Nonlinear Hypothesis Testing of Geometrical Object Properties of Shapes Applied to Hippocampi

Jörn Schulz · Stephen M. Pizer · J.S. Marron · Fred Godtliebsen

Abstract This paper presents a novel method to test mean differences of geometrical object properties (GOPs). The method is designed for data whose representations include both Euclidean and non-Euclidean elements. It is based on advanced statistical analysis methods such as backward means on spheres. We develop a suitable permutation test to find global and local morphological differences between two populations based on the GOPs. To demonstrate the sensitivity of the method, an analysis exploring differences between hippocampi of first episode schizophrenics and controls is presented. Each hippocampus is represented by a discrete skeletal representation (s-rep). We investigate important model properties using the statistics of populations. These properties are highlighted by the s-rep model that allows accurate capture of the object interior and boundary while, by design, being suitable for statistical analysis of populations of objects. By supporting non-Euclidean GOPs such as direction vectors, the proposed hypothesis test is novel in the study of morphological shape differences. Suitable difference measures are proposed for each GOP. Both global and local analyses showed statistically significant differences between the first episode schizophrenics and controls.

J. Schulz
Department of Mathematics and Statistics, University of Tromsø, Norway Tel.: +47 45696867
E-mail: jorn.schulz@uit.no

F. Godtliebsen
Department of Mathematics and Statistics, University of Tromsø, Norway, E-mail: fred.godtliebsen@uit.no

Stephen M. Pizer
Department of Computer Science, University of North Carolina at Chapel Hill (UNC), USA, E-mail: smp@cs.unc.edu

J.S. Marron
Department of Statistics & Operations Research, UNC, USA, E-mail: marron@unc.edu

Keywords Hippocampus · Hypothesis test · Permutation test · Principal nested sphere · Schizophrenia · Skeletal representation

1 Introduction

Statistical analysis of anatomical shape differences has been broadly reported in the literature (e.g., [2,5,10, 15]). In medical settings, the study of morphological changes of human organs and body structures is of great interest. An important subfield in medical imaging is to understand neuroanatomical structures of the human brain (e.g., [12,14,42]). Morphological changes of brain structures can provide the physician with information about neuropsychiatric diseases such as Alzheimer's and schizophrenia. A common interest of medical shape analysis is to test for morphological differences between healthy and diseased populations. In addition, the study of drug effects is of high interest in epidemiology. Volumetric measurements often can not distinguish between brain structure differences of two populations [45]. Therefore, sophisticated mathematical shape models with properties that support an accurate statistical analysis are required.

Shape differences can be quantified by hypothesis tests. A statistical hypothesis test requires a null hypothesis H_0 and an alternative hypothesis H_1; a standard null hypothesis assumes no differences between the populations. In this paper, we propose a novel approach for a hypothesis test on geometrical object properties (GOPs) of shapes with application to the hippocampus of the human brain.

Such a hypothesis test of populations of medical shapes depends on 1) the type of medical data, 2) extraction of the object and the following shape representation by a model, 3) selecting object properties for the shape model, 4) statistics necessary to perform population comparison of the models and, 5) a

method for constructing a hypothesis test based on given difference measures.

We use a discrete *skeletal representation*, abbreviated as s-rep [36] as a shape model. The amenities of s-reps relative to other shape representations are described in Section 3. The s-reps are fit to a set of binary images of the hippocampus extracted from a magnetic resonance imaging (MRI) data set. All skeletal shape models have Euclidean as well as non-Euclidean components. Thus, a hypothesis test based on skeletal models must support the two different types of features. The approach presented in this paper allows a sensitive hypothesis test between the components of s-reps. By using this approach, local and global shape differences of the hippocampi between schizophrenia and control populations are investigated.

The hypothesis test requires *i)* fair correspondence between all skeletal models within a population and across populations, *ii)* a method to compute means of populations of skeletal models, *iii)* a test statistic with appropriate distance measures for the Euclidean and non-Euclidean components of the means, *iv)* a method to calculate a test statistic and the empirical distribution of the test statistic, and *v)* a procedure to correct for multiple comparison of local and global testing of GOPs.

The paper is presented as follows. The data set for the schizophrenia study describing two shape populations is presented in Section 2. The skeletal model is discussed in Section 3 in addition to required statistical properties for shape analysis of populations. Section 4 introduces the method composite principal nested great spheres (CPNG), which allows statistical analysis of the Euclidean and non-Euclidean components of skeletal models such as the calculation of means. Section 5 describes the model fitting procedure for the two shape populations, which produces the statistical properties required for each model. A permutation test is introduced in Section 6 and specified for skeletal models together with required statistics. Finally in Section 7, hypothesis test results of the hippocampus study are reported.

2 Schizophrenia study data set

The data consist of MRI assessments of hippocampi from patients with schizophrenia and a similar set from a healthy control group as described in [28,40]. In the original study, 238 first-episode schizophrenics and 56 controls were enrolled. First-episode schizophrenia patients have not received medical treatment prior to the MRI assessment. The hippocampi were segmented from the aligned MRI scans with an automated atlas based segmentation tool developed at the University of North Carolina [16,28].

Statistical analysis must be performed on either the left or right hippocampus as a combined analysis could bias the result. Accordingly, the left hippocampus is evaluated in this paper. Records of the the left hippocampus were not available for 17 patients from the schizophrenia group. Therewith, the data set consists of 221 first-episode schizophrenia cases (**SG**) and 56 control cases (**CG**) and is represented by binary files which reflect the segmented hippocampi. In the data provided, the hippocampi have been normalized in volume but the original volumes were reported as separate scaling features.

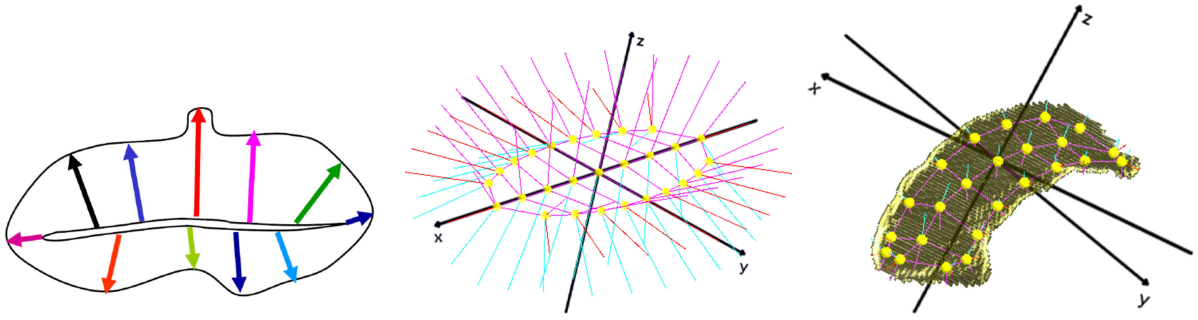
3 Object representation

The representation by a shape model allows calculation of shape statistics of the hippocampus. The type of model, chosen to compare two shape populations, should capture a rich collection of GOPs presented in the data. In addition, small deformations in objects should be reflected by small deformations in the models. Finally, the model should not introduce artificial variation across a population which is not present in the objects themselves.

As discussed by [36], a model that fulfills these three properties is an interior-filling s-rep as depicted in Figure 1. Starting with the continuous case, this model and its properties will be discussed in the following.

The desired GOPs of the model can be categorized into three groups. The first group (G1) should capture locational information of the object boundary. The second group (G2) should reflect the local surface curvature by incorporating directional information into the shape model. The shape model should accurately depict the local orientation of the object. The third group (G3) should describe how the object boundary is connected by the interior in order to reflect the relationship across the interior of the object. The thickness of an object is one property of the interior among others. *Skeletal models* are designed to obtain these geometric properties.

The family of skeletal models has been widely studied in computer vision and medical image analysis. In Section 3 of [41] it is shown that the medial locus [4] of an object $\Omega \subseteq \mathbb{R}^n$ can be described by an inward "grassfire" that starts at the boundary and dies out at a folded version of the medial locus called M_Ω. Given a folded medial locus M_Ω, the medial representation of an object Ω is determined by a set of *spoke* directions from points of M_Ω to the corresponding points of tangency on the boundary $\partial\Omega$. The collection of spoke end points capture locational information of the object boundary as postulated in (G1). The second group (G2) is captured by the directional information



(a) skeletal representation (b) ellipsoid (c) hippocampus

Fig. 1: Continuous skeletal representation and fitted s-reps. (a) Sectional view of a two-sided skeletal 3D object representation. Colored spokes emanate from the skeletal sheet (which is not medial) to the surface. In the continuous form there is a spoke on each point on the skeletal sheet. (b) Discrete s-rep of a non-deformed 3D ellipsoid. (c) Discrete s-rep of a hippocampus.

of the spokes. The points of M_Ω describe the inherent symmetry of an object and therewith (G3) above.

Strictly medial representations are limited by the fact that every protruding boundary kink results in additional medial branches. Thus, two versions of the same object with small noise can have drastically different medial representations. Skeletal models achieve additional stability by relaxing the medial constraint.

Figure 1a visualizes a sectional view of a two-sided skeletal object representation in $\mathbb{R}^3$ composed of a *skeletal sheet* and spokes which emanate from a *skeletal position* on the skeletal sheet to the surface. The skeletal sheet is close to midway but is not medial. An exactly medial representation of this object would require the set M_Ω to include an additional long branch. Elimination of such branches in M_Ω is the goal of the skeletal representation. Figures 1b and 1c will be discussed later.

Stability in the branching structure and stability in the skeletal sheet ensure structural case-by-case stability of the model and thus good correspondence across the samples in the full data set. The branching constraint can be tightened for specific classes of objects where the shape is known. For an ellipsoid-like object shape, such as the hippocampus, the constraint of no branching is reasonable and is adopted. Yet, we want to retain as much as possible the medial properties, such as orthogonal spokes to the boundary, as equal as possible skeletal positions and approximately equal spoke length on both sides of the skeletal sheet. Therefore, the family of skeletal models is restricted by the class of interior-filling s-reps that are modeled as medial as possible [36].

In addition to the case-by-case stability, we require population stability to avoid artificial variance across a population that is solely an artifact of the individual s-rep fittings; such variance is not connected to the objects themselves. Population stability can be achieved by a re-fitting step of the s-rep to the object using an estimated shape probability distribution of the population. The re-fitting step reduces the variance of the s-rep population as described in Section 5. Both case-by-case stability and population stability ensure that the shape models have improved correspondence of both the spokes and the skeletal locations between objects, which will support accurate statistics across a population.

A discrete s-rep, as required for the numerical analysis of slab-shaped objects, consists of a two-sided (folded) sheet of skeletal positions sampled as a grid of atoms, whose skeletal positions are depicted as small spheres in Figures 1b and 1c. On each side of the sheet, there is a spoke, a vector with direction and length on the top and on the bottom connecting the skeletal sheet to the boundary. Also, for each edge grid point there is an additional spoke vector connecting the skeletal sheet folded to the crest of the slab. The sheet is close to midway consistent to the fixed branching constraint between the two sides of the slab, and the spokes are approximately orthogonal to the object boundary. Each discrete s-rep is described by a feature vector

$$\mathbf{s} = (p_1, \ldots, p_{n_a}, r_1, \ldots, r_{n_s}, u_1, \ldots, u_{n_s}) \tag{1}$$

with $n_a = n_a^{ext} + n_a^{int}$ the number of atoms and $n_s = 3n_a^{ext} + 2n_a^{int}$ is the number of radii and spoke directions. A slabular s-rep consists of n_a^{ext} exterior (edge grid points) and n_a^{int} interior atoms. An interior atom consists of a skeletal position $p \in \mathbb{R}^3$, two spoke directions $u \in S^2$ and two spoke lengths $r \in \mathbb{R}_+$ (top, bottom) where $S^2 = \{x \in \mathbb{R}^3 \mid \|x\| = 1\}$ is the unit sphere. An exterior atom consists of a skeletal position $p \in \mathbb{R}^3$, three spoke directions $u \in S^2$ and three spoke lengths $r \in \mathbb{R}_+$ (top, crest, bottom). As a result, the shape space of $\mathbf{s} \in \mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s} \times (S^2)^{n_s}$ is a product of Euclidean and non-Euclidean spaces. Each s-rep can

also be described in the space $\mathbb{R}^{n_s+1} \times S^{3n_a-4} \times (S^2)^{n_s}$ together with a scaling factor $\gamma \in \mathbb{R}_+$. This representation is derived from a pre-shape space [22] as discussed in Section 4.

Another popular class of modeling 3D objects is a boundary point distribution model (PDM) where a solid object is defined by the positions of the sampled surface points [8,10,24]. In general, normal directions can be attached on the surface points of a PDM but to the best of our knowledge, it has not been used in practice. In addition, deformation-of-atlas models are well known, wherein the shape changes of an object in images are modeled by the deformations of a template image [34,38]. Such models can capture the local orientation of an object. Nevertheless, both approaches are less suitable for shape statistics of populations by the lack of the interior description of an object.

We restrict our analysis to discrete slabular s-reps which are organized into a (3×8) grid of skeletal positions, i.e., each s-rep consists of 24 atoms. The choice of the grid size defines how exactly the binary images can be described by the s-rep model. We have chosen a grid of 3×8 atoms as a trade-off between capturing important object features, avoiding an overfitting and keeping the dimension of the shape space low. A hippocampus example with bumps that are not tightly described by a (3×8) grid is visualized in Section 1.1 of the Supplementary Material. However, we do not look at individual s-reps that may not be perfectly correct but rather at differences between groups which are not biased versus the other.

4 CPNG analysis

A hypothesis test on mean differences requires a method to calculate means from populations of shape models. The method should incorporate all geometrical components of such models. We have presented in Section 3 an s-rep as a suitable model with Euclidean components and components which live on spheres. This section will discuss an approach to produce means, in addition to shape distributions of populations of s-reps.

First of all, we need to understand the shape space of a discrete s-rep to apply a proper statistical analysis. Each discrete s-rep is described by a feature vector $\mathbf{s} \in \mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s} \times (S^2)^{n_s}$ as defined in (1), and lives in a product of Euclidean and non-Euclidean spaces. Each element of $\mathbf{s}$ corresponds across the population. The points $X_p = (p_1, \ldots, p_{n_a})' \in \mathbb{R}^{3n_a}$ form an $(n_a \times 3)$ matrix and a PDM that can be centered and normalized at the origin by $Z_H = HX_p/\|HX_p\|$ with H a Helmert sub-matrix which removes the origin [10,22]. H is an $((n_a - 1) \times n_a)$ matrix with row $i-1$ defined by the vector

$$(H)_{i-1} = (d_i, \ldots, d_i, -id_i, 0, \ldots, 0)$$

with $d_i = (i(i+1))^{-\frac{1}{2}}, i = 2, \ldots, n_a$ where d_i is repeated i-times. Z_H is called a pre-shape with information of location and scale removed. Therewith, the Cartesian product of $p_i \in \mathbb{R}^3, i = 1, \ldots, n_a$ can be described by the pre-shape Z_H and by a scaling term $\gamma = \|HX_p\|$. The pre-shape Z_H lives on the $(3n_a - 4)$ dimensional unit sphere $S^{3n_a-4} \subset \mathbb{R}^{3n_a-3}$. Each spoke direction $u_i, i = 1, \ldots, n_s$ of $\mathbf{s}$ lives on the unit sphere S^2. The radii $r_i \in \mathbb{R}_+, i = 1, \ldots, n_s$ and scale factor $\gamma \in \mathbb{R}_+$ are log-transformed to the Euclidean space $\mathbb{R}$. Thus, a discrete s-rep $\mathbf{s}$ can be described in the shape space $\mathbb{R}^{n_s+1} \times S^{3n_a-4} \times (S^2)^{n_s}$ composed of several spheres and a real space. Jung et al. [20] and Pizer et al. [36] have proposed a method to analyze a population of s-reps that are living in such an abstract space. This method is called *composite principal nested spheres* (CPNS) and will be discussed in the following.

Suppose we have a population of N s-reps. In order to analyze the covariance structure of such a population, we have to find a common coordinate system. CPNS consists of two main parts. First, the spherical parts are analyzed by *principal nested spheres* (PNS) [18,19], which analyzes data on spheres in decreasing dimension, i.e., using a *backward view*. Therewith, the pre-shape Z_{Hj} and each u_{ij} can be mapped to a Euclidean space with $j = 1, \ldots, N$. Second, the Euclideanized variables are concatenated with the $\log r_i$ and $\log \gamma$ to give a matrix Z_{comp} and an array of scale factors to make all variables commensurate as discussed in detail in [36]. Finally, the structure of the covariance is investigated from the scaled matrix Z_{comp}. PNS is a novel method to estimate the joint probability distribution of data on a d-dimensional sphere S^d by a backward view along the dimensions. The backward view allows dealing with one dimension at a time and thus produces better probability distributions.

In Euclidean space, the forward and backward approaches to *principal component analysis* (PCA) are equivalent, which is not true in general non-Euclidean spaces, such as the d-dimensional unit sphere S^d. Damon and Marron [9] have studied generalizations (e.g., PNS) of PCA across a variety of contexts, and have shown that backwards is generally more amenable to analysis, because it is equivalent to a simple adding of constraints.

PNS is a fully backward approach that fits the best lower dimensional subsphere in each dimension starting with S^d. The subsphere can be great (a sphere with radius 1) or small (less than 1). Figure 2 visualizes the method which takes into account variation along small circles (non-geodesic variations) as well as variation along geodesics. Thus, the decomposition

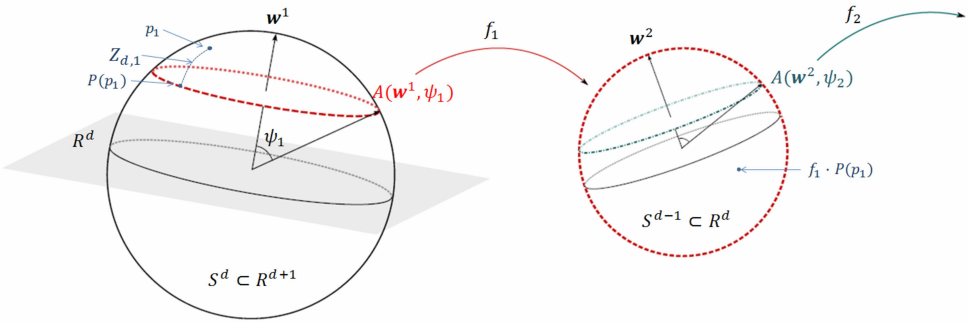


Fig. 2: Backwards PNS by computation of the nearest lower dimensional subsphere. In addition, the projection $P(p_1)$ of a data point p_1 with score $Z_{d,1}$ on the subsphere $A(w^1, psi_1)$ is depicted.

is non-geodesic. Each subsphere $A(w^l, \psi_l)$ of S^{d-l+1} is defined by an axis (polar position) w^l with latitude angle $\psi_l, l = 1, \ldots, d$. The lowest principal nested sphere $A(w^d, \psi_d)$ is a circle. The Fréchet mean [13,21] on the remaining 1-dimensional subsphere can be seen as the best-fitting 0-dimensional subsphere (a point) of the data. Finally, this mean is projected back to S^d resulting in a backward Fréchet mean.

CPNS has been shown to be powerful in the analysis of a single s-rep population. In this study, our hypothesis test of local or global shape differences involves the comparison of two or more CPNS statistics, which is facilitated by stable statistics and correspondence between populations, using a common coordinate system. In fact, PNS is a non-geodesic method which fits small and great subspheres. The fitting of small spheres has advantages in describing the amount of data variation inside a population with fewer principal components but two or more populations can have different decompositions into small and great spheres, which introduce additional variation across the populations, e.g., reflected by a larger variation between CPNS means of several populations. Therefore, all CPNS analyses in this paper are constrained by fitting principal nested great spheres, called PNG and CPNG respectively. Therewith, we avoid additional variation, ensure correspondence and a common coordinate system across several populations. A preliminary simulation study on permuted populations of s-reps confirmed the improvements of CPNG compared to CPNS. Notice, CPNG is identical to [17] in the two dimensional case. We are leaving a commensurate CPNS analysis for populations using small spheres for future work and discuss a possible approach in Section 1.2 of the Supplementary Material.

5 Model fitting

The application of the proposed hypothesis test to the hippocampi study, introduced in Section 2, requires a procedure to generate s-rep fittings with statistical object and population properties as discussed in Section 3. This section will introduce such a procedure.

Assume we have a set of binary images and a set of corresponding signed distance images. The distance images are used during the fitting process as the target data to which reference models are fitted. Following [36], the fitting procedure can be described with 5 consecutive steps: initial alignment, atom stage, spoke stage, CPNG stage and the final spoke stage.

Initial alignment: A reference s-rep is translated, rotated and scaled into the space of distance images by matching of moments to the boundary.

Atom stage: The atom stage defines the geometry of the object and accordingly, the case-by-case stability. Each atom, i.e., each skeletal grid point and its set of spokes are fit one by one with multiple iterations through these atoms. For each atom, an objective function is optimized [36]. The objective function reflects the goodness of the fit and is calculated by a weighted sum of different optimization criteria. The function penalizes factors which are making the s-rep structurally improper, such as irregularity in the grid and crossing of adjacent spokes. In addition, it penalizes the spoke ends deviating from the object boundary and their directions deviating from the boundary normal (both implied by the input distance image). The spokes are further penalized from failing to match the geometry of the crest implied by the distance image. The penalties are summed over spokes which are interpolated from the original s-rep.

Spoke stage: The spoke stage optimizes the spoke lengths to match the object boundary more closely. The skeletal grid points and the resulting geometry of the s-rep will not be changed during this stage. The atom and spoke stage provide appropriate s-rep fittings to the data with case-by-case stability.

CPNG stage: The CPNG stage is designed to provide improved correspondence across a population of s-reps. The fits of the spoke stage are used to calculate CPNG statistics as described in Section 4. Improved correspondence is achieved by restricting the fits to a

shape space which results from the CPNG analysis. CPNG estimates a mean s-rep from the population but also yields eigenmodes and modes of variation [18]. Consequently, any s-rep can be expressed by the score of the eigenmodes in the CPNG space. Hence, correspondence across a population is achieved by initialization of each fitting with the CPNG mean s-rep and by restricting the shape space to the CPNG space. In addition, s-rep candidates from the CPNG space are penalized by the Mahalanobis distance between the candidate and the CPNG mean. As a result improved correspondence is achieved.

Final spoke stage: The final spoke stage adjusts the spokes of the CPNG stage fits to match closely the object boundary. Consequently, s-reps can be generated which are not an element of the CPNG space.

The first three stages form a *preliminary stage* in the fitting procedure. The fitting stages are implemented in a software called Pablo, developed at the University of North Carolina. The program is available at [32].

6 Multiple hypothesis testing

A sensitive hypothesis test is useful for the quantification of shape differences, both to compare populations globally and locally. The introduction of a suitable shape model in Section 3, a method to calculate means from populations in Section 4 and a procedure to generate s-rep fittings in Section 5 provide us with tools to generate models and means that contain the desired properties for a sensitive hypothesis test. An important challenge is that the geometric object elements of each model are spatially correlated. Furthermore, a suitable hypothesis test should correct for multiple comparisons.

6.1 An overview of multiple comparison corrections

The problem of false positives with multiple statistical tests is well recognized. Statistical shape analysis must deal with a large number of hypotheses, each derived from a GOP element, for example of the s-rep. Two common categories of multiple comparison correction are familywise error rate (FWER) and false discovery rate (FDR) [3]. Let V be the number of rejected hypotheses when the null is true (type 1 error), and S the number rejected hypotheses when the null is false. The FWER is defined as the probability of at least one type 1 error by $\mathrm{P}(V \geq 1)$. The FDR is defined as the expected proportion of type 1 errors among the total number of rejected hypotheses by $\mathrm{E}(V/(S+V)$ with $V/(S+V) = 0$ if $(S+V) = 0$. There are several approaches to control FWER and FDR. A commonly used one is the Bonferroni correction. Another approach is using typical wavelet coefficient selection methods [1,6,44]. In addition, variable selection based on threshold random field theory (RFT) have been used [7,23,33]. Permutation tests allow multiple comparison correction by estimating the empirical null-distribution and the covariance structure of the test statistics [30,35,43]. This paper uses multiple comparison correction by FWER.

The Bonferroni correction has several major drawbacks; the Bonferroni threshold can be conservative if the GOPs are dependent of each other. In particular, spatial autocorrelations result in fewer effective variables. Spatial correlation can be expected between neighbor spokes and skeletal positions of an s-rep. In addition, the Bonferroni correction reduces the power of a test as the probability of false negatives increases, because it controls only the probability of false positives. RFT requires strong assumptions such as the same parametric distribution at each spatial location (e.g., multivariate Gaussian), sufficient smoothness as well as stationarity. The assumption of a parametric distribution can not be fulfilled in case of an s-rep model and the assumption of stationarity can also be doubtful.

Permutation tests have advantages over the approaches above that make them particularly suitable for s-reps. S-reps are defined on a product of Euclidean and non-Euclidean spaces with unknown probability distributions of the geometric object elements. In contrast to standard parametric methods such as Bonferroni and RFT, a permutation test is a non-parametric approach using the data to estimate the sampling distribution of the test statistic under the null-hypothesis H_0. Permutation tests are also adaptive to underlying correlation patterns in the data.

A minimal assumption of permutation testing is the exchangeability under H_0 such as identical distributions of populations 1 and 2. The underlying idea of a permutation test is that any permutation of the observations has the same probability to occur under the assumption H_0. Given the permuted populations, a common test statistic measures differences between population means. The test statistic may calculate feature by feature differences or combine features to measure differences between GOPs. The permuted populations can be used to estimate the distribution of the test statistic as well as to estimate the correlation structure.

6.2 A permutation test for s-reps

Suppose we have two populations of s-reps described by a set $\tilde{A}_1 = \{\tilde{\mathbf{s}}_{11}, \ldots, \tilde{\mathbf{s}}_{1N_1}\}$ of N_1 s-reps and a set $\tilde{A}_2 = \{\tilde{\mathbf{s}}_{21}, \ldots, \tilde{\mathbf{s}}_{2N_2}\}$ of N_2 s-reps with $\tilde{\mathbf{s}}_{il}$ as defined in (1). We assume without loss of generality $N_1 \geq N_2$.

The permutation test for populations of s-reps can be divided into four steps.

First, observed and permuted population CPNG means are generated as described in Section 6.2.2. Second, appropriate Euclidean or non-Euclidean GOP differences are calculated between the means of the observed populations $\tilde{A}_1$ and $\tilde{A}_2$, and between the means of corresponding permuted populations as described in Sections 6.2.3 and 6.2.4. Third, p-values are calculated for each GOP difference as described in Section 6.2.5. Each of these p-value is uniformly distributed and mapped by probability integral transformations to standard normal distributed variables. Hence, the GOPs can be mapped from a non-linear to a linear space with the same coordinate system for each GOP. Finally, the covariance matrix of these standard normal distributed variables is estimated, in order to incorporate the true multivariate nature of the data and the correlation between the GOPs as described in Section 6.2.6. As a result, the partial tests of the GOPs can be combined into a single summary statistic by the Mahalanobis distance. In addition, a feature-by-feature test can be constructed as described in Section 6.2.7.

6.2.1 Pre-processing

In a first pre-processing step, global translational and rotational variations should be removed from all s-reps in order to analyze only shape variations. To make the alignment unbiased with respect to the population, the overall backwards CPNG mean $\tilde{\mu}$ is estimated from the set union

$$\tilde{A} = \tilde{A}_1 \bigcup \tilde{A}_2 = \{\tilde{\mathbf{s}}_{11}, \ldots, \tilde{\mathbf{s}}_{1N_1}, \tilde{\mathbf{s}}_{21}, \ldots, \tilde{\mathbf{s}}_{2N_2}\}.$$

The CPNG mean $\tilde{\mu}$ is translationally aligned by the subtraction of the mean of the locational components. In addition, the eigenvectors of the second moments about the center of the skeletal positions yields a rotational alignment to the x, y and z-axis. The translationally and rotationally aligned CPNG mean $\tilde{\mu}$ is called μ. Afterwards, each s-rep $\tilde{\mathbf{s}} \in \tilde{A}$ is translated, rotated and scaled to μ by standard Procrustes alignment (see [10]) based on the hub-positions of each s-rep. For each aligned s-rep $\mathbf{s}$, the scaling factor $\tau \in \mathbb{R}_+$ is kept as a variable. The global translation and rotation information is not considered of interest in the shape analysis of hippocampi. Moreover, we have chosen to use features which can be understood by the user (e.g., physicians). Therefore, the skeletal positions are considered in $\mathbb{R}^{3n_a}$ instead on S^{3n_a-4} as in Section 4. Thus, each aligned s-rep is described by a feature vector $\mathbf{t} = (\tau, \mathbf{s})$, where $\mathbf{t}$ contains $n = 1 + n_a + 2n_s$ features and is an element of the shape space $\mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times (S^2)^{n_s}$. Set $A_1 = \{\mathbf{t}_{11}, \ldots, \mathbf{t}_{1N_1}\}$, $A_2 = \{\mathbf{t}_{21}, \ldots, \mathbf{t}_{2N_2}\}$ and $A = A_1 \bigcup A_2$.

6.2.2 Generation of observed and permuted sample means

First, a method to calculate means for the observed and permuted samples of the two populations is required in order to create a hypothesis test of mean differences.

Observed sample means. For each set A_i, $i = 1, 2$ the observed sample mean is $\hat{\mu}_i = (\bar{\tau}_i, \bar{\mu}_i) \in \mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times (S^2)^{n_s}$. The component $\bar{\mu}_i$ is a CPNG backwards mean as described in Section 4. The mean scaling factor $\bar{\tau}_i \in \mathbb{R}_+$ is computed as a geometric mean (which is natural for scaling factors) by

$$\bar{\tau}_i = \exp\left(\frac{1}{N_i}\sum_{j=1}^{N_i} \log(\tau_{ij})\right), \qquad i = 1, 2. \tag{2}$$

In fact, the CPNG backwards mean $\bar{\mu}_i$ consists of n_s+1 PNG backwards means, one for the skeletal position and n_s for the spoke directions, and n_s means for the spoke lengths respective to (2).

Permuted sample means. The number of all possible permutations of the index set $I = \{1, \ldots, N_1+N_2\}$ is

$$\binom{N_1 + N_2}{N_1} = \frac{(N_1 + N_2)!}{N_1!N_2!}.$$

Random sample sets $I_l, l = 1, \ldots, P$ of $P = 30,000$ permutations of the index set I were generated, a number comparable to the suggested number in [11] and [26]. Larger numbers of permutations increase the accuracy of the p-values but require more computation time. The permutation group $A_{1l} \subset A$ contains all s-reps indexed by the first N_1 indices of I_l. The group $A_{2l} = A \setminus A_{1l}$ contains the remaining N_2 s-reps. For each permutation I_l the means $\hat{\nu}_{il}$, $i = 1, 2$ are estimated by $\hat{\nu}_{1l} = (\bar{\kappa}_{1l}, \bar{\nu}_{1l})$ for the group A_1, $\hat{\nu}_{2l} = (\bar{\kappa}_{2l}, \bar{\nu}_{2l})$ for the group A_2. $\bar{\nu}_{il}$ is estimated by the CPNG backwards mean and $\bar{\kappa}_{il}$ is the mean scaling factor of the corresponding permutation respective to (2).

6.2.3 Test statistics

Equality of distributions between populations A_1 and A_2 can be tested by a nonparametric combination of a finite number of dependent *partial tests* as proposed in Pesarin [35]. The global null hypothesis is given by $H_0 : \{A_1 \stackrel{d}{=} A_2\}$, where $\stackrel{d}{=}$ denotes the equality in distribution. Let H_1 be the global alternative hypothesis. In general, the test requires the definition of a *statistic* T in testing a null hypothesis. A natural test statistic is

$$T(A_1, A_2) = d(\hat{\mu}_1, \hat{\mu}_2), \tag{3}$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the observed sample means as defined in Section 6.2.2 and d is a difference measure on the nonlinear manifold describing the GOPs. The test statistic T consists of K different partial tests depending on the difference measure. Thus, the global null hypothesis can be written in terms of K subhypotheses $H_0 : \{\bigcap_{k=1}^{K} H_{0k}\}$ and the alternative as $H_1 : \{\bigcup_{k=1}^{K} H_{1k}\}$. Usually, the dependence relation among partial tests are unknown even though they are functions of the same data. Pesarin [35] has shown that a suitable combining function (described in Section 6.2.6) will produce an unbiased test for the global hypothesis H_0 against H_1 if all partial tests are assumed to be marginally unbiased, consistent and significant for large values. The partial tests $T_k, k = 1, \ldots, K$ are defined by the *partial difference measures*. Therewith, a hypothesis test for identical statistical distribution of two s-rep populations is given by mean differences,

$$H_0 : \{\mu_1 = \mu_2\} \text{ versus } H_1 \ : \ \{\mu_1 > \mu_2\} \tag{4}$$

for a one-sided test in case the difference measures are unsigned and

$$H_0 : \{\mu_1 = \mu_2\} \text{ versus } H_1 \ : \ \{\mu_1 \neq \mu_2\} \tag{5}$$

for a two-sided test in case of signed differences.

The hypothesis H_0 will be rejected if the probability of observing $T(A_1, A_2)$ under H_0 from the empirical distribution is smaller than a chosen significance level α; otherwise we do not reject. The significance level describes the probability of type 1 error, i.e., H_0 is wrongly rejected. Alternatively, the type 2 error occurs when H_0 is not rejected but it is in fact false.

6.2.4 Difference measures

This section defines a signed difference measure d^2 for the test statistic (3). An alternative unsigned difference measure d^1 is defined in Section 1.3 of the Supplementary Material. Suppose we have two s-reps

$$\mathbf{t}_i = (\tau_i, p_{i1}, \ldots, p_{in_a}, r_{i1}, \ldots, r_{in_s}, u_{i1}, \ldots, u_{in_s})',$$

$i = 1, 2$ with the skeletal positions $p_{ij}, \in \mathbb{R}^3$ and the scale factors $\log(\tau_i), \log(r_{ij}) \in \mathbb{R}$ as Euclidean GOPs and the spoke directions $u_{ij} \in S^2$ as non-Euclidean GOPs. Thus, a suitable difference measure is required as defined in the following.

The measure d^2 is a vector of differences

$$\begin{aligned} d^2(\mathbf{t}_1, \mathbf{t}_2) := (&d_1(\tau_1, \tau_2), \\ &d_2(p_{11}, p_{21}), \ldots, d_2(p_{1n_a}, p_{2n_a}), \\ &d_3(r_{11}, r_{21}), \ldots, d_3(r_{1n_s}, r_{2n_s}), \\ &d_4(u_{11}, u_{21}), \ldots, d_4(u_{1n_s}, u_{2n_s}))' \end{aligned} \tag{6}$$

with appropriate partial difference measures: d_1 for the scaling factors τ_i, d_2 for the positions p_{ik}, d_3 for the spoke lengths r_{ij} and d_4 for the spoke directions u_{ij} with $i = 1, 2$, $k = 1, \ldots, n_a$ and $j = 1, \ldots, n_s$ by

$$\begin{aligned} d_1(\tau_1, \tau_2) &= \log(\tau_2) - \log(\tau_1), \\ d_2(p_{1k}, p_{2k}) &= p_{2k} - p_{1k}, \\ d_3(r_{1j}, r_{2j}) &= \log(r_{2j}) - \log(r_{1j}), \\ d_4(u_{1j}, u_{2j}) &= d_{gs}(u_{1j}, u_{2j}). \end{aligned}$$

The partial difference measure d_{gs} is defined by longitude and latitude differences of the spoke directions (u_{1j}, u_{2j}) using a normalization by the shift of the geodesic mean as explained in the next paragraph. The components of

$$\begin{aligned} d^2 : &(\mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times (S^2)^{n_s}) \times \\ &(\mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times (S^2)^{n_s}) \longrightarrow \mathbb{R}^{3n_a+3n_s+1} \end{aligned}$$

are not metrics because they can take on negative values.

Shift by the geodesic mean. The spoke directions $(u_{1j}, u_{2j}) \in S^2 \times S^2$ can be mapped by spherical parametrization to latitudes ϕ_{1j}, ϕ_{2j} and longitudes θ_{1j}, θ_{2j} in the base coordinate system of all aligned hippocampi. The spherical mapping can be defined by

$$\begin{aligned} \phi_{ij}(u_{ij}) &= \text{atan2}(u_{ij3}, \sqrt{u_{ij1}^2 + u_{ij2}^2}), \\ \theta_{ij}(u_{ij}) &= \text{atan2}(u_{ij2}^2, u_{ij1}^2), \end{aligned}$$

with $\phi_{ij} \in [-\pi/2, \pi/2]$ and $\theta_{ij} \in (-\pi, \pi]$; the two-argument function $\text{atan2}(x_2, x_1) \in (-\pi, \pi]$ is the signed angle between two vectors $e_1 = (1, 0)'$ and $(x_1, x_2)' \in \mathbb{R}^2$. The longitude ϕ is measured from the x-y plane.

The spherical mapping is not uniquely defined in general. Furthermore, it does not establish an appropriate correspondence. Two points close to the equator with identical geodesic distance as two points close to the north pole have different latitude and longitude differences, and are therefore not commensurate. For that reason, longitude and latitude pair differences will be normalized by shifting the geodesic mean of (u_{1j}, u_{2j}) along its meridian to the equator by a rotation about an axis $c \in S^2$ with rotation angle $\psi \in [0, \pi/2)$. Then, the directions (u_{1j}, u_{2j}) are rotated along small circles on the sphere about the same axis c with the same rotation angle ψ towards the equator.

In more detail, consider a pair (u_{1j}, u_{2j}) of spoke directions on S^2 with northpole $N_p = (0, 0, 1)'$. At first, find its geodesic mean by

$$\mu_g(u_{1j}, u_{2j}) = \frac{u_{1j} + u_{2j}}{\|u_{1j} + u_{2j}\|}.$$

We assume $\text{acos}(|\mu_g' N_p|)) > 1e - 3$; otherwise choose a different northpole. Given a rotation matrix $R_1 :=$

$R_1(c,\psi)$, the rotation of μ_g along its meridian to the equator is $\tilde{\mu}_g = R_1\mu_g$ with

$$R_1(c,\psi) = I_3 + \sin\psi[c]_\times + (1-\cos\psi)(cc' - I_3), \quad (7)$$

where I_3 is the three-dimensional unit matrix and $[c]_\times$ is the cross product matrix satisfying $[c]_\times v = c \times v$ for any $v \in \mathbb{R}^3$. To avoid discontinuity problems between $-\pi$ and π for θ_{ij}, let R_2 be the rotation matrix as defined by (7) that rotates $\tilde{\mu}_g$ towards $(1,0,0)'$, i.e., $R_2\tilde{\mu}_g = (1,0,0)'$. Now, shift each pair (u_{1j}, u_{2j}) by applying $\tilde{u}_{1j} = R_2R_1u_{1j}$ and $\tilde{u}_{2j} = R_2R_1u_{2j}$. Finally, we calculate the latitudes $\phi_{1j}(\tilde{u}_{1j}), \phi_{2j}(\tilde{u}_{2j})$ and longitudes $\theta_{1j}(\tilde{u}_{1j}), \theta_{2j}(\tilde{u}_{2j})$ and define the differences of the transported spoke directions by the delta latitude $\Delta\phi_j = \phi_{2j} - \phi_{1j}$ and delta longitude $\Delta\theta_j = \theta_{2j} - \theta_{1j}$. Therewith, the difference measure d_{gs} is defined by

$$d_{gs}(u_{1j}, u_{2j}) := (\Delta\phi_j, \Delta\theta_j).$$

6.2.5 Mapping of GOP differences to standard normally distributed variables

Suppose we have the test statistic $T_0 := T(A_1, A_2)$ of the underlying observed sample. The idea is to estimate the sampling distribution of the statistic T_0 from test statistics of the permuted samples

$$T_l := T(A_{1l}, A_{2l}), \quad l = 1, \ldots, P.$$

The test statistic measures the GOP differences in different units. The vector $T_l = (T_{l1}, \ldots, T_{lK})$ contains K partial tests, where K is the number of components of the difference measure d^2. The elements of the vector T_l are not commensurate as required for the estimation of the covariance structure. Thus, the GOP differences must be normalized and mapped to a common coordinate system in a way that preserves the multivariate dependence structure between the GOPs. The procedure is explained in the following and depicted in Figure 3 on the basis of a selected GOP using distance measure d^2. The figure is discussed further in the text.

Calculating p-values for GOP differences. After the calculation of T_l, we estimate for each GOP difference $k = 1, \ldots, K$ the empirical cumulative distribution function (CDF) by

$$C_k(T_{lk}) = \frac{1}{P}\sum_{l'=1}^{P} I(T_{l'k}, T_{lk})$$

$$\text{with } I(T_{l'k}, T_{lk}) = \begin{cases} 1 & : \quad T_{l'k} \le T_{lk}, \\ 0 & : \quad \text{otherwise.} \end{cases}$$

Respectively, we can calculate $C_k(T_{0k})$.

Mapping of p-values to $\mathcal{N}(0,1)$. By construction the p-values have a uniform distribution. Thus, the GOP differences can be represented as

$$U_{lk} = \Phi^{-1}\left(\tilde{C}_k(T_{lk})\right), \quad (8)$$

where Φ^{-1} is the inverse standard Gaussian CDF,

$$\tilde{C}_k(T_{lk}) = \frac{sc-2}{sc}C_k(T_{lk}) + \frac{1}{sc}$$

and $k = 1, \ldots, K$, $l = 1, \ldots, P$. The inverse standard Gaussian CDF requires values greater than 0 or less than 1; otherwise $U_{lk} = \pm\infty$. Therefore, all p-values are scaled by $\tilde{C}_k(T_{lk})$ with $sc = 10000$. Simulations have shown numerical instabilities for larger values of sc. The marginal distribution of U_{lk} is standard Gaussian for every k, i.e, $U_{lk} \sim \mathcal{N}(0,1)$.

Using the estimated inverse empirical CDF C_k, the observed GOP differences T_{0k} are mapped to U_{0k}, respectively.

6.2.6 Global test with multivariate comparisons correction

Given $U_{lk} \sim \mathcal{N}(0,1)$, the $K \times K$ covariance matrix Σ_U of the $P \times K$ matrix $U = (U_1, \ldots, U_P)'$ with $U_l = (U_{l1}, \ldots, U_{lK})$, $l = 1, \ldots, P$ is estimated by

$$\hat{\Sigma}_U = \frac{1}{K-1}U'U.$$

A corrected test statistic is then given by the Mahalanobis distances

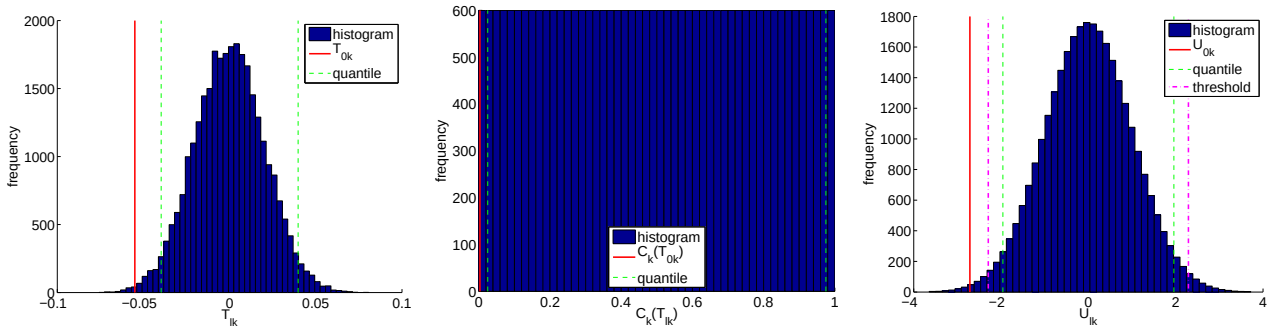
$$M_0 = U_0'\hat{\Sigma}_U^{-1}U_0, \quad M_l = U_l'\hat{\Sigma}_U^{-1}U_l, \quad l = 1, \ldots, P,$$

which defines a suitable combining function [35, Section 6.2.4] that includes the GOP correlation structure. The sampling distribution of the final test statistic under the null-hypothesis H_0 can be estimated from M_l by an empirical CDF. The probability of observing M_0 under H_0 from the empirical null-distribution is given by

$$p(M_0) = \frac{1}{P}\sum_{l=1}^{P} H(M_l, M_0), \quad (9)$$

$$\text{with } H(M_l, M_0) = \begin{cases} 1 & : \quad M_l \ge M_0, \\ 0 & : \quad M_l < M_0. \end{cases}$$

Equation (9) defines the p-value of the final global test by rejecting H_0 if $p(M_0) < \alpha$.



(a) GOP difference (b) p-value (c) std with corrected threshold

Fig. 3: Mapping of GOP differences to standard normally distributed variables using the example of a skeletal z-position ($k = 3$) with distance measure d^2. The 2.5% and 97.5% quantiles are visualized in each plot. (a) GOP differences T_{lk} of the permuted samples with $l = 1, \ldots, 30,000$ and T_{0k} of the underlying observed sample. (b) Calculated p-values $C_k(T_{lk})$ and $C_k(T_{0k})$ using the empirical cumulative distribution function. (c) Standard normal distributed variables U_{lk} and U_{0k}. The dotted-dashed line depict the corrected threshold λ for the GOP as described in Section 6.2.7.

6.2.7 Feature-by-feature test with multivariate comparisons correction

The global shape analysis in the previous section can not indicate local shape differences which motivates the introduction of an FWER threshold correction for a feature-by-feature test. The permutation test approach on each variable T_{lk} yields an empirical distribution C_k, dependent standard Gaussian variables U_{lk} and the empirical covariance matrix $\hat{\Sigma}_U$. As a result, $U_l = (U_{l1}, \ldots, U_{lK})$ is approximately distributed as $\mathcal{N}_K(0, \hat{\Sigma}_U)$, where $\mathcal{N}_K$ is a multivariate Gaussian distribution with mean 0, covariance $\hat{\Sigma}_U$ and density function ψ such that each marginal is $U_{lk} \sim \mathcal{N}(0, 1)$.

Because each random variable U_{lk} is standard Gaussian, the threshold for each standard Gaussian variable should be the same. Thus, given a significance level α, we wish to find the threshold λ such that

$$P(U_{l1} < \lambda, \ldots, U_{lK} < \lambda) = 1 - \frac{\alpha}{2}.$$

The function P is a multiple integral from $-\infty$ to λ in each variable of $U_l \sim \mathcal{N}_K(0, \hat{\Sigma}_U)$ and can be understood as a function $g(\lambda)$ of the single variable λ. The function $g(\lambda)$ is monotonic increasing with asymptotes at 0 and 1. The numerical calculation of the p-values is based on the approximation over an appropriate interval of λ. Recall that $\lambda \geq \lambda_{corr}$ with

$$\lambda_{corr} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

is the threshold for a single standard Gaussian variable. Let $l \in \{1, \ldots, P\}$ be fixed, the threshold λ_{corr} is applicable if all $U_{\cdot k}$ are perfectly correlated. Furthermore, we know that $\lambda \leq \lambda_{indep}$ with

$$\lambda_{indep} = \Phi^{-1}\left(\left(1 - \frac{\alpha}{2}\right)^{1/K}\right)$$

because the threshold λ_{indep} is applicable if all $U_{\cdot k}$ are independent. The desired level $1 - \alpha/2$ will be rather near 1. Thus, the function $g(\lambda)$ will be concave downward in the interval $[\lambda_{corr}, \lambda_{indep}]$.

The values $g(\lambda_{corr})$ and $g(\lambda_{indep})$ can be estimated from a large number N_{Samp} of random samples $Y_n \sim \mathcal{N}_K(0, \hat{\Sigma}_U)$ with $n = 1, \ldots, N_{Samp}$ by

$$\hat{g}(\lambda) = \frac{\sum_{n=1}^{N_{Samp}} I_\lambda(\psi(y_{n1}, \ldots, y_{nK}))}{\sum_{n=1}^{N_{Samp}} \psi(y_{n1}, \ldots, y_{nK})}$$

$$\text{with } I_\lambda(\psi(y_n)) = \begin{cases} \psi(y_n) & : \quad \psi(y_n) < \lambda, \\ 0 & : \quad \text{otherwise}, \end{cases}$$

and $y_n = (y_{n1}, \ldots, y_{nK})$. We have chosen a number of $N_{Samp} = 200,000$ samples.

The computation of $g(\lambda_{indep})$ requires the comparison of y_n values only for those identified as not in the accepted subset for the smaller value λ_{corr} and adding into the accumulated sum for the newly accepted samples. Finally, the standard regula falsi method can be used to iteratively solve the equation $g(\lambda) = 1 - \alpha/2$ with initial evaluations $g(\lambda_{corr})$ and $g(\lambda_{indep})$.

The dashed-dotted line in Figure 3c shows the corrected threshold λ for a selected GOP.

7 Results

7.1 Fitting of s-reps to hippocampi

The hippocampus data set consists of binary images of 221 first-episode schizophrenia cases and 56 control cases as described in Section 2. Antialiased distance images were generated from the binary images according to [31]. Based on the distance images, appropriate preliminary fits by an initial alignment and

an atom and spoke stage are produced as described in Section 5. This preliminary stage is described in detail in Section 1.4 of the Supplementary Material. In order to control the manual work during the preliminary stage, we considered only the first 96 of 221 cases of SG as discussed in the Supplementary Material. Let $\tilde{A}_1$ be the set of 96 preliminary fits for SG and $\tilde{A}_2$ be the set of 56 preliminary fits for CG. All preliminary fits were translated and rotated to the CPNG mean of the set union $\tilde{A}_1 \cup \tilde{A}_2$ by standard Procrustes alignment [10] in order to remove global variation from the preliminary fits. Let $\bar{A}_1$ be the set of 96 aligned SG preliminary fits and $\bar{A}_2$ the set of 56 aligned CG preliminary fits. Finally, CPNG statistics were calculated for the s-rep populations as described in Sections 4 and 5.

A challenging question is the appropriate estimation of the shape distributions of both populations (SG and CG) during the CPNG stage. An option, is to calculate the CPNG statistic of each population ($\bar{A}_1$ and $\bar{A}_2$) resulting in two means and shape distributions. Another option, is to calculate the CPNG statistic of the pooled population ($\bar{A}_1 \cup \bar{A}_2$) resulting in a single mean and shape distribution. The use of two individual shape distributions result in independent fittings between the two populations. On the other hand, the fittings should not be biased and have good correspondence between the populations, which is provided by a pooled shape distribution. A pooled CPNG statistic also removes possible bias from the manual adjustments during the preliminary stage.

The final fitting results obtained from two separate shape distributions showed extraordinary high separation properties and indicated a large bias. Thus, the main focus was the analysis of fittings using a pooled CPNG statistic from $\bar{A}_1 \cup \bar{A}_2$. In addition, we have generated a second group of final fittings derived from CPNG stages using a pooled shape distribution, two individual shape distributions and two individual interchanged shape distributions. The second group is a compromise between independence and a small bias, and is discussed in Section 3 of the Supplementary Material.

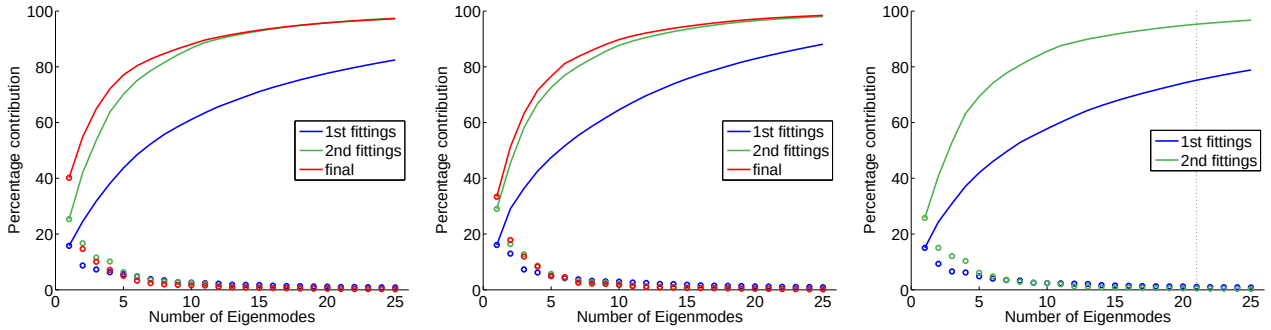
Each CPNG statistic contains a backward mean, the eigenmodes and the corresponding CPNG scores. Figure 4 shows the explained amount of variation by the first 25 eigenmodes for the aligned preliminary fittings after atom and spokes stages (1st fittings), i.e., for $\bar{A}_1$ (subset of SG), $\bar{A}_2$ (CG) and $\bar{A}_1 \cup \bar{A}_2$. The number of eigenmodes was selected to describe more than 75% of the total cumulative variance. This number compromises on capturing enough shape variation while limiting the shape space in order to avoid overfitting. Accordingly, the first 21 eigenmodes of the pooled shape distribution were selected for the CPNG stage describing 75.2% total variance. 18 eigenmodes are required to describe 75.3% of the total cumulative variance of $\bar{A}_1$, and 15 eigenmodes to describe 75.7% for $\bar{A}_2$.

In the CPNG stage, the obtained backward mean of $\bar{A}_1 \cup \bar{A}_2$ was translationally and rotationally aligned to the 221 SG cases and the 56 CG cases. An additional scaling of the means would bias the CPNG statistic because the principal components already contain size information. Afterwards, the aligned means were optimized inside the CPNG shape space and under the penalty of a Mahalanobis distance match term. A high penalty term leads to better correspondence between cases but to less accurate fits. An appropriate penalty term was chosen by a simulation study, the report of which is omitted. At the end, the final spoke stage was performed to ensure that the spoke directions match the boundary.

Figure 4 also shows CPNG analyses for the obtained 2nd fittings of the corresponding cases to $\bar{A}_1$, $\bar{A}_2$ and $\bar{A}_1 \cup \bar{A}_2$ using a pooled shape space during the CPNG stage. The respective numbers of eigenmodes explain an increased amount of variation compared to the first fittings as a result of improved correspondence across the populations. Now, 18 eigenmodes describe 94.9% of the total variance for the subset of SG, 15 eigenmodes describe 93.5% for CG and 21 eigenmodes describe 95.3% for the pooled group.

The final fittings were re-scaled into a world coordinate system (units of mm) with the stored scaling factor from the normalization step described in Section 2. We denote the re-scaled sets of final fittings by the set A_1 of 221 s-reps for SG and the set A_2 of 56 s-reps for CG. The total cumulative variance of the CPNG analysis of A_1 (SG) and A_2 (CG) is depicted in Figure 4 (final). Now, 18 eigenmodes describe 94.8% of the total variance of SG and 15 eigenmodes describe 94.5% for CG. More than 75% of the total cumulative variance of CPNG shape space is now described by using only 5 eigenmodes compared to 18 and 15 as shown previously.

The average volume in mm^3 (and standard deviation) of the final fittings was $3,036$ (343) for SG and $3,137$ (295) for CG. The observed hippocampal volume reduction for schizophrenia patients is consistent with previous studies (e.g., [25]). The average volume overlap between fittings and binary images was 94% for SG and CG (depicted in Section 3 of the Supplementary Material) which is fairly accurate. The percent-volume overlap was measured by the Dice coefficient as defined in the Supplementary Material. The variance of the Dice coefficient is small for both groups. Nevertheless, a larger variance inside SG is observed. Schizophrenia is a heterogeneous disease and also contains hippocampi variations between healthy patients.



(a) SG (b) CG (c) joint group

Fig. 4: CPNG analysis of s-reps before the CPNG stage (1st fittings), after CPNG and final spoke stage (2nd fittings) and after scaling into the world coordinate system (final). The variance contribution of the first 25 eigenmodes are depicted together with the cumulative variance for the CPNG analysis of the (a) SG group, (b) CG group and (c) joint group. The set of 1st and 2nd fittings consist of 96 s-reps for SG, 56 s-reps for CG and 152 s-reps for the pooled group. The set of final fittings consist of 221 s-reps for SG and 56 s-reps for CG. The dotted vertical line in (c) depicts the number 21 of used eigenmodes for the description of the shape space during the CPNG stage.

In Figure 5, the distributions of the SG and CG fittings are visualized by the projections of the scaled CPNG scores matrix Z_{Comp} (see Section 4) onto the distance weighted discrimination (DWD) direction. DWD is a discrimination method which avoids the data piling problems of support vector machine [27, 37]. The projected distributions of SG and CG fittings for the pooled class are estimated by kernel density estimates (KDEs). The different areas under the CG and SG curves are due to unbalanced population sizes (56 for CG compared to 221 for SG). A difference between the populations is visible but not very strong. Thus, it is an interesting question whether the proposed hypothesis test in Section 6.2 will be able to find significant differences between SG and CG for both fittings classes.

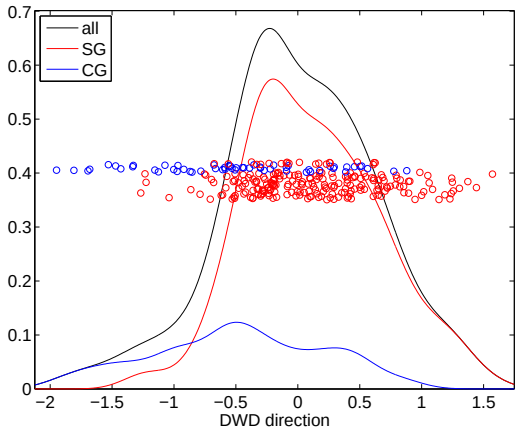


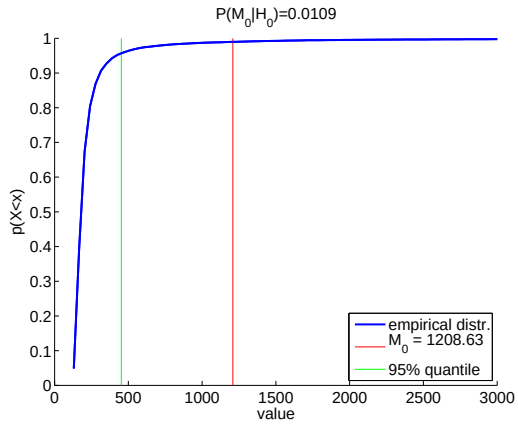
Fig. 5: Jitterplot and KDEs showing the distributions of final SG and CG fittings projected onto the DWD direction. Additionally, the KDE of the pooled distribution of SG and CG is shown (all). A difference between the populations is visible but not very strong.

7.2 Global test results

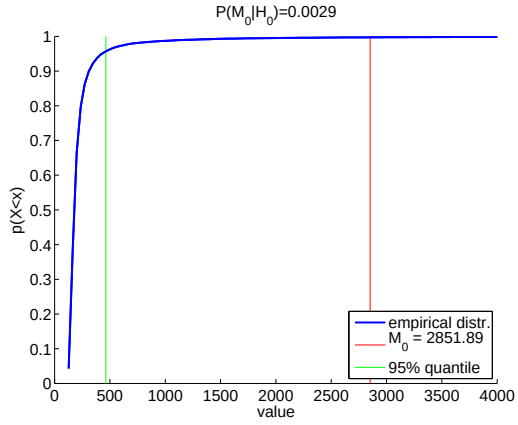
The obtained final fittings were used to test the hypothesis (5) by the proposed procedure in Section 6.2 with a significance level of $\alpha = 0.05$. An alternative pre-processing step (called PP2) is applied in addition to the pre-processing described in Section 6.2.1 (called PP1 in the following). PP2 translates and rotates each s-rep $\tilde{\mathbf{s}} \in \tilde{A}$ to an overall CPNG backward mean μ without scaling. Thus, each aligned s-rep is described by a feature vector $\mathbf{t} = \mathbf{s}$. The global scaling information was previously described by the feature τ in PP1. In contrast, this is captured by the skeletal positions and spoke length using PP2.

Figure 6 shows the global test results for the difference measure d^2 using PP1 and PP2. The global hypothesis of equal sample means is rejected and a statistically significant difference between the shape distribution of SG and CG is established ($p = 0.0109$ for PP1 and $p = 0.0029$ for PP2 with $p = P(M_0|H_0)$). The smaller p-value for PP2 seems to be due to the volume information being spread into the 24 skeletal positions instead of into a single feature. The feature-by-feature test will highlight this fact in the next section. Intermediate results of the proposed hypothesis test procedure are shown in Figure 3 on the basis of a selected GOP. Further visualizations of the procedure can be found in the Supplementary Material.

A detailed power and simulation study is beyond the scope of this paper and left for future work. However, the power of the proposed hypothesis test is



(a) PP1



(b) PP2

Fig. 6: Global test results using PP1 in (a) and PP2 in (b). The empirical distribution of $M_l, l = 1, \ldots, 30,000$ is shown together with M_0 and the 95% quantile of the empirical distribution.

demonstrated on the basis of a real data example. Furthermore, the results are compared with a direction projection permutation (DiProPerm) based mean hypothesis test [46]. The DiProPerm test is based on the evaluation of the scaled CPNG scores matrix Z_{Comp} (see Section 4). The CPNG scores matrix is calculated for SG and CG using both pre-processing methods. Thus, the DiProPerm test is calculated in Euclidean space using standard Euclidean statistics in contrast to the proposed hypothesis test, which is performed in the non-Euclidean s-rep space using the CPNG backward means. An interesting open problem is to extend a method such as DiProPerm in an intrinsic way: in other words to perform DiProPerm using Manifold geodesic distances.

Table 1 summarizes all global test results. We used 30,000 permutations in all settings to be consistent with Section 6.2.2. The DiProPerm test does not require such a high number of permutations in contrast to the proposed global test. Simulations, reported in the Supplementary Material, reveals that a large permutation size is needed to obtain stable results because of the Mahalanobis distance. DiProPerm was carried out using a mean difference (MD) test statistic as recommended in [46]. The DWD-DiProPerm performance was comparable to the Mahalanobis distance results. The support vector machine (SVM) results of DiProPerm were less powerful, probably due to data pilling effects. All results are statistically significant at the level of $\alpha = 0.05$.

Table 1: Empirical p-value results using difference measure d^2 for the proposed global hypothesis test in comparison with results obtained by DiProPerm. Two different pre-processing steps were applied: (PP1) Full Procrustes alignment with scaling. (PP2) Full Procrustes alignment without scaling. Three different projection directions were used for DiProPerm.

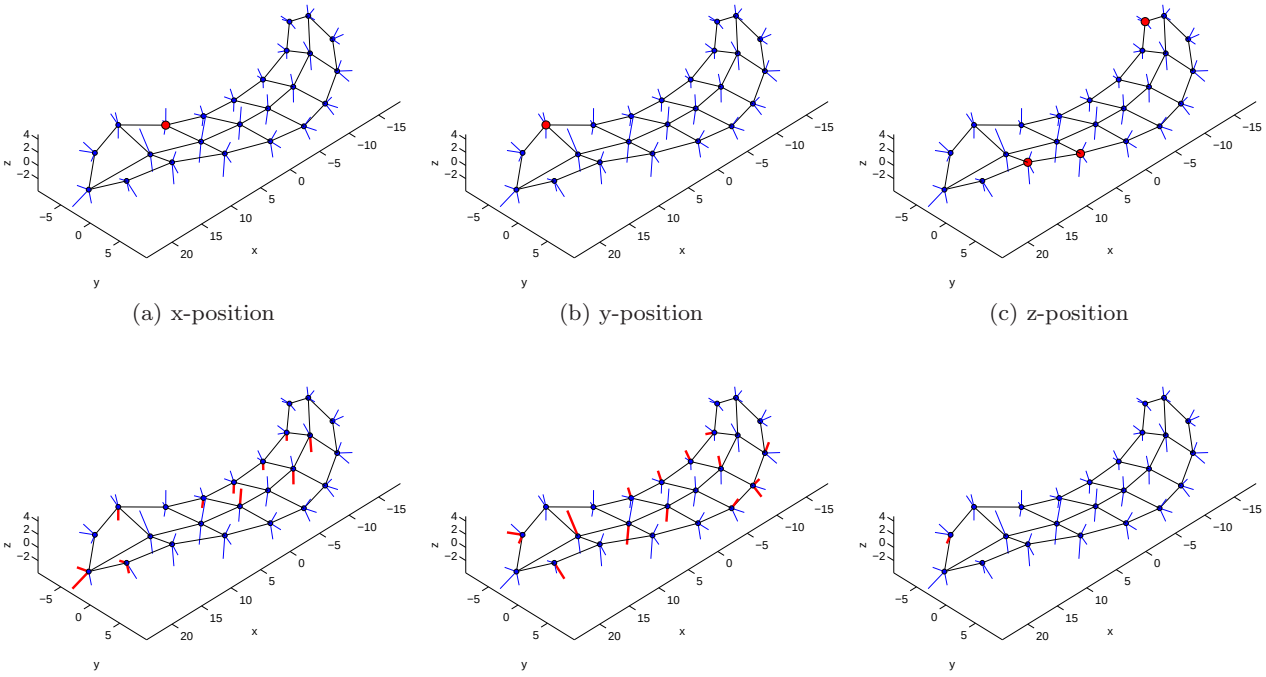
method	empirical p-value PP1	PP2
Mahalanobis distance		
difference measure d^2	0.0109	0.0029
DiProPerm using MD-statistic		
DWD direction vector	0.0074	0.0038
SVM direction vector	0.0119	0.0136

7.3 Single GOP test results

The global shape analysis of hippocampi in the previous section can not indicate local shape differences. Interesting structural changes of the surface are often reflected by a few GOPs, e.g., the local bending of an area. Therefore, the proposed threshold correction for a feature-by-feature test in Section 6.2.7 is useful. Such a feature-by-feature test is not available from DiProPerm.

As our feature-by-feature test approach is novel for nonlinear hypotheses, there is no competing method to compare with. However, a method to evaluate the test is needed. The performance of the feature-by-feature test was evaluated using Receiver Operating Characteristic (ROC) curves. Selected examples of this analysis are reported in Section 2.4 of the Supplementary Material. For each permutation, an ROC curve was generated from the cumulative histograms of the two permuted samples which results in an envelope under the null distribution. In addition, an ROC curve between the two true observed samples was obtained. A significant feature is indicated if the ROC curve of the observed data is close to the boundary or outside the envelope, otherwise not. A comparison of the hypothesis test results to this reveals the high quality of the proposed method.

Figures 7 and 8 visualize the feature-by-feature test results for PP1. Test results are shown on the basis of the skeletal grid given by the CPNG back-



(a) x-position (b) y-position (c) z-position

(d) latitude direction (e) longitude direction (f) spoke length

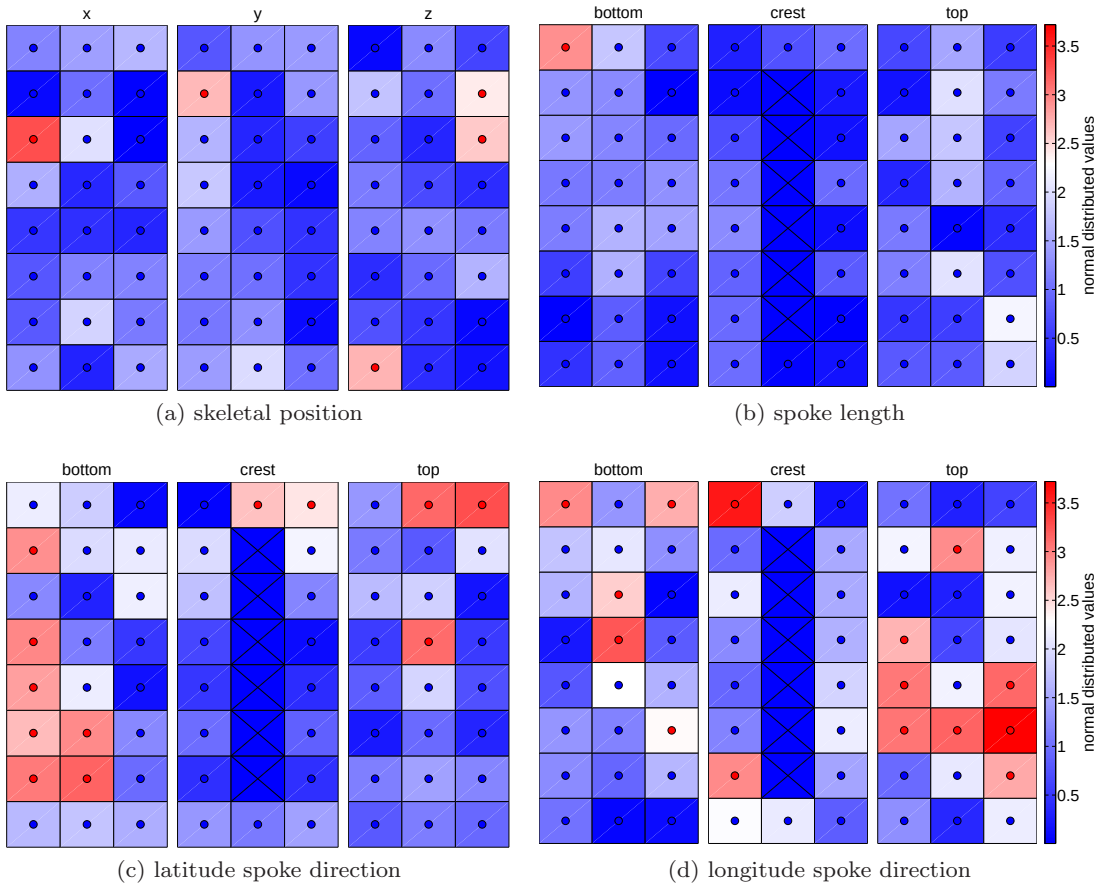
Fig. 7: Significant GOPs using PP1 based on the 3×8 skeletal sheet of the SG CPNG mean. Test results are shown in (a)-(c) for the skeletal x, y and z-positions, in (d) for the latitude spoke directions, in (e) for the longitude spoke directions and in (f) for the spoke lengths. Non-significant skeletal positions are marked by small blue points and significant skeletal positions are marked by large red points. Similar, non-significant spoke directions and lengths are marked by small blue lines whereas significant spoke directions and lengths are marked by wide red lines.

ward mean of SG. Recall that each discrete slabular s-rep is organized into 24 atoms in a 3×8 grid (see Section 3). This results in 271 GOPs with 72 GOPs corresponding to the skeletal positions of the s-rep (x, y and z-positions), 66 GOPs for the latitude spoke directions (bottom, crest and top), 66 GOPs for the longitude spoke directions (bottom, crest and top), 66 GOPs for the spoke lengths (bottom, crest and top) and 1 GOP for the global scaling factor. The corrected threshold is $\lambda = 2.2917$ as defined in Section 6.2.7.

Figure 7 indicates local shape changes by highlighting local parts of the s-rep. Red points mark significant skeletal x, y and z-positions in the Figures (a)-(c). Non-significant skeletal positions are marked by smaller blue points in these figures. Five significant skeletal positions can be observed at the crest of the sheet, one in the x and y-direction and three in the z-direction. Moreover, significant spoke directions and lengths are marked by wide red lines and non-significant by thinner blue lines in the Figures (d)-(f). Several latitude and longitude spoke directions indicate locally significant deformations between the two groups in the Figures (d)-(e). The most latitude differences are statistically significant on the bottom side of the skeletal sheet whereas more longitude differences are significant on the top side. Furthermore, we observe no spoke direction with simultaneously significant latitude and longitude. This behavior should be investigated in future studies. The latitude and longitude differences could indicate local bending around the y and z-axis, respectively. Figure (f) highlights one significant spoke length on the front bottom side of the skeletal sheet.

In addition to the results presented in Figures 7, the global scaling factor τ between SG and CG was found statistically significant. The GOP $|U_{0K}|$ was 2.7627 where the index K corresponds to the global scale factor.

These observations and results are also emphasized by Figure 8 which shows the magnitude of significance of all GOPs except the scaling factor. In order to simplify the visualization all standard normal values $U_{0k}, k = 1, \ldots, K-1$ are presented in absolute values. The color map is non-linear defined from blue to white to red. The corrected threshold λ defines the color white. Blue and red visualize non-significant and significant values, respectively. The blue small circle inside a block marks whether a U_{0k} is less than or equal to the threshold λ. Red small circles mark if a



(a) skeletal position (b) spoke length

(c) latitude spoke direction (d) longitude spoke direction

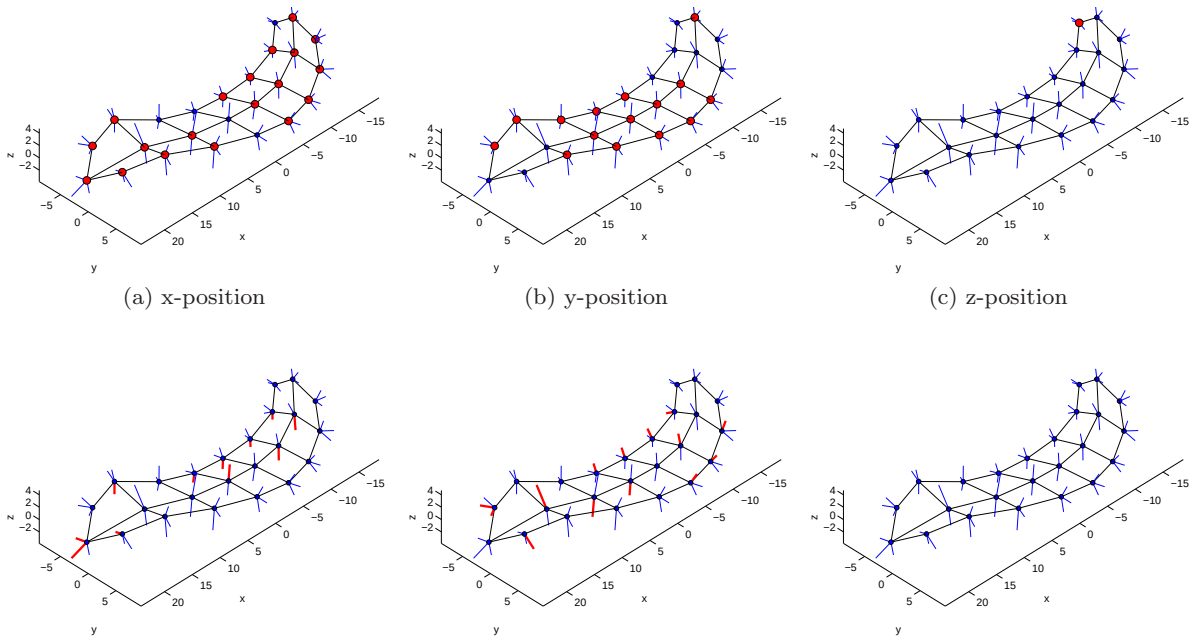
Fig. 8: Colored significant map of U_{0k} with a corrected threshold $\lambda = 2.2917$ using PP1. Each box represents a GOP which correspond to a skeletal atom. The color map on the left side is non-linear and has a range from blue (not significant) to white (λ) to red (significant). The circle inside each box marks whether an U_{0k} is less or equal than the threshold λ (symbolized by blue) or if an U_{0k} is greater than the threshold λ (symbolized by red).

U_{0k} is greater than the threshold λ and thus statistically significant, showing up as red in Figure 7. Particularly, several latitude and longitude spoke directions show a highly significant magnitude in Figure 8.

Figures 9 and 10 are identical to the two previous figures except for the use of PP2 instead of PP1. Several skeletal x and y-positions are statistically significant in contrast to Figures 7 and 8 with only one significant skeletal x and y-position. The volume difference between the two populations is reflected by the skeletal x and y-positions using PP2. Thus, the significant skeletal x and y-positions show rather significant differences from a global deformation than from local deformations. However, we observe only one statistically significant skeletal z-position because the skeletal sheet of the hippocampus is rather flat, as medial as possible and therefore located close to the x-y plane, where $z = 0$. As a result, several skeletal z-coordinates are scaling invariant.

Nevertheless, the observation of only one significant skeletal z-position in addition to no observed statistically significant spoke length in Figure 9f implies that we only observe statistically significant volume differences in the x-y direction but not in the z-direction. Skeletal x and y-positions equal to $x = 0$ and $y = 0$ are scaling invariant in the x and y-directions respectively. As a result, no statistically significant x-positions can be observed close to $x = 0$ in Figure 9a. Moreover, Figures 10c and 10d show only small differences compared to Figures 8c and 8d. Similar results between spoke directions are expected because of the scaling invariance of $u_{ij} \in S^2$. The slightly different color scheme is also due to a different threshold.

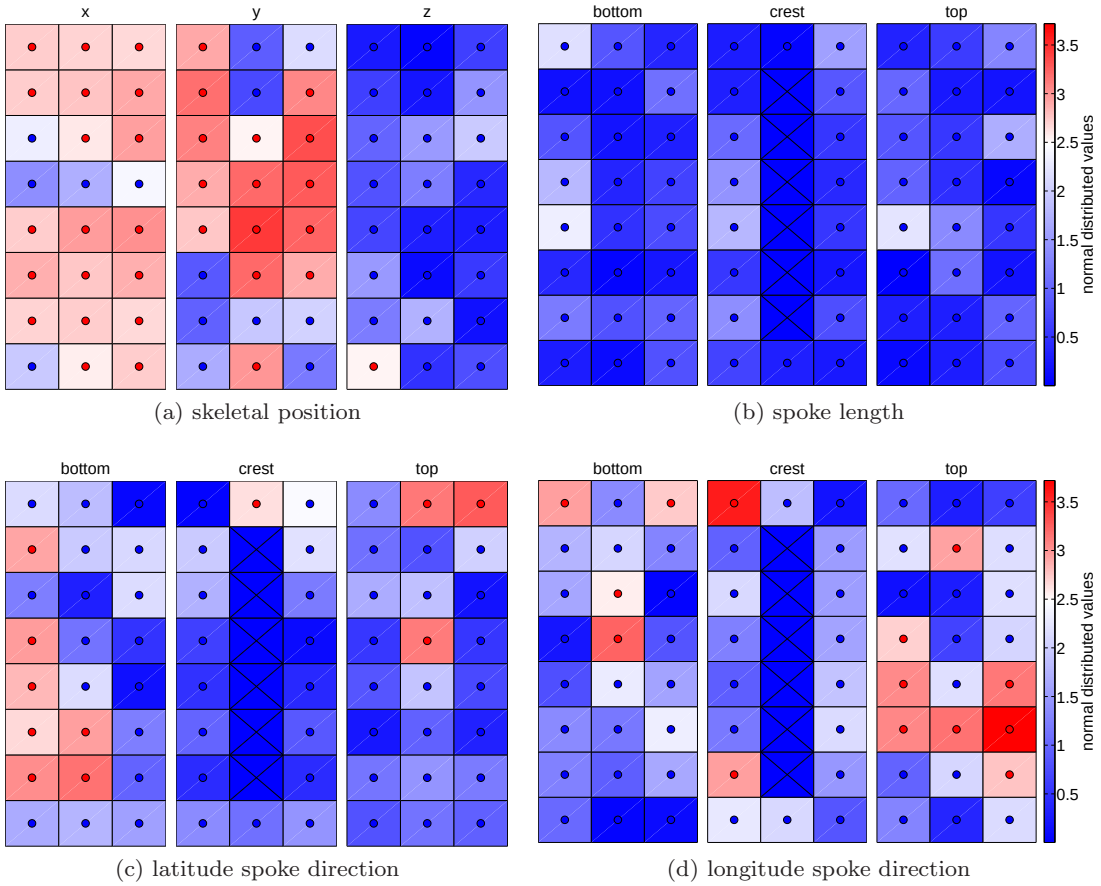
Additional computations and results are shown in the Supplementary Material. Section 2.5 of the Supplementary Material presents results using an alternative measure d^1 defined by a vector of unsigned partial differences such as the Euclidean distance between two skeletal positions. That difference measure changes the GOPs, i.e., how the single s-rep features are combined to GOPs. The difference measure d^2 closely reflects each s-rep feature. The choice of an appropriate difference measure depends on the nature of the medical



(a) x-position (b) y-position (c) z-position

(d) latitude direction (e) longitude direction (f) spoke length

Fig. 9: As Figure 7, now based on PP2.



(a) skeletal position (b) spoke length

(c) latitude spoke direction (d) longitude spoke direction

Fig. 10: As Figure 8, now based on PP2 with a corrected threshold $\lambda = 2.4837$.

research question. In addition, hypothesis test results using a second group of final fittings are presented, derived from 5 independent CPNG stages in Section 5 by using a pooled shape distribution, two individual shape distributions and two individual interchanged shape distributions. The second group of final fittings is described in detail in Section 3 of the Supplementary Material.

8 Discussion

This paper proposes a novel method to test global and local hypotheses on Euclidean and non-Euclidean data. Important requirements of shape models are pointed out in order to test for population differences. Furthermore, suitable statistical methods are proposed to analyze the Euclidean and non-Euclidean elements of the models. In addition, the estimation of appropriate shape distributions of populations is worked out. Finally, the analysis of first episode schizophrenia patients compared to controls demonstrated the power of the hypothesis test given a proper pre-processing. The effect of different pre-processings of the data are highlighted. The developed feature-by-feature test is novel and important for physicians in order to understand local shape changes. The method can easily be adapted for desired GOPs depending on underlying research questions. A difference measure for the analysis of s-reps is proposed. The visualization of local shape changes is of great interest for the study of local rotational deformations [39] which is a subject of future studies.

The s-rep model, statistics and the fitting procedure resulted in accurate fittings with a high concentration of variance in relatively few eigenmodes. This reflects the high correspondence between the s-reps. The introduced test found significant differences between the two populations. First, a statistically significant loss of hippocampal volume was observed by the global scaling factor which is in agreement with [25,28,29]. Second, a significant volume difference was observed in the x and y-directions but not in the z-direction for the aligned hippocampi. Third, several spoke directions were found as statistically significant.

This study is the first study that examines directional information using s-reps. The significant differences of several spoke directions confirm the importance of our contribution in the research of morphological shape changes and encourages further research. Later studies should more deeply investigate if spoke direction differences are due to independent local deformation of GOPs or due to local rotational deformation. Styner et al. [42] indicated a potential local bending of the hippocampi between the two groups. Furthermore, this study is the first study that could identify directions driving the volume change.

In general, results are challenging to compare between studies of brain morphology because of different models, features and metrics. Narr et al. [29] calculated a radial distance measure in addition to a measure that examined the signal intensity on the basis of a surface based mesh modelling method. Also, Mamah et al. [25] used a triangulated graph representation of the hippocampi. Such models are limited compared to s-reps because the interior of an object is not described by the model itself. The model representation in McClure et al. [28] is a skeleton type which leads to less correspondence between populations and contains further disadvantages, e.g., all spoke length are identical on each atom. Furthermore, McClure et al. [28] applied an FDR based test approach in contrast to the FWER based approach proposed in this paper. The FDR is a less strict multiple testing criteria than the FWER. However, the discussed results are consistent between the studies. The s-rep model provides a relatively rich description of an object. Moreover, the proposed test procedure offers global and local nonlinear hypothesis tests based on Euclidean and non-Euclidean GOPs. Thus, the test supports more consistent and sensitive interpretations of morphological changes.

This paper motivates several areas of further research. 1) A simplification of the s-rep fitting procedure is desirable that depends on correct choice of several fitting parameters. The choice of a large number of parameters might be an avoidable difficulty in the use of s-reps in clinical practice. 2) The definition of an adaptive s-rep model that finds an optimal skeletal grid could be of relevance for the future. The grid need not to be rectangular but must correspond across cases of a population. 3) The hypothesis test might be extended by including image intensities in addition to morphological features. An interesting research question is the study of correlation between morphological changes and intensities. 4) An alternative combining function might decrease the required large number of permutations for a global test. 5) A power study on the basis of simulated data to elaborate further the proposed method. 6) A comparison of hippocampi between two treatment groups of first episode schizophrenia. 7) Extension of the method to hippocampi from longitudinal data. In addition, a similar hypothesis test based on the sample variance instead of the sample mean could be of future interest.

Acknowledgements The following researchers have also contributed to this work: Jared Vicory (UNC) gave advice on running Pablo and provided earlier fits of 62 hippocampi, Juan Carlos Prieto (CREATIS-INSA, France) provided the implementation of a crest interpolation term in Pablo and removed bugs from the program, Sungkyu Jung

(University of Pittsburgh, USA) provided Figure 2, program code and additional discussions about CPNS, Martin Styner (UNC) provided the hippocampus data set and answered questions. The first author acknowledges support from the Norwegian Research Council through grant 176872/V30 in the eVita program and additional support from the Tromsø Telemedicine Laboratory and the Department of Electrical Engineering and Computer Science at the University of Stavanger, Norway.

References

1. Abramovich, F., Benjamini, Y.: Adaptive thresholding of wavelet coefficients. Comput. Statist. Data Anal. **22**(4), 351–361 (1996)
2. Albertson, R.C., Streelman, J.T., Kocher, T.D.: Genetic basis of adaptive shape differences in the cichlid head. J. Hered. **94**(4), 291–301 (2003)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. **57**(1), 289–300 (1995)
4. Blum, H., Nagel, R.: Shape description using weighted symmetric axis features. Pattern Recognit. **10**(3), 167–180 (1978)
5. Bookstein, F.L.: Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Image Anal. **10**(3), 225–243 (1996)
6. Bullmore, E., Fadili, J., Breakspear, M., Salvador, R., Suckling, J., Brammer, M.: Wavelets and statistical analysis of functional magnetic resonance images of the human brain. Stat. Methods Med. Res. **12**(5), 375–399 (2003)
7. Chumbley, J.R., Friston, K.J.: False discovery rate revisited: FDR and topological inference using Gaussian random fields. NeuroImage **44**(1), 62–70 (2009)
8. Cootes, T.F., Taylor, C., Cooper, D., Graham, J.: Training models of shape from sets of examples. In: D. Hogg, R. Boyle (eds.) Proc. British Machine Vision Conference, pp. 9–18. Berlin. Springer-Verlag (1992)
9. Damon, J., Marron, J.S.: Backwards principal component analysis and principal nested relations. J. Math. Imaging Vision (2013). DOI: 10.1007/s10851-013-0463-2
10. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. John Wiley & Sons, Chichester (1998)
11. Edgington, E.: Randomization Tests, 3rd edn. Dekker, New York (1995)
12. Ferrarini, L., Palm, W.M., Olofsen, H., van Buchem, M.A., Reiber, J.H., Admiraal-Behloul, F., et al.: Shape differences of the brain ventricles in Alzheimer's disease. NeuroImage **32**(3), 1060–1069 (2006)
13. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Ann. Inst. Henri Poincaré **10**, 215–310 (1948)
14. Gerig, G., Styner, M., Shenton, M.E., Lieberman, J.A.: Shape versus size: improved understanding of the morphology of brain structures. MICCAI pp. 24–32 (2001)
15. Goodall, C.: Procrustes methods in the statistical analysis of shape. J. R. Stat. Soc. Ser. B Stat. Methodol. **53**(2B), 285–339 (1991)
16. Gouttard, S., Styner, M., Joshi, S., Gerig, G.: Subcortical structure segmentation using probabilistic atlas prior. In: Proceedings of the SPIE Medical Imaging, vol. 65122, pp. J1–J11 (2007)
17. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. Statist. Sinica **20**(1), 1–58 (2010)
18. Jung, S., Dryden, I.L., Marron, J.S.: Analysis of principal nested spheres. Biometrika **99**(3), 551–568 (2012)
19. Jung, S., Foskey, M., Marron, J.S.: Principal arc analysis on direct product manifolds. Ann. App. Statist. **5**(1), 578–603 (2011)
20. Jung, S., Liu, X., Marron, J.S., Pizer, S.M.: Generalized PCA via the backward stepwise approach in image analysis. In: J.A. et al. (ed.) 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*, vol. 83, pp. 111–123 (2010)
21. Karcher, H.: Riemannian center of mass and mollifier smoothing. Comm. Pure Appl. Math. **30**(5), 509–541 (1977)
22. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley, Chichester (1999)
23. Kilner, J.M., Kiebel, S.J., Friston, K.J.: Applications of random field theory to electrophysiology. Neurosci. Lett. **374**, 174–178 (2005)
24. Kurtek, S., Ding, Z., Klassen, E., Srivastava, A.: Parameterization-invariant shape statistics and probabilistic classification of anatomical surfaces. Inf. Process. Med. Imaging **22**, 147–158 (2011)
25. Mamah, D., Harms, M.P., Barch, D.M., Styner, M.A., Lieberman, J., Wang, L.: Hippocampal shape and volume changes with antipsychotics in early stage psychotic illness. Front Psychiatry **3**(96), 1–10 (2012)
26. Marozzi, M.: Some remarks about the number of permutations one should consider to perform a permutation test. Statistica **64**(1), 193–202 (2004)
27. Marron, J.S., Todd, M.J., Ahn, J.: Distance weighted discrimination. J. Amer. Statist. Assoc. **102**(480), 1267–1271 (2007)
28. McClure, R.K., Styner, M., Maltbie, E., Liebermann, J.A., Gouttard, S., Gerig, G., Shi, X., Zhu, H., et al.: Localized differences in caudate and hippocampal shape are associated with schizophrenia but not antipsychotic type. Psychiatry Res. Neuroimaging **211**(1), 1–10 (2013)
29. Narr, K.L., Thompson, P.M., Szeszko, P., Robinson, D., Jang, S., Woods, R.P., Kim, S., Hayashi, K.M., Asunction, D., Toga, A.W., Bilder, R.M.: Regional specificity of hippocampal volume reductions in first-episode schizophrenia. NeuroImage **21**(4), 1563–1575 (2004)
30. Nichols, T.E., Hayasaka, S.: Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. **12**(5), 419–446 (2003)
31. Niethammer, M., Juttukonda, M.R., Pizer, S.M., Saboo, R.R.: Anti-aliasing slice-segmented medical images via Laplacian of curvature flow. In preparation (2013)
32. Nitrc: S-rep fitting, statistics, and segmentation. http://www.nitrc.org/projects/sreps (2013)
33. Pantazis, D., Nichols, T.E., Baillet, S., Leahy, R.M.: A comparison of random field theory and permutation methods for the statistical analysis of MEG data. NeuroImage **25**(2B), 383–394 (2005)
34. Pennec, X.: Statistical computing on manifolds: from Riemannian geometry to computational anatomy. Emerging Trends in Visual Computing **5416**, 347–386 (2008)
35. Pesarin, F.: Multivariate Permutation Tests with Applications to Biostatistics. John Wiley & Sons, Chichester (2001)
36. Pizer, S.M., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J.N., Huckemann, S., Marron, J.S.: Nested sphere statistics of skeletal models. In: Innovations for Shape Analysis: Models and Algorithms, Lecture Notes in Comput. Sci., pp. 93–115. Springer (2013)

37. Qiao, X., Zhang, H.H., Liu, Y., Todd, M.J., Marron, J.S.: Weighted distance weighted discrimination and its asymptotic properties. J. Amer. Statist. Assoc. **105**(489), 401–414 (2010)
38. Rohde, G.K., Ribeiro, A.J.S., Dahl, K.N., Murphy, R.F.: Deformation-based nuclear morphometry: capturing nuclear shape variation in HeLa cells. Cytometry A **73**(4), 341–350 (2008)
39. Schulz, J., Jung, S., Huckemann, S., Pierrynowski, M., Marron, J.S., Pizer, S.M.: Analysis of rotational motion from directional data. Submitted (2013)
40. Shi, X., Ibrahim, J.G., Lieberman, J., Styner, M., Li, Y., Zhu, H.: Two-stage empirical likelihood for longitudinal neuroimaging data. Ann. Appl. Stat. **5**(2B), 1132–1158 (2011)
41. Siddiqi, K., Pizer, S.: Medial Representations: Mathematics, Algorithms and Applications, 1 edn. Computational Imaging and Vision, Vol. 37. Springer, Dordrecht, Netherlands (2008)
42. Styner, M., Lieberman, J., Pantazis, D., Gerig, G.: Boundary and medial shape analysis of the hippocampus in schizophrenia. Med. Image Anal. **8**(3), 197–203 (2004)
43. Terriberry, T., Joshi, S., Gerig, G.: Hypothesis Testing with Nonlinear Shape Models. In: G. Christensen, M. Sonka (eds.) Information Processing in Medical Imaging, *Lecture Notes in Computer Science*, vol. 3565, pp. 15–26. Springer Berlin Heidelberg (2005)
44. Van De Ville, D., Blu, T., Unser, M.: Integrated wavelet processing and spatial statistical testing of fMRI data. NeuroImage **23**(4), 1472–1485 (2004)
45. Wang, L., Joshi, S.C., Miller, M.I., Csernansky, J.G.: Statistical analysis of hippocampal asymmetry in schizophrenia. NeuroImage **14**(3), 531–545 (2001)
46. Wei, S., Lee, C., Wichers, L., Li, G., Marron, J.S.: Direction-projection-permutation for high dimensional hypothesis tests (2013). ArXiv:1304.0796

Supplementary Material

Nonlinear Hypothesis Testing of Geometrical Object Properties of Shapes Applied to Hippocampi

Jörn Schulz · Stephen M. Pizer · J.S. Marron · Fred Godtliebsen

1 Model fitting and statistics

1.1 Limitation of a 3×8 grid of skeletal positions

A hippocampus example with bumps which are not tightly described by a (3×8) grid is visualized in Figure 1. An s-rep model with a larger number of skeletal positions, i.e., with a finer grid could solve such problems. The example depicts a limitation only in specific cases since the shape of the hippocampus differs from person to person. Furthermore, we do not look at individual s-reps that may not be perfectly correct but rather at differences between groups which are not biased versus the other.

1.2 Discussion on CPNS analysis across populations

In Section 4 in the main article, we have pointed out the difference between CPNS and CPNG. CPNG uses only great subsphere fittings whereas the best fitting subspheres can be small or great in CPNS. We have observed an increased variance of the CPNS means across several populations, e.g., for a large number of permutation sets as used in the proposed hypothesis test. Jung et al. [2] pointed out a potential overfitting of the data because PNS tends to find smaller spheres than great spheres. Therefore, a sequential test was proposed in [2, Section 3]. This section will propose a modification of the test in [2] and refers to the paper for detailed descriptions. The sequential test procedure consists of a likelihood ratio test and a parametric bootstrap test in order to test the significance of a "small" subsphere fitting as explained in the following.

1. Test $H_{0a} : r = \pi/2$ versus $H_{1a} : r < \pi/2$ by the likelihood ratio test where $r = \pi/2$ indicates a great sphere and $r < \pi/2$ a small sphere. If H_{0a} is accepted, then fit a great sphere with $r = \pi/2$ and proceed to the next layer.
2. If H_{0a} is rejected, then test the isotropy of the distribution by the parametric bootstrap test with $H_{0b} : F_X$ is an isotropic distribution with a single mode, versus $H_{0b} :$ not H_{0b} (i.e., anisotropic) given a distribution function $F_X, X \in S^d$. If H_{0b} is accepted, then use great spheres for all further subsphere fittings.

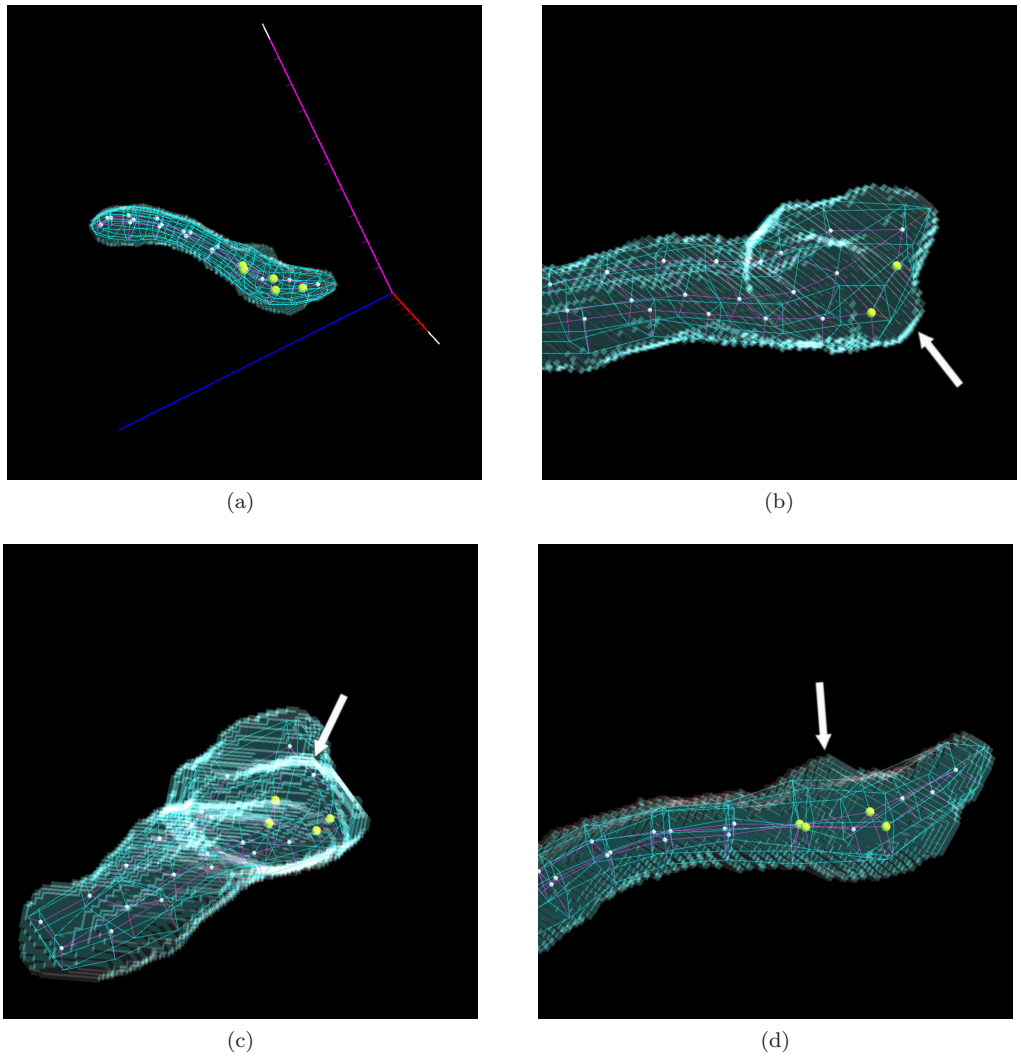
In calculation of CPNS statistics for several populations, the sequential test will be carried out independently for each population leading to potential different decompositions. Thus, the test must be modified, because the analysis of CPNS means across populations requires commensurate coordinate systems. Suppose we have two populations G_1 and G_2 with samples on S^d and P permutations of the set union $G_1 \bigcup G_2$. Each permuted set union can be split into two subgroups G_{1l} and G_{2l} with the same number of elements as G_1

J. Schulz
Department of Mathematics and Statistics, University of Tromsø, Norway Tel.: +47 45696867
E-mail: jorn.schulz@uit.no

F. Godtliebsen
Department of Mathematics and Statistics, University of Tromsø, Norway, E-mail: fred.godtliebsen@uit.no

Stephen M. Pizer
Department of Computer Science, University of North Carolina at Chapel Hill (UNC), USA, E-mail: smp@cs.unc.edu

J.S. Marron
Department of Statistics & Operations Research, UNC, USA, E-mail: marron@unc.edu



(a) (b) (c) (d)

Fig. 1: Final fit of a hippocampus with bumps that are not well described by an s-rep based on a (3×8) grid. (a) Entire 3D view to the s-rep with corresponding coordinate system. (b) Bump on the side located between two hub positions. (c-d) Bump on the top located between four hub positions.

and G_2, $l = 1, \ldots, P$. In order to analyze mean difference, the CPNS mean must be calculated for each permutation group $G_{il}, i = 1, 2$. We propose a modified sequential test by the following procedure.

1. Test $H_{0a} : \bigcap_i \bigcap_l H_{0a}^{i,l}$ versus $H_{1a} : \bigcup_i \bigcup_l H_{1a}^{i,l}$ by the likelihood ratio test with $i = 1, 2$ and $l = 1, \ldots, P$, whereas $H_{0a}^{i,l}$ is the sub-hypothesis for the lth permutation of group i. If H_{0a} is accepted, then fit a great sphere with $r = \pi/2$ and proceed to the next layer.
2. If H_{0a} is rejected, then test the isotropy of the distribution by the parametric bootstrap test. If $H_{0b} : \bigcap_i \bigcap_l H_{0b}^{i,l}$ is accepted, then use great spheres for all further subsphere fittings.

The implementation of such a test is left for future work. In this article we have used CPNG to analyze populations of s-reps.

1.3 An alternative unsigned difference measure d^1

This section introduce an alternative difference measure d^1 in addition to d^2 as described in Section 6.2.4 in the main article. The measure d^2 is defined by signed differences whereas the measure d^1 is defined by unsigned differences which turning each GOP into a single non-negative value. Suppose we have two s-reps

$$\mathbf{t}_i = (\tau_i, p_{i1}, \ldots, p_{in_a}, r_{i1}, \ldots, r_{in_s}, u_{i1}, \ldots, u_{in_s})',$$

$i = 1, 2$ with the skeletal positions $p_{ij}, \in \mathbb{R}^3$ and the scale factors $\log(\tau_i), \log(r_{ij}) \in \mathbb{R}$ as Euclidean GOPs and the spoke directions $u_{ij} \in S^2$ as non-Euclidean GOPs. The vector d^1 of differences is defined by

$$\begin{aligned} d^1(\mathbf{t}_1, \mathbf{t}_2) := (&d_1(\tau_1, \tau_2),\ d_2(p_{11}, p_{21}), \ldots, d_2(p_{1n_a}, p_{2n_a}),\ d_3(r_{11}, r_{21}), \ldots, d_3(r_{1n_s}, r_{2n_s}), \\ &d_4(u_{11}, u_{21}), \ldots, d_4(u_{1n_s}, u_{2n_s}))' \end{aligned} \tag{1}$$

with appropriate partial difference measures: d_1 for the scaling factors τ_i, d_2 for the positions p_{ik}, d_3 for the spoke lengths r_{ij} and d_4 for the spoke directions u_{ij} with $i = 1, 2$, $k = 1, \ldots, n_a$ and $j = 1, \ldots, n_s$ by

$$\begin{aligned} d_1(\tau_1, \tau_2) &= |\log(\tau_2) - \log(\tau_1)|, \\ d_2(p_{1k}, p_{2k}) &= \left(\sum_{m=1}^{3} (p_{2km} - p_{1km})^2 \right)^{1/2}, \\ d_3(r_{1j}, r_{2j}) &= |\log(r_{2j}) - \log(r_{1j})|, \\ d_4(u_{1j}, u_{2j}) &= d_g(u_{1j}, u_{2j}) = \arccos(u'_{1j} u_{2j}). \end{aligned}$$

The geodesic distance function $d_g : S^2 \times S^2 \to [0, \pi]$ is defined by the arc length of the shortest great circle segment joining $u_{1j}, u_{2j} \in S^2$ and is invariant to rotation. The Euclidean metric $d_2 : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_+$ is invariant to translation and $d_1, d_3 : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ are invariant to scale. All GOP differences of

$$d^1 : (\mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times S^{2n_s}) \times (\mathbb{R}^{3n_a} \times \mathbb{R}_+^{n_s+1} \times S^{2n_s}) \longrightarrow \mathbb{R}_+^{n_a+n_s+1} \times [0, \pi]^{n_s}$$

are single non-negative values. Therewith, the hypothesis test of identical statistical distributions of two s-rep populations is given by an one-sided test,

$$H_0 : \{\mu_1 = \mu_2\} \text{ versus } H_1 \ : \ \{\mu_1 > \mu_2\}. \tag{2}$$

Given d^1, we can calculate the p-values $C_k(T_{lk})$ as described in Section 6.2.5 in the main article. In the case of a one-sided test by using difference measure d^1, we map the p-values $C_k(T_{lk})$ to the positive half of a standard Gaussian CDF by

$$\widetilde{U}_{lk} = \Phi^{-1} \left(0.5 + 0.5 \tilde{C}_k(T_{lk}) \right), \tag{3}$$

where Φ^{-1} is the inverse standard Gaussian CDF,

$$\tilde{C}_k(T_{lk}) = \frac{sc - 2}{sc} C_k(T_{lk}) + \frac{1}{sc}$$

and $sc = 10000$, $k = 1, \ldots, K$, $l = 1, \ldots, P$ similar to Section 6.2.5 in the main article.

An open problem is a sensitive mapping of $\widetilde{U}_{lk}$ to a full multivariate distribution that preserve the correlation structure of the variables. Given an appropriate mapping, the global and feature-by-feature test can be applied as described in Section 6.2.6 and 6.2.7 of the main article.

The results presented in Section 2.5 below use random signs $\tau_{lk} \in \{-1, 1\}$ that are generated for each permutation and GOP in order to map $\tilde{C}_k(T_{lk})$ to a full multivariate distribution by $U_{lk} = \tau_{lk} \widetilde{U}_{lk}$ with standard normal marginals. Thereby, we dot not preserve the correlation structure between the GOPs which results in a conservative test.

1.4 Preliminary fitting stage of s-reps to hippocampi

The hippocampus data set consists of binary images of 221 first-episode schizophrenia cases and 56 control cases as described in Section 2 in the main article. Antialiased distance images were generated from the binary images according to [4]. We selected the first 96 of the 221 SG cases to control manual work as described in the following. Based on the distance images, we used the 96 cases of SG and all cases of CG to produce appropriate preliminary fits.

Two different models were used as initializations of the fitting procedure. The first initial model $\mathbf{m}_1$ was a CPNG backwards mean of 62 hippocampus fits presented in [6]. In addition, the second initial model $\mathbf{m}_2$ was derived from the CPNG backwards mean of manually adjusted fits of the control group. The initial models $\mathbf{m}_1$ and $\mathbf{m}_2$ were pre-aligned by translation and rotation, and fit to the hippocampi of CG and SG followed by an atom and spoke stage. As a results, two fittings corresponding to $\mathbf{m}_1$ and $\mathbf{m}_2$ are obtained

for each hippocampus. The fitting with the lowest objective function were selected for further processing. The objective function value is provided by the fitting software Pablo [5] and measures the goodness-of-fit of each s-rep model to the binary data.

The 96 SG and 56 CG fits were manually evaluated and adjusted when necessary. The adjusted fittings were refit by the second atom and spoke stage in order to minimize influence of the manual adjustment on the final fittings and to ensure that all spokes match the object boundary. Let $\tilde{A}_1$ be the set of 96 fits for SG and $\tilde{A}_2$ be the set of 56 fits for CG.

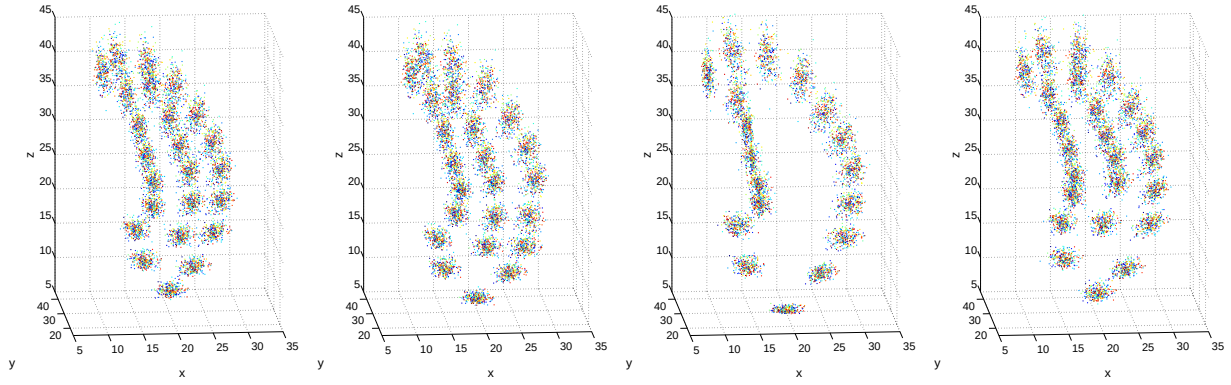
Correspondence across population is achieved by calculation of CPNG statistics. As a pre-processing step the obtained fittings must be aligned, otherwise the CPNG statistics would reflect undesirable rotational variations of the data. Therefore, the CPNG mean of the set union $\tilde{A}_1 \cup \tilde{A}_2$ was calculated. Afterwards, all fittings were translated and rotated to the mean by standard Procrustes alignment [1]. The alignment was based on the skeletal positions and not on the spoke ends, due to the CPNG analysis of the skeletal positions in a pre-shape space as described in Section 4 in the main article. Let $\bar{A}_1$ be the set of 96 aligned SG fits and $\bar{A}_2$ the set of 56 aligned CG fits. Finally, CPNG statistics were calculated for the s-rep populations $\bar{A}_1$, $\bar{A}_2$ and the pooled population $\bar{A}_1 \cup \bar{A}_2$.

2 Additional data analysis on fittings using a pooled shape distribution

The presented results in the main article are based on fittings obtained by the use of a pooled shape distribution during the CPNG stage (see Sections 7.1 in the main article). This section will present additional analyses and plots based on the same data.

2.1 Procrustes alignment of final fittings

Let $\tilde{A}$ be the obtained fittings of s-reps after the CPNG stage, final spoke stage and re-scaling into a world coordinate system as described in Section 7.1 in the main article. Figure 2 visualizes the skeletal positions and the spoke tail ends of $\tilde{A}$. Each spoke tail end is defined by the corresponding skeletal position, spoke direction and length.



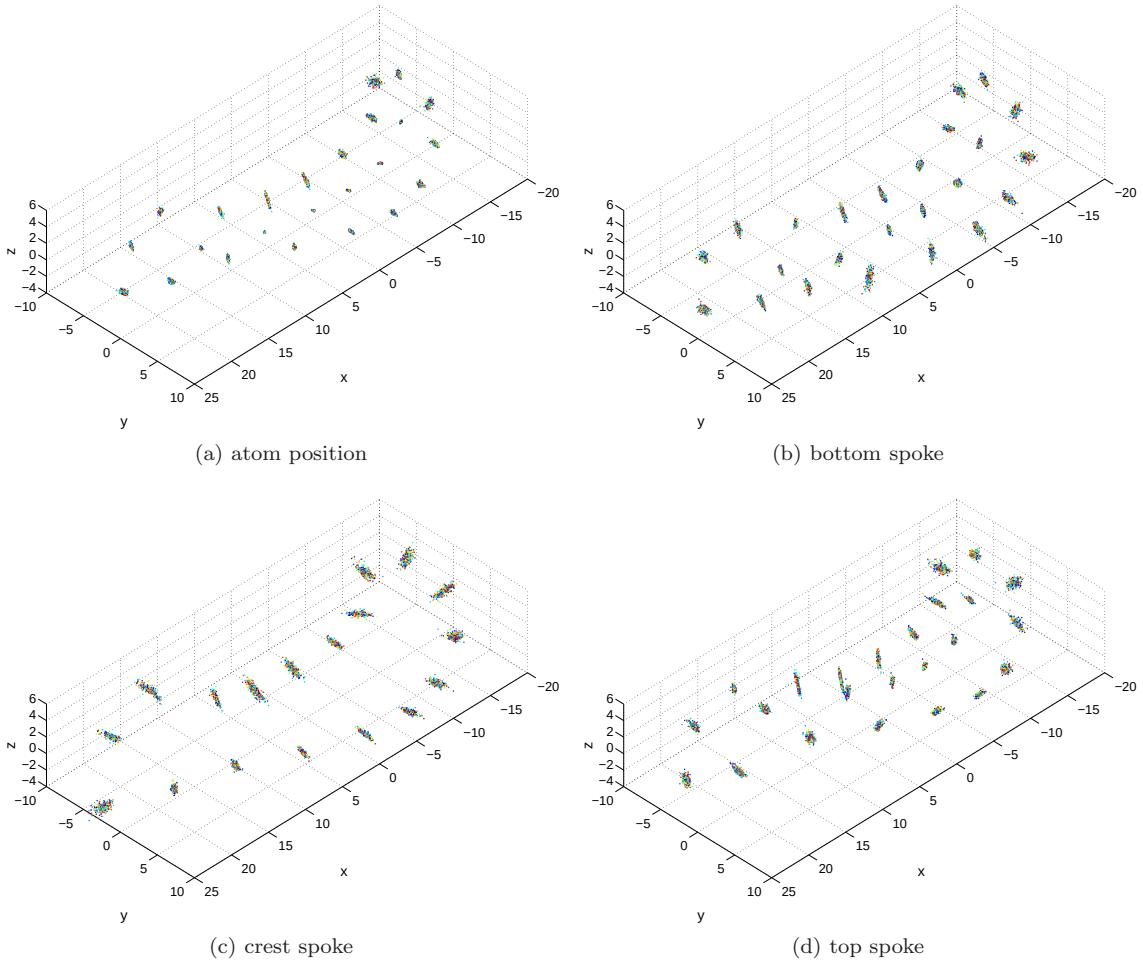
(a) atom position (b) bottom spoke (c) crest spoke (d) top spoke

Fig. 2: Final obtained s-rep fittings after the final spoke stage and re-scaling into a world coordinate system. Skeletal positions are depicted in (a). Bottom, crest and top spoke directions and lengths are depicted in (b-d) by the spoke tail ends based on the corresponding skeletal positions. The 277 fittings are represented by individual colors.

As discussed in Section 6.2.1 in the main article, an appropriate pre-processing of the data is required for a reasonable interpretation of the differences, e.g., between the latitude, longitude, x, y and z-coordinate using d^2. Let $\tilde{\mu}$ the overall backwards CPNG mean, estimated from the set union $\tilde{A}$ of obtained final fittings with

$$\tilde{A} = \tilde{A}_1 \bigcup \tilde{A}_2 = \{\tilde{\mathbf{s}}_{11}, \ldots, \tilde{\mathbf{s}}_{1N_1}, \tilde{\mathbf{s}}_{21}, \ldots, \tilde{\mathbf{s}}_{2N_2}\}.$$

The CPNG mean $\tilde{\mu}$ is translationally aligned by the subtraction of the mean of the locational components. In addition, the eigenvectors of the second moments about the center of the skeletal positions yields a rotational alignment to the x, y and z-axis. The translationally and rotationally aligned CPNG mean $\tilde{\mu}$ is called μ. Figure 3 depicts the translated, rotated and scaled s-reps of $\tilde{A}$ to μ using a standard Procrustes alignment [1], based on the skeletal positions of each s-rep $\tilde{\mathbf{s}} \in \tilde{A}$. The pre-processing removed undesirable variation from the data and enabled a meaningful interpretation for later analysis. This is highlighted by Figure 3 which shows considerable reduced variation compared to Figure 2.



(a) atom position (b) bottom spoke

(c) crest spoke (d) top spoke

Fig. 3: S-reps fittings are visualized after standard Procrustes alignment with translation, rotation and scaling based on the skeletal positions. The aligned skeletal positions are depicted in (a). Bottom, crest and top spoke directions and lengths are depicted in (b-d) by the spoke tail ends based on the corresponding skeletal positions. The 277 fittings are represented by individual colors.

2.2 Visualization of generated permutations

The distribution of $P = 1000$ permuted sample means $\hat{\nu}_{1l}$ for SG and $\hat{\nu}_{2l}$ for CG (see Section 6.2.2 in the main article) is visualized in Figure 4, $l = 1, \ldots, P$. The permuted sample means are depicted by the projections of the scaled CPNG scores matrix Z_{Comp} of $\{\hat{\nu}_{1l}, \hat{\nu}_{2l} \mid l = 1, \ldots, P\}$ (see Section 4 in the main article) onto the distance-weighted discrimination (DWD) direction and the first three orthogonal directions to the DWD direction as described in Marron et al. [3] and Qiao et al. [7]. Red circles depict permuted SG means and blue circles permuted CG means. The larger variance of CG is due to the unbalanced group size (SG contains 221 cases and CG 56 cases). The observed Gaussian distributions indicate appropriate permutation sets.

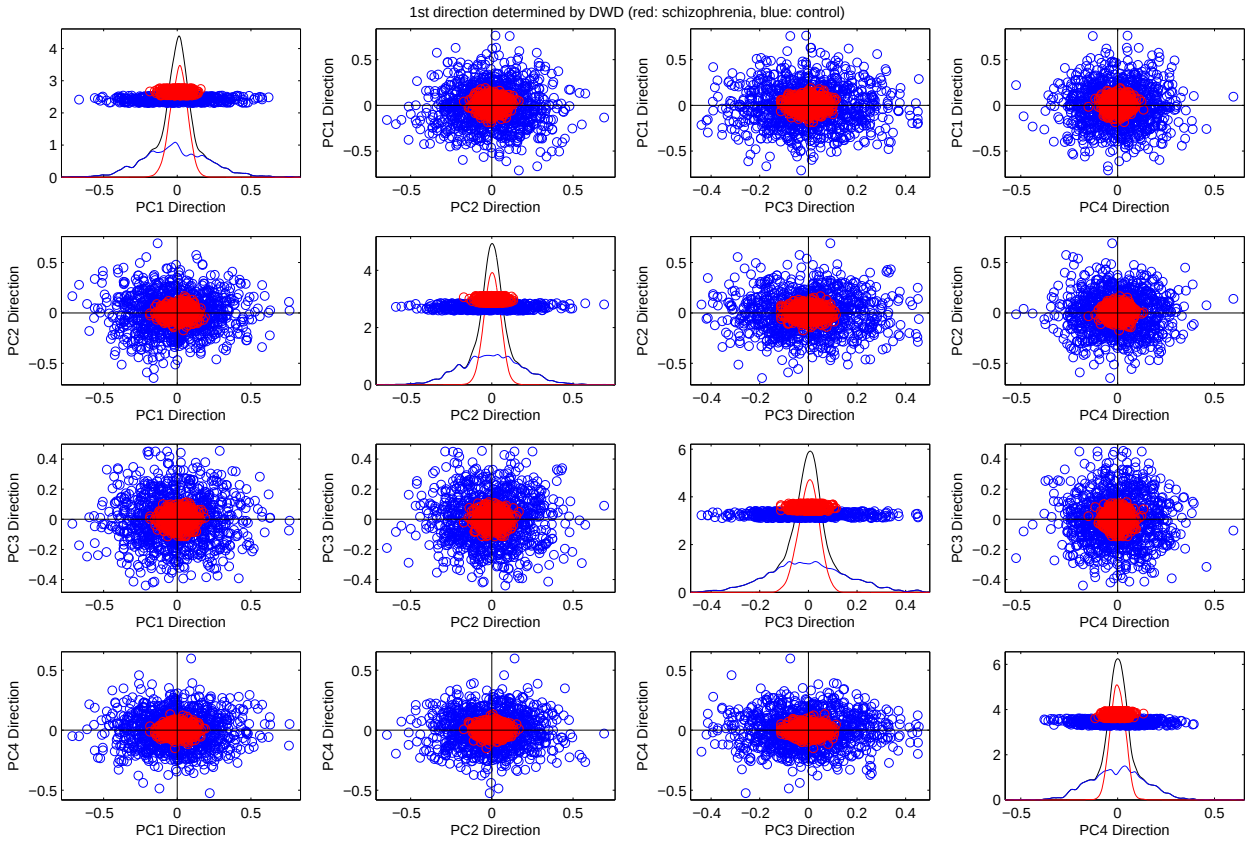


Fig. 4: Scatter plots and jitterplots (diagonal) with KDE are showing the distribution of permuted sample means projected on the DWD direction and the first three orthogonal directions to the DWD direction. Additionally, the KDE of the pooled distribution of SG and CG is shown in the jitterplots. Red circles depict permuted SG means and blue circles permuted CG means.

2.3 DiProPerm results using a MD test statistic and a DWD projection direction

Figure 5 visualizes the DiProPerm test reported in Table 1 in Section 7.2 in the main article using a mean difference (MD) test statistic and DWD as the projection direction. The DiProPerm test is based on the evaluation of the scaled CPNG scores matrix Z_{Comp} as described in Section 4 in the main article. The DiProPerm test is a global test and the hypothesis of identical mean between the two populations was rejected given a significance level $\alpha = 0.05$.

2.4 ROC analysis compared to feature-by-feature test results using distance measure d^2 and PP1

This section evaluates the performance of the feature-by-feature test by Receiver Operating Characteristic (ROC) curves. The ROC analysis gives a curve lying in $[0,1]\times[0,1]$, which quantifies the amount of "overlap" of each GOP between the samples of the two populations. The ROC curve resulting from the observed data is visualized by a red line in the following plots. In addition, for each permutation a ROC curve is generated, represented by a blue line, which results in an envelope under the null distribution. In the following, each envelope is visualized by the first $1,000$ of the $30,000$ permutations. A ROC curve of the observed data close to the boundary of this envelope indicates a significant feature. The comparison is done using the distance measure d^2 and the standard pre-processing of the data as described in Section 6.2.1 in the main article. The GOPs that represent latitude and longitude of the spoke direction are normalized corresponding to the mean shift as explained in Section 6.2.4 in the main article.

The feature-by-feature test results are reported in Figures 7 and 8 in Section 7.3 in the main article. Several GOPs were tested as statistically significant including the global scaling factor $|U_{0K}| = 2.7627$ given

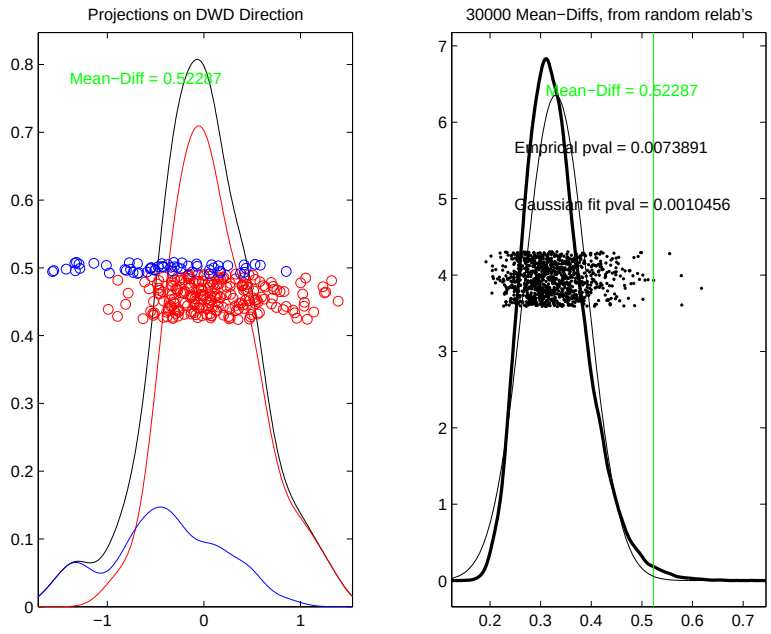


Fig. 5: The DiProPerm hypothesis test of mean differences based on the scaled CPNG scores matrix Z_{Comp} of the final fittings after pre-processing by PP1. DiProPerm is a two sample mean hypothesis test. The left plot shows a jitterplot by the projection of the data on the DWD direction together with the kernel density estimates (KDEs) of the distribution of SG (red circles), CG (blue circles) and the set union SG∪CG. The right plot shows a jitterplot of the mean differences of the 30,000 permutations, a KDE of the distribution of the MD test statistic in addition to the MD between the observed population SG and CG (green line).

a corrected threshold $\lambda = 2.2917$. Figure 6 depicts the ROC curve for the global scaling factor (red) together with the envelope (blue) obtained from the permutations. A major part of the red curve is located close to the boundary of the envelope. Thus, Figure 6 indicates a significant GOP in agreement with the obtained feature-by-feature test result.

The area under the curve (AUC) value is a simple numerical summary which is useful for a comparison of several ROC curves, e.g., a comparison of the ROC curves between the figures below.

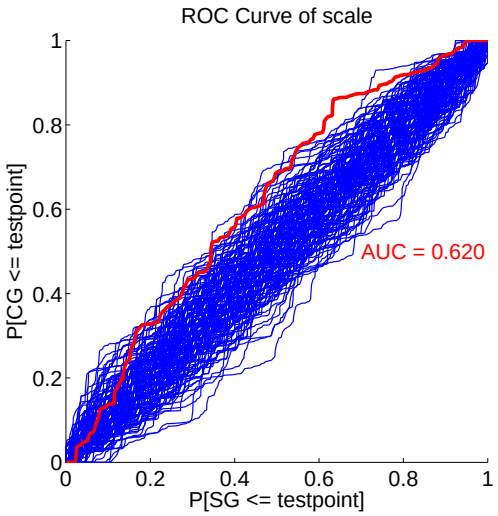
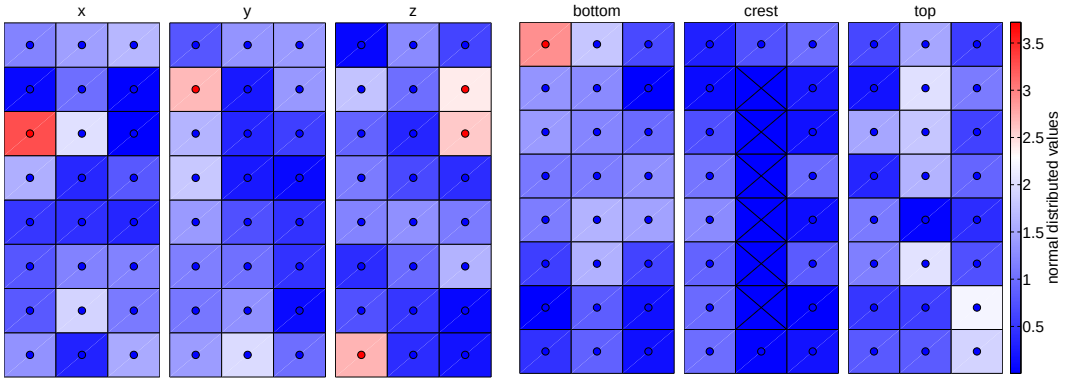
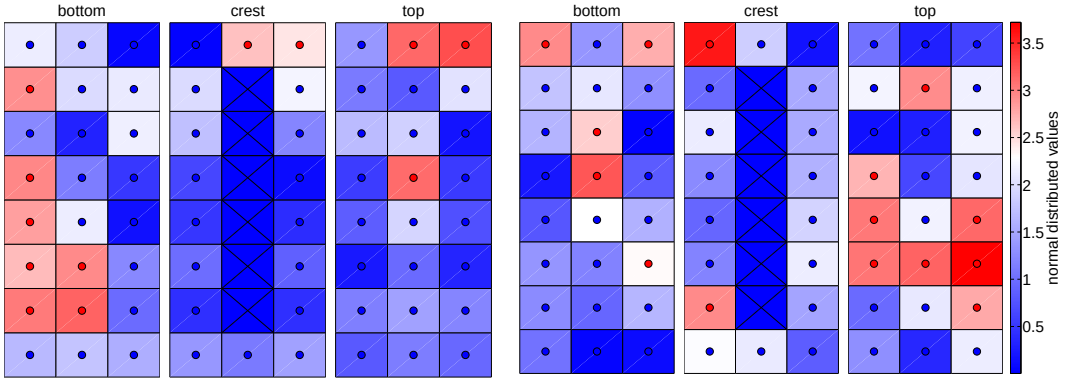


Fig. 6: The ROC curve of the global scaling factor (red) is visualized together with the envelope (blue) obtained from the permutations.

Figure 7 below is identical to Figure 8 in the main article and shows the magnitude of significance of each GOP using the difference measure d^2. In order to simplify the visualization all standard normal values $U_{0k}, k = 1, \ldots, K$ are presented in absolute values. The color map is non-linear defined from blue to white to red. The corrected threshold $\lambda = 2.2917$ defines the color white, blue and red visualize non-significant and significant values, respectively. Blocks which show a white color have U_{0k} around the threshold λ. The blue small circles inside each block mark whether a U_{0k} is less than or equal to the threshold λ. Red small circles mark if an U_{0k} is greater than the threshold λ and therewith statistical significant.



(a) skeletal position (b) spoke length



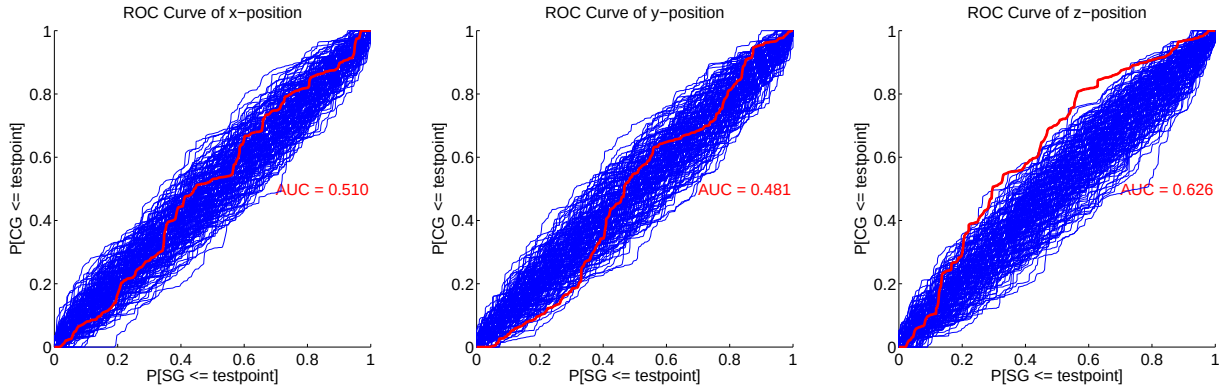
(c) latitude spoke direction (d) longitude spoke direction

Fig. 7: Colored significant map of U_{0k} using difference measure d^2 with a corrected threshold $\lambda = 2.2917$. Each box corresponds to a GOP. The color map on the left side is non-linear and has a range from blue (not significant) to white (λ) to red (significant). The circle inside each box marks whether an U_{0k} is less or equal than the threshold λ (symbolized by blue) or if an U_{0k} is greater than the threshold λ (symbolized by red).

The results are presented on the basis of the 3×8 skeletal sheet such as the 24 skeletal x-positions in Figure 7a. The skeletal sheet is numbered from bottom to top and from left to right, i.e., atom 1 correspond to the the left bottom block, atom 8 to the left top block, atom 9 to the middle bottom block, atom 16 to the middle top block, atom 17 to the right bottom block and finally, atom 24 correspond to the right top block. In the following, we compare results for selected GOPs from Figure 7 with the ROC analysis.

Figure 8 visualizes the ROC curve of the skeletal x, y and z-position of atom 22. Figure 7a indicates the z-position of atom 22 as statistically significant. The x and y-position are not statistically significant whereas the x-position shows a lower value than the y-position of atom 22. These results are reflected in Figure 8 by the ROC analysis. The ROC curve for the x-position of atom 22 is located close to the center of the envelope, the ROC curve for the y-position is located closer to the boundary of the envelope in some regions whereas the ROC curve for the z-position is close to the boundary in major parts of the envelope.

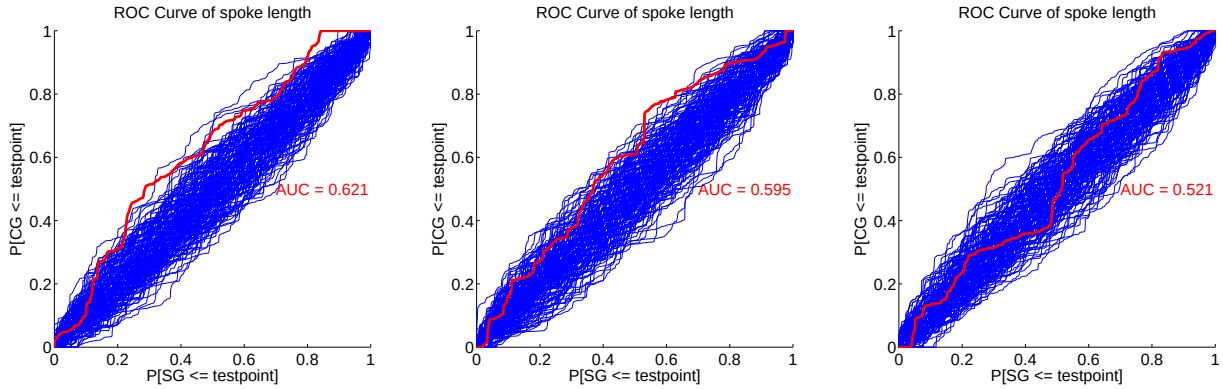
Figure 9 visualizes the ROC curve of the bottom spoke lengths of atom 8, 16 and 24. Figure 7b indicates the bottom spoke length of atom 8 as statistically significant whereas the bottom lengths of atom 16 and 24 are not significant. Furthermore, atom 24 shows a lower value than atom 16. These observations are reflected in the ROC analysis and the AUC values in Figure 9. The ROC curve in Figure 9a is located closer to



(a) atom 22 (x-position) (b) atom 22 (y-position) (c) atom 22 (z-position)

Fig. 8: ROC curves are visualized for (a) the x-position, (b) the y-position and (c) the z-position of atom 22 from the skeletal 3×8 sheet. The blue lines depict the ROC curves from the permutations and define an envelope. The red line depicts the ROC curve between the observed samples of two populations.

the boundary of the envelope than the ROC curve in Figure 9b, and again more than the ROC curve in Figure 9c.



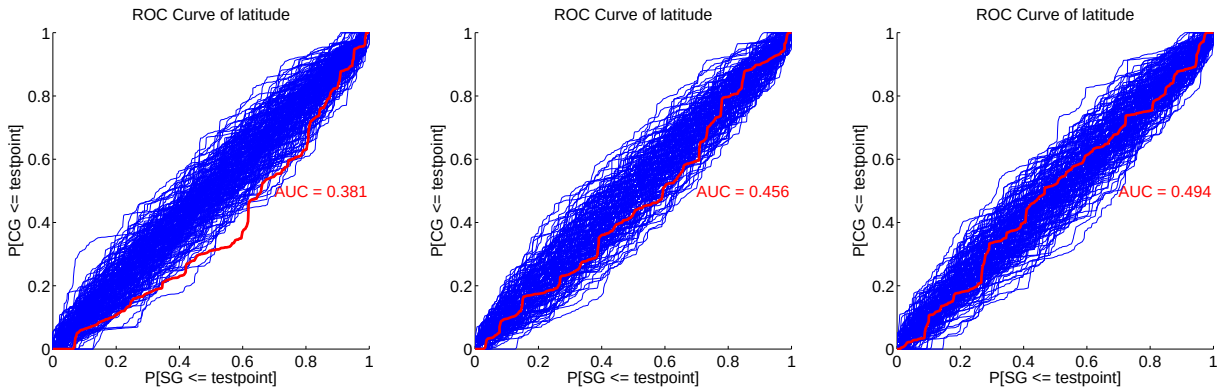
(a) atom 8 (bottom) (b) atom 16 (bottom) (c) atom 24 (bottom)

Fig. 9: ROC curves are visualized for the spoke lengths for (a) atom 8, (b) atom 16 and (c) atom 24 on the bottom side of the skeletal 3×8 sheet. The blue lines depict the ROC curves from the permutations and define an envelope. The red line depicts the ROC curve between the observed samples of two populations.

Figure 10 visualizes the ROC curve of the latitude spoke directions of atom 3 on the bottom, crest and top of skeletal sheet. Figure 7c indicates the latitude spoke direction of atom 3 on the bottom of the skeletal sheet as statistically significant whereas the latitude spoke direction on the crest and top are not significant. The box color of the top latitude spoke direction of atom 3 reflects a smaller value than the crest latitude spoke direction of atom 3. As above, all observations are reflected by the corresponding ROC curves in Figure 10.

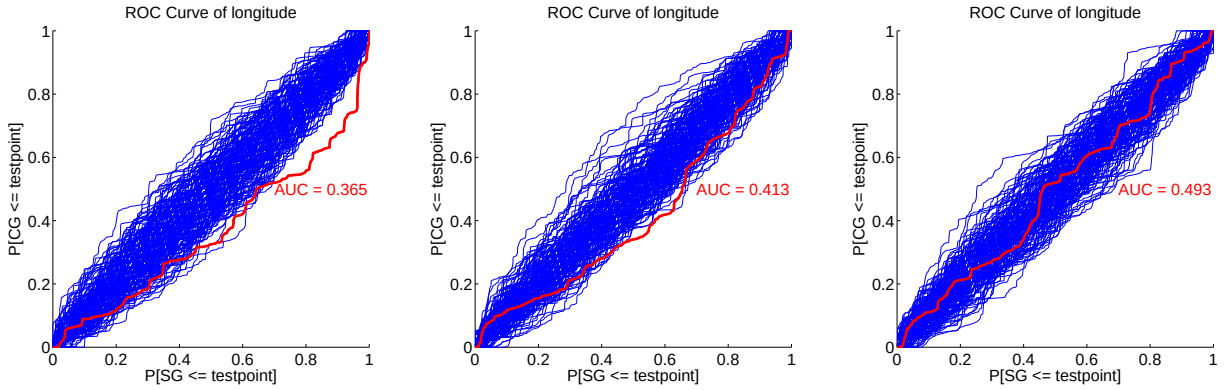
Finally, Figure 11 visualizes the ROC curve of the longitude spoke direction on the crest of atom 8, 16 and 24. Figure 7d indicates a statistically significant longitude spoke direction of atom 8 on the crest of the skeletal sheet whereas the longitude spoke direction on the crest of atom 16 and 24 are not significant. The color for atom 24 reflects a considerably smaller value than for atom 16. A comparison with Figure 11 confirms these observations. The ROC curve in Figure 11a is mostly located outside or close to the boundary of the envelope wheres the ROC curve of Figure 11c is close to the center of the envelope.

The observations described in this section verify the correctness of the feature-by-feature test results on the basis of selected GOPs. The ROC visualization of all 271 GOPs described by the distance measure d^2 was omitted for the purpose of clarity of this article .



(a) atom 6 (bottom) (b) atom 6 (crest) (c) atom 6 (top)

Fig. 10: ROC curves are visualized for the spoke latitude directions for atom 3 on (a) the bottom, (b) the crest and (c) the top of the skeletal 3×8 sheet. The blue lines depict the ROC curves from the permutations and define an envelope. The red line depicts the ROC curve between the observed samples of two populations.



(a) atom 8 (crest) (b) atom 16 (crest) (c) atom 24 (crest)

Fig. 11: ROC curves are visualized for the spoke longitude directions for (a) atom 8, (b) atom 16 and (c) atom 24 on the crest of the skeletal 3×8 sheet. The blue lines depict the ROC curves from the permutations and define an envelope. The red line depicts the ROC curve between the observed samples of two populations.

2.5 Test results for the unsigned difference measure d^1

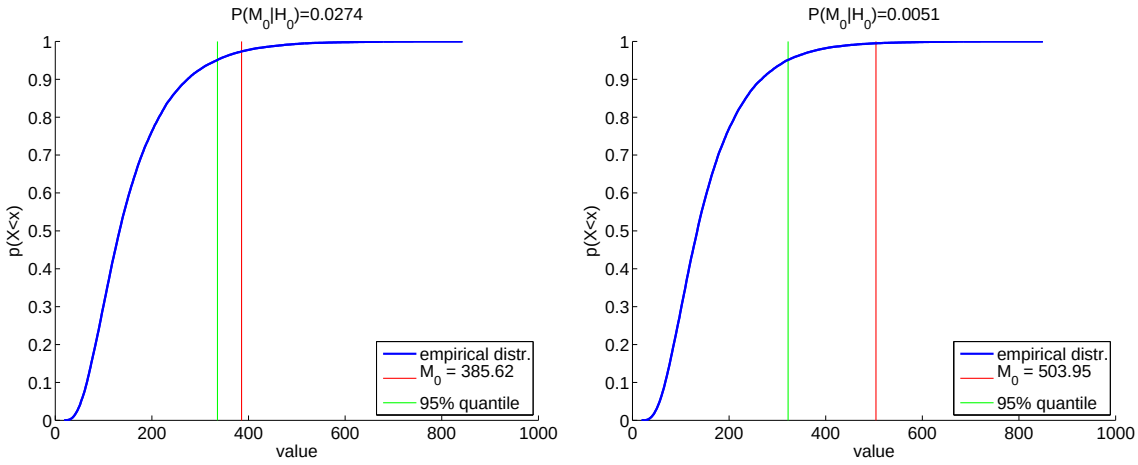
This section reports hypothesis test results using distances measure d^1 as described in Section 1.3. Results are based on the pre-processing methods PP1 and PP2 as described in Section 7.2 in the main article.

2.5.1 Global test results using d^1

Figure 12 shows the global test results for difference measures d^1 using PP1 and PP2. The global hypothesis of equal sample means is rejected and a statistical significant difference between the shape distribution of SG and CG is established ($p = 0.0274$ for PP1 and $p = 0.0051$ for PP2 with $p = P(M_0|H_0)$). These results correspond to the results using d^2 ($p = 0.0109$ for PP1 and $p = 0.0029$ for PP2) as presented in Section 7.2 in the main article. The larger p-values for d^1 are due to less information is being used for the unsigned differences, because the correlation structure between the GOPs was removed after the applied mapping to a full multivariate Gaussian as described in Section 1.3. Thus, results presented in the main article are quantified by the conservative test results in this section.

2.5.2 Single GOP test results using d^1

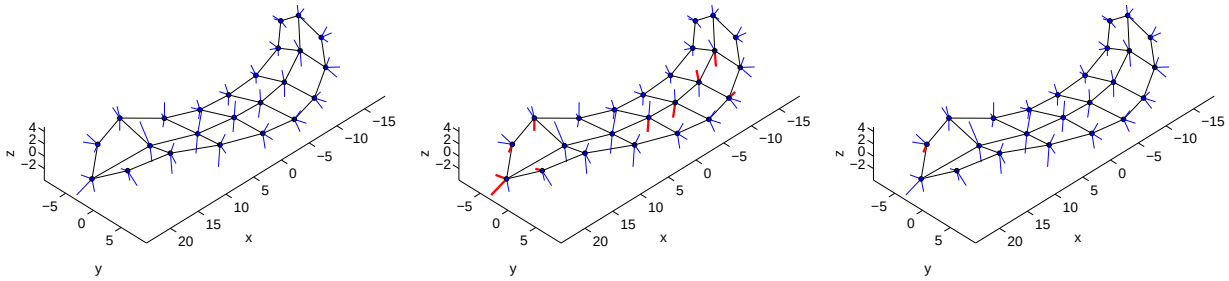
Figures 13 and 14 visualize the feature-by-feature test results for the difference measure d^1 using PP1. Recall that each discrete slabular s-rep is organized into 24 atoms by a 3×8 grid. Thereby, the measure d^1 (see



(a) PP1 (b) PP2

Fig. 12: Global test results using PP1 in (a) and PP2 in (b). The empirical distribution of $M_l, l = 1, \ldots, 30,000$ is shown together with M_0 and the 95% quantile of the empirical distribution.

Section 1.3) results in 157 GOPs with 24 GOPs corresponding to the skeletal position of each atom, 66 GOPs for the spoke directions (bottom, crest and top), 66 GOPs for the spoke lengths (bottom, crest and top) and 1 GOP for the global scaling factor. Figure 14 shows the magnitude of significance as described for Figure 7 in Section 2.4. The corrected threshold from the feature-by-feature test is $\lambda = 2.5532$.

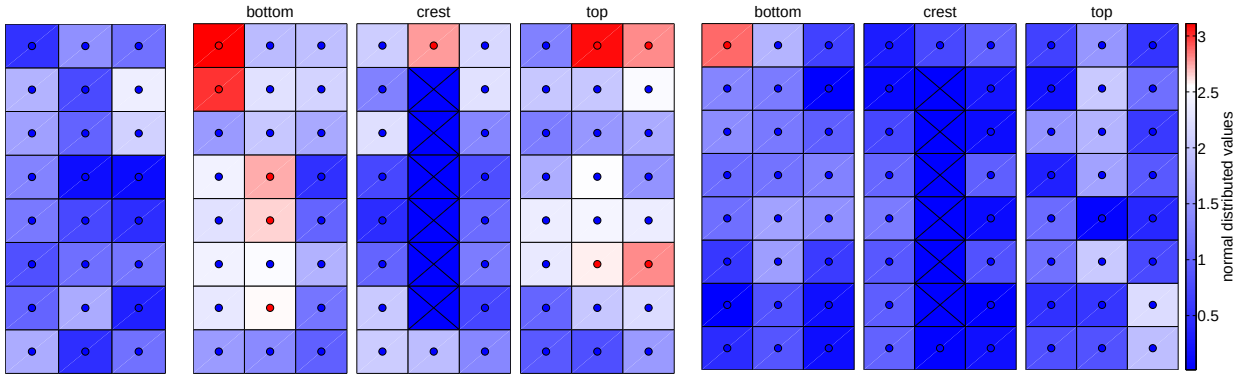


(a) skeletal position (b) spoke direction (c) spoke length

Fig. 13: Significant GOPs using PP1 and difference measure d^1 based on the 3×8 skeletal sheet of the SG CPNG mean. Test results are shown in (a) for the skeletal positions, in (b) for the spoke directions and in (c) for the spoke lengths. No skeletal position is statistically significant where non-significant skeletal positions are marked by small blue circles and significant skeletal positions are marked by large red circles. Similar, non-significant spoke directions and lengths are marked by small blue lines whereas significant spoke directions and lengths are marked by wide red lines.

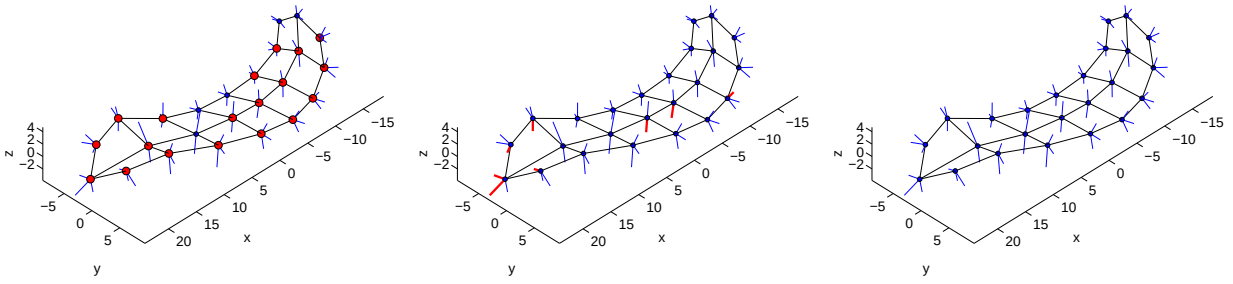
Figures 13 and 14 show several statistically significant GOPs. No skeletal position but one spoke length and 10 spoke directions are statistically significant. Moreover, the global scaling factor τ between SG and CG was found statistically significant by the GOP $|U_{0K}| = 2.7704$.

Figures 15 and 16 are identical to both previous figures except for the use of PP2 instead of PP1. Several skeletal positions are statistically significant in contrast to Figures 13a and 14a with no statistically significant skeletal position. The volume difference between the two populations is reflected by the skeletal positions using d^1 and PP2. Thus, Figures 15a and 16a show rather significant differences from a global deformation than from local deformations. Figures 14c and 16c show only small differences, which reveals that the global volume information is described by scaling of the skeletal grid. The spoke lengths are designed to capture only local differences whereas the skeletal position captures global scale differences. Similar results between spoke directions are expected because of the scaling invariance of $u_{ij} \in S^2$.



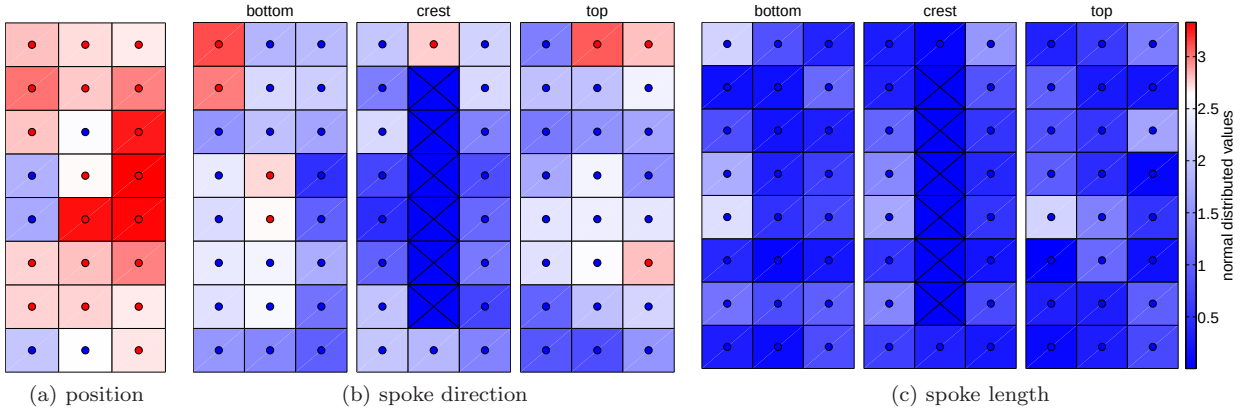
(a) position (b) spoke direction (c) spoke length

Fig. 14: Colored significant map of U_{0k} with a corrected threshold $\lambda = 2.5532$ using PP1 and difference measure d^1. Each box represents a GOP which correspond to a skeletal atom. The color map on the left side is non-linear and has a range from blue (not significant) to white (λ) to red (significant). The circle inside each box marks whether an U_{0k} is less or equal than the threshold λ (symbolized by blue) or if an U_{0k} is greater than the threshold λ (symbolized by red).



(a) skeletal position (b) spoke direction (c) spoke length

Fig. 15: As Figure 13, now based on PP2 and difference measure d^1.



(a) position (b) spoke direction (c) spoke length

Fig. 16: As Figure 14, now based on PP2 and difference measure d^1 with a corrected threshold $\lambda = 2.6368$.

A comparison of the results in this section with Section 7.3 in the main article leads to very similar observations and conclusions. Thereby, the results in the main article are quantified by the conservative test results presented in this section which not use the correlation structure between the GOPs (see Section 1.3). This is reflected by less significant GOPs, in particular for the spoke directions.

Using difference measure d^2 a significant volume difference was observed in the x and y-directions but not in the z-direction for the aligned hippocampi. Thus, we could obtain additional information using d^2 compared to d^1.

2.6 Asymptotic behavior of the global test for the two difference measures d^1 and d^2

This section will study the asymptotic behavior of the global test (described in Section 6.2 in the main article) for an increasing permutation size using PP1. The reported empirical p-values are 0.0274 for d^1 and 0.0109 for d^2 using $30,000$ permutations and given a significance level of $\alpha = 0.05$.

We have randomly selected subsets of $P = 500, 1000, 1500, 2000, 2500, \ldots, 29500$ from the set of 30000 permutations and applied the proposed testing procedure of Section 6.2 in the main article. Figure 17 visualizes the results and indicates a stabilization of the p-value from the global test after around $10,000$ permutations. Surprisingly, we observe a p-value equal to zero for a very small permutation size. This section will show the Mahalanobis space as the cause on the basis of distance measure d^1.

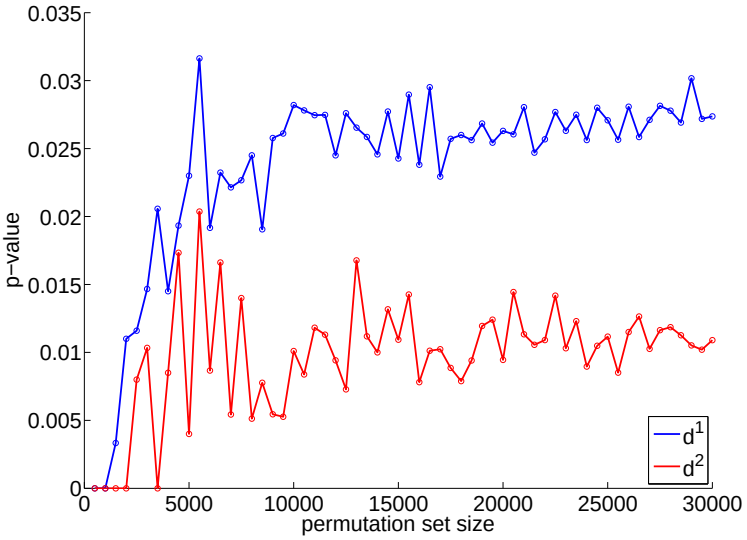


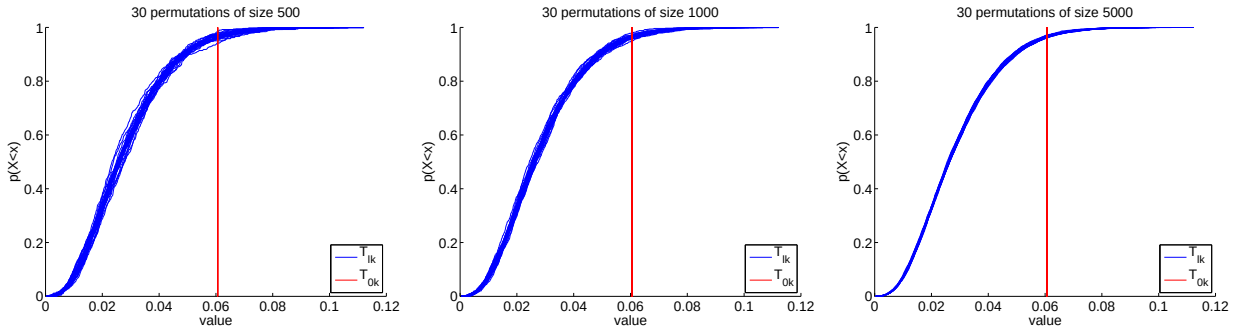
Fig. 17: The p-values are plotted against the number of permutations using difference measures d^1 and d^2. 30000 permutation were generated. The hypothesis test was calculated on randomly chosen subsets with $500, 1000, 1500, 2000, 2500, \ldots, 29500$ permutations.

In order to elaborate the convergence behavior of d^1, we have generated 30 random permutation sets with 500, 1000 and 5000 permutations for each permutation set. Afterwards, we applied the proposed testing procedure of Section 6.2 in the main article.

First, we calculated the difference measure $T_l = d^1(\mathbf{t}_{1l}, \mathbf{t}_{2l})$ (see Section 6.2.5 in the main article) between the s-reps $\mathbf{t}_{1l}$ and $\mathbf{t}_{2l}$, $l = 1, \ldots, P$ where P is the number of permutations. Each blue line in Figure 18 shows the cumulative empirical distribution for the chosen element $k = 22$ from the 157 dimensional GOP d^1-difference vector T_l. The selected element describes the atom position 22 from the 3×8 skeletal grid. Each plot contains 30 cumulative empirical distributions (blue lines) corresponding to each permutation set. We observe a higher variance of the envelope for a smaller permutation set size. $T_0 = d^1(\mathbf{t}_1, \mathbf{t}_2)$ is identical for all 30 permutation sets.

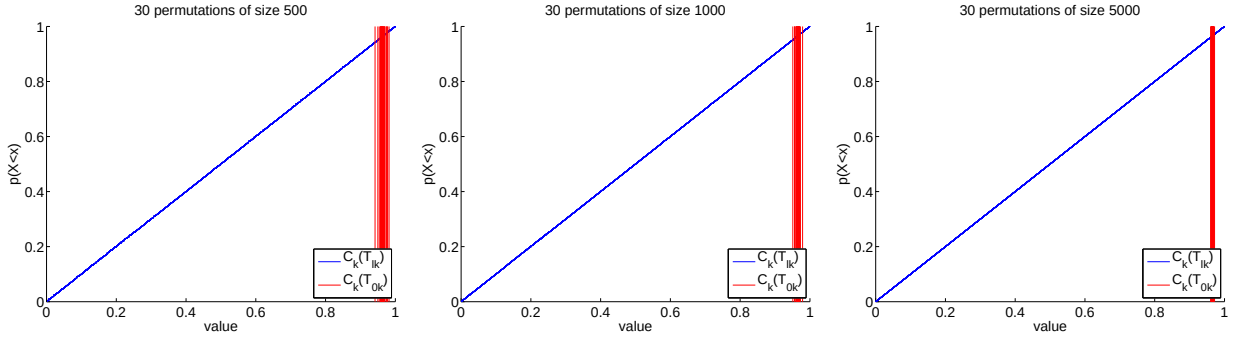
Afterwards, we estimated the empirical cumulative functions C_k for $k = 1, \ldots, K$ partial tests following to Section 6.2.5 in the main article. As a result, we obtained for each GOP difference a p-value $C_k(T_{lk})$, and $C_k(T_{0k})$ respectively. The cumulative empirical distribution of the calculated p-values are depicted in Figure 19. The p-values of the 30 permutations sets have by construction a uniform distribution. Therefore, no variance is visible between the blue line in Figures 19a-19c. However, we observe a larger variance of the red line for smaller permutation set size. The cumulative function C_k bases on the empirical distribution which shows larger variation for a smaller permutation set size in Figure 18. Therefore, the observed larger variance between $C_k(T_{0k})$ (red line) can be expected.

Subsequently, we calculated standard normal distributed variables from the uniformly distributed p-values by the inverse cumulative normal distribution function as described in the previous Section 1.3. Figure 20 visualizes the calculated standard normal distributed variables U_{lk} (blue) and U_{0k} (red). The blue and red lines show a larger variance for smaller permutation set size. However, the mean of T_{0k}, $C_k(T_{0k})$ and U_{0k} is similar for different permutation set size.



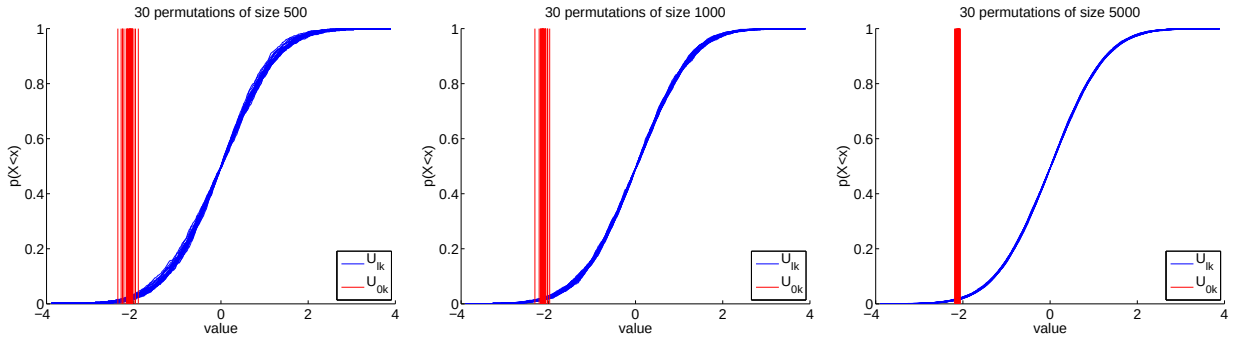
(a) 500 (b) 1000 (c) 5000

Fig. 18: The cumulative empirical distributions of GOP differences are depicted for a selected GOP using difference measure d^1. Each plot visualizes 30 random permutation sets of sizes 500, 1000 and 5000 (corresponding to 30 blue lines in each plot). The selected GOP is the atom position 22.



(a) 500 (b) 1000 (c) 5000

Fig. 19: The cumulative empirical distributions of the p-values $C_k(T_{lk})$ (blue) are depicted together with $C_k(T_{0k})$ (red) using difference measure d^1. Each plot visualizes 30 random permutation sets of sizes 500, 1000 and 5000. The selected GOP is the atom position 22.

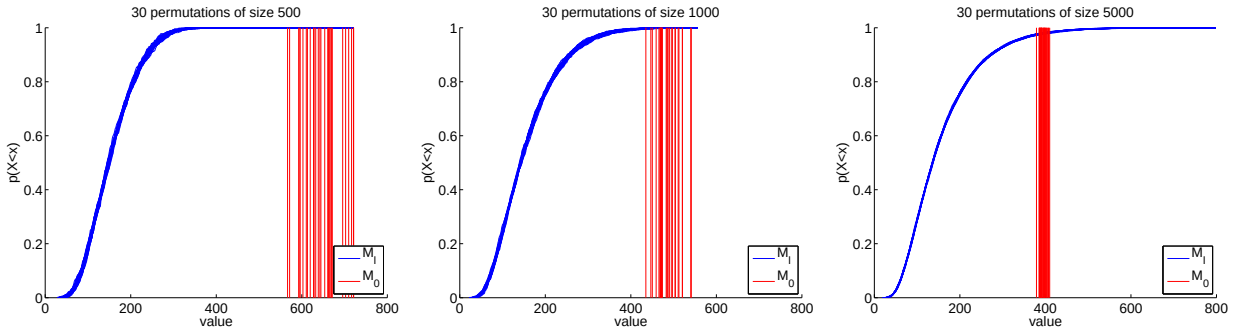


(a) 500 (b) 1000 (c) 5000

Fig. 20: The cumulative empirical distributions of the standard normal variables U_{lk} (blue) are visualized together with U_{0k} (red) using difference measure d^1. Each plot visualizes 30 random permutation sets of sizes 500, 1000 and 5000. The selected GOP is the atom position 22.

Finally, the p-values of the global tests were obtained by the estimation of the covariance matrix $\hat{\Sigma}_U$ from U_{lk} and the Mahalanobis distance as a combining function (see Section 6.2.6 in the main article). For each permutation $l = 1, \ldots, P$, we obtained the Mahalanobis distance M_l in addition to M_0 between the two populations SG and CG. Figure 21 shows the Mahalanobis distance for the three different permutation

set sizes. A smaller permutation set size strongly increase the variance of M_0. In addition, the blue curves indicate a smaller slope for higher permutation set size. In contrast to the previous figures, we observe a change in the mean value of M_0 with a larger value for smaller permutation set size. As a result, $p(M_0)$ is 0 (see equation (9) in the main article) using a small permutation set size such as 500 because $H(M_l, M_0) = 0$ for all $l = 1, \ldots, P$.



(a) 500 (b) 1000 (c) 5000

Fig. 21: Cumulative empirical distributions of Mahalanobis distances M_l (blue) are visualized together with M_0 (red) using difference measure d^1. Each plot visualizes 30 random permutation sets of sizes 500, 1000 and 5000. The selected GOP is the atom position 22.

Figures 18 to 21 and additional simulations on the covariance matrix found the covariance matrix as the reason for the convergence behavior in Figure 17. The Mahalanobis distance combines all GOPs to a corrected global test by the covariance matrix $\hat{\Sigma}_U$. A smaller permutation set size increases the magnitude of the elements of the covariance matrix, i.e., leads to a larger variance between the matrix elements of $\hat{\Sigma}_U$. As a result, the covariance matrix assigns different weights to the GOPs by the Mahalanobis distance.

Therefore, we recommend a permutation set size greater than $10,000$ for the proposed global hypothesis test. The study of an alternative combining functions for the global hypothesis test is left for future research.

3 Data analysis on an alternative group of final fittings

Besides the obtained final fittings using a joint shape distribution during the CPNG stage as described in Section 7.1 in the main article, we have generated a second group of final fittings derived from CPNG stages using a pooled shape distribution (FG1), two individual shape distributions (FG2) and two individual interchanged shape distributions (FG3). Interchanged shape distributions means the use of the estimated individual CG shape distribution for the re-fitting of the SG population during the CPNG stage, and the individual SG shape distribution for the re-fitting of the CG population. In each CPNG stage, the obtained backward mean was translational and rotational aligned to the data, i.e, the alignment of the CPNG backward mean of

1. $\bar{A}_1 \cup \bar{A}_2$ to the 221 and 56 CG cases for FG1,
2. $\bar{A}_1$ to the 221 SG cases and of $\bar{A}_2$ to the 56 CG cases for FG2,
3. $\bar{A}_2$ to the 221 SG cases and of $\bar{A}_1$ to the 56 CG cases for FG3.

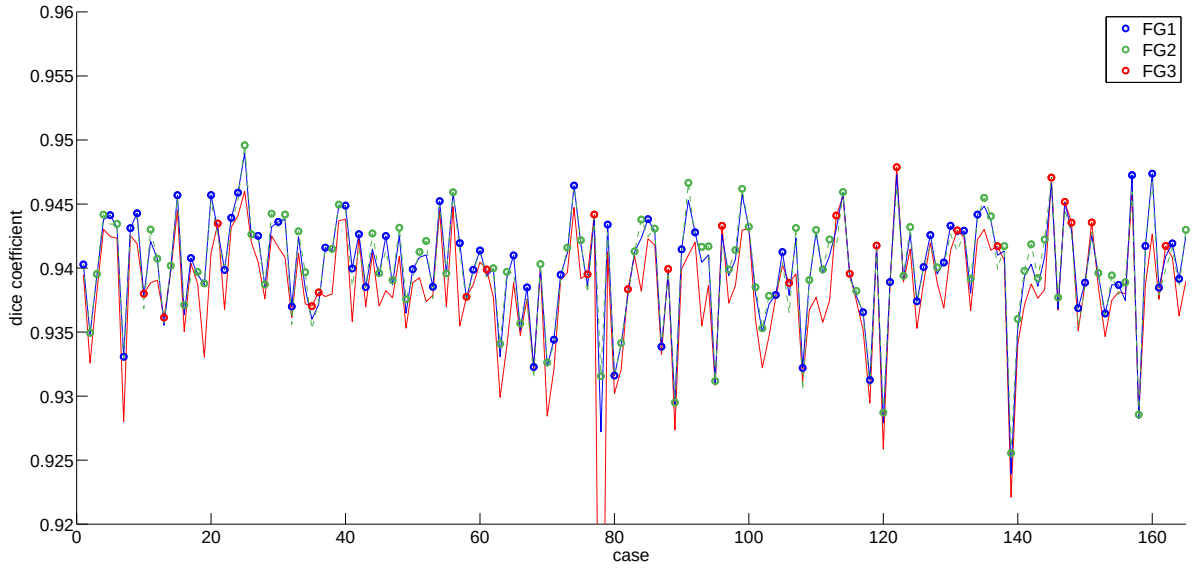
Afterwards, the means were optimized inside the CPNG shape space with an additional final spoke stage (see Section 5 in the main article). As a result, we obtained three fittings for each hippocampus. We chose the fitting with the largest Dice similarity coefficient. The Dice coefficient is a measure of the volume and was calculated between the original binary image B_1 and the binary image B_2 generated from each fitting. The coefficient is defined by

$$d_{vol}(B_1, B_2) = 2\frac{|B_1 \cap B_2|}{|B_1| + |B_2|} \tag{4}$$

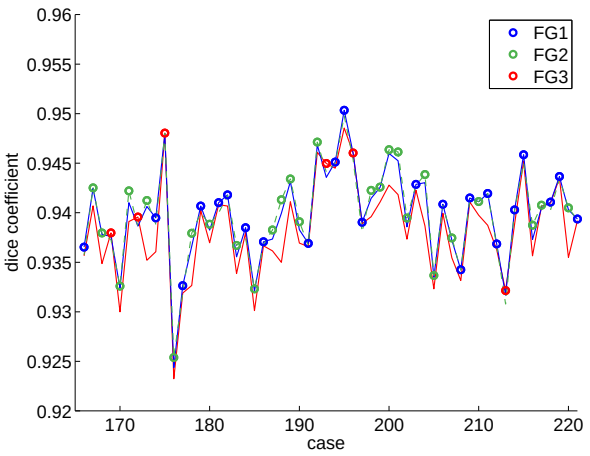
where $|\cdot|$ denotes the number of voxels that describe hippocampal tissue. Figure 22 shows the Dice coefficients of SG and CG for all three fitting types. Accordingly, the second group of final SG fittings consist of 84

fittings from FG1, 107 fittings from FG2 and 30 fittings from FG3. The second group of final CG fittings consist of 18 fittings from FG1, 21 fittings from FG2 and 17 fittings from FG3.

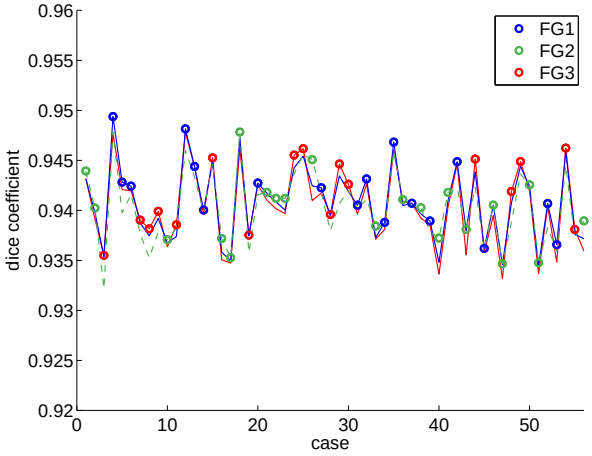
Figure 22 shows also an average volume overlap of 94% for both groups which indicates accurate fittings. We observe an outlier for case 73 of SG for FG3 due to a poor fitting result. The variance of the Dice coefficient is small for both groups. Nevertheless, a larger variance inside SG can be observed. Moreover, we can observe that FG1 and FG2 leads to a comparable Dice coefficient. The Dice coefficient of FG3 is inferior to FG1 and FG2 for SG but comparable for CG. There are two reasons for this observation. First, schizophrenia is a heterogeneous disease and also contains hippocampi variations between healthy patients. Therefore, the interchanged shape distribution from the schizophrenia cases can also describe the control cases. Second, both populations have an unbalanced size with a higher number of schizophrenics.



(a) SG (cases 1-165)



(b) SG (cases 166-221)

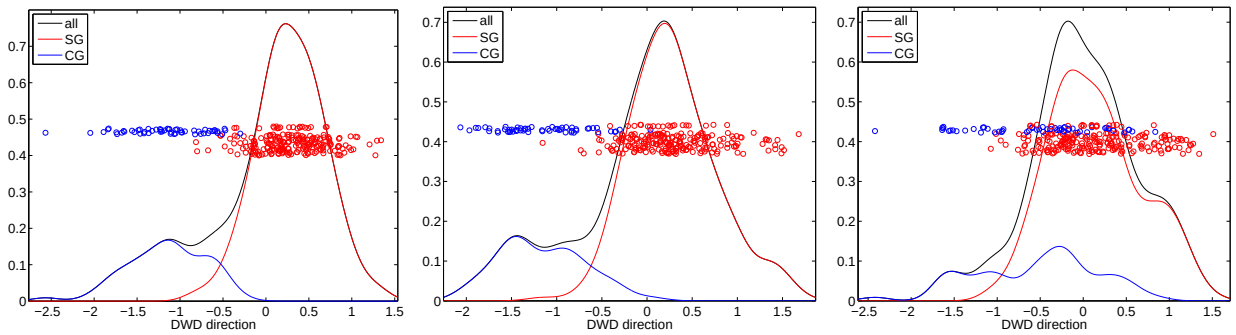


(c) CG

Fig. 22: Dice coefficient between the final fittings for (a-b) SG and (c) CG. The coefficient is depicted for the three types of obtained fittings by using a pooled shape distribution (FG1), two individual distributions (FG2) and two interchanged individual distributions (FG3) during the CPNG stage. The maximal Dice coefficient is depicted by a circle for each case colored by the corresponding class. The solid and dashed lines connect all points of the corresponding classes and depict the variance. SG shows larger variance than CG in correspondence with the heterogeneous character of the schizophrenia disease.

In addition to Figure 5 in the main article, Figure 23 shows the distribution of of SG and CG fittings obtained from (a) two individual distributions during the CPNG stage, (b) two interchanged individual

distributions and (c) of SG and CG fittings selected by the Dice criteria. The distributions are visualized by the projections of the CPNG score matrix Z_{Comp} on the DWD direction. Figures 23a and 23b show high separation properties between SG and CG. In contrast, a difference between the populations is not very strongly visible in Figure 23c which visualizes the second group of final fittings. The group is a compromise between independent fittings and a small bias as discussed in Section 7.1 in the main article.



(a) individual (b) interchanged (c) selected

Fig. 23: Jitterplot and KDEs show the distribution of SG and CG fittings projected onto the DWD direction. SG and CG fittings are obtained by using (a) two individual distributions during the CPNG stage, (b) two interchanged individual distributions during the CPNG stage and (c) by a selection of the final fittings using the Dice criteria. Additionally, the KDE of the pooled distribution of SG and CG is shown (all). A difference between the populations is visible for (a) and (b) but not very strong in (c).

The obtained second group of final fittings were used to test each of the hypotheses

$$H_0 : \{\mu_1 = \mu_2\} \text{ versus } H_1 \ : \ \{\mu_1 > \mu_2\} \qquad \text{(one-sided)} \tag{5}$$

for a one-sided test in case the difference measure is unsigned (e.g., d^1) and

$$H_0 : \{\mu_1 = \mu_2\} \text{ versus } H_1 \ : \ \{\mu_1 \neq \mu_2\} \qquad \text{(two-sided)} \tag{6}$$

for a two-sided test in case the difference measure is signed (e.g., d^2). The hypotheses are tested by the proposed global and feature-by-feature test in Section 6.2 in the main article at a significance level of $\alpha = 0.05$.

3.1 Global test results

Table 1 shows the global test results for the difference measures d^1 and d^2 for the two different pre-processing methods. Both difference measures rejected the hypothesis of equal population means and established a statistical significant difference between the two populations. In addition, DiProPerm results are reported in Table 1. All reported values are consistent with the results obtained from fittings using a pooled shape distribution, see Table 1 in the main article and Section 2.5 above. We observe an overall improved p-value in Table 1, particularly for the difference measure d^2. Thus, the second group of final fittings reveals an improved separation of the two populations, schizophrenics and controls.

3.2 Single GOP test results

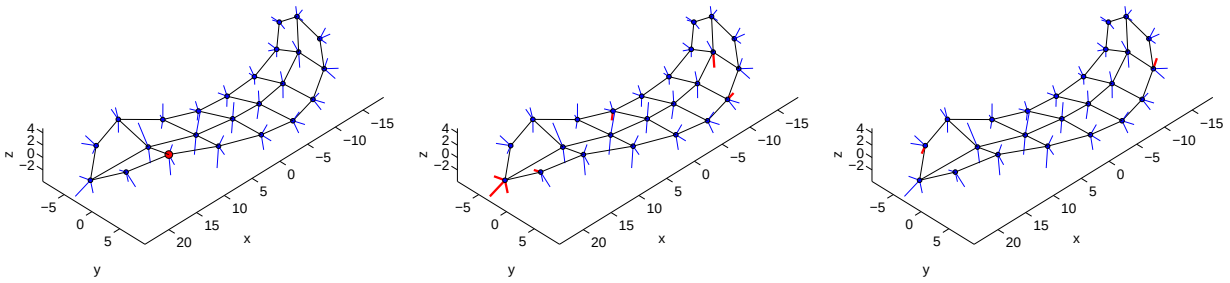
This section presents feature-by-feature test results for the two distance measures d^1 and d^2 using PP1. We have left out additional results for PP2 because neither additional informations nor conclusions would be added to this section.

Figures 24 and 25 visualize the feature-by-feature test results for the difference measure d^1 and correspond to Figures 13 and 14 above. The corrected threshold is $\lambda = 2.5632$. The measure d^1 results in 157 GOPs with

Table 1: Empirical p-value results using difference measures d^1 and d^2 for the proposed global hypothesis test in comparison with results obtained by DiProPerm. Two different pre-processing steps were applied: (PP1) Full Procrustes alignment with scaling. (PP2) Full Procrustes alignment without scaling. Three different projection directions were used for DiProPerm.

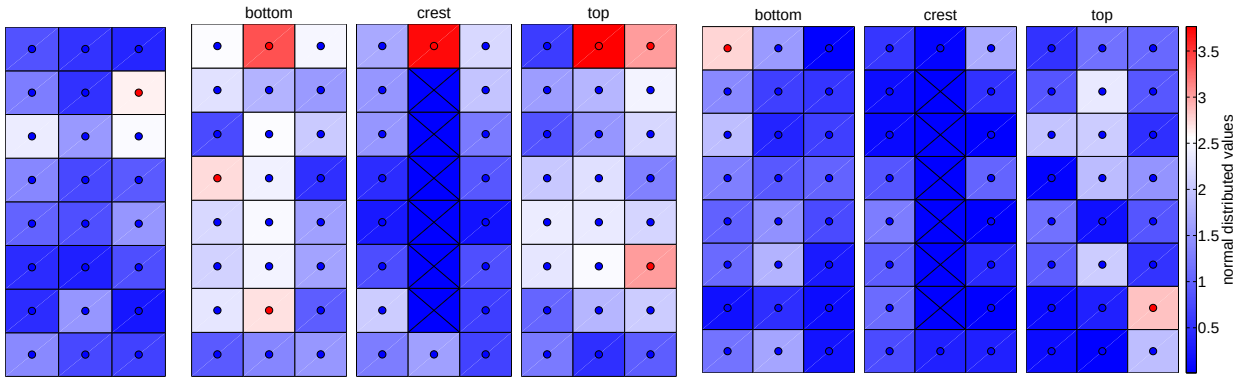
method	empirical p-value PP1	PP2
Mahalanobis distance		
difference measure d^1	0.0245	0.0043
difference measure d^2	0.0013	0.0009
DiProPerm using MD-statistic		
DWD direction vector	0.0018	0.0011
SVM direction vector	0.0039	0.0051

24 GOPs corresponding to the skeletal position of each atom, 66 GOPs for the spoke directions (bottom, crest and top), 66 GOPs for the spoke lengths (bottom, crest and top) and one GOP for the global scaling factor. Figure 24 and 25 show statistically significant GOPs. One skeletal position, two spoke lengths and 7 spoke directions are statistically significant compared to Figure 13 above where no skeletal position but one spoke length and 10 spoke directions are statistically significant. Moreover, the global scaling factor τ between SG and CG was found statistically significant.



(a) skeletal position (b) spoke direction (c) spoke length

Fig. 24: As Figure 13, now based on PP1, difference measure d^1 and the alternative group final fittings.



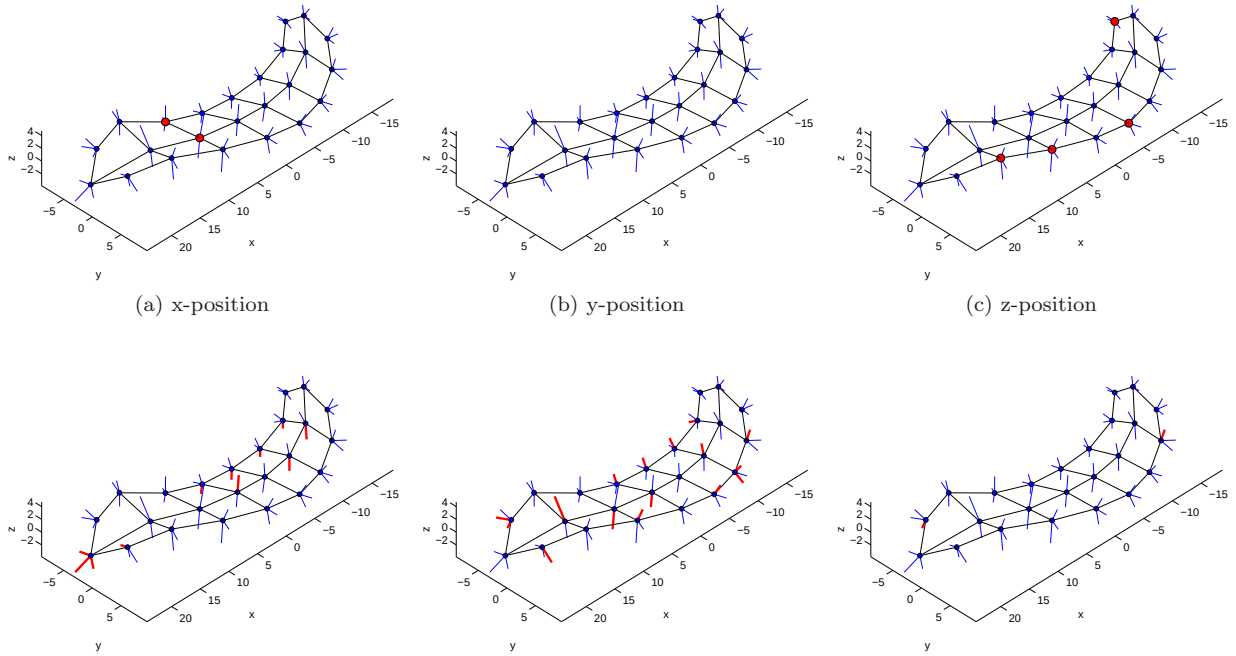
(a) position (b) spoke direction (c) spoke length

Fig. 25: As Figure 14, now based on PP1, difference measure d^1 and the alternative group final fittings with a corrected threshold $\lambda = 2.5632$.

Figure 25 shows the magnitude of significance as described for Figure 7 in the previous Section 2.4. The corrected threshold from the feature-by-feature test is $\lambda = 2.5632$. The GOP $|U_{0K}| = 2.7388$ is statistically significant where the index K corresponds to the global scale factor τ. A comparison of Figure 25 with

Figure 14 above shows a very similar pattern between the colored significant maps except the pattern between the bottom spoke directions. In the previous Figure 14, we observe two significant atoms 7 and 8 (top right of the skeletal sheet) and two significant atoms 12 and 13 (center middle) which are not significant in Figure 25. A detailed interpretation of this observation is left as an open question for the future. However, the second group of the final s-reps reflects tighter fittings based on the Dice coefficient. Therefore, the two populations are better separated, which decreases noise artifacts and yields a larger threshold $\lambda = 2.5632$ compared to $\lambda = 2.5532$ in Section 2.5.2.

Figures 26 and 27 visualize the feature-by-feature test results for the difference measure d^2 and correspond to Figures 7 and 8 in the main article. The measure d^2 results in 271 GOPs with 72 GOPs corresponding to the skeletal position of each atom (x, y and z-position), 66 GOPs for the latitude spoke directions (bottom, crest and top), 66 GOPs for the longitude spoke directions (bottom, crest and top), 66 GOPs for the spoke lengths (bottom, crest and top) and one GOP for the global scaling factor. The corrected threshold is $\lambda = 2.5214$. Figure 26 and 27 show statistically significant GOPs. Two skeletal x-positions, no y-position, 4 z-positions, one bottom, no crest and one top spoke lengths, 7 bottom, one crest and three top latitude spoke directions, 5 bottom, two crest and 9 top latitude spoke directions are statistically significant. Moreover, the GOP $|U_{0K}|$ is 2.7198 and is statistically significant where the index K corresponds to the global scale factor τ.



(a) x-position (b) y-position (c) z-position

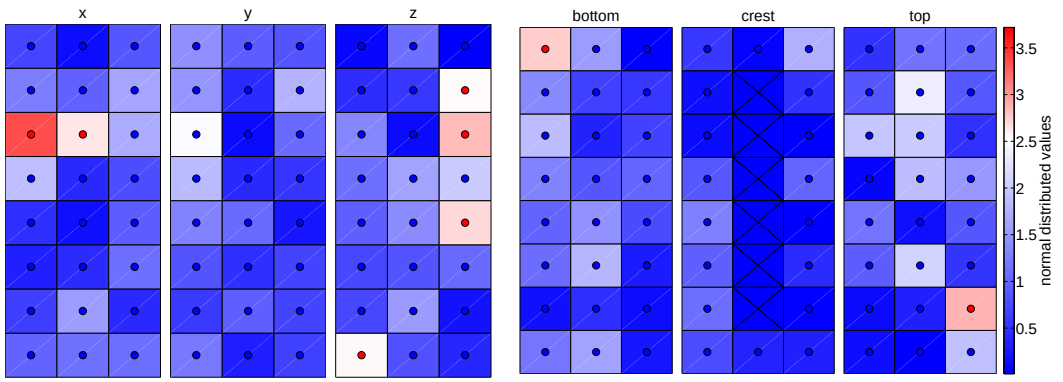
(d) latitude direction (e) longitude direction (f) spoke length

Fig. 26: As Figure 13, now based on PP1, difference measure d^2 and the alternative group final fittings.

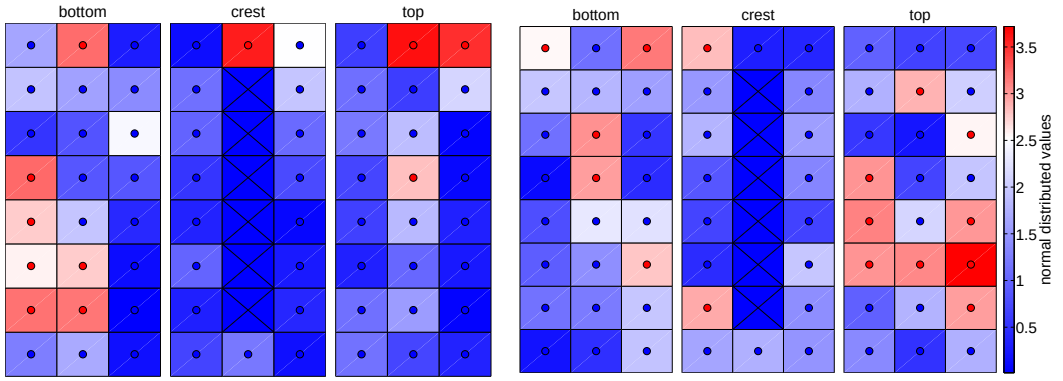
As before, a comparison of Figure 27 with Figure 8 in the main article shows a very similar pattern between the colored significant maps. The lower color intensity for several boxes in Figure 27 is due to a larger threshold $\lambda = 2.5214$ compared to $\lambda = 2.2917$ in the main article.

3.3 Conclusion

The additional data analysis by the second group of final fittings in this section confirms the results and conclusions of the main article and Section 2 above. The global test results establish smaller p-values compared to the results from the first group of final fittings. This indicates a better separation of the two populations by the second group of final fittings. The feature-by-feature test show similar patterns between the colored significant maps and demonstrate therewith the sensitivity of the proposed test in the case of less separated fittings.



(a) skeletal position (b) spoke length



(c) latitude spoke direction (d) longitude spoke direction

Fig. 27: As Figure 14, now based on PP1, difference measure d^2 and the alternative group final fittings with a corrected threshold $\lambda = 2.5214$.

References

1. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. John Wiley & Sons, Chichester (1998)
2. Jung, S., Dryden, I.L., Marron, J.S.: Analysis of principal nested spheres. Biometrika **99**(3), 551–568 (2012)
3. Marron, J.S., Todd, M.J., Ahn, J.: Distance weighted discrimination. J. Amer. Statist. Assoc. **102**(480), 1267–1271 (2007)
4. Niethammer, M., Juttukonda, M.R., Pizer, S.M., Saboo, R.R.: Anti-aliasing slice-segmented medical images via Laplacian of curvature flow. In preparation (2013)
5. Nitrc: S-rep fitting, statistics, and segmentation. http://www.nitrc.org/projects/sreps (2013)
6. Pizer, S.M., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J.N., Huckemann, S., Marron, J.S.: Nested sphere statistics of skeletal models. In: Innovations for Shape Analysis: Models and Algorithms, Lecture Notes in Comput. Sci., pp. 93–115. Springer (2013)
7. Qiao, X., Zhang, H.H., Liu, Y., Todd, M.J., Marron, J.S.: Weighted distance weighted discrimination and its asymptotic properties. J. Amer. Statist. Assoc. **105**(489), 401–414 (2010)