JOURNAL OF ALGEBRAIC STATISTICS
Vol. 6, No. 1, 2015, 1-16
ISSN 1309-3452 – www.jalgstat.com

A Family of Quasisymmetry Models

Maria Kateri[1], Fatemeh Mohammadi[2,*], Bernd Sturmfels[3]

[1] *Institute of Statistics, RWTH Aachen University, 52056 Aachen, Germany*
[2] *Institut für Mathematik, TU Berlin, 10623 Berlin, Germany*
[3] *Department of Mathematics, University of California, Berkeley, CA 94720, USA*

Abstract. We present a one-parameter family of models for square contingency tables that interpolates between the classical quasisymmetry model and its Pearsonian analogue. Algebraically, this corresponds to deformations of toric ideals associated with graphs. Our discussion of the statistical issues centers around maximum likelihood estimation.

2000 Mathematics Subject Classifications: 05E40, 13P25, 62H17

Key Words and Phrases: Square contingency tables, Algebraic statistics, Toric models, Linear models, Maximum likelihood estimation, ϕ-divergence.

1. Introduction

Consider a square contingency table with commensurable row and column classification variables X and Y. Such tables can arise from cross-classifying repeated measurements of a categorical response variable. They are common in panel and social mobility studies. One of the most cited examples, taken from Stuart [12], is shown in Table 1. It cross-classifies 7477 female subjects according to the distance vision levels of their right and left eyes.

	Left Eye Grade			
Right Eye Grade	best	second	third	worst
best	1520	266	124	66
second	234	1512	432	78
third	117	362	1772	205
worst	36	82	179	492

Table 1: Cross classification of 7477 women by unaided distance vision of right and left eyes.

The most parsimonious model for such tables is the symmetry (S) model, due to Bowker [3]. While the S model is easy to interpret, it is too restrictive and rarely fits well. An important model that often fits better is the quasi-symmetry (QS) model of Caussinus [4]. Kateri and Papaioannou [7] studied the QS model from the information-theoretic point of view and generalized it to a family of models based on the ϕ-divergence [10]. In their framework, classical QS is closest to the S model under the Kullback-Leibler divergence. However, by changing the divergence used to measure proximity of distributions, alternative QS models are found. For instance, the Pearsonian divergence yields the Pearsonian QS model. For the data in Table 1, Bishop *et al.* [2] applied the QS model, while Kateri and Papaioannou [7] applied the Pearsonian QS model, and here these two lead to estimates of similar fit. However, there are other data sets where only one of them performs well. Our goal is to link these two models. We

*Corresponding author.

Email addresses: maria.kateri@rwth-aachen.de (M. Kateri), fatemeh.mohammadi716@gmail.com (F. Mohammadi), bernd@berkeley.edu (B. Sturmfels)

shall construct a one-parameter family of QS models that connects these two. In this way, more options for data analysis are available. In case of a single square contingency table, the optimal choice of this model parameter would be of interest. However, the more interesting practical application lies in analyzing and comparing independent square tables of the same set-up, when they cannot be modeled adequately all by the same (classical or Pearsonian) QS model. For example, consider the same panel study carried out at two independent centers, with one of them being modeled only by the classical QS and the other only by the Pearsonian QS. In this scenario, the two fitted models are not as comparable as we would like. Our approach furnishes in-between compromise models. Our family exhibits interesting properties when viewed from the perspective of algebraic statistics [5]. It interpolates between two fundamental classes of discrete variable models, namely, toric models and linear models [9, §1.2]. Indeed, the QS model is toric, and its Markov basis is well-known, by work of Rapallo [11] and Latunszynski-Trenado [5, §6.2]. The Pearsonian QS model reduces to a linear model, specified by the second factors in (3). Its ML degree is the number of bounded regions in the arrangement of hyperplanes $\{a_i - a_j = 1\}$, by Varchenko's formula [9, Theorem 1.5].

This paper is organized as follows. Our parametric family of QS models is introduced in Section 2. In Section 3 we derive the implicit representation of our model by polynomial equations in the cell entries. That section is written in the algebraic language of ideals and varieties. It will be of independent interest to scholars in combinatorial commutative algebra [8, 13]. Maximum likelihood estimation (MLE) and the fit of the model are discussed in Section 4. Section 5 examines a natural submodel given by independence constraints. Section 6 discusses statistical applications and presents computations with concrete data sets. Section 7 offers an information-theoretic characterization in terms of ϕ-divergence, following [7] and [10].

2. Quasisymmetry Models

We consider models for square contingency tables of format $I \times I$. Probability tables $\mathbf{p} = (p_{ij})$ are points in the simplex Δ_{I^2-1}. Here p_{ij} is the probability that an observation falls in the (i, j) cell. We write $\mathbf{n} = (n_{ij})$ for the table of observed frequencies. The model of symmetry (S) is

$$p_{ij} = s_{ij} \quad \text{with parameters} \quad s_{ij} = s_{ji} \quad \text{for } 1 \le i \le j \le I. \tag{1}$$

Here, and in what follows, the table (s_{ij}) is non-negative and its entries sum to 1. Geometrically, the S model is a simplex of dimension $\binom{I+1}{2} - 1$ inside the ambient probability simplex Δ_{I^2-1}. The classical *QS model* can be defined, as a model of divergence from S, by

$$p_{ij} = s_{ij} \frac{2c_i}{c_i + c_j} , \quad i, j = 1, \ldots, I, \tag{2}$$

where $c_i > 0$, $i = 1, \ldots, I$. The *Pearsonian QS model* is defined by the parametrization

$$p_{ij} = s_{ij}(1 + a_i - a_j) , \quad i, j = 1, \ldots, I, \tag{3}$$

with parameters a_i being constrained by $|a_i - a_j| \le 1$ for $1 \le i < j \le n$. Both models are semialgebraic subsets of dimension $\binom{I+1}{2} + I - 2$ in the simplex Δ_{I^2-1}. The parameters c_i and a_i, in (2) and (3) respectively, are expressing the departure from symmetry due to category i. Their role and nature will be clarified later, after (9). The S model is the subset obtained respectively for $c_1 = \cdots = c_I$ in (2) or $a_1 = \cdots = a_I$ in (3).

We here study the following quasisymmetry model (QS_t), where $t \in [0, 1]$ is a parameter:

$$p_{ij} = s_{ij} \left(1 + \frac{(1+t)(a_i - a_j)}{2 + (1-t)(a_i + a_j)} \right) , \quad i \ne j, \quad i, j = 1, \ldots, I. \tag{4}$$

In all three models, the matrix entries on the diagonal are set to $p_{ii} = s_{ii}$ for $i = 1, \ldots, I$. For $t = 1$, the model (4) specializes to the Pearsonian QS model (3). For $t = 0$, it specializes to the QS model (2), if we set $a_i = c_i - 1$. The parameters a_i will be assumed to satisfy the restriction

$$t \cdot \max_i a_i - \min_i a_i \leq 1. \tag{5}$$

Since we had assumed $0 \leq s_{ij} \leq 1/2$, the constraint (5) on the a_i ensures that the p_{ij} are probabilities (i.e. lie in the interval $[0, 1]$). Furthermore, if we change the parameters via

$$s_{ii} = x_{ii} \quad \text{for } i = j, \quad \text{and} \quad s_{ij} = x_{ij}\left(1 + (1-t)\frac{a_i + a_j}{2}\right) \quad \text{for } i \neq j,$$

then the model (QS_t), defined in (4), is rewritten in the simpler form

$$p_{ij} = x_{ij}(1 + a_i - t a_j)\ , \quad i \neq j, \quad i, j = 1, \ldots, I. \tag{6}$$

Note that $x_{i+} = \sum_{j=1}^I x_{ij} = \sum_{j=1}^I x_{ji} = x_{+i}$, since the table (x_{ij}) is symmetric. For $t = 1$, the probabilities defined by (6) satisfy $\sum_{i,j} p_{ij} = 1$. For $t \neq 1$, the 'weighted sum to zero' constraint

$$\sum_{i=1}^{I} (x_{i+} - x_{ii}) a_i = 0 \tag{7}$$

is required in order to ensure that the cell probabilities in (6) satisfy $\sum_{i,j} p_{ij} = 1$.

The expressions (4) and (6) are equivalent. Whether one or the other is preferred is a matter of convenience. Maximum likelihood estimation is easier with (4), since the MLEs of the s_{ij} are rational functions of the observed frequencies n_{ij}. The estimates of the a_i depend algebraically on $\mathbf{n}$, and they generally have to be computed by an iterative method. In the formulation (6), none of the parameters have estimates that are rational in $\mathbf{n}$. We shall see this in Section 4. On the other hand, for our algebraic analysis of the QS_t model, it is more convenient to use (6).

Example 1. Fix $I = 3$. For any fixed t, the model (6) is a hypersurface in the simplex Δ_8 of all 3×3 probability tables. This hypersurface is the zero set of the cubic polynomial

$$\begin{array}{c}(1 + t + t^2)(p_{12}p_{23}p_{31} - p_{21}p_{32}p_{13}) + \\ t(p_{12}p_{23}p_{13} + p_{12}p_{32}p_{31} + p_{21}p_{23}p_{31} - p_{12}p_{32}p_{13} - p_{21}p_{23}p_{13} - p_{21}p_{32}p_{31}).\end{array} \tag{8}$$

For $t = 0$, we recover the familiar binomial relation that encodes the cycle of length three [5, §6.2]. Thus, our family of QS_t models represents a deformation of that Markov basis:

$$p_{12}p_{23}p_{31} - p_{21}p_{32}p_{13} + O(t).$$

The generalization of the relation (8) to higher values of I will be presented in Section 3. $\diamondsuit$

Another characteristic model for square tables with commensurable classification variables is the model of *marginal homogeneity* (MH). This is specified by the equations

$$p_{i+} = p_{+i} \quad \text{for } \ i = 1, \ldots, I. \tag{9}$$

The model of symmetry S implies MH and QS, *i.e.* (2) with $c_1 = \cdots = c_I$. By [2, §8.2.3], if the models MH and QS hold simultaneously, then S is implied. In symbols, $\mathrm{S} = \mathrm{MH} \cap \mathrm{QS}$. This identity is important in that it underlines the role of the parameters c_i in the QS model. These express the contribution of the classification category i to marginal inhomogeneity. We shall prove next that the same identity holds for our generalized QS_t model.

Proposition 1. *For any $t \in [0, 1]$, we have $S = MH \cap QS_t$.*

Proof. It is straightforward to verify that S implies MH and QS_t with $a_i = 0$, for all i, which leads to $p_{ij} = x_{ij} = s_{ij}$, for all i, j. On the other hand, under QS_t as defined by (6), we have

$$p_{i+} - p_{+i} \;=\; (1+t)\left(a_i(x_{i+} - x_{ii}) - \sum_{j\neq i} a_j x_{ij}\right) \qquad \text{for} \;\; i = 1, \ldots, I. \tag{10}$$

Combining this with MH as in (9), and setting $y_i := x_{ii} - x_{i+}$, the equation (10) implies

$$\sum_{j\neq i} a_j x_{ij} + a_i y_i \;=\; 0 \qquad \text{for} \;\; i = 1, \ldots, I. \tag{11}$$

This can be written in the matrix form $\mathbf{Ba} = \mathbf{0}$, where $\mathbf{a} = (a_1, \ldots, a_I)^T$, $\mathbf{x} = (x_{ij})$, and

$$\mathbf{B} \;= \mathbf{x} - \mathrm{diag}(\mathbf{x1}) \;= \begin{bmatrix} & & & x_{1I} \\ & \tilde{\mathbf{B}} & & \vdots \\ & & & x_{I-1,I} \\ x_{I1} & x_{I2} & \ldots & y_I \end{bmatrix}.$$

The matrix $\tilde{\mathbf{B}}$ is strictly diagonally dominant, provided $|y_i| = x_{i+} - x_{ii} > \sum_{j\neq i}^{I-1} x_{ij}$. This is ensured if all x_{il} are positive, as in Remark 1; otherwise a separate argument is needed.

By the Levy-Desplanques Theorem, the matrix $\tilde{\mathbf{B}}$ is invertible and $\mathrm{rank}(\tilde{\mathbf{B}}) = I - 1$. Hence $\mathrm{rank}(\mathbf{B}) = I - 1$, since $\mathbf{B1} = 0$. Therefore, all solutions of $\mathbf{Ba} = \mathbf{0}$ have the form $\mathbf{a} = a\mathbf{1}$ for some $a \in \mathbb{R}$. For $t = 1$, equation (6) now implies $p_{ij} = x_{ij} = s_{ij}$, for all i, j. For $t \neq 1$, combining (7) with the positivity of $x_{i+} - x_{ii}$, we get $a = 0$. Hence symmetry S holds and the proof is complete.

Remark 1. Contingency tables with structural zeros, i.e., cells of zero probability, are rare. If they exist, they usually have a specific pattern (zero diagonal, triangular table). In our set-up it is realistic to assume that there exists an index j such that $p_{ij} > 0$ for all $i = 1, \ldots, I$. Thus, without loss of generality, we can assume that $p_{iI} > 0$ and therefore $x_{iI} > 0$ for all $i = 1, \ldots, I$.

Example 2. ($I = 3$) Marginal homogeneity defines a linear space of codimension 2, via

$$\begin{aligned} p_{11} + p_{12} + p_{13} &= p_{11} + p_{21} + p_{31}, \\ p_{21} + p_{22} + p_{23} &= p_{12} + p_{22} + p_{32}, \\ p_{31} + p_{32} + p_{33} &= p_{13} + p_{23} + p_{33}. \end{aligned}$$

Inside that linear subspace, the cubic (8) factors into a hyperplane, which is the S model $\{p_{12} = p_{21},\ p_{13} = p_{31},\ p_{23} = p_{32}\}$, and a quadric having no points with positive coordinates. $\diamondsuit$

In the light of Proposition 1, the parameter a_i of the QS_t model can be interpreted as the contribution of each category i to the *marginal inhomogeneity*. By this we mean the difference of a_i minus the weighted average of all a_i's. This is the parenthesized expression in the identity

$$p_{i+} - p_{+i} \;=\; (1+t)x_{i+}\left(a_i - \sum_j \frac{x_{ij}}{x_{i+}} a_j\right), \qquad i, j = 1, \ldots, I. \tag{12}$$

3. Implicit Equations

We now examine the quasisymmetry models QS_t through the lens of algebraic statistics [5, 9, 11]. To achieve more generality and flexibility, we fix an undirected simple graph G with vertex set $\{1, 2, \ldots, I\}$. Let $\mathcal{I}_G$ denote the prime ideal of algebraic relations among the

quantities $p_{ij} = x_{ij}(1+a_i-ta_j)$ in (6), where $\{i,j\}$ runs over the edge set $E(G)$ of the graph G. The ideal $\mathcal{I}_G$ lives in the polynomial ring $\mathbb{K}[\,p_{ij}, p_{ji} : \{i,j\} \in E(G)\,]$. Here we take $\mathbb{K} = \mathbb{Q}[[t]]$ to be the local ring of formal Laurent series in one unknown t.

Our main result in this section is the derivation of generators for the ideal $\mathcal{I}_G$. One motivation for studying $\mathcal{I}_G$ is the constrained formulation of the MLE problem in Section 4. The model in Section 2 corresponds to the complete graph on I nodes, denoted $G = K_I$. In particular, for $I = 3$, the ideal $\mathcal{I}_{K_3}$ is the principal ideal generated by the cubic in (8). Here we work with arbitrary graphs G, not just K_I, so as to allow for sparseness in the models. We disregard the 'weighted sum to 0' constraint (7), as this does not affect the homogeneous relations in $\mathcal{I}_G$.

Let $\mathbb{E}(G)$ denote the set of oriented edges of G. For each edge $\{i,j\}$ in $E(G)$ there are two edges ij and ji in $\mathbb{E}(G)$. So we have $|\mathbb{E}(G)| = 2|E(G)|$. An *orientation* of G is the choice of a subset $\mathcal{O} \subset \mathbb{E}(G)$ such that, for each edge $\{i,j\}$ in $E(G)$, *either* ij or ji belongs to $\mathcal{O}$. An orientation of G is called *acyclic* if it contains no directed cycle.

Let C denote the undirected n-cycle, with $E(C) = \{\{1,2\},\{2,3\},\ldots,\{n,1\}\}$. Then C has 2^n orientations, shown in Figure 1 for $n = 3$. Precisely two of these orientations are *cyclic*. These two directed cycles are denoted by o_C and $\bar{o}_C$. Their edge sets are $\mathbb{E}(o_C) = \{12, 23, \ldots, n1\}$ and $\mathbb{E}(\bar{o}_C) = \{21, 32, \ldots, 1n\}$. Any orientation δ_C of C defines a monomial of degree n via

$$p^{\delta_C} \;=\; \prod_{ij \in \mathbb{E}(\delta_C)} p_{ij}.$$

We also define the integer $\mathrm{c}(\delta_C) = 2|\mathbb{E}(o_C) \cap \mathbb{E}(\delta_C)| - n$. Note that $\mathrm{c}(o_C) = n$ and $\mathrm{c}(\bar{o}_C) = -n$.

We associate with the n-cycle C the following polynomial of degree n with 2^n terms:

$$P^C \;=\; \sum_{\delta_C} \mathrm{coeff}(\delta_C) \cdot p^{\delta_C}. \tag{13}$$

The sum is over all orientations δ_C of C, and the coefficients are the scalars in $\mathbb{K}$ defined by

$$\mathrm{coeff}(\delta_C) \;=\; \begin{cases} \frac{\mathrm{c}(\delta_C)}{|\mathrm{c}(\delta_C)|} \cdot \left(t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}} + t^{r+2-\frac{|\mathrm{c}(\delta_C)|}{2}} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|}{2}-2}\right) & \text{if} \quad n = 2r, \\ \frac{\mathrm{c}(\delta_C)}{|\mathrm{c}(\delta_C)|} \cdot \left(t^{r-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + t^{r+1-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|-1}{2}-1}\right) & \text{if} \quad n = 2r-1. \end{cases}$$

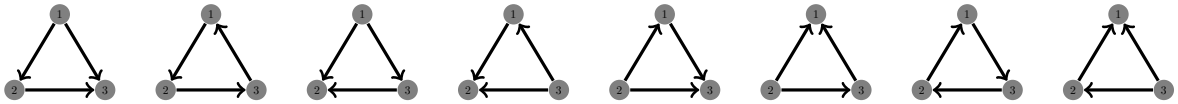

Figure 1: The eight orientations $\delta_1, \delta_2, \ldots, \delta_8$ of $C = K_3$.

Example 3. We consider the cycle $C = K_3$ of length $n = 3$. It has eight orientations, depicted in Figure 1. The corresponding monomials and their coefficients are as follows:

$$\begin{array}{lll} p^{\delta_1} = p_{12}p_{23}p_{13} & \mathrm{c}(\delta_1) = 1 & \mathrm{coeff}(\delta_1) = t \\ p^{\delta_2} = p_{12}p_{23}p_{31} & \mathrm{c}(\delta_2) = 3 & \mathrm{coeff}(\delta_2) = 1+t+t^2 \\ p^{\delta_3} = p_{12}p_{32}p_{13} & \mathrm{c}(\delta_3) = -1 & \mathrm{coeff}(\delta_3) = -t \\ p^{\delta_4} = p_{12}p_{32}p_{31} & \mathrm{c}(\delta_4) = 1 & \mathrm{coeff}(\delta_4) = t \\ p^{\delta_5} = p_{21}p_{23}p_{13} & \mathrm{c}(\delta_5) = -1 & \mathrm{coeff}(\delta_5) = -t \\ p^{\delta_6} = p_{21}p_{23}p_{31} & \mathrm{c}(\delta_6) = 1 & \mathrm{coeff}(\delta_6) = t \\ p^{\delta_7} = p_{21}p_{32}p_{13} & \mathrm{c}(\delta_7) = -3 & \mathrm{coeff}(\delta_7) = -1-t-t^2 \\ p^{\delta_8} = p_{21}p_{32}p_{31} & \mathrm{c}(\delta_8) = -1 & \mathrm{coeff}(\delta_8) = -t \end{array}$$

Thus, the polynomial P^C defined in (13) is the cubic (8) seen in Example 1. $\diamondsuit$

We define the *classical QS model* on the graph G by the parametrization (2) where $\{i,j\}$ runs over the set $E(G)$ of edges of G. We write $\mathcal{T}_G$ for the ideal of this model. This is a toric ideal whose Markov basis is obtained from the cycle polynomials P^C by setting $t=0$:

Lemma 1. *The ideal $\mathcal{T}_G$ has a universal Gröbner basis consisting of the binomials*

$$P^C|_{t=0} \;=\; p^{o(C)} - p^{\bar{o}(C)} \qquad \text{for all cycles } C \text{ in } G. \tag{14}$$

Proof. The identity in (14) is straightforward from the definition of $c(\delta_C)$ and $\mathrm{coeff}(\delta_C)$. It was shown in [5, §6.2] that the binomials $p^{o(C)} - p^{\bar{o}(C)}$ form a Markov basis for QS. Since the underlying model matrix is totally unimodular, the Markov basis is also a Graver basis, and hence it is a universal Gröbner basis, by [13, Propositions 4.11 and 8.11].

Example 4. For $I=4$, the model QS_t corresponds to the complete graph K_4. This graph has seven undirected cycles C, four of length 3 and three of length 4. Its defining prime ideal $\mathcal{I}_{K_4}$ is generated by four cubics and three quartics, all of the form P^C. For $t=0$, we recover the binomials corresponding to the seven moves that are listed in [11, §5.4, page 395]. ♢

This example is explained by the following theorem, which is our main result in Section 3.

Theorem 1. *The prime ideal $\mathcal{I}_G$ of the quasisymmetry model associated with an undirected graph G is generated by the cycle polynomials P^C where C runs over all cycles in G.*

Proof. We begin by proving that P^C lies in $\mathcal{I}_G$. The image of P^C under the substitution $p_{ij} \mapsto x_{ij}(1+a_i-ta_j)$ can be written as $Q^C \times \prod_{\{i,j\}\in E(C)} x_{ij}$, where Q^C is a polynomial in $\mathbb{K}[a_1,\ldots,a_n]$. Since each term p^{δ_C} of P^C is divisible by either p_{1n} or p_{n1}, we can write

$$Q^C \;=\; (1+a_1-ta_n)T_{1n} + (1+a_n-ta_1)T_{n1}. \tag{15}$$

We need to show that Q^C is zero. To do this, we shall establish the following identities:

$$\begin{aligned} T_{1n} &= (-1)^{[\frac{n-1}{2}]+1}(t+1)^{2r-2}(1+a_n-ta_1)\textstyle\prod_{i=2}^{n-1}(1+a_i-ta_i) \\ \text{and} \qquad T_{n1} &= (-1)^{[\frac{n-1}{2}]}(t+1)^{2r-2}(1+a_1-ta_n)\textstyle\prod_{i=2}^{n-1}(1+a_i-ta_i). \end{aligned}$$

To prove these, we shall use the decompositions

$$\begin{aligned} T_{1n} &= (1+a_1-ta_2)T_{1n,12} + (1+a_2-ta_1)T_{1n,21} \\ \text{and} \qquad T_{n1} &= (1+a_1-ta_2)T_{n1,12} + (1+a_2-ta_1)T_{n1,21}. \end{aligned}$$

With this notation, we claim that the following holds for a suitable integer r:

(i) $T_{1n,12} \;=\; (-1)^{[\frac{n-2}{2}]}t(t+1)^{2r-3}(a_2-a_n)\prod_{i=3}^{n-1}(1+a_i-ta_i)$,

(ii) $T_{1n,21} \;=\; (-1)^{[\frac{n-2}{2}]}(t+1)^{2r-3}(t^2a_2-t-a_n-1)\prod_{i=3}^{n-1}(1+a_i-ta_i)$.

Let C' be the cycle $2-3-\cdots-n-2$. In analogy to (15), we write

$$Q^{C'} = (1+a_2-ta_n)S_{2n} + (1+a_n-ta_2)S_{n2}.$$

Note that for any orientation δ_C of C in which $1n$ and 12 belong to $\mathbb{E}(\delta_C)$, we have

$$c(\delta_C) \;=\; \begin{cases} c(\delta_{C'})-1 & \text{if} \quad n2 \in \mathbb{E}(\delta_{C'}), \\ c(\delta_{C'})+1 & \text{if} \quad 2n \in \mathbb{E}(\delta_{C'}). \end{cases}$$

Also note that $\frac{c(\delta_C)}{|c(\delta_C)|} = \frac{c(\delta_{C'})}{|c(\delta_{C'})|}$. In order to prove (i) we consider the following two cases:

Case 1. $n = 2r-1$ is an odd number: We claim that $T_{1n,12} = t(S_{n2} + S_{2n})$. Note that C' is an even cycle with $n-1 = 2(r-1)$. The coefficient for δ_C can be written as

$$t \times \frac{\mathrm{c}(\delta_C)}{|\mathrm{c}(\delta_C)|}\big((t^{r-1-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + t^{r+1-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|-1}{2}-2}) \quad + \\ (t^{r-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + t^{r+2-\frac{|\mathrm{c}(\delta_C)|-1}{2}} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|-1}{2}-3})\big).$$

The first summand corresponds to the orientation $\delta_{C'}$ with $n2 \in \mathbb{E}(\delta_{C'})$. The second summand corresponds to the orientation $\delta_{C'}$ with $2n \in \mathbb{E}(\delta_{C'})$. By induction on n, we have

$$\begin{aligned} & S_{2n} &=& \ (-1)^{[\frac{n-2}{2}]+1}(t+1)^{2r-4}(1+a_n-ta_2)\textstyle\prod_{i=3}^{n-1}(1+a_i-ta_i), \\ \text{and} \quad & S_{n2} &=& \ (-1)^{[\frac{n-2}{2}]}(t+1)^{2r-4}(1+a_2-ta_n)\textstyle\prod_{i=3}^{n-1}(1+a_i-ta_i). \end{aligned}$$

Since $-(1+a_n-ta_2)+(1+a_2-ta_n) = (1+t)(a_2-a_n)$, the claim (i) holds for n odd.

Case 2. $n = 2r$ is an even number: We will first show that $T_{1n,12} = t(S_{n2}+S_{2n})/(1+t)^2$. Here C' is an odd cycle on $n-1 = 2r-1$ vertices. The coefficient for δ_C equals

$$\frac{t}{(1+t)^2} \times \frac{\mathrm{c}(\delta_C)}{|\mathrm{c}(\delta_C)|}\big(t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}-1} + 2t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}} + \cdots + 2t^{r+\frac{|\mathrm{c}(\delta_C)|}{2}-2} + t^{r+\frac{|\mathrm{c}(\delta_C)|}{2}-1}\big).$$

This sum can be decomposed as

$$(t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}-1} + t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|}{2}-1}) + (t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}} + t^{r-\frac{|\mathrm{c}(\delta_C)|}{2}+1} + \cdots + t^{r+\frac{|\mathrm{c}(\delta_C)|}{2}-2}),$$

where the first summand corresponds to the orientation $\delta_{C'}$ with $n2 \in \mathbb{E}(\delta_{C'})$, and the second summand corresponds to the orientation $\delta_{C'}$ with $2n \in \mathbb{E}(\delta_{C'})$. Therefore $T_{1n,12} = \frac{t(S_{n2}+S_{2n})}{(1+t)^2}$. By induction on n, we have

$$\begin{aligned} & S_{2n} &=& \ (-1)^{[\frac{n-2}{2}]+1}(t+1)^{2r-2}(1+a_n-ta_2)\textstyle\prod_{i=3}^{n-1}(1+a_i-ta_i) \\ \text{and} \quad & S_{n2} &=& \ (-1)^{[\frac{n-2}{2}]}(t+1)^{2r-2}(1+a_2-ta_n)\textstyle\prod_{i=3}^{n-1}(1+a_i-ta_i) \end{aligned}$$

Since $-(1+a_n-ta_2)+(1+a_2-ta_n) = (1+t)(a_2-a_n)$, the result holds for even n as well.

By a similar argument one can prove (ii). Now applying (i) and (ii) and the equality

$$-(1+a_2-ta_2)(1+a_n-ta_1)(1+t) = (1+a_1-ta_2)(a_2-a_n)t + (1+a_2-ta_1)(t^2a_2-t-a_n-1),$$

we obtain

$$T_{1n} \ = \ (-1)^{[\frac{n-2}{2}]+1}(t+1)^{2r-2}(1+a_n-ta_1)\prod_{i=2}^{n-1}(1+a_i-ta_i)\ .$$

The identity for T_{n1} is analogous. It follows that $P^C \in \mathcal{I}_G$ for all cycles of G.

It remains to be shown that the P^C generate the homogeneous ideal $\mathcal{I}_G$. Recall that, by Lemma 1, the images of the P^C generate this ideal after we tensor, over the local ring $\mathbb{K}$, with the residue field $\mathbb{Q} = \mathbb{K}/\langle t\rangle$. Hence, by Nakayama's Lemma, the P^C generate $\mathcal{I}_G$.

Remark 2. In Theorem 1 we can replace the local ring $\mathbb{K} = \mathbb{Q}[[t]]$ with the polynomial ring $\mathbb{Q}[t]$ because no t appears in the leading forms $(P^C)|_{t=0}$. This ensures that $\mathbb{Q}[t][p_{ij}]$ modulo the ideal $\langle P^C : C \,\text{cycle in}\, G\rangle$ is torsion-free, hence free, and therefore flat over $\mathbb{Q}[t]$.

In statistical applications, the quantity t will always take on a particular real value. In the remainder of this paper, we assume $t \in \mathbb{R}$, and we identify $\mathcal{I}_G$ with its image in $\mathbb{R}[p_{ij}]$.

Corollary 1. *For any $t \in \mathbb{R}$, the cycle polynomials P^C generate the ideal $\mathcal{I}_G$ in $\mathbb{R}[p_{ij}]$.*

Theorem 1 furnishes a (flat) degeneration from $\mathcal{I}_G$ to the toric ideal $\mathcal{T}_G$. Geometrically, we view this as a degeneration of varieties (or semialgebraic sets) from $t > 0$ to $t = 0$. Lemma 1 concerns further degenerations from the toric ideal $\mathcal{T}_G$ to its initial monomial ideals $\mathcal{M}_G$. Any such $\mathcal{M}_G$ is squarefree and serves as a combinatorial model for both $\mathcal{T}_G$ and $\mathcal{I}_G$.

We describe one particular choice and draw some combinatorial conclusions. Fix a term order on $\mathbb{R}[p_{ij}]$ with the property that $p_{ij} \succ p_{k\ell}$ whenever $i < k$, or $i = k$ and $j < \ell$. For any cycle C, we label the two directed orientations o_C and $\bar{o}_C$ so that $p^{o(C)} \succ p^{\bar{o}(C)}$. Fix a spanning tree T of G. Let $\mathfrak{P}_T$ denote the monomial prime ideal generated by all unknowns p_{ij} where $\{i,j\} \in E(G)\backslash E(T)$ and p_{ij} divides p^{o_C}, where C is the unique cycle in $E(T) \cup \{\{i,j\}\}$. The squarefree monomial ideal

$$\mathcal{M}_G \;=\; \mathrm{in}_{\succ}(\mathcal{T}_G) \;=\; \bigl\langle\, p^{o_C} \,:\, C \text{ cycle in } G \,\bigr\rangle \;=\; \bigcap_T \mathfrak{P}_T \,, \tag{16}$$

is obtained by taking the intersection over all spanning trees T of G. The simplicial complex with *Stanley-Reisner ideal* $\mathcal{M}_G$ is a regular triangulation of the *Lawrence polytope* of the graph G. This triangulation is shellable and hence our ideals are Cohen-Macaulay. We record the following fact.

Proposition 2. *The ideals $\mathcal{M}_G$, $\mathcal{T}_G$ and $\mathcal{I}_G$ define varieties of dimension $|E(G)| + I - 1$ in affine space, and their common degree is the number of spanning trees of the graph G.*

Proof. Each of the components $\mathfrak{P}_T$ in (16) has codimension $|E(G)\backslash E(T)| = |E(G)|-I+1$.

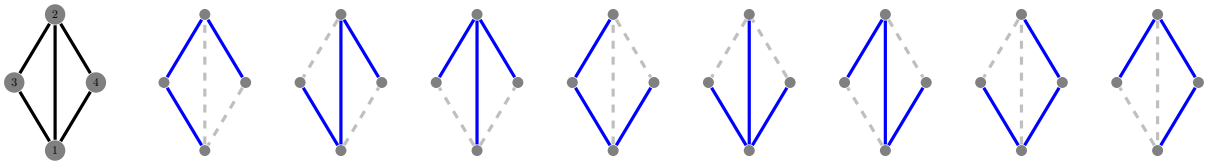

Figure 2: A graph G on $I = 4$ nodes and its eight spanning trees T

Example 5. Consider the graph G depicted in Figure 2. The associated toric ideal equals

$$\mathcal{T}_G \;=\; \langle\, \underline{p_{12}p_{23}p_{31}} - p_{21}p_{32}p_{13}\,,\; \underline{p_{12}p_{24}p_{41}} - p_{21}p_{42}p_{14}\,,\; \underline{p_{13}p_{32}p_{24}p_{41}} - p_{31}p_{23}p_{42}p_{14}\,\rangle.$$

This has codimension 2 and degree 8. Its (underlined) initial monomial ideal $\mathcal{M}_G$ equals

$$\langle p_{12}, p_{13}\rangle \cap \langle p_{12}, p_{32}\rangle \cap \langle p_{12}, p_{24}\rangle \cap \langle p_{12}, p_{41}\rangle \cap \langle p_{23}, p_{41}\rangle \cap \langle p_{23}, p_{24}\rangle \cap \langle p_{24}, p_{31}\rangle \cap \langle p_{31}, p_{41}\rangle.$$

These eight monomial prime ideals correspond to the spanning trees in Figure 2. The ideal $\mathcal{I}_G$ has three generators, two cubics with 8 terms and one quartic with 16 terms, as in (13). These are obtained from the Markov basis of $\mathcal{T}_G$ by adding terms that are divisible by t. ♢

4. Maximum Likelihood Estimation

A data table $\mathbf{n} = (n_{ij})$ of format $I \times I$ can arise either by multinomial sampling or by sampling from I^2 independent Poisson distributions, one for each of its cells. In both cases, the log-likelihood function, up to an additive constant, is equal to

$$\ell_{\mathbf{n}}(\mathbf{p}) \;\;=\;\; \sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij} \cdot \log(p_{ij}). \tag{17}$$

Maximum likelihood estimation (MLE) is the problem of maximizing $\ell_{\mathbf{n}}$ over all probability tables $\mathbf{p} = (p_{ij})$ in the model of interest. For us, that model is the quasisymmetry model (QS_t), where t is a fixed constant in the interval $[0, 1]$. This optimization problem can be expressed in either constrained form or in unconstrained form. The *constrained MLE problem* is written as

$$\text{Maximize } \ell_{\mathbf{n}}(\mathbf{p}) \quad \text{subject to} \quad \mathbf{p} \in V(\mathcal{I}_G) \cap \Delta_{I^2-1}, \tag{18}$$

where $G = K_I$ is the complete graph on I nodes, and $V(\mathcal{I}_G)$ is the zero set of the cycle polynomials P^C constructed in Section 3. The *unconstrained MLE problem* is written as

$$\text{Maximize } \ell_{\mathbf{n}}(\mathbf{a}, \mathbf{s}). \tag{19}$$

The decision variables in (19) are the vector $\mathbf{a} = (a_1, \ldots, a_I)$ and the symmetric probability matrix $\mathbf{s} = (s_{ij})$. The objective function in (19) is obtained by substituting (4) into (17). We shall discuss both formulations, starting with a simple numerical example for the formulation (18).

Example 6. Let $I = 3$, $t = 2/3$ and consider the data table

$$\mathbf{n} = \begin{bmatrix} 2 & 3 & 5 \\ 11 & 13 & 17 \\ 19 & 23 & 29 \end{bmatrix} \qquad \text{with sample size } n_{++} = 122.$$

Our aim is to maximize $\ell_{\mathbf{n}}(\mathbf{p})$ subject to the cubic equation (8) and $p_{11} + p_{12} + \cdots + p_{33} = 1$. Using Lagrange multipliers for these two constraints, we derive the *likelihood equations* by way of [5, Algorithm 2.29]. These polynomial equations in the nine unknowns p_{ij} have 15 complex solutions. Two of the complex solutions are non-real. Of the 13 real solutions, 12 have at least one negative coordinate. Only one solution lies in the probability simplex Δ_8:

$$\begin{array}{lllllllll} \hat{p}_{11} & = & 1/61, & \hat{p}_{12} & = & 0.0286294, & \hat{p}_{13} & = & 0.0376289, \\ \hat{p}_{21} & = & 0.0861247, & \hat{p}_{22} & = & 13/122, & \hat{p}_{23} & = & 0.1446119, \\ \hat{p}_{31} & = & 0.1590924, & \hat{p}_{32} & = & 0.1832569, & \hat{p}_{33} & = & 29/122. \end{array} \tag{20}$$

This is the global maximum of the constrained MLE problem for this instance. ♢

The benefit of the constrained formulation is that we can take advantage of the combinatorial results in Section 3, and we do not have to deal with issues of identifiability and singularities arising from the map (4). On the other hand, most statisticians would prefer the unconstrained formulation because this corresponds more directly to the fitting of model parameters to data.

To solve the unconstrained MLE problem (19), we take the partial derivations of the objective function $\ell_n(\mathbf{a}, \mathbf{s})$ with respect to all model parameters a_i and s_{ij}. The resulting system of equations decouples into a system for $\mathbf{a}$ and a system for $\mathbf{s}$. The latter is trivial to solve. Using the requirement that the entries of $\mathbf{s}$ sum to 1, it has the closed form solution

$$\hat{s}_{ij} \;=\; \frac{n_{ij} + n_{ji}}{2n_{++}}, \quad i, j = 1, \ldots, I. \tag{21}$$

After dividing by $1 + t$, the partial derivatives of $\ell_{\mathbf{n}}(\mathbf{a}, \mathbf{s})$ with respect to $a_1, a_2, \ldots, a_I$ are

$$\sum_{\substack{j=1 \\ j \neq i}}^{I} \frac{(1 + a_j - ta_j)[n_{ij}(1 + a_j - ta_i) - n_{ji}(1 + a_i - ta_j)]}{(1 + a_i - ta_j)(1 + a_j - ta_i)[2 + (1 - t)(a_i + a_j)]} \qquad \text{for } i = 1, 2, \ldots, I. \tag{22}$$

This system of equations has infinitely many solutions, because the model QS_t is not identifiable. The general fiber of the map (4) is a line in $\mathbf{a}$-space. Hence only $I - 1$ of the I parameters a_i can be estimated. One way to fix this is to simply add the constraint $\hat{a}_I = 0$.

Example 7. Let us return to the numerical instance in Example 6. Here we have

$$\hat{s}_{11} = 1/61,\ \hat{s}_{12} = 7/122,\ \hat{s}_{13} = 6/61,\ \hat{s}_{22} = 13/122,\ \hat{s}_{23} = 10/61,\ \hat{s}_{33} = 29/122. \tag{23}$$

The equations (22) can be solved in a computer algebra system by clearing denominators and then saturating the ideal of numerators with respect to those denominators. As before, there are precisely 15 complex solutions, of which 13 are real. The MLE is given by

$$\hat{a}_1 = -0.65948848999731861332,\ \hat{a}_2 = -0.13818331109451658084,\ \hat{a}_3 = 0. \tag{24}$$

These are floating point approximations to algebraic numbers of degree 15 over $\mathbb{Q}$. An exact representation is given by their minimal polynomials. For the first coordinate, this is

$$\begin{array}{c} 62031304a_1^{15} + 2201861910a_1^{14} + 30829909776a_1^{13} + 206135547000a_1^{12} + 528436383696a_1^{11} \\ -1126661553720a_1^{10} - 9740892273264a_1^9 - 4305524252579a_1^8 + 26533957305582a_1^7 \\ +88281552626154a_1^6 + 44254830057030a_1^5 - 76332701171853a_1^4 - 83490498412056a_1^3 \\ +1857597611688a_1^2 + 29825005557312a_1 + 9354112703280 \quad = \ 0. \end{array}$$

With this, the second coordinate $\hat{a}_2$ is a certain rational expression in $\mathbb{Q}(\hat{a}_1)$. By plugging (23) and (24) into (4) with $t = 2/3$, we recover the estimated probability table in (20). ♢

For larger cases, solutions to the likelihood equations (22) are computed by iterative numerical methods, such as the *unidimensional Newton's method*. The updating equations at the q-th step of this iterative method are

$$a_i^{(q)} \ = \ a_i^{(q-1)} - \frac{\partial \ell_{\mathbf{n}}(\mathbf{a})/\partial a_i}{\partial^2 \ell_{\mathbf{n}}(\mathbf{a})/\partial a_i^2}\Big|_{\mathbf{a}=\mathbf{a}^{(q-1)}} \quad \text{for} \quad i = 1, \ldots, I-1,\ q = 1, 2, \ldots\ . \tag{25}$$

We find it convenient to rewrite the first derivatives (22) as

$$\frac{\partial \ell_{\mathbf{n}}(\mathbf{a})}{\partial a_i} = (1+t)\sum_{j=1}^{I} \frac{s_{ij}}{2+(1-t)(a_i+a_j)}\left(1 - \frac{1-t}{1+t}c_{ij}\right)\left(\frac{n_{ij}}{p_{ij}} - \frac{n_{ji}}{p_{ji}}\right). \tag{26}$$

The second derivative equals

$$\begin{aligned} \frac{\partial^2 \ell_{\mathbf{n}}(\mathbf{a})}{\partial a_i^2} \ = \ & -(1+t)\sum_{j=1}^{I} \frac{2(1-t)s_{ij}}{[2+(1-t)(a_i+a_j)]^2}\left(1 - \frac{1-t}{1+t}c_{ij}\right)\left(\frac{n_{ij}}{p_{ij}} - \frac{n_{ji}}{p_{ji}}\right) \\ & -(1+t)\sum_{j\neq i} \frac{(1+t)s_{ij}^2}{[2+(1-t)(a_i+a_j)]^2}\left(1 - \frac{1-t}{1+t}c_{ij}\right)^2\left(\frac{n_{ij}}{p_{ij}^2} + \frac{n_{ji}}{p_{ji}^2}\right). \end{aligned} \tag{27}$$

Here $i = 1, \ldots, I-1$, the p_{ij} are the expressions in (4), and

$$c_{ij} = \frac{(1+t)(a_i - a_j)}{2+(1-t)(a_i+a_j)}\ .$$

We believe that the numerical solution found by this iteration is always the global maximum in (19). This would be implied by the following conjecture, which holds for $t = 0$ and $t = 1$.

Conjecture 2. *The Hessian* $\mathbf{H}(\mathbf{a}) = \left(\frac{\partial^2 \ell_{\mathbf{n}}(\mathbf{a})}{\partial a_i \partial a_j}\right)$ *is negative definite for all* $\mathbf{a} \in \mathbb{R}^I$ *with* (5).

We verified this conjecture for many examples with $t \in (0, 1)$. In each case, we also ran our iterative algorithm for many starting values, and it always converged to the same solution.

The diagonal entries of the Hessian matrix are given in (27), while the non-diagonal are

$$\frac{\partial^2 \ell_{\mathbf{n}}(\mathbf{a})}{\partial a_i \partial a_j} \ = \ \frac{2(1-t)^2 s_{ij} c_{ij}}{[2+(1-t)(a_i+a_j)]^2}\left(\frac{n_{ij}}{p_{ij}} - \frac{n_{ji}}{p_{ji}}\right) \tag{28}$$

$$+ \frac{(1+t)^2 s_{ij}^2}{[2+(1-t)(a_i+a_j)]^2}\left[1-\left(\frac{1-t}{1+t}c_{ij}\right)^2\right]\left(\frac{n_{ij}}{p_{ij}^2}+\frac{n_{ji}}{p_{ji}^2}\right).$$

In the iterative algorithm described above, we had fixed the last parameter a_I at zero. This ensures identifiability, and it is done for simplicity. The constraint $a_I = 0$ defines a reference point for the other parameters $a_1, \ldots, a_{I-1}$. Under this constraint, (12) leads to

$$a_i = \frac{1}{1+t}\left(\frac{p_{i+}-p_{+i}}{x_{i+}} - \frac{p_{I+}-p_{+I}}{x_{I+}}\right) \quad \text{for} \quad i = 1, \ldots, I-1.$$

This means that the contribution of category i to marginal inhomogeneity is compared to the last category's contribution. Hence, in view of (12), a reasonable alternative constraint could be $\sum_{j=1}^{I} \frac{x_{ij}}{x_{i+}} a_j = 0$. This constraint calibrates each category's contribution to marginal inhomogeneity relative to the weighted average of all I categories.

Remark 3. The iterative procedure described above for fitting the $\mathrm{QS_t}$ models was implemented by us in `R`. The algorithm works regardless of whether we impose the restriction $a_I = 0$ or not. We noticed that when imposing this constraint, the algorithm requires more iterations to converge. The convergence is also affected by the initial values $\mathbf{a}^{(0)}$ we used. A classical choice would be $a_i = 0$ for all i, as this corresponds to complete symmetry. However, we observed that for $\mathbf{a}^{(0)}$ with coordinates $\frac{n_{i+}-n_{+i}}{n_{i+}+n_{+i}}$, $i = 1, \ldots, I$, the convergence is faster.

Remark 4. Here we consider the model parameter t as fixed. Alternatively, it could be estimated from the data, as for the power-divergence logistic regression model in [6].

5. Quasisymmetric Independence

A natural submodel of (1) is the symmetric independence model (SI), which is given by

$$p_{ij} = s_i s_j \ , \quad i, j = 1, \ldots, I. \tag{29}$$

The I parameters s_i are non-negative and sum to 1. The corresponding probability tables $\mathbf{p} = (p_{ij})$ are symmetric and have rank 1. The models of quasisymmetric independence (QSI_t) can be defined analogously to the QS_t models, by measuring departure from (29). Namely, replacing the symmetric probabilities s_{ij} in (4) by the factored form in (29), we get

$$p_{ij} = s_i s_j \left(1 + \frac{(1+t)(a_i - a_j)}{2+(1-t)(a_i+a_j)}\right) \ , \quad i \neq j, \quad i, j = 1, \ldots, I. \tag{30}$$

The MLEs of the parameters of the SI model in (29) are

$$\hat{s}_i = \frac{n_{i+}+n_{+i}}{2n} \quad \text{for} \quad i = 1, \ldots, I. \tag{31}$$

These are also the MLEs of the s_i parameters in the QSI_t model. The likelihood equations for $\mathbf{a}$ are as before, but with p_{ij}'s in (26) as defined in (29) and (30). Their numerical solution can be computed with the iterative procedure described in Section 4, adjusted accordingly.

Remark 5. In Proposition 1, if we replace the models S and QS_t by SI and QSI_t, then an analogous statement holds. Thus, we have $\mathrm{SI} = \mathrm{MH} \cap \mathrm{QSI}_t$ for each $t \in [0, 1]$.

Following the discussion in Section 3, it would be interesting to derive the implicit equations for the model QSI_t. At present, we have a complete solution only for the special case $t = 1$. The quasisymmetric independence model QSI_1 is defined by the parametrization

$$p_{ij} \ = \ s_i s_j \cdot (1 + a_i - a_j), \qquad 1 \leq i, j \leq I. \tag{32}$$

Alternatively, $\{i, j\}$ could range over the edges of a graph G, as in Section 3. In the following result, whose proof we omit, we restrict ourselves to the case of the complete graph K_I.

Proposition 3. *The prime ideal of the* QSI_1 *model in (32) is generated by the following homogeneous quadratic polynomials (for any choices of indices* i, j, k, ℓ *among* $1, \ldots, I$*):*

- $(p_{ij} + p_{ji})^2 - 4p_{ii}p_{jj}$,
- $p_{kk}(p_{ij} - p_{ji}) + p_{ki}p_{jk} - p_{ik}p_{kj}$,
- $(p_{ij} - p_{ji})(p_{jk} - p_{kj}) + 4(p_{jj}p_{ki} - p_{ji}p_{kj})$,
- $p_{\ell i}(p_{jk} - p_{kj}) + p_{\ell j}(p_{ki} - p_{ik}) + p_{\ell k}(p_{ij} - p_{ji})$,
- $p_{i\ell}(p_{jk} - p_{kj}) + p_{j\ell}(p_{ki} - p_{ik}) + p_{k\ell}(p_{ij} - p_{ji})$.

The general case where $t < 1$ differs from the $t = 1$ case in that the prime ideal of QSI_1 is no longer generated by quadrics. Even for $I = 3$, a minimal generator of degree 3 is needed:

Example 8. Fix $I = 3$. For general $t \in \mathbb{R}$, we consider the model (30) with $p_{ii} = s_i s_i$ for $i = 1, 2, 3$. Its ideal is minimally generated by 7 polynomials: 6 quadrics and one cubic. ♢

6. Fitting the Models to Data

We next illustrate the new models and their features on some characteristic data sets. The goodness-of-fit of a model is tested asymptotically by the likelihood ratio statistic. The associated degrees of freedom for QS_t and QSI_t are $df(\mathrm{QS}_t) = (I-1)(I-2)/2$ and $df(\mathrm{QSI}_t) = (I-1)^2$, respectively. As we shell see, the models in each family can perform either quite similar or differ significantly, depending on the specific data under consideration.

A case of similar behavior is the classical vision example of Table 1. The model of QS ($t = 0$) has been applied on this data often in the literature, while [7] applied Pearsonian QS. Both models provide a quite similar fit, namely ($G^2 = 7.27076$, p-value $= 0.06375$) for QS_0 and ($G^2 = 7.26199$, p-value $= 0.06340$) for QS_1. Here, $df = 3$.

The behavior of the QS_t models for $t \in (0, 1)$ is similar. The log-likelihood values vary from -16388.11444 ($t = 0$) to -16388.11006 ($t = 1$) while the saturated log-likelihood is -16384.47906 (see Figure 3, left). Table 2 gives the MLEs of the expected cell frequencies under the models QS_0, QS_1 and $QS_{2/3}$. For $t = 2/3$ we get $G^2 = 7.26234$, with p-value $= 0.06399$.

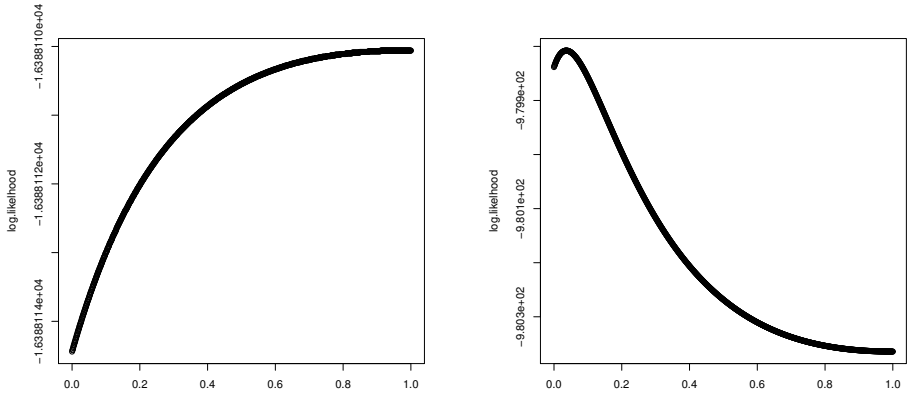

Figure 3: Log-likelihood values of QS_t for t in $[0, 1]$ for data in Tables 1 (left) and 3(c) (right).

Examples for which the members of the QS_t family are not of similar performance are the two 3×3 tables of [7, Tables 3 and 4], displayed in Table 3 (a) and (b). Here, the models QS_0 and QS_1 differ considerably in their fit. In particular, the data in Table 3 (a) are modeled well by QS_0 but not by QS_1 ($G_0^2 = 0.18572$ and $G_1^2 = 5.29006$), while the opposite holds for Table 3 (b), since $G_0^2 = 6.29035$ and $G_1^2 = 0.29215$.

In such situations, the question arises whether some t is appropriate for both data sets. Finding t such that QS_t works for two or more $I \times I$ tables of the same set-up is of special

	Left Eye Grade			
Right Eye Grade	best	second	third	worst
best	1520 –	266 (263.38[a]/ 263.38[b]/ 263.39[c])	124 (133.58/ 133.59/ 133.60)	66 (59.04/ 59.09/ 59.09)
second	234 (236.62/ 236.62/ 236.61)	1512 –	432 (418.99/ 418.90/ 418.90)	78 (88.39/ 88.40/ 88.40)
third	117 (107.42/ 107.40/ 107.40)	362 (375.01/ 375.10/ 375.10)	1772 –	205 (201.57/ 201.58/ 201.58)
worst	36 (42.96/ 42.91/ 42.91)	82 (71.61/ 71.60/ 71.60)	179 (182.43/ 182.42/ 182.42)	492 –

Table 2: Unaided distance vision of right and left eyes for 7477 women. Parenthesized values are ML estimates of the expected frequencies under models (*a*) QS_0, (b) $QS_{2/3}$, and (*c*) QS_1.

(a)

	1	**2**	**3**
1	28	10	15
1	122	126	102
1	49	22	26

(b)

	1	**2**	**3**
1	38	128	36
1	5	119	43
1	12	88	31

(c)

	1	**2**	**3**
1	28	12	25
1	122	126	102
1	49	22	26

Table 3: Simulated 3×3 examples of [7], generated by the models (a) QS_0 and (b) QS_1 (their Tables 3 and 4, respectively). A toy example in (c).

interest in the study of stratified tables. Using the same model on all strata makes parameter estimates among models comparable. This is a major advantage of the proposed family.

Models that lie 'in-between' the two extreme cases ($t = 0$ and $t = 1$) may lead to a consensus. Even if that consensus model does not perform as well as QS_0 and QS_1 on each table separately, it can provide a reasonable fit for both tables. To visualize this, Figure 4 (left) shows the p-values of the fit of the QS_t models with $t \in [0, 1]$, for Tables 3 (a) and (b), by solid and dashed curves, respectively, along with the significance level of $\alpha = 0.05$. The consensus model QS_t would have $t \in (0.061, 0.302)$. Among these models, we propose $\mathrm{QS}_{0.14}$, since the intersection of the two curves happens around $t = 0.137$. The fit of this model for Table 3 (a) is $G^2 = 2.27614$ (p-value=0.1314) while for (b) it is $G^2 = 2.16744$ (p-value=0.1409). The vector of MLEs for parameters a_i is $(-0.5458, 1.8555, 0)$ and $(2.1247, -0.5406, 0)$, respectively. We note that, in deriving the consensus model, the G^2 values could have been used as an alternative to the p-values in Figure 4.

In all examples treated so far, the log-likelihood under QS_t was monotone in t (see Figure 3, left, and Figure 5, upper), suggesting that the 'best' model will be achieved at either $t = 0$ or $t = 1$. This is not always the case. For example, for the data in Table 3 (c), the best fit occurs for $t = 0.036$ (see also Figure 3, right), giving $G^2 = 1.742943 \cdot 10^{-6}$ (p-value=0.9989) while for $t = 0$ and $t = 1$, it is $G^2 = 0.0610$ (p-value= 0.8049) and $G^2 = 1.1131$ (p-value=0.2914), respectively. Furthermore, even when the best model is for $t = 0$ or $t = 1$, we may still want to use some $t \in (0, 1)$, *e.g.* for stratified tables with different optimal model at each level of the stratifying variable, as explained above.

Applying the quasisymmetric independence models to Tables 3 (a) and (b), we observe that QSI_0 fits well on Table 3 (a) but not on (b), while model QSI_1 is of acceptable fit for both data sets. Indeed, we have $G^2_a(QSI_0) = 1.3600$ (p-value=0.8511), $G^2_b(QSI_0) = 11.8622$ (p-value=0.0184), $G^2_a(QSI_1) = 6.4643$ (p-value=0.1671) and $G^2_b(QSI_1) = 5.8640$ (p-value=0.2095). For the performance of the QSI_t model for $t \in [0, 1]$, see Figure 4 (right) and Figure 5 (lower). For $t = 0.532$, the p-value of the fit of the model is equal to 0.1983 for both data sets.

All examples of this section were worked out with `R` functions we developed for fitting the QS_t and QSI_t models via the unidimensional Newton's method. The adopted inferential approach is asymptotic. In cases of small sample size, exact inference can be carried out via an algebraic computations along the lines described in Section 3, and demonstrated in Examples 6 and 7.

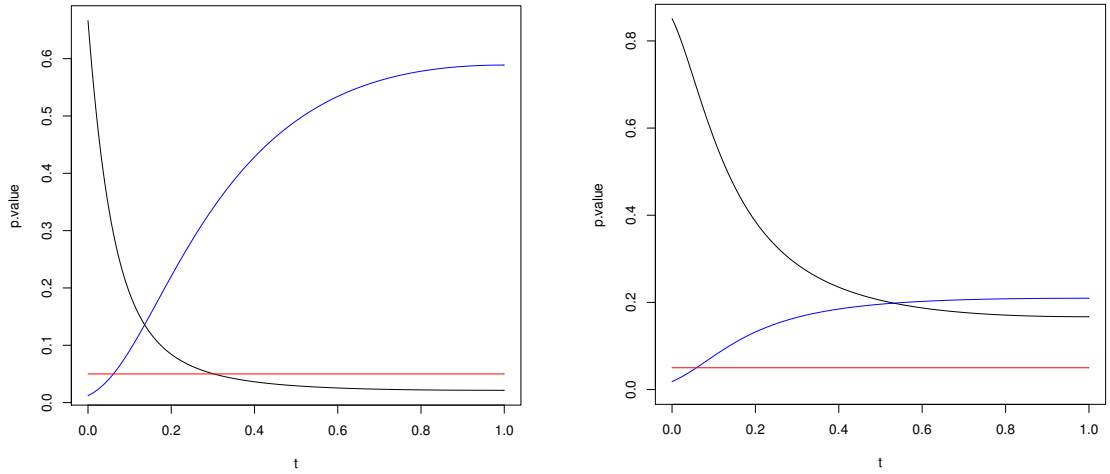

Figure 4: p-values for the G^2 goodness-of-fit test of QS_t (left) and QSI_t (right) for $t \in [0,1]$, along with the significance level $\alpha = 0.05$. Data are from Table 3: (a) black and (b) blue.

7. Divergence Measures

The one-parameter family of QS models we proposed, QS_t, $t \in [0,1]$, connects the classical QS model ($t = 0$) and the Pearsonian QS model ($t = 1$). These two belong both to a broader class of generalized QS models that are derived using the concept of *ϕ–divergence* [7, 10]. Measures of divergence quantify the distance between two probability distributions and play an important role in information theory and statistical inference. A well known divergence measure is the Kullback-Leibler (KL) divergence. However there exist broader classes of divergences. Such a class, including the KL as a special case, is the ϕ-divergence. In the framework of two-dimensional contingency tables, this class is defined as follows.

Let $\mathbf{p} = (p_{ij})$ and $\mathbf{q} = (q_{ij})$ be two discrete bivariate probability distributions. The *ϕ–divergence* between $\mathbf{p}$ and $\mathbf{q}$ (or Csiszar's measure of information in $\mathbf{q}$ about $\mathbf{p}$) is defined by

$$D_\phi(\mathbf{p},\mathbf{q}) \;=\; \sum_{i,j} q_{ij}\phi(p_{ij}/q_{ij}). \tag{33}$$

Here $\phi : [0,\infty) \to \mathbb{R}^+$ is a convex function such that $\phi(1) = \phi'(1) = 0$, $0 \cdot \phi(0/0) = 0$, and $0 \cdot \phi(x/0) = x \cdot \lim_{u\to\infty} \phi(u)/u$. For $\phi(u) = u\log(u) - u + 1$ and $\phi(u) = (u-1)^2/2$, the divergence (33) becomes the KL and the Pearson's divergence, respectively. We adopt the notation in [10]. For properties of ϕ-divergence, as well as a list of well-known divergences belonging to this family, we refer to [10, Section 1.2]. The differential geometric structure of the Riemannian metric induced by such a divergence function is studied by Amari and Cichock [1].

The generalized QS models introduced by Kateri and Papaioannou [7] are based on the ϕ-divergence and are characterized by the fact that each model in this class is the closest model to symmetry S, when the distance is measured by the corresponding divergence measure. The classical QS model corresponds to the KL divergence, while the Pearsonian QS corresponds to Pearson's distance. We shall prove in Theorem 3 that the other members of the QS_t family, *i.e.* for $t \in (0,1)$, are ϕ-divergence QS models as well, and we identify the corresponding ϕ function.

Theorem 3. *Fix $t \in (0,1)$ and consider the class of models that preserve the given row (or column) marginals p_{i+} (or p_{+i}) for $i = 1,\ldots,I$, and also preserve the given sums $p_{ij} + p_{ji} = 2s_{ij}$ for $i,j = 1,\ldots,I$. In this class, the QS_t model (4) is the closest model to the complete symmetry*

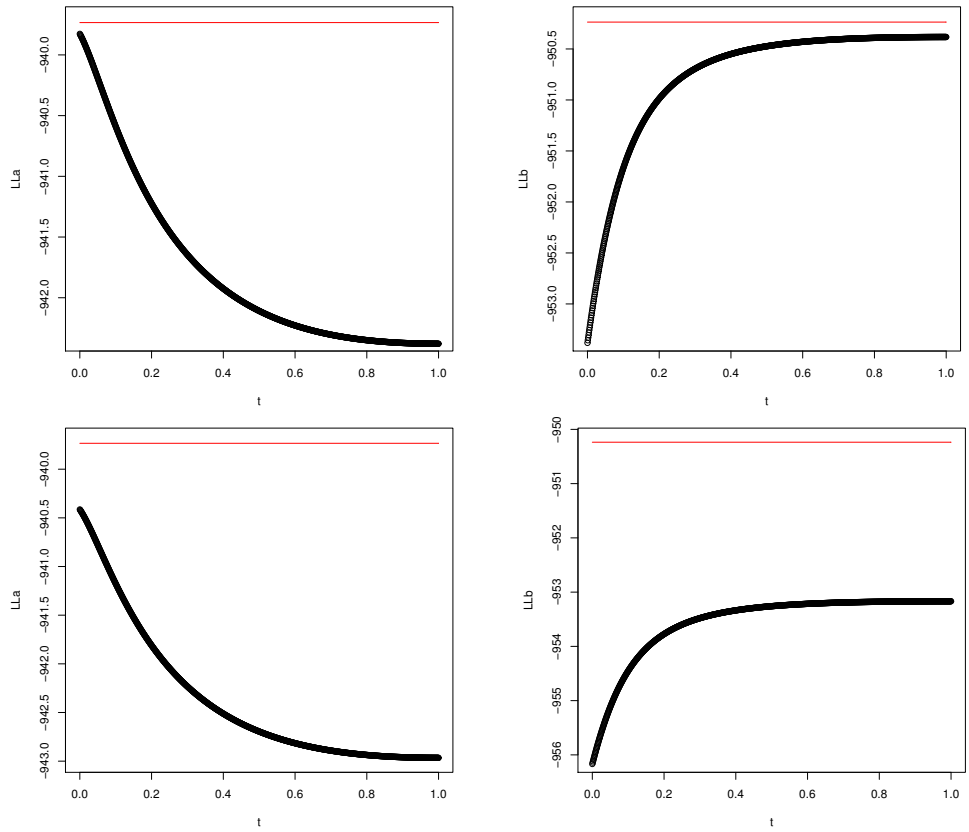

Figure 5: Log-likelihood values of QS_t (upper) and QSI_t (lower) with $t \in [0,1]$ for the data in Table 3 (a, left) and (b, right). The straight line marks the saturated log-likelihood value.

model S in (1)*, where 'closest' refers the ϕ-divergence defined by*

$$\begin{array}{rcl} \phi(u) & = & f_t(u) - f_t(1) - f_t'(1)(u-1), \\ \textit{where} \quad f_t(u) & = & (u + \frac{2t}{1-t}) \log(u + \frac{2t}{1-t}). \end{array} \tag{34}$$

Proof. We set $F_t(u) = \phi'(u) = \log(u + \frac{2t}{1-t}) - \ell_t$, where $\ell_t = \log(1 + \frac{2t}{1-t})$ is just a constant for given t. This choice of constant ensures $\phi'(1) = 0$. Then the inverse function to F_t is

$$F_t^{-1}(x) = (\frac{-2t}{1-t}) + e^{x+\ell_t}.$$

With this, we can write

$$p_{ij} \;=\; s_{ij} F_t^{-1}(\alpha_i + \gamma_{ij}) \;=\; s_{ij}(\frac{-2t}{1-t} + e^{\alpha_i + \gamma_{ij} + \ell_t}) \;=\; s_{ij}\big(\frac{-2t}{1-t} + \frac{\beta_i(\frac{2(1+t)}{1-t})}{\beta_i + \beta_j}\big)\ ,$$

where

$$\beta_i \;=\; e^{\alpha_i + \ell_t} \quad \text{and} \quad e^{\gamma_{ij}} \;=\; \frac{\frac{2(1+t)}{1-t}}{e^{\alpha_i + \ell_t} + e^{\alpha_j + \ell_t}}\ .$$

We next rewrite p_{ij} as

$$p_{ij} \;=\; s_{ij}\big(1 + \frac{-(1+t)}{1-t} + \frac{\beta_i(\frac{2(1+t)}{1-t})}{\beta_i + \beta_j}\big) = s_{ij}\big(1 + \frac{\frac{(1+t)}{1-t}(\beta_i - \beta_j)}{\beta_i + \beta_j}\big)\ .$$

Setting $\beta_i = 1 + (1-t)a_i$ and $\beta_j = 1 + (1-t)a_j$, this translates into our parametrization (4). Now the result follows from [7, Theorem 1]. For a probability table $\mathbf{s}$ with symmetry S, the quantity $D_\phi(\mathbf{p}, \mathbf{s})$ is minimized when $\mathbf{p}$ is the probability table satisfying QS_t.

The fact that the QS_t models are ϕ-divergence QS models implies that they share all the desirable properties of the ϕ-divergence QS models [7]. This includes the properties that highlight the physical interpretation issues of these models. As far as we know, the ϕ-divergence for

the parametric ϕ_t function (34) has not been considered so far. Its study can be the subject of further research. Such a future project has the potential to build a bridge between information geometry [1] and algebraic statistics [5].

Acknowledgements

Fatemeh Mohammadi was supported by the Alexander von Humboldt Foundation. Bernd Sturmfels was supported by the NSF (DMS-0968882) and DARPA (HR0011-12-1-0011).

References

[1] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[2] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, 2007.

[3] Albert H. Bowker. A test for symmetry in contingency tables. *Journal of the american statistical association*, 43(244):572–574, 1948.

[4] Henri Caussinus. Contribution à l'analyse statistique des tableaux de corrélation. In *Annales de la Faculté des Sciences de Toulouse*, volume 29, pages 77–183. Université Paul Sabatier, 1965.

[5] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*. Springer Science & Business Media, 2008.

[6] Maria Kateri and Alan Agresti. A generalized regression model for a binary response. *Statistics & Probability Letters*, 80(2):89–95, 2010.

[7] Maria Kateri and Takis Papaioannou. Asymmetry models for contingency tables. *Journal of the American Statistical Association*, 92(439):1124–1131, 1997.

[8] Ezra Miller and Bernd Sturmfels. *Combinatorial commutative algebra*, volume 227 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2005.

[9] Lior Pachter and Bernd Sturmfels. *Algebraic statistics for computational biology*, volume 13. Cambridge University Press, 2005.

[10] Leandro Pardo. *Statistical inference based on divergence measures*. CRC Press, 2005.

[11] Fabio Rapallo. Algebraic markov bases and mcmc for two-way contingency tables. *Scandinavian journal of statistics*, 30(2):385–397, 2003.

[12] Alan Stuart. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, pages 105–110, 1953.

[13] Bernd Sturmfels. Gröbner bases and convex polytopes, University Lecture Series, vol. 8, Providence, RI. 1996.