Computational Algebraic Statistics

The American Institute of Mathematics

This is a hard–copy version of a web page available through http://www.aimath.org
Input on this material is welcomed and can be sent to workshops@aimath.org
Version: Sat Mar 27 11:54:56 2004

Table of Contents

Chapter A: Open problems

This document was prepared by Seth Sullivant.

Problems suggested during talks

Persi Diaconis

Fixed First-Order Summaries.

The basic question is to try to understand how to understand "permutation data." Suppose that n voters rank p statements in order of preference. The data from such a survey consists of a function

$$f : S_p \longrightarrow \mathbb{N}$$

from the symmetric group on p letters to he nonnegative integers. Note that $\sum_{\sigma \in S_p} f(\sigma) = n$ The *first order summary* of f is a $p \times p$ table S of nonnegative integers. The (i, j) entry of S, $S_{i,j}$ is equal to the number of people who ranked candidate i in position j. If we assume no higher-order interactions between rankings, this describes a log-linear model.

Problem 1. Explicitly describe a Markov basis for tables with fixed first-order summary. That is, describe a set of moves which will connect any two surveys that have the same first order summary.

In a purely mathematical language: describe a generating set for the toric ideal given by the $p^2 \times p!$ matrix whose columns are the $p!$ $p \times p$ permutation matrices.

Contingency tables with quadratic statistics.

Problem 2. Find a Markov basis for studying the set $\mathcal{X}$ of 2-way tables with a fixed quadratic statistic:

$$\mathcal{X} = \{T | \sum_j T_{ij} = R_i, \sum_i T_{ij} = C_j \text{ and } \sum_{i,j} (i-2)(j-2)T_{ij} = Q\}.$$

For 3×4 tables, the matrix that computes the sufficient statistics (i.e. the matrix which represents the log-linear model) is the following:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & -2 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 2 \end{pmatrix}.$$

Polyhedral Cones and Testing.

Observe

$$X_{ij} = \{ \begin{matrix} 1 \\ 0 \end{matrix} | P(X_{ij}|X_{ij}) = Z^{-1}(\theta) i) e^{x_{ij}\theta_j} \}$$

Test $\theta_i = \theta \forall i$ against $A \cdot \theta \geq 0$ with A fixed.

Example 3. Possible conditions enforced by $A \cdots \theta \geq 0$ are

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_I$$

or

$$\theta_i \geq \theta_I \forall i.$$

There should be a nice interaction between statistics and polyhedral combinatorics in this problem.

Compare "Competing" Techniques.

A. Classical Asymptotics
B. Edgeworth Corrections
C. StatXact
D. Sequential Importance Sampling
E. Others

How do these different techniques for estimating p-values relate to each other?

With respect to sequential importance sampling: what are possible multi-way generalizations of the Gale-Ryser theorem which gives necessary and sufficient conditions for the existence of a 0/1 matrix with fixed row and column sums?

Stephen Fienberg

Questions about Odds-Ratios.

Question 4. For a 2×2 table of probabilities

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

: does it makes sense to fix values of two of the odds ratios $\alpha = \frac{p_{00}p_{11}}{p_{01}p_{10}}$ and $\alpha^* = \frac{p_{00}p_{01}}{p_{11}p_{10}}$ and look at the locus of possible values for the third odds-ratio $\alpha^{**} = \frac{p_{00}p_{10}}{p_{01}p_{11}}$?

Describe the curve (in the probability simplex) obtained by intersecting two of these quadric surfaces.

Question 5. What is the algebro-geometric structure of passing from local to global odds ratios? Note that *local odds ratio* is obtained by looking at a 2^k subtable of a given table whereas a *global odds ratio* is obtained by looking at the odds-ratio in a 2^k collapsing of a given table.

Question 6. For multi-way tables, what are the complete and partial specifications that come from specifying sets of odds-ratios?

Existence of Maximum Likelihood Estimates.

Question 7. What conditions on the structures of zeros in a table of counts will guarantee that the maximum likelihood estimates do not exist? In this setting, a maximum likelihood estimate is required to have all strictly positive probabilities.

As Alessandro Rinaldo, later pointed out in his talk, this is equivalent to giving a combinatorial description of the facets of the cone of all feasible margins. This polyhedral cone is obtained by taking of the positive hull of the columns of the matrices A_G which appear in the working questions on graphical models by Chris Meek and Serkan Hoşten.

Conjecture 8. *Maximum likelihood estimates for an $I \times J \times K$ table under the no 3-way interaction model do not exist if and only if they do not exist for all collapsings of the table to a $2 \times 2 \times 2$ table.*

Later in the week, Nicholas Eriksson showed that this conjecture is false already for $3 \times 3 \times 3$ tables.

Mathias Drton asked "How can we incorporate structural zeroes into this problem?"

Stephen Roehrig asked "Is it possible to relate the MLE solutions for decomposable models to decomposable network flow problems?"

Akimichi Takemura

Question 9. Do $I \times J \times K$ tables under the no 3-way interaction model always have a unique minimal Markov basis?

Bernd Sturmfels asked, "What about problems for 3-way tables that have structural zeros coming from a very regular design?"

Henry Wynn asked, "Does the group structure of circuits play a role in the solution of this problem?"

Emily Gamundi

Question 10. Are there always only 0, 1, or 2 statistically interesting solutions to the blocking probability equations in network tomography?

Ruriko Yoshida

Ruriko Yoshida discussed short encodings of the set of points in polytopes by means of rational generating functions, and the applications of rational generating functions for giving short encodings of Markov bases.

Ian Dinwoodie asked, "Is there a way to extract (random) monomials from the generating functions?"

Akimichi Takemura asked, "Is there a way to extract only the absolutely essential Markov basis elements that are needed to connect a given fiber of the form $\{A\mathbf{x} = \mathbf{b} | \mathbf{x} \in \mathbb{N}^d\}$?

Ruriko answered,"We are still attempting to solve these problems."

Henry Wynn asked, " Can these techniques be used to approximate the number of integral points inside an arbitrary convex set?"

Mathias Drton

Mathias Drton discussed multimodality of the likelihood function for the model of seemingly unrelated regression (SUR). These can be thought of as conditional independence models induced by the chain graph with two sets of nodes, $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ and directed edges from $X_i \to Y_i$ and undirected edges between any pair $Y_i - Y_j$.

Question 11. Can algebraic techniques be used to see the unimodality of the likelihood function in a MANOVA model. These models correspond to the chain graphs as described above but with the extra edges $X_i \to Y_j$ for each i and j.

Akimichi Takemura asked,"As the number of samples grows, how quickly does the likelihood converge to a unimodal function in the bivariate SUR model?"

Aleksandra Slavkovic

Aleksandra Slavkovic spoke about disclosure limitation problems associated with the releasse of marginals and conditionals. Since fixing conditionals is a linear constraint, the theory of Markov bases can be applied to the problem of deciding how many tables have the given fixed conditionals.

Question 12. What can be done with Markov bases when their are floating point approximations made in marginals or conditionals?

Question 13. What is the family of distributions that has margins and conditionals as sufficient statistics? What should the stationary distribution be for MCMC?

Jon Forster asked, "Is it safe to release full conditionals that have been perturbed?"

Aleksandra replied, "Probably not."

Luis Garcia

Luis Garcia spoke about the algebraic structure of the probability distributions which arise from graphical models with hidden variables.

Chris Meek asked, "What is the algebraic characterization of the Verma constraints? (non-independence constraints)"

Mathias Drton asked, "Are these the same constraints described by Richardson and Spirtes (Annals of Statistics 2002)?"

Henry Wynn asked, "How do inequalities play a role in this problem?"

Henry Wynn

Henry Wynn spoke about formulae relating cumulants to moments.

Problem 14. Use the transformations from probabilities to moments to cumulants to describe independence models/ hierarchical models in terms of polynomial functions in the cumulants.

Question 15. What is the cumulant ideal of the binary four-cycle model?

Russell Steele

Russell Steele spoke about mixture models, model selection, and the Bayesian Information Criterion (BIC).

Question 16. Computing normalization constants and posterior distributions for continuous random variables involves the computation of integrals involving rational and transcendental funtions. Can algebraic techniques be used to give exact expressions for these integrals or help give approximations for them?

Question 17. Why does the BIC work for model fitting with mixture models? What is being approximated (and why is it good to approximate these things?)

This question raised some controversy during the discussion following the talk which was later resolved. The main issue that is interesting in this situation is that the likelihood function for mixture models does not satisfy the necessary regularity conditions to apply the classical results about the BIC.

Question 18. Can algebra be used to better approximate the location of the modes in a mixture?

Question 19. Can algebraic techniques be used to better imputing missing categorical variable?

Question 20. Can algebraic techniques be used to collapse levels of categorical variables in order to improve the quality of multiple imputation without destroying the quality of inference?

Frantisek Matus

Frantisek Matus spoke about the representability of conditional independence (CI) structures.

Conjecture 21. *Every p-representable CI-structure can be p-represented by a random vector living on a finite probability space endowed with the uniform distribution.*

Conjecture 22. *There exists a connected matroid whose CI-structure is p-representable but not multilinear.*

Suppose independence vatieties are properly defined, also for the CI constraints involving functional independences, over the field of complex numbers.

Conjecture 23. *When P belongs to such a zero dimensional independence variety then P is the distribution of a random vector with its CI-structure equal to the CI-structure of a matroid.*

Donald Richards

Donald Richards spoke on the problem of determining the number of solutions to generic maximum likelihood problems and the use of computational algebra, homotopy continuation, and mixed volume in these calculations.

Problem 24 (Behrens-Fischer Problem)**.** Solve the maximum likelihood equations for the mixture of two multivariate Gaussians with the same mean and different covariance matrices.

Problem 25. Solve the maximum likelihood equations for a multivariate guassian with arbitrary missing data patterns.

Question 26. What are the possible implications of semi-algebraic geometry to maximum likelihood estimation?

Seth Sullivant

Seth Sullivant spoke about Markov bases for hierarchical models and theoretical results about the structure of minimal Markov basis elements.

Question 27. What types of "finiteness" statements can be made about Markov bases for hierarchical models as more than one set of levels is allowed to vary?

Problem 28. Characterize the binary hierarchical models which have Markov bases consisting of moves of degree 4 or less.

Elizabeth Allman

Problem 29. Compute the ideal of the secant varieties $Sec^k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1})$ for $k \geq 4$.

Question 30. Is the ideal obtained by taking the invariants arising from all 3-dimensional flattenings equal to the ideal of all invariants for a phylogenetic tree?

Problem 31. Determine the image of the stochastic parameterization (as opposed to the complex parameterization).

Problem 32. "Finding good invariants": Do some of the invariants have more of a statistical/biological significance than others?

Question 33. How can we efficiently decide which trees fit the data best? Which model of mutation is most accurate?

Problem 34. Find invariants for other models with rate variation among sites: secant varieties of the phylogenetic variaties.

Problem 35. Develop techniques to use invariants in combination with quartet methods.

Question 36. Can we develop algebraic invariant techniques for identifying clades?

Shmuel Onn

Shmuel Onn spoke about the spectra of hierarchical models. The spectra is defined to be the set of all sets of possible cell entries in for a multiway nonnegative integral table with fixed marginal totals.

Problem 37. Find the spectra of various models and classes of models.

Question 38. With respect to some distribution is the set of tables whose range of cell entries are intervals dense among all tables?

Joseph Landsberg

Ilias Kotsireas

Challenge problem in Gröbner bases of polynomial ideals.

The system of polynomial equations given in 3 different formats at
http://www.cargo.wlu.ca/hi/had7.cocoa
http://www.cargo.wlu.ca/hi/had7.maple
http://www.cargo.wlu.ca/hi/had7.reduce
has 41 equations in the 27 variables $a_1, \ldots, a_{27}$.

Prove that this system of polynomial equations, does not have any solutions.

This would constitute an algebraic proof of the combinatorial result of L. D. Baumert: **there are no Hadamard matrices of order** 28 **with one circulant core**.

Moreover, it is of interest to write down the effective Nullstellensatz for this ideal, i.e. write down explicitly a linear combination of (some of) the generators of the ideal, which is equal to 1.

Similar algebraic proofs could then be devised for Hadamard matrices with one circulant core of bigger order.

Reference: Ilias Kotsireas, Christos Koukouvinos, Jennifer Seberry, *Hadamard ideals and Hadamard matrices with one circulant core.* Submitted, November 2003.

Ideas that were suggested during the workshop:

- (Bernd Sturmfels) Devise some kind of incremental process to write down the effective Nullstellensatz. e.g. break the system into pieces, write the effective Nullstellensatz for each piece separately and then combine the results to obtain the effective Nullstellensatz for the whole system;
- (Beth Arnold) Use CoCoA and her own mod p Gröbner implementation, to see if the Gröbner basis can be computed.
- (Nick Eriksson) Use 4ti2, to see if the Grobner basis can be computed.

Design efficient heuristics in binary trees.

Given sets of binary vectors in a set of binary trees with given depths, design efficient heuristics to predict the positions of the analogous vectors in binary trees of bigger depth. The adjective "analogous" can take on various meanings. In our context, it designates solutions of the associated polynomial system. Ideas that were suggested during the workshop: (Persi Diaconis) first of all obtain graphical visualizations and then analyze the data with DFT and other transforms. Work in progress.

Indicator function approach for Hadamard Equivalence.

Check data sets of Hadamard matrices of big orders to identify the inequivalent Hadamard matrices. Ideas that were suggested during the workshop: (Maria Piera Rogantin) A JSPI paper by Fontana-Pistone-Rogantin contains a description of the indicator function approach to check Hadamard equivalence. In general, one needs to compute the indicator function which has exponentially many monomials. However, this approach is especially efficient for Hadamard matrices with some structure, e.g. with one circulant core. The main reason for this is that the exponentially many monomials fall naturally in a few categories and we only need to check equality of the coefficients of one representative from these categories. Work in progress.

First Open Problem Session

Question 39 (Sturmfels)**.** In a graphical model with hidden variables how many (nonnegative) real modes might exist in solutions to the likelihood equations?

Question 40 (Wynn)**.** Which nodes being hidden imply the multimodality/ unimodality of the likelihood? Given a particular graph, what nodes must be observed to ensure unimodality?

Question 41 (Hoşten)**.** How many complex roots to the likelihood equations can there be? What is the statistical significance of complex roots/ negative real roots or other roots that do not yield a valid statistical model (e.g. covariance matrices that are not positive definite)?

Question 42 (Richards)**.** What can be said about the number and structure of solutions to the likelihood equations which arise in the Behrens-Fischer problem?

Question 43 (Sturmfels)**.** Which is more important to solve: fixing the model and increasing the number of levels in the model, or fixing binary random variables and letting the number of random variables get large?

Meek: It depends on the type of problems you are interested in studying.

Roehrig: You can exchange a solution to problems on binary random variables to a problem on nonbinary random variables.

Question 44 (Landsberg)**.** Why is flattening used so frequently in statistics? Might other linear transformations on categorical data be more useful? Why not project onto the irreducible representations of the symmetric group? What about other forms of collapsing data that takes into account group actions?

Problem 45 (Fienberg)**.** Computing the full distributions for observables using Bayesian extensions. Predictive distributions.

Dinwoodie: How might you use algebra to pursue Bayesian directions in model fit criteria?

Sturmfels: How to you compute this in Macaulay 2? What is Bayesian algebraic geometry?

Fienberg: Taking mixtures of things the different models we have discussed at this conference.

Meek: Finding the modes of the posterior distributions could be an interesting research problem.

Pistone: You could use moment generating functions and cumulant generating functions for these problems.

Fienberg: Identifiability is a big issue in the Bayesian context.

Wynn: How may Edgeworth corrections play a role in this problem?

Question 46 (Steele)**.** What might be the application of these techniques to nonparametric models?

Question 47. [Pistone] Take a design of N distinct rational vectors $D = \{v_1, \ldots, v_N\} \subseteq \mathbb{Q}^d$ and $x^{\alpha_1}, \ldots, x^{\alpha_N}$ are monomials in $\mathbb{Q}[x_1, \ldots, x_n]$ which are linearly independent modulo the design ideal $I(D)$. Further suppose that these monomials are equal to the set of standard monomials of some monomial ideal M.

Suppose that $M = \langle m_1, \ldots, m_r \rangle$. Since the monomials $x^{\alpha_1}, \ldots, x^{\alpha_N}$ are a basis for a quotient space, there is a unique representation of the m_i as linear combinations of the x^{α_j}. This produces r polynomials $f_1, \ldots, f_r$ which belong to $I(D)$. Question: are the only points which vanish on these polynomials the design points; that is, does $V(f_1, \ldots, f_r) = D$?

Question 47 was answered in the negative by Bernd Sturmfels at the conference.

Question 48 (Meek)**.** Is there a polynomial time algorithm for describing the Markov basis of an arbitrary hierarchical model? Can you randomly choose elements from the Markov basis in polynomial time?

Sullivant: Is the number of symmetry classes a polynomial in the size of the problem?

Question 49 (Wynn)**.** Why polynomials? What can be said about algebraic methods when applied to other rings, like rings of rational functions, or rings where Fourier analysis can be used?

Question 50 (Kuhnt)**.** What about adding continuous random variables into these problems? Might an algebraic framework be useful?

Dinwoodie: Can we try to apply algebraic techniques to microarray data?

Meek: Gaussian (or similar) assumptions on the random variables might be possible to handle with algebraic techniques.

Question 51 (Meek)**.** Is there a polynomial time algorithm for describing the constraints on the joint probability distribution on the observed variables in a graphical model with hidden variables?

Problem 52 (Pachter)**.** Try applying the algebraic techniques to more restrictive families of models. For example, models with homogeneous transition matrices (that is, the same set of parameters is used for more than one edge/ clique in the graph), or genomic models.

Question 53 (Fienberg)**.** How can we make the algorithms and existing theoretical results apply to problems of sizes that are of practical interest? Scaling up to 5^{10} tables is a hopeful goal.

Question 54 (Fienberg)**.** How can we make it easier to identify "new" interesting details among long complicated computer output?

Karr: can we take advantage of sparse representations of problems?

Second Open Problem Session

Problem 55 (Karr)**.** Find a set of strictly positive (or strictly > 5) margins for which the set of tables which has these margins has an arbitrarily large gap in a cell entry.

Dinwoodie: Are there examples of this phenomenon in logistic regression?

Problem 56 (Fienberg)**.** Consider 3-way table missing data problem. This consists of a 3-way table with holes and 2-way margins which also have holes. . How can you estimate the likelihood and "glue" this information together into a good global picture of the whole table? Can this problem be formulated as an algebraic variety?

Problem 57. Gaussian missing data problem. Consider a continuous Gaussian population and a random sample of size n but only some of the characteristics of each individual are observed. Estimate the mean and covariance matrix.

Fienberg: First approach the problem using a block structure.

Steele: Look at a sequence of regressions as a sequence of polynomials. Can you solve the structure explicitly?

Problem 58 (Sturmfels)**.** Compute a mixture of the S_4 model from Diaconis' talk. What are the equations defining this secant variety? Is the likelihood multimodal?

MLE Project

On Thursday afternoon, Bernd Sturmfels, Serkan Hoşten, and Mathias Drton led a section on algebraic approaches to maximum likelihood estimations. The purpose was to introduce some algebraic techniques and possible small instances of statistical models to which one could apply these techniques. The eventual goal of this project is to try to come

to a better understanding of the maximum likelihood estimation problem using algebraic means.

Algebraic Methods for Optimization:

A. Singular: Gröbner bases and numerical eigenvalues for companion matrices or lexicographic solving
B. Lagrange multipliers
    1. Transformations out of the simplex?
    2. Exponential representation for problems
C. Resultants
D. Homotopy Methods (e.g. PHCpack)
E. Sum of Squares (e.g. SOStools)

Small Instances of Statistical Models

A. $\sigma_3(\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^2)$, that is, the hidden variable naive Bayes model with three observed variables, where all the random variables have three states. This model has 20 parameters.
B. Seemingly unrelated regression (SUR) model with three random variables. See questions related to Mathias Drton's talk. This model has 3 parameters.
C. Hidden Markov model with two hidden nodes. Hidden variables have 2 state and observed variables have 3 states. This model has 6 parameters.
D. An unconditional independence model for 4 binary random variables. $X_1 \perp X_2$ and $X_3 \perp X_4$. This model has 13 parameters.
E. Perfect tables on 3 random variables. $X_1 \perp X_2$, $X_2 \perp X_3$, and $X_1 \perp X_3$.

Questions from the Working Sessions

Graphical Models

Question 59 (Hoşten)**.** What can algebra say about MLE's for models with hidden variables?

Question 60 (Sullivant)**.** For a log-linear model, what can be said about the structure of the set of b with $Ax = b$ for some b and the solution space to the maximum likelihood equations is not zero-dimensional?

Hoşten: Look at the leading coefficients of polynomials in the elimination ideals (with parameters) and determine when these are zero.

Question 61 (From graphical models notes)**.** Is there any useful statistical measure that can be obtained from evaluating the polynomials that vanish on probabilities coming from a given model at an empirical distribution? What about

$$\sum |f(\widehat{p})|?$$

Linear Polynomial Models

A design D is a finite subset of $\mathbb{R}^d$. The design ideal $I(D)$ is the vanishing ideal of polynomial functions vanishing on D. The set of standard monomials modulo $I(D)$ given a term order τ is denoted Est_τ. A linear polynomial model L is a list of monomials.

Question 62 (Riccomagno)**.** Is there an algebraic method to decide which Est_τ factor is best when working with perturbed designs?

Problem 63 (Riccomagno)**.** Given a model L find a design D with special statistical properties such that the monomials in L are linearly independent modulo $I(D)$.

Question 64 (Laubenbacher)**.** Choosing models modulo the design ideal depends on the term order chosen. How can we eliminate this choice of term order from the problem of finding a model that fits the data?

Richards: Could you use random search or genetic algorithms to solve this problem?

Question 65 (Laubenbacher)**.** The model selection problem does not seem to be very effective when only one time series is used but does seem to work well when multiple knockout time series are used in combination. Why?

Question 66 (Laubenbacher)**.** The model selection problem over finite fields is very sensitive to noise. Is there some sort of "least squares" over finite fields to deal with noisy data?

Question 67 (Laubenbacher)**.** What is the complexity of the algorithms used for time series modelling? How can we make the algorithms faster?

Question 68 (Sturmfels)**.** In many of the problems for analyzing time series there are very few data points in a high dimensional space. How can techniques for reducing the dimensionality be used?

Software for Algebraic Statistics

Question 69 (Roehrig)**.** Why are there so many different systems for computational algebra? Couldn't they be add-ins to other systems?

Stillman: Not likely to work since those other systems (Mathematica, Maple, etc.) do not give access to their kernels.

Question 70 (Slavkovic)**.** Are there scripts which convert between input formats for the different computational algebra systems?

Question 71 (Sturmfels)**.** What should be the first algebraic functions available in R?

Dinwoodie: Toric Markov bases

Riccomagno: Ideal of points

Pachter: A general R interface with algebra packages

Riccomagno: Gröbner bases over fraction fields, with algebraic numbers and parameters.

Pachter: When do we find out when these things have been implemented?

Allman: Improve the documentation!