Algebraic Statistics progress report

Joe Neeman

December 11, 2008

1 A model for biochemical reaction networks

We consider a model introduced by Craciun, Pantea and Rempala [2] for identifying biochemical reaction networks. A motivation for the model can be found in [2]; we will focus only on its properties. Suppose we are given $x_1, \dots, x_m \in \mathbb{R}^d$ in convex position such that there is some hypersurface containing all of the x_i. Then every subset $\sigma \subset [d]$, $|\sigma| = d$ defines a polyhedral cone $C_\sigma \subset \mathbb{R}^d$. We will assume for the moment that all $\binom{m}{d}$ of these cones are full-dimensional (which is the case for generic x_i) and denote the set of these cones by $\mathcal{R}_d$.

Let $\{S_i; i = 1, \dots, n\}$ be the set of full-dimensional cones that can be obtained by intersecting elements of $\mathcal{R}_d$. These will be the states of our model and we will refer to them as *chambers*. The model is parameterized by $\theta \in \mathbb{R}^m$ and the weight (ie. the non-normalized probability) of the i-th state is defined to be

$$g_i(\theta) = \mathrm{vol}(S_i) \sum_{\substack{C_\sigma \in \mathcal{R}_d \\ C_\sigma \supset S_i}} \frac{1}{\mathrm{vol}(C_\sigma)} \prod_{j \in \sigma} \theta_j. \tag{1}$$

The article [2] used a hill-climbing method to solve the maximum likelihood equations in this model. We will study the case $d = 3$, $m = 5$ using algebraic techniques in some detail and we hope, by doing so, to gain a better understanding of the structure of the model.

For this particular example, we can give a detailed combinatorial description of the cones C_σ and S_i. There are 10 cones C_σ and they come in 2 different flavors. Assume the x_i are numbered in a clockwise order. Then, working modulo 5, we have either $\sigma \equiv \{a, a+1, a+2\}$ or $\sigma \equiv \{a, a+2, a+3\}$. There are five instances of each flavor, corresponding to the five choices of a. So that our notation is explicit regarding the flavor of each of these cones, we write $A_a = C_{\{a-1,a,a+1\}}$ and $B_a = C_{\{a,a+2,a+3\}}$.

Turning to the cones S_i, there are 11 cones in three flavors. Five of the S_i share a 2-dimensional face with the cone generated by the x_j; each of these cones is contained in three C_σ cones and we define $T_{ab} = S_i$ if the 2-dimensional face in question is the cone generated by x_a and x_b. Five of the S_i share only a 1-dimensional face with the cone generated by the x_j; each of these cones is contained in four C_σ cones and we define

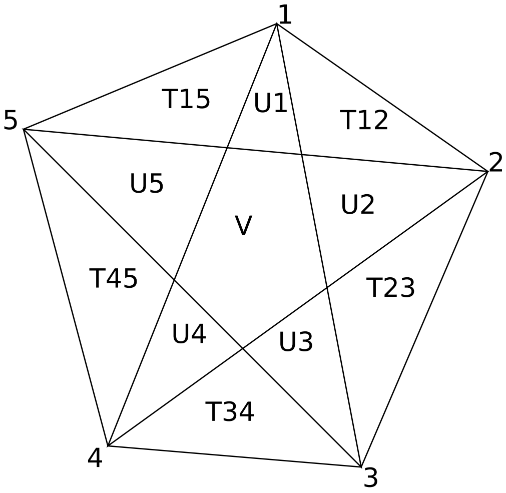

Figure 1: The chambers of our model and their labels.

$U_a = S_i$ if this 1-dimensional face is the ray generated by x_a. The final S_i cone lies entirely in the interior of the cone generated by the x_j and is contained in five C_σ cones. We define V to be this cone. This labeling is shown in Figure 1.

We can write down an implicit form of the model using a computer algebra package (in this case, `Singular`). The ideal generating the model in $\mathbb{P}^{10}$ is

$$\Big\langle \frac{p_5}{\mathrm{vol}(S_5)} - \frac{p_6}{\mathrm{vol}(S_6)} - \frac{p_{10}}{\mathrm{vol}(S_{10})} + \frac{p_{11}}{\mathrm{vol}(S_{11})}, \tag{2}$$

$$\frac{p_4}{\mathrm{vol}(S_4)} - \frac{p_9}{\mathrm{vol}(S_9)} - \frac{p_{10}}{\mathrm{vol}(S_{10})} + \frac{p_{11}}{\mathrm{vol}(S_{11})}, \tag{3}$$

$$\frac{p_3}{\mathrm{vol}(S_3)} - \frac{p_4}{\mathrm{vol}(S_4)} - \frac{p_8}{\mathrm{vol}(S_8)} + \frac{p_{10}}{\mathrm{vol}(S_{10})}, \tag{4}$$

$$\frac{p_2}{\mathrm{vol}(S_2)} - \frac{p_7}{\mathrm{vol}(S_7)} - \frac{p_8}{\mathrm{vol}(S_8)} + \frac{p_{11}}{\mathrm{vol}(S_{11})}, \tag{5}$$

$$\frac{p_1}{\mathrm{vol}(S_1)} - \frac{p_6}{\mathrm{vol}(S_6)} - \frac{p_7}{\mathrm{vol}(S_7)} + \frac{p_{11}}{\mathrm{vol}(S_{11})}, \tag{6}$$

$$q(p_1, \ldots, p_{11})\Big\rangle \tag{7}$$

where $(p_1, \ldots, p_{11})$ are the probabilities assigned to $(T_{12}, T_{23}, T_{34}, T_{45}, T_{51}, U_1, \ldots, U_5, V)$ respectively and q is a homogeneous polynomial of degree 6 with 16 terms. We don't give q explicitly because its coefficients depend in a complex way on the volumes of the cones S_i and C_σ (each coefficient is a degree 16 polynomial in $1/\mathrm{vol}(S_i)$ and $1/\mathrm{vol}(C_\sigma)$). Note that the coefficients of the first 5 generators of the model have a very simple dependence on the volumes of the cones. We will see that this is true in general for any m and d.

We can write out an incidence matrix showing which S_i are contained in which C_σ:

	A_1	A_2	A_3	A_4	A_5	B_1	B_2	B_3	B_4	B_5
T_{12}	1	1	0	0	0	0	0	0	1	0
T_{23}	0	1	1	0	0	0	0	0	0	1
T_{34}	0	0	1	1	0	1	0	0	0	0
T_{45}	0	0	0	1	1	0	1	0	0	0
T_{15}	1	0	0	0	1	0	0	1	0	0
U_1	1	0	0	0	0	1	0	1	1	0
U_2	0	1	0	0	0	0	1	0	1	1
U_3	0	0	1	0	0	1	0	1	0	1
U_4	0	0	0	1	0	1	1	0	1	0
U_5	0	0	0	0	1	0	1	1	0	1
V	0	0	0	0	0	1	1	1	1	1

Let M be this matrix and let $\tilde{M}$ be the matrix obtained from M by multiplying each row by the volume of its associated cone and dividing each column by the volume of its associated cone. Then the parametrization (1) is given by

$$\begin{pmatrix} g_1(\theta) \\ \vdots \\ g_{10}(\theta) \end{pmatrix} = \tilde{M} \begin{pmatrix} \theta_5\theta_1\theta_2 \\ \theta_1\theta_2\theta_3 \\ \theta_2\theta_3\theta_4 \\ \theta_3\theta_4\theta_5 \\ \theta_4\theta_5\theta_1 \\ \theta_1\theta_3\theta_4 \\ \theta_2\theta_4\theta_5 \\ \theta_3\theta_5\theta_1 \\ \theta_4\theta_1\theta_2 \\ \theta_5\theta_2\theta_3 \end{pmatrix}.$$

It follows that the model lives in $\tilde{M}\mathbb{R}^{10} \subset \mathbb{R}^{11}$. We can compute that $\operatorname{rank}(\tilde{M}) = \operatorname{rank}(M) = 6$ (since to obtain $\tilde{M}$ from M, we only multiplied rows and columns by non-zero numbers) and that the first five generators of (2) are linearly independent and orthogonal to every column of $\tilde{M}$. That is, all of the linear constraints in the ideal (2) are explained by the fact that our model is the linear transformation of some other ideal in $\mathbb{R}^{10}$, where the linear transformation has reduced rank.

The rank of the matrix M was studied in [1]. Specifically, the rank of M is $\binom{m}{2}$ for generic points $x_1, \ldots, x_m$ and the general formula for the rank of M is given by

$$\binom{m}{2} + \sum_{k=1}^{d-1} (-1)^k \sum_{Q_k} \binom{m(Q_k) - 1}{d} \tag{8}$$

where Q_k ranges over all k-dimensional affine subspaces spanned by subsets of $x_1, \ldots, x_m$ and $m(Q_k)$ is the number of points in $\{x_1, \ldots, x_m\}$ that lie in Q_k. Also, for the above formula to make sense, we define $\binom{a}{b}$ to be zero whenever $a < b$. For generic points, $m(Q_k) = k$ for all affine subspaces Q_k and therefore (8) reduces to $\binom{m}{2}$.

In addition to studying the rank, [1] also describes explicitly the linear relations between the rows of M (at least, under the assumption that every extreme point of every chamber, with the exception of the original points $x_1, \ldots, x_m$ lies on exactly n distinct affine hyperplanes generated by subsets of $x_1, \ldots, x_m$). Rather

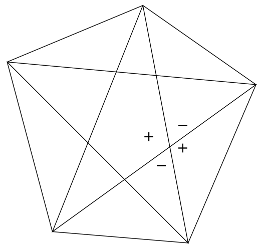

Figure 2: Linear relations between the rows of M

than describe these linear relations in detail, we will just say that all alternating sums of adjacent chambers are zero and give an explanatory picture (Figure 2).

2 Maximum likelihood estimation

Suppose we observe a large number of points in $\mathbb{R}^3$ and we generate a vector of counts $(u_1, \ldots, u_{11}) \in \mathbb{N}^{11}$ where u_i is the number of data points in the cone S_i. We would like to estimate the parameter θ by maximizing the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{11} u_i \log g_i(\theta)$$

subject to the constraints $\theta_i \geq 0$ and $\sum_i g_i(\theta) = 1$. There are, broadly speaking, two exact approaches to this maximization problem: an explicit approach using the parameterization g and an implicit approach using generators of the ideal cut out by g. Neither of these approaches appears to be computationally feasible for this model (they ran out of memory before terminating). Nevertheless, we will briefly describe the two methods.

Our attempt to maximize the likelihood implicitly used algorithm 2.2.9 from [3]. Unfortunately, `Singular` was unable to compute the kernel of the the Jacobian $J(P)$ modulo P (where P is the ideal of the model) even in smaller, degenerate cases.

A naïve explicit approach would undoubtably fare even worse. We can write out the Lagrangian of ℓ and take its partial derivates. These are of the form

$$\frac{\partial}{\partial \theta_j} L(\theta, \lambda) = \sum_{i=1}^{11} u_i \frac{\frac{\partial}{\partial \theta_j} g_i(\theta)}{g_i(\theta)} + \lambda \sum_{i=1}^{11} \frac{\partial}{\partial \theta_j} g_i(\theta).$$

Maximizing the Lagrangian involves taking the polynomial ideal generated by $\prod_i g_i(\theta) \frac{\partial}{\partial \theta_j} L(\theta, \lambda)$ for $j = 1, \ldots, 5$ and saturating it by $\prod_i g_i(\theta)$, which is a degree 33 polynomial with 2375 terms in 5 unknowns.

Instead, we could try to take advantage of the fact that our model is the linear image of a toric ideal in $\mathbb{R}^{10}$. That is, define V to be the toric variety parametrized by

$$(\theta_1, \ldots, \theta_5) \mapsto (\theta_1\theta_2\theta_3, \theta_1\theta_2\theta_4, \ldots, \theta_3\theta_4\theta_5)$$

and attempt to maximize

$$\ell(\theta) = \sum_{i=1}^{11} u_i \log(\tilde{M}x)_i$$

subject to the constraints $x \in V$ and $\sum_i (\tilde{M}x)_i = 1$. The first step is to find an ideal generating V so that we can write down the constraints on x. It turns out that V is cut out by the quadratic binomial ideal

$$I = \langle b_3b_2 - a_1a_4, b_1b_2 - b_4a_4, a_5b_5 - a_1a_4, b_1b_5 - a_2a_4, b_4b_5 - a_2b_2, \\ a_5a_3 - b_4a_4, b_3a_3 - a_2a_4, a_1a_3 - b_4b_5, b_4b_3 - a_2a_5, a_1b_1 - b_4b_3\rangle$$

where we have used the notation $x = (a_1, \ldots, a_5, b_1, \ldots, b_5)$ to emphasize the fact that the first 5 coordinates correspond to the cones A_1 through A_5 and the last 5 coordinates correspond to the cones B_1 through B_5.

Then the ML estimate can be written as the solution of a constrained optimization problem in 21 variables ($a_1, \ldots, b_5$, 10 Lagrange multipliers for the 10 binomial constraints in V and 1 Lagrange multiplier for the constraint that the probabilities sum to 1). Although there has been an increase in the number of parameters, there has been a substantial decrease in the degree of the polynomials involved. Indeed, each partial derivative of the Lagrangian has the form

$$\frac{\partial L}{\partial a_i} = \frac{u_i}{a_i + a_{i+1} + b_{i+3}} + \frac{u_{i-1}}{a_{i-1} + a_i + b_{i+2}} + \frac{u_{i\oplus 5}}{a_i + b_i + b_{i+2} + b_{i+3}} + \text{ constraints} \tag{9}$$

or

$$\frac{\partial L}{\partial b_i} = \frac{u_{i+2}}{b_i + a_{i+2} + a_{i+3}} + \frac{u_{i\oplus 5}}{a_i + b_i + b_{i+2} + b_{i+3}} + \frac{u_{i\oplus 7}}{a_{i+2} + b_i + b_{i+2} + b_{i-1}} \\ + \frac{u_{i\oplus 8}}{a_{i+3} + b_i + b_{i+1} + b_{i+3}} + \frac{u_{11}}{b_1 + b_2 + b_3 + b_4 + b_5} + \text{ constraints} \tag{10}$$

where $i \pm x$ is taken modulo 5 and $i \oplus x$ is defined to be $5 + (i + x \bmod 5)$. The constraints in all cases are sums of quadratic terms where exactly one unknown in each term is a Lagrange multiplier. If we clear denominators, we end up with a system of polynomial equations of which 5 have degree 5, 5 have degree 7, 10 have degree 2 and one has degree 1. Using the linear relations described in Figure 2, we can eliminate 4 of the equations of degree 7, but the resulting system is still intractable.

3 Future work

Given that classical algebraic methods are unable to maximize the likelihood on this model, other methods need to be tried. Craciun et al. proposed a coordinate ascent algorithm for maximizing the likelihood (since the parameterization is multi-linear, it is easy to maximize the likelihood on one coordinate of the parameter). They didn't use this method in their numerical simulations however, because it proved to be too slow (they used a generic hill-climbing algorithm with random restarts). One might hope to improve the situation by using the nice structure of the model (ie. the fact that it is a low-rank linear transformation of a toric variety) to make an initial guess for the starting point of a hill-climbing algorithm. Even better would be a proof that the maximum likelihood estimator lies in a particular region (ie. some sort of "Varchenko's formula"-type result). Then a hill-climbing algorithm could be restricted to that region. For example, if we were using random restarts to avoid getting stuck in local maxima, we could direct the restarts to the regions where the maximizer is allowed to be.

References

[1] Tatiana V. Alekseyevskaya. Combinatorial bases in systems of simplices and chambers. *Discrete Mathematics*, 157(1–3):15–37, 1996.

[2] Gheorghe Craciun, Casian Pantea, and Grzegorz A. Rempala. Algebraic methods for inferring biochemical networks: a maximum likelihood approach, 2008.

[3] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics.* Springer, 2009.