Statistical and Machine Learning Methods for Neuroimaging: Examples, Challenges, and Extensions to Diffusion Imaging Data

Lauren J. O'Donnell and Thomas Schultz

Abstract In neuroimaging research, a wide variety of quantitative computational methods enable inference of results regarding the brain's structure and function. In this chapter, we survey two broad families of approaches to quantitative analysis of neuroimaging data: statistical testing and machine learning. We discuss how methods developed for traditional scalar structural neuroimaging data have been extended to diffusion magnetic resonance imaging data. Diffusion MRI data have higher dimensionality and allow the study of the brain's connection structure. The intended audience of this chapter includes students or researchers in neuroimage analysis who are interested in a high-level overview of methods for analyzing their data.

1 Introduction

The study of the human brain was originally performed by expert dissection of fixed brains. Now, with the advent of structural and functional neuroimaging, we can apply quantitative computational analyses to study and model the brain in vivo. Neuroimaging analyses have important scientific and clinical applications that include the study or diagnosis of disease, the measurement of change, the detection of neural activation, and the modeling of anatomy. In this chapter, we aim to provide a general overview of analysis approaches for neuroimaging data, including some specific examples of neuroimaging studies.

Much of the research in the neuroimage analysis field has focused on the analysis of scalar data, such as structural magnetic resonance imaging (MRI) or computed tomography (CT), where a single scalar value is present at each voxel. Another large body of analysis research focuses on detection of neural activations using

L.J. O'Donnell (✉)
Harvard Medical School, Boston, MA, USA
e-mail: odonnell@bwh.harvard.edu

T. Schultz
University of Bonn, Bonn, Germany
e-mail: schultz@cs.uni-bonn.de

I. Hotz, T. Schultz (eds.), *Visualization and Processing of Higher Order Descriptors for Multi-Valued Data*, Mathematics and Visualization,
DOI 10.1007/978-3-319-15090-1_15

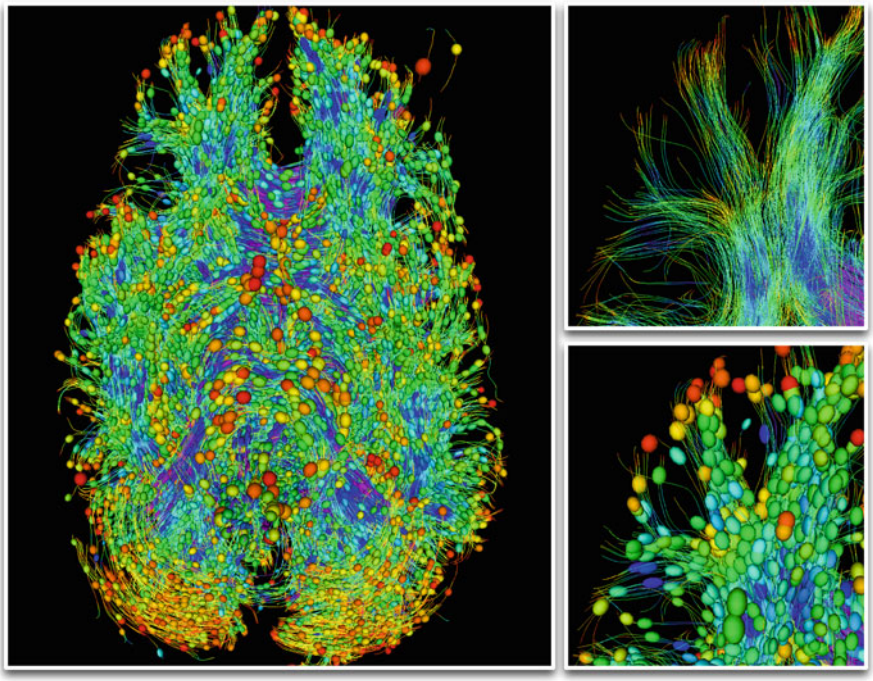

Fig. 1 Example diffusion MRI data, including fiber tract trajectories from tractography with selected (randomly sampled) ellipsoids to visualize diffusion tensors along the tracts. At *left*, the whole brain is shown in an inferior view. At *right*, zoomed images show fiber tract trajectories (*top*) plus ellipsoids (*bottom*). The tracts and ellipsoids are colored by fractional anisotropy (FA), a popular scalar measure for statistical analyses of diffusion MRI. *Blue* and *purple* are high FA, *green* and *yellow* are intermediate values, and *red* is low FA

timecourse data: the blood-oxygen-level dependent (BOLD) signal of functional magnetic resonance imaging (fMRI). Of particular interest in this chapter is the analysis of diffusion MRI, the only non-invasive scan for measurement of the brain's connectional structure. The traditional representation of diffusion MRI data is not scalar. Rather, it is a tensor (specifically, a 3×3 symmetric, positive-definite matrix) at each voxel. Data employing the tensor representation are called diffusion tensor MRI or DTI. In current diffusion MRI research, higher-order models (as well as connectivity data) may also be reconstructed from the scan [58]. Figure 1 shows both tensor and connectivity (fiber tract) data from a diffusion MRI scan of a healthy human brain.

Because different types of neuroimaging data have different data dimensionalities as well as vastly different interpretations in the context of the brain, and because neuroimaging studies have many possible designs, the analyses developed for neuroimaging data are manifold. Analyses have been developed for scalar data, for timecourse data, for tensor-valued data, and for many other types of

data representations such as measurements from regions of interest or along image skeletons.

To organize this chapter, we categorize quantitative neuroimaging analyses into two groups according to their overall philosophy: statistical testing or machine learning. In statistical testing, the goal is to obtain a result that is statistically significant: unlikely to have arisen by chance. Most often, these approaches are applied to measure a result, such as a functional or structural difference, between groups of subjects. Statistical testing methods are also in regular clinical use to detect functional brain activations in individual patients. The supervised machine learning methods that are treated in the second part of this chapter learn computational models that estimate or predict the values of unobserved variables. During learning, they are given access to labeled training data, for which the value of the variable of interest is known, such as images categorized into healthy control and patient images, or annotated with subject age. In a second step, the respective quantity—such as disease state or age—is estimated based on other brain scans for which it is unknown. Statistical and machine learning methods for exploratory data analysis, such as clustering or Principal Component Analysis (PCA), are outside of our main focus, even though a use of PCA as part of a predictive model is discussed in Sect. 3.4.

In the rest of this chapter, we survey examples from the scalar neuroimaging field, and where possible we describe extensions or new methods developed for the analysis of diffusion MRI data. The chapter is divided into two parts: first, the more traditional statistical testing approaches, and second, the more recent machine learning approaches.

2 Methods for Neuroimaging Analysis That Use Statistical Tests

In neuroimaging research, statistical tests are used in many scenarios. Examples include: to find regions of significant difference between two populations in a clinical neuroimaging study, to find regions of neural activation in fMRI, or to detect abnormal regions that differ from a model of the healthy brain. In the rest of this section, we first describe basic concepts, then we give examples of popular methods that employ statistical testing in neuroimaging data, and finally we describe extensions of the statistical testing frameworks that have been proposed for analysis of diffusion MRI data.

2.1 Basic Concepts and Potential Problems

We begin this section with a simple example that motivates the vocabulary and the basic concepts used in statistical hypothesis testing. Readers familiar with this may wish to skip ahead to the overview of methods that have been developed for diffusion MRI.

In the statistical hypothesis testing framework, there is generally a *null hypothesis* H_0, such as "There is no difference between the two study groups, thus their data have the same mean." A corresponding *alternative hypothesis* H_a^1 could be, in this simple example, that the means of the data from the two groups are different. (Another H_a^2 could be, for example, that the mean of one group is larger than the other.)

To assess this possible difference, a *test statistic* is chosen. In our example, the test statistic should be a quantity related to the difference in means, such as the popular t-statistic [64]. The *null distribution* is a probability distribution that gives the probability, under the null hypothesis, of observing values of the test statistic. The null distribution can be known or estimated from the data. In our example, armed with the null distribution and an observed test statistic, the researcher will determine the conditional probability of observing the test statistic if both groups have the same mean (the null hypothesis).

If the observed test statistic is found to have low probability under the null hypothesis, the reasoning is that the observed test statistic is unlikely to have occurred by chance. Thus there may be an experimental finding: it may be possible to reject the null hypothesis in support of the alternative hypothesis. To decide whether to reject the null hypothesis, the *statistical significance* of the observed test statistic is determined by calculating a *p-value*, the probability of observing a statistic at least as extreme as the observed statistic (under the null distribution). Here, the word "extreme" refers to the tails of the null distribution, where the probability of observing the test statistic values is low: For H_a^1, the first alternative hypothesis mentioned above, both tails would be considered to be extreme ("two-tailed test"). For H_a^2, only the tail corresponding to larger values would be taken into account ("one-tailed test"). If the calculated p-value falls below a predetermined threshold or *alpha level*, such as 0.05, the result may be considered significant. Alternatively, for a given alpha level, the test statistic can be compared to a threshold for which 5 % of the area of the null distribution is located under the tail(s).

Potential problems in hypothesis testing have been widely discussed, for example in the book "The Cult of Statistical Significance" [76]. Issues include incorrect rejection of a true null hypothesis, called *type I error,* or false positive error. In the context of neuroimaging, this type of error would lead to publication of a false finding. Type I errors are typically controlled at an alpha level of 0.05, which means that statistical tests commonly used in neuroimaging have a 5 % chance of rejecting a true null hypothesis due to chance. A second issue is *type II error* or false negative error. This means that statistical significance of an effect, even though it is true, cannot be shown based on the acquired data.

It is clear that performing multiple tests (*multiple comparisons*) is dangerous: eventually, one of the tests will produce a significant value. If this is not correctly accounted for, the overall chance of a type I error can increase drastically. An infamous illustration of this was given by an fMRI experiment in which activation was found in the brain of a dead salmon [8]. Popular strategies for correcting this potential source of error are mentioned in the next section.

2.2 Popular Neuroimaging Analyses

Here we give a brief overview of two main approaches to data analysis: voxel-based, where data are measured and statistics are performed in a large number of voxels throughout the brain, and region-based, where data measurement and statistical analyses are restricted to neuroanatomical regions generated by image segmentation. We note that neuroimage analysis methods may also be categorized according to the number of subjects analyzed. Often, analyses employ a population, or a neuroimaging dataset that includes data from multiple subjects. However, some analyses are inherently single-subject, such as fMRI activation detection in neurosurgical patients.

Voxel-Based Statistics

There is a large and sophisticated body of literature on voxel-based morphometry (VBM) and statistical parametric mapping (SPM) in structural and functional imaging [26]. These approaches use the general linear model (GLM) framework, a linear regression model that incorporates covariates and any indicator variables reflecting study design [28]. The overall idea is that parameters of interest are calculated from the GLM, then a parametric statistical test is applied at each voxel, such as the t-test or F-test. In VBM, traditionally the gray matter is segmented and smoothed, giving a map of gray matter concentration that is compared across groups [4]. In fMRI analysis, where the per-voxel information is a vector of time-course data, traditionally the GLM approach uses regression to obtain a single scalar parameter for univariate statistical testing [29]. The voxel-based approach assumes that anatomy corresponds across subjects at the voxel level, and thus smoothing and image registration play important roles. Statistical analyses called deformation-based or tensor-based morphometry generally analyze the Jacobian determinants of the vector-valued deformation fields generated by image registration [5].

In voxel-based analyses, multiple comparisons arise naturally because the tests are performed at many anatomical locations within the brain. Several statistical methods may be employed to correct for multiple comparisons, including the stringent Bonferroni correction, where the threshold for statistical significance is adjusted to account for the multiple tests. The Bonferroni correction assumes tests are independent, which is not the case in spatially smooth image data, and leads

to an overly conservative correction, reducing statistical power. Thus, the theory of Gaussian random fields is employed in SPM to correct for multiple comparisons [4]. An alternative that controls the expected proportion of false positives within a statistical map, rather than the probability that any part of the map includes a false positive, is the false discovery rate (FDR) [7, 30]. However, simple application of FDR does not take into account the fact that voxels are spatially contiguous and represent continuous data [11]. In another approach, a summary or maximal test statistic (such as maximum suprathreshold cluster size) may be used to summarize information from multiple statistical tests across voxels, and the null distribution may be estimated for this new, overall test statistic. This strategy may be used in combination with permutation testing for computation of the null distribution [47]. Permutation tests are increasingly used because they are powerful, non-parametric, and simple to perform by repeatedly randomizing the labels of the data. However, they can be computationally intensive.

Region-Based Statistics

In the case where there is a hypothesis about the likely region of an effect (for example, if the corpus callosum is hypothesized to differ between groups), a region of interest (ROI) can be created for measurement. This may be done via a manual or automated image segmentation procedure. Scalar measurements are made, such as the ROI's volume or the mean value of image voxels within the ROI. This approach can avoid the multiple comparisons problem, if only one ROI is measured, and only one type of information is measured from that ROI. More typically, data from more than one ROI are measured, and Bonferroni or FDR correction would be appropriate. Traditional t-tests and ANOVA are very commonly used in the neuroimaging literature to identify possible differences between groups in ROI-based studies, for example [62].

2.3 Extension of Analyses to Diffusion MRI

We give examples of analyses in the voxel-based and region-based frameworks, as well as methods where statistical tests have been developed to deal with unique types of data from diffusion MRI. We begin with voxel-based and region-based methods that operate on scalar values derived from diffusion MRI, most commonly the fractional anisotropy (FA). Next we describe statistical methods that have been developed for diffusion MRI tracts, followed by methods for vector and tensor data estimated from diffusion MRI. This is by no means an exhaustive list of references from the field; rather, we intend to provide examples illustrating the main concepts.

Voxel-Based Statistics Proposed for Diffusion MRI

At this point, standard VBM studies are not often performed on diffusion MRI data. It has been shown that results are highly sensitive to the size of the smoothing kernel [36] and that image registration often fails to match the high FA core of the white matter tracts [63]. Furthermore, there are issues with non-normally distributed residuals after fitting a GLM model [36].

The most popular voxel-based analysis of diffusion MRI data was designed to address these issues. Though it is a voxel-based method, it is called Tract-Based Spatial Statistics [63]. In this method, to ameliorate registration difficulties and to restrict analyses to the presumed core of the tract, locally high FA values are projected onto voxels of a group FA skeleton. After this procedure, the voxels of the groupwise skeleton have been attributed with data from every subject in the study, and standard GLM analyses may be used.

Methods have also been investigated for diagnostic analysis of diffusion MRI on the single subject level. Diffusion MRI is of particular interest as a sensitive marker for traumatic brain injury, where a quantitative marker is desired to help in diagnosis and prognosis. Initially, standard VBM techniques were applied to investigate brain changes by comparing an individual to controls [42]. Then alternative voxel-based analyses were designed to detect abnormal regions within the single subject, based on comparison to a model of normal diffusion that employs data from multiple control subjects. An FA-based method that employs bootstrap methods for estimating control population variance and corrects for covariates such as age and gender has been developed to assess departure from the normal model using z-scores [43].

Region-Based Statistics Proposed for Diffusion MRI

Existing image segmentation and measurement pipelines may be applied to any scalar data derived from diffusion MRI. Additionally, many diffusion-MRI specific methods exist for defining white matter tracts, including deterministic and probabilistic tractography methods for estimation of white matter connections. For a recent overview of tractography segmentation methods, see [49]. Many diffusion MRI analysis pipelines use atlases derived from tractography, such as the Mori atlas [69] to define regions of interest in individual subjects. Once tract ROIs have been defined, they can be used for measurement of quantities such as the average FA within the tract. This enables region-based statistical analyses.

After ROI definition, measurement and statistical analysis are the same as for any imaging ROI study, except for the fact that there are many scalar parameters that may be measured from one diffusion MRI scan. For a basic diffusion tensor reconstruction, scalars can include FA, mean diffusivity (MD), and more. Thus the multiple comparisons problem may be more severe in diffusion MRI studies.

Tract-Based Statistics Proposed for Diffusion MRI

In more sophisticated data analyses than average measurement within an entire tract region, fiber tracts have been used for structure-specific statistical mapping. Since tracts can be considered to have a linear structure (connecting brain region A to brain region B), one option is to analyze data along the tract. This style of analysis measures data versus arc length along a tract [13]. Methods for measurement and analysis have been proposed by many authors. Simple averaging of data at points along the tract and use of permutation testing found significant differences across hemispheres [48]. Authors have proposed more sophisticated machinery, such as using an extension of multivariate statistics called functional regression analysis [31, 75]. Fiber tractography in a DTI atlas was employed to define and parameterize tracts in conjunction with the Hotelling T^2 statistic to analyze both FA and tensor norm [31]. Analysis of data along tracts has been shown to have advantages over simple averaging of the data, which may mask differences [12, 48]. Related approaches have proposed analysis over the entire tract surface, representing it as a sheet, rather than attributing a single trajectory with data [74].

Eigenvector and Tensor Statistical Tests Proposed for Diffusion MRI

Some disagreement exists regarding an appropriate manifold for diffusion tensors. A Riemannian metric between diffusion tensors was proposed [3, 6, 23, 24, 41] in order to restrict analyses to the space of positive definite symmetric matrices. However, others believe that a Euclidean metric is more appropriate for actual diffusion MRI data [51]. Several groups have investigated geodesics for interpolation of diffusion tensors [22, 38]. However, recent work on smoothing may indicate that the metric between tensors has little practical effect for data analyses [67]. Each metric may be useful for certain computational tasks: in registration, the log-Euclidean metric may be used for reducing blurring when averaging, while the Euclidean metric performs well for the objective function [37].

Limited work exists on statistical testing for group differences in principal diffusion directions (major eigenvectors) and in entire diffusion tensors. A statistical method based on the bipolar Watson distribution was proposed to test whether the principal diffusion direction had the same mean in two groups of subjects [60]. This test was shown to detect differences that were invisible to a more standard FA analysis [60]. Additional work by the same author investigated tensor statistics [59] and gave further insight into FDR correction for the eigenvector testing [61]. Another group investigated the application of several multivariate statistical tests directly to the components of the full diffusion tensor [73].

3 Regression and Classification in Neuroimaging

Given samples from one or multiple populations, statistical hypothesis testing allow us to infer statements about parameters describing those populations. In the context of neuroimaging, frequent examples of populations in the statistical sense include groups of subjects, voxels, or time steps.

Recently, methods from machine learning are increasingly being used to make statements about individual samples, rather than populations. Example applications include supporting the diagnosis of disease based on examples of both healthy and diseased subjects [25], estimating a person's brain maturity [21, 27], detecting which class of object a person is currently looking at [14], whether or not he or she is telling the truth [19], or predicting behavior [32].

Building a system that facilitates such predictions requires selecting a suitable machine learning method, extracting mathematical descriptors ("features") on which further analysis can be based, and selecting features that are particularly relevant to the task. Obtaining a reliable estimate of a method's accuracy can pose serious and surprising pitfalls. Finally, it is desirable, though unfortunately difficult, to gain insight on how the machine learning method arrived at its final estimate.

In this section, we will elaborate on each of these steps. Since the field is young, new methods are evolving rapidly, and no widely used standards have been established so far. Therefore, we cannot hope to provide a final and exhaustive overview, but rather focus on general principles and examples of solutions that have been found to be effective on more than a single dataset and, ideally, by different groups. We are particularly interested in examples involving diffusion MRI and multimodal imaging, which have been excluded from an earlier, related overview [52].

3.1 Methods for Classification

In the context of neuroimaging, classification is the assignment of a subject or a cognitive state to a specific class, such as recognizing that a subject suffers from a specific disease, or is currently looking at an example from a certain class of objects. Mathematically, classification is performed by a function $f(x)$ that maps an instance $x \in \mathcal{X}$, the subject or cognitive state, to a discrete output variable ("label") y, which encodes the different classes. In practice, x is usually represented by an m-dimensional feature vector $\mathbf{x} \in \mathbb{R}^m$.

Training a classifier amounts to learning the function f from a training dataset $\{(\mathbf{x}_i, y_i)\}$, $i = 1, \ldots, n$ so that $f(\mathbf{x}_i) = y_i$ for as many training examples as possible. At the same time, f should be as "simple" as possible, in a sense that can be made mathematically precise [57], to maximize the chance that it will produce correct results also for novel inputs $\tilde{\mathbf{x}}$ which have not been part of the training data.

In neuroimaging applications, it is common to have a high-dimensional feature space, but relatively little training data ($m \gg n$). Support Vector Machines (SVM) are widely used as a classifier, since they are known to be able to deal with this situation relatively well [57]. They are based on finding a hypersurface in the feature space that correctly separates as many of the training samples as possible, while also maximizing the distance of the decision boundary to the samples that are correctly classified.

SVMs can be generalized to nonlinear classification by implicitly mapping the features into an abstract higher-dimensional space using the "kernel trick" [57]. While LaConte et al. [40], working with very high-dimensional feature vectors to begin with, do not find a clear benefit from mapping them to an even higher-dimensional space, Wee et al. [71] report a noticeable increase in accuracy when using nonlinear kernels with moderately sized feature vectors, and Rasmussen et al. [53] construct an example in which a nonlinear kernel aids classification even in high-dimensional space. Ultimately, no single kernel is optimal for all applications, and classification accuracy can often be increased by trying different alternatives.

Aside from support vector machines, the machine learning literature offers a wide range of classifiers that are occasionally used in neuroimaging, including Fisher Linear Discriminant Analysis (LDA) [25] and maximum uncertainty Linear Discriminant Analysis (MLDA) [18], naive Bayesian classifiers [45], k Nearest Neighbor classifiers [70], neural networks [2], and random forests [1]. For more detailed explanations of these methods and further pointers to the machine learning literature, we refer the reader to [9].

Sometimes, it is desirable to combine the results from multiple classifiers. For example, in multimodal imaging, a separate classifier might be created for each modality, and a single prediction y has to be derived from their outputs. In the simplest case, it can be based on a majority vote [35]. A natural improvement of this is to weight the impact of each classifier by its estimated accuracy [18]. In adaptive boosting (AdaBoost), this idea is combined with an iterative training of classifiers on re-weighted training samples, so that classifiers trained at later stages focus on examples misclassified previously [44].

3.2 Methods for Regression

Regression differs from classification mainly in the fact that the output variable y is continuous, such as age or brain maturity [21, 27], rather than discrete. Many methods for classification have a closely related variant that can be used for regression. An example is support vector regression [57] which, like support vector classification, produces a function f that can be written in terms of a subset of the training data, the so-called support vectors. Relevance vector regression, as it was used in [27], generally provides an even sparser representation of a similar form and at similar accuracy, at the cost of a more difficult and time consuming training process.

Many aspects of learning a function $f(x)$ that will be discussed in the remainder of this section are common to classification and regression. In this case, we will refer to methods that create such functions as "learning machines".

3.3 Feature Extraction

Feature extraction is the process of producing a feature vector **x** from the image data. It will subsequently represent a subject or cognitive state and contain information relevant to the classification or regression task. Initially, the individual images often undergo the same preprocessing that would be used for voxel-based statistical analysis, as it was explained in section "Voxel-Based Statistics". This includes normalization to a standard space, so that each voxel position (approximately) corresponds to the same anatomical structure, often followed by smoothing to reduce image noise and to compensate for residual misalignment.

At this point, each voxel could in principle be turned into an entry of the feature vector [40]. Often, a shorter feature vector is desired and is achieved by averaging values over larger blocks of voxels [19] or over predefined functional regions [18, 21], whose selection may be informed by prior knowledge on the regions involved in specific tasks or conditions [20].

In the context of diffusion MRI, feature extraction often makes use of the pipeline developed for Tract-Based Spatial Statistics (TBSS), which was explained in section "Voxel-Based Statistics Proposed for Diffusion MRI". In this case, the features are given by the values on the TBSS skeleton [33, 56], sometimes averaged over predefined white matter regions [15].

A more complex way of deriving feature vectors from diffusion MRI involves a brain connectivity graph constructed using tractography. To this end, Wee et al. [71] first parcellate the brain into anatomical regions of interest and detect which of them are connected by a deterministic full-brain tractography. The resulting graph is represented as an adjacency matrix, where edges are alternatively weighted by fiber count, Fractional Anisotropy, Mean Diffusivity, or any of the three diffusion tensor eigenvalues, and the resulting six matrices are vectorized and concatenated to form the final feature vector. In a follow-up work, these dMRI-based connectivity matrices have been combined with ones constructed from correlations in resting-state fMRI [72].

3.4 Feature Selection and Feature Weighting

Even though many learning machines are in principle able to operate on high-dimensional feature spaces, their effectiveness can be reduced when feature vectors include components whose variation does not carry any information about the desired output y, especially when, in addition, little training data is available. This

is particularly relevant for some of the feature vectors described in the previous subsection, which can be very high-dimensional ($m \approx 10^6$), and often include information from all regions of the brain, even if only some small specialized area may be affected by a disease or relevant to a task.

Initially, one often attempts to give similar influence to all features ("feature normalization"), for example by subtracting the mean and dividing by the standard deviation, or by linearly rescaling all features to some fixed interval [15]. Subsequently, a crucial step in most applications of machine learning in neuroimaging is to reduce the impact of features which are less relevant to the task at hand.

Feature selection methods attempt to find a subset of features which is particularly well suited for building a learning machine. In order to arrive at an optimal solution, one would have to evaluate each possible combination of features, which is infeasible in most cases. Therefore, a frequently used strategy is to first rank features according to their expected utility, and to include the top k features in the final feature vector.

In neuroimaging, the Fisher score (as it would be used in an F-test [15]), the t score (as it would be used in Student's t-test [71]), and, in case of regression, the Pearson correlation coefficient [21], are particularly popular for ranking features, possibly due to their ubiquitous use for statistical testing on the same type of data. As an alternative to these straightforward methods, a family of heuristics known as Relief, ReliefF, and RReliefF [54] is occasionally used [33, 50], and offers the advantage of being able to detect nonlinear dependencies between features and labels, as well as providing a higher rating of features that are only useful when used in combination, whereas the simple methods rate each feature in isolation.

Once a ranking has been achieved, the number k of features that should be included can be found by cross-validation [71], which will be explained in greater detail in Sect. 3.5. As a computationally less demanding alternative, sometimes only features are used whose difference between labels is statistically significant [18], or the number of retained features is simply set to some constant value [21].

Traditional techniques for dimensionality reduction such as Principal Component Analysis have also been used [27], but have occasionally been found to perform worse than other feature selection schemes [71]. This might be explained by the fact that, unlike all methods described above, they only consider the feature vectors $\mathbf{x}_i$ in isolation, and do not account for their relationship to the labels y_i.

An alternative to feature selection is feature weighting, which assigns a greater influence to some features than to others, rather than eliminating features completely. For example, Schmidt-Wilcke et al. [56] and Schlaffke et al. [55] scale all features by their corresponding F score, which avoids the need to decide how many features to retain.

Related to the idea of feature weighting are multiple-kernel SVMs, which are based on a weighted sum of several distance measures (kernels) between the $\mathbf{x}_i$, each of which might depend only on a certain subset of features. A natural application of this concept is multimodal imaging, where each modality is represented by a separate kernel [72].

3.5 Validation and Parameter Tuning

Once a function $f(x)$ has been trained for classification or regression, its accuracy can be estimated by applying it to a set of test data $\{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}$, and measuring the difference between the predictions $f(\tilde{\mathbf{x}}_i)$ and the true $\tilde{y}_i$. In order to ensure that the resulting estimate is unbiased (i.e., not overly optimistic), it is essential that the test data may not overlap with the data that has been used for training.

When data is available from only relatively few subjects, as it is quite common in neuroimaging, setting part of it aside for testing only allows us to evaluate accuracy on very few examples, leading to estimates that may be unbiased (on average, we do not overestimate accuracy), but have high variance (individual estimates of accuracy are highly uncertain). This problem can be reduced by applying cross validation, in which the learning machine is trained repeatedly on part of the data. In particular, in n-fold cross validation, the data is distributed equally between n sets ("folds"). Based on these, the learning machine is trained n times, each time using data from $n-1$ folds, and evaluating the result on the data from the remaining fold. The final estimate of accuracy is obtained by averaging the results from all n iterations. A special case of this is leave-one-out cross validation, in which the number of folds coincides with the size of the available training dataset, so that, in each iteration, only one sample $(\mathbf{x}_i, y_i)$ is left out of the training set.

Most learning machines have parameters that need to be set, such as choosing a kernel and setting a regularization parameter in support vector machines, or deciding how many features to retain in feature selection. While some authors simply use fixed default settings [21], results can often be improved greatly by evaluating alternative settings using cross-validation, and using the one that led to the highest estimated accuracy.

When cross-validation is used for parameter tuning, obtaining a reliable estimate of the final accuracy requires nested cross-validation, so that an outer cross-validation loop, which is responsible for estimating the overall accuracy, separates the data into a training and a testing set, and the inner loop, which performs the parameter tuning, may only access the training data set from the outer loop.

A subtle consequence of this, which is sometimes overlooked, is that in order for cross-validation to be effective, *only the training data* may be used for feature selection. A pragmatic safeguard against accidental double dipping is to attempt classification of the same data with randomly permuted labels y_i, to repeat this a large number of times, and to observe the resulting distribution of accuracies. Dosenbach et al. [21] perform this experiment as a part of validating their method; Schmidt-Wilcke et al. [56] use it as a permutation-based test to assess significance of their classification results.

Since a classifier cannot be expected to predict random labels with larger-than-random accuracy, an unbiased estimate should, on average, result in the same accuracy as a random guess. This is illustrated in Fig. 2: It is based on 100 iterations in which random group labels have been assigned to 18 healthy subjects, and a support vector machine with feature weighting has been trained to predict those

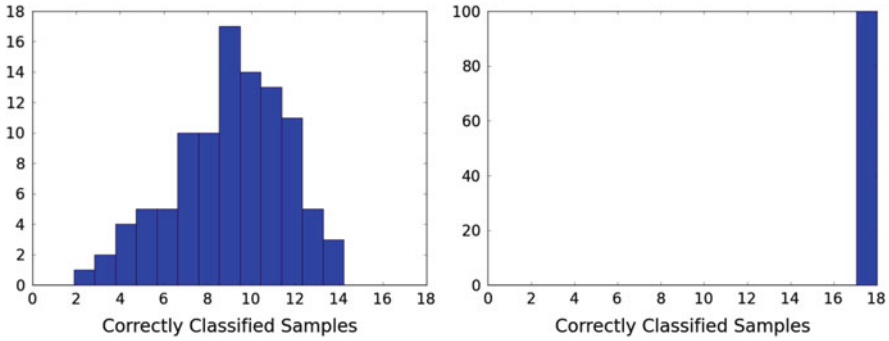

Fig. 2 On average, trying to predict 100 different sets of random labels cannot lead to better-than-random accuracy (*left*). However, if feature weighting is performed outside of the cross validation loop, the classifier is erroneously reported to achieve perfect results in each case (*right*)

random labels from MRI data. On the left, feature weighting is done correctly, within the leave-one-out cross validation. As expected, on average, the classifier does not achieve higher accuracy than a random guess. On the right, the same classification is attempted using the same method; the only difference is that, similar to [33, 50], feature weighting has now been performed as a pre-process on all data, including the test data. This leads to the misleading estimate that random labels can be predicted with perfect accuracy in all cases.

This surprising pitfall can be explained by the fact that we are given few data points with a huge number of features, many of which take on random values. This means that, given arbitrary class assignments, the feature vector includes features that happen to separate the data into those classes by pure chance. Performing feature selection on the whole dataset allows the classifier to operate on those features, without having any independent data left to check whether or not they actually contained legitimate information, or were correlated with the labels only by chance. This is similar to the fact that, after selecting a region of interest (ROI) based on correlations with another random variable, it is no longer meaningful to perform a statistical test on those correlations within that same ROI [68].

In summary, Fig. 2 illustrates that performing feature selection or feature weighting outside the cross validation loop can bias estimates so severely that they lose all meaning. While many works have avoided this problem by a correct setup [21, 27, 35, 71], some others merely acknowledge that performing feature selection as a pre-process on all data might lead to results that are "too optimistic, probably related to some degree of over-fitting" [33] or that "validation in an independent sample will be essential to determine how robust the current approach is when applied to a fully independent dataset" [50], which does not appreciate the full severity of the problem. Importantly, one should never attempt to compare accuracies from a correct setup with those reported after doing feature selection on the full dataset.

The fact that this section has illustrated one difficulty in correctly applying machine learning to neuroimaging data should not be taken as an indication that these techniques are fundamentally flawed: They rest on a solid statistical foundation [66] and, when applied correctly, they have already led to results that could be reproduced across different datasets [21, 35], and learning machines trained on one scanner have been tested successfully on data from other scanners [27].

3.6 Interpretation and Visualization

Since a fundamental goal in most neuroimaging studies is to better understand how the structure of the brain and its activity relate to specific functions, or to factors such as gender, age, and disease, it is desirable to obtain not only a classification or regression result from applying a learning machine to the data, but to gain at least some level of understanding of how it arrives at its prediction, e.g., which regions of the brain were most important for detecting a specific disease.

Most machine learning methods are designed to achieve the highest possible accuracy, whereas interpretability by a human operator is usually not a primary design goal. One common way to still glean some insight is to consider the results of feature selection. For example, if each feature corresponds to the average over some region of interest (ROI), the selected features indicate which regions were used to achieve the classification. When cross validation is used, different features might be selected in each iteration, and it is common to only report the features that are selected most frequently [15, 71] or even in all cases [18, 21].

Closely related to this, some authors compare the accuracies that can be achieved when making different parts of their data available to the classifier. For example, in the context of diffusion tensor MRI, this may indicate whether Fractional Anisotropy, Mean Diffusivity, or individual eigenvalues allow for more reliable detection of a certain disease [71].

As part of their training, linear classifiers, such as linear support vector machines or linear discriminant analysis, compute a weight with which each feature contributes to the final result. If features were appropriately normalized, this makes it natural to inspect the weight vectors as an indicator of feature importance. In fact, when features correspond to individual voxels or small ROIs, weight vectors can be visualized as spatial maps, similar to activation maps from mass-univariate statistical analysis [40]. Support vector machines have recently been extended to increase spatial regularity of the resulting maps, with the goal of making them more interpretable [17].

However, an important caveat in the interpretation of weight vectors is that classifiers may put significant weight on features that are unrelated to the given task or disease, and that the largest weights do not necessarily correspond to the features which are most strongly related to the label. In particular, Haufe et al. [34] provide examples in which features are only included to cancel out artifacts that are also present in truly informative features and might obtain an even greater weight.

Even though weight vectors have been found to agree with prior knowledge about abnormalities in Alzheimer's disease [17, 39] and have indicated neuroanatomically plausible regions in cases where mass univariate analysis failed to detect significant differences [16], Haufe et al. [34] conclude that the only truly firm conclusion that can be drawn from weight vectors of a successful classifier is that at least one of the features with non-zero weight is associated with the given condition or task.

When support vector machines are used with a nonlinear kernel, the weight vector is defined in an abstract higher-dimensional space, and generally cannot be mapped back to the original feature vector [57]. However, sensitivity analysis [40, 53] can still quantify how much impact each feature has on the classification. In the linear case, sensitivities amount to the squared feature weights [53], so they suffer from the same limitations with respect to their interpretability.

4 Main Challenges and Conclusions

As discussed in section "Voxel-Based Statistics", spatially contiguous regions play an important role in maintaining statistical power while correcting for multiple comparisons in mass-univariate statistical testing. In contrast, most learning machines act on abstract feature vectors, and are oblivious of the underlying spatial structure. Even though attempts have recently been made to increase accuracy and interpretability of classifiers by spatial regularization [17], it is still widely unexplored how to best account for spatial and anatomical structures when training learning machines, and how much is to be gained from it. Taken to the extreme, Honorio et al. [35] have demonstrated that, on several datasets with a limited number of subjects each, classification based on a single discriminative region of interest outperformed some widely used multivariate methods that were found to make use of a larger number of scattered voxels.

While there is hope that the multivariate analysis afforded by machine learning techniques will lead to an understanding of interactions and dependencies that would remain hidden to mass-univariate approaches, interpretation and visualization of what allows a learning machine to perform successful classification or regression remains a difficult task [34], and merits further work.

Applications of machine learning to diffusion MRI have so far mostly been based on features derived from the second-order diffusion tensor model. However, it is now common to acquire more complex diffusion MR data that requires higher-order models, including High Angular Resolution Diffusion Imaging (HARDI), Diffusion Spectrum Imaging (DSI), and multi-shell data. Only few initial works exist on extracting features suitable for machine learning from such models, based on spherical deconvolution [10], or a spherical harmonics expansion of apparent diffusivities [46]. There is still a need to explore alternative features based on such rich and complex data, and to evaluate their power and reliability in a range of applications.

Finally, training highly accurate learning machines and obtaining a realistic impression of their performance requires more data than is typically acquired for traditional statistical analysis. Currently, relatively few groups have the opportunity to apply machine learning to sufficiently uniform datasets that include hundreds of subjects [21, 27]. However, larger datasets, such as the ones from the Human Connectome Project [65], are currently becoming available to the general research community, and are about to open up new horizons for the development and evaluation of machine learning on neuroimaging data.

Acknowledgements This work has resulted from a series of breakout sessions at Dagstuhl seminar 14082. We thank Anna Vilanova (TU Delft, The Netherlands) for her collaboration in those sessions, and for her help in organizing the LaTeX structure of this chapter. Author LJO thanks NIH grant support R01MH074794, P41EB015902, R21CA156943, P41EB015898, and U01NS083223.

References

1. Anderson, A., Dinov, I.D., Sherin, J.E., Quintana, J., Yuille, A.L., Cohen, M.S.: Classification of spatially unaligned fMRI scans. NeuroImage **49**(3), 2509–2519 (2010)
2. Arribas, J.I., Calhoun, V.D., Adalı, T.: Automatic bayesian classification of healthy controls, bipolar disorder and schizophrenia using intrinsic connectivity maps from fMRI data. IEEE Trans. Biomed. Eng. **57**(12), 2850–2860 (2010)
3. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magn. Reson. Med. **56**(2), 411–421 (2006)
4. Ashburner, J., Friston, K.J.: Voxel-based morphometry—the methods. NeuroImage **11**(6), 805–821 (2000)
5. Ashburner, J., Hutton, C., Frackowiak, R., Johnsrude, I., Price, C., Friston, K., et al.: Identifying global anatomical differences: deformation-based morphometry. Hum. Brain Mapp. **6**(5–6), 348–357 (1998)
6. Batchelor, P., Moakher, M., Atkinson, D., Calamante, F., Connelly, A.: A rigorous framework for diffusion tensor calculus. Magn. Reson. Med. **53**(1), 221–225 (2005)
7. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) **57**, 289–300 (1995)
8. Bennett, C.M., Baird, A.A., Miller, M.B., Wolford, G.L.: Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction. J. Serendipitous Unexpected Results **1**, 1–5 (2010)
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
10. Bloy, L., Ingalhalikar, M., Eavani, H., Roberts, T.P.L., Schultz, R.T., Verma, R.: HARDI based pattern classifiers for the identification of white matter pathologies. In: Fichtinger, G., Martel, A., Peters, T. (eds.) Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 6892, pp. 234–241. Springer, Berlin (2011)
11. Chumbley, J.R., Friston, K.J.: False discovery rate revisited: FDR and topological inference using gaussian random fields. NeuroImage **44**(1), 62–70 (2009)
12. Colby, J.B., Soderberg, L., Lebel, C., Dinov, I.D., Thompson, P.M., Sowell, E.R.: Along-tract statistics allow for enhanced tractography analysis. Neuroimage **59**(4), 3227–3242 (2012)
13. Corouge, I., Fletcher, P.T., Joshi, S., Gouttard, S., Gerig, G.: Fiber tract-oriented statistics for quantitative diffusion tensor mri analysis. Med. Image Anal. **10**(5), 786–798 (2006)
14. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage **19**(2 Pt 1), 261–270 (2003)

15. Cui, Y., Wen, W., Lipnicki, D.M., Beg, M.F., Jin, J.S., Luo, S., Zhu, W., Kochan, N.A., Reppermund, S., Zhuang, L., Raamana, P.R., Liu, T., Trollor, J.N., Wang, L., Brodaty, H., Sachdev, P.S.: Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach. NeuroImage **59**, 1209–1217 (2012)
16. Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Colliot, O.: Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Med. Image Anal. **15**(5), 729–737 (2011)
17. Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O.: Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 682–696 (2013)
18. Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., He, Y.: Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (m3). NeuroImage **59**, 2187–2195 (2012)
19. Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M.: Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. NeuroImage **28**(3), 663–668 (2005)
20. Deshpande, G., Li, Z., Santhanam, P., Coles, C.D., Lynch, M.E., Hamann, S., Hu, X.: Recursive cluster elimination based support vector machine for disease state prediction using resting state functional and effective brain connectivity. PLOS One **5**(12), e14277 (2010)
21. Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett J.R., Jr., Barch, D.M., Petersen, S.E., Schlaggar, B.L.: Prediction of individual brain maturity using fMRI. Science **329**, 1358–1361 (2010)
22. Dryden, I.L., Koloydenko, A., Zhou, D.: Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. **3**(3), 1102–1123 (2009)
23. Fillard, P., Pennec, X., Arsigny, V., Ayache, N.: Clinical dt-mri estimation, smoothing, and fiber tracking with log-euclidean metrics. IEEE Trans. Med. Imaging **26**(11), 1472–1482 (2007)
24. Fletcher, P.T., Joshi, S.: Principal geodesic analysis on symmetric spaces: statistics of diffusion tensors. In: Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, pp. 87–98. Springer, Berlin (2004)
25. Ford, J., Farid, H., Makedon, F., Flashman, L.A., McAllister, T.W., Megalooikonomou, V., Saykin, A.J.: Patient classification of fMRI activation maps. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 2879, pp. 58–65. Springer, Berlin (2003)
26. Frackowiak, R.S., Friston, K.J., Frith, C.D., Dolan, R.J., Price, C.J., Zeki, S., Ashburner, J.T., Penny, W.D.: Human brain function. Academic, New York (2004)
27. Franke, K., Luders, E., May, A., Wilke, M., Gaser, C.: Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. NeuroImage **63**, 1305–1312 (2012)
28. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. **2**(4), 189–210 (1994)
29. Friston, K.J., Holmes, A.P., Poline, J., Grasby, P., Williams, S., Frackowiak, R.S., Turner, R.: Analysis of fmri time-series revisited. NeuroImage **2**(1), 45–53 (1995)
30. Genovese, C.R., Lazar, N.A., Nichols, T.: Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage **15**(4), 870–878 (2002)
31. Goodlett, C.B., Fletcher, P.T., Gilmore, J.H., Gerig, G.: Group analysis of dti fiber tract statistics with application to neurodevelopment. NeuroImage **45**(1), S133–S142 (2009)
32. Grosenick, L., Greer, S., Knutson, B.: Interpretable classifiers for fMRI improve prediction of purchases. IEEE Trans. Neural Syst. Rehabil. Eng. **16**(6), 539–548 (2008)
33. Haller, S., Nguyen, D., Rodriguez, C., Emch, J., Gold, G., Bartsch, A., Lovblad, K.O., Giannakopoulos, P.: Individual prediction of cognitive decline in mild cognitive impairment

using support vector machine-based analysis of diffusion tensor imaging data. J Alzheimers Dis. **22**(1), 315–327 (2010)
34. Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage **87**, 96–110 (2014)
35. Honorio, J., Tomasi, D., Goldstein, R.Z., Leung, H.C., Samaras, D.: Can a single brain region predict a disorder? IEEE Trans. Med. Imaging **31**(11), 2062–2072 (2012)
36. 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), 546–554 (2005)
37. 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**, 153–163 (2013)
38. Kindlmann, G., Estepar, R.S.J., Niethammer, M., Haker, S., Westin, C.F.: Geodesic-loxodromes for diffusion tensor interpolation and difference measurement. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2007, pp. 1–9. Springer, Heidelberg (2007)
39. Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.: Automatic classification of MR scans in alzheimer's disease. Brain **131**(3), 681–689 (2008)
40. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fMRI data. NeuroImage **26**(2), 317–329 (2005)
41. Lenglet, C., Rousson, M., Deriche, R., Faugeras, O.: Statistics on the manifold of multivariate normal distributions: theory and application to diffusion tensor mri processing. J. Math. Imaging Vision **25**(3), 423–444 (2006)
42. Lipton, M.L., Gellella, E., Lo, C., Gold, T., Ardekani, B.A., Shifteh, K., Bello, J.A., Branch, C.A.: Multifocal white matter ultrastructural abnormalities in mild traumatic brain injury with cognitive disability: a voxel-wise analysis of diffusion tensor imaging. J Neurotrauma **25**(11), 1335–1342 (2008)
43. Lipton, M.L., Kim, N., Park, Y.K., Hulkower, M.B., Gardin, T.M., Shifteh, K., Kim, M., Zimmerman, M.E., Lipton, R.B., Branch, C.A.: Robust detection of traumatic axonal injury in individual mild traumatic brain injury patients: intersubject variation, change over time and bidirectional changes in anisotropy. Brain Imaging Behav. **6**(2), 329–342 (2012)
44. Martínez-Ramón, M., Kltchinskii, V., Heileman, G.L., Posse, S.: fMRI pattern classification using neuroanatomically constrained boosting. NeuroImage **31**(3), 1129–1141 (2006)
45. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X.: Learning to decode cognitive states from brain images. Mach. Learn. **57**, 145–175 (2004)
46. Nagy, Z., Alexander, D.C., Thomas, D.L., Weiskopf, N., Sereno, M.I.: Using high angular resolution diffusion imaging data to discriminate cortical regions. PLOS One **8**(5), e63842 (2013)
47. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. **15**(1), 1–25 (2002)
48. O'Donnell, L., Westin, C., Golby, A.: Tract-based morphometry for white matter group analysis. NeuroImage **45**(3), 832–844 (2009)
49. O'Donnell, L.J., Golby, A.J., Westin, C.F.: Fiber clustering versus the parcellation-based connectome. NeuroImage **80**, 283–289 (2013)
50. O'Dwyer, L., Lamberton, F., Matura, S., Scheibe, M., Miller, J., Rujescu, D., Prvulovis, D., Hampel, H.: White matter differences between healthy young ApoE4 carriers and non-carriers identified with tractography and support vector machines. PLOS One **7**(4), e36024 (2012)
51. Pasternak, O., Sochen, N., Basser, P.J.: The effect of metric selection on the analysis of diffusion tensor mri data. NeuroImage **49**(3), 2190–2204 (2010)
52. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fmri: a tutorial overview. NeuroImage **45**(1 Suppl.), S199–S209 (2009)

53. Rasmussen, P.M., Madsen, K.H., Lund, T.E., Hansen, L.K.: Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage **55**, 1120–1131 (2011)
54. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. **53**(1–2), 23–69 (2003)
55. Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., Schmidt-Wilcke, T., Brüne, M.: Shared and non-shared neural networks of cognitive and affective theory-of-mind: a neuroimaging study using cartoon picture stories. Hum. Brain Mapp. (2014). Early View. doi: 10.1002/hbm.22610
56. 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 Clin. (2014). DOI 10.1016/j.nicl.2014.07.001
57. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Massachusetts (2002)
58. Schultz, T., Fuster, A., Ghosh, A., Deriche, R., Florack, L., Lim, L.H.: Higher-order tensors in diffusion imaging. In: Westin, C.F., Vilanova, A., Burgeth, B. (eds.) Visualization and Processing of Tensors and Higher Order Descriptors for Multi-valued Data, pp. 129–161. Springer, Berlin (2014)
59. Schwartzman, A.: Random ellipsoids and false discovery rates: statistics for diffusion tensor imaging data. Ph.D. thesis, Stanford University (2006)
60. Schwartzman, A., Dougherty, R.F., Taylor, J.E.: Cross-subject comparison of principal diffusion direction maps. Magn. Reson. Med. **53**(6), 1423–1431 (2005)
61. Schwartzman, A., Dougherty, R.F., Taylor, J.E.: False discovery rate analysis of brain diffusion direction maps. Ann. Appl. Stat. **2**(1), 153–175 (2008)
62. Shenton, M.E., Kikinis, R., Jolesz, F.A., Pollak, S.D., LeMay, M., Wible, C.G., Hokama, H., Martin, J., Metcalf, D., Coleman, M., et al.: Abnormalities of the left temporal lobe and thought disorder in schizophrenia: a quantitative magnetic resonance imaging study. N. Engl. J. Med. **327**(9), 604–612 (1992)
63. Smith, S., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T., Mackay, C., Watkins, K., Ciccarelli, O., Cader, M., Matthews, P., et al.: Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. NeuroImage **31**(4), 1487–1505 (2006)
64. Student: The probable error of a mean. Biometrika **6**(1), 1–25 (1908)
65. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K.: The WU-Minn human connectome project: an overview. NeuroImage **80**, 62–79 (2013)
66. Vapnik, V.: The Nature of Statistical Learning Theory, 2nd edn. Information Science and Statistics. Springer, New York (1999)
67. Viswanath, V., Fletcher, E., Singh, B., Smith, N., Paul, D., Peng, J., Chen, J., Carmichael, O.: Impact of dti smoothing on the study of brain aging. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 94–97. IEEE, New York (2012). doi: 10.1109/EMBC.2012.6345879
68. Vul, E., Harris, C., Winkielman, P., Pashler, H.: Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. **4**(3), 274–290 (2009)
69. Wakana, S., Jiang, H., Nagae-Poetscher, L.M., Van Zijl, P.C., Mori, S.: Fiber tract–based atlas of human white matter anatomy 1. Radiology **230**(1), 77–87 (2004)
70. Wang, X., Hutchinson, R., Mitchell, T.M.: Training fMRI classifiers to detect cognitive states across multiple human subjects. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) Proceedings of Neural Information Processing Systems, pp. 709–716 (2003)
71. Wee, C.Y., Yap, P.T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Enriched white matter connectivity networks for accurate identification of MCI patients. NeuroImage **54**, 1812–1822 (2011)
72. Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Identification of MCI individuals using structural and functional connectivity networks. NeuroImage **59**, 2045–2056 (2012)
73. Whitcher, B., Wisco, J.J., Hadjikhani, N., Tuch, D.S.: Statistical group comparison of diffusion tensors via multivariate hypothesis testing. Magn. Reson. Med. **57**(6), 1065–1074 (2007)

74. Yushkevich, P.A., Zhang, H., Simon, T.J., Gee, J.C.: Structure-specific statistical mapping of white matter tracts. NeuroImage **41**(2), 448–461 (2008)
75. Zhu, H., Styner, M., Tang, N., Liu, Z., Lin, W., Gilmore, J.H.: Frats: functional regression analysis of dti tract statistics. IEEE Trans. Med. Imaging **29**(4), 1039–1049 (2010)
76. Ziliak, S.T., McCloskey, D.N.: The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor (2008)