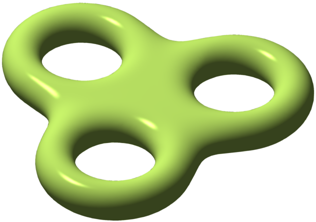
Learning Algebraic Varieties from Samples

Sara Kališnik

November 22, 2018



With Paul Breiding, Bernd Sturmfels and Madeleine Weinstein

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

Research

The theory, algorithms, and software of linear algebra are familiar tools across mathematics, the applied sciences, and engineering. This ubiquity of linear algebra masks the fairly recent growth of nonlinear algebra in mathematics and its application. The proliferation of nonlinear methods, notably for systems of multivariate polynomial equations, has been fueled by recent theoretical advances, efficient software, and an increased awareness of these tools. This connects to numerous branches in the mathematical sciences, as highlighted in the description of the SIAM Journal of Applied Algebra and Geometry.

The Nonlinear algebra group at MPI Leipzig works on fundamental problems in algebra, geometry and combinatorics that are relevant for nonlinear models. This involves algebraic geometry (complex and real), commutative algebra, combinatorics, polyhedral geometry, and more. On the applications side, we are especially interested in statistics, optimization and the life sciences.

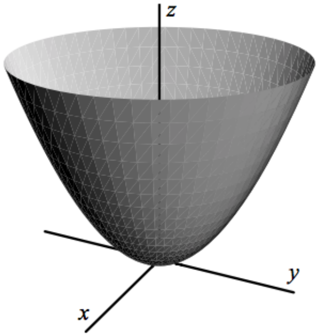
Short overview:

1. Introduction to Varieties (Basic Definitions, Examples, Applications)
2. Extracting Information from Samples: Dimension Estimates
3. Extracting Information from Samples: Persistent Homology
4. Extracting Information from Samples: Computing Polynomials, using Algebraic Geometry Software

Introduction to Varieties-Basic Definitions

Algebraic Varieties

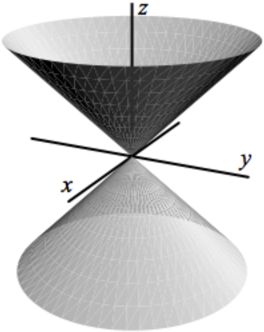
Given polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$, their common zero set is an algebraic variety V. It lives in $\mathbb{R}^n$ or $\mathbb{C}^n$ (affine variety).



The zero set of $z - x^2 - y^2$.

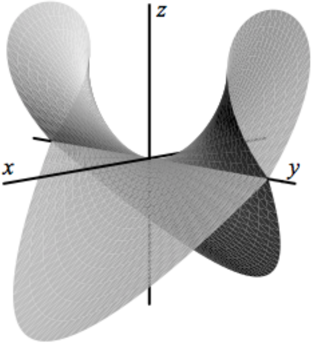
Ideals, varieties and algorithms by Cox, Little, O'Shea.

Introduction to Varieties-Basic Definitions



The zero set of $z^2 - x^2 - y^2$.

Introduction to Varieties-Basic Definitions



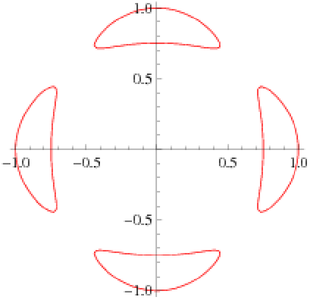
The zero set of $x^2 - y^2z^2 + z^3$.

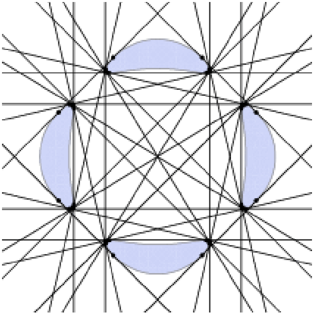
In these last two examples the surfaces are not smooth everywhere: the cone has a sharp point at the origin, and the last example intersects itself along the whole y-axis. These are examples of singular points.

Introduction to Varieties-Basic Definitions

Trott Curve

$$12^2(x^4+y^4)-15^2(x^2+y^2)+350x^2y^2+81=0.$$





A generic plane quartic over the complex projective plane has precisely 28 bitangent lines. The original proof was given by Cayley in 1879.

Introduction to Varieties-Basic Definitions

Rotation Matrices

The group $\mathrm{SO}(3)$ consists of all 3×3-matrices $X = (x_{ij})$ with $\det(X) = 1$ and $X^T X = \mathrm{Id}_3$. The last constraint translates into 9 quadratic equations:

$$
\begin{array}{cc}
x_{11}^2 + x_{21}^2 + x_{31}^2 - 1 & x_{11}x_{12} + x_{21}x_{22} + x_{31}x_{32} \\
x_{11}x_{12} + x_{21}x_{22} + x_{31}x_{32} & x_{12}^2 + x_{22}^2 + x_{32}^2 - 1 \\
x_{11}x_{13} + x_{21}x_{23} + x_{31}x_{33} & x_{12}x_{13} + x_{22}x_{23} + x_{32}x_{33}
\end{array}
$$

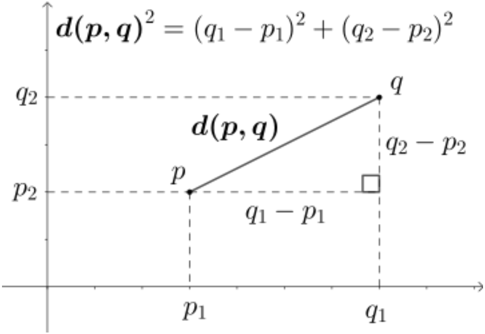
$$
\begin{array}{c}
x_{11}x_{13} + x_{21}x_{23} + x_{31}x_{33} \\
x_{12}x_{13} + x_{22}x_{23} + x_{32}x_{33} \\
x_{13}^2 + x_{23}^2 + x_{33}^2 - 1
\end{array}
$$

These quadrics say that X is an orthogonal matrix. Adding the cubic $\det(X) - 1$ gives 10 polynomials that define $\mathrm{SO}(3)$ as a variety in $\mathbb{R}^9$.

Introduction to Varieties-Basic Definitions

For data in $\mathbb{R}^n$ we use the Euclidean metric:

$$\|u - v\| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}.$$



Introduction to Varieties-Basic Definitions

Projective Space

The set of all lines in $\mathbb{R}^{n+1}$ passing through the origin $\mathbf{0} = (0, ..., 0)$ is called the n-dimensional real projective space and is denoted by $\mathbb{P}^n_{\mathbb{R}}$. We can also identify it with

$$(\mathbb{R}^{n+1} \setminus \mathbf{0})/\mathbb{R}^*,$$

where $(x_1, \ldots, x_{n+1}) \simeq (\lambda x_1, \ldots, \lambda x_{n+1})$ for all $\lambda \in \mathbb{R}^*$, i.e. , two points in $\mathbb{R}^{n+1} \setminus \{0\}$ are equivalent if they are on the same line through the origin.

Introduction to Varieties-Basic Definitions

Homogeneous Coordinates

An element of $\mathbb{P}^n_{\mathbb{R}}$ is called a point. If P is a point, then any $(n+1)$-tuple $(a_1, \ldots, a_{n+1})$ in the equivalence class P is called a set of homogeneous coordinates for P. Equivalence classes are often denoted by $P = [a_1 : \ldots : a_{n+1}]$ to distinguish from the affine coordinates. Note that

$$[a_1 : \ldots : a_{n+1}] = [\lambda a_1 : \ldots : \lambda a_{n+1}]$$

for all $\lambda \in \mathbb{R}^*$.

Introduction to Varieties-Basic Definitions

Projective Varieties

Given homogeneous polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$, their common zero set is an projective variety V. It lives in a projective space $\mathbb{P}^n_{\mathbb{R}}$ or $\mathbb{P}^n_{\mathbb{C}}$.

Low Rank Matrices

Consider the set of $m \times n$-matrices of rank $\leq r$. This is the zero set of $\binom{m}{r+1}\binom{n}{r+1}$ polynomials, namely the $(r+1) \times (r+1)$-minors. These equations are homogeneous of degree $r+1$. Hence this variety lives naturally in the projective space $\mathbb{P}^{mn-1}_{\mathbb{R}}$.

Introduction to Varieties-Basic Definitions

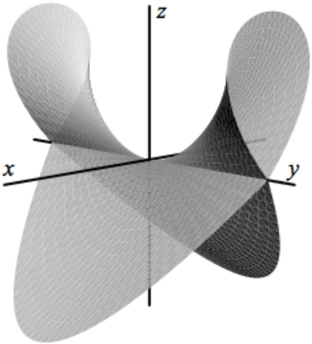
For data in $\mathbb{P}^n_{\mathbb{R}}$ we use the Fubini-Study metric. Points u and v in $\mathbb{P}^n_{\mathbb{R}}$ are represented by their homogeneous coordinate vectors. The Fubini-Study distance between u and v is the angle between the lines spanned by u and v:

$$\mathrm{dist}_{\mathrm{FS}}(u, v) \;=\; \arccos \frac{|\langle u, v\rangle|}{\|u\|\|v\|}.$$

Introduction to Varieties-Basic Definitions

Implicit Representation of Affine Varieties

Let $\mathbf{V}(f_1, \ldots, f_s)$ be a variety. These defining equations $f_1 = \ldots = f_s = 0$ of V are called an **implicit representation** of V.



Introduction to Varieties-Basic Definitions

Parametrizations of Affine Varieties

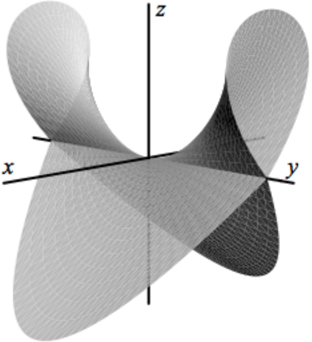
Let $\mathbf{V}(f_1, \ldots, f_s)$ be a variety. A **rational parametric representation** of V consists of rational functions $r_1, \ldots, r_n \in \mathbb{R}(t_1, \ldots, t_m)$ such that the points given by

$$\begin{aligned} x_1 &= r_1(t_1, \ldots, t_m), \\ x_2 &= r_2(t_1, \ldots, t_m), \\ &\vdots \\ x_n &= r_n(t_1, \ldots, t_m) \end{aligned}$$

lie in V. We also require that V be the "smallest" variety containing these points.

In many situations, we have a parametrization of a variety V, where $r_1, \ldots, r_n$ are polynomials rather than rational functions. This is what we call a **polynomial parametric representation** of V.

Introduction to Varieties-Basic Definitions



This picture was not plotted using the implicit representation

$$x^2 - y^2 z^2 + z^3 = 0.$$

Rather, we used the parametric representation given by

$$\begin{aligned} x &= t(u^2 - t^2), \\ y &= u, \\ z &= u^2 - t^2 \end{aligned}$$

There are two parameters t and u since we are describing a surface, and the above picture was drawn using t, u in the range $-1 \leq t, u \leq 1$.

Introduction to Varieties-Basic Definitions

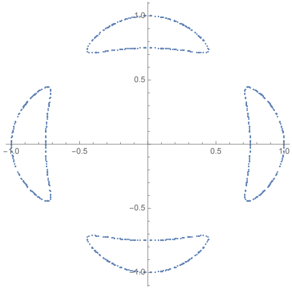
Does every variety have such a parametrization?
No, smooth plane curves of degree ≥ 3 do not.

In fact, most affine varieties cannot be parametrized in the sense described here. Those that can are called unirational. In general, it is difficult to tell whether a given variety is unirational or not.

Introduction to Varieties-Basic Definitions

Unirational Varieties

If V is a unirational variety with given rational parametrization, then it is easy to create a finite subset Ω of V. One selects parameter values at random and plugs these into the parametrization.



While most varieties are not unirational, those arising in applications often are.

Sampling Real Algebraic Varieties for Topological Data Analysis by E. Dufresne, P.Edwards, H. Harrington, J. Hauenstein

Sampling from the uniform distribution on an algebraic manifold by P. Breiding, O. Marigliano

Introduction to Varieties-Basic Definitions

Ideal, Prime Ideal

A subset $I \subset \mathbb{R}[x_1, \ldots, x_n]$ is an **ideal** if it satisfies:

1. $0 \in I$.
2. If $f, g \in I$, then $f + g \in I$.
3. If $f \in I$ and $h \in \mathbb{R}[x_1, \ldots, x_n]$, then $hf \in I$.

An ideal $I \subset \mathbb{R}[x_1, \ldots, x_n]$ is **prime** if whenever $f, g \in \mathbb{R}[x_1, \ldots, x_n]$ and $fg \in I$, then either $f \in I$ or $g \in I$.

Let $V \subset \mathbb{R}^n$ be an affine variety. Then we set

$$I(V) = \{f \in \mathbb{R}[x_1, \ldots, x_n] : f(a_1, \ldots, a_n) = 0 \text{ for all } (a_1, \ldots, a_n) \in V\}.$$

$I(V)$ **is an ideal!** We call $I(V)$ the ideal of V.

Introduction to Varieties-Basic Definitions

Irreducibility

An affine variety $V \subset \mathbb{R}^n$ is irreducible if whenever V is written in the form $V = V_1 \cup V_2$, where V_1 and V_2 are affine varieties, then either $V_1 = V$ or $V_2 = V$.

$\mathbf{V}(xz, yz)$ is not an irreducible variety.

Characterization of Irreducibility

Let $V \subset \mathbb{R}^n$ be an affine variety. Then V is irreducible if and only if $I(V)$ is a prime ideal.

Decomposing a Variety

Let $V \subset \mathbb{R}^n$ be an affine variety. Then V can be written as a finite union $V = V_1 \cup \ldots \cup V_m$, where each V_i is an irreducible variety.

Introduction to Varieties-Basic Definitions

The most important invariant of a linear subspace of affine space is its dimension.

Dimension

The dimension d of V is the maximum integer such that there exist

$$V_0 \subset V_1 \subset \ldots \subset V_d = V,$$

where all of the subsets are proper and all of the sets V_i are irreducible varieties.

Introduction to Varieties-Basic Definitions

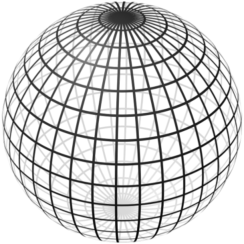
Degree

The degree of an affine or projective variety of dimension d is the number of intersection points of the variety with d hyperplanes in general position (for an algebraic set, the intersection points must be counted with their intersection multiplicity).

Examples of Varieties

Hypersurfaces

The most basic varieties are defined by just one polynomial.



For example, the equation

$$x_1^2 + x_2^2 + \cdots + x_n^2 - 1 = 0$$

defines an algebraic hypersurface of dimension $n-1$ in the Euclidean space of dimension n.

Examples of Varieties

Factor Analysis

In factor analysis, correlated continuous variables are modeled as conditionally independent given hidden (latent) variables that are called factors. In many applications the focus is on interpreting the factors as unobservable theorized concepts. In fact, the desire to explain observed correlations between individuals' exam performances by the concept of intelligence was the driving force in the original development of factor analysis (Spearman, 1904, 1927).

Examples of Varieties

Factor Analysis

Spearman was trying to discover the hidden structure of human intelligence. His observation was that schoolchildren's grades in different subjects were all correlated with each other. He went beyond this to observe a particular pattern of correlations, which he thought he could explain as follows: the reason grades in math, English, history, etc., are all correlated is performance in these subjects is all correlated with something else, a general or common factor, which he named "general intelligence", for which the natural symbol was of course g or $\mathbf{G}$.

Examples of Varieties

Factor Analysis

If X is our data matrix, with n rows for the different observations (students) and p columns for the different variables (X_{ij} is the value of variable j in observation i)(school subjects), then Spearman's model becomes:

$$X = \epsilon + \mathbf{G}w,$$

where ϵ is an error or residual term, where G is an $n \times 1$ matrix and w is a $1 \times p$ matrix. If we assume that the features and common factor are all centered to have mean 0, and that there is no correlation between ϵ_{ij} and G_i for any j, then the correlation between the jth feature, $X_{\cdot j}$, and G is just w_j.

Examples of Varieties

Factor Analysis

Under these assumptions, it follows that the correlation between the ith feature and the jth feature is just the product of the factor loadings:

$$v_{ij} = cov[X_{\cdot i}, X_{\cdot j}] = w_i w_j.$$

Up to this point, this is all so much positing and assertion and hypothesis. What Spearman did next, though, was to observe that this hypothesis carried a very strong implication about the ratios of correlation coefficients. Pick any four distinct features, i, j, k, l. Then, if the model is true,

$$\begin{aligned} \frac{v_{ij}/v_{kj}}{v_{il}/v_{kl}} &= \frac{w_i w_j / w_k w_j}{w_i w_l / w_k w_l} \\ &= \frac{w_i/w_k}{w_i/w_k} \\ &= 1 \end{aligned}$$

Examples of Varieties

Factor Analysis

The relationship $v_{ij}v_{kl} = v_{il}v_{kj}$ is called the "tetrad equation". In Spearman's model, this is one tetrad equation for every set of four distinct variables. Spearman found that the tetrad equations held in his data on school grades (to a good approximation), and concluded that a single general factor of intelligence must exist. This was, of course, logically fallacious. Later work, using large batteries of different kinds of intelligence tests, showed that the tetrad equations do not hold in general, or more exactly that departures from them are too big to explain away as sampling noise.

Examples of Varieties

Factor Analysis

Tetrads have played a major role throughout the history of factor analysis and also appear in recent research. While tetrads are ubiquitous in the literature, there has been very little work attempting to find invariants of models with more than one factor. The work by Kelley (1935) who derived the pentad, a fifth degree polynomial vanishing over covariance matrices from two-factor models, constitutes the exception.

Today

Algebraic Statistics is the use of algebra to advance statistics.

M. Drton, B. Sturmfels, S. Sullivant. Lectures on Algebraic Statistics, Springer 2009.

L. Pachter and B. Sturmfels. Algebraic Statistics for Computational Biology. Cambridge University Press 2005.

Examples of Varieties

Rank Constraints

Consider $m \times n$-matrices with linear entries having rank $\leq r$. We saw the $r = 1$ case earlier. A *rank variety* is the set of all matrices of fixed size and rank that satisfy some linear constraints. The constraints often take the simple form that two entries are equal. This includes symmetric matrices, Hankel matrices, Toeplitz matrices, Sylvester matrices, etc. Many classes of structured matrices generalize naturally to tensors.

Examples of Varieties

Any $n \times n$ matrix A of the form

$$A = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & \dots & a_{n-1} \\ a_1 & a_2 & & & & \vdots \\ a_2 & & & & & \vdots \\ \vdots & & & & & a_{2n-4} \\ \vdots & & & & a_{2n-4} & a_{2n-3} \\ a_{n-1} & \dots & \dots & a_{2n-4} & a_{2n-3} & a_{2n-2} \end{bmatrix}$$

is a **Hankel matrix**. If the i, j element of A is denoted $A_{i,j}$, then we have

$$A_{i,j} = A_{j,i} = a_{i+j-2}.$$

Examples of Varieties

Any $n \times n$ matrix A of the form

$$A = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \dots & \dots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \dots & \dots & a_2 & a_1 & a_0 \end{bmatrix}$$

is a **Toeplitz matrix**. If the i,j element of A is denoted $A_{i,j}$, then we have

$$A_{i,j} = A_{i+1,j+1} = a_{i-j}.$$

Examples of Varieties

The distance geometry problem (DGP) is that of finding the coordinates of a set of points by using the distances between some pairs of such points. There exists nowadays a large community that is actively working on this problem, because there are several real-life applications that can lead to the formulation of a DGP:

- **Determine the location of sensors in telecommunication networks.**

  In such a case, the positions of some sensors are known (which are called anchors) and some of the distances between sensors (which can be anchors or not) are also known: the problem is to identify the positions in space for all sensors.

- **Determine the conformation of a given molecule.**

  Experimental techniques are able to estimate distances between pairs of atoms of a given molecule, and the problem becomes the one of identifying the three-dimensional conformation of the molecule, i.e. the positions of all its atoms. In this field, the main interest is on proteins, because discovering their three-dimensional conformation allows us to get clues about the function they are able to perform.

Examples of Varieties

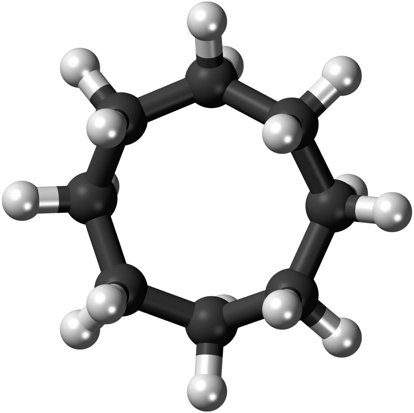
In *distance geometry*, one encodes a metric space with p points in the matrix M

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

Here d_{ij} is the squared distance between points i and j. This symmetric matrix is positive semidefinite if and only if the metric space is Euclidean, and its embedding dimension is the rank of M. Hence the rank varieties of these matrices encode the finite Euclidean metric spaces. We will now take a look at the conformation space of the cyclo-octane, which corresponds to the case $p = 8$ and $r = 3$.

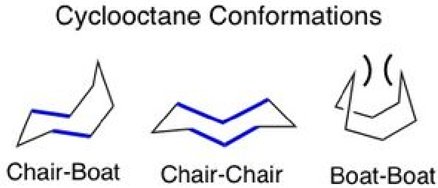
Examples of Varieties

Conformation Space of Cyclo-Octane



Examples of Varieties

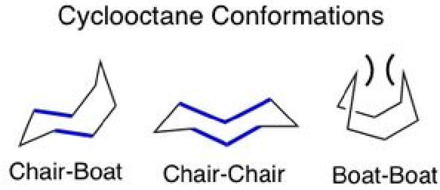
Conformation Space of Cyclo-Octane



Cyclo-octane consists of 8 carbon atoms arranged in a ring and each bonded to a pair of hydrogen atoms. The location of the hydrogen atoms is determined by that of the carbon atoms due to energy minimization. Hence, the conformation space of a cyclo-octane consists of all possible spatial arrangements, up to rotation and translation, of the ring of carbon atoms. Each datum is a point in $\mathbb{R}^{24} = \mathbb{R}^{8 \cdot 3}$, which represents the coordinates for $\{z_0, \ldots, z_7\} \subset \mathbb{R}^3$, the locations of carbon atoms in cyclo-octane.

Examples of Varieties

Conformation Space of Cyclo-Octane



Each carbon atom z forms an isosceles triangle with its two neighbors in a way that that the angle at z is $\frac{2\pi}{3}$. Therefore (after scaling), the squared distances $d_{i,j} = \|z_i - z_j\|^2$ satisfy

$$d_{i,i+1} = 1 \quad \text{and} \quad d_{i,i+2} = \frac{8}{3}$$

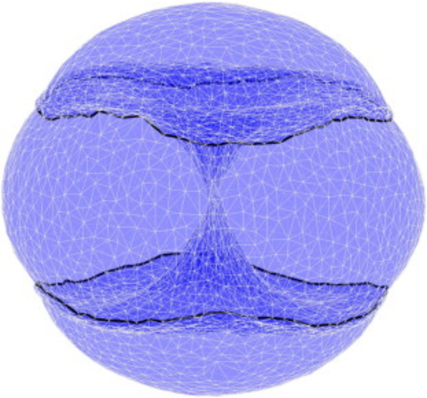
for all i (here i is modulo 8).

Examples of Varieties

Conformation Space of Cyclo-Octane

M.W. Brown, S. Martin et. al: *Algorithmic dimensionality reduction for molecular structure analysis*

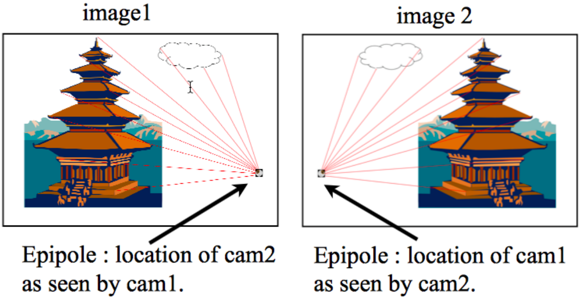
S. Martin, A. Thompson, E. A. Coutsias, and J. P. Watson: *Topology of cyclo-octane energy landscape*.



Examples of Varieties

Distortion varieties

We have two views of a scene taken from different viewpoints. We see an image point $\mathbf{p_0}$, which is the projection of a 3D point.



Given the $\mathbf{p_0}$ in the first image where can the corresponding point $\mathbf{p_1}$ in the second image be?

Examples of Varieties

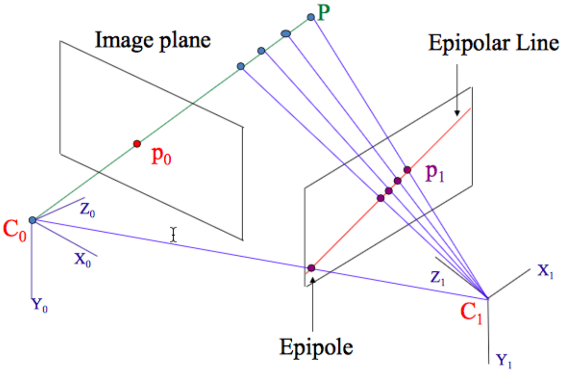
Distortion varieties

The essential matrices are 3×3 matrices that "encode" the epipolar geometry of two views.

Motivation: Given a point in one image, multiplying by the essential matrix will tell us which epipolar line to search along in the second view.

Examples of Varieties

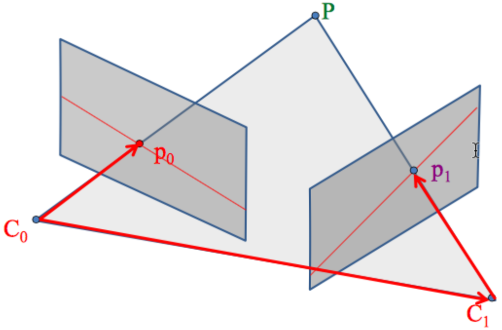
Distortion varieties



Examples of Varieties

Distortion varieties

The optical centers of the two cameras, a point P, and the image points $\mathbf{p_0}$ and $\mathbf{p_1}$ of P all lie in the same plane (**epipolar plane**).



These vectors are co-planar: $\vec{C_0p_0}, \vec{C_1p_1}, \vec{C_0C_1}$.

Examples of Varieties

Distortion varieties

Another way to write the fact they are co-planar is

$$(\vec{C_0 p_0}) \cdot (\vec{C_1 p_1} \times \vec{C_0 C_1}) = 0.$$

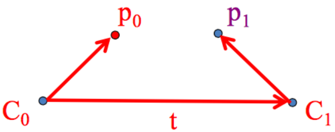
Examples of Varieties

Distortion varieties

So we can write the coplanar constraint as

$$\mathbf{p_0} \cdot (\mathbf{t} \times \mathbf{R}\mathbf{p_1}) = 0,$$

where $\mathbf{R}$ is the rotation of camera 1 with respect to camera 0 and $\mathbf{t}$ is the translation of the camera 1 origin with respect to camera 0.



$$\overrightarrow{\mathbf{C}_0\mathbf{p}_0} \cdot \left(\overrightarrow{\mathbf{C}_0\mathbf{C}_1} \times \overrightarrow{\mathbf{C}_1\mathbf{p}_1}\right) = 0$$

Examples of Varieties

Distortion varieties

The cross product of a vector $\mathbf{a}$ with a vector $\mathbf{b}$, $\mathbf{a} \times \mathbf{b}$, can be represented as a 3×3 matrix times the vector $\mathbf{b}$:

$$\begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \mathbf{b} = \mathbf{a} \times \mathbf{b}.$$

The matrix on the left is a skew-symmetric matrix and we denote it by $[\mathbf{a}]_\times$.

Examples of Varieties

Distortion varieties

We rewrite the epipolar constraint as a matrix product:

$$\mathbf{p_0}^T[\mathbf{t}]_\times \mathbf{R}\mathbf{p_1} = 0.$$

Then

$$E = [\mathbf{t}]_\times \mathbf{R}$$

is the 3×3 matrix called the **essential matrix**. It relates the image of a point in one camera to its image in the other camera, given a translation and rotation.

Play the Fundamental Matrix Song

The variety of essential matrices is defined by ten cubics, known as the *Démazure cubics*.

J. Kileel, Z. Kukelova, T. Pajdla and B. Sturmfels: Distortion varieties, Foundations of Computational Mathematics, (2018).

W. Hoff: EGGN 512 lectures.