From Markov moves in contingency tables to linear model estimability

Roberto Fontana, Fabio Rapallo and Maria Piera Rogantin

Abstract The aim of this work is to highlight some interesting connections between contingency tables analysis and Design of Experiments. In particular, we consider two-way tables in correspondence to two-factor designs. A condition that characterizes the estimability of the independence model for all saturated fractions is provided.

1 Introduction

We consider contingency tables under the classical theory of log-linear models. Given two categorical random variables X and Y, a sample is summarized in an $I \times J$ contingency table. Under the Poisson sampling scheme, the counts of the cells are independent Poisson-distributed random variables $N_{i,j}$ with mean parameters $\mu_{i,j} > 0$. The independence model is therefore defined through the system of equations:

$$\log(\mu_{i,j}) = \lambda + \lambda_i^{(X)} + \lambda_j^{(Y)} .$$

Such a model has $p = I + J - 1$ parameters. For a detailed presentation of the independence model and its parametrizations, we refer to [1].

An $I \times J$ contingency table can be viewed also as a 2-factor experiment where the variables X and Y are the factors. In analogy with the independence model, we

Roberto Fontana
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, e-mail: roberto.fontana@polito.it

Fabio Rapallo
Dipartimento DISIT, Università del Piemonte Orientale, Viale Teresa Michel 11, 15121 Alessandria, e-mail: fabio.rapallo@unipmn.it

Maria Piera Rogantin
Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, 16146 Genova, e-mail: rogantin@dima.unige.it

consider linear models with the constant and the simple effects estimated through *saturated* fractions with $p = I + J - 1$ points.

The connections between tables and designs have been already explored in [3], where the focus was on the generation of all sudoku games. Here, we explore a different kind of connection, studying the estimability of saturated models.

2 Results

The design matrix of the independence model for $I \times J$ tables, under a suitable parametrization, is a full-rank matrix with dimensions $IJ \times (I + J - 1)$:

$$A = (a_0 \mid r_1 \mid \ldots \mid r_{I-1} \mid c_1 \mid \ldots \mid c_{J-1}) ,$$

where a_0 is a column vector of 1's, $r_1, \ldots, r_{I-1}$ are the indicator vectors of the first $(I-1)$ rows, and $c_1, \ldots, c_{J-1}$ are the indicator vectors of the first $(J-1)$ columns. For instance, in the case of 3×3 tables, the design matrix is:

$$A = \begin{matrix} (1,1) \\ (1,2) \\ (1,3) \\ (2,1) \\ (2,2) \\ (2,3) \\ (3,1) \\ (3,2) \\ (3,3) \end{matrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} .$$

As the parameter vector is a point of the space $\mathbb{R}^p$, the minimum number of points needed to estimate the parameters is p. The problem is therefore to determine the subsets $\mathscr{S}$ with exactly p cells that yield a non-singular submatrix. This problem is not trivial. For instance, let us consider the following 3×3 configurations with $p = I + J - 1 = 5$ cells, where $\star$ stands for a chosen cell.

$$\mathscr{S}_1 = \begin{bmatrix} \star & \star & - \\ \star & \star & - \\ - & - & \star \end{bmatrix} \qquad \mathscr{S}_2 = \begin{bmatrix} \star & \star & - \\ - & \star & - \\ - & \star & \star \end{bmatrix} .$$

$\mathscr{S}_1$ and $\mathscr{S}_2$ have a different behavior. In fact, the corresponding submatrices are:

$$A_{\mathscr{S}_1} = \begin{matrix} (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \\ (3,3) \end{matrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad A_{\mathscr{S}_2} = \begin{matrix} (1,1) \\ (1,2) \\ (2,2) \\ (3,2) \\ (3,3) \end{matrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

with $\det(A_{\mathscr{S}_1}) = 0$ and $\det(A_{\mathscr{S}_1}) = -1$. The difference between the two configurations is that the former contains a cycle, while the latter does not.

Definition 1. A k-cycle ($k \geq 2$) is a subset of $2k$ cells in a $k \times k$ subtable such that there are exactly 2 cells in each row and in each column.

The k-cycles have a special meaning in Algebraic Statistics in order to enumerate all tables with fixed margins (i.e., the tables in the Fréchet class). Recall that a Markov basis is a set of moves which makes connected each pair of tables with the same margins. It is well known that the basic moves of the form $\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}$ for all 2×2 submatrices of the table form a Markov basis, and their supports are just the 2-cycles. It is easy to see a 2-cycle in the configuration $\mathscr{S}_1$ above.

Moreover, filling a k-cycle with appropriate $+1$'s and -1's we obtain a move which preserves the marginal totals. For further details on the relations between the cycles and the Markov bases for the independence model, see [2] and [5].

The connections between the cycles and the factorial designs are established in the following results. We recall the definition of Orthogonal Array, see [4], as a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D} \equiv \mathscr{D}_1 \times \ldots \times \mathscr{D}_m$, where each factor $\mathscr{D}_i$ has n_i levels, $i = 1, \ldots, m$.

Definition 2. A fraction $\mathscr{F}$ of a design $\mathscr{D}$ is a *mixed orthogonal array* of strength t if it factorially projects onto any I-factors, $I = \{i_1, \ldots, i_t\}$, with $\#I = t$. *Factorially projects onto I factors* means that the projections of the fraction $\mathscr{F}$ over the I factors contain each t-tuple of $\mathscr{D}_{i_1} \times \ldots \times \mathscr{D}_{i_t}$ the same number $\alpha_I > 0$ of times.

We denote a fraction $\mathscr{F}$ that satisfies Definition 2 and such that $\#\mathscr{F} = n$ by $OA(n, n_1 \times \ldots \times n_m, t)$. We get the following proposition.

Proposition 1. *A k-cycle ($k \geq 2$) is:*

- *an $OA(2k, k \times k, t)$ where $t = 2$ if $k = 2$ and $t = 1$ if $k \geq 3$;*
- *the union of two disjoint orthogonal arrays $OA(k, k \times k, 1)$.*

The relation between the k-cycles and the non-estimability of linear models is established in the following theorem.

Theorem 1. *A subset $\mathscr{S}$ with p points yields a non-singular design matrix if and only if it does not contains cycles.*

3 Examples and discussion

We illustrate the above theory by a simple example. Let us consider the following configuration $\mathscr{S}$ for a 5×5 table. It contains a 4-cycle in the first 4 rows and the first 4 columns, hence it defines a singular design matrix:

$$\mathscr{S} = \begin{bmatrix} \star & - & \star & - & - \\ - & \star & - & \star & - \\ \star & \star & - & - & - \\ - & - & \star & \star & - \\ - & - & - & - & \star \end{bmatrix} .$$

Filling the 4-cycle with suitable $+1$'s and -1's, we obtain a move. Such move can be decomposed in the sum of its positive and negative part:

$$\begin{bmatrix} +1 & 0 & -1 & 0 \\ 0 & -1 & 0 & +1 \\ -1 & +1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{bmatrix} = \begin{bmatrix} +1 & 0 & 0 & 0 \\ 0 & 0 & 0 & +1 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & +1 & 0 \\ 0 & +1 & 0 & 0 \\ +1 & 0 & 0 & 0 \\ 0 & 0 & 0 & +1 \end{bmatrix} .$$

The left hand side corresponds to an $OA(8, 4 \times 4, 1)$, while the right hand side corresponds to two $OA(4, 4 \times 4, 1)$, namely:

$$\{(1,1),(2,4),(3,2),(4,3)\} \cup \{(1,3),(2,2),(3,1),(4,4)\} .$$

Finally, we notice that proportion of singular designs is not negligible. Approximately, for $I = J = 3$ we obtain a singular design in 36% of cases, for $I = J = 4$ in 64% of cases and for $I = J = 5$ in 81% of cases. Hence, the characterization of non-singular designs, as given in Theorem 1, is useful from an algorithmic point of view, because the random choice of a subset of $I + J - 1$ points does not appear an efficient procedure.

Acknowledgements Part of this paper has been written while FR was visiting BCAM (Basque Center for Applied Mathematics) in Bilbao, Spain. FR is partially supported by the PRIN2009 grant number 2009H8WPX5.

References

1. Agresti, A.: Categorical Data Analysis, 2 edn. Wiley, New York (2002)
2. Drton, M., Sturmfels, B., Sullivant, S.: Lectures on Algebraic Statistics. Birkhauser, Basel (2009)
3. Fontana, R., Rapallo, F., Rogantin, M.P.: Markov bases for sudoku grids. In: A. Di Ciaccio, M. Coli, J.M. Angulo Ibanez (eds.) Advanced Statistical Methods for the Analysis of Large Data-Sets, Studies in Theoretical and Applied Statistics. Springer, Berlin (2012). In press.
4. Hedayat, A.S., Sloane, N.J.A., Stufken, J.: Orthogonal arrays. Theory and applications. Springer Series in Statistics. Springer-Verlag, New York (1999)
5. Rapallo, F.: Markov bases and structural zeros. J. Symbolic Comput. **41**(2), 164–172 (2006)