Algebraic Geometry in Algebraic Statistics and Geometric Modeling

Texas Advanced Research Project Grant
15 May 2008 — 14 May 2010

Luis García-Puente & Frank Sottile, co-PIs

Summary

Geometric modeling builds computer models for industrial design and manufacture from basic units, called patches. Many patches, including Bezier curves and surfaces, are pieces of toric varieties, which are objects from algebraic geometry. Statistical models are families of probability distributions used in statistical inference to study the distribution of observed data. Many statistical models, including the log-linear or discrete exponential models used to analyze discrete data, are also pieces of toric varieties. While these connections of geometric modeling and algebraic statistics to algebraic geometry were known, direct connections between these applied subjects were only recently discovered by the PIs [arXiv:/0706.2116]. For example, iterative proportional fitting (IPF) from statistics can be used to compute patches, and linear precision in geometric modeling is related to maximum likelihood estimation in statistics.

Not only are these basic objects the same, but an important tool, the algebraic moment map of a toric variety, is also shared. In geometric modeling it offers a prefered parametrization of a patch, and in algebraic statistics it is the expectation map of a log-linear model. These same objects arise in dynamical systems; in [arXiv:0708.3431] the authors introduce toric dynamical systems whose space of systems is a toric variety with the algebraic moment map giving the steady state solution.

We will exploit these connections to transfer ideas and techniques between these three fields. Krasauskas's multi-sided toric patches [Adv. Comput. Math. 17 (2002)] generalize the classical Bezier patches. Several important problems, including the tuning of a toric patch to achieve linear precision, are amenable to techniques from algebraic statistics. We will also develop IPF into a tool to compute and manipulate toric patches.

While these fields use objects from algebraic geometry, the application of methods from algebraic geometry to these fields is in its infancy. We will deepen these applications. For example, there is a dictionary between toric degenerations and control polyhedra of patches that we will elucidate and use. Another direction is to find applications in modeling and statistics of theoretical work of Sottile on bounds (both lower and upper) on numbers of real solutions to equations. We forsee many avenues of research, including several which are suitable for drawing undergraduates into research and others which could form the foundation for a Masters or PhD thesis.

Student Involvement

This project will link students at SHSU with the rest of Sottile's group at TAMU. We plan both individual supervised research projects for students and vertically integrated research projects between undergraduates, graduate students, postdocs and faculty at both institutions. This will train junior members of our group in the art of collaborative scientific research, in the relevant mathematics, and in the use of computers in mathematical research. It will provide SHSU students, many of whom are the first in their families to attend college or come from underrepresented groups, with the opportunity of being involved in research at a major research university. Senior members will get experience in mentoring.

This project will fund our joint research, two undergraduate students and a Masters student at SHSU and a PhD student and postdoc at TAMU.

A. Research objectives

Recent work in geometric modeling [4], dynamical systems [2], and algebraic statistics [9] has revealed a mathematical object common to these fields—the positive part of a toric variety and its algebraic moment map, μ. In geometric modeling, the inverse of μ is a preferred parametrization of a toric patch, in algebraic statistics it is the maximum likelihood estimate of a log-linear model, and for toric dynamical systems it is the corresponding steady state. We will exploit this connection to transfer ideas and techniques between these fields. This project will involve collaboration between teams of students and researchers at Sam Houston State University and at Texas A&M University. Our team-based, experimental approach represents a new methodology for mathematical research. The outcomes will include the training of junior researchers, published results, and the submission of proposals for federally-funded research.

The problems we propose below are designed to advance our understanding of these fields across a broad front. They will deepen the mathematical foundations of these areas, as well as further the development of tools and introduce new objects which may eventually prove useful in actual applications. Some will also impact pure mathematics as the specific needs of applications often highlight new mathematical structures to investigate.

1. The algebraic moment map. An integer vector $\mathbf{a} = (\alpha_1, \ldots, \alpha_d)$ corresponds to a monomial

$$x^{\mathbf{a}} \ := \ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

for $x_i \neq 0$. (Typically, $x_i \in \mathbb{C}^\times (:= \mathbb{C} - \{0\})$, but we may have $x_i > 0$ so that $x_i \in \mathbb{R}_>$.)

Lists $c = (c_0, c_1, \ldots, c_n) \in \mathbb{R}_>^{n+1}$ of positive numbers (called a *weight*) and $\mathcal{A} = \{\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n\}$ of integer vectors together define a map $\varphi_{c,\mathcal{A}} \colon (\mathbb{C}^\times)^d \to \mathbb{P}^n$ (n-dimensional projective space) via

$$\varphi_{c,\mathcal{A}}(x) \ := \ [c_0 x^{\mathbf{a}_0},\, c_1 x^{\mathbf{a}_1},\, \ldots,\, c_n x^{\mathbf{a}_n}]\,. \tag{1}$$

The closure of the image $\varphi_{c,\mathcal{A}}((\mathbb{C}^\times)^d)$ is the *toric variety* $cX_{\mathcal{A}}$. The closure of $\varphi_{c,\mathcal{A}}(\mathbb{R}^d_>)$ is the *positive part* $cX^+_{\mathcal{A}}$ of $cX_{\mathcal{A}}$, which consists of those points of $cX_{\mathcal{A}}$ with nonnegative coordinates.

The *algebraic moment map* [10] $\mu_{\mathcal{A}} \colon cX_{\mathcal{A}} \to \mathbb{C}^d$ (or $\mu_{\mathcal{A}} \colon \mathbb{P}^n \to \mathbb{C}^d$) is defined by

$$cX_{\mathcal{A}} \ni y = [y_0, y_1, \ldots, y_n] \ \longmapsto \ \mu_{\mathcal{A}}(y) \ := \ \frac{\sum_{i=0}^n \mathbf{a}_i y_i}{\sum_{i=0}^n y_i}\ .$$

The image of $cX^+_{\mathcal{A}}$ is $\Delta_{\mathcal{A}}$, the convex hull of the vectors $\mathbf{a}_i \in \mathcal{A}$. The following result is fundamental.

Theorem. *The algebraic moment map $\mu_{\mathcal{A}}$ is a homeomorphism between $cX^+_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$.*

2. Algebraic Statistics. In algebraic statistics, the positive part Δ_n of $\mathbb{P}^n$ is identified with the probability simplex $\{(p_0,\ldots,p_n) \mid p_i \geq 0 \text{ and } p_0+\cdots+p_n = 1\}$. Algebraic subsets X of Δ_n are *statistical models*. These usually come with a parametrization from a set of meaningful parameters to the family of probability distributions X. A fundamental problem is statistical inference:

Given a point $q \in \Delta_n$ and a model X, find the point $p \in X$ which 'best' agrees with q.

Typically, 'best' means the point corresponding to the maximum likelihood estimate (MLE) for q. *Toric models* have the form $cX^+_{\mathcal{A}}$. For these, the MLE for $q \in \Delta_n$ is the point $p \in cX_{\mathcal{A}}$ such that

$$\mu_{\mathcal{A}}(p) \;=\; \mu_{\mathcal{A}}(q) \qquad \text{and hence} \qquad p \;=\; \mu_{\mathcal{A}}^{-1}(\mu_{\mathcal{A}}(q))\,.$$

Iterative proportional fitting (IPF) [3] is a numerical algorithm to compute $\mu_{\mathcal{A}}^{-1}$ and thus the MLE.

Geometrically, $\mu_{\mathcal{A}}^{-1}(x)$ is the intersection of a linear subspace $E_{\mathcal{A},x}$ of $\mathbb{P}^n$ with $cX^+_{\mathcal{A}}$. Catanese, *et. al.* [1] defined the *maximum likelihood degree* of the model $cX^+_{\mathcal{A}}$. This is the number of intersection points of $cX_{\mathcal{A}}$ and $E_{\mathcal{A},x}$ (outside of a base locus, for general x). Models with maximum likelihood degree 1 are those for which the MLE p is a rational function of the data q. IPF reaches the desired values after one single iteration of the algorithm for such models.

Problem 1. *How does the convergence of IPF depend upon the maximum likelihood degree?*

Problem 2. *Classify toric models with maximum likelihood degree* 1.

The restriction that $\mathcal{A} \subset \mathbb{Z}^d$ is unnatural for statistics; any set $\mathcal{A} \subset \mathbb{R}^d$ of vectors gives a map $\varphi_{c,\mathcal{A}}$ (1) defined on $\mathbb{R}^d_>$ whose image $cX^+_{\mathcal{A}}$ is a *log-linear model*. While maximum likelihood degree is not defined for $cX^+_{\mathcal{A}}$, we will investigate the following extension of problem 1 to irrational $\mathcal{A}$.

Problem 3. *Determine how the convergence of IPF depends upon $\mathcal{A}$ and the weight c.*

3. Geometric modeling. Geometric modeling builds computer models for industrial design and manufacture from basic units, called *patches*. While the classical Bézier patches are widely used, special applications require more flexible multi-sided patches. Krasauskas [8] generalized Bézier patches to multi-sided *toric patches* to fill this niche. Mathematically, a toric patch is the composition of a parametrization $\Delta_{\mathcal{A}} \to cX^+_{\mathcal{A}}$ with a linear projection π_b given by control points (Figure 1). The parametrization determines the internal structure

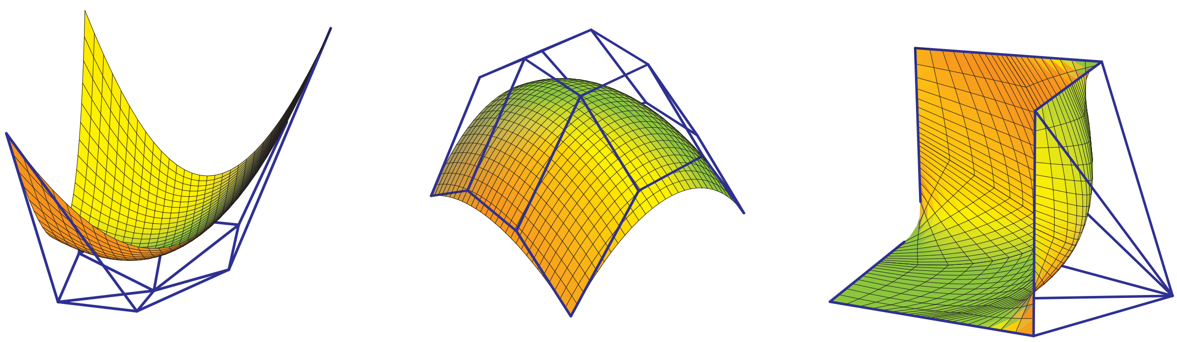

Figure 1. Toric patches and their control polytopes

of the patch and the control points determine its shape. While toric patches share many properties with Bézier patches, it is not known which possess *linear precision*, which is the ability of a patch to replicate affine functions. The PIs showed [4] that every toric patch has a unique reparameterization with linear precision, namely $\mu_{\mathcal{A}}^{-1}$. In particular IPF provides an algorithm to compute toric patches with linear precision.

Problem 4. *Develop IPF as a tool to compute toric patches. Compare its complexity and numerical stability to other (e.g. subdivision) methods* [6]*.*

Integral exponents $\mathcal{A}$ allow parameterizations $\Delta_{\mathcal{A}} \to cX_{\mathcal{A}}^{+}$ by rational functions. Using IPF to compute parameterizations removes this restriction, allowing irrational exponents, $\mathcal{A} \subset \mathbb{R}^d$.

Problem 5. *Develop the theory and properties of irrational patches, those with $\mathcal{A} \subset \mathbb{R}^d$.*

An important open question is when $cX_{\mathcal{A}}$ has *rational linear precision* (i.e. $\mu_{\mathcal{A}}^{-1}$ is rational).

Problem 6. *Classify toric patches $cX_{\mathcal{A}}$ with rational linear precision.*

This is *almost* equivalent to Problem 2 (simplifications we made mask the distinction). While maximum likelihood degree 1 and rational linear precision do differ (we have examples), they do so in a subtle way that we do not yet fully understand. Examples show that it is sometimes possible to 'tune' (alter $\mu_{\mathcal{A}}$) a patch $cX_{\mathcal{A}}$ so that $\mu_{\mathcal{A}}^{-1}$ is a rational function.

Problem 7. *Investigate when a patch $cX_{\mathcal{A}}$ may be tuned to achieve rational linear precision.*

Relatively little of the structure of toric varieties has been exploited in geometric modeling. We plan to change this. For example, every toric variety enjoys a degeneration to a polyhedral complex. Under the map corresponding to control points, this complex becomes the *control polygon/polytope* of the patch (which is manipulated to change the shape of the patch), and the degeneration corresponds to changing weights on the control points.

Problem 8. *Develop and exploit this dictionary between toric degenerations and control polytopes.*

4. Approximation theory and beyond. More generally, a patch $\beta = \{\beta_{\mathbf{a}} \mid \mathbf{a} \in \mathcal{A}\}$ is a collection of nonnegative functions with domain $\Delta_{\mathcal{A}}$. When $\mathcal{A}$ consists of the vertices of $\Delta_{\mathcal{A}}$, the patch has linear precision precisely when β forms a set of *barycentric coordinates* for $\Delta_{\mathcal{A}}$. Thus patches with linear precision are a general form of barycentric coordinates. We believe that, like barycentric coordinates, these more general toric patches may have a role to play in high (> 2) dimensional modeling or as special finite elements in numerical schemes.

Problem 9. *Study the feasibility of these new applications of toric patches.*

When the functions in a patch β are rational (as in toric patches), they define a map $\beta\colon \mathbb{C}^d \to \mathbb{P}^n$ whose image X_β is an algebraic variety. By the geometric analysis of [4], this has rational linear precision when X_β has a remarkably singular intersection (called a *base locus*) with a canonical linear subspace of $\mathbb{P}^n$. Exploiting this unusual geometry is one approach to Problems 2, 6, and 7. Wachspress and then Warren [11] defined rational barycentric coordinates for arbitrary polytopes. Examples show that their base loci have interesting properties.

Problem 10. *Study the base loci of Wachspress coordinates.*

While we do not propose any problems involving toric dynamical systems, we remark that Sottile and Craciun have applied a result from dynamical systems to develop a criterion on control points which guarantees a patch has no self-intersections. This line of research is in its infancy.

B. Methodology

Some problems we propose are amenable to immediate successful completion as the way forward is clear—they are designed to help train the junior members of our group in these topics and in the art of mathematical research. Learning the common background of toric varieties will give the students a foundation for their future studies as toric varieties are fundamental in computational algebraic geometry and ubiquitous in the applications of algebraic geometry. Several problems are ideal for direct investigation by undergraduates. While different problems are suitable for different members of our team, we will create a vertically integrated collaborative research environment that involves several members on most projects. Our emphasis on vertically integrated collaboration, computation, and experimentation is uncommon in mathematics, but we believe it forms an effective paradigm for research and training in mathematics.

Once we develop IPF as a tool for geometric modeling (Problem 4) it is clear how to proceed on Problem 5, as our preliminary investigations have been fruitful and they point the way forward. We will simply need to write a paper and solve the problems which will arise in that venture. Another straightforward project should be Problem 8, as Sottile has used the technique of toric degenerations several times in his pure-mathematical research.

In contrast, Problem 9 is much more open-ended and it will require us to learn more about approximation theory. The workshop on algebraic geometry and approximation theory that Garcia-Puente is organizing this April should be useful in this regard. Both PIs have colleagues who are experts in this area and we plan to make use of their expertise.

Problems 6 and 2 are perhaps the hardest. For example, the partial results [7] on Problem 6 when $d = 2$ used the classification of birational maps of the plane $\mathbb{P}^2$. Very little is known in general about birational maps of $\mathbb{P}^d$ when $d \geq 3$, so treating this case, which is important for high-dimensional modeling, may require us to develop some additional theory about such birational maps.

While these problems require more advanced knowledge, the remaining problems do not and have a completely different flavor. For example, Problem 4 will begin with our developing software tools for using IPF to manipulate patches, and will proceed via a mixture of theory and experimentation to help us understand how this method to compute patches compares to other methods.

Problems 1 and 3 are similar; they involve the dependence of a numerical algorithm (IPF) on the geometrical properties of a statistical model (maximum likelihood degree, or c and $\mathcal{A}$). These projects start with numerical experimentation that should reveal the nature of the dependence of one upon the other, and thus the theorems we will prove.

The last two Problems (7 and 10) will start with an exhaustive study of elementary examples. This approach is particularly well-suited for directed investigation by undergraduates. For Problem 7, understanding how many simple patches may be tuned should lead to a general theory for tuning all patches which we will develop. For Problem 10, the base loci of Wachspress coordinates should depend upon the polytope in an interesting manner. Again, this will be best understood by looking at many examples, which can be generated by the undergraduate members of our team.

C. Research personnel

Frank Sottile, a professor of mathematics at Texas A&M, is known internationally for his work on the applications of algebraic geometry, as well as combinatorics, real algebraic geometry, and Schubert calculus. Computation and computer experiments are central to his work. His research has been continuously funded by the National Science Foundation since 2000, including a CAREER award (2002–2007), and he was a Clay Mathematical Institute Senior Researcher and a Kavli Fellow of the US National Academy of Sciences. Since receiving his PhD in 1994, he has had over 60 research publications, held visiting positions at institutions in North America and Europe, organized about two dozen scientific conferences, given 20 plenary talks and over 150 invited presentations. Sottile is known for his collaborations and mentoring with 60 coauthors, half of whom were postdocs or students. He has supervised 8 postdocs and 5 graduate students.

Luis Garcia-Puente is an assistant professor of mathematics at Sam Houston State University and works on computational algebraic geometry and its applications, including computational statistics, computational biology, systems biology, phylogenetics, and geometric modeling. He wrote a foundational paper in algebraic statistics [5], was a postdoc at the Mathematical Sciences Research Institute and Texas A&M, and has given invited talks at scientific meetings in the US, Mexico, Germany, and Spain. He has has served as a research mentor for undergraduate research programs in Puerto Rico (2001) and in Texas A&M (2005, 2006), directing five research projects. He is a co-organizer of an international workshop on algebraic geometry and approximation theory at Towson University in April 2008.

D. Student involvement

A main goal of this proposal is the training of students, from undergraduates through postdocs. We will assemble a research team consisting of three undergraduate students, and one each of a Masters, PhD, and postdoctoral student. Garcia-Puente will recruit the undergraduates and Masters student from the student body at SHSU, over half of whom are the first in their family to attend college and many of whom come from groups underrepresented in Science. Sottile will either attract one of his current graduate students to work on this research or recruit a new student from among the 80 PhD students in mathematics at TAMU. Likewise, either a current or new postdoc working with him will join our team. His department hires 6-8 postdocs each year.

The undergraduate members of our team will work full-time in the summers on research projects and part-time during the school year. While this pattern of term-time work is unusual for mathematics students involved in research, it is common in other sciences at research universities and we feel necessary to keep them engaged in our project. It will also relieve them of the necessity of taking less engaging work (eg. stacking library books or flipping burgers.)

In addition to the regular direct supervision provided by the PIs, we will hold monthly group meetings, alternating between our institutions which are 56 miles apart. This will expose students from both institutions to each other, not only providing students from SHSU with an experience of the research culture at Texas A&M, but also giving team members from TAMU insight into the culture at SHSU, thereby broadening the education of all.

The different problems we propose to study are designed to appeal to different members of our team. In particular Problems 7 and 10 will enable undergraduate students to begin working on relevant and useful research questions, while numbers 1, 3, 4, and 8 provide a route into this subject for more advanced students. We intend to work collaboratively among ourselves, and each of the problems we identified could lead to joint publications.

E. Institutional commitment and sources of additional support

Texas A&M supports research in the areas covered by this proposal. The Departments of Mathematics, Statistics and Computer Science have major investments in personnel and facilities in areas such as algebraic geometry, computational statistics, geometric modeling, and approximation theory, all of which would benefit from our research, as well as help us in the investigations. Garcia-Puente also has colleagues with expertise in approximation theory. In addition, his department will provide basic institutional support for the students involved, and will fund a *third* undergraduate student, as well as provide some travel support.

We expect to leverage our ARP funds with Federal agencies. Garcia-Puente plans to submit a National Security Agency Young Investigators Grant proposal next year and a National Science Foundation RUI (Research in Undergraduate Institutions) grant under the program in Algebra of the Division of Mathematical Sciences the year after. While Sottile will use some resources from his current NSF grant to support this project, he also plans to apply next year for additional NSF funding to develop numerical software for use in the theory and applications of real algebraic geometry. Part of that proposal will build upon our investigations into the uses of IPF.

F. Budget justification

The yearly stipend for a PhD student at TAMU is $1,850/month, for 11 months. We plan to provide 3 months of summer support at $5,000/month to a postdoc at TAMU during this grant. The stipend for a Masters student at SHSU is $12,000 per academic year and $4,000 during the Summer. The stipend for an undergraduate student at SHSU is $2,500 per academic year and $3,500 during the Summer.

The computational and experimental nature of this research requires that each member be equipped with a portable laptop. With accessories, we expect this to cost no more than $1,500 per machine for 7 machines. The software we need is either free, or can be obtained for a nominal fee using University licenses. We will use PI discretionary funds for software.

We will budget almost $2,000 per member for travel per year. The SHSU students and Garcia-Puente will attend the yearly SACNAS (Society for Advancement of Chicanos and Native Americans in Science) and the AMS/MAA Joint Mathematics Meetings. These are important forums for presenting undergraduate research, and for experiencing the culture of US Science. This will also help Garcia-Puente and the TAMU members attend one or more research conferences in this subject each year. For example, there is a special year at SAMSI (Statistical and Applied Mathematical Sciences Institute) in 2008-9 on algebraic methods in systems biology and statistics and at least yearly conferences in Europe on interactions between algebraic geometry and geometric modeling (Europe is the center of activity on the interactions of algebraic geometry and geometric modeling).

G. Bibliography

References

[1] F. Catanese, S. Hoşten, A. Khetan, and B. Sturmfels, *The maximum likelihood degree*, Amer. J. Math. **128** (2006), no. 3, 671–697.

[2] G. Craciun, A. Dickenstein, A. Shiu, and B. Sturmfels, *Toric dynamical systems*, 2007, `arXiv:0708.3431`.

[3] J. N. Darroch and D. Ratcliff, *Generalized iterative scaling for log-linear models*, Ann. Math. Statist. **43** (1972), 1470–1480.

[4] L. Garcia and F. Sottile, *Linear precision for parametric patches*, 2007, `arXiv:0706.2116`.

[5] L. Garcia, M. Stillman, and B. Sturmfels, *Algebraic geometry of Bayesian networks*, J. Symbolic Comput. **39** (2005), no. 3-4, 331–355.

[6] R. Goldman, *Pyramid algorithms: A dynamic programming approach to curves and surfaces for geometric modeling*, Morgan Kaufmann, 2002.

[7] H.-C. Graf van Bothmer, K. Ranestad, and F. Sottile, *Linear precision for toric surface patches*, 2007, in progress.

[8] R. Krasauskas, *Toric surface patches*, Adv. Comput. Math. **17** (2002), no. 1-2, 89–133, Advances in geometrical algorithms and representations.

[9] L. Pachter and B. Sturmfels (eds.), *Algebraic statistics for computational biology*, Cambridge U. Press, 2005.

[10] F. Sottile, *Toric ideals, real toric varieties, and the algebraic moment map*, Topics in algebraic geometry and geometric modeling, Contemp. Math., vol. 334, Amer. Math. Soc., Providence, RI, 2003, pp. 225–240.

[11] J. Warren, *Barycentric coordinates for convex polytopes*, Adv. Comput. Math. **6** (1996), no. 2, 97–108.