ESTIMATING LATENT LINEAR CORRELATIONS FROM FUZZY CONTINGENCY TABLES

Antonio Calcagnì [1]

[1] DPSS, University of Padova, Italy (e-mail: antonio.calcagni@unipd.it)

ABSTRACT: In this contribution, we describe a method to estimate polychoric correlations when data are available in the form of fuzzy frequency tables. A simulation study is used to assess the characteristics of the proposed approach. Fuzzy polychoric correlations can be of particular utility, for instance, in studies involving covariance structural analysis (e.g., CFA) and dimensionality reduction techniques (e.g., EFA).

KEYWORDS: fuzzy frequencies, polychoric correlations, fuzzy classification

1 Introduction

The latent linear correlation (LLC), also called polychoric correlation, is a measure of linear association which is usually adopted when dealing with categorical variables or statistics such as frequency or contingency tables. Given a set of J variables, LLC is computed pairwise for each pair (j,k) of variables by considering their joint frequencies $\mathbf{N}_{R\times C}^{(j,k)} = (n_{11}^{(j,k)},\ldots,n_{rc}^{(j,k)},\ldots,n_{RC}^{(j,k)})$ over a $R_{jk}\times C_{jk}$ partition space of the variables' domain. The general idea is to adopt a bivariate Gaussian distribution with correlation ρ_{jk} as a latent statistical model underlying the observed frequency table $\mathbf{N}_{R\times C}^{(j,k)}$, which maps the $R_{jk}\times C_{jk}$ space to the real domain of the bivariate density via a threshold-based approach. There are several contexts in which LLCs have been applied, including covariance structural analysis (e.g., CFA) and dimensionality reduction techniques (e.g., PCA, EFA). In this contribution, we generalize the problem of estimating polychoric correlations from fuzzy frequency tables, which are of particular utility when observed data are classified using fuzzy categories as done, for example, in socio-economic studies, images/videos classification, and content analysis. In all these cases, the $R_{jk}\times C_{jk}$ space of the variables' domain constitutes a fuzzy partition and observed counts in $\mathbf{N}_{R\times C}^{(j,k)}$ are no longer natural numbers. In order to deal with this issue, in this paper we describe a novel way to compute fuzzy frequency tables and provide a way to estimate ρ_{jk} when observed frequencies are fuzzy. In what follows, we will set $R = C$ and $J = 2$ for the sake of simplicity.

2 Fuzzy frequencies

A fuzzy subset $\tilde{A}$ of a universal set $\mathcal{A}$ is defined by means of its characteristic function $\xi_{\tilde{A}} : \mathcal{A} \to [0,1]$. Let $\mathcal{A} \subset \mathbb{R}$ without loss of generality and consider (X,Y) a pair of random variables taking values on $\mathcal{A}$. Then $\mathcal{A}$ can conveniently be partitioned into a collection of fuzzy subsets, namely $\mathcal{C}_j = \{\tilde{C}_1,\ldots,\tilde{C}_r,\ldots,\tilde{C}_R\}$ and $\mathcal{C}_k = \{\tilde{C}_1,\ldots,\tilde{C}_c,\ldots,\tilde{C}_C\}$. The random realizations $\mathbf{x} = (x_1,\ldots,x_I)$ and $\mathbf{y} = (y_1,\ldots,y_I)$ can partially or fully be classified into $\mathcal{C}_j$ or $\mathcal{C}_k$. The evaluation of the amount of sample realizations over $\tilde{C}_j$ or $\tilde{C}_k$ is called *cardinality*. This is a natural number or crisp count (i.e., $n_{rc} \in \mathbb{N}_0$) when the observations fully belong to subsets of $\tilde{C}_j$ or $\tilde{C}_k$. On the opposite case, it is a fuzzy natural number $\tilde{n}_{rc} \in \mathcal{F}(\mathbb{N})$, with $\mathcal{F}(\mathbb{N})$ being the set of all *generalized natural numbers* (Bodjanova & Kalina, 2008). Let $\tilde{C}_{rc}$ be an element of the fuzzy Cartesian product $\tilde{C}_j \;\tilde{\times}\; \tilde{C}_k$. Then a fuzzy count $\tilde{n}_{rc}$ is a fuzzy set with membership function $\xi_{\tilde{n}_{rc}} : \mathbb{N}_0 \to [0,1]$ being computed as follows: $\xi_{\tilde{n}_{rc}}(n) = \min(\nu_{rc}(n), \mu_{rc}(n))$, with $\nu_{rc}(n) = \mathrm{FGC}(\boldsymbol{\varepsilon}_{rc})$ and $\mu_{rc}(n) = \mathrm{FLC}(\boldsymbol{\varepsilon}_{rc})\ \forall\, n \in \{0,1,\ldots,I\} \subset \mathbb{N}_0$. In this context, FGC(.) and FLC(.) are the fuzzy counting functions as defined by Zadeh (1983) whereas $\boldsymbol{\varepsilon}_{rc} = \min(\xi_{\tilde{C}_r}(\mathbf{x}_j), \xi_{\tilde{C}_c}(\mathbf{y}_k))$ contains the joint degrees of inclusion of the sample observations $\mathbf{x}$ and $\mathbf{y}$ w.r.t. the fuzzy categories. More details can be found in Bodjanova & Kalina (2008). Finally, the fuzzy frequency table $\widetilde{\mathbf{N}}_{R\times C}$ can be computed by applying the above calculus over $r = 1,\ldots,R$ and $c = 1,\ldots,C$.

3 LLCs for fuzzy frequency tables

The latent statistical model underlying the sample realizations is bivariate Gaussian $(X^*,Y^*) \sim \mathcal{N}(\mathbf{0},\rho)$ under the constraints that $(X \in \tilde{C}_r) \;\wedge\; (Y \in \tilde{C}_c)$ iif $(X^*,Y^*) \in (\tau^X_{r-1},\tau^X_r] \times (\tau^Y_{c-1},\tau^Y_c] \subset \mathbb{R}^2$ for all $r = 1,\ldots,R$ and $c = 1,\ldots,C$. The thresholds $\boldsymbol{\tau}^X$ and $\boldsymbol{\tau}^Y$ are defined so that $\tau_0 = -\infty$ and $\tau_R = \infty$ for both X and Y variables. Note that (X^*,Y^*) are unobserved pairs of latent variables. Following Olsson (1979), the parameters $\boldsymbol{\theta} = \{\rho,\boldsymbol{\tau}^X,\boldsymbol{\tau}^Y\} \in [-1,1] \times \mathbb{R}^{R-1} \times \mathbb{C}^{C-1}$ can be estimated using a two step-approach. In particular, given the filtered counts at the current iteration, thresholds are estimated using the cumulative marginals of $\widehat{\mathbf{N}}_{R\times C}$ (first step). Then, ρ is estimated by maximizing the log-likelihood implied by the model conditioned on $\hat{\boldsymbol{\tau}}^X$ and $\hat{\boldsymbol{\tau}}^Y$ (second step):

$$\ln \mathcal{L}(\boldsymbol{\theta};\mathbf{N}) \propto \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc} \ln \int_{\tau^X_{r-1}}^{\tau^X_r} \int_{\tau^Y_{c-1}}^{\tau^Y_c} \phi(x,y;\rho)\, dxdy \qquad (1)$$

with $\phi(x,y;\rho)$ being the bivariate Gaussian density centered at zero. In what follows, we will focus on estimating ρ as estimation of thresholds follows straightforwardly from Olsson (1979). As we observe fuzzy frequencies $\widetilde{\mathbf{N}}_{R\times C}$, we solve the maximization problem via the fuzzy EM algorithm proposed by Denoeux (2011), which in this case requires the computation of the following quantity:

$$\mathbb{E}_{\boldsymbol{\theta}'}\left[\ln \mathcal{L}(\boldsymbol{\theta};\mathbf{N})|\widetilde{\mathbf{N}}\right] \propto \sum_{r=1}^{R}\sum_{c=1}^{C} \mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right] \ln \int_{\tau^X_{r-1}}^{\tau^Y_r}\int_{\tau^Y_{c-1}}^{\tau^Y_c} \phi(x,y;\rho)\, dxdy \quad (2)$$

given a candidate estimate $\boldsymbol{\theta}'$. The quantity $N_{rc}|\tilde{n}_{r,c}$ is a random variable conditioned on a fuzzy event:

$$\mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right] = \sum_{n\in\mathbb{N}_0} \frac{\xi_{\tilde{n}_{rc}}(n) f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta}))}{\sum_{n\in\mathbb{N}_0}\xi_{\tilde{n}_{rc}}(n) f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta}))}\, n \quad (3)$$

where $f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta})) = \mathcal{B}in(n;\pi_{rc}(\boldsymbol{\theta}))$, with $\pi_{rc}(\boldsymbol{\theta}) = \int_{\tau^X_{r-1}}^{\tau^Y_r}\int_{\tau^Y_{c-1}}^{\tau^Y_c} \phi(x,y;\rho)\, dxdy$. Note that $\hat{n}_{rc} = \mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right]$ denotes the reconstructed rc-th count. The fuzzy EM algorithm proceeds by alternating between the computation of Eq. (3) and the maximization of Eq. (1) once $\hat{n}_{rc}$ has been obtained.

4 Simulation study

The aim of this Monte Carlo study is twofold. First, we will evaluate the performances of fuzzy-EM estimator for ρ_{jk} when fuzzy frequency data are available. Second, we will assess whether the standard maximum likelihood estimator for polychoric correlations performs as good as the proposed method if applied on max-based and mean-based defuzzified data. The case $J = 2$ was considered for the sake of simplicity.

Design. The design involved two factors, namely (i) $I \in \{150, 250, 500\}$, and (ii) $\rho \in \{0.15, 0.50, 0.85\}$, which were varied in a complete factorial design. For each combination, $B = 5000$ samples were generated.

Data generation. For each condition of the simulation design, data were generated according to a two-step procedure. First, a crisp frequency table $\mathbf{N}_{R\times C}$ was computed using the approximation $n_{rc} = I \cdot \pi_{rc}$ ($r = 1,\ldots,R$; $c = 1,\ldots,C$), with $\boldsymbol{\tau}^X = \boldsymbol{\tau}^Y = (-2,-1,0,1,2)$. Second, each element of $\mathbf{N}_{R\times C}$ was fuzzified via the following probability-possibility transformation: $\boldsymbol{\xi}_{\tilde{n}_{rc}} = f_{\mathcal{G}_\mathrm{d}}(\mathbf{n};\alpha_{rc},\beta_{rc}) \big/ \max f_{\mathcal{G}_\mathrm{d}}(\mathbf{n};\alpha_{rc},\beta_{rc})$, $\alpha_{rc} = 1 + m_1\beta_{s_1}$, $\beta_{s_1} = 1 + (m_1 + m_1^2 +$

$4s_1^2)^{\frac{1}{2}}/2s_1^2$, $\beta_{rc} = (m_1 + m_1^2 + 4s_1^2)^{\frac{1}{2}}/2s_1^2$, $m_1 \sim \mathcal{G}amma_{\mathrm{d}}(\alpha_{m_1}, \beta_{m_1})$ where $\alpha_{m_1} = 1 + n_{rc}\beta_{m_1}$, $\beta_{m_1} = (n_{rc} + n_{rc}^2 + 4s_1^2)^{\frac{1}{2}}/2s_1^2$, $s_1 \sim \mathcal{G}amma_{\mathrm{d}}(\alpha_{s_1}, \beta_{s_1})$, $\alpha_{s_1} = 1 + m_0\beta_{s_1}$, $\beta_{s_1} = (m_0 + m_0^2 + 4s_0^2)^{\frac{1}{2}}/2s_0^2$, $m_0 = 1$ and $s_0 = 0.15$. Note that $f_{\mathcal{G}_{\mathrm{d}}}$ is the density of the discrete Gamma random variable $\mathcal{G}amma_{\mathrm{d}}$.

Outcome measures. For each condition of the simulation design, sample results were evaluated using bias of estimates and root mean square error.

Results. Table 1 shows the results of the study. As expected, fEM outperformed standard ML applied on both max-based and mean-based defuzzified data in terms of bias and root mean square errors. This is mainly due to the fact that ρ_{fEM} estimator weights the observed fuzzy data $\xi_{\tilde{n}_{rc}}$ with the probabilistic model for the unobserved n_{rc}.

	fEM		dML (max)		dML (mean)	
	bias	*rmse*	*bias*	*rmse*	*bias*	*rmse*
$\rho = 0.15$						
$I = 150$	0.0358	0.0881	-0.0105	0.1142	-0.0402	0.0846
$I = 250$	0.0043	0.0514	-0.0284	0.0817	-0.0403	0.0683
$I = 500$	0.0099	0.0297	0.0020	0.0416	-0.0082	0.0335
$\rho = 0.50$						
$I = 150$	0.0103	0.0747	-0.0933	0.1545	-0.1797	0.1956
$I = 250$	-0.0363	0.0626	-0.1216	0.1488	-0.1706	0.1800
$I = 500$	-0.0006	0.0264	-0.0457	0.0689	-0.0828	0.0903
$\rho = 0.85$						
$I = 150$	0.0013	0.0441	-0.2150	0.2525	-0.3274	0.3354
$I = 250$	-0.0028	0.0269	-0.1707	0.1967	-0.2580	0.2642
$I = 500$	-0.0009	0.0145	-0.1034	0.1211	-0.1630	0.1672

Table 1. *Monte Carlo study: Estimating ρ via fuzzy-EM (fEM) and standard ML (dML) on max-based and mean-based defuzzified frequency tables.*

References

BODJANOVA, SLAVKA, & KALINA, MARTIN. 2008. Cardinalities of Granules of Vague Data. *Pages 63–70 of:* MAGDALENA, L., OJEDA-ACIEGO, M., & VERDEGAY, J.L. (eds), *Proceedings of IPMU2008, Torreliminos (Malaga), June 22-27 2008.*

DENOEUX, THIERRY. 2011. Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy sets and systems*, **183**(1), 72–91.

OLSSON, ULF. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, **44**(4), 443–460.

ZADEH, LOTFI A. 1983. A computational approach to fuzzy quantifiers in natural languages. *Pages 149–184 of: Computational linguistics*. Elsevier.