)
30. Gil, M.A., López-Díaz, M., Ralescu, D.A.: Overview on the development of fuzzy random variables. *Fuzzy Sets Syst.* **157**(19), 2546–2557 (2006)
31. Greenacre, M.: Fuzzy coding in constrained ordinations. *Ecology* **94**(2), 280–286 (2013)
32. Grzegorzewski, P.: Distribution-free tests for vague data. In: *Soft Methodology and Random Information Systems*, pp. 495–502. Springer, Berlin (2004)
33. Grzegorzewska, P., Romaniuka, M.: Epistemic bootstrap for fuzzy data. In: *19th World Congress of the International Fuzzy Systems Association (IFSA), 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and 11th International Summer School on Aggregation Operators (AGOP)*, pp. 538–545. Atlantis Press (2021)
34. Hanss, M.: *Applied Fuzzy Arithmetic*. Springer, Berlin (2005)
35. Higham, N.J.: Computing the nearest correlation matrix—a problem from finance. *IMA J. Numer. Anal.* **22**(3), 329–343 (2002)
36. Hryniewicz, O.: Goodman-Kruskal  $\gamma$  measure of dependence for fuzzy ordered categorical data. *Comput. Stat. Data Anal.* **51**(1), 323–334 (2006)
37. Inés, C., Dubois, D.: Statistical reasoning with set-valued information: ontic vs. epistemic views. *Int. J. Approx. Reason.* **55**(7), 1502–1518 (2014)
38. Jadon, R.S., Chaudhury, S., Biswas, K.K.: A fuzzy theoretic approach for video segmentation using syntactic features. *Pattern Recogn. Lett.* **22**(13), 1359–1369 (2001)
39. Jin, S., Yang-Wallentin, F.: Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika* **82**(1), 67–85 (2017)
40. Jones, P.N., Geoffrey, M.J.: Improving the convergence rate of the em algorithm for a mixture model fitted to grouped truncated data. *J. Stat. Comput. Simul.* **43**(1–2), 31–44 (1992)
41. Jöreskog, K.G.: On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika* **59**(3), 381–389 (1994)
42. Kahraman, C., Bozdog, C.E., Ruan, D., Fahri Özok, A.: Fuzzy sets approaches to statistical parametric and nonparametric tests. *Int. J. Intell. Syst.* **19**(11), 1069–1087 (2004)
43. Kirilenko, A.P., Stepchenkova, S.: Inter-coder agreement in one-to-many classification: fuzzy kappa. *PLoS ONE* **11**(3), e0149787 (2016)
44. Knol, D.L., ten Berge, J.M.F.: Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika* **54**(1), 53–61 (1989)
45. Kolenikov, S., Angeles, G.: Socioeconomic status measurement with discrete proxy variables: is principal component analysis a reliable answer? *Rev. Income Wealth* **55**(1), 128–165 (2009)
46. Lee, S., Lee, J.-H., Lee, K.-M., Youn, H.Y.: Fuzzy category and fuzzy interest for web user understanding. In: *International Conference on Computational Science and Its Applications*, pp. 1149–1158. Springer (2005)
47. Lee, S.-Y., Lam, M.-L.: Estimation of polychoric correlation with elliptical latent variables. *J. Stat. Comput. Simul.* **30**(3), 173–188 (1988)
48. Lee, S.-Y., Poon, W.-Y.: Two-step estimation of multivariate polychoric correlation. *Commun. Stat. Theory Methods* **16**(2), 307–320 (1987)
49. Lee, S.-Y., Shi, J.-Q.: Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics* **57**(3), 787–794 (2001)
50. Lee, C.-T., Zhang, G., Edwards, M.C.: Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivar. Behav. Res.* **47**(2), 314–339 (2012)
51. Lorenzo-Seva, U., Ferrando, P.J.: Not positive definite correlation matrices in exploratory item factor analysis: causes, consequences and a proposed solution. *Struct. Equ. Model. Multidiscip. J.* **28**(1), 138–147 (2021)
52. Lotfi Asker Zadeh: Probability measures of fuzzy events. *J. Math. Anal. Appl.* **23**(2), 421–427 (1968)
53. Louis, T.A.: Finding the observed information matrix when using the em algorithm. *J. R. Stat. Soc. Ser. B (Methodological)* **44**(2), 226–233 (1982)
54. McLachlan, G.J., Krishnan, T.: *The EM Algorithm and Extensions*, vol. 382. Wiley, Hoboken (2007)
55. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, Hoboken (2004)

56. Monroe, S.: Contributions to estimation of polychoric correlations. *Multivar. Behav. Res.* **53**(2), 247–266 (2018)
57. Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**(1), 115–132 (1984)
58. Muthén, B.O., Satorra, A.: Technical aspects of muthén's liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika* **60**(4), 489–503 (1995)
59. Olsson, U.: Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**(4), 443–460 (1979)
60. Petry, K., Kuppens, S., Vos, P., Maes, B.: Psychometric evaluation of the dutch version of the mood, interest and pleasure questionnaire (mipq). *Res. Dev. Disabil.* **31**(6), 1652–1658 (2010)
61. Price, P.B., Jones, E.E.: Examining the alliance using the psychotherapy process q-set. *Psychother. Theory Res. Pract. Train.* **35**(3), 392 (1998)
62. Quost, B., Denoeux, T.: Clustering and classification of fuzzy data using the fuzzy em algorithm. *Fuzzy Sets Syst.* **286**, 134–156 (2016)
63. Roscino, A., Pollice, A.: A generalization of the polychoric correlation coefficient. In: *Data Analysis, Classification and the Forward Search*, pp. 135–142. Springer (2006)
64. Rosseel, Y.: Lavaan: an r package for structural equation modeling and more: version 0.5–12 (beta). *J. Stat. Softw.* **48**(2), 1–36 (2012)
65. Sherwani, R.A.K., Iqbal, S., Abbas, S., Aslam, M. and AL-Marshadi, A.H.: A new neutrosophic negative binomial distribution: properties and applications. *J. Math.* **2021** (2021)
66. Shiina, K., Ueda, T., Kubo, S.: Polychoric correlations for ordered categories using the em algorithm. In: *The Annual Meeting of the Psychometric Society*, pp. 247–259. Springer (2017)
67. Silvia, A.B.M.N.R., Osmetti, A.: Reliability measurement for polytomous ordinal items: the empirical polychoric ordinal alpha. *Quaderni di Statistica*, **14** (2012)
68. Song, X.-Y., Lee, S.-Y.: Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivar. Behav. Res.* **38**(1), 57–79 (2003)
69. Taheri, S.M., Hesamian, G., Viertl, R.: Contingency tables with fuzzy information. *Commun. Stat. Theory Methods* **45**(20), 5906–5917 (2016)
70. Tóth, Z.E., Jónás, T., Dénes, R.V.: Applying flexible fuzzy numbers for evaluating service features in healthcare-patients and employees in the focus. *Total Qual. Manag. Bus. Excel.* **30**(sup1), S240–S254 (2019)
71. Trutschnig, W.: A strong consistency result for fuzzy relative frequencies interpreted as estimator for the fuzzy-valued probability. *Fuzzy Sets Syst.* **159**(3), 259–269 (2008)
72. Viertl, R.: *Statistical Methods for Fuzzy Data*. Wiley, Hoboken (2011)
73. Vovan, T., Lethithu, T.: A fuzzy time series model based on improved fuzzy function and cluster analysis problem. *Commun. Math. Stat.* 1–16 (2020)
74. Wolodzko, T.: extraDistr: Additional Univariate and Multivariate Distributions, 2020. R package version 1.9.1
75. Wygralak, M.: Questions of cardinality of finite fuzzy sets. *Fuzzy Sets Syst.* **102**(2), 185–210 (1999)
76. Yager, R.R.: Generalized probabilities of fuzzy events from fuzzy belief structures. *Inf. Sci.* **28**(1), 45–62 (1982)
77. Yang, N.: East Asia in transition: re-examining the east Asian welfare model using fuzzy sets. *J. Asian Public Policy* **10**(1), 104–120 (2017)
78. Yang-Wallentin, F., Jöreskog, K.G., Luo, H.: Confirmatory factor analysis of ordinal variables with misspecified models. *Struct. Equ. Model.* **17**(3), 392–423 (2010)
79. Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. In: *Computational Linguistics*, pp. 149–184. Elsevier (1983)
80. Zhi-Gang, S., Wang, P.-H., Li, Y.-G., Zhou, Z.-K.: Parameter estimation from interval-valued data using the expectation-maximization algorithm. *J. Stat. Comput. Simul.* **85**(2), 320–338 (2015)
81. Zumbo, B.D., Gadermann, A.M., Zeisser, C.: Ordinal versions of coefficients alpha and theta for likert rating scales. *J. Mod. Appl. Stat. Methods* **6**(1), 4 (2007)