Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
B. Kozlíková, L. Linsen, and P.-P. Vázquez (Editors)

Interactive Formation of Statistical Hypotheses in Diffusion Tensor Imaging

A. Abbasloo[1], V. Wiens[1], T. Schmidt-Wilcke[2], P. Sundgren[3], R. Klein[1], and T. Schultz[1]

[1]University of Bonn, Bonn, Germany
[2]St. Mauritius Hospital, Meerbusch and University of Düsseldorf, Düsseldorf, Germany
[3]University of Michigan Health Care, Ann Arbor, MI, USA and Lund University, Lund, Sweden

Abstract

When Diffusion Tensor Imaging (DTI) is used in clinical studies, statistical hypothesis testing is the standard approach to establish significant differences between groups, such as patients and healthy controls. However, diffusion tensors contain six degrees of freedom, and the most commonly used univariate tests reduce them to a single scalar, such as Fractional Anisotropy. Multivariate tests that account for the full tensor information have been developed, but have not been widely adopted in practice. Based on analyzing the limitations of existing univariate and multivariate tests, we argue that it is beneficial to use a more flexible, steerable test. Therefore, we introduce a test that can be customized to include any subset of tensor attributes that are relevant to the analysis task at hand. We also present a visual analytics system that supports the exploratory task of customizing it to a specific scenario. Our system closely integrates quantitative analysis with suitable visualizations. It links spatial and abstract views to reveal clusters of strong differences, to relate them to the affected anatomical structures, and to visually compare the results of different tests. A use case is presented in which our system leads to the formation of several new hypotheses about the effects of systemic lupus erythematosus on water diffusion in the brain.

CCS Concepts

• ***Visualization application domains*** → *Visual analytics;* • ***Life and medical sciences*** → *Health informatics;*

1. Introduction

Diffusion Tensor Imaging (DTI) provides a symmetric second-order tensor field that encodes the direction and strength of water molecule diffusion inside biological tissue. It is widely used for clinical studies of brain disease, because it makes it possible to measure parameters that relate to tissue microstructure and that correlate with factors such as the integrity of neural tracts [BP96].

In DTI, statistical hypothesis tests are the standard tool for showing significant differences between specific populations, such as between patients suffering from some type of disease and healthy controls [Cer10, OS15]. In current practice, univariate tests which only account for a single parameter, such as Fractional Anisotropy (FA) or Mean Diffusivity (MD), are by far the most widely used.

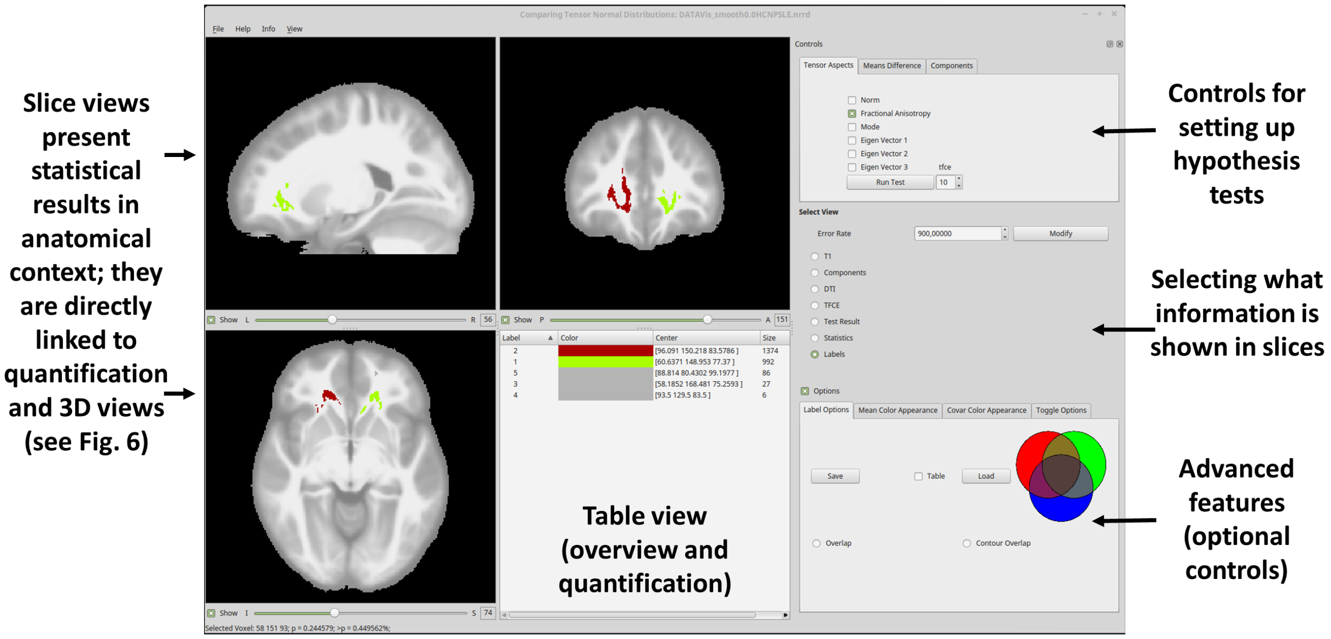
Diffusion tensors, however, contain six degrees of freedom, and reducing them to just one or two scalars could miss some relevant group differences. In order to account for the full information, multivariate statistical methods have been developed based on a tensor normal distribution [BP03], on interpoint distances [BF04, WWHT07], or on eigenvalues and eigenvectors [SDT10].

So far, such methods have not been widely adopted in practice. We believe that an important reason for this is their reduced statistical power. Intuitively, one might hope that, compared to testing for specific differences in FA or MD, testing for all possible types of differences would detect more or larger regions of differences. As we will demonstrate in Section 3, the opposite can happen when dealing with limited sample sizes, as it is usually the case in clinical studies. This is because including degrees of freedom that do not contain a strong group difference will make it more difficult to reach the agreed-upon threshold for statistical significance.

We propose to address this by tailoring the null hypothesis to the expected group differences: If we know which degrees of freedom in the diffusion tensor are affected by the specific type of disease under study, including precisely those should result in a test with optimal power. This strategy requires a “steerable” hypothesis test for tensor fields which, to the best of our knowledge, is not available in the literature. As our first contribution, in Section 4, we introduce such a test, building on the invariant gradients and rotation tangents framework by Kindlmann et al. [KEWW07].

A practical challenge in using such a customizable hypothesis test is having to decide which aspects of the diffusion tensor to include: If we already knew how exactly a disease affects the diffusion tensors, there would not be a need to do a clinical study in the first place. Therefore, our second contribution is a visual analytics system for statistical hypothesis formation, shown in Figure 1, and

DOI: 10.2312/vcbm.20191229



Slice views present statistical results in anatomical context; they are directly linked to quantification and 3D views (see Fig. 6)

Controls for setting up hypothesis tests

Selecting what information is shown in slices

Advanced features (optional controls)

Table view (overview and quantification)

Comparing Tensor Normal Distributions: DATAVis_smooth0.0HCNPSLE.nrrd

Label ▲	Color	Center	Size
2	🟥	[96.091 150.218 83.5786]	1374
1	🟩	[60.6371 148.953 77.37]	992
5	▨	[88.814 80.4302 99.1977]	86
3	▨	[58.1852 168.481 75.2593]	27
4	▨	[93.5 129.5 83.5]	6

Controls

Tensor Aspects | Means Difference | Components

□ Norm
☒ Fractional Anisotropy
□ Mode
□ Eigen Vector 1
□ Eigen Vector 2
□ Eigen Vector 3

tfce 10

Select View

Error Rate 900,00000

○ T1
○ Components
○ DTI
○ TFCE
○ Test Result
○ Statistics
● Labels

☒ Options

Label Options | Mean Color Appearance | Covar Color Appearance | Toggle Options

□ Table

○ Overlap
○ Contour Overlap

Selected Voxel: 58 151 93; p = 0.244579; >p = 0.449562%;

Figure 1: *The main window of our visual analytics system includes slice views, a table view (here, used to highlight two specific regions with significant group differences in red and light green), and controls that allow the user to perform different hypothesis tests, as well as to explore and compare their results with a range of visualization tools.*

described in more detail in Section 5. It allows the analyst to interactively explore a wide range of interpretable null hypotheses that can span the full spectrum between focusing on individual degrees of freedom to including all available information.

When using our system, it is crucial to adhere to the distinction between exploratory and confirmatory data analysis: Statistical testing is usually done for confirmatory analysis, and our proposed steerable test can be used for that purpose. On the other hand, hypothesis formation is an exploratory task, and our visual analytics system is solely a tool for exploratory analysis.

In practice, this means that a hypothesis that has been found with our system should consequently be tested on independent data: Trying out many hypotheses with our tool and selectively reporting the most sensitive one, on the same data and without correcting for the number of tests that have been run, would amount to “p-hacking”, and would represent a misuse of our system.

The intended, legitimate, and statistically valid use of our system is to perform a secondary analysis of already published data, in order to come up with novel hypotheses for future studies. We report such a use case in Section 6, based on data from a study of systemic lupus erythematosus [SWCW*14]. Due to the immense effort of such clinical studies, the corresponding confirmatory analysis has to be left for future work.

2. Related Work

Since our visual analytics system addresses the comparison of groups of tensor fields, the most closely related works are the Tender glyphs, which explicitly encode differences between two symmetric positive definite tensors [ZSL*16], and followup works that have extended this encoding and combined it with complementary visualizations to enable visualization of tensor field ensembles [ZCH*17], as well as a visual comparison of groups of tensor fields at multiple levels of detail [ZHC*17]. Unlike this existing approach, which focuses on a visual assessment, our goal is to closely integrate visual with quantitative statistical analysis. In this sense, our work is similar to a previous one on visualizing tensor normal distributions [AWHS16] which, however, did not include any mechanisms for group comparison.

The body of literature on hypothesis testing in DTI and neuroimaging in general is far too large to exhaustively survey here, and has been summarized elsewhere [Paj11, OS15]. We will provide details on the most relevant related works on DTI analysis as needed throughout the paper. In particular, our proposed tool makes use of state-of-the-art techniques for spatial normalization [ZAY*07, ZYRG07] and enhancement of statistical maps [SN09]. We also make use of an interpretable reparametrization of diffusion tensor differences [KEWW07] that has been used for tensor field processing and visualization previously [SBW06, AWHS16]. To our knowledge, we are the first to use it for hypothesis testing.

3. Motivation

In this section, we provide a more detailed motivation of our approach by illustrating and analyzing the limitations of existing univariate and multivariate tests.

Intuitively, statistical hypothesis tests decide on the significance

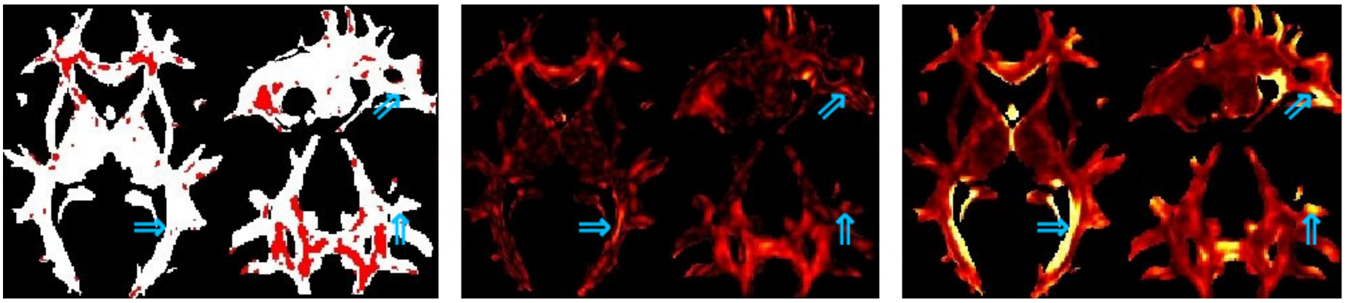
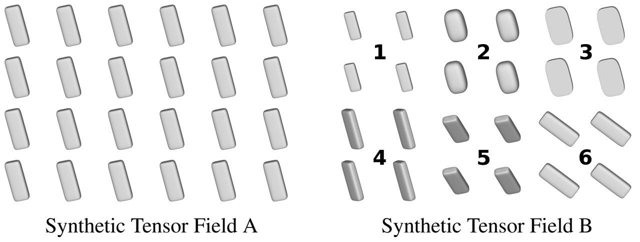


Figure 2: *A group difference at a particular location is considered statistically significant (marked red in the left subfigure) if the magnitude of the group difference (color coded in the center), is large relative to the local variance (color coded on the right). Blue arrows highlight some regions in which a strong difference fails to be significant due to a large variance.*



Synthetic Tensor Field A

Synthetic Tensor Field B

Figure 3: *The two tensor fields A and B have been designed to differ with respect to one of their attributes in each of the six regions. A commonly used statistical hypothesis test based on FA alone would only detect the difference in region 2.*

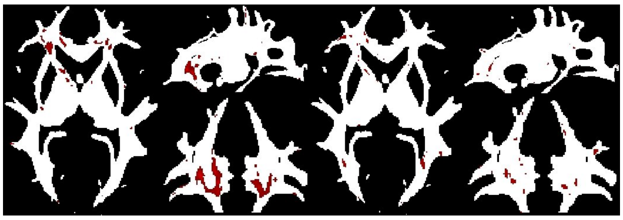


Figure 4: *Compared to a univariate test of tensor changes associated with FA (left), a multivariate test for arbitrary changes in the tensor (right) produces fewer, rather than more, significant results.*

of specific group differences by comparing their magnitude to the amount of variance in the data. This is illustrated in Figure 2: The red regions in the left part, in which statistically significant differences aligned with FA have been detected, have a clear correlation with regions of strong group difference, which is color coded in the central part. However, there are regions with a strong group difference, some of them highlighted with blue arrows, that are not significant in a statistical sense. This happens where the variance, color coded in the right part of the figure, is so large that even large group differences might have arisen due to chance.

Most clinical studies perform hypothesis testing on only one or two scalar values derived from the tensor field, such as Fractional Anisotropy (FA) or Mean Diffusivity (MD). To illustrate the need for multivariate hypothesis tests, Figure 3 presents two synthetic tensor fields. It is obvious from their visualization using superquadric glyphs [Kin04] that the two fields differ at all locations, with respect to all six degrees of freedom in 3 × 3 symmetric tensors: Reduced overall magnitude, as measured by reduced Frobenius norm (1); reduced degree of directional dependence, as measured by reduced FA (2); change from a more linear towards a more planar type of anisotropy as measured by reduced tensor mode, while keeping FA constant (3), as well as rotations around three orthogonal axes (4–6).

In this example, the widely used univariate hypothesis test based on FA would be blind to all differences except those in region 2. In contrast to this, a multivariate test that accounts for all available information in the tensor should be able to detect all different types of variation. Accordingly, multivariate testing is often applied in the hope that it will be able to detect more and larger regions of significant differences. For example, as a key result of their proposed multivariate Cramér test, Whitcher et al. report that they observed a 169% increase "in the volume of a significant cluster compared to the univariate FA test" [WWHT07].

Disappointingly, in our own experiments on clinical data, we often observed a decrease, rather than an increase of the overall number of significant voxels when replacing the widely used univariate t-tests with its multivariate counterpart, Hotelling's T^2 test. An example is shown in Figure 4. The tests will be explained in more detail in Section 4.1. For now, we observe on the left that clear and extended clusters (red) have been detected when looking for changes along one specific axis in tensor space, namely changes in Fractional Anisotropy (FA). One might hope that a test that accounts for the full tensor information would highlight the same regions, plus others. Unfortunately, it is clear from the result on the right that this is not the case. We note that even the 169% increase in volume reported in [WWHT07] pertains to *a single selected cluster;* the authors do not report whether, and by how much, the overall volume of all clusters in their data increased.

To better understand why, depending on the data characteristics, a multivariate test may or may not be more powerful than a simple univariate test, Figure 5 presents two toy examples. In the first one, the blue squares are sampled from a multivariate Gaussian distribu-

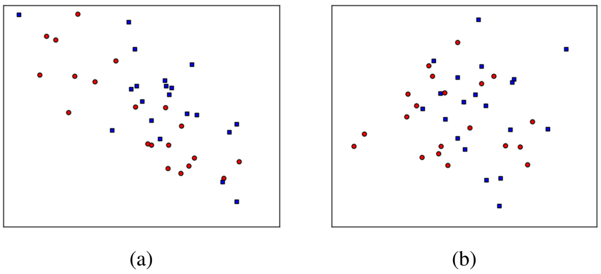


Figure 5: *In (a), a multivariate Hotelling test that accounts for both axes is more powerful for detecting the difference between red circles and blue squares than univariate t tests along either axis. In (b), a univariate test along the horizontal axis is more powerful than a multivariate test. Axes are not labeled, since the effect is unaffected by shifting and uniform scaling.*

tion whose center is at the top right compared to the center of the distribution from which the red circles are taken. Variance along both axes is so high that, with a univariate t test, neither horizontal nor vertical distance alone is sufficient to detect a significant difference ($p > 0.11$ and $p > 0.43$, respectively). In contrast, by accounting for the differences in both dimensions simultaneously, a Hotelling test achieves a clearly significant result ($p < 0.004$).

In the second example, the center of the blue squares is to the right compared to the red circles, but at the same height. Here, a t test along the horizontal axis shows a significant difference ($p < 0.03$), but the Hotelling test fails to produce a significant result ($p > 0.06$). The reason is that including the second dimension adds pure noise, which reduces the power. Consequently, a larger number of samples would be needed for statistical significance. In this illustrative 2D case, the difference in p values is relatively small, and only affects the final outcome because it happens close to the threshold for significance. However, including a larger number of noisy dimensions will amplify this effect and lead to highly relevant changes in practice, such as those that were observed in Figure 4.

Since clinical studies often face difficulties in finding a large number of patients that can be included, and due to the high cost of brain imaging, it is beneficial to pick the null hypothesis such that sensitivity of the resulting test is optimized. Therefore, we introduce a novel hypothesis test for DTI data that can be customized in the sense of deciding which tensor attributes should be part of its null hypothesis. We acknowledge that hypothesis testing as such has some widely known limitations [WL16] and that alternatives such as Bayesian analysis have legitimate use cases [WJP*09]. We believe that our manuscript is not the right place to survey this long-standing and controversial discussion. Our current focus on hypothesis testing is motivated by the fact that, *de facto,* it represents the predominant paradigm in clinical studies.

4. Steerable Statistical Hypothesis Testing

Our first contribution is to derive a steerable statistical hypothesis test for diffusion tensor imaging. This is achieved by suitably combining the multivariate Hotelling test [Sri02] with a specific reparametrization of tensor differences [KEWW07].

4.1. Multivariate Testing with Meaningful Projections

The most common way to identify group differences in imaging studies is *mass univariate testing.* This amounts to running a large number of statistical tests, each accounting only for a single value at a single location of the brain. Spatially mapping locations where the null hypothesis was rejected highlights regions in which the groups differ in a statistically significant manner [OS15].

Tensors are intrinsically multivariate, reflecting not just a single property (such as amount of anisotropy), but also the overall amount of diffusion, type of anisotropy, and preferred diffusion directions. Our framework accounts for this by testing more complex null hypotheses, stating that the mean tensors in the two groups are the same with respect to multiple or even all attributes. The Hotelling test provides the corresponding extension of the widely used t test [Sri02]. Its test statistic is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{d}}_1 - \bar{\mathbf{d}}_2)^T \mathbf{S}_p^{-1} (\bar{\mathbf{d}}_1 - \bar{\mathbf{d}}_2), \tag{1}$$

where $\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_2$ are group mean vectors and $\mathbf{S}_p$ is a pooled estimate of the covariance matrix, $\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$, where n_i and $\mathbf{S}_i$ are the number of subjects and sample covariance for group i, respectively. To apply this test to symmetric 3×3 diffusion tensors with coefficients D_{ij}, we embed them isometrically into $\mathbb{R}^6$ by setting

$$\mathbf{d} := \left[D_{11}, D_{22}, D_{33}, \sqrt{2}D_{12}, \sqrt{2}D_{13}, \sqrt{2}D_{23} \right]^T. \tag{2}$$

As we demonstrated in Section 3, blindly applying a multivariate test to all available degrees of freedom can lead to a loss of sensitivity due to a swamping with noise. To avoid this, we introduce a steerable test, which allows the analyst to interactively formulate null hypotheses that relate to a meaningful subset of the information contained in the tensor fields, and that deliberately exclude other aspects. Unfortunately, the components in Eq. (2) are not suitable for this task, since they depend on the chosen frame of reference, and lack an intuitive interpretation.

We address this by expressing deviations from the local mean tensor in the so-called IGRT ("invariant gradient and rotation tangent") basis, an orthonormal basis that can be constructed from the gradients of tensor invariants such a tensor norm and Fractional Anisotropy, and from the tangent vectors corresponding to infinitesimal rotations [KEWW07]. The IGRT basis is specific to each voxel, since its construction depends on the grand mean tensor $\bar{\mathbf{d}} = \left(n_1\bar{\mathbf{d}}_1 + n_2\bar{\mathbf{d}}_2\right)/(n_1+n_2)$ at that voxel. After expressing differences $\mathbf{d} - \bar{\mathbf{d}}$ between individual tensor values $\mathbf{d}$ and $\bar{\mathbf{d}}$ in the IGRT basis, we can interpret them as being related to

1. Changes in Frobenius norm ($R_1 = \|\mathbf{D}\|_F$)
2. Changes in the amount of anisotropy, as measured by $R_2 = \mathrm{FA}$
3. Changes in the type of anisotropy, as measured by $R_3 = \mathrm{mode}$

4–6. Rotations around the three eigenvectors

We allow the analyst to build interpretable null hypotheses by interactively selecting arbitrary combinations of these six degrees of freedom, which are illustrated in Figure 3. More specifically, the six values correspond to the magnitude of diffusion tensor differences that are associated with changes in norm, FA, etc. As such, they all have the physical units of diffusivity. A seemingly more straightforward alternative to reparametrizing tensor *differences* would be to reparametrize the tensor itself (e.g., similar to [ZSL*16]) and to subsequently measure differences between the resulting parameters. In the latter case, one would have to be very careful about singularities: For example, for isotropic tensors, one would have to be careful to disregard changes in tensor mode or rotations. When starting with differences in the tensor itself, this happens automatically: Since rotating an isotropic tensor does not change it, projecting that zero difference on a rotation tangent correctly indicates no tensor difference associated with rotation.

When including all six degrees of freedom in our test, it will produce the same result as a Hotelling test without the reparametrization. This is due to two facts: First, the Hotelling test is invariant under non-singular linear transformations. Second, despite singularities in the involved quantities, the IGRT basis can always be kept orthonormal [SBW06], and its use therefore corresponds to a non-singular linear transformation. However, testing tensor differences associated with FA changes within our framework can produce slightly different results than the more widely used practice of testing changes in pre-computed FA, because the two are non-linearly related.

Kindlmann et al. [KEWW07] describe two variants of the IGRT basis: One is derived from a cylindrical coordinate system, whose invariants are written as K_i and include tensor trace, which is proportional to the widely used Mean Diffusivity (MD). In this work, we use the second one, which is derived from a spherical coordinate system, written as R_i, and includes the widely used Fractional Anisotropy (FA). Unfortunately, the gradients of MD and FA are non orthogonal, making it impossible to combine both into a common orthonormal IGRT frame. We have chosen the one that includes FA rather than MD to facilitate a better comparison with a previous work that has tested FA on the same data [SWCW*14].

4.2. Cluster Enhancement

When designing a statistical hypothesis test, we have to control its type I error rate α. In our context, a type I error amounts to reporting a group difference in a region where there appears to be one in the sample – e.g., due to noise or due to the specific choice of subjects – even though there is none in the underlying populations. In neuroimaging, it is widely accepted to use $\alpha = 0.05$.

Performing a statistical test in a huge number of voxels greatly increases the *family-wise error,* i.e., the probability that the null hypothesis will be falsely rejected in at least one voxel. This could be addressed by reducing the α of each individual test, a strategy known as Bonferroni correction. However, this approach is overly conservative, since nearby voxels are usually highly correlated. Therefore, it leads to an unnecessarily drastic reduction of α that no longer allows us to detect any of the true group differences.

A common way to correct for family-wise errors while preserving a useful amount of sensitivity is to rely on the size of contiguous regions in which an effect is observed, since large regions are less likely to occur due to chance [NH03]. Our work uses threshold-free cluster enhancement (TFCE) [SN09], a state-of-the-art variant of this idea. TFCE automatically tries out a wide range of possible thresholds to form contiguous regions. For each voxel, it then takes a weighted sum of the cluster volumes that the voxel belongs to at different thresholds. This way, voxels that belong to larger neighborhoods receive greater values without the need to select any specific cluster forming treshold.

For confirmatory analysis, TFCE is used within a permutation based statistical test [NH02], which involves computing the test statistic and its TFCE transformation for a large number of randomly permuted group labels. Based on this procedure, one can determine a value at which the TFCE map should be thresholded to yield a statistical test with a given family-wise error rate. Unfortunately, even when approximating the permutation test with a random subsample of permutations, it can take many hours on a standard workstation, and is thus much too time-consuming for use in an interactive visual analytics framework.

An exploratory analysis provides a more complete picture if we do not just use a single fixed setting of α, but explore different “levels of significance”. Therefore, we simply decide not to report p values associated with TFCE analysis in our tool, which serves as an additional reminder that our tool is solely meant for exploratory analysis, not for “p-hacking”. For confirmatory analysis, p values can easily be obtained in an off-line process. During data exploration, TFCE maps can be thresholded at arbitrary values, guided by a cumulative histogram that visualizes how the number of super-threshold voxels depends on the threshold value.

5. Visual Analytics for Statistical Hypothesis Formation

This section describes the requirements analysis, design, and implementation of our visual analytics system. It is written in C++ with Qt for the user interface, Teem (teem.sf.net) for standard tensor visualization including fiber tracking, and OpenGL for 3D graphics. Data preprocessing, such as transforming all tensors into the IGRT basis (Section 4.1), or computation of means and covariances of each subgroup, was performed using Python scripts.

5.1. Requirements Analysis

In discussions with our clinical co-authors, we identified four requirements that a visual analytics system for hypothesis formation in diffusion tensor imaging should meet:

R1 The system should make it easy to specify different hypotheses, and to start exploring the results after a minimal delay.

R2 The system should offer the two-dimensional slice views that clinical researchers are used to and that are frequently found in clinical publications.

R3 The system should also provide three-dimensional views to help assess anatomical structures, e.g., via fiber tracking. Compared to 2D views, tractography makes it easier and faster to identify the white matter tract in which a difference occurred.

R4 The system should support the direct comparison of the results from different null hypotheses, to judge the extent to which the corresponding regions might spatially overlap. When dealing with newly discovered differences, it is relevant to see whether or not they affect the same regions as previously known differences.

5.2. Hypothesis Specification and Testing

The steerable test that was introduced in Section 4.1 can be used to specify 63 different null hypotheses, i.e., all 2^6 possible combinations of the six degrees of freedom, minus the configuration in which none of them is included, which does not result in a meaningful test. Six checkboxes in the top right part of our user interface (cf. Figure 1) can be used to specify which tensor attributes should be included, and to run the corresponding test.

In our case study, the selected hypothesis tests could be computed within a few seconds on a standard workstation. The most time-consuming part is the threshold-free cluster enhancement (Section 4.2). Therefore, to help meet requirement **R1,** we allow the user to modify its level of discretization, i.e., the number of thresholds over which cluster volumes are averaged. A low setting can be used to quickly screen a larger number of possible hypotheses, while a high setting can be used for a high-quality investigation of a specific promising one.

A prerequisite of comparing all tensor fields on a per-voxel basis is to bring them into spatial correspondence by nonlinear registration to a common template. As part of this process, tensors have to be rotated according to the incurred changes of their frame of reference. We used the publicly available Diffusion Tensor Imaging ToolKit (DTI-TK) [ZAY*07, ZYRG07] to achieve this.

5.3. Spatial Views

After specifying and running a hypothesis test, the analyst would like to investigate the anatomical location, spatial extent, and shape of the regions in which a group difference was detected. According to requirement **R2,** axis-aligned slice views play a prominent role in our user interface. We follow the radiological convention, i.e., the patient's right hemisphere is shown on the left hand side of each picture (e.g., see labels for left and right in Figure 7).

Interpreting the results of a statistical test requires viewing them within a proper anatomical context. By default, we provide this context by superimposing them on an averaged anatomical MR image which has been transformed into the same reference space as the DTI data (also in Figure 7). On demand, a more exact assessment of the affected tract can be made by displaying an XYZ-RGB color encoding of the principal diffusion direction [PP99].

Our system meets requirement **R3** by seeding a streamline-based tractography algorithm [BPP*00] in an affected region, and providing a three-dimensional view of the result, with orthogonal slices as optional anatomical context. In order not to bias the tractography towards any of the involved subjects, we run it on an average of all coregistered tensor fields. The average is taken in Log-Euclidean space [AFPA06], which is known to minimize blurring [KZR*13]. We found that this strategy preserves the characteristic shapes of the major bundles. For example, parts of the forceps minor, anterior thalamic radiation, inferior fronto-occipital fasciculus, and uncinate fasciculus are well-recognizable in Figure 6.

5.4. Table View

In neuroimaging, contiguous sets of voxels in which a significant difference was detected are referred to as clusters. To allow for their quick and objective assessment, we present the number and volume of all detected clusters, sorted by cluster size, in a table. It can be seen in the immediate neighborhood of the slice views in Figure 1. The table also includes spatial information (center of gravity of each cluster) and is linked to the slices, so that clicking on a table row moves the slices to the respective cluster center.

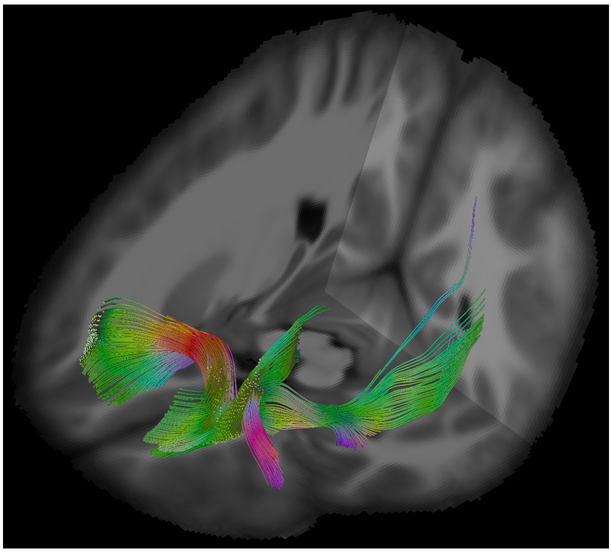


Figure 6: *Seeding streamline tractography in a region of differences between the tensor fields allows the analyst to quickly identify the affected fiber tracts based on their characteristic shapes and anatomical context.*

To highlight and better distinguish clusters of particular interest, the user may assign a color to them in the table view. In this mode, unselected clusters are shown in gray, to make them less visually salient than the more relevant ones, while still distinguishing them from the rest of the brain and the background. In Figure 1, two clusters are shown in maroon and green, respectively.

5.5. Comparative View

To address requirement **R4,** our framework supports the comparison of results from different tests, thresholds, or levels of data smoothing. After running a test, optionally with cluster enhancement, one can save the resulting binary mask to a file. Up to three different test results can then be loaded into an overlay map, shown in Figure 7.

It is a difficult task to visually encode the overlap of different classes in a manner that is intuitive and easy to interpret. After trying out several alternatives, we chose an encoding that has been proposed in a very different application context, namely, in the "splatterplot" approach to overcoming overdraw in scatter plots [MG13]. This encoding combines specific rules for color blending and modulation with contouring.

Colors that encode regions of overlap are obtained in two steps: First, the colors that represent the overlapping classes are averaged in the CIE Lab color space to obtain a color that has approximately

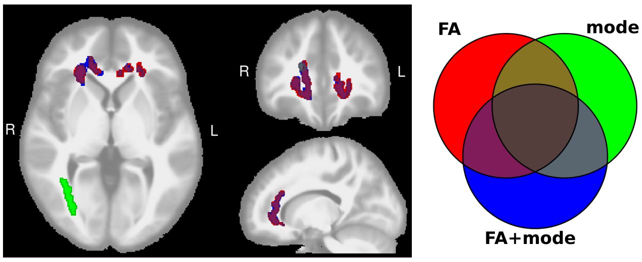


Figure 7: *A multivariate test FA+mode is compared to the two respective univariate tests by mapping the results of each to a different color, mixing them in a perceptual color space, and delineating them with contours. A Venn diagram serves as a color legend for regions of overlap.*

the same perceptual distance to all involved classes. Second, in order to more clearly indicate the presence of multiple classes, lightness and chroma of the mixed color are attenuated in the LCH color space, with an increasing effect for an increasing number of overlapping classes. The exact equations can be found in [MG13], and the result can be observed in Figure 7. To further facilitate interpretation of the overlay map, our user interface displays a Venn diagram as a color legend.

Despite the advanced color blending and additional visual cues from contours, which we draw in slightly darker shades of the respective color, we came across cases in which the underlying maps were so complex that trying to understand the exact relationship of all three from a single image remained challenging. For these cases, our interface allows the user to temporarily hide some of the classes in order to build a better understanding in an interactive and iterative manner.

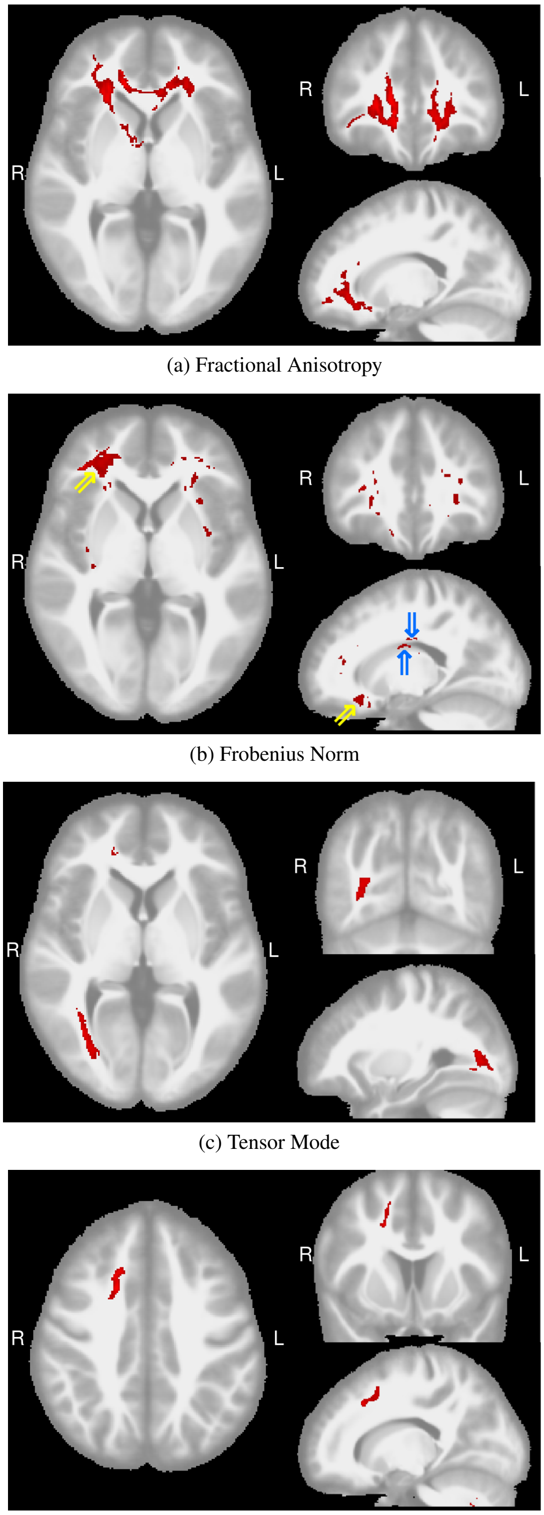
The exact number of voxels included in each mask, and the number of voxels in the overlapping regions, are displayed as a tooltip of the Venn diagram. We also update the table view to reflect the connected components of the regions where the results of all three tests overlap. Specific use cases, and interpretation of Figure 7, are discussed in Sections 6.2 and 6.3.

6. Case Study on SLE

During the design and implementation of our system, we continuously accounted for feedback from our clinical co-authors. In this section, we report results from a specific case study, which used DTI data from a clinical study of systemic lupus erythematosus (SLE). A standard analysis of this data has been published previously [SWCW*14], along with details about the data acquisition, clinical parameters, and criteria for inclusion. The study included 56 subjects. Our analysis used 37 of them, comparing 19 SLE patients with neuropsychiatric symptoms (NPSLE) to 18 subjects in a healthy control group.

6.1. Hypothesis Formation on Clinical Data

As it was explained in Section 4.1, our framework tests *changes in tensor values* associated with certain tensor invariants or rotations. This makes a univariate test of Fractional Anisotropy (FA)



(a) Fractional Anisotropy

(b) Frobenius Norm

(c) Tensor Mode

(d) Rotation around main axis

Figure 8: *In data from a study on SLE, results for Fractional Anisotropy differences between healthy controls and NPSLE patients agree with previously published results, which validates our approach. Additional hypothesis tests can be constructed using our tool, and lead to interesting new observations.*

in our tool similar, but not equivalent to traditional hypothesis tests, which are based on *changes in (pre-computed) FA*. Therefore, comparing the results to those from a traditional analysis can serve as a valuable initial validation. Figure 8 (a) shows the result from comparing healthy controls to NPSLE patients in our framework, based on FA-related changes and threshold-free cluster enhancement. The detected regions are remarkably similar to those from a previous analysis of the same data with standard methods [SWCW*14]. Seeding tractography in them also reveals most of the fibers that have been reported in a previous tractography-based study [SSW*16] and have been ranked highly in terms of their predictiveness [KSWS*17], including the genu of the corpus callosum, left and right inferior fronto-occipital fasciculus, and left uncinate fasciculus (see also Figure 6).

The remaining parts of Figure 8 show the results from other univariate hypothesis tests that are enabled by our framework, and for which no similar results have been reported previously. Subfigure (b) shows regions in which the Frobenius norm of the diffusion tensors is increased in the patients, indicating increased overall diffusivity. The largest cluster (yellow arrows) is immediately adjacent to, and partly overlapping with, the previously detected changes in FA. Only a small part of the second largest cluster is visible in Figure 8 (b) (blue arrows).

Subfigure (c) shows a cluster in which tensor mode is decreased in the patients, which means that diffusion is restricted more to a plane than a line. This might indicate a larger degree of fiber spread in the patients. A tractography result seeded in this cluster indicates that the difference extends over parts of the forceps major and inferior fronto-occipital fasciculus.

The only result we observed when testing for differences in orientation was a cluster in the corona radiata of the right hemisphere, where tensors are systematically twisted (rotated around their principal eigenvector) in the patient group compared to the controls. It is unclear whether this is an artifact of the specific sample or the registration algorithm, or if it corresponds to a true difference between the populations, especially given that the cluster is small.

In summary, our framework was successfully used to produce new hypotheses that should be tested in a confirmatory setting in future clinical studies.

6.2. Comparing the Results of Different Tests

One question that the comparative view from Section 5.5 allows us to answer is how the results from two univariate tests, e.g., ones that compare NPSLE patients to healthy controls based on FA or tensor mode alone, differ from a single multivariate test that combines both. Figure 7 shows the corresponding overlay.

As it was mentioned in Section 4.2, our software does not convert the results from TFCE to p values, since this would require time consuming permutation testing. In order to still compare univariate and multivariate tests in a meaningful manner, Figure 7 was created using the Hotelling test without TFCE. In this case, the null distribution has a parametric form [Sri02], which allows us to compute p values within interactive runtimes. To compensate for the fact that testing FA and mode separately amounts to performing twice as many tests as a single combined test, we made the thresholds for the univariate tests twice as restrictive ($p_{\text{uncorr}} < 0.005$ vs. $p_{\text{uncorr}} < 0.01$).

In this example, the comparative view revealed that the volume highlighted by the multivariate test ($7.8\,\text{cm}^3$) mostly agrees with the union of the two univariate tests ($4.3\,\text{cm}^3$ for FA and $4.0\,\text{cm}^3$ for mode), with almost no overlap between FA and mode. This leads us to the hypothesis that testing mode in future studies of SLE may result in new findings that complement the ones from FA.

6.3. Visualizing the Effect of Data Smoothing

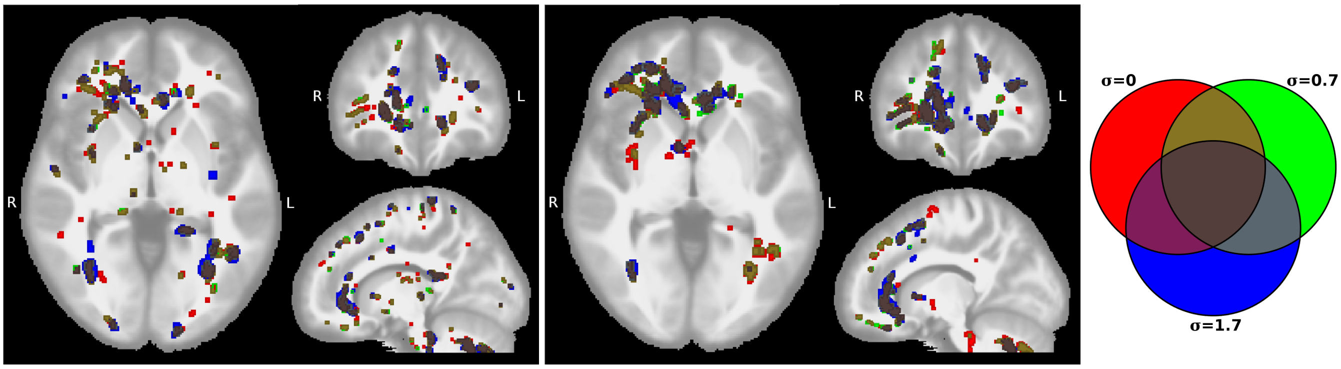
There has been some controversy about whether to smooth diffusion tensor data in preparation for statistical analysis. On one hand, spatial smoothing helps to compensate for anatomical misalignment that may remain after registration, it smoothes out noise that might otherwise lead to false positive detections, it makes the data distribution more Gaussian and, when the bandwidth is matched to the spatial scale of regions of difference, it can boost statistical power. On the other hand, the ideal spatial scale is usually not known in advance, and there is no principled approach for selecting the bandwidth parameter, which might have a substantial effect on the results [JSCH05]. Moreover, it has been argued that, since smoothing amounts to artificially decreasing the image resolution, it is counter-productive for DTI in particular, which specifically measures diffusion in order to overcome the limits of image resolution and to infer tissue parameters at a microscale [SJJB*06].

For these reasons, Tract-Based Spatial Statistics (TBSS), a widely used method for statistical hypothesis testing of diffusion tensor fields, avoids spatial smoothing and instead corrects for residual misalignment by projecting FA values onto a so-called white matter skeleton, a medial surface representation of the major fiber tracts [SJJB*06]. However, TBSS has been designed for scalar invariants, and does not include any mechanisms for tensor re-orientation, which might be required for correctly aligning the full tensors. Therefore, recent studies that have analyzed the full tensor have still employed smoothing [BNH*16].

We have used the comparative view from Section 5.5 to explore the effects of data smoothing, as shown in Figure 9. In this experiment, we compared the healthy and NPSLE populations with respect to the full tensor information. As in [BNH*16], we compare the spatial regions that result from different processing pipelines by showing the 5% “most significant” voxels, rather than setting some *a priori* threshold.

Figure 9 (a) compares results on the original data (red) to data that has been smoothed with bandwidth $\sigma = 0.7$ voxels (green) and $\sigma = 1.7$ voxels (blue), respectively. Many isolated voxels and small regions are shown in red, indicating that they are removed by smoothing, while many of the larger ones grow, as indicated by blue halos around them. Optionally, the analyst can focus on the dark brown regions, indicating differences that are considered significant irrespective of the amount of smoothing.

Figure 9 (b) shows results from the same experiment, but additionally uses threshold-free cluster enhancement (TFCE). Even without any smoothing, TFCE eliminates many of the small clus-



(a) Results without TFCE

(b) Results with TFCE

Figure 9: *Our decision to use threshold-free cluster enhancement (TFCE) is based on comparing results with (b) and without it (a). Results without any smoothing are shown in red, with $\sigma = 0.7$ in green, with $\sigma = 1.7$ in blue. Regions of overlap are encoded as indicated in the Venn diagram on the right. We observe that TFCE reduces the effect, and also the necessity, of pre-smoothing.*

ters, and leads to larger connected ones. It also reduces the overall effect of smoothing, leading to a 10% increase in the number of consensus voxels. Based on these results, we decided to omit smoothing, but use TFCE in the experiments in Section 6.1.

6.4. User Feedback

We had a two-hour in-person meeting in which we provided a live demo of the system to one of our collaborators and went through several examples together, discussing both the design of our software and the clinical implications of our findings.

Our collaborator particularly liked the fact that our system tightly integrates statistical and quantitative data analysis with visualization, which is not achieved in the standard software packages available to him, and which he found to encourage data exploration and hypothesis generation. He considered this particularly helpful for DTI analysis, due to the complexity inherent in the different tensor properties, and the opportunity to directly seed tractography from regions in which significant differences were found. However, he also mentioned that he would like to see similar systems for other modalities, such as functional MRI or Voxel Based Morphometry.

In preparation for the meeting, we reorganized our original user interface by placing controls that we considered to be less crucial to the main workflow into an optional “advanced” view, as indicated in Figure 1. Our collaborator found that, despite this effort, a certain level of complexity remained. However, he thought it was commensurate with the complexity of the analysis task and told us that a certain amount of training is also needed to become familiar with the standard software packages in the field.

7. Conclusion

Hypothesis testing based on data from diffusion tensor MRI is widely used to study how different brain diseases affect white matter microstructure. The multivariate nature of the diffusion tensor makes it challenging to formulate suitable statistical tests. Most available options are either so simple that they might miss important group differences, or so complex that reduced statistical power and interpretability limit their practical utility.

In this paper, we have argued for a middle ground in between these extremes, by introducing a steerable hypothesis test which can be customized to include any combination of the six degrees of freedom in the diffusion tensor, after reparametrizing tensor differences so that they become interpretable.

We also present a visual analytics system that supports the exploratory task of deciding which tensor attributes to include in the test. Closely following the requirements that have been agreed upon with domain experts, it provides a visual interface to specify different null hypotheses and to quickly explore the resulting regions within their anatomical context, based on the assumption that larger connected clusters are more likely to be relevant. Linking them to three-dimensional fiber tractography helps to identify the affected bundles. Finally, overlay views make it easy to compare the results from different tests, thresholds, or levels of smoothing.

We demonstrate our system in a case study on data from a clinical study on systemic lupus erythematosus. In an exploratory analysis, we arrived at several new hypotheses, in particular, that the disease goes along with changes in overall diffusivity, as well as with changes in tensor mode, in regions that are disjoint from the ones in which Fractional Anisotropy changes. We emphasize that exploratory and confirmatory analysis have to be done on independent cohorts. In our case study, the confirmatory part had to be left to a future work.

There are many interesting opportunities for future extensions of our system: First, adding visualizations that provide additional insights on which tensor attributes most strongly determined the result of a multivariate test, and how they correlate, would allow for a more in-depth analysis. Second, even though the diffusion tensor model is still dominant in clinical studies, it has long been known to be insufficient for modeling multiple fiber directions within the same voxel. High-angular resolution diffusion imaging is now considered to be a state-of-the-art alternative for fiber tractography [JDML19], and multi-shell models, such as diffusional kurtosis [JH10] or NODDI [ZSWKA12], provide even more detailed quantitative information. So far, only very few visualization techniques are available for these [BZGV14, SV19]. Third, a quantitative user study based on synthetic data could assess the agreement between user reported insights and true (simulated) group differences [ZZZK18]. Finally, despite the prevalence of hypothesis testing in clinical studies, Bayesian data analysis is certainly an interesting alternative [WJP*09].

References

[AFPA06] ARSIGNY V., FILLARD P., PENNEC X., AYACHE N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine 56*, 2 (2006), 411–421. 6

[AWHS16] ABBASLOO A., WIENS V., HERMANN M., SCHULTZ T.: Visualizing tensor normal distributions at multiple levels of detail. *IEEE Trans. on Visualization and Computer Graphics 22*, 1 (2016), 975–984. 2

[BF04] BARINGHAUS L., FRANZ C.: On a new multivariate two-sample test. *Journal of multivariate analysis 88*, 1 (2004), 190–206. 1

[BNH*16] BOUCHON A., NOBLET V., HEITZ F., LAMY J., BLANC F., ARMSPACH J.-P.: Which is the most appropriate strategy for conducting multivariate voxel-based group studies on diffusion tensors? *NeuroImage 142* (2016), 99–112. 8

[BP96] BASSER P. J., PIERPAOLI C.: Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *Journal of Magnetic Resonance, Series B 111* (1996), 209–219. 1

[BP03] BASSER P. J., PAJEVIC S.: A normal distribution for tensor-valued random variables: applications to diffusion tensor MRI. *IEEE Trans. on Medical Imaging 22*, 7 (2003), 785–794. 1

[BPP*00] BASSER P. J., PAJEVIC S., PIERPAOLI C., DUDA J., ALDROUBI A.: In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine 44* (2000), 625–632. 6

[BZGV14] BISTA S., ZHUO J., GULLAPALLI R. P., VARSHNEY A.: Visualization of brain microstructure through spherical harmonics illumination of high fidelity spatio-angular fields. *IEEE Trans. on Visualization and Computer Graphics 20*, 12 (2014), 2516–2525. 10

[Cer10] CERCIGNANI M.: Strategies for patient-control comparison of diffusion MR data. In *Diffusion MRI*, Jones D., (Ed.). Oxford University Press, 2010, pp. 485–499. 1

[JDML19] JEURISSEN B., DESCOTEAUX M., MORI S., LEEMANS A.: Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine 32*, 4 (2019), e3785. 10

[JH10] JENSEN J. H., HELPERN J. A.: MRI quantification of non-gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine 23*, 7 (2010), 698–710. 10

[JSCH05] JONES D. K., SYMMS M. R., CERCIGNANI M., HOWARD R. J.: The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage 26*, 2 (2005), 546–554. 8

[KEWW07] KINDLMANN G., ENNIS D. B., WHITAKER R. T., WESTIN C.-F.: Diffusion tensor analysis with invariant gradients and rotation tangents. *IEEE Trans. on Medical Imaging 26*, 11 (2007), 1483–1499. 1, 2, 4, 5

[Kin04] KINDLMANN G.: Superquadric tensor glyphs. In *EG/IEEE Symposium on Visualization (SymVis)* (2004), pp. 147–154. 3

[KSWS*17] KHATAMI M., SCHMIDT-WILCKE T., SUNDGREN P., ABBASLOO A., SCHÖLKOPF B., SCHULTZ T.: BundleMAP: anatomically localized classification, regression, and statistical analysis in diffusion MRI. *Pattern Recognition 63* (2017), 593–600. 8

[KZR*13] KEIHANINEJAD S., ZHANG H., RYAN N. S., MALONE I. B., MODAT M., CARDOSO M. J., CASH D. M., FOX N. C., OURSELIN S.: An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to alzheimer's disease. *NeuroImage 72* (2013), 153–163. 6

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE Trans. on Visualization and Computer Graphics 19*, 9 (2013), 1526–38. 7

[NH02] NICHOLS T. E., HOLMES A. P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping 15*, 1 (2002), 1–25. 5

[NH03] NICHOLS T., HAYASAKA S.: Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research 12*, 5 (2003), 419–446. 5

[OS15] O'DONNELL L. J., SCHULTZ T.: Statistical and machine learning methods for neuroimaging: examples, challenges, and extensions to diffusion imaging data. In *Visualization and Processing of Higher Order Descriptors for Multi-Valued Data*. Springer, 2015, pp. 299–319. 1, 2, 4

[Paj11] PAJEVIC S.: Statistical issues in diffusion tensor MRI. In *Diffusion MRI: Theory, Method, and Applications*, Jones D. K., (Ed.). Oxford University Press, 2011, ch. 20, pp. 331–353. 2

[PP99] PAJEVIC S., PIERPAOLI C.: Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain. *Magnetic Resonance in Medicine 42*, 3 (1999), 526–540. 6

[SBW06] SCHULTZ T., BURGETH B., WEICKERT J.: Flexible segmentation and smoothing of DT-MRI fields through a customizable structure tensor. In *Advances in Visual Computing (Proc. ISVC)* (2006), Bebis G., Boyle R., Parvin B., Koracin D., Remagnino P., Nefian A. V., Gopi M., Pascucci V., Zara J., Molineros J., Theisel H., Malzbender T., (Eds.), vol. 4291 of *LNCS*, pp. 455–464. 2, 5

[SDT10] SCHWARTZMAN A., DOUGHERTY R. F., TAYLOR J. E.: Group comparison of eigenvalues and eigenvectors of diffusion tensors. *Journal of the American Statistical Association 105*, 490 (2010), 588–599. 1

[SJJB*06] SMITH S. M., JENKINSON M., JOHANSEN-BERG H., RUECKERT D., NICHOLS T. E., MACKAY C. E., WATKINS K. E., CICCARELLI O., CADER M. Z., MATTHEWS P. M., ET AL.: Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage 31*, 4 (2006), 1487–1505. 8

[SN09] SMITH S. M., NICHOLS T. E.: Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage 44*, 1 (2009), 83–98. 2, 5

[Sri02] SRIVASTAVA M. S.: *Methods of Multivariate Statistics*. Wiley, 2002. 4, 8

[SSW*16] SHASTRI R. K., SHAH G. V., WANG P., CAGNOLI P., SCHMIDT-WILCKE T., MCCUNE J., HARRIS R., SUNDGREN P. C.: MR diffusion tractography to identify and characterize microstructural white matter tract changes in systemic lupus erythematosus patients. *Academic Radiology 23*, 11 (2016), 1431–1440. 8

[SV19] SCHULTZ T., VILANOVA A.: Diffusion MRI visualization. *NMR in Biomedicine 32*, 4 (2019), e3902. 10

[SWCW*14] SCHMIDT-WILCKE T., CAGNOLI P., WANG P., SCHULTZ T., LOTZ A., MCCUNE W. J., SUNDGREN P. C.: Diminished white matter integrity in patients with systemic lupus erythematosus. *NeuroImage: Clinical 5* (2014), 291–297. 2, 5, 7, 8

[WJP*09] WOOLRICH M. W., JBABDI S., PATENAUDE B., CHAPPELL M., MAKNI S., BEHRENS T., BECKMANN C., JENKINSON M., SMITH S. M.: Bayesian analysis of neuroimaging data in FSL. *NeuroImage 45*, 1 (2009), S173–S186. 4, 10

[WL16] WASSERSTEIN R. L., LAZAR N. A.: The ASA's statement on p-values: context, process, and purpose. *The American Statistician 70*, 2 (2016), 129–133. 4

[WWHT07] WHITCHER B., WISCO J. J., HADJIKHANI N., TUCH D. S.: Statistical group comparison of diffusion tensors via multivariate hypothesis testing. *Magnetic Resonance in Medicine 57*, 6 (2007), 1065–1074. 1, 3

[ZAY*07] ZHANG H., AVANTS B. B., YUSHKEVICH P. A., WOO J. H., WANG S., MCCLUSKEY L. F., ELMAN L. B., MELHEM E. R., GEE J. C.: High-dimensional spatial normalization of diffusion tensor images improves the detection of white matter differences: an example study using amyotrophic lateral sclerosis. *IEEE Trans. on Medical Imaging 26*, 11 (2007), 1585–1597. 2, 6

[ZCH*17] ZHANG C., CAAN M., HÖLLT T., EISEMANN E., VILANOVA A.: Overview + detail visualization for ensembles of diffusion tensors. *Computer Graphics Forum 36*, 3 (2017), 121–132. 2

[ZHC*17] ZHANG C., HÖLLT T., CAAN M., EISEMANN E., VILANOVA A.: Comparative visualization for diffusion tensor imaging group study at multiple levels of detail. In *EG Workshop on Visual Computing for Biology and Medicine* (2017), Bruckner S., Hennemuth A., Kainz B., (Eds.), pp. 53–62. 2

[ZSL*16] ZHANG C., SCHULTZ T., LAWONN K., EISEMANN E., VILANOVA A.: Glyph-based comparative visualization for diffusion tensor fields. *IEEE Trans. on Visualization and Computer Graphics 22*, 1 (2016), 797–806. 2, 5

[ZSWKA12] ZHANG H., SCHNEIDER T., WHEELER-KINGSHOTT C. A., ALEXANDER D. C.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage 61*, 4 (2012), 1000–1016. 10

[ZYRG07] ZHANG H., YUSHKEVICH P. A., RUECKERT D., GEE J. C.: Unbiased white matter atlas construction using diffusion tensor images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2007), Springer, pp. 211–218. 2, 6

[ZZZK18] ZGRAGGEN E., ZHAO Z., ZELEZNIK R., KRASKA T.: Investigating the effect of the multiple comparisons problem in visual analysis. In *Proc. Conf. on Human Factors in Computing Systems (CHI)* (2018), p. 479. 10