Interesting Mathematics from Biological Structures

1. Distributions of 3D motifs in structural RNA

K. Sargsyan, C. Lim, Arrangement of 3D structural motifs in ribosomal RNA, Nucleic Acids Research 38 (11), 3512 (2010)

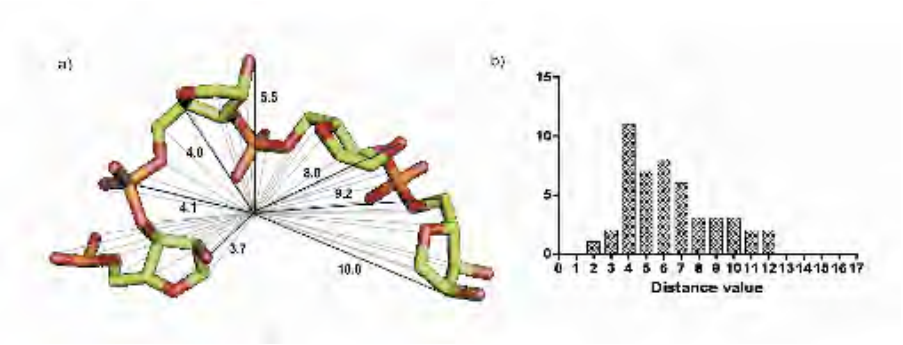
2. GeoPCA: Principal Component Geodesics

K. Sargsyan, J. Wright, C. Lim, GeoPCA: a new tool for multivariate analysis of dihedral angles based on principal component geodesics, Nucleic Acids Research 40 (3), e25 (2012)

Distributions of 3D motifs in structural RNA

- Similarity measure for 3D motifs in RNA.

- The structure of a short RNA fragment is described using the shape histogram. Only backbone atoms of the RNA fragment are considered. For a given RNA fragment, the centroid with respect to the phosphorus atoms of the fragment and its distance to each backbone atom are computed. Next, the set of distances are then rounded to the nearest. The frequency of each integer distance value is plotted as a 2D histogram, and represented by a histogram vector h = (h_1,..., h_n), where h_i is the frequency of the integer distance). The shape histogram of a RNA fragment, which is a distribution of Euclidean distances of the RNA backbone atoms from the centroid, can be considered as a signature of the fragment structure.



- Figure: The shape histogram of a given RNA fragment. (a) The backbone atoms of a RNA fragment 1794-1797 and some of their distances to a centroid. (b) The shape histogram of a given RNA fragment represented by the frequency of an integer distance in Å.

- The measure of similarity $$Cos(h,g) = \frac{\sum_i h_i g_i}{\sqrt{\sum_i h_i^2}\sqrt{\sum_i g_i^2}}$$

- Idea is taken from: Apostolico, G. Ciriello, C. Guerra, C. Heitsch, C. Hsiao, L. Williams (2009) Finding 3D Motifs in Ribosomal RNA Structures, *Nucleic Acids Research,* doi: 10.1093/nar/gkn1044.

Distributions of 3D motifs in structural RNA

- RNA fragment as a "word".

- A RNA fragment can be considered as a "word" containing structural information, which is described by a shape histogram. However, a word has fixed spelling and discrete letters, so 2 words are either equal or not equal, whereas the *Cos* and *RMSD*, describing the difference between two RNA fragments, are continuous variables.

- Therefore, a combination of *Cos* and *RMSD* thresholds was used to assign the structural similarity of two RNA fragments (equivalence of 2 words). In this way, motifs are treated as discrete objects like words. For a RNA consisting of *N* nucleotides, there will be *N*-*l*+1 RNA fragments composed of *l*-nucleotides (*l*-mer or "*l*-letter word") starting from each nucleotide. The *l*-mer fragment starting at position *i* was compared with all the other *N*-*l* *l*-mer fragments starting at position *i*+1, *i*+2, …, *N*-*l*+1, 1, 2, 3, …, *i*−1, and a match was recorded by the position of the matching fragment, denoted by the position of its first nucleotide, a_i.

- If there are >10 matches, then the *l*-mer fragment starting at position *i* is a potential motif, whose distribution along the RNA chain is described by the positions of the "word" along the text, $S_i = (a_1, a_2, a_3 \ldots.)$. Note that an *l*-mer RNA fragment that repeats ≤10 along the chain was not considered as a motif for distribution analyses due to the lack of statistics.

- What follows next is according to: P Carpena, P Bernaola-Galván, M Hackenberg, AV Coronado, JL Oliver, "Level statistics of words: Finding keywords in literary texts and symbolic sequences", Physical Review E 79 (3), 035102 (2009)

Distributions of 3D motifs in structural RNA

- Evaluating if a 3D motif distribution is random or not.

- Let $d_i = a_i - a_{i-1}$ denote the separation of the same consecutive motifs ("words") along the RNA chain ("text"), and $D = (d_1, d_2, \ldots, d_N)$ denote the set of integer distances. The probability of distance d, $P(d)$, in set D in the case of random placement of the motif along the RNA chain (or a word along the text) is described by the geometric distribution:

$$P(d) = p(1-p)^{d-1}$$

- The set of D values that differs from those described by the geometric distribution was estimated by σ, the ratio of the normalized mean square deviations for elements of set D to those in the geometric distribution:

$$\sigma = \frac{\sqrt{\langle d^2 \rangle - \langle d \rangle^2}}{\langle d \rangle \sqrt{1-p}}$$

- A value of σ = 1 indicates a given 3D motif occurs randomly along the RNA chain, whereas $\sigma > 1$ or < 1 indicates the l-mers of a given 3D motif attract or repel each other, respectively, along the chain.

- Since σ and thus $P(\sigma)$ depends on the occurrence frequency of the motif, n, artificial clustering caused by fluctuations is possible for motifs that occur infrequently along the RNA chain. Thus, the deviation of the distribution of D values from a geometric distribution was also evaluated using a Z-score measure, which depends on the self-attraction/repulsion of a l-mer motif and its frequency:

$$C(\sigma, n) = \frac{\sigma - \langle \sigma \rangle(n)}{sd(\sigma)(n)} \qquad \langle \sigma \rangle(n) = \frac{2n-1}{2n+2} \qquad sd(\sigma) = \frac{1}{\sqrt{n}\left(1 + 2.8n^{-0.865}\right)}$$

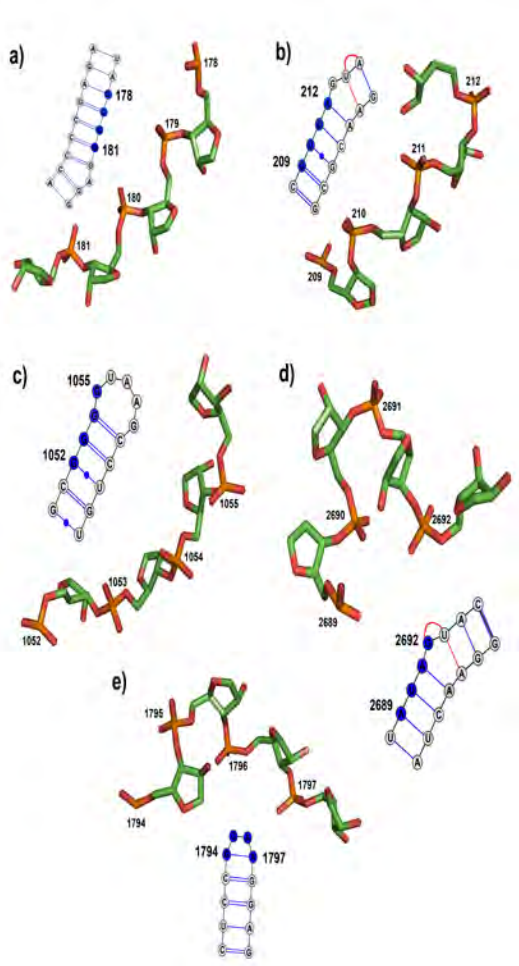
- A value of C = 0 indicates a given 3D motif occurs randomly along the RNA chain, whereas $C > 0$ or < 0 indicates the l-mers of a given 3D motif attract or repel each other, respectively, along the chain.

Distributions of 3D motifs in structural RNA

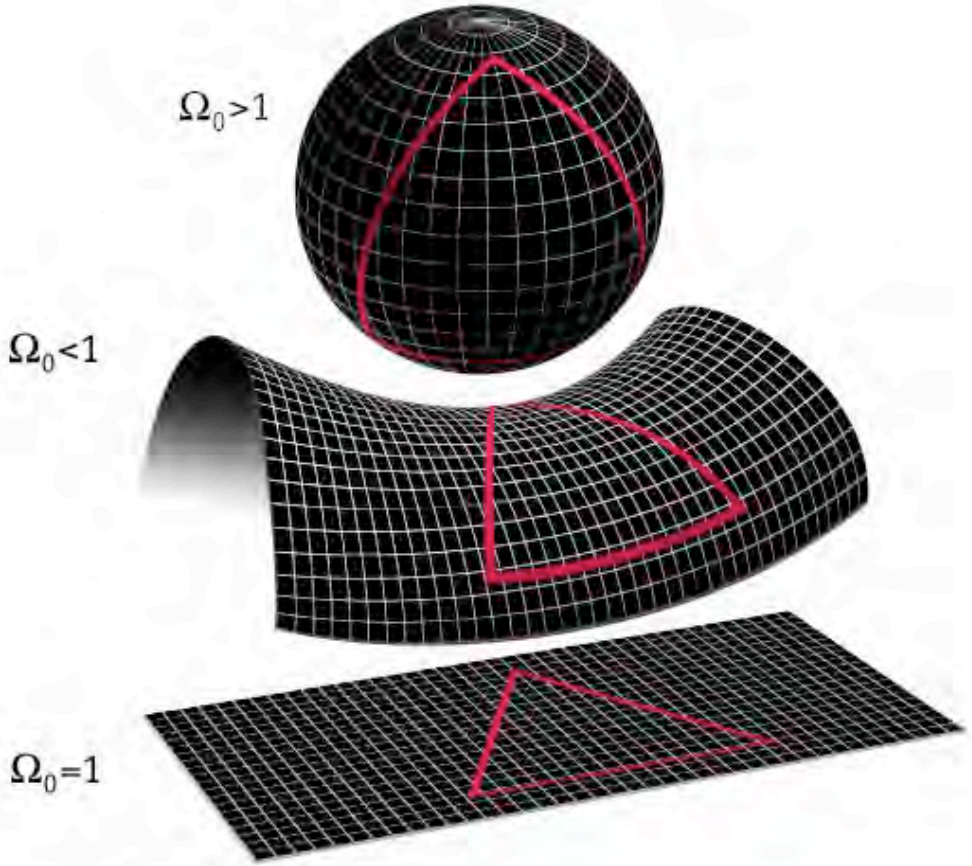
- Results for 4-mers and binding.

- On Figure: 3D backbone and secondary structures corresponding to the representative 4–mer motifs found: (a) helical motif, 178-181, (b) part of an internal loop, 1052-1055, (c) tetraloop motif, 1794-1797, (d) part of an internal loop, 209-212, and (e) part of an internal loop, 2689-2692. Secondary structures were prepared by the program, VARNA.

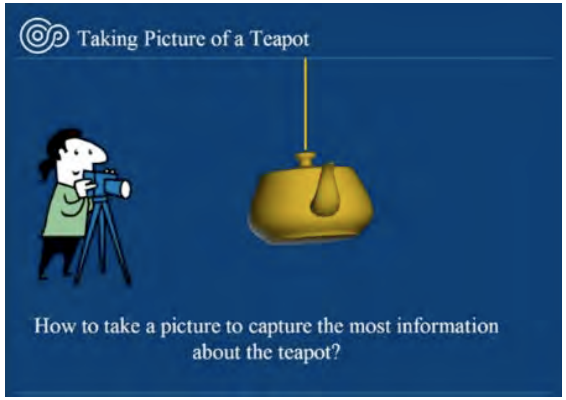
- To reveal whether the 3, 4, and 5-mer motifs, which deviate significantly from a random distribution and persist under changes of the threshold of similarity measure, play a role in binding, the fraction of "binding" residues in these motifs was compared to that in all l–mers. According to this fraction motifs distributed non randomly tend to participate in binding more than others.

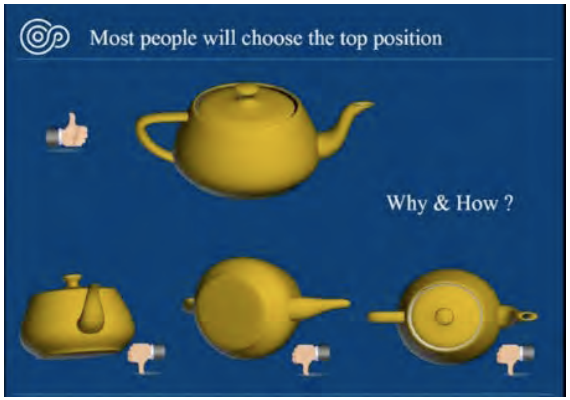


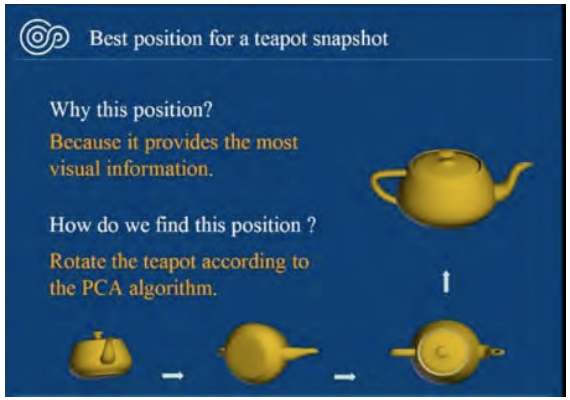
Multivariate Analysis of Dihedral Angles based on Principal Component Geodesics.

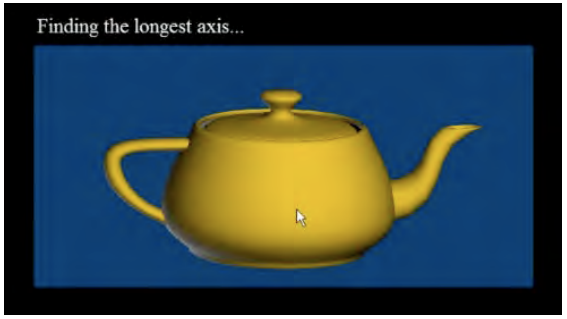


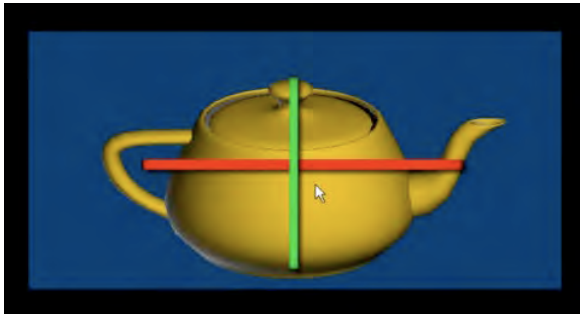
Principal Component Analysis on Teapot









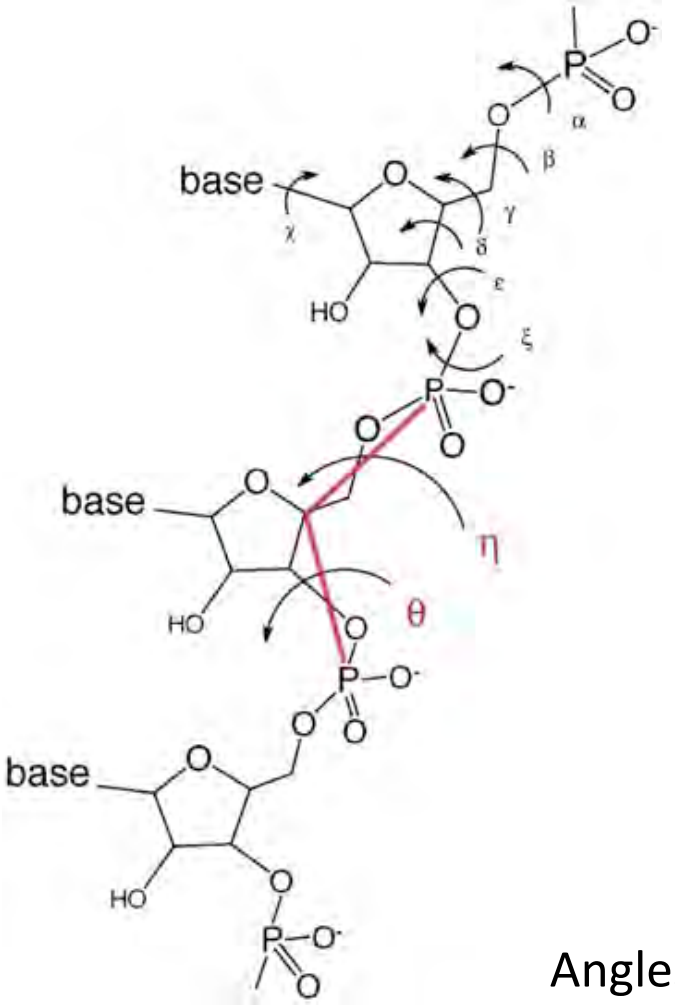


Mathematical Aspects of Principal Component Analysis

- Data points are in Euclidean space and subject of Euclidean geometry.
- PCA projects the data along the directions (straight lines) where the data varies the most.
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.
- The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.

It doesn’t work if topology is not Euclidean!

Angles? Are They Interesting?



Using dihedral angles alone could drastically reduce the number of degrees of freedom present in large biomolecules such as proteins and nucleic acids.

The conformation of nicotinamide adenine dinucleotide composed of 44 atoms can be described by 11 dihedral angles, thus reducing the internal degrees of freedom ($3\times44 - 6 = 126$) by > 90%

Conformation of RNA nucleotide is given by 7 dihedral angles. It is possible to simplify description to 2 “pseudotorsions”

Angles are useful. How to do multivariate analysis on them?

PCA for Angular Data

1. Transformation of angular data using *cosine* (*cos*) and *sine* (*sin*) values in PCA analysis

   Yuguang Mu, Phuong H. Nguyen, and Gerhard Stock, Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis, PROTEINS: Structure, Function, and Bioinformatics 58:45–52 (2005)

   Problem: Neglected correlation $\boldsymbol{cos^2 + sin^2 = 1}$

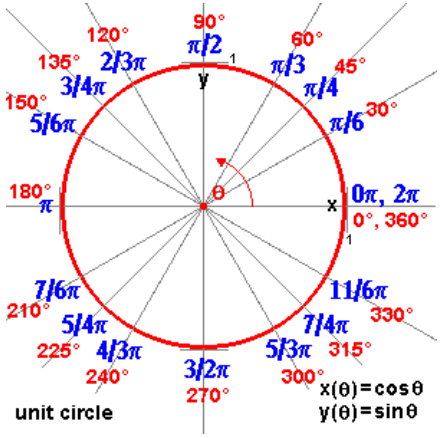
   Konrad Hinsen, Comment on: "Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis", PROTEINS: Structure, Function, and Bioinformatics 64:795–797 (2006)

2. Employment of circular means and circular covariance/correlation matrix

   T.H Reijmers, R Wehrens and L.M.C Buydens, "Circular effects in representations of an RNA nucleotides data set in relation with principal components analysis", Chemometrics and Intelligent Laboratory Systems, Volume 56, Issue 2, 30 May 2001, Pages 61-71

   Problem: Result depends on representation ([0,360] or [-180,180])

M-Sphere Representation



Let $P = (p_1, p_2, \ldots, p_n)$ denote a set of torsion angles describing a molecule,

p_i is a set of angular values/observations $=(a_1^i, a_2^i, \ldots, a_m^i)$

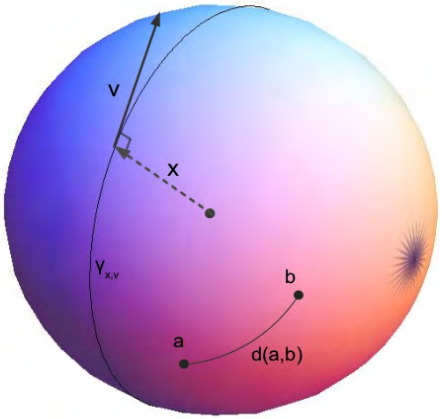
$p_i =(a_1^i, a_2^i, \ldots, a_m^i)$ can be treated as points on the unit *m*-sphere.

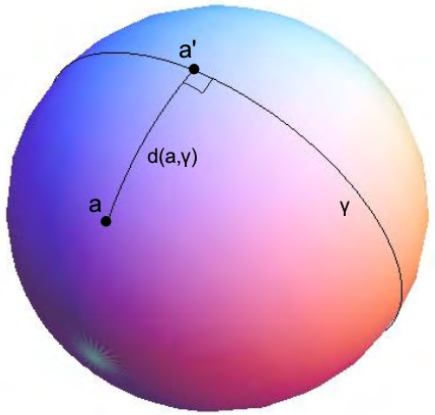
we consider *m*-sphere in (*m*+1) dimensional Euclidean space. As an advantage, inner product on *m*-sphere becomes equal to scalar product in (*m*+1) dimensional Euclidean space. Thus, *m*-sphere is defined as

$$\phi(x) := \langle x, x \rangle - 1 = 0$$

where *x* are points in (*m*+1) dimensional Euclidean space.

Principal Geodesics





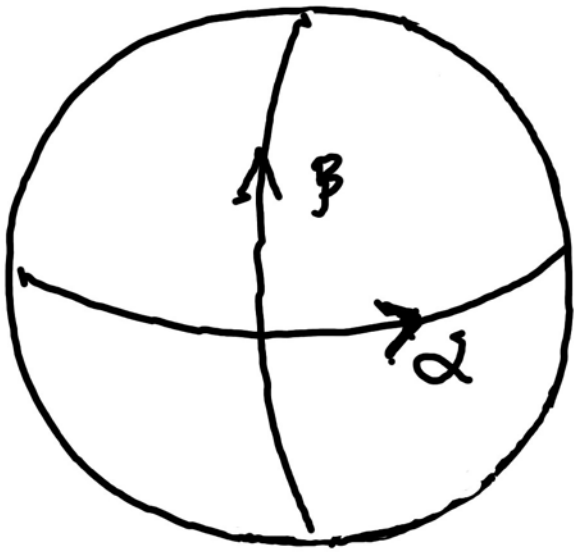
$$d(a,b) = \arccos\langle a,b\rangle$$

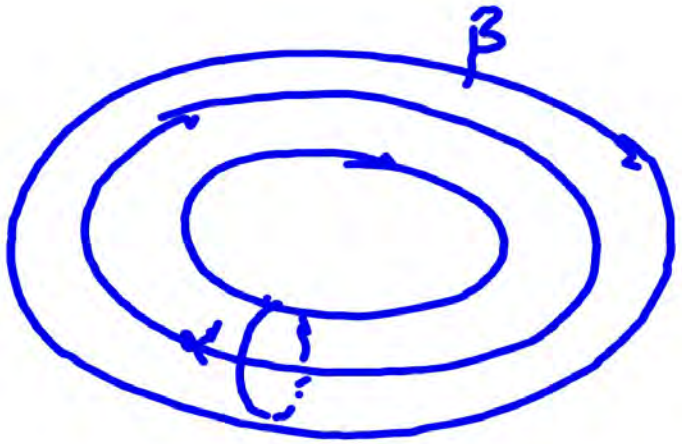
$$d(a,\gamma_{x,v}) = \arccos\sqrt{\langle a,x\rangle^2 + \langle a,v\rangle^2}$$

$$F(x,v) = \sum_{i=1}^{n} d(p_i,\gamma_{x,v})^2$$

$$a' = \frac{\langle x,a\rangle x + \langle v,a\rangle v}{\sqrt{\langle x,a\rangle^2 + \langle v,a\rangle^2}}$$

Possible Topologies for points given by two angular coordinates α and β





Two different topologies yield different connectivity between points. While geometry on the torus is Euclidean, some unnatural restrictions are applied to possible changes of angles

ClustAngles Software

clusta.limlab.ibms.sinica.edu.tw:9100/clusta

Search

中央研究院 ACADEMIA SINICA

Search

Login | Preferences | Help/Guide | About Trac

Wiki | Timeline | Roadmap | Browse Source | View Tickets | New Ticket | Search

wiki: WikiStart

Start Page | Index | History

ClustAngles

ClustAngles: R and Python Packages for multivariate analysis of circular (angular data) with particular attention to needs of Structural Bioinformatics. Packages include such tools as Principle Geodesics Analysis, Spherical and Toric versions of K-means clustering, Spherical and Toric versions of Hierarchical clustering, linear regression for circular data and many other useful functions.

Download ClustAngles 0.1 R version from wiki:InstallR

Download ClustAngles 0.1 Python version from wiki:InstallPython

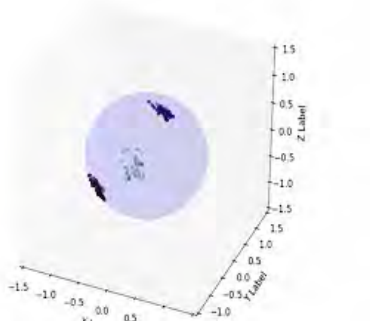
Documentation and tutorial are available at wiki:DocumentationPython and wiki:DocumentationR

contact: karsar at ibms dot sinica dot edu dot tw

In case of any bug or willingness to contribute you may create a Ticket on the web site.

Minor update (no change in package functionality) 9 Oct. 2013

Last modified 15 months ago



Thank You!