JOURNAL OF ALGEBRAIC STATISTICS
Issue in honor of S. E. Fienberg
Volume 10, No. 1, 2019, p.30-50
https://jalgstat.library.iit.edu/
ISSN 1309-3452

Inference for Ordinal Log-Linear Models Based on Algebraic Statistics

Thi Mui Pham[1], Maria Kateri [1,*]

[1] *Institute of Statistics, RWTH Aachen University, Aachen, Germany*

Abstract. Tools of algebraic statistics combined with MCMC algorithms have been used in contingency table analysis for model selection and model fit testing of log-linear models. However, this approach has not been considered so far for association models, which are special log-linear models for tables with ordinal classification variables. The simplest association model for two-way tables, the uniform (U) association model, has just one parameter more than the independence model and is applicable when both classification variables are ordinal. Less parsimonious are the row (R) and column (C) effect association models, appropriate when at least one of the classification variables is ordinal. Association models have been extended for multidimensional contingency tables as well. Here, we adjust algebraic methods for association models analysis and investigate their eligibility, focusing mainly on two-way tables. They are implemented in the statistical software R and illustrated on real data tables. Finally the algebraic model fit and selection procedure is assessed and compared to the asymptotic approach in terms of a simulation study.

2000 Mathematics Subject Classifications: 13P10, 62H17, 62H15

Key Words and Phrases: Sparse contingency tables, Association models, Model selection, Diaconis-Sturmfels algorithm, Markov Chain Monte Carlo.

1. Introduction

Steve Fienberg was admirable for his broad research interests, the ability to perfectly combine the development of sound statistical methodology and its sophisticated and inspiring application in practice, as well as for his talent to detect intriguing contemporary statistical problems in diverse fields (such as social sciences, justice, official statistics) that generated stimulating research questions. He was also admirable for the natural way he switched among diverse approaches for statistical inference. Not only was he a frequentist and a Bayesian but he also set his signature to the analysis of contingency tables via algebraic statistical methods. In his editorial note of the Statistica Sinica (2007) special

*Corresponding author.

Email addresses: thi.mui.pham@posteo.de (T.M. Pham), maria.kateri@rwth-aachen.de (M. Kateri)

volume on algebraic statistics he stated: *'I have a fondness for contingency table problems, and many of them utilize algebraic geometry representations, often in multiple forms'.*

Log-linear models are undoubtedly the predominant tool for analyzing contingency tables. The books of Bishop, Fienberg & Holland [4] and Fienberg [11] have inducted generations of statisticians into the analysis of contingency tables and are established as timeless fundamental references. The usual inferential approach to log-linear models is asymptotic. However, in the case of small sample sizes (or sparse tables) exact inference is typically more appropriate while under sparseness the existence of the maximum likelihood estimators (MLEs) of the model parameters is not always ensured. In this setting, given a model and a collection of sufficient statistics, analyzing the exact conditional distribution of a contingency table can be carried out using algebraic methods as revealed by the pioneering work of Diaconis and Sturmfels [6]. They proposed an algorithm for sampling from a set of tables with given marginals, using a Markov basis, a kind of lattice basis that can be obtained by computing a Gröbner basis of a specially designed ideal. Following up on this work, Aoki and Takemura [3] and Rapallo [24, 25] derived Gröbner bases for some classical log-linear models that take structural zeros into account. At the same time, Fienberg and co-authors dealt with the problem of existence of MLEs for log-linear models. Erikson et al. [10] provided a polyhedral description of the conditions for the existence of MLEs for hierarhical log-linear models. Fienberg and Rinaldo [12] derived necessary and sufficient conditions for the existence of the MLEs of log-linear models' parameters when sampling zeros are observed. They further studied the geometric properties of log-linear models and provided algorithms for extended maximum likelihood estimation. Fienberg also introduced approaches for disclosure limitation in multidimensional tables to protect the confidentiality of individual responses and considered the problem of inference for log-linear models under disclosure limitation (see [7] and references therein).

Standard log-linear models treat all the classification variables of a contingency table as nominal. In practice, it is common to use these models even in scenarios where one or more variables do have a natural ordering to their levels (i.e. some are ordinal), ignoring information that can be leveraged for better inferential procedures and more sensible modeling. In such cases, ordinal log-linear models, also known as association models, are more appropriate (cf. Goodman [13, 14]), which capture this ordering by assigning scores to the categories of the classification variables. They impose a structure on the underlying association and are thus more parsimonious than usual log-linear models. Simultaneously, they provide sound physical interpretation for the local associations in the table, expressed in terms of the local odds ratios and based on the differences between the scores of successive classification categories.

The goal of this paper is the application of algebraic methods to model fit and selection in association models. We will focus mainly on two-way contingency tables. Using known algebraic techniques for log-linear models, we elaborate on how to conduct hypothesis tests for association models with sparse tables and on model selection for the class of association models. The corresponding algorithms are implemented in `R` [23].

The paper is structured as follows. In Section 2, association models are presented. Subsequently, in Section 3 algebraic techniques for analyzing contingency tables are briefly

reviewed. In Section 4, Markov bases for association models are constructed. The related algorithms are demonstrated on three examples in Section 5 while the algebraic p-value simulation is compared with the traditional asymptotically approximated p-value computation in Section 6. The results are summarized in the final section of the paper.

2. Association Models

Consider an $I \times J$ contingency table of observed frequencies $\mathbf{u} = (u_{ij})$ that cross-classifies two categorical variables X and Y of I and J levels, respectively, for a sample of fixed size $n = \sum_{i,j} u_{ij}$. Let U_{ij} be the random number of observations in cell (i, j). For a given distribution $\mathbf{p} = (p_{ij})$ on the table, we denote the expected cell frequencies by $m_{ij} = np_{ij}$.

Provided the table does not exhibit any structural zeroes, the possible hierarchical log-linear models for $\mathbf{u}$ are (i) that of independence (I) between X and Y

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \ , \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \tag{1}$$

where λ, λ_i^X and λ_j^Y denote the overall mean, the ith row and jth column main effects, respectively, and (ii) the saturated (S) model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \ , \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \tag{2}$$

with λ_{ij}^{XY} being the interaction parameter between the ith row and jth column. Log-linear models are over-parameterized and their parameters are subject to some identifiability constraints (usually the parameters corresponding to the last (or first) level of the classification variables are considered as redundant and set equal to 0). For $\lambda_I^X = \lambda_J^Y = 0$, the non-redundant parameters of model (1) are λ, $\lambda_1^X, \ldots, \lambda_{I-1}^X$, $\lambda_1^Y, \ldots, \lambda_{J-1}^Y$. Thus, the degrees of freedom for model (1) are $df(\mathrm{I}) = (I-1)(J-1)$ while (2) has $IJ-1$ non-redundant parameters and hence $df(\mathrm{S}) = 0$. Te derivation of the degrees of freedom of a model will be further considered in Section 3. For more details on log-linear models and related inference, we refer to Agresti [2].

These models treat both classification variables as nominal, i.e. the parameter estimates and consequently the models' goodness-of-fit statistics are invariant to reordering of the row or column categories. Thus, they ignore important information if the measurement scale of at least one classification variable is ordinal. However, ordinal data are very common in many application fields. Association models are special models for contingency tables that incorporate this additional information by assigning ordered scores to the rows and columns of the table. They impose a structure on the association among the classification variables, leading thus to models that are more parsimonious than the saturated model while at the same time providing meaningful interpretations.

The simplest association model for two-way tables is that of Linear-by-Linear association (LL model)

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j \ , \tag{3}$$

where u_i, $i = 1, \ldots, I$, and v_j, $j = 1, \ldots, J$, are known row and column scores with $u_1 \leq u_2 \leq \cdots \leq u_I$ ($u_1 < u_I$) and $v_1 \leq v_2 \leq \cdots \leq v_J$ ($v_1 < v_J$). Usually, the scores are set to satisfy the sum-to-zero and the sum of squares-to-one identifiability constraints (see Kateri [17, Chapter 6]). The independence model (1) is nested in (3), which has just one parameter more, the intrinsic association parameter β. Hence model (3) has $df(\text{LL}) = (I-1)(J-1)-1$. If the scores are equidistant for successive categories of rows as well as of columns, then under model (3) all local odds ratios of the table are equal, i.e.

$$\log \theta_{ij} = \frac{p_{ij}p_{i+1.j+1}}{p_{i+1.j}p_{i.j+1}} = c \ , \quad i = 1, \ldots, I-1, \ j = 1, \ldots, J-1, \tag{4}$$

and it is called the Uniform association model (U model) .

The LL model has the advantage of being very parsimonious and easy to interpret. However, it requires the direct assignment of scores to the row and column categories, which is not always an easy task (see Agresti [1, Chapter 6] for details). Thus, it is useful to have models that allow for parametric scores. If we consider the row scores in (3) to be unknown while the column scores are known, (3) can equivalently be expressed as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j \ , \tag{5}$$

where $\mu_i = \beta u_i$ (parameter β is redundant and thus absorbed in the parametric row scores). Model (5) is called the Row effects association model (R model) and assumes Y to be ordinal. X can be nominal (since μ_i are not necessarily ordered) or ordinal with unknown scores. The R model has $I-1$ additional parameters than model (3), corresponding to the row scores. Thus, the associated degrees of freedom of model (5) equal $df(\text{R}) = (I-1)(J-2)$. Analogously, the Column effects association model (C model)

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + u_i \nu_j \ , \tag{6}$$

with $df(\text{C}) = (I-2)(J-1)$, considers known row scores and parametric column scores $\nu_j = \beta v_j$, treating X as ordinal and Y as nominal (or ordinal with unknown scores). Parametric row or column scores are also subject to identifiability constraints (see Kateri [17]).

The association models considered above are special log-linear models that fill the gap between the two extreme log-linear models (1) and (2), and have been mainly developed by Goodman (cf. [13, 14, 15]). For a detailed introduction to association models, their features, inference and an overview of the related literature, we refer to [17, Chapters 6 and 7]. At this point we shall mention only one basic property of association models (see Remark 2.1 below) that will be needed in the sequel.

Remark 2.1.

1. Inferentially, association models are invariant under linear transformations of the row and column scores. For this, important are not the scores themselves but the distances between scores of successive categories. In the case of binary classification variables (i.e. one distance), score assignment does not have any inferential impact on the association model; any two distinct values can be used and are thus considered as prefixed.

2. As already noted, parametric scores are not necessarily ordered for successive categories and can be applied also to tables having the corresponding classification variable nominal. In this case, interpretation requires some caution [17, Section 6.4].
3. We have assumed above that the sample size n is known a priori and the random table of counts is multinomially distributed, e.g. $\mathbf{U} = (U_{ij}) \sim \mathcal{M}ult(n, \mathbf{p})$. Alternatively, it could be assumed that each cell count is Poisson distributed, e.g. $U_{ij} \overset{ind.}{\sim} \mathcal{P}(m_{ij})$. In this case, the total sample size n is random. Upon observing the sample and conditioning on its size n, the kernel of the likelihood is the same for both sampling schemes, hence they lead to the same MLEs.

Association models are also applicable for contingency tables of higher dimension. They can be derived by replacing one or more of the interaction terms of any hierarchical log-linear model by multiplicative terms based on scores, leading thus to more parsimonious models that impose a certain structure on the interactions. For example, in the presence of a third classification variable of K levels (layers), consider for the $I \times J \times K$ contingency table the log-linear model of conditional independence of Y and Z given X

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \ , \tag{7}$$

denoted by (XY, XZ) in standard hierarchical log-linear model notation. If Z is ordinal and w_k, $k = 1, \ldots, K$ are known scores for its categories, then the most parsimonious association model of conditional independence of Y and Z given X is

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta^{XY} u_i v_j + \beta^{XZ} u_i w_k \ , \tag{8}$$

having just two parameters more than the model of complete independence (X, Y, Z). Both interaction terms of model (8) are of LL-type. Model (8) is denoted $(XY_{\mathrm{LL}}, XZ_{\mathrm{LL}})$ and is illustrated in Example 5.3 below. Less parsimonious association models are derived by considering some of the scores in (8) as parametric.

3. Inference for Log-Linear Models

For inferential purposes it is advantageous to use the generalized linear model (GLM) formulation of log-linear models. For this, contingency tables of counts are expanded to vectors. Consider a random, non-negative integer valued k-way contingency table $\mathbf{U} \in \mathbb{N}^{m_1 \times \cdots \times m_k}$, formed by cross-classifying n independent and identically distributed realizations of the categorical variables $X_1, \ldots, X_k$, that take values in the finite sets of categories $[m_1], \ldots [m_k]$, respectively, with $[m_i] := \{1, \ldots, m_i\}$. Elongating the cells of the contingency table into a vector (subject to an ordering), we denote by X the corresponding discrete random vector with sample space $\mathcal{X} = [m]$, $m = \prod_i m_i$. A probability distribution on $\mathcal{X}$ is then defined by the values of the parameters $p_x = \mathrm{P}[X = x]$ for $x \in \mathcal{X}$. In the following, we use $\boldsymbol{p} = (p_x, x \in \mathcal{X}) \in \Delta_{m-1}$ for a m-dimensional probability vector, where Δ_{m-1} denotes the standard probability simplex. A statistical model $\mathcal{M}$ is a subset of the probability simplex $\mathcal{M} \subseteq \Delta_{m-1}$, and, with a common abuse of notation, a parametric

model is a model $\mathcal{M} = \text{Im}(\boldsymbol{p})$ where $\boldsymbol{p} : B \to \Delta_{m-1}$ and B is called the parameter space. In this article, assume $B = \mathbb{R}^{d_{\mathcal{M}}}$.

Elongating for example the $I \times J$ random table of counts $\mathbf{U}$ by rows, we set $\mathbf{u} = (U_{11}, \ldots, U_{1J}, \ldots, U_{I1}, \ldots, U_{IJ})^T$. The I model (1) can then equivalently be expressed as a Poisson/multinomial GLM in terms of the non-redundant parameters as

$$\log Y = \log(m_{11}, \ldots, m_{1J}, \ldots, m_{I1}, \ldots, m_{IJ})^T = D\beta \ , \tag{9}$$

with $d_{\text{I}} = I + J - 1$, $\beta = (\lambda, \lambda_1^X, \ldots, \lambda_{I-1}^X, \lambda_1^Y, \ldots, \lambda_{J-1}^Y)^T$ and

$$D = \begin{pmatrix} \mathbf{1} & \mathbf{1}^{(1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{1}^{(2)} & \mathbf{I}^* \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{1}^{(I-1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{0}_{J\times(I-1)} & \mathbf{I}^* \end{pmatrix}$$

with $\mathbf{1}$ the $J \times 1$ matrix of 1's, $\mathbf{1}^{(i)}$ the $J \times (I-1)$ matrix with 1's at the i-th column and 0's in all other entries, $\mathbf{0}_{s\times t}$ the $s \times t$ matrix of 0's and

$$\mathbf{I}^* = \begin{pmatrix} E_{J-1} \\ \mathbf{0}_{1\times(J-1)} \end{pmatrix} \ ,$$

where E_s is the $s \times s$ identity matrix. D is known as the design (or model) matrix. The degrees of freedom of model $\mathcal{M}$ equals $df(\mathcal{M}) = \dim\ker D = m - \text{rank}D$. The design matrix of a GLM expressed in terms of non-redundant parameters is of full rank, i.e $\text{rank}D = d_{\mathcal{M}}$.

Poisson/multinomial GLMs share the very convenient property of GLMs with canonical link function, namely their design matrices specify directly the sufficient statistics of the model. In particular, if $T : \mathcal{X} \to \mathbb{N}^{d_{\mathcal{M}}}$ with $T = (T_1, \ldots, T_{d_{\mathcal{M}}})$ the minimal sufficient statistic, then the $(i,j)^{th}$ element of D^T is given by $D^T(i,j) = T_j(x_i)$, $i = 1, \ldots, m$, $j = 1, \ldots, d_{\mathcal{M}}$, so that in aggregate the vector of minimal sufficient statistics is $D^T\mathbf{u}$. Practically, this means that all possible contingency tables of the same sample size n that yield the same T lead to the same estimated table of expected frequencies under the corresponding model. This set of tables is crucial for exact conditional hypothesis (model) testing, as explained next.

The standard goodness of fit test of a log-linear model $\mathcal{M}$ is based on Pearson's X^2 or the likelihood ratio statistic G^2, which are asymptotically equivalent. Under $\mathcal{M}$,

$$G^2(\mathcal{M}) = 2\sum_i x_i \log\left(\frac{x_i}{\hat{y}_i}\right) \ ,$$

where $\hat{y}_i$ is the maximum likelihood estimate of the expected frequency y_i under model $\mathcal{M}$, is under mild conditions asymptotically $\chi^2_{df(\mathcal{M})}$ distributed. Whenever asymptotic inferential methods are not accurate, the test is based on the conditional distribution of G^2 (or X^2), given $T = t$, i.e. conditioning on the sufficient statistic. This exact conditional

distribution can be explicitly computed using combinatorial methods only for very small sample sizes; however, it can be approximated to arbitrarily high accuracy using the algebraically-enabled Monte Carlo strategy introduced by Diaconis and Sturmfels [6].

3.1. Model Fitting using Markov bases

In algebraic statistics, a log-linear model $\mathcal{M}$ is usually expressed in a power product model formulation

$$p_x = c(\beta)\exp\left(\sum_{i=1}^{d}\beta_i T_i(x)\right) \propto \prod_{i=1}^{d}\theta_i^{T_i(x)}, x \in \mathcal{X},$$

with $\theta_i = \exp(\beta_i)$. Considering $p_x = p_x(\theta)$, the new parameter space is $\Theta = \mathbb{R}^d_{\geq 0}$, with $d > d_{\mathcal{M}}$.

For depicting the connection between modeling contingency tables and computational algebra, log-linear models are connected to matrices possessing the property of *homogeneity*.

Definition 3.1. A matrix $A \in \mathbb{N}^{d\times m}$ is called homogeneous if there exists $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\boldsymbol{\theta}^T A = (1,\ldots,1)$.

This leads to an algebraic representation of a log-linear model.

Definition 3.2. Let $A \in \mathbb{N}^{d\times m}$ be a homogeneous matrix. The log-linear model associated with A is defined by

$$\mathcal{M}_A := \{\boldsymbol{p} = (p_x)_{x\in\mathcal{X}} \in \text{int}(\Delta_{m-1}) : \log \boldsymbol{p} \in \text{rowspan}(A)\}, \tag{10}$$

where rowspan(A) is the linear space spanned by the rows of A.

The matrix A specifies a sufficient statistic function, so that A is sometimes referred to as the sufficient statistic generating matrix. However, due to the parameterization, the sufficient statistics vector $A\mathbf{u}$ is not minimal (rank$A = d_{\mathcal{M}} < d$ while in the GLM formulation usually rank$D = d_{\mathcal{M}}$). For example, for the I model, the minimal sufficient statistics vector is $D^T\mathbf{u} = (n, U_{1+},\ldots,U_{I-1.,+},U_{+1},\ldots,U_{+.J-1})$ while the sufficient statistics vector specified by A is $A\mathbf{u} = (U_{1+},\ldots,U_{I+},U_{+1},\ldots,U_{+J})$ subject to $\sum_{i,j}U_{ij} = n$, where $U_{i+} = \sum_j U_{ij}$ and $U_{+j} = \sum_i U_{ij}$, for $i = 1,\ldots,I$, $j = 1,\ldots,J$, with $d = I + J = d_{\text{I}} + 1$.

Remark 3.3.

1. Note that the matrix A in Definition 1.1.9 of Drton et al. [8] is a special case of A in Definition 3.2 above, since every matrix whose columns all sum to the same value is also homogeneous.

2. For the log-linear independence model (1), the sufficient statistic generating matrix $A \in \mathbb{N}^{(I+J)\times IJ}$ always fulfills the homogeneity assumption since $\theta^T A = (1,\ldots,1)$ with $\theta = (\mathbf{1}_{1\times I}, 0,\ldots,0)$.

The set of contingency tables of the same total size n that lead the same value of the sufficient statistic vector for a log-linear model $\mathcal{M}_A$ is denoted by

$$\mathcal{F}_A(\boldsymbol{u}) := \{\boldsymbol{v} \in \mathbb{N}^m \,|\, A\boldsymbol{v} = A\boldsymbol{u}\} = \{\boldsymbol{v} \in \mathbb{N}^m \,|\, \boldsymbol{v} - \boldsymbol{u} \in \ker_{\mathbb{Z}}(A)\}$$

and is called the **fiber** of the contingency table $\boldsymbol{u} \in \tau(n) = \{\boldsymbol{u} \in \mathbb{N}^m : \sum_i u_i = n\}$ with respect to the log-linear model $\mathcal{M}_A$.

For $\boldsymbol{b} \in \ker_{\mathbb{Z}} A$ we denote the positive elements of $\boldsymbol{b}$ by $\boldsymbol{b}^+ = (b_i^+)_i = (\max(0, b_i))_{i=1,\dots,m}$ and the negative part by $\boldsymbol{b}^- = (b_i^-)_i = -\min(0, b_i)_{i=1,\dots,m}$. Then $\boldsymbol{b}$ can be written as $\boldsymbol{b} = \boldsymbol{b}^+ - \boldsymbol{b}^-$.

The Diaconis-Sturmfels Markov chain algorithm [6] is used to sample from $\mathcal{X}$ conditional on a sufficient statistic, i.e. under the associated assumption for $\boldsymbol{p}(\theta)$. It is based on constructing Markov basis, which comprises moves between tables in a fiber ensuring that every pair of tables in this fiber is connected. The computation of Markov bases for specific log-linear models has been considered by Rapallo [24]. For a review on the concept of the Diaconis-Sturmfels algorithm along with a compact and smooth presentation of the mapping of statistical probabilities to polynomials, we refer to Riccomagno [27]. The log-linear representation is connected to parametric (and therefore binomial) toric models as discussed in Rapallo [26]. It turns out that the toric ideal $\mathcal{I}_T$ corresponding to the related binomial toric model has the following connection to the sufficient statistic generating matrix.

Proposition 3.4. *Consider a log-linear model associated to a sufficient statistic generating matrix $A \in \mathbb{N}^{d\times m}$ and let $\theta_1^{T_1(x)} \cdots \theta_d^{T_d(x)}$ for $x \in \mathcal{X}$ be the power products of its parametrization. The toric ideal of the toric model can then be described as*

$$\mathcal{I}_T = \langle \boldsymbol{p}^u - \boldsymbol{p}^v \,|\, Au = Av \ \text{ for } \ u, v \in \mathbb{N}^m \rangle. \tag{11}$$

Note that if A is the sufficient statistic generating matrix, i.e. $A = (a_{ix})_{\substack{i=1,\dots,d,\\ x\in\mathcal{X}}}$ then a set of generators for $\mathcal{I}_T$ corresponds to a Markov basis $\mathcal{B} \subseteq \ker_{\mathbb{Z}}(A)$ of A.

The computation of Markov bases is related to the problem of computing the corresponding toric ideals. This connection is characterized exactly by the following theorem of Diaconis and Sturmfels [6, Theorem 3.1].

Theorem 3.5. (Fundamental Theorem of Markov Bases - FTMB)
A finite subset $\mathcal{B} \subseteq \ker_{\mathbb{Z}}(A)$ is a Markov Basis for A if and only if the collection of binomials $\{p^{\boldsymbol{b}^+} - p^{\boldsymbol{b}^-} \,|\, \boldsymbol{b} \in \mathcal{B}\}$ is a generating set for the toric ideal $\mathcal{I}_T(A)$, i.e.

$$\langle p^{\boldsymbol{b}^+} - p^{\boldsymbol{b}^-} \,|\, \boldsymbol{b} \in \mathcal{B}\rangle = \mathcal{I}_T(A).$$

The property of homogeneity for matrices can be translated directly to ideals, as stated next.

Proposition 3.6. *[28]*
The toric ideal $\mathcal{I}_T(A)$ is homogeneous if and only if the corresponding matrix A fulfills the assumption of homogeneity.

Remark 3.7. It can be shown that the minimal sets of generators of $\mathcal{I}_T$ have the same number of elements (see Kreuzer and Robbiano [19, Proposition 4.1.22]). Note that Buchberger's algorithm for computing Gröbner bases of ideals (see [5]) preserves homogeneity. Hence, with a minimal set of generators of $\mathcal{I}_T$ we can deduce a minimal Markov basis and the FTMB reduces the problem of determining minimal Markov bases for the Metropolis algorithm to computing a minimal generating set of a corresponding toric ideal.

3.2. Model Selection

Whenever asymptotic inferential approaches are not applicable, the algebraic methods discussed so far can be used for model selection, as demonstrated by Krampe and Kuhnt [18]. However, when considering numerous different models the Diaconis-Sturmfels algorithm has to be applied several times, making the model selection computational very expensive. The computational costs can be reduced when the models considered are downsized to sequences of nested models by exploiting their structures.

Consider two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with $\mathcal{M}_1$ being nested in $\mathcal{M}_2$, i.e. $\mathcal{M}_1 \subseteq \mathcal{M}_2$. Then $G^2(\mathcal{M}_1) \geq G^2(\mathcal{M}_2)$. If $\mathcal{M}_2$ is acceptable for modeling the observed data, the decision whether the simpler model $\mathcal{M}_1$ could be adopted is based on the conditional testing of model $\mathcal{M}_1$, given that $\mathcal{M}_2$ holds. Asymptotically,

$$G^2(\mathcal{M}_1|\mathcal{M}_2) = G^2(\mathcal{M}_1) - G^2(\mathcal{M}_2) \sim \chi^2_{df(\mathcal{M}_1)-df(\mathcal{M}_2)} \ ,$$

providing a test that is more powerful than the unconditional based on $G^2(\mathcal{M}_1)$.

It follows that the sufficient statistic $T_{\mathcal{M}_2}$ for the parameters of model $\mathcal{M}_2$ contains the sufficient statistics $T_{\mathcal{M}_1}$. Let $\mathcal{F}_t$ be the set of all contingency tables with the same sufficient statistic value $T = t$. Model $\mathcal{M}_2$ comprises of more restrictions than model $\mathcal{M}_1$. Therefore, the set $\mathcal{F}_{t_{(\mathcal{M}_1)}}$ contains $\mathcal{F}_{t_{(\mathcal{M}_2)}}$, i.e. $\mathcal{F}_{t_{(\mathcal{M}_2)}} \subseteq \mathcal{F}_{t_{(\mathcal{M}_1)}}$. This hierarchical structure is passed to the corresponding Gröbner bases.

Theorem 3.8. *[18]*
Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two log-linear models with $\mathcal{M}_1 \subseteq \mathcal{M}_2$. Following the Diaconis-Sturmfels approach, if $\mathcal{I}_{\mathcal{M}_1}$ and $\mathcal{I}_{\mathcal{M}_2}$ are the corresponding elimination ideals, $\mathcal{I}_{\mathcal{M}_1} \supseteq \mathcal{I}_{\mathcal{M}_2}$.

For a sequence of nested models

$$\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots \subseteq \mathcal{M}_k$$

Theorem 3.8 allows us to generate in the Metropolis algorithm only one Markov chain for the most parsimonious model $\mathcal{M}_1$. For every other model $\mathcal{M}_i$, $i = 2, \ldots, k$, chains can be extracted from this Markov chain.

From a practical perspective, it should be mentioned that the accuracy of the approximation of this method depends on the number of simulated tables from $\mathcal{F}_{t_{(\mathcal{M}_1)}}$ with $T_{\mathcal{M}_i} = t_{\mathcal{M}_i}, i = 1, \ldots, k$. This number is bounded by the number of steps of the Markov chain on $\mathcal{F}_{t_{(\mathcal{M}_1)}}$ and can decrease extremely for models $\mathcal{M}_i$, $i > 1$, as their complexity (i.e. $d_{\mathcal{M}_i}$) increases. Furthermore, the larger the sample size and therefore the larger the number of possible tables that could be generated, the smaller becomes this number. Hence,

the benefit of algebraic model selection can only be ensured if the length of the simulated Markov chain is adjusted. If the adjusted chain length turns out to be too large, it might be more efficient to conduct several distinct MCMC simulations. For more details on this algebraic model selection approach and its performance we refer to [18].

4. Algebraic Inference for Association Models

Since the association models considered in Section 2 are special log-linear models, the algebraic methods developed for log-linear models goodness-of-fit testing are applicable also to them. Thus, the corresponding p-values can be simulated by a Metropolis algorithm, which will be based on conditioning on the sufficient statistics and will use Markov bases of the respective model to perform a random walk on the fiber. Hence, in order to apply the algorithm, we need to specify the sufficient statistic generating matrix for the association models.

4.1. Markov Bases for Association Models

For the LL model the row and column score vectors $u = (u_1, \ldots, u_I)^T$, $v = (v_1, \ldots, v_J)^T$, respectively, are fixed and the joint probability distribution is $\mathbf{p} = (p_{ij})$ with

$$p_{ij} = \frac{\exp(\lambda)}{n} \exp(\lambda_i^X) \exp(\lambda_j^Y) \exp(\beta u_i v_j) \propto x_i y_j z^{u_i v_j} \ ,$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. In order to apply algebraic methods we need $u_i v_j \in \mathbb{N}$. It can be easily verified that the sufficient statistics of the LL model are given by $\mathbf{U}^X = (U_{1+}, \ldots, U_{I+})$, $\mathbf{U}^Y = (U_{+1}, \ldots, U_{+J})$ and $\sum_i \sum_j u_i v_j U_{ij}$. Let the sufficient statistic vector be

$$T_{\mathrm{LL}} = (\mathbf{U}^X, \mathbf{U}^Y, \sum_i \sum_j u_i v_j U_{ij}).$$

An explicit form of the corresponding sufficient statistic generating matrix A_{LL}, $A_{\mathrm{LL}} \in \mathbb{N}^{(I+J+1)\times IJ}$, can be given using the Kronecker product notation:

$$A_{\mathrm{LL}} = \begin{pmatrix} E_I \otimes \mathbf{1}_J^T \\ \mathbf{1}_I^T \otimes E_J \\ u^T \otimes \boldsymbol{v}^T \end{pmatrix} \tag{12}$$

where E_m denotes an $m \times m$ identity matrix and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ denotes the n-dimensional vector consisting of 1s. The rank of A_{LL} is equal to the dimension of $\mathrm{Im}(T_{\mathrm{LL}})$ which is equal to the number of linearly independent parameters. Note that $\mathbf{U}^X$ and $\mathbf{U}^Y$ are not minimal, since $\sum_i U_{i+} = \sum_j U_{+j} = n$. Thus $\mathrm{rank} A_{\mathrm{LL}} = I + J$ and hence $df(\mathrm{U}) = \dim \ker A_{\mathrm{LL}} = IJ - \mathrm{rank} A_{\mathrm{LL}} = (I-1)(J-1) - 1$. For $I, J > 2$ there exist non-trivial elements in the matrix kernel that could serve as moves in a Markov basis. The corresponding binomial ideal is

$$\mathcal{I} = \langle p_{ij} - x_i y_j z^{u_i v_j}, i = 1, \ldots, I, j = 1, \ldots, J \rangle.$$

It can be shown that the reduced Gröbner Basis of the elimination ideal

$$\mathcal{I}_T(A_{\text{LL}}) = \mathcal{I} \cap \mathbb{R}[p_{ij}, i = 1, \ldots, I, j = 1, \ldots, J]$$

is generated from binomials too. By Remark 3.7 its minimal sets of generators have the same number of elements (and therefore there exist corresponding minimal Markov bases) if the ideal $\mathcal{I}$ is homogeneous. By Proposition 3.6 the toric ideal is homogeneous if and only if the sufficient statistic generating matrix fulfills the homogeneity assumption. Hence, the question arises whether the matrix associated to the LL model is homogeneous.

Proposition 4.1. *Consider a* LL *model with row and column scores vectors* u *and* v*, respectively. The associated matrix* A_{LL} *given by* (12) *fulfills the assumption of homogeneity.*

Proof. It needs to be shown that there exists a real vector $\boldsymbol{\theta} \in \mathbb{R}^{(I+J+1)}$ such that $A_{\text{LL}}^T \boldsymbol{\theta} = \mathbf{1}^T$ where $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^{IJ}$. Consider the sufficient statistic generating matrix A_{I} for the independence model that is the $(I + J) \times IJ$ submatrix of A_{LL} where the last row $u^T \otimes v^T$ is not included. By Remark 3.3, the matrix is homogenous. Indeed, if we choose $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{(I+J)}$ with $\tilde{\boldsymbol{\theta}}_{I+J}$ arbitrary and $\tilde{\boldsymbol{\theta}}_i = -\tilde{\boldsymbol{\theta}}_{I+J} + 1, \tilde{\boldsymbol{\theta}}_{I+j} = \tilde{\boldsymbol{\theta}}_{I+J}$ for $i = 1, \ldots, I, j = 1, \ldots, J$, we get $A_{\text{I}}^T \tilde{\boldsymbol{\theta}} = \mathbf{1}$. Then for $\boldsymbol{\theta}^T = (\tilde{\boldsymbol{\theta}}^T, 0)$ it holds $A_{\text{LL}}^T \boldsymbol{\theta} = \mathbf{1}^T$ and thus A_{LL} is homogeneous.

Since homogeneity is preserved when computing the Gröbner basis (see [19]), a minimal Markov basis of $\ker_{\mathbb{Z}}(A)$ is well-defined and can be computed by the FTMB (Theorem 3.5).

The R model is based on known and fixed column scores $v = (v_1, \ldots, v_J)^T$ and parametric row scores $\mu = (\mu_1, \ldots, \mu_I)$. Similarly to the LL model, the joint probability of the R model is determined by

$$p_{ij} = \frac{\exp(\lambda)}{n} \exp(\lambda_i^X) \exp(\lambda_j^Y) \exp(\mu_i v_j) = z_0 x_i y_j w_i^{v_j} \propto x_i y_j w_i^{v_j} \ .$$

The sufficient statistic is given by

$$T_{\text{R}} = (\mathbf{U}^X, \mathbf{U}^Y, \mathbf{U}^R),$$

where $\mathbf{U}^R = (\sum_j v_j U_{1j}, \ldots, \sum_j v_j U_{Ij})$. The corresponding sufficient statistic generating matrix A_{R}, $A_{\text{R}} \in \mathbb{N}^{(2I+J)\times IJ}$, and the binomial ideal are

$$A_{\text{R}} = \begin{pmatrix} E_I \otimes \mathbf{1}_J^T \\ \mathbf{1}_I^T \otimes E_J \\ E_I \otimes v^T \end{pmatrix} \text{ and } \ \mathcal{I} = \langle p_{ij} - x_i y_j w_i^{v_j}, i = 1, \ldots, I, j = 1, \ldots, J \rangle. \tag{13}$$

Matrix A_{R} has $\text{rank} A_{\text{R}} = 2I + J - 2$ and thus $df(\text{R}) = \dim \ker A_{\text{R}} = (I - 1)(J - 2)$.

Similarly to the LL model, the corresponding sufficient statistic generating matrix of the R model is homogeneous. The proof is analogous to that of Proposition 4.1, with $\boldsymbol{\theta}^T = (\tilde{\boldsymbol{\theta}}, \mathbf{0}_I)$ where $\mathbf{0}_I$ is the I-dimensional vector of zeroes.

Proposition 4.2. *Consider an* R *model with fixed column scores* v *and parametric row scores* $\boldsymbol{\mu}$*. The associated matrix* A *given by* (13) *fulfills the assumption of homogeneity.*

It is straightforward to obtain the analogous results for the C model, for which $df(\mathrm{C}) = \dim\ker A_{\mathrm{C}} = (I-2)(J-1)$, with

$$A_{\mathrm{C}} = \begin{pmatrix} E_I \otimes \mathbf{1}_J^T \\ \mathbf{1}_I^T \otimes E_J \\ \mathbf{u}^T \otimes E_J \end{pmatrix} \quad \text{and} \quad \mathcal{I} = \langle p_{ij} - x_i y_j z_j^{u_i}, i = 1, \ldots, I, j = 1, \ldots, J \rangle \ . \tag{14}$$

4.2. Model Selection for Association Models

For two-way tables and the association models considered we have the following possible sequences of nested models

$$\mathrm{I} \subseteq \mathrm{U}\ (\text{or LL}) \subseteq \mathrm{R}\ (\text{or C}) \ , \tag{15}$$

provided the known column (row) scores of the R (C) model are the same to the corresponding scores of the U (or LL) model. The model selection procedures discussed in Section 3.2 applies directly and Theorem 3.8 leads to the following result.

Lemma 4.3. *Let* $\mathcal{I}_{\mathrm{I}}, \mathcal{I}_{\mathrm{LL}}, \mathcal{I}_{\mathrm{R}}, \mathcal{I}_{\mathrm{C}}$ *denote the elimination ideals from the Diaconis-Sturmfels algorithm for the* I, LL, R, C *model, respectively. Then it holds*

(i) $\mathcal{I}_I \supseteq \mathcal{I}_{\mathrm{LL}} \supseteq \mathcal{I}_{\mathrm{R}}$ *and*

(ii) $\mathcal{I}_I \supseteq \mathcal{I}_{\mathrm{LL}} \supseteq \mathcal{I}_C$.

It is straightforward to adjust these results to nested association models for tables of higher dimension. Pham [22] considered model selection for association models for three-way tables.

5. Examples

We shall apply the algebraic statistics approach on well–known examples of the contingency tables literature, two two-dimensional and one three–dimensional. They are all worked out in `R` using functions for association models fitting by [17] and the `R` package `algstat` by Kahle, Garcia-Puente and Yoshida ([16]). Though `R` itself has no base support for symbolic computation, `algstat` provides some functionality for algebraic statistics in `R` having ports to Macaulay2, Bertini, LattE-integrale and 4ti2. Technical details are omitted and can be found in Pham [22].

5.1. Boys' Disturbed Dreams

Consider the 5×4 table of Boys' Disturbed Dreams (Maxwell [20, p.70]), depicted in Table 1. The study cross-classified boys by their age and the severity of their disturbed

Table 1: Boys' Disturbed Dreams by Age. In parentheses are provided the MLEs of the expected frequencies under the C model.

	Degree of Suffering (ordinal)				
Age	Not severe (1)	(2)	(3)	Very severe (4)	Totals
5-7	7 (4.047)	4 (5.712)	3 (4.476)	7 (6.765)	21
8-9	10 (14.252)	15 (11.878)	11 (10.395)	13 (12.475)	49
10-11	23 (20.564)	9 (10.120)	11 (9.891)	7 (9.425)	50
12-13	28 (31.930)	9 (9.279)	12 (10.128)	10 (7.664)	59
14-15	32 (29.208)	5 (5.012)	4 (6.110)	3 (3.671)	44
Totals	100	42	41	40	223

dreams. The variables age (X) and degree of suffering (Y) are measured on a five-level interval scale and a four-level ordinal scale, respectively.

We fitted the Independence (I) and the LL, R and C models to the data in Table 1. Scores for interval scales can be decided naturally as the midpoints of the intervals. Since in our set-up scores need to be integers, we consider the row scores $u = (6, 8, 10, 12, 14)$. Scores for ordinal scales are subjective. For the column scores we set $v = (1, 2, 3, 10)$, emphasizing the last category. The G^2 goodness-of-fit tests, along with the asymptotic and the algebraic approximated exact p-values, are given in Table 2.

For this example, the asymptotic and algebraic strategies provide similar results, as probably expected, since, for example for the C model, the estimated expected frequencies in 15% ($< 20\%$) of the cells are smaller than 5. However, in tables with small cell entries, the adequacy of the chi-squared approximation for G^2 can not always be guaranteed by the well-known 20% condition of Cochran. It is difficult to give guidelines that cover all cases of possibly poor chi-squared approximations and is thus suggested to apply small-sample methods as well, whenever the approximation is doubtful [2, Section 3.2.3].

There is a strong evidence against independence. The model that best describes these data is the C model. At significance level $\alpha = 0.10$, this is the only acceptable model. Conditional testing of the LL model, given that the C model holds, is based on

Table 2: G^2 goodness-of-fit tests for the models applied in Table 1.

			p-value	
Model	G^2	d.f.	algebraic	asympt.
I	32.457	12	0.0019	0.0012
LL	18.308	11	0.0950	0.0747
R	16.524	8	0.0396	0.0355
C	7.459	9	0.6179	0.5895

$G^2(\text{LL}\,|\text{C}\,) = 10.849$, which is asymptotically χ^2_2 distributed and gives a p-value of 0.0044. The corresponding algebraic approximated p-value equals 0.0909. Thus, for $\alpha = 0.01$, the

algebraic model selection suggests the simpler LL model while the asymptotic approach the C model.

5.2. Pathologists' Diagnosis of Carcinoma

The 4×4 table, given in Table 3 can be found in Agresti [2, Section 11.5]. It cross-classifies the ratings by two pathologists, labeled A and B, who separately classified 118 slides regarding the presence and extent of carcinoma of the uterin cervix. Their common rating scale has the ordered categories (1) negative, (2) atypical squamous hyperplasia,(3) carcinoma in situ,(4) squamous or invasive carcinoma. This data set is a raters agreement

Table 3: Diagnosis of Carcinoma Data. In parentheses are provided the MLEs of the expected frequencies under the LL model.

	Pathologist B				
Pathologist A	1	2	3	4	Total
1	22 (21.608)	2 (3.428)	2 (0.963)	0 (0.001)	26
2	5 (5.189)	7 (6.461)	14 (14.255)	0 (0.095)	25
3	0 (0.201)	2 (1.966)	36 (34.048)	0 (1.785)	38
4	0 (0.002)	1 (0.145)	17 (19.734)	10 (8.119)	28
Total	27	12	69	10	118

problem and can be analyzed by models for matched pairs data. Agresti fitted in his analysis the quasi-independence (QI) model but concluded that the quasi-symmetry (QS) model fits better, based on asymptotic testing [2, Section 11.5.3]. For testing these models, Rapallo [24] provided the algebraic approximation of the corresponding p-values (QI: p-value $= 0.008$, QS: p-value $= 1$). We shall apply the association models and and test the model fit via algebraic statistics, since the table is sparse.

We assigned the scores $u = v = (1, 2, 3, 4)$ to the categories of the classification variables and fitted the models I, U, R and C. Notice that the last column contains three (out of four) zero cells (sampling zeros). Thus, we suspect that there might cause problems in estimating the last column score parameter ν_4 for the C model. Indeed, in the estimation procedure in `R`, though estimates for the model parameters are provided, their standard errors are huge alarming for the correctness of the MLEs (implication of the sparsity of the table). It can be verified that one of the likelihood equations of the C model, reduces to $u_{44} = \hat{m}_{44}$. Thus, the cell $(4, 4)$ is in fact fixed. Fixed cells are treated analogous to structural zeros, i.e. they are set equal to the fixed value and the model's *df* are corrected accordingly. In this case, since $u_{+4} = u_{44}$, the whole last column turns out to be fixed.

Algebraic hypothesis testing of the independence (I) and the association models U, R and C leads to the following table of analysis of association (ANOAS) based on the likelihood-ratio test statistic. Note the corrected degrees of freedom for the C model. The table shows the model, the corresponding likelihood-ratio statistic G^2, the asymptotic and algebraic simulated p-value.

Model	Test statistic	d.f.	p-value (alg.)	p-value (approx.)
1 I	117.9569	9	0.00000	0.0000
2 LL	8.8422	8	0.13830	0.3558
3 R	7.8447	6	0.05853	0.2497
4 C	2.2235	4	0.56155	0.6947
5 I|LL	109.1147	1	0.00000	0.0000
6 I|R	110.1121	3	0.00000	0.0000
7 I|C	115.7333	5	0.00000	0.0000
8 LL|R	0.9974	2	0.63752	0.6073
9 LL|C	6.6186	4	0.12331	0.1575

The advantage of the association models over the QS model for this data set lies on the fact that under QS the diagonal cells are not modeled (all are kept fixed), which excludes in this case almost 20% of the data.

5.3. FDA Toxicology Study

The data for this example are from the U.S. Food and Drug Administration (FDA) and can be found for example in [21]. Animals were treated with four dose levels of a carcinogen (Y) and then observed (at necropsy) for the presence or absence of a tumor type (X). The data is stratified by survival time (Z: in weeks) into four time intervals $0-50, 51-80, 81-104$ and terminal sacrifice. As there were no tumors found in the first time interval, this stratum is not included in this analysis. The data for the remaining three strata are displayed in Table 4.

Table 4: FDA Animal Toxicology Data. In parentheses are provided the MLEs of the expected frequencies under the $(XY_{\text{LL}}, XZ_{\text{LL}})$ model.

	Dose of Carcinogen				
	None	1 unit	5 units	50 units	Total
Disease Status	Stratum 1: 51-80 weeks of survival				
Tumor Present	0 (0.148)	0 (0.168)	0 (0.227)	1 (0.5)	1
Tumor Absent	7 (8.015)	10 (7.819)	6 (7.911)	8 (7.212)	31
	Stratum 2: 81-104 weeks of survival				
Tumor Present	0 (0.272)	1 (0.308)	0 (0.417)	1 (0.916)	2
Tumor Absent	11 (12.191)	9 (11.892)	13 (12.033)	14 (10.970)	47
	Stratum 3: Sacrificed at the end of 104 weeks				
Tumor Present	1 (0.718)	1 (0.811)	1 (1.100)	2 (2.415)	5
Tumor Absent	29 (26.656)	26 (26.003)	28 (26.311)	20 (23.986)	103

We tested the model $(XY_{\text{LL}}, XZ_{\text{LL}})$, i.e. whether the survival time is conditionally independent from the dose, given the disease status. The column and layer scores for this model shall be given by $v = (1, 2, 4, 10)$ and $w = (1, 2, 3)$. Note that intuitively the

column scores would be chosen as $(0, 1, 5, 50)$ or $(1, 2, 6, 51)$, respectively. However, large scores can cause difficulties in Markov basis computation as the corresponding polynomials ideals would contain polynomial with a high degree. For the latter choice of scores, the Markov basis computation using the `algstat` package in `R` does not terminate even after 7 hours. Therefore, we decided to use smaller scores that still reflect the different doses of Carcinogen. The sufficient statistic generating matrix corresponding to this model is given by

$$A = \begin{pmatrix} E_I \otimes \mathbf{1}_J^T \otimes \mathbf{1}_K^T \\ \mathbf{1}_I^T \otimes E_J \otimes \mathbf{1}_K^T \\ \mathbf{1}_I^T \otimes \mathbf{1}_J^T \otimes E_K \\ E_I \otimes \mathbf{v}^T \otimes \mathbf{1}_K^T \\ E_I \otimes \mathbf{1}_J^T \otimes \mathbf{w}^T \end{pmatrix}.$$

The observed test statistic is $G^2 = 8.082$ with $df = 15$. Thus, the asymptotic p-value equals 0.9205. The corresponding algebraic simulated p-value is 0.9998.

6. Simulation Study

We have seen in the examples above that for sparse tables the algebraically approximated p-value can be smaller or larger from the corresponding asymptotic one. In order to evaluate the performance of algebraic p-value and compare them to the asymptotic ones, performed a small simulation study. We simulated 4×4 data tables from unconditional distributions of U models. We assumed fixed row and column scores for all simulated tables given by $(1, 2, 3, 4)$. The parameters $\lambda, \lambda_i^X, \lambda_j^Y$ and z for $i, j = 1, \ldots, 4$ are uniformly sampled from a range $(0, 0.1]$ and then used to compute the expected cell frequencies. The chosen range ensures that the sample size of the simulated data is not too large. The sample size distribution of the data tables is depicted in Figure 1. Assuming the independent Poisson sampling scheme, we randomly draw the cell frequencies U_{ij} by using the R function `rpois(s, mu)` where `s` is the number of random samples to be returned and `mu` the vector of (non-negative) expected values. We restrict our simulation study to tables with at most 3 zero cell frequencies. By doing that, we expect simulation tables with defined maximum likelihood estimators.

The results of the algebraic simulated and approximate p-values are plotted in a diagram and depicted in Figure 2. In the diagram on the left, the dotted lines correspond to a significance level $\alpha = 0.05$. The diagram on the right shows the p-values that are smaller or equal to $\alpha = 0.05$.

The fourth quadrant (with respect to the dotted lines of the right diagram of Figure 2) depicts the p-values where we would come to a different test decision at a significance level $\alpha = 0.05$. In fact, in 24 out of 1000 data sets the null hypothesis would be rejected by the approximate test whereas it would not be rejected by the algebraic method. In these cases the approximate p-value is smaller than then significance level $\alpha = 0.05$ which leads to rejecting the null hypothesis whereas they would not be rejected by the algebraic method as the respective algebraic simulated p-values are larger than α.

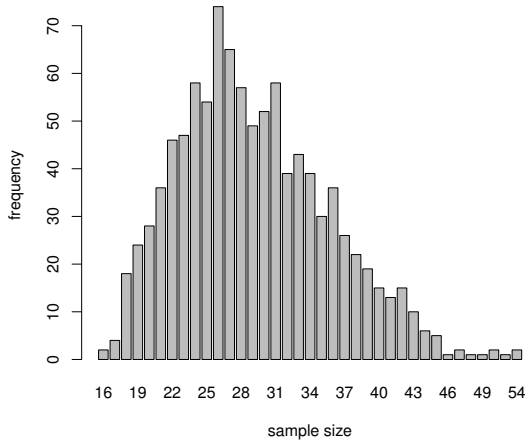

Figure 1: Frequency distribution of sample sizes of simulated data tables.

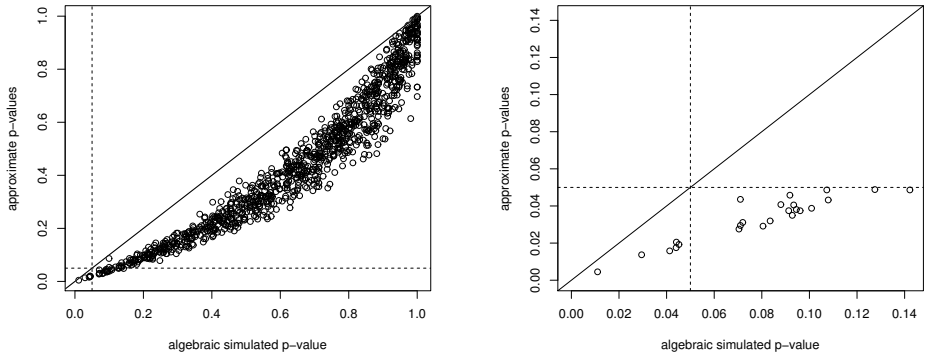

Figure 2: Comparison of simulated and approximate p-values from the Likelihood Ratio test for data simulated from U model with various sample sizes. The *dashed line* represents $\alpha = 0.05$.

We next examine the type-I-error of the two methods. In 6 out of 1000 data sets the algebraic method rejects the null hypothesis incorrectly whereas the traditional approximate procedure rejects 24 out of 1000 data tables falsely. Both methods, and especially the algebraic one, seem to be too conservative, since their actual sizes (i.e. 0.006 and 0.024) are far below the nominal size (0.05). The conservatism issue is partly unavoidable in case of small samples or sparse contingency tables [2, Sections 3.5, 16.5]. For this, it is suggested to base inference on adjustments of the p-value (like the mid p-value).

The maximum difference between the respective p-values is 0.3981, achieved for the following frequency table of sample size $n = 18$

```
      [,1] [,2] [,3] [,4]
[1,]     1    1    1    1
```

```
[2,]    1    1    2    1
[3,]    1    0    3    0
[4,]    0    2    1    2
```

For this table the $G^2(\mathrm{U}) = 8.078$, while the asymptotic p-value ($= 0.4259$) is inappropriate, since all the cell frequencies are very small (≤ 3). The algebraic p-values equals 0.8240. The MLEs of the expected cell frequencies under this model are

```
        1         2         3         4
1 0.9377356 1.0204964 1.4171104 0.6246576
2 0.9321875 1.1806791 1.9081901 0.9789433
3 0.5795058 0.8542480 1.6068351 0.9594112
4 0.5505711 0.9445766 2.0678644 1.4369879
```

Theoretically in such cases, the asymptotic p-values tend to be too small (see Agresti [2]), which can be confirmed in our simulation.

7. Discussion

In this paper we considered a class of log-linear models for contingency tables, applicable when at least one of the classification variables is ordinal, known as association models. Targeting in developing algebraic exact inference procedures for goodness of fit testing, we defined association models as toric models. Special emphasis is given in three models of this class for two-way tables, namely the linear by linear (LL), the row effect (R) and the column effect (C) model. The most popular uniform (U) association model is a LL model with equidistant row and column scores for successive classification categories.

Algebraic goodness of fit testing of a model is in practice enabled due to the Diaconis-Sturmfels algorithm, based on computing Markov bases for this model. A key-result states that Markov bases are equivalent to generating sets of toric ideals and the problem of computing a Markov basis reduces to the problem of finding a Gröbner basis for toric ideals. Determining minimal Markov bases is based on setting up matrices of sufficient statistics that fulfill the assumption of homogeneity.

For the association models mentioned above we derived their sufficient statistic generating matrices (A), which specify their sufficient statistics, and proved that they are homogeneous. The algorithm is then implemented in the `R`-package `algstat` and illustrated on three data sets with ordered categories (two two-way tables and one three-way). The algebraic p-values are compared with the corresponding asymptotic ones. This comparison is further highlighted by a small simulation study. The conservatism of algebraic inferential methods in case of small samples or sparse tables needs to be further studied.

In designing the corresponding algorithms, attention must also be paid on their efficiency. The efficiency of computing Markov basis relies mainly on the efficiency of Gröbner basis computation which is determined by the computational complexity of Buchberger's algorithm. Hence, Markov basis computation may not be efficient, especially for polynomials with high total degrees and hence matrices of sufficient statistics with large entries (Dubé [9]). This can be the case very easily for association models as the ordered scores

assigned to the classification variables categories have to be integers. Especially for scores with large distances between successive categories the computation of Markov bases can become prohibitive. It is worth to study further the algebraic structure of their corresponding varieties and toric ideals in order to obtain simplification for Gröbner basis computation.

With regard to model selection among nested models, recall that in order to reduce the computational costs, the adopted approach in Section 3.2 constructs only one Markov chain (corresponding to the simplest model) and the chains for the other models are the appropriate subsets of this chain that fulfill the additional constraints imposed by their additional sufficient statistics. However, the length (N) of the simulated Markov chain needs to be sufficiently large to ensure that the sub-chains are also sufficiently large. For instance, in Example 5.1 the independence model is the simplest one and thus serves as the model for which the Markov chain is built. For this example we have realized that the length of selected chains are very small even for large number of iterations for the Markov chain of the independence model. In fact, for $N = 1000000$ the selected chain for the LL model has length 4. Therefore, in this case we simulated an individual Markov chain for each model considered. Thus, further investigation on how to reduce computation costs for algebraic p-value simulation for ordinal association models is needed.

Finally note that we have not considered here the association model having parametric both, the row and the column scores. In this case the model, called multiplicative Row-Column association model (RC model), is multiplicative in its parameters and thus log-nonlinear. It would be interesting to develop an algebraic inferential approach for the RC model.

Acknowledgements

The authors thank the referees for their constructive and very useful comments that improved the presentation of the paper.

References

[1] Alan Agresti. *Analysis of Ordinal Categorical Data*. Wiley, Hoboken, 2nd edition, 2010.

[2] Alan Agresti. *Categorical Data Analysis*. Wiley, Hoboken, 3rd edition, 2013.

[3] Satoshi Aoki and Akimichi Takemura. Minimal basis for a connected markov chain over $3 \times 3 \times$k contingency tables with fixed two-dimensional marginals. *Australian & New Zealand Journal of Statistics*, 45(2):229–249, 2003.

[4] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass., 1975.

[5] Bruno Buchberger. Gröbner-bases: An algorithmic method in polynomial ideal theory. In Nirmal K. Bose, editor, *Multidimensional Systems Theory*, pages 184–232. Reidel Publishing Company, Dodrecht, Holland, 1985.

[6] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26:363–397, 1998.

[7] Adrian Dobra, Stephen E. Fienberg, Alessandro Rinaldo, Aleksandra Slavkovic, and Yi Zhou. Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In *Emerging Applications of Algebraic Geometry*, pages 63–88. Springer, New York, 2009.

[8] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser, Basel, 2009.

[9] Thomas W. Dubé. The structure of polynomial ideals and gröbner bases. *SIAM J. Comput.*, 19:750–773, 1990.

[10] Nicholas Eriksson, Stephen E. Fienberg, Alessandro Rinaldo, and Seth Sullivant. Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *J. Symbolic. Comput.*, 41:222–233, 2006.

[11] Stephen E. Fienberg. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, Mass., 1980.

[12] Stephen E. Fienberg and Alessandro Rinaldo. Maximum likelihood estimation in log-linear models. *Ann. Statist.*, 40:996–1023, 2012.

[13] Leo A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74:537–552, September 1979.

[14] Leo A. Goodman. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.*, 13:10–69, 1985.

[15] Leo A. Goodman. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Int. Stat. Rev.*, 54:243–270, 1986.

[16] Kahle, David and Garcia-Puente, Luis. `algstat 0.0.2` - Algebraic statistics in `R`. `https://cran.r-project.org/web/packages/algstat/algstat.pdf`, 2014.

[17] Maria Kateri. *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser/Springer, New York, 2014.

[18] Anne Krampe and Sonja Kuhnt. Model selection for contingency tables with algebraic statistics. In Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn, editors, *Algebraic and Geometric Methods in Statistics*, pages 83–98. Cambridge University Press, 2009.

[19] Martin Kreuzer and Lorenzo Robbiano. *Computational Commutative Algebra 2*. Springer, Heidelberg, 2005.

[20] Albert E. Maxwell. *Analysing Qualitative Data*. Methuen, New York, 1961.

[21] Cyrus R. Mehta and Nitin R. Patel. Exact inference for categorical data. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*, pages 1411–1422. Wiley, Chichester, 1998.

[22] Thi Mui Pham. Torische statistische modelle: parametrische und binomiale repräsentationen. *manuscript*, 2015.

[23] R Core Team. R: A language and environment for statistical computing. `https://www.R-project.org/`, 2018.

[24] Fabio Rapallo. Algebraic markov bases and mcmc for two-way contingency tables. *Scand. J. Statist.*, 30:385–397, 2003.

[25] Fabio Rapallo. Markov bases and structural zeros. *J. Symbolic Comput.*, 41:164–172, 2006.

[26] Fabio Rapallo. Toric statistical models: parametric and binomial representations. *Ann. Inst. Statist. Math.*, 59:727–740, 2007.

[27] Eva Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.

[28] Bernd Sturmfels. *Gröbner Bases and Convex Polytopes*. Univ. Lecture Series, No. 8. American Mathematical Society, Providence, Rhode Island, 1996.