Outliers and patterns of outliers in contingency tables with Algebraic Statistics

Fabio Rapallo
Department DISTA, Università del Piemonte Orientale
Viale Teresa Michel, 11
15121 Alessandria, Italy
fabio.rapallo@mfn.unipmn.it

February 7, 2011

Abstract

In this paper we provide a definition of pattern of outliers in contingency tables within a model-based framework. In particular, we make use of log-linear models and exact goodness-of-fit tests to specify the notions of outlier and pattern of outliers. The language and some techniques from Algebraic Statistics are essential tools to make the definition clear and easily applicable. Some numerical examples show how to use our definitions.

Key words: Algebraic Statistics; goodness-of-fit tests; log-linear models; toric models.

1 Introduction

The detection of outliers is one of the most important problems in Statistics and it is a current research topic in the field of contingency tables and categorical data. Some recent developments in this direction can be found in Kuhnt (2004), where the author describes a procedure to identify outliers based on the tails of the Poisson distribution and discusses the use of different estimators to compute the expected counts under the null hypothesis. A model-based approach to the detection of unexpected cell counts is the Configural Frequency Analysis (CFA), where the outlying counts are called "types" or "antitypes" if they are significantly higher or smaller with respect to the expected counts under a suitable model. The use of log-linear

models for CFA was presented in Kieser and Victor (1999) and reanalyzed in von Eye and Mair (2008).

The difficulties behind the definition of outlying cell in contingency tables is proved by the number of different approaches (see the references below). About these difficulties, and more generally on the old question: "What a contingency table does say", an interesting discussion is presented in Kateri and Balakrishnan (2008).

The notion of outlier for univariate and multivariate continuous distributions is a well known fact. For example, in the univariate case the outliers are usually detected through the boxplot or the comparison of the standardized values with respect to the quantiles of the normal distribution. It should be noted that there is no unique mathematical definition of outlier, see for instance Barnett and Lewis (1994). Notice also that the notion of outlier should be considered as outlier with respect to a specified probability model. For instance, in the continuous univariate case, it is usual to consider outliers with respect to the Gaussian distribution, leading to the well known three-sigma criterion.

The notion of outlier for contingency tables has a less clear meaning. In fact, the random variables we consider are categorical and the cells of the table are counts. When we consider contingency tables, we do not define the outliers among the subjects, but among the counts. As the counts can be modelled in a simple Poisson sampling scheme, one would use the quantiles of the Poisson distribution in order to detect the outliers in a contingency table. Using a different approach, the detection of outliers can also be deduced from the analysis of the adjusted residuals. This approach has been presented in Fuchs and Kenett (1980), while the algorithm in Simonoff (1988) is based on the analysis of the adjusted residuals and their contribution to the chi-squared Pearson's test statistics.

In the past decade, Algebraic Statistics has been a very growing research area, with major applications to the analysis of contingency tables. Algebraic Statistics now provides an easy description of complex log-linear models for multi-way tables and it represents the natural environment to define statistical models for contingency tables with structural zeros, through the notion of toric models. Moreover, non-asymptotic inference is now more actual via the use of Markov bases and the Diaconis-Sturmfels algorithm. As general references on the use of Algebraic Statistics for contingency tables, see Pistone *et al.* (2001), Pachter and Sturmfels (2005) and Drton *et al.* (2009). Some specific statistical models to study complex structures in contingency tables can be found in Rapallo (2005) and Carlini and Rapallo (2010) and Carlini and Rapallo (2011), with rel-

evant applications in the detection of special behavior of some subsets of cells (quasi-independence models, quasi-symmetry models, weakened independence models).

In this paper, we use the dictionary, the reasoning and some techniques from Algebraic Statistics in order to study the notion of outliers in contingency tables. The outliers are defined in terms of goodness-of-fit tests for tables with fixed cell counts. Then, we investigate the main properties of the outliers and we show how Algebraic Statistics is a useful tool both to make exact inference for goodness-of-fit tests, and to easily describe complex structures of outliers.

The material is organized as follows. In Section 2 we recall some definitions and basic results about toric models, while in Section 3 we show how to study a single outlying cell in the framework of toric models and we describe explicitly the Monte Carlo test using Markov bases. In Section 4 we present the notions of sets and patterns of outliers, and we analyze a real-data example. Finally, Section 5 contains some concluding remarks and pointers to future works.

2 Some recalls about log-linear and toric models

A probability distribution on a finite sample space $\mathcal{X}$ with K elements is a normalized vector of K non-negative real numbers. Thus, the most general probability model is the simplex

$$\Delta = \left\{ (p_1, \ldots, p_K) \ : \ p_k \geq 0 \ , \ \sum_{k=1}^{K} p_k = 1 \right\} .$$

A statistical model $\mathcal{M}$ is therefore a subset of Δ.

A classical example of finite sample space is the case of d-way contingency table where the cells are the joint counts of d random variables with a finite number of levels each. In the case of two-way contingency tables, where the sample space is usually written as a cartesian product of the form $\mathcal{X} = \{1, \ldots, I\} \times \{1, \ldots, J\}$. We will consider this case extensively in the next sections.

A wide class of statistical models for contingency tables are the log-linear models, see e.g. Agresti (2002). A model is log-linear if the log-probabilities lie in an affine subspace of the vector space $\mathbb{R}^K$. Given s real parameters $\alpha_1, \ldots, \alpha_s$, a log-linear model is described, apart from normalization,

through the equations:

$$\log(p_k) = \sum_{r=1}^{s} A_{k,r}\alpha_r \tag{1}$$

for $k = 1, \ldots, K$, where A is the design matrix, see (Pistone *et al.*, 2001, Ch.6). Exponentiating Eq. (1), we obtain the expression of the corresponding toric model

$$p_k = \prod_{r=1}^{s} \zeta_r^{A_{k,r}} \tag{2}$$

for $k = 1, \ldots, K$, where $\zeta_r = \exp(\alpha_r)$, $r = 1, \ldots, s$, are the new non-negative parameters. It follows immediately that the design matrix A is also the matrix representation of the minimal sufficient statistic of the model.

Notice that the model representations in Eq. (1) and (2) are equivalent on the open simplex, but the toric representation allows us to consider also the boundary, and therefore the tables with structural zeros. This issue will be essential in our definition of outliers. The matrix representation of the toric models as in equation (2) is widely discussed in e.g. Rapallo (2007) and Drton *et al.* (2009).

To obtain the implicit equations of the model, it is enough to eliminate the ζ parameters from the system in Eq. (2). In this paper, we will make use of the following ingredients from Algebraic Statistics:

(i) the toric ideal $\mathcal{I}_A$ of a statistical toric model with design matrix A;

(ii) the variety $\mathcal{V}_A$ of the model;

(iii) the Markov basis $\mathcal{M}_A$ of the model.

However, to keep the exposition simple, we have collected the formal definitions of these objects and some basic results on them in the Appendix.

As an example in the two-way setting, the independence model for 3×3 tables is represented by the matrix

$$A_{\text{ind}} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix},$$

while the quasi-independence model, which encodes independence of the two random variables except for the diagonal cells is represented by

$$A_{\mathrm{q-ind}} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

The last three columns of $A_{\mathrm{q-ind}}$ force the diagonal cells to be fitted exactly. For further details on the quasi-independence models, see Bishop *et al.* (1975). The equations of the independence model with design matrix A_{ind} is the set of all 2×2 minors of the table of probabilities, i.e.,

$$\begin{aligned} \mathcal{I}_{A_{\mathrm{ind}}} = \mathrm{Ideal}(&p_{1,1}p_{2,2} - p_{1,2}p_{2,1},\ p_{1,1}p_{2,3} - p_{1,3}p_{2,1},\ p_{1,1}p_{3,2} - p_{1,2}p_{3,1}, \\ &p_{1,1}p_{3,3} - p_{1,3}p_{3,1},\ p_{1,2}p_{2,3} - p_{1,3}p_{2,2},\ p_{1,2}p_{3,3} - p_{3,2}p_{2,3}, \\ &p_{2,1}p_{3,2} - p_{3,1}p_{2,2},\ p_{2,1}p_{3,3} - p_{3,1}p_{2,3},\ p_{2,2}p_{3,3} - p_{3,2}p_{2,3}), \end{aligned} \tag{3}$$

while for the quasi-independence model from the matrix $A_{\mathrm{q-ind}}$ we have only one binomial:

$$\mathcal{I}_{A_{\mathrm{q-ind}}} = \mathrm{Ideal}(p_{1,2}p_{2,3}p_{3,1} - p_{1,3}p_{3,2}p_{2,1}).$$

Notice that, from the point of view of the statistical models, a fixed cell count has the same behavior as a structural zero. See Rapallo (2006) for a discussion on this issue. Thus, we start guessing that outliers can be modelled in the framework of statistical models with structural zeros, as we will make precise in the following section. The use of structural zeros to model contingency tables with complex structure is presented in Consonni and Pistone (2007) under the point of view of Bayesian inference.

Remark 2.1. *In the special case of independence model for two-way tables, the use of* 2×2 *minors as in Eq.* (3) *to detect outliers was implemented in Kotze and Hawkins (1984). We also mention that the connections between the implicit equations of the model and the adjusted residuals are known at least in the simple case of the independence model for two-way table, see for instance Tsumoto and Hirano (2007).*

3 Outliers

Let us consider the following synthetic contingency table:

$$f = \begin{pmatrix} 7 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 3 & 2 & 2 & 2 \end{pmatrix}. \tag{4}$$

Under the independence model, it seems that the cell $(1, 1)$ could be an outlier.

With the approach presented in Fuchs and Kenett (1980), the observed contingency table f is the realization of a multinomial distribution and the authors analyze the adjusted residuals under the independence model

$$Z_{i,j} = \frac{f_{i,j} - f_{i,+}f_{+,j}/N}{\sqrt{f_{i,+}(N - f_{i,+})f_{+,j}(N - f_{+,j})/N^3}}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where N is the sample size and $f_{i,+}$ and $f_{+,j}$ are the row and column sums, respectively. To check the presence of outlying cells, the authors use the test statistics $Z = \max_{i,j} |Z_{i,j}|$ and they find suitable approximations for the two-sided α-level critical value. The use of the adjusted residuals to detect outliers was first described in Haberman (1973). However, we warn that the test in Fuchs and Kenett (1980) is a global test and it is not useful to detect the position of the outliers in the table.

On the other hand, the approach described in Kuhnt (2004) is based on the computation of the MLE (or L_1) estimate of the mean of the Poisson distributions for the cell counts, and then a cell is declared as outlier if the actual count falls in the tails of the appropriate Poisson distribution.

Let us analyze the observed table f above under the two approaches described here. Using the adjusted residuals as in Fuchs and Kenett (1980), the value of the test statistics is $z = 1.5670$ (the highest adjusted residuals), while the critical value at the $\alpha = 5\%$ level is 3.1594, showing that there is no evidence of outlying cells. Under the Poisson approach as in Kuhnt (2004), we find that the observed value in the cell $(1, 1)$ does not belong to the 5% tails of the Poisson distribution with mean parameter 4.7895 (i.e., the expected cell count under independence).

As mentioned in the previous section, we adopt here a different point of view to set up the definition and the detection of the outliers in a contingency table. We define them using a model-based approach with appropriate goodness-of-fit tests for the comparison of two nested models. The

starting point is similar to the definition of types and antitypes in CFA, see Kieser and Victor (1999), but after the first definitions we will use Algebraic Statistics to understand and generalize the notion of outlier.

Given a contingency table with K cells, let us consider a statistical toric model for the table. The model has the expression:

$$p_k = \prod_{r=1}^{s} \zeta_r^{A_{k,r}} \tag{5}$$

for all $k = 1, \ldots, K$. This model with matrix representation A will be named as the base model. Moreover, let $\alpha \in (0, 1)$.

Definition 3.1. *The cell h, $h \in \{1, \ldots, K\}$ is a α-level outlier with respect to the base model if the model*

$$p_k = \begin{cases} \prod_r \zeta_r^{A_{k,r}} & \text{for } \ k \neq h \\ \\ \prod_r \zeta_r^{A_{h,r}} \zeta_h^{(s)} & \text{for } \ k = h \end{cases} \tag{6}$$

is significantly better than the base model at level α, where $\zeta_h^{(s)}$ is a new non-negative parameter.

This means that we compare two toric models:

- the base model in Equation (5) with matrix representation A;
- the model in Equation (6), whose design matrix is

$$\tilde{A} = [A \mid I_h]$$

  where I_h is the indicator vector of the cell h: I_h is a vector of length K with all components equal to 0 but the h-th component equal to 1.

Notice that we do not test the goodness-of-fit of the model in Equation (6), but we only compare it with the base model.

To avoid trivialities in Definition 3.1, we suppose that the cell h is not a component of the sufficient statistic of the base model, i.e., we suppose that matrices $\text{rank}(\tilde{A}) = \text{rank}(A) + 1)$. In fact, if $\text{rank}(\tilde{A}) = \text{rank}(A)$, then the count in the cell h is already a component of the sufficient statistic of the base model and the goodness-of-fit test becomes useless.

From the point of view of toric models, the new parameter $\zeta_h^{(s)}$ imposes the exact fit of the candidate outlier h. Using the basic facts about elimination ideals reported in the Appendix, it is easy to show that, given generators

of the ideal $\mathcal{I}_A$ for the basic model, the elimination algorithm also identifies the ideal $\mathcal{I}_{\tilde{A}}$ with a simple step:

$$\mathcal{I}_{\tilde{A}} = \mathrm{Elim}(p_h, \mathcal{I}_A)\,.$$

In terms of varieties, the variety $\mathcal{V}_A$ is a subset of $\mathcal{V}_{\tilde{A}}$. This follows from the proposition below. We will use it also in the next section, thus we state the result in a general setting.

Proposition 3.2. *Let A_1 and A_2 be two integer non-negative matrices with K rows, and let $\mathrm{Im}(A_1)$ and $\mathrm{Im}(A_2)$ be their images, as vector spaces in $\mathbb{R}^K$. If $\mathrm{Im}(A_1) \subset \mathrm{Im}(A_2)$, then $\mathcal{V}_{A_1} \subset \mathcal{V}_{A_2}$.*

Proof. By virtue of Proposition A.9 in Appendix, we have to show that $\mathcal{I}_{A_2} \subset \mathcal{I}_{A_1}$. Let $g \in \mathcal{I}_{A_2}$. Then,

$$g = r_1 g_1 + \ldots + r_\ell g_\ell$$

where $\{g_1, \ldots, g_\ell\}$ is a system of generators of $\mathcal{I}_{A_2}$ and $r_1, \ldots, r_\ell$ are polynomials. From A.7 in the Appendix, $g_1, \ldots, g_\ell$ are binomials and their log-vectors (see Definition A.6) $m_1, \ldots, m_\ell$ are in $\ker(A_2)$. As $\ker(A_2) \subset \ker(A_1)$, we have also that $g \in \mathcal{I}_{A_1}$. This proves the result. □

Now, the inclusion $\mathcal{V}_A \subset \mathcal{V}_{\tilde{A}}$ follows from Proposition 3.2 with $A_1 = A$ and $A_2 = \tilde{A}$.

To actually check if a cell is an outlier, it is enough to implement the goodness-of-fit test in Definition 3.1. This test can be done using the log-likelihood ratio statistic, see e.g. (Agresti, 2002, p.591). The test statistic has the expression

$$G^2 = 2 \sum_{k=1}^{K} f_k \log\left(\frac{\hat{f}_{1k}}{\hat{f}_{0k}}\right),$$

where $\hat{f}_{0k}$ and $\hat{f}_{1k}$ are the maximum likelihood estimates of the cell counts under the base model with design matrix A and the model with design matrix $\tilde{A}$, respectively. The value of G^2 must be compared with the appropriate quantiles of the chi-square distribution with 1 df.

Alternatively one can make exact inference via Markov bases and the Diaconis-Sturmfels algorithm, see (Drton *et al.*, 2009, Ch.1).

Given an observed contingency table $f \in \mathbb{N}^K$ and a Markov basis $\mathcal{M}_A$ for the base model, one can apply the Diaconis-Sturmfels algorithm by sampling B contingency tables from the fiber

$$\mathcal{F}_t = \left\{ f' \in \mathbb{N}^K \;:\; A(f') = A(f) \right\}$$

with the hypergeometric distribution $\mathcal{H}(f')$. This is actually implemented through a Metropolis-Hastings Markov chain starting from the observed table. At each step:

1. let f be the current table;
2. choose with uniform probability a move $m \in \mathcal{M}_A$ and a sign $\epsilon = \pm 1$ with probability 1/2 each;
3. define the candidate table as $f_+ = f + \epsilon m$;
4. generate a random number u with uniform distribution over $[0, 1]$. If $f_+ \geq 0$ and
$$\min\left\{1, \frac{\mathcal{H}(f_+)}{\mathcal{H}(f)}\right\} > u$$
then move the chain in f_+; otherwise stay at f.

The proportion of sampled table with test statistics greater than or equal to the test statistic of the observed one is the Monte Carlo approximation of p-value of the log-likelihood ratio test.

In our numerical example, with a Monte Carlo approximation based on $10,000$ tables we obtain an approximated p-value 0.1574, showing that there is no evidence to conclude that the cell $(1, 1)$ is an outlier. In this example, the asymptotic p-value based on the chi-squared approximation is 0.0977, with a noteworthy difference with respect to the Monte Carlo approach. Notice that in similar problems the asymptotic approximation dramatically fails. To see this, consider the observed table

$$f' = \begin{pmatrix} 0 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 3 & 2 & 2 & 2 \end{pmatrix}.$$

This table differs from the first example in Eq. (4) only in the first cell. Here, the cell $(1, 1)$ is an antitype with an observed count less than the expected under independence, while in Eq. (4) the cell $(1, 1)$ was a type. For this table f', the Monte Carlo p-value is 0.1856, while the corresponding asymptotic approximation is 0.0522. All the simulations presented in this paper has been performed in `R`, see R Development Core Team (2010)

Finally, we remark that in many cases the computation of a Markov basis $\mathcal{M}_A$ for the base model does not need explicit symbolic computations. In fact, for several statistical models, such as independence, symmetry, quasi-independence, a Markov basis has been computed theoretically,

see Drton *et al.* (2009) and Rapallo (2003). For instance, our numerical example in this section considers the independence model as base model and a suitable Markov basis is formed by the 36 basic moves of the form $\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$ for all 2×2 minors of the table.

From the point of view of Geometry, this example is quite simple. The variety of the base model is described by the vanishing of all 2×2 minors of the table of probabilities, while the variety of the model with one outlier is described by the vanishing of 2×2 minors. They are exactly the 27 minors not involving the $(1, 1)$ cell.

4 Sets and patterns of outliers

Definition 3.1 can be easily extended to a set of outliers.

Definition 4.1. *The cells $h_1, \dots, h_m$ form a α-level set of outliers with respect to the base model if the model*

$$p_k = \begin{cases} \prod_r \zeta_r^{A_{k,r}} & \text{for } \; k \neq h_1, \dots, h_m \\ \\ \prod_r \zeta_r^{A_{k,r}} \zeta_k^{(s)} & \text{for } \; k = h_1, \dots, h_m \end{cases} \tag{7}$$

is significantly better than the base model at level α, where $\zeta_{h_1}^{(s)}, \dots, \zeta_{h_m}^{(s)}$ are m new non-negative parameters.

In analogy with our previous analysis, notice that the model in Equation (7) has matrix representation

$$\tilde{A} = [A \mid I_{h_1} \mid \cdots \mid I_{h_m}],$$

where $I_{h_1}, \dots, I_{h_m}$ are the indicator vectors of the cell $h_1, \dots, h_m$ respectively.

Also in this definition, to avoid trivialities, we suppose that the cells $h_1, \dots, h_m$ are not components of the sufficient statistic of the base model, i.e., we suppose that $\operatorname{rank}(\tilde{A}) > \operatorname{rank}(A)$. It is clear that the difference $\operatorname{rank}(\tilde{A}) - \operatorname{rank}(A)$ is just the number of degrees of freedom of the goodness-of-fit test. The test procedure can be performed with the same technique as for a single outlier. The algorithm is essentially the same as in Section 3 for a single outlier.

Example 4.2. *Let us consider the independence model for* 5×5 *tables as the base model and the* 10 *cells on the diagonal and the anti-diagonal as the set of outliers. The ideal of the base model is generated by the* 100 2×2 *minors of the table of probabilities, while the elimination of the* 10 *variables* $p_{1,1}, \ldots, p_{5,5}, p_{1,5}, \ldots, p_{5,1}$ *gives an ideal generated by the following binomials:* 10 *binomials of degree* 2

$$
\begin{aligned}
&-p_{1,4}p_{3,2} + p_{1,2}p_{3,4}, -p_{1,3}p_{5,2} + p_{1,2}p_{5,3},\\
&-p_{1,4}p_{5,2} + p_{1,2}p_{5,4}, -p_{1,4}p_{5,3} + p_{1,3}p_{5,4},\\
&-p_{2,5}p_{3,1} + p_{2,1}p_{3,5}, -p_{2,3}p_{4,1} + p_{2,1}p_{4,3},\\
&-p_{2,5}p_{4,1} + p_{2,1}p_{4,5}, -p_{2,5}p_{4,3} + p_{2,3}p_{4,5},\\
&-p_{3,5}p_{4,1} + p_{3,1}p_{4,5}, -p_{3,4}p_{5,2} + p_{3,2}p_{5,4},
\end{aligned}
$$

and 18 *binomials of degree* 3

$$
\begin{aligned}
p_{3,5}p_{4,3}p_{5,2} - p_{3,2}p_{4,5}p_{5,3},\\
-p_{2,5}p_{3,4}p_{5,3} + p_{2,3}p_{3,5}p_{5,4},\\
-p_{1,2}p_{3,4}p_{5,3} + p_{1,3}p_{3,2}p_{5,4},\\
p_{3,1}p_{4,3}p_{5,2} - p_{3,2}p_{4,1}p_{5,3},\\
-p_{2,1}p_{3,4}p_{5,3} + p_{2,3}p_{3,1}p_{5,4},\\
p_{3,4}p_{4,1}p_{5,3} - p_{3,1}p_{4,3}p_{5,4},\\
-p_{1,4}p_{3,5}p_{4,3} + p_{1,3}p_{3,4}p_{4,5},\\
-p_{1,2}p_{3,5}p_{4,3} + p_{1,3}p_{3,2}p_{4,5},\\
-p_{2,1}p_{3,5}p_{4,3} + p_{2,3}p_{3,1}p_{4,5},\\
p_{1,3}p_{2,5}p_{3,4} - p_{1,4}p_{2,3}p_{3,5},\\
p_{2,3}p_{3,1}p_{5,2} - p_{2,1}p_{3,2}p_{5,3},\\
-p_{1,3}p_{2,5}p_{3,2} + p_{1,2}p_{2,3}p_{3,5},\\
p_{1,3}p_{3,2}p_{4,1} - p_{1,2}p_{3,1}p_{4,3},\\
-p_{1,4}p_{2,3}p_{3,1} + p_{1,3}p_{2,1}p_{3,4},\\
-p_{1,2}p_{2,3}p_{3,1} + p_{1,3}p_{2,1}p_{3,2},\\
-p_{1,3}p_{3,4}p_{4,1} + p_{1,4}p_{3,1}p_{4,3},\\
-p_{2,3}p_{3,5}p_{5,2} + p_{2,5}p_{3,2}p_{5,3},\\
-p_{3,4}p_{4,5}p_{5,3} + p_{3,5}p_{4,3}p_{5,4}\,.
\end{aligned}
$$

As mentioned in the Introduction, one among the key points of Algebraic Statistics lies in the possibility to make the description and the meaning of log-linear models easier. Thus, we can enrich the base model in many ways.

Definition 4.3. *The cells $h_1, \ldots, h_m$ form a pattern of outliers with respect to the base model if the model*

$$p_k = \begin{cases} \prod_r \zeta_r^{A_{k,r}} & \text{for } \; k \neq h_1, \ldots, h_m \\ \\ \prod_r \zeta_r^{A_{k,r}} \zeta^{(p)} & \text{for } \; k = h_1, \ldots, h_m \end{cases}$$

is significantly better than the base model, where $\zeta^{(p)}$ is a new non-negative parameter.

To avoid trivialities in Definition 4.3, we suppose that the indicator vector of the cells $h_1, \ldots, h_m$ is not a component of the sufficient statistic of the base model, i.e., we suppose that the matrices $\tilde{A}$ and A satisfy: $\text{rank}(\tilde{A}) = \text{rank}(A) + 1$).

Remark 4.4. *Notice that in Definition 4.3 the outlying cells are characterized by a single parameter $\zeta^{(p)}$. This means that we assume a common behavior of that cells.*

Proposition 4.5. *Let $h_1, \ldots, h_m$ be m cells. The model with $h_1, \ldots, h_m$ as a set of outliers contains the model with $h_1, \ldots, h_m$ as a pattern of outliers.*

Proof. It is enough to apply Proposition 3.2. □

As a consequence, the definition of set of outliers in 4.1 is stronger than the definition of pattern of outliers. On the other hand, the notion of pattern of outliers may help in finding parsimonious models.

The definitions of set of outliers and pattern of outliers are very flexible and can be combined in many ways. In order to show this feature, we reconsider the following data analyzed in von Eye and Mair (2008) about the size of social network. The sample is formed by 516 individuals, classified by marital status ($M = 1$ married, $M = 2$ not married), gender ($G = 1$ male; $G = 2$ female), and size of social network ($S = 1$ small, $S = 2$ large). The 8 cell counts are listed in Table 4.

As a base model, we use the complete independence model, which can be written in log-linear form (with the usual log-linear notation) as:

$$\log p_{i,j,k} = \lambda + \lambda_i^{(M)} + \lambda_j^{(G)} + \lambda_k^{(S)} .$$

M	G	S	f
1	1	1	48
1	1	2	87
1	2	1	5
1	2	2	14
2	1	1	78
2	1	2	45
2	2	1	130
2	2	2	109

Table 1: Data on social network size.

The ideal of this base model is:

$$\begin{aligned}\text{Ideal}(&p_{1,2,1}p_{2,1,1} - p_{1,1,1}p_{2,2,1}, p_{1,2,1}p_{2,1,2} - p_{1,1,2}p_{2,2,1},\\ &-p_{1,2,2}p_{2,2,1} + p_{1,2,1}p_{2,2,2}, -p_{2,1,2}p_{2,2,1} + p_{2,1,1}p_{2,2,2},\\ &-p_{1,1,2}p_{2,1,1} + p_{1,1,1}p_{2,1,2}, p_{1,2,2}p_{2,1,1} - p_{1,1,2}p_{2,2,1},\\ &p_{1,2,2}p_{2,1,2} - p_{1,1,2}p_{2,2,2}, -p_{1,1,2}p_{2,2,1} + p_{1,1,1}p_{2,2,2},\\ &-p_{1,1,2}p_{1,2,1} + p_{1,1,1}p_{1,2,2}\big)\,.\end{aligned}$$

Thus, a Markov basis for this model is formed by 9 moves. A quick inspection of the residuals suggests that the cells $(1,1,2)$ and $(2,2,1)$ are potential types, while the cells $(1,2,1)$, $(1,2,2)$ and $(2,1,2)$ are potential antitypes.

If one would run a test for each of the previous cells as in Definition 3.1, the approximated Monte Carlo p-values are 0 in all cases. Notice also that in this example the definition of set of outliers as in 4.1 is not helpful, as the corresponding model become saturated. However, if we run the Monte Carlo test as in Definition 4.3 with these 5 cells as a unique pattern of outliers, we obtain a p-value 0.1411, showing that the 5 cells do not have a common behavior, but the test with two patterns of outliers, namely the potential types and antitypes separately, exhibits a p-value 0.0001, with strong evidence that the cells in the two patterns $\{(1,1,2),(2,2,1)\}$ and $\{(1,2,1),(1,2,2),(2,1,2)\}$ have a homogeneous pattern in deviating from

the base model. The design matrix for this model is

$$\tilde{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and we are able to describe the outlying cells with only two additional parameters. We note that the model with two patterns of outliers has a less clear geometric description with respect to the base model. In fact, the corresponding ideal is:

$$\begin{aligned} \text{Ideal}(&-p_{1,2,2}^2 p_{2,1,1}^2 + p_{1,1,1}p_{1,2,1}p_{2,1,2}p_{2,2,2}, \\ &-p_{1,1,2}p_{1,2,1}p_{2,1,1}^2 + p_{1,1,1}^2 p_{2,1,2}p_{2,2,1}, p_{1,1,1}p_{1,2,2}^2 p_{2,2,1} - p_{1,1,2}p_{1,2,1}^2 p_{2,2,2}, \\ &p_{1,2,2}^4 p_{2,1,1}^2 p_{2,2,1} - p_{1,1,2}p_{1,2,1}^3 p_{2,1,2}p_{2,2,2}^2). \end{aligned}$$

5 Final remarks

In this paper, we have shown how Algebraic Statistics is useful in addressing the problem of outliers in contingency tables. In particular, we have shown the efficacy of this approach in two directions: (a) the use of non-asymptotic inference for statistical models to recognize outliers; (b) a simple and practical description of such statistical models from the point of view of Geometry.

In particular, we have shown that Algebraic Statistics allows us to a simple definition of set of outliers, patterns of outliers, and their combinations.

Of course, the theory presented here does not exhaust all the research themes on this topic. Many questions remain still open, and among these problems we mention: the need for procedures and algorithms for the recognition of outliers; the problems of the choice of the α-level for multiple tests, using Bonferroni-type techniques. These problems are widely discussed in many articles cited above, see e.g. Kieser and Victor (1999).

From the perspective of Algebraic Statistics, some interesting issues are yet to be explored:

- The connections between the models studied here and the mixture models. Mixture models for the special case of outliers on the main diagonal are already considered in Bocci *et al.* (2010);

- The complete study of the of Markov bases for models with outliers and patterns of outliers. In the case of a single pattern of outliers, some results are presented in Hara *et al.* (2009).

Acknowledgments

We acknowledge the help and support of Enrico Carlini (Politecnico di Torino, Italy), who has provided several suggestions for a precise and clear algebraic presentation.

A Basic definitions and tools from Algebraic Statistics

In this appendix we collect some basic facts about toric ideals and statistical toric models. A more detailed presentation of these results can be found in Drton *et al.* (2009). For some basic algebraic definitions we also refer to Pistone *et al.* (2001).

Let $\mathbb{R}[p, \zeta] = \mathbb{R}[p_1, \ldots, p_K, \zeta_0, \zeta_1, \ldots, \zeta_s]$ be the polynomial ring in the variables $p_1, \ldots, p_K, \zeta_1, \ldots, \zeta_s$ with real coefficients.

Definition A.1 (Polynomial ideal)**.** *An ideal $\mathcal{I}$ in $\mathbb{R}[p, \zeta]$ is a set of polynomials such that for all $g, h \in \mathcal{I}$, $g + h \in \mathcal{I}$ and for all $g \in \mathcal{I}, h \in \mathbb{R}[p, \zeta]$, $gh \in \mathcal{I}$.*

The Hilbert's basis theorem states that every polynomial ideal $\mathcal{I}$ as in A.1 has a finite set of generators $\{g_1, \ldots, g_\ell\}$, i.e., for all $g \in \mathcal{I}$, there exist $r_1, \ldots, r_\ell \in \mathbb{R}[p, \zeta]$ with $g = r_1 g_1 + \ldots + r_\ell g_\ell$. In such a case, we write

$$\mathcal{I} = \text{Ideal}(g_1, \ldots, g_\ell)\,.$$

Let A be a non-negative integer matrix with K rows and s columns.

Definition A.2 (Toric model)**.** *The toric model associated to A is the set of probability distributions on $\{1, \ldots, K\}$ satisfying*

$$p_k = \zeta_0 \prod_{h=1}^{s} \zeta_h^{A_{k,h}}$$

for all $k = 1, \ldots, K$.

In the definition above, the parameter ζ_0 acts as a normalizing constant. As noticed in Section 2, a toric model is the extension of a log-linear model and the matrix A is the matrix representation of the minimal sufficient statistics.

Now, define the ideal $\mathcal{J}_A$ as the ideal generated by the set of binomials

$$\left\{ p_k - \prod_{h=1}^{s} \zeta_h^{A_{k,h}} \ : \ k = 1, \ldots, K \right\} .$$

Eliminating the ζ parameters, i.e., intersecting the ideal $\mathcal{J}_A$ with the polynomial ring $\mathbb{R}[p] \subset \mathbb{R}[p, \zeta]$, we define the toric ideal associated to A.

Definition A.3. *The toric ideal $\mathcal{I}_A$ associated to A is*

$$\mathcal{I}_A = \mathrm{Elim}(\zeta, J_A) = \mathcal{J}_A \cap \mathbb{R}[p] \, . \tag{8}$$

It is known that the toric ideal in Eq. (8) is generated by a finite set of pure homogeneous binomials $\mathcal{B}_A = \{b_1, \ldots, b_\ell\}$. To actually compute the set of generators $\mathcal{I}_A$ one can use Computer Algebra softwares such as CoCoA together with the command `Elim`, see CoCoATeam (2009). For toric ideals, specific algorithms are implemented in `4ti2`, see 4ti2 team (2008).

The toric ideal $\mathcal{I}_A$ has two major meanings in Algebraic Statistics. From the combinatorial side, the binomials $b_1, \ldots, b_\ell$ specify a Markov basis for the statistical model, while from a geometric point of view they describe the statistical model.

Definition A.4. *Let $f \in \mathbb{N}^K$ be a contingency table with K cells, and let A be a $K \times s$ matrix A. The reference set (or fiber) of f under A is:*

$$\mathcal{F}_t = \left\{ f' \in \mathbb{N}^k \ : \ A(f') = A(f) \right\} .$$

Definition A.5 (Markov basis)**.** *A set of tables $\mathcal{M}_A = \{m_1, \ldots, m_\ell\}$, $m_j \in \mathbb{Z}^K$, is a Markov basis for the reference set $\mathcal{F}_t$ if $Am_j = 0$ for all j and for any $f', f'' \in \mathcal{F}_t$ there exist a sequence of moves $(m_{j_1}, \ldots, m_{j_A})$ and a sequence of signs $(\epsilon_i)_{i=1}^A$ with $\epsilon_i = \pm 1$ such that*

$$f'' = f' + \sum_{i=1}^{A} \epsilon_i m_{j_i} \qquad \textit{and} \qquad f' + \sum_{i=1}^{a} \epsilon_i m_{j_i} \geq 0$$

for all $1 \leq a \leq A$. The elements of a Markov basis are called moves.

Definition A.6 (log-vector)**.** *Given a binomial in* $\mathbb{R}[p]$

$$b = \prod_{k=1}^{K} p_k^{m^+(k)} - \prod_{k=1}^{K} p_k^{m^-(k)} ,$$

its log-vector is:

$$m = m^+ - m^- \in \mathbb{Z}^K .$$

Theorem A.7 (Diaconis-Sturmfels)**.** *A set of vectors* $m_1, \ldots, m_\ell$ *is a Markov basis for the toric model associated to* A *if and only if the corresponding binomials* $b_1, \ldots, b_\ell$ *generate the toric ideal* $\mathcal{I}_A$*.*

Now, we show how the toric ideal $\mathcal{I}_A$ identifies the statistical toric model.

Definition A.8. *The set of points*

$$\mathcal{V}_A = \{p = (p_1, \ldots, p_K) \ : \ g(p) = 0 \ \ \textit{for all} \ \ g \in \mathcal{I}_A\}$$

is the variety associated to A*.*

To actually determine the variety $\mathcal{V}_A$, it is enough to solve the polynomial system $b_1(p) = 0, \ldots, b_\ell(p) = 0$, where $b_1, \ldots, b_\ell$ is a system of generators of $\mathcal{I}_A$.

The relations between the ideal $\mathcal{I}_A$ and the variety $\mathcal{V}_A$ imply that the computational algorithms for the Markov bases and for the varieties are just the same. Moreover, the following fundamental result holds.

Proposition A.9. *Let* $\mathcal{I}_{A_1}$ *and* $\mathcal{I}_{A_2}$ *be two toric ideals. Then:*

$$\mathcal{I}_{A_1} \subset \mathcal{I}_{A_2} \iff \mathcal{V}_{A_2} \subset \mathcal{V}_{A_1}$$

Finally, the statistical toric model is formed by the probability distributions in $\mathcal{V}_A$, i.e., the statistical toric model is simply $\mathcal{V}_A \cap \Delta$.

References

4ti2 team (2008). 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley, 2 ed.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. New York: John Wiley and Sons, 3 ed.

Bishop, Y. M., Fienberg, S., and Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Bocci, C., Carlini, E., and Rapallo, F. (2010). Geometry of diagonal-effect models for contingency tables. In M. A. Viana and H. P. Wynn (Eds.), *Algebraic Methods in Statistics and Probability II*, American Mathematical Society, vol. 516 of *Contemporary Mathematics*. 61–73.

Carlini, E. and Rapallo, F. (2010). Algebraic modelling of category distinguishability. In P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn (Eds.), *Mathematics Explorations in Contemporary Statistics*, Cambridge University Press. 111–122.

Carlini, E. and Rapallo, F. (2011). A class of statistical models to weaken independence in two-way contingency tables. *Metrika*, *73*, 1–22.

CoCoATeam (2009). CoCoA: a system for doing Computations in Commutative Algebra. Available at `http://cocoa.dima.unige.it`.

Consonni, G. and Pistone, G. (2007). Algebraic bayesian analysis of contingency tables with possibly zero-probability cells. *Statist. Sinica*, *17*, 1355–1370.

Drton, M., Sturmfels, B., and Sullivant, S. (2009). *Lectures on Algebraic Statistics*. Basel: Birkhauser.

Fuchs, C. and Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *J. Amer. Statist. Assoc.*, *75*(370), 395–398.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*(1), 205–220.

Hara, H., Takemura, A., and Yoshida, R. (2009). Markov bases for two-way subtable sum problems. *J. Pure Appl. Algebra*, *21*(3), 1507–1521.

Kateri, M. and Balakrishnan, N. (2008). Statistical evidence in contingency tables analysis. *J. Statist. Plann. Inference*, *138*, 873–887.

Kieser, M. and Victor, N. (1999). Configural Frequency Analysis (cfa) revisited – a new look at an old approach. *Biom. J.*, *41*(8), 967–983.

Kotze, T. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Appl. Statist.*, *33*(2), 215–223.

Kuhnt, S. (2004). Outlier identification procedures for contingency tables using maximum likelihood and L_1 estimates. *Scand. J. Statist.*, *31*, 431–442.

Pachter, L. and Sturmfels, B. (2005). *Algebraic statistics for computational biology*. New York: Cambridge University Press.

Pistone, G., Riccomagno, E., and Wynn, H. P. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Boca Raton: Chapman&Hall/CRC.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rapallo, F. (2003). Algebraic Markov bases and MCMC for two-way contingency tables. *Scand. J. Statist.*, *30*(2), 385–397.

Rapallo, F. (2005). Algebraic exact inference for rater agreement models. *Stat. Methods Appl.*, *14*(1), 45–66.

Rapallo, F. (2006). Markov bases and structural zeros. *J. Symbolic Comput.*, *41*(2), 164–172.

Rapallo, F. (2007). Toric statistical models: Parametric and binomial representations. *Ann. Inst. Statist. Math.*, *59*(4), 727–740.

Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backward stepping. *Technometrics*, *30*(3), 339–345.

Tsumoto, S. and Hirano, S. (2007). Characteristic of Pearson residuals in a contingency matrix. In D. Zhang, Y. Wang, and W. Kinsner (Eds.), *Proc. 6th IEEE Int. Conf. on Cognitive Informatics (ICCI'07)*, IEEE. 195–204.

von Eye, A. and Mair, P. (2008). A functional approach to Configural Frequency Analysis (CFA) revisited – a new look at an old approach. *Austrian Journal of Statistics*, *37*(2), 161–173.