Estimating latent linear correlations from fuzzy frequency tables

Antonio Calcagnì

DPSS, University of Padova

13th Scientific Meeting of the
Classification and Data Analysis Group (CLADAG)

September, 9-11 2021

Università degli Studi di Padova

Introduction

The latent linear correlation (a.k.a. **polychoric correlation**) is a measure of linear association commonly used when data are arranged in terms of **contingency tables**.

LLCs are frequently adopted for categorical data with the purpose of computing a **correlation** statistic useful for further analyses (e.g., **CFA**, **SEM**).

Unlike other association measures like Goodman-Kruskal's γ or Kendall's τ, the polychoric measure ρ uses a **latent probabilistic model** (e.g., *Gaussian*) as a back-end representation onto which the observed joint frequencies $\mathbf{N}$ of two categorical variables X and Y are mapped via the *Muthen's thresholds-based approach* [4].

Università degli Studi di Padova

Introduction

Sometimes contingency tables can show some degree of **fuzziness**.

This is most common when precise data are classified into imprecise categories (e.g., images or scenes classification, content analysis, human-based assessments) or, less common, when fuzzy data are classified using either precise or imprecise categories.

In all these cases, the observed counts $\mathbf{N} = (n_{11}, \ldots, n_{rc}, \ldots, n_{RC})$ in the classification grid are no longer natural numbers, but rather fuzzy numbers.

UNIVERSITÀ DEGLI STUDI DI PADOVA

Introduction

Consequently, estimating the association ρ between two variables (X, Y) given a fuzzy contingency table $\widetilde{\mathbf{N}}$ requires an appropriate generalization.

In this presentation, we will generalize the maximum likelihood-based polychoric estimator to deal with fuzzy frequency tables. We will focus on estimating ρ from a pair (X, Y) of variables (the generalization to a set of J variables is straightforward).

More technical details and extended results are available in [2].

Università degli Studi di Padova

To set the problem, let (X, Y) be a pair of real random variables with $\{(x, y)_1, \ldots, (x, y)_I\}$ being a sample of length I.

Then, consider two collections of **imprecise categories**

$$\mathcal{C}_X = (\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_r, \ldots, \tilde{\mathcal{C}}_R) \quad \text{and} \quad \mathcal{C}_Y = (\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_c, \ldots, \tilde{\mathcal{C}}_C)$$

through which the observed sample is subsequently classified. The categories are represented as **fuzzy numbers** (e.g., trapezoidal) over the support $\mathcal{A} \subset \mathbb{R}$ of (X, Y) via their membership functions, e.g. $\xi_{\tilde{\mathcal{C}}_r} : \mathcal{A} \to [0, 1]$.

Note: $\mathcal{C}_X \tilde{\times} \mathcal{C}_Y$ constitute a **fuzzy partition** of $\mathcal{A}$ in the sense of Ruspini [1].

Università degli Studi di Padova

The process of counting how many observations fall in the joint category $(\tilde{\mathcal{C}}_r, \tilde{\mathcal{C}}_c)$ give raise to a fuzzy set $\tilde{n}_{rc}$ with membership function $\xi_{\widetilde{n}_{rc}} : \mathbb{N}_0 \to [0,1]$.

This is a **generalized natural numbers** [6].

UNIVERSITÀ DEGLI STUDI DI PADOVA

Data

Fuzzy counts as generalized natural numbers

$\xi_{\widetilde{n}_{rc}}$ is defined following Bodjanova and Kalina's findings [1], which revolve around the **Zadeh's counting functions** [7]:

$$\xi_{\widetilde{n}_{rc}}(n) = \min\left(\mu_{\text{FLC}}(n), \mu_{\text{FGC}}(n)\right) \qquad n = 0, 1, \dots$$

$\mu_{\text{FLC}}(n) = \text{FLC}(\boldsymbol{\epsilon}_{rc})$ possibility that at least n observations are classified in $(\tilde{\mathcal{C}}_r, \tilde{\mathcal{C}}_c)$

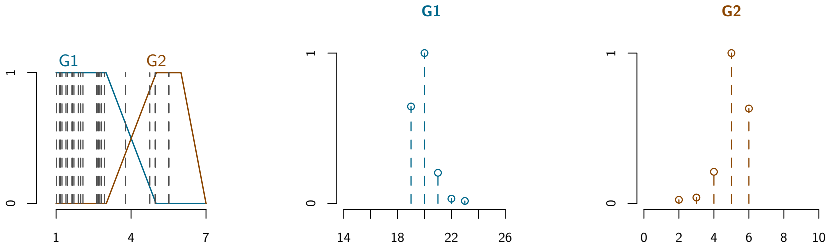
$\mu_{\text{FGC}}(n) = \text{FGC}(\boldsymbol{\epsilon}_{rc})$ possibility that at most n observations are classified in $(\tilde{\mathcal{C}}_r, \tilde{\mathcal{C}}_c)$

where $\boldsymbol{\epsilon}_{rc}$ is the array for the joint degree of inclusion of the observations w.r.t. $(\tilde{\mathcal{C}}_r, \tilde{\mathcal{C}}_c)$.

Note: The degree of inclusion $\epsilon_{\tilde{A},\tilde{B}}$ between two fuzzy sets $\tilde{A}$ and $\tilde{B}$ is computed as:
$\epsilon_{\tilde{A},\tilde{B}} = \text{card}\left(\min_x \xi_{\widetilde{A}}(x), \xi_{\widetilde{B}}(x)\right) / \max\left(1, \text{card}(\tilde{A})\right)$

Data

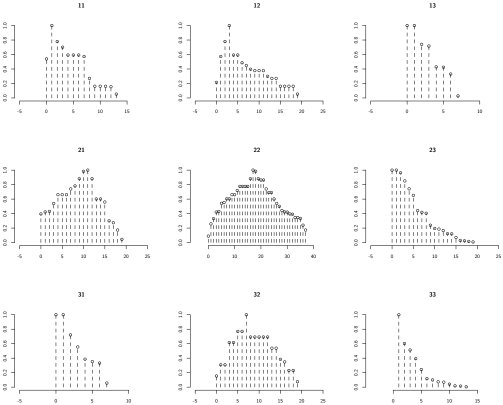
An example of fuzzy counts



Leftmost panel: Crisp observations (dashed black lines) along with two fuzzy categories `G1` and `G2`. **Center/Rightmost panels**: Fuzzy counts for both `G1` and `G2` categories. Note that in all the panels, fuzzy membership functions are represented along the vertical axes.

Data

An example of fuzzy contingency table



Graphical representation of a 3×3 **fuzzy contingency table** for a pair of variables.

Statistical model

Model definition

As for the non-fuzzy case, the standard LLC-based statistical model is mean-zero unit-variance **bivariate Normal** with correlation ρ:

$$(X^*, Y^*) \sim \mathcal{N}_2(x, y; \rho)$$

which relates to the observed sample through the following condition:

$$\underbrace{(x_i^{\text{obs}} \in \tilde{\mathcal{C}}_r) \wedge (y_i^{\text{obs}} \in \tilde{\mathcal{C}}_c)}_{\text{fuzzy counting}} \iff (X^*, Y^*) \in \underbrace{(\tau_{r-1}^X, \tau_r^X] \times (\tau_{c-1}^Y, \tau_c^Y]}_{\text{rectangles on the latent domain}}$$

with $\tau_0^X = \tau_0^Y = -\infty$ and $\tau_R^X = \tau_C^Y = \infty$ for $r = 1, \ldots, R$ and $c = 1, \ldots, C$.

Parameters to be estimated: $\boldsymbol{\theta} = \{\rho, \boldsymbol{\tau}^X, \boldsymbol{\tau}^Y\} \in [-1, 1] \times \mathbb{R}^{R-1} \times \mathbb{R}^{C-1}$

UNIVERSITÀ DEGLI STUDI DI PADOVA

The estimation procedure is performed by coupling the **fuzzy-EM algorithm** [3] to the **Olsson's two-stage** ML procedure [5]. The log-likelihood function is:

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) \propto \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc} \ln \int_{\tau_{r-1}^{X}}^{\tau_r^X} \int_{\tau_{c-1}^{Y}}^{\tau_c^Y} f_{X,Y}(x, y; \rho) \; dxdy$$

The estimation procedure is performed by coupling the **fuzzy-EM algorithm** [3] to the **Olsson's two-stage** ML procedure [5] on the likelihood function:

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) \propto \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc} \ln \int_{\tau_{r-1}^{X}}^{\tau_{r}^{X}} \int_{\tau_{c-1}^{Y}}^{\tau_{c}^{Y}} f_{X,Y}(x, y; \rho)\, dxdy$$

Given a candidate $\boldsymbol{\theta}'$, the algorithm iterates between:

- **E-step**
  Computing $\mathbb{E}_{\boldsymbol{\theta}'}\left[\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) | \widetilde{\mathbf{N}}\right]$ with

  $\hat{n}_{rc} = \mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc} | \tilde{n}_{rc}\right] = \sum_{n \in \mathbb{N}_0} n \frac{\xi_{\tilde{n}_{rc}}(n) f_{N_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}{\sum_{n \in \mathbb{N}_0} \xi_{\tilde{n}_{rc}}(n) f_{N_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}$ (fitered counts)

- **M-step**
  Maximizing $\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N})$ by replacing $\mathbf{N}$ with $\hat{\mathbf{N}}$

Università degli Studi di Padova

Statistical model

Parameter estimation

The estimation procedure is performed by coupling the **fuzzy-EM algorithm** [3] to the **Olsson's two-stage** ML procedure [5] for the likelihood function:

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) \propto \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc} \ln \int_{\tau_{r-1}^X}^{\tau_r^X} \int_{\tau_{c-1}^Y}^{\tau_c^Y} f_{X,Y}(x, y; \rho)\, dxdy$$

Given a candidate $\boldsymbol{\theta}'$, the algorithm iterates between:

- **E-step**

  Computing $\mathbb{E}_{\boldsymbol{\theta}'}\left[\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{N}) | \widetilde{\mathbf{N}}\right]$ with

$$\hat{n}_{rc} = \mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc} | \tilde{n}_{rc}\right]$$

$$= \sum_{n \in \mathbb{N}_0} n \boxed{\frac{\xi_{\widetilde{n}_{rc}}(n) f_{N_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}{\sum_{n \in \mathbb{N}_0} \xi_{\widetilde{n}_{rc}}(n) f_{N_{rc}}(n; \pi_{rc}(\boldsymbol{\theta}))}}$$

Density conditioned on fuzzy numbers

$$f_{N_{rc}}(n; \pi_{rc}(\boldsymbol{\theta})) = \mathcal{B}in(n; \pi_{rc}(\boldsymbol{\theta}))$$

A **simulation study** was run to assess the performances of the fuzzy-EM estimator for $\boldsymbol{\theta}$ against two naive Olsson's estimators based on mean-based and max-based defuzzification of the data.

- **Design**
  Three factors $I \in \{150, 250, 500\}$, $\rho \in \{0.15, 0.50, 0.85\}$, $R = C \in \{4, 6\}$ were varied in a complete factorial design with $B = 5000$ samples. Thresholds $\boldsymbol{\tau}^X = \boldsymbol{\tau}^Y$ were defined to be equidistant from -2 to 2.

UNIVERSITÀ DEGLI STUDI DI PADOVA

Simulation study

Design and procedure

- **Design**
  Three factors $I \in \{150, 250, 500\}$, $\rho \in \{0.15, 0.50, 0.85\}$, $R = C \in \{4, 6\}$ were varied in a complete factorial design with $B = 5000$ samples. Thresholds $\boldsymbol{\tau}^X = \boldsymbol{\tau}^Y$ were defined to be equidistant from -2 to 2.

- **Data generation**
  Two-step procedure:

  1. Non-fuzzy counts were generated via $n_{rc} = I\pi_{rc}(\boldsymbol{\theta})$
  2. Counts were fuzzified using a probability-possibility transformation based on discrete Gamma densities:
     $\boldsymbol{\xi}_{\tilde{n}_{rc}} = f_{\mathcal{G}_d}(\mathbf{n}; \alpha_{rc}, \beta_{rc}) \big/ \max f_{\mathcal{G}_d}(\mathbf{n}; \alpha_{rc}, \beta_{rc})$ *further details in [2]

Università degli Studi di Padova

- **Design**
  Three factors $I \in \{150, 250, 500\}$, $\rho \in \{0.15, 0.50, 0.85\}$, $R = C \in \{4, 6\}$ were varied in a complete factorial design with $B = 5000$ samples. Thresholds $\boldsymbol{\tau}^X = \boldsymbol{\tau}^Y$ were defined to be equidistant from -2 to 2.

- **Data generation**
  Two-step procedure:

  1. Non-fuzzy counts were generated via $n_{rc} = I\pi_{rc}(\boldsymbol{\theta})$
  2. Counts were fuzzified using a probability-possibility transformation based on discrete Gamma densities:
     $\boldsymbol{\xi}_{\tilde{n}_{rc}} = f_{\mathcal{G}_{\mathrm{d}}}(\mathbf{n}; \alpha_{rc}, \beta_{rc}) \big/ \max f_{\mathcal{G}_{\mathrm{d}}}(\mathbf{n}; \alpha_{rc}, \beta_{rc})$ *further details in [2]

- **Outcome measures**
  Bias of estimates and RMSE.

Università degli Studi di Padova

Simulation study

A sketch of the results

$R = C = 4$	fEM		dML-max		dML-mean	
	bias	*rmse*	*bias*	*rmse*	*bias*	*rmse*
$\rho = 0.15$						
$I = 150$	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
$\rho = 0.50$						
$I = 150$	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
$\rho = 0.85$						
$I = 150$	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

Results for ρ in the $R = C = 4$ case. Note that `fEM` indicates the fuzzy estimator whereas `dML-max` and `dML-mean` indicate the naive estimators.

Università degli Studi di Padova

Simulation study

A sketch of the results

Overall, results indicate that the fuzzy estimator `fEM` outperformed the naive estimators `dML-max` and `dML-mean` in terms of bias and RMSE.

The results for the condition $R = C = 6$ largely resembled those obtained for simplest $R = C = 4$ case.

All the approaches showed similar results in estimating the thresholds $\{\boldsymbol{\tau}^X, \boldsymbol{\tau}^Y\}$ (note that the primary interest laid on estimating ρ).

For extended results, see [2].

Università degli Studi di Padova

Conclusions

- When data are represented in terms of fuzzy contingency tables, the standard ML estimators should be generalized to cope with this type of data.
- The proposed fuzzy-EM estimator works with both crisp observations/fuzzy categories and fuzzy observations/crisp or fuzzy categories. In this sense, it encompasses the standard crisp observations/crisp categories as a special case.
- Real-world applications of the fuzzy-EM estimator for polychoric correlations (e.g., inter-rater agreement) are further discussed in [2].

Università degli Studi di Padova

[1] Bodjanova, S., and Kalina, M.
Cardinalities of granules of vague data.
In *Proceedings of IPMU2008, Torreliminos (Malaga), June 22-27 2008* (2008), L. Magdalena, M. Ojeda-Aciego, and J. Verdegay, Eds., pp. 63–70.

[2] Calcagnì, A.
Estimating latent linear correlations from fuzzy frequency tables.
arXiv:2105.03309[stat.ME].

[3] Denœux, T.
Maximum likelihood estimation from fuzzy data using the em algorithm.
Fuzzy Sets and Systems 183, 1 (2011), 72–91.

[4] Muthén, B. O., and Satorra, A.
Technical aspects of muthén's liscomp approach to estimation of latent variable relations with a comprehensive measurement model.
Psychometrika 60, 4 (1995), 489–503.

[5] Olsson, U.
Maximum likelihood estimation of the polychoric correlation coefficient.
Psychometrika 44, 4 (1979), 443–460.

[6] Wygralak, M.
Questions of cardinality of finite fuzzy sets.
Fuzzy Sets and Systems 102, 2 (1999), 185–210.

[7] Zadeh, L. A.
A computational approach to fuzzy quantifiers in natural languages.
In *Computational linguistics*. Elsevier, 1983, pp. 149–184.

antonio.calcagni@unipd.it

Università degli Studi di Padova