Torus Principal Component Analysis with an Application to RNA Structures

Benjamin Eltzner[1,*], Stephan Huckemann[1,*], Kanti V. Mardia[2]

November 17, 2015

Abstract

There are several cutting edge applications needing PCA methods for data on tori and we propose a novel torus-PCA method with important properties that can be generally applied. There are two existing general methods: tangent space PCA and geodesic PCA. However, unlike tangent space PCA, our torus-PCA honors the cyclic topology of the data space whereas, unlike geodesic PCA, our torus-PCA produces a variety of non-winding, non-dense descriptors. This is achieved by deforming tori into spheres and then using a variant of the recently developed principle nested spheres analysis. This PCA analysis involves a step of small sphere fitting and we provide an improved test to avoid overfitting. However, deforming tori into spheres creates singularities. We introduce a data-adaptive pre-clustering technique to keep the singularities away from the data. For the frequently encountered case that the residual variance around the PCA main component is small, we use a post-mode hunting technique for more fine-grained clustering. Thus in general, there are three successive interrelated key steps of torus-PCA in practice: pre-clustering, deformation, and post-mode hunting. We illustrate our method with two recently studied RNA structure (tori) data sets: one is a small RNA data set which is established as the benchmark for PCA and we validate our method through this data. Another is a large RNA data set (containing the small RNA data set) for which we show that our method provides interpretable principal components as well as giving further insight into its structure.

Keywords: Statistics on manifolds, tori deformation, directional statistics, dimension reduction, dihedral angles, angular clustering, fitting small spheres, principle nested spheres analysis.

[1] Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georg-August-University Göttingen

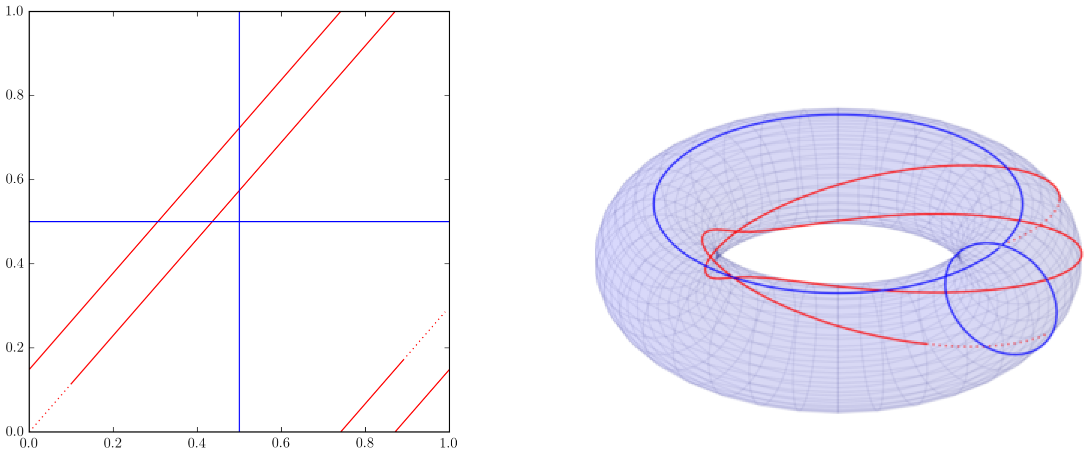
[2] Department of Statistics, University of Oxford and Department of Statistics, University of Leeds

[*] The authors gratefully acknowledge DFG CRC 755 and the Niedersachsen Vorab of the Volkswagen Foundation

1 Introduction

With the rise of the internet, large biomolecule databases, see, for example, Berman et al. (2000), have become publicly available and further the increased computational power has led to a surge in statistical evaluation. In particular, there are cutting edge applications in structural bioinformatics needing PCA methods for data on a torus, for examples, for RNA structural data (see, for example, Sargsyan et al. (2012)) and for protein structural data (see, for example, Altis et al. (2008)). However dimension reduction on non-Euclidean manifolds with PCA-like methods has been a challenging task.

There are two usually successful categories of methods which have been developed in the last decades: tangent space PCA (extrinsic approach, see, for example, Fletcher et al. (2004); Boisvert et al. (2006); Arsigny et al. (2006)), and geodesic PCA (intrinsic approach, see, for example, Huckemann and Ziezold (2006); Sommer (2013)). However, for the very simple non-Euclidean case of the flat and compact space of a torus (a direct product space of two or more angles), these approaches are not adequate. Namely, tangent space PCA fails to take into account the periodicity of the torus and, even worse, geodesic PCA is completely inapplicable because almost all geodesics are densely winding around as seen in Figure 1.



(a) Flat torus as square in $\mathbb{R}^2$ with edges identified.

(b) Curved torus embedded in $\mathbb{R}^3$.

Figure 1: *Flat (1a) and curved (1b) torus representation. Except for horizontal and vertical geodesics (blue) in 1a all other geodesics wind around. All geodesics (red) with an irrational slope in 1a are dense.*

In this paper we propose the novel tool of torus-PCA (T-PCA), which does not suffer from these defects. This is achieved by deforming tori into spheres and then using a variant of the recently developed principle nested spheres analysis (PNS) of Jung et al. (2012). This PNS analysis involves a step of small sphere fitting and we provide an improved test to avoid overfitting. However, deforming the geometry of the torus into that of a sphere – locally glued to itself (to honor periodicity) – creates singularities. We introduce a data-adaptive pre-clustering technique to keep the singularities away from the data. We then apply the torus deformation to clusters separately. Further, mode hunting is utilized to deal with the case of large variance explained by the 1D PC. To sum up, our full T-PCA algorithm (Section 4) consists of three successive steps: pre-clustering, deformed torus PNS (DT-PNS = torus deformation with altered PNS) and post-clustering,

We illustrate the power of our method, using two important RNA data sets. Indeed, the data sets stem from analyses of RNA folding which is believed to be a centerpiece in within-cell communication, see, for example, Chapman et al. (1998); Chakrabarti et al. (2011); Brewer (2013). The folding structure is usually described by *dihedral angles* between neighboring planes, each spanned by three adjacent atoms, similar to pages of an opened book (see Appendix for an illustration). Each nucleic base corresponds to a backbone segment described by 6 angles and one angle for the base, giving a total of 7 angles. Understanding the distribution of these 7 angles over large samples of RNA strands is an intricate problem that has drawn some attention, e. g. Murray et al. (2003); Schneider et al. (2004); Wadley et al. (2007); Richardson et al. (2008); Frellsen et al. (2009).

Simulation studies are frequently used to model and understand interactions of RNA strands with proteins occurring in cells, see Hermann and Westhof (1999); Magee and Warwicker (2005); Zhao et al. (2006); Estarellas et al. (2015). As the computational complexity of full molecular dynamics simulations is very high, there is a large demand for concisely reduced models obtained from investigations of the RNA conformation space. One way of reducing complexity consists in representing the data in a lower dimensional subspace as done by PCA. Another power of PCA lies in providing continuity to a discretely sampled conformational space as in Frellsen et al. (2009). For lack of satisfactory torus PCA-methods, previous studies of RNA residue geometry have made use of the two pseudo-torsion angles η and θ (see Figure 3b), to accomplish a lower dimensional data representation. These η–θ plots (see Figures 4a and 10a), projecting a two-dimensional torus onto the plane, are called Ramachandran plots for example by Duarte and Pyle (1998).

Some torus-specific PCA approaches have been developed apart from tangent space PCA and geodesic PCA. Using wrapped normals, Kent and Mardia (2009) circumvent the problem of winding geodesics and provide for an intrinsic parametric model with the same number of degrees of freedom as classical PCA, which, as discussed in Huckemann and Eltzner (2015), is less than the number of degrees of freedom for our type of approach. The PCA used by Altis et al. (2008) is a particular case of Kent and Mardia (2009). Allowing geodesics only that wind around at most once, as proposed by Kent and Mardia (2015), further reduces the degrees of freedom.

It seems that Sargsyan et al. (2012) have been the first and only to treat toroidal data describing RNA structures in a spherical geometry. In their construction, they halved the corresponding seven torus angles and treated these as polar angles from a seven-dimensional sphere, thus taking only a very first step towards T-PCA. On this seven-dimensional sphere they investigated a test data set consisting of 190 residues. However, Sargsyan et al. (2012) did neither discuss nor exploit the drastic change of geometry and only applied geodesic PCA, see Huckemann and Ziezold (2006), maximizing projected variance and not minimizing residual variance. Incidentally, some pitfalls of using projected variance for compact manifolds have been pointed out in Huckemann et al. (2010).

In our applications, first we use the *small RNA data set* used by Sargsyan et al. (2012) as a benchmark for our T-PCA method. We find that T-PCA retrieves the underlying clusters in an effective way. Then we analyze a classical data set consisting of 8301 residues, subsequently called the *large RNA data set*, which was carefully selected for high experimental X-ray precision (0.3 nanometers) by Duarte and Pyle (1998); Wadley et al. (2007) and analyzed by them and others, for example Murray et al. (2003); Richardson et al. (2008). The small RNA data set is a subset of the large RNA data set consisting of neighborhoods of three known cluster centers in the η–θ-plot (as in Figure 4a). We compare our method to tangent space PCA and show that T-PCA captures much more data variation in low dimensional subspaces, explaining at least 80% of the data variance in the one-dimensional representation whereas in contrast tangent space PCA requires at least two dimensions. Beyond two rather well known clusters we identify a new cluster which has not been found previously.

The plan of the paper is as follows: In Section 2 we introduce DT-PNS, which is the center piece of our methodology. After reviewing the auxiliary clustering methods in Section 3, we present our torus PCA algorithm in Section 4. In Section 5 we apply our method to the small and large RNA data sets and review the results. The paper ends with a discussion. A brief overview of our abbreviations and technical terms used throughout this paper is given in the Appendix.

2 Deformed Torus PNS

2.1 Torus Deformation Schemes

Let $T^D = (\mathbb{S}^1)^{\times D}$ be the D-dimensional unit torus and $\mathbb{S}^D = \{x \in \mathbb{R}^{D+1} : \|x\| = 1\}$ the D-dimensional unit sphere, $D \in \mathbb{N}$. The definition of the data-adaptive deformation mapping $P : T^D \longrightarrow \mathbb{S}^D$ defined in this section is based on a comparison of Riemannian squared line elements. If $\psi_k \in \mathbb{S}^1 = [0, 2\pi]/\sim$ $(k = 1, \ldots, D)$ where $\sim$ denotes the usual identification of

0 with 2π, the squared line element of T^D is given by the Euclidean

$$ds^2 = \sum_{k=1}^{D} d\psi_k^2 \,.$$

For $\mathbb{S}^D$, in polar coordinates $\phi_k \in [0,\pi]$ for $k = 1,\dots,D-1$ and $\phi_D \in [0,2\pi]/\sim$, whose relation to embedding space coordinates x_k is elaborated in the Appendix, the spherical squared line element is given by

$$ds^2 = d\phi_1^2 + \sum_{k=2}^{D} \left(\prod_{j=1}^{k-1} \sin^2 \phi_j \right) d\phi_k^2 \,. \tag{1}$$

In fact, this squared line element is not defined for the full sphere but only for $\phi_k \in (0,\pi)$ ($k = 1,\dots,D-1$). The singularities at $\phi_k = 0,\pi$ will account for singularities of P which form a subtorus of dimension $D-2$ (or a union of self-glued subtori). Because in (1), $d\phi_1^2$ comes with the factor 1, no deformation at all occurs for ϕ_1, i.e. this angle corresponds to spherical distances without distortion. In the summation for $k = 2$ we have a factor $\sin^2 \phi_1$ of $d\phi_2^2$, which shows how the angle ϕ_1 distorts the angle ϕ_2 and finally the deformation factor $\prod_{j=1}^{D-1} \sin^2 \phi_j$ of $d\phi_D^2$ reflects the distortions of ϕ_D by all other angles. For this reason, in the following, we will refer to ϕ_D as the *innermost angle* and to ϕ_1 as the *outermost angle.*

Remark 2.1. *At this point note that near the equatorial great circle given by $\phi_k = \frac{\pi}{2}$ ($k = 1,\dots,D-1$) this squared line element is nearly Euclidean. Distortions occur whenever leaving the equatorial great circle. More precisely, distortions are higher when angles ϕ_k with low values of the index k are close to zero, than when angles ϕ_k with high values of the index k are close to zero.*

Definition 2.2 (Torus to Sphere Deformation)**.** *With a data-driven permutation p of $\{1,\dots,D\}$, data-driven central angles μ_k ($k = 1,\dots,D$) and data-driven scalings α_k, all of which are described below, set*

$$\phi_k = \frac{\pi}{2} + \alpha_{p(k)}(\psi_{p(k)} - \mu_{p(k)}), \quad k = 1,\dots,D \tag{2}$$

where $p(k)$ is the index k permuted by p and the difference $(\psi_{p(k)} - \mu_{p(k)})$ is taken modulo 2π such that it is in the range $(-\pi,\pi]$.

In general, the scalings are restricted to the choices $\alpha_{k'} = 1/2$ and $\alpha_{k'} = 1$, $k' = p(k)$. If all of the k'-th torus angles of the data are within an interval of length π, choose $\alpha_{k'} = 1$ ($k' = 1,\dots,D-1$) leading to *unscaled* (U) angles. Else choose $\alpha_{k'} = 1/2$ ($k' = 1,\dots,D-1$) leading to *halved* (H) angles. The innermost angle will always remain unscaled, $\alpha_D = 1$. In practical situations, the torus data are often spread out over more than half circles for several angles. Then we choose (H) angles. In rare cases where data is concentrated we can choose (U) angles.

The central angles (μ_k) will be chosen such that data points come to lie near the equatorial great circle and omit the singularities. Two plausible choices are:

(i) with the circular intrinsic mean $\overline{\psi}_{k,\mathrm{intr}}$ (we use the fast algorithm from Hotz and Huckemann (2014)), set $\mu_k = \overline{\psi}_{k,\mathrm{intr}}$ to obtain *mean centered* (MC) data

(ii) with $\psi_{k,\mathrm{gap}}$, the center of the largest gap between neighboring ψ_k values of data points and $\psi^*_{k,\mathrm{gap}}$ its antipodal point, define $\mu_k = \psi^*_{k,\mathrm{gap}}$ to obtain *gap (antipode) centered* (GC) data.

MC data has the merit that the torus mean of the data is mapped to the equatorial great circle and thus, in that sense, deformation of the data is minimized. For a strongly skewed data distribution, spread out over a half circle, halved GC data will still be confined to a $\pi/2$ neighborhood of the equator while halved MC data will touch the singularities, leading to high distortion there. For data sets with outliers, GC centering may not be robust, making MC more favorable.

The choice of the permutation (p) is driven by analyses of the *data spread*

$$\sigma_k^2 = \sum_{i=1}^{n} (\psi_{k,n} - \mu_k)^2 \tag{3}$$

for each angle, where $\psi_{k,i} \in \mathbb{S}^1$ are the torus data and n is the number of data points on T^D. If the angles are ordered by increasing data spread, such that $\sigma^2_{p(1)}$ is minimal and $\sigma^2_{p(D)}$ is maximal, in view of Remark 2.1, the change of distances between data points caused by the deformation factors $\sin^2 \phi_j$ in Equation (1) is minimized. We call this case *spread inside* (SI), because variation is concentrated on the inner angles of the sphere. The opposite ordering is called *spread outside* (SO). We will restrict consideration to these two options.

Due to periodicity on the torus, $\psi_k = 0$ is identified with $\psi_k = 2\pi$ for all $k = 1, \dots, D$. In contrast, for all angles $\phi_k = 0$, with $k = 1, \dots, D-1$, denotes spherical locations different from $\phi_k = \pi$. For an invariant representation respecting the torus' topology, however, it is necessary to identify these locations accordingly, which results in a *self-gluing* of the $\mathbb{S}^D$ as elaborated in more detail in the Appendix.

2.2 Linking the Torus' Deformation to PNS

For data sets on a torus, we apply a deformation as detailed in Section 2.1, in particular a data driven choice of scalings ($\alpha_k = 1/2$ if the data in one of the angles except for the innermost is spread out on more than a half circle or else $\alpha_k = 1$) and of centering (MC or GC) is performed. On the resulting self-glued $\mathbb{S}^D$ we use an alteration of principal nested sphere analysis (PNS) by Jung et al. (2010, 2012) for dimension reduction.

The PNS iteration leads to a sequence of small subspheres

$$\mathbb{S}^D \supset S^{D-1} \supset \cdots \supset S^2 \supset S^1 \supset \{\mu\}. \tag{4}$$

The ultimate point μ is called the *nested mean.* For real data applications, with probability one, the S^d are not great subspheres but proper small subspheres ($d = 1, \ldots, D-1$), the radii of which decrease monotone (as discussed further in Section 2.4). At each reduction step, the residues are given as signed distances: points lying inside the small subsphere receive a positive distance, points lying outside a negative distance.

The PNS algorithm consists of two parts which alternate, namely the fitting of a subsphere S^{d-1} and the projection to this subsphere $\pi_d : S^d \to S^{d-1}$ ($d = D, D-1, \ldots, 1$). If $\mathbb{S}^D$ is glued to itself, in the fitting step as well as in the projection step, distances through the glued part (which may be shorter than the spherical distance) can be used instead of spherical distances only, as in classical PNS. We find from experiments that the fitting procedure when taking into account gluing is numerically badly behaved and tends toward local minima, even more so if we use $\tilde{\delta}$ introduced below in (5). For this reason we alter classical PNS by taking into account the topological identifications only in the projection step. Thus, data fidelity is preserved while the simplified choice of subspace sequences is supported by the resulting good low dimensional description of the data.

2.3 Comparing Variances

In Euclidean spaces, PCA variances are additive with monotone decrements leading to a convex variance plot as a property of the metric. In non-Euclidean spaces, this is no longer the case (see the discussion for various definitions of intrinsic variances in Huckemann et al. (2010)). Even worse, comparing variances of different clusters is further complicated by each cluster being analyzed in its own data-adaptive deformed torus. In order to perform a meaningful comparison of variances, we propose the following calculation of residual variances as measures for the quality of the fit.

Assume a cluster $\mathcal{C}$ and a corresponding adaptive deformation $P_{\mathcal{C}} : T^D \to \mathbb{S}^D$. Using the inverse deformation $P_{\mathcal{C}}^{-1}$ (which is well defined except for the singularities) and the torus metric

$$\delta : T^D \times T^D \to \mathbb{R}^{\geq 0} \qquad (p, q) \mapsto \left(\sum_{i=1}^{D} \min\left(|p_i - q_i|^2, (2\pi - |p_i - q_i|)^2 \right) \right)^{\frac{1}{2}}$$

we define the following function on the sphere

$$\tilde{\delta} : \mathbb{S}^D \times \mathbb{S}^D \to \mathbb{R}^{\geq 0} \qquad (x, y) \mapsto \delta\big(P_{\mathcal{C}}^{-1}(x), P_{\mathcal{C}}^{-1}(y)\big)\,. \tag{5}$$

This is a metric when we take into account the topological identifications. Recall that PNS yields a sequence of subspaces $\mathbb{S}^D \supset S^{D-1} \supset \cdots \supset S^1 \supset \{\mu\}$ with projections $\pi_d : S^{d+1} \to$

$S^d \subset S^{d+1}$, $\pi_0 : S^1 \to \{\mu\}$. From these we define the iterated projections

$$\Pi_d = \pi_d \circ \pi_{d+1} \circ \cdots \circ \pi_{D-1}$$

and finally the residual variances (variance not explained by S^d)

$$V_{\mathcal{C},P_{\mathcal{C}},d} = \sum_{q \in \mathcal{C}} \tilde{\delta}^2(q, \Pi_d(q)) \, .$$

Due to nestedness, these sequences are non-increasing with d. However, the decrements $V_{\mathcal{C},P_{\mathcal{C}},d-1} - V_{\mathcal{C},P_{\mathcal{C}},d}$ ($d = 1, \ldots, D$) are not necessarily non-increasing, so the resulting curve in the variance plot need not be convex as seen in Figure 7 (discussed in Section 5.2). This is in contrast to the Euclidean case, where the plot is always convex because decrements correspond to the non-increasingly ordered eigenvalues of the corresponding covariance matrix. In order to honor differences of densities over different clusters, we normalize all residual variances by dividing by the common scale given by the *total variance*

$$V_0 := \min_{p \in T^D} \sum_{q \in \mathcal{Z}} \tilde{\delta}^2(q, p) \tag{6}$$

over the full data set $\mathcal{Z}$. If we would individually normalize the residual variances of a cluster $\mathcal{C}$ by its total variance or by $V_{\mathcal{C},P_{\mathcal{C}},0}$ then a concentrated and isotropic cluster would yield a nearly linear residual variance plot, suggesting that the data may be high dimensional. Normalizing by the common scale, however, still yields nearly a line, now starting well below 100% at the zero-dimensional approximation, more realistically suggesting that the data is zero-dimensional (see Figures 7 and 8).

2.4 Improved PNS: Avoiding Overfitting

In the PNS algorithm a cluster of points concentrated around a single center may still be best fitted by a very small subsphere. As this obvious overfitting is undesirable, Jung et al. (2011, 2012) would rather fit a great subsphere in such cases and give tests for this purpose. We propose an improved test based on geometrically better hypotheses and a likelihood ratio, detailed in the Appendix.

We carried out a simulation study to compare the tests. Test data conforming to the null hypothesis (a cluster leading to a great sphere fit) have been generated, by simulating isotropically normal distributed points in a tangent plane with standard deviations σ uniform in $[0.1, 0.45]$ truncated to 2σ and projecting these points orthogonally to the sphere. Test data for the alternative (small sphere) has a uniform angular distribution and a non-centered normal radial distribution with means μ_r uniform in $[0.1, 0.5]$ and σ uniform in $[0.01\mu_r, 0.5\mu_r]$. The resulting distribution is then truncated to the unit circle. The results are displayed in Table 1. In effect of modeling specific hypotheses (among others a ring with a central cluster as detailed in the Appendix), the errors of the first kind of the tests by Jung et al. (2011, 2012) are unacceptably high, whereas for our test, these are not as high. Our test features approximately the same order for the errors of the second kind, whereas for Jung et al. (2011, 2012), they are much smaller. Our test is clearly an improvement, approximately giving an equal error rate. However, its fine tuning and robust extension to more general data models warrants further investigation beyond the scope of this paper.

Table 1: *Errors of three test; of the first kind (falsely rejecting the null hypothesis of a great sphere) and of the second kind (falsely accepting the alternative of a small sphere) in a simulation with* 1000 *test clusters.*

	our test		Jung et al. (2011)		Jung et al. (2012)	
Cluster size	1st kind	2nd kind	1st kind	2nd kind	1st kind	2nd kind
30	18.6%	21.3%	93.8%	1.0%	52.3%	9.3%
100	8.9%	18.0%	84.8%	3.3%	71.5%	5.1%
300	10.9%	12.3%	80.6%	2.6%	95.7%	1.6%
1000	23.6%	10.3%	84.3%	4.1%	100.0%	0.6%

3 Pre- and Post-Clustering

3.1 Single Linkage Pre-Clustering

We show in Figure 2 a toy example on T^2 which highlights two clusters, each of curved one-dimensional data, but entangled. Without treating each cluster separately, DT-PNS fails to discover the one-dimensional structures and suggests a two-dimensional approximation of the joint data. Hence, we need a clustering method that specifically uncovers entangled curved clusters and allows to separate them. It turns out that hierarchical single linkage clustering (also called "nearest neighbor clustering"), e.g. Mardia et al. (1979), is a suitable method for this task.

Single linkage clustering joins nearest neighbors recursively to a binary tree structure which is pruned by branch cuts. Each proper node of the tree carries a value denoting the minimal distance between leaves of its two branches, such that node values increase when approaching the root.

Determining suitable branch cuts for such a cluster tree is a delicate issue that has received attention recently. Especially, methods for data adaptive branch cutting instead of cutting at a fixed level have been proposed, e.g. by Langfelder et al. (2008); Obulkasim et al. (2015). We have designed a similar data adaptive branch cutting recursive procedure

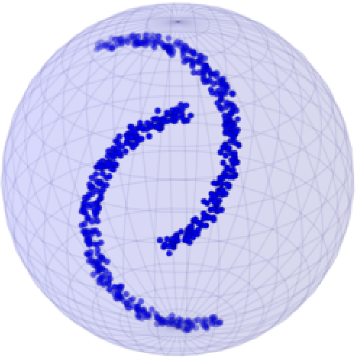


Figure 2: *An example of two entangled half circles.*

detailed in the Appendix. This returns clusters of decreasing density, populating surprisingly low dimensional subspaces of the data space, often entangled as sketched in Figure 2.

3.2 Post-Clustering: Mode Hunting

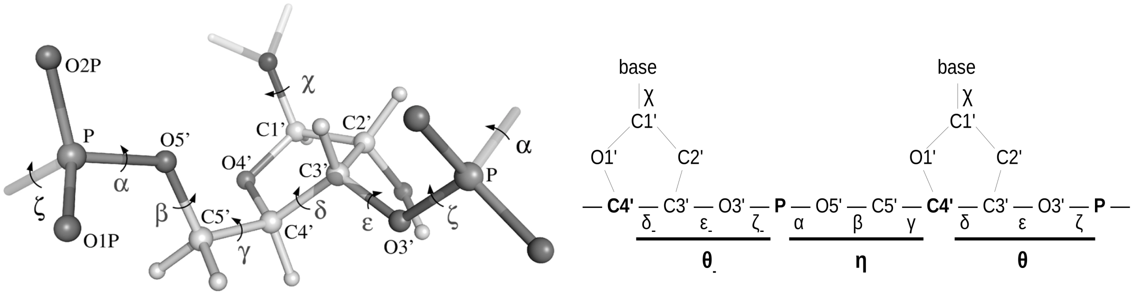
As a final step, we analyze the one-dimensional projection of the data using the multiscale method described by Dümbgen and Walther (2008) for mode hunting. Although this method was originally defined for the real line, its numerical implementation for circular data is even simpler. Since modes are separated by minima, we use this method to identify regions in which minima are located with a certain confidence level. (Throughout the applications, we use a fixed confidence level of 95%.) For circular data, we use a wrapped Gaussian kernel smoother for the one-dimensional projections of the points. For every region with minimal smoothed density we increase the kernel's width until there is exactly one minimum of the smoothed distribution left. Here, we separate the modes.

4 Torus PCA: A Brief Overview

To a given data set $\mathcal{Z}$ we first apply DT-PNS as described in Section 2 for all deformations (centering with MC or GC and permuting via SI or SO). If for none of these deformations the residual variance to the penultimate small sphere (actually a 1D small circle) is below a threshold of 20%, pre-clustering as described in Section 3.1 is performed and DT-PNS as above is applied again for each cluster until the threshold is reached. If none of the deformations achieve a residual variance below 20%, the cluster is declared final; otherwise, we apply mode hunting as detailed in Section 3.2. If mode hunting detects new sub-clusters, these are added to the list of clusters and DT-PNS is performed again on each new sub-cluster. A flow chart for the T-PCA algorithm is included in the Appendix.

5 Application to RNA Structure

For our applications, we need some background on RNA molecules, which is provided in the following. RNA molecules, like DNA, consist of three building blocks, a phosphate group (O3'-P-O5'), a 5-carbon sugar (pentose) and a nucleic base, see Figure 3b. The phosphate groups and sugars form the *backbone* in an alternating sequence, where a nucleic base is attached to each sugar, see Figure 3b. Each triple is called a *residue* (nucleotide), where, however, for reasons of symmetry, boundaries are not put at phosphate groups beginnings/endings, rather the parts between two consecutive phosphor atoms are considered. The sugar molecules are strictly directed, the C5' to C3' atoms being part of the backbone and the C4' to C1' atoms forming a ring. The nucleic base is attached to the C1' atom, which is furthest from the backbone. Four different standard bases exist, namely Adenine (A), Cytosine (C), Guanine (G) and Uracil (U); while A, C, and G are also common in DNA but U is replaced by Thymine (T) in DNA. In distinction to the DNA backbone, an oxygen atom is attached to the C2' atom, as displayed in Figure 3a. As a result, the C3'-endo sugar pucker (non-planar sugar ring) is energetically preferred and thus by far more common for RNA. For more details we refer the reader to Egli and Saenger (1984).



(a) *3D structure of an RNA residue.*

(b) *2D scheme of an RNA residue.*

Figure 3: *Part of an RNA backbone (Phosphate groups followed by sugars to which a nucleic base it bound). Dihedral angles (Greek letters) are defined by three bonds, the central bond carries the label; pseudo-torsion angles (bold Greek letters) are defined by the pseudo-bonds between bold printed atoms (Figure 3b). The precise definition with same canonical atom notation are given in Table 2. O denotes oxygen, C carbon and P phosphor. The subscript "−" denotes angles of the neighboring residue. Figure 3a is from Frellsen et al. (2009).*

In further contrast to DNA, which usually takes a double-stranded helical conformation, RNA is usually single-stranded and the single strand interacts with itself, forming complex shapes. This means that the geometry is much more variable even on the scale of single atoms. Each nucleic base corresponds to a backbone segment described by 6 dihedral angles and one angle for the base, giving a total of 7 angles. Understanding the distribution of these 7 angles over large samples of RNA strands is an intricate problem that has drawn some attention, see Murray et al. (2003); Schneider et al. (2004); Wadley et al. (2007); Richardson et al. (2008); Frellsen et al. (2009). Figure 3a details a segment of the RNA backbone with seven angles for each residue giving the 3D folding structure and Table 2 gives the canonical names of the atoms involved in the definition of each angle.

Table 2: *Seven dihedral angles (α, β, γ, δ, ϵ, ζ, χ) and two pseudo-torsion angles (η, θ) in terms of their corresponding four atoms. Figure 3a shows the geometry of these atoms. (N denotes nitrogen.)*

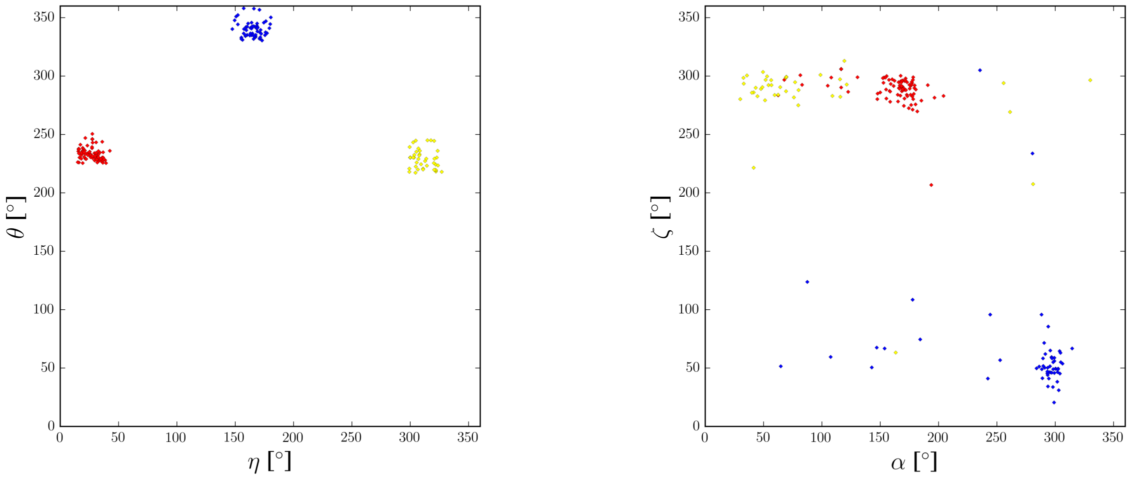
α	$O3'-P-O5'-C5'$
β	$P-O5'-C5'-C4'$
γ	$O5'-C5'-C4'-C3'$
δ	$C5'-C4'-C3'-O3'$
ϵ	$C4'-C3'-O3'-P$
ζ	$C3'-O3'-P-O5'$
χ	$O4'-C1'-N1-C2$ for pyrimidine (monocyclic) bases $O4'-C1'-N9-C4$ for purine (bicyclic) bases
η	$C4'-P-C4'-P$
θ	$P-C4'-P-C4'$

An approximation of the geometric folding structure on the level of single residues is given by the two *pseudo-torsion angles* η and θ (Figure 3b and Table 2). These provide at once a two-dimensional visualization (Figure 4a), see e.g. Duarte and Pyle (1998); Wadley et al. (2007). Clustering and structure investigation based on the purely backbone torsion angles δ_{-}, ϵ_{-}, ζ_{-}, α, β, γ and δ (see Figure 3b) has been performed by Murray et al. (2003); Richardson et al. (2008).

5.1 The Small RNA Data Set

This small RNA data set has been carefully selected by Sargsyan et al. (2012) to validate their method. They took clusters labeled I (blue, 59 points), II (red, 88 points) and V (yellow, 43 points) by Wadley et al. (2007) totaling 190 data points, which form three clusters in the η–θ plot as shown in Figure 4a. While clusters I and II correspond to distinct structural elements featuring base stacking, the residues in cluster V belong to a wider variety of structural elements.

As challenge, however, this distinction cannot be readily seen in the 7D space of all torsion angles. Figure 4b depicts the most discriminant angle pair (α, ζ): The yellow cluster is not very concentrated and parts of it are very close to the red cluster, which is twice as



(a) *η–θ plot*

(b) *α–ζ plot*

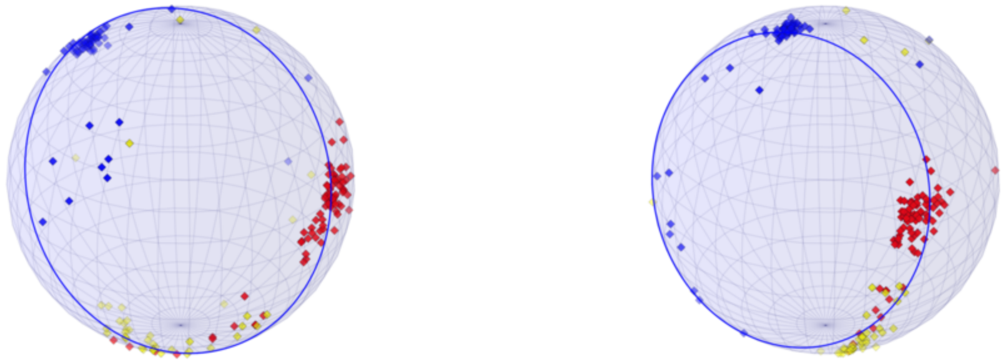
Figure 4: *4a: The small RNA data set of Sargsyan et al. (2012) with their three preselected clusters in the η–θ plot. 4b: The small RNA data set plotted in the two most discriminant (α, ζ) out of the seven dihedral angles.*

big. In fact, upon close inspection, due to periodicity, the red and yellow clusters are also rather close in the η–θ plot in Figure 4a.

Without pre-clustering and without post-mode hunting, we have applied T-PCA to all seven angles and depict the two-dimensional representation both for SI (Figure 5a) and SO (Figure 5b) ordering in Figure 5. The data are, in fact, very well approximated by the best fit circle. Using the same coloring for Figure 5 as Figure 4 shows that the three preselected clusters can be rather well distinguished by eye with slightly better distinction for SO ordering.

When mode-hunting (see Section 3.2) is applied to the one-dimensional T-PCA representation, for SI-ordering we only find two clusters which correspond roughly to the original blue cluster and the union of the original red and yellow cluster in the η–θ plot; recall that the latter are also nearby in the η–θ plot. Using mode hunting with SO ordering the relative distance of the red and yellow cluster is visibly enhanced and we find again three clusters which are only slightly different from the preselected ones. Figure 6 assigns colors to clusters found by T-PCA with mode-hunting and depicts these in the η–θ plot. As some outliers occur in the 7-dimensional representation, we chose mean centered angles.

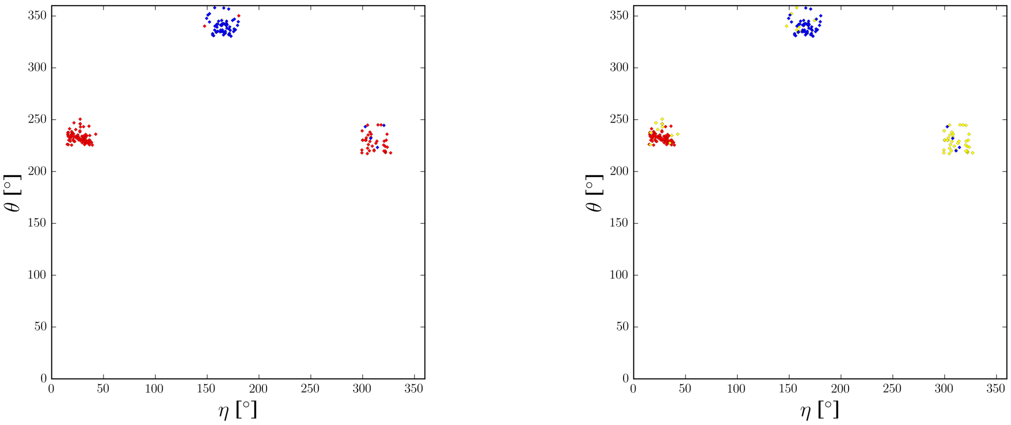
This result illustrates the power of backward dimension reduction methods going significantly beyond the analysis of Sargsyan et al. (2012). Not only can the preselected clusters be separated but the data are very accurately approximated by their projection to a circle. Additionally, some points can be identified, which stray so far from the bulk of their designated



(a) *2D approximation, SI*

(b) *2D approximation, SO*

Figure 5: *Two-dimensional T-PCA approximation of the small RNA data set with SI (5a) and SO (5b) ordering. (Colors representing the same clusters as in Figure 4).*



(a) *SI η–θ plot*

(b) *SO η–θ plot*

Figure 6: *The small RNA data set in pseudo-torsion angles with clusters obtained from T-PCA and labelled with colors red, blue and yellow. 6a: mode hunting and SI. 6b: mode hunting and SO.*

cluster in the 7D representation that they are attributed to other clusters by T-PCA (e.g. the blue and yellow points in the left middle in Figure 5a). Indeed, applying our T-PCA with pre-clustering, we can identify these points as outliers not belonging to any of the three clusters, as elaborated in the Appendix.

5.2 The Large RNA Data Set

The large RNA data set consists of 8301 data points. These data spread out widely in almost all seven dihedral angles, so gluing effects must be taken into account for a T-PCA analysis with neither pre- nor post-clustering. To the end of minimizing the effect of these topological degeneracies we use gap centered angles in this case. The residual variance plot in Figure 7 of the full data indicates that at least 4 or 5 dimensions are necessary to obtain at most 20% residual variance (residual variance is defined in Section 2.3). To achieve better dimension reduction we thus need to pre-cluster the data.

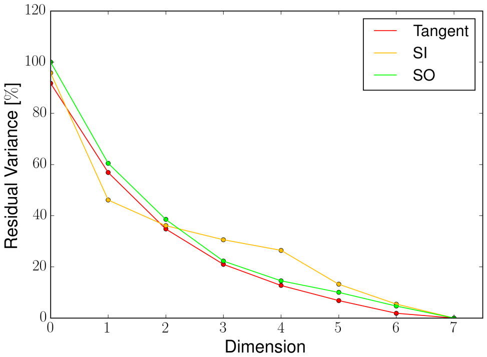


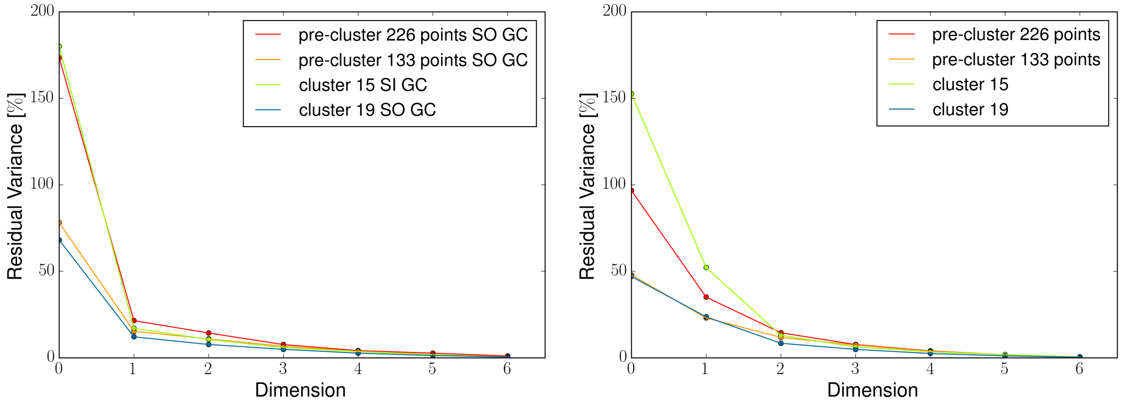
Figure 7: *Large RNA data set: Residual variances for T-PCA (SI and SO without clustering).*

5.2.1 PCAs of Clustered Data

T-PCA method: By pre-clustering we find 15 clusters, which will in the following be called *pre-clusters* and are listed in the Appendix. About 10% of the data are characterized as outliers. In Figure 8a we give an illustration of the residual variances for pre-clusters 5, 8, 11 and 12 that shows the low residual variance of the one-dimensional T-PCA representation. For almost all pre-clusters, the one-dimensional T-PCA representation has less than 20% residual data variance in relation to total data variance. The percentage at dimension zero for each pre-cluster gives its total variance relative to the large RNA data set's total variance as detailed in Section 2.3, in order to make non-Euclidean variances comparable with one-another.

Due to low residual variances of the one-dimensional projections of pre-clusters, we can use post-mode hunting and meaningfully interpret the found modes as clusters. This yields 22 *final clusters* with overall decreased variance and dimensionality; the smallest final cluster contains 28 points. Of the pre-clusters used in Figure 8a, pre-cluster 5 is decomposed into final clusters 11, 15 and 17 and pre-cluster 8 is decomposed into final clusters 14 and 18 while pre-clusters 11 and 12 remain unchanged as final clusters 16 and 19, respectively. Several of the final clusters, especially final clusters of low density, have apparently not been described before. Final clusters are also listed in the Appendix.

Comparison with tangent space PCA: We find that for the pre-clusters 5, 8, 11 and 12, as used above for the T-PCA, tangent space PCA leaves more than 30% residual data variance in the one-dimensional representation (see Figure 8b), whereas the one-dimensional T-PCA representation has less than 20% residual data variance. The root of the success of T-PCA over tangent space PCA can be seen by investigating those pre-clusters which are much better approximated by T-PCA. All of these pre-clusters have distinct non-linear shapes, so they are badly fitted by linear subspaces. Figure 9 displays two-dimensional projections of pre-clusters 5 and 12, whose residual variances are displayed in Figure 8. In particular the non-linear shapes are preserved in the two-dimensional representation and one can clearly see that they are one-dimensionally much better fit by small circles which is impossible (let alone the two-dimensional spherical representation) via a tangent space approach.



(a) DT-PCA variances

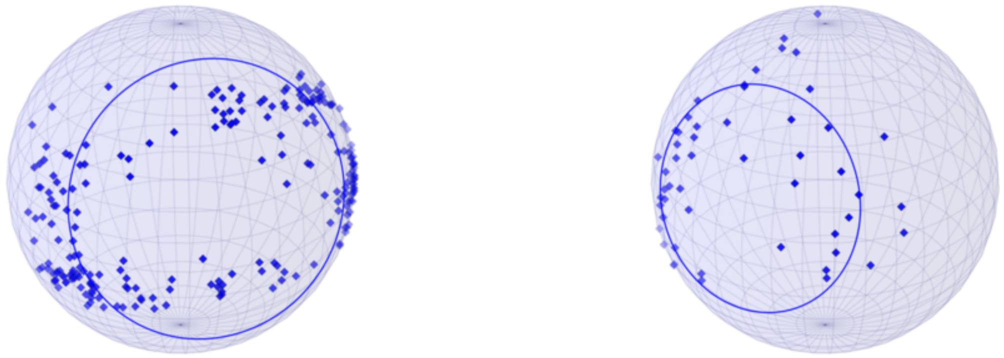
(b) Tangent space PCA variances

Figure 8: *Residual variance plots of pre-clusters: T-PCA (8a) versus tangent space PCA (8b). These plots include only pre-clusters where the results differ markedly.*

5.2.2 Locating a New Low Density Cluster

We now illustrate the power of our method by example of using three final clusters from the large RNA data set. These clusters are numbers 1, 2 and 7 out of the 22 final clusters listed in the Appendix (Tables 1 and 2); these contain a total of 5625 out of 8301 data points. These three clusters have been selected because they strongly overlap and are inseparable in the 2D pseudo-torsion representation. To large parts, the high density clusters 1 and 2 have been described by Richardson et al. (2008). Cluster 7 has low density and could only be found by T-PCA.

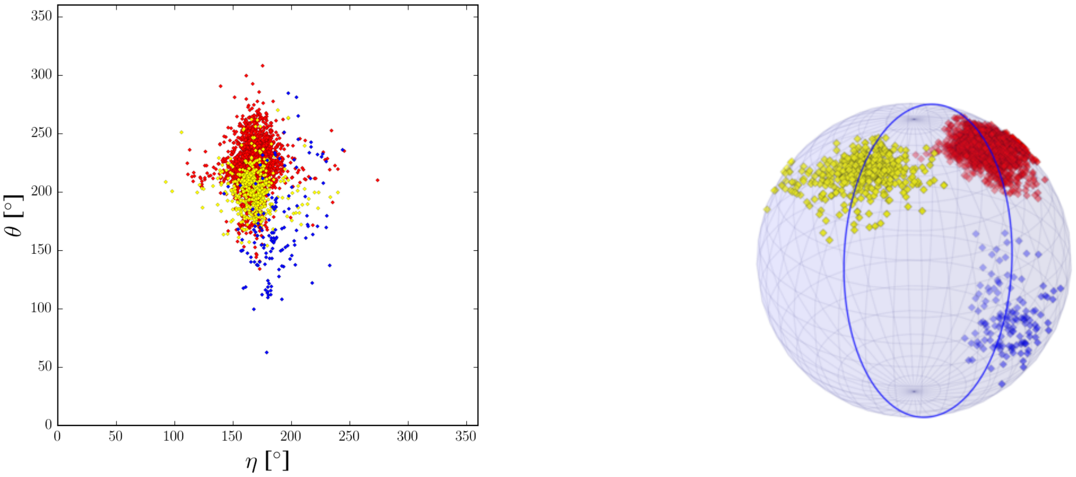
The molecular properties assembled in Tables 3 show that cluster 7 (blue in Figure 10) is very different from the other two clusters (cluster 1 in red and cluster 2 in yellow in Figure 10). One of the most striking differences, see Table 3a, is that clusters 1 and 2 consist only



(a) *Pre-cluster 5 with 226 points*

(b) *Pre-cluster 12 with 52 points*

Figure 9: *Two-dimensional approximations of pre-clusters 5 and 12 from the large RNA data set for which T-PCA without mode hunting yields much better one-dimensional representations than tangent space PCA. The circles on the spheres illustrate the best fit circles found by T-PCA, along which mode hunting is performed.*



(a) η–θ plot

(b) $2d$ T-PCA subsphere plot

Figure 10: *Clusters* 1 *in red,* 2 *in yellow and* 7 *in blue. They overlap in the* η–θ *plot (10a) but can easily be separated in the 7D torus of* α-β-γ-δ-ϵ-ζ-χ *by T-PCA, as illustrated by the two-dimensional projection displayed in 10b.*

Table 3: *Properties of clusters* 1 *(red),* 2 *(yellow) and* 7 *(blue) of Figure 10. The "# Bonds" section of Table 3a gives the relative number of bases bound to* 1, 2 *or* 3 *other bases. Base pair types in Table 3b are denoted by five letters representing in this order: 1. bond region of the base, 2. bond region of its partner, 3. cis/trans bond geometry, 4. type of the base, 5. type of its partner. The bond regions are denoted as follows: W for Watson-Crick, H for Hoogsteen, S for Sugar. For example, "**SHtGA**" means that the sugar edge (**S**) of a cluster residue is bound to the Hoogsteen edge (**H**) of another residue with trans (**t**) alignment, where the cluster residue is a Guanine (**G**) base and its bond partner is an Adenine base (**A**). The information in the "# Bonds" section of Table 3a and all of Table 3b was extracted using the proprietary RNAview software, see Yang et al. (2003).*

(a) General information

Cluster #	1	2	7
# Points	4921	477	137
C3'-endo	100%	100%	3.62%
Bases			
A	19.89%	16.35%	25.36%
C	31.64%	24.74%	11.59%
G	32.55%	46.75%	42.75%
U	15.91%	12.16%	20.29%
# Pairs			
1	71.88%	67.92%	44.20%
2	17.84%	18.66%	30.43%
3	3.13%	5.03%	4.35%

(b) Bond information

Cluster #	1	2	7
# Points	4921	477	137
Bonds			
WWcCG	26.91%	20.13%	5.07%
WWcGC	23.15%	38.16%	7.25%
WWcAU	7.88%	4.61%	1.45%
WWcUA	8.01%	4.40%	0.72%
SHtGA	1.95%	1.68%	16.67%
SHcAA	0.08%	0.00%	5.07%
HWtAU	0.33%	1.68%	4.35%
WWcGU	2.84%	4.19%	0.72%
none	6.77%	8.18%	21.01%

of residues with C3'-endo sugar pucker, while cluster 7 features mostly the alternate C2'-endo sugar pucker. Furthermore, residues in cluster 7 have mostly irregular or no bonds (all bonds not of base pair type "WWc$\cdots$" i.e. below the first four in Table 3b are irregular), and display an abundance of bases connected to more than one other base. Clusters 1 and 2 are very similar in terms of bond patterns although these differ most strikingly by the excess of Guanine residues in cluster 2 of 47% versus 33%, see Table 3a, which is mirrored by an excess of corresponding bonds.

The only dihedral angles for which nested mean values differ significantly among cluster 1 and 2 are α and γ while the nested mean of cluster 7 deviates from that of cluster 1 in the angles δ and ζ. For cluster 1, $\alpha \approx 297°$, $\gamma \approx 53°$, $\delta \approx 81°$ and $\zeta \approx 293°$, while for cluster 2, $\alpha \approx 154°$ and $\gamma \approx 176°$ and for cluster 7, $\delta \approx 147°$ and $\zeta \approx 149°$. Thus, residues from both clusters 2 and 7 can be regarded as kinked versions of the residues of cluster 1 as illustrated in Figure 11. Our findings concerning δ seem well in accordance with previous results, e.g. Richardson et al. (2008) note that the C3'-endo pucker corresponds to a mean δ between

78° and 90° while the C2'-endo pucker leads to mean δ between 140° and 152°. The sugar puckers explain that cluster 2 is larger than cluster 7, see Table 3a.

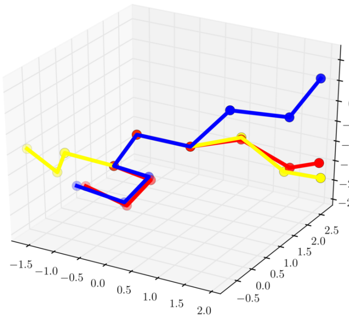


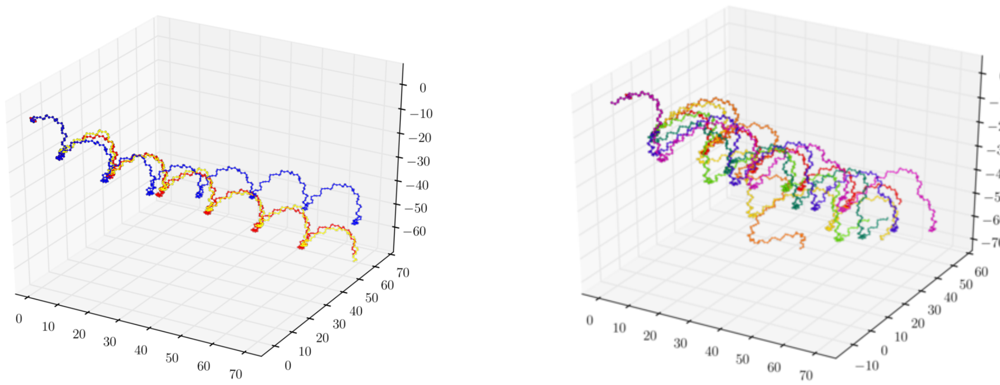
Figure 11: *Atomic geometries of the backbone pieces corresponding to the DT-PNS nested means for cluster* 1 *(red),* 2 *(yellow) and* 7 *(blue) with same colors as in Figure 10. This display is analogous to Figure 3a. The first three and last three atoms represent the phosphate groups and the chains are aligned along the three carbon atoms from the sugar to visualize similarities between the structures. (For clear vizualization we have used constant bond length.)*

5.3 Helical Structures

We continue with the three clusters of Section 5.2.2. To describe the typical residue geometry of a cluster, we use its nested mean, as defined in Section 2.2, in the following. A backbone consisting only of typical residues from cluster 1 takes a helical shape with approximately 11 residues forming one turn, which is typical for the A-helix conformation frequently found in RNA structures, see Figure 12a and Duarte and Pyle (1998); Wadley et al. (2007); Richardson et al. (2008). It is not possible to form such structures consisting of typical residues either alone from cluster 2 or alone from cluster 7 as these result in a tightly wound helix, which is incompatible with bases attached to the residues. Interspersing a backbone of typical residues from cluster 1, however, with typical residues of cluster 2 leads to a variety of bent or loose helical shapes, see Figure 12a. This is consistent with the often irregular shapes of RNA molecules, see e.g. Wadley et al. (2007). Typical residues from cluster 7 when interspersed in a backbone consisting mostly of typical residues from cluster 1 result in very irregular strands, some of which are depicted in Figure 12b. These seem not frequently observed as larger secondary structure, however. This suggests that such mixed conformations are rather unusual in long strands.

Indeed, this is confirmed by our investigation of the RNA chains of the large RNA data set we use, where we identify contiguous sequences of residues from the three clusters investigated here. The sequences containing residues from cluster 7 are much shorter on average than those containing residues from cluster 2. Furthermore, the residues from cluster 7 are much less likely to be located in the middle of a sequence than those from cluster 2, indicating that they are not usually part of rather regular helical regions.

In summary, clusters 1 and 2 can be clearly associated with helical structure elements, while cluster 7, which is clearly a distinct cluster in the data, does not correspond alone to a typical structural element.



(a) Helical conformations

(b) Irregular strands

Figure 12: *Strands of typical residues of clusters* 1*,* 2 *and* 7*. (Coloring scheme here is independent of Figures 10 and 11.) 12a: a sequence solely of cluster 1 residues (red); a sequence of mostly cluster 1 residues, every sixth residue being a cluster 2 residue (yellow); a sequence of mostly cluster 1 residues, every 11th residue being a cluster 2 residue (blue). 12b: Conformations of cluster 1 residue sequences interspersed with cluster 7 residues.*

6 Discussion

We have provided a novel framework for torus PCA to perform PCA-like dimension reduction for angular data. Previous attempts have not been satisfactory, because, on the one hand, the geometry featuring dense geodesics lead to severe restrictions for geodesic approaches while, on the other hand, Euclidean approximations disregard periodicity. We have used an adaptive deformation to a benign geometry, whilst at the same time preserving periodicity. For T-PCA to be fully effective it needs pre- and post-clustering. In application to dihedral angles of RNA structures we validated our method using a small classical benchmark data set. On a large classical data set, we go well beyond results achieved by analysis of 2-dimensional pseudo-torsion angles or recent 7-d clustering methods. Also we provide moderately sized clusters, half of them of size between 59 and 139 points. In fact, we have identified several clusters of low density which have not been located before. Some clusters have been examined in relation to helical conformations. Our method is widely applicable and can be used for geometrical analysis of biomolecular strands such as proteins, DNA and others.

Acknowledgements

We are grateful to Thomas Hamelryck, John Kent and J.S. Marron for helpful discussions. We thank Jes Frellsen for his comments on a draft of this article and for pointing to the RNAview program for calculation of RNA bonds. Further, we wish to thank Karen Sargsyan for providing details on GeoPCA and on the RNA data used by them in their paper.

A Supplementary material

A.1 Additional Illustrations

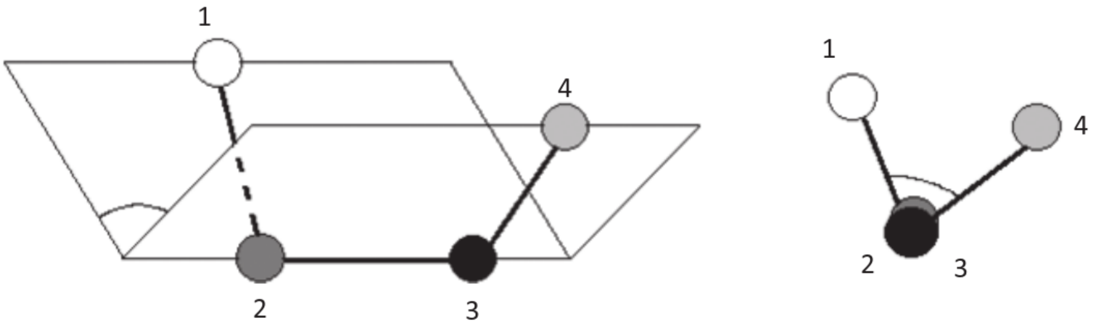


Figure 13: *Illustration of a dihedral angle defined by four atoms or three bonds, it is the opening angle between to pages of a book. (Reproduced from Mardia (2013).)*

A.2 Polar Coordinates for Higher Dimensions

Assuming the embedding $\mathbb{S}^D = \{x \in \mathbb{R}^{D+1} : \|x\| = 1\}$, the coordinates of the embedding space x_k are related to angular coordinates ϕ_k as follows

$$\begin{aligned} x_1 &= \cos\phi_1 \\ \forall 2 \le k \le D \,:\, x_k &= \left(\prod_{j=1}^{k-1} \sin\alpha_j\right)\cos\phi_k \\ x_{D+1} &= \left(\prod_{j=1}^{D} \sin\phi_j\right). \end{aligned}$$

A.3 Flow Chart of the T-PCA Algorithm

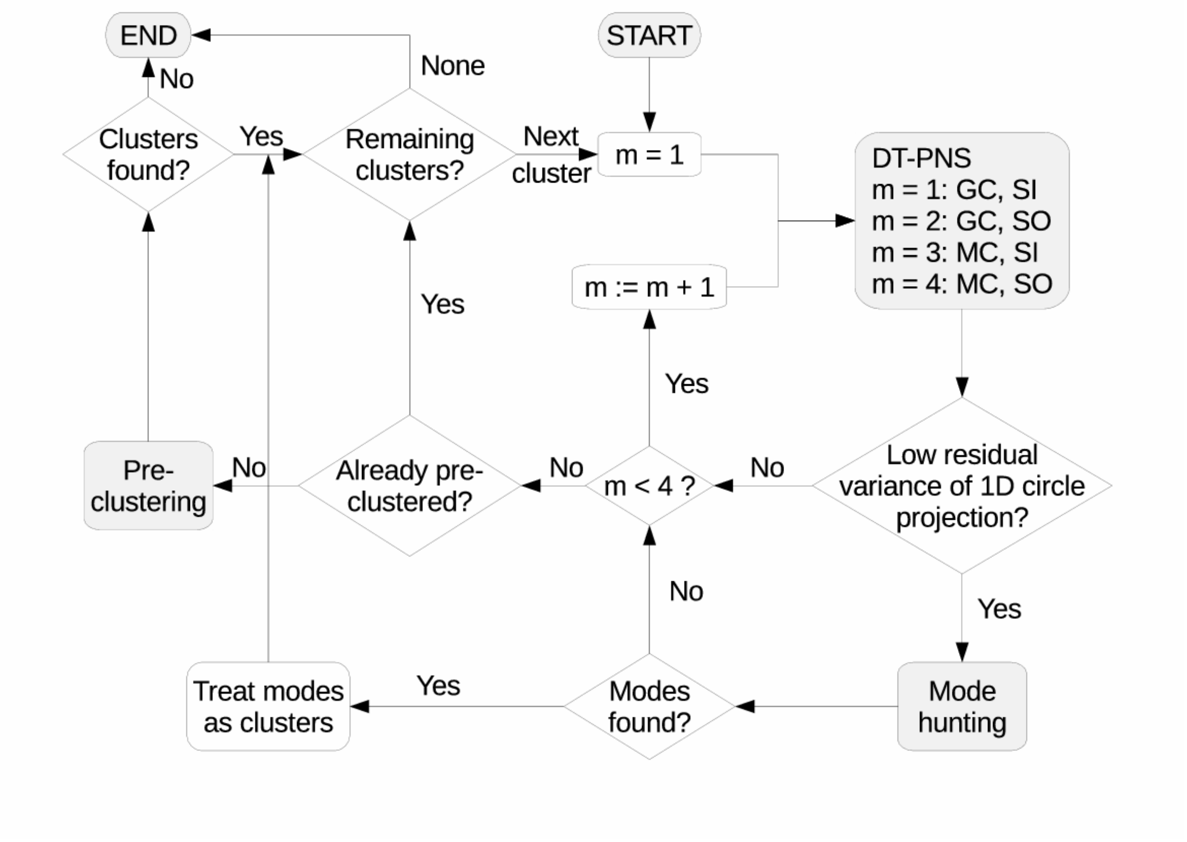


Figure 14: *The torus-PCA algorithm including DT-PNS, pre-clustering and mode hunting.*

A.4 Abbreviations and Technical Terms

We give a brief overview over abbreviations and technical terms used throughout this paper:
T-PCA: Torus Principal Component Analysis. The dimension reduction method via geometrical deformation presented in the present article.
DT-PNS: Deformed Torus Principal Nested Spheres, an alteration of PNS by Jung et al. (2012). A backwards method for dimension reduction on spheres. At each step, one finds the small subsphere with codimension 1 which best fits the data.
MC: Mean Centered, **GC**: Gap Centered, **SI**: Spread Inside, **SO**: Spread Outside,
H: Halved angles, **U**: Unscaled angles, see Subsection 2.1.
Codimension: The codimension of k-dimensional subspace of a d-dimensional space is defined as $d - k$.
Residual: Statistically unexplained data variation, see Subsection 2.3.
Residue: RNA molecule segment corresponding to a single nucleic base, see Subsection 5.
Nested Mean: The ultimate point μ of the sequence of small subspheres $\mathbb{S}^D \supset S^{D-1} \supset \cdots \supset S^2 \supset S^1 \supset \{\mu\}$ found by PNS, see subsection 2.2.

A.5 Topological Details

Due to periodicity on the torus, $\psi_k = 0$ is identified with $\psi_k = 2\pi$ for all $k = 1, \dots, D$. In contrast, for all angles $\phi_k = 0$ denotes spherical locations different from $\phi_k = \pi$. In case of halving (H), except for the innermost angle ($1 \leq k < D$), for an invariant representation respecting torus distance, however, it is necessary to identify these locations accordingly, which results in a self-gluing of $\mathbb{S}^D$ along specific codimension two great subspheres, see Figures 15 and 16.

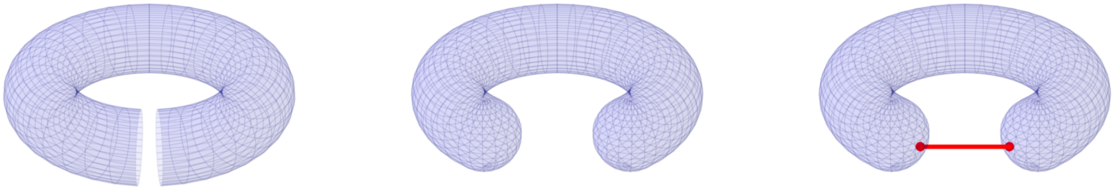


Figure 15: *Gluing in the case of halved angles for T^2: the introduced curvature due to embedding T^2 in $\mathbb{R}^3$, for illustration's sake, should not be mistaken for the torus deformation of Section 2.1*

.

In the rare case of angles concentrated to an interval of length π, using unscaled angles (U) this interval is mapped to $[0, \pi]$ without any distortion.

Here is an illustration for the gluing effects in the case of halving.

Example A.1. *For $D = 3$, on $\mathbb{S}^3$ we have the squared line element*

$$ds^2 = d\phi_1^2 + \sin^2\phi_1 \left(d\phi_2^2 + \sin^2\phi_2 d\phi_3^2\right) .$$

where the angle ranges are $\phi_1, \phi_2 \in (0, \pi)$, $\phi_3 \in [0, 2\pi)$. When using halved angles, for ϕ_1 and ϕ_2 we have the identification $0 \equiv \pi$. For ϕ_1 this is an identification of two points. For ϕ_2 this is an identification of the points $(\phi_1, 0, 0)$ and $(\phi_1, \pi, 0)$ for all $\phi_1 \in (0, \pi)$, which means that pairs of points are glued together along half circles. The example $D = 2$ is illustrated in Figures 15 and 16.

A.6 Improved Hypothesis Test

Let S^d be a fitted small subsphere, $2 \leq d \leq D$. For ease of notation, assume that $S^d = \mathbb{S}^d$ and that $p \in \mathbb{S}^d$ is the center of the fitted small S^{d-1}. For simplicity, we restrict attention to probability distributions $q \mapsto g(q;p)$ which depend only on the angular distance $r = d(p,q)$ for any data point $q \in \mathbb{S}^d$. Let γ be a curve along any great circle connecting p with its

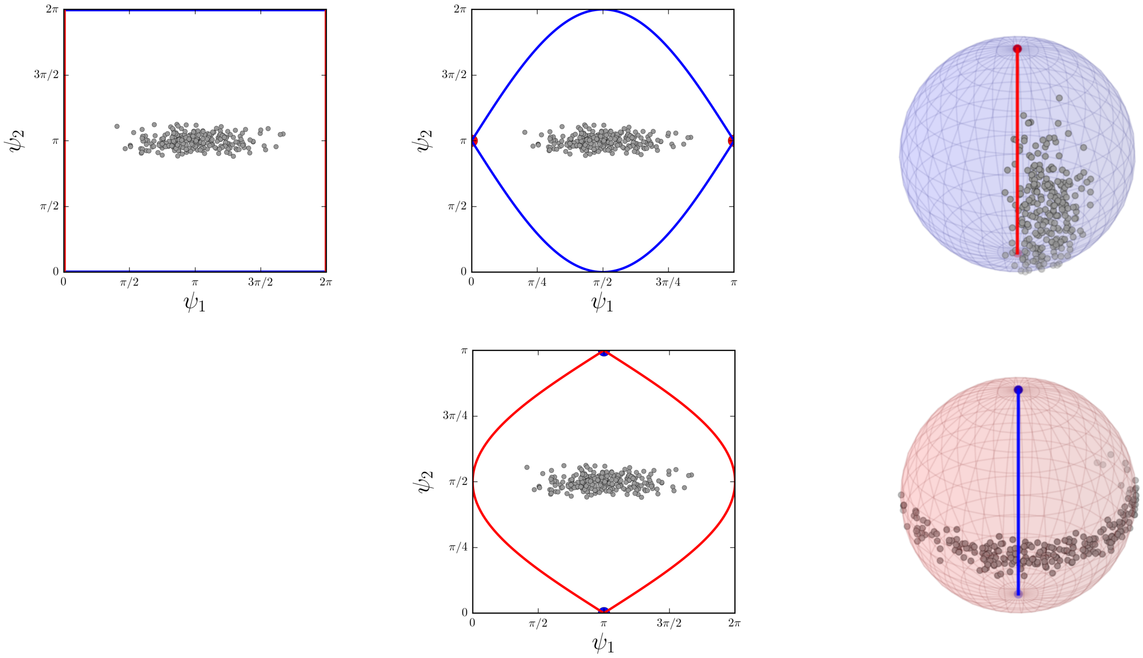


Figure 16: *Two possibilities for gluing in the case of halved angles for T^2. Top row: data on a torus in flat representation and the effect of halving ψ_1; due to the torus' periodicity, top and bottom blue arcs are identified; due to collapsing of the identified red lines to points (the singularity set), north and south pole of the sphere are identified (shown by the red line). Bottom row: the effect of halving ψ_2 with the roles of red and blue reversed.*

antipodal and let γ be parametrized by $r \in [0, \pi]$ such that $\forall r : d(p, \gamma(r)) = r$. Then, due to its symmetry, g is fully characterized by the function

$$h(r; p) := \mathrm{vol}_{\mathbb{S}^{d-1}} \cdot g(\gamma(r); p)$$

on $[0, \pi]$. Using the spherical volume element $d_{\mathbb{S}^d}\Omega(q)$ at $q = \gamma(r)$ we note

$$1 = \int g(q; p) d_{\mathbb{S}^d}\Omega(q) = \int \frac{h(r; p)}{\mathrm{vol}_{\mathbb{S}^{d-1}}} d_{\mathbb{S}^d}\Omega(q) = \int_0^\pi h(r; p) \sin^{d-1}(r) dr$$

which means that h is a marginal distribution with respect to the measure

$$d\mu_h(r) = \sin^{d-1}(r) dr \, .$$

The marginal distribution with respect to the Lebesgue measure on $[0, \pi]$ is defined as

$$f(r; p) := \sin^{d-1}(r) h(r; p) \, , \qquad \int_0^\pi f(r; p) dr = 1$$

For the following, consider the so-called folded normal distribution

$$\begin{aligned}\mathcal{F}(r;\rho,\sigma) :=&\frac{1}{\sqrt{2\pi}\sigma}\left(\exp\left(-\frac{(r-\rho\sigma)^2}{2\sigma^2}\right)+\exp\left(-\frac{(r+\rho\sigma)^2}{2\sigma^2}\right)\right)\\ =&\frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{r^2}{2\sigma^2}-\frac{\rho^2}{2}\right)\cosh\left(\frac{r\rho}{\sigma}\right),\ r\geq 0\,.\end{aligned}$$

For $\rho \to \infty$ this tends to a usual normal distribution centered at $\rho\sigma$, while it becomes a halved normal distribution for $\rho \to 0$. Visualizing as a surface of revolution over $\mathbb{R}^2$, in polar coordinates $(r,\vartheta) \mapsto \mathcal{F}(r;\rho,\sigma)\,\frac{1}{2\pi}$, the former case yields a ring while the latter case yields a symmetric Gaussian distribution. Due to its smoothness it is a good candidate for a test distribution to distinguish concentrated clusters from ring shapes.

With the above marginals we therefore define

$$h(r;p,\rho,\sigma)=\frac{\sqrt{2\pi}\sigma}{\mathcal{C}(\rho,\sigma)}\mathcal{F}(r;\rho,\sigma),\quad f(r;p,\rho,\sigma)=\frac{\sqrt{2\pi}\sigma}{\mathcal{C}(\rho,\sigma)}\sin^{d-1}(r)\mathcal{F}(r;\rho,\sigma)\,,$$

whose normalization $\mathcal{C}(\rho,\sigma)$ can be easily determined numerically. We can determine the MLEs for ρ and σ using standard numerical optimization. Although a numerical integral has to be calculated for normalization in each optimization step, the optimizations usually converge very quickly. If $\rho_{\text{MLE}} < 1$, the distribution has its maximum at $r = 0$ and the small subsphere hypothesis can be readily rejected in favor of a great subsphere fit.

If $\rho_{\text{MLE}} > 1$ we apply a likelihood ratio test. For a fixed p, given as the center of the best fit small subsphere, q_i the data and $r_i = d(p,q_i)$ the spherical distances,

$$\begin{aligned}\ell(\rho,\sigma|\{r_i\}_{i=1}^n)=&-n\ln\mathcal{C}(\rho,\sigma)+(d-1)\sum_{i=1}^n\ln\sin(r_i)\\ &-\frac{n\rho^2}{2}+n\ln(2)+\sum_{i=1}^n\left(-\frac{r_i^2}{2\sigma^2}+\ln\cosh\left(\frac{r_i\rho}{\sigma}\right)\right)\end{aligned}$$

is the log-likelihood for $f(r;p,\rho,\sigma)$ given a data set $\{r_i\}$. As null hypothesis we assume $\rho = 1$, which means that the data form a dense cluster. The alternative hypothesis $\rho > 1$ means that the data are better approximated by a small circle. Let

$$\begin{aligned}\lambda(\{r_i\}_{i=1}^n)=&2\sup\{\ell(\rho,\sigma|\{r_i\}_{i=1}^n):\rho\in(1,\infty),\,\sigma\in\mathbb{R}^+\}\\ &-2\sup\{\ell(\rho,\sigma|\{r_i\}_{i=1}^n):\rho=1,\,\sigma\in\mathbb{R}^+\}\end{aligned}$$

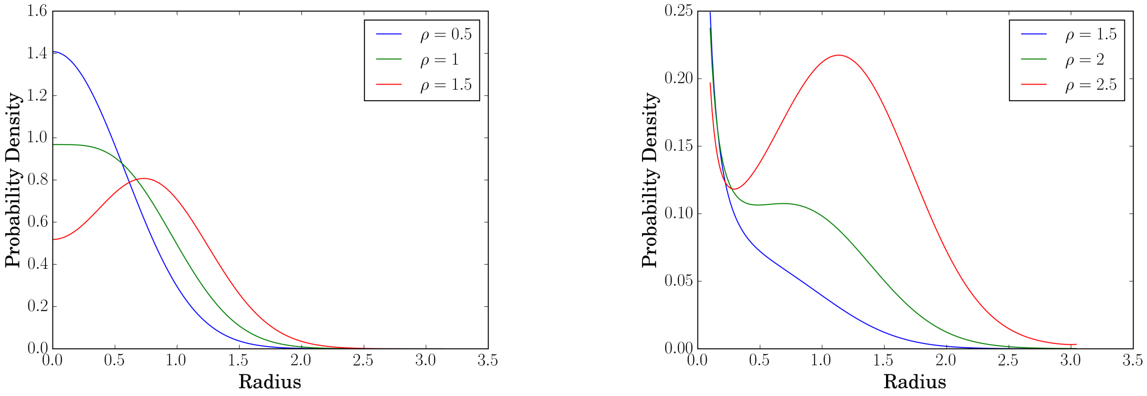
be the usual test statistic for a likelihood ratio test. Then, due to Wilks' theorem, see Wilks (1938) and (van der Vaart, 1998, Chapter 16), λ is asymptotically χ_1^2 distributed for $n \to \infty$. The null hypothesis is rejected with 95% confidence if $\lambda > \chi^2_{1,95}$ in which case we keep the fitted small subsphere. If the null hypothesis is not rejected, we perform a great subsphere fit.

The functions f and h are briefly discussed in Jung et al. (2011) but they use the folded normal distribution defined on $\mathbb{R}^+$ for f, i. e.

$$f^*(r; p, \rho, \sigma) = \mathcal{F}(r; \rho, \sigma), \quad h^*(r; p, \rho, \sigma) = \frac{1}{\sin^{d-1}(r)} \mathcal{F}(r; \rho, \sigma)\,,$$

leading to a singularity of the probability density h^* and thus of g at p. The small circle is accepted, if the probability distribution exhibits a ring shaped local maximum, which is the case for $\rho > 2$. A singularity at p, however, is an undesirable feature of the distribution h^* (resulting in a frequent rejection of a projected Gaussian in the tangent space as null hypothesis as seen in Table 1 in the original article), which we have avoided as above. The functions $h(r)$ for our distribution and $h^*(r)$ for the one used by Jung et al. (2011) are illustrated in Figure 17.

In Jung et al. (2012) among others, null hypothesis and alternative are both modeled via von Mises-Fisher distributions and a student t-like test statistic of distances to the estimated center point is used. However, the von Mises-Fischer distribution has a heavier tail and thus a higher standard deviation than the truncated Gaussians used here. In consequence, the test statistic is considerably larger for our more concentrated clusters and thus the null hypothesis is frequently rejected, especially for large sample sizes (see Table 1 in the original article).



(a) *The probability distribution h for any d*

(b) *The probability distribution h^* for $d = 2$*

Figure 17: *The probability densities for $\sigma = 0.5$ along a geodesic in $\mathbb{S}^d$ for the distribution h used here and h^* used by Jung et al. (2011). Displaying a value for ρ below the respective threshold, at the threshold and above the threshold – which is* 1 *in our approach and* 2 *in case of Jung et al. (2011).*

A.7 Single Linkage Branch Cutting Algorithm

Let P with $|P| =: n$ be the set of data points to cluster, m a lower bound for the cluster size and $d_{\max}$ a maximal outlier distance. Define the minimal cluster size $S_P = \sqrt{n + m^2}$. Then, we iteratively first store outliers to a list R and cluster the rest of the data to a list C as follows.

1. Start with empty lists R and C.

2. Compute the cluster tree of P.

3. Perform a branch cut at distance $d_{\max}$, i. e. removing all nodes with values above $d_{\max}$. For all nodes with less than m points add their points to the list R and remove them from P.

4. Compute S_P and the cluster tree for P.

5. Perform the branch cut,

   (a) Starting from the root, follow the branch containing more points at each fork.

   (b) If the smaller branch at a fork has more than S_P points, store it to a list L.

   (c) At the last fork, where the smaller branch has size at least S_P store also the larger branch to L.

   (d) For the largest cluster in L remove its points from P and store the cluster to a list C. If L is empty, add all remaining points to C as one cluster.

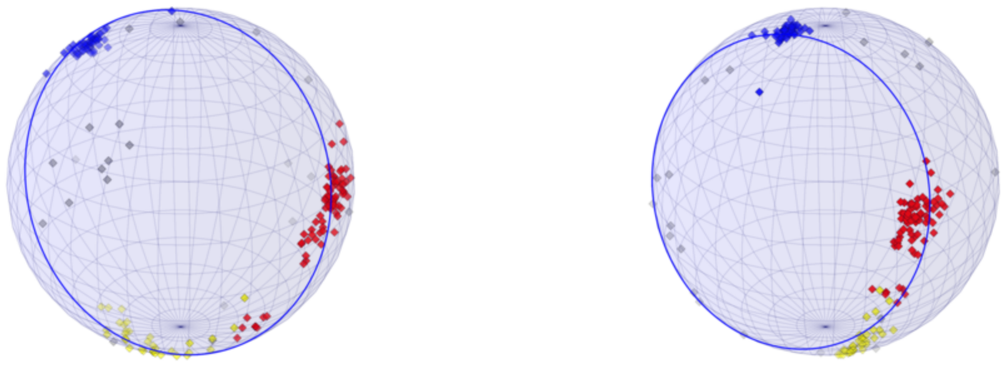
6. If points remain, go to 2.

7. Return the list of clusters C and the outliers R.

For our analysis we chose $m = 15$ and $d_{\max} = 50°$.

A.8 Outliers in the Small Data Set

The small data set has been devised by Sargsyan et al. (2012) to feature three clusters. These are found by T-PCA. Additionally including preclustering in our method, outliers are found whose small scale geometry is very different from the majority of the residues of any of the three clusters. Clusters and outliers are depicted in Figure 18.



(a) *2D approximation, SI*

(b) *2D approximation, SO*

Figure 18: *Two dimensional T-PCA approximation of the pre-clusters found in the small RNA data set with SI (18b) and SO (18b) ordering. Red, blue and yellow represent the same clusters as in Figure 4, gray points have been classified as outliers.*

A.9 List of All Found Clusters

Table 4: *Nested means of the 22 clusters found by T-PCA.*

			nested mean						
cluster #	# points	method	α	β	γ	δ	ϵ	ζ	χ
1	4921	SO MC	297.95	173.73	52.47	81.02	202.96	291.83	197.96
2	477	SO MC	162.62	203.96	171.62	83.72	232.57	280.81	181.95
3	232	SO GC	306.29	128.46	53.09	84.23	205.07	293.40	225.13
4	211	SO MC	157.78	182.50	51.04	84.70	212.65	292.49	195.22
5	145	SO GC	293.52	173.23	53.80	82.88	208.23	71.47	207.56
6	139	SI GC	56.76	160.75	51.56	83.65	216.26	289.84	189.47
7	137	SO GC	304.39	162.78	59.02	146.49	228.62	157.06	243.20
8	134	SI GC	294.62	173.66	53.57	83.41	227.20	204.37	204.05
9	125	SO GC	304.31	164.04	45.01	145.04	247.30	76.24	227.76
10	122	SO GC	210.51	115.51	160.16	85.67	225.35	280.39	184.57
11	107	SO GC	99.00	177.58	57.58	144.18	268.68	319.98	237.95
12	85	SO GC	303.40	165.94	59.13	142.65	260.53	292.92	237.45
13	84	SI MC	146.77	221.93	170.04	80.20	245.21	223.11	189.16
14	78	SI MC	79.50	188.62	182.21	84.75	214.53	293.24	201.44
15	60	SO MC	152.44	136.18	51.17	146.56	277.10	116.79	226.51
16	60	SI GC	280.79	199.65	174.56	90.66	217.55	271.56	210.26
17	59	SO GC	67.15	180.96	52.17	149.34	246.12	62.31	249.72
18	55	SI MC	50.61	184.82	286.25	97.58	205.34	306.97	202.51
19	52	SO GC	291.86	184.92	53.69	95.52	30.43	166.71	230.74
20	46	SO MC	254.42	199.20	80.36	97.36	302.41	233.82	219.45
21	33	SO GC	278.08	253.81	280.18	86.93	203.81	294.46	193.98
22	28	SI MC	290.03	199.31	55.09	87.25	71.04	283.36	226.96

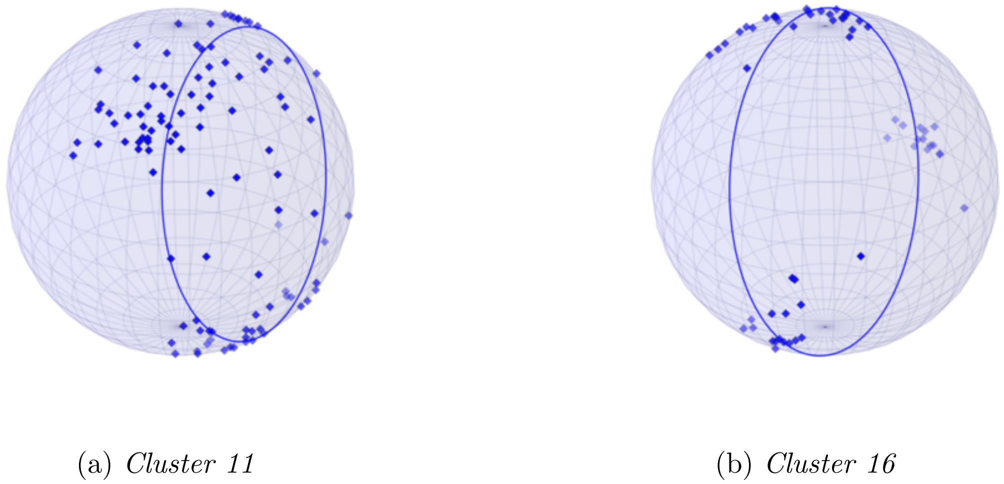
Table 5: *Residual variances of the projections of the 22 clusters found by T-PCA for all correspondingly dimensional subspheres.*

cluster #	0 D	1 D	2 D	3 D	4 D	5 D	6 D
1	653.67	394.26	277.06	176.02	93.88	38.47	24.38
2	2002.29	717.43	369.59	225.49	114.48	55.06	24.84
3	7548.59	2794.06	1647.43	1089.96	770.95	283.91	39.87
4	2540.55	649.73	334.17	213.59	118.32	60.94	33.17
5	3563.49	1096.97	568.65	342.73	221.13	134.87	66.42
6	4361.17	1436.75	846.12	598.74	302.46	112.93	16.78
7	3370.20	1279.69	735.58	522.58	267.93	151.08	69.19
8	1340.83	525.74	212.90	110.07	53.82	18.01	5.76
9	3513.08	1474.97	945.99	522.21	303.78	138.61	38.67
10	4377.74	1273.44	762.57	459.16	261.26	97.23	53.49
11	7065.66	1510.25	942.27	641.22	377.90	126.61	53.06
12	1977.41	919.47	580.16	367.15	232.95	113.00	46.77
13	7622.07	2997.64	1683.83	1098.01	613.00	236.14	106.52
14	4570.60	2194.72	1365.03	664.37	398.73	176.38	18.33
15	4379.46	1273.97	761.20	505.75	204.07	99.97	30.46
16	24749.45	2334.71	1455.38	809.90	492.65	242.71	56.44
17	7771.96	2046.17	1304.37	807.84	467.27	202.97	66.31
18	2357.09	1325.03	766.96	435.02	283.81	161.77	49.54
19	9360.10	1660.47	1056.75	668.63	368.23	173.33	66.23
20	11420.75	3021.12	1762.57	952.27	420.14	218.83	97.05
21	4395.15	1127.39	479.15	232.76	146.39	108.96	44.76
22	2926.13	1538.61	986.94	613.66	288.46	157.12	70.01

Table 6: *The* 15 *pre-clusters and their decomposition into the* 22 *clusters found by mode hunting.*

pre-cluster #	# points	cluster #s
1	5055	1, 8
2	492	5, 7, 9, 12
3	477	2
4	232	3
5	226	11, 15, 17
6	211	4
7	139	6
8	133	14, 18
9	122	10
10	84	13
11	60	16
12	52	19
13	46	20
14	32	21
15	28	22

Visual inspection of the one- or two-dimensional projections shows distinct subgroups for some clusters, see Figure 19. These subgroups have not been found by T-PCA due to the 95% confidence bound for mode hunting; in particular in Figure 19b the number of points is too small to attain significance.



(a) *Cluster 11*

(b) *Cluster 16*

Figure 19: *Two dimensional T-PCA approximations of two clusters which appear to be composed of several distinct clusters.*

References

Altis, A., Otten, M., Nguyen, P. H., Rainer, H., and Stock, G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *The Journal of Chemical Physics*, 128(24):245102.

Arsigny, V., Commowick, O., Pennec, X., and Ayache, N. (2006). A log-euclidean framework for statistics on diffeomorphisms. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 924–931. Springer.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, 28:235–242.

Boisvert, J., Pennec, X., Labelle, H., Cheriet, F., and Ayache, N. (2006). Principal spine shape deformation modes using riemannian geometry and articulated models. In *Articulated Motion and Deformable Objects*, pages 346–355. Springer.

Brewer, J. W. (2013). Regulatory crosstalk within the mammalian unfolded protein response. *Cellular and Molecular Life Sciences*, 71(6):1067–1079.

Chakrabarti, A., Chen, A. W., and Varner, J. D. (2011). A review of the mammalian unfolded protein response. *Biotechnology and Bioengineering*, 108(12):2777–2793.

Chapman, R., Sidrauski, C., and Walter, P. (1998). Intracellular Signaling from the Endoplasmic Reticulum to the Nucleus. *Annual Review of Cell and Developmental Biology*, 14(1):459–485.

Duarte, C. M. and Pyle, A. M. (1998). Stepping through an RNA structure: a novel approach to conformational analysis. *Journal of Molecular Biology*, 284(5):1465–1478.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *The Annals of Statistics*, 36(4):1758–1785.

Egli, M. and Saenger, W. (1984). *Principles of Nucleic Acid Structure*. Advanced Texts in Chemistry. Springer.

Estarellas, C., Otyepka, M., Koa, J., Ban, P., Krepl, M., and poner, J. (2015). Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(5):1072–1090.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. C. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *i3eTransMedIm*, 23(8):995–1005.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K. V., Ferkinghoff-Borg, J., and Hamelryck, T. (2009). A Probabilistic Model of RNA Conformational Space. *PLoS Comput Biol*, 5(6):e1000406.

Hermann, T. and Westhof, E. (1999). Simulations of the dynamics at an RNA-protein interface. *Nature Structural Biology*, 6(6):540–544.

Hotz, T. and Huckemann, S. (2014). Intrinsic means on the circle: uniqueness, locus and asymptotics. *Annals of the Institute of Statistical Mathematics*, 67(1):177–193.

Huckemann, S., Hotz, T., and Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 1(20):1–58.

Huckemann, S. and Ziezold, H. (2006). Principal component analysis for riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 2(38):299–319.

Huckemann, S. F. and Eltzner, B. (2015). Polysphere pca with applications. *Proceedings of the Leeds Annual Statistical Research (LASR) Workshop 2015*.

Jung, S., Dryden, I. L., and Marron, J. S. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

Jung, S., Foskey, M., and Marron, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. Appl. Stat.*, 5(1):578–603.

Jung, S., Liu, X., Marron, J., and Pizer, S. M. (2010). Generalized PCA via the backward stepwise approach in image analysis. In Angeles, J. et al., editors, *Brain, Body and Machine: Proceedings of an International Symposium on the 25th Anniversary of McGill University Centre for Intelligent Machines, Advances in Intelligent and Soft Computing*, volume 83 of *Body and Machine*, pages 111–123. Springer.

Kent, J. T. and Mardia, K. V. (2009). Principal component analysis for the wrapped normal torus model. *Proceedings of the Leeds Annual Statistical Research (LASR) Workshop 2009.*

Kent, J. T. and Mardia, K. V. (2015). The winding number for circular data. *Proceedings of the Leeds Annual Statistical Research (LASR) Workshop 2015.*

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720.

Magee, J. and Warwicker, J. (2005). Simulation of non-specific protein-mRNA interactions. *Nucleic Acids Research*, 33(21):6694–6699.

Mardia, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):487–514.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Probability and mathematical statistics. Academic Press.

Murray, L. J. W., Arendall, W. B. I., Richardson, D. C., and Richardson, J. S. (2003). RNA backbone is rotameric. *Proc. Natl Acad. Sci. USA*, 100(24):13904–13909.

Obulkasim, A., Meijer, G. A., and Wiel, M. A. v. d. (2015). Semi-supervised adaptive-height snipping of the hierarchical clustering tree. *BMC Bioinformatics*, 16(1):15.

Richardson, J. S., Schneider, B., Murray, L. W., Kaprall1, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., and Berman, H. M. (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, 14:465–481.

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

Schneider, B., Morvek, Z., and Berman, H. M. (2004). RNA conformational classes. *Nucleic Acids Research*, 32(5):1666–1677.

Sommer, S. (2013). Horizontal dimensionality reduction and iterated frame bundle and development. In *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 76–83.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wadley, L. M., Keating, K. S., Duarte, C. M., and Pyle, A. M. (2007). Evaluating and learning from RNA pseudotorsional space: Quantitative validation of a reduced representation for RNAstructure. *Journal of Molecular Biology*, 372(4):942–957.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13):3450–3460.

Zhao, Y., Kormos, B. L., Beveridge, D. L., and Baranger, A. M. (2006). Molecular dynamics simulation studies of a protein-RNA complex with a selectively modified binding interface. *Biopolymers*, 81(4):256–269.