

Annotated Bibliography— Seth Sullivant

- 27.** (with N. Beerenwinkel) Markov models for accumulating mutations. Submitted, 2007.

We introduce and analyze a waiting time model for the accumulation of genetic changes. The continuous time conjunctive Bayesian network is defined by a partially ordered set of mutations and by the rate of fixation of each mutation. The partial order encodes constraints on the order in which mutations can fixate in the population, shedding light on the mutational pathways underlying the evolutionary process. We study a censored version of the model and derive equations for an EM algorithm to perform maximum likelihood estimation of the model parameters. We also show how to select the maximum likelihood poset. The model is applied to genetic data from different cancers and from drug resistant HIV samples, indicating implications for diagnosis and treatment.

- 26.** (with S. Hogten) Algebraic complexity of maximum likelihood estimation in bivariate missing data problems. Submitted, 2007.

We study the problem of maximum likelihood estimation for general patterns of bivariate missing data for normal and multinomial random variables, under the assumption that the data is missing at random (MAR). For normal data, the score equations have nine complex solutions, at least one of which is real and statistically significant. Our computations suggest that the number of real solutions is related to whether or not the MAR assumption is satisfied. In the multinomial case, all solutions to the score equations are real and the number of real solutions grows exponentially in the number of states of the underlying random variables, though there is always precisely one statistically significant local maxima.

- 25.** Conditional independence for Gaussian random variables has no finite complete characterization. Submitted, 2007.

We show that there can be no finite list of conditional independence relations which can be used to deduce all conditional independence implications among Gaussian random variables. To do this, we construct, for each $n > 3$ a family of n conditional independence statements on n random variables which together imply that $X_1 \perp\!\!\!\perp X_2$, and such that no subset have this same implication. The proof relies on binomial primary decomposition.

- 24.** Algebraic geometry of Gaussian Bayesian networks. To appear in *Advances in Applied Mathematics*, 2007.

Conditional independence models in the Gaussian case are algebraic varieties in the cone of positive definite covariance matrices. We study these varieties in the case of Bayesian networks, with a view towards generalizing the recursive factorization theorem to situations with hidden variables. In the case when the underlying graph is a tree, we show that the vanishing ideal of the model is generated by the conditional independence statements implied by graph. We also show that the ideal of any Bayesian network is homogeneous with respect to a multigrading induced by a collection of upstream random variables. This has a number of important consequences for hidden variable models. Finally, we relate the ideals of Bayesian networks to a number of classical constructions in algebraic geometry including toric degenerations of the Grassmannian, matrix Schubert varieties, and secant varieties.

- 23.** (with J. Sidman) Prolongations and computational algebra. To appear in *Canadian Journal of Mathematics*, 2006.

We explore the geometric notion of prolongations in the setting of computational algebra, extending results of Landsberg and Manivel which relate prolongations to equations for secant varieties. We also develop methods for computing prolongations which are combinatorial in nature. As an application, we use prolongations to derive a new family of secant equations for the binary symmetric model in phylogenetics.

22. (with M. Drton) Algebraic statistical models. To appear in *Statistica Sinica* Special Issue on Algebraic Statistics and Computational Biology, 2006.

Many statistical models are algebraic in that they are defined in terms of polynomial constraints, or in terms of polynomial or rational parameterizations. The parameter spaces of such models are typically semi-algebraic subsets of the parameter space of a reference model with nice properties, such as for example a regular exponential family. This observation leads to the definition of an ‘algebraic exponential family’. In this paper we review the ingredients to this definition and illustrate in examples how computational algebraic geometry can be used to solve problems arising in statistical inference in algebraic models.

21. Combinatorial symbolic powers. To appear in *Journal of Algebra*, 2006.

Symbolic powers of ideals are studied in the combinatorial context of monomial ideals. When the ideals are generated by quadratic squarefree monomials, the generators of the symbolic powers are obstructions to vertex covering in the associated graph and its blow-ups. As a result, perfect graphs play an important role in the theory, dual to the role played by perfect graphs in the theory of secants of monomial ideals. Among the applications are a new, unified approach to the Gröbner bases of symbolic powers of determinantal and Pfaffian ideals.

20. (with B. Sturmfels) Toric geometry of cuts and splits. To Appear in *Michigan Mathematical Journal*, 2006.

Associated to any graph is a toric ideal whose generators record relations among the cuts of the graph. We study these ideals and the geometry of the corresponding toric varieties. Our theorems and conjectures relate the combinatorial structure of the graph and the corresponding cut polytope to algebraic properties of the ideal. Cut ideals generalize toric ideals arising in phylogenetics and the study of contingency tables.

19. Toric fiber products. To appear in *Journal of Algebra*, 2006.

We introduce and study the toric fiber product of two ideals in polynomial rings that are homogeneous with respect to the same multigrading. Under the assumption that the set of degrees of the variables form a linearly independent set, we can explicitly describe generating sets and Gröbner bases for these ideals. This allows us to unify and generalize some results in algebraic statistics.

18. (with M. Drton and B. Sturmfels) Algebraic factor analysis: tetrads, pentads, and beyond. *Probab. Theory Related Fields* **138** (2007), no. 3-4, 463–493.

Factor analysis refers to a statistical model in which observed variables are conditionally independent given fewer hidden variables, known as factors, and all the random variables follow a multivariate normal distribution. The parameter space of a factor analysis model is a subset of the cone of positive definite matrices. This parameter space is studied from the perspective of computational algebraic geometry. Gröbner bases and resultants are applied to compute the ideal of all polynomial functions which vanish on the parameter space. These polynomials, known as model invariants, arise from rank conditions on a symmetric matrix under elimination of the diagonal entries of the matrix. Besides revealing the geometry of the factor analysis model, the model invariants also furnish useful statistics for testing goodness-of-fit.

- 17.** (with S. Hoşten) A finiteness theorem for Markov bases of hierarchical models. *J. Combin. Theory Ser. A* **114** (2007), no. 2, 311–321.

We show that the complexity of the Markov bases of multidimensional tables stabilizes eventually if a single table dimension is allowed to vary. In particular, if this table dimension is greater than a computable bound, the Markov bases consist of elements from Markov bases of smaller tables. We give an explicit formula for this bound in terms of Graver bases. We also compute these Markov and Graver complexities for all $K \times 2 \times 2 \times 2$ tables.

- 16.** Compressed polytopes and statistical disclosure limitation. *Tohoku Mathematical Journal* **58** (2006), 433–445.

We provide a characterization of the compressed lattice polytopes in terms of their facet defining inequalities and prove that every compressed lattice polytope is affinely isomorphic to a 0/1-polytope. As an application, we characterize those graphs whose cut polytopes are compressed and discuss consequences for studying linear programming relaxations in statistical disclosure limitation.

- 15.** (with B. Sturmfels) Combinatorial secant varieties. *Quarterly Journal of Pure and Applied Mathematics* **2** (2006) 285–309.

The construction of joins and secant varieties is studied in the combinatorial context of monomial ideals. For ideals generated by quadratic monomials, the generators of the secant ideals are obstructions to graph colorings, and this leads to a commutative algebra version of the Strong Perfect Graph Theorem. Given any projective variety and any term order, we explore whether the initial ideal of the secant ideal coincides with the secant ideal of the initial ideal. For toric varieties, this leads to the notion of delightful triangulations of convex polytopes.

- 14.** (with Y. Chen and I. Dinwoodie) Sequential importance sampling for multiway tables. *Annals of Statistics*. **34** (2006) No. 1, 523–545

We describe an algorithm for the sequential sampling of entries in multiway contingency tables with given constraints. The algorithm can be used for computations in exact conditional inference. To justify the algorithm, a theory relates sampling values at each step to properties of the associated toric ideal using computational commutative algebra. In particular, the property of interval cell counts at each step is related to exponents on lead indeterminates of a lexicographic Grbner basis. Also, the approximation of integer programming by linear programming for sampling is related to initial terms of a toric ideal. We apply the algorithm to examples of contingency tables which appear in the social and medical sciences. The numerical results demonstrate that the theory is applicable and that the algorithm performs well.

- 13.** (with N. Eriksson, S. E. Fienberg, and A. Rinaldo) Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *Journal of Symbolic Computation: special issue on Computational Algebraic Statistics*, **41** (2006), 222–233.

We provide a polyhedral description of the conditions for the existence of the maximum likelihood estimate (MLE) for a hierarchical log-linear model. The MLE exists if and only if the observed margins lie in the relative interior of the marginal cone. Using this description, we give an algorithm for determining if the MLE exists. If the tree width is bounded, the algorithm runs in polynomial time. We also perform a computational study of the case of three random variables under the no three-factor effect model.

- 12.** (with A. Slavkovic) The space of compatible full conditionals is a unimodular toric variety. *Journal of Symbolic Computation: special issue on Computational Algebraic Statistics*, **41** (2006), 196–209.

The set of all m -tuples of compatible full conditional distributions on discrete random variables is an algebraic set whose defining ideal is a unimodular toric ideal. We identify the defining polynomials of these ideals with closed walks on a bipartite graph. Our algebraic characterization provides a natural generalization of the requirement that compatible conditionals have identical odds ratios and holds regardless of the patterns of zeros in the conditional arrays.

11. Toric Ideals in Algebraic Statistics. Ph. D. Thesis. Department of Mathematics, UC Berkeley, 2005.

Algebraic statistics is the study of the algebraic varieties which arise in discrete multivariate statistics. These varieties are called algebraic statistical models and nonnegative real points on these varieties represent probability distributions on random variables that take a finite number of states. Such statistical models/ algebraic varieties make frequent appearances in the social and biological sciences.

This thesis is concerned with the study of a wide range of examples and problems in algebraic statistics, including: the study of Markov bases and conditional inference, statistical disclosure limitation and linear optimization, phylogenetics, and compatibility problems in conditionally specified models. While the statistical models and problems we will see in this work are quite varied, there are two themes which underlie the entire work. First, the study and the structure of the ideals of functions that define these models play a crucial role in algebraic statistics. Second, toric varieties and their defining ideals arise over and over again in disparate contexts. Understanding broad principles in the analysis of toric varieties and toric ideals can be useful for deriving specific results in algebraic statistics.

10. (with N. Eriksson, K. Ranestad, and B. Sturmfels) Phylogenetic algebraic geometry. In *Projective Varieties with Unexpected Properties*, edited by C. Ciliberto, et al., Walter de Gruyter, Berlin, 2005.

Phylogenetic algebraic geometry is concerned with certain complex projective algebraic varieties derived from finite trees. Real positive points on these varieties represent probabilistic models of evolution. For small trees, we recover classical geometric objects, such as toric and determinantal varieties and their secant varieties, but larger trees lead to new and largely unexplored territory. This paper gives a self-contained introduction to this subject and offers numerous open problems for algebraic geometers.

9. (with M. Casanellas) The strand symmetric model. Chapter in *Algebraic Statistics for Computational Biology*, (eds. L. Pachter and B. Sturmfels), 2005.

Base pairing in DNA implies that a surviving mutation in a DNA sequence will be accompanied by a symmetric mutation on the opposite strand of DNA. We study the resulting phylogenetic models that arise when imposing this strand symmetric assumption. This model possesses features of both the group-based models and the general Markov model. After applying a suitable Fourier transform to the parametrization, we study the phylogenetic invariants of the model using techniques of Allman and Rhodes for the general Markov model.

8. (with M. Casanellas and L. Garcia) Catalog of small trees. Chapter in *Algebraic Statistics for Computational Biology*, (eds. L. Pachter and B. Sturmfels), 2005.

We describe the results of a project to create a webpage to make phylogenetic invariants publicly available.

7. Small contingency tables with large gaps. *SIAM Journal on Discrete Mathematics*. **18** (2005), no. 4, 787-793.

We construct examples of contingency tables on n binary random variables where the gap between the linear programming lower/upper bound and the true integer lower/upper bounds on cell entries is exponentially large. These examples provide evidence that linear programming may not be an effective heuristic for detecting disclosures when releasing margins of multi-way tables.

- 6.** (with B. Sturmfels) Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, **12** no. 2 (2005), 204-228.

Statistical models of evolution are algebraic varieties in the space of joint probability distributions on the leaf colorations of a phylogenetic tree. The phylogenetic invariants of a model are the polynomials which vanish on the variety. Several widely used models for biological sequences have transition matrices that can be diagonalized by means of the Fourier transform of an abelian group. Their phylogenetic invariants form a toric ideal in the Fourier coordinates. We determine generators and Grbner bases for these toric ideals. For the Jukes-Cantor and Kimura models on a binary tree, our Grbner bases consist of certain explicitly constructed polynomials of degree at most four.

- 5.** (with A. Dobra) A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Computational Statistics* **19** (2004), 347-366.

We describe a divide and conquer technique for generating a Markov basis that connects all tables of counts having a fixed set of marginal totals. This procedure is based on decomposing the independence graph induced by these marginals. We discuss the practical imports of using this method in conjunction with other algorithms for determining Markov bases.

- 4.** (with S. Hoşten) Ideals of adjacent minors. *Journal of Algebra* **277** (2004), 615-642.

We give a description of the minimal primes of the ideal generated by the 2×2 adjacent minors of a generic matrix. We also compute the complete prime decomposition of the ideal of adjacent $m \times m$ minors of an $m \times n$ generic matrix when the characteristic of the ground field is zero. A key intermediate result is the proof that the ideals which appear as minimal primes are, in fact, prime ideals. This introduces a large new class of mixed determinantal ideals that are prime.

- 3.** (with M. Develin) Markov bases of binary graph models. *Annals of Combinatorics*, **7** (2003), 441-466.

This paper is concerned with the topological invariant of a graph given by the maximum degree of a Markov basis element for the corresponding graph model for binary contingency tables. We describe a degree four Markov basis for the model when the underlying graph is a cycle and generalize this result to the complete bipartite graph $K_{2,n}$. We also give a combinatorial classification of degree two and three Markov basis moves as well as a Buchberger-free algorithm to compute moves of arbitrary given degree. Finally, we compute the algebraic degree of the model when the underlying graph is a forest.

- 2.** Algebraic Geometry and Combinatorics of Hierarchical Models. M. A. thesis, San Francisco State University, 2002.

This thesis is about algebraic, geometric, and combinatorial aspects of hierarchical models arising in statistical analysis. In particular, we consider the matrix $A_{\Delta,d}$ of the linear transformation that computes the marginals of a k -way tables, and the vector configuration gotten from the columns of this matrix. In Chapter 2, we study the structure of the polyhedral cone that is the positive hull of the columns of this matrix, and give a complete description of this cone for binary cyclic models. In Chapter 3, we study the toric ideal arising from the matrix $A_{\Delta,d}$ and describe a method to construct Gröbner bases for these ideals when the underlying simplicial complex is reducible. We conclude with some open problems in this area.

1. (with S. Hoşten) Gröbner bases and polyhedral geometry of reducible and cyclic models. *Journal of Combinatorial Theory: Series A* **100** (2002) no. 2, 277-301.

This article studies the polyhedral structure and combinatorics of polytopes that arise from hierarchical models in statistics. It also shows how to construct Gröbner bases of toric ideals associated to a certain subset of such models.