Riemannian Variance Filtering: An Independent Filtering Scheme for Statistical Tests on Manifold-valued Data

Ligang Zheng[§] Hyunwoo J. Kim[†] Nagesh Adluru[†] Michael A. Newton[†] Vikas Singh[†]

[§]Guangzhou University [†]University of Wisconsin-Madison
zlg@gzhu.edu.cn hwkim@cs.wisc.edu adluru@wisc.edu {newton,vsingh}@biostat.wisc.edu

Abstract

Performing large scale hypothesis testing on brain imaging data to identify group-wise differences (e.g., between healthy and diseased subjects) typically leads to a large number of tests (one per voxel). Multiple testing adjustment (or correction) is necessary to control false positives, which may lead to lower detection power in detecting true positives. Motivated by the use of so-called "independent filtering" techniques in statistics (for genomics applications), this paper investigates the use of independent filtering for manifold-valued data (e.g., Diffusion Tensor Imaging, Cauchy Deformation Tensors) which are broadly used in neuroimaging studies. Inspired by the concept of variance of a Riemannian Gaussian distribution, a type of non-specific data-dependent Riemannian variance filter is proposed. In practice, the filter will select a subset of the full set of voxels for performing the statistical test, leading to a more appropriate multiple testing correction. Our experiments on synthetic/simulated manifold-valued data show that the detection power is improved when the statistical tests are performed on the voxel locations that "pass" the filter. Given the broadening scope of applications where manifold-valued data are utilized, the scheme can serve as a general feature selection scheme.

1. Introduction

Statistical analysis focused on identifying group level differences (e.g., between healthy controls and individuals with a clinical condition) is an important task in neuroimaging. For example, given a set of "co-registered" (i.e., in a common coordinate system) brain images of 100 individuals who are cognitively healthy and 100 individuals who suffer from dementia, one can perform a statistical test at *each brain voxel* to assess if the distribution of the measurements is different across healthy/diseased groups. When the distributions are different (i.e., the null hypothesis is rejected), we can obtain a map of (uncorrected) p-values showing brain regions likely to be affected by the disease. This voxel-by-voxel analysis is very common across neuroscience, and widely deployed on both structural and functional brain imaging data.

Multiple testing. An important step that was omitted in the foregoing discussion is multiple testing. Since the voxel-specific tests yield a voxel-specific statistic which is "uncorrected", we need to take into account the number of times, say N, that the test was conducted. For instance, if the test is repeated at 1M different voxels, we must perform a correction to control the number of false positives. For example, many null hypotheses will produce small p-values purely by chance. As a result, a large number of false positives (or type-1 error) will occur when repeating a test 1M times — so the p-values cannot simply be compared to a conventional threshold, such as $p < 0.05$ or $p < 0.01$. Individual p-values, e.g., 0.01, no longer correspond to significant findings. Therefore, adjusting the p-value threshold by taking into account the number of times a test was performed is important before we can assess the statistical significance of our findings and control the experiment-wide error. In statistics, there are well-established procedures for such control, such as Bonferroni correction [30].

Multiple testing in imaging studies: Problems. Multiple testing adjustment provides a rigorous control on the extent to which false positives occur in our experimental analysis. This topic is very well studied. Unfortunately, when deployed in the analysis of high-dimensional brain imaging data, we typically encounter a serious practical issue. Observe that such control comes at the cost of reduced power of detecting the true positives because we seek to *avoid* false positives. The situation will be more severe as the number of tests increases. For example, in brain imaging, if we perform millions of tests (the so-called massive multiple comparisons (MCP) problem [25]), after a conservative cor-

rection, only a small region (where the group-wise signal is very strong) will survive. Many other regions that are specific to the disease *may indeed* have small p-values — but not small enough to survive the correction. Therefore, we will be unable to reject the null hypothesis at these voxels. When the sample sizes are small or the effect sizes are poor, the power of detecting the true positives may turn out to be quite low.

A practical feature selection scheme. The above issue is ubiquitous in brain imaging. As a result, a practitioner may often resort to region-of-interest based analysis — essentially, focusing the entire analysis on a few brain regions. This clearly reduces the number of tests that will be performed, thereby reducing how strict the multiple testing correction is. One difficulty is that this may lead to an increase in the number of false negatives by *inadvertently leaving out* some regions where there is a disease specific signal. Separate from this "domain-knowledge" based feature selection, it is not uncommon to find situations where a heuristic scheme based on feature selection is adopted. Essentially, in a pre-processing stage, some statistic is calculated and voxels discarded based on a pre-specified criterion. Then, in the next step, the actual analysis is conducted on a smaller subset of voxels. The pitfall of this procedure is that if the feature selection scheme in the **first** step is *not independent* of the statistical testing in the **second** step, this selection can, in fact, change the null distribution [4]. The interpretation of all subsequent p-value calculations may turn out to be problematic. An elegant solution to this problem was presented in [4] (also see references therein) which shows a mechanism to construct a "filtering" criteria (i.e., feature selection step) that is provably independent of the statistic being calculated in the second step. This allows avoiding two sub-optimal alternatives: **(1)** heuristics that are practically sensible but theoretically flawed and **(2)** choosing a conservative multiple testing correction scheme (with no feature selection) and risking finding no meaningful reportable result from the analysis.

Some related work. This idea of filtering has been studied in other forms in the literature, but is less widely used in machine learning and neuroimaging. For example, several papers have studied how *filtering* can reduce the impact that multiple testing adjustment has on detection power [4, 17, 35]. Many filtering schemes have been proposed for bioinformatics applications which suffer the same massive multiple comparisons issue as in brain imaging. Bourgon et al. proposed a general filtering scheme [4], in which filter pairs are marginally independent under the null hypothesis and dependent under the alternative hypothesis. The filtering scheme can increase the detection power while not losing type-1 error control. In [27], the authors proposed using principal component based-filtering to improve the detection power for Affymetrix gene expression arrays. In [14], the authors present an independent spectral enrichment filter for gene set testing. Independent hypothesis weighting [19] can increase power while controlling the false discovery rate. Broadly speaking, a "filtering scheme" should be thought of a **two stage approach**. In the first stage, a filter is used to filter out some *non-informative* items (or tests). In the second stage, a multiple testing is performed based only on the number of items that pass the filter.

Manifold-valued setting. Various scientific disciplines routinely acquire measurements where data is manifold-valued. For instance, the response variable may be a probability distribution function, a parametric family such as a multinomial, a covariance matrix or samples drawn from a high dimensional unit sphere. Such data arise routinely in machine learning [22, 18, 6, 32], medical imaging [5, 24] and computer vision [33, 28, 7, 38]. Even when performing a basic statistical analysis on such datasets, vector-space operations (such as addition and multiplication) cannot be applied because the manifold is not a vector space. Driven by these motivations, there is a rapidly developing body of theoretical and applied work which generalizes classical tools from multivariate statistics to the Riemannian manifold setting. Various statistical constructs have been successfully extended to Riemannian manifolds: these include regression [39, 21], classification [37], interpolation/convolution/filtering [15], dictionary learning [18, 6], canonical correlation [20] and principal geodesic analysis [13, 31]. While these results expand the operating range of multivariate statistics to the Riemannian manifold setting, simple feature selection schemes (e.g., independent filtering) to facilitate multiple testing have not been studied much.

The **main contribution** of this paper is to investigate the effectiveness of independent filtering for manifold-valued data before group-difference analysis (and multiple testing). We show promising preliminary results via synthetic experiments — such a scheme is simple yet can enable detecting a reasonable group-specific signal in various situations where standard multiple testing correction is too conservative. Our procedure is a two stage hypothesis testing scheme. In the first stage, some voxels are filtered based on a novel Riemannian Variance Filter (RVF). The idea of RVF is inspired by the Riemannian Gaussian distribution. In the second stage, a standard test is conducted on voxels (each with a manifold-valued measurement) passing the first stage filter. Our experimental results show the effectiveness of the filtering scheme. The benefit of using filtering is two fold. First, filtering helps to improve the number of rejections while keeping false positive at a reasonable level. Second, the filtering makes the multiple testing more computationally efficient, especially when using the permutation based testing for manifold-valued data.

1.1. Hypothesis Testing on Manifold-valued Data

A focus of this work will be conducting statistical tests on diffusion tensor imaging (DTI) data, and this will serve as a target application throughout this paper. The literature on statistical analysis of DTI is sizable, so we simply describe a few common schemes of performing hypothesis tests on DTI. Diffusion tensor images have a symmetric positive definite matrix $p_i \succeq 0$ at each voxel location i in the image. A simple approach is to compute a scalar-valued summary measure for each tensor p_i and then use univariate tests, such as a standard t-test, or permutation test to compute the desired statistic. For example, fractional anisotropy or the mean diffusivity of the tensors are typical candidate voxel-wise summary measures [26, 16]. Such procedures reduce the runtime complexity of the hypothesis testing, especially for large brain images. However, compressing the tensor into a scalar value will lose information which may lead to poor testing sensitivity.

Instead of using univariate testing with fractional anisotropy or the mean diffusivity, various alternatives include applying multivariate hypothesis testing procedure to the diffusion tensor. Considering the tensor as a multivariate variable and using Hotelling's T^2 test is an option that has been used successfully [36]. This method does not consider the manifold property of the tensors. An alternative method is to use matrix logarithm transformed forms of the tensors (into Euclidean space), and then use multivariate Hotelling's T^2 test [23].

Both the standard t-test and Hotelling T^2 test assume that the null distribution of variables is a Gaussian or multivariate Gaussian. However, this normality assumption is a potential limitation in analysis of diffusion tensor images. A permutation test is an alternative to the parametric test methods which makes no distributional assumptions. Various papers have used permutation test for manifold-valued data – for example, mean-based and dispersion-based permutation testing for data on nonlinear manifolds [9].

2. Multiple Testing Adjustment

For a single test, the conventional threshold protects us with a probability of $p < 0.05$ from wrongly declaring a voxel as significantly modulated when there is no disease effect. However, when dealing with a very large number of tests simultaneously, the number of wrongly rejected null hypotheses will become very large, entirely by chance. We therefore need a multiple testing correction procedure to adjust our statistical confidence measures based on the number of tests performed. There are a number of well-known correction procedures in the literature. The family-wise error rate (FWER) is the probability of at least one false conclusion in a series of hypothesis tests. In other words, it is the probability of making at least one type 1 error. The most commonly used method which controls FWER at level α is called the Bonferroni's method. The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of α/m, where α is the desired overall significance level and m is the number of hypotheses. When we perform a "filtering step" before the statistical test, the value m accordingly becomes the expectation of the number of hypotheses passing the filters. Separately, the false discovery rate (FDR) is another widely used scheme for controlling the rate of type-1 errors [3].

3. Preliminaries

We first briefly introduce some basic *differential geometry* notations and basic operations on symmetric positive definite (SPD) manifolds that we will use. Note that while the ideas in this paper are generally applicable, to make the presentation concrete, we will utilize the SPD manifold as an example. For more details on some of the notations below, we refer the reader to [10, 8].

Let $\mathcal{M}$ be a *differentiable (smooth) manifold* in arbitrary dimensions. A differentiable manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure. A *Riemannian manifold* $(\mathcal{M}, g)$ is a differentiable manifold $\mathcal{M}$ equipped with a smoothly varying inner product g. The family of inner products on all tangent spaces is known as the *Riemannian metric*, which defines various geometric notions on curved manifolds such as the length of a curve etc. A *geodesic curve* is a locally shortest path, which is analogous to a straight line in $\mathbb{R}^d$. Unlike the Euclidean space, note that there may exist multiple geodesic curves between two points on a curved manifold. So the *geodesic distance* between two points on $\mathcal{M}$ is defined as the length of the *shortest* geodesic curve connecting two points (i.e., SPD matrices). Formally, the distance between p and q is defined as

$$\mathrm{d}(p,q) := \inf_{\gamma} \int_a^b \sqrt{g_\gamma(t)(\dot\gamma(t),\dot\gamma(t))}dt \qquad (1)$$

where $\gamma(a) = p$ and $\gamma(b) = q$. The *tangent space* at $p \in \mathcal{M}$ (denoted by $T_p\mathcal{M}$) is the vector space, which consists of the tangent vectors of *all* possible curves passing through p. The geodesic curve from y_i to y_j is parameterized by a tangent vector in the tangent space anchored at y_i with an exponential map $\mathrm{Exp}(y_i, \cdot) : T_{y_i}\mathcal{M} \to \mathcal{M}$. The inverse of the exponential map is the logarithm map, $\mathrm{Log}(y_i, \cdot) : \mathcal{M} \to T_{y_i}\mathcal{M}$. The exponential map and its inverse logarithm map are denoted by $\mathrm{Exp}(p, x)$ and $\mathrm{Log}(p, v)$ respectively, where $p, x \in \mathcal{M}$ and $v \in T_p\mathcal{M}$. They are usually denoted $\exp_p(x)$ and $\log_p(v)$ in most differential geometry books. These two operations move us back and forth between the manifold and the tangent space. Separate from the above notations, matrix exponential (i.e, $\exp(X) := \sum \frac{1}{k!}X^k$, where $0! = 1$

and $X^0 = I$) and matrix logarithm are denoted by as $\exp(\cdot)$ and $\log(\cdot)$.

Intrinsic mean. Let $d(\cdot,\cdot)$ define the distance between two points. The intrinsic (or Karcher) mean is the minimizer to

$$\bar{y} = \arg\min_{y\in\mathcal{M}} \sum_{i=1}^{N} d(y, y_i)^2, \tag{2}$$

which may be an arithmetic, geometric or harmonic mean depending on $d(\cdot,\cdot)$. A Karcher mean is a local minimum to (2) and a global minimum is referred as a Fréchet mean. On manifolds, the Karcher mean satisfies $\sum_{i=1}^{N} \mathrm{Log}_{\bar{y}} y_i = 0$. This identity implies the first order necessary condition of (2), i.e., $\bar{y}$ is a local minimum with a zero norm gradient. In general, on manifolds, the existence and uniqueness of th.e Karcher mean is not guaranteed unless we assume, for uniqueness, that the data is in a small neighborhood.

3.1. Geometry of SPD Manifolds

Symmetric positive definite matrices are widely used in neuroimaging, e.g., in diffusion imaging. Let SPD(n) be a manifold for symmetric positive definite matrices of size $n \times n$. This forms a quotient space $GL(n)/O(n)$, where $GL(n)$ denotes the general linear group (the group of $(n \times n)$ nonsingular matrices) and $O(n)$ is the orthogonal group (the group of $(n \times n)$ orthogonal matrices). The inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is given by

$$\langle u, v\rangle_p = \mathrm{tr}(p^{-1/2} u p^{-1} v p^{-1/2}) \tag{3}$$

This plays the role of the Fisher-Rao metric in the statistical model of multivariate distributions. Here, $T_p\mathcal{M}$ is a tangent space at p (which is a vector space) is the space of symmetric matrices of dimension $(n+1)n/2$. The geodesic distance is $d(p,q)^2 = \mathrm{tr}(\log^2(p^{-1/2} q p^{-1/2}))$.

The exponential map and logarithm map are given as

$$\begin{aligned} \mathrm{Exp}(p, v) &= p^{1/2} \exp(p^{-1/2} v p^{-1/2}) p^{1/2}, \\ \mathrm{Log}(p, q) &= p^{1/2} \log(p^{-1/2} q p^{-1/2}) p^{1/2}. \end{aligned} \tag{4}$$

and the geodesic distance w.r.t. the affine invariant metric is given by

$$\mathrm{d}(p, q)^2 = \mathrm{tr}(\log^2(p^{-1} q)). \tag{5}$$

4. Method

The independent filtering [4] is proposed for univariate measurements and we extend it for manifold-valued measurements (e.g., diffusion tensors). The independent filtering is a two stage procedure comprised of filtering and hypothesis test over variables which pass the filter. Depending on the final hypothesis test, the *criterion for filtering should be properly chosen so that the null distribution after filtering is still invariant*. Such a filtering is called independent filtering. In this section, we discuss multiple combinations for group difference analysis. For univariate measurements, the independent filtering [4] suggested filtering by variance and hypothesis test by Student's t-test. We adopt this idea and develop a similar procedure for manifold-measurements. At the first stage, we filter voxels by dispersion (corresponding to variance) of manifold-valued data. And then, hypothesis tests are performed by nonparametric tests (e.g., Mean-based permutation test, Cramér's test). The two stages will be explained for both fractional anisotropy and diffusion tensors.

4.1. Filtering

In general, filtering means that some meta test will be set up according to a pre-specified criterion, and (ideally) will reduce the number of tests in the follow-up step. For univariate data, the overall mean and overall variance - computed across all arrays, are generally used in genomics research [4]. It is easy to rank the overall mean or overall variance (scalar value) and take a threshold to perform "filtering". For multivariate data, the situation may become a bit more complex. Various strategies have been proposed as briefly described above, for instance, the principal components of the covariance matrix as a filter [27]. For tensors (in diffusion imaging), we usually need to consider its intrinsic structure, which makes the problem more difficult. As described below, we find that an analogous criterion for manifold-valued data can be obtained from a generalization of Gaussian distribution on manifolds.

Let $\boldsymbol{\mu} \in \mathcal{M}$ and $\sigma \in \mathbb{R}_+$. One generalization of the Gaussian distribution on Riemannian manifolds is given by

$$\begin{aligned} f(X; \boldsymbol{\mu}, \sigma) &= \frac{1}{\zeta(\sigma)} \exp\left(-\frac{\mathrm{d}(X, \boldsymbol{\mu})^2}{2\sigma^2}\right) \\ &\text{where} \\ \zeta(\sigma) &= \int_{\mathcal{M}} \exp\left(-\frac{d(X, \boldsymbol{\mu})^2}{2\sigma^2}\right) dX. \end{aligned} \tag{6}$$

$d(\cdot,\cdot)$ denotes the geodesic distance between two manifold-valued data points. On SPD(n), we use the affine-invariant Riemannian metric for $d(\cdot,\cdot)$. $\boldsymbol{\mu} \in \mathcal{M}$ and $\sigma \in \mathbb{R}_+$ corresponds to the mean and variance. We call σ dispersion, which is used to perform filtering on manifold-valued variables. Multiple generalizations of Gaussian distributions are discussed in [1, 12]. $\zeta(\sigma)$ is the normalization factor to make the integration of the PDF in the space of SPD(n) work. It is known that $\zeta(\sigma)$ is not functionally dependent on $\boldsymbol{\mu}$ in Riemannian symmetric spaces [12]. However, it is difficult to calculate the normalization factor in practice [29]. This results in a non-trivial maximum likelihood estimation (MLE) of dispersion (σ).

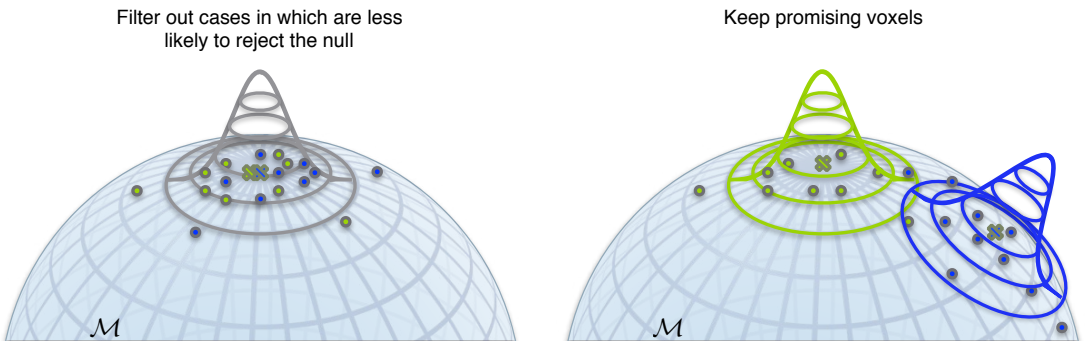


Figure 1. Our framework is a two-stage method. The figure shows the first stage. At the first stage, we filter out voxels which are less likely to reject null hypothesis in the left figure whereas the right figure shows that the filtering lets voxels pass where they may have significant group diffenrence. Note that the filtering step does not use any group informtion. Further the filtering criterion should be independent from conditional test statistics.

The empirical mean will be denoted by $\overline{X}$ and its MLE can be obtained by the least squares estimation w.r.t. the geodesic distance on SPD(n) by minimizing the energy function ($\mathcal{E}_n$) given as

$$\mathcal{E}_n(\overline{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathrm{d}(\overline{X}, X_i)^2. \tag{7}$$

The MLE of the dispersion parameter σ can also be estimated by maximizing the log-likelihood function. The first order necessary condition w.r.t. σ given by

$$\sigma^3 \frac{d}{d\sigma} \log \zeta(\sigma) = \mathcal{E}_n(\overline{X}). \tag{8}$$

The solution to (8) can be written as

$$\widehat{\sigma} = \phi(\mathcal{E}_n(\overline{X})) = \phi(\frac{1}{n}\sum_{i=1}^{n} d^2(\overline{X}, X_i)), \tag{9}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function.

The estimation of $\widehat{\sigma}$ involves ϕ. Note that we filter voxels by rankings based on a criterion. So, as long as it keeps the order of variables unchanged, we can use a much simpler ϕ for our procedure. In this work, we replace $\phi(\cdot)$ with the identity function and use it as our criterion, which is $\mathcal{E}_n(\overline{X})$ and the voxels which have a relatively larger $\mathcal{E}_n(\overline{X})$ pass the filter, e.g., top 10% of voxels. We call the filtering by $\mathcal{E}_n(\overline{X})$ the **Riemannian Variance Filter (RVF)**. This filter is used for diffusion tensors in the experiment section.

4.2. Hypothesis Tests for Group Difference Analysis

Hypothesis test for group difference can be performed using various test statistics and null distributions. In [4], Student's t-test is used. We discuss nonparametric hypothesis tests: mean-based permutation test and Cramér's test. We would like to note that we only make standard assumptions (such as pixel independence) and do not make any additional assumptions not common in neuroimage analysis [2]. Furthermore, each subject is assumed to be independent, so the use of permutation testing is sensible.

Mean-based permutation test: The mean-based permutation test uses the distance between means of two groups, i.e., $\Delta = \mathrm{d}(\overline{X}, \overline{Y})$. Using permutation tests, we simulate the null distribution of Δ. The iterative procedure for computing the Fréchet mean of diffusion tensors is computationally expensive for a large number of permutation tests. For faster estimation, log-Euclidean metric [1] can be used as

$$\overline{X} \approx \exp(\frac{1}{n}\sum_{i=1}^{n} \log X_i). \tag{10}$$

There are also some other choices to compute the mean through the spectral decomposition of the tensors [8], which may provide better decoupling between orientation and anisotropy but these strategies were not utilized here.

Cramér's test: We use a two sample Cramér's test as a unified hypothesis test for group difference analysis. It requires only pairwise distances and group label information. The distance can be either Euclidean distance or geodesic distance. So, this test is directly applicable for univariate, multivariate and even manifold-valued variables. Also, the null distribution is simulated by the sampling distribution of test statistic when the null hypothesis is true. The test statistic is given by

$$\delta_{n_1,n_2} = \frac{n_1 n_2}{n_1+n_2}\left(\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathrm{d}(X_i, Y_j) - \frac{1}{2n_1^2}\sum_{i=1}^{n_1}\sum_{i'=1}^{n_1} \mathrm{d}(X_i, X_{i'}) - \frac{1}{2n_2^2}\sum_{j=1}^{n_2}\sum_{j'=1}^{n_2} \mathrm{d}(Y_j, Y_{j'})\right) \tag{11}$$

where $\mathrm{d}(\cdot,\cdot)$ is a distance metric for samples.

5. Experimental Evaluations

In our experiments, we use two different sets of synthetic experiments using simulated normal (healthy) and the abnormal (diseased) brain image population, where each DTI image was composed of 3×3 DTI tensors at each voxel. The two groups are generated from a set of reference tensors and a set of transformed tensors. Given a reference tensor, the following three different geometric transformation methods are studied in our experiments: (1) change the eigenvalues (2) change the orientations (3) change both eigenvalues and orientations. We assume that these changes are associated with the clinical phenomena under study. The size of the transformed patches (i.e., where the disease specific signal is assumed to be localized) is 10×10, 15×15 and 15×15 respectively.

Each group (diseased, healthy) is assumed to be comprised of 15 subjects. To keep the experiments simple (since the computation time for permutation testing can be significant), we assume that the images are of size 50×50, which means there are 2500 voxels for each subject. This corresponds to a total of 2500 hypothesis tests where we perform the statistical test at each voxel seeking to identify if there are group-wise differences (induced as a result of the class membership). A normal and a diseased subject are shown using a glyphviewer in DTITK in Fig. 2.

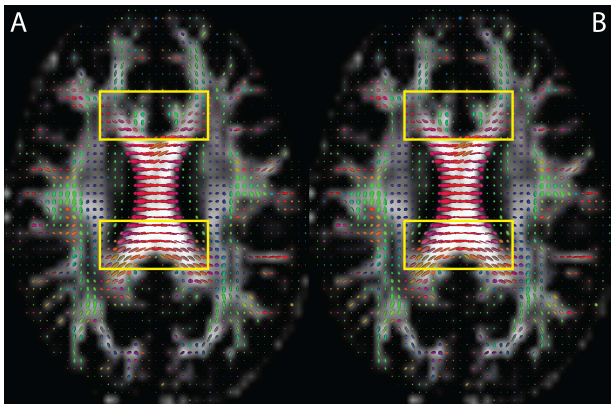


Figure 2. Each image is a representative of the "mean" healthy control and diseased subject. The disease effect has been simulated to be restricted to the region indicated in a yellow box. In regions outside the yellow box, there are no group-wise differences.

In our experiments, we use the true positives (TP) and false positives (FP) to evaluate the effectiveness of our algorithm. Our motivation is to investigate the effectiveness of adding a filtering step in hypothesis testing for manifold-valued data. We specifically evaluate whether there is a performance gain when compared to the no filtering scheme. What is also important to evaluate is whether this scheme improves TP while controlling FP (type-1 error).

For filtering, we consider the overall RVF of a set of tensors at a voxel, which means that we do not consider the class label in the filtering step (else, the evaluations will be meaningless). The locations where RVF are in the lowest portion $\theta \in [0, 1]$ are filtered out. In stage 2, both a standard two sample t-test and a permutation test are used for hypothesis testing. The FDR is used for multiple testing correction, in the standard way. The significance level is set as a standard value $\alpha = 0.05$.

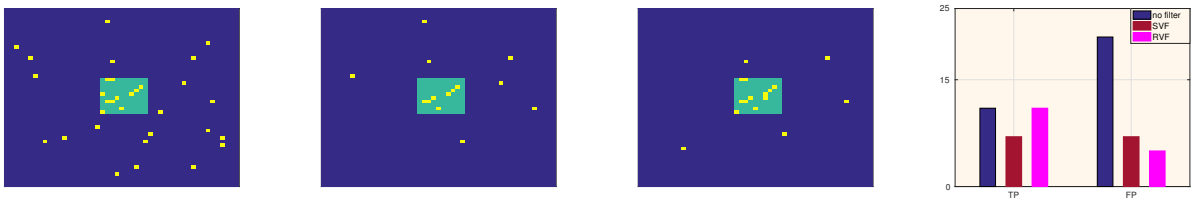
Results for standard t-test: The results of standard two sample t-test for the tensors generated by methods 3 and method 1 are given in Fig. 3 and Fig. 4 respectively. The t-test shows the results when we use the standard two sample t-test without filtering. The scalar variance filter ("SVF") shows the result from using a scalar variance filter ($\theta = 0.7$) before two sample t-test while the "RVF" shows the results from using Riemannian variance filter ($\theta = 0.7$) in stage 1. We can see from the result that filtering increases the TP while controlling the type-1 error (false positives). From the TP and FP sub-figures, we can see that using the RVF filter, the false positives (type-1 error) are controlled while the detection of the true positives are not reduced. In fact, filtering increases the true positive in many situations. Both SVF and RVF based t-tests fail in detecting any true positives when changing the orientation with 2° (disease effect) while keeping the eigenvalues unchanged.

Results for permutation testing approach: Fig. 6 shows the results of permutation testing for the three tensor generating methods. In the filtering stage, our RVF filtering scheme is used. In the multiple testing stage, we used four methods as shown in Table. 1 including FA, MFA, Cramér and LeMean. The "no filter" corresponds to the 0 filtering for each method. When changing the orientation by 2° (disease effect), both FA and MFA methods fail in detecting any true positives, so we only show the results of Cramér and LeMean permutations in Fig. 5(c).

Table 1. notations in permutation testing

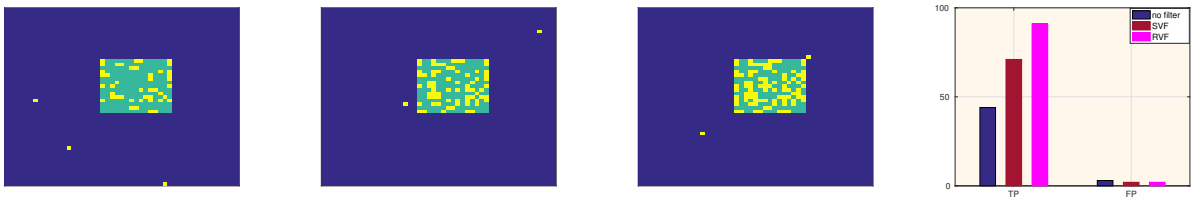
FA	$\triangleq$	SVF + FA
MFA	$\triangleq$	RVF + FA
Cramér	$\triangleq$	RVF + Cramér test
LeMean	$\triangleq$	RVF + Log-Euclidean Mean

Results for synthetic brain imaged data: We used both standard t-test and permutations testing methods including MFA, Cramér and LeMean. For standard t-test, we used both scalar variance and Riemannian variance based filtering methods, which are shown in Fig. 6(a) and Fig. 6(b). The results of permutation tests including MFA, Cramér and LeMean are respectively given in Fig. 6(c), Fig. 6(d) and Fig. 6(e). We can see from the results that the standard t-test using the scalar variance filter only detects a very small number of true positives. The other methods including permutation methods and t-test achieve a comparable performance or more true positives while largely reducing



(a) t-test (b) SVF + t-test (c) RVF + t-test (d) TP and TP

Figure 3. The results of standard t-test for scalar variance and Riemannian variance filter on simulated data sets generated by method 3. (a) only using the t-test (b) using scalar variance filter and t-test (c) using Riemannian variance filter and t-test (d) the TP and FP performance of all the three methods.



(a) t-test (b) SVF + t-test (c) RVF + t-test (d) TP and FP

Figure 4. Similar to Fig. 3, but on simulated data sets generated by method 1. (a)only using the t-test (b) using scalar variance filter and t-test (c) using Riemannian variance filter and t-test (d) the TP and FP performance of all the three methods.

the false positives.

6. Additional Analysis of Experiments

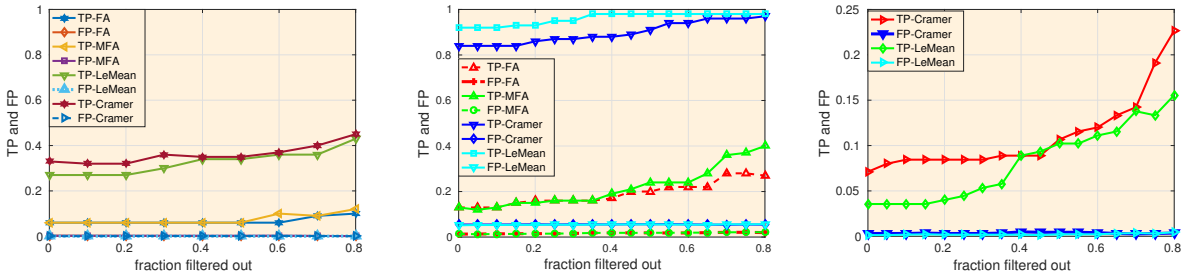
SVF vs. RVF: Fractional anisotropy is a popular scalar summary measure used in diffusion imaging. When using this measure for multiple testing, it is easy to perform filtering before the second stage testing. According to [4], the overall variance and standard t-test pair will be an ideal choice. However, fractional anisotropy does not directly use the orientation information of the tensors, which may lead to poor testing power even using an independent filtering strategy. The experiment results in Fig. 3(b) and Fig .4(b) also support this statement empirically.

However, the testing power of fractional anisotropy will be improved when using the Riemannian variance filter in the first stage, which may due to the fact that the filter considers the nonlinear nature of the tensors. For the standard t-test, RVF improves TP while controlling FP when compared with SVF and the zero filtering setup. This statement is supported by sub-figures from Fig. 3 and Fig. 4. For permutation test, we can get a similar conclusion from Fig. 6. In some situations, the scalar variance filter does not reduce the testing power, even slightly improves the testing power, see Fig. 5(b). But it has a lower increase in rejections.

Filtering increase discoveries: The motivation to introduce filtering in the first stage is to reduce the impact of multiple testing adjustment on detection power. Filtering can reduce the number of hypothesis tests, which has an effect on how conservative the correction is at the subsequent stage. Compared to scalar variance filtering method, Riemannian variance filter increases the "rejections" for both standard t-test and permutation test improving our ability to see disease specific effects. From Fig. 6, we can see that all four second stage testing methods improve the number of rejections compared with 0 filtering at the 0 point in the x-axis. Even using the simple scalar variance (fractional anisotropy), there are some small improvements in the number of rejections.

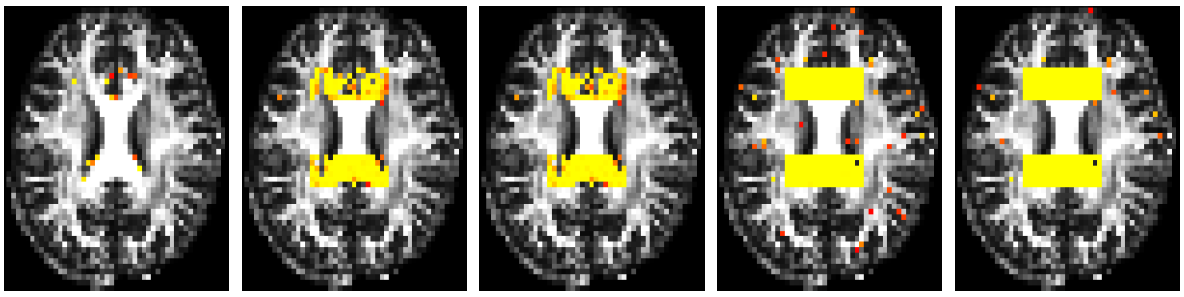
Type I error: In the previous section, we show that using nonspecific variance filter RVF will increase the number of true positives (rejections). However, this increase will be meaningless if the false positive control is compromised. Therefore, the ideal situation is when filtering improves the detection power but also when the type-1 error (false positive) is controlled. From both simulated data and brain image data, we can see from the Fig. 3(d), Fig. 4(d) and Fig. 6 that the false positives (type-1 error) are appropriately controlled.

RVF vs. random filtering: A random filter, which arbitrarily selects and removes a proportion of 'locations', was also considered in our experiments. The random filtering can reduce the number of hypothesis, however, this filter will also reduce a lot of voxels that are specific to te disease effect. This issue will become more severe as the number of voxels filtered out increases. In fact, the normal and diseased voxels have a theoretically equal chance of being filtered out. This may or may not help the second stage to



(a) Changing both eigenvalues and orientations(method 3) (b) Changing eigenvalues (method 1) (c) Changing orientation (method 2)

Figure 5. The x-axis cooresponds to fraction of locations filtered out, the y-axis is the TP and FP rate. TP-FA, TP-MFA, TP-Cramér and TP-LeMean are TP rate of using FA, MFA, Cramér and LeMean as permutation testing. FP-FA, FP-MFA, FP-Cramér and FP-LeMean are FP rate of using FA, MFA, Cramér and LeMean as permutation testing. 0 in x-axis corresponds to the "no filter" in the stage 1. FA uses the scalar variance filter in the first stage while all other schemes use Riemannian variance filter.



(a) SVF + t-test (b) RVF + t-test (c) RVF + MFA (d) RFV + Cramér (e) RVF + LeMean

Figure 6. Brain image experiments on filter followed by test statistics.

improve the testing power. Our experiments also show that random filtering does not help to improve the testing power consistently. In fact, randomly filtering usually performs poorly. Therefore, we see that only a carefully chosen filter can improve the detection power.

7. Discussion

Multiple testing adjustment provides control on false positives, but such control comes at the the cost of reduced power to detect true positives. In particular, the situation become more severe when the number of hypothesis increases. For univariate and multivariate data, various papers have proposed effective filtering schemes to reduce the impact of multiple testing adjustments on detection power. But for manifold-valued data, there is no such filtering scheme currently available. As the manifold-valued data have a nonlinear intrinsic structure, it is relatively difficult to rank and filter without changing the true null distribution.

In this paper, we propose to use RVF as a filter based on the Riemannian Gaussian distribution. The filter is a mechanism to measure the mean-based dispersion of the data samples, which is similar to the notion of variance or covariance matrix used in univariate and multivariate filters. The proposed scheme is tested on a set of synthetic tensors. According to our preliminary experimental results, the scheme can improve the rejections while controlling the type-1 error.

Finally, we want to note a few caveats. First, as with most statistical analysis methods on neuroimaging data, our procedure is only designed to infer *if* the groups are different. Confounding variables are not considered in our procedure. Second, our use of permutation testing is justified with two-group comparisons when the subjects are exchangeable on the null hypothesis; This is a typical assumption in statistical testing [11]. The potentially more complex dependencies between voxels typically do not invalidate voxel-specific p-values computed from exchangeable subjects [34]. Going forward, advances in filtering methodology may further improve the power to detect subtle distributional changes [19].

Acknowledgments: This work was partially supported by China Scholarship Council, NSFC 61300205 and Guangzhou BEY 1201630355 while LZ was a visiting scholar at UW-Madison. The research was also supported in part by BRAIN Initiative R01EB022883, Waisman IDDRC U54HD090256, UW CPCP AI117924, and NSF CAREER award 1252725 (VS).

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006. 4, 5

[2] J. Ashburner and K. J. Friston. Voxel-based morphometrythe methods. *Neuroimage*, 11(6):805–821, 2000. 5

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995. 3

[4] R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551, 2010. 2, 4, 5, 7

[5] H. E. Cetingul and R. Vidal. Sparse Riemannian manifold clustering for HARDI segmentation. In *ISBI*, pages 1750–1753, 2011. 2

[6] A. Cherian and S. Sra. Generalized dictionary learning for SPD matrices with application to nearest neighbor retrieval. In *ECML*, pages 318–332, 2011. 2

[7] A. Cherian and S. Sra. Riemannian sparse coding for positive definite matrices. In *ECCV*, pages 299–314, 2014. 2

[8] A. Collard, S. Bonnabel, C. Phillips, and R. Sepulchre. An anisotropy preserving metric for dti processing. *arXiv preprint arXiv:1210.2826*, 2012. 3, 5

[9] A. Collard, C. Phillips, and R. Sepulchre. Statistical tests for group comparison of manifold-valued data. In *52nd IEEE Conference on Decision and Control*, pages 1144–1149, Dec 2013. 3

[10] M. P. Do Carmo. *Riemannian geometry*. Springer, 1992. 3

[11] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003. 8

[12] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International journal of computer vision*, 105(2):171–185, 2013. 4

[13] P. T. Fletcher, C. Lu, et al. Principal geodesic analysis for the study of nonlinear statistics of shape. *TMI*, 23(8):995–1005, 2004. 2

[14] H. R. Frost, Z. Li, F. W. Asselbergs, and J. H. Moore. An independent filter for gene set testing based on spectral enrichment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(5):1076–1086, Sept 2015. 2

[15] A. Goh, C. Lenglet, P. M. Thompson, and R. Vidal. A nonparametric Riemannian framework for processing HARDI. In *CVPR*, pages 2496–2503, 2009. 2

[16] C. B. Goodlett, P. T. Fletcher, J. H. Gilmore, and G. Gerig. Group analysis of dti fiber tract statistics with application to neurodevelopment. *Neuroimage*, 45(1):S133–S142, 2009. 3

[17] A. J. Hackstadt and A. M. Hess. Filtering for increased power for microarray data analysis. *BMC bioinformatics*, 10(1):11, 2009. 2

[18] J. Ho, Y. Xie, and B. C. Vemuri. On a nonlinear generalization of sparse coding and dictionary learning. In *ICML*, pages 1480–1488, 2013. 2

[19] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580, 2016. 2, 8

[20] H. Kim, N. Adluru, B. B. Bendlin, S. C. Johnson, B. C. Vemuri, and V. Singh. Canonical correlation analysis on riemannian manifolds and its applications. In *Proceedings of European Conference on Computer Vision (ECCV)*, October 2014. 2

[21] H. Kim, N. Adluru, M. D. Collins, M. K. Chung, B. B. Bendlin, S. C. Johnson, R. J. Davidson, and V. Singh. Multivariate general linear models (mglm) on riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[22] G. Lebanon. *Riemannian geometry and statistical machine learning*. PhD thesis, 2005. 2

[23] A. D. Lee, N. Lepore, F. Lepore, F. Alary, P. Voss, Y. Chou, C. Brun, M. Barysheva, A. W. Toga, and P. M. Thompson. Brain differences visualized in the blind using tensor manifold statistics and diffusion tensor imaging. In *Frontiers in the Convergence of Bioscience and Information Technologies*, pages 470–476. IEEE, 2007. 3

[24] C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *JMIV*, 25(3):423–444, 2006. 2

[25] M. A. Lindquist and A. Mejia. Zen and the art of multiple comparisons. *Psychosomatic medicine*, 77(2):114, 2015. 1

[26] Z. Liu, H. Zhu, B. L. Marks, L. M. Katz, C. B. Goodlett, G. Gerig, and M. Styner. Voxel-wise group analysis of dti. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 807–810. IEEE, 2009. 3

[27] J. Lu, R. T. Kerns, S. D. Peddada, and P. R. Bushel. Principal component analysis-based filtering improves detection for affymetrix gene expression arrays. *Nucleic acids research*, 39(13):e86–e86, 2011. 2, 4

[28] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on Lie algebra. In *CVPR*, pages 728–735, 2006. 2

[29] S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, April 2017. 4

[30] P. Sedgwick. Multiple hypothesis testing and bonferroni's correction. *BMJ: British Medical Journal (Online)*, 349, 2014. 1

[31] S. Sommer, F. Lauze, and M. Nielsen. Optimization over geodesics for exact principal geodesic analysis. *Advances in Computational Mathematics*, pages 283–313, 2013. 2

[32] S. Sra and R. Hosseini. Geometric optimisation on positive definite matrices with application to elliptically contoured distributions. In *NIPS*, pages 2562–2570, 2013. 2

[33] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *CVPR*, pages 1–8, 2007. 2

[34] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004. 8

[35] D. Tritchler, E. Parkhomenko, and J. Beyene. Filtering genes for cluster and network analysis. *BMC bioinformatics*, 10(1):193, 2009. 2

[36] B. Whitcher, J. J. Wisco, N. Hadjikhani, and D. S. Tuch. Statistical group comparison of diffusion tensors via multivariate hypothesis testing. *Magnetic Resonance in Medicine*, 57(6):1065–1074, 2007. 3

[37] Y. Xie, B. C. Vemuri, et al. Statistical analysis of tensor fields. In *MICCA*, pages 682–689. 2010. 2

[38] L. Zheng, G. Qiu, H. Fu, and J. Huang. Salient covariance for near-duplicate image and video detection. In *Proceedings of International Conference on Image Processing*, pages 2585–2588, 2011. 2

[39] H. Zhu, Y. Chen, J. G. Ibrahim, et al. Intrinsic regression models for positive-definite matrices with applications to DTI. *JASA*, 104(487), 2009. 2