I study algebraic statistics, the application of commutative algebra and algebraic geometry to statistical modeling. Earlier in my PhD, my focus was on pure algebraic geometry, but I have since discovered that I find real-world applications highly motivating, so in 2011 I switched to algebraic statistics and began working with Bernd Sturmfels. I now consider myself a mathematical statistician, with a strong background in algebraic geometry to help me along the way.

In section 1, I'll give a brief introduction to algebraic statistics in the form of a simple example. In section 2, I'll outline my three research publications in applied algebra: one on hidden Markov models (used in machine learning), one with with J. Morton on matrix product quantum states (used in condensed matter physics), and one with M. Brunelli and M. Fedrizzi on the analytic hierarchy process (used in group decision-making). Sections 3 and 4 cover my current and proposed research projects, and section 5 describes my software development experience and some research-related organizational initiatives I've undertaken during my PhD.

1 What is algebraic statistics?

Many of the most commonly used statistical models are finite-dimensional families of probability distributions parametrized by algebraic maps, which are called *algebraic statistical models*. Gaussian models, exponential families, hidden Markov models, phylogenetic tree models, directed and undirected graphical models, structural equation models, and deep belief networks are all algebraic statistical models. For an introduction and overview of some biological applications, see *Algebraic Statistics for Computational Biology* (Pachter and Sturmfels, 2005).

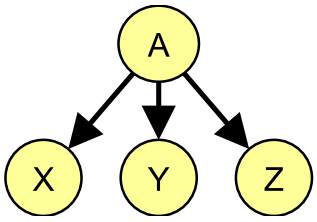


Figure 1: Naive Bayes with four observables

Example. Consider a coin A and parameters $a_i = \Pr(A = i)$ for $i = 0, 1$ specifying its distribution. Suppose the outcome of A determines the distributions of three other coins, X, Y, and Z before they are flipped. That is, parameters $x_{ij} = \Pr(X = j|A = i)$, $y_{ik} = \Pr(Y = k|A = i)$, and $z_{i\ell} = \Pr(Z = \ell|A = i)$, determine the *effect* of A on each of the other coins. These dependencies are depicted in the graph at the left, called a *causal diagram*. When this process runs, each outcome $(A, X, Y, Z) = (i, j, k, \ell)$ has some probability $p_{i,j,k,\ell}$ of occurring, given by the polynomial expression

$$\begin{aligned} p_{ijk\ell} &= \Pr(A = i, X = j, Y = j, Z = \ell) \\ &= a_i x_{ij} y_{ik} z_{i\ell} \end{aligned} \tag{1}$$

Since $a_0 + a_1 = 1$ and $x_{j0} + x_{j1} = 1$, etc., if we choose the seven parameters a_1, x_{j1}, y_{k1}, and $z_{\ell 1}$ freely in $[0, 1]$, then the other seven parameters are uniquely determined. As we vary the free parameters, we obtain different probability tables p according to (1), thus defining a polynomial map $\phi : [0, 1]^7 \to \mathbb{R}^{2\times2\times2\times2}$. The set of $2 \times 2 \times 2 \times 2$ probability tables p which can be *explained* or *modeled* as arising from such a causal diagram of coins is hence the image of ϕ.

Many statistical properties of this model translate to algebraic or geometric properties of the map ϕ. For example, pardoning jargon for the moment, we have the following dictionary of non-trivial equivalences:

Algebraic geometry	**Statistics**
ϕ is injective	The parameters can always be learned with sufficient data.
ϕ has smooth fibres	The Bayesian Information Criterion (BIC) will accurately penalize this model in model selection algorithms.
The signed topological Euler characteristic of $\overline{\text{image}}(\phi)$ is 1	There is 1 critical point in maximum likelihood estimation from generic data.

Most statistical models have more complicated geometry than this one, and have correspondingly more subtle statistical behavior. Such is the case for hidden Markov models, which I'll describe next.

2 Recent papers

2.1 Binary hidden Markov models and varieties

Hidden Markov models (HMM) are machine learning models with diverse applications, including natural language processing, gesture recognition, genomics, and Kalman filtering of physical measurements in robotics and aeronautics. An HMM treats a series of observed phenomena, such as words being recorded by a microphone, as arising from a series of *hidden* or *unobserved* variables, such as the English text in the mind of a speaker that she is hoping to produce on-screen. An HMM-based learning algorithm updates its past beliefs about hidden variables based on present measurements of observables. For example, if a computer thinks you've said "ice cream", and you then say "loudly", for grammatical reasons, it may update its previous opinion to "I scream" while you are still speaking.

What exactly is a hidden Markov model? An HMM of length n involves n hidden *nodes* (random variables) H_i and n visible nodes V_i. We will write $\text{HMM}(\ell, m, n)$ for an HMM of length n where the hidden nodes have ℓ states and the visible nodes have m states. In any HMM, the variables *affect* each other according to a causal diagram with some parameter matrices π, T, and E, respectively of size $1 \times \ell$, $\ell \times \ell$, and $\ell \times m$. A diagram with $n = 4$ is depicted in Figure 2. The parameter matrices determine how the probabilistic effects work, according to the formulae

$$\pi_i = \Pr(H_i = 0), \quad T_{ij} = \Pr(H_t = j \,|\, H_{t-1} = i), \quad E_{ij} = \Pr(V_t = j \,|\, H_t = i)$$

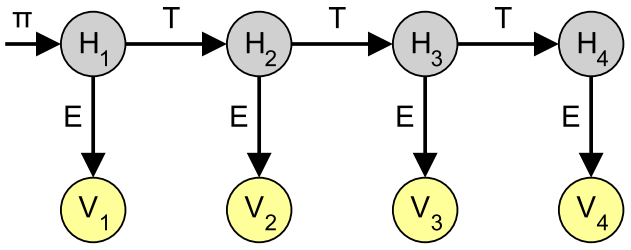


Figure 2: An HMM of length 4

Given the parameter matrices π, T, and E, any particular m-ary string $v = (v_1, \ldots, v_n)$ has a certain probability $p_v = P(V = v \mid \pi, T, E)$ of being *observed*, so we obtain an $m \times m \times \cdots \times m$ table of probabilities, p. The set of all tables p which can arise from such a causal process is denoted by $\text{HMM}(\ell, m, n)$. The entries of such p are forced to satisfy some implicit polynomial equations, and one can ask for constructive description of the set of all such equations, which is called an *ideal*.

What did I prove? Since the initial work of Bray and Morton [1], it has remained an open question to construct this ideal of polynomial equations satisfied by a given HMM. The ideal of $\text{HMM}(2, 2, 3)$ was

determined by Schönhuth [10], and in the process of solving the next simplest case, I discovered:

Theorem 1 (C-, 2012) *All but a measure-zero subset of* $\mathrm{HMM}(2,m,n)$ *can be parametrized by a single generically injective polynomial map* $U \to \Delta_p^{2^n-1}$ *with an explicitly known, rational inverse formula, where* $U \subseteq \mathbb{R}^5$ *is a* 5*-dimensional open set cut out by known algebraic inequalities. In geometric terms, the Zariski closure of* $\mathrm{HMM}(2,m,n)$ *in* $\mathbb{P}^{2^n-1}$ *is a rational projective variety.*

The proof makes use of classical invariant theory results of Sibirskii [11] and others on so-called *trace algebras*. Using this new parametrization, along with a new coordinate system called *cumulant coordinates* developed by Sturmfels and Zwiernik [13], I determined

Theorem 2 (C-, 2012) *Inside the hyperplane* $\sum_v p_v = 1$*, the ideal of polynomial equations satisfied by every* $p \in \mathrm{HMM}(2,2,4)$ *is minimally generated by 21 homogeneous quadric and 29 homogeneous cubic equations. Each of these* 50 *equations can be used to derive* $(n-3)(n-6)\cdots(n-3\lfloor\frac{n}{3}\rfloor)2^{m-1}$ *polynomial equations satisfied by* $\mathrm{HMM}(2,m,n)$ *for each* $m \geq 2$ *and* $n \geq 4$*.*

Why is this important? HMM are an important test case for the application of algebraic techniques to statistical modeling. In particular, parameter estimation and model selection methods used in applications of HMM do not explicitly take into account their algebraic constraints, so there may be significant performance gains to be achieved by considering their geometry in this way.

2.2 Constraints on matrix product state models of entangled qubits

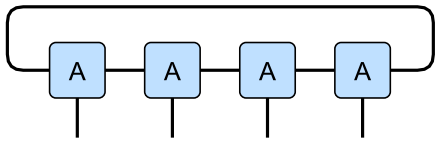


Figure 3: An MPS of length 4 with periodic boundary conditions

Matrix product state (MPS) models are used in condensed matter physics to express an entangled quantum state tensor in terms of a combination of simpler tensors connected by *virtual bonds*. For example, in the left hand diagram, if A is a $d \times D \times D$ tensor, the resulting MPS ψ is a $d \times d \times d \times d$ tensor Ψ — one d for each free "wire" – whose entries are given by

$$\psi_{i_1 i_2 i_3 i_4} = \sum_{j \in \{0,1\}^4} A^{j_1}_{i_1 j_2} A^{j_2}_{i_2 j_3} A^{j_3}_{i_3 j_4} A^{j_4}_{i_4 j_1}.$$

MPS can be used to represent "stable" states of matter, and so classifying such states reduces to understanding the set of tensors representable as MPS [3]. The closure of this set is an algebraic variety, and in this paper, we study its geometry.

What did I prove? Using a reparametrization technique similar to that used on HMM above, I derived some polynomial constraints which must be satisfied by MPS of various formats.

Theorem 3 (C-, Morton, 2012) *A four-qubit state* Ψ *is a limit of binary periodic translation invariant MPS if and only if the following irreducible polynomial in its entries vanishes:*

$$\begin{aligned}
&\psi_{1010}^2\psi_{1100}^4 - 2\psi_{1100}^6 - 8\psi_{1000}\psi_{1010}\psi_{1100}^3\psi_{1110} + 12\psi_{1000}\psi_{1100}^4\psi_{1110} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1110}^2 \\
&+ 2\psi_{0000}\psi_{1010}^3\psi_{1110}^2 + 16\psi_{1000}^2\psi_{1010}\psi_{1100}\psi_{1110}^2 - 4\psi_{0000}\psi_{1010}^2\psi_{1100}\psi_{1110}^2 - 16\psi_{1000}^2\psi_{1100}^2\psi_{1110}^2 \\
&+ 4\psi_{0000}\psi_{1010}\psi_{1100}^2\psi_{1110}^2 - 4\psi_{0000}\psi_{1100}^3\psi_{1110}^2 - 4\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1110}^3 + 8\psi_{0000}\psi_{1000}\psi_{1100}\psi_{1110}^3 \\
&- \psi_{0000}^2\psi_{1110}^4 + 2\psi_{1000}^2\psi_{1010}^3\psi_{1111} - \psi_{0000}\psi_{1010}^4\psi_{1111} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1100}\psi_{1111} + 4\psi_{1000}^2\psi_{1010}\psi_{1100}^2\psi_{1111} \\
&+ 2\psi_{0000}\psi_{1010}^2\psi_{1100}^2\psi_{1111} - 4\psi_{1000}^2\psi_{1100}^3\psi_{1111} + \psi_{0000}\psi_{1100}^4\psi_{1111} - 4\psi_{1000}^3\psi_{1010}\psi_{1110}\psi_{1111} \\
&+ 4\psi_{0000}\psi_{1000}\psi_{1010}^2\psi_{1110}\psi_{1111} + 8\psi_{1000}^3\psi_{1100}\psi_{1110}\psi_{1111} - 8\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1100}\psi_{1110}\psi_{1111} \\
&- 2\psi_{0000}\psi_{1000}^2\psi_{1110}^2\psi_{1111} + 2\psi_{0000}^2\psi_{1010}\psi_{1110}^2\psi_{1111} - \psi_{1000}^4\psi_{1111}^2 + 2\psi_{0000}\psi_{1000}^2\psi_{1010}\psi_{1111}^2 - \psi_{0000}^2\psi_{1010}^2\psi_{1111}^2.
\end{aligned}$$

Theorem 4 (C-, Morton, 2012) *The ideal of constraints on binary translation invariant MPS with periodic boundary conditions is minimally generated by 3 quartics, 27 sextics, and possibly some higher degree polynomials.*

Why is this important? Aside from implicitly classifying states of matter, the proofs illustrate a connection between HMM and MPS first suggested by my coauthor Jason Morton, which we hope will begin a transfer of techniques between graphical statistical modeling and condensed matter physics.

2.3 Proportionality of consistency indices in the analytic hierarchy process

The analytic hierarchy process (AHP) is a procedure for organizing and making complex decisions, especially group decisions, which has been refined and studied extensively since the it was developed in the 1970s by Saaty [9]. The penultimate step of the process involves evaluating the consistency of judgements made by participants, which can be represented by a *pairwise comparison matrix*. Various measures of consistency are used, and our paper shows the equivalence of two pairs of such measures.

What did I prove? My main contribution to [2] was a proof that two consistency measures called ρ and the *geometric consistency index* (GCI) are proportional. To compare their traditional definitions, one needs various changes of variables which I will bypass here. Given an $n \times n$ skew-symmetric matrix $Q = (q_{ij})$, arising as the entry-wise logarithm of a pairwise comparison matrix, up to a scalar factor, one can define ρ and GCI by letting $m_i = \text{mean}_j(q_{ij})$, and then

$$\text{GCI}(Q) := n \cdot \binom{n}{3}^{-1} \sum_{1 \le i < j \le n} (q_{ij} + m_j - m_i)^2, \quad \rho(Q) := \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} (q_{ij} + q_{jk} - q_{ik})^2$$

Theorem 5 (Brunelli, C-, Fedrizzi, 2010) *As defined here,* $\text{GCI}(Q) = 4\rho(Q)$ *for all* n, Q.

The proof is elementary but surprisingly non-trivial, requiring nearly two pages of careful term collecting.

Why is this important? If one finds using the GCI that expert A is more consistent than expert B, one *should not* consider an evaluation of ρ to give independent evidence in favor of A; they will always yield the same result! As well, the GCI is computable in $O(n^2)$ time, compared to $O(n^3)$ for ρ, which could prove useful in assessing the consistency of AI-based experts expressing huge numbers of preferences.

3 Current projects

3.1 Tensor representations of discrete data

A binary tree model is a directed tree on a set of binary variables, where each variable is *affected* by its parent according to a 2×2 conditional probability table, in a way similar to HMM. It is typical in applications that one can only observe the leaves of the tree, so one is interested in the joint marginal probability distribution induced on the leaves, called the *observed distribution*. Smith and Zwiernik [12] defined new coordinates called *tree cumulants* which allow for symbolically efficient expression of the observed distribution in terms of linear regression coefficients associated with the interior of the tree.

This spring, I discovered that a very similar parametrization could work for ternary and n-ary trees if the parameters were represented as entries of certain tensors. Since then I have been joined by S. Lin, P. Zwiernik, and L. Weihs on a project to determine the computational and theoretical implications of this method. We have already observed significant performance gains in symbolic computations, and plan to finish the corresponding paper by January, 2013.

3.2 Parameter estimation degrees for noisy-OR models

Human beings are naturally skilled at noticing causal relationships from small numbers of first-person observations. This process is called causal induction. Griffiths and Tenenbaum [6] found that on causal induction tasks, humans behave very similarly to a graphical model selection criterion which assumes only simple causal relationships called *noisy-OR* and *noisy-AND-NOT*.

It is therefore interesting to ask whether parameter estimation is easy or difficult for these models, as a possible explanation of human performance. There are various ways to make this precise, one of which is to compute the *maximum likelihood (ML) degree* of noisy-OR models: the number of critical points arising in the non-convex optimization problem of finding the parameters which make the observed data maximally likely, given the model.

I have been meeting to discuss this problem with Professor Tom Griffiths in the Psychology department here at Berkeley. When only one or two causes are present for a given effect, the ML degree is small: respectively 1 or 2. But when three causes are present, the ML degree takes on the larger value 10, which suggests that human heuristics for causal inference may not be as well suited to dealing with larger numbers of causes. This is not at all conclusive, however, and more research is necessary to determine if there are lower-degree estimators which perform reasonably well for these models. I hope to conclude this work in the spring semester and include the findings in my May 2013 dissertation.

4 Proposed Projects

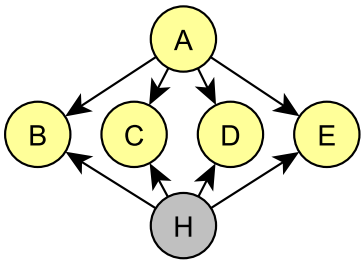
4.1 Algebraic geometry and non-interventional causal inference

Many real-world causal inference problems are so important that we are unable to perform an experiment to test them. For example, given that there is an association between lack of sleep and obesity, for people wishing to control these variables in their own lives, it is essential to know to what extent (1) lack of sleep causes obesity, (2) obesity causes lack of sleep, or (3) other factors cause both. And precisely *because* these variables are so important, we could probably never arrange for 500 randomly selected individuals to deprive themselves of sleep to see if obesity results.

There has been much work in the past decade toward understanding when and how causal inference from purely observational data is possible, due in part to Judea Pearl's seminal textbook, *Causality: Models, Reasoning, and Inference* (2000). The American Institute for Mathematics (AIM) hosted a 5-day workshop in 2010 on "Parameter identification in graphical models", organized by Matthias Drton and Seth Sullivant, which identified 14 open problems on the algebra, geometry, and combinatorics of causal modeling, and I hope to resolve some of them. My background is in algebraic geometry, but increasingly I

view myself as a statistician as I approach these problems, with improvement of existing software packages being a near-term goal. Real-world applications are especially motivating to me considering that 30 years ago non-interventional causal inference was widely considered impossible.

What algebraic geometry has to offer. When all relevant variables are observable, the simplest criteria for testing causal hypotheses without interventions are *conditional independence tests*. When hidden variables are present — and in reality, they usually are — these tests are not enough. However, when one specifies the number of states each hidden variable can attain, sometimes there are algebro-geometric constraints on the set of probability distributions the model can explain, which can be used to test it.



For example, consider the causal structure at the left, where A through E are observable binary variables, and H is a hidden (unobserved) variable. If H is allowed to have any number of states, this model cannot be falsified. However, if H is binary, the set of probability distributions it can generate on the observed variables is an 18-dimensional semi-variety inside the standard 31-dimensional simplex. It therefore satisfies at least 13 independent polynomial constraints, and probably many more, which could be used to test it against purely observational data.

In general, it is not known how to write down the polynomial constraints satisfied by an arbitrary discrete causal model with hidden variables, and the asymptotics of the various statistical tests they imply are are not fully understood. The former is a problem I have some experience with in the case of hidden Markov models, and the latter I have been discussing in correspondence with Xin Gao at York University.

As well, at a subsequent 5-day AIM workshop in 2011 on a new development in algebraic statistics called *Singular Learning Theory*, I met York professor Hélène Massam, and have since corresponded with her on a common interest in how bounds on tensor rank can inform model selection. At the same workshop I also met McGill professor Russell Steele, with whom I share the interest of implementing the so-called *singular Bayesian information criterion* for model selection with real-world models and data.

Parameter identification problems are also amenable to commutative algebra. Sullivant, Garcia-Puente, and Spielvogel [14] explain how to identify causal effects by computing Gröbner bases; see also Meshkat, Eisenberg, and DiStefano [8] for a striking application to ODE models in the biosciences. I would like to spend at least the next several years of my career investigating such methods as they apply to causal inference, and working closely with statisticians to develop software of immediate practical value.

4.2 Noether's problem in algebraic statistics

I was able to compute the defining equations for the binary hidden Markov model of length 4 using a reparametrization which I derived by diagonalizing a certain group action. In this case it turned out that the model was *rational*, meaning that, after omitting a lower-dimensional subset, it can be parametrized by a polynomial map which is generically injective on its entire domain. This has the convenient statistical interpretation that almost every observable distribution has a "unique explanation".

For other graphical models with hidden variables, it is not known whether this is possible. However, there is a rich literature within algebraic geometry on *Noether's problem*, the question of whether the quotient of a vector space by a finite group action is rational. In particular, a great deal is known about quotients by the symmetric groups S_n where $n \leq 4$. This is convenient, because S_4 acts on may phylogenetic tree models by permuting the 4 nucleotides adenine, thymine, guanine and cytosine. Therefore, I would like to

involve more algebraic geometers with a knowledge of Noether's problem in settling rationality questions for these statistical models.

5 Other research activities

5.1 Macaulay2 development experience

Macaulay2 is a software system for research in commutative algebra and algebraic geometry. In spring 2012, I was funded by the DARPA Deep Learning program as a graduate student researcher, and during this time, I learned to use Macaulay2 to investigate algebraic properties of statistical models. In August 2012 at Wake Forest University, I attended the annual Macaulay2 developers' workshop, where I studied its source code and how to create new packages. There I began developing a new package called Tensors.m2 with Claudiu Raicu for manipulating spaces of tensors and data arrays algebro-geometrically. I hope that it will be useful for examining questions relating tensor network models and rank questions to statistics.

5.2 Organizational initiatives

I am generally energetic about organizing seminars and events relating to my research and related interests. From spring 2011 through spring 2012, I co-organized the Berkeley student algebraic geometry seminar, first with Charley Crissman, and later with Andrew Dudzik. Concurrently, in fall 2011, I co-organized the Berkeley algebraic statistics seminar with postdoc Shaowei Lin, which was attended by a diverse audience of graduate students in mathematics, statistics, electrical engineering and computer science. In February 2012, with psychology PhD student Michael Pacer, I organized and spoke in the cognitive science graduate students causal inference symposium, which sparked my current collaboration with cognitive science professor Tom Griffiths on noisy-OR models. During my postdoc, I plan to continue organizing stimulating research activies of this sort, within mathematics and statistics and hopefully crossing other departmental boundaries as well.

5.3 Applied rationality

Between 2011 and 2012, I helped form the Center for Applied Rationality (CFAR) in Berkeley, which is now a registered non-profit organization being run full-time by Julia Galef, Anna Salamon, and Michael Smith. Observational research on human cognitive and behavioral biases — that is, deviation from mathematical norms of inference and optimization — has been booming since the work of Kahneman and Tversky [7]. CFAR's goal is to rapidly develop the art and science of teaching non-experts to *overcome* these biases. Its formation is part of a broader "applied rationality" movement aimed at increasing awareness of human biases, and how to improve upon them with an empirical understanding of the strengths and weaknesses of intuition.

To begin testing curriculum for this purpose, in June 2012, we ran a randomized controlled intervention in the form of a 4-day intensive workshop for one half of a group of forty interested participants. Its effect on various everyday aspects of participants' lives, in comparison with the control group, will be assessed in June 2013. Without control groups, we also ran two 8-day workshops and one other 4-day workshop this past summer, each with over 20 participants, and curriculum continues to be developed.

I am also working with professor Saul Perlmutter, recent recipient of the Nobel Prize in Physics, to develop an undergraduate course called "Sense, Sensibility, and Science", on similar content. Although the mathematical principles of science and reasoning are very precise, helping people to integrate them in a meaningful way with their everyday lives is a highly qualitative and non-trivial task. Nonetheless, recent and quantifiable advances in cognitive science make us better equipped than ever to help people achieve this. I consider developing workshops and courses for this purpose to be an extremely important initiative, and will hopefully continue contributing to these efforts throughout my academic career.

References

[1] Bray, N. and J. Morton (2005). Equations defining hidden Markov models. In *Algebraic Statistics for Computational Biology*, Chapter 11. Cambridge Univerisy Press.

[2] Brunelli, M., A. J. Critch, and M. Fedrizzi (2010). A note on the proportionality between some consistency indices in the AHP. arXiv:1203.6431v1.

[3] Chen, X., Z. Gu, and X. Wen (2011). Classification of gapped symmetric phases in one-dimensional spin systems. *Physical Review B 83*(3), 035107.

[4] Critch, A. J. (2012). Binary hidden Markov models and varieties. arXiv:1206.0500.

[5] Critch, A. J. and J. Morton (2012). Polynomial constraints on representing entangled qubits as matrix product states. arXiv:1210.2812.

[6] Griffiths, T. and J. Tenenbaum (2005). Structure and strength in causal induction. *Cognitive psychology 51*(4), 334–384.

[7] Kahneman, D. and A. Tversky (1973). On the psychology of prediction. *Psychological review 80*(4), 237.

[8] Meshkat, N., M. Eisenberg, and J. J. DiStefano (2009). An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using Gröbner bases. *Mathematical Biosciences 222*(2), 61 – 72.

[9] Saaty, T. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology 15*, 234–281.

[10] Schönhuth, A. (2011). Generic identification of binary-valued hidden Markov processes. arXiv:1101.3712.

[11] Sibirskii, K. (1968). Algebraic invariants for a set of matrices. *Siberian Mathematical Journal 9*, 115–124.

[12] Smith, J. Q. and P. Zwiernik (2010). Tree cumulants and the geometry of binary tree models. arXiv:1004.4360v3.

[13] Sturmfels, B. and P. Zwiernik (2011). Binary cumulant varieties. arXiv:1103.0153.

[14] Sullivant, S., L. D. Garcia-Puente, and S. Spielvogel (2010). Identifying causal effects with computer algebra. Proceedings of the 26th Conference of Uncertainty in Artificial Intelligence.