A constructive method to minimize couple matchings

Pierre Bertrand, Michel Broniatowski, Jean-François Marcotorchino

Highlights

A constructive method to minimize couple matchings

Pierre Bertrand, Michel Broniatowski, Jean-François Marcotorchino

- Logical Indeterminacy minimizes couple matchings
- Logical Indeterminacy sums up as a mixture of three independent couplings
- Janson Vegelius correlation coefficient computes a deviation to Logical Indeterminacy

A constructive method to minimize couple matchings

Pierre Bertrand, Michel Broniatowski, Jean-François Marcotorchino

[a]*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France*
[b]*Laboratoire de Probabilités, Statistique et Modélisation & CNRS UMR 8001, Sorbonne Université, Paris, France*
[c]*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France*

Abstract

This paper provides constructive procedures for the indeterminacy coupling between two marginal distributions, an alternative to independence coupling. It also introduces a drawing under indeterminacy into a mixture of three independent couplings. Leveraging on this decomposition it states that indeterminacy optimally reduces couple matchings, minimizing the expected number of equal couples drawn in a row. Besides it is seen that the Janson Vegelius coefficient is nothing but a deviation to indeterminacy and it is shown that it tends to 0 when the number of modalities increases.

Keywords: Mathematical Relational Analysis, Correlation, Logical Indeterminacy, Coupling Functions

1. Introduction

A correlation criterion usually computes a deviation with respect to some equilibrium. Theoretical considerations lead to consider independence and logical indeterminacy as the only two possible "natural" equilibria or discrete coupling functions, as can be seen making use of a work of Csizar Csiszár et al. (1991), a summarized version of which is expressed in Bertrand et al. (2022).

A discrete coupling is a function C operating on two discrete marginal laws $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ and which defines a probability law π on the product space:

$$\pi_{u,v} = C(\mu_u, \nu_v),\ \forall\ 1 \leq u \leq p,\ 1 \leq v \leq q$$

We respectively quote both those above mentioned couplings $C^{\times}$ (independence) and C^{+} (indeterminacy); this last notion has been initially introduced by J.-F. Marcotorchino in his seminal paper Marcotorchino (1984)); formulas will be reintroduced later on in section 2.

Their usefulness arises in statistical applications: namely, most of our usual statistical deviation criteria for contingency analysis are expressed in terms of deviations from one of the

Email addresses: pierre.bertrand@ens-cachan.fr (Pierre Bertrand), michel.broniatowski@sorbonne-universite.fr (Michel Broniatowski), jfmmarco3@gmail.com (Jean-François Marcotorchino)

two couplings (see Conde-Céspedes (2013), which gathers a classification of such criteria); the famous χ^2 index, widely used in practice, computes a deviation to the independence coupling of the empirical margins. Symmetrically the Janson-Vegelius coefficient, initially introduced in Janson and Vegelius (1977) as a contingency association index, measures a deviation to indeterminacy; we shall detail this point in section 4.

Indeterminacy appears as a poorly known coupling, whose properties have been rarely presented in an explicit way. Its property in relation to minimization of couple matchings (definition 1) that it is useful in the Guessing or in the Task Partitioning problem is noticed in Bertrand et al. (2021). The present paper, is precisely dedicated to the properties implied by indeterminacy: *i)* we recall that indeterminacy aims at minimizing couple matching occurrences *ii)* we estimate the probability for a couple of margins uniformly and independently drawn to be eligible for an indeterminacy coupling (property 4) *iii)* we decompose an indeterminacy coupling into a mixture of three independent couplings leading to a constructive drawing; it enables to explain the couple matching minimization (property 5) *iv)* we analyze the Janson Vegelius correlation coefficient whose expression is nothing but a deviation to indeterminacy.

The paper is structured as follows. Section 2 gathers a summarized version of the construction of indeterminacy. The first part of section 3 provides the measure of the space of margins eligible for an indeterminacy coupling. A second part is dedicated to the decomposition of indeterminacy. This decomposition is new, to the best of our knowledge, and conveys an interpretation of the initial formula. Section 4 gathers an analysis of the Janson Vegelius coefficient.

2. Construction of indeterminacy

Although being the most natural, independence is not, by far, the only existing available coupling method; actually, as introduced by Sklar (1973), any copula function will lead to a coupling function acting on two cumulative distribution functions.

In the discrete case, two probability measures $\mu = \mu_1 \dots \mu_p$ and $\nu = \nu_1 \dots \nu_q$ represent the initial margins to be coupled. The first one belongs to the simplex S_p of dimension p while the second belongs to S_q of dimension q. A coupling π of μ and ν is as an element of S_{pq} whose margins are μ and ν, meaning:

$$\sum_{u=1}^{p} \pi_{u,v} = \nu_v, \ \forall 1 \leq v \leq q \qquad (1) \qquad\qquad \sum_{v=1}^{q} \pi_{u,v} = \mu_u, \ \forall 1 \leq u \leq p \qquad (2)$$

We quote $\mathcal{L}_{\mu,\nu}$ the subset of S_{pq} whose elements obey Equation (1) and Equation (2). It defines the space of couplings of μ and ν.

2.1. Reducing the information conveyed by the coupling

Among $\mathcal{L}_{\mu,\nu}$, some couplings π convey more information onto the margins than others. We suppose we want to reduce the available information one can extract out of realizations from π, while picking π as close to the uniform measure as possible. Applications for such an hypothesis can be found in Bertrand et al. (2021).

A constant, necessarily the uniform law $\mathbb{U}^{pq}$ would convey not information. However, it obviously does not respect margins. Therefore, let us force π to belong to $\mathcal{L}_{\mu,\nu}$ while being as close as possible to $\mathbb{U}^{pq}$. The square distance is a natural choice, actually motivated by the mean square error decomposition. We end up looking at:

Problem 1 (Minimal Trade Model)**.** $\min_{\pi\in\mathcal{L}_{\mu,\nu}} \sum_{u=1}^{p}\sum_{v=1}^{q}\left(\pi_{u,v}-\mathbb{U}^{pq}_{u,v}\right)^2$

It happens that we can compute the exact form of the solution (see Bertrand et al. (2022)). It is given by the so-called indeterminacy coupling π^+:

$$\pi^+_{u,v} = (\mu\oplus\nu)_{u,v} = \frac{\mu_u}{q}+\frac{\nu_v}{p}-\frac{1}{pq},\ \forall 1\le u\le p,\ \forall 1\le v\le q \tag{3}$$

This formula is positive if and only if the following inequality holds:

$$\frac{\mu_0}{q}+\frac{\nu_0}{p}-\frac{1}{pq}\ge 0 \tag{4}$$

where $\mu_0=\min_{\forall 1\le u\le p}\mu_u$ and $\nu_0=\min_{\forall 1\le v\le q}\nu_u$.

Inequality (4) considerably reduce the choice of the margins. In subsection 3.1 we describe a projection which aims at changing any couple of margins into a couple of margins respecting Inequality (4). Furthermore, we measure the proportion of margins eligible for an indeterminacy coupling.

2.2. Couple matchings minimization

The cost function in Problem 1 leads to minimize:

$$\sum_{u=1}^{p}\sum_{v=1}^{q}\pi^2_{u,v} \tag{5}$$

A first remark is that substituting $\mathbb{U}^{pq}_{u,v}$ by any constant in Problem 1 would have led to the same simplification. Though, interpreting the constant as a probability measure requires its value to be $\frac{1}{pq}$.

Definition 1 (Couple Matching)**.** *Consider to independent draws (U_1,V_1) and (U_2,V_2) of π, a probability law in the simplex S_{pq}. A couple matching occurs when $U_1=U_2$ and $V_1=V_2$.*

Equation (5) is nothing than probability that $\mathbb{P}((U_1,V_1)=(U_2,V_2))$, the probability of matching. Therefore π^+ minimizes the chances of couple matchings under any π in $\mathcal{L}_{\mu,\nu}$.

3. Properties of indeterminacy

3.1. Measuring the subset of margins eligible for indeterminacy

In this section, we estimate the impact of Inequality (4) on the margins. We begin with a simple case: to construct π^+ coupling μ with itself, the pair (μ,μ) must satisfy (4) which here writes $\mu_0\ge\frac{1}{2p}$. We estimate the probability that such an event happens. For this, consider the uniform distribution on S_p, the simplex of all laws on p values and compute the normalized Lebesgue measure of the eligible subset of S_p.

Proposition 1. *The proportion of μ in S_p such that (μ, μ) respects Inequality (4) is $\frac{1}{2^{p-1}}$.*

Proof. By (4) impose restricted bounds on the integrals constructing μ. The eligible set has measure

$$\int_{\frac{1}{2p}}^{1-\frac{p-1}{2p}} \int_{\frac{1}{2p}}^{1-\frac{p-2}{2p}-x_1} \dots \int_{\frac{1}{2p}}^{1-\frac{1}{2p}-\sum_{i=1}^{p-2} x_i} \mathrm{d}x_1 \dots \mathrm{d}x_{p-1}.$$

With the successive changes of variables $x_i \leftarrow x_i + \frac{1}{2p}$, this integral writes:

$$\begin{aligned} & \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-x_1} \dots \int_0^{\frac{1}{2}-\sum_{i=1}^{p-2} x_i} . \, \mathrm{d}x_1 \dots \mathrm{d}x_{p-1} \\ = \; & \frac{1}{2^{p-1}} \int_0^1 \int_0^{1-y_1} \dots \int_0^{1-\sum_{i=1}^{p-2} y_i} . \, \mathrm{d}y_1 \dots \mathrm{d}y_{p-1} \\ = \; & \frac{1}{2^{p-1}}. \end{aligned}$$

□

Remark 1. *The previous result is not surprising. A constructive method exists to build a valid μ. Indeed, by Inequality (4), μ_u is greater than $\frac{1}{2p}$ for all u. We deduce: $\mu_u = \frac{1}{2p} + \frac{r_u}{2}$ where r is an arbitrary probability law on p elements.*

Following the same assumptions as before, we draw μ and ν uniformly and independently among the probability laws, therefore among S_p and S_q respectively. The proposition below characterizes eligible couples for an indeterminacy coupling.

Proposition 2 (Construction of eligible margins, discrete case)**.**
The couple of margins (μ, ν) respects Inequality (4) if and only if there exists a positive real α such that:

$$\forall 1 \leq u \leq p, \; \mu_u \geq \frac{\alpha}{p} \qquad (6) \qquad\qquad \forall 1 \leq v \leq q, \; \nu_v \geq \frac{1-\alpha}{q} \qquad (7)$$

Proof. First, consider Inequality (4) is satisfied and define defining $\alpha = p\mu_0 \in [0, 1]$ then $\forall 1 \leq v \leq q$:

$$\frac{\mu_0}{q} + \frac{\nu_v}{p} \geq \frac{1}{pq}$$

which rewrites:

$$\nu_v \geq \frac{1-\alpha}{q}$$

Now, if such an α exists then $\forall 1 \leq u \leq p$ and $\forall 1 \leq v \leq q$,

$$\frac{\mu_u}{q} + \frac{\nu_v}{p} \geq \frac{\alpha}{pq} + \frac{1-\alpha}{pq} = \frac{1}{pq}$$

□

Remark 2. *Since μ_0 is the minimum of a set of p elements summing to 1, α as well as $\beta = 1 - \alpha$ belong to $[0, 1]$. It symmetrically implies that $\nu_0 \geq \frac{\beta}{q}$.*

Remark 3. *The introduction of the variable p in the definition of α it saves the symmetry between μ and ν by ensuring that all the values α of $[0, 1]$ are eligible regardless of p or q.*

Through a generalization of remark 1, Property 2 gives the existence of two probability laws r and s on p and q elements such as:

$$\forall 1 \leq u \leq p,\ \mu_u = \frac{\alpha}{p} + (1-\alpha) r_u \quad (8) \qquad \forall 1 \leq v \leq q,\ \nu_v = \frac{1-\alpha}{q} + \alpha s_v \quad (9)$$

Proposition 3 (Constructive eligible margins)**.** *A couple of probability laws $(\mu, \nu) \in S_p \times S_q$ respects Inequality (4) if and only it it exists a real $\alpha \in [0, 1]$ and a couple of probability laws $(r, s) \in S_p \times S_q$ such that Equations (8) and (9) are satisfied.*

For fixed α, the the space of eligible μ appears as a $(1 - \alpha)$-contraction of S_p whereas that of ν is an α-contraction of S_q. Since the two laws are drawn independently, the measure of the eligible space in $S_p \times S_q$ is given by:

$$\int_{\alpha=0}^{1} \alpha^{p-1} (1-\alpha)^{q-1} \, \mathrm{d}\alpha = \frac{(p-1)!(q-1)!}{(p+q-2)!} \tag{10}$$

The eligibility results are summarized in the following proposition:

Proposition 4 (Valid proportion)**.** *If μ is drawn in the simplex S_p uniformly, the probability that the pair (μ, μ) respects the Inequality (4) is $\frac{1}{2^{p-1}}$. Then, there exists a probability law r in the simplex S_p such that μ satisfies:*

$$\forall u, \mu_u = \frac{1}{2p} + \frac{r_u}{2}. \tag{11}$$

If additionally ν is drawn in S_q, independently upon μ then, the probability that the pair (μ, ν) respects Inequality(4) is $\frac{(p-1)!(q-1)!}{(p+q-2)!}$. In this case, there exists a real α, a probability law r in the simplex S_p and a probability law s in the simplex S_q such that:

$$\forall u,\ \mu_u = \frac{\alpha}{p} + (1-\alpha) r_u \qquad (12) \qquad\qquad \forall v,\ \nu_v = \frac{1-\alpha}{q} + \alpha s_v \qquad (13)$$

In addition, the previous writings characterize compliance to Inequality (4).

Remark 4 (Different shapes)**.** *We notice that the expression of the eligible proportion depends on whether we are interested in the coupling of μ with itself or with a second and independent law ν: the formula (10) of the second case does not catch up with the one in Property 1 by simply setting $p = q$. The difference comes from independency only holding in the second case.*

3.2. Indeterminacy as a mixture of three independent couplings

The formula which defines indeterminacy given in Equation (3) does not provide as such an efficient way to draw under indeterminacy nor any interpretation of the meaning of it. We propose to rewrite this formula so as to view indeterminacy as a classic mixture of three independent couplings. Our starting point is the usual form of an indeterminacy coupling.

$$\pi^+_{u,v} = \frac{\mu_u}{q} + \frac{\nu_v}{p} - \frac{1}{pq},\ \forall 1 \leq u \leq p,\ \forall 1 \leq v \leq q$$

Quoting $\mu_0 = \min_u \mu_u$ and $\nu_0 = \min_u \nu_u$ it rewrites:

$$\pi^+_{u,v} = \left[\frac{\mu_u - \mu_0}{q}\right] + \left[\frac{\nu_v - \nu_0}{p}\right] + \left[\frac{\mu_0}{q} + \frac{\nu_0}{p} - \frac{1}{pq}\right]$$

First let us remark that the three square brackets are positive since (4) is satisfied. Thus, we renormalize them to extract probability laws. Formally:

$$\pi^+_{u,v} = (1 - p\mu_0)\left[\frac{\mu_u - \mu_0}{q(1 - p\mu_0)}\right] + (1 - q\nu_0)\left[\frac{\nu_v - \nu_0}{p(1 - q\nu_0)}\right] + (p\mu_0 + q\nu_0 - 1)\left[\frac{1}{pq}\right] \tag{14}$$

Remark 5 (Tight case)**.** *In case any of the two first brackets equals* 0 *it means* μ *or* ν *is uniform. In that case indeterminacy and independence couplings are the same so that an interpretation of indeterminacy is trivial. Anticipating on the action of* T *defined below, when equality holds in (4) then no uniform component exists leading to* $R = 3$ *never happening.*

We now define a transformation T acting on a probability law by:

Definition 2. *Given a probability law* $s = s_1, \ldots, s_r$ *on* r *elements, we quote* s_0 *its minimum. The transformation* T^r *generates a new law on the same elements by:*

$$\begin{aligned} T^r \ &: \ S_r \to S_r \\ &\quad (s_i)_{1\leq i\leq r} \mapsto \left(\frac{s_i - s_0}{1 - r s_0}\right)_{1\leq i\leq r} \end{aligned}$$

We shall quote T *the transformation acting on any* S_r *through* $T|_{S_r} = T^r$*.*

We notice that T actually removes as much uniform part as possible from the probability law it operates on. $T(s)$ will concentrates its realizations on the modes of s. With this notation, Equation (14) rewrites:

$$\pi^+_{u,v} = (1 - p\mu_0)\frac{1}{q}T(\mu)_u + (1 - q\nu_0)\frac{1}{p}T(\nu)_v + (p\mu_0 + q\nu_0 - 1)\mathbb{U}^{pq}_{u,v} \tag{15}$$

Reading Equation (15), we are able to decompose an indeterminacy draw as stated in proposition 5.

Proposition 5 (Indeterminacy drawing decomposition)**.** *We introduce a random variable R on 3 modalities $1, 2, 3$ with respective probabilities $1 - p\mu_0$, $1 - q\nu_0$ and $p\mu_0 + q\nu_0 - 1$. Realizations under indeterminacy eventually decomposes as a mixture of three straightforward drawings:*

1. *draw R;*
2. *if $R = 1$ then (u, v) is drawn under the independence coupling of $T(\mu)$ and $\mathbb{U}^q$;*
3. *if $R = 2$ then (u, v) is drawn under the independence coupling of $\mathbb{U}^p$ and $T(\mu)$;*
4. *if $R = 3$ then (u, v) is drawn under the independence coupling of $\mathbb{U}^p$ and $\mathbb{U}^q$ (i.e. $\mathbb{U}^{pq}$).*

Under this form, it appears that π^+ exhausts the uniform part of each margin. It is definitely consistent with indeterminacy being the projection of $\mathbb{U}^{pq}$ on $\mathcal{L}_{\mu,\nu}$. $T(\mu)$ is more concentrated on the modes of μ than μ itself. Consequently when $R = 1$, U is concentrated on the mode of μ, far from the uniform: this is the price for the margin being μ. On any other value of R, U is uniformly drawn. Symmetrically, for V, the concentration on modes of ν happens when $R = 2$.

Eventually, Proposition 5 justifies the method induced by indeterminacy to reduce couple matchings. If $R = 1$ a couple matching is rare since $U_1 = U_2$ is prevented by U being drawn uniformly under U^p; if $R = 2$ then V is drawn uniformly; if $R = 3$ then both are drawn uniformly.

4. Janson Vegelius coefficient

In statistical analysis, given independent realizations $(U_1, V_1), \ldots, (U_n, V_n)$, the categorization of n individuals under two measures (for instance the city they live in, their socio-professional category, their ages, ...), how do we measure the correlation between U and V? A solution is to use a deviation-to-independence coefficient, for instance: the χ_2. To compute its value, from the n realizations of (U, V), we deduce an empirical margin π counting the proportion of individuals in each couple of modalities:

$$\pi_{u,v} = \frac{\#\{i \ / \ U_i = u \ \& \ V_i = v\}}{n}, \ \forall 1 \leq u \leq p, \ \forall 1 \leq v \leq q; \tag{16}$$

similarly, denote μ and ν the empirical margins. The empirical χ^2 index, denoted χ^2_n is then defined:

$$\chi^2_n(U, V) = \sum_{u=1}^{p} \sum_{v=1}^{q} \frac{(\pi_{u,v} - (\mu \otimes \nu)_{u,v})^2}{(\mu \otimes \nu)_{u,v}}; \tag{17}$$

which obviously happens to be null if and only if the empirical distribution π of the observed data is an independence coupling of the empiric margins. Obviously, such an event never happens even under independence.

Using a symmetric idea, a lesser known criterion, called Janson-Vegelius Index, after the name of the inventors of this coefficient (see Janson and Vegelius (1977), Janson and Vegelius

(1978) or Janson and Vegelius (1982)) writes as a deviation to indeterminacy:

$$JV_n(U,V) = \sum_{u=1}^{p}\sum_{v=1}^{q} \frac{\left(\pi_{u,v} - (\mu \oplus \nu)_{u,v}\right)^2}{\sqrt{\frac{p-2}{p}\left(\sum_{u=1}^{q}\mu_u^2 + 1\right)}\sqrt{\frac{q-2}{q}\left(\sum_{v=1}^{q}\nu_v^2 + 1\right)}}; \tag{18}$$

and obviously is equal to zero if and only if the empirical π is an indeterminacy coupling of the empirical margins as defined in Equation (3). We omit the subscript n in the following. The JV index is actually just a classical cosine coefficient when rewritten in the "Mathematical Relational Analysis" Space. The relational analysis space no longer encodes modalities but links between individuals. Two matrices X and Y of size $n \times n$ respectively associated to variables U and V are introduced as shown in Definition 3.

Definition 3 (Mathematical Relational Analysis notations)**.** *Let* $(U_1, \ldots, U_n)$ *and* $(V_1, \ldots, V_n)$ *be two* n *probabilistic draws of* U *and* V *respectively. We define two associated symmetric* $n \times n$ *matrices* X *and* Y *by:*

$$X_{i,j} = \mathbb{1}_{U_i = U_j},\ \forall 1 \leq i,j \leq n \tag{19}$$

$$Y_{i,j} = \mathbb{1}_{V_i = V_j},\ \forall 1 \leq i,j \leq n \tag{20}$$

To understand the notation, let us begin with some remarks about Definition 3. Basically, the two $\{0,1\}$ matrices X and Y represent agreements and disagreements between the two variables on a same draw of size n; they are symmetric with 1 values on their diagonal. As expected, one can pass from the relational encoding to the usual contingency encoding as well as in the reciprocal way; those transfer formulas are demonstrated in the mentioned articles and enable us to write JV as a cosine in the relational space:

$$JV(U,V) = JV(X,Y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)\left(Y_{i,j} - \frac{1}{q}\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{i,j} - \frac{1}{p}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{i,j} - \frac{1}{q}\right)^2}} \tag{21}$$

Calculations leading to Equation (21) from Equation (18) can be found in Marcotorchino and Michaud (1979) or Marcotorchino and El Ayoubi (1991). Additionally, key features about relational analysis can be found in Marcotorchino (1984), Messatfa (1990), Opitz and Paul (2005), Marcotorchino (1986), Marcotorchino (1991) and Ah-Pine (2010).

4.1. Average value of JV through simulation

We simulate random probability laws π uniformly in S_{p^2} to compute the distribution of the criterion JV according to p. We first propose Figure 1 which presents the distribution of the criterion. One element strikes immediately: values concentrate around 0 as p increases. We start proving it in the case $\pi = \mu \otimes \mu$ for which the formula is simplified before demonstrating the general case.

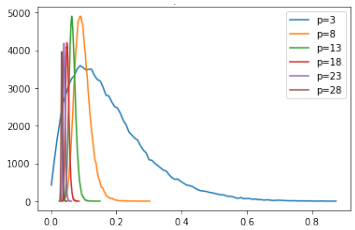

Figure 1: Distribution of the JV when π is uniform in S_{p^2}

4.2. Average value of JV through computations

This section gives a proof (in the case of the independent coupling of μ with itself) of the limit property noted in the previous section. It is stated in Proposition 6.

Proposition 6 (JV limit, independence case)**.**
If μ is uniform in S_p then $\lim_{p\to\infty} \mathbb{E}_\mu \left[JV(\mu \otimes \mu)\right] = 0$

Proof. We start by using a sequence of inequalities to extract an upper bound of JV.

$$\begin{aligned} JV(\mu\otimes\mu) &= \frac{p^2 \sum_{u=1}^{p} \left(\mu_u - \frac{1}{p}\right)^2 \sum_{u=1}^{p} \left(\mu_u - \frac{1}{p}\right)^2}{p(p-2)(\sum_{u=1}^{p} \mu_u^2) + 1} \\ &= \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p(p-2)(\sum_{u=1}^{p} \mu_u^2) + 1} \leq \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p^2(p-2)\frac{1}{p^2} + 1} = \frac{p^2 \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2}{p-1} \\ &\leq 2p \left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2 \end{aligned} \tag{22}$$

To demonstrate the convergence, we introduce Dirichlet law:

Definition 4 (Dirichlet law)**.** *The density of the Dirichlet law of parameter $(\alpha_1, \ldots, \alpha_p) \in (\mathbb{R}^{+*})^p$ on the simplex S_p is expressed as follows:*

$$f(\mu_1, \ldots, \mu_p, \alpha_1, \ldots, \alpha_p) \prod_{k=1}^{p} \mathrm{d}\mu_k = \frac{1}{B(\alpha)} \prod_{u=1}^{p} \mu_u^{\alpha_u - 1} \prod_{u=1}^{p} \mathrm{d}\mu_u \mathbb{1}_{(\mu_1,\ldots,\mu_p)\in S_p}$$

where, if we denote Γ the usual gamma function, B is the mutinomial beta function:

$$B(\alpha) = \frac{\prod_{u=1}^{p} \Gamma(\alpha_u)}{\Gamma\left(\sum_{u=1}^{p} \alpha_u\right)}.$$

The particular case when $\alpha_u = 1,\ \forall 1 \leq u \leq p$ expresses a uniform law on S_p whose density is given by:

$$f(\mu_1, ..., \mu_p) \prod_{k=1}^{p} \mathrm{d}\mu_k = (p-1)! \prod_{k=1}^{p} \mathrm{d}\mu_k \mathbb{1}_{(\mu_1,\ldots,\mu_p)\in S_p}$$

We specify the moments of the Dirichlet law to deduce the exact calculation of the expectation of the upper bound.

Proposition 7 (Dirichlet law moments)**.** *Given $\mu \in S_p$ drawn according to Dirichlet law with parameter $(\alpha_1, \ldots, \alpha_p)$. Denote $\alpha_0 = \sum_{u=1}^{p} \alpha_u$. Then for all p-uplet $\beta_1, \ldots, \beta_p$ of positive integers, we have the formula (with $\beta_0 = \sum_{u=1}^{p} \beta_u$):*

$$\mathbb{E}\left(\prod_{u=1}^{p} \mu_u^{\beta_u}\right) = \frac{\Gamma\left(\sum_{u=1}^{p} \alpha_u\right)}{\Gamma\left(\sum_{u=1}^{p} \alpha_u + \beta_u\right)} \prod_{u=1}^{p} \frac{\Gamma(\alpha_u + \beta_u)}{\Gamma(\alpha_u)} = \frac{\Gamma\left(\alpha_0\right)}{\Gamma\left(\alpha_0 + \beta_0\right)} \prod_{u=1}^{p} \frac{\Gamma(\alpha_u + \beta_u)}{\Gamma(\alpha_u)}$$

We now develop the upper-bound and we evaluate all terms making use of Proposition 7 with $\alpha_1 = \ldots = \alpha_p = 1$.

$$\mathbb{E}\left[\left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2\right] = \sum_{1 \leq u,v \leq p} \mathbb{E}(\mu_u^2 \mu_v^2) - \frac{2}{p} \sum_{u=1}^{p} \mathbb{E}(\mu_u^2) + \frac{1}{p^2}$$

For any $u \neq v$, it holds

$$\mathbb{E}(\mu_u^2 \mu_v^2) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \beta_0)} \frac{\Gamma(\alpha_u + 2)\Gamma(\alpha_v + 2)}{\Gamma(\alpha_u)\Gamma(\alpha_v)} = \frac{4}{p(p+1)(p+2)(p+3)}. \tag{23}$$

For $u = v$ then

$$\mathbb{E}(\mu_u^4) = \frac{2 * 3 * 4}{p(p+1)(p+2)(p+3)} = \frac{24}{p(p+1)(p+2)(p+3)}. \tag{24}$$

Finally,

$$\mathbb{E}(\mu_u^2) \quad = \quad \frac{2}{p(p+1)} \tag{25}$$

Summing up, by Equations (23), (24) and (25), we get:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{u=1}^{p} \mu_u^2 - \frac{1}{p}\right)^2\right] &= \frac{4p(p-1)}{p(p+1)(p+2)(p+3)} + \frac{24p}{p(p+1)(p+2)(p+3)} - \frac{2 * 2p}{p(p+1)} + \frac{1}{p^2} \\
&= \frac{4(p-1)}{(p+1)(p+2)(p+3)} + \frac{24}{(p+1)(p+2)(p+3)} - \frac{4p}{(p+1)} + \frac{1}{p^2} \\
&= \frac{4p(p+5) - 4(p+2)(p+3)}{p(p+1)(p+2)(p+3)} + \frac{1}{p^2} = \frac{-24}{p(p+1)(p+2)(p+3)} + \frac{1}{p^2}
\end{aligned}$$

Therefore by Inequality (22) we know that $\mathbb{E}_\mu\left[JV(\mu \otimes \mu)\right] = \mathcal{O}(\frac{1}{p})$ which ends the proof of Proposition 6. □

We now prove the following more general result.

Proposition 8 (JV limit)**.**
If π is uniform in S_{pq} then $\lim_{pq\to\infty} \mathbb{E}_\pi\left[JV(\pi)\right] = 0$

Proof. For all $1 \leq u \leq p$ and all $1 \leq v \leq q$, we denote $\pi_{u,\cdot} = \sum_{v=1}^{q} \pi_{u,v}$ and $\pi_{\cdot,v} = \sum_{u=1}^{p} \pi_{u,v}$. As in the case when $\pi = \mu \otimes \mu$, it holds:

$$JV(\pi) \leq \frac{4}{\sqrt{pq}} \left(pq \sum_{u,v} (\pi_{u,v}^2) - p \sum_{u} (\pi_{u,\cdot}^2) - q \sum_{v} (\pi_{\cdot,v}^2) + 1 \right)$$

We evaluate the expectation of the above upper bound making use of Proposition 7 with a uniform Dirichlet law on the simplex S_{pq}. Therefore consider $(\pi_{u,v})_{1\leq u\leq p, 1\leq v\leq q}$ a vector of size pq. Then, for the first term, $\mathbb{E}(\pi_{u,v}^2) = \frac{2}{pq(pq+1)}$ and for the second term,

$$\mathbb{E}(\pi_{u,\cdot}^2) = \mathbb{E}\left(\sum_{v,v'} \pi_{u,v}\pi_{u,v'} \right) = (q^2 - q)\frac{1}{pq(pq+1)} + q\frac{2}{pq(pq+1)} = \frac{q+1}{p(pq+1)} \tag{26}$$

By combining both equations and playing with symmetry for the third term, we obtain:

$$\begin{aligned} \mathbb{E} \quad & \left(pq \sum_{u,v} (\pi_{u,v}^2) - p \sum_{u} (\pi_{u,\cdot}^2) - q \sum_{v} (\pi_{\cdot,v}^2) + 1 \right) \\ & = p^2q^2 \frac{2}{pq(pq+1)} - p^2 \frac{q+1}{p(pq+1)} - q^2 \frac{p+1}{q(pq+1)} + 1 \\ & = \frac{2pq}{pq+1} - \frac{p(q+1)}{pq+1} - \frac{q(p+1)}{pq+1} + 1 = \frac{-p-q}{pq+1} + 1. \end{aligned}$$

Eventually, $\mathbb{E}_\pi\left[JV(\pi)\right] = \mathcal{O}\left(\frac{1}{\sqrt{pq}}\right)$. □

Remark 6. *Since JV is positive we have also shown that it tends towards* 0 *in probability. Besides, The propensity of the JV to approach* 0 *assumes that we use it sparingly before assuming indeterminacy.*

5. Conclusion

The main innovation of this paper is the decomposition of indeterminacy. It enables us, first to efficiently generate a drawing, second to interpret it as a mixture of three straightforward drawings and last but not least to explain how it reduces couple matchings while respecting the forced margins. Since indeterminacy cannot be defined on all margins, the paper also computes the proportion of eligible margins. Furthermore, it proposes a constructive method to transform any couple into an eligible couple. Eventually, the 0-limit of the Janson Vegelius coefficient helps us to mind when defining a threshold to conclude to indeterminacy.

References

Ah-Pine, J., 2010. On aggregating binary relations using 0-1 integer linear programming, in: ISAIM, pp. 1–10.

Bertrand, P., Broniatowski, M., Marcotorchino, J.F., 2021. Minimization with respect to divergences and applications, in: International Conference on Geometric Science of Information, Springer. pp. 818–828.

Bertrand, P., Broniatowski, M., Marcotorchino, J.F., 2022. Independence versus indetermination: basis of two canonical clustering criteria. Advances in Data Analysis and Classification , 1–25.

Conde-Céspedes, P., 2013. Modélisations et extensions du formalisme de l'analyse relationnelle mathématique à la modularisation des grands graphes. Ph.D. thesis. Paris 6.

Csiszár, I., et al., 1991. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. The annals of statistics 19, 2032–2066.

Janson, S., Vegelius, J., 1977. Correlation coefficients for nominal scales. Uppsala: Department of Statistics .

Janson, S., Vegelius, J., 1978. On the applicability of truncated component analysis based on correlation coefficients for nominal scales. Applied Psychological Measurement , 135–145.

Janson, S., Vegelius, J., 1982. The J-index as a measure of nominal scale response agreement. Applied Psychological Measurement , 111–121.

Marcotorchino, J.F., 1984. Utilisation des comparaisons par paires en statistique des contingences. Publication du Centre Scientifique IBM de Paris et Cahiers du Séminaire Analyse des Données et Processus Stochastiques Université Libre de Bruxelles , 1–57.

Marcotorchino, J.F., 1986. Maximal association theory as a tool of research. Classification as a tool of research , W.Gaul and M. Schader editors, North Holland Amsterdam .

Marcotorchino, J.F., 1991. Seriation problems:an overview. Applied Stochastic Models and Data Analysis 7, 139–151.

Marcotorchino, J.F., El Ayoubi, N., 1991. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. Revue de Statistique Appliquée 39, 25–46.

Marcotorchino, J.F., Michaud, P., 1979. Optimisation en Analyse Ordinale des Données. Ed Masson, Paris.

Messatfa, H., 1990. Maximal association for the sum of squares of a contingency table. Revue RAIRO, Recherche Opérationnelle 24, 29–47.

Opitz, O., Paul, H., 2005. Aggregation of ordinal judgements based on condorcet's majority rule. Data Analysis and Decision Support. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg .

Sklar, A., 1973. Random variables, joint distribution functions, and copulas. Kybernetika 9, 449–460.