Signature Curves

In Classifying DNA Supercoils

Cheri Shakiban

IMA

University of MN
What is a signature curve?

Signature curves are unique curves which are invariant under Euclidean transformations such as rotation.
Why Signature Curves?

• Signature curves are useful in the study of computer vision applications.

• They hold much promise for future developments in the field of artificial intelligence.

• Signature curves can be used in medical imaging devices such as CAT or MRI scans.

• Applications in the study of DNA.

• Applications to the Military.
How do we calculate the Signature of a curve in $\mathbb{R}^2$?

The Euclidean Signature Curve associated with a plane curve $C = \{x(s), y(s)\}$, parametrized by arclength $s$

$$s(t) = \int_0^t \sqrt{x_t^2 + y_t^2} \, dt$$

is the curve $S = \{(K(s), K_S(s))\}$. Where

- $K$ is curvature
- $K_S$ is derivative with respect to arclength

The curvature $K$ at a point $P$ is $1/r$ where $r$ is the radius of osculating circle.

The formal definition of curvature is: $K = |dT/ds|$
Where $T$ is the unit tangent vector and $s$ is arc length.

[Diagram of Signature Curve]
Theorem: (É. Cartan) Two smooth curves $C$ and $\bar{C}$ can be mapped to each other by a proper Euclidean transformation, if and only if their signature curves are identical i.e. $S = -\bar{S}$. 

$S = -\bar{S}$
Example:
The polar Curve $r = 3 + \frac{1}{10} \cos \theta$

The Original Curve

Euclidean Signature Curve

Approximate 2 fold symmetry
Example:

\[ x = \cos t + \frac{1}{5} \cos^2 t \]
\[ y = \frac{1}{2} x + \sin t + \frac{1}{10} \sin^2 t \]

The Original Curve

Euclidean Signature Curve

Approximate 3 fold symmetry
Observation:
If a curve has three-fold rotational symmetry then its signature curve is not a simple closed curve, but goes around three times.
( has winding number 3)
• It is **tedious** to calculate the Signature curves directly even if we know the equations for the curve.

• For Most Images we don’t know what the equation of the curve is.
Numerical approximation formula for Euclidean Signature curves were first introduced by.

1st approach:
Differential and numerically invarint signature curves applied to object recognition.

2nd approach:
Numerically invarint signature curves
Mireille Boutin
Let $A, B, C$ be three points on the curve $C$.

$$a = d(A,B) \quad b = d(B,C) \quad c = d(A, C)$$

$r$ = radius of the circle passing through $A, B, C$.

$\kappa(B)$ curvature of the curve at the point $B = \frac{1}{r}$

$\kappa(A,B,C)$ curvature of the circle passing through $A, B, C = 4 \frac{\Delta}{abc}$

$\Delta$ = Area of the triangle $A B C$
Heron's Formula:

Let \( \Delta = \text{Area of the triangle } \triangle ABC \)

\[
\Delta = \sqrt{s(s - a)(s - b)(s - c)}
\]

\[
s = \frac{1}{2} (a+b+c)
\]

\[
\kappa(A, B, C) = 4 \frac{\sqrt{s(s - a)(s - b)(s - c)}}{abc}
\]
To approximate $\kappa_s = \frac{d\kappa}{ds}$ we use five points:

$$\tilde{\kappa}_s(p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2}) = \frac{\tilde{\kappa}(p_i, p_{i+1}, p_{i+2}) - \tilde{\kappa}(p_{i-2}, p_{i-1}, p_i)}{3(2a + 2b + d + g)}$$
Numerical Euclidean Signature in 2D

50 points

250 points
Numerically Invariant Euclidean Signature for a space curve (3D):

**Theorem:** The signature curve of a space Curve in $\mathbb{R}^3$ is given by
$(\kappa, \kappa_s, \tau)$, where
$\kappa =$ curvature
$\tau =$ torsion : defined as the derivative of the angle of the osculating plane with respect to arc-length.
*
(Pre-print)
Approximations for $\kappa$ and $\kappa_s$ remain the same as in the planar case.
Approximations for $\tau$

$$\tilde{\tau}(p_i) = 6 \frac{H}{d e f \tilde{k}}$$

$H$ is the height of the tetrahedron with sides $a, b, c, d, e$. 
The Invariant Signature Curve associated with a regular space curve
Creating a Signature Curve

For an Image

**Step 1:** Determine the set of 2D points that make up the object
- Use edge detection to determine points
- Roughly 100-350 points (n)
- The points outline the figure in a closed loop
Creating Discrete Signature Curve

- **Step 2:** Calculate the signature curve for the original points
  - Compute the curvature and derivative of curvature (and torsion for 3D) using successive sets of 3 points from the initial object
    - The first signature point relies on points 1, 2, 3 in the object
    - The second signature point relies on points 2, 3, 4
    - So on
Creating a Signature Curve

- **Step 3:** Use a cubic spline routine to interpolate between signature curve points creating a smooth curve.
How can we compare Signature Curves??

Maple leaf

Buckthorn leaf
What Is Latent Semantic Analysis?

**LSA**

- A Statistical Method that provides a way to describe the underlying structure of texts
- Used in author recognition, search engines, detecting plagiarism, and comparing texts for similarities
- The contexts in which a certain word exists or does not exist determine the similarity of the documents
- Closely models human learning, especially the manner in which people learn a language and acquire a vocabulary
Step #1: Construct the term-document matrix; TDM
- One column for each document
- One row for every word
- The value of cell \((i, j)\) is the frequency of word \(i\) in document \(j\)

Example:

<table>
<thead>
<tr>
<th></th>
<th>Story #1</th>
<th>Story #2</th>
<th>Story #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>rabid</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>spread</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>disease</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>was</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>put</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>down</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The Method of LSA

- **Step #2: Normalize the elements**
  - Each cell in the frequency table receives a value between 0 & 1
  - Takes the length of a document into account

- **Step #3: Weight functions**
  - Each element is replaced with the product of a local weight function and a global weight function
  - LWF considers the frequency of a word within a particular text
  - GWF examines a term’s frequency across all the documents
The Method of LSA

- **Step #4: Singular value decomposition (SVD)**
  - Removes noise or infrequent words that do not help
  - SVD factors the initial $m \times n$ matrix into three new matrices. The center matrix contains the **singular values**, while the matrices on the left and right are orthogonal.
Step #5: Build a correlation matrix

- The Final Symmetric Matrix with the various documents taking up the rows and columns.
- Cell \((i, j)\) gives the correlation between the \(i\)th and \(j\)th documents.
- Similarity between documents A & B is given by the cosine of the angle between column vectors \(a\) and \(b\) in the transformed matrix:

\[
\cos[\theta] = \frac{a \ast b}{\|a\| \ast \|b\|}
\]

Example:
Sigature Curves & LSA

- LSA is proposed as a means to identify signature curves
- Traditional LSA uses categories of texts -
- We want LSA rely on categories of objects

**Proposition:** The data of signature curves should be broken down and sorted into a TDM that can undergo traditional transformations to compare new objects
LSA for Sorting Signature Curves- 2 D

▪ Step #1: Create the signature curves for at least 2 distinct, clearly defined categories of objects

▪ Step #2: Build the Square-object matrix, SOM (variant of term-document matrix)
  ▪ Each object occupies a column

▪ We can use a nxn grid centered at the origin to cover the signature curves.

▪ The (i, j) cell gives the amount of signature curve points in the ith square for the jth object
LSA for Sorting Signature Curves - 2D Illustration
Step #3: Normalize

Step #4: Data compression:
  - Keeps only rows with important info on objects
  - Phase 1: Clear near-zero-rows
    - Rows that have mostly zeros only create noise
  - Phase 2: Clear rows with low standard deviations
    - A row with a low standard deviation means that the data is fairly uniform for that cube across all objects
    - A high standard deviation means that frequencies are high for some objects and low for others
  - Phase 3: Drop rows with low entropies
    - A low entropy suggests that all objects have a similar, low frequency for that cube
- **Step #5:** Round remaining cells
- **Step #6:** Use the LWF and GWF
- **Step #7:** Use a trigonometric weight function
  - Gives higher values to points that lie farther from the origin
  - The outer points or “shell” of the signature curve is the unique portion for every signature and must be emphasized

\[
\begin{align*}
TWF = 0 \quad &\Rightarrow T(i, j) = 1 \\
TWF = 1 \quad &\Rightarrow T(i, j) = \sqrt{a^2 + b^2 + c^2} \\
TWF = 2 \quad &\Rightarrow T(i, j) = (a^2 + b^2 + c^2) \\
TWF = 3 \quad &\Rightarrow T(i, j) = \frac{4}{\sqrt{a^2 + b^2 + c^2}}
\end{align*}
\]
The Optimal Transformations

- The best combination of transformations for LSA (combined with sig curves) is desired
- The number of possible combinations is large
- New measures
  - Condition number from SVD plot
    - Condition number = (largest singular value) / (smallest singular value)
    - A lower condition number means the plot does not change dramatically, and SVD will not denoise the matrix further
  - Correlation matrix quality:
    - M = avg strong correlation
    - N = avg weak correlation

\[ L = \frac{M - N}{1 - M} \]
An example with 2 categories of objects: maple leaves and buck thorn leaves
Leaf Example: Overview

- 100 pictures were used, 50 of each leaf
- Sig curves are in 2D

Partitioning schemes

- Each curve had between 9,000 and 19,000 points
- 10,000 squares, side 1
- 10,000 squares, side 2
- 10,000 squares, side 0.5
- 40,000 squares, side 0.25

- Cells rounded to nearest 0.10 after being normalized
Leaf Example: Correlation Matrices

Trial #1
- Squares: 50 x 1
- Clearing Rows
  - Percent Filled: 10%
  - Standard Dev: 95%
  - Entropy: 35%
- Transformation: (1,3,2,7)
- $L = 3.8956$
- Compression: 4,046 – 12 KB
- Accuracy: 90%

- Correlation Matrices
  - Upper-left is for maple leaves, very strong
  - Lower-right is for buck thorn leaves, strong
  - Dark blue areas indicate small cross-category correlations
Leaf Example: Correlation Matrices

**Trial #2**
- Squares: 50 x 1
- Clearing Rows
  - Percent Filled: 10%
  - Standard Dev: 75%
  - Entropy: 40%
- Transformation: (1,3,2,8)
- $L = 1.2042$
- Compression: 4,046 – 47 KB
- Accuracy: 98%
DNA has the ability to twist, bend, and knot.

A supercoiled DNA molecule has sharp bends in its form and appears to fold and twist over itself.

Supercoiling occurs when the two sugar-phosphate backbones in the DNA molecule are over wound or under wound.

A DNA molecule that exists in a closed ring, called a plasmid, often becomes supercoiled.

Rods of DNA could also become supercoiled, but the ends of the DNA strand must be fixed to prevent the molecule from unwinding and returning to a relaxed state.

Supercoils contain potential energy and tension that can drastically alter the DNA molecule’s function.

Mathews, van Holde, and Ahern
DNA Supercoil Model

The figure reveals several forms of supercoiling.
- The linking number increases in the strands from left to right.
- The molecule on the far left has a large quantity of twists but no evident writhes.
- The strand second from the left contains several writhes.
- The molecule on the far right is tightly wound with a high number of twists and writhes.
DNA Supercoil Model

John Maddocks group: EPFL
Circular DNA
Supercoil Molecule

Signature Curve
(10,305 points), Zoom Factor 1

Discrete Signature Curve Example
DNA Models
Irregular DNA Chain

Signature Curve (10,053 points), Zoom Factor 1
DNA Experiment

- Purpose: to use LSA to classify 3D signature curves
- 100 DNA molecules—54 circular, 46 irregular
- The signature curves of circular DNA strands are extremely dense at the origin and along the axes
- The signatures of irregular chains have tendrils that project away from the origin

Circular DNA Sig Curve  Irregular DNA Sig Curve
LSA for Signature Curves- 3D

Step #1: Create the signature curves for at least 2 distinct, clearly defined categories of objects

Step #2: Build the cube-object matrix COM
  - Each object occupies a column, and each row is a cube address
  - The 4x4x4 box centered at the origin is broken into thousands of tiny cubes. Each cube is assigned an address (a, b, c)

The (i, j) cell gives the amount of signature curve points in the ith cube for the jth object
**LSA for Signature Curves**

**Step #3:** Remove the cubes that are popular for all signature curves.
- All cubes within a certain radius from the origin are erased.
- Cubes near the axes may be eliminated

**Step #4:** Apply a trigonometric weight function to give higher values to points that lie farther from the origin. The outer points or “shell” of the signature curve is the unique portion for every signature and must be emphasized

\[
\begin{align*}
\text{TWF} = 0 \quad &\rightarrow \quad T(i, j) = 1 \\
\text{TWF} = 1 \quad &\rightarrow \quad T(i, j) = \sqrt{a^2 + b^2 + c^2} \\
\text{TWF} = 2 \quad &\rightarrow \quad T(i, j) = (a^2 + b^2 + c^2) \\
\text{TWF} = 3 \quad &\rightarrow \quad T(i, j) = 4\sqrt{a^2 + b^2 + c^2}
\end{align*}
\]
LSA for Signature Curves

- **Step #5**: Normalize the matrix
- **Step #6**: Apply a local weight function and a global weight function
- **Step #7**: Use singular value decomposition to strip away unnecessary information
- **Step #8**: Build the correlation matrix
Correlation matrices produced with center-thinning of radius 0.6:

- All the matrices have a checkerboard pattern.
- The blue crosshairs prove that LSA was able to recognize two distinct classes of signature curves.

Correlations among the circular strands are in the upper-left corner.
- The circular DNA molecules are very highly correlated.

Correlations for the irregular chains are in the lower-right corner.
- The irregular strands are weakly related.
- The two categories are somewhat correlated.
Future Research

- More experimentation with a range of objects is needed
- Must find the optimum transformation combination for signature curves
- The situations in which the method works best must be uncovered
- The ability for LSA to label a new object with replicas already existing in the database should be developed
- Create a system for forming the signature curves for 3D surfaces

Thank you!