Statistica Sinica **17**(2007), 1261-1272

Editorial

Expanding the Statistical Toolkit with Algebraic Statistics[1]

1. Evolution of Activities in Algebraic Statistics

Polynomials abound in the specification of statistical models and inferential methods. In particular, many common statistical procedures involve finding the solution to polynomial equations. Thus, in retrospect, we should not be surprised at the sudden emergence of a wide array of papers linking statistical methodology to modern approaches to computational algebraic geometry. But the fact is that these connections are a relatively recent development in the statistical literature and they have led to the use of the terminology "algebraic statistics" to describe this linkage.

There were two stimuli to much of the work in algebraic statistics that has taken place over the past decade. One of these was a paper by Pistone and Wynn (1996) using ideas for Gröbner bases to address the issue of confounding in experimental design. This paper spawned a line of research which ultimately led to the publication of the book on algebraic statistics by Pistone, Riccomagno, and Wynn (2001), which primarily addresses problems in experimental design. The other stimulus was a paper by Diaconis and Sturmfels (1998) on algebraic methods for discrete exponential family distributions. This paper influenced much statistical activity in other parts of statistics, including much of my own work on the topic, and led to the volume edited by Pachter and Sturmfels (2005) which focused on links between algebraic statistics and computational biology, the latter being the topic of the Clay Institute workshop that spawned much of the present special issue. The authors of both of these papers and their students have been active in much of the subsequent research (see the theses of Riccomagno (1997) and Sullivant (2005)).

[1] The research reported here was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences and by Army contract DAAD19-02-1-3-0389 to CyLab at Carnegie Mellon University. Section 2 is based on material developed collaboratively with Alessandro Rinaldo. I thank Mathias Drton, Alessandro Rinaldo, and Seth Sullivant for many edits.

A variety of workshops and conferences ensued on both sides of the Atlantic Ocean (e.g., GROSTAT workshops). A week-long workshop at the American Institute of Mathematics in Palo Alto, California, in December 2003, led to a 2006 special issue on the topic in the *Journal of Symbolic Computation*. Subsequently, there was a major component of the activities focusing on applications of algebraic statistics and computational biology supported by the Institute of Mathematics and Its Applications at the University of Minnesota during its Thematic Year on Applications of Algebraic Geometry, including a workshop in March 2007.

The availability of computer software for carrying out algebraic tasks has also advanced many of the activities in algebraic statistics. Programs such as CoCoA (CoCoATeam (2007)), Macaulay 2 (Grayson and Stillman (2006)), Singular (Greuel et al. (2005)), 4ti2 (Hemmeke, et al. (2005)), Latte (de Loera et al. (2003)), and Polymake (Gawrilow and Joswig (2005)) allow statisticians to do surprisingly complex algebraic calculations and derive new mathematical results.

In the next section, I describe how the tools of algebraic geometry can be used to characterize problems in the analysis of contingency tables, and then turn to the papers comprising this special issue of *Statistica Sinica*.

2. Algebraic Statistics for Contingency Tables

I have a fondness for contingency table problems, and many of them utilize algebraic geometry representations, often in multiple forms. Stimulated by the Diaconis and Sturmfels paper, I and several of my students have pursued these representations. See for example the Ph.D. theses of Dobra (2002), Slavkovic (2004) and Rinaldo (2005). Here I describe some aspects of the related algebraic statistics literature.

Contingency tables are arrays of non-negative integers arising from cross-classifying N objects based on a set of criteria specified by the categorical variables of interest (Bishop et al. (1975); Lauritzen (1996)). A contingency table $\mathbf{n}$ can be represented as a vector of non-negative integers, each indicating the number of times a given configuration of classifying criteria has been observed in the sample. Our work has focused on three interrelated classes of problems: (1) geometric characterization of log-linear models for cell probabilities in contingency tables, (2) estimation of cell probabilities under log-linear models, and (3) disclosure limitation

strategies associated with contingency tables which protect against the identification of individuals associated with counts in the tables.

Log-linear models are statistical models for the vector **p** of cell probabilities that are fully specified by a 0-1 design matrix, in the sense that, for each **p** in the model, log **p** belongs to the row span of A. If **n** follows a multinomial distribution with parameters N and **p**, then the likelihood function (ignoring the multinomial coefficient) takes the form of a monomial: $\prod P_i^{n_i}$.

The vector **t** = A **n** of marginal tables for the highest-order terms in the model is a minimal sufficient statistics for the underlying parameters (Haberman (1974); Bishop et al. (1975)). One strategy for disclosure limitation is to release only a subset of lower-dimensional margins for multi-way tables (Dobra and Fienberg (2003); Fienberg and Slavkovic (2005)). It is precisely because released margins are also sufficient statistics that disclosure limitation techniques based on the release of marginal totals and inference for log-linear models are structurally linked and share much of the same statistical and mathematical formalism.

2.1 Parameter Space

In algebraic statistics, we represent hierarchical log-linear models by polynomial maps. The parameter space (i.e., set of probability distributions implied by such models) is a smooth hyper-surface of points satisfying polynomial equations, referred to as a *toric variety* (Sturmfels (1995)). We illustrate this fundamental notion using the 2×2 table with cell probabilities

p_{11}	p_{12}
p_{21}	p_{22}

and the model of independence, specified by the design matrix

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

The parameter space for this model is the set of all points $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ in the simplex such that log **p** belongs to the row span of A. Because the kernel of A is spanned by the vector $(+1,-1,+1,-1)$, this model is defined by the linear constraint $\log p_{11} + \log p_{22} - \log p_{12} - \log p_{21} = 0$. Taking the exponential transformation, this constraint is equivalent to an odds ratio of 1, i.e. $\frac{p_{11}p_{22}}{p_{12}p_{21}} = 1$. The toric variety for this model is the set of non-negative probability distributions **p** satisfying one polynomial equation

$$p_{11}p_{22} - p_{12}p_{21} = 0 \quad . \tag{1}$$

The solution set of this equation, depicted here in Figure 1, is a doubly-ruled surface known in statistics as the *surface of independence* (Fienberg and Gilbert (1970)), and in algebraic geometry as a *Segre variety*. See Slavkovic (2004) and Carlini and Rapallo (2005) for additional details.

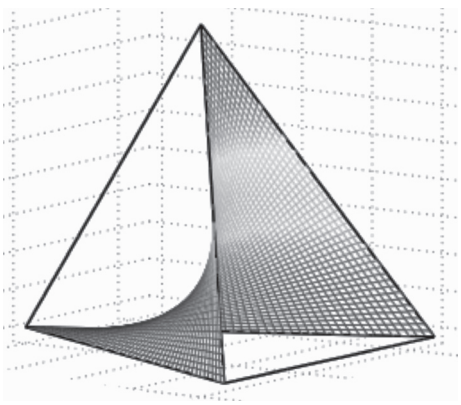

Figure 1 : Surface of independence for the 2×2 table. The tetrahedron represents the set of all probability distributions for a 2×2 table, while the enclosed surface identifies the probability distributions satisfying equation (1), i.e. the toric variety for the model of independence.

This algebraic geometry representation can be generalized to any hierarchical log-linear model, where the expression (1) becomes a (possibly very large) system of polynomial equations fully specified by the kernel of the design matrix A. Then a non-negative probability vector **p** is compatible with the log-linear model determined by A when it is a solution of that system.

Log-linear models represent log **p** as a point in the vector space spanned by the rows of A , but all permissible points **p** must be strictly positive. In the above example, only strictly positive probability distributions can, in fact, give an odds ratio of 1. In contrast, distributions for which $p_{11} = p_{21} = 0$, will satisfy (1). In fact, the algebraic geometry representation enjoys the crucial advantage of naturally providing an explicit representation of the closure of the parameter space. This closure consists of set of points in the simplex that belong to the toric variety and have some zero coordinates. It is the possibility of identifying these points, both analytically and geometrically, that allows for a full description of all possible patterns of sampling zeros leading to a non-existence of the maximum likelihood estimation (MLE) and alternative inferential procedures in such circumstances (Rinaldo (2005)).

2.2 Sample Space

We can represent virtually all data-dependent objects encountered in the study of the log-linear models as closed convex sets defined by linear inequalities. In particular, for a given log-linear model and a set of margins **t**, consider the convex bounded set

$$P_t = \{\mathbf{x} \text{ real, non-negative} : \mathbf{t} = A\,\mathbf{x}\}$$

of all real-valued non-negative tables having the same margins **t**, computed using the design matrix A . From the geometric point of view, P_t is a polytope that we can explicitly describe as a finite intersection of half-spaces. The set of all integer points inside P_t is the *fiber* of **t** and A . From the statistical point of view, the fiber is the portion of the sample space associated with the same set of sufficient statistics (margins). The fiber is, in fact, the support of the conditional distribution of tables given the margins, often known as the *exact distribution* and its characterization is fundamental to three statistical tasks:

(1) Counting. The simplest indication of the complexity of the fiber is its size, the number of integer-valued tables with prescribed margins, i.e., the lattice points in the polytope. Counting is of particular interest to those who want to explore issues of confidentiality and disclosure limitation (Fienberg and Slavkovic (2005)).

(2) Sampling. The support of the conditional distribution can be quite big, in fact so big that enumerating the points in the fiber may be an unrealistic task. An alternative solution is to perform a random walk over the space of tables with fixed margins by

means of Markov moves, integer valued vectors in the kernel of A that, added to the current table, will produce a new one with the same margins (Diaconis and Sturmfels (1998)). A Markov basis is a smallest set of moves that preserve connectedness in the fiber. Using Markov bases, it is possible to build a simple Metropolis sampler to explore in a stochastic fashion the fiber and to estimate the conditional distribution of the tables given the margins. In most cases, Markov bases can only be computed with algebraic symbolic software, such as 4ti2 (Hemmecke at al. (2005)), using algorithms that do not scale with the dimension of the problem and are not practical even for tables of modest size. For example, the Markov basis for a $4 \times 4 \times 4$ table and the model of no second-order interaction consists of 148,968 moves, obtained with considerable computational effort (Hemmecke and Malkin (2006)). Furthermore, the results of De Loera and Onn (2006) indicate that there is little hope for an efficient computation of Markov bases for large problems. Nevertheless, from the theoretical point of view, Markov bases are of extreme importance, since they characterize the complexity and geometry of the fiber. For example, De Loera and Onn also show in a constructive way that the fiber can be largely (in fact, arbitrarily) disconnected.

Since Markov bases provide the minimal conditions required to guarantee, for every point in the fiber, a positive probability of being sampled, any algorithm for sampling from the exact distribution must necessarily rely, even only indirectly, on them. Otherwise there is the potential that we may ignore a very large portion of the fiber. We are thus left with a mixed challenge of statistical and computational algebra in this domain.

(3) Optimizing. An important task of great relevance for disclosure limitation techniques is integer linear programming over the fiber. In a limited sense, the technique solves the so-called "table entry data security problem," that is, the computation of sharp lower and upper bounds for the individual table entries given a set of margins. The knowledge of these bounds will allow for an immediate assessment of disclosure risk associated with the release of a set of given marginals (Fienberg and Slavkovic (2005)). Integer linear programming involves maximizing a linear function over the polyhedron P_t, with the additional constraint that the solution has to be integral. Unfortunately, the linear programming solution is not guaranteed to be correct, as it may produce bounds that are fractional and not sharp (e.g., Dobra et al. (2003) and Sullivant (2005)). Algebraic conditions on the size of the difference between the linear programming and integer linear programming solution, called the *integer gap*, are given by Hoşten and Sturmfels (2007). Dobra (2002) provides an

algorithm for computing bounds − the *generalized shuttle algorithm*, which produces, as special cases with minimal computation, the formulas for the sharp bounds in Dobra and Fienberg (2000, 2001, 2003) for margins corresponding to decomposable graphs, as well as their device for reducing computation when the margins correspond to regular graphs (where the decomposable components are not necessarily fully connected). Because this algorithm is substituting for the traversal of all lattice points in the convex polytope, and because this involves aspects of the *exact distribution* without the probabilities, it is not surprising that there are links with the issues of MLE, c.f. the result on decomposable models in Geiger, Meeks, and Sturmfels (2006).

The convex hull of all the possible margins **t** that could be observed for a given design matrix A is called the *marginal cone*, and it is an unbounded (here N is allowed to be any integer number) convex set consisting of all the linear combinations of the columns of A with nonnegative coefficients,

$$C_A = \{\mathbf{y} : \mathbf{y} = A\,\mathbf{x},\ \mathbf{x} \text{ real, non-negative}\}.$$

Since the margins are sufficient statistics, the marginal cone provides the most efficient and parsimonious representation of the entire sample space. In particular, it allows for a constructive characterization for when the MLE $\hat{p}$ of the cell probability vector $\hat{p}$ lies in the interior of the parameter space where the cell probabilities are strictly less than 1 and strictly greater than 0 (c.f., Rinaldo (2005); Eriksson et al. (2006); Fienberg and Rinaldo (2007)).

In summary, we use the design matrix A and the marginal tables **t** to obtain geometric representations of the parameter and sample space for log-linear models. On one hand, A determines a system of polynomial equations that encode the dependencies among the random variables in the table. The solution set of these equations is the hyper-surface representing the parameter space as a compact subset of the simplex. On the other hand, **t** belongs to the marginal cone C_A and determines the polytope P_t, which in turn contains the fiber, i.e., part of the parameter space that is relevant for both statistical inference and the enumeration and bounding problems given a set of marginal totals.

There are many other related contributions expanding upon and extending these ideas, some of which have appeared in the past five years and several of which are in various stages of preparation, review, and publication.

3. The Special Issue

In the preceding section, I introduced some of the basic ideas in the algebraic statistics literature relevant for contingency table analysis. Four papers in this special issue tie in directly to different aspects of this work.

(1) Elizabeth Allman and John Rhodes describe an algebraic representation for a class of problems in molecular phylogenetics. Their problem can be interpreted as one involving inference in tree-structured latent class models for manifest contingency tables. Thus their results have value in the exploration of extensions of the work described in section 2, which are also special cases of Garcia et al.'s (2005) Bayesian networks representation.

(2) Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels deal with the computational biology problem of epistasis and shapes of fitness landscapes. This does in fact have a contingency table representation and the models they explore in depth, and to which they give a novel geometric representation, relate to explorations of the 2nd- and higher-order interaction structures in contingency tables.

(3) Guido Consonni and Giovanni Pistone provide extensions of the basic algebraic statistics representation for contingency tables in Diaconis and Sturmfels (1998) to allow for possible zero-probability cells in a Bayesian framework.

(4) Lawrence Cox examines the class of contingency tables of what he describes as the network type, which includes many decomposable log-linear models and others with square-free Markov bases such as the Rasch model. This work builds on his 2002 and 2003 papers, and in many senses on the earlier work of Dobra and Fienberg (2000, 2001, 2003), Dobra (2003), and Dobra and Sullivant (2004). [It does not contain direct comparisons with the generalized shuttle algorithm of Dobra (2003) referred to above in Section 2.]

The other four papers that are a part of this special issue deal with other facets of algebraic statistics including experimental design, continuous multivariate problems, and generalizations of classes of models previously considered.

(5) Max Buot, Serkan Hoşten, and Donald Richards extend the approach in Buot and Richards (2006), which uses Gröbner bases for counting and locating the solutions of polynomial systems of maximum likelihood equations to deal with extensions to the basic Behrens-Fisher problem involving inference for the differences in means in the presence of unequal variance structures.

(6) Mathias Drton and Seth Sullivant define the statistical class of algebraic exponential families in an effort to unify the work on discrete (categorical) variable problems and continuous ones.

(7) Sergi Elizalde and Kevin Woods deal with bounds on the number of inference functions of a graphical model, and deal with algebraic structures for Bayesian networks related to the work of Garcia et al. (2005).

(8) Maruri-Aguilar, Roberto Notari, and Eva Riccomagno extend the work on experimental design to deal with the description and identifiability analysis of experiments with mixtures.

This collection of papers should provide an important entry point to the statistical literature on algebraic statistics for an expanded group of statisticians and mathematicians.

References

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. Reprinted (2007), Springer, New York.

Buot, M. and Richards, D. (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations, I. *J. Symbolic Comput.* **41**, 234-244.

Carlini, E. and Rapallo, F. (2005). The geometry of statistical models for two-way contingency tables with fixed odds ratios. *Rendiconti dell'Istituto di Matematica dell'Universitá di Trieste* **37**, 71-84.

CoCoATeam (2007). CoCoA: A system for doing Computations in Commutative Algebra. http://cocoa.dima.unige.it .

Cox, L. H. (2002). Bounds on entries in 3-dimensional contingency tables subject to given marginal totals. *Inference Control in Statistical Databases.* In J. Domingo-Ferrer (Ed.). Springer LNCS 2316, 21-33.

Cox, L. H. (2003). On properties of multi-dimensional statistical tables. *J. Statist. Plann. Inference* **117**, 251-273.

De Loera, A., Hemmecke, R., Tauzer, J. and Yoshida, R. (2003). Effective lattice point counting in rational convex polytopes. *http://www.math.ucdavis.edu/~latte/pdf/lattE.pdf* .

De Loera, J. A. and Onn, S. (2006). Markov bases of 3-way tables are arbitrarily complicated. *J. Symbolic Comput.* **41**, 173-181.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distribution. *Ann. Statist.* **26**, 363-397.

Dobra, A. (2002). *Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables*. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.

Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli* **9**, 1-16.

Dobra, A., Erosheva, E. A. and Fienberg, S. E. (2003). Disclosure limitation methods based on bounds for large contingency tables with application to disability data. *Statistical Data Mining and Knowledge Discovery.* In H. Bozdogan (Ed.). Chapman and Hall/CRC Press, New York, 93-116.

Dobra, A. and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Natl. Acad. Sci.* **97**, 11885-11892.

Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statist. J. UN ECE* **18**, 363-371.

Dobra, A. and Fienberg, S. E. (2003). Bounding entries in multi-way contingency tables given a set of marginal totals. *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000.* In Y. Haitovsky, H. R. Lerche and Y. Ritov (Eds.). Physica-Verlag, 3-16.

Dobra, A. and Sullivant, S. (2004). A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Comput. Statist.* **19**, 347-366.

Eikland K. and Notebaert, P. *lp_solve* (Manual and program code). http://lpsolve.sourceforge.net/5.5/ .

Eriksson, N., Fienberg, S. E., Rinaldo, A. and Sullivant, S. (2006). Polyhedral conditions for the non-existence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.* **41**, 222-233.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *J. Amer. Statist. Assoc.* **66**, 694-701.

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137**, 3430-3445.

Fienberg, S. E. and Slavkovic, A. B. (2005). Preserving the confidentiality of categorical databases when releasing information for association rules. *Data Min. Knowl. Discov.* **11**, 155-180.

Garcia, L., Stillman, M. and Sturmfels, B. (2005). Algebraic geometry of Bayesian networks. *J. Symbolic Comput.* **39**, 331-355.

Gawrilow, E. and Joswig, M. (2005). Geometric reasoning with polymake. *arXiv:math.CO/0507273* .

Geiger, D. Meek, C. and Sturmfels, B. (2006). On the toric algebra of graphical models. *Ann. Statist.* **34**, 1463-1492.

Grayson, D. R. and Stillman, M. E. (2006). Macaulay 2, a software system for research in algebraic geometry. http://www.math.uiuc.edu/Macaulay2/ .

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

Haberman, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press, Chicago, IL.

Hemmecke, R., Hemmecke, R. and Malkin, P. (2005). 4ti2 Version 1.2─Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. http://www.4ti2.de .

Hemmecke, R. and Malkin, P. (2006). Computing generating sets of lattice ideals. http://arxiv.org/abs/math/0508359 .

Hoşten, S. and Sturmfels, B. (2007). Computing the integer programming gap. *Combinatorica* **27**, 367-382.

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

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

Pistone, G. and Wynn, H. P. (1996). Generalised confounding with Gröbner bases. *Biometrika* **83**, 653-666.

Pistone, G. and Riccomagno, E. and Wynn, H. P. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics,* Chapman and Hall, New York.

Riccomagno, E. M. (1997). *Algebraic Geometry in Experimental Design and Related Fields.* Ph.D. Dissertation, Department of Statistics, University of Warwick.

Rinaldo, A. (2005). *Maximum Likelihood Estimation for Log-linear Models.* Ph.D.

Dissertation, Department of Statistics, Carnegie Mellon University.

Slavkovic, A. B. (2004). *Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables*. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.

Sturmfels, B. (1995). *Gröbner Bases and Convex Polytope,* American Mathematical Society, University Lecture Series 8, Providence, RI.

Sullivant, S. (2005). *Toric Ideals in Algebraic Statistics.* Ph. D. dissertation, University of California, Berkeley.

--Stephen E. Fienberg