Statistical Inference on Large Contingency Tables: Convergence, Testability, Stability

Marianna Bolla

Institute of Mathematics, Budapest University of Technology and Economics
H-1111. Egry József u. 1, Budapest, Hungary, *marib@math.bme.hu*

Abstract. Convergence of rectangular arrays with nonnegative, bounded entries is defined together with the limit object and cut distance. A statistic defined on a contingency table is testable if it can be consistently estimated based on a smaller, but still sufficiently large table which is selected randomly from the original one in an appropriate manner. By the above randomization, classical multivariate methods can be carried out on a smaller part of the array. This fact becomes important when our task is to discover the structure of large and evolving arrays, like genetic maps, social, and communication networks. Special block structures behind large tables are also discussed from the point of view of stability and spectra.

Keywords: convergence of contingency tables, testable contingency table parameters, block matrices, spectrum and stability

1 Introduction

In order to discover the structure of large rectangular arrays, e.g., microarrays, social, economic, or communication networks, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes. For basic notions see Section 2.

For the above causes, convergence and distance of general normalized arrays is introduced in Section 3. Roughly speaking, a squence of contigency tables converges if their global structure becomes more and more similar which fact will be formulated in terms of the convergence of homomorphism densities of maps taking small 0-1 "probe" tables into the large one. The limit object is a measurable, bounded function on $[0,1]^2$ which can be regarded as generalization of graph limits, cf. Lovász and Szegedy (2006). As such a convergent sequence of contingency tables is also a Cauchy sequence in the so-called cut metric, we are able to define distance between contingency tables of different sizes. Relation to the Aldous–Hoover Representation Theorem (see Diaconis and Janson (2008)) is also discussed.

In Section 4, testable contingency table parameters are defined. In fact, they are statistics that can be consistently estimated based on a fairly large sample. Most parameters based on spectral and balanced classification properties of the table are testable. Hence, classical methods of variance, factor,

or cluster analysis can be carried out on a smaller part of the table, obtained by an appropriate random selection of the rows and columns.

In Section 5, we generalize the famous Szemerédi's Regularity Lemma (see Frieze and Kannan (1999), Borgs et al. (2008)) to rectangular arrays. In this form, the theorem states that any $m \times n$ rectangular array can be approximated by a matrix having a special block structure, where the number of blocks does not depend on m and n, it merely depends on the accuracy of the approximation. If the number of blocks is relatively small, both the original and the (correspondence) transformed table will have as many structural singular values as the rank of the block matrix, see Bolla et al. (2010). In the $m = n$, but not necessarily symmetric case, it is true for the number of structural eigenvalues too, the real parts of which determine the stability of the system, cf. May (1972), Érdi and Tóth (1990), Juhász (1996).

2 Preliminaries

Let $C = C_{m\times n}$ be a contingency table of row set $Row_C = \{1, \dots, m\}$ and column set $Col_C = \{1, \dots, n\}$. The nonnegative, real entries c_{ij}'s are interactions between the rows and columns, and they are normalized such that $0 \le c_{ij} \le 1$. Sometimes we have *binary* tables of entries 0 or 1. We may assign positive weights $\alpha_1, \dots, \alpha_m$ to the rows and $\beta_1, \dots, \beta_n$ to the columns expressing individual importance of the categories embodied by the rows and columns. (In correspondence analysis, these are the row- and column-sums.) A contingency table is called *simple* if all the row- and column-weights are equal to 1. Assume that C does not contain identically zero rows or columns, moreover C is dense in the sense that the number of nonzero entries is comparable with mn. Let $\mathcal{C}$ denote the set of such tables (with any natural numbers m and n).

Consider a simple binary table $F_{a\times b}$ and maps $\Phi : Row_F \to Row_C$, $\Psi : Col_F \to Col_C$; further

$$\alpha_\Phi := \prod_{i=1}^{a} \alpha_{\Phi(i)}, \quad \beta_\Psi := \prod_{j=1}^{b} \beta_{\Psi(j)}, \quad \alpha_C := \sum_{i=1}^{m} \alpha_i, \quad \beta_C := \sum_{j=1}^{n} \beta_j.$$

Definition 1. The $F \to C$ homomorphism density is

$$t(F, C) = \frac{1}{(\alpha_C)^a (\beta_C)^b} \sum_{\Phi,\Psi} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}.$$

If C is simple, then $t(F, C) = \frac{1}{m^a n^b} \sum_{\Phi,\Psi} \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}$. In addition, if C is binary too, then $t(F, C)$ is the probability that a random map $F \to C$ is a homomorphism (preserves the 1's). The maps Φ and Ψ correspond to sampling a rows and b columns out of Row_C and Col_C with replacement, respectively. In case of simple C it means uniform sampling, otherwise the rows and columns are selected with probabilities proportional to their weights.

To sampling without replacement, injective maps Φ, Ψ correspond.

Definition 2. The injective and induced homomorphism densities of $F \to C$ are

$$t_{inj}(F,C) = \frac{1}{(\alpha)_a a!(\beta)_b b!} \sum_{\Phi,\Psi\ inj.} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)} \quad \text{and}$$

$$t_{ind}(F,C) = \frac{1}{(\alpha)_a a!(\beta)_b b!} \sum_{\Phi,\Psi\ inj.} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)} \prod_{f_{ij}=0} (1 - c_{\Phi(i)\Psi(j)}),$$

where $(\alpha)_a$ and $(\beta)_b$ denote the ath and bth elementary symmetric polynomials of $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_n$, respectively.

For simple C, $(\alpha)_a = \binom{m}{a}$ and $(\beta)_b = \binom{n}{b}$. Clearly, $t_{inj}(F,C)$ and $t_{ind}(F,C)$ are zeroes if $a \geq m$ or $b \geq n$. Typically, a and b are much smaller than m and n. As most maps into a large table are injective, $t(F,C)$ and $t_{inj}(F,C)$ are very close to each other. Namely, for simple C, $|t(F,C) - t_{inj}(F,C)| \leq \frac{ab}{m+n}$ that tends to zero for fixed a and b as $m, n \to \infty$. For not simple C the above difference also tends to zero if we assume that there are not dominant rows and columns in C in the sense that $\max_i \frac{\alpha_i}{\alpha_C} \to 0$ and $\max_j \frac{\beta_j}{\beta_C} \to 0$ as $m \to \infty$ and $n \to \infty$, respectively.

The following simple binary random table $\xi(a \times b, C)$ will play an important role in proving the theorems of Section 4. Select a rows and b columns of C with replacement, with probabilities α_i/α_C $(i = 1, \ldots, m)$ and β_j/β_C $(j = 1, \ldots, n)$, respectively. If the ith row and jth column of C are selected, they will be connected by 1 with probability c_{ij} and 0, otherwise, independently of the other selected row–column pairs, conditioned on the selection of the rows and columns. For large m and n, $\mathbb{P}(\xi(a \times b, C) = F)$ is very close to $t_{ind}(F,C)$ that is reminiscent of a likelihood function.

3 Convergence of contingency tables

Definition 3. We say that the sequence $(C_{m\times n})$ of contingency tables is convergent if the sequence $t(F, C_{m\times n})$ converges for any simple binary table F as $m, n \to \infty$.

In view of Section 2, the convergence of $t(F, C_{m\times n})$ is equivalent to the convergence of $t_{inj}(F, C_{m\times n})$ and $t_{ind}(F, C_{m\times n})$, as well. The convergence means that the tables $C_{m\times n}$ become more and more similar in small details as they are probed by smaller 0-1 tables $(m, n \to \infty)$.

The limit object is a measurable function $U : [0,1]^2 \to [0,1]$ and we call it *contingon*. In the $m = n$ and symmetric case, C can be regarded as the weight matrix of an edge- and node-weighted graph (the row-weights are equal to the column-weights, loops are possible) and the limit object was introduced as graphon, see Lovász and Szegedy (2006). The step-function contingon

U_C is assigned to C in the following way: the sides of the unit square are divided into intervals $I_1, \dots, I_m$ and $J_1, \dots, J_n$ of lengths $\alpha_1/\alpha_C, \dots, \alpha_m/\alpha_C$ and $\beta_1/\beta_C, \dots, \beta_n/\beta_C$, respectively; then over the rectangle $I_i \times J_j$ the step-function takes on the value c_{ij}.

In fact, the above convergence of contingency tables can be formulated in terms of a special distance. First we define it for contingons.

Definition 4. The cut distance between the contingons U and V is

$$\delta_\square(U, V) = \inf_{\mu,\nu} \|U - V^{\mu,\nu}\|_\square \tag{1}$$

where the cut norm of the contingon U is defined by

$$\|U\|_\square = \sup_{S,T\subset[0,1]} \left| \iint_{S\times T} U(x,y)\, dx\, dy \right|,$$

and the infimum in (1) is taken over all measure preserving bijections $\mu, \nu : [0,1] \to [0,1]$, while $V^{\mu,\nu}$ denotes the transformed V after performing the measure preserving bijections μ and ν on the sides of the unit square, respectively.

An equivalence relation is defined over the set of contingons: two contingons belong to the same class if they can be transformed into each other by measure preserving map, i.e., their cut distance is zero. In the sequel, we consider contingons modulo measure preserving maps, and under contingon we understand the whole equivalence class. By a theorem of Borgs et al. (2008), the equivalence classes form a compact metric space with the $\delta_\square$ metric.

Definition 5. The cut distance between the contingency tables $C, C' \in \mathcal{C}$ is

$$\delta_\square(C, C') = \delta_\square(U_C, U_{C'}).$$

By the above remarks, the distance of C and C' is indifferent to permutations of the rows or columns of C and C'. In the special case when C and C' are of the same size, $\delta_\square(C, C')$ is $\frac{1}{mn}$ times the usual cut distance of matrices, cf. Frieze and Kannan (1999).

The following reversible relation between convergent contingency table sequences and contingons also holds, as a rectangular analogue of a theorem of Borgs et al. (2008).

Theorem 1. *For any convergent sequence $(C_{m\times n}) \subset \mathcal{C}$ there exists a contingon such that $\delta_\square(U_{C_{m\times n}}, U) \to 0$ as $m, n \to \infty$. Conversely, any contingon can be obtained as the limit of a sequence of contingency tables in $\mathcal{C}$. The limit of a convergent contingency table sequence is essentially unique: if $C_{m\times n} \to U$, then also $C_{m\times n} \to U'$ for precisely those contingons U' for which $\delta_\square(U, U') = 0$.*

It also follows that a sequence of contingency tables in $\mathcal{C}$ is convergent if, and only if it is a Cauchy sequence in the metric $\delta_\square$.

A simple binary random $a \times b$ table $\xi(a \times b, U)$ can also be randomized based on the contingon U in the following way. Let $X_1, \dots, X_a$ and $Y_1, \dots, Y_b$ be i.i.d., uniformly distributed random numbers on [0,1]. The entries of $\xi(a \times b, U)$ are indepenent Bernoully random variables, namely the entry in the ith row and jth column is 1 with probability $U(X_i, Y_j)$ and 0, otherwise. It is easy to see that the distribution of the previously defined $\xi(a \times b, C)$ and that of $\xi(a \times b, U_C)$ is the same. Further, $\delta_\square(C_{m\times n}, \xi(a \times b, C_{m\times n}))$ tends to 0 in probability, for fixed a and b as $m, n \to \infty$. This fact also plays an important role in proving the theorems of Section 4.

Note, that in the above way, we can as well randomize an infinite simple binary table $\xi(\infty \times \infty, U)$ out of the contingon U by generating countably infinitely many i.i.d. uniform random numbers on [0,1]. The distribution of the infinite binary array $\xi(\infty \times \infty, U)$ is denoted by $\mathbb{P}_U$. Because of the symmetry of the construction, this is an *exchangeable* array in the sense that the joint distribution of its entries is invariant under permutations of the rows and colums. Moreover, any exchangeable binary array is a mixture of such $\mathbb{P}_U$'s. More precisely, the Aldous–Hoover Representation Theorem (see Diaconis and Janson (2008)) states that for every infinite exchangeable binary array ξ there is a probability distribution μ (over the contingons) such that $\mathbb{P}(\xi \in A) = \int \mathbb{P}_U(A)\, \mu(dU)$.

4 Testable contingency table parameters

A function $f : C \to \mathbb{R}$ is called a *contingency table parameter* if it is invariant under isomorphism and scaling of the rows/columns. In fact, it is a statistic evaluated on the table, and hence, we are interested in contingency table parameters that are not sensitive to minor changes in the entries of the table.

Definition 6. A contingency table parameter f is testable if for every $\varepsilon > 0$ there are positive integers a and b such that if the row- and column-weights of C satisfy

$$\max_i \frac{\alpha_i}{\alpha_C} \le \frac{1}{a}, \qquad \max_j \frac{\beta_j}{\beta_C} \le \frac{1}{b}, \tag{2}$$

then

$$\mathbb{P}(|f(C) - f(\xi(a \times b, C))| > \varepsilon) \le \varepsilon.$$

Consequently, such a contingency table parameter can be consistently estimated based on a fairly large sample. Now, we introduce some equivalent statements of the testability, indicating that a testable parameter depends continuously on the whole table. This is the generalization of a theorem of Borgs et al. (2008) applicable to simple graphs.

Theorem 2. *For a testable contingency table parameter f the following are equivalent:*

- *For every $\varepsilon > 0$ there are positive integers a and b such that for every contingency table $C \in \mathcal{C}$ satisfying the condition (2),*

$$|f(C) - \mathbb{E}(f(\xi(a \times b, C)))| \leq \varepsilon.$$

- *For every convergent sequence $(C_{m\times n})$ of contingency tables with no dominant row- or columnn-weights, $f(C_{m\times n})$ is also convergent ($m, n \to \infty$).*
- *f is continuous in the cut distance.*

For example, in case of simple binary tables the singular spectrum is testable, as $C_{m\times n}$ can be regarded as part of the adjacency matrix of a bipartite graph on $m + n$ vertices, where Row_C and Col_C are the two independent vertex sets; further, the ith vertex of Row_C and the jth vertex of Col_C are connected by an edge if and only if $c_{ij} = 1$. The non-zero real eigenvalues of the symmetric $(m + n) \times (m + n)$ adjacency matrix of this bipartite graph are the numbers $\pm s_1, \ldots, \pm s_r$, where $s_1, \ldots, s_r$ are the non-zero singular values of C, and $r \leq \min\{m, m\}$ is the rank of C. Consequently, the convergence of adjacency spectra implies the convergence of the singular spectra. Therefore, by Theorem 2, any property of a large contingency table based on its singular value decomposition (e.g., correspondence decomposition) can be concluded from a smaller part of it. In Section 5, testability of some balanced classification properties is discussed.

5 Homogeneous partitions, spectra, and stability

Now, we shall prove that special blown up tables burdened with a general kind of noise are convergent.

Definition 7. The $m \times n$ random matrix E is a noise matrix if its entries are independent, uniformly bounded random variables of zero expectation.

Theorem 3. *The cut norm of any sequence $(E_{m\times n})$ of noise matrices tends to zero as $m, n \to \infty$, almost surely.*

Definition 8. The $m \times n$ real matrix B is a blown up matrix, if there is an $a \times b$ so-called *pattern matrix* P with entries $0 \leq p_{ij} \leq 1$, and there are positive integers $m_1, \ldots, m_a$ with $\sum_{i=1}^{a} m_i = m$ and $n_1, \ldots, n_b$ with $\sum_{i=1}^{b} n_i = n$, such that the matrix B, after rearranging its rows and columns, can be divided into $a \times b$ blocks, where block (i, j) is an $m_i \times n_j$ matrix with entries all equal to p_{ij} ($1 \leq i \leq a$, $1 \leq j \leq b$).

Let us fix the matrix $P_{a\times b}$, blow it up to obtain matrix $B_{m\times n}$, and let $A_{m\times n} = B + E$, where $E_{m\times n}$ is a noise matrix. If the block sizes grow proportionally, the following almost sure statements are proved in Bolla et. al (2010): the noisy matrix A has as many structural (protruding) singular values of order $\sqrt{mn}$ as the rank of the pattern matrix, all the other singular

values are of order $\sqrt{m+n}$; further, by representing the rows and columns by means of the singular vector pairs corresponding to the structural singular values, the a- and b-variances of the representatives tend to 0 as $m, n \to \infty$. Conversely, in the presence of structural singular values, with some additional conditions for the representatives, the block structure can be recovered.

Theorem 4. *Let the block sizes of the blown up matrix $B_{m\times n}$ are $m_1, \dots, m_a$ horizontally, and $n_1, \dots, n_b$ vertically ($\sum_{i=1}^{a} m_i = m$ and $\sum_{j=1}^{b} n_j = n$). Let $A_{m\times n} := B + E$ and $m, n \to \infty$ is such a way that $m_i/m \to r_i$ $(i = 1, \dots, a)$, $n_j/n \to q_j$ $(j = 1, \dots, b)$, where r_i's and q_j's are fixed ratios. Under these conditions, the "noisy" sequence $(A_{m\times n})$ converges almost surely.*

In many applications we are looking for clusters of the rows and columns of a rectangular array such that the the densities within the cross-products of the clusters be homogeneous. E.g., in microarray analysis we are looking for clusters of genes and conditions such that genes of the same cluster equally influence conditions of the same cluster. The following theorem ensures the existence of such a structure with possibly many clusters. However, the number of clusters does not depend on the size of the array, it merely depends on the accuracy of the approximation.

Theorem 5. *For every $\varepsilon > 0$ and $C_{m\times n} \in \mathcal{C}$ there exists a blown up matrix $B_{m\times n}$ of an $a \times b$ pattern matrix with $a + b \le 4^{1/\varepsilon^2}$ (independently of m and n) such that $\delta_\square(C, B) \le \varepsilon$.*

The theorem is a consequence of the Szemerédi's Regularity Lemma (see Frieze and Kannan (1999), Borgs et al. (2008)) and can be proved by embedding C into the adjacency matrix of an edge-weighted bipartite graph. The statement of the theorem is closely related to the testability of the following contingency table parameter:

$$S^2_{a,b}(C) = \min \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k\in A_i} \sum_{l\in B_j} (c_{kl} - \bar{c}_{i,j})^2, \quad \bar{c}_{i,j} = \frac{1}{|A_i| \cdot |B_j|} \sum_{k\in A_i} \sum_{l\in B_j} c_{kl},$$

where the minimum is taken over balanced a- and b-partitions $A_1, \dots, A_a$ and $B_1, \dots, B_b$ of Row_C and Col_C, respectively; further, instead of c_{kl} we may take $\alpha_k \beta_l c_{kl}$ in the row- and column-weighted case, provided there are no dominant rows/columns.

We applied our spectral partitioning algorithm for mixture of noisy data. Figure 1 shows the original 300×500 contingency table (a); the 1500×2500 blown up table close to the limit, with rows and columns sorted with respect to their cluster memberships obtained by k-means algorithm (b); eventually, the colour illustration of the average densities of the blocks formed by SVD (c).

The Gardner–Ashby's connectance c_n of a not necessarily symmetric array $A_{n\times n}$ is the percentage of nonzero entries in the matrix, that is the ratio

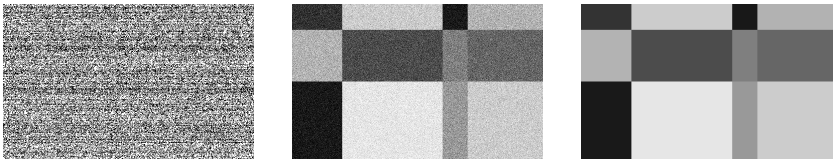

Fig. 1. noisy table (a); table close to the limit (b); approximation by SVD (c).

of actual row-column interactions to all possible ones in the network. In social and ecological models, a random array $A_{n\times n}$ of independent entries is considered. Suppose that the entries have symmetric distribution (consequently, zero expectation) and common variance σ_n^2, where σ_n is called average interaction strength. The stability of the system is characterized by the stability of the equilibrium solution 0 of the differential equation $dx/dt = A_{n\times n}x$ (sometimes this is achieved by linearization techniques in the neighbourhood of the equilibrium solution). Based on Wigner's famous semicircle law, May (1972) proves that the equilibrium solution is stable in the $\sigma_n^2 nc_n < 1$, and unstable in the $\sigma_n^2 nc_n > 1$ case; further, the transition region between stability and instability becomes narrow as $n \to \infty$. Hence, it seems that high connectance and high interaction strength destroy stability, but only in this simple model. If $A_{n\times n}$ is a block matrix, like a noisy matrix before, it has some structural, possibly complex eigenvalues, cf. Juhász (1996). If all their real parts are negative, the system is stable, see Érdi and Tóth (1990). In fact, in many natural ecosystems and other networks the interactions are arranged in blocks, at least an approximation of Theorem 5 works.

References

BOLLA, M., FRIEDL, K., and KRÁMLI, A. (2010): Singular value decomposition of large random matrices (for two-way classification of microarrays). *Journal of Multivariate Analysis 101, 434-446.*

BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T., and VESZTERGOMBI, K. (2008): Convergent sequences of dense graphs I, subgraph frequences, metric properties and testing. *Advances in Mathematics 219, 1801-1851.*

DIACONIS, P. and JANSON, S. (2008): Graph limits and exchangeable random graphs. *Rendiconti di Matematica 28 (Serie VII), 33-61.*

ÉRDI, P. and TÓTH, J. (1990): What is and what is not stated by the May-Wigner theorem? *J. Theor. Biol. 145, 137-140.*

FRIEZE, A. and KANNAN, R. (1999): Quick approximation to matrices and applications. *Combinatorica 19, 175-220.*

JUHÁSZ, F. (1996): On the structural eigenvalues of block random matrices. *Linear Algebra and Its Applications 246, 225-231.*

LOVÁSZ, L. and SZEGEDY, B. (2006): Limits of dense graph sequences. *J. Comb. Theory B 96, 933-957.*

MAY, R. M. (1972): Will a large complex system be stable? *Nature 238, 413-414.*