Contingency tables with structural zeros and discrete copulas

Tabelle di contingenza con zeri strutturali e copule discrete

Roberto Fontana[a], Elisa Perrone[b], and Fabio Rapallo[c]

[a]Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy, roberto.fontana@polito.it

[b]Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands, e.perrone@tue.nl

[c]Dipartimento DIEC, Università di Genova, via Vivaldi 5, 16126 Genova, Italy, fabio.rapallo@unige.it

Abstract

In this work, we analyze the connection between contingency table analysis and copulas in a discrete framework. We focus on the impact of structural zeros on the general theory presented by Geenens (2020) based on a new idea of copula models for discrete variables. Through examples, we investigate the pros and cons of applying the theory developed by Geenens (2020) and discuss some open questions for future research.

In questo lavoro si analizza la connessione tra l'analisi delle tabelle di contingenza e le copule per variabili casuali discrete. Ci si concentra in particolare sull'impatto degli zeri strutturali sulla teoria presentata da Geenens (2020), che è basata su una nuova idea di modelli copula per variabili discrete. Si illustrano alcuni esempi per indagare i pro e i contro dell'applicazione di tale teoria e per discutere alcuni problemi aperti di interesse per future ricerche.

Keywords: Bubble plot, Categorical data analysis, Discrete copulas, Iterated Proportional Fitting (IPF), Structural zeros

1. Introduction

Contingency tables appear in many applied fields, such as biology, health care, and social science. Due to their importance in application, the analysis of such tables was studied in statistics, where researchers developed methodological tools to extract information about the relation between the variables involved (9). Recent work by Geenens (4) highlights interesting connections between standard methods in contingency table analysis and *copulas* in a discrete setting. Copulas are popular tools to model dependencies between random variables. Their popularity is due to Sklar's theorem (11), which states that, for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$, the joint distribution function $F_{\mathbf{X}}$ of any d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can be written as $F_{\mathbf{X}}(x_1, \ldots, x_d) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$, where the function $C : [0,1]^d \to [0,1]$ is a *d-dimensional copula* and $F_{X_1}, \ldots, F_{X_d}$ are univariate marginal distributions. In a nutshell, copulas can be seen as joint probability distributions with uniform margins in $[0,1]$. When the random vector $\mathbf{X}$ is discrete, the copula associated with $F_{\mathbf{X}}$ is uniquely defined on the Range$(F_{X_1}) \times \cdots \times$ Range(F_{X_d}). Thus, it is possible to associate any contingency table with the

restriction of a full-domain copula on a grid domain, i.e., a so-called *discrete copula*; see, e.g., (6; 7) and references therein. Any discrete copula can be extended into a full-domain copula by simply spreading the probability mass on each hyper-rectangle of their grid domain. However, the extension is not unique (1), which leads to problems while applying copula theory in discrete settings. In this work, we elaborate on a novel approach to using copulas in contingency table analysis presented in (4). In Sect. 2, we recall basic definitions (in the bivariate case) and results on the topic with special attention to contingency tables with structural zeros, and we discuss the connections with some classical log-linear models for incomplete tables. Some interesting examples are illustrated in Sect. 3, where we also discuss some open questions for future research on the topic.

2. Background

In this section, we provide the mathematical framework and notation to work with discrete copulas. We consider $R \in \mathbb{Z}_{>0}$ and denote $I_R = \{0, 1/R, \ldots, (R-1)/R, 1\}$, $[R] = \{1, \ldots, R\}$, and $\langle R \rangle = \{0, \ldots, R\}$. For R and S in $\mathbb{Z}_{>0}$, we define $U_R = \{u_0 = 0, u_1, \ldots, u_{R-1}, u_R = 1\}$, $u_0 < \ldots < u_R$ and $V_S = \{v_0 = 0, v_1, \ldots, v_{S-1}, v_S = 1\}$, $v_0 < \ldots < v_S$ as two finite grid partitions of the unit interval. A discrete copula C_{U_R,V_S} is a function defined on $U_R \times V_S$ that satisfies the properties of a copula function on the grid domain $U_R \times V_S$. As highlighted in (6), there are interesting connections between the space of discrete copulas and convex polytopes called *transportation polytopes*, which are also linked with contingency tables analysis (2) . Namely, considering two vectors $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_R) \in \mathbb{R}^R_{>0}$ and $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_S) \in \mathbb{R}^S_{>0}$, we can define the transportation polytope $\mathcal{T}(\tilde{u}, \tilde{v})$ as the convex polytope in the RS variables $x_{i,j}$ satisfying, for all $i \in [R]$ and $j \in [S]$, the following conditions: $x_{i,j} \geq 0$, $\sum_{h=1}^{S} x_{i,h} = \tilde{u}_i$, $\sum_{\ell=1}^{R} x_{\ell,j} = \tilde{v}_j$. The two vectors $\tilde{u}$ and $\tilde{v}$ are called the margins of $\mathcal{T}(\tilde{u}, \tilde{v})$. In (8), the authors show that any discrete copula C_{U_R,V_S} corresponds to a matrix within a transportation polytope $\mathcal{T}(\tilde{u}, \tilde{v})$, and viceversa. Intuitively, the transportation matrix directly relates to the probability mass function of the discrete random vector, while the corresponding discrete copula relates to the cumulative distribution function.

We now show how to derive the discrete copula associated with a given contingency table. We consider an example taken from (10) whose data is reported in Table 1. The table shows the cross-classification of a father's and his son's occupational status categories in Japan in 1955. There are four categories (1: Professional and Managers; 2: Clerical and Sales; 3: Skilled manual and Semiskilled manual; 4: Unskilled manual and Farmers). Since this table is analyzed in (10) under quasi-symmetry models, we have removed the diagonal, because the diagonal cells are fitted exactly, and thus there is no variability. We get $N = 799$ values. In this example, $R = S = 4$, the vectors $\tilde{u}$ and $\tilde{v}$ are the margins of the contingency table, i.e., $\tilde{u} = (128, 136, 144, 391)$ and $\tilde{v} = (139, 301, 264, 95)$, while the defining grids of the corresponding discrete copula are $U_R = \frac{1}{N}\{0, \tilde{u}_1, \tilde{u}_1 + \tilde{u}_2, \ldots\} = \{0, 0.16, 0.33, 0.51, 1\}$ and $V_S = \{0, 0.17, 0.55, 0.88, 1\}$. The entries of the discrete copula $C_1 = C_{U_R,V_S} = (c_{i,j})$, $i \in [R]$ and $j \in [S]$ are computed from the entries of the contingency table $(x_{i,j})$ by summing up and normalizing, i.e., $c_{i,j} = \frac{1}{N} \sum_{\ell=1}^{i} \sum_{h=1}^{j} x_{\ell,h}$, while $c_{0,0} = c_{i,0} = c_{0,j} = 0$, for $i \in [R]$ and $j \in [S]$.

In (4), the author highlights the difficulty of drawing conclusions on the dependence from tables that have non-uniform margins as the one reported in Table 1. Therefore, in the spirit of copula theory for continuous random variables, the author suggests searching for a representative $\bar{\mathbf{p}}$ of all $(R \times S)$ probability distributions that (1) preserves the inter-dependencies of a contingency table in terms of odds-ratios, and (2) has uniform margins equal to $1/R$ and $1/S$. In discrete copula terms, this is equivalent to searching for a discrete copula defined on the rectangular grid $I_R \times I_S$ which preserves the dependence structure of the original discrete copula computed from a given contingency table. Looking at the example above, we would search for a discrete copula $\tilde{C}_1$ with support $I_4 \times I_4$ and margins I_4 associated with C_1 in a meaningful way. A natural question that arises is whether or not such an element exists and is unique. The answer to this question is given in the theorem below presented in (4). We here report the cases that are relevant to our examples while the original formulation of the theorem is more general and presents more scenarios. The cardinality of a set A is denoted by $|A|$.

Theorem 1. *Let* $\mathbf{p}$ *be in the set* $\mathcal{P}_{R\times S}$ *of all* $(R \times S)$ *probability distributions. We define Supp*$(\mathbf{p}) = \{(i,j) \in [R]\times[S]$ *s.t.* $p_{i,j} > 0\}$ *and N*$(\mathbf{p}) = \{(v_X \times v_Y) : v_X \subset [R], v_Y \subset [S] s.t. \sum_{(i,j)\in v_X \times v_Y} p_{i,j} = 0\}$, *the set of rectangular subset of* $[R] \times [S]$ *where* $\mathbf{p}$ *is null.*

1. *Suppose that for all* $(v_X \times v_Y) \in N(\mathbf{p})$, $\frac{|v_X|}{R} + \frac{|v_Y|}{S} < 1$, *then there exists a unique* $\bar{\mathbf{p}}$ *with uniform margins, same odds-ratio structure of* $\mathbf{p}$, *and associated with a discrete copula* $C_{I_R\times I_S}$.
2. *Suppose that there exists* $\tilde{\tilde{v}}_X \times \tilde{\tilde{v}}_Y \in N(\mathbf{p})$ *such that* $\frac{|\tilde{\tilde{v}}_X|}{R} + \frac{|\tilde{\tilde{v}}_Y|}{S} > 1$. *Then there is no element* $\bar{\mathbf{p}}$ *with uniform margins such that it has same odds-ratio structure of* $\mathbf{p}$ *and is associated with a discrete copula* $C_{I_R\times I_S}$.

The element $\bar{\mathbf{p}}$ can be obtained by using the iterated proportional fitting (IPF) procedure, which is a standard method in contingency table analysis for a meaningful comparison of tables with different margins and same dependence structure in terms of corresponding odds-ratios (4; 9). Theorem 1 states that zero-count cells in a contingency table impact the existence of the element $\bar{\mathbf{p}}$ of interest.

When a contingency table contains structural zeros, classical statistical methods become difficult to apply. Some of the odds-ratios can not be defined, and the standard independence model can not be used, because it implies a sample space in the form of a Cartesian product $[R] \times [S]$. For the analysis of incomplete square tables with structural zeros on the main diagonal, as in the example in Table 2, one can use quasi-independence or quasi-symmetry log-linear models. The latter is defined by

$$\log(p_{i,j}) = \lambda + \lambda_i^{(X)} + \lambda_j^{(Y)} + \lambda_{i,j}^{(XY)} \tag{1}$$

for $(i,j) \in [R] \times [S]$, where λ is a mean parameter, $\lambda_i^{(X)}$ are the row-parameters, $\lambda_j^{(Y)}$ are the column-parameters, and $\lambda_{i,j}^{(XY)}$, with the constraints $\lambda_{i,j}^{(XY)} = \lambda_{j,i}^{(XY)}$, measure the symmetry beyond the marginal contributions. Although the expression of the model is simple, the practical interpretation of the values of the λ parameters is not easy, and usually, log-linear models of this kind are used only for a global goodness-of-fit test. Using the discrete copulas and their graphical visualization through the bubble plots as the ones reported in Fig. 1, we will be able to focus on the dependence described by the $\lambda_{i,j}^{(XY)}$ parameters. In the next section, we further explore this aspect by analyzing two examples of contingency tables with special zero-count cell structures.

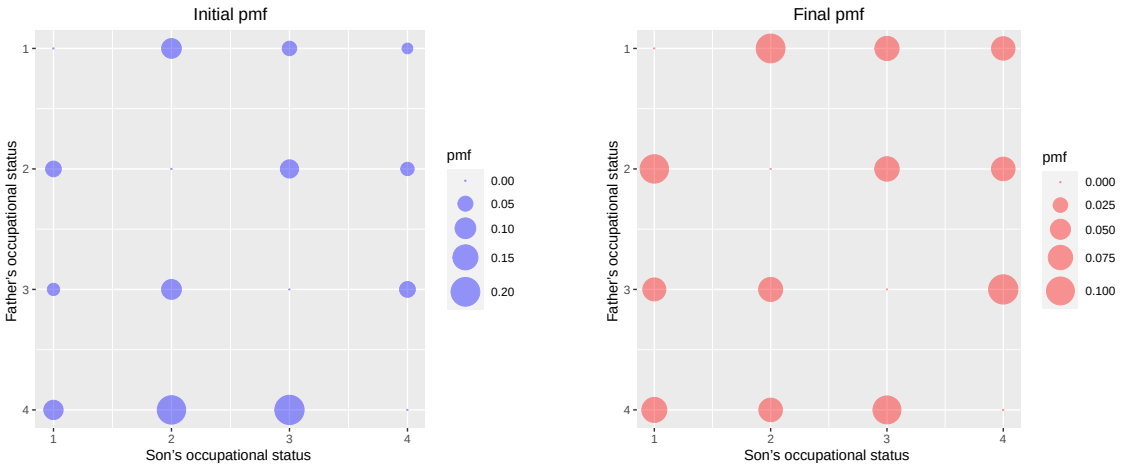
3. Examples and discussion

The first example we consider here is taken from (10) and we have already described it in Section 2. The corresponding data are reported in Table 1. We have $R = S = 4$, $\mathrm{N}(\mathbf{p}) = \{(v_X \times v_Y) : v_X = v_Y = \{i\}, i = 1,\ldots,4\}$ and for all $(v_X \times v_Y) \in N(\mathbf{p})$, $\frac{|v_X|}{R} + \frac{|v_Y|}{S} = 1/4+1/4 < 1$. This example falls under case 1 of Theorem 1. As expected, we are able to find a representative element $\bar{\mathbf{p}}$ with uniform margins

Table 1: Father's and son's occupational status categories in Japan in 1955, adapted from (10). Observed data in the left panel (dashes denote structural zeros), and associated discrete copula in the right panel.

	Son				
Father	1	2	3	4	Total
1	–	72	37	19	128
2	44	–	61	31	136
3	26	73	–	45	144
4	69	156	166	–	391
Total	139	301	264	95	799

$$C_1 = \begin{pmatrix} 0 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.09 & 0.14 & 0.16 \\ 0 & 0.06 & 0.15 & 0.27 & 0.33 \\ 0 & 0.09 & 0.27 & 0.39 & 0.51 \\ 0 & 0.17 & 0.55 & 0.88 & 1.00 \end{pmatrix}$$



(a) observed pmf

(b) final pmf as output of IPF

Figure 1: Bubble plots of the probability mass functions

$1/4$ and a corresponding discrete copula $\tilde{C}_1$ defined on the uniform grid I_4^2. The results are reported in Table 2. Fig. 1 displays the bubble plots for this example, a graphical representation of the probability mass function (pmf) for the observed data and the transformed pmf obtained through IPF. In the right-hand panel, exploiting the IPF algorithm, we have removed the effects of marginal non-homogeneity and the red plot unveils the symmetry structure, which is masked in the blue panel by the different values of the marginal frequencies. The role of the frequencies in the right-hand panel, i.e., to unveil symmetry beyond marginal non-homogeneity, is the same role of the residuals in the log-linear quasi-symmetry model in Eq. (1). Nevertheless, the graphical representation of the discrete copula is immediate to read and provides an easy way to detect symmetries in the data table. In this example, we can observe that the bubbles in symmetric cells have nearly the same radius, and thus the data table has a strong symmetry.

The second example is taken from (3), and the data are displayed in Table 3. The table here differ from previous case, as here we do not have a square table. The data concern the relationship between the locular composition (the number of locules of the ovary with odd or even numbers of ovules) and radial symmetry (root mean square deviation of the number of ovules from the mean number in the individual ovary) for the fruit of the American Bladder Nut, *Staphylea trifolia*, which has three locules per ovary. Some of the combinations are biologically impossible, which implies that the table has structural zeros. Note that after a reordering of the rows and columns, the observed table in Table 3 can be split into two separate complete sub-tables without structural zeros. This approach has been developed in, e.g., (5). We have retained the original structure since ordinal random variables are relevant in the copula framework.

This example falls under the assumption of case 2 in Theorem 1. Indeed, we have $R = 4$, $S = 9$ and with $v(X) = \{1, 4\}$, $v(Y) = \{2, 3, 5, 6, 8\}$ we get a sub-table with structural zeros with $|v(X)|/R +$

Table 2: IPF output of the data reported in Table 1 in the left panel, and corresponding discrete copula in the right panel

	Son				
Father	1	2	3	4	Total
1	0.0000	0.1069	0.0739	0.0693	0.25
2	0.1040	0.0000	0.0757	0.0703	0.25
3	0.0666	0.0730	0.0000	0.1105	0.25
4	0.0794	0.0702	0.1004	0.0000	0.25
Total	0.25	0.25	0.25	0.25	1

$$\tilde{C}_1 = \begin{pmatrix} 0 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.11 & 0.18 & 0.25 \\ 0 & 0.10 & 0.21 & 0.36 & 0.50 \\ 0 & 0.17 & 0.35 & 0.50 & 0.75 \\ 0 & 0.25 & 0.50 & 0.75 & 1.00 \end{pmatrix}$$

Table 3: Locular composition versus radial symmetry, from (3). Observed data in the top panel dashes denote structural zeros), IPF output in the central panel, and IPF output of the transposed table in the bottom panel.

	Radial symmetry									
Locular c.	.00	.47	.82	.94	1.25	1.41	1.63	1.70	1.89	Total
3 even 0 odd	462	–	–	130	–	–	2	–	1	595
2 even 1 odd	–	614	138	–	21	14	–	1	–	788
1 even 2 odd	–	4413	95	–	22	8	–	5	–	4543
0 even 3 odd	103	–	–	35	–	–	1	–	0	139
Total	565	5027	233	165	43	22	3	6	1	6065

$|v(Y)|/S = 2/4 + 5/9 > 1$. Therefore, the existence of a unique discrete copula is not guaranteed. Running the IPF algorithm on the original table, we obtain the probabilities in Table 4, top panel, where only the row variable is uniform. Running the IPF algorithm on the transposed of the observed table, we obtain the probabilities in Table 4, bottom panel, where, again, only the row variable is uniform. Thus, we have two solutions. Both of them share the same odd-ratio structure of the observed table due to the properties of the IPF algorithm. However, it is not possible to construct a unique discrete copula on $I_4 \times I_9$ (or on $I_9 \times I_4$) which is associated to the table. This implies that the identifiability issue of a unique copula model in a discrete setting is not completely solved by the approach presented in (4). More research is needed to shed light on such cases and further inquire into other invariant properties of the discrete copulas that can be relevant for more scenarios.

In this work we have only discussed two examples. Though, the theory is far more rich and several interesting examples do not fall within the cases of Theorem 1. For instance, we have not explored here the situation where $\frac{|v_X|}{R} + \frac{|v_Y|}{S} \leq 1$ for all $(v_X \times v_Y) \in N(\mathbf{p})$, but there are $\tilde{v}_X, \tilde{v}_Y$ such that $\frac{|\tilde{v}_X|}{R} + \frac{|\tilde{v}_Y|}{S} = 1$. This case happens for instance in the framework of triangular tables, which are often observed in applications. A complete description of the copulas with different patterns of structural zeros will be the subject of further research.

References

[1] de Amo, E., Díaz Carrillo, M., Durante, F., Fernández Sánchez, J.: Extensions of subcopulas. J. Math. Anal. Appl. **452**(1), 1–15 (2017)

[2] De Loera, J.A., Kim, E.D.: Combinatorics and geometry of transportation polytopes: An update. In: Discrete Geometry and Algebraic Combinatorics, *Contemporary Mathematics*, vol. 625, pp. 37–76. American Mathematical Society, Providence, RI (2014)

[3] Fienberg, S.E.: The analysis of incomplete multi-way contingency tables. Biometrics **28**(1), 177–202 (1972)

[4] Geenens, G.: Copula modeling for discrete random vectors. Dependence Modeling **8**(1), 417–440 (2020)

[5] Goodman, L.A.: The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing cells. J. Amer. Statist. Ass. **63**, 1091–1131 (1968)

[6] Perrone, E.: Polytopes of discrete copulas and applications. In: L.A. García-Escudero, A. Gordaliza, A. Mayo, M.A. Lubiano Gomez, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (eds.) Building Bridges between Soft and Statistical Methodologies for Data Science, pp. 319–325. Springer International Publishing, Cham (2023)

[7] Perrone, E., Durante, F.: Extreme points of polytopes of discrete copulas. In: Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th In-

Table 4: Locular composition versus radial symmetry, from (3). IPF output in the top panel, and IPF output of the transposed table in the bottom panel.

	Radial symmetry									
Locular c.	.00	.47	.82	.94	1.25	1.41	1.63	1.70	1.89	Total
3 even 0 odd	0.0451	0.0000	0.0000	0.0401	0.0000	0.0000	0.0259	0.0000	0.1111	0.2222
2 even 1 odd	0.0000	0.0208	0.0785	0.0000	0.0681	0.0826	0.0000	0.0277	0.0000	0.2778
1 even 2 odd	0.0000	0.0903	0.0326	0.0000	0.0430	0.0285	0.0000	0.0834	0.0000	0.2778
0 even 3 odd	0.0660	0.0000	0.0000	0.0710	0.0000	0.0000	0.0852	0.0000	0.0000	0.2222
Total	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	1

	Locular composition				
Radial s.	3 even 0 odd	2 even 1 odd	1 even 2 odd	0 even 3 odd	Total
.00	0.0507	0.0000	0.0000	0.0743	0.125
.47	0.0000	0.0188	0.0812	0.0000	0.100
.82	0.0000	0.0707	0.0293	0.0000	0.100
.94	0.0451	0.0000	0.0000	0.0799	0.125
1.25	0.0000	0.0613	0.0387	0.0000	0.100
1.41	0.0000	0.0744	0.0256	0.0000	0.100
1.63	0.0292	0.0000	0.0000	0.0958	0.125
1.70	0.0000	0.0249	0.0751	0.0000	0.100
1.89	0.1250	0.0000	0.0000	0.0000	0.125
Total	0.25	0.25	0.25	0.25	1

ternational Summer School on Aggregation Operators (AGOP), pp. 596–601. Atlantis Press (2021)

[8] Perrone, E., Solus, L., Uhler, C.: Geometry of discrete copulas. J Multivariate Anal **172**, 162 – 179 (2019)

[9] Rudas, T.: Lectures on categorical data analysis. Springer (2018)

[10] Saigusa, Y., Fukumoto, N., Nakagawa, T., Tomizawa, S.: Measure of departure from conditional partial symmetry for square contingency tables. Journal of Mathematics and Statistics **18**, 138–142 (2022). DOI 10.3844/jmssp.2022.138.142

[11] Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de Paris **8**, 229–231 (1959)