Maximum Likelihood Estimation in Latent Class Models for Contingency Table Data

Stephen E. Fienberg
Department of Statistics, Machine
Learning Department and Cylab
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Patricia Hersh
Department of Mathematics
Indiana University
Bloomington, IN 47405-7000 USA

Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Yi Zhou
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Abstract

Statistical models with latent structure have a history going back to the 1950s and have seen widespread use in the social sciences and, more recently, in computational biology and in machine learning. Here we study the basic latent class model proposed originally by the sociologist Paul F. Lazarfeld for categorical variables, and we explain its geometric structure. We draw parallels between the statistical and geometric properties of latent class models and we illustrate geometrically the causes of many problems associated with maximum likelihood estimation and related statistical inference. In particular, we focus on issues of non-identifiability and determination of the model dimension, of maximization of the likelihood function and on the effect of symmetric data. We illustrate these phenomena with a variety of synthetic and real-life tables, of different dimensions and complexities. Much of the motivation for this work stems from the "100 Swiss Franks" problem, which we introduce and describe in detail.

Keywords: latent class model, algebraic geometry, variety, Segre variety, secant variety, effective dimension.

Contents

1 Introduction 1

2 Latent Class Models for Contingency Tables 2

3 Geometric Description of Latent Class Models 4

4 Examples Involving Synthetic Data 8

4.1 Effective Dimension and Polynomials 8

4.2 The 100 Swiss Franks Problem 10

4.2.1 Introduction 10

4.2.2 Global and Local Maxima 12

4.2.3 Unidentifiable Space 13

4.2.4 Plotting the Log-likelihood Function 16

4.2.5 Further Remarks and Open Problem 16

5 Two Applications 26

5.1 Example: Michigan Influenza 26

5.2 Data From the National Long Term Care Survey 27

6 On Symmetric Tables and the MLE 29

6.1 Introduction and Motivation 31

6.2 Preservation of Marginals and Some Consequences 34

6.3 The 100 Swiss Franks Problem 37

7 Conclusions 39

8 Acknowledgments 39

A Algebraic Geometry 40

A.1 Polynomial Ring, Ideal and Variety 40

A.2 Segre Variety and Secant Variety 44

B Symbolic Software of Computational Algebra 47

B.1 Computing the dimension of the image variety 47

B.2 Solving Polynomial Equations 49

B.3 Plotting Unidentifiable Space 52

B.4 `Surf` Script 53

C Proof of the Fixed Points for 100 Swiss Franks Problem 55

D Matlab Codes 57

1 Introduction

Latent class (LC) or latent structure analysis models were introduced in the 1950s in the social science literature to model the distribution of dichotomous attributes based on a survey sample from a population of individuals organized into distinct homogeneous classes according to an unobservable attitudinal feature. See Anderson (1954), Gibson (1955), Madansky (1960) and, in particular, Henry and Lazarfeld (1968). These models were later generalized in Goodman (1974), Haberman (1974), Clogg and Goodman (1984) as models for the joint marginal distribution of a set of manifest categorical variables, assumed to be conditionally independent given an unobservable or latent categorical variable, building upon the then recently developed literature on log-linear models for contingency tables. More recently, latent class models have been described and studied as special cases of a larger class of directed acyclic graphical models with hidden nodes, sometimes referred to as Bayes nets, Bayesian networks, or causal models, e.g., see Lauritzen (1996), Cowell et al. (1999), Humphreys and Titterington (2003) and, in particular, Geiger et al. (2001). A number of recent papers have established fundamental connections between the statistical properties of latent class models and their algebraic and geometric features, e.g., see Settimi and Smith (1998, 2005), Smith and Croft (2003), Rusakov and Geiger (2005),Watanabe (2001) and Garcia et al. (2005).

Despite these recent important theoretical advances, the fundamental statistical tasks of estimation, hypothesis testing and model selection remain surprisingly difficult and, in some cases, infeasible, even for small latent class models. Nonetheless, LC models are widely used and there is a "folklore" associated with estimation in various computer packages implementing algorithms such as EM for estimation purposes, e.g., see Uebersax (2006a,b).

The goal of this article is two-fold. First, we offer a simplified geometric and algebraic description of LC models and draw parallels between their statistical and geometric properties. The geometric framework enjoys notable advantages over the traditional statistical representation and, in particular, offers natural ways of representing singularities and non-identifiability problems. Furthermore, we argue that the many statistical issues encountered in fitting and interpreting LC models are a reflection of complex geometric attributes of the associated set of probability distributions. Second, we illustrate with examples, most of which quite small and seemingly trivial, some of the computational, statistical and geometric challenges that LC models pose. In particular, we focus on issues of non-identifiability and determination of the model dimension, of maximization of the likelihood function and on the effect of symmetric data. We also show how to use symbolic software from computational algebra to obtain a more convenient and simpler parametrization and for unravelling the geometric features of LC models. These strategies and methods should carry over to more complex latent structure models, such as in Bandeen-Roche et al. (1997).

In the next section, we describe the basic latent class model and its statistical properties and, in Section 3, we discuss the geometry of the models. In Section 4, we turn to our examples exemplifying the identifiability issue and the complexity of the likelihood function, with a novel focus on the problems arising from symmetries in the data. Finally, we present some computational results for two real-life examples, of small and very large dimension.

2 Latent Class Models for Contingency Tables

Consider k categorical variables, $X_1, \ldots, X_k$, where each X_i takes value on the finite set $[d_i] \equiv \{1, \ldots, d_i\}$. Letting $\mathcal{D} = \bigotimes_{i=1}^{k} [d_i]$, $\mathbb{R}^{\mathcal{D}}$ is the vector space of k-dimensional arrays of the format $d_1 \times \ldots \times d_k$, with a total of $d = \prod_i d_i$ entries. The cross-classification of N independent and identically distributed realizations of $(X_1, \ldots, X_k)$ produces a random integer-valued vector $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, whose coordinate entry $\mathbf{n}_{i_i,\ldots,i_k}$ corresponds to the number of times the label combination $(i_1, \ldots, i_k)$ was observed in the sample, for each $(i_1, \ldots, i_k) \in \mathcal{D}$. The table $\mathbf{n}$ has a Multinomial$(N, \mathbf{p})$ distribution, where $\mathbf{p}$ is a point in the $(d-1)$-dimensional probability simplex Δ_{d-1} with coordinates

$$p_{i_1,\ldots,i_k} = Pr\left\{(X_1, \ldots, X_k) = (i_1, \ldots, i_k)\right\}, \qquad (i_1, \ldots, i_k) \in \mathcal{D}.$$

Let H be an unobservable latent variable, defined on the set $[r] = \{1, \ldots, r\}$. In its most basic version, also known as the *naive Bayes model*, the LC model postulates that, conditional on H, the variables $X_1, \ldots, X_k$ are mutually independent. Specifically, the joint distributions of $X_1, \ldots, X_k$ and H form the subset $\mathcal{V}$ of the probability simplex Δ_{dr-1} consisting of points with coordinates

$$p_{i_1,\ldots,i_k,h} = p_1^{(h)}(i_1) \ldots p_k^{(h)}(i_k)\lambda_h, \qquad (i_1, \ldots, i_k, h) \in \mathcal{D} \times [r], \tag{1}$$

where λ_h is the marginal probability $Pr\{H = h\}$ and $p_l^{(h)}(i_l)$ is the conditional marginal probability $Pr\{X_l = i_l | H = h\}$, which we assume to be strictly positive for each $h \in [r]$ and $(i_1, \ldots, i_k) \in \mathcal{D}$.

The log-linear model specified by the polynomial mappings (1) is a decomposable graphical model (see, e.g, Lauritzen, 1996) and $\mathcal{V}$ is the image set of a homeomorphism from the parameter space

$$\begin{aligned} \Theta &\equiv \left\{\theta \colon \theta = (p_1^{(h)}(i_1) \ldots p_k^{(h)}(i_k), \lambda_h), (i_1, \ldots, i_k, h) \in \mathcal{D} \times [r]\right\} \\ &= \bigotimes\nolimits_i \Delta_{d_i - 1} \times \Delta_{r-1}, \end{aligned}$$

so that global identifiability is guaranteed. The remarkable statistical properties of this type of models and the geometric features of the set $\mathcal{V}$ are well understood. Statistically, equation (1) defines a linear exponential family of distributions, though not in its natural parametrization. The maximum likelihood estimates, or MLEs, of λ_h and $p_l^{(h)}(i_l)$ exist if and only if the minimal sufficient statistics, i.e., the empirical joint distributions of (X_i, H) for $i = 1, 2, \ldots, k$, are strictly positive and are given in closed form as rational functions of the observed two-way marginal distributions between X_i and H for $i = 1, 2, \ldots, k$. The log-likelihood function is strictly concave and the maximum is always attainable, possibly on the boundary of the parameter space. Furthermore, the asymptotic theory of goodness-of-fit testing is fully developed.

Geometrically, we can obtain the set $\mathcal{V}$ as the intersection of Δ_{dr-1} with an affine variety (see, e.g., Cox et al., 1996) consisting of the solutions set of a system of $r \prod_i \binom{d_i}{2}$ homogeneous square-free polynomials. For example, when $k = 2$, each of these polynomials take the form of quadric equations of the type

$$p_{i_1,i_2,h} p_{i'_1,i'_2,h} = p_{i'_1,i_2,h} p_{i_1,i'_2,h}, \tag{2}$$

with $i_1 \neq i'_1$, $i_2 \neq i'_2$ and for each fixed h. Provided the probabilities are strictly positive, equations of the form (2) specify conditional odds ratio of 1, for every pair $(X_i, X_{i'})$ given $H = h$. Furthermore, for each given h, the coordinate projections of the first two coordinates of the points satisfying (2) trace the surface of independence inside the simplex Δ_{d-1}. The strictly positive points of $\mathcal{V}$ form a smooth manifold whose dimension is $r \prod_i (d_i - 1) + (r - 1)$ and whose co-dimension corresponds to the number of degrees of freedom. The singular points of $\mathcal{V}$ all lie on the boundary of the simplex Δ_{dr-1} and identify distributions with degenerate probabilities along some coordinates. More generally, the singular locus of $\mathcal{V}$ can be described similarly in terms of stratified components of $\mathcal{V}$, whose dimensions and co-dimensions can also be computed explicitly.

Under the LC model, the variable H is unobservable and the new model $\mathcal{H}$ is a r-class mixture over the exponential family of distributions prescribing mutual independence among the manifest variables $X_1, \ldots, X_k$. Geometrically, $\mathcal{H}$ is the set of probability vectors in Δ_{d-1} obtained as the image of the marginalization map from Δ_{dr-1} onto Δ_{d-1} which consists of taking the sum over the coordinate corresponding to the latent variable. Formally, $\mathcal{H}$ is made up of all probability vectors in Δ_{d-1} with coordinates satisfying the *accounting equations* (see, e.g., Henry and Lazarfeld, 1968)

$$p_{i_1,\ldots,i_k} = \sum_{h \in [r]} p_{i_1,\ldots,i_k,h} = \sum_{h \in [r]} p_1^{(h)}(i_1) \ldots p_k^{(h)}(i_k) \lambda_h, \tag{3}$$

where $(i_1, \ldots, i_k, h) \in \mathcal{D} \times [r]$.

Despite being expressible as a convex combination of very well-behaved models, even the simplest form of the LC model (3) is far from being well-behaved and, in fact, shares virtually none of the properties of the standard log-linear models. In particular, the latent class models specified by equations (3) do not define exponential families, but instead belong to a broader class of models called stratified exponential families (see Geiger et al., 2001), whose properties are much weaker and less well understood. *The minimal sufficient statistics for an observed table* $\mathbf{n}$ *are the observed counts themselves and we can achieve no data reduction via sufficiency.* The model may not be identifiable, because for a given $\mathbf{p} \in \Delta_{d-1}$ satisfying (3), there may be a subset of Θ, known as the *non-identifiable space*, consisting of parameter points all satisfying the same accounting equations. The non-identifiability issue has in turn considerable repercussions on the determination of the correct number of degrees of freedom for assessing model fit and, more importantly, on the asymptotic properties of standard model selection criteria (e.g. likelihood ratio statistic and other goodness-of-fit criteria such as BIC, AIC, etc), whose applicability and correctness may no longer hold.

Computationally, maximizing the log-likelihood can be a rather laborious and difficult task, particularly for high dimensional tables, due to lack of concavity, the presence of local maxima and saddle points, and singularities in the observed Fisher information matrix. Geometrically, $\mathcal{H}$ is no longer a smooth manifold in the relative interior of Δ_{d-1}, with singularities even at probability vectors with strictly positive coordinates, as we show in the next section. The problem of characterizing the singular locus of $\mathcal{H}$ and of computing the dimensions of its stratified components (and of the tangent spaces and tangent cones of its singular points) is of statistical importance: singularity points of $\mathcal{H}$ are probability distributions of lower complexity, in the sense that they are specified by lower-dimensional subsets of Θ, or, loosely speaking, by less parameters. Because the sample space is discrete, although the singular locus of $\mathcal{H}$ has typically Lebesgue measure zero, there is nonetheless a positive probability that the maximum likelihood estimates end up being either a singular point in the relative interior of the simplex Δ_{d-1} or a point on the boundary. In both cases, standard asymptotics for hypothesis testing and model selection fall short.

3 Geometric Description of Latent Class Models

In this section, we give a geometric representation of latent class models, summarize existing results and point to some of the relevant mathematical literature. For more details, see Garcia et al. (2005) and Garcia (2004).

The latent class model defined by (3) can be described as the set of all convex combinations of r-tuple of points lying on the surface of independence inside Δ_{d-1}. Formally, let

$$\begin{array}{rccc} \sigma\colon & \Delta_{d_1-1}\times\ldots\times\Delta_{d_k-1} & \to & \Delta_{d-1} \\ & (p_1(i_1),\ldots,p_k(i_k)) & \mapsto & \prod_j p_j(i_j) \end{array}$$

be the map that sends the vectors of marginal probabilities into the k-dimensional array of joint probabilities for the model of complete independence. The set $\mathcal{S}\equiv\sigma(\Delta_{d_1-1}\times\ldots\times\Delta_{d_k-1})$ is a manifold in Δ_{d-1} known in statistics as the surface of independence and in algebraic geometry (see, e.g. Harris, 1992) as (the intersection of Δ_{d-1} with) the Segre embedding of $\mathbb{P}^{d_1-1}\times\ldots\times\mathbb{P}^{d_k-1}$ into $\mathbb{P}^{d-1}$. The dimension of $\mathcal{S}$ is $\prod_i(d_i-1)$, i.e., the dimension of the corresponding decomposable model of mutual independence. The set $\mathcal{H}$ can then be constructed geometrically as follows. Pick any combination of r points along the hyper-surface $\mathcal{S}$, say $\mathbf{p}^{(1)},\ldots,\mathbf{p}^{(r)}$, and determine their convex hull, i.e. the convex set consisting of all points of the form $\sum_h \mathbf{p}^{(h)}\lambda_h$, for some choice of $(\lambda_1,\ldots,\lambda_r)\in\Delta_{r-1}$. The coordinates of any point in this new subset satisfy, by construction, the accounting equations (3). In fact, the closure of the union of all such convex hulls is precisely the latent class model $\mathcal{H}$. In algebraic geometry, $\mathcal{H}$ would be described as the intersection of Δ_{d-1} with the r-th secant variety of the Segre embedding mentioned above. More about the Segre and secant varieties can be found in the appendix A.2.

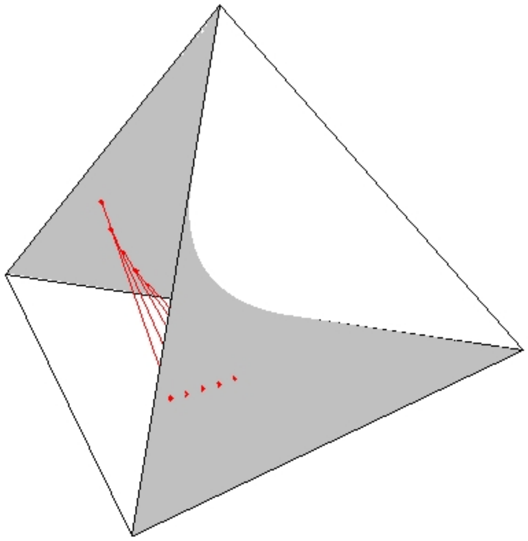


Figure 1: Surface of independence for the 2×2 table with 3 secant lines.

Example 3.1 *The simplest example of a latent class model is for a 2×2 table with one latent variable with $r = 2$. The surface of independence, i.e. the intersection of the simplex Δ_3 with the Segre variety, is shown in Figure 1. The secant variety for this latent class models is the union of all the secant lines, i.e. the lines connecting any two distinct points lying on the surface of independence. Figure 1 displays three such secant lines. It is not too hard to picture that the union of all such secant lines is the enveloping simplex Δ_3 and, therefore, $\mathcal{H}$ fills up all the available space (for formal arguments, see Catalisano et al., 2002, Proposition 2.3).*

The model $\mathcal{H}$ is not a smooth manifold. Instead, it is a semi-algebraic set (see, e.g., Benedetti, 1990), clearly singular on the boundary of the simplex, but also at strictly positive points along the $(r-1)$st secant variety, (both of Lebesgue measure zero). This means that the model is singular at all points in $\mathcal{H}$ which satisfy the accounting equations with one or more of the λ_h's equal to zero. In Example 3.1 above, the surface of independence is a singular locus for the latent class model. From the statistical viewpoint, singular points of $\mathcal{H}$ correspond to simpler models for which the number of latent classes is less than r (possibly 0). As usual, for these points one needs to adjust the number of degrees of freedom to account for the larger tangent space.

Unfortunately, we have no general closed-form expression for computing the dimension of $\mathcal{H}$ and the existing results only deal with specific cases. Simple considerations allow us to compute an upper bound for the dimension of $\mathcal{H}$, as follows. As Example 3.1 shows, there may be instances for which $\mathcal{H}$ fills up the entire simplex Δ_{d-1}, so that $d-1$ is an attainable upper bound. Counting the number of free parameters in (3), we can see that this dimension cannot exceed $r\sum_i(d_i-1)+r-1$, (c.f. Goodman, 1974, page 219). This number, the *standard dimension*, is the dimension of the fully observable model of conditional independence. Incidentally, this value can be determined mirroring the geometric construction of $\mathcal{H}$ as follows (c.f. Garcia, 2004). The number $r\sum_i(d_i-1)$ arises from the choice of r points along the $\sum_i(d_i-1)$-dimensional surface of independence, while the term $r-1$ accounts for the number of free parameters for a generic choice of $(\lambda_1,\ldots,\lambda_r)\in\Delta_{r-1}$. Therefore, we conclude that the dimension of $\mathcal{H}$ is bounded by

$$\min\left\{d-1, r\sum_i(d_i-1)+r-1\right\}, \tag{4}$$

a value known in algebraic geometry as the *expected dimension* of the variety $\mathcal{H}$.

Cases of latent class models with dimension strictly smaller than the expected dimension have been known for a long time, however. In the statistical literature, Goodman (1974) noticed that the latent class models for 4 binary observable variables and a 3-level latent variable, whose expected dimension is 14, has dimension 13. In algebraic geometry, secant varieties with dimension smaller than the expected dimension (4) are called *deficient* (e.g., see Harris, 1992). In particular, Exercise 11.26 in Harris (1992) gives an example of a deficient secant variety, which corresponds to a latent class model for a 2-way table with a binary latent variable. In this case, the deficiency is 2, as is demonstrated below in equation (5). The true or *effective* dimension of a latent class model, i.e. the dimension of the semi-algeraic set $\mathcal{H}$ representing it, is crucial for establishing identifiability and for computing correctly the number of degrees of freedom. In fact, if a model is deficient, then the pre-image of each probability array in $\mathcal{H}$ arising from the accounting equations is a subset (in fact, a variety) of Θ called the *non-dentifiable subspace*, with dimension exactly equal to the deficiency itself. Therefore, a deficient model is non-identifiable, with adjusted degrees of freedom equal to number of degrees of freedom for the observable graphical model plus the value of the deficiency.

The effective dimension of $\mathcal{H}$ is equal to the maximal rank of the Jacobian matrix for the polynomial mapping from Θ into $\mathcal{H}$ given coordinatewise by (3). Geiger et al. (2001) showed that this value is equal to the dimension of $\mathcal{H}$ almost everywhere with respect to the Lebsegue measure, provided the Jacobian is evaluated at strictly positive parameter points θ, and used this result to devise a simple and efficient algorithm to compute numerically the effective dimension. We include the Matlab codes for computing the numberical rank of the Jacobian in the appendix D.

Recently, in the algebraic-geometry literature, Catalisano et al. (2002, 2003) have obtained explicit formulas for the effective dimensions of some secant varieties which are of statistical interest. In particular, they show that for $k = 3$ and $r \leq \min\{d_1, d_2, d_3\}$, the latent class model has the expected dimension and is identifiable. On the other hand, assuming $d_1 \leq d_2 \leq \ldots \leq d_k$, $\mathcal{H}$ is deficient when $\prod_{i=1}^{k-1} d_i - \sum_{i=1}^{k-1}(d_i - 1) \leq r \leq \min\left\{d_k, \prod_{i=1}^{k-1} d_i - 1\right\}$. Finally, under the same conditions, $\mathcal{H}$ is identifiable when $\frac{1}{2}\sum_i(d_i - 1) + 1 \geq \max\{d_k, r\}$. In general, obtaining bounds and results of this type is highly non-trivial and is an open area of research. Refer to the appendix A.2 to see more results on the effective dimension of secant varieties.

In the remainder of the paper, we will focus on simpler latent class models for tables of dimension $k = 2$ and illustrate with examples the results mentioned above. For latent class models on two-way tables, there is an alternative, quite convenient way of describing $\mathcal{H}$ by representing each $\mathbf{p}$ in Δ_{d-1} as a $d_1 \times d_2$ matrix and by representing the map σ as a vector product. In fact, each point $\mathbf{p}$ in $\mathcal{S}$ is a rank one matrix obtained as $\mathbf{p}_1\mathbf{p}_2^\top$, where $\mathbf{p}_1 \in \Delta_{d_1-1}$ and $\mathbf{p}_2 \in \Delta_{d_1-2}$ are the appropriate marginal distributions of X_1 and X_2. Then, the accounting equations for a latent class models with r-level become

$$\mathbf{p} = \sum_h \mathbf{p}_1^{(h)}(\mathbf{p}_2^{(h)})^\top \lambda_h, \qquad (\mathbf{p}_1, \mathbf{p}_2, (\lambda_1, \ldots, \lambda_r)) \in \Delta_{d_1-1} \times \Delta_{d_2-1} \times \Delta_{r-1}$$

i.e. the matrix $\mathbf{p}$ is a convex combination of r rank 1 matrices lying on the surface of independence. Therefore all points in $\mathcal{H}$ are non-negative matrices with entries summing to one and with rank at most r. This simple observation allows one to compute the effective dimension of $\mathcal{H}$ for 2-way table as follows. In general, a real valued $d_1 \times d_2$ matrix has rank r or less if and only if the homogeneous polynomial equations corresponding to all of its $(r+1) \times (r+1)$ minors all vanish. Provided $k < \min\{d_1, d_2\}$, on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, the zero locus of all such equations form a *determinantal* variety of co-dimension $(d_1 - r)(d_2 - r)$ (c.f. Harris, 1992, Proposition 12.2) and hence has dimension $r(d_1 + d_2) - r^2$. Subtracting this value from the expected dimension (4), and taking into account the fact that all the points lie inside the simplex, we obtain

$$r(d_1 + d_2 - 2) + r - 1 - \left(r(d_1 + d_2) - r^2 - 1\right) = r(r - 1). \tag{5}$$

This number is also the difference between the dimension of the (fully identifiable, i.e. of expected dimension) graphical model of conditional independence X_1 and X_2 given H, and the deficient dimension of the latent class model obtained by marginalizing over the variable H.

The study of higher dimensional tables is still an open area of research. The mathematical machinery required to handle larger dimensions is considerably more complicated and relies on the notions higher-dimensional tensors, rank tensors and non-negative rank tensors, for which only partial results exist. See Kruskal (1975), Cohen and Rothblum (1993) and Strassen (1983) for details. Alternatively, Mond et al. (2003) conduct an algebraic-topological investigation of the topological properties of stochastic factorization of stochastic matrices representing models of conditional independence with one hidden variable and

Allman and Rhodes (2006, 2007) explore an overlapping set of problems framed in the context of trees with latent nodes and branches.

The specific case of k-way tables with 2 level latent variables is a fortunate exception, for which the results for 2-way tables just described apply. In fact, Landsberg and Manivel (2004) show that that these models are the same as the corresponding model for any two-dimensional table obtained by any "flattening" of the $d_1 \times \ldots \times d_k$-dimensional array of probabilities $\mathbf{p}$ into a two-dimensional matrix. Flattening simply means collapsing the k variables into two new variables with f_1 and f_2 levels, and re-organizing the entries of the k-dimensional tensor $\mathbf{p} \in \Delta_{d-1}$ into a $f_1 \times f_1$ matrix accordingly, where, necessarily, $f_1 + f_2 = \sum_i d_i$. Then, $\mathcal{H}$ is the determinantal variety which is the zero set of all 3×3 sub-determinants of the matrix obtained by any such flattening. The second example in Section 4.1 below illustrates this result.

4 Examples Involving Synthetic Data

We further elucidate the non-identifiability phenomenon from the algebraic and geometric point of view, and the multi-modality of the log-likelihood function issue using small synthetic examples. In particular, in the "100 Swiss Franks" problem below, we embark on an exhaustive study of a table with symmetric data and describe the effects of such symmetries on both the parameter space and the log-likelihood function. Although those examples treat simplest cases of LC models, they already exhibit considerable statistical and geometric complexity.

4.1 Effective Dimension and Polynomials

We show how it is possible to take advantage of the polynomial nature of equations (3) to gain further insights into the algebraic properties of distributions obeying latent class models. All the computations that follow were made in `SINGULAR` (Greuel et al., 2005) and are described in details, along with more examples, in Zhou (2007). Although in principle symbolic algebraic software allows one to compute the set of polynomial equations that fully characterize LC models and their properties, this is still a rather difficult and costly task that can be accomplished only for smaller models.

The accounting equations (3) determine a polynomial mapping $f \colon \Theta \to \Delta_{d-1}$ given by

$$(p_1(i_1) \ldots p_k(i_k), \lambda_h) \mapsto \sum_{h \in [r]} p_1(i_1) \ldots p_k(i_k) \lambda_h, \tag{6}$$

so that the latent class model is analytically defined as its image, i.e. $\mathcal{H} = f(\Theta)$. Then, following the geometry-algebra dictionary principle (see, e.g., Cox et al., 1996), the problem of computing the effective dimension of $\mathcal{H}$ can in turn be geometrically cast as a problem of computing the dimension of the image of a polynomial map. We illustrate how this representation offers considerable advantages with some small examples.

Consider a $2 \times 2 \times 2$ table with $r = 2$ latent classes. From Proposition 2.3 in Catalisano et al. (2002), the latent class models with 2 classes and 3 manifest variables are identifiable. The standard dimension, i.e. the dimension of the parameter space Θ is $r \sum_i (d_i - 1) + r - 1 = 7$, which coincides with the dimension of the enveloping simplex Δ_7. Although this condition implies that the number of parameters to estimate is no larger than the number of cells in the table, a case which, if violated, would entail non-identifiability, it does not guarantee that the effective dimension is also 7. This can be verified by checking that the symbolic rank of the Jacobian matrix of the map (6) is indeed 7, almost everywhere with respect to the Lebesgue measure. Alternatively, one can determine the dimension of the non-identifiable subspace using computational symbolic algebra. First, we consider the ideal of polynomials generated by the 8 equations in (6) in the polynomial ring in which the (redundant) 16 indeterminates are the 8 joint probabilities in Δ_7 and the 3 pairs of marginal probabilities in Δ_1 for the observable variables, and the marginal probabilities in Δ_1 for the latent variable. Then we use implicization (see, e.g., Cox et al., 1996, Chapter 3) to eliminate all the marginal probabilities and to study the Groebner basis of the resulting ideal in which the indeterminates are the joint probabilities only. There is only one element in the basis,

$$p_{111} + p_{112} + p_{121} + p_{122} + p_{211} + p_{212} + p_{221} + p_{222} = 1,$$

which gives the trivial condition for probability vectors. This implies the map (6) is surjective, so that $\mathcal{H} = \Delta_7$ and the effective dimension is also 7, showing identifiability, at least for positive distributions.

Next, we consider the $2 \times 2 \times 3$ table with $r = 2$. For this model, Θ has dimension 9 and the symbolic rank of the associated Jacobian matrix is 9 as well, so that the model is identifiable. Alternatively, using the same route as in the previous example, we see that, in this case, the image of the polynomial mapping (6) is the variety associated to the ideal whose Groebner basis consists of the trivial equation

$$p_{111} + p_{112} + p_{113} + p_{121} + p_{122} + p_{123} + p_{211} + p_{212} + p_{213} + p_{221} + p_{222} + p_{223} = 1,$$

and four polynomials corresponding to the determinants

$$
\begin{gathered}
\begin{vmatrix} p_{121} & p_{211} & p_{221} \\ p_{122} & p_{212} & p_{222} \\ p_{123} & p_{213} & p_{223} \end{vmatrix} \\
\\
\begin{vmatrix} p_{1+1} & p_{211} & p_{221} \\ p_{1+2} & p_{212} & p_{222} \\ p_{1+3} & p_{213} & p_{223} \end{vmatrix} \\
\\
\begin{vmatrix} p_{+11} & p_{121} & p_{221} \\ p_{+12} & p_{122} & p_{222} \\ p_{+13} & p_{123} & p_{223} \end{vmatrix} \\
\\
\begin{vmatrix} p_{111} & p_{121}+p_{211} & p_{221} \\ p_{112} & p_{122}+p_{212} & p_{222} \\ p_{113} & p_{123}+p_{213} & p_{223} \end{vmatrix}
\end{gathered}
\tag{7}
$$

where the subscript symbol "+" indicates summation over that coordinate. The SINGULAR codes for computing the Groebner basis are in the appendix B.1. The zero set of the above determinants coincide with the determinantal variety specified by the zero set of all 3×3 minors of the $3{\times}4$ matrix

$$
\begin{pmatrix} p_{111} & p_{121} & p_{211} & p_{221} \\ p_{112} & p_{122} & p_{212} & p_{222} \\ p_{113} & p_{123} & p_{213} & p_{223} \end{pmatrix}
\tag{8}
$$

which is a flattening of the $2 \times 2 \times 3$ array of probabilities describing the joint distribution for the latent class model under study. This is in accordance with the result in Landsberg and Manivel (2004) mentioned above. Now, the determinantal variety given by the vanishing locus of all the $3{\times}3$ minors of the matrix (8) is the latent class model for a 3×4 table with 2 latent classes, which, according to (5), has deficiency equal to 2. The effective dimension of this variety is 9, computed as the standard dimension, 11, minus the deficiency. Therefore, the effective dimension of the model we are interested is also 9 and we conclude that the model is identifiable.

Table 10 summarizes some of our numerical evaluations of the different notions of dimension for a different LC models. We computed the effective dimensions by evaluating with MATLAB the numerical rank of the Jacobian matrix, based on the algorithm of Geiger et al. (2001) and also using SINGULAR, for which only computations involving small models were feasible.

4.2 The 100 Swiss Franks Problem

4.2.1 Introduction

Now we study the problem of fitting a non-identifiable 2-level latent class model to a two-way table with symmetry counts. This problem was suggested by

Table 1: Different dimensions of some latent class models. The Complete Dimension is the dimension $d-1$ of the envoloping probability simplex Δ_{d-1}. See also Table 1 in Kocka and Zhang (2002).

Latent Class Model		Effective Dimension	Standard Dimension	Complete Dimension	Deficiency
Δ_{d-1}	r				
2×2	2	3	5	3	0
3×3	2	7	9	8	1
4×5	3	17	23	19	2
$2 \times 2 \times 2$	2	7	7	7	0
$2 \times 2 \times 2$	3	7	11	7	0
$2 \times 2 \times 2$	4	7	15	7	0
$3 \times 3 \times 3$	2	13	13	26	0
$3 \times 3 \times 3$	3	20	20	26	0
$3 \times 3 \times 3$	4	25	27	26	1
$3 \times 3 \times 3$	5	26	34	26	0
$3 \times 3 \times 3$	6	26	41	26	0
$5 \times 2 \times 2$	3	17	20	19	2
$4 \times 2 \times 2$	3	14	17	15	1
$3 \times 3 \times 2$	5	17	29	17	0
$6 \times 3 \times 2$	5	34	44	35	1
$10 \times 3 \times 2$	5	54	64	59	5
$2 \times 2 \times 2 \times 2$	2	9	9	15	0
$2 \times 2 \times 2 \times 2$	3	13	14	15	1
$2 \times 2 \times 2 \times 2$	4	15	19	15	0
$2 \times 2 \times 2 \times 2$	5	15	24	15	0
$2 \times 2 \times 2 \times 2$	6	15	29	15	0

Bernd Sturmfels to the participants of his postgraduate lectures on Algebraic Statistics held at ETH Zurich in the Summer semester of 2005 (where he offered 100 Swiss Franks for a rigorous solution), and is described in detail as Example 1.16 in Pachter and Sturmfels (2005). The observed table is

$$n = \begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix} \tag{9}$$

and the 100 Swiss Franks problem requires proving that the three tables in Table 2 **a)** are global maxima for the the basic LC model with one binary latent variable. For this model, the standard dimension of $\Theta = \Delta_3 \times \Delta_3 \times \Delta_1$ is $2(3+3)+1 = 13$ and, by (5), the deficiency is 2. Thus, the model is not identifiable and the pre-image of each point $\mathbf{p} \in \mathcal{H}$ by the map (6) is a 2-dimensional surface in Θ. To keep the notation light, we write α_{ih} for $p_1^{(h)}(i)$ and β_{jh} for $p_2^{(h)}(j)$, where $i, j = 1, \ldots, 4$ and $\alpha^{(h)}$ and $\beta^{(h)}$ for the conditional marginal distribution of X_1 and X_2 given $H = h$, respectively. The accounting equations for the points in $\mathcal{H}$ become

$$p_{ij} = \sum_{h \in \{1,2\}} \lambda_h \alpha_{ih} \beta_{jh}, \qquad i, j \in [4] \tag{10}$$

and the log-likelihood function, ignoring an irrelevant additive constant, is

$$\ell(\theta) = \sum_{i,j} n_{ij} \log \left(\sum_{h \in \{1,2\}} \lambda_h \alpha_{ih} \beta_{jh} \right), \qquad \theta \in \Delta_3 \times \Delta_3 \times \Delta_1.$$

It is worth emphasizing, as we did above and as the previous display clearly shows, that the observed counts are minimal sufficient statistics.

Alternatively, we can re-parametrize the log-likelihood function using directly points in $\mathcal{H}$ rather than the points in the parameter space Θ. Recall from our discussion in section 3 that, for this model, the 4×4 array $\mathbf{p}$ is in $\mathcal{H}$ if and only if each 3×3 minor vanishes. Then, we can write the log-likelihood function as

$$\ell(\mathbf{p}) = \sum_{i,j} n_{ij} \log p_{ij}, \qquad \mathbf{p} \in \Delta_{15}, \ \det(\mathbf{p}^*_{ij}) = 0 \text{ all } i, j \in [4], \tag{11}$$

where $\mathbf{p}^*_{ij}$ is the 3×3 sub-matrix of $\mathbf{p}$ obtained by erasing the ith row and the jth column.

Although the first order optimality conditions for the Lagrangian corresponding to the parametrization (11) are algebraically simpler and can be given the form of a system of a polynomial equations, in practice, the classical parametrization (10) is used in both the EM and the Newton-Raphson implementations in order to compute the maximum likelihood estimate of $\mathbf{p}$. See Goodman (1979), Haberman (1988), and Redner and Walker (1984) for more details about these numerical procedures. Codes for solving equation 11 in SINGULAR are given in the appendix B.2.

4.2.2 Global and Local Maxima

Using both EM and Newton-Raphson algorithm with several different starting points, we found 7 local maxima of the log-likelihood function, reported in Table 2. The maximal value of the log-likelihood function was found experimentally to be $-20.8074 + const.$, where *const.* denotes the additive constant stemming from the multinomial coefficient. The maximum is achieved by the three tables of fitted values Table 2 **a)**. The remaining four tables are local maxima of $-20.8616 + const.$, close in value to the actual global maximum. Using SINGULAR (see (Greuel et al., 2005)), we checked that the tables found satisfy the first order optimality conditions (11). After verifying numerically the second order optimality conditions, we conclude that those points are indeed local maxima. As noted in Pachter and Sturmfels (2005), the log-likelihood function also has a few saddle points.

Table 2: Tables of fitted value corresponding to the 7 maxima of the likelihood equation for the 100 Swiss Franks data shown in (9). **a)**: global maximua (log-likelihood value -20.8079). **b)**: local maxima (log-likelihood value -20.8616).

a)

$$\begin{pmatrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{pmatrix} \quad \begin{pmatrix} 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \\ 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \end{pmatrix} \quad \begin{pmatrix} 3 & 2 & 2 & 3 \\ 2 & 3 & 3 & 2 \\ 2 & 3 & 3 & 2 \\ 3 & 2 & 2 & 3 \end{pmatrix}$$

b)

$$\begin{pmatrix} 8/3 & 8/3 & 8/3 & 2 \\ 8/3 & 8/3 & 8/3 & 2 \\ 8/3 & 8/3 & 8/3 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix} \quad \begin{pmatrix} 8/3 & 8/3 & 2 & 8/3 \\ 8/3 & 8/3 & 2 & 8/3 \\ 2 & 2 & 4 & 2 \\ 8/3 & 8/3 & 2 & 8/3 \end{pmatrix}$$

$$\begin{pmatrix} 8/3 & 2 & 8/3 & 8/3 \\ 2 & 4 & 2 & 2 \\ 8/3 & 2 & 8/3 & 8/3 \\ 8/3 & 2 & 8/3 & 8/3 \end{pmatrix} \quad \begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 8/3 & 8/3 & 8/3 \\ 2 & 8/3 & 8/3 & 8/3 \\ 2 & 8/3 & 8/3 & 8/3 \end{pmatrix}$$

A striking feature of the global maxima in Table 2 is their invariance under the action of the symmetric group on four elements acting simultaneously on the row and columns. Different symmetries arise for the local maxima. We will give an explicit representation of these symmetries under the classical parametrization (10) in the next section.

Despite the simplicity and low-dimensionality of the LC model for this table and the strong symmetric features of the data, we have yet to provide a purely mathematical proof that the three top arrays in Table 2 correspond to a global maximum of the likelihood function. We view the difficulty and complexity of the 100 Swiss Franks problem as a consequence of the inherent difficulty of even small LC models and perhaps an indication that the current theory has still many open, unanswered problems. In Section 6, we present partial results towards the completion of the proof.

4.2.3 Unidentifiable Space

It follows from equation (5) that the non-identifiable subspaces are a two-dimensional subsets of Θ. We give an explicit algebraic description of this space, which we will then use to obtain interpretable plots of the profile likelihood.

Firstly, we focus on the three global maxima in Table 2 **a)**. By the well-known properties of the EM algorithm (see, e.g., Pachter and Sturmfels, 2005, Theorem 1.15), if the vector of parameters θ is a stationary point in the maximization step of the EM algorithm, then θ is a critical point and hence a good

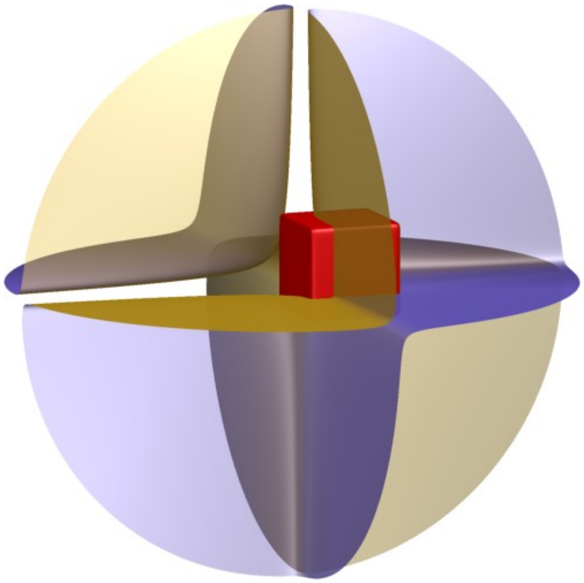


Figure 2: The 2-dimensional surface defined by equation (13), when evaluated over the ball in $\mathbb{R}^3$ of radius 3, centered at the origin. The inner box is the unit cube $[0,1]^3$ and its intersection with the surface corresponds to solutions points defining probability distributions.

candidate for a local maximum. Using this observation, it is possible to show (see Zhou, 2007) that any point in Θ satisfying the equations

$$\begin{array}{l} \alpha_{1h} = \alpha_{2h},\ \alpha_{3h} = \alpha_{4h} \quad h = 1,2 \\ \beta_{1h} = \beta_{2h},\ \beta_{3h} = \beta_{4h} \quad h = 1,2 \\ \sum_h \lambda_h \alpha_{1h}\beta_{1h} = \sum_h \lambda_h \alpha_{3h}\beta_{3t} = 3/40 \\ \sum_h \lambda_h \alpha_{1h}\beta_{3h} = \sum_h \lambda_h \alpha_{3h}\beta_{1t} = 2/40 \end{array} \tag{12}$$

is a stationary point. Notice that the first four equations in (22) require $\alpha^{(h)}$ and $\beta^{(h)}$ to each have the first and second pairs of coordinates identical, for $h = 1,2$. The equation (22) defines a 2-dimensional surface in Θ. Using SINGULAR, we can verify that, holding, for example, α_{11} and β_{11} fixed, determines all of the other parameters according to the equations

$$\begin{cases} \lambda_1 = \frac{1}{80\alpha_{11}\beta_{11} - 20\alpha_{11} - 20*\beta_{11} + 6} \\ \lambda_2 = 1 - \lambda_1 \\ \alpha_{21} = \alpha_{11} \\ \alpha_{31} = \alpha_{41} = 0.5 - \alpha_{11} \\ \alpha_{12} = \alpha_{22} = \frac{10\beta_{11} - 3}{10(4\beta_{11} - 1)} \\ \ \alpha_{32} = \alpha_{42} = 0.5 - \alpha_{12} \\ \beta_{21} = \beta_{11} \\ \beta_{31} = \beta_{41} = 0.5 - \beta_{11} \\ \beta_{12} = \beta_{22} = \frac{10\alpha_{11} - 3}{10(4\alpha_{11} - 1)} \\ \beta_{32} = \beta_{42} = 0.5 - \beta_{12}. \end{cases}$$

Derivation of these equations can be found in the appendix B.3. Using the *elimination* technique (see Cox et al., 1996, Chapter 3) to remove all the variables in the system except for λ_1, we are left with one equation

$$80\lambda_1\alpha_{11}\beta_{11} - 20\lambda_1\alpha_{11} - 20\lambda_1\beta_{11} + 6\lambda_1 - 1 = 0. \tag{13}$$

Without the constraints for the coordinates of α_{11}, β_{11} and λ_1 to be probabilities, (13) defines a two-dimensional surface in $\mathbb{R}^3$, depicted in Figure 2. Notice that the axes do not intersect this surface, so that zero is not a possible value for α_{11}, β_{11} and λ_1. Because the non-identifiable space in Θ is 2-dimensional, equation (13) actually defines a bijection between α_{11}, β_{11} and λ_1 and the rest of the parameters. Then, the intersection of the surface (13) with the unit cube $[0,1]^3$, depicted as a red box in Figure 2, is the projection of the non-identifiable subspace into the 3-dimensional unit cube where α_{11}, β_{11} and λ_1 live. Figure 3 displays two different views of this projection. In the appendices B.3 and B.4, we show how to draw these figures using SINGULAR's graphics engine, the programme `SURF`.

The preceding arguments hold unchanged if we replace the symmetry conditions in the first two lines of equation (22) with either of these other two conditions, requiring different pairs of coordinates to be identical, namely

$$\alpha_{1h} = \alpha_{3h},\ \alpha_{2h} = \alpha_{4h},\ \ \beta_{1h} = \beta_{3h},\ \beta_{2h} = \beta_{4h} \tag{14}$$

and

$$\alpha_{1h} = \alpha_{4h},\ \alpha_{2h} = \alpha_{3h},\ \ \beta_{1h} = \beta_{4h},\ \beta_{2h} = \beta_{3h}, \tag{15}$$

where $h = 1, 2$.

The non-identifiable surfaces inside Θ corresponding each to one of the three pairs of coordinates held fixed in equations (22), (14) and (15), produce the three distinct tables of maximum likelihood estimates reported in Table 2 **a)**. Figure 3 shows the projection of the non-identifiable subspaces for the three MLEs in Table 2 **a)** into the three dimensional unit cube for λ_1, α_{11} and β_{11}. Although each of these three subspaces are disjoint subsets of Θ, their lower dimensional projections comes out as unique. By projecting onto the different coordinates

λ_1, α_{11} and β_{21} instead, we obtain two disjoint surfaces for the first, and second and third MLE, shown in Figure 4.

Table 3 presents some estimated parameters using the EM algorithm. Though these estimates are hardly meaningful, because of the non-identifiability issue, they show the symmetry properties we pointed out above and implicit in equations (22), (14) and (15), and they explain the invariance under simultaneous permutation of the fitted tables. In fact, the number of global maxima is the number of different configurations of the 4 dimensional vectors of estimated marginal probabilities with two identical coordinates, namely 3. This phenomenon, entirely due to the strong symmetry in the observed table (9), is completely separate from the non-identrifiability issues, but just as problematic.

By the same token, we can show that vectors of marginal probabilities with 3 identical coordinates also produce stationary points for the EM algorithms. This type of stationary points trace surfaces inside Θ which determine the local maxima of Table 2 **b)**. The number of these local maxima corresponds, in fact, to the number of possible configurations of 4-dimensional vectors with 3 identical coordinates, namely 4. Figure 5 depicts the lower dimensional projections into λ_1, α_{11} and β_{11} of the non-identifiable subspaces for the first MLE in Table 2 **a)**, the first three local maxima and the last local maxima in Table 2 **b)**.

We can summarize our finding as follows: the maxima in Table 2 define disjoint 2-dimensional surfaces inside the parameter space Θ, the projection of one of them being depicted in Figure 3. While non-identifiability is a structural feature of these models which is independent of the observed data, the multiplicity and invariance properties of the maximum likelihood estimates and the other local maxima is a phenomenon caused by the symmetry in the observed table of counts.

4.2.4 Plotting the Log-likelihood Function

Having determined that the non-identifiable space is 2-dimensional and that there are multiple maxima, we proceed with some plots of the profile log-likelihood function. To obtain a non-trivial surface, we need to consider three parameters. Figures 9 and 7 display the surface and contour plot respectively of the profile log-likelihhod function for α_{11} and α_{21} when α_{31} is one of the fixed parameters. Both Figures show clearly the different maxima, each lying on the top of "ridges" of the log-likelihood surface which are placed symmetrically with respect to each others. The position and shapes of these ridges reflect, once again, the symmetric properties of the estimated probabilities and parameters.

4.2.5 Further Remarks and Open Problem

We conclude this section with some observations and pointers to open problems.

One of the interesting aspects we came across while fitting the table (9) was the proximity of the values of the local and global maxima of the log-likelihood function. Furthermore, although these values are very close, the fitted tables

Table 3: Estimated parameters by the EM algorithm for the three global maxima in Table 2 **a)**.

Estimated Means	**Estimated Parameters**		
	$\widehat{\alpha}^{(1)} = \widehat{\beta}^{(1)}$	$\widehat{\alpha}^{(2)} = \widehat{\beta}^{(2)}$	$\widehat{\lambda}$
$\begin{pmatrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{pmatrix}$	$\begin{pmatrix} 0.3474 \\ 0.3474 \\ 0.1526 \\ 0.1526 \end{pmatrix}$	$\begin{pmatrix} 0.1217 \\ 0.1217 \\ 0.3783 \\ 0.3783 \end{pmatrix}$	$\begin{pmatrix} 0.5683 \\ 0.4317 \end{pmatrix}$
$\begin{pmatrix} 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \\ 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \end{pmatrix}$	$\begin{pmatrix} 0.3474 \\ 0.1526 \\ 0.3474 \\ 0.1526 \end{pmatrix}$	$\begin{pmatrix} 0.1217 \\ 0.3783 \\ 0.1217 \\ 0.3783 \end{pmatrix}$	$\begin{pmatrix} 0.5683 \\ 0.4317 \end{pmatrix}$
$\begin{pmatrix} 3 & 2 & 2 & 3 \\ 2 & 3 & 3 & 2 \\ 2 & 3 & 3 & 2 \\ 3 & 2 & 2 & 3 \end{pmatrix}$	$\begin{pmatrix} 0.3474 \\ 0.1526 \\ 0.1526 \\ 0.3474 \end{pmatrix}$	$\begin{pmatrix} 0.1217 \\ 0.3783 \\ 0.3783 \\ 0.1217 \end{pmatrix}$	$\begin{pmatrix} 0.5683 \\ 0.4317 \end{pmatrix}$

corresponding to global and local maxima are remarkably different. Even though the data (9) are not sparse, we wonder about the effect of cell sizes. Figure 8 show the same profile log-likelihood for the table (9) multiplied by 10000. While the number of global and local maxima, the contour plot and the basic symmetric shape of the profile log-likelihood surface remain unchanged after this rescaling, the peaks around the global maxima have become much more pronounced and so has the difference between of the values of the global and local maxima.

We have looked at a number of variations of table (9), focussing in particular on the symmetric data. We report only some of our results and refer to Zhou (2007) for a more extensive study. Table 4 shows the values and number of local and global maxima for a the 6×6 version of (9). As for the 4×4 case, we notice strong invariance features of the various maxima of the likelihood function and a very small difference between the value of the global and local maxima.

Fitting the same model to the table

$$\begin{pmatrix} 1 & 2 & 2 & 2 \\ 2 & 1 & 2 & 2 \\ 2 & 2 & 1 & 2 \\ 2 & 2 & 2 & 1 \end{pmatrix}$$

we found 6 global maxima of the likelihood function (their value was $-77.2927+$ *const.*), which return as many maximum likelihood estimates, all obtainable via simultaneous permutation of rows and columns of the table

$$\begin{pmatrix} 7/4 & 7/4 & 7/4 & 7/4 \\ 7/4 & 7/4 & 7/4 & 7/4 \\ 7/4 & 7/4 & 7/6 & 7/3 \\ 7/4 & 7/4 & 7/3 & 7/6 \end{pmatrix}.$$

Based on the various cases we have investigated, we have the following conjecture, which we verified computationally up to dimension $k = 50$:

Conjecture: The MLEs For the $n \times n$ table with values x along the diagonal and values $y \leq x$ for off the diagonal elements, the maximum likelihood estimates for the latent class model with 2 latent classes are the 2×2 block diagonal matrix of the form $\begin{pmatrix} A & B \\ B' & C \end{pmatrix}$ and the permutated versions of it, where A, B, and C are

$$\begin{aligned} A &= \left(y + \tfrac{x-y}{p}\right) \cdot \mathbf{1}_{p\times p}, \\ B &= y \cdot \mathbf{1}_{p\times q}, \\ C &= \left(y + \tfrac{x-y}{q}\right) \cdot \mathbf{1}_{q\times q}, \end{aligned}$$

and $p = \left\lfloor \frac{n}{2} \right\rfloor$, $q = n - p$.

We also noticed other interesting phenomena, which suggest the need for further geometric analysis. For example, consider fitting the (non-identifiable) latent class model with 2 levels to the table of counts (suggested by Bernd Sturmfels)

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 6 & 2 \\ 1 & 2 & 6 \end{pmatrix}.$$

Based on our computations, the maximum likelihood estimates appear to be unique, namely the table of fitted values

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 4 & 4 \end{pmatrix}. \tag{16}$$

Looking at the non-identifiable subspace for this model, we found that the MLEs (16) can arise from combinations of parameters some of which can be 0, such as

$$\alpha^{(1)} = \beta^{(1)} = \begin{pmatrix} 0.7143 \\ 0.1429 \\ 0.1429 \end{pmatrix}, \quad \alpha^{(2)} = \beta^{(2)} = \begin{pmatrix} 0 \\ 0.5 \\ 0.5 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 0.3920 \\ 0.6080 \end{pmatrix}.$$

This finding seems to indicate the possibility of singularities besides the obvious ones given by marginal probabilities for H containing 0 coordinates (which have the geometric interpretation as lower order secant varieties) and by points $\mathbf{p}$ along the boundary of the simplex Δ_{d-1}.

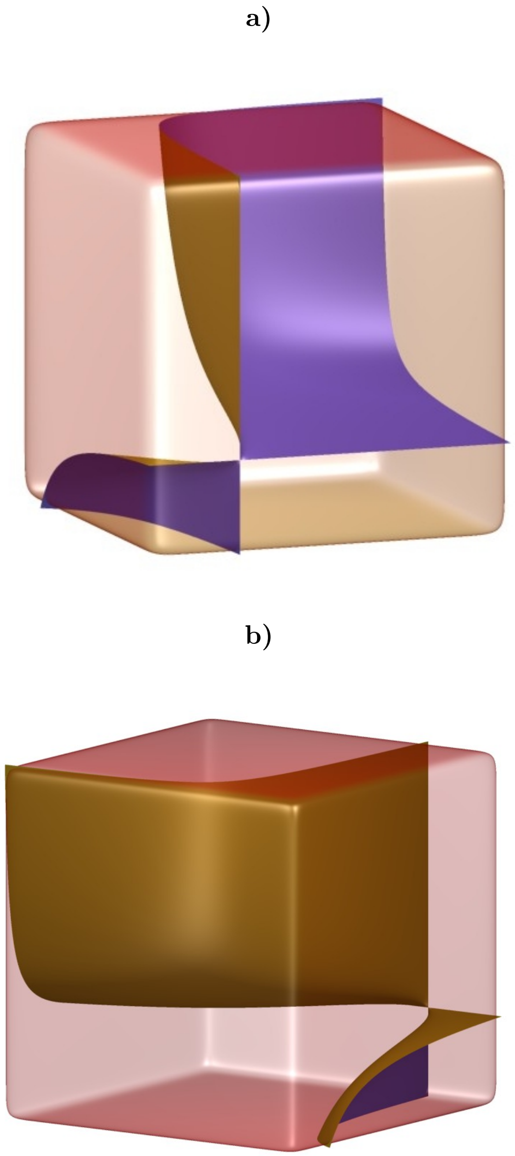


Figure 3: Intersection of the surface defined by equation (13) with the unit cube $[0,1]^3$, different views obtained using `surf` in **a)** and in **b)**.

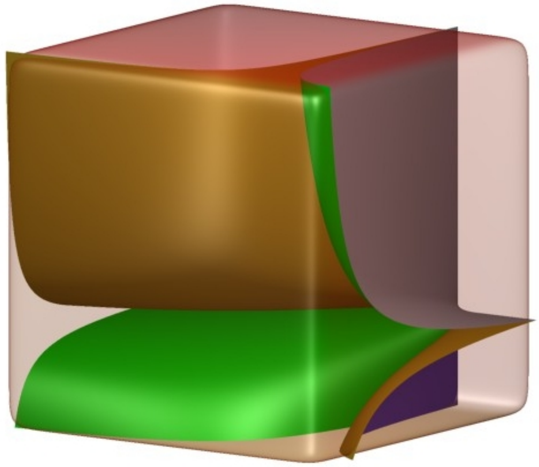


Figure 4: Projection of the non-identifiable subspaces corresponding to the first and second and third MLE from Table 2 **a)** into the 3-dimensional unit cube where λ_1, α_{11} and β_{21} take values.

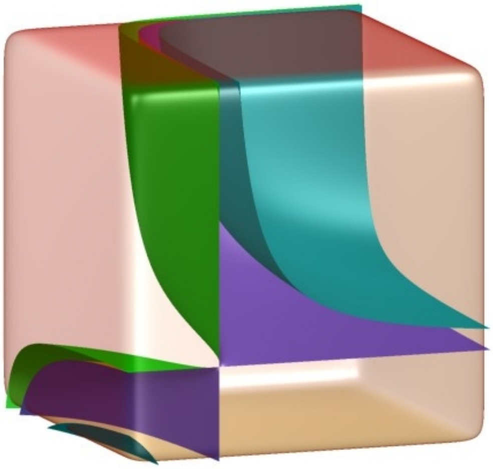


Figure 5: Projection of the non-identifiable subspaces corresponding to the first MLE in Table 2 **a)**, the first three local maxima and the last local maxima in Table 2 **b)** into the 3-dimensional unit cube where λ_1, α_{11} and β_{11} take values. In this coordinate system, the projection of non-identifiable subspaces for the first three local maxima in Table 2 **b)** results in the same surface; in order to obtain distinct surfaces, it would be necessary to change the coordinates over which the projections are made.

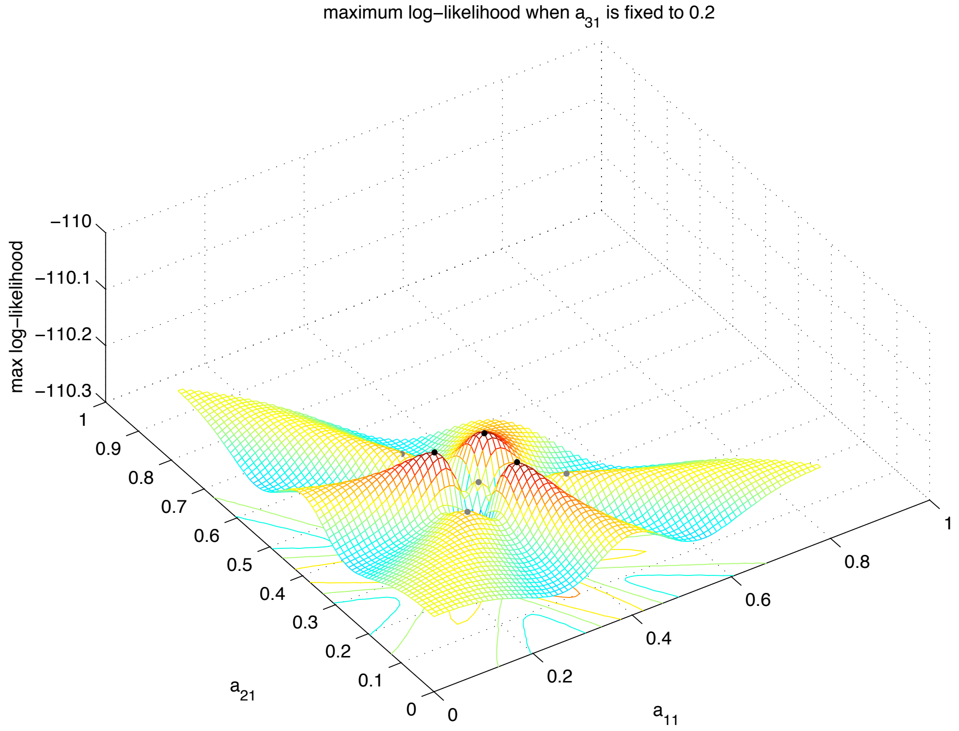


Figure 6: The plot of the profile likelihood as a function of α_{11} and α_{21} when α_{31} is fixed to 0.2. There are seven peaks: the three black points are the MLEs and the four gray diamonds are the other local maxima.

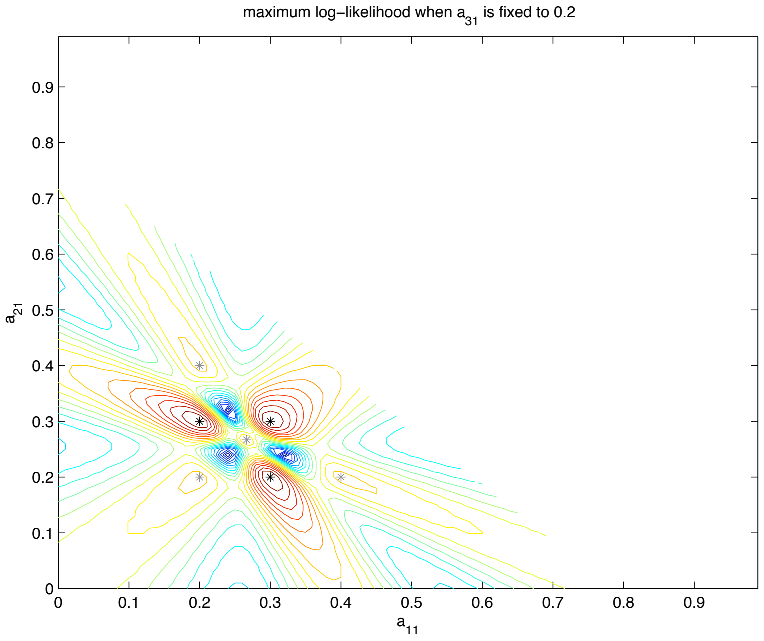


Figure 7: The contour plot of the profile likelihood as a function of α_{11} and α_{21} when α_{31} is fixed. There are seven peaks: the three black points are the MLEs and the four gray points are the other local maxima.

Table 4: Stationary points for the 6×6 version of the table (9). All the maxima are invariant under simultaneous permutations of the rows and columns of the corresponding fitted tables.

Fitted counts	Log-likelihood
$\begin{pmatrix} 4 & 2 & 2 & 2 & 2 & 2 \\ 2 & 12/5 & 12/5 & 12/5 & 12/5 & 12/5 \\ 2 & 12/5 & 12/5 & 12/5 & 12/5 & 12/5 \\ 2 & 12/5 & 12/5 & 12/5 & 12/5 & 12/5 \\ 2 & 12/5 & 12/5 & 12/5 & 12/5 & 12/5 \\ 2 & 12/5 & 12/5 & 12/5 & 12/5 & 12/5 \end{pmatrix}$	$-300.2524 + const.$
$\begin{pmatrix} 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 29/15 & 29/15 & 29/15 & 44/15 & 44/15 \\ 7/3 & 29/15 & 29/15 & 29/15 & 44/15 & 44/15 \end{pmatrix}$	$-300.1856 + const.$
$\begin{pmatrix} 3 & 3 & 2 & 2 & 2 & 2 \\ 3 & 3 & 2 & 2 & 2 & 2 \\ 2 & 2 & 5/2 & 5/2 & 5/2 & 5/2 \\ 2 & 2 & 5/2 & 5/2 & 5/2 & 5/2 \\ 2 & 2 & 5/2 & 5/2 & 5/2 & 5/2 \\ 2 & 2 & 5/2 & 5/2 & 5/2 & 5/2 \end{pmatrix}$	$-300.1729 + const.$
$\begin{pmatrix} 8/3 & 8/3 & 8/3 & 2 & 2 & 2 \\ 8/3 & 8/3 & 8/3 & 2 & 2 & 2 \\ 8/3 & 8/3 & 8/3 & 2 & 2 & 2 \\ 2 & 2 & 2 & 8/3 & 8/3 & 8/3 \\ 2 & 2 & 2 & 8/3 & 8/3 & 8/3 \\ 2 & 2 & 2 & 8/3 & 8/3 & 8/3 \end{pmatrix}$	$-300.1555 + const.$ (MLE)
$\begin{pmatrix} 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \end{pmatrix}$	$-301.0156 + const.$
$\begin{pmatrix} 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 35/9 & 35/18 & 35/18 & 35/18 & 35/18 \\ 7/3 & 35/18 & 175/72 & 175/72 & 175/72 & 175/72 \\ 7/3 & 35/18 & 175/72 & 175/72 & 175/72 & 175/72 \\ 7/3 & 35/18 & 175/72 & 175/72 & 175/72 & 175/72 \\ 7/3 & 35/18 & 175/72 & 175/72 & 175/72 & 175/72 \end{pmatrix}$	$-300.2554 + const.$

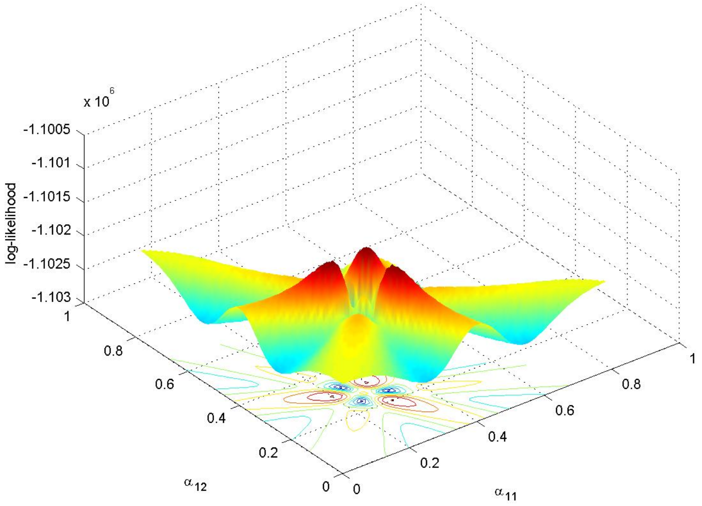


Figure 8: The contour plot of the profile likelihood as a function of α_{11} and α_{21} when α_{31} is fixed for the data (9) multiplied by 10000. As before, there are seven peaks: three global maxima and four identical local maxima.

5 Two Applications

5.1 Example: Michigan Influenza

Monto et al. (1985) present data for 263 individuals on the outbreak of influenza in Tecumseh, Michigan during the four winters of 1977-1981: (1) Influenza type A (H3N2), December 1977–March 1978; (2) Influenza type A (H1N1), January 1979–March 1979; (3) Influenza type B, January 1980–April 1980 and (4) Influenza type A (H3N2), December 1980–March 1981. The data have been analyzed by others including Haber (1986) and we reproduce them here as Table 5. This table is characterized by a large count for the cell corresponding to lack of infection from any type of influenza.

Table 5: Infection profiles and frequency of infection for four influenza outbreaks for a sample of 263 individuals in Tecumseh, Michigan during the winters of 1977-1981. A value of 0 in the first four columns codes the lack of infection. Source: Monto et al. (1985). The last column is the values fitted by the naive Bayes model with $r = 2$.

Type of Influenza				Observed Counts	Fitted Values
(1)	(2)	(3)	(4)		
0	0	0	0	140	139.5135
0	0	0	1	31	31.3213
0	0	1	0	16	16.6316
0	0	1	1	3	2.7168
0	1	0	0	17	17.1582
0	1	0	1	2	2.1122
0	1	1	0	5	5.1172
0	1	1	1	1	0.4292
1	0	0	0	20	20.8160
1	0	0	1	2	1.6975
1	0	1	0	9	7.7354
1	0	1	1	0	0.5679
1	1	0	0	12	11.5472
1	1	0	1	1	0.8341
1	1	1	0	4	4.4809
1	1	1	1	0	0.3209

The LC model with one binary latent variable (which is identifiable by Theorem 3.5 in Settimi and Smith, 2005) fits the data extremely well, as shown in Table 5. We also conducted a log-linear model analysis of this dataset and concluded that there is no indication of second or higher order interaction among the four types of influenza. The best log-linear model selected via both Pearson's chi-squared and the likelihood ratio statistic was the model of conditional independence of influenza of type (2), (3) and (4) given influenza of type (1) and was outperformed by the LC model.

Despite the reduced dimensionality of this problem and the large sample size, we report on the instability of the Fisher scoring algorithm implemented in the R package gllm, e.g., see Espeland (1986). As the algorithm cycles through, the evaluations of the expected Fisher information matrix become increasing ill-conditioned and eventually produce instabilities in the estimated coefficients and, in particular, in the standard errors. These problems disappear in the modified Newton-Raphson implementation, originally suggested by Haberman (1988), based on an inexact line search method known in the convex optimization literature as the Wolfe condition.

5.2 Data From the National Long Term Care Survey

Erosheva (2002) and Erosheva et al. (2007) analyze an extract from the National Long Term Care Survey in the form of a 2^{16} contingency table that contains data on 6 activities of daily living (ADL) and 10 instrumental activities of daily living (IADL) for community-dwelling elderly from 1982, 1984, 1989, and 1994 survey waves. The 6 ADL items include basic activities of hygiene and personal care (eating, getting in/out of bed, getting around inside, dressing, bathing, and getting to the bathroom or using toilet). The 10 IADL items include basic activities necessary to reside in the community (doing heavy housework, doing light housework, doing laundry, cooking, grocery shopping, getting about outside, travelling, managing money, taking medicine, and telephoning). Of the 65,536 cells in the table, 62,384 (95.19%) contain zero counts, 1,729 (2.64%)contain counts of 1, 499 (0.76%) contain counts of 2. The largest cell count, corresponding to the $(1, 1, \ldots, 1)$ cell, is 3,853.

Erosheva (2002) and Erosheva et al. (2007) use an individual-level latent mixture model that bears a striking resemblance to the LC model. Here we report on analyses with the latter.

We use both the EM and Newton-Raphson algorithms to fit a number of LC models with up to 20 classes, which can be shown to be all identifiable in virtue of Proposition 2.3 in Catalisano et al. (2002). Table 6 reports the maximal values of the log-likelihood function and the values of the BIC, which seem to indicate that larger LC models with many levels are to be preferred. To provide a better sense of how well these models fit the data, we show in Table 7 the fitted values for the six largest cells, which, as mentioned, deviate considerably from most of the cell entries. We have also considered alternative model selection criteria such as AIC and modifications of it. AIC (with and without a second order correction) points to $k > 20$! (An ad-hoc modification of AIC due to Anderson et al. (1994) for overdispersed data gives rather bizarre results.) The dimensionality of a suitable LC model for these data appears to be much greater than for the individual level mixture model in Erosheva et al. (2007).

Because of its high dimensionality and remarkable degree of sparsity, this example offers an ideal setting for testing the relative strengths and disadvantages of the EM and Newton-Raphson algorithms. In general, the EM algorithm, as a hill-climbing method, moves steadily towards solutions with higher values of

Table 6: BIC and log-likelihood values for various values of r for the NLTCS dataset.

r	Dimension	Maximal log-likelihood	BIC
2	33	-152527.32	305383.97
3	50	-141277.14	283053.25
4	67	-137464.19	275597
5	84	-135272.97	271384.21
6	101	-133643.77	268295.46
7	118	-132659.70	266496.96
8	135	-131767.71	264882.63
9	152	-131367.70	264252.25
10	169	-131033.79	263754.09
11	186	-130835.55	263527.24
12	203	-130546.33	263118.46
13	220	-130406.83	263009.09
14	237	-130173.98	262713.04
15	254	-129953.32	262441.37
16	271	-129858.83	262422.04
17	288	-129721.02	262316.06
18	305	-129563.98	262171.63
19	322	-129475.87	262165.07
20	339	-129413.69	262210.34

the log-likelihood, but converges only linearly. On the other hand, despite its faster quadratic rate of convergence, the Newton-Raphson method tends to be very time and space consuming when the number of variables is large, and may be numerically unstable if the Hessian matrices are poorly conditioned around critical points, which again occurs more frequently in large problems (but also in small ones, such as the Michigan Influenza examples above).

For the class of basic LC models considered in this paper, the time complexity for one single step of the EM algorithm is $\mathcal{O}\left(d \cdot r \cdot \sum_i d_i\right)$, while the space complexity is $\mathcal{O}\left(d \cdot r\right)$. In contrast, for the Newton-Raphson algorithm, both the time and space complexity are $\mathcal{O}\left(d \cdot r^2 \cdot \sum_i d_i\right)$. Consequently, for the NLTCS dataset, when r is bigger than 4, Newton-Raphson is sensibly slower than EM, and when r goes up to 7, Newton-Raphson needs more than 1G of memory. Another significant drawback of the Newton-Raphson method we experienced while fitting both the Michigan influenza and the NLTCS dataset is its potential numerical instability, due to the large condition numbers of the Hessian matrices. As remarked at the end of the previous section, following Haberman (1988), a numerically convenient solution is to modify the Hessian matrices so that they remain negative definite and then approximate locally the log-likelihood by a quadratic function. However, since the log-likelihood is neither concave nor quadratic, these modifications do not necessarily guarantee its values increases at each iteration step. As a result, the algorithm may ex-

Table 7: Fitted values for the largest six cells for the NLTCS dataset for various r.

r	Fitted values					
2	826.78	872.07	6.7	506.61	534.36	237.41
3	2760.93	1395.32	152.85	691.59	358.95	363.18
4	2839.46	1426.07	145.13	688.54	350.58	383.19
5	3303.09	1436.95	341.67	422.24	240.66	337.63
6	3585.98	1294.25	327.67	425.37	221.55	324.71
7	3659.80	1258.53	498.76	404.57	224.22	299.52
8	3663.02	1226.81	497.59	411.82	227.92	291.99
9	3671.29	1221.61	526.63	395.08	236.95	294.54
10	3665.49	1233.16	544.95	390.92	237.69	297.72
11	3659.20	1242.27	542.72	393.12	244.37	299.26
12	3764.62	1161.53	615.99	384.81	235.32	260.04
13	3801.73	1116.40	564.11	374.97	261.83	240.64
14	3796.38	1163.62	590.33	387.73	219.89	220.34
15	3831.09	1135.39	660.46	361.30	261.92	210.31
16	3813.80	1145.54	589.27	370.48	245.92	219.06
17	3816.45	1145.45	626.85	372.89	236.16	213.25
18	3799.62	1164.10	641.02	387.98	219.65	221.77
19	3822.68	1138.24	655.40	365.49	246.28	213.44
20	3836.01	1111.51	646.39	360.52	285.27	220.47
Observed	3853	1107	660	351	303	216

perience a considerable slowdown in the rate of convergence, which we in fact observed with the NLTCS data. Table 8 shows the condition numbers for the true Hessian matrices evaluated at the numerical maxima, for various values of r. This table suggests that, despite full identifiability, the log-likelihood has a very low curvature around the maxima and that the log-likelihood may, in fact, look quite flat. To further elucidate this point, we show in Figure 9 the profile log-likelihood plot for the parameter α_{12} in the simplest LC model with $r = 2$. The actual profile log-likelihood is shown in red and is obtained as the upper envelop of two distinct, smooth curves, each corresponding to local maxima of the log-likelihood. The location of the optimal value of α_{12} is displayed with a vertical line. Besides illustrating multimodality, the log-likelihood function in this example is notable for its relative flatness around its global maximum.

6 On Symmetric Tables and the MLE

In this section, inspired by the 100 Swiss Franks problem (9), we investigate in detail some of the effects that invariance to row and column permutations of the observed table have on the MLE. In particular, we study the seemingly simple problem of computing the MLE for the basic LC model when the observed table is square, symmetric and has dimension bigger than 3.

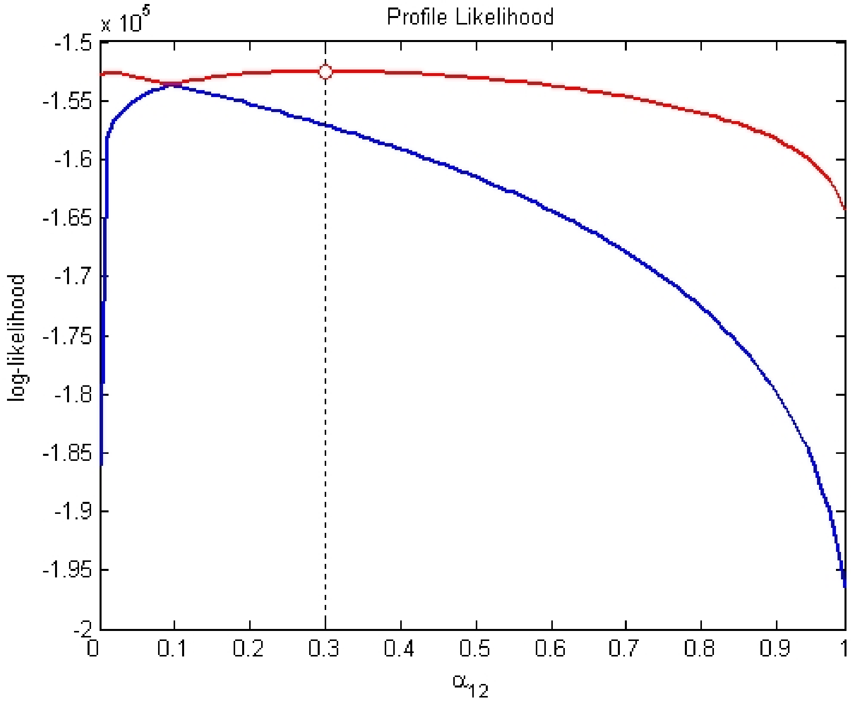


Figure 9: The plot of the profile likelihood for the NLCST dataset, as a function of α_{12}. The vertical line indicates the location of the maximizer.

Table 8: Condition numbers of Hessian matrices at the maxima for the NLTCS data.

r	Condition number
2	$2.1843e+03$
3	$1.9758e+04$
4	$2.1269e+04$
5	$4.1266e+04$
6	$1.1720e+08$
7	$2.1870e+08$
8	$4.2237e+08$
9	$8.7595e+08$
10	$8.5536e+07$
11	$1.2347e+19$
12	$3.9824e+08$
13	$1.0605e+20$
14	$3.4026e+18$
15	$3.9783e+20$
16	$3.2873e+09$
17	$1.0390e+19$
18	$2.1018e+09$
19	$2.0082e+09$
20	$2.5133e+16$

We show how symmetry in the data allows one to symmetrize, via averaging, local maxima of the likelihood function and to obtain critical points that are more symmetric. In various examples we looked at, these have larger likelihood than the tables from which they are obtained. We also prove that if the aforementioned averaging process always causes likelihood to go up, then among the 4×4 matrices of rank 2, the ones maximizing the log-likelihhod function for the 100 Swiss Franks problem (9) are given in Table 2 **a)**.

We will further simplify the notation and write L for the likelihood function, which can be expressed as

$$L(M) = \frac{\prod_{i,j} M_{i,j}^{n_{i,j}}}{(\sum_{i,j} M_{i,j})^{\sum_{i,j} n_{i,j}}}, \tag{17}$$

where $n_{i,j}$ is the count for the (i,j) cell and M is a square matrix with positive entries at which L is evaluated. The denominator is introduced as a matter of convenience to projectivize, i.e. ensuring that multiplying the entire matrix by a scalar will not change L.

6.1 Introduction and Motivation

A main theme in this section is to understand in what ways symmetry in data forces symmetry in the global maxima of the likelihood function. One question is whether our ideas can be extended at all to nonsymmetric data by suitable scaling. We prove that nonsymmetric local maxima will imply the existence of more symmetric points which are critical points at least within a key subspace and are related in a very explicit way to the nonsymmetric ones. Thus, if the EM algorithm leads to a local maximum which lacks certain symmetries, then one may deduce that certain other, more symmetric points are also critical points (at least within certain subspaces), and so check these to see if they give larger likelihood. There is numerical evidence that they do, and also a close look at our proofs shows that for "many" data points this symmetrization process is guaranteed to increase the value of the likelihood, by virtue of a certain single-variable polynomial encoding of the likelihood function often being real-rooted.

Here is an example of our symmetrization process. Given the data

$$\begin{matrix} 4 & 2 & 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 4 & 2 & 2 & 2 \\ 2 & 2 & 2 & 4 & 2 & 2 \\ 2 & 2 & 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 2 & 2 & 4 \end{matrix},$$

one of the critical points located by the EM algorithm is

$$\begin{matrix} 7/3 & 7/3 & 7/3 & 7/3 & 7/3 & 7/3 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 13/5 & 13/5 & 13/5 & 29/15 & 29/15 \\ 7/3 & 29/15 & 29/15 & 29/15 & 44/15 & 44/15 \\ 7/3 & 29/15 & 29/15 & 29/15 & 44/15 & 44/15 \end{matrix}.$$

One way to interpret this matrix is that $M_{i,j} = 7/3 + e_i f_j$ where

$$\mathbf{e} = \mathbf{f} = (\mathbf{0}, \mathbf{2}/\sqrt{\mathbf{15}}, \mathbf{2}/\sqrt{\mathbf{15}}, \mathbf{2}/\sqrt{\mathbf{15}}, -\mathbf{3}/\sqrt{\mathbf{15}}, -\mathbf{3}/\sqrt{\mathbf{15}}).$$

Our symmetrization process suggests replacing the vectors $\mathbf{e}$ and $\mathbf{f}$ each by the vector

$$(1/\sqrt{15}, 1/\sqrt{15}, 2/\sqrt{15}, 2/\sqrt{15}, -3/\sqrt{15}, -3/\sqrt{15})$$

in which two coordinates are averaged; however, since one of the values being averaged is zero, it is not so clear whether this should increase likelihood. However, repeatedly applying such symmetrization steps to this example, does converge to a local maximum. Now let us speak more generally. Let M be an n by n matrix of rank at most two which has row and column sums all equalling kn, implying (by results of Section 6.2) that we may write $M_{i,j}$ as $k + e_i f_j$ where e, f are each vectors whose coordinates sum to 0.

We are interested in the following general question:

Question 6.1 *Suppose a data matrix is fixed under simultaneously swapping rows and columns i, j. Consider any M as above, i.e. with $M_{i,j} = k + e_i f_j$.*

Does $e_i > e_j > 0, f_i > f_j > 0$ (or similarly $e_i < e_j < 0, f_i < f_j < 0$) imply that replacing e_i, e_j each by $\frac{e_i+e_j}{2}$ and f_i, f_j each by $\frac{f_i+f_j}{2}$ always increases the likelihood?

Remarks The weaker conditions $e_i > e_j = 0$ and $f_i > f_j = 0$ (resp. $e_i < e_j = 0, f_i < f_j = 0$) do not always imply that this replacement will increase likelihood. However, one may consider the finite list of possibilities for how many zeroes the vectors $\mathbf{e}$ and $\mathbf{f}$ may each have; an affirmative answer to Question 6.1 would give a way to find the matrix maximizing likelihood in each case, and then we could compare this finite list of maxima to find the global maximum.

Question 6.2 *Are all real-valued critical points of the likelihood function obtained by setting some number of coordinates in the $\mathbf{e}$ and $\mathbf{f}$ vectors to zero and then averaging by the above process so that the eventual vectors $\mathbf{e}$ and $\mathbf{f}$ have all positive coordinates equal to each other and all negative coordinates equal to each other? This seems to be true in many examples.*

One may check that the example discussed in Chapter 1 of Pachter and Sturmfels (2005) gives another instance where this averaging approach leads quickly to what appears to be a global maximum. Namely, given the data matrix

$$\begin{matrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{matrix}$$

and a particular starting point, the EM algorithm converges to the saddle point

$$\frac{1}{48}\begin{matrix} 4 & 2 & 3 & 3 \\ 2 & 4 & 3 & 3 \\ 3 & 3 & 3 & 3 \\ 3 & 3 & 3 & 3 \end{matrix},$$

whose entries may be written as $M_{i,j} = 1/48(3 + a_i b_j)$ for $\mathbf{a} = (-\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{0})$ and $\mathbf{b} = (-\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{0})$. Averaging -1 with 0 and 1 with the other 0 simultaneously in $\mathbf{a}$ and $\mathbf{b}$ immediately yields the global maximum directly by symmetrizing the saddle point, i.e. rather than finding it by running the EM algorithm repeatedly from various starting points.

An affirmative answer to Question 6.1 would imply several things. It would yield a (positive) solution to the 100 Swiss Franks problem, as discussed in Section 6.3. More generally, it would explain in a rather precise way how certain symmetries in data seem to impose symmetry on the global maxima of the maximum likelihood function. Moreover it would suggest good ways to look for global maxima, as well as constraining them enough that in some cases they can be characterized, as we demonstrate for the 100 Swiss Franks problem. To make this concrete, one thing it would tell us for an n by n data matrix which is fixed by the S_n action simultaneously permuting rows and columns in the same way, is that any probability matrix maximizing likelihood for such a data matrix will have at most two distinct types of rows.

We do not know the answer to this question, but we do prove that this type of averaging will at least give a critical point within the subspace in which e_i, e_j, f_i, f_j may vary freely but all other parameters are held fixed. Data also provide evidence that the answer to the question may very well be yes. At the very least, this type of averaging appears to be a good heuristic for seeking local maxima, or at least finding a way to continue to increase maximum likelihood beyond what it is at a critical point one reaches. Moreover, while real data are unlikely to have these symmetries, perhaps it could come close, and this could still be a good heuristic to use in conjunction with the EM algorithm.

6.2 Preservation of Marginals and Some Consequences

Proposition 6.1 *Given a two-way table in which all row and column sums (i.e. marginals) are equal, then for M to maximize the likelihood function among matrices of a fixed rank, the row and column sums of M must be equal.*

We prove the case mentioned in the abstract, which should generalize by adjusting exponents and ratios in the proof. It may very well also generalize to distinct marginals and tables with more rows and columns.

Proof Let R_1, R_2, R_3, R_4 be the row sums of M. Suppose $R_1 \geq R_2 \geq R_3 > R_4$; other cases will be similar. Choose δ so that $R_3 = (1+\delta)R_4$. We will show that multiplying row 4 by any $1+\epsilon$ with $0 < \epsilon < \min(1/4, \delta/2)$ will strictly increase L, giving a contradiction to M maximizing L. The result for column sums follows by symmetry.

Let us write $L(M')$ for the new matrix M' in terms of the variables $x_{i,j}$ for the original matrix M, so as to show that $L(M') > L(M)$. The first inequality below is proven in Lemma 6.1.

$$\begin{aligned}
L(M') &= \frac{(1+\epsilon)^{10}(\prod_{i=1}^4 x_{i,i})^4(\prod_{i\neq j} x_{i,j})^2}{R_1+R_2+R_3+(1+\epsilon)R_4)^{40}} \\
&> \frac{(1+\epsilon)^{10}(\prod_{i=1}^4 x_{i,i})^4(\prod_{i\neq j} x_{i,j})^2}{[(1+1/4(\epsilon-\epsilon^2))(R_1+R_2+R_3+R_4)]^{40}} \\
&= \frac{(1+\epsilon)^{10}(\prod_{i=1}^4 x_{i,i})^4(\prod_{i\neq j} x_{i,j})^2}{[(1+1/4(\epsilon-\epsilon^2))^4]^{10}[R_1+R_2+R_3+R_4]^{40}} \\
&= \frac{(1+\epsilon)^{10}(\prod_{i=1}^4 x_{i,i})^4(\prod_{i\neq j} x_{i,j})^2}{[1+4(1/4)(\epsilon-\epsilon^2)+6(1/4)^2(\epsilon-\epsilon^2)^2+\cdots+(1/4)^4(\epsilon-\epsilon^2)^4]^{10}[\sum_{i=1}^4 R_i]^{40}} \\
&\geq \frac{(1+\epsilon)^{10}}{(1+\epsilon)^{10}} \cdot L(M)
\end{aligned}$$

Lemma 6.1 *If $\epsilon < \min(1/4, \delta/2)$ and $R_1 \geq R_2 \geq R_3 = (1+\delta)R_4$, then $R_1 + R_2 + R_3 + (1+\epsilon)R_4 < (1 + 1/4(\epsilon - \epsilon^2))(R_1 + R_2 + R_3 + R_4)$.*

Proof It is equivalent to show $\epsilon R_4 < (1/4)(\epsilon)(1-\epsilon)\sum_{i=1}^4 R_i$. However,

$$\begin{aligned}
(1/4)(\epsilon)(1-\epsilon)(\sum_{i=1}^4 R_i) &\geq (3/4)(\epsilon)(1-\epsilon)(1+\delta)R_4 + (1/4)(\epsilon)(1-\epsilon)R_4 \\
&> (3/4)(\epsilon)(1-\epsilon)(1+2\epsilon)R_4 + (1/4)(\epsilon)(1-\epsilon)R_4 \\
&= (3/4)(\epsilon)(1+\epsilon-2\epsilon^2)R_4 + (1/4)(\epsilon-\epsilon^2)R_4 \\
&= \epsilon R_4 + [(3/4)(\epsilon^2) - (6/4)(\epsilon^3)]R_4 - (1/4)(\epsilon^2)R_4 \\
&= \epsilon R_4 + [(1/2)(\epsilon^2) - (3/2)(\epsilon^3)]R_4 \\
&\geq \epsilon R_4 + [(1/2)(\epsilon^2) - (3/2)(\epsilon^2)(1/4)]R_4 \\
&> \epsilon R_4.
\end{aligned}$$

Corollary 6.1 *There exist vectors (e_1, e_2, e_3, e_4) and (f_1, f_2, f_3, f_4) such that $\sum_{i=1}^4 e_i = \sum_{i=1}^4 f_i = 0$ and $M_{i,j} = K + e_i f_j$. Moreover, K equals the average entry size.*

In particular, this tells us that L may be maximized by treating it as a function of just six variables, namely $e_1, e_2, e_3, f_1, f_2, f_3$, since e_4, f_4 are also determined by these; changing K before solving this maximization problem simply has the impact of multiplying the entire matrix M that maximizes likelihood by a scalar.

Let E be the *deviation matrix* associated to M, where $E_{i,j} = e_i f_j$.

Question 6.3 *Another natural question to ask, in light of this corollary, is whether the matrix of rank at most r maximizing L is expressible as the sum of a rank one matrix and a matrix of rank at most $r-1$ that maximizes L among matrices of rank at most $r-1$.*

Remarks When we consider matrices with fixed row and column sums, then we may ignore the denominator in the likelihood function and simply maximize the numerator.

Corollary 6.2 *If M which maximizes L has $e_i = e_j$, then it also has $f_i = f_j$. Consequently, if it has $e_i \neq e_j$, then it also has $f_i \neq f_j$.*

Proof One consequence of having equal row and column sums is that it allows the likelihood function to be split into a product of four functions, one for each row, or else one for each column; this is because the sum of all table entries equals the sum of those in any row or column multiplied by four, allowing the denominator to be written just using variables from any one row or column. Thus, once the vector e is chosen, we find the best possible f for this given e by solving four separate maximization problems, one for each f_i, i.e. one for each column. Setting $e_i = e_j$ causes the likelihood function for column i to coincide with the likelihood function for column j, so both are maximized at the same value, implying $f_i = f_j$.

Next we prove a slightly stronger general fact for matrices in which rows and columns i, j may simultaneously be swapped without changing the data matrix:

Proposition 6.2 *If a matrix M maximizing likelihood has $e_i > e_j > 0$, then it also has $f_i > f_j > 0$.*

Proof Without loss of generality, say $i = 1, j = 3$. We will show that if $e_1 > e_3$ and $f_1 < f_3$, then swapping columns one and three will increase likelihood, yielding a contradiction. Let

$$L_1(e_1) = (1/4 + e_1 f_1)^4 (1/4 + e_1 f_2)^2 (1/4 + e_1 f_3)^2 (1/4 + e_1 f_4)^2$$

and

$$L_3(e_3) = (1/4 + e_2 f_1)^2 (1/4 + e_2 f_2)^2 (1/4 + e_3 f_3)^4 (1/4 + e_3 f_4)^2,$$

namely the contributions of rows 1 and 3 to the likelihood function. Let

$$K_1(e_1) = (1/4 + e_1 f_3)^4 (1/4 + e_1 f_2)^2 (1/4 + e_1 f_1)^2 (1/4 + e_1 f_4)^2$$

and

$$K_3(e_3) = (1/4 + e_3 f_3)^2 (1/4 + e_3 f_2)^2 (1/4 + e_3 f_1)^4 (1/4 + e_3 f_4)^2,$$

so that after swapping the first and third columns, the new contribution to the likelihood function from rows one and three is $K_1(e_1)K_3(e_3)$. Since the column swap does not impact that contributions from rows 2 and 4, the point is to show $K_1(e_1)K_3(e_3) > L_1(e_1)L_3(e_3)$. Ignoring common factors, this reduces to showing

$$(1/4 + e_1 f_3)^2 (1/4 + e_3 f_1)^2 > (1/4 + e_1 f_1)^2 (1/4 + e_3 f_3)^2,$$

in other words

$$(1/16 + 1/4(e_1 f_3 + e_3 f_1) + e_1 e_3 f_1 f_3)^2 > (1/16 + 1/4(e_1 f_1 + e_3 f_3) + e_1 e_3 f_1 f_3)^2,$$

namely $e_1 f_3 + e_3 f_1 > e_1 f_1 + e_3 f_3$. But since $e_3 < e_1, f_1 < f_3$, we have $0 < (e_1 - e_3)(f_3 - f_1) = (e_1 f_3 + e_3 f_1) - (e_1 f_1 + e_3 f_3)$, just as needed.

Question 6.4 *Does having a data matrix which is symmetric with respect to transpose imply that matrices maximizing likelihood will also be symmetric with respect to transpose?*

Perhaps this could also be verified again by averaging, similarly to what we suggest for involutions swapping a pair of rows and columns simultaneously.

6.3 The 100 Swiss Franks Problem

We use the results derived to far to show how to reduce the 100 Swiss Franks problem to Question 6.1. Thus, an affirmative answer to Question 6.1 would provide a mathematical proof formally that the three tables in 2 **a)** are global maxima of the log-likelihood function for the basic LC model with $r = 2$ and data given in (9).

Theorem 6.1 *If the answer to Question 6.1 is yes, then the 100 Swiss Franks problem is solved.*

Proof Proposition 6.1 showed that for M to maximize L, M must have row and column sums which are all equal to the quantity which we call $R_1, R_2, R_3, R_4, C_1, C_2, C_3,$ or C_4 at our convenience. The denominator of L may therefore be expressed as $(4C_1)^{10}(4C_2)^{10}(4C_3)^{10}(4C_4)^{10}$ or as $(4R_1)^{10}(4R_2)^{10}(4R_3)^{10}(4R_4)^{10}$, enabling us to rewrite L as a product of four smaller functions using distinct sets of variables.

Note that letting S_4 simultaneously permute rows and columns will not change L, so let us assume the first two rows of M are linearly independent. Moreover, we may choose the first two rows in such a way that the next two rows are each nonnegative combinations of the first two. Since row and column sums are all equal, the third row, denoted v_3, is expressible as $xv_1 + (1-x)v_2$ for v_1, v_2 the first and second rows and $x \in [0,1]$. One may check that M does not have any row or column with values all equal to each other, because if it had one, then it would have the other, reducing to a three by three problem which one may solve, and one may check that the answer does not have as high of likelihood as

$$\begin{matrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{matrix} .$$

Proposition 6.3 will show that if the answer to Question 6.1 is yes, then for M to maximize L, we must have $x = 0$ or $x = 1$, implying row 3 equals either row 1 or row 2, and likewise row 4 equals one of the first two rows. Proposition 6.4 shows M does not have three rows all equal to each other, and therefore must have two pairs of equal rows. Thus, the first column takes the form $(a, a, b, b)^T$, so it is simply a matter of optimizing a and b, then noting that the optimal choice will likewise optimize the other columns (by virtue of the way we broke L into a product of four expressions which are essentially the same, one for each column). Thus, M takes the form

$$\begin{matrix} a & a & b & b \\ a & a & b & b \\ b & b & a & a \\ b & b & a & a \end{matrix}$$

since this matrix does indeed have rank two. Proposition 6.5 shows that to maximize L one needs $2a = 3b$, finishing the proof.

Proposition 6.3 *If the answer to Question 6.1 is yes, then row 3 equals either row 1 or row 2 in any matrix M which maximizes likelihood. Similarly, each row i with $i > 2$ equals either row 1 or row 2.*

Proof $M_{3,3} = xM_{1,3}+(1-x)M_{2,3}$ for some $x \in [0,1]$, so $M_{3,3} \leq \max(M_{1,3}, M_{2,3})$. If $M_{1,3} = M_{2,3}$, then all entries of this column are equal, and one may use calculus to eliminate this possibility as follows: either M has rank one, and then we may replace column three by $(c, c, 2c, c)^T$ for suitable constant c to increase likelihood, since this only increases rank to at most two, or else the column space of M is spanned by $(1,1,1,1)^T$ and some (a_1, a_2, a_3, a_4) with $\sum a_i = 0$; specifically, column three equals $(1/4, 1/4, 1/4, 1/4) + x(a_1, a_2, a_3, a_4)$ for some x, allowing its contribution to the likelihood function to be expressed as a function of x whose derivative at $x = 0$ is nonzero, provided that $a_3 \neq 0$, implying that adding or subtracting some small multiple of $(a_1, a_2, a_3, a_4)^T$ to the column will make the likelihood increase. If $a_3 = 0$, then row three is also constant, i.e. $e_3 = f_3 = 0$. But then, an affirmative answer to the second part of Question 6.1 will imply that this matrix does not maximize likelihood.

Suppose, on the other hand, $M_{1,3} > M_{2,3}$. Our goal then is to show $x = 1$. By Proposition 6.1 applied to columns rather than rows, we know that $(1,1,1,1)$ is in the span of the rows, so each row may be written as $1/4(1,1,1,1) + cv$ for some fixed vector v whose coordinates sum to 0. Say row 1 equals $1/4(1,1,1,1)+kv$ for $k = 1$. Writing row three as $1/4(1,1,1,1)+lv$, what remains is to rule out the possibility $l < k$. However, Proposition 6.2 shows that $l < k$ and $a_1 < a_3$ together imply that swapping columns one and three will yield a new matrix of the same rank with larger likelihood.

Now we turn to the case of $l < k$ and $a_1 \geq a_3$. If $a_1 = a_3$ then swapping rows one and three will increase likelihood. Assume $a_1 > a_3$. By Corollary 6.1, we have (e_1, e_2, e_3, e_4) with $e_1 > e_3$ and (f_1, f_2, f_3, f_4) with $f_1 > f_3$. Therefore, if the answer to Question 6.1 is yes, then replacing e_1, e_3 each by $\frac{e_1+e_3}{2}$ and f_1, f_3 each by $\frac{f_1+f_3}{2}$ yields a matrix with larger likelihood, completing the proof.

Proposition 6.4 *In any matrix M maximizing L among rank 2 matrices, no three rows of M are equal to each other.*

Proof Without loss of generality, if M had three equal rows, then M would take the form

$$\begin{array}{cccc} a & c & e & g \\ b & d & f & h \\ b & d & f & h \\ b & d & f & h \end{array}$$

but then the fact that M maximizes L ensures $d = f = h$ and $c = e = g$ since L is a product of four expressions, one for each column, so that the second, third and fourth columns will all maximize their contribution to L in the same way. Since all row and column sums are equal, simple algebra may be used to show that all entries must be equal. However, we have already shown that such matrices do not maximize L.

Proposition 6.5 *To maximize M requires a, b related by $2a = 3b$.*

Proof We must maximize $\frac{a^6 b^4}{(8a+8b)^{10}}$. We may assume $a+b = 1$ since multiplying the entire matrix by a constant does not change L, so we maximize $(1/8)^{10} a^6 b^4$ with $b = 1 - a$; in other words, we maximize $f(a) = a^6(1-a)^4$. But solving $f'(a) = 0 = 6a^5(1-a)^4 + a^6(4)(1-a)^3(-1) = a^5(1-a)^3[6(1-a) - 4a]$ yields $6(1-a) - 4a = 0$, so $a = 6/10$ and $b = 4/10$ as desired.

7 Conclusions

In this paper we have reconsidered the classical latent class model for contingency table data and studied its geometric and statistical properties. We have exploited theoretical and computational tools from algebraic geometry to display the complexities of the latent class model. We have focused on the problem of maximum likelihood estimation and have studied the singularities arising from symmetries in the contingency table data and the multiple maxima that appear to result from these. We have given an informal characterization of this problem, but a strict mathematical proof of the existence of identical multiple maxima has eluded us; we describe elements of a proof in a separate section.

We have also applied LC models to synthetic data and to data arising from two real-life applications. In one, the model is quite simple and maximum likelihood estimation poses little problems, whereas in the other high-dimensional example various issues, computational as well as model-based, arise. From the computational standpoint, both the EM and the Newton-Raphson algorithm are especially vulnerable to problems of multimodality and provide little in the way of clues regarding the dimensionality difficulties associated with the underlying structure of LC models. Furthermore, the seemingly singular behavior of the Fisher information matrix at the MLE that we observe even for well-behaved, identifiable models is an additional element of complexity.

Based on our work, we would advise practitioners to exercise caution in applying LC models, especially to sparse data. They have a tremendous heuristic appeal and in some examples provide a clear and convincing description of the data. But in many situations, the kind of complex behavior explored in this paper may lead to erroneous inferences.

8 Acknowledgments

This research was supported in part by the National Institutes of Health under Grant No. R01 AG023141-01, by NSF Grant DMS-0631589, and by a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program, all to the Department of Statistics to Carnegie Mellon University, and by NSF Grant DMS-0439734 to the Institute for Mathematics and Its Application at the University of Minnesota. We thank Bernd Sturmfels for introducing us to the 100 Swiss Franks problem, which motivated much of this work, and for his valuable comments and feedback.

A Algebraic Geometry

A.1 Polynomial Ring, Ideal and Variety

In this section, we review some basic concepts and definitions in algebraic geometry and we draw connections between algebraic geometry and statistics. We begin with some concepts in abstract algebra. In mathematics, a *ring* is an algebraic structure in which addition and multiplication are defined and have some properties.

Definition A.1 (Ring) *A ring is a set $\mathcal{R}$ equipped with two binary operations $+ : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ and $\cdot : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$, called addition and multiplication, such that:*

- $(\mathcal{R}, +)$ *is an abelian group with identity element* 0*, so that* $\forall a, b, c \in \mathcal{R}$*, the following axiom hold:*
  - $a + b \in \mathcal{R}$
  - $(a + b) + c = a + (b + c)$
  - $0 + a = a + 0 = a$
  - $a + b = b + a$
  - $\exists - a \in \mathcal{R}$ *such that* $a + (-a) = (-a) + a = 0$
- $(\mathcal{R}, \cdot)$ *is a monoid with identity element* 1*, so that* $\forall a, b, c \in \mathcal{R}$*, the following axioms hold:*
  - $a \cdot b \in \mathcal{R}$
  - $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
  - $1 \cdot a = a \cdot 1 = a$
- *Multiplication distributes over addition:*
  - $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$
  - $(a + b) \cdot c = (a \cdot c) + (b \cdot c)$

The set of integer numbers $\mathbb{Z}$, the set of real numbers $\mathbb{R}$, and the set of rational numbers $\mathbb{Q}$ all are rings with the common addition and multiplication defined for numbers. Algebraic geometry is interested in polymonials and hence the polymonial rings. A *polynomial ring* is the set of polynomials in one or more unknowns with coefficients in a ring, for example, the set of polynomials with one variable in real numbers $\mathbb{R}[x]$ or the set of polynomials with two variables in rational numbers $\mathbb{Q}[x, y]$.

An *ideal* is a special subset of a ring. The ideal concept generalizes in an appropriate way some important properties of integers like “even number” or “multiple of 3”.

Definition A.2 (Ideal, generating set) *An ideal $\mathcal{I}$ is a subset of a ring $\mathcal{R}$ satisfying:*

- *$f + g \in \mathcal{I}$ if $f \in \mathcal{I}$ and $g \in \mathcal{I}$, and*
- *$pf \in \mathcal{I}$ if $f \in \mathcal{I}$ and $p \in \mathcal{R}$ is an arbitrary element.*

In other words, an ideal is a subset of a ring which is closed under addition and multiplication by elements of the ring. Let $\mathcal{I} = \langle \mathcal{A} \rangle$ denote the ideal $\mathcal{I}$ generated by the set $\mathcal{A}$, this means any $f \in \mathcal{I}$ is of the form $f = a_1 r_1 + \cdots + a_n r_n$ where each $a_i \in \mathcal{A}$ and $r_i \in \mathcal{R}$. If $\mathcal{A}$ is finite then $\mathcal{I}$ is a finitely generated ideal and if $\mathcal{A}$ is a singleton then $\mathcal{I}$ is called a principal ideal.

From now on, we only talk about the polynomial rings and ideals in the polynomial rings. For an ideal, we can consider the *generating set* of the ideal and a particular kind of generating set is called *Groebner basis*. Roughly speaking, a polynomial f is in the ideal if and only if the reminder of f with respect with the Groebner basis is 0. But here, the division algorithm requires a certain type of *ordering* on the monomials. So Groebner basis is stated relative to some monomial order in the ring and different orders will result in different bases. Later, we will give some of examples on the Groebner basis.

The following terms and notation are present in the literature of Groebner basis and will be useful later on.

Definition A.3 (degree, leading term, leading coefficient, power product) *A power product is a product of indeterminants $\left\{ x_1^{\beta_1} \cdots x_n^{\beta_n} : \beta_i \in \mathbb{N}, 1 \leq i \leq n \right\}$. The degree of a term of polynomial f is the sum of exponents of the term's power product. The degree of a polynomial f, denoted $\deg(f)$, is the greatest degree of terms in f. The leading term of f, denoted $\mathrm{lt}(f)$, is the term with the greatest degree. The leading coefficient of f is the coefficient of the leading term in f while the power product of the leading term is the leading power product, denoted $\mathrm{lp}(f)$.*

But sometimes there are many terms in the polynomial which all have the greatest degree, therefore to make the *leading term* well-defined, we need a well-defined *term order*. Below is one kind of term ordering.

Definition A.4 (Degree Reverse Lexicographic Ordering) *Let $x > y > z$ be a lex ordering and $\mathbf{u}^\alpha = x^{\alpha_1} y^{\alpha_2} z^{\alpha_3}$. Then $\mathbf{u}^\alpha < \mathbf{u}^\beta$ if and only if one of the following is true:*

- $\alpha_1 + \alpha_2 + \alpha_3 < \beta_1 + \beta_2 + \beta_3$
- *$\alpha_1 + \alpha_2 + \alpha_3 = \beta_1 + \beta_2 + \beta_3$ and the first coordinates α_i and β_i from the right which are different satisfy $\alpha_i > \beta_i$.*

For example, consider the polynomial $f = x^3 z - 2x^2 y^2 + 5y^2 z^2 - 7yz$. Then the degree reverse lexicographic ordering produces $x^2 y^2 > x^3 z > y^2 z^2 > yz$. So the leading term of f is $\mathrm{lt}(f) = -2x^2 y^2$ and the leading power product is $\mathrm{lp}(f) = x^2 y^2$. Now we can introduce the definition of *Groebner basis*.

Definition A.5 (Groebner basis) *A set of polynomials $\mathcal{G}$ contained in an ideal $\mathcal{I}$ is called a Groebner basis for $\mathcal{I}$ if the leading term of any polynomial in $\mathcal{I}$ is divisible by some polynomial in $\mathcal{G}$.*

Equivalent definitions for Groebner basis can be given according to the below theorem.

Theorem A.1 *Let $\mathcal{I}$ be an ideal and $\mathcal{G}$ be a set contained in $\mathcal{I}$. Then the following statements are equivalent:*

(a) $\mathcal{G}$ is a Groebner basis of $\mathcal{I}$.

(b) The ideal given by the leading terms of polynomials in $\mathcal{I}$ is itself generated by the leading terms of $\mathcal{G}$.

(c) The reminder of the division of any polynomial in the ideal $\mathcal{I}$ by $\mathcal{G}$ is 0.

(d) The reminder of the division of any polynomial in the ring $\mathcal{R}$ in which the ideal $\mathcal{I}$ is defined by $\mathcal{G}$ is unique.

Now that we can obtain a Groebner basis, we would like to obtain a simple and probably unique basis. The concept of *minimal Groebner basis* ensures the simplicity of the basis in some sense.

Definition A.6 (Minimal Groebner basis) *A Groebner basis $\mathcal{G}$ is minimal if for $\forall g \in \mathcal{G}$, the leading coefficient of g is 1 and for $\forall g_1 \neq g_2 \in \mathcal{G}$, the leading power product of g_1 does not divide the leading power product of g_2.*

A minimal Groebner basis has the least number of polynomials among the Groebner bases. But a minimal Groebner basis is not unique. For example if our basis is $\{y^2+yx+x^2, y+x, y, x^2, x\}$ for the ideal $\{y^2+yx+x^2, y+x, y\}$ with the lex $y > x$ term order then both $\{y, x\}$ and $\{y+x, x\}$ are minimal Groebner bases. To obtain a unique Groebner basis, we need to put further restrictions on the basis.

Definition A.7 (Reduced Groebner basis) *A Groebner basis is reduced if for $g \in \mathcal{G}$ the leading coefficient of g is 1 and g is reduced with respect to other polynomials in $\mathcal{G}$.*

By the definition, in our previous example $\{y, x\}$ is a reduced Groebner basis. Every non-zero ideal $\mathcal{I}$ has a unique reduced Groebner basis with respect to a fixed term order. In algebraic geometry, Buchberger's algorithm is the most commonly used algorithm for compute the Groebner basis and it can be viewed as a generalization of the Euclidean algorithm for univariate Greatest Common Divisor computation and of Gaussian elimination for linear systems. The basic version of Buchberger's algorithm does not guarantee the resulting basis to be minimal and reduced, but there are many variants of the basic algorithm to produce a minimal or reduced basis.

Now let's talk about the varieties. A variety is indeed a hyper-surface or a manifold in the enveloping space where it is defined. It is essentially a finite or infinite set of points where a polynomial in one or more variables attains, or a set of such polynomials all attain, a value of zero. The ideal arising from a variety is just the set of all polynomials attaining zero on the variety. For example, the surface of independence for the 2 × 2 table is a variety, and the ideal of this variety is generated by the set $\{p_{11}p_{22} - p_{12}p_{21}\}$ (Groebner basis). As a geometric object, we can consider the dimension of a variety. The dimension of a variety and the dimension of its ideal is the same thing, as the ideal dimension is the dimension of the intersection of its projective topological closure with the infinite hyperplane. And we will show later the way we compute the dimension of a variety is by computing the dimension of the ideal arising from it. The dimension of a variety may be less than the dimension of its enveloping space. Again, take the surface of independence as an example. The dimension of this variety is 2 while the dimension of the enveloping space, the probability simplex, is 3.

Definition A.8 (Variety) *A variety is zero sets of systems of polynomial equations in several unknowns.*

Definition A.9 (Ideal of variety) *The ideal of an variety is the set of polynomial vanishing on the variety.*

Algebraic geometry studies the polynomials and varieties. And the models we are working on, the traditional log-linear models and the latent class models, are all stated with polynomials! That's why concepts in statistics and concepts in algebraic geometry connects with each other. For example, in Pachter and Sturmfels (2005), Pachter and Sturmfels drawed the connections between some basic concepts of statistics and algebraic geometry, and we summarized them in table 9.

Statistics		**Algebraic Geometry**
independence	=	Segre variety
log-linear model	=	toric variety
curved exponential family	=	manifold
mixture model	=	joint of varieties
MAP estimation	=	tropicalization
……	=	……

Table 9: A glimpse of the statistics - algeobraic geometry dictionary.

Algebraic geometry views statistical models as varities, for example, the model of independence is related to the surface of independence. And here we like to refer to another figure in Pachter and Sturmfels (2005), which we show here in figure 10, to illustrate the connection between models and varieties. The model of interest here corresponds to the polynomial mapping f and the image of f which is a variety in the probability simplex. The observed data is a point in the probability simplex. Thus, maximum likelihood estimation is to find a point $\hat{p}$ in the image of the mapping f, which maps back to $\hat{\theta}$ in the parameter space, *closest to* the observed data point.

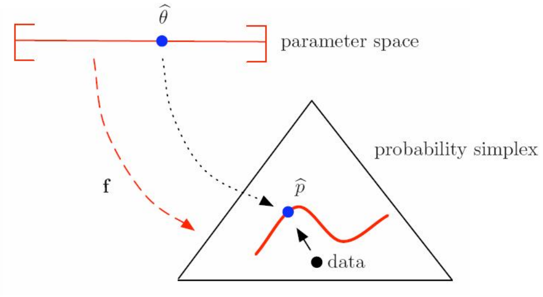


Figure 10: The geometry of maximum likelihood estimation.

In table 9, we can see that specific models are corresonded to specific varieties. Here we want to talk more about the Segre variety and the secant variety because they are related to the log-linear models and the latent class models.

A.2 Segre Variety and Secant Variety

Let's begin by setting up the basic notations and concepts. let $\mathbb{R}^{n+1}$ be a $(n+1)$-dimensional vector space on the real field. Then the n-dimensional *projective space* $\mathbb{P}^n = \mathbb{P}(\mathbb{R}^{n+1})$ of $\mathbb{R}^{n+1}$ is a set of elements constructed from $\mathbb{R}^{n+1}$ such that a distinct element of the projective space consists of all non-zero vectors which are equal up to a multiplication by a non-zero scalar. The projective space $\mathbb{P}^n$ is isomorphic to the n-dimensional simplex.

Definition A.10 (Segre map) *The Segre map σ is a map from the product space of two projective space $\mathbb{P}^n \times \mathbb{P}^m$ to a higher dimensional projective space $\mathbb{P}^{(n+1)(m+1)-1}$, such that for $\forall \mathbf{x} = (x_0, x_1, \ldots, x_n) \in \mathbb{P}^n$, $\forall \mathbf{y} = (y_0, y_1, \ldots, y_m) \in \mathbb{P}^m$,*

$$\sigma : \quad (\mathbf{x},\ \mathbf{y}) \mapsto \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} y_0, & y_1, & \cdots, & y_m \end{pmatrix}$$

The *Segre varieties* are the varieties $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_t}$ embedded in $\mathbb{P}^N$, $N = \prod(n_i + 1) - 1$, by Segre mapping, and the Segre embedding is based on the canonical multilinear map:

$$\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_t} \to \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_t}$$

where $\otimes$ is the *tensor product*, a.k.a. outer product. Now we denote the enveloping space $\mathbb{P}(\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_t})$ by $\mathbb{P}^N$ and denote the embedded Segre variety $\mathbb{P}^{n_1} \otimes \cdots \otimes \mathbb{P}^{n_t}$ as $\mathbb{X}_n$. Then, with this point of view:

- the Segre variety $\mathbb{X}_n$ is the set of all classes of *decomposable tensors*, i.e. classes of tensors (i.e. multi-dimensional arrays) in $\mathbb{P}(\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_t})$ of the form $v_1 \otimes \cdots \otimes v_t$.

- the *secant variety*, $Sec_r(\mathbb{X}_n)$, is the closure of the set of classes of those tensors which can be written as the sum of $\leq r+1$ decomposable tensors.

Now let's consider the 2-dimensional tensors, which are actually matrices. In such case, $\mathbb{P}^{n_1}$ is the set of (n_1+1)-dimensional vectors, $\mathbb{P}^{n_2}$ is the set of (n_2+1)-dimensional vectors, and $\mathbb{P}^N$ is the set of $(n_1+1) \times (n_2+1)$ matrices, all under the projective equivalence. Then, the Segre variety $\mathbb{P}^{n_1} \otimes \mathbb{P}^{n_2}$ consists of all the rank 1 matrices in $\mathbb{P}^N$. And the r-secant variety $Sec_r(\mathbb{P}^{n_1} \otimes \mathbb{P}^{n_2})$ is the set of matrices having rank $\leq r+1$ because a matrix has rank $\leq r+1$ if and only if it is a sum of $\leq r+1$ matrices of rank 1.

For example, consider the embedding of $\mathbb{P}^2 \otimes \mathbb{P}^2$ in $\mathbb{P}^8$, where $\mathbb{P}^8$ is the projective space of 3×3 matrices under projective equivalence. The ideal of 2×2 minors of the generic matrix of size 3×3 defines $\mathbb{P}^2 \otimes \mathbb{P}^2$ and the determinant of the generic matrix gives the equation of $Sec_1(\mathbb{P}^2 \otimes \mathbb{P}^2)$. *The Segre variety $\mathbb{P}^2 \otimes \mathbb{P}^2$ corresponds to the no 2nd-effect log-linear model for the 3×3 table and the secant variety $Sec_1(\mathbb{P}^1 \otimes \mathbb{P}^2)$ corresponds to the 2-level latent class model for the 3×3 table.*

Back to the former notations, we have $\mathbb{X}_n = \mathbb{P}^{n_1} \otimes \cdots \otimes \mathbb{P}^{n_t}$. What is the dimension of the secant variety $Sec_r(\mathbb{X}_n)$? There is an *expected dimension* by counting parameters:

$$\min\{N, (r+1)\prod_i n_i + r\}$$

which is only an upper bound of the actual dimension of $Sec_r(\mathbb{X}_n)$. If the actual dimension is different from the expected dimension, the secant variety is *deficient*. Computing the dimension of secant varieties has been a challenge problem in algebraic geometry. We summarize some results in the following theorems.

For the case of two factors, we have a complete answer for the actual dimension of the secant variety.

Theorem A.2 (Proposition 2.3 in Catalisano etc.'s Catalisano et al. (2002)) *For the case of two factors, for all r, $1 \leq r < \min(n_1, n_2)$ the secant varieties $Sec_r(\mathbb{X}_n)$ all have dimension less than the expected dimension. Moreover, the least integer for which $Sec_r(\mathbb{X}_n)$ fills its enveloping space is $r = n_1$.*

When it comes to the case of three factors, the dimension of the secant variety is still an open problem in general. But for some special varieties, there are beautiful results. The below two theorems are for $n = (n_1, n_2, n_3)$.

Theorem A.3 (Proposition 2.3 in Catalisano etc.'s Catalisano et al. (2002)) *If $n = (n_1, n_2, n_3)$ and $r \leq min(n_1, n_2, n_3)$, then $Sec_r(\mathbb{X}_n)$ has the expected dimension.*

As a direct proposition from theorem A.3, we have a complete answer for 2-level latent class model for 3×3 tables.

Theorem A.4 *When $n = (n_1, n_2, n_3)$, the secant line variety for any Segre variety has the expected dimension.*

Remarks Theorem A.3 and A.4 says that 2-level and "small" latent class models for 3×3 tables have the dimension

$$\min\{(n_1+1)(n_2+1)(n_3+1)-1, (r+1)(n_1+n_2+n_3)+r\}$$

Note that the first term is the free dimension of the observed table and the second term is the dimension of underlining parameter space. And obviously, theorem A.4 can be directly applied to our conjecture about $2{\times}2{\times}K$ models.

For more factors, the dimension of some special varieties can still be derived.

Theorem A.5 (Proposition 3.7 in Catalisano etc.'s Catalisano et al. (2002)) *Let $n = (n_1, \dots, n_t)$ and let $t \geq 3$, $n_1 \leq n_2 \leq \cdots \leq n_t$,*

$$\left\lceil \frac{n_1+n_2+\cdots+n_t+1}{2} \right\rceil \geq \max(n_t+1, r+1)$$

Then $\dim Sec_r(\mathbb{X}_n) = (r+1)(n_1+n_2+\cdots+n_t)+r$.

Another result concerning about higher secant varieties is from coding theory when the dimensions of the Segre varieties are equal, that is, $n_1 = n_2 = \cdots = n_t = q-1$.

Theorem A.6 (Example 2.4 in Catalisano etc.'s Catalisano et al. (2002))

(i) Let k be any positive integer, $q = 2$, $t = 2^k - 1$, $r = 2^{t-k}$. For these numbers the Segre embedding

$$\mathbb{X}_t = \underbrace{\mathbb{P}^1 \times \cdots \times \mathbb{P}^1}_{t} \rightarrow \mathbb{P}^{2^t-1}$$

we have $Sec_{r-1}(\mathbb{X}_t) = \mathbb{P}^{2^t-1}$ and these secant varieties fit "exactly" into theor enveloping space.

(ii) We can make families of similar examples for products of $\mathbb{P}^2$, $\mathbb{P}^3$, $\mathbb{P}^4$, $\mathbb{P}^7$, $\mathbb{P}^8$, ..., $\mathbb{P}^{q-1}$ where q is a prime power. Given such a q, for any integer $k \geq 1$ we take $t = (q^k-1)/(q-1)$ copies of $\mathbb{P}^{q-1}$, which gets embedded in $\mathbb{P}^{q^t-1}$. Then for $r = q^{t-k}$ we get

$$Sec_{r-1}(\underbrace{\mathbb{P}^{q-1} \times \cdots \times \mathbb{P}^{q-1}}_{t\text{-times}}) = \mathbb{P}^{q^t-1}$$

B Symbolic Software of Computational Algebra

Unlike many numerical softwares we use in machine learning, by which we get the answer for a particular set of values of the variables of interest, symbolic softwares provide us an algebraic answer for all possible values of the variables. The symbolic computation can fill up the machine very quickly. So current symbolic softwares can only deal with limited-scale problems. Here we use some examples to show some symbolic computation relevant to the problems we have been discussed so far. We have been using various symbolic softwares for different purposes and here we will talk about the software Singular because it is the software we need to do the computations related to our problems in this paper.

B.1 Computing the dimension of the image variety

Let's take the $2 \times 2 \times 3$ table with 2 latent classes as an example, to see how to compute the dimension of the image variety defined by the polynomial mapping f:

$$\begin{array}{rrcl} f: & \Delta_1 \times \Delta_1 \times \Delta_1 \times \Delta_2 & \to & \Delta_{11} \\ & (a_t, x_{it}, y_{jt}, z_{kt}) & \mapsto & p_{ijk} = \sum_t a_t x_{it} y_{jt} z_{kt} \end{array}$$

where Δ_n is the n-dimensional probability simplex. The first step is to get the ideal arising from the model that is only defined on the probabilities $\{p_{ijk}\}$. In Singluar, we define a polynomial ring r on the unknowns p_{ijk} which stand for cell probabilities and the unknowns $a_t, x_{it}, y_{jt}, z_{kt}$ which stand for the conditional probabilities. The ideal I on the ring r is defined by the model equalities (the first 12 polynomials) and sum 1 constraints of the probabilties (the last 7 polynomials).

```
ring r=0, (a1,x11,x21,y11,y21,z11,z21,z31,a2,x12,x22,
y12,y22,z12,z22,z32,p111,p112,p113,p121,p122,p123,p211,
p212,p213,p221,p222,p223), lp;
ideal I=p111-a1*x11*y11*z11-a2*x12*y12*z12,
p112-a1*x11*y11*z21-a2*x12*y12*z22,
p113-a1*x11*y11*z31-a2*x12*y12*z32,
p121-a1*x11*y21*z11-a2*x12*y22*z12,
p122-a1*x11*y21*z21-a2*x12*y22*z22,
p123-a1*x11*y21*z31-a2*x12*y22*z32,
p211-a1*x21*y11*z11-a2*x22*y12*z12,
p212-a1*x21*y11*z21-a2*x22*y12*z22,
p213-a1*x21*y11*z31-a2*x22*y12*z32,
p221-a1*x21*y21*z11-a2*x22*y22*z12,
p222-a1*x21*y21*z21-a2*x22*y22*z22,
p223-a1*x21*y21*z31-a2*x22*y22*z32,
a1+a2-1,
x11+x21-1,
x12+x22-1,
y11+y21-1,
y12+y22-1,
z11+z21+z31-1,
z12+z22+z32-1;
```

But the ideal I defined as above is on all the unknowns, including both the cell probabilities and the conditional probabilities. So the next step is to eliminate the unknowns $a_t, x_{it}, y_{jt}, z_{kt}$ and then to get the image variety where p_{ijk} lies. To use the elimination functions in SINGULAR, we need to include the library "ELIM.LIB".

```
LIB "elim.lib";
ideal J=elim1(I, a1*x11*x21*y11*y21*z11*z21*z31*a2*x12*x22
*y12*y22*z12*z22*z32);
J;
===>
J[1]=p121*p212*p223-p121*p213*p222-....;
J[2]=p112*p211*p223+p112*p212*p223-p112*p213*p221-....;
J[3]=p112*p121*p223+p112*p122*p223-p112*p123*p221-....;
J[4]=p112*p121*p213+p112*p121*p223+p112*p122*p213+....;
J[5]=p111+p112+p113+p121+p122+p123+p211+p212+p213+p221+p222+p223-1;
```

Now we can see the image variety is defined by five polynomials of ideal J. And the first four polynomials are right the determinants in equation 18 and the last one corresponds to the sum 1 constaint. We can also get the five polynomials by computing Groebner basis.

$$
\begin{gathered}
\begin{vmatrix} p_{121} & p_{211} & p_{221} \\ p_{122} & p_{212} & p_{222} \\ p_{123} & p_{213} & p_{223} \end{vmatrix} \\[1em]
\begin{vmatrix} p_{1+1} & p_{211} & p_{221} \\ p_{1+2} & p_{212} & p_{222} \\ p_{1+3} & p_{213} & p_{223} \end{vmatrix} \\[1em]
\begin{vmatrix} p_{+11} & p_{121} & p_{221} \\ p_{+12} & p_{122} & p_{222} \\ p_{+13} & p_{123} & p_{223} \end{vmatrix} \\[1em]
\begin{vmatrix} p_{111} & p_{121}+p_{211} & p_{221} \\ p_{112} & p_{122}+p_{212} & p_{222} \\ p_{113} & p_{123}+p_{213} & p_{223} \end{vmatrix}
\end{gathered}
\tag{18}
$$

```
ideal J=groebner(I);
```

Using the above command "GROEBNER", we will get an ideal J defined by 184 polynomials. Among them, the first five polynomials only involve the variable p_{ijk} and they are the five polynomials we have got before. When using the "GROEBNER" command, please be aware that the resulting basis is subject to the monomial ordering you choose for defining the ring.

To compute the dimension of the ideal, we need to define another ring r_1 only with unknowns p_{ijk} and then an ideal (which we also call J) defined by the above five polynomials. Note that the dimension of the ideal and the size of the Groebner basis for the ideal are different things.

```
ring r1=0, (p111,p112,p113,p121,p122,p123,p211,p212,p213,p221,p222,
p223), lp;
ideal J;
J[1]=p121*p212*p223-p121*p213*p222-....;
J[2]=p112*p211*p223+p112*p212*p223-p112*p213*p221-....;
J[3]=p112*p121*p223+p112*p122*p223-p112*p123*p221-....;
J[4]=p112*p121*p213+p112*p121*p223+p112*p122*p213+....;
J[5]=p111+p112+p113+p121+p122+p123+p211+p212+p213+p221+p222+p223-1;
dim(groebner(J));
===> 7
```

Table 10 lists the effective dimenions of some latent class models which have been considered so far. In Kocka and Zhang (2002), Kocha and Zhang have showed that the maximal numerical rank of the Jacobian of the polynomial mapping equals to the symbolic rank and the numerical rank reaches the maximal rank almost surely. Therefore, although it is impossible to compute the symbolic rank of the Jacobian or to compute the dimension of the image variety, we can calculate the numerical rank of the Jacobian at many points to find the possible maximal rank.

B.2 Solving Polynomial Equations

Singular can also be used to solve the polynomial equations. For example, in the 100 Swiss Franks Problem, we need to solve the optimization problem in equation 19.

$$\ell(\mathbf{p}) = \sum_{i,j} n_{ij} \log p_{ij}, \qquad \mathbf{p} \in \Delta_{15},\ \det(\mathbf{p}^*_{ij}) = 0 \text{ all } i,j \in [4], \tag{19}$$

where $\mathbf{p}^*_{ij}$ is the 3×3 sub-matrix of $\mathbf{p}$ obtained by erasing the ith row and the jth column. Using Lagrange multipliers method, the objective becomes finding all the local extrema of the below function $H(\cdot)$.

$$H(p_{ij}, h_0, h_{ij}) = \sum_{i,j} n_{ij} \log p_{ij} + h_0 \left(\sum_{i,j} p_{ij} - 1 \right) + h_{ij} \det \mathbf{p}^*_{ij} \tag{20}$$

Taking the derivative of $H(\cdot)$ with respect to p_{ij}, h_0 and h_{ij}, we get a system of 33 polynomial functions. In Singular, we can define the ideal generated by these 33 polynomials.

Latent class model		Effective dimension	
dim of table	num of latent class	dim of image variety	max numerical rank of Jacobi
2×2	$r = 2$	3	3
3×3	$r = 2$	7	7
4×5	$r = 3$	17	17
$2 \times 2 \times 2$	$r = 2$	7	7
$2 \times 2 \times 2$	$r = 3$	7	7
$2 \times 2 \times 2$	$r = 4$	7	7
$3 \times 3 \times 3$	$r = 2$	N/A	13
$3 \times 3 \times 3$	$r = 3$	N/A	20
$3 \times 3 \times 3$	$r = 4$	N/A	25
$3 \times 3 \times 3$	$r = 5$	N/A	26
$3 \times 3 \times 3$	$r = 6$	N/A	26
$5 \times 2 \times 2$	$r = 3$	N/A	17
$4 \times 2 \times 2$	$r = 3$	N/A	14
$3 \times 3 \times 2$	$r = 5$	N/A	17
$6 \times 3 \times 2$	$r = 5$	N/A	34
$10 \times 3 \times 2$	$r = 5$	N/A	54
$2 \times 2 \times 2 \times 2$	$r = 2$	N/A	9
$2 \times 2 \times 2 \times 2$	$r = 3$	N/A	13
$2 \times 2 \times 2 \times 2$	$r = 4$	N/A	15
$2 \times 2 \times 2 \times 2$	$r = 5$	N/A	15
$2 \times 2 \times 2 \times 2$	$r = 6$	N/A	15

Table 10: Effective dimensions of some latent class models. 'N/A' means it is computationally infeasible.

```
ring r=0, (p11,p21,p31,p41,p12,p22,p32,p42,p13,p23,p33,p43,p14,p24,p34,p44,
h11,h21,h31,h41,h12,h22,h32,h42,h13,h23,h33,h43,h14,h24,h34,h44,h0), lp;
ideal I=4+h0*p11+h23*p11*p32*p44-h23*p11*p34*p42+h24*p11*p32*p43 ...,
2+h0*p21+h13*p21*p32*p44-h13*p21*p34*p42+h14*p21*p32*p43 ...,
2+h0*p31-h13*p31*p22*p44+h13*p31*p24*p42-h14*p31*p22*p43 ...,
2+h0*p41+h13*p41*p22*p34-h13*p41*p24*p32+h14*p41*p22*p33 ...,
2+h0*p12-h23*p31*p12*p44+h23*p41*p12*p34-h24*p31*p12*p43 ...,
4+h0*p22-h13*p22*p31*p44+h13*p41*p22*p34-h14*p22*p31*p43 ...,
2+h0*p32+h13*p32*p21*p44-h13*p41*p24*p32+h14*p32*p21*p43 ...,
2+h0*p42-h13*p42*p21*p34+h13*p42*p31*p24-h14*p42*p21*p33 ...,
2+h0*p13+h24*p42*p31*p13-h24*p41*p13*p32-h21*p32*p13*p44 ...,
2+h0*p23+h14*p42*p31*p23-h14*p41*p23*p32-h11*p32*p23*p44 ...,
4+h0*p33-h14*p42*p21*p33+h14*p41*p22*p33+h11*p22*p33*p44 ...,
2+h0*p43+h14*p32*p21*p43-h14*p22*p31*p43-h11*p22*p34*p43 ...,
2+h0*p14+h23*p31*p14*p42-h23*p41*p14*p32+h21*p32*p14*p43 ...,
2+h0*p24+h13*p42*p31*p24-h13*p41*p24*p32+h11*p32*p24*p43 ...,
2+h0*p34-h13*p42*p21*p34+h13*p41*p22*p34-h11*p22*p34*p43 ...,
4+h0*p44+h13*p32*p21*p44-h13*p22*p31*p44+h11*p22*p33*p44 ...,
p22*p33*p44-p22*p34*p43-p32*p23*p44+p32*p24*p43+p42*p23*p34-p42*p24*p33,
p12*p33*p44-p12*p34*p43-p32*p13*p44+p32*p14*p43+p42*p13*p34-p42*p14*p33,
p12*p23*p44-p12*p24*p43-p22*p13*p44+p22*p14*p43+p42*p13*p24-p42*p14*p23,
p12*p23*p34-p12*p24*p33-p22*p13*p34+p22*p14*p33+p32*p13*p24-p32*p14*p23,
p21*p33*p44-p21*p34*p43-p31*p23*p44+p31*p24*p43+p41*p23*p34-p41*p24*p33,
p11*p33*p44-p11*p34*p43-p31*p13*p44+p31*p14*p43+p41*p13*p34-p41*p14*p33,
p11*p23*p44-p11*p24*p43-p21*p13*p44+p21*p14*p43+p41*p13*p24-p41*p14*p23,
p11*p23*p34-p11*p24*p33-p21*p13*p34+p21*p14*p33+p31*p13*p24-p31*p14*p23,
p21*p32*p44-p21*p34*p42-p31*p22*p44+p31*p24*p42+p41*p22*p34-p41*p24*p32,
p11*p32*p44-p11*p34*p42-p31*p12*p44+p31*p14*p42+p41*p12*p34-p41*p14*p32,
p11*p22*p44-p11*p24*p42-p21*p12*p44+p21*p14*p42+p41*p12*p24-p41*p14*p22,
p11*p22*p34-p11*p24*p32-p21*p12*p34+p21*p14*p32+p31*p12*p24-p31*p14*p22,
p21*p32*p43-p21*p33*p42-p31*p22*p43+p31*p23*p42+p41*p22*p33-p41*p23*p32,
p11*p32*p43-p11*p33*p42-p31*p12*p43+p31*p13*p42+p41*p12*p33-p41*p13*p32,
p11*p22*p43-p11*p23*p42-p21*p12*p43+p21*p13*p42+p41*p12*p23-p41*p13*p22,
p11*p22*p33-p11*p23*p32-p21*p12*p33+p21*p13*p32+p31*p12*p23-p31*p13*p22,
p11+p21+p31+p41+p12+p22+p32+p42+p13+p23+p33+p43+p14+p24+p34+p44-1;
```

By using the routine ’SOLVE’ in Singular we can find the numerical solutions to the system of polynomial equations.

```
LIB ’solve.lib’;
solve(I, 6, 0 , ’nodisplay’);
```

Unfortunately, the system we want to solve beyond what Singular can handle. But we can check whether a given table $\{p_{ij}\}$ is a solution to the system or not, by substituting the values of p_{ij} into the ideal I. And if the resulting ideal is not an empty set, then $\{p_{ij}\}$ is a solution to the system.

```
LIB "poly.lib"
ideal v=p11,p21,p31,p41,p12,p22,p32,p42,p13,p23,p33,p43,p14,p24,p34,p44;
ideal p=3/40,3/40,2/40,2/40,3/40,3/40,2/40,2/40,2/40,2/40,3/40,3/40,
2/40,2/40,3/40,3/40;
ideal J=substitute(I,v,p);
dim(std(J));
===> 28
```

It should be noted that the reason we get a dimension 28 is that the ideal v and p are defined on the ring r which has additional 17 unknowns other than p_{ij}. No matter what the number is, the positiveness of the number means p is a solution for p_{ij}. Otherwise, if it is zero, p is not a solution for p_{ij}.

B.3 Plotting Unidentifiable Space

For the 100 Swiss Franks problem, we know that

$$\frac{1}{40}\begin{pmatrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{pmatrix}$$

is one MLE for the 2-level latent class model, that is, the MLE maximizing the equation 19. And we also know there is a 2-dimensional subspace in the parameter space of conditional probabilities corresponding to this MLE. Now we show how to find the equations defining this unidentifiable space. In the below code, w_t's are the marginal probabilities of the latent variable, a_{it}'s and b_{jt}'s are the conditional probabilities of the observed variables given the latent variable. Then we define an ideal I, in which the first 5 polynomials corresponding to the sum 1 constraints and the last 16 polynomials corresponding to the model equalities $p_{ij} = \sum_t w_t a_{it} b_{jt}$ for the MLE.

```
ring r=0, (w1,a11,a21,a31,a41,b11,b21,b31,b41,
w2,a12,a22,a32,a42,b12,b22,b32,b42), lp;
ideal I=w1+w2-1,
a11+a21+a31+a41-1,
a12+a22+a32+a42-1,
b11+b21+b31+b41-1,
b12+b22+b32+b42-1,
w1*a11*b11+w2*a12*b12-3/40,
w1*a11*b21+w2*a12*b22-3/40,
w1*a11*b31+w2*a12*b32-2/40,
w1*a11*b41+w2*a12*b42-2/40,
w1*a21*b11+w2*a22*b12-3/40,
w1*a21*b21+w2*a22*b22-3/40,
w1*a21*b31+w2*a22*b32-2/40,
w1*a21*b41+w2*a22*b42-2/40,
w1*a31*b11+w2*a32*b12-2/40,
w1*a31*b21+w2*a32*b22-2/40,
w1*a31*b31+w2*a32*b32-3/40,
w1*a31*b41+w2*a32*b42-3/40,
w1*a41*b11+w2*a42*b12-2/40,
w1*a41*b21+w2*a42*b22-2/40,
w1*a41*b31+w2*a42*b32-3/40,
w1*a41*b41+w2*a42*b42-3/40;
dim(std(I));
===> 2
```

Now we can see the dimension of the ideal I is really 2. Then we can eliminate the unknowns other than w_1, a_{11}, b_{11} from the ideal I, thus we get the equation for the projection of the 2-dimensional unidentifiable subspace in (w_1, a_{11}, b_{11}) coordinates.

```
ideal J=elim1(I, a21*a31*a41*b21*b31*b41*w2*a12*a22*a32*a42
*b12*b22*b32*b42);
J;
===> J[1]=80*w1*a11*b11-20*w1*a11-20*w1*b11+6*w1-1;
```

The resulting ideal J has a one-to-one correspondence to the identifiable space. This is because the unidentifiable space is 2-dimensional, thus once the values of w_1, a_{11} and b_{11} are known so do the other paramters.

```
LIB "surf.lib";
ring r2=0, (w1, a11, b11), lp;
ideal J=80*w1*a11*b11-20*w1*a11-20*w1*b11+6*w1-1;
plot(J);
```

Singular calls the programme `surf` to draw real pictures of plane curves and surfaces in 3-D space. If you load library "SURF.LIB" in Singular and execute the "PLOT" command to show the vanishing surface of the ideal J, you will get a picture in figure 11.

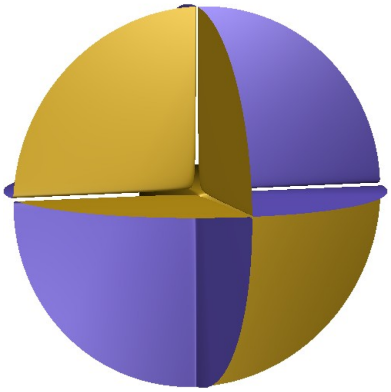
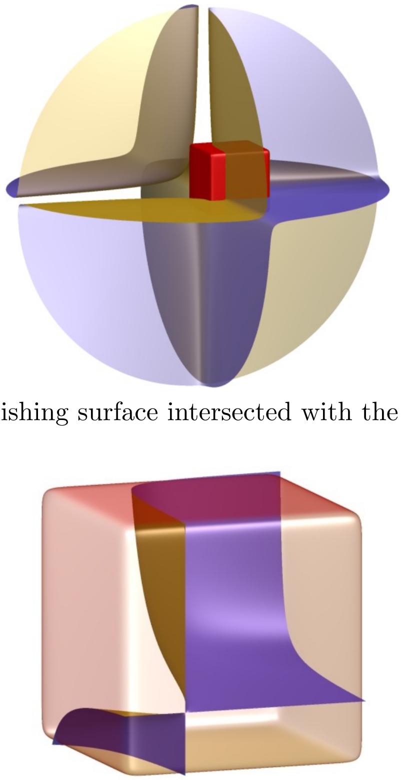


Figure 11: The surface that the ideal J is vanishing.

But the surface showed in figure 11 doesn't guarantee w_1, a_{11}, b_{11} to be within 0 and 1. If we want to plot more sophisticated surfaces, we can use the stand-alone programme `surf`. The unidentifiable space is the intersection of the vanishing surface and the $[0, 1]^3$ cube, which is shown in figure 12. We include the script used in `surf` to draw the pictures in the next section.

B.4 `Surf` Script

Below is the script used in `surf` to draw the pictures in figure 12-(b).



(a) the vanishing surface intersected with the unit cube.

(b) the vanishing surface inside the unit cube.

Figure 12: The intersection of the vanishing surface for ideal J and the $[0,1]^3$ cube.

```
width = 500;
height = 500;
double pi = 3.1415926;
double ss = 0.15;

origin_x = -0.5;
origin_y = -0.5;
origin_z = 0;

clip = cube;
radius = 0.5;
center_x = 0.5;
center_y = 0.5;
center_z = 0.5;
```

```
scale_x = ss;
scale_y = ss;
scale_z = ss;

rot_x = pi / 180 * 10;
rot_y = - pi / 180 * 20;
rot_z =  pi / 180 * 0;

antialiasing = 4;
antialiasing_threshold = 0.05;
antialiasing_radius = 1.5;

surface2_red = 255;
surface2_green = 0;
surface2_blue = 0;

inside2_red = 255;
inside2_green = 0;
inside2_blue = 0;

transparence = 0;
transparence2 = 70;

illumination = ambient_light + diffuse_light + reflected_light + transmitted_light;

surface = 80*x*y*z - 20*x*z - 20*y*z + 6*z -1;
surface2 = (x-0.500)^30 + (y-0.500)^30+(z-0.500)^30 - (0.499)^30;

clear_screen;
draw_surface;
```

C Proof of the Fixed Points for 100 Swiss Franks Problem

In this section, we show that when maximizing the log-likelihood function of 2-level latent class model for the 100 Swiss Franks problem, the table

$$f = \frac{1}{40}\begin{pmatrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{pmatrix} \tag{21}$$

is a fixed point in the Expectation Maximization algorithm. Here the observed table is

$$p = \frac{1}{40}\begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix}$$

Under the conditional independence of the latent structure model, we have

$$f_{ij} = \sum_{t \in \{0,1\}} \lambda_t \alpha_{it} \beta_{jt}$$

where $\sum_t \lambda_t = \sum_i \alpha_{it} = \sum_j \beta_{jt} = 1$, $\lambda_t \geq 0$, $\alpha_{it} \geq 0$ and $\beta_{jt} \geq 0$.

Now, we show that if we start with the values such that

$$\begin{array}{l} \alpha_{1t} = \alpha_{2t},\ \alpha_{3t} = \alpha_{4t} \\ \beta_{1t} = \beta_{2t},\ \beta_{3t} = \beta_{4t} \\ \sum_t \lambda_t \alpha_{1t} \beta_{1t} = \sum_t \lambda_t \alpha_{3t} \beta_{3t} = 3/40 \\ \sum_t \lambda_t \alpha_{1t} \beta_{3t} = \sum_t \lambda_t \alpha_{3t} \beta_{1t} = 2/40 \end{array} \tag{22}$$

then the EM will stay in these values and the fitted table is right the one in equation 21. In fact, in the E step, the posterior probability is updated by

$$\pi_{ijt}^{AB\bar{X}} = P(X = t | A = i, B = j) = \frac{\lambda_t \alpha_{it} \beta_{jt}}{f_{ij}}$$

Then in the M step, the parameters are updated by

$$\begin{array}{rcl} \hat{\lambda}_t & = & \sum_{i,j} p_{ij} \pi_{ijt}^{AB\bar{X}} \\ & = & \sum_{i,j} p_{ij} \frac{\lambda_t \alpha_{it} \beta_{jt}}{f_{ij}} \\ & = & \lambda_t + \frac{1}{3}[\alpha_{1t}\beta_{1t} + \alpha_{2t}\beta_{2t} + \alpha_{3t}\beta_{3t} + \alpha_{4t}\beta_{4t}] \\ & & -\frac{1}{3}[\alpha_{1t}\beta_{2t} + \alpha_{2t}\beta_{1t} + \alpha_{3t}\beta_{4t} + \alpha_{4t}\beta_{3t}] \\ & = & \lambda_t \\ \hat{\alpha}_{it} & = & \sum_j p_{ij} \pi_{ijt}^{AB\bar{X}} / \hat{\lambda}_t \\ & = & \alpha_{it} \sum_j p_{ij} \beta_{jt} / f_{ij} \\ & = & \left\{ \begin{array}{ll} \alpha_{it}[1 + \frac{1}{3}\beta_{1t} - \frac{1}{3}\beta_{2t}], & i = 1 \\ \alpha_{it}[1 + \frac{1}{3}\beta_{2t} - \frac{1}{3}\beta_{1t}], & i = 2 \\ \alpha_{it}[1 + \frac{1}{3}\beta_{3t} - \frac{1}{3}\beta_{4t}], & i = 3 \\ \alpha_{it}[1 + \frac{1}{3}\beta_{4t} - \frac{1}{3}\beta_{3t}], & i = 4 \end{array} \right. \\ & = & \alpha_{it} \\ \hat{\beta}_{jt} & = & \sum_i p_{ij} \pi_{ijt}^{AB\bar{X}} / \hat{\lambda}_t \\ & = & \beta_{jt} \sum_i p_{ij} \alpha_{it} / f_{ij} \\ & = & \left\{ \begin{array}{ll} \beta_{jt}[1 + \frac{1}{3}\alpha_{1t} - \frac{1}{3}\alpha_{2t}], & j = 1 \\ \beta_{jt}[1 + \frac{1}{3}\alpha_{2t} - \frac{1}{3}\alpha_{1t}], & j = 2 \\ \beta_{jt}[1 + \frac{1}{3}\alpha_{3t} - \frac{1}{3}\alpha_{4t}], & j = 3 \\ \beta_{jt}[1 + \frac{1}{3}\alpha_{4t} - \frac{1}{3}\alpha_{3t}], & j = 4 \end{array} \right. \\ & = & \beta_{jt} \end{array}$$

Thus, we have proved that the starting point given by equation 22 is a fixed point in the EM algorithm. And this fixed point will give us the fitted table f in equation 21. However, this is not the only fixed points for the EM. In fact, according to the proof above, we can also show that the points

$$\alpha_{1t} = \alpha_{3t},\ \alpha_{2t} = \alpha_{4t},\ \ \beta_{1t} = \beta_{3t},\ \beta_{2t} = \beta_{4t}$$

and

$$\alpha_{1t} = \alpha_{4t},\ \alpha_{2t} = \alpha_{3t},\ \ \beta_{1t} = \beta_{4t},\ \beta_{2t} = \beta_{3t}$$

are fixed points too. And the two points will lead to the tables

$$\frac{1}{40}\begin{pmatrix} 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \\ 3 & 2 & 3 & 2 \\ 2 & 3 & 2 & 3 \end{pmatrix}$$

and

$$\frac{1}{40}\begin{pmatrix} 3 & 2 & 2 & 3 \\ 2 & 3 & 3 & 2 \\ 2 & 3 & 3 & 2 \\ 3 & 2 & 2 & 3 \end{pmatrix}$$

Similarly, we can show that the table

$$\frac{1}{40}\begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 8/3 & 8/3 & 8/3 \\ 2 & 8/3 & 8/3 & 8/3 \\ 2 & 8/3 & 8/3 & 8/3 \end{pmatrix}$$

and its permutations are also the fixed points in the EM algorithm.

D Matlab Codes

Here we include the two matlab subroutines which are used to compute the Jacobian of the polynomial mapping $f\colon \Delta_{d_1-1} \times \cdots \times \Delta_{d_k-1} \times \Delta_{r-1} \to \Delta_{d-1}$ $(d = \prod_i d_i)$ in equation 23 and its numerical rank for latent class models.

$$(p_1(i_1)\ldots p_k(i_k), \lambda_h) \mapsto \sum_{h\in[r]} p_1(i_1)\ldots p_k(i_k)\lambda_h, \qquad (23)$$

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [J,f,x,w,a] = jacob_lcm(T, I)

% -------------------------------------------------------------------------
% JACOB_LCM computes the Jacobian of the latent class model.
% For example:
%       [J, f, x, w, a] = jacob_lcm(2, [3,3,3]);
```

```
%

w = sym('', 'real');
a = sym('', 'real');
for t=1:T
    w(end+1) = sym(['w', int2str(t)], 'real');
    for k=1:length(I)
        for i=1:I(k)
            a{k}(i,t) = sym(['a', int2str(i), int2str(t), int2str(k)], 'real');
        end
    end
end
w(end) = 1 - sum(w(1:end-1));
x = w(1:end-1);
for k=1:length(I)
    for t=1:T
        a{k}(end,t) = 1 - sum(a{k}(1:end-1,t));
        x = [x, a{k}(1:end-1,t)'];
    end
end

% get the mapping from parameters to table
f = sym('', 'real');
for idx=1:prod(I)
    subv = ind2subv(I, idx);
    val = sym('0');
    for t=1:T
        temp = w(t);
        for k=1:length(I)
            temp = temp * a{k}(subv(k),t);
        end
        val = val + temp;
    end
    f(end+1) = val;
end

% get the Jacobian
J = jacobian(f, x);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function r = rank_lcm(J, w, a)

% ------------------------------------------------------------------------
% RANK_LCM computes the numberical rank of the sybotical matri 'J', which
% is a function of 'w' and 'a'. It is used after calling the funtion JACOB_LCM.
% For example,
%       [J,f,x,w,a] = jacob_lcm(2, [2,2,2,2]);
%       rank_lcm(J,w,a);
%

T = length(w);
I = zeros(1, length(a));
for k=1:length(a)
    I(k) = size(a{k},1);
end

% compute the numberical rank
v = unifrnd(0,1,1,T);
v = v ./ sum(v);
for t=1:T
    for k=1:length(I)
        b{k}(:,t) = unifrnd(0,1,I(k),1);
        b{k}(:,t) = b{k}(:,t) ./ sum(b{k}(:,t));
    end
end

JJ = zeros(size(J));
for  i=1:size(J,1)
    for j=1:size(J,2)
        cc = char(J(i,j));
        for t=1:T
            cc = strrep(cc, char(w(t)), num2str(v(t)));
            for k=1:length(I)
                for p=1:I(k)
                    cc = strrep(cc, char(a{k}(p,t)), num2str(b{k}(p,t)));
                end
            end
        end
        JJ(i,j) = eval(cc);
    end
end

r = rank(JJ);
```

Here are the EM and Newton-Raphson codes for maximum likelihood estimation in latent class models.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [nhat, m, b, se, llk, retcode, X] = LCM_newton(n, T, maxiter, eps, m, X, verbose)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
```

```
% INPUT:
%       n(required):          observed table, a multi-dimensional array
%       T(required):          number of latent classes
%       maxiter(required):    maximum number of iterations
%       eps(required):        converge threshold
%  m(optional):   initial value for the mean vector
% X(optional):   design matrix
%       verbose(optional):   display results if true
% OUTPUT:
%       nhat:       estimated observed table
%       m:          estimated probability for the full table
%       b:          estimated parameter
%       se:         standard error of mle
%       llk:        log-likelihood values in iterations
%       retcode:    1, if the algorithm terminates normally; 0, otherwise
% X: design matrix
%

dbstop if warning;
dbstop if error;

% 1. initialize
y = n(:);                                   % observed table
k = length(y);                              % number of cells
dim = size(n);                              % dimensions of observed table
s = catrep(2, T, [1:k]);
S = zeros(T*k, k);                          % scatter matrix ===> S'm = nhat
for i=1:k
    idx = find( s==i );
    S(idx, i) = 1;
end
z = S * inv(S'*S) * y;                      % observed full table ===> S'z = y
fulldim = [dim, T];                         % dimensions of full table

if nargin < 7
    verbose = 1;
end
if nargin < 6
    X = [];
end
if nargin < 5
    m = [];
end

if isempty(X)
X = zeros(T*k, 1+(T-1)+sum(dim-1)+sum((T-1)*(dim-1)));  % design matrix
```

```
for idx=1:prod(fulldim)
    % for main effect
    xrow = 1;
    % for first order effect
    G = {};
    subv = ind2subv(fulldim, idx);
    for i=1:length(subv)
        if subv(i)==fulldim(i)
            G{i} = - ones(fulldim(i)-1, 1);
        else
            G{i} = zeros(fulldim(i)-1, 1);
            G{i}(subv(i)) = 1;
        end
        xrow = [xrow, G{i}'];
    end
    % for second order effect
    for i=1:length(subv)-1
        temp = G{end} * G{i}';
        xrow = [xrow, temp(:)'];
    end
    %
    if length(xrow)~=size(X,2)
        keyboard;
    end
    X(idx,:) = xrow;
end
end

if isempty(m)
    b = unifrnd(-1, 1, size(X,2), 1);           % initial value of the parameter
    m = exp(X*b);                               % estimated mean counts
else
    b = inv(X'*X) * (X' * log(m));
    m = exp(X*b);
end

% 2. newton-raphson
llk = sum(y .* log(S' * m ./ sum(m)));
retcode = 1;
for i=1:maxiter
    % Jacobi
    A = S'*diag(m)*S;
    if min(diag(A))<eps     % A is diagonal
        disp('maxtrix A for the Jacobi is singular.');
        disp('the algorithm stops without converging.');
        retcode = 0;
        break;
    end
    A = inv(A);
    P = S * A * S';
    J = (z-m)' * P * diag(m) * X;

    % Hessian
    C = X' * (diag(z' * P) * diag(m) - diag(m) * (S * diag(y) * (A^2) * S') * diag(m)) * X;
    D = X' * diag(m) * X;
    H = C - D;
    if  max(eig(H)) >= 0
       H = -D;
    end
    [eigvec, eigval] = eig(H);
    eigval = diag(eigval);
    if  min(eigval) >= 0
        disp('the hessian matrix is non-negative definite.');
        retcode = 0;
        break;
    end
    eigval(find(eigval<0)) = 1 ./ eigval(find(eigval<0));
    eigval(find(eigval>=0)) = 0;
    db = eigvec * diag(eigval) * eigvec' * J';
    ss = 1;
    b = b - ss * db;

    m = exp(X*b);
    % log-likelihood
    llk(end+1) = sum(y .* log(S' * m ./ sum(m)));
    %if abs(llk(end)-llk(end-1))<eps
    if max(abs(J)) < eps
        disp(['algorithm convergs in ', int2str(i), ' steps.']);
        break;
    end
end

% log-likelihood
llk = llk;
% fitted table
nhat = S'* (m ./ sum(m)) * sum(n(:));
% standard errors
se = sqrt(-diag(inv(H)));


% 3. show results
if verbose
```

```
    disp('the fitted and observed counts:');
    disp([nhat, n(:)]);
    disp('mle and stand error of the parameter:');
    disp([b, se]);
    plot(llk);
    axis tight;
    xlabel('iteration');
    ylabel('log-likelihood');
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [f, m, llk, llr, df, c, p, devbuf, c00, p00] = em_lsm(n, T, maxiter, eps, c0, p0)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% EM algorithm for latent class model
%
% input:
%       n(required): obserbed table. multi-dimensional array
%       T(required): number of latent classes
%       maxiter(required): maximum number of iterations
%       eps(required): converge threshold
%       c0(optional): initial value for class probabilities
%       p0(optional): initial value for conditional probabilities
% output:
% f: fitted table
% m: expected mean vector
% llk: log-likelihoods
% llr: likelihood ratio statistic
% df: degree of freedoms
%       c: class probabilities
%       p: conditional probabilities
%       devbuf:  maximum deviations of the estimates in iterations
% c00: initial class probabilties
% p00:  initial conditional probabilities
%

dbstop if warning;

f0 = n;
n = n / sum(n(:));
sz = size(n);
if nargin < 6
    p0 = cell(1, length(sz));
    for i=1:length(p0)
        A = rand(sz(i), T);
        A = A ./ kron(ones(sz(i),1), sum(A, 1));
        p0{i} = A;
    end
end
if nargin < 5
    c0 = rand(1,T);
    c0 = c0 ./ sum(c0);
end
c00 = c0;
p00 = p0;

nn = zeros([sz, T]);
c = c0;
p = p0;
iter = 0;
devbuf = [];
llk = 0;
while iter < maxiter
    % E step
    for idx=1:prod(size(nn))
        subv = ind2subv(size(nn), idx);
        nn(idx) = c(subv(end));
        for i=1:length(sz)
            nn(idx) = nn(idx) * p{i}(subv(i), subv(end));
        end
    end
    nnhat = sum(nn, length(sz)+1);
    nnhat = catrep(length(sz)+1, T, nnhat);
    nnhat = nn ./ nnhat;

    % M step
    for t=1:T
        A = subarray(length(sz)+1, t, nnhat);
        A = n .* A;
        c(t) = sum(A(:));
        for i=1:length(sz)
            for k=1:sz(i)
                B = subarray(i, k, A);
                p{i}(k, t) = sum(B(:)) / c(t);
            end
        end
    end

    % mle of counts
    f = zeros([sz, T]);
    for idx=1:prod(size(f))
        subv = ind2subv(size(f), idx);
        f(idx) = c(subv(end));
        for i=1:length(sz)
            f(idx) = f(idx) * p{i}(subv(i), subv(end));
        end
    end
    f = sum(f, length(sz)+1);
    llk(end+1) = sum( f0(:) .* log(f(:)) );

    % if converged
    maxdev = max(abs(c-c0));
    for i=1:length(p)
        A = abs(p{i}-p0{i});
        maxdev = max(maxdev, max(A(:)));
    end
    devbuf = [devbuf, maxdev];
    if maxdev < eps
        disp(['algorithm converges in ', int2str(iter), ' steps.']);
        break;
    end

    c0 = c;
    p0 = p;
    iter = iter + 1;
end

% frequencies estimation
f = zeros([sz, T]);
for idx=1:prod(size(f))
    subv = ind2subv(size(f), idx);
    f(idx) = c(subv(end));
    for i=1:length(sz)
        f(idx) = f(idx) * p{i}(subv(i), subv(end));
    end
end
m = f;      % full table
f = sum(f, length(sz)+1);
f = f .* sum(f0(:));

% likelihood ratio test statistics
f0 = f0(:);
f1 = f(:);
llr = f0./f1;
llr( find(llr==0) ) = 1;
llr = 2 * sum( f0.*log(llr) );

% degree of freedom
```

```
df = (prod(size(n))-1) - (T-1+T*sum(size(n)-1));

llk = llk(2:end);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function C = catrep(dim, n, A)
str = [’C = cat(’, int2str(dim), ’,’];
for i=1:n
    str = [str, ’A,’];
end
str = [str(1:end-1), ’);’];
eval(str);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function subv = ind2subv(siz, idx)
fn = ’[’;
for k=1:length(siz)
    fn = [fn, ’subv(’, num2str(k), ’),’];
end
fn = [fn(1:length(fn)-1), ’] = ind2sub(siz, idx);’];
eval(fn);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function ind = subv2ind(siz, subv)
fn = ’ind = sub2ind(siz, ’;
for k=1:length(siz)
    fn = [fn, ’subv(’, num2str(k), ’),’];
end
fn = [fn(1:length(fn)-1), ’);’];
eval(fn);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function C = subarray(dim, idx, A)
str = ’C = A(’;
for i=1:length(size(A))
    if i==dim
        str = [str, int2str(idx), ’,’];
    else
        str = [str, ’:,’];
    end
end
str = [str(1:end-1), ’);’];
eval(str);
squeeze(C);
```

References

Allman, E.S. and Rhodes, J.A. (2006). Phylogenetic invariants for stationary base composition, *Journal of Symbolic Computation*, 41, 138–150.

Allman, E. and Rhodes, J.A. (2007). Phylogenetic ideals and varieties for the general Markov model, *Advances in Applied Mathematics*, to appear.

Anderson, D.R., Burham, K.P., and White, G.C. (1994). AIC model selection in overdispersed capture-recature data. *Ecology*, 75, 1780–1793.

Anderson, T.W. (1954). On estimation of parameters in latent structure analysis, *Psychometrika,* 19, 1–10.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S., and Rathouz, P.J. (1997). Latent variable regressionfor multiple discrete outcomes, *Journal of the American Statistical Association*, 92, 1375–1386.

Benedetti, R. (1990). *Real algebraic and semi-algebraic sets,* Hermann.

Catalisano, M.V., Geremita, A.V. and Gimigliano, A. (2002). Ranks of tensors, secant varieties of Segre varieties and fat points, *Linear Algebra and Its Applications,* 355, 263–285.

Catalisano, M.V., Geremita, A.V. and Gimigliano, A. (2003). Erratum to: "Ranks of tensors, secant varieties of Segre varieties and fat points," *Linear Algebra and Its Applications,* 367, 347–348.

Clogg, C. and Goodman, L. (1984). Latent Structure Analysis of a Set of Multidimensional Contingency Tables, *Journal of the American Statistical Association,* 79, 762–771.

Cohen, J.E. and Rothblum, U.G. (1993). Nonnegative rank, decompositions and factorisations of nonnegative matrices. *Linear Algebra and Its Applications,* 190, 149–168.

Cox, D.A., Little, J. and O'Shea, D. (1996). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra,* Springer-Verlag.

Cowell, R.G., Dawid, P.A., Lauritzen, S.L. and Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems,* Springer-Verlag.

Erosheva, E.A. (2002). Grade of Membership and Latent Structure Models with Application to Disability Survey Data. PhD thesis, Department of Statistics, Carnegie Mellon University.

Erosheva, E.A. (2005). Comparing latent structures of the grade of membership, Rasch and latent class models, *Psychometrika*, 70, 619–626.

Erosheva, E.A., Fienberg, S.E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data, *Annals of Applied Statistics*, 1, in press.

Espeland, M. A. (1986). A general class of models for discrete multivariate data, *Communications in Statistics: Simulation and Computation,* 15, 405–424.

Garcia, L.D. (2004). Algebraic Statistics in Model Selection, *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04),* 177–18, AUAI Press.

Garcia, L., Stillman, M. and Sturmfels, B. (2005). Algebraic Geometry of Bayesian Networks, *Journal of Symbolic Computation*, 39, 331–355.

Geiger, D., Heckerman, D., King, H. and Meek, C. (2001). Stratified Exponential Families: Graphical Models and Model Selection, *Annals of Statistics,* 29(2), 505–529.

Gibson, W.A. (1955). An extension of Anderson's solution for the latent structure equations, *Psychometrika,* 20, 69–73.

Goodman, L. (1979). On the estimation of parameters in latent structure analysis, *Psychometrika*, 44(1), 123–128.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, 61, 215–231.

Greuel, G.-M. , Pfister, G. and H. Schönemann. (2005). SINGULAR 3.0. A Computer Algebra System for Polynomial Computations. Centre for Computer Algebra, University of Kaiserslautern.
`http://www.singular.uni-kl.de`.

Haber, M. (1986). Testing for pairwise independence, *Biometrics*, 42, 429–435.

Haberman, S.J. (1974). Log-linear models for frequency tables derived by indirect obsertations: maximum likelihood equations, *Annals of Statistics,* 2, 911–924.

Haberman, S.J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation, *Sociological Methodology,* 18, 193–211.

Harris, J. (1992). *lgebraic Geometry: A First Course,* Springer-Verlag.

Henry, N.W.. and Lazarfeld, P.F. (1968). *Latent Structure Analysis,* Houghton Mufflin Company.

Kocka, T. and Zhang, N. L. (2002). Dimension correction for hierarchical latent class models, P*Proceeding of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, 267–274, Morgan Kaufmann.

Kruskal, J.B. (1975). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling, *Psychometrica,* 41, 281–293.

Landsberg, J.M. and Manivel, L. (2004). On the ideals of secant varieties of Segre varieties, *Foundations of Computational Mathematics,* 4, 397–422.

Lauritzen, S.L. (1996). *Graphical Models,* Oxford University Press.

Madansky, A. (1960). Determinantial methods in latent class analysis, *Psychometrica,* 25, 183–198.

Mond, D.M.Q., Smith, J.Q. and Van Straten, D. (2003) Stochastic factorisations, sandwiched simplices and the topology of the space of explanations, *Proceeding of the Royal Society of London, Series A*, 459, 2821–2845.

Monto, A.S., Koopman, J.S., and Longini, I.M. (1985). Tecumseh study of illness. XIII. Influenza infection and disease. American *Journal of Epidemiology*, 121, 811–822.

Pachter, L. and Sturmfels, B., eds. (2005). *Algebraic Statistics for Computational Biology,* Cambridge University Press.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review,* 26, 195–239.

Rusakov, D. and Geigerm, D. (2005). Asymptotic model selection for naive Bayesian networks, *Journal of Machine Learning Research,* 6, 1–35.

Settimi, R. and Smith, J.Q. (2005). Geometry, moments and conditional independence trees with hidden variables, *Annals of Statistics,* 28, 1179-1205.

Settimi, R. and Smith, J.Q. (1998). On the geometry of Bayesian graphical models with hidden variables, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intel ligence,* 479–472, Morgan Kaufmann Publishers.

Smith, J.Q. and Croft, J. (2003). Bayesian networks for discrete multivariate data: an algebraic approach to inference, *Journal of Multivariate Analysis*, 84, 387–402.

Strassen, V. (1983). Rank and optimal computation of generic tensors, *Linear Algebra and Its Applications,* 52/53, 654–685.

Humphreys, K. and Titterington, D.M. (2003). Variational approximations for categorical causal modeling with latent variables. *Psychometrika*, 68, 391–412.

Uebersax, J. (2006a). LCA Frequently Asked Questions (FAQ).
`http://ourworld.compuserve.com/homepages/jsuebersax/faq.htm`

Uebersax, J. (2006b). Latent Class Analysis, A web-site with bibliography, software, links and FAQ for latent class analysis. `http://ourworld.compuserve.com/homepages/jsuebersax/index.htm`

Watanabe, S. (2001). Algebraic analysis for nonidentifiable learning machines, *Neural Computation,* 13, 899933.

Zhou, Y. (2007). Maximum Likelihood Estimation in Latent Class Models. Manuscript.