Supplementary Material

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

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

November 17, 2013

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 of 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

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

F. Godtliebsen
Department of Mathematics and Statistics, UiT The Arctic University of Norway, 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

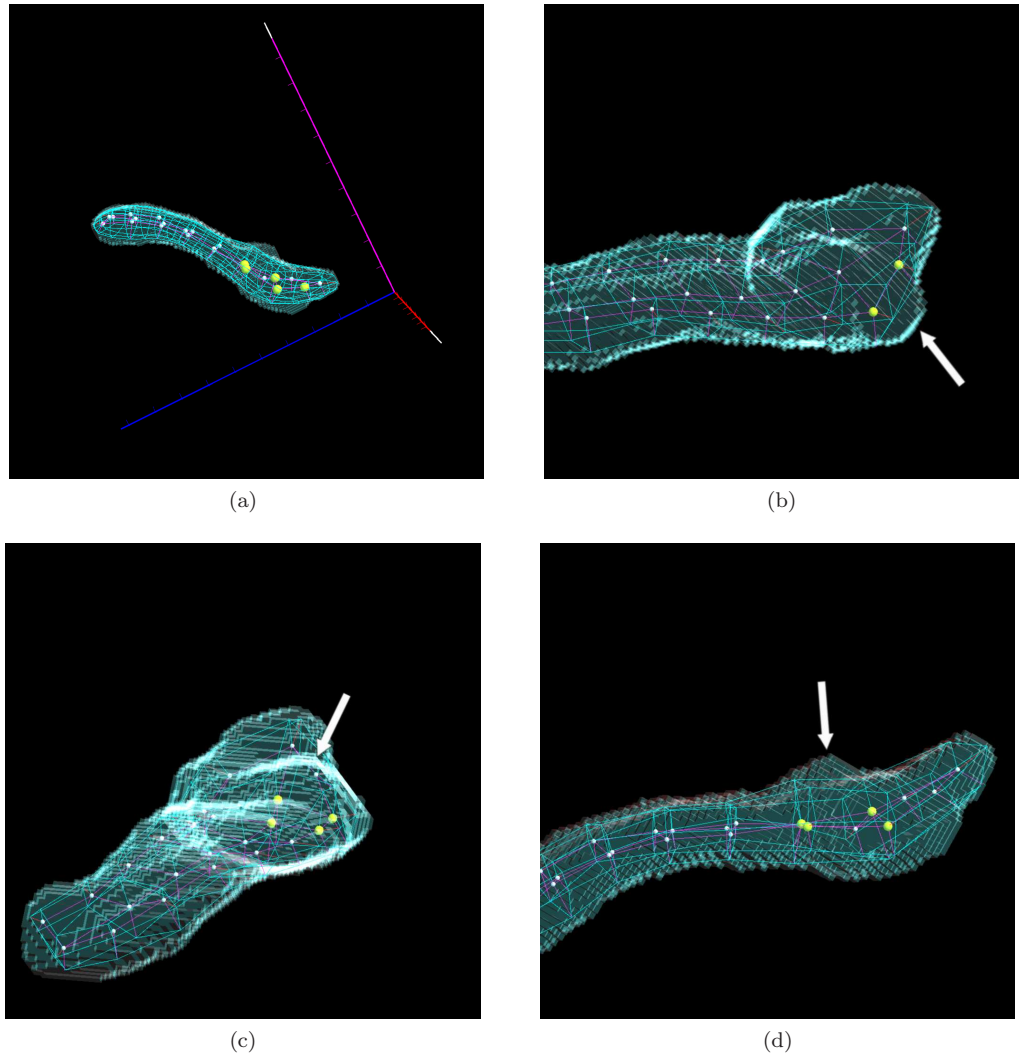


Fig. 1: Final fit of a hippocampus with bumps that are not tightly described by an s-rep based on a (3×8) grid. (a) Entire 3D view to the s-rep with corresponding coordinate system. (b) Bump on the side located between two highlighted hub positions. (c-d) Bump on the top located between four highlighted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

permuted set union can be split into two subgroups G_{1l} and G_{2l} with the same number of elements as G_1 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 do not preserve the correlation structure between the GOPs, which results in a conservative test.

An alternative to the mapping of $\widetilde{U}_{lk}$ to a full multivariate distribution is the calculation of a corrected threshold for the p-values $C_k(T_{lk})$ using Copulas. A Copula is a multivariate distribution function $C \colon [0, 1]^K \to [0, 1]$ with uniform marginals in $[0, 1]$. The implementation of such a procedure is left for the future.

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 result, two fittings corresponding to $\mathbf{m}_1$ and $\mathbf{m}_2$ were 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.

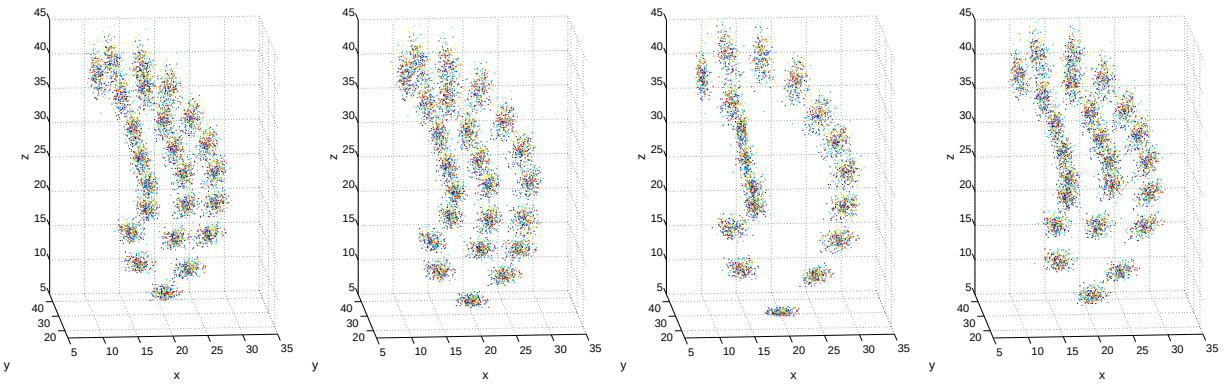
Correspondence across population is achieved by calculation of CPNG statistics. As a pre-processing step the obtained fittings must be aligned, otherwise the CPNG statistics would reflect undesirable rotational variations of the data. Therefore, the CPNG mean of the set union $\tilde{A}_1 \cup \tilde{A}_2$ 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.



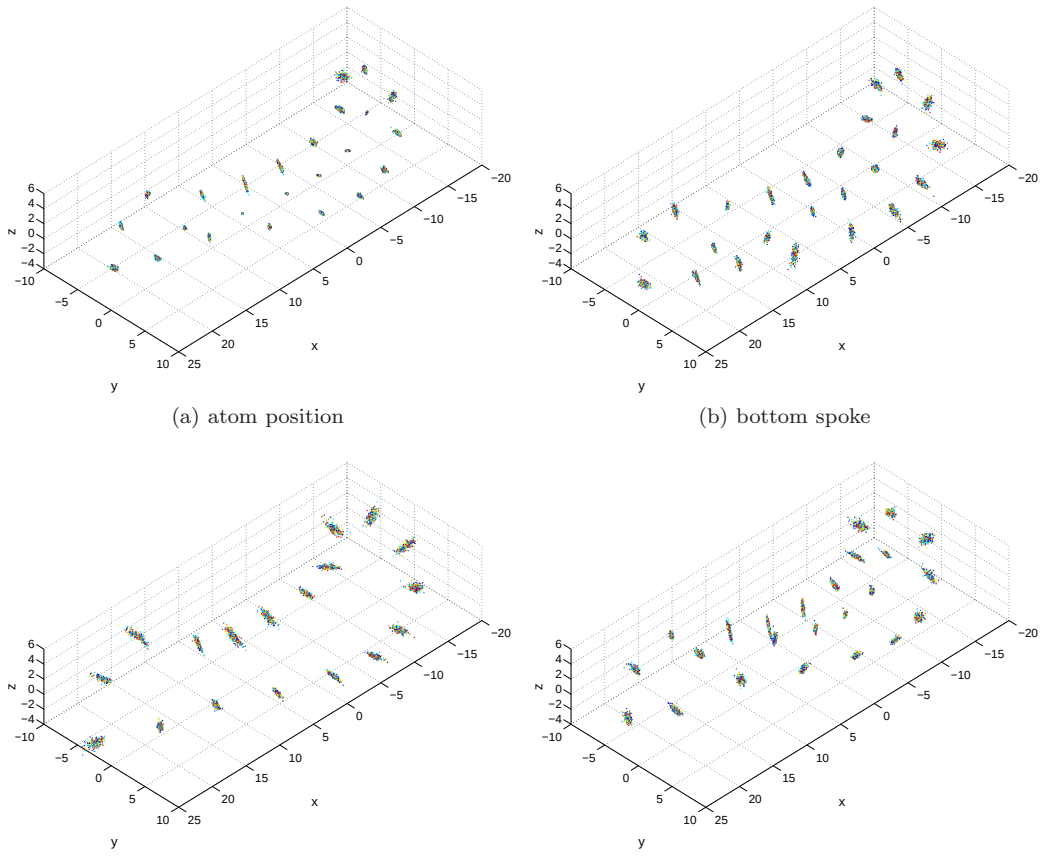
(a) atom position (b) bottom spoke (c) crest spoke (d) top spoke

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.

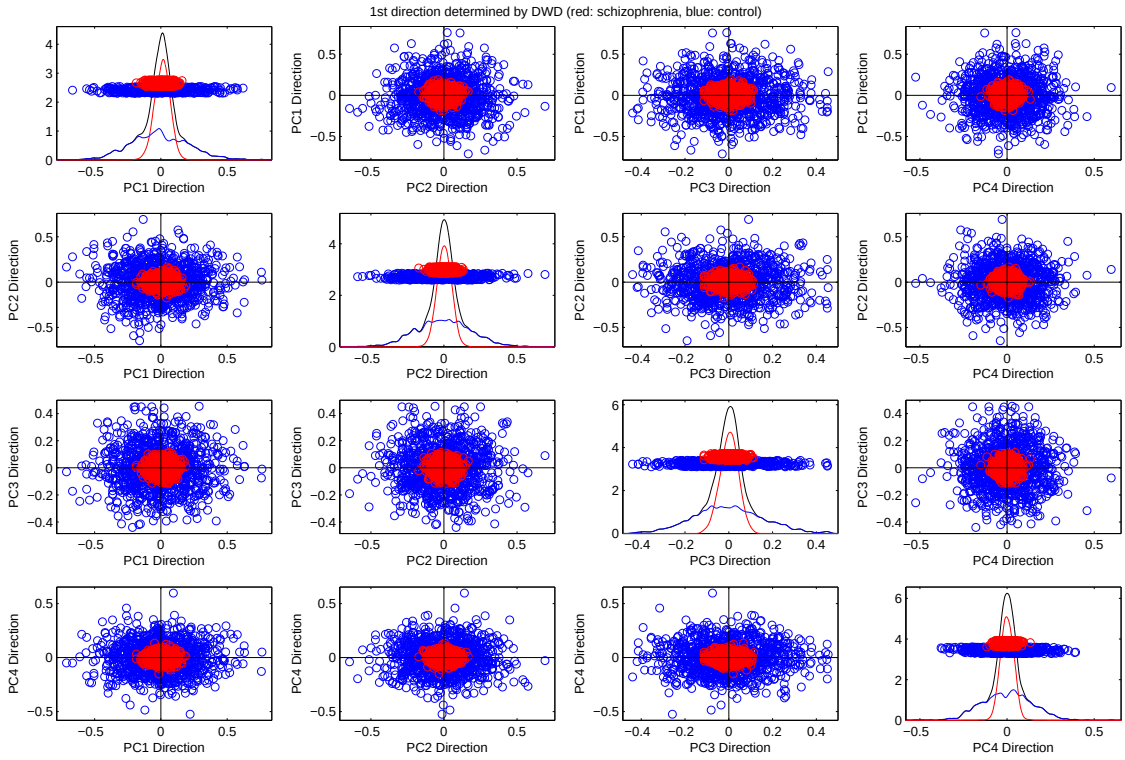


Fig. 4: Scatter plots and jitterplots (diagonal) with KDE are showing the distribution of permuted sample means projected on the DWD direction and the first three orthogonal directions to the DWD direction. Additionally, the KDE of the pooled distribution of SG and CG is shown in the jitterplots. Red 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 [8] reported for PP1 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. 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.

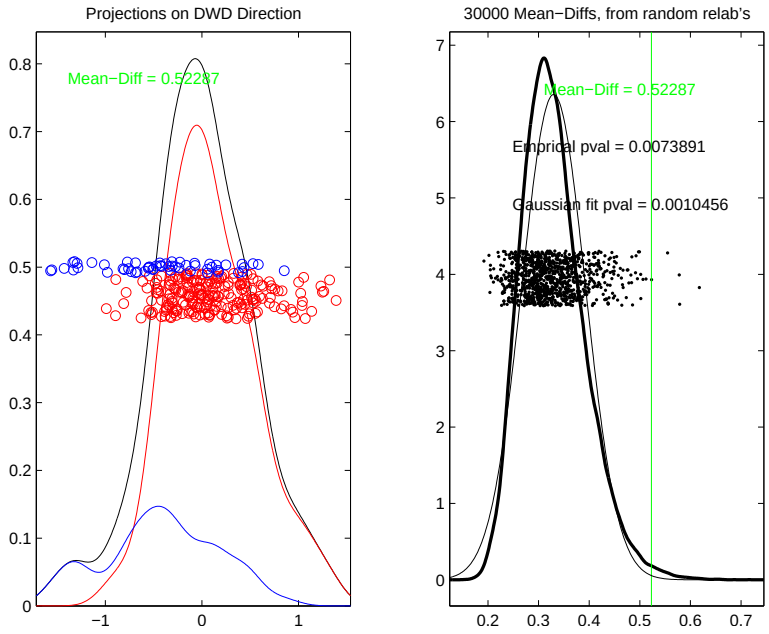


Fig. 5: The DiProPerm hypothesis test of mean differences based on the scaled CPNG scores matrix Z_{Comp} of the final fittings after pre-processing by PP1. DiProPerm is a two sample mean hypothesis test. The left plot shows a jitterplot by the projection of the data on the DWD direction together with 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).

The GOPs that represent latitude and longitude of the spoke direction are normalized corresponding to the shift by the geodesic mean 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 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.

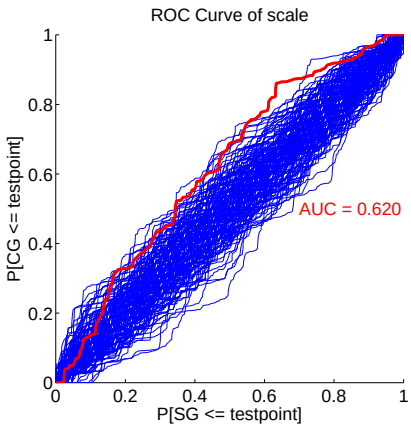
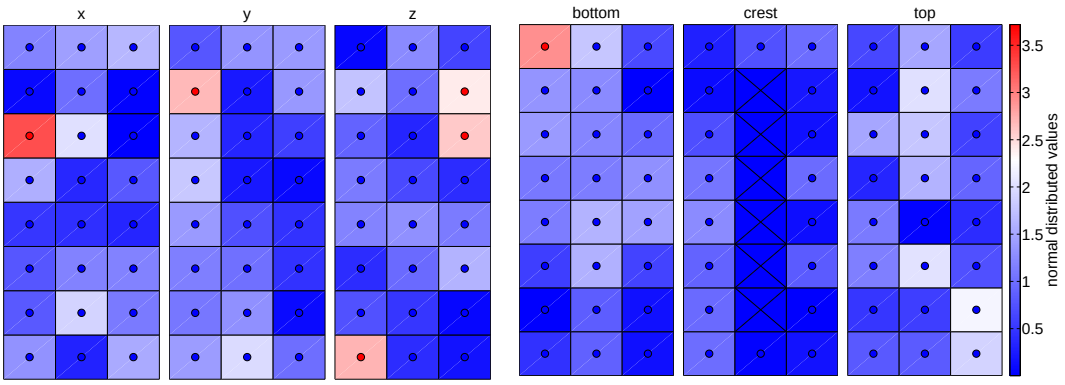
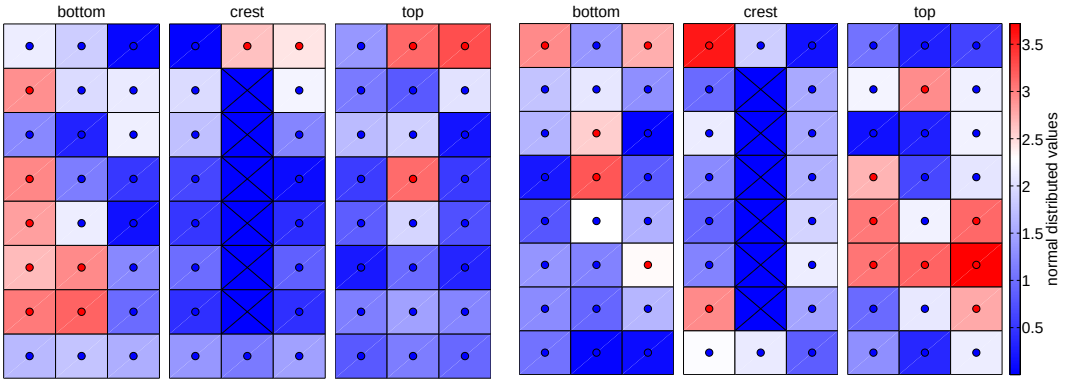


Fig. 6: The ROC curve of the global scaling factor (red) is visualized together with the envelope (blue) obtained from the permutations. The red curve is close to the boundary of the envelope and indicates thereby a significant GOP.

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



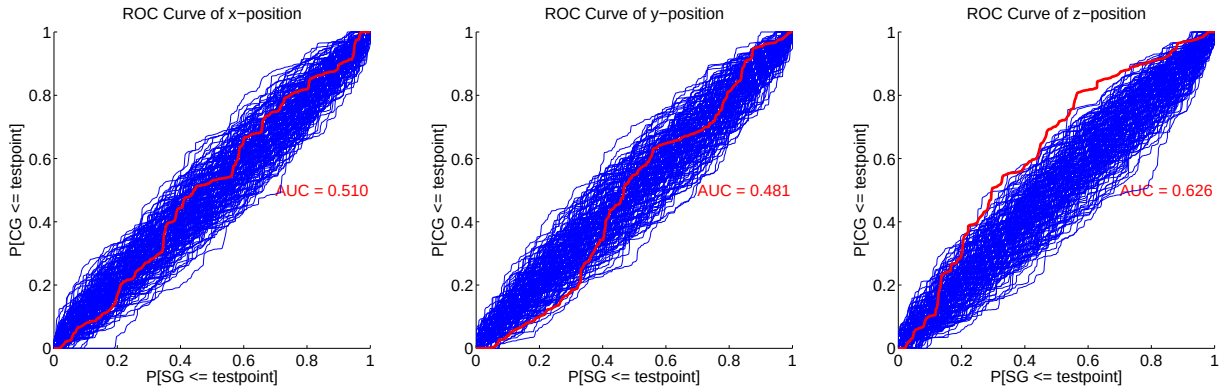
(c) latitude spoke direction (d) longitude spoke direction

Fig. 7: Colored significance 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 corresponds to 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 corresponds 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 confirmed 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.

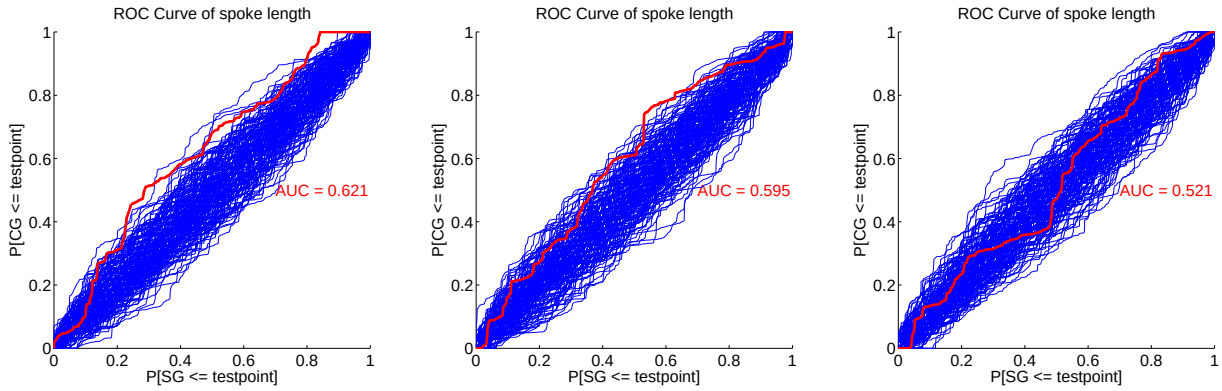
Figure 9 visualizes the ROC curve of the bottom spoke lengths of atom 8, 16 and 24. Figure 7b indicates the bottom spoke length of atom 8 as statistically significant, whereas the bottom lengths of atom 16 and 24 are not significant. Furthermore, atom 24 shows a lower value than atom 16. These observations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



(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 two population samples. The z-position in (c) corresponds to a significant GOP in Figure 7, whereas (a) and (b) correspond to non-significant GOPs.

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



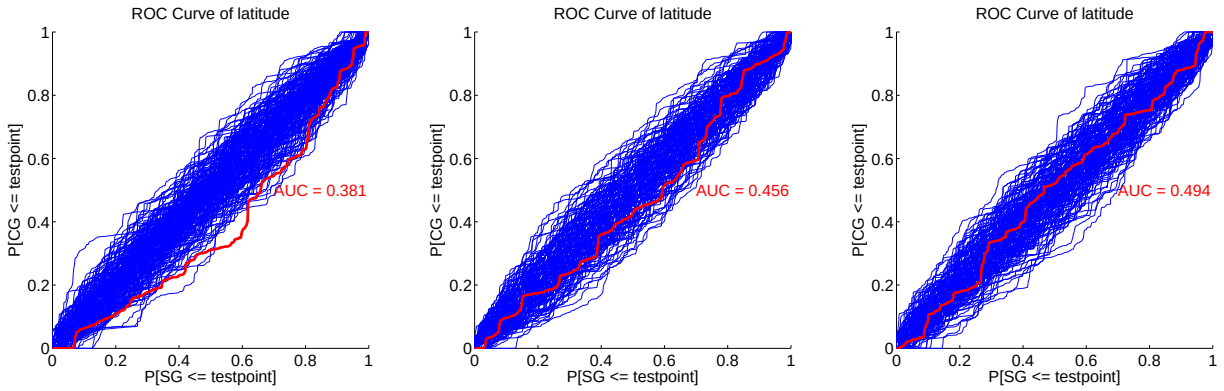
(a) atom 8 (bottom) (b) atom 16 (bottom) (c) atom 24 (bottom)

Fig. 9: As Figure 8, now the 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 figure (a) corresponds to a significant GOP in Figure 7, whereas (b) and (c) correspond to non-significant GOPs.

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 confirmed by the corresponding ROC curves in Figure 10.

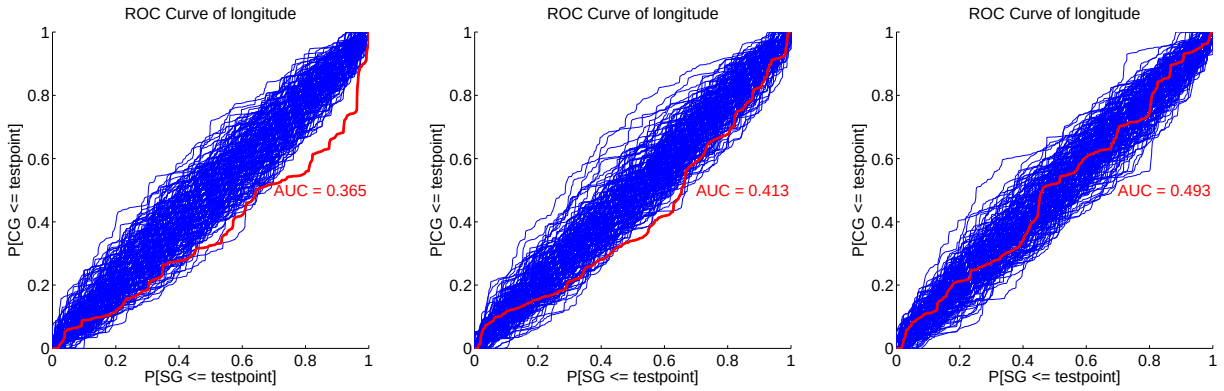
Finally, Figure 11 visualizes the ROC curve of the longitude spoke direction on the crest of atom 8, 16 and 24. Figure 7d indicates a statistically significant longitude spoke direction of atom 8 on the crest of the skeletal sheet, whereas the longitude spoke direction on the crest of atom 16 and 24 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: As Figure 8, now the 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 figure (a) corresponds to a significant GOP in Figure 7, whereas (b) and (c) correspond to non-significant GOPs.



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

Fig. 11: As Figure 8, now the 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 figure (a) corresponds to a significant GOP in Figure 7, whereas (b) and (c) correspond to non-significant GOPs.

2.5 Test results for the unsigned difference measure d^1

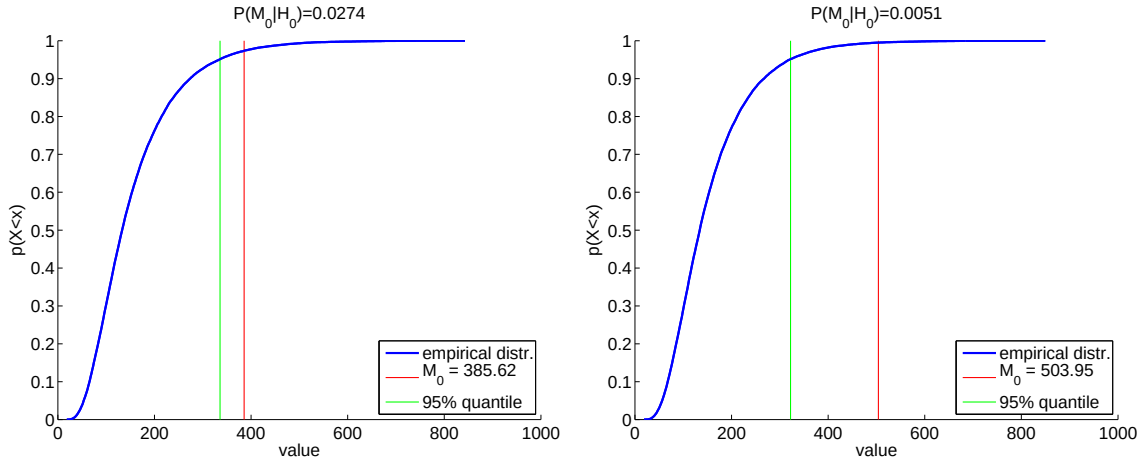
This section reports hypothesis test results using distances measure d^1 as described in Section 1.3. Results are based on the pre-processing methods PP1 and PP2 as described in Section 7.2 in the main article.

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 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

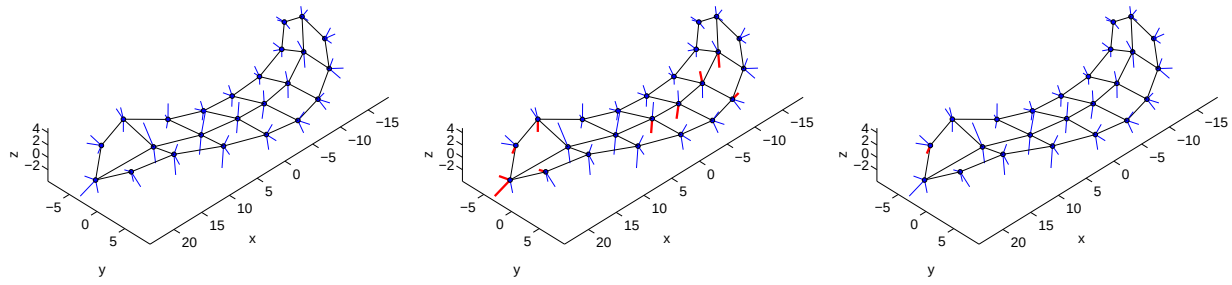
Figures 13 and 14 visualize the feature-by-feature test results for the difference measure d^1 using PP1. Recall that each discrete slabular s-rep is organized into 24 atoms by a 3×8 grid. Thereby, the measure d^1 (see



(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 $\lambda = 2.5532$.

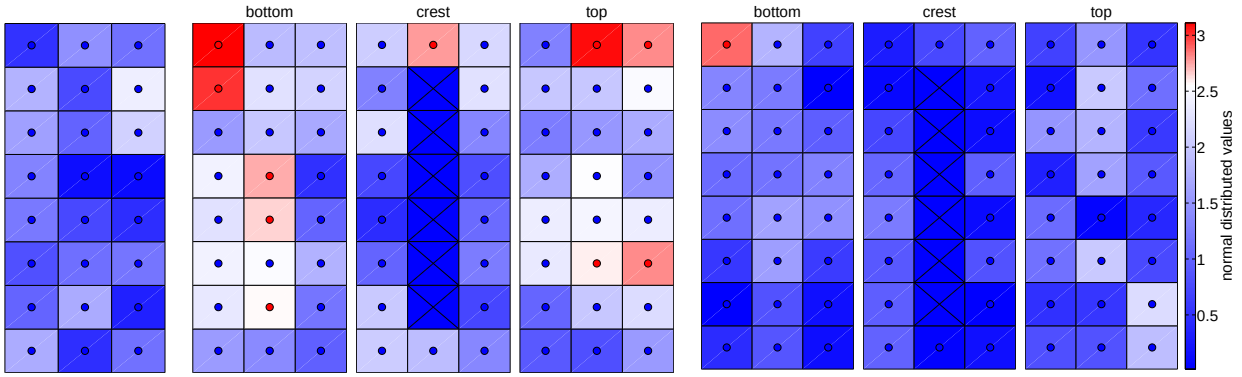


(a) skeletal position (b) spoke direction (c) spoke length

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.

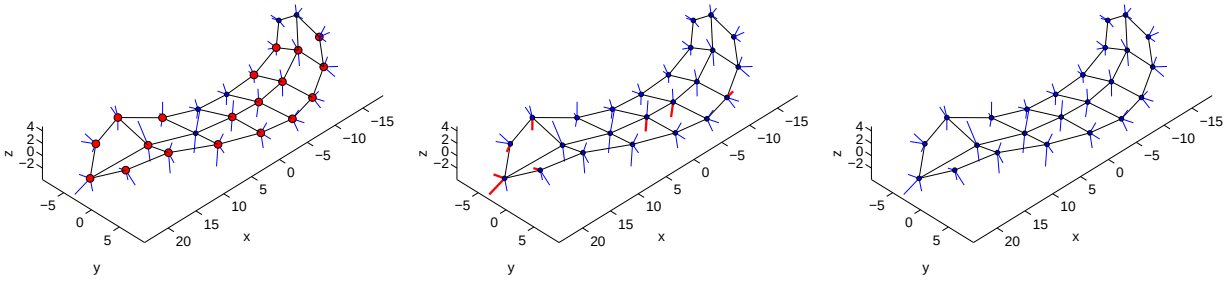
Figures 13 and 14 show several statistically significant GOPs. No skeletal position but one spoke length and 10 spoke directions are statistically significant. Moreover, the global scaling factor τ between SG and CG was found statistically significant by the GOP $|U_{0K}| = 2.7704$.

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 $u_{ij} \in S^2$.



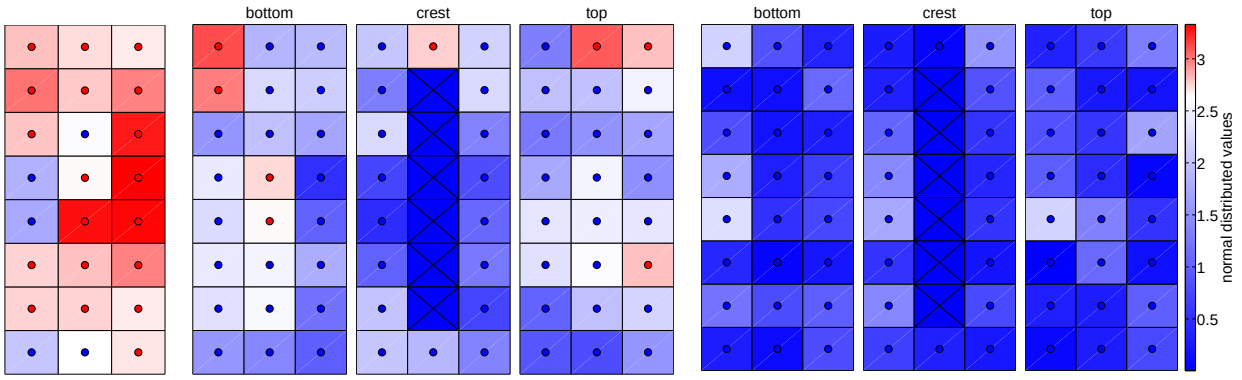
(a) position (b) spoke direction (c) spoke length

Fig. 14: Colored significance map of U_{0k} with a corrected threshold $\lambda = 2.5532$ using PP1 and difference measure d^1. Each box represents a GOP which corresponds 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 fewer 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 Number of permutations for the global test using the two difference measures d^1 and d^2

This section will study the impact of the number of permutations on the global test (described in Section 6.2 in the main article) 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 independent 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 when using distance measure d^1.

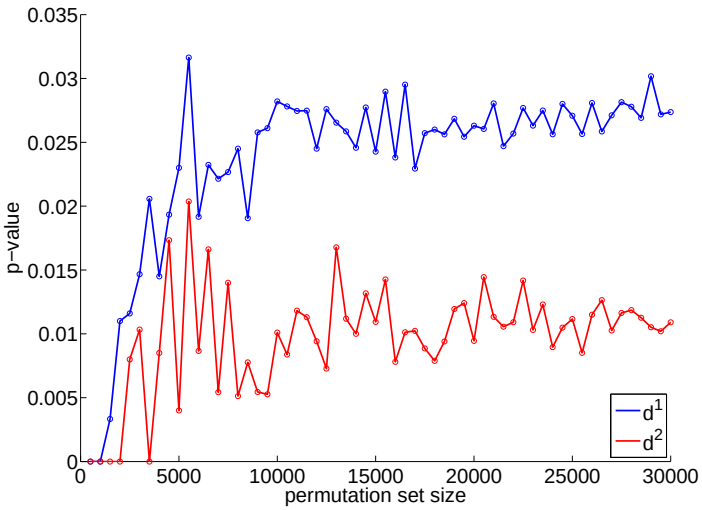


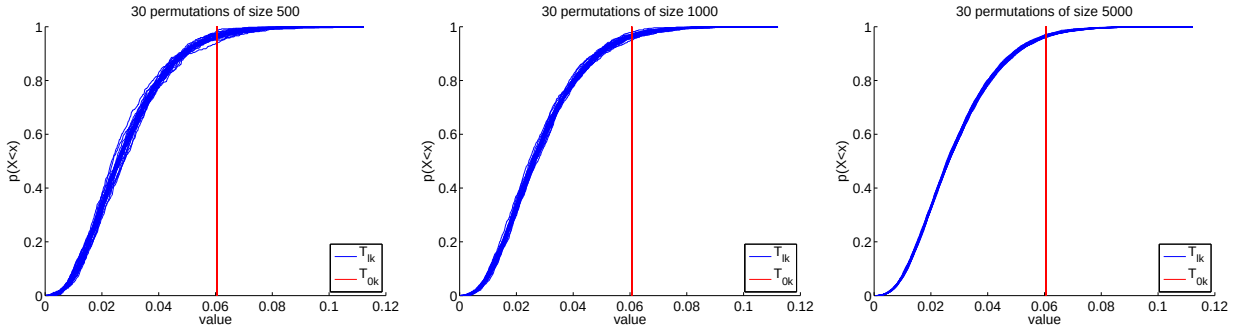
Fig. 17: The p-values are plotted against the number of permutations using difference measures d^1 and d^2. 30000 permutation were generated. The hypothesis test was calculated on randomly chosen subsets with $500, 1000, 1500, 2000, 2500, \ldots, 29500$ permutations.

In order to study the convergence behavior of d^1, we have generated 30 independent 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.

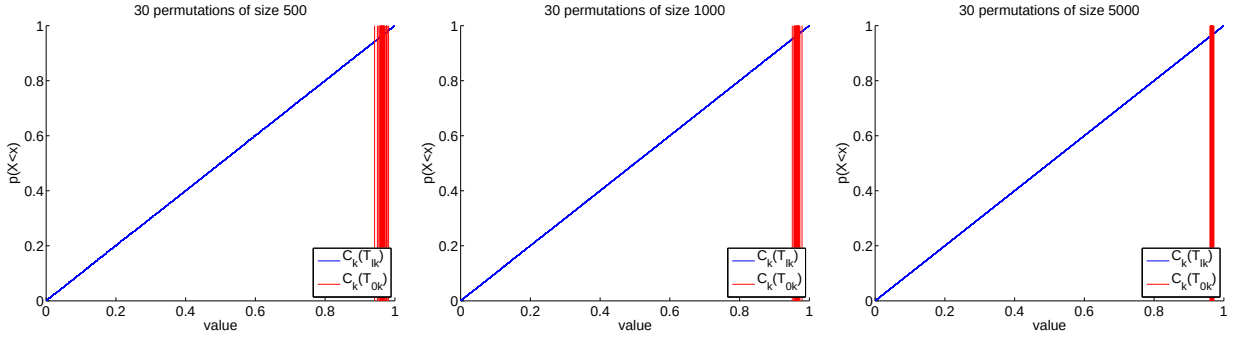
Afterwards, we estimated the empirical cumulative functions C_k for $k = 1, \ldots, K$ partial tests following to Section 6.2.5 in the main article. As a result, we obtained for each GOP difference a p-value $C_k(T_{lk})$, and $C_k(T_{0k})$ respectively. The cumulative empirical distribution of the calculated p-values are depicted in Figure 19. The p-values of the 30 permutation 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 is based 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.



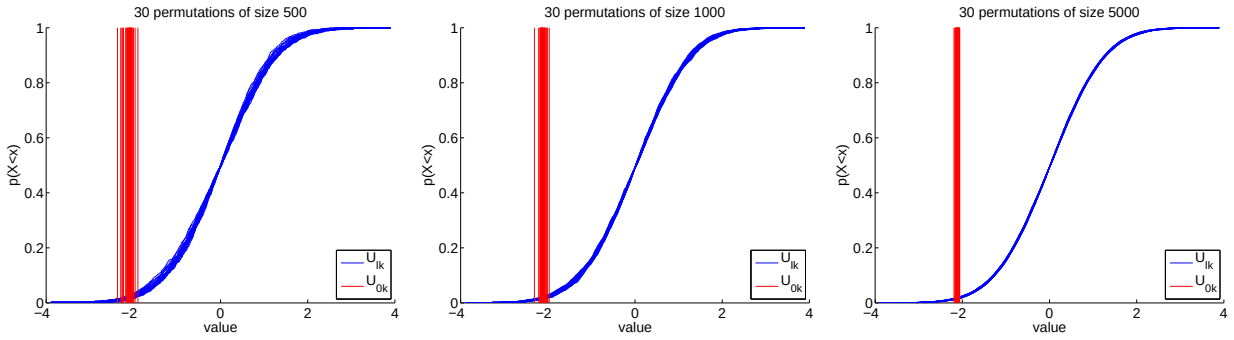
(a) 500 (b) 1000 (c) 5000

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.

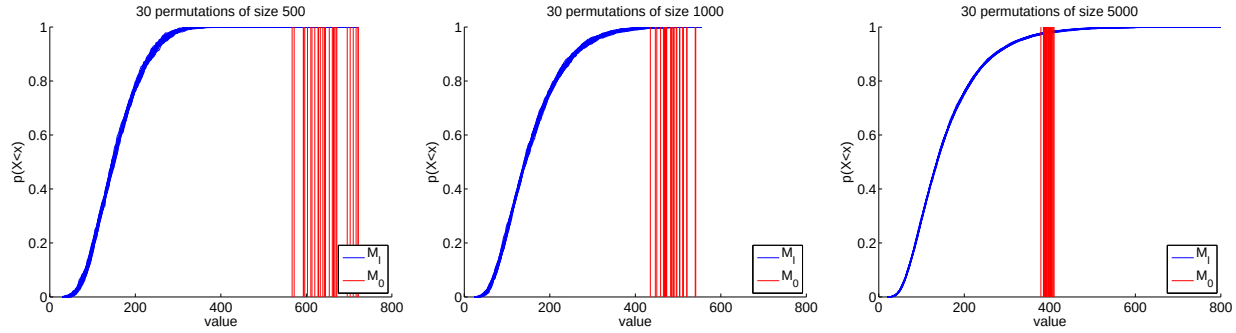


(a) 500 (b) 1000 (c) 5000

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 function 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 pooled 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 use the estimated individual CG shape distribution for the re-fitting of the SG population during the CPNG stage, and use the individual SG shape distribution for the re-fitting of the CG population. In each CPNG stage, the obtained backward mean was translationally and rotationally 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.

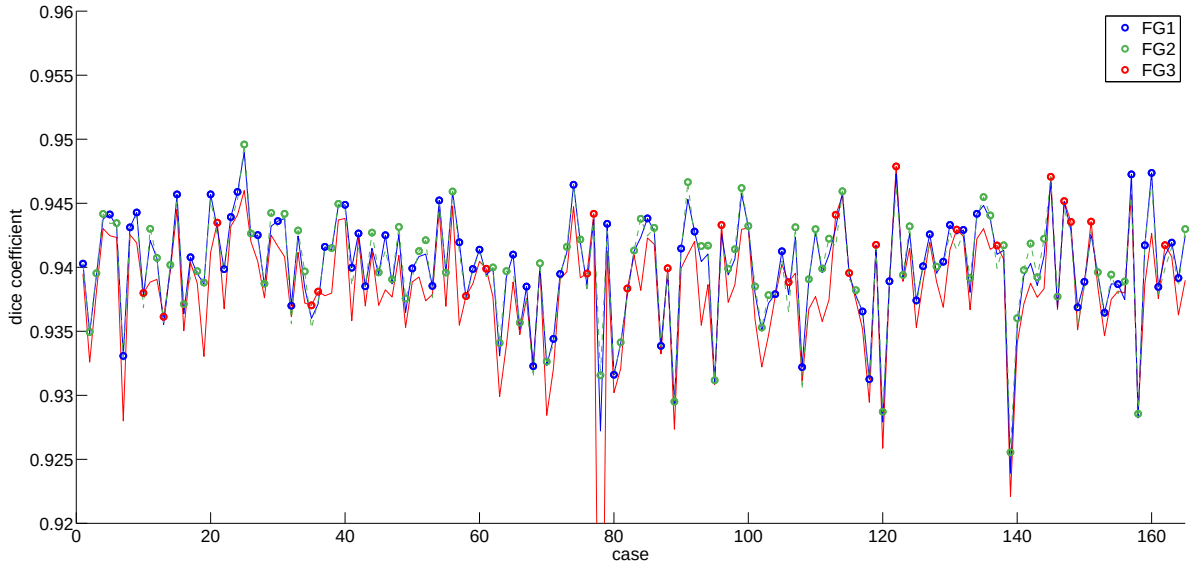
Afterwards, the means were optimized inside the CPNG shape space with an additional final spoke stage (see Section 5 in the main article). As a result, we obtained three fittings for each hippocampus. We chose the fitting with the largest Dice similarity coefficient. The Dice coefficient is a measure of volume overlap 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

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

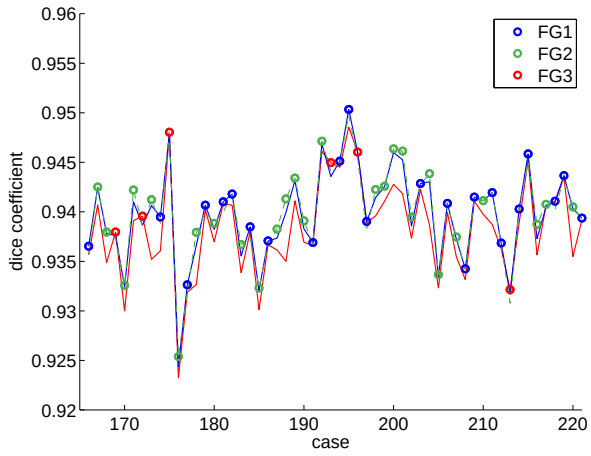
where $|\cdot|$ denotes the number of voxels that describe hippocampal tissue. Figure 22 shows the Dice coefficients of SG and CG for all three fitting types. Accordingly, the second group of final SG fittings consists of 84

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

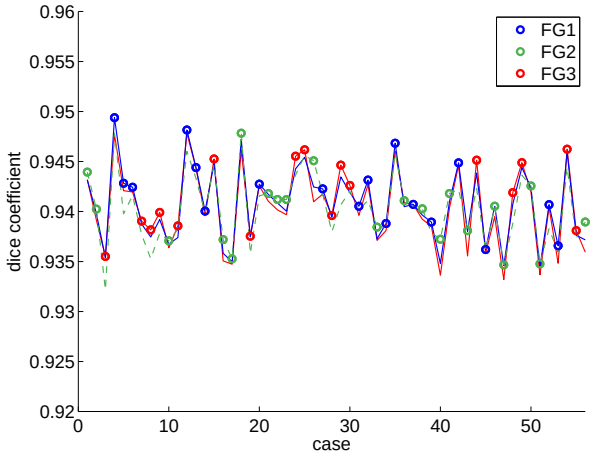
Figure 22 also shows an average volume overlap of 94% for both groups, which indicates accurate fittings. We observe an outlier for case 73 of SG for FG3 due to a poor fitting result. The variance of the Dice coefficient is small for both groups. Nevertheless, a larger variance inside SG can be observed. Moreover, we can observe that FG1 and FG2 leads to comparable Dice coefficients. The Dice coefficient of FG3 is inferior to FG1 and FG2 for SG but comparable for CG. There are two reasons for this observed behavior. 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)

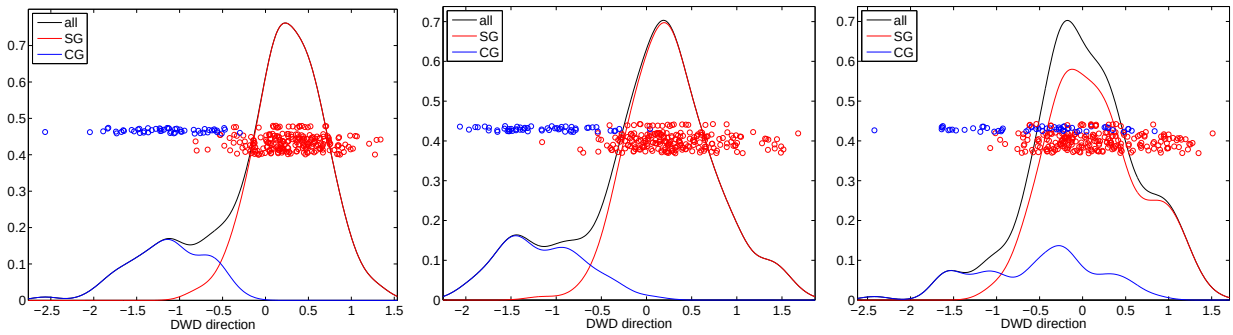


(c) CG

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 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) 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 smaller p-values in Table 1 compared to fittings using a pooled shape distribution, 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.

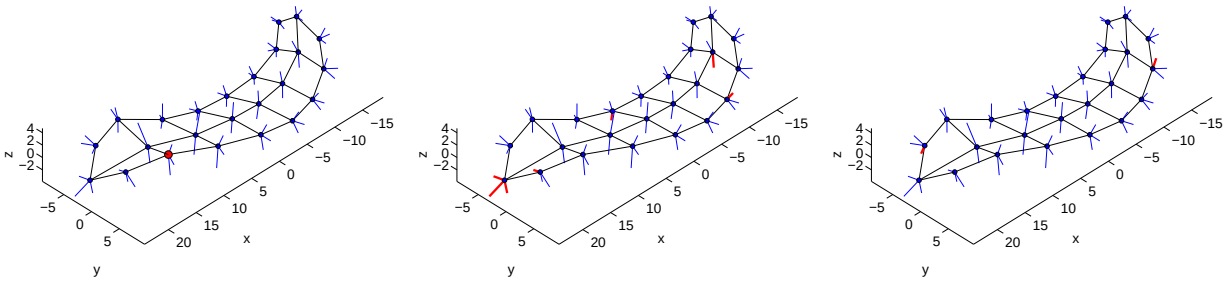
3.2 Single GOP test results

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 information nor conclusions would be added to this section by a repeated comparison of PP1 and PP2 as presented for fittings using a pooled shape distribution.

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.

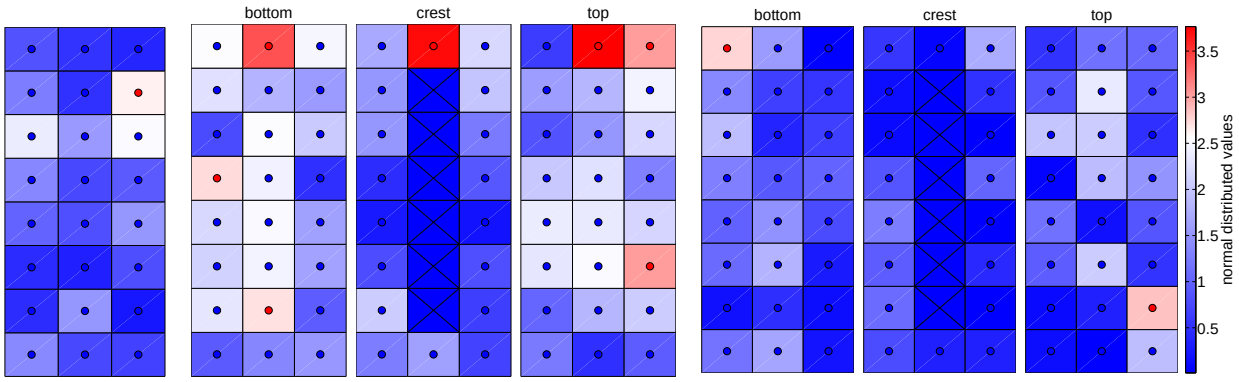
method	empirical p-value PP1	PP2
Mahalanobis distance		
difference measure d^1	0.0245	0.0043
difference measure d^2	0.0013	0.0009
DiProPerm using MD-statistic		
DWD direction vector	0.0018	0.0011
SVM direction vector	0.0039	0.0051

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 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. Figures 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.



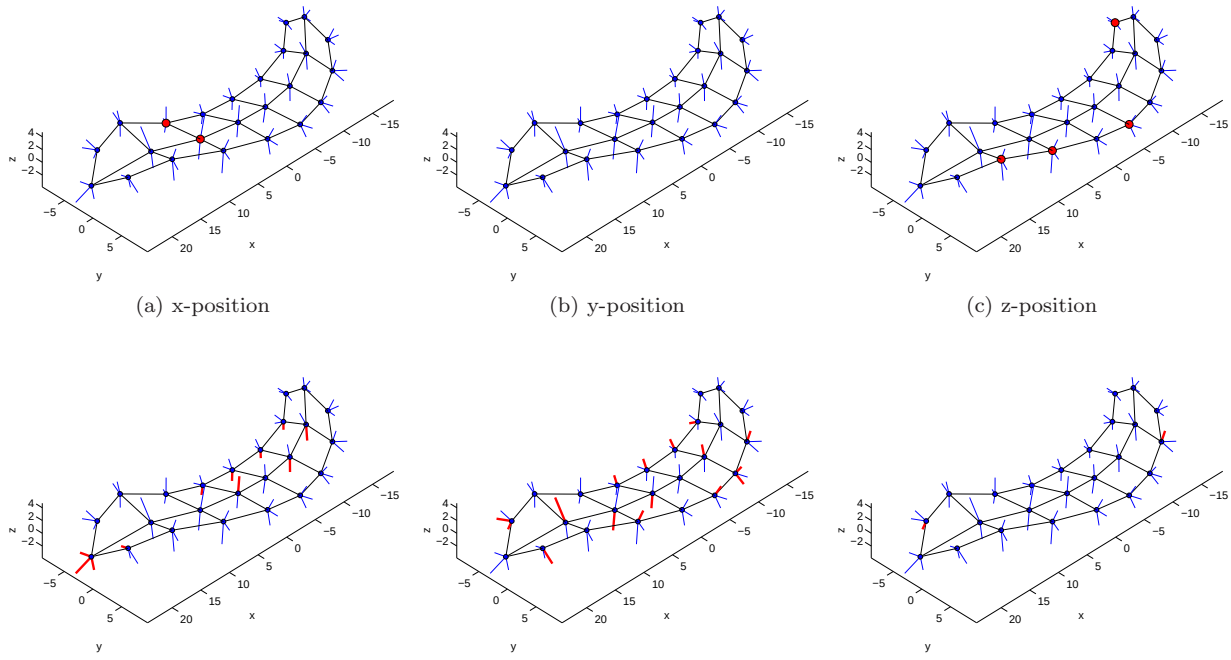
(a) position (b) spoke direction (c) spoke length

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 significance maps except for 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) that 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 $\lambda = 2.5214$. Figures 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

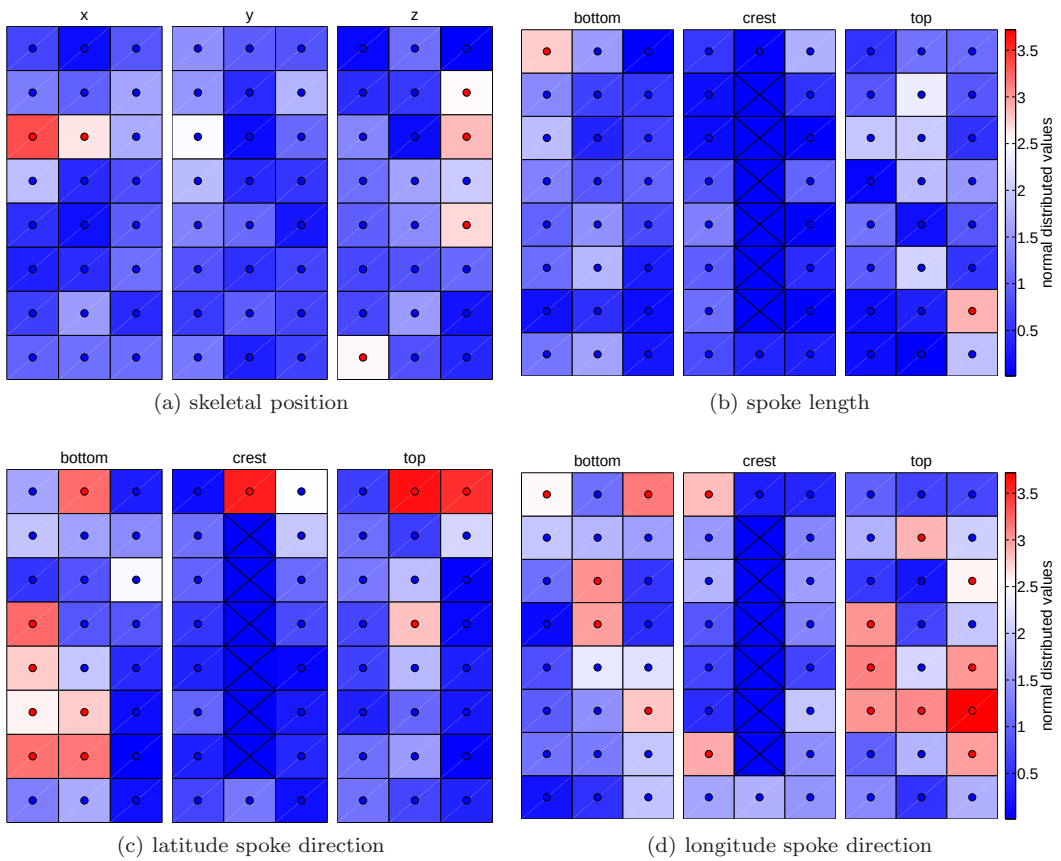
(d) latitude direction (e) longitude direction (f) spoke length

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 significance 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 significance 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)
8. Wei, S., Lee, C., Wichers, L., Li, G., Marron, J.S.: Direction-projection-permutation for high dimensional hypothesis tests (2013). ArXiv:1304.0796