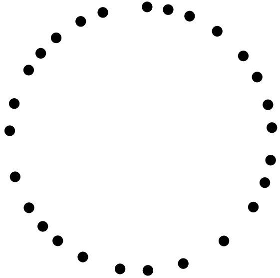
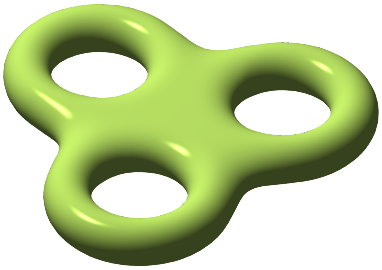
Learning Algebraic Varieties from Samples

Bernd Sturmfels
MPI Leipzig and UC Berkeley





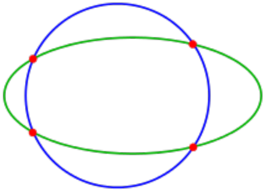
With Paul Breiding, Sara Kališnik and Madeleine Weinstein

TAGS workshop

Wednesday, February 21, 2018

Varieties

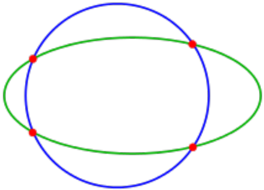
Given polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$, their common zero set is an *algebraic variety* V. It lives in $\mathbb{R}^n$ or $\mathbb{C}^n$. If the f_i are homogeneous then V lives in a projective space $\mathbb{P}^{n-1}_{\mathbb{R}}$ or $\mathbb{P}^{n-1}_{\mathbb{C}}$.



Linear spaces are varieties. Linear Algebra $\hookrightarrow$ Non-Linear Algebra.

Varieties

Given polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$, their common zero set is an *algebraic variety* V. It lives in $\mathbb{R}^n$ or $\mathbb{C}^n$. If the f_i are homogeneous then V lives in a projective space $\mathbb{P}^{n-1}_{\mathbb{R}}$ or $\mathbb{P}^{n-1}_{\mathbb{C}}$.



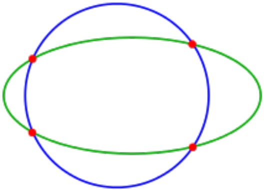
Linear spaces are varieties. Linear Algebra $\hookrightarrow$ Non-Linear Algebra.

The word variety is **not scary**. Data scientists are invited to use it interchangeably with manifold, model, or space.

A line L in $\mathbb{P}^2_{\mathbb{R}}$ is the variety of a linear form $\alpha x + \beta y + \gamma z$.

Varieties

Given polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$, their common zero set is an *algebraic variety* V. It lives in $\mathbb{R}^n$ or $\mathbb{C}^n$. If the f_i are homogeneous then V lives in a projective space $\mathbb{P}^{n-1}_{\mathbb{R}}$ or $\mathbb{P}^{n-1}_{\mathbb{C}}$.



Linear spaces are varieties. Linear Algebra $\hookrightarrow$ Non-Linear Algebra.

The word variety is **not scary**. Data scientists are invited to use it interchangeably with manifold, model, or space.

A line L in $\mathbb{P}^2_{\mathbb{R}}$ is the variety of a linear form $\alpha x + \beta y + \gamma z$.

Quiz: *How many connected components does* $\mathbb{P}^2_{\mathbb{R}} \backslash L$ *have?*

What happens if we take a conic instead of a line?

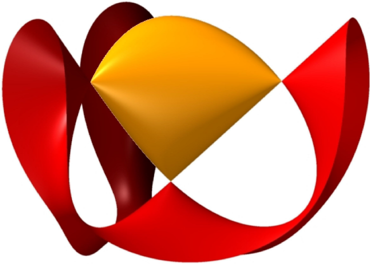
Dimension and Degree

The variety V depends only on the ideal $I = \langle f_1, \ldots, f_r \rangle$.
Algorithms for ideals, e.g. Gröbner bases, reveal geometric features.

Quiz: How to define *dimension* of V?

$$x^2 + y^2 + z^2 - 2xyz - 1$$

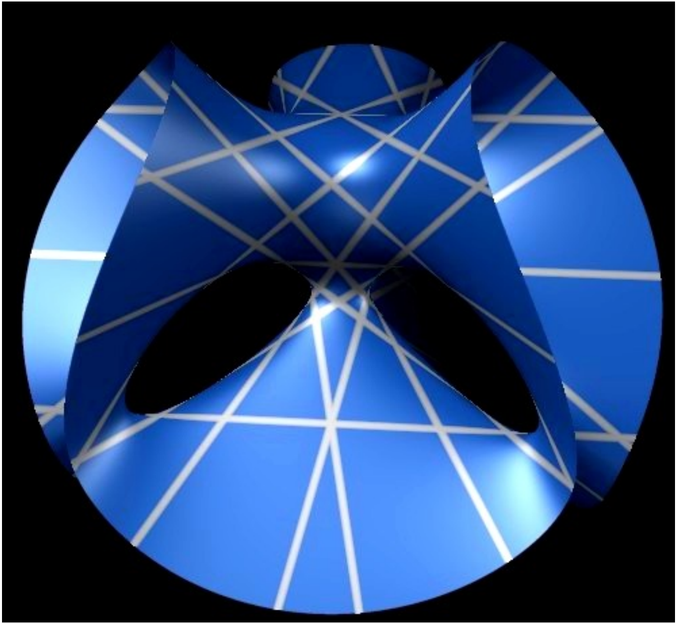
a surface of degree three



Cox, Little, O'Shea: *Ideals, Varieties, and Algorithms*,
Springer Undergraduate Texts in Mathematics, 1993.

The *singular locus* $\mathrm{Sing}(V)$ is a proper subvariety of V, defined by minors of the Jacobian $\left(\partial f_i / \partial x_j\right)$. Hence $V \backslash \mathrm{Sing}(V)$ is a manifold.

27 Lines on the Cubic Surface



The Data

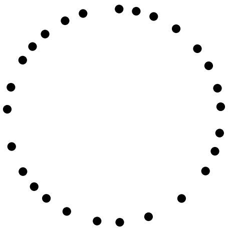
We are given a finite set of points in $\mathbb{R}^n$ or $\mathbb{P}^{n-1}_{\mathbb{R}}$

$$\Omega = \{u^{(1)}, u^{(2)}, \ldots, u^{(m)}\}$$

These are sampled from an unknown variety V.

Exact data? Approximate data?

Goal: Learn the variety V from Ω.



The Data

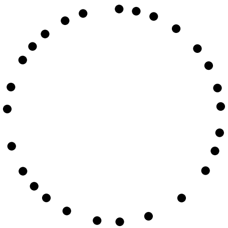
We are given a finite set of points in $\mathbb{R}^n$ or $\mathbb{P}^{n-1}_{\mathbb{R}}$

$$\Omega = \{u^{(1)}, u^{(2)}, \ldots, u^{(m)}\}$$

These are sampled from an unknown variety V.

Exact data? Approximate data?

Goal: Learn the variety V from Ω.



$\longrightarrow$ $(x-3)^2 + (x-5)^2 - 100$

First Question: What is the *dimension* of V?

Sampling

If V is presented by a polynomial parametrization then it is easy to sample. **Quiz**: Does every variety have such a parametrization?

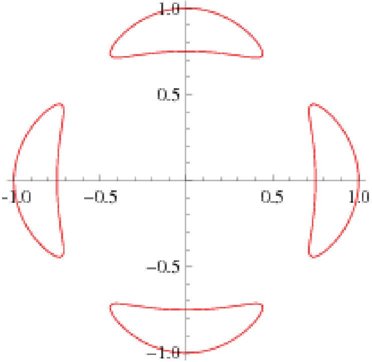
Sampling

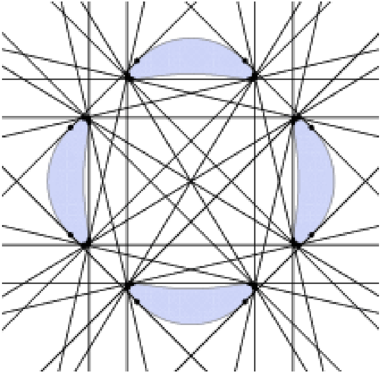
If V is presented by a polynomial parametrization then it is easy to sample. **Quiz**: Does every variety have such a parametrization?

No, smooth plane curves of degree ≥ 3 do not.

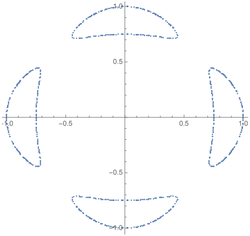
The *Trott curve* is the plane quartic defined

$$12^2(x^4+y^4) - 15^2(x^2+y^2) + 350x^2y^2 + 81.$$

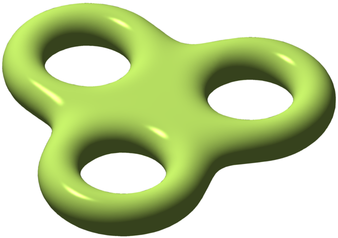


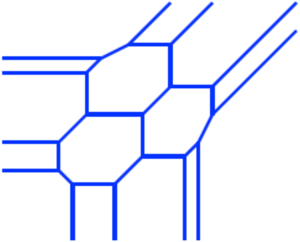


Sampling



How to draw exact samples from a plane quartic ?





A quartic curve.

Good News: While most varieties are not unirational, those arising in applications often are. Nice parametrizations exist.

Three Running Examples

Example 1: The Trott curve.

Example 2: The group of rotations $SO(3)$.
Parametrization by quaternions.
Ideal generated by ten quadrics: $X^T X = \mathrm{Id}, \det(X) = 1.$

Rotations arise in many applications, including computer vision and structural biology.

Quiz: *What is the dimension and degree of this variety?*
How many linearly independent quadrics vanish on $SO(3)$?

Three Running Examples

Example 1: The Trott curve.

Example 2: The group of rotations $SO(3)$.
Parametrization by quaternions.
Ideal generated by ten quadrics: $X^T X = \mathrm{Id}, \det(X) = 1$.

Rotations arise in many applications, including computer vision and structural biology.

Quiz: *What is the dimension and degree of this variety?*
How many linearly independent quadrics vanish on $SO(3)$?

Example 3: The variety of $m \times n$-matrices of rank 1.
Parametrization by "column vector times row vector".
Ideal generated by quadrics, namely the 2×2-minors.

Known to algebraic geometers as Segre variety, and to statisticians as independence model.

Quiz: *Dimension and degree?*

How about tensors?

Our Problem Illustrated

Input: A sample Ω of forty points in $\mathbb{R}^6$:

$(0, -2, 6, 0, -1, 12)$	$(-4, 5, -15, -12, -5, 15)$	$(-4, 2, -3, 2, 6, -1)$	$(0, 0, -1, -6, 0, 4)$
$(12, 3, -8, 8, -12, 2)$	$(20, 24, -30, -25, 24, -30)$	$(9, 3, 5, 3, 15, 1)$	$(12, 9, -25, 20, -15, 15)$
$(0, -10, -12, 0, 8, 15)$	$(15, -6, -4, 5, -12, -2)$	$(3, 2, 6, 6, 3, 4)$	$(12, -8, 9, 9, 12, -6)$
$(2, -10, 15, -5, -6, 25)$	$(5, -5, 0, -3, 0, 3)$	$(-12, 18, 6, -8, 9, 12)$	$(12, 10, -12, -18, 8, -15)$
$(1, 0, -4, -2, 2, 0)$	$(4, -5, 0, 0, -3, 0)$	$(12, -2, 1, 6, 2, -1)$	$(-5, 0, -2, 5, 2, 0)$
$(3, -2, -8, -6, 4, 4)$	$(-3, -1, -9, -9, -3, -3)$	$(0, 1, -2, 0, 1, -2)$	$(5, 6, 8, 10, 4, 12)$
$(2, 0, -1, -1, 2, 0)$	$(12, -9, -1, 4, -3, -3)$	$(5, -6, 16, -20, -4, 24)$	$(0, 0, 1, -3, 0, 1)$
$(15, -10, -12, 12, -15, -8)$	$(15, -5, 6, 6, 15, -2)$	$(-2, 1, 6, -12, 1, 6)$	$(3, 2, 0, 0, -2, 0)$
$(24, -20, -6, -18, 8, 15)$	$(-3, 3, -1, -3, -1, 3)$	$(-10, 0, 6, -12, 5, 0)$	$(2, -2, 10, 5, 4, -5)$
$(4, -6, 1, -2, -2, 3)$	$(3, -5, -6, 3, -6, -5)$	$(0, 0, -2, 3, 0, 1)$	$(-6, -4, -30, 15, 12, 10)$

Task: Learn the variety V.

Our Problem Illustrated

Input: A sample Ω of forty points in $\mathbb{R}^6$:

$(0,-2,6,0,-1,12)$	$(-4,5,-15,-12,-5,15)$	$(-4,2,-3,2,6,-1)$	$(0,0,-1,-6,0,4)$
$(12,3,-8,8,-12,2)$	$(20,24,-30,-25,24,-30)$	$(9,3,5,3,15,1)$	$(12,9,-25,20,-15,15)$
$(0,-10,-12,0,8,15)$	$(15,-6,-4,5,-12,-2)$	$(3,2,6,6,3,4)$	$(12,-8,9,9,12,-6)$
$(2,-10,15,-5,-6,25)$	$(5,-5,0,-3,0,3)$	$(-12,18,6,-8,9,12)$	$(12,10,-12,-18,8,-15)$
$(1,0,-4,-2,2,0)$	$(4,-5,0,0,-3,0)$	$(12,-2,1,6,2,-1)$	$(-5,0,-2,5,2,0)$
$(3,-2,-8,-6,4,4)$	$(-3,-1,-9,-9,-3,-3)$	$(0,1,-2,0,1,-2)$	$(5,6,8,10,4,12)$
$(2,0,-1,-1,2,0)$	$(12,-9,-1,4,-3,-3)$	$(5,-6,16,-20,-4,24)$	$(0,0,1,-3,0,1)$
$(15,-10,-12,12,-15,-8)$	$(15,-5,6,6,15,-2)$	$(-2,1,6,-12,1,6)$	$(3,2,0,0,-2,0)$
$(24,-20,-6,-18,8,15)$	$(-3,3,-1,-3,-1,3)$	$(-10,0,6,-12,5,0)$	$(2,-2,10,5,4,-5)$
$(4,-6,1,-2,-2,3)$	$(3,-5,-6,3,-6,-5)$	$(0,0,-2,3,0,1)$	$(-6,-4,-30,15,12,10)$

Task: Learn the variety V.

Output: For each data point $(x_1, x_2, \ldots, x_6)$, the 2×3-matrix

$$\begin{pmatrix} x_1 & x_2 & x_5 \\ x_4 & x_6 & x_3 \end{pmatrix}$$

has rank 1. Three nice quadrics like $x_1x_3 - x_4x_5$ vanish on Ω.

Hence V is the Segre variety $\mathbb{P}^1 \times \mathbb{P}^2$ in $\mathbb{P}^5$. In statistics, this is the independence model for two random variables: binary and ternary.

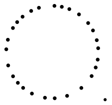
Estimating Dimension

How to use the existing literature on intrinsic dimension ?

Key point: Our sample size $m = |\Omega|$ is fixed and relatively small.

There are *various estimators* $\dim_\bullet(\Omega, \epsilon)$. These depend on a parameter $\epsilon > 0$ and they produce positive real numbers.

Key point: ϵ does <u>not</u> tend to 0. This would be meaningless.



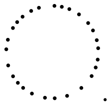
Estimating Dimension

How to use the existing literature on intrinsic dimension ?

Key point: Our sample size $m = |\Omega|$ is fixed and relatively small.

There are *various estimators* $\dim_\bullet(\Omega, \epsilon)$. These depend on a parameter $\epsilon > 0$ and they produce positive real numbers.

Key point: ϵ does <u>not</u> tend to 0. This would be meaningless.



Gold Standard: Principal Component Analysis (PCA)

$$T_1 \cdot \Omega \cdot T_2 = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \quad \text{where} \quad \lambda_1 \geq \cdots \geq \lambda_n \geq 0$$

Take the index k for which the jump from $\log(\sigma_{k-1})$ to $\log(\sigma_k)$ is largest.

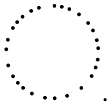
Estimating Dimension

How to use the existing literature on intrinsic dimension ?

Key point: Our sample size $m = |\Omega|$ is fixed and relatively small.

There are *various estimators* $\dim_\bullet(\Omega, \epsilon)$. These depend on a parameter $\epsilon > 0$ and they produce positive real numbers.

Key point: ϵ does <u>not</u> tend to 0. This would be meaningless.



Gold Standard: Principal Component Analysis (PCA)

$$T_1 \cdot \Omega \cdot T_2 = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \quad \text{where} \ \ \lambda_1 \geq \cdots \geq \lambda_n \geq 0$$

Take the index k for which the jump from $\log(\sigma_{k-1})$ to $\log(\sigma_k)$ is largest.

We define the *Nonlinear PCA dimension* by using ϵ to cluster Ω. Then $\dim_{\mathrm{npca}}(\Omega, \epsilon)$ is the average value of k over all clusters.

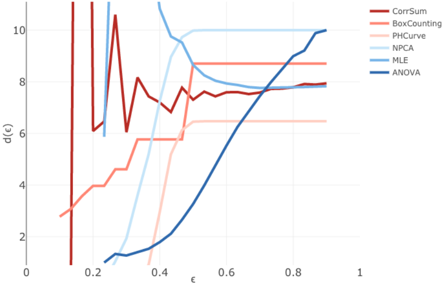
Dimension Diagrams

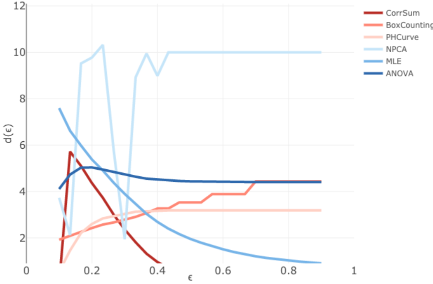
Let $\dim_\bullet$ be one of these six dimension estimators:

- Correlation dimension
- Box counting dimension
- Persistent homology curve dimension
- Nonlinear PCA dimension
- Bickel-Levina dimension
- ANOVA dimension

The *dimension diagram* of the sample Ω is the graph of the map

$$(0,1) \to \mathbb{R}_{\geq 0}\,, \quad \epsilon \mapsto \dim_\bullet(\Omega,\epsilon).$$





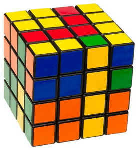
Correlation Dimension

Regard Ω as a finite metric space using the Euclidean metric on $\mathbb{R}^n$ or the *Fubini-Study metric* on $\mathbb{P}^{n-1}_{\mathbb{R}}$, which is defined by

$$\mathrm{dist}_{\mathrm{FS}}(u, v) \;=\; \arccos \frac{|\langle u, v\rangle|}{\|u\|\|v\|} \qquad \text{for } u, v \in \mathbb{R}^n.$$

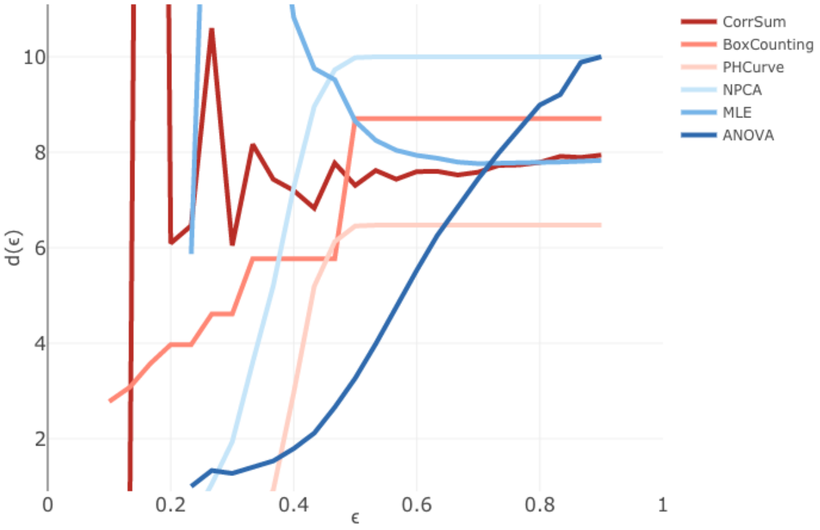
Write $C(\epsilon)$ for the fraction of pairs $\{u^{(i)}, u^{(j)}\}$ having distance $\leq \epsilon$. We set

$$\dim_{\mathrm{cor}}(\Omega, \epsilon) \;:=\; \frac{\log(C(\epsilon))}{|\log(\epsilon)|}.$$



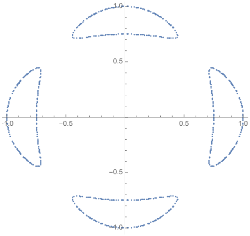
Box Counting Dimension is based on the fraction of boxes occupied by the samples Ω.

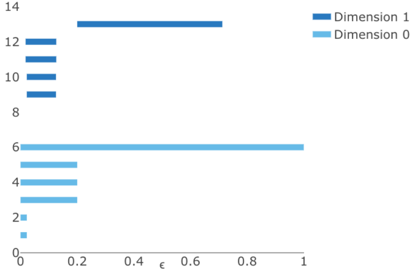
Six Hundred 3×4 Matrices of Rank 2

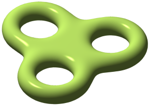


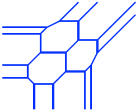
A 9-dimensional variety in $\mathbb{P}^{11}$ of degree 6 defined by four cubics.

Persistent Homology









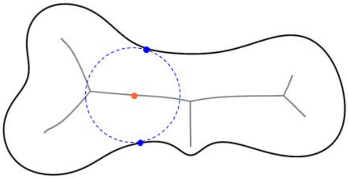
A quartic curve.

Topology of real and complex algebraic varieties is well-studied.
This offers an excellent testing ground for persistent homology.

Reaching the Reach

Niyogi, Smale and Weinberger (2006) give conditions under which a sample Ω reveals the true homology of V, provided V is a compact manifold.

A key ingredient is the *reach* of V.



The *medial axis* of V is the set M_V of points $u \in \mathbb{R}^n$ such that minimum distance from V to u is attained by two distinct points.

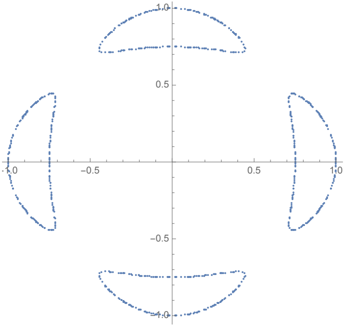
The *reach* $\tau(V)$ is the shortest distance from V to its medial axis M_V. These objects are hard to compute. But it doesn't hurt to try.

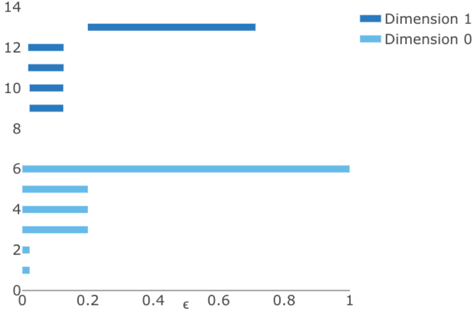
Punchline: *M_V is a variety and $\tau(V)$ is an algebraic number.*

Maddie's poster: Horobeț and Weinstein (2018) deduce

Algebraicity of Persistent Homology

Bar Hopping





$$\frac{1}{8} = 0.125\,, \quad \frac{\sqrt{24025 - 217\sqrt{9889}}}{248} = 0.19941426\ldots\,, \quad \frac{3}{4} = 0.75$$

Tangent Spaces and Ellipsoids

Suppose we know some polynomials that vanish on Ω and V.

Using their Jacobian, we can estimate the tangent space of V at each point $u^{(i)}$. We use ϵ-ellipsoids that are adjusted to these tangent spaces instead of ϵ-balls when computing the Vietoris-Rips complex for persistent homology in `Eirene`.

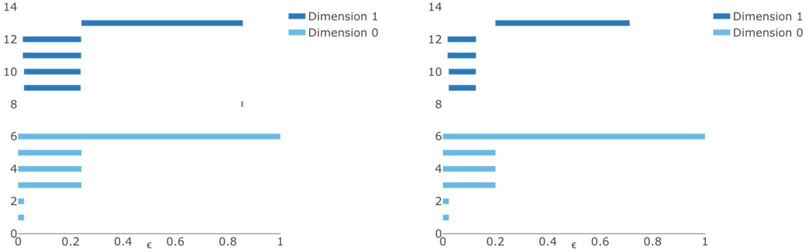


Figure: The left picture shows the ellipsoid-driven barcodes for the Trott curve. The topological features persist longer than when balls are used.

Finding Equations

Let $\mathcal{M}$ be a set of monomials in $S = \mathbb{R}[x_1, \ldots, x_n]$. Write $S_{\mathcal{M}}$ for the subspace with basis $\mathcal{M}$. Examples are all monomials of degree d resp. $\leq d$. The corresponding subspaces $S_{\mathcal{M}}$ satisfy

$$\dim(S_d) = \binom{n+d-1}{d} \quad \text{and} \quad \dim(S_{\leq d}) = \binom{n+d}{d}.$$

Write $U_{\mathcal{M}}(\Omega)$ for the multivariate Vandermonde matrix of format $m \times |\mathcal{M}|$: in the ith row are the values of the monomials in $\mathcal{M}$ at the point $u^{(i)}$. For example, if $n = 1$, $m = 3$, $\Omega = \{u, v, w\}$ then

$$U_{\leq d}(\Omega) = \begin{pmatrix} u^d & u^{d-1} & \cdots & u^2 & u & 1 \\ v^d & v^{d-1} & \cdots & v^2 & v & 1 \\ w^d & w^{d-1} & \cdots & w^2 & w & 1 \end{pmatrix}.$$

Remark: The kernel of $U_{\mathcal{M}}(\Omega)$ is the space $I_\Omega \cap S_{\mathcal{M}}$ of $\mathbb{R}$-linear combinations of $\mathcal{M}$ and that vanish on Ω.

Goal: Learn the ideal I_V of the unknown variety V.

Numerical Linear Algebra

Desirable properties in making an educated guess for $\mathcal{M}$:

(a) *The ideal I_V is generated by its subspace $I_V \cap S_{\mathcal{M}}$.*

(b) *Inclusion of $I_V \cap S_{\mathcal{M}}$ in $I_\Omega \cap S_{\mathcal{M}} = \ker(U_{\mathcal{M}}(\Omega))$ is an equality.*

Note: If $\mathcal{M}$ is too small then (a) fails. If $\mathcal{M}$ is too large then (b) fails.

Requirement (b) imposes a lower bound on the sample size:

$$m \geq |\mathcal{M}| - \dim(I_V \cap S_{\mathcal{M}}).$$

Example: It takes $m \geq \binom{n+2}{2}$ samples to learn quadrics in I_V.

Numerical Linear Algebra

Desirable properties in making an educated guess for $\mathcal{M}$:

(a) *The ideal I_V is generated by its subspace $I_V \cap S_{\mathcal{M}}$.*

(b) *Inclusion of $I_V \cap S_{\mathcal{M}}$ in $I_\Omega \cap S_{\mathcal{M}} = \ker(U_{\mathcal{M}}(\Omega))$ is an equality.*

Note: If $\mathcal{M}$ is too small then (a) fails. If $\mathcal{M}$ is too large then (b) fails.

Requirement (b) imposes a lower bound on the sample size:

$$m \geq |\mathcal{M}| - \dim(I_V \cap S_{\mathcal{M}}).$$

Example: It takes $m \geq \binom{n+2}{2}$ samples to learn quadrics in I_V.

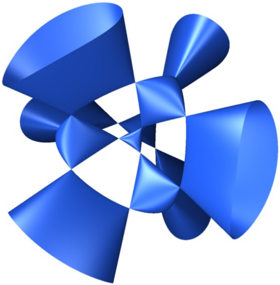
We implemented

three methods for the kernel of the Vandermonde matrix $U_{\mathcal{M}}(\Omega)$

SVD	accurate, fast, but returns orthonormal and hence dense basis.
QR	slightly less accurate and fast than SVD, yields some sparsity.
RREF	no accuracy guarantees, not as fast as the others, gives sparse basis.

Computational Algebraic Geometry

We now have a set $\mathcal{P}$ of polynomials that vanish on Ω, and we hope that it defines the true variety V. What to do with $\mathcal{P}$?



Use symbolic or numerical methods to answer these questions:

1. What is the dimension of V ?
2. What is the degree of V ?
3. Find the irreducible components of V.
   Determine their dimensions and degrees.

Primary Decomposition

A sample of 500 points in $\mathbb{R}^6$ is drawn from a generative model V. The kernel of the 500×210-matrix $U_{\leq 4}(\Omega)$-matrix is 2-dimensional:

$$\mathcal{P} = \{ acf^2 + ad^2f - 2ade^2 - b^2f^2 + 2bd^2e - c^2df + c^2e^2 - cd^3, \\ a^2df - a^2e^2 + ac^2f - acd^2 - 2b^2cf + b^2d^2 + 2bc^2e - c^3d \}$$

Primary Decomposition

A sample of 500 points in $\mathbb{R}^6$ is drawn from a generative model V.
The kernel of the 500×210-matrix $U_{\leq 4}(\Omega)$-matrix is 2-dimensional:

$$\mathcal{P} \;=\; \{\, acf^2 + ad^2 f - 2ade^2 - b^2 f^2 + 2bd^2 e - c^2 df + c^2 e^2 - cd^3,\\ a^2 df - a^2 e^2 + ac^2 f - acd^2 - 2b^2 cf + b^2 d^2 + 2bc^2 e - c^3 d \,\}$$

There are two components of codimension 2, of degrees 3 and 10.
Since $3+10 \neq 16$, the ideal $\langle \mathcal{P} \rangle$ is not radical. Back to finding equations.

Primary Decomposition

A sample of 500 points in $\mathbb{R}^6$ is drawn from a generative model V. The kernel of the 500×210-matrix $U_{\leq 4}(\Omega)$-matrix is 2-dimensional:

$$\mathcal{P} = \{ acf^2 + ad^2f - 2ade^2 - b^2f^2 + 2bd^2e - c^2df + c^2e^2 - cd^3, \\ a^2df - a^2e^2 + ac^2f - acd^2 - 2b^2cf + b^2d^2 + 2bc^2e - c^3d \}$$

There are two components of codimension 2, of degrees 3 and 10. Since $3+10 \neq 16$, the ideal $\langle \mathcal{P} \rangle$ is not radical. Back to finding equations.

The kernel of $U_5(\Omega)$ yields two new quintics and we get $\sqrt{\langle \mathcal{P} \rangle}$. Finally, the kernel of the 500×452-matrix $U_6(\Omega)$ suffices. The prime ideal I_V is generated by 2 quartics, 2 quintics and 4 sextics.

Primary Decomposition

A sample of 500 points in $\mathbb{R}^6$ is drawn from a generative model V. The kernel of the 500×210-matrix $U_{\leq 4}(\Omega)$-matrix is 2-dimensional:

$$\mathcal{P} \,=\, \{\, acf^2 + ad^2f - 2ade^2 - b^2f^2 + 2bd^2e - c^2df + c^2e^2 - cd^3, \\ a^2df - a^2e^2 + ac^2f - acd^2 - 2b^2cf + b^2d^2 + 2bc^2e - c^3d \,\}$$

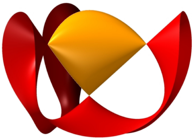
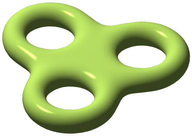
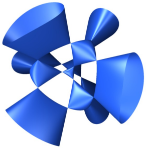
There are two components of codimension 2, of degrees 3 and 10. Since $3{+}10 \neq 16$, the ideal $\langle \mathcal{P} \rangle$ is not radical. Back to finding equations.

The kernel of $U_5(\Omega)$ yields two new quintics and we get $\sqrt{\langle \mathcal{P} \rangle}$. Finally, the kernel of the 500×452-matrix $U_6(\Omega)$ suffices. The prime ideal I_V is generated by 2 quartics, 2 quintics and 4 sextics.

The mystery variety $V \subset \mathbb{R}^6$ is 4×4 Hankel matrices of rank 2 whose antidiagonal entry has been deleted:

$$\begin{bmatrix} a & b & c & x \\ b & c & x & d \\ c & x & d & e \\ x & d & e & f \end{bmatrix} \;=\; \begin{bmatrix} s_1^3 & s_2^3 \\ s_1^2 t_1 & s_2^2 t_2 \\ s_1 t_1^2 & s_2 t_2^2 \\ t_1^3 & t_2^3 \end{bmatrix} \begin{bmatrix} s_1^3 & s_1^2 t_1 & s_1 t_1^2 & t_1^3 \\ s_2^3 & s_2^2 t_2 & s_2 t_2^2 & t_2^3 \end{bmatrix}.$$

Beauty

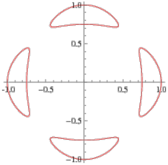
Equations are Beautiful:

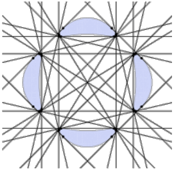
$$
\begin{aligned}
&p_{16}p_{25}p_{34} - p_{15}p_{26}p_{34} - p_{16}\;p_{24}p_{35} + p_{14}p_{26}p_{35} + p_{15}p_{24}p_{36} \\
&-p_{14}p_{25}p_{36} + p_{16}p_{23}p_{45} - p_{13}p_{26}p_{45} + p_{12}p_{36}p_{45} - p_{15}p_{23}p_{46} \\
&+p_{13}p_{25}p_{46} - p_{12}p_{35}p_{46} + p_{14}p_{23}p_{56} - p_{13}p_{24}p_{56} + p_{12}p_{34}p_{56}
\end{aligned}
$$

$$
\begin{gathered}
x_{110}^2x_{001}^2+x_{100}^2x_{011}^2+x_{010}^2x_{101}^2+x_{000}^2x_{111}^2 + 4x_{000}x_{110}x_{011}x_{101}+4x_{010}x_{100}x_{001}x_{111} \\
-2x_{100}x_{110}x_{001}x_{011} - 2x_{010}x_{110}x_{001}x_{101} - 2x_{010}x_{100}x_{011}x_{101} \\
-2x_{000}x_{110}x_{001}x_{111} - 2x_{000}x_{100}x_{011}x_{111} - 2x_{000}x_{010}x_{101}x_{111}
\end{gathered}
$$

$$
\begin{bmatrix}
2d_{1p} & d_{1p}+d_{2p}-d_{12} & d_{1p}+d_{3p}-d_{13} & \cdots & d_{1p}+d_{p-1,p}-d_{1,p-1} \\
d_{1p}+d_{2p}-d_{12} & 2d_{2p} & d_{2p}+d_{3p}-d_{23} & \cdots & d_{2p}+d_{p-1,p}-d_{2,p-1} \\
d_{1p}+d_{3p}-d_{13} & d_{2p}+d_{3p}-d_{23} & 2d_{3p} & \cdots & d_{3p}+d_{p-1,p}-d_{3,p-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
d_{1p}+d_{p-1,p}-d_{1,p-1} & d_{2p}+d_{p-1,p}-d_{2,p-1} & d_{3p}+d_{p-1,p}-d_{3,p-1} & \cdots & 2d_{p-1,p}
\end{bmatrix}
$$

Real Degree and Volume





Theorem (Kinematic formula)

Let V be a smooth projective variety of dimension d in $\mathbb{P}^{n-1}_{\mathbb{R}}$. Then its volume is the volume of $\mathbb{P}^{d}_{\mathbb{R}}$ times the real degree:

$$\mathrm{vol}(V) \;=\; \frac{\pi^{\lceil \frac{d+1}{2} \rceil}}{2\Gamma(\frac{d+1}{2})} \cdot \deg_{\mathbb{R}}(V)$$

$$\textit{where} \quad \deg_{\mathbb{R}}(V) \;=\; \int_{L \in \mathrm{Gr}(n-d-1, \mathbb{P}^{n-1}_{\mathbb{R}})} \#(L \cap V)\, \mathrm{d}\nu$$

is the expected number of intersection points with a linear space.

Example ($n = 2, k = 1$)

The real degree of the projective Trott curve V in $\mathbb{P}^2_{\mathbb{R}}$ equals

$$\deg_{\mathbb{R}}(V) \;=\; 1.88364$$

Multiply with $\mu(\mathbb{P}^1_{\mathbb{R}}) = \pi$ to learn that the length of V is 5.91763.

Software and Experiments

All algorithms are implemented in our `Julia` package

`LearningAlgebraicVarieties`

Please try it out !!

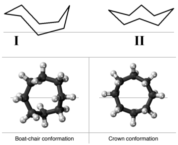
Case Study: Data set with $n = 24$ and $m = 6040$.
Samples $u^{(i)}$ are configurations of 8 points in 3-space.
These represent conformations of cyclo-octane C_8H_{16}.

Constraints:

$$d_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 = \begin{cases} 1 & \text{if } j = i+1, \\ 8/3 & \text{if } j = i+2. \end{cases}$$

Implicit representation: *the* 7×7 *Cayley-Menger matrix has rank* 3:

$$\begin{bmatrix} 2d_{18} & d_{18}+d_{28}-d_{12} & d_{18}+d_{38}-d_{13} & \cdots & d_{18}+d_{78}-d_{17} \\ d_{18}+d_{28}-d_{12} & 2d_{28} & d_{28}+d_{38}-d_{23} & \cdots & d_{28}+d_{78}-d_{27} \\ d_{18}+d_{38}-d_{13} & d_{28}+d_{38}-d_{23} & 2d_{38} & \cdots & d_{38}+d_{78}-d_{37} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{18}+d_{78}-d_{17} & d_{28}+d_{78}-d_{27} & d_{38}+d_{78}-d_{37} & \cdots & 2d_{78} \end{bmatrix}$$



Raising the Bar

The conformation space is the union of a sphere with a Klein bottle glued along two circles. Mod 2 Betti numbers are **1,2,1**.

[Brown *et al.* 2008]   [Martin *et al.* 2010]   [Tausz *et al.* 2014]

Our software confirms this. First finding equations helps a lot:

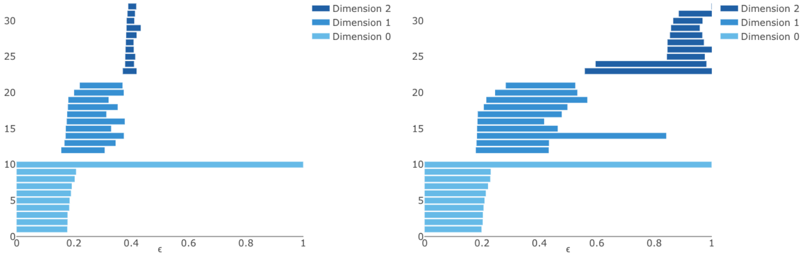
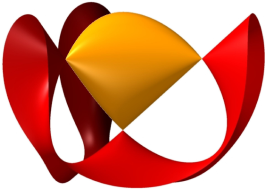
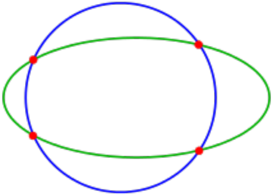


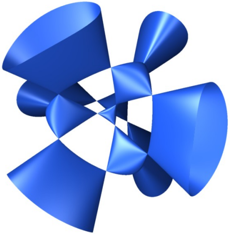
Figure: Persistent homology: standard (left) versus ellipsoid-driven (right)

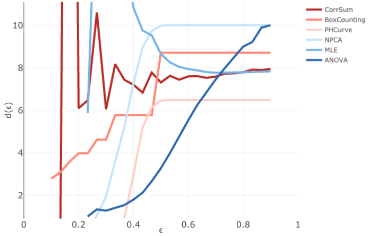
Conclusion

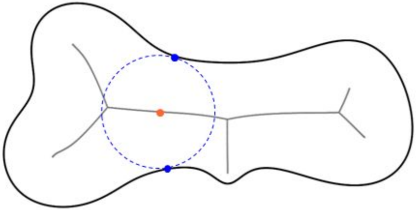
Learning Algebraic Varieties from Samples in one ingredient in Linking **T**opology to **A**lgebraic **G**eometry and **S**tatistics











Many Thanks for Listening

Hanuta Prize

Let $n = 6$, $m = 100$, and $\Omega =$

(1,4,8,4,13,20),(1,14,3,7,2,7),(1,16,6,1,1,10),(1,20,6,16,5,4),(1,20,14,2,2,12),
(2,1,5,3,17,1),(2,3,2,8,12,10),(2,3,6,1,4,3),(2,4,1,18,6,3),(2,5,2,6,4,4),
(2,6,1,20,8,14),(2,6,15,2,8,9),(2,7,1,7,3,7),(2,8,5,18,15,15),(2,11,12,7,10,13),
(2,12,17,6,10,9),(3,2,1,17,13,3),(3,5,4,16,14,2),(3,5,5,14,17,5),(3,8,2,4,4,8),
(3,11,2,2,5,17),(3,11,7,8,7,7),(3,15,3,3,4,17),(3,18,12,5,4,4),(4,7,5,4,4,2),
(4,12,14,10,12,1),(4,19,1,17,3,10),(4,20,8,15,10,20),(4,20,12,13,9,6),(5,1,1,6,16,2),
(5,3,1,5,5,2),(5,6,3,8,14,12),(5,7,1,6,8,10),(5,8,7,10,15,10),(5,9,9,3,13,18),
(5,9,11,9,16,9),(5,11,5,6,5,5),(5,13,13,3,8,13),(5,17,2,19,4,6),(5,20,15,17,15,9),
(6,3,18,2,20,4),(6,4,10,1,10,5),(6,5,7,5,19,10),(6,6,3,8,5,1),(6,8,13,2,7,5),
(6,9,3,17,11,8),(6,9,6,1,6,8),(6,9,12,1,12,16),(6,13,16,14,20,6),(6,16,7,16,13,16),
(6,18,3,8,5,11),(7,5,2,3,4,2),(7,5,16,1,6,2),(7,10,18,4,17,14),(7,12,5,16,9,4),
(7,17,5,1,4,9),(7,18,5,18,12,18),(7,19,15,7,7,4),(7,20,10,19,13,10),(8,7,8,1,8,6),
(8,12,16,3,16,18),(8,15,2,12,8,12),(8,15,4,9,12,18),(8,16,8,11,9,7),(9,4,2,8,13,4),
(9,8,1,20,7,4),(9,8,2,2,5,4),(9,13,2,15,3,1),(9,19,16,18,18,6),(10,4,5,12,20,2),
(10,6,6,3,8,3),(10,10,20,4,15,7),(11,4,12,3,20,4),(12,4,3,12,18,3),(12,6,4,9,16,5),
(12,7,2,15,18,8),(12,7,7,1,13,7),(12,13,1,6,6,6),(12,16,20,3,12,11),(12,16,20,10,17,6),
(12,19,8,3,12,17),(13,11,16,2,10,6),(13,14,4,5,7,6),(13,16,6,20,14,8),(14,10,8,5,11,5),
(15,10,10,7,10,2),(15,18,6,14,18,16),(15,20,16,5,10,8),(16,7,1,5,3,1),(16,12,1,20,3,1),
(16,15,17,12,20,6),(16,20,14,4,18,19),(17,5,18,2,14,2),(17,13,1,20,12,8),(19,3,4,5,13,1),
(19,4,7,7,17,1),(19,9,10,1,18,8),(19,10,11,4,12,4),(20,10,3,20,18,6),(20,20,15,12,15,6)

`https://math.berkeley.edu/~bernd/hanuta.html`

Task: Name the projective variety V in $\mathbb{P}^5_{\mathbb{R}}$.

The first correct answer wins a **prize**: Ten Hanuta **bars**

Students and coauthors of Bernd are not eligible to win. But they are encouraged to help others.