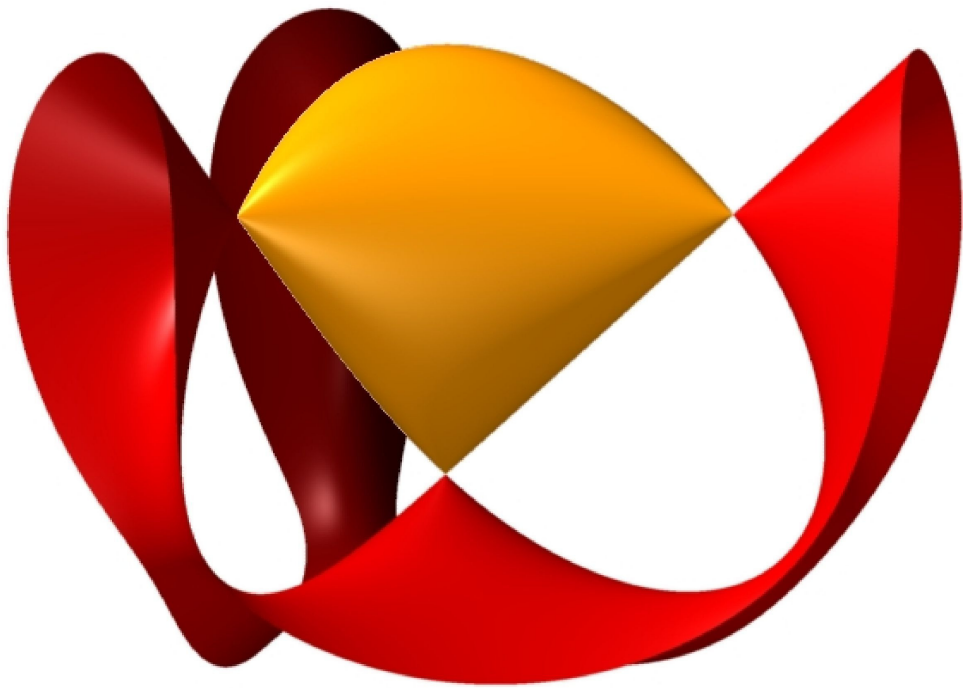
BEYOND LINEAR ALGEBRA

Bernd Sturmfels
University of California at Berkeley



Renaissance Technologies Colloquium, March 25, 2014

Undergraduate Linear Algebra

All undergraduate students learn about *Gaussian elimination*, a general method for solving linear systems of algebraic equations:

Input:

$$\begin{aligned} x + 2y + 3z &= 5 \\ 7x + 11y + 13z &= 17 \\ 19x + 23y + 29z &= 31 \end{aligned}$$

Undergraduate Linear Algebra

All undergraduate students learn about *Gaussian elimination*, a general method for solving linear systems of algebraic equations:

Input:

$$\begin{aligned} x + 2y + 3z &= 5 \\ 7x + 11y + 13z &= 17 \\ 19x + 23y + 29z &= 31 \end{aligned}$$

Output:

$$\begin{aligned} x &= -35/18 \\ y &= 2/9 \\ z &= 13/6 \end{aligned}$$

Solving very large linear systems is central to applied mathematics.

Undergraduate Non-Linear Algebra

Lucky undergraduate students also learn about *Gröbner bases*, a general method for non-linear systems of algebraic equations:

Input:

$$\begin{aligned} x^2 + y^2 + z^2 &= 2 \\ x^3 + y^3 + z^3 &= 3 \\ x^4 + y^4 + z^4 &= 4 \end{aligned}$$

Undergraduate Non-Linear Algebra

Lucky undergraduate students also learn about *Gröbner bases*, a general method for non-linear systems of algebraic equations:

Input:

$$\begin{aligned} x^2 + y^2 + z^2 &= 2 \\ x^3 + y^3 + z^3 &= 3 \\ x^4 + y^4 + z^4 &= 4 \end{aligned}$$

Output:

$$3z^{12} - 12z^{10} - 12z^9 + 12z^8 + 72z^7 - 66z^6 - 12z^4 + 12z^3 - 1 = 0$$

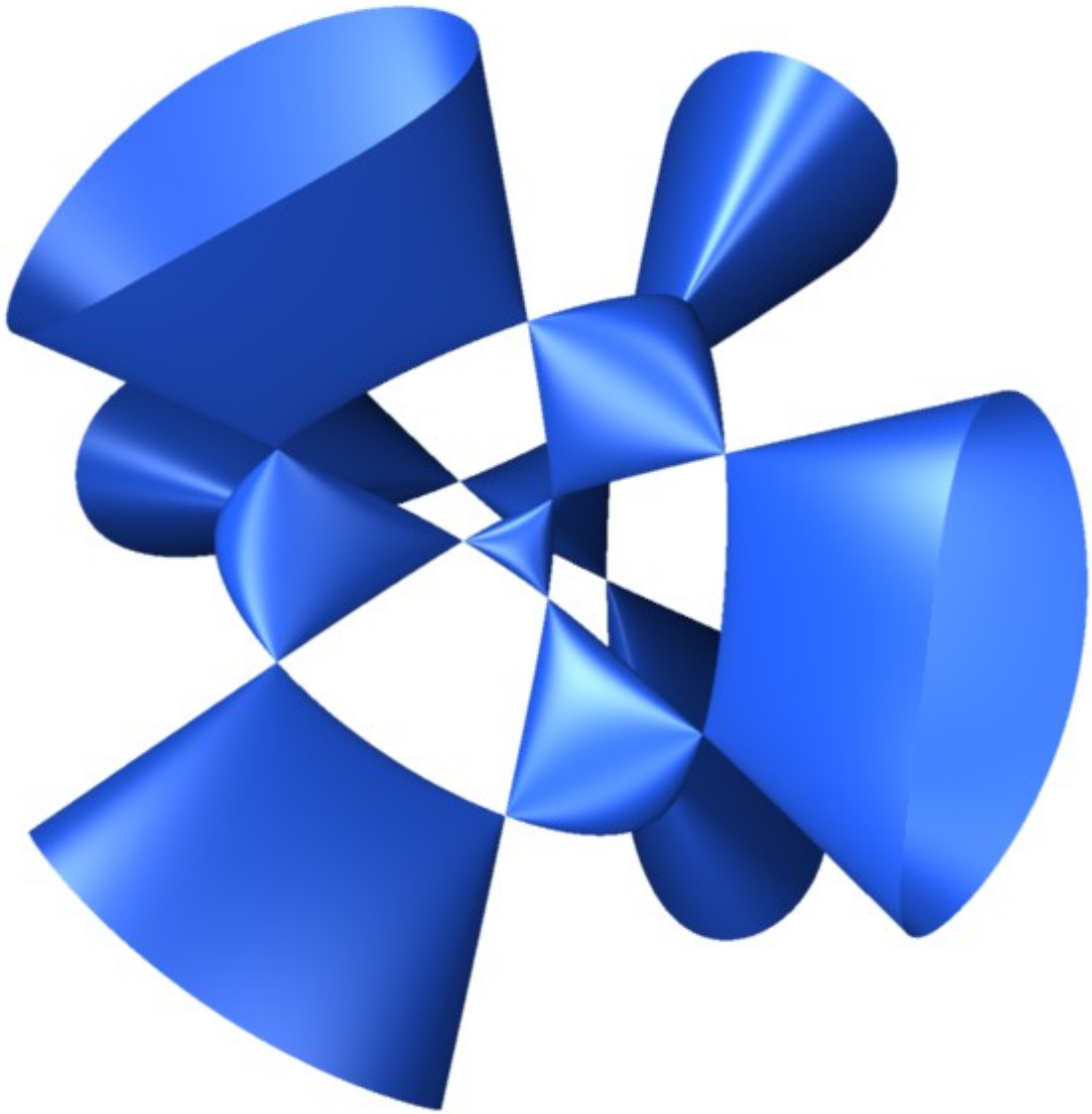
$$\begin{aligned} 4y^2 + (&36z^{11} + 54z^{10} - 69z^9 - 252z^8 - 216z^7 + 573z^6 + 72z^5 \\ &- 12z^4 - 99z^3 + 10z + 3) \cdot y + 36z^{11} + 48z^{10} - 72z^9 \\ &- 234z^8 - 192z^7 + 564z^6 - 48z^5 + 96z^4 - 96z^3 + 10z^2 + 8 = 0 \end{aligned}$$

$$\begin{aligned} 4x + 4y + &36z^{11} + 54z^{10} - 69z^9 - 252z^8 - 216z^7 \\ &+ 573z^6 + 72z^5 - 12z^4 - 99z^3 + 10z + 3 = 0 \end{aligned}$$

Non-linear equations can be intimidating, but ....

Truth and Beauty

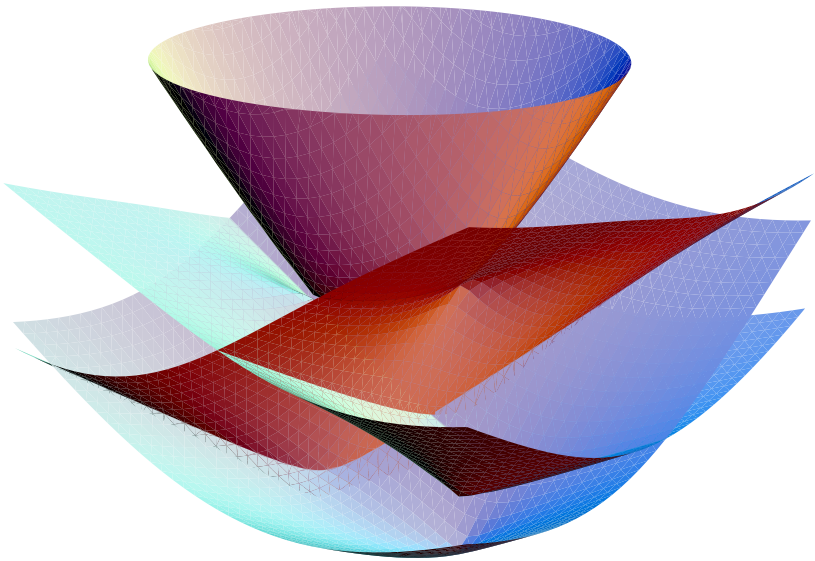
Many models in the sciences and engineering are characterized by polynomial equations. Such a set is an algebraic variety $X \subset \mathbb{R}^n$.

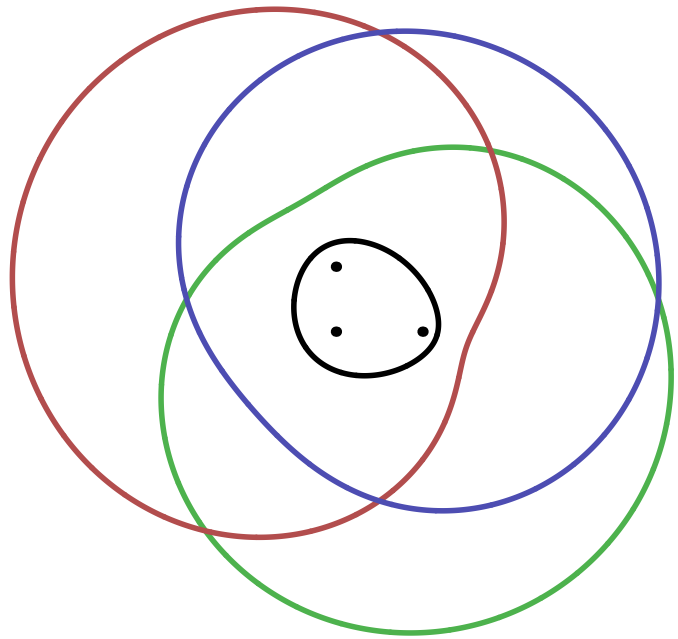


This Lecture

What I shall speak about:

- Tensor Decomposition
- Polynomial Optimization
- Algebraic Statistics



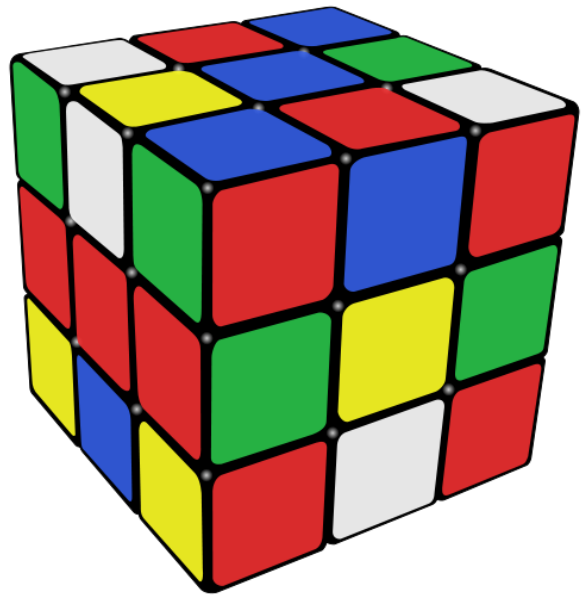


Linear algebra is the foundation of scientific computing and its numerous applications. Yet, **the world is non-linear**. We argue that it pays off to work with models described by non-linear polynomials, while still taking advantage of the power of numerical linear algebra. This leads us to **applied algebraic geometry**. We present a glimpse of this area, by discussing recent advances in tensor decomposition, polynomial optimization, and algebraic statistics.

Topic 1: TENSOR DECOMPOSITION

A tensor is an n-dimensional array of numbers $(x_{i_1 i_2 \cdots i_n})$.
For $n = 1$ this is a vector, and for $n = 2$ this is a matrix.

The vector space of $3\times3\times3$*-tensors has dimension* 27*:*



A tensor has rank one if it is the outer product of n vectors.

$3\times3\times3$*-tensors of rank* 1 *have the form*

$$x_{ijk} = a_i b_j c_k \qquad \text{for } 1 \le i, j, k \le 3.$$

Book: JM Landsberg: *Tensors: Geometry and Applications*, 2012.

Does Watching Soccer Cause Hair Loss?

296 subjects aged 40 to 50 were asked about their hair length and how many hours per week they watch soccer on TV. The data are summarized in a 3×3 matrix. Is it close to having rank 1?

$$U = \begin{array}{c} \\ \leq 2 \text{ hrs} \\ 2\text{–}6 \text{ hrs} \\ \geq 6 \text{ hrs} \end{array} \begin{array}{c} \begin{array}{ccc} \text{lots of hair} & \text{medium hair} & \text{little hair} \end{array} \\ \begin{pmatrix} 51 & 45 & 33 \\ 28 & 30 & 29 \\ 15 & 27 & 38 \end{pmatrix} \end{array}$$

Is there a correlation between watching soccer and hair loss?

Does Watching Soccer Cause Hair Loss?

296 subjects aged 40 to 50 were asked about their hair length and how many hours per week they watch soccer on TV. The data are summarized in a 3×3 matrix. Is it close to having rank 1?

$$U = \begin{array}{c} \\ \leq 2 \text{ hrs} \\ 2\text{–}6 \text{ hrs} \\ \geq 6 \text{ hrs} \end{array} \begin{array}{c} \begin{array}{ccc} \text{lots of hair} & \text{medium hair} & \text{little hair} \end{array} \\ \begin{pmatrix} 51 & 45 & 33 \\ 28 & 30 & 29 \\ 15 & 27 & 38 \end{pmatrix} \end{array}$$

Is there a correlation between watching soccer and hair loss?

Not really. There is a hidden random variable, namely *gender*. The table is the sum of a table for 126 males and one for 170 females:

$$U = \begin{pmatrix} 3 & 9 & 15 \\ 4 & 12 & 20 \\ 7 & 21 & 35 \end{pmatrix} + \begin{pmatrix} 48 & 36 & 18 \\ 24 & 18 & 9 \\ 8 & 6 & 3 \end{pmatrix}.$$

Both matrices have rank 1, hence U has rank 2. We cannot reject

H_0 : *Soccer on TV and Hair Growth are Independent Given Gender.*

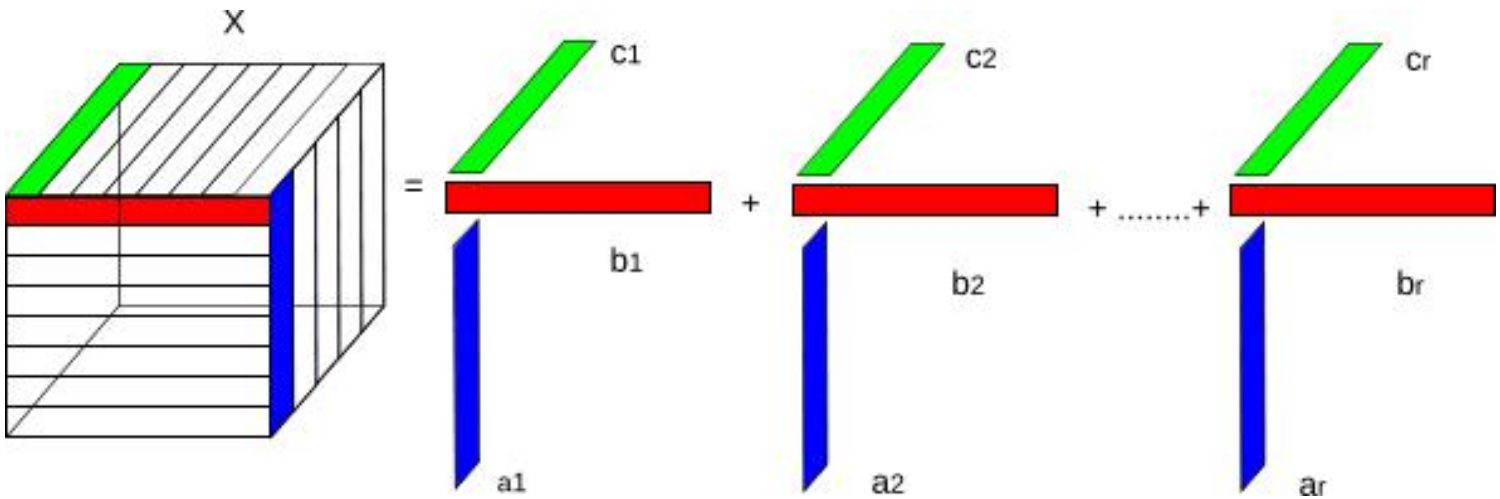
Decomposition and Rank

The soccer-hair example illustrates the importance of decomposing a matrix as a sum of rank 1 *matrices.*

Tensor decomposition:

- Express a given tensor as a sum of rank one tensors.
- Use as few summands as possible.

A tensor has rank r if it is the sum of r tensors of rank one (not fewer).



A nonnegative tensor has nonnegative rank r if it is the sum of r nonnegative tensors of rank one (but not fewer).

Henri Poincaré said ...

Mathematics is the art of giving the same name to different things.

Aren't the following eight things "different"?

- the set of $4\times4\times4$ tensors of rank ≤ 4,
- $x_{ijkl} = a_{1i}b_{1j}c_{1k}d_{1l} + a_{2i}b_{2j}c_{2k}d_{2l} + a_{3i}b_{3j}c_{3k}d_{3l} + a_{4i}b_{4j}c_{4k}d_{4l}$.
- the fourth mixture model for 3 independent random variables,
- the naive Bayes model with four classes,
- the conditional independence model $[\, X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 \mid Y \,]$,
- the fourth secant variety of the Segre variety $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$,
- the general Markov model for the phylogenetic tree $K_{1,3}$,
- superpositions of four pure states in quantum systems.

Allman's Salmon Problem: What are equations describing this set?

Solved recently by Friedland-Gross and Bates-Oeding.

Rank Two

A tensor can be written as a matrix by aggregating indices.

Such a matrix is a flattening of the tensor.

Example: the flattenings of a $2\times3\times5\times7$-tensor are matrices of formats 2×105, 3×70, 5×42, 7×30, 6×35, 10×21 and 14×15.

Theorem (Landsberg-Manivel, Raicu)

A tensor (of any format) has rank ≤ 2 if and only if all its matrix flattenings have rank ≤ 2.

Fine print: this is true up to closure, over the complex numbers.

Theorem (Allman-Rhodes-St-Zwiernik)

A nonnegative tensor (of any format) has nonnegative rank ≤ 2 if and only if it has rank 2 and it is supermodular,

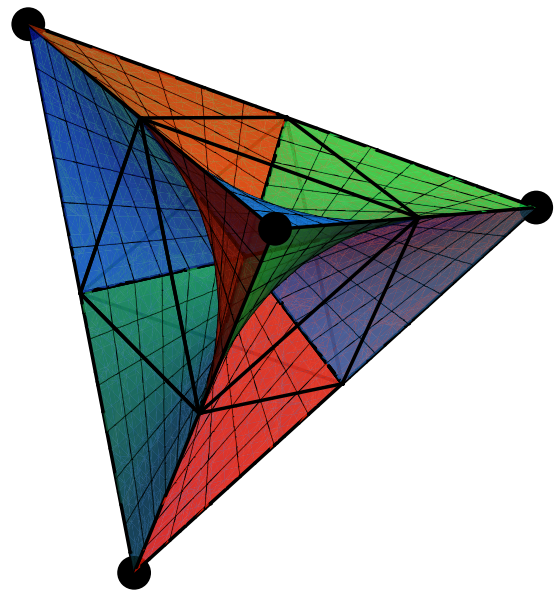
i.e. it satisfies certain quadratic inequalities like $x_{111}x_{222} \geq x_{122}x_{211}$.

Higher Rank

... is more complicated:

Theorem (Strassen 1983)

A $3 \times 3 \times 3$*-tensor has rank* ≤ 4 *if and only if a certain explicit polynomial of degree* 9 *vanishes.*



... but still finite:

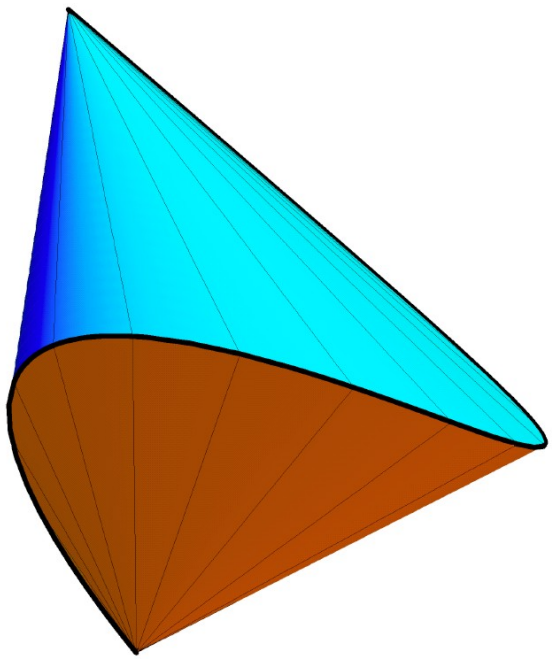
Theorem (Draisma-Kuttler 2014)

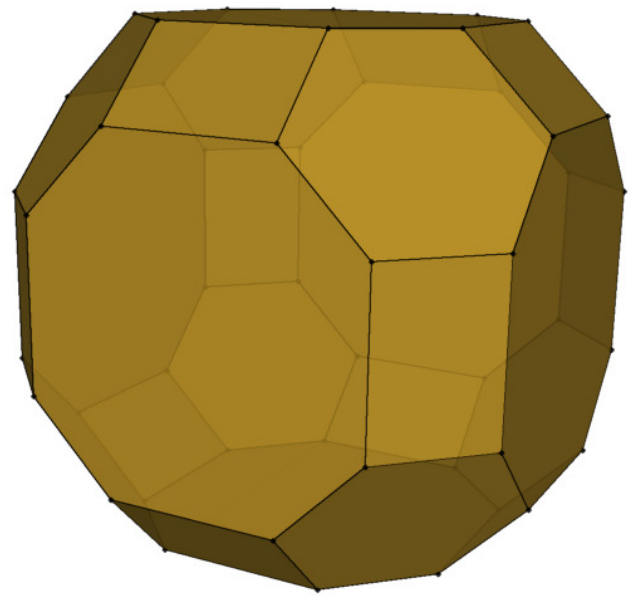
For any given r *there exists an integer* $D(r)$ *such that a tensor has rank* $\leq r$ *if and only if certain polynomials of degree* $\leq D(r)$ *vanish.*

Topic 2: POLYNOMIAL OPTIMIZATION

A **spectrahedron** is the intersection of the cone of positive semidefinite matrices with a linear space.

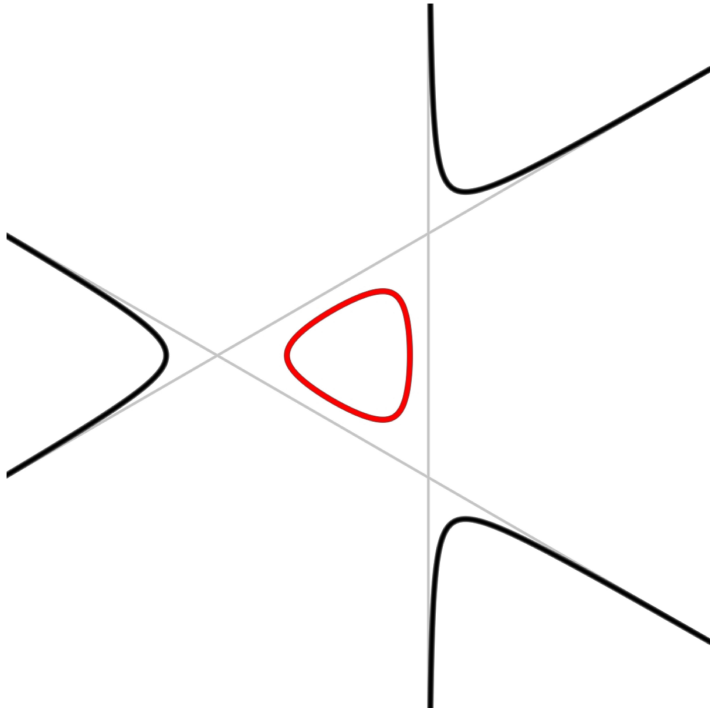
Semidefinite programming is the problem of minimizing a linear function over a spectrahedron. Can be done efficiently.

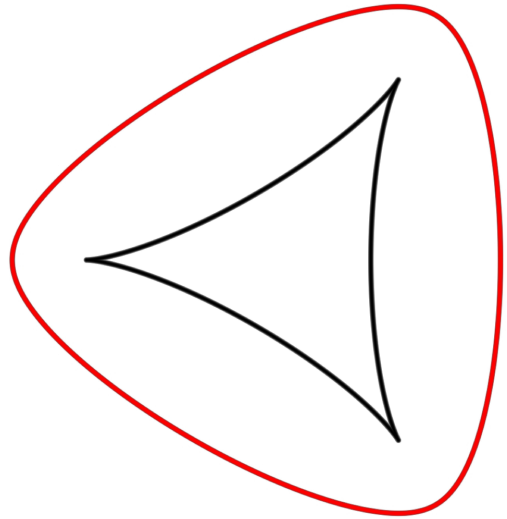


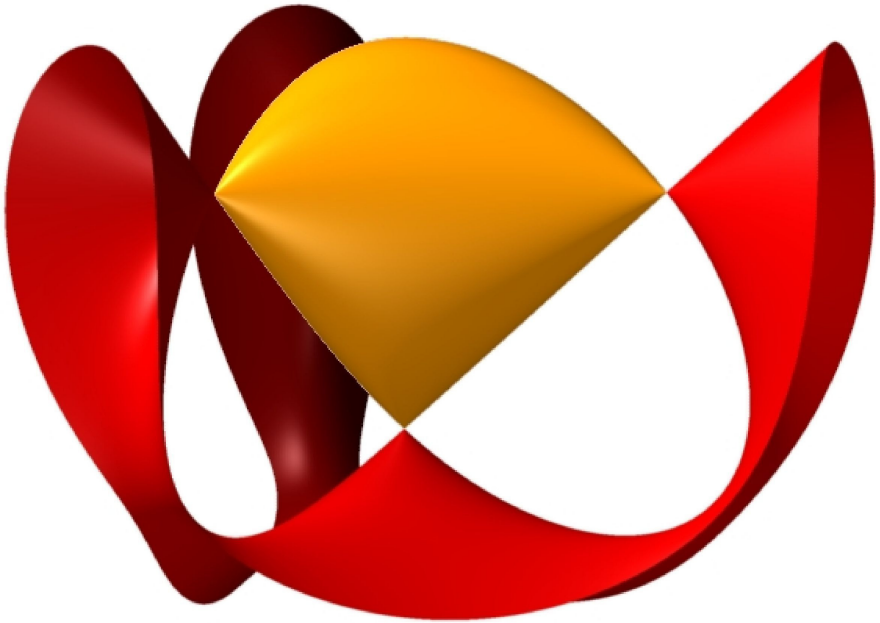


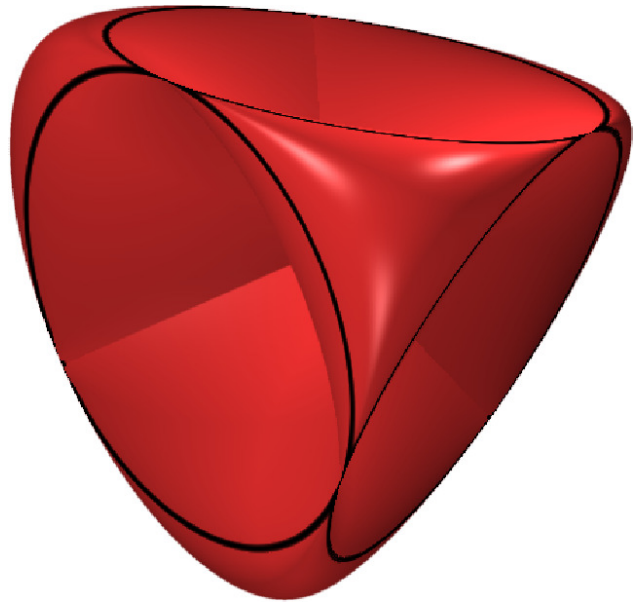
For diagonal matrices: **polyhedron** and **linear programming**.

Duality









Sums of Squares

Let $f(x_1, \ldots, x_m)$ be a polynomial of even degree $2d$.
Goal: compute the global minimum x^* of $f(x)$ on $\mathbb{R}^m$.

This optimization problem is **hard**. It is equivalent to

Maximize λ such that $f(x) - \lambda$ is non-negative on $\mathbb{R}^m$.

Sums of Squares

Let $f(x_1, \ldots, x_m)$ be a polynomial of even degree $2d$.
Goal: compute the global minimum x^* of $f(x)$ on $\mathbb{R}^m$.

This optimization problem is **hard**. It is equivalent to

Maximize λ such that $f(x) - \lambda$ is non-negative on $\mathbb{R}^m$.

The following relaxation gives a lower bound:

Maximize λ such that $f(x) - \lambda$ is a sum of squares of polynomials.

This is much easier. It is a semidefinite program.

The optimal value of the SDP often agrees with the global minimum, and optimal point x^* can be recovered [Parrilo-St 2003].

Book: G. Blekherman, P. Parrilo, R. Thomas:
Semidefinite Optimization and Convex Algebraic Geometry, 2013

SOS Example

Let $m = 1$, $d = 2$ and $f(x) = 3x^4 + 4x^3 - 12x^2$. Then

$$f(x) - \lambda \quad = \quad \begin{pmatrix} x^2 & x & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & \mu - 6 \\ 2 & -2\mu & 0 \\ \mu - 6 & 0 & -\lambda \end{pmatrix} \begin{pmatrix} x^2 \\ x \\ 1 \end{pmatrix}$$

Maximize λ over (λ, μ) s.t. the 3×3-matrix is positive semidefinite.

SOS Example

Let $m = 1$, $d = 2$ and $f(x) = 3x^4 + 4x^3 - 12x^2$. Then

$$f(x) - \lambda \quad = \quad \begin{pmatrix} x^2 & x & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & \mu - 6 \\ 2 & -2\mu & 0 \\ \mu - 6 & 0 & -\lambda \end{pmatrix} \begin{pmatrix} x^2 \\ x \\ 1 \end{pmatrix}$$

Maximize λ over (λ, μ) s.t. the 3×3-matrix is positive semidefinite.

The solution to this semidefinite program is

$$(\lambda^*, \mu^*) \;=\; (-32, -2).$$

Cholesky factorization reveals the sum of squares representation

$$f(x) - \lambda^* \quad = \quad \left((\sqrt{3}\,x - \frac{4}{\sqrt{3}}) \cdot (x+2)\right)^2 \;+\; \frac{8}{3}(x+2)^2.$$

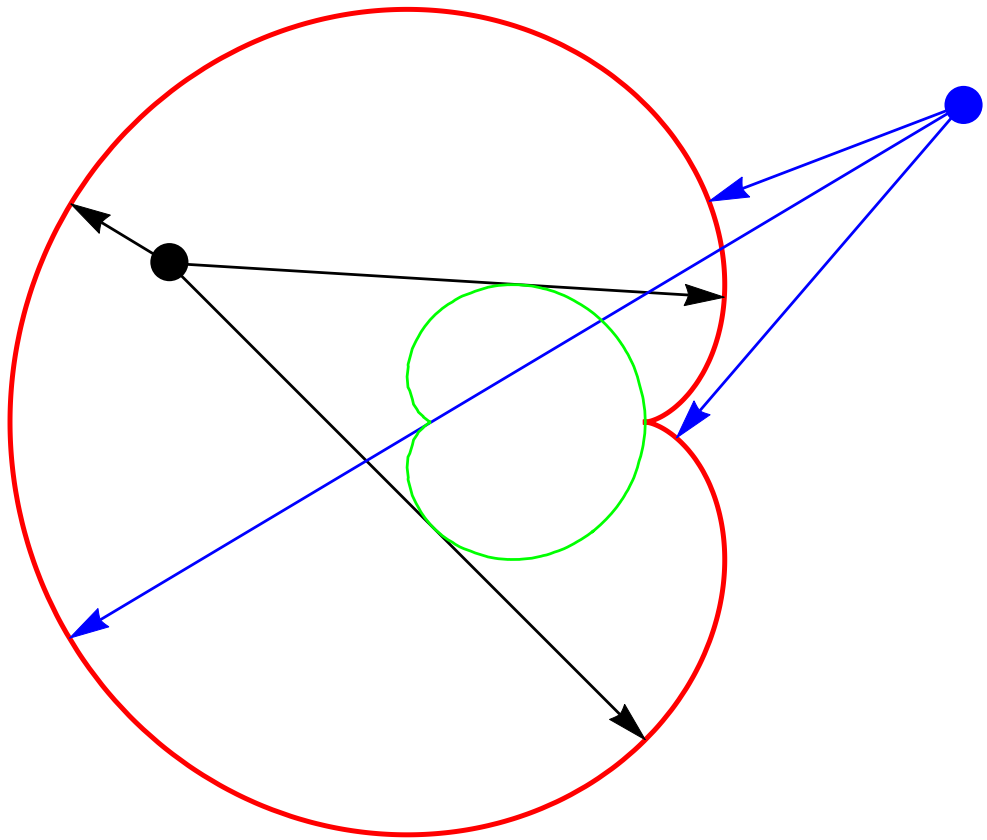
We conclude that the global minimum is $x^* = -2$.

This approach works for a wide range of optimization problems.

Distance Minimization

For any variety X, we study the following optimization problem:

for any data point $u \in \mathbb{R}^n$, find $x \in X$ that minimizes the Euclidean distance function $x \mapsto \sum_{i=1}^{n}(u_i - x_i)^2$.



[Draisma-Horobeț-Ottaviani-St-Thomas:
The Euclidean Distance Degree of an Algebraic Variety, 2013]

Topic 3: ALGEBRAIC STATISTICS

What is a "statistical model" ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parametrized* **set of probability distributions** *of the form* $\{P_\theta \,|\, \theta \in \Theta\}$.

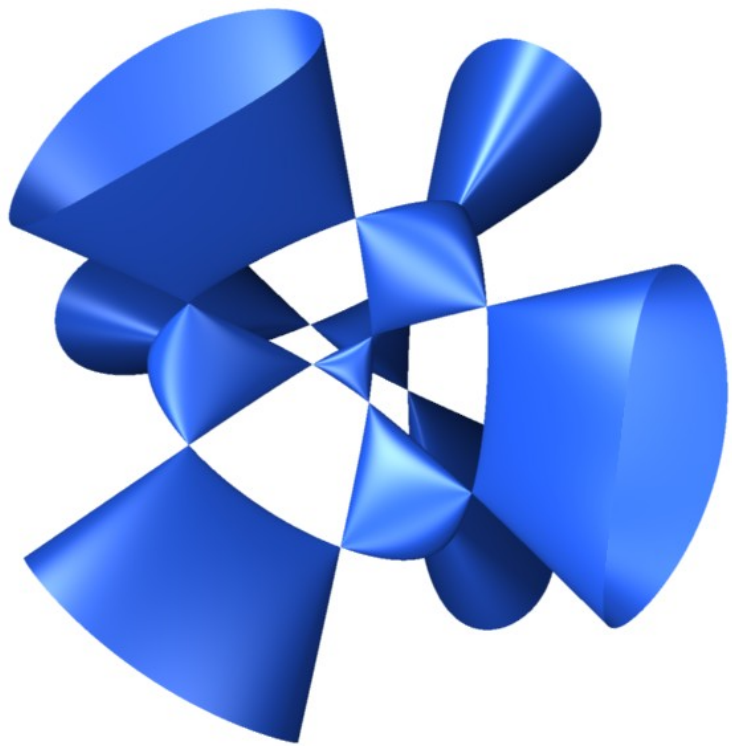
Topic 3: ALGEBRAIC STATISTICS

What is a "statistical model" ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parametrized* **set of probability distributions** *of the form* $\{P_\theta \,|\, \theta \in \Theta\}$.

Geometrically, think of this **set** as

- topological space
- differentiable manifold
- algebraic variety



This leads to subjects such as

- *Topological Data Analysis*
- *Information Geometry*
- *Algebraic Statistics*

(Conditional) Independence

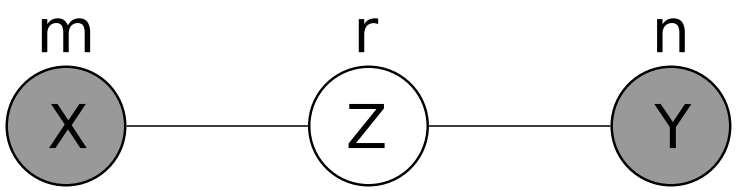
Consider two random variables X and Y having m and n states.
Their joint probability distribution is an $m \times n$-matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix}.$$

whose entries are non-negative and sum to 1.

Let $\mathcal{M}_r$ be the *manifold* of rank r matrices in the simplex Δ_{mn-1}.

Matrices P in $\mathcal{M}_1$ represent independent distributions.



The model $\mathcal{M}_r$ comprises *mixtures* of r independent distributions.
Its elements P represent conditionally independent distributions.

Maximum Likelihood

Suppose i.i.d. samples are drawn from an unknown distribution.
We summarize these data also in a matrix

$$U \quad = \quad \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix} \qquad \text{e.g. soccer/hair}$$

The *likelihood function* is the monomial

$$\ell_U \quad = \quad \prod_{i=1}^{m} \prod_{j=1}^{n} p_{ij}^{u_{ij}}.$$

Maximum Likelihood Estimation: Maximize $\ell_U(P)$ subject to $P \in \mathcal{M}_r$.

The solution $\widehat{P}$ is a rank r matrix. This is the *MLE* for data U.

[Hauenstein-Rodriguez-St: Maximum likelihood for matrices with rank constraints, Journal of Algebraic Statistics, 2014]

3×3-Matrices of Rank 2

Optimization Problem:

Maximize $p_{11}^{u_{11}} p_{12}^{u_{12}} p_{13}^{u_{13}} p_{21}^{u_{21}} p_{22}^{u_{22}} p_{23}^{u_{23}} p_{31}^{u_{31}} p_{32}^{u_{32}} p_{33}^{u_{33}}$ subject to

$$\det(P) = \begin{matrix} p_{11}p_{22}p_{33} - p_{11}p_{23}p_{32} - p_{12}p_{21}p_{33} \\ +p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31} \end{matrix} = 0 \quad \text{and}$$

$$p_{++} = p_{11}+p_{12}+p_{13}+p_{21}+p_{22}+p_{23}+p_{31}+p_{32}+p_{33} = 1.$$

Equations for the Critical Points:

$$\det(P) = 0 \quad \text{and} \quad p_{++} = 1$$

and the rows of the following matrix are linearly dependent:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{21} & u_{22} & u_{23} & u_{31} & u_{32} & u_{33} \\ p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} & p_{31} & p_{32} & p_{33} \\ p_{11}a_{11} & p_{12}a_{12} & p_{13}a_{13} & p_{21}a_{21} & p_{22}a_{22} & p_{33}a_{33} & p_{31}a_{31} & p_{32}a_{32} & p_{33}a_{33} \end{bmatrix}$$

where $a_{ij} = \frac{\partial \det(P)}{\partial p_{ij}}$. These equations have **10** complex solutions.

ML Degree

The ML degree of a statistical model (or an algebraic variety) is the number of critical points of the likelihood function for generic data.

Theorem

The known values for the ML degrees of the rank varieties $\mathcal{V}_r$ are

	$(m,n) =$ $(3,3)$	$(3,4)$	$(3,5)$	$(4,4)$	$(4,5)$	$(4,6)$	$(5,5)$
$r=1$	1	1	1	1	1	1	1
$r=2$	10	26	**58**	**191**	**843**	**3119**	**6776**
$r=3$	1	1	1	**191**	**843**	**3119**	**61326**
$r=4$				1	1	1	**6776**
$r=5$							1

Use the numerical algebraic geometry software **Bertini** to compute all critical points and hence all local maxima.

Jose Rodriguez: Duality Theory for Maximum Likelihood.

A Symmetric 3×3-Matrix

Consider the symmetric matrix model with data

$$U = \begin{bmatrix} 10 & 9 & 1 \\ 9 & 21 & 3 \\ 1 & 3 & 7 \end{bmatrix}$$

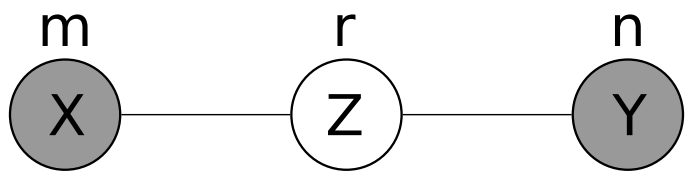
All **6** critical points of the likelihood function are real and positive:

p_{11}	p_{12}	p_{13}	p_{22}	p_{23}	p_{33}	$\log \ell_U(p)$
0.1037	0.3623	0.0186	0.3179	0.0607	0.1368	−82.18102
0.1084	0.2092	0.1623	0.3997	0.0503	0.0702	−84.94446
0.0945	0.2554	0.1438	0.3781	0.4712	0.0810	−84.99184
0.1794	0.2152	0.0142	0.3052	0.2333	0.0528	−85.14678
0.1565	0.2627	0.0125	0.2887	0.2186	0.0609	−85.19415
0.1636	0.1517	0.1093	0.3629	0.1811	0.0312	−87.95759

The first three points are local maxima in Δ_5 and the last three points are local minima. Coordinates can be written in radicals !!

Expectation Maximization

Practitioners use *expectation-maximization* (EM) for maximizing the likelihood function ℓ_U of a hidden variable model, such as



This is a local algorithm in the space Θ of model parameters.

Mathematical problems include
non-identifiability, singularities, local maxima, other fixed points,...

Geometric study in [Kubjas-Robeva-St: Fixed Points of the EM Algorithm and Nonnegative Rank Boundaries, 2013]

Example: The matrix $\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ has rank 3 but nonnegative rank 4.

Conclusion

Think Non-Linearly!