**Article:**
Eltzner, B, Huckemann, S and Mardia, KV orcid.org/0000-0003-0090-6235 (2018) Torus principal component analysis with applications to RNA structure. Annals of Applied Statistics, 12 (2). pp. 1332-1359. ISSN 1932-6157

https://doi.org/10.1214/17-AOAS1115

*Submitted to the Annals of Applied Statistics*

# TORUS PRINCIPAL COMPONENT ANALYSIS WITH APPLICATIONS TO RNA STRUCTURE

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

[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,*

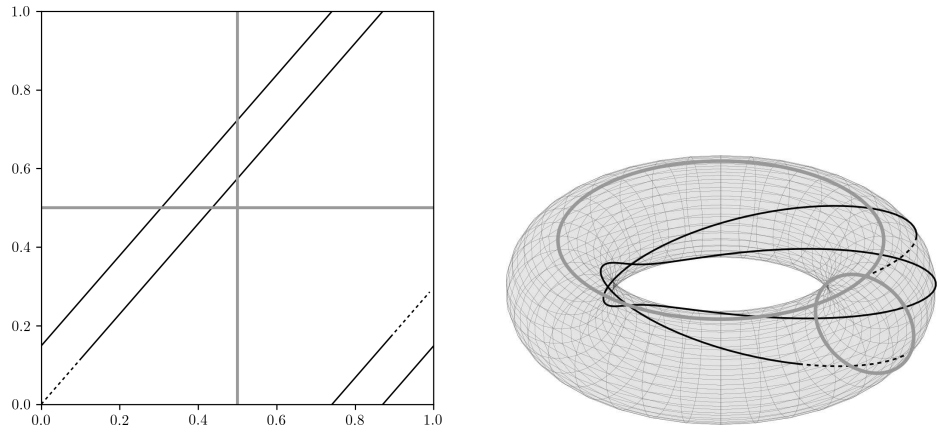
There are several cutting edge applications needing PCA methods for data on tori and we propose a novel torus-PCA method that adaptively favors low-dimensional representations while preventing overfitting by a new test, both of which can be generally applied and address shortcomings in two previously proposed PCA methods: Unlike tangent space PCA, our torus-PCA features structure fidelity by honoring the cyclic topology of the data space, and, unlike geodesic PCA, produces non-winding, non-dense descriptors. These features are achieved by deforming tori into spheres with self-gluing and then using a variant of the recently developed principal nested spheres analysis. This PCA analysis involves a step of subsphere fitting and we provide a new test to avoid overfitting. We validate our torus-PCA by application to an RNA benchmark data set. Further, using a larger RNA data set, torus PCA recovers previously found structure, now globally at the one-dimensional representation, which is not accessible via tangent space PCA.

**1. Introduction.** Dimension reduction on non-Euclidean manifolds with PCA-like methods has been a challenging task for which two usually successful categories of methods have been developed in the last decades: extrinsic (tangent space) approaches, e.g. Gower (1975); Fletcher et al. (2004); Boisvert et al. (2006); Arsigny et al. (2006), and intrinsic (geodesic) ones, e.g. Huckemann and Ziezold (2006). A critical review of PCA methods has been given in Huckemann, Hotz and Munk (2010); Sommer (2013) is another recently developed intrinsic PCA method. 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,

*The authors gratefully acknowledge DFG HU 1575/4, DFG CRC 755 and the Niedersachsen Vorab of the Volkswagen Foundation.

*Keywords and phrases:* Statistics on manifolds, tori deformation, directional statistics, dimension reduction, dihedral angles, fitting small spheres, principal nested spheres analysis

tangent space PCA (TS-PCA) fails to take into account the periodicity of the torus and, even worse, geodesic PCA is completely inapplicable because almost all geodesics densely wind around, as in Figure 1.




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

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

Fig 1: *Flat (1a) and curved (1b) torus representation. Except for horizontal and vertical geodesics (grey) in (1a), and diagonal ones, all other geodesics wind around ((1a) and (1b)). All geodesics (black) with an irrational slope in (1a) are dense.*

In this paper, we propose the novel tool of torus-PCA (T-PCA), which not only removes these defects, but also more flexibly adapts to low dimension, in a statistically controlled way to guard against overfitting. This is achieved by transforming the "geometrically benign" structure of the torus into a *statistically benign geometry*, namely one that does not allow for dense geodesics. We note that these dense geodesics are in the closure of the non-dense geodesics, which in turn can be viewed as 1D subtori, and so, an attempt for *principal nested tori* still suffers from the statistically non benign geometry. Specifically, we deform tori into spheres by choosing a nearby statistically benign geometry with clever mapping and self-gluing, and then use a modification of the recently developed principal nested spheres analysis (PNS) of Jung, Dryden and Marron (2012). In particular, this PNS analysis involves a step of small sphere fitting and we provide a new 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 (where curvature is unbounded). Notably, although locally respecting the flat geometry of the torus, ignoring periodicity, TS-PCA introduces in fact

not only geometric but also topological singularities (the tangent space is not homeomorphic to the torus).

At this point we recall that within a sphere of radius $r > 0$, every subsphere with the same radius $r$ is a *great subsphere* and one of smaller radius is a *proper small subsphere*. In this paper we speak of *small subspheres* to include great and proper small subspheres.

Some torus-specific PCA approaches have been developed apart from TS-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. The PCA used by Altis et al. (2008) is a particular case of Kent and Mardia (2009). Allowing only geodesics that wind around at most once, as proposed by Kent and Mardia (2015), further reduces the degrees of freedom. As discussed in Huckemann and Eltzner (2015) for classical PCA in $\mathbb{R}^n$ the space of $k$-dimensional affine subspaces ($0 \leq k \leq n$) has dimension $(n-k)(k+1)$; in contrast for PNS in the $n$-dimensional sphere, the space of $k$-dimensional small subspheres has dimension $(n-k)(k+2)$ ($1 \leq k \leq n-1$). For this reason (building on PNS), T-PCA more flexibly favors lower dimensional representations than TS-PCA, while this flexibility is better controlled against overfitting than in classical PNS.

Sargsyan, Wright and Lim (2012) may have been the first to treat toroidal data describing RNA structures in a spherical geometry. In their construction, they halved the corresponding seven torus angles defined below and treated them 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 which we call the *benchmark data*. However, Sargsyan, Wright and Lim (2012) neither discussed nor exploited the drastic change of geometry, let alone amended by self-gluing, 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 were noted in Huckemann, Hotz and Munk (2010).

**RNA structure analysis and challenges: a bigger picture.** The last decades have witnessed finding an unexpected variety of RNA shape and function, and this variety is ever increasing. Base sequences, also called *primary structures* and consisting of polymers of four different *nucleotides,* are nowadays easily accessible by high throughput sequencing and it is one ultimate goal to link these sequences to biological function. Biological function, however, is highly dependent on the 3D structure (or *fold*) which manifests at different levels (e.g. Chapman, Sidrauski and Walter (1998); Chakrabarti,

Chen and Varner (2011); Seetin and Mathews (2012); Brewer (2013)). At the bottom level is the *single residue* geometry usually described by *dihedral angles* between neighboring planes, each spanned by three adjacent atoms, similar to pages of an open book (Figure 2). The structure of each nucleotide can be described by 6 angles for the polymeric backbone and one angle for the nucleotide's *base*, giving a total of 7 angles (Figure 3 and Table 1). *Secondary structure* is given by self-interaction within the RNA molecule via base pairing and other interactions, forming specific patterns such as A-helices, hairpin loops, and others. At the top level, *tertiary* and higher order structure arises from interacting lower order structure patterns via further base and backbone bindings.

In contrast to primary structure, the 3D structure is not easily accessible but needs to be reconstructed by elaborate technology such as X-ray crystallography. However, experimental structures are prone to misinterpretation and various errors. For example, backbone inconsistencies, where different reconstructed atoms occupy the same spatial location, frequently occur during reconstruction Richardson et al. (2008); Jain, Richardson and Richardson (2015). To avoid or correct such errors, the space of possible 3D structures is often restrained or constrained to previously observed structures. This is typically done at the nucleotide or paired nucleotide level Yang et al. (2003); Schneider, Morvek and Berman (2004); Wadley et al. (2007); Čech et al. (2013). Specifically, use is made of so-called *rotamers* describing empirical modes of probability distributions of nucleotide or nucleotide pair conformations. As these distributions are relatively peaked, limiting the conformational space to such rotamers avoids the introductions of incorrect conformations by limiting the conformational space to previously observed 3D patterns.

Among the many challenges along this path, we discuss two specific ones: data reduction methods and alignment strategies.

To the end of backbone reconstruction, single residue conformation space is explored and *dimension reduction methods* are applied to identify errors in experimental structures, provided among others by the popular free software of Davis et al. (2007). For example, removing inconsistencies, Murray et al. (2003) have found that RNA backbone is *rotameric locally at hemi-nucleotide level*, i.e. among others, when reducing the 7D single residue space to a 3D backbone angular space, involving angles on only one side of the base (cf. Table 1 and Figure 3), conformer groups of each of the two sugar puckers (explained in Section 3), follow essentially one angle only. In our second application below, we revisit the data corresponding to one sugar pucker and generalize the result to finding a 1D structure common to all

conformer groups.

On the one hand, matching RNA strands requires elaborate registration and alignment strategies (e.g. Mardia (2013)), building on statistical (e.g. Dryden and Mardia (2016); Srivastava and Klassen (2016)) and Bayesian (e.g. Green and Mardia (2006)) shape technology including non-Euclidean averaging and elastic curve representations (e.g. Liu, Srivastava and Zhang (2011); Laborde et al. (2013)). On the other hand, averaging and exploring the 7D single residue space can be achieved via dynamically simulating similar structures (e.g. Duarte and Pyle (1998); Chen and García (2013); Estarellas et al. (2015)), and probabilistic approaches to this end require *dimension reduction methods* (e.g. Frellsen et al. (2009)). In this context, also for higher order structure prediction, it is necessary to explore not only the variation of single residue geometries typical for specific secondary structure elements but also single residue geometries for intermediate and transition regions between structure elements (e.g. Dunbrack and Karplus (1994); Jain, Richardson and Richardson (2015)).

**Applying torus-PCA to RNA structure analysis** we provide for a novel dimension reduction method at residue level and we apply it within the focus of current research to single residue geometries. However, it readily generalizes to simultaneous analysis of geometries of residue sequences (7n angles for n residues) but such an extension is left for future research. We measure effectively the statistical performance of our method by dimension reduction and faithfulness in terms of preserving previously known structure.

All of the angles used in our applications are defined in Table 1 and displayed in Figure 3. First we use the *benchmark data set* of Sargsyan, Wright and Lim (2012) which consists of neighborhoods of three known cluster centers in the $\eta$–$\theta$-plot (as in Figure 7a, the pseudo-torsion angles $\eta$, $\theta$ are depicted in Figure 3b, cf. also Table 1). We find that T-PCA retrieves the underlying clusters in an effective way. This benchmark data set is a subset of a large RNA data set carefully selected for high experimental X-ray precision (0.3 nanometers) by Duarte and Pyle (1998), updated by Wadley et al. (2007) and analyzed by them and others, for example, Murray et al. (2003); Richardson et al. (2008). Next we use another subset of this large RNA data set with C2'-endo sugar pucker (this and the other sugar pucker are explained fully in Section 3), subsequently called the *C2 data set*, where we compare our method to TS-PCA and show that T-PCA captures not only much more variance in the one-dimensional subspace, also the wrong topology in TS-PCA hides and tears apart subtle structural similarities.

In contrast, T-PCA provides *structure fidelity*, as global and local structural similarities are naturally preserved, most of it already visible in the

1D T-PCA representation, generalizing the above finding of Murray et al. (2003) that RNA backbone is locally rotameric at heminucleotide level, to:

*These RNA conformers are* rotameric *at full residue level, possibly in a non-linear sense, however.*

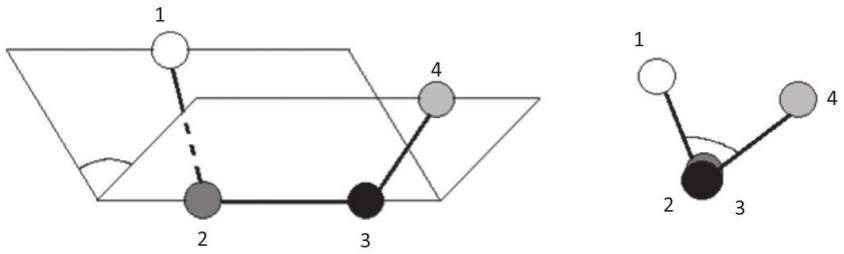
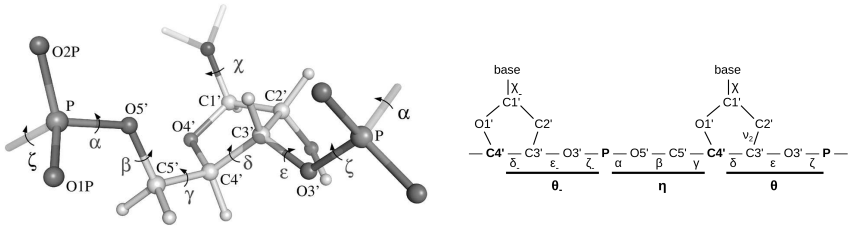



Fig 2: *Illustration of a dihedral (torsion) angle defined by four atoms or three bonds, it is the opening angle between two pages of a book. (Reproduced from Mardia (2013).)*




(a) *3D structure of an RNA residue.* (b) *2D scheme of an RNA residue.*

Fig 3: *Part of an RNA backbone (Phosphate groups with central atom denoted by P, followed by sugar rings that connect along the atoms labeled by C4' and C3', to which a nucleic base is 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). Underlying each pseudo torsion angle are three heminucleotide angles. The precise definitions with same canonical atom notation are given in Table 1. The subscript "−" denotes angles of the neighboring residue. Figure 3a is reproduced from Frellsen et al. (2009).*

In Section 2 we introduce torus PCA, which is the center piece of our methodology. In Section 3 we apply our method to the benchmark and C2 data sets, and review the results. The paper ends with a discussion

TABLE 1

*Atom bonds (2nd column) defining angles (1st column) with description (3rd column). The two sets of heminucleotide angles (each of which can be approximated by a pseudo torsion angle) define the backbone, which in conjunction with the base angle $\chi$ define a residue. Figure 3a shows the geometry of these atoms. (N denotes nitrogen.)*

| | | |
|---|---|---|
| $\alpha$ | $O3'-\ P\ -O5'-C5'$ | heminucleotide angles |
| $\beta$ | $P\ -O5'-C5'-C4'$ | |
| $\gamma$ | $O5'-C5'-C4'-C3'$ | |
| $\delta$ | $C5'-C4'-C3'-O3'$ | heminucleotide angles |
| $\epsilon$ | $C4'-C3'-O3'-\ P$ | |
| $\zeta$ | $C3'-O3'-\ P\ -O5'$ | |
| $\chi$ | $O4'-C1'-N1-C2$<br>$O4'-C1'-N9-C4$ | for pyrimidine (monocyclic) bases<br>for purine (bicyclic) bases |
| $\eta$ | $C4'-\ P\ -C4'-\ P$ | pseudo torsion angles |
| $\theta$ | $P\ -C4'-\ P\ -C4'$ | |
| $\nu_2$ | $C1'-C2'-C3'-C4'$ | sugar pucker angle |

and further illustrations in Supplement A. An implementation of our T-PCA method and the RNA data sets we use are included as supplementary material Supplement B, Supplement C and can be found under http://www.stochastik.math.uni-goettingen.de/SFB755_B8.

**Residues and residual variance.** To avoid confusion, we clarify that the biochemical term *residue* denotes a RNA molecule segment corresponding to a single nucleic base (Section 3) whereas the statistical term *residual variance* denotes unexplained variation (Section 2.3).

**2. Torus PCA.** Our dimension reduction procedure proceeds in two steps. First, the data space is deformed from a torus to a sphere with self-gluing, i.e. parts of the sphere are topologically identified with themselves, see Figures 4 and 5. Several degrees of freedom are present in the deformation map we propose and we discuss consequences of specific parameter choices. The second step is the dimension reduction for which we use a well established procedure for dimension reduction on spheres with some extensions to take into account the original torus geometry and the self-gluing of the sphere.

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 comparing squared Riemannian 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 $2\pi$, the squared line element of $T^D$

is given by the squared Euclidean line element

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

For $\mathbb{S}^D$, in polar coordinates $\phi_k \in [0, \pi]$ for $k = 1, \ldots, D-1$ and $\phi_D \in [0, 2\pi]/\sim$, whose relation to embedding Euclidean coordinates $x_k$ is given by

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

the spherical squared line element is given by

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

In fact, this squared line element is not defined for the full sphere but only for $\phi_k \in (0, \pi)$ $(k = 1, \ldots, D-1)$, i.e. the singularities of $\phi_k = 0, \pi$ are excluded. The singularities at $\phi_k = 0, \pi$ will account for singularities of $P$ which results in a self-gluing as explained below.

**Angular distortions in a spherical geometry.** Following colloquial usage, we use "distortion" synonymous with "deformation" in the following. Because in (1), $d\phi_1^2$ comes with the factor 1, no deformation at all occurs for $\phi_1$, 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 $\phi_1$ distorts the angle $\phi_2$ and finally the deformation factor $\prod_{j=1}^{D-1} \sin^2\phi_j$ of $d\phi_D^2$ reflects the distortions of $\phi_D$ by all other angles. For this reason, in the following, we will refer to $\phi_D$ as the *innermost angle* and to $\phi_1$ as the *outermost angle*.

We now make an important note for later use.

Remark 2.1. *Near the equatorial great circle given by $\phi_k = \frac{\pi}{2}$ $(k = 1, \ldots, D-1)$ the squared line element $ds^2$ is nearly Euclidean. Distortions occur whenever leaving the equatorial great circle. More precisely, distortions*

*are higher when angles $\phi_k$ with low values of the index $k$ (outer angles) are close to zero or $\pi$, than when angles $\phi_k$ with high values of the index $k$ (inner angles) are close to zero or $\pi$.*

DEFINITION 2.2 (Torus to Sphere Deformation). *With a data-driven permutation $p$ of $\{1,\dots,D\}$, data-driven central angles $\mu_k$ ($k=1,\dots,D$) and data-driven scalings $\alpha_k$, 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 $2\pi$ such that it is in the range $(-\pi,\pi]$.*

We now explain in detail how the choices are data-driven. Further illustration including practical advice is given in Supplement A. First, we comment on the general applicability of T-PCA.

REMARK 2.3. *The singularity set introduced, forms a subtorus of dimension $D-2$. In consequence, T-PCA is applicable, whenever there is a structural data gap in all angles except for at most two; the larger the gap, the higher the structural fidelity.*

**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 $\pi$, choose $\alpha_{k'} = 1$ ($k' = 1,\dots,D-1$) leading to *unscaled* (U) angles. Otherwise, we choose $\alpha_{k'} = 1/2$ ($k' = 1,\dots,D-1$) leading to *halved* (H) angles. In practical situations, the torus data are often spread out over more than half circles for several angles. Then we choose (H) angles. In fact, for all of the analyses below, we chose (H) angles and discuss below only the gluing effects corresponding to (H) angles. Notably, the innermost angle $\phi_D$ always remains unscaled: $\alpha_D = 1$. This is depicted in the second row of Figure 5.

**The central angles** $\mu_k$ will be chosen such that the mapped data points come to lie near the equatorial great circle and omit the singularities. Two plausible choices are:

(i) with the circular intrinsic mean $\overline{\psi}_{k,\text{intr}}$, set $\mu_k = \overline{\psi}_{k,\text{intr}}$ to obtain *mean centered* data;
(ii) with $\psi_{k,\text{gap}}$, the center of the largest gap between neighboring $\psi_k$ values of data points and $\psi^*_{k,\text{gap}}$ its antipodal point, define $\mu_k = \psi^*_{k,\text{gap}}$ to obtain *gap (antipode) centered* data.

While the implementation for (ii) is straightforward, for (i) we have used the fast algorithm from Hotz and Huckemann (2014). Mean-centered data has the merit that the intrinsic means for each angle $\phi_k$ are mapped to the equatorial great circle thus minimizing deformation of the data.

For a strongly skewed data distribution, say, spread out over a half circle, mean centered data using halved angles may touch the singularities, leading to high distortion there, while gap centered data will still be confined to a $\pi/2$ neighborhood of the equator. On the other hand, for data sets with outliers, gap centered centering may be less robust than mean centered, making the latter more favorable, as depicted in Figures 5c and 5e.

REMARK 2.4. Robustness w.r.t. outliers *is surprisingly different on a compact space than on the usually considered non-compact spaces. Specific loci of outliers occurring nearly antipodal to the data bulk do not much affect the location of the mean, the largest data gap, however, is much more sensitive to these loci.*

**The choice of the permutation** $p_k$ is driven by analyses of the *data spread*

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

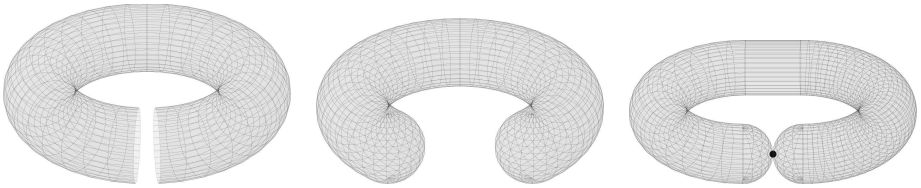
for each angle, where $\psi_{k,i} \in \mathbb{S}^1$ are the torus data and $n$ is the number of data points on $T^D$. If the angles are ordered by increasing data spread, such that $\sigma^2_{p(1)}$ is minimal and $\sigma^2_{p(D)}$ is maximal, in view of Remark 2.1, the change of distances between data points caused by the deformation factors $\sin^2 \phi_j$ in Equation (1) is minimized. We call this ordering *spread inside* (SI), because variation is concentrated on the inner angles of the sphere. The opposite ordering is called *spread outside* (SO). Figure 5 illustrates different effects of SI and SO ordering of angles. We will restrict our considerations to these two options.

**Self-gluing in case of halved angles: "From a donut to a sausage".** In the following, we give a brief overview of this procedure for (H) halved-angles (not for (U) angles for the reasons given above).

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 $\phi_k$ ($k = 1, \dots, D-1$), $\phi_k = 0$ denotes spherical locations different from $\phi_k = \pi$. For a representation respecting the torus' topology, however, it is necessary to identify these locations accordingly. Due to the spherical geometry, each of those regions is of dimension $D - j - 1$, in which all angles vary except for $j$ of the

$\phi_1, \ldots, \phi_{D-1}$ which are set to fixed values in $\{0, \pi\}$. In the topology of the torus, all those regions with a specific choice of fixed angles are identified with one-another. In particular, there are $2(D-1)$ such regions of highest dimension $D-2$ on the sphere (where only one angle is fixed to 0 or $\pi$), two of which are pairwise identified in the topology of the torus. In fact, in the topology of the torus, each of these $D-1$ regions of highest dimension $D-2$ itself carries the topology of a torus of dimension $D-2$, each glued to each other torus along a subtorus of dimension $D-3$, and so on. Thus the *self-gluing* of $S^D$ giving the topology of $T^D$ can be iteratively achieved along a topological subsphere of dimension $D-2$ which is suitably divided into $2(D-1)$ regions that are pairwise identified by way of a torus, sharing common boundaries which correspond to lower dimensional tori.

Example 2.5 details the case $D = 3$ and Figures 4 and 5 illustrate the case $D = 2$ as well as different choices for the permutation $p$.



(a) *Cutting open along a circle, giving two circles.* (b) *Separately collapsing each circle to a point.* (c) *Identifying the points restores the cyclic topology.*

Fig 4: *Self-gluing of $T^2$: From a donut to a sausage. These operations are only topological, Figure 5 reflects the changes in geometry.*

EXAMPLE 2.5. 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 $\phi_1, \phi_2 \in [0, \pi]$, $\phi_3 \in [0, 2\pi)$.

Due to the spherical geometry in the region determined by $\phi_1 = 0 \bmod \pi$ or $\phi_2 = 0 \bmod \pi$, the circle $\phi_3 \in [0, 2\pi)$ is a single point, say, $\phi_3 = 0$. This region is a topological circle on $S^3$ comprising four arcs

$$\begin{aligned} A_1 &= \{(0, \phi_2, 0) : 0 \le \phi_2 < \pi\} , & A_2 &= \{(\pi, \phi_2, 0) : 0 \le \phi_2 < \pi\} , \\ A_3 &= \{(\phi_1, 0, 0) : 0 \le \phi_1 < \pi\} , & A_4 &= \{(\phi_1, \pi, 0) : 0 \le \phi_1 < \pi\} . \end{aligned}$$

Imposing the topology of the torus, when using halved angles, for $\phi_1$ and $\phi_2$ we also have the identification $0 \equiv \pi$ which results in the identification of

$A_1$ with $A_2$ and of $A_3$ with $A_4$ with endpoints identified as one single point, forming a topological figure eight.

2.2. *Linking the Torus' Deformation to PNS.* For data sets on a torus, having applied a deformation on the resulting self-glued $\mathbb{S}^D$ (see Section 2.1), we modify principal nested sphere analysis (PNS) by Jung et al. (2010); Jung, Dryden and Marron (2012) for dimension reduction.

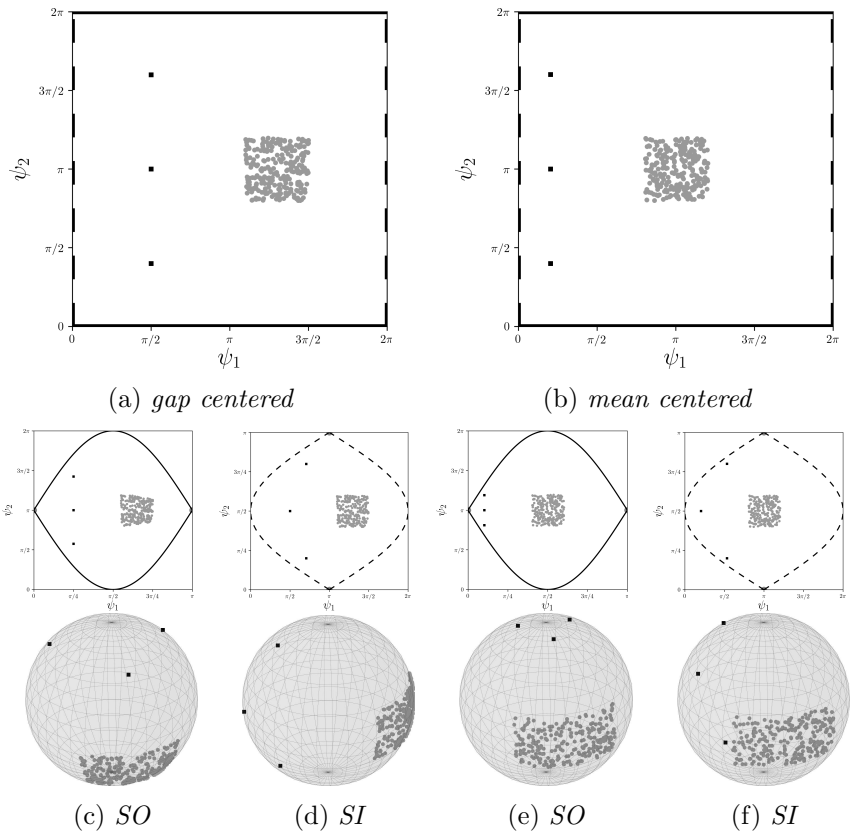
Assume a $d$-dimensional sphere $S^d \subset \mathbb{R}^{D+1}$ with center $x \in \mathbb{R}^{D+1}$ and radius $r > 0$, and an affine $d$-dimensional plane $A^d \subset \mathbb{R}^{D+1}$ with distance $s < r$ from $x$. For $d \geq 2$ then the intersection $S^d \cap A^d \subset \mathbb{R}^{D+1}$ is a $(d-1)$-dimensional subsphere $S^{d-1}$ of $S^d$ with radius $r = \sqrt{1-s^2}$. If $r = 1$ (i.e. $s = 0$) this subsphere is a *great subsphere*, otherwise it is a *proper small subsphere*. For $d = 1$ we pick just one point $\mu$, writing in expedient abuse of notation: $S^0 = \{\mu\}$. In order to include all, great, proper small subspheres and the ultimate point, we call these *small subspheres*.

The PNS iteration leads to a sequence of small subspheres

$$\mathbb{S}^D \supset S^{D-1} \supset \cdots \supset S^2 \supset S^1 \supset S^0 = \{\mu\}\,, \tag{4}$$

where the ultimate point $\mu$ is called the *nested mean*. Each $S^d$ $(d = 1, \ldots, D)$ is a $d$-dimensional sphere, the radii of which decrease monotonically with decreasing dimension (due to nesting). At each reduction step, the *residual variances not explained* by the corresponding subsphere are given as signed distances: points lying inside the small subsphere – if it is a proper small sphere – receive a positive distance, points lying outside a negative distance. Indeed, for most realistic data applications, with probability one, all subspheres are proper small subspheres. However, to avoid overfitting, we want to ensure that the "small subsphere" is not too small but rather a great subsphere is fitted; see Section 2.4. In this case the direction of positive distance is picked at random. Similarly, we pick the direction of positive distance at random for the reduction from $d = 1$ to $d = 0$.

The classical PNS algorithm consists of two parts which alternate, namely the fitting of a subsphere $S^d$ and the projection to this subsphere $\pi_d : S^{d+1} \to S^d$ $(d = D-1, \ldots, 0)$ giving the *fitted values explained* by this subsphere. As $\mathbb{S}^D$ is glued to itself in T-PCA, distances through the glued part can be shorter than spherical distances. In such cases, these distances are used in the fitting step as well as in the projection step. More precisely, our fitting procedure is done in two steps to avoid local minima. In the first step, we minimize the sum of squares of spherical distances. The resulting subsphere is taken as a starting point for the second step.




(a) *gap centered* (b) *mean centered*

(c) *SO* (d) *SI* (e) *SO* (f) *SI*

Fig 5: *All possibilities for gluing for $T^2$, illustrated by a data set uniform in a square with three outliers. Using mean centered (5b), the square is near the equatorial great circle ($\psi_1 = \pi$ for SO (5e) and $\psi_2 = \pi$ for SI (5f)) and thus the square suffers little distortion, in comparison to the outliers. For gap centered (5a), the outliers are less distorted and for SO (5c) the square is particularly distorted because the equatorial great circle ($\psi_1 = \pi$) is then between outliers and square. In both cases, SO decreases the spread of the outliers, SI increases it, more drastically for mean centered. Due to the torus' periodicity, lines of same type in the flat torus angle plots (top row, 5a and 5b) are identified. The respective outer angle is halved, the respective inner angle is unscaled (middle row, (5c) and (5d)). Due to collapsing of some identified lines to points (the singularity set, in Figure 4 this is the circle along which the donut is cut), north and south pole of each sphere are identified (bottom row, (5e) and (5f)).*

For the second step, we use the torus metric

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

Assuming a data set $\mathcal{A}$ and a corresponding adaptive deformation $P_{\mathcal{A}} : T^D \to \mathbb{S}^D$ we define the following function on the sphere

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

using the inverse deformation $P_{\mathcal{A}}^{-1}$, which is well-defined except for the singularities which are of dimension $D-2$. This is a metric when we take into account the topological identifications. To considerably lower computational speed for data analyses, we orthogonally project data to lower dimensional subspheres using the spherical geometry only. On the deformed torus this can be viewed as a non-orthogonal projection. For the minimization in the second step, however, we use $\tilde{\delta}$ as the distance function.

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 because decrements correspond to the non-increasingly ordered eigenvalues of the corresponding covariance matrix. This means that every component can be thought of as contributing a fixed amount of variance and thus the sum of such individual variances can be understood as *explained variance*. If one views the principal components as defining a nested sequence of subspaces, the amount of variance which is not explained by the components spanning the subspace is equal to the *residual variance* of data around the subspace. Explained variance and residual variance add to 1 and thus yield equivalent descriptions of data variance.

In non-Euclidean spaces, linear PCA is not applicable and non-linear dimension reduction methods do not come with a similar notion of additive variance (see the discussion for various definitions of intrinsic variances in Huckemann, Hotz and Munk (2010)). This means that explained variance can no longer be defined in a straightforward way. However, residual variance is still a well-defined notion, therefore we use residual variances in the following to define cumulative variances, and to compare results of different approaches.

Recall that T-PCA just as PNS yields a sequence of subspaces $\mathbb{S}^D \supset S^{D-1} \supset \cdots \supset S^1 \supset S^0 = \{\mu\}$ with projections $\pi_d : S^{d+1} \to S^d \subset S^{d+1}$ $(d = 0, \ldots, D-1)$. 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 $S^d$) of a data set $\mathcal{A}$

$$V_{\mathcal{A},P_{\mathcal{A}},d} = \sum_{q\in\mathcal{A}} \tilde{\delta}^2(q, \Pi_d(q))\,, \quad d = 0,\dots,D-1$$

and $V_{\mathcal{A},P_{\mathcal{A}},D} = 0$, where $\tilde{\delta}$ is from (5). Due to nestedness, these sequences are non-increasing with $d$. However, the decrements $V_{\mathcal{A},P_{\mathcal{A}},d-1} - V_{\mathcal{A},P_{\mathcal{A}},d}$ $(d = 1,\dots,D)$ are not necessarily non-increasing, so the resulting curve in the variance plot need not be convex. Still, this allows to define that $\{\mu\}, S^1,\dots,S^d$ explain the *cumulative variance* up to dimension $d$

$$V_{\mathcal{A},P_{\mathcal{A}},0} - V_{\mathcal{A},P_{\mathcal{A}},d}\,, \quad d = 0,\dots,D$$

which is non-decreasing in $d$.

2.4. *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 overfitting is obviously undesirable, Jung, Foskey and Marron (2011); Jung, Dryden and Marron (2012) have fitted a great subsphere in such cases: Jung, Foskey and Marron (2011) have given a decision rule whereas Jung, Dryden and Marron (2012) have given a test for this purpose. We propose the following new test based on a geometrically better suited model and highlight its attractive properties, in particular we show how robust is our test under the null model of Jung, Dryden and Marron (2012), that is a misspecified model for our case. We also indicate some limitations of the two previous procedures.

**New model.** Let $S^d$ be a fitted small subsphere, $2 \le d < D$. For ease of notation, we now move and rescale $S^d$ to the unit sphere $\mathbb{S}^d$, without loss of generality, and $p \in \mathbb{S}^d$ is the center of the, also moved and rescaled, fitted small subsphere $S^{d-1} \subset \mathbb{S}^d$. For our purpose, we can restrict our probability model for $q \in \mathbb{S}^d$, say, $g(q;p)$, to depend only on the angular distance $r = d(p,q) \in [0,\pi]$. Further suppose that $\mathrm{vol}_{\mathbb{S}^d}$ denotes the surface volume of the $d$-dimensional unit sphere. Then, due to symmetry, $g$ fully characterizes the *spherical angular marginal density* of $r$

$$h(r;p) := \mathrm{vol}_{\mathbb{S}^{d-1}} \cdot g(\gamma(r);p), \quad r \in [0,\pi]\,. \tag{6}$$

Here, $\gamma$ is any curve along a great circle connecting $p$ with its antipodal, parametrized by $r \in [0,\pi]$ such that $\forall r : d(p,\gamma(r)) = r$. Using the spherical volume element $d_{\mathbb{S}^d}\Omega(q)$ at $q = \gamma(r)$ we note that

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

which means that $h(\cdot; p)$ is indeed a marginal density with respect to the spherical angular measure

$$d\mu(r) = \sin^{d-1}(r)dr, \quad r \in [0, \pi]\,.$$

Then the *Lebesgue angular marginal density* $f(\cdot; p)$ of $r$ is defined as

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

since it gives the marginal density corresponding to $h(\cdot; p)$ with respect to the Lebesgue measure on $[0, \pi]$.

Note that these densities are well studied for $d = 2$ where the angle $r$ is called colatitude (see for example, Mardia and Jupp (2000)); the uniform distribution in polar coordinates for any $d$ on which this discussion is based, see, for example, Mardia, Kent and Bibby (1979).

For the following, we will need the density of the "folded normal distribution" on $[0, \infty)$:

$$\mathcal{F}(r; \rho, \sigma) := \frac{1}{\sqrt{2\pi}\sigma}\left(\exp\left(-\frac{(r-\rho\sigma)^2}{2\sigma^2}\right) + \exp\left(-\frac{(r+\rho\sigma)^2}{2\sigma^2}\right)\right)\,.$$

That is, we have

$$\mathcal{F}(r; \rho, \sigma) = \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{r^2}{2\sigma^2} - \frac{\rho^2}{2}\right)\cosh\left(\frac{r\rho}{\sigma}\right),\ r \geq 0\,. \tag{7}$$

This density has two positive parameters, $\rho$ and $\sigma$. Note that here $r$ is on $[0, \infty)$ so it is not restricted to $[0, \pi]$, a fact which will be of importance later on where we will truncate this distribution. For $\rho \to \infty$ this tends to a usual normal distribution centered at $\rho\sigma$, while it becomes a halved normal distribution (of doubled height) for $\rho \to 0$. For $\rho \leq 1$ the mode stays fixed at the origin, for $\rho > 1$ it moves to the right.

With the above discussion on the marginals we therefore choose $g \propto \mathcal{F}$ yielding the spherical angular marginal density $h$ and the Lebesgue angular marginal density $f$:

$$\begin{aligned} h(r; p, \rho, \sigma) &:= \frac{\sqrt{2\pi}\sigma}{\mathcal{C}(\rho, \sigma)}\mathcal{F}(r; \rho, \sigma), \\ f(r; p, \rho, \sigma) &:= \frac{\sqrt{2\pi}\sigma}{\mathcal{C}(\rho, \sigma)}\sin^{d-1}(r)\mathcal{F}(r; \rho, \sigma), \quad r \in [0, \pi]\,, \end{aligned} \tag{8}$$

where we have truncated $\mathcal{F}(r;\rho,\sigma)$ from (7) and $\mathcal{C}(\rho,\sigma)$ is the normalization. These will be referred to as *h- and f-distribution*, respectively, in the following.

Subsequently, it will be important to note the following property of these distributions, for dimension $d = 2$, 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 case $\rho > 1$ yields a ring while the case $\rho = 0$ yields a symmetric Gaussian distribution. Due to its smoothness it is a good candidate for a test distribution for the angular spherical marginal density (6) to distinguish "just" concentrated data near $p$ ($p$ is at $r = 0$) from concentrated data along a distinct subsphere (a ring in 2D) around $p$.

**Likelihood ratio test.** Suppose we are given the sample $\{q_1,\dots,,q_n\}$ from the $f$-distribution with the spherical distances $r_i = d(p,q_i)$ $(i = 1,\dots,n)$ where the center $p$ of the subsphere is known. If $\rho \le 1$, the $h$ distribution has its maximum at $r = 0$, i.e. there is no proper small spherical structure about the center $p$. If $\rho > 1$, there is a proper small spherical structure about the center $p$. Thus, $\rho = 1$ forms the boundary between the two cases.

Therefore, we can formulate our hypotheses as follows for testing for a great subsphere.

$$H_0 : \rho = 1 \text{ (great subsphere) vs. } H_1 : \rho > 1 \text{ (small subsphere)}\,. \tag{9}$$

The log likelihood up to a constant is given by

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

Note that the normalization $\mathcal{C}(\rho,\sigma)$ can be easily computed numerically so we can determine the MLEs for $\rho$ and $\sigma$ using standard numerical optimization. For $H_1$, the MLEs need to be constrained under $\rho > 1$. Then twice the log of the likelihood ratio (with negative sign) is given by

$$\begin{aligned}\lambda =& 2\sup\{\ell(\rho,\sigma|\{r_i\}_{i=1}^n) : \rho \in (1,\infty),\, \sigma \in \mathbb{R}^+\}\\ &- 2\sup\{\ell(\rho,\sigma|\{r_i\}_{i=1}^n) : \rho = 1,\, \sigma \in \mathbb{R}^+\}\,.\end{aligned} \tag{10}$$

From Wilks' theorem, the statistic $\lambda$, under $H_0$, is asymptotically distributed as $\chi_1^2$. We use a 5% significance level for our test, which means that when $H_0$ is rejected, we keep the fitted small subsphere if $\lambda > \chi^2_{1,0.95} \approx 3.84$; otherwise, we perform a great subsphere fit.

**Comparison with the decision rule of Jung, Foskey and Marron (2011).** This rule is based on another type of angular $h$ and $f$ (versus our angular $f$ and $h$ given by (8))

$$h_{Jung}(r; p, \rho, \sigma) := \frac{1}{\sin^{d-1}(r)} \mathcal{F}(r; \rho, \sigma), \quad f_{Jung}(r; p, \rho, \sigma) := \mathcal{F}(r; \rho, \sigma)\,.$$

In their decision rule, a great sphere is fitted if the probability distribution does not exhibit a ring-shaped local maximum, which is the case if $\rho \leq 2$. But this model leads to a singularity of the density $h_{Jung}$ at $p$, which is not a desirable feature. In contrast, our $h$-distribution leads to a smooth distribution on the sphere as illustrated by the above considerations about the surface of revolution. Our $h$ distribution is compared with the $h_{Jung}$ distribution in Figure 6 for appropriate values of $\rho$. Our $h$ distribution is the same for all $d$ but for illustration, we have used $d = 2$ for $h_{Jung}$ which depends on $d$.

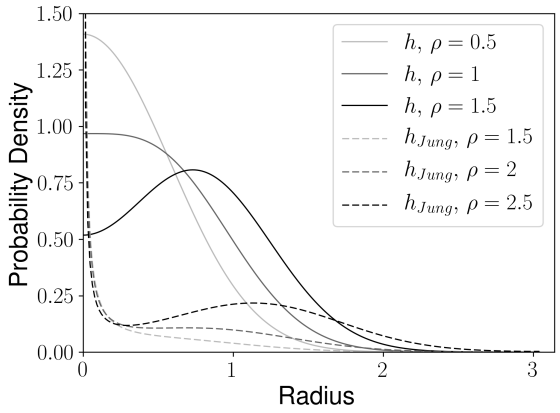



Fig 6: *The probability densities for $\sigma = 0.5$ along the geodesic $\gamma$ in $\mathbb{S}^d$ from (6) for our $h$ (invariant under $d$) and the $h_{Jung}$ (for $d = 2$) distribution. Displaying a value for $\rho$ below the respective boundary, at the boundary and above the boundary; namely, $\rho = 1$ for our $h$ and $\rho = 2$ for the $h_{Jung}$ distribution.*

**In validation of our test** we carried out two simulation studies.

$D_0$**:** We simulate data under $H_0$ in (9) by choosing $\rho = 1$ in (8) and average over the nuisance parameter $\sigma$ by 1000 samples, where in each $\sigma$ is uniform in $[0.1, 0.4]$.

$D_1$**:** We simulate data under $H_1$ in (9) by choosing various combinations of $\rho \in \{1.2, 1.5, 2, 3\}$ and $\sigma \in \{0.15, 0.2, 0.5\}$ in (8), and for each we average using 1000 samples.

TABLE 2

*Type 1 errors (rejecting $H_0$) for $D_0$ and Type 2 errors (accepting $H_0$) for $D_1$ for our test with various parameter values in a simulation with* 1000 *repetitions and asymptotic level of* 5%*, i.e. rejecting for $\lambda > \chi^2_{1,0.95} \approx 3.84$ with $\lambda$ from (10).*

| | Type 1 ($D_0$) | Type 2 ($D_1$) | | | | |
|---|---|---|---|---|---|---|
| Sample size | $\rho = 1$ | $\rho = 1.2$ $\sigma = 0.15$ | $\rho = 1.5$ $\sigma = 0.2$ | $\rho = 2$ $\sigma = 0.2$ | $\rho = 2$ $\sigma = 0.5$ | $\rho = 3$ $\sigma = 0.15$ |
| 100 | 7.4% | 80.4% | 41.2% | 3.4% | $< 0.1\%$ | $< 0.1\%$ |
| 200 | 5.5% | 73.2% | 20.2% | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ |
| 500 | 5.0% | 59.7% | 1.0% | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ |
| 1000 | 4.9% | 34.7% | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ |

TABLE 3

*We estimate the asymptotic level for our test leading to a true level of* 5% *i.e. achieving a Type 1 error (rejecting $H_0$) for $D_0$ of* 5%*. The table gives the asymptotic level and Type 2 errors (accepting $H_0$) for $D_1$ for our test with various parameter values in a simulation with* 1000 *repetitions.*

| | ($D_0$) | Type 2 ($D_1$) | | | | |
|---|---|---|---|---|---|---|
| Sample size | asymptotic level $\rho = 1$ | $\rho = 1.2$ $\sigma = 0.15$ | $\rho = 1.5$ $\sigma = 0.2$ | $\rho = 2$ $\sigma = 0.2$ | $\rho = 2$ $\sigma = 0.5$ | $\rho = 3$ $\sigma = 0.15$ |
| 100 | 3.0% | 85.0% | 54.4% | 5.1% | $< 0.1\%$ | $< 0.1\%$ |
| 200 | 4.4% | 76.4% | 24.2% | 0.1% | $< 0.1\%$ | $< 0.1\%$ |
| 500 | 5.0% | 59.7% | 1.0% | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ |
| 1000 | 5.0% | 34.7% | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ | $< 0.1\%$ |

The results in Table 2 show that our test at *asymptotic level* of 5%, i.e. it rejects a small sphere when $\lambda > \chi^2_{1,0.95} \approx 3.84$ with $\lambda$ from (10), holds asymptotically the level and that the Type 2 error asymptotically decays to zero, very quickly for larger $\rho$. Since for $N = 100, 200$ the true levels are above 5%, we have estimated the asymptotic levels yielding a true level of 5% in Table 3 and display the corresponding Type 2 error there also. This estimation is a matter of minutes for $N = 100$ and below one hour for $N = 1000$. Based on these simulations we recommend to use our test for sample sizes at least around $N = 200$. This is the case for our application to the C2 data set with $N = 649$ (cf. Section 3.2) and almost the case for the benchmark data set with $N = 181$ (cf. Section 3.1). For both data sets

we have used our test against overfitting a small sphere at asymptotic level of 5%.

**Assessment of robustness of our test under the null distribution of Jung, Dryden and Marron (2012).** We now assess the robustness of our test under a misspecified model, namely, the von Mises-Fisher distribution which is the null distribution of Jung, Dryden and Marron (2012). To carry this out, we note the following points related to their test. First, we note that they have translated their null hypothesis of a compact cluster into fitting by a great subsphere through a von Mises-Fisher distribution. The parameters of this distribution are estimated via MLE. Then a Student $t$-like test statistic of distances to the estimated center point is used as their test statistic. Next, we note that for their test statistic, they simulate bootstrap quantiles from the von Mises-Fisher distribution with parameters given by the MLE. However, Jung, Dryden and Marron (2012) have given neither a theoretical result – like we have the asymptotic p-value of our test statistics $\lambda$ – nor a simulation study to assess their test statistics under their null hypothesis. We have reimplemented their data driven procedure so as to use their null hypothesis and have carried out the following simulation study.

$D_0'$**:** Here, we directly simulate spherical samples leading to a great circle, from the null hypothesis of the test of Jung, Dryden and Marron (2012), namely from a von Mises-Fisher distribution with density in $x$ proportional to $e^{\kappa\mu^T x}$ with a high value of the concentration parameter $\kappa = 10$ to give a fair chance. We average over 1000 samples with $\mu$ uniform on the sphere.

TABLE 4

*Type 1 errors (rejecting the null hypothesis of Jung, Dryden and Marron (2012) which is a von Mises-Fisher distribution) for the test of Jung, Dryden and Marron (2012) and errors under this misspecified model for our test, with concentration parameter $\kappa = 10$ in a simulation with 1000 repetitions. For their test we use a simulated level of 5% and for our test we use an asymptotic level of 5%.*

| Sample size | Jung, Dryden and Marron (2012) | our test |
|---|---|---|
| 100 | 17.0 % | 1.0 % |
| 200 | 13.4 % | $< 0.1$ % |
| 500 | 8.4 % | $< 0.1$ % |
| 1000 | 5.9 % | $< 0.1$ % |

As shown in Table 4, we note that our test is more conservative on the null hypothesis of the test of Jung, Dryden and Marron (2012). Further, the true level of the test of Jung, Dryden and Marron (2012) also decreases with

sample size, and almost reaches the simulated level for $N = 1000$. In passing, we note that estimating the simulated level leading to a true level of 5% for the test by Jung, Dryden and Marron (2012), however, is impractical, as for $N = 100$ already, estimation takes weeks.

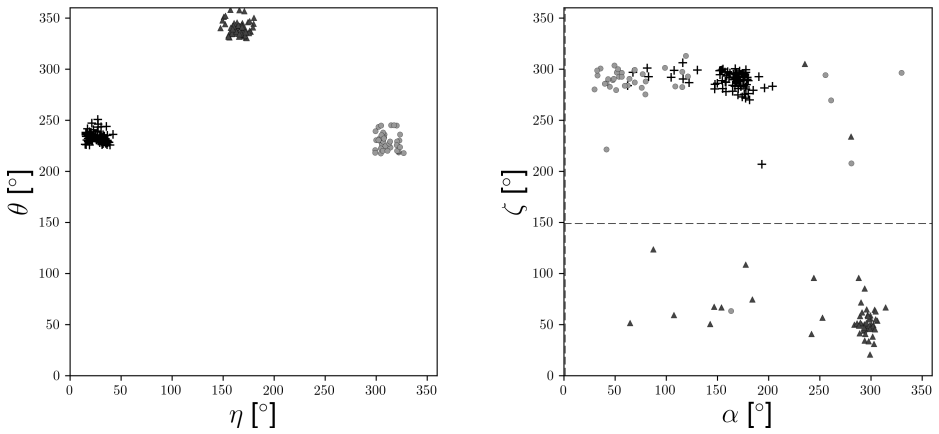
**3. Application to RNA Structure.** RNA is usually single-stranded and the single strand interacts with itself, forming complex shapes (this is in contrast to DNA which usually takes a double-stranded helical conformation). This means that the geometry is rather variable even on the scale of single atoms. As described in Section 1, 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, cf. Table 1 and Figure 3. The distribution of these 7 angles over large samples of RNA strands have been studied in detail, see Murray et al. (2003); Schneider, Morvek and Berman (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. An approximation of the geometric folding structure on the level of single residues is given by the two *pseudo-torsion angles* $\eta$ and $\theta$ (Figure 3b). These two (dihedral) angles provide at once a two-dimensional visualization (Figure 7a), see e.g. Duarte and Pyle (1998); Wadley et al. (2007).

Finally, the dihedral angle $\nu_2$ (Figure 3b and Table 1) quantifies the folding (pucker) of the sugar ring. Only two modes of folding are geometrically and energetically possible, which are characterized by either C3' or C2' being outside the plane spanned by C1'-O1'-C4' and towards the direction of O5'. If C2' lies outside the plane then $\nu_2 \approx 325°$, this is called *C2'-endo* sugar pucker, whereas if C3' lies outside the plane then $\nu_2 \approx 35°$, this is called *C3'-endo* sugar pucker. The hydroxy group attached to the C2' atom in RNA causes the C3'-endo sugar pucker to be energetically preferred (see e.g. Egli, Portmann and Usman (1996)) and thus this is about 10 times more abundant than the C2'-endo sugar pucker in the large RNA data set of Duarte and Pyle (1998) and Wadley et al. (2007).

For our application below we use two subsets of a large classical data set (8301 residues) which was carefully selected for high experimental X-ray precision (0.3 nanometers) by Duarte and Pyle (1998), updated by Wadley et al. (2007) and analyzed by them and others, for example, Murray et al. (2003); Richardson et al. (2008).

3.1. *The Benchmark Data Set.* This benchmark data set has been carefully selected by Sargsyan, Wright and Lim (2012) to validate their method. From the C3'-endo sugar pucker they took clusters labeled I ("triangles",

59 points), II ("crosses", 83 points) and V ("disks", 39 points) by Wadley et al. (2007) totaling 181 data points, which form three clusters in the $\eta$–$\theta$ plot as shown in Figure 7a. 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.




(a) $\eta$–$\theta$ *plot* (b) $\alpha$–$\zeta$ *plot*

Fig 7: *7a: The benchmark data set of Sargsyan, Wright and Lim (2012) with their three preselected clusters in the $\eta$–$\theta$ plot. 7b: The benchmark data set plotted for the two most discriminant angles ($\alpha$, $\zeta$) chosen out of the seven dihedral angles; in the "donut to sausage" transformation along the dashed lines the corresponding angles are collapsed to a single point.*

Visualization is obviously not possible in the 7D space of all torsion angles. However, we find that the angle pair ($\alpha$, $\zeta$) is the most discriminatory and a plot is given in Figure 7b: The "disks" cluster is not very concentrated, in contrast to the "crosses" cluster which is twice as big, and parts of the "disks" are very close to the "crosses" cluster. In fact, upon close inspection, due to periodicity, the "triangles" and "crosses" clusters are also rather close in the $\eta$–$\theta$ plot in Figure 7a.

We have applied T-PCA to all seven angles and depict the two-dimensional representation for SI ordering in Figure 8a (which is hardly visually distinguishable from SO ordering). To see that the data are, in fact, very well approximated by the best fit circle we use a planar representation of the first two T-PCs in Figure 8b. Using the same symbols for Figure 8 as in Figure 7 shows that the three preselected clusters can be rather well distinguished by eye. We note that the first component explains 84 % of data

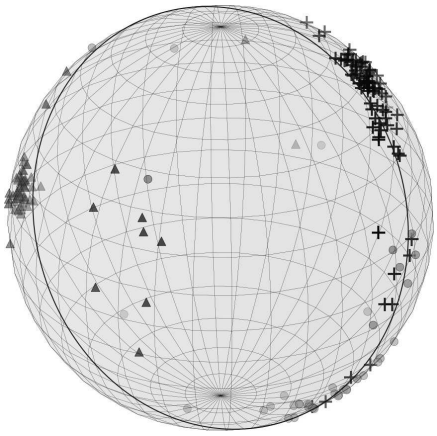
variation. In comparison in Figure 8c we adapt Figure 6 from Sargsyan, Wright and Lim (2012). Again the clusters can be well discriminated along the first GeoPC (horizontal in the 2D approximation in Figure 8b). In contrast to T-PCA, however, the data are not well approximated by the first GeoPC, as the projections to the second GeoPC component (vertical in the 2D approximation in Figure 8c), feature maximal data range. In fact, both GeoPCs explain roughly similar amounts of data variation.

Thus Figure 8 illustrates the power of T-PCA going significantly beyond the analysis of Sargsyan, Wright and Lim (2012). Not only can the preselected clusters be separated but the data are very accurately approximated by their projection to the 1D component.

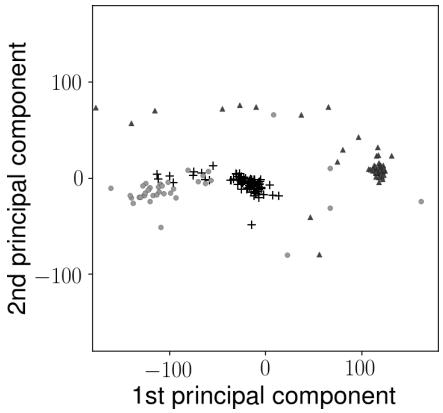
3.2. *The 1D structure of C2 Data Set.* We now describe in detail how our C2 data set is extracted from the large RNA data set. Notably, some of the RNA structures in this data set are only short pieces adhering to a protein or another RNA structure. Therefore, we prune by removing residues further than $50^\circ$ in torus distance from their nearest neighbor. This leads to 7544 residues and 649 of these are residues with C2'-endo sugar pucker. i.e. $\nu_2 \in [300^\circ, 350^\circ]$. This produces a moderately large data set to analyze (in contrast to the very large data set of all other residues including C3'-endo sugar pucker).

Murray et al. (2003) noted that this data set is locally *rotameric*, as, among others, conformer clusters essentially extend along the $\beta$ angle, considering only the 3 *heminucleotide* angles $\alpha - \beta - \gamma$ (Figure 9a). Already in this heminucleotide space, these individual 1D cluster patterns compete with the group spread along the $\alpha$ angle and in full 7D residual space, there are more competing features, which, in the 2D TS-PCA plot involving all 7 angles, manifest as 3 diffused stripe shaped clusters (Figure 9b). Here the 1D pattern of the largest conformer group can be traced along the shifted second diagonal. The two conformer groups next in size, which are close in heminucleotide angles, are ripped apart in TS-PCA due to its wrong topology, because they are far from the base point of the tangent space that is controlled by the dominating cluster. Notably, the correct topology could not even be forced onto that plot because, due to the winding effects illustrated in Figure 1, boundary loci correspond to different torus loci.

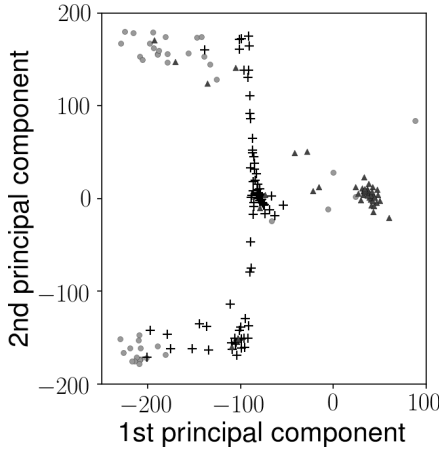
Due to its larger flexibility and higher fidelity, T-PCA recovers a 1D pattern as the overall dominating structure, reflecting the proximity of the second and third largest cluster in the 2nd component (Figure 9c, and Figure 9d in planar representation for better illustration, which is, of course, periodic). Notably, according to Remark 2.3, structural fidelity can be expected



(a) *2D approximation, T-PCA (SI)*




(b) *2D approximation, T-PCA (SI) in planar coordinates*

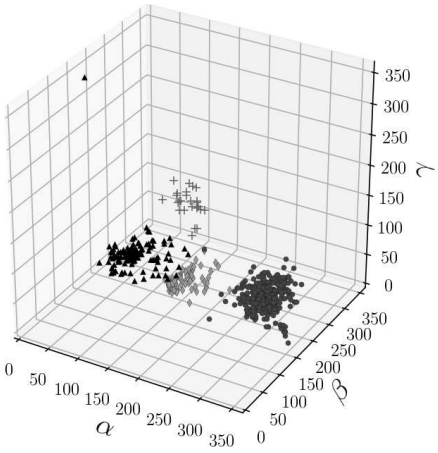



(c) *2D approximation, GeoPCA*

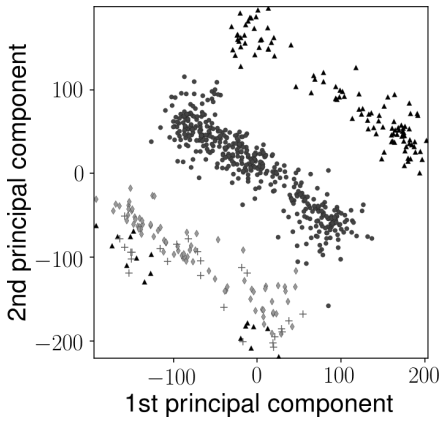
Fig 8: *Two-dimensional PCA approximations of the benchmark data set via T-PCA with SI ordering in natural spherical coordinates (8a), in planar coordinates (8b) and GeoPCA adapted from (Sargsyan, Wright and Lim, 2012, Figure 6) (8c). The symbols represent the same clusters as in Figure 7.*

due to the large gaps in the $\beta$ and $\gamma$ angles, cf. Figure 9a. Using T-PCA, we generalize the finding of a locally rotameric structure by Murray et al. (2003) to

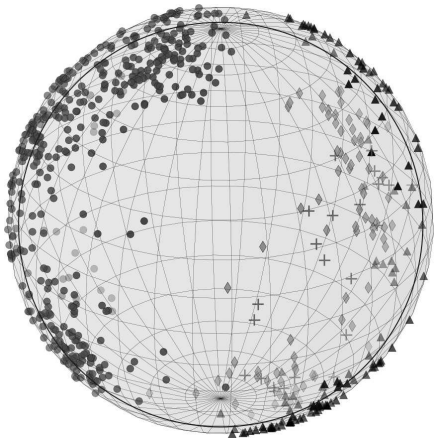
*In full 7D angular space, the RNA residue conformers are* rotameric,



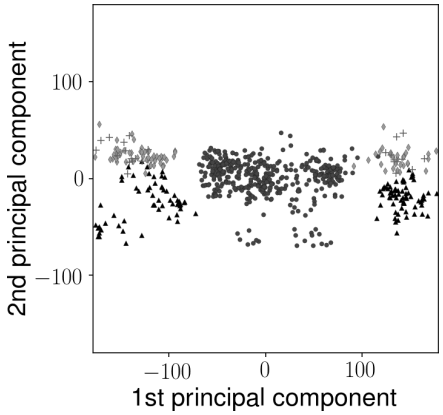

(a) *Conformer clusters in the $\alpha-\beta-\gamma$ plot adapted from (Murray et al., 2003, Figure 4.b)*




(b) *TS-PCA*



(c) *T-PCA (SI)*




(d) *T-PCA (SI), planar view*

Fig 9: *Residues with C2'-endo sugar pucker with clustering following Murray et al. (2003). Three-dimensional heminucleotide angles (9a); two-dimensional TS-PCA (9b) approximation; two-dimensional T-PCA (SI) approximation, the small circle gives the 1D approximation (9c); two-dimensional T-PCA (SI) approximation in planar representation (9d).*

*essentially following a single angle that is a non-linear combination of the original ones, however.*

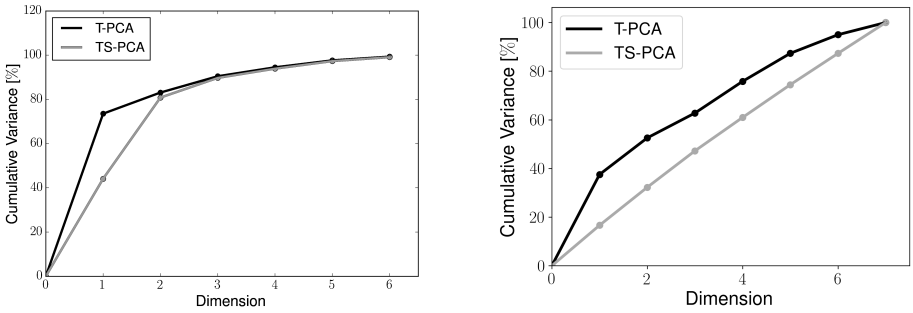
Upon yet closer inspection, the fine clustering along the 1D component re-

flects the clustering in the complementary heminucleotide $\delta - \epsilon - \zeta$ angles from (Murray et al., 2003, Figure 4.c, rear part).

3.3. *Comparing T-PCA with TS-PCA.* We summarize our use of T-PCA and TS-PCA using all 7 angles for the C2 data in Table 5a and Figure 10a. In 1D, T-PCA captures 73 % of the variance whereas TS-PCA captures only 44 % of the variance. Only when adding a second dimension TS-PCA captures more variance (81 %) than the 1D component of T-PCA. Higher order PCs, both for T-PCA and TS-PCA, explain roughly the same amount of data variance.

To highlight the differences between the two PCA methods, let us consider the example of three points. There is an exactly fitting small circle used by T-PCA. Indeed, if applied to the $\eta$–$\theta$ plot (Figure 7a), T-PCA would reduce the three clusters rather accurately to a 1D circle. In contrast, TS-PCA approximates three points only along a straight line in the tangent space and such an approximation is only possible if data lie favorably such as in the $\eta$–$\theta$ plot, see (Figure 7a). The $\alpha$–$\zeta$ plot (Figure 7b), however, illustrates that a 1D approximation for all seven angles is not possible for TS-PCA, while it is possible for T-PCA (Figure 8b).

In fact, usually T-PCA requires one dimension less than TS-PCA because $k$ points in general position span a $k$-dimensional affine subspace, which is detected by TS-PCA, and the surface of a $(k-1)$-dimensional sphere, which is detected by T-PCA. We illustrate this using a *simulated simplex data set* with points in general position, namely, 800 7D angles distributed independently at one of 8 simplex vertices, $\pi$ apart with Gaussian noise of variance $(\pi/3)^2$. The results are displayed in Table 5b and Figure 10b. If there are affine data dependencies, however, this advantage of T-PCA over TS-PCA by one dimension is lost. Indeed the C2 data set features such affine dependencies between angles, which is already visible in Figure 9a, and hence in Figure 10a, T-PCA outperforms TS-PCA in terms of explained variance only in dimension one.




(a) *C2 data* (b) *Simulated simplex data*

Fig 10: *Scree plots of cumulative variances for T-PCA (SI) compared to TS-PCA.*

TABLE 5
*Cumulative variances for T-PCA (SI) and for TS-PCA.*

(a) *C2 data*

| Dimension | T-PCA (SI) | TS-PCA |
|---|---|---|
| 1 | 74% | 44% |
| 2 | 83% | 81% |
| 3 | 90% | 90% |
| 4 | 95% | 94% |
| 5 | 98% | 97% |
| 6 | 99% | 99% |

(b) *Simulated simplex data*

| Dimension | T-PCA (SI) | TS-PCA |
|---|---|---|
| 1 | 39% | 18% |
| 2 | 50% | 34% |
| 3 | 63% | 48% |
| 4 | 77% | 62% |
| 5 | 89% | 75% |
| 6 | 95% | 88% |

**4. 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 leads to severe restrictions for geodesic approaches while, on the other hand, Euclidean approximations disregard periodicity. We have used an adaptive deformation to a statistically benign geometry, allowing for increased and statistically controlled flexibility whilst at the same time guaranteeing structure fidelity. In application to dihedral angles of RNA structures we have validated our method using a classical benchmark data set. Using a C2'-endo sugar pucker residue data set we have given evidence on how T-PCA is better and more meaningful than TS-PCA, and we have illustrated that the *significant interdependence* found by Murray et al. (2003) in a 3D representation is seen by T-PCA remarkably in 1D.

There are several benefits coming with dimension reduction to 1D. In view of data clustering it allows to build on powerful and well established sta-

tistical 1D methods for mode detection (e.g. Dümbgen and Walther (2008); Schmidt-Hieber, Munk and Dümbgen (2013); Huckemann et al. (2016)), and this challenge will be taken up in future research.

**Acknowledgments.** We are grateful to John Kent and J. S. Marron for helpful discussions. We thank Thomas Hamelryck and Jes Frellsen for their valuable comments on this paper 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 her and her collaborators.

## SUPPLEMENTARY MATERIAL

**Supplement A: Data**
(doi: COMPLETED BY THE TYPESETTER; .tar.gz). An illustration how to choose data-driven parameters for torus PCA.

**Supplement B: Data**
(doi: COMPLETED BY THE TYPESETTER; .tar.gz). RNA residue data used for the analysis in this paper.

**Supplement C: Implementation**
(doi: COMPLETED BY THE TYPESETTER; .tar.gz). Source code of the T-PCA implementation used for this paper.

## 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** 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* 924–931. Springer.

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* 346–355. Springer.

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

Chakrabarti, A., Chen, A. W. and Varner, J. D. (2011). A review of the mammalian unfolded protein response. *Biotechnology and Bioengineering* **108** 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** 459–485.

Chen, A. A. and García, A. E. (2013). High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **110** 16820–16825.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. et al. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research* **35** W375–W383.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

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

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

Dunbrack, R. L. and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Structural & Molecular Biology* **1** 334–340.

Egli, M., Portmann, S. and Usman, N. (1996). RNA hydration: a detailed look. *Biochemistry* **35** 8489–8494.

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** 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** 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** e1000406.

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika* **40** 33–51.

Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* 235–254.

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

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

Huckemann, S., Hotz, T. and Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica* **1** 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** 299–319.

Huckemann, S., Kim, K.-R., Munk, A., Rehfeldt, F., Sommerfeld, M., Weickert, J., Wollnik, C. et al. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli* **22** 2113–2142.

Jain, S., Richardson, D. C. and Richardson, J. S. (2015). Computational Methods for RNA Structure Validation and Improvement. In *Structures of Large RNA Molecules and Their Complexes*, (S. A. Woodson and F. H. Allain, eds.) **558** 181–212. Academic Press.

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

Jung, S., Foskey, M. and Marron, J. S. (2011). Principal arc analysis on direct product manifolds. *The Annals of Applied Statistics* **5** 578–603.

Jung, S., Liu, X., Marron, J. S. and Pizer, S. M. (2010). Generalized PCA via the Backward Stepwise Approach in Image Analysis. In *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. Body and*

*Machine* **83** 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*.

Laborde, J., Robinson, D., Srivastava, A., Klassen, E. and Zhang, J. (2013). RNA global alignment in the joint sequence–structure space using elastic shape analysis. *Nucleic acids research* **41** e114–e114.

Liu, W., Srivastava, A. and Zhang, J. (2011). A mathematical framework for protein structure comparison. *PLoS Comput Biol* **7** e1001075.

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** 487–514.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, New York.

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** 13904–13909.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral1, 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** e25.

Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale Methods for Shape Constraints in Deconvolution: Confidence Statements for Qualitative Features. *Ann. Statist.* **41** 1299–1328.

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

Seetin, M. G. and Mathews, D. H. (2012). RNA structure prediction: an overview of methods. *Bacterial Regulatory RNA: Methods and Protocols* 99–122.

Sommer, S. (2013). Horizontal Dimensionality Reduction and Iterated Frame Bundle and Development. In *Geometric Science of Information. Lecture Notes in Computer Science* **8085** 76–83.

Srivastava, A. and Klassen, E. P. (2016). *Functional and shape data analysis*. Springer.

Čech, P., Kukal, J., Černỳ, J., Schneider, B. and Svozil, D. (2013). Automatic workflow for the classification of local DNA conformations. *BMC bioinformatics* **14** 205.

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** 942–957.

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** 3450–3460.