Published in final edited form as:

IEEE Trans Med Imaging. 2016 October ; 35(10): 2243–2257. doi:10.1109/TMI.2016.2554554.

Transformation Invariant Control of Voxel-Wise False Discovery Rate

Junning Li, Yonggang Shi, and Arthur W. Toga

Laboratory of Neuro Imaging (LONI), Mark Neuroimaging and Informatics Institute, Keck School of Medicine, the University of Southern California. 2001 North Soto Street - Room 102, Los Angeles, CA 90032, United States

Arthur W. Toga: toga@loni.usc.edu

Abstract

Multiple testing for statistical maps remains a critical and challenging problem in brain mapping. Since the false discovery rate (FDR) criterion was introduced to the neuroimaging community a decade ago, many variations have been proposed, mainly to enhance detection power. However, a fundamental geometrical property known as transformation invariance has not been adequately addressed, especially for the voxel-wise FDR. Correction of multiple testing applied after spatial transformation is not necessarily equivalent to transformation applied after correction in the original space. Without the invariance property, assigning different testing spaces will yield different results. We find that normalized residuals of linear models with Gaussian noises are uniformly distributed on a unit high-dimensional sphere, independent of t-statistics and F-statistics. By defining volumetric measure in the hyper-spherical space mapped by normalized residuals, instead of the image's Euclidean space, we can achieve invariant control of the FDR under diffeomorphic transformation. This hyper-spherical measure also reflects intrinsic "volume of randomness" in signals. Experiments with synthetic, semi-synthetic and real images demonstrate that our method significantly reduces FDR inconsistency introduced by the choice of testing spaces.

I. Introduction

Mapping functional activation or genetic influence to brain anatomy involves hypothesis tests carried out at numerous voxel locations. Due to its stochastic nature, it is impossible to completely eliminate statistical errors, but at best to keep an appropriate balance between false positives and false negatives. If the threshold is too conservative, then true hints may be overlooked; if it is too liberal, then spurious claims may flood in, and accompanied with publication bias [1] towards positive results they may later mislead the research community to spend extra resources on hypotheses that did not hold in the first place.

A good threshold should be selected according to an error rate criterion and the amount of simultaneous tests. Though an arbitrary uncorrected threshold may yield more "promising" results, its misgauged uncertainty may turn into pitfalls. A functional study in 2009 [2] illustrated that "significant" activation can be "detected" with an uncorrected threshold even in a dead salmon's brain. Two widely used criteria are (1) the family wise error rate (FWER) [3] which is the probability that at least one error occurs in the results, and (2) the false

discovery rate (FDR) [4, 5] which is the expected portion of false discoveries among reported discoveries. Rather than prohibiting even a single error, the FDR provides a direct trade-off between detection power and the error ratio. Since its debut in 1995 [4] the FDR has been actively adopted.

Though multiple testing affects a broad range of modern science, from hunting genetic causes of complex diseases to correlating social factors with economic decisions, its application in statistical mapping is uniquely challenging because it involves both randomness and geometry. First, it essentially involves an infinite and uncountable number of tests residing on a continuous space, unlike a finite number of tests in genetic scans. Even though in practice it deals with a finite number of voxels, the mathematical nature of the problem domain is a continuous manifold composed of an infinite number of points. Second, its relationship with geometric operations and topological properties, such as spatial transformation and connected components, should also be considered. Moreover, spatial dependence among signals makes the problem highly complicated.

We are particularly interested in transformation invariance in voxel-wise FDR control. It is a popular practice to warp subjects' images to a template space for statistical mapping, for example, the Talairach atlas [6], the LPBA40 atlas [7], the ICBM152 atlas [8] or customized atlases built with various software, such as ANTS, AIR, ART, Diffeomorphic Demons, or FNIRT. As statistical mapping may take place in so many possible image spaces, it naturally raises a question, as illustrated in Figure 1: if we threshold a statistical map in two diffeomorphic spaces with the same FDR level, will their results be consistent under spatial transformation? If our methods are transformation variant, then our results are subject to the bias introduced by our arbitrary choice of atlases or registration software and parameters, or researchers may happen to select a particular image space favoring or against a certain hypothesis, and produce different results even with the same statistical map. Therefore, transformation invariance stands as a required property for objective thresholding. Since Genovese, Lazar, and Nichols [9] introduced the FDR to the neuroimaging community a decade ago, various adaptations and improvements have been made. However, transformation invariance, this highly desirable feature, has not been adequately addressed. To the best of our knowledge, this is the first paper dedicated to this topic.

In Genovese et al.'s work (2002) [9], the FDR is defined as the volumetric ratio of false positive regions against positive regions, and controlled by Benjamini and Hochberg's step-up procedure [4, 10] applied to voxel p -values. Following this voxel-wise definition of the FDR, many other methods have been developed for statistical mapping. Srikanth et al. [11] (2006) applied wavelet denoising to statistical maps before FDR thresholding. Chen et al. [12] (2009), later commented by Reiss et al. [13] (2012), used Storey's positive FDR [5] and Benjamini et al.'s two-stage adaptive FDR [14] to enhance detection power by estimating the portion of positive regions from data. Schwartzman et al. [15] (2009) used Efron's framework [16] to estimate a null distribution from data instead of using the theoretical one, and Pendse et al. [17] (2009) specifically used the mixed Gaussian distribution as the empirical null distribution. Lawyer et al. [18] (2009) applied Efron et al.'s [19] local FDR method which defines the FDR as the empirical probability of a null hypothesis given a test value. Langers et al. [20] (2007) proposed the Searchlight FDR method to limit the impact

from remote regions on a local region. Zhang et al. [21] (2011) proposed median-filters over p -values to increase the detection power. Nguyen et. al. (2014) [22] revisited Genovese et al.'s method [9], and reduced speckles with a Markov random field. All these methods are based on the voxel-wise FDR definition and they share an obvious problem: if the image is spatially transformed, its volumetric measure will change and so does the FDR, as elaborated in Section II-C.

Pacifico et al. (2004) [23] pioneered to control the FDR for the number of clusters (which formally are connected excursion sets), instead of the volumetric ratio. He classified a cluster as false positive if its false positive volume exceeds a pre-defined ratio. Heller et. al. (2006) [24] then proposed a two-stage method: in the first stage, functional activated clusters are segmented with pilot data; then in the second stage, FDR control is applied to p -values derived from clusters' average signals. Benjamini and Heller (2007) [25] later extended the method in [24] by weighting each cluster with its volumetric size. Though the number of clusters does not change under diffeomorphic transformation, these methods all partially rely on volumetric measure which is transformation variant: classification of false positive clusters by volumetric ratio in Pacifico et al.'s method [23], taking clusters' average signals in Heller et. al.'s method [24], and weighting clusters with their sizes in Benjamini and Hochberg's method [25].

Chumbley and Friston (2009) [26] and Chumbley et. al. (2010) [27] worked on random field theory. According to Gaussian random field theory, they calculated cluster p -values by either their sizes [26] or peak signals [27], and then they applied Benjamini and Hochberg's step-up procedure [4, 10] for FDR correction. Their methods have been incorporated into software Statistical Parametric Mapping (SPM) [28]. Though these methods enjoy transformation invariance in their cluster-based FDR definition and p -value calculation, they introduced another problem: the number of clusters entering into the FDR stage, even pre-screened with a fixed threshold, becomes a random variable, while Benjamini and Hochberg's methods [4, 10] demand a fixed number of clusters. These methods then require the number of pre-screened clusters to be independent of cluster p -values, an assumption whose proof is unclear.

To achieve transformation invariance, we shall either define the FDR based on topological indices such as the number of clusters, or local properties such as local maxima, or if volumetric integration is involved, use a measure that does not change under transformation. Here we explore defining the measure in the signal space instead of directly in the image space. Although it is natural to define the measure directly in the image's Euclidean space, spatial transformation and atlas choice affect this measure. On the other hand, the signal space is intrinsic and unaffected, as elaborated in Section II-D. We find that normalized residuals of a Gaussian multi-variate linear regression model uniformly distribute on a unit hyper-sphere independent of its regression statistics such as the t -statistics, F -statistics and p -values. Therefore, the image and the normalized residuals form a mapping from a Euclidean space to a unit hyper-sphere. We can define the volumetric measure as the "area" spanned by signals on the unit hyper-sphere. Because volumetric measure is involved in many aforementioned FDR methods [9, 11, 12, 15, 17, 18, 20–25], embedding our hyper-spherical measure could equip them with transformation-invariance.

The flexibility in the methods used to analyze neuroimaging datasets is known to be one of the key causes of publication biases and false findings, as suggested by Ioannidis [29] and experimentally confirmed by researchers such as Joshua Carp [30]. Despite being widely recognized, this issue has seldom been tackled. This paper addresses one of these sources of flexibility, namely the choice of a reference frame. With the proposed method, researchers will be able to freely choose reference templates without worrying that such a choice may impact their findings.

This paper extends and significantly enhances our previously presented conference article [31] with detailed derivations, intuitive illustrations, three new experiments, and in-depth discussions and explanations. In Section II, we elaborate our method, including the unit hyper-sphere of normalized residuals, the weighted FDR, and the voxel-wise FDR using signal space measure. In Section III, we demonstrate with synthetic, semi-synthetic and real images that our method significantly reduces the bias introduced by the choice of test spaces. In Section IV, we discuss signals' spatial dependence and factors affecting transformation invariance.

II. Methods

Multiple testing in statistical mapping involves linear regressions, geometric measures, and correction of p -values. First, we present probabilistic properties of linear regressions with Gaussian noises. The property playing a central role is the uniform distribution of normalized residual vectors on a high-dimensional sphere. Next, we define the voxel-wise FDR for statistical mapping, both the original version and the weighted version. Then, we explain why the FDR varies under spatial transformations if its regional measures are defined directly in image spaces. Afterwards, we derive how transformation invariance can be achieved by defining regional measures in signal domains. All these are formulated in continuous spaces, not discretized images. We close this section with the numerical implementation of our method on discrete grids.

A. Intrinsic Spherical Signal Space

In this subsection, we study a linear regression at an image point individually and the conclusions are limited to a single image point. For the sake of simplicity, we do not indicate the point location, because spatial information is not involved in this subsection. Spatial relationships among linear regressions, such as structured noises, are discussed in Section II-D.

Let us consider a linear model $Y = X\beta + \varepsilon$ where X is an m -by- n ($m > n$) full column-rank matrix, Y is a column vector of length m , β is a column vector of length n , and ε is a noise vector whose elements ε_i ($i = 1, \dots, m$) are identical and independently distributed (i.i.d.) -variables following a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. In research applications, Y may be a functional MRI (fMRI) time serials at a voxel and X may be the design matrix of functional tasks, or in a tensor-based morphometry (TBM) study Y may be subjects' Jacobian determinants at a voxel and X may be a factor matrix of age, gender or disease states, etc. Please note:

- We assume that X is full column-rank so that the inverse of $X^T X$ exists.
- We assume that $m > n$ so it is possible to estimate σ^2 .

The maximum-likelihood and unbiased estimate of β is

$$\hat{\beta} \equiv (X^T X)^{-1} X^T Y = \beta + (X^T X)^{-1} X^T \varepsilon.$$

The residuals and the unbiased estimate of σ^2 are

$$\begin{cases} \hat{\varepsilon} & \equiv Y - X\hat{\beta} = \left[I - X(X^T X)^{-1} X^T \right] \varepsilon, \\ \hat{\sigma}^2 & \equiv \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{df}, \text{ where } df = m - n. \end{cases}$$

Let us focus on the probabilistic and geometric properties of $\hat{\varepsilon}$. We find:

- i. It follows a σ^2 -variance isotropic multi-variate Gaussian distribution embedded in the null space of X 's columns. More specifically, there exists an m -by- df matrix Z which satisfies $X^T Z = 0$ and $Z^T Z = I$, such that $\tilde{\varepsilon} \equiv Z^T \hat{\varepsilon} \sim N(0, \sigma^2 I_{df \times df})$, $\tilde{\varepsilon} = Z \tilde{\varepsilon}$, and $\hat{\varepsilon}^T \hat{\varepsilon} = \tilde{\varepsilon}^T \tilde{\varepsilon}$.
- ii. It is independent of $\hat{\beta}$ and β .
- iii. Its normalized vector $\hat{u} \equiv \hat{\varepsilon} / |\hat{\varepsilon}|$ uniformly distributes on a unit hyper-sphere in the null space of X 's columns, independent of σ^2 and $\hat{\sigma}^2$.

These properties imply that at an image point the normalized residual vector $\hat{u}$ stochastically uniformly distributes on a unit hyper-sphere, and is independent of statistics derived from $\hat{\beta}$ and $\hat{\sigma}^2$, such as the t-statistic and the F-statistic. Its distribution is also unaffected by the true value β and the true variance σ^2 . These statistical features make the unit sphere where $\hat{u}$ resides on a suitable space for metric definition. We outline the proof as follows:

1. Define $\tilde{\varepsilon} \equiv Z^T \hat{\varepsilon}$ where Z is an m -by- df matrix whose columns are a set orthonormal bases of the null space of X 's columns.
2. The mapping established by Z between $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ is linear, bijective, and isometric:
 - $\tilde{\varepsilon} \equiv Z^T \hat{\varepsilon}$ and $\hat{\varepsilon} = Z \tilde{\varepsilon}$,
 - $\hat{\varepsilon}^T \hat{\varepsilon} = \tilde{\varepsilon}^T \tilde{\varepsilon}$.
3. $\tilde{\varepsilon}$ is a df -element random vector following a Gaussian distribution $N(0, \sigma^2 I_{df \times df})$.
4. $\tilde{\varepsilon}$ is independent of β because by definition it does not involve β .
5. $\tilde{\varepsilon}$ and $\hat{\beta}$ are independent because they are both Gaussian and $\text{Cov}(\hat{\beta}, \tilde{\varepsilon}) = 0$.

6. $\tilde{u} \equiv \tilde{\varepsilon} / |\tilde{\varepsilon}|$ stochastically uniformly distributes on a high-dimensional sphere S^{df-1} , independent of σ^2 and $\hat{\sigma}^2 \equiv \hat{\varepsilon}^\top \hat{\varepsilon} / df$, because $\tilde{\varepsilon}$ is iostropically distributed, and the normalization removes its dependence on σ^2 and $\hat{\sigma}^2$.
7. Because of property (2), $\hat{\varepsilon}$ inherits properties (3), (4) and (5) from $\tilde{\varepsilon}$, and $\hat{u} \equiv \hat{\varepsilon} / |\hat{\varepsilon}|$ inherits property (6) from $\tilde{u}$, except that they are embedded in the null space of X 's columns.

The key of the proof is the linear, bijective and isometric mapping between $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ established in Step (2). For details, please refer to Appendix A in the supplementary materials. Even though $\hat{\varepsilon} \in R^m$ and $\tilde{\varepsilon} \in R^{df}$ are represented with different numbers of elements, they essentially have the same geometric structure. Similar is the relationship between $\hat{u}$ and $\tilde{u}$. In the remaining of this paper, we will just say that “ $\hat{u}$ uniformly distributes on S^{df-1} ”, for the sake of simplicity.

B. Weighted FDR

Let R_{pos} denote the detected region, R_{tru} the underlying truth, and $/\bullet/$ the volume of a region. The volume-based FDR is defined as follows:

$$FDR \equiv E \left[\frac{|R_{pos} \setminus R_{tru}|}{|R_{pos}|} \right] \text{ where } \frac{0}{0} \equiv 0. \quad (1)$$

Genovese, Lazar, and Nichols [9] defined the volumetric measure in the image space, and counted it as the number of voxels. Benjamini and Hochberg's step-up procedure [4] was applied to control the FDR. Given a user specified FDR level q^* , the step-up procedure finds

$$p^*(q^*) = \max \left\{ p_{(k)} \mid \frac{p_{(k)} N}{k} \leq q^* \right\},$$

where $p_{(k)}$ is the k -th smallest voxel p -value, and N is the number of voxels. This step-up procedure is able to handle positive dependence among tests, as Benjamini and Yekutieli discussed in [10]. For more general dependence among tests, please refer to [10].

Benjamini and Hochberg (1997) [32] upgraded the step-up procedure to a weighted version as follows:

$$FDR \equiv E \left[\frac{\sum_{i \in R_{pos} \setminus R_{tru}} w_i}{\sum_{i \in R_{pos}} w_i} \right], \quad (2a)$$

$$p^*(q^*) = \max \left\{ p_{(k)} \mid \frac{p_{(k)} \sum_{i=1}^N w_{(i)}}{\sum_{i=1}^k w_{(i)}} \leq q^* \right\}, \quad (2b)$$

$$q_{(k)} = \min \left\{ \frac{p_{(j)} \sum_{i=1}^N w_{(i)}}{\sum_{i=1}^j w_{(i)}} \mid p_{(j)} \geq p_{(k)} \right\}, \quad (2c)$$

where w_i is the weight associated with a voxel i and $q_{(k)}$ is the corresponding FDR level if voxels are thresholded at $p_{(k)}$. Derived from Eq. (2b), Eq. (2c) calculates for each voxel the q -value if the statistical map is thresholded at its p -value. With this $p \rightarrow q$ transform, we can later more flexibly threshold at any arbitrary FDR level, instead of just finding the p^* to produce a binarized mask of the positive region associated with a particular FDR threshold q^* .

The continuous counterpart of Eq. (2) is:

$$\text{FDR} \equiv E \left[\frac{\int_{R_{\text{pos}} \setminus R_{\text{tru}}} w(x) dx}{\int_{R_{\text{pos}}} w(x) dx} \right], \quad (3a)$$

$$p^*(q^*) = \sup \left\{ p \mid \frac{p \int_{R_{\text{all}}} w(x) dx}{\int_{R_{\leq p}} w(x) dx} \leq q^* \right\}, \quad (3b)$$

$$q(x) = \inf \left\{ \frac{p \int_{R_{\text{all}}} w(x) dx}{\int_{R_{\leq p}} w(x) dx} \mid p \geq p(x) \right\}, \quad (3c)$$

where $w(x)$, $p(x)$ and $q(x)$ are respectively the weight density, the p -value and the corresponding q -value at a point x , and R_p is the region whose p -value supremum is p , that is $R_p \equiv \{x \mid p(x) \leq p\}$. In the continuous case, there are an infinite number of p -values residing in a continuous space, so Eq. (3) replaces maxima, minima, subsets and summations in Eq. (2) respectively with suprema, infima, sublevel sets and integrals. We consider Eq. (3) the theoretical formulation, and Eq. (2) its discretization.

C. FDR Weighted with Image Space Measures

The definition of FDR in Eq. (3a) and the conversion from p -values to q -values in Eq. (3c) not only involve p -values but also regional weights. Let $w(R)$ denote the weight of a region

R : $w(R) = \int_R w(x)dx$, then Eq. (3) involves $w(R_{pos}|R_{tru})$, $w(R_{pos})$, $w(R_{all})$ and $w(R_p)$. If regional weights are defined as their volume measured directly in image spaces, then they vary as images are spatially transformed, as shown by the following derivation.

Given two diffeomorphic manifolds A and B , there is a smooth map Φ with a smooth inverse establishing correspondence between them. Φ by definition satisfies the following properties.

- Given any point $x \in A$, we have $y = \Phi(x) \in B$.
- Given any point $y \in B$, we have $x = \Phi^{-1}(y) \in A$.
- Both Φ and Φ^{-1} are differentiable. (If Φ and Φ^{-1} both are d -order continuously differentiable, then Φ is called a C^d diffeomorphism.)

Given a region R in space A , its volumetric weight calculated in space A is Eq. (4a). When space A is mapped to space B , region R becomes $\Phi \circ R \equiv \{y = \Phi(x) | x \in R\}$ and correspondingly its volumetric weight calculated in space B becomes Eq. (4b).

$$\text{vol} w_A(R) \equiv \int_R 1dx = \int_{\Phi(R)} |\det J_{\Phi^{-1}}(y)| dy \quad (4a)$$

$$\text{vol} w_B(\Phi \circ R) \equiv \int_{\Phi \circ R} 1dy = \int_R |\det J_{\Phi}(x)| dx \quad (4b)$$

In Eq. (4), $J_{\Phi}(x)$ and $J_{\Phi^{-1}}(y)$ denote the Jacobian matrices of Φ and Φ^{-1} at x and y respectively. If $|\det J_{\Phi}(x)|$ is a constant value, for example, in affine transformations, then this constant ratio between $\text{vol} w_A(R)$ and $\text{vol} w_B(\Phi \circ R)$ applies to both the numerators and denominators in Eq. (3), and is canceled. In this case, the FDR results do not change under transformations. However, if $|\det J_{\Phi}(x)|$ depends on x (and so does $|\det J_{\Phi^{-1}}(y)|$ on y), which is typical in nonlinear transformations, then the FDR results are not guaranteed to be consistent between space A and space B . Increase of the ratio $\text{vol} w(R_p)/\text{vol} w(R_{all})$ will likely enhance detection sensitivity, while decrease of the ratio will likely reduce sensitivity.

One may argue that “As long as we already know the correspondence between R and $\Phi \circ R$, we can enforce $\text{vol} w_B(\Phi \circ R) \equiv \text{vol} w_A(R)$ ”. However this argument in general is unsatisfactory for the following reasons. First, given a manifold A , there might be an infinite number of diffeomorphisms to spatially transform it. For example, for a unit cube $A = (0, 1)^3$, any smooth velocity field on A can generate a diffeomorphism via geodesic shooting [33]. Second, given two diffeomorphic manifolds A and B , we do not have a clear criteria to favor one over another. B can be regarded as a warp of the “standard” A through Φ , but vice versa A can also be regarded as a warp of the “standard” B through Φ^{-1} . Enforcing $\text{vol} w_B(\Phi \circ R) \equiv \text{vol} w_A(R)$ essentially favors space A over space B . Third, composition of many small diffeomorphisms can generate large diffeomorphic deformations. Let $\Phi_1, \Phi_2 \dots$ be a series of small diffeomorphisms, and $A_1 = \Phi_1 \circ A_0, A_2 = \Phi_2 \circ A_1 \dots$ be a series of spaces generated by them. Although each diffeomorphism may only introduce a small deformation, the difference between space A_0 and space A_k might be very large

when k is large enough. In summary, if regional weights are defined as their volume measured directly in image spaces, we face an infinite number of choices of spaces and each of them may yield a different thresholding result for the same FDR threshold.

D. FDR Weighted with Signal Space Measures

Section II-A focuses on properties of the normalized residual vector $\hat{u}$ at an individual image point and reveals that it stochastically uniformly distributes on a high-dimensional hypersphere S^{df-1} . In this section, we show how to employ the field composed of normalized residual vectors to achieve transformation invariance. Please note:

- We assume that the design matrix X is the same for every point in the image, so all the normalized residual vectors are embedded in the column null-space of the same X .
- We do not require the model parameter β or the noise variance σ^2 to be the same for every point in the image. Different point locations may have different β s and σ^2 s.

After linear regressions, each point x is associated with a normalized residual vector, and all normalized residual vectors in a spatial domain A form a mapping $f: x \in A \rightarrow \hat{u} \in S^{df-1}$. Because spatial transformations deform A but not its mapped areas on S^{df-1} , we can achieve transformation invariance by defining FDR weights in the signal domain S^{df-1} , as illustrated in Figure 2(a) and formulated as follows.

Let $f \circ R \equiv \{f(x)/x \in R\}$ denote the area on S^{df-1} where a region R is mapped to through f . Instead of defining regional weight $w(R)$ as the image volume of R , we can define it as the area size of $f \circ R$, as $\text{sig } w_A(R)$ in Eq. (5a), where $\text{sig } w(\bullet)$ denotes the signal-space measure of region $\bullet$ in space S^{df-1} .

$$\text{sig } w_A(R) \equiv \int_{f \circ R} 1 d\hat{u} = \int_R \sqrt{\det [J_f^\top(x) J_f(x)]} dx \quad (5a)$$

$$\begin{aligned} \text{sig } w_B(\Phi \circ R) &\equiv \int_{(f \circ \Phi^{-1}) \circ (\Phi \circ R)} 1 d\hat{u} \\ &= \int_{\Phi \circ R} \sqrt{\det [J_{f \circ \Phi^{-1}}^\top(y) J_{f \circ \Phi^{-1}}(y)]} dy \end{aligned} \quad (5b)$$

$$\because f \circ R = (f \circ \Phi^{-1}) \circ (\Phi \circ R) \quad (5c)$$

$$\therefore \text{sig } w_A(R) = \text{sig } w_B(\Phi \circ R) \quad (5d)$$

When space A is mapped to space B via transformation Φ , region R becomes $\Phi \circ R \equiv \{y = \Phi(x)/x \in R\}$, and at the same time the mapping from the image space to $\mathcal{S}^{df-1}$ becomes $f \circ \Phi^{-1}$. Consequently, the regional weight of $\Phi \circ R$ calculated via the space-to-signal mapping becomes $\text{sig } w_B(\Phi \circ R)$ in Eq. (5b). In Eq. (5), $J_A(x)$ and $J_{f \circ \Phi^{-1}}(y)$ denote the Jacobian matrices of f and $f \circ \Phi^{-1}$ at point x and y respectively.

As shown in Eq. (5c), the area where $\Phi \circ R$ mapped to through $f \circ \Phi^{-1}$ is actually the same as the area where R mapped to through f . Therefore, $\text{sig } w_B(\Phi \circ R)$ equals $\text{sig } w_A(R)$. This is the key to achieve transformation invariance.

Please note that as long as we have got normalized residual vectors in space B , the calculation of $\text{sig } w_B$ does not need to refer to space A again via Φ^{-1} , because the normalized residual vectors in space B have already directly established the mapping $f \circ \Phi^{-1}$. For this reason, when we conduct FDR correction, we just need to directly calculate regional weights through normalized residuals, and the results will be consistent with those in other diffeomorphically transformed spaces, even though we do not necessarily know the transformations connecting them.

The proposed method does not require homogeneous noises. Actually, the property of transformation invariance in its very nature accommodates inhomogeneity. Homogeneous noises in one image space will become inhomogeneous after nonlinear deformation into another image space. A transformation-invariant method works in both the original space and the transformed one, deal with both homogeneous and inhomogeneous noises, and produce consistent results. Therefore, our method does not require homogeneous noises, as demonstrated in the experiment in Section III-A. For clarification, please refer to Appendix B in the supplementary materials.

Reasons for weighting voxels with random residuals—The answer lies in the fact that spatial correlation of residuals is not purely noisy but reflects the “effective volume of randomness” held by a region. First, even though residuals at a single voxel are noises of its own linear regression model, residuals of neighboring voxels may still correlate, forming spatial pattern due to spatial correlation. The spatial correlation is not a result of randomness, but essentially due to the smoothness of signals. For example, fMRI signals are more often than not spatially smooth, so be the residuals after regression to activation tasks. Though spatial correlation estimated from residuals is not deterministic, its variability decreases as the degree of freedom increases. With a large degree of freedom, we expect a reliable estimate of spatial correlation.

Second, spatially correlated noises ε lead to spatially correlated p -values. A p -value, as a random variable uniformly distributes in $[0, 1]$ under the null hypothesis, sources its randomness from the noises. If the noises at neighboring voxels correlate, their p -values also correlate. The residuals, as the projection of ε to the column null space of X , partially captures the spatial correlation in ε . The higher the ratio df/m is, the more information in ε is kept in the residuals.

Third, two voxels with highly correlated p -values should not be treated equally as those with independent p -values. It is the principle of multiple testing that the more independent tests there are, the more stringent the threshold should be. If the noises of two voxels correlate perfectly, their p -values will always be the same, and they should not be treated as two independent tests, but just one. In the signal space, neighboring voxels with higher correlation will be placed closer than those with lower correlation, reflecting their lower degree of independence, as illustrated in Figure 2(b). The larger $\text{sig}w$ is, the more randomness is covered by a region, and the more trustful positive results in it are. In this sense, $\text{sig}w$ relates to multiple testing in an essential way. Therefore, $\text{sig}w$ stands for the “effective volume of randomness”, similar to the concept of “effective number of independent tests” exploited in genetic association studies [34, 35].

E. Numerical Implementation on Discrete Grids

The weight of a voxel on a discrete grid is calculated from the finite difference between its normalized residual vector and those of its neighboring voxels. For a voxel on a D -dimensional grid, let $d-$ and $d+$ ($d=1, \dots, D$), as shown in Figure 3(a), denote the voxels immediately next to it by respectively decreasing or increasing one index along the d -th dimension, and let u and $u^\dagger$ ($\dagger \in \{d-, d+ \mid d=1, \dots, D\}$) respectively denote the normalized residual vectors associated with the central voxel and a neighboring voxel. The weight $\text{sig}w$ of the central voxel is calculated as follows.

1. Map $u^\dagger$ ($\dagger \in \{d-, d+ \mid d=1, \dots, D\}$) to the tangent plane at u as a velocity vector $v^\dagger$:

$$v^\dagger \equiv \log_u(u^\dagger) \equiv \frac{u^\dagger - u \cos \varphi}{\|u^\dagger - u \cos \varphi\|} \varphi,$$

where $\varphi \equiv \arccos \langle u, u^\dagger \rangle$.

Because normalized residual vectors reside on a unit hyper-sphere S^{D-1} , we do not simply take their linear differences. Instead, we keep u as the center, and map $u^\dagger$ to its tangent plane as a velocity vector $v^\dagger$, as shown in Figure 3(b). The length of $v^\dagger$ is the same as the great arc connecting u and $u^\dagger$, and the direction of $v^\dagger$ is the initial direction of the great arc starting at u .

2. Average the Jacobian determinants derived from all possible combinations of forward and backward finite differences:

$$\text{sig}w = \frac{1}{2^D} \sum_{v^1 \in \{v^{1-}, v^{1+}\}} \dots \sum_{v^D \in \{v^{D-}, v^{D+}\}} \sqrt{\det \left\{ [v^1 \dots v^D]^\top [v^1 \dots v^D] \right\}}.$$

The velocity vectors of the $2D$ (2 times D , not two-dimensional) neighboring voxels can be divided into D groups each along one dimension: $\{v^{d-}, v^{d+}\}$ where $d=1, \dots, D$. A Jacobian matrix can be

composed by picking one velocity vector from each of the group. In total, there are 2^D possible combinations and the weight of the central voxel is calculated as the average Jacobian determinant of all the possible combinations.

We used the Kahan summation algorithm [36] to sum up regional weights. It can reduce numeric errors in adding up a large number of small values represented as finite precision floating point numbers.

III. Experiments

The proposed method is compared with Genovese, Lazar, and Nichols' [9] voxel-wise FDR method which uses the image space measure. For convenience, we denote the two methods as sigFDR and volFDR respectively. We choose the method in [9] for comparison because it is the most straightforward adaptation of the FDR to statistical mapping, and many variants [12, 15, 18, 20–22] are derived from it. In this way, we minimize influence introduced by other factors, such as Markov random field smoothing [22] or cluster pre-screening [24], but solely focus on the impact of different volumetric measures.

The experiments are conducted with synthetic, semi-synthetic and real brain images. Homogeneous and inhomogeneous noises are used in synthetic images. Inhomogeneous noise derived from real images are used in semi-synthetic images. In synthetic and semi-synthetic experiments, signals are inserted to specific regions as the ground truth to verify whether the sigFDR method is still able to control the FDR around user specified levels. The synthetic and semi-synthetic experiments by no means imply real applications, but demonstrate the phenomena of transformation invariance or variance. Real 3D images of the Philadelphia Neurodevelopmental Cohort (PNC) [37] and the Alzheimer's Disease Neuroimaging Initiative (ADNI) are also used in the experiments. With real brain images, FDR invariance are tested across multiple transformations derived from real images, with real inhomogeneous noises and real but unknown signals.

A. Experiment with Exact Transformations

1) Experiment Design—In this experiment, FDR correction in two 2D image spaces are compared. The two image spaces are connected with an analytic transformation, so the transformation and its inverse are exactly accurate. Correction results, q -values obtained in the two spaces, after transformed to one of the spaces, are compared to check consistency. The experiment is performed at different image resolutions. As the resolution becomes finer and finer, interpolation error is minimized, and the continuous formulations in Section II-D is approximated.

2) Data—Experiment data are shown in Figure 4. Subfigures (a) and (b) show the two spaces A and B respectively. Space A is a disk with radius 50mm centered at the origin. The transformation Φ from space A to space B is

$$y = \Phi(x) = 50 \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \frac{\ln(|x|/10+1)}{\ln 6} \frac{x}{|x|},$$

where $x \in A$ and $y \in B$. Signals and noises are generated in space A . The signals, as shown in subfigure (c) have the shape of a Gaussian distribution with radius 10 centered at the origin. The peak of the signals is normalized to be 1. White noises with amplitude 1.0 are generated on discrete grids spaced by 1.00mm. The white noises are then smoothed with anisotropic diffusion to produce spatial inhomogeneity. The diffusion conductivity is shown in Subfigure (d), and the diffusion time is 0.5. Subfigure (e) shows the noise amplitude after the anisotropic diffusion. Subfigure (f) shows average correlation coefficient with nearest neighbors. Regions with higher diffusion conductivity in subfigure (d) show lower noise amplitude and higher correlation coefficients with neighbors. 50 noise images are generated independently, so the length of a noise vector is 50. Experiments are repeated at resolutions of 2.00mm, 1.00mm, 0.5.0mm and 0.25mm. Images at resolutions other than 1mm are produced by interpolation of those at 1.00mm.

3) Results—Figure 5 shows the experiment results presented in space A , where the upper row is for the volFDR method and the lower row is for the sigFDR method. The left panel shows the difference between q -values obtained in the two image spaces, in the logarithmic scale. For the volFDR method, the difference of $\ln(q)$ between the two image spaces does not decrease as the image resolutions become finer and finer, from 2.00mm to 0.25mm. The difference keeps at almost the same level. For the sigFDR method, the difference of $\ln(q)$ between the two image spaces diminishes to almost zero when the image resolution reaches 0.25mm. The right panel shows the $-\ln(q)$ map at the resolution of 0.25mm. For the volFDR method, its results in the two spaces are apparently different, while for the sigFDR method, its results are very similar. This clearly demonstrates that the sigFDR method as formulated with the continuous equations in Section II-D, is transformation invariant in its nature.

B. Experiment with Synthetic 2D Images

1) Experiment Design—The experiments are performed in the following framework: vector-valued images are generated in its original space which we call the fixed image space, and then warped with linear interpolation to a new space which we call the moving image space. After FDR thresholding is applied at the 5% level in both the fixed and the moving image spaces, the results in the two image spaces are compared. To evaluate the FDR, positive signals are added to some regions, so the ground truth is known in the image generation stage and the evaluation stage, but blind to the FDR thresholding stage. This procedure randomly repeats several times to estimate the expected value and standard deviation of performance indices. We span the signal to noise ratio (SNR) in a range to provide a broad view of the performance. The evaluation indices include:

- Image-wise consistency: Let P_{fix} and P_{mov} respectively denote the detected binary image (where 1 stands for positive and 0 for negative) in the fixed and the moving image spaces, and $P_{mov} \circ \Phi$ the transformation of P_{mov} to the fixed image space with linear interpolation followed by

thresholding at 0.5. Image-wise consistency, measured with the Dice index between P_{fix} and $P_{mov} \circ \Phi$, is

$$\text{Dice} \equiv \frac{2 \sum_{i \in \text{brain}} P_{fix}(i) \wedge P_{mov} \circ \Phi(i)}{\sum_{i \in \text{brain}} [P_{fix}(i) + P_{mov} \circ \Phi(i)]}.$$

- Image-wise detection power: Let T denote the binary image of the ground truth, and P a detection result in the same space. Image-wise detection power is

$$\text{Power} = \frac{\sum_{i \in \text{brain}} P(i) \wedge T(i)}{\sum_{i \in \text{brain}} T(i)}.$$

P_{fix} and $P_{mov} \circ \Phi$ are respectively used as P and the truth image in the fixed image space is used as T .

- FDR: For P_{fix} and $P_{mov} \circ \Phi$, their FDR is calculated with Eq. (1).
- Voxel-wise consistency: The detection of a voxel in repetitive trials builds a binary vector. The detection consistency at a voxel location is the Dice index between the two binary vectors derived from P_{fix} and $P_{mov} \circ \Phi$.
- Voxel-wise detection power: The detection power at a voxel is the ratio of its detection in the repetitive trials.

2) Data—50 T1-weighted MR images are selected from the PNC [37] to construct an 3D atlas. An axial slice of the atlas is extracted as the fixed image, and high-sampled to the 0.5mm resolution. A nonlinear deformation is generated by registration to a 2D slice extracted from an atlas constructed from 9 elder subjects. For details on how the deformation field is generated, please refer to Appendix C in the supplementary materials. In each trial, 40 images filled with unit-variance white noises are simulated and then smoothed with a 4-voxel wide Gaussian kernel. Full width at half maximum (FWHM) of the smoothing kernel is $2\sqrt{2 \ln 2} \times (0.5 \times 4/2) \approx 2.35\text{mm}$. Then, single are added in the regions shown in Figure 6. The SNR ranges from 0.1 to 0.2. Voxel p -values are calculated with the two-sided t-test. 50 random trials are repeated.

3) Results—The image-wise FDR, detection power and consistency are shown in Figure 7. Both the sigFDR (green) and the volFDR (red) methods control the FDR, as measured in the fixed image space, around the target 5%. However, the volFDR shows considerable difference in its FDR and detection power between detection directly in the fixed image space (solid curves) and that warped back from the moving image space (dashed curves). The sigFDR shows much less difference. This is also evidenced by the Dice index plot.

The voxel-wise detection power and consistency at SNR = 0.16 are shown in Figure 8. The power images of the volFDR method show considerable difference between the two image spaces, while the sigFDR results are much more consistent. The Dice images show evidence of the sigFDR method's improved consistency.

C. Experiment with Semi-Synthetic 3D Images

This experiment follows the same design described in Section III-B1.

1) Data—50 subjects are randomly selected from the PNC [37]. The sampled subjects include 29 males (average age = 14.069 with std. = 3.741), and 21 females (average age = 15.4286 with std. = 3.9060). A 3D T1-weighted atlas is constructed and a normalized Jacobian determinant map are derived from TBM analysis for each subject. To build null hypotheses, we remove the linear effects contributed by both age and gender from the normalized Jacobian maps, but just utilize the residual images. Please note that the residual images still hold the inhomogeneous spatial dependence in the real data. Voxels in smooth regions will have similar residuals, and those in coarse regions will have weak correlation in their residuals. 50 random trials are generated with bootstrap as follows. In each trial, 50 subjects are randomly re-sampled to replace the original residuals, and then signals (SNR=0.1~0.2) correlated with their ages are added to the hippocampus region, as shown in Figure 9. Voxel p -values are calculated with the two-sided t -test to detect the effect of age. The hippocampus region is shrunk to produce moving images. For details on the deformation field, please refer to Appendix D in the supplementary materials.

2) Results—The image-wise FDR, detection power and consistency are shown in Figure 10. The sigFDR (green) controls the FDR closer to the target value 5% than the volFDR (red) does. The detection power of the volFDR decreases by about 7% in the moving image space (dashed curves), while that of the sigFDR only changes slightly. The sigFDR also gains 7%~20% absolute increase in the Dice consistency. Readers should note that it is possible for the volFDR to achieve higher detection power if the regions of interest expand instead of shrink. The key advantage of the sigFDR is its consistency under spatial transformation.

The voxel-wise detection power and consistency at SNR=0.15 are shown in Figure 11. The detection power of the volFDR decreases considerably in the moving image space, while the sigFDR method does not show significant changes. The sigFDR also achieves high voxel-wise consistency.

D. Experiment with PNC Real 3D Images

1) Experiment Design—With real 3D images, we cannot evaluate the realized FDR due to the unknown truth, but we can estimate the consistency of detection under different spatial transformations following the procedure illustrated in Figure 12. Given a set of N_{images} data images in the same space A , and a set of N_{trans} random spatial transformations and their inverses, we conduct the experiment as follows. Each time we randomly sample a subset of N_{samples} from the data images. Then we warp them together with each of the N_{trans} transformations and in the warped image spaces conduct statistical mapping and calculate q -values with Eq. (3c). To compare the results obtained in different spaces, we inverse-warp their $-\ln(q)$ images back to space A and calculate voxel-wise statistics (such as mean and standard deviation) about $-\ln(q)$ across different transformations. We repeat this procedure N_{trials} times and average the statistics.

This experiment design features two kinds of randomization and quantitative comparison of $-\ln(q)$. The first is transformation randomization: statistical mapping and FDR control are conducted in multiple randomly warped spaces, not just one fixed image space and one moving image space. The second is data randomization: each time a subset of data images are randomly sampled and this procedure is repeated. All statistics, such as standard deviations, are first calculated across the N_{trans} spatial transformations, and then averaged over the N_{trials} repetitive random trials. Statistics are calculated across spatial transformations because checking transformation invariance is the interest of the experiment; statistics are averaged over repetitive random trials to estimate the expected performance. Moreover, instead of analyzing binary masks at a particular FDR level such as 5%, we compare the logarithm of q -values calculated with Eq. (3c). Because q -value images allow thresholding at any arbitrary FDR level, this comparison provides a “continuous” rather than “binary” view on consistency.

2) Data—The data images are normalized Jacobian determinant maps of the $N_{\text{images}}=976$ subjects in the PNC dataset. The normalized Jacobian maps are derived by nonlinearly registering subjects’ T1 MR images to the 3D template constructed in Section III-B. Transformations of $N_{\text{trans}}=24$ randomly sampled subjects to the template are used for validating invariance. Advanced Normalization Tools (ANTS) [38] with its default configuration were used for the nonlinear registration and the atlas construction. Each time the Jacobian determinant maps of $N_{\text{samples}}=100$ subjects are randomly sampled as signal images, and then linearly regressed to subjects’ age with gender as the cofactor, followed by two-sided t-tests. This procedure is repeated for $N_{\text{trials}}=20$ times.

The 976 subjects are aged between 8 and 21, and there are 464 males whose average age is 13.72 years (std. = 3.40) and 512 females whose average age is 14.34 years (std. = 3.51). In the 24 randomly selected subjects serving as the image spaces for multiple testing, there are 14 males and 10 females and their average age is 12.67 years (STD= 4.09).

3) Results—Figure 13 shows the performance statistics on PNC real 3D images. Subfigures (a) for the volFDR and (b) for the sigFDR show standard deviations s of $s = \text{std}[-\ln(q)]$, with degrees of freedom equal $N_{\text{trans}} - 1 = 23$. The standard deviations s_{volFDR} of the volFDR method are not only higher than those of the sigFDR, but also cover a broader region, forming structural patterns. On the other hand, the standard deviations s_{sigFDR} of the sigFDR method in general distribute sporadically. To examine whether the differences are statistically significant, s_{volFDR} and s_{sigFDR} are compared with the F-test, and the statistics in logarithms and the p -values in negative logarithms are shown in subfigures (c) and (d) respectively. Extensively in the white matter region, the p -values of the F-test reaches the range $[e^{-12.5} = 3.73 \times 10^{-6}, e^{-25} = 1.39 \times 10^{-11}]$, evidencing the statistical significance of the differences.

Subfigures (e) and (f) show expected values g_{volFDR} and g_{sigFDR} of $-\ln(q)$ respectively for the volFDR and the sigFDR methods. The two subfigures are visually very similar. Subfigure (g) shows the contrast g_{volFDR} and g_{sigFDR} . In most areas, the difference is negligible, except a few small regions.

Subfigure (h) shows the t-statistics of correlation between Jacobian determinants and age, with gender as the cofactor. Consistent with previous studies [39, 40] on teenagers, their white matter areas grow and gray matter areas shrink. Regions with large t-statistics overlap with the regions where the volFDR shows considerable variances of $-\ln(q)$ under spatial transformations, as shown in subfigure (a). We do not consider the overlap as just a coincidence, but explain with the following two reasons. First, for regions with large regression t-statistics regarding age, the denominator term $w(R_{\leq p}) \equiv \int_{R_{\leq p}} w(x) dx$ in Eq. (3c) varies considerably across the applied spatial transformations because they are sampled from teenager subjects at different ages. Second, the denominator term $w(R_{\leq p})$ is small with a small p (or equivalently, a large absolute value of the t-statistic), so the ratio term $\int_{R_{\text{all}}} w(x) dx / \int_{R_{\leq p}} w(x) dx$ in Eq. (3c) is more sensitive to its variance across different spaces.

E. Experiment with ADNI Real 3D Images

This experiment follows the same design described in Section III-D1, but replaces data images with brain atrophy rate maps derived from images of the ADNI. Data used in this section were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

1) Data—1.5T T1-weighted MR images are selected. This subset includes images of 274 males and 200 females. Among them, 97 are diagnosed with Alzheimer's disease, 99 with late mild cognitive impairment, 112 with early mild cognitive Impairment, and 166 are normal subjects. Their average age at recruitment was 77.46 years (STD=6.56).

For each subject, his/her image acquired 24 months after the initial scan at his/her recruitment was nonlinearly registered to the initial scan. Logarithms of Jacobian determinants are derived from the transformations and divided by 24 as monthly brain atrophy rate maps. These brain atrophy rate maps are originally in each subject's native space at the initial scan.

To validate transformation invariance, $N_{\text{trans}}=25$ brain images are randomly sampled and used as the common spaces for statistical mapping. The brain atrophy rate maps are transformed to each to the $N_{\text{trans}}=25$ spaces and then linearly regressed to subjects' disease states, with age and gender as the cofactors, followed by two-sided t-tests to contrast Alzheimer's Disease patients over normal elders. Corrected FDR q -values are then transformed to an atlas image for comparison. In total there are $N_{\text{images}}=474$ brain atrophy rate maps. Each time, $N_{\text{samples}}=100$ maps are randomly sampled as data images for regression, and this procedure is repeated for $N_{\text{trials}}=20$ times.

The transformations between images are established by nonlinear registration of T1-weight structural brain images to an atlas image. The atlas image is constructed from 1.5T T1-weighted MR images of 40 normal elders. Nonlinear registrations and atlas construction are conducted with the Advanced Normalization Tools (ANTs) [38] with the cross-correlation energy function and the “SyN” transformation model.

2) Results—Figure 14 shows the performance statistics on ADNI real 3D images. Subfigures (a) for the volFDR and (b) for the sigFDR show standard deviations s of $-\ln(q)$ obtained across different image spaces, with degrees of freedom equal $N_{\text{trans}} - 1 = 24$. The standard deviations of the volFDR method are much higher and broader than those of the sigFDR, involving temporal lobes, ventricles, and parietal lobes. We compare s_{volFDR} and s_{sigFDR} statistically with the F-test, and the statistics in logarithms and the p -values in negative logarithms are shown in subfigures (c) and (d) respectively. The ratio of s_{volFDR} over s_{sigFDR} reaches $\sqrt{e^5} \approx 12.18$, and the p -values of the F-test reaches the range $[e^{-10} = 4.54 \times 10^{-5}, e^{-25} = 1.39 \times 10^{-11}]$. The decrease of standard deviations by the sigFDR method is not a result of randomness, but is statistically significant.

Subfigure (e) for the volFDR and subfigure (f) for the sigFDR show $g = E[-\ln(q)]$ obtained across different image spaces. The two subfigures are visually very similar. The contrast of g_{volFDR} and g_{sigFDR} is shown in subfigure (g). g_{volFDR} is larger in ventricles, but mostly smaller elsewhere.

Subfigure (h) shows t-statistics contrasting Alzheimer’s Disease patients over normal elders. Ventricles of the patients expand, and temporal lobes and parietal lobes shrink. These regions overlaps where the volFDR method produces larger standard deviations of $-\ln(q)$, as shown in subfigure (a). Because volumes of these regions experience greater longitudinal changes, their volFDR also have greater variances if corrected in different image spaces. As discussed in Section II-C, regional expansion and shrinkage may also offer the volFDR method different regional sensitivity other than the sigFDR method, for instance, the increase in ventricles and the decrease in temporal lobes and parietal lobes.

IV. Discussions and Conclusions

We have shown in theory that defining volumetric measure in the intrinsic signal space can achieve transformation invariance in FDR control. Our experiments evidence the significant gain in consistency across thresholding taking place in different image spaces. Because the volumetric measure is a fundamental component in FDR control, it can benefit many variants of spatial FDR methods, for instance, the positive FDR [5, 12], the adaptive FDR [12, 14], the local FDR [18, 19], the regional FDR [20], the FDR with Markov random field smoothing [22], the cluster-based FDR [24, 25], as we have already shown with Genovese, Lazar, and Nichols’ method [9]. Therefore, the relationship between our method and other methods are complementary rather than competitive.

Volumetric measure in signal space is closely related to signals’ “effective volume of randomness”. Generally, the more tests are conducted simultaneously, the more stringent the p -values should be adjusted. Due to dependence among the tests, it may be over-

conservative to adjust with the number of tests, but more suitable with the effective number of independent tests, as successfully demonstrated in many genetic studies [34, 35]. The volumetric measure in signal space indicates the dispersion of signals on the unit hyper-sphere, reflecting the effective volume of randomness contained in the signals.

Worsley et al. and Taylor [41–43] used resels to adjust the FWER in random field theory, under the assumption that the image space can be “flattened” to make signals’ spatial dependence homogeneous. Our method does not rely on such an assumption, but consider the normalized residual image as a mapping from a Euclidean space to a unit hyper-sphere. Worsley and Taylor’s works were dedicated to anisotropic signals and the FWER, while ours is dedicated to transformation invariance and the FDR. Moreover, the FWER prohibits even a single error while the FDR balances between false positives and sensitivity, as discussed in Section I. The FWER is primarily related to the total integral of resels over the whole image space, while the FDR is also affected by the regional integral of spherical weights over positive areas, so the FDR is more sensitive to the definition of volumetric measures than the FWER does. Putting these two works together, it is very clear that transformation invariance and spatial dependence are two closely related issues.

The proposed method, as formulated in continuous cases, is able to achieve invariance under diffeomorphic transformations. However, with voxelized images in practice, the following factors also influence realization of transformation invariance. First, interpolation with discretized images introduces errors. If original images are not collected at a spatial resolution higher than the Nyquist frequency, then interpolation errors are inevitable and areas with large deformations tend to suffer from this type of errors. Second, composition of multiple discretized transformations accumulates errors, because composition also involves interpolation and discretization. For example, composition of a transformation and its inverse does not always perfectly yield the identity transformation. To minimize these two types of errors, we use an analytic transformation with an exact inverse and high-sample images to 0.25mm in Section III-A. The results in Figure 5 clearly demonstrate the transformation invariance of the sigFDR method when the continuous case is approached.

In the experiments on the PNC and ADNI datasets, cross-space standard deviations of the volFDR method’s logarithmic FDR values, as shown in Figures 13(a) and 14(a), roughly reaches 0.3. If we assume that logarithmic FDR values follow a Gaussian distribution across spaces, then for an expected FDR value of 5%, it will fall in the range of 3.05%~8.19% in 90% of the subject spaces. Such a bias range is acceptable. Therefore, standard practices in brain imaging of collecting statistics in a well-defined atlas space, such as the MNI152, make it unlikely that large bias would occur with the volFDR method.

The properties of regression residuals in Section II-A is proved for one-level fixed-effect models where the effect of interest β is assumed to be deterministic though unobserved. Group analysis of fMRI often deals with two-level mixed-effect models [44, 45] and usually makes inference in the summary statistic approach [44, 45] at the group level, for example, $\hat{\beta} = X_g \beta_g + \varepsilon_g + \varepsilon_s$ where $\hat{\beta}$ is activation coefficients individually estimated at the subject-level, X_g is the group design matrix, β_g is the group effects of interest, the random effect ε_g accounts for inter-subject variability, and ε_s is the unobserved intra-subject errors of $\hat{\beta}$. In

this case, the proof in Section II-A does not hold, but after all the ad hoc inference procedure of β_g [45] provides an adaptation in this situation. After estimating the mixed variance $\hat{V}$ for $\varepsilon = \varepsilon_g + \varepsilon_s$ (which possibly is not in the format $I\sigma^2$), the model is converted to

$\hat{V}^{-1/2}\hat{\beta} = \hat{V}^{-1/2}X_g\beta_g + \hat{V}^{-1/2}\varepsilon$. This conversion spherizes ε to $\hat{V}^{-1/2}\varepsilon$, then in the manner similar to the fixed effect model, though not rigorously, t-statistics and F-statistics can be derived for β_g [45], so be the normalized residual $\hat{u}$. If the intra-subject error ε_s follows the same distribution for every subject or diminishes when the number of intra-subject time points increases, as assumed in the SPM package [28], then the mixed-effect model reduces to the random-effect model [44, 45] and hence rigorously satisfies the conditions of the proposed method.

Group analysis of fMRI may encounter another issue by warping individual subject's image to a template space. If voxel values represent the average signals within voxel cubes, then values mapped from subjects whose corresponding regions are larger will have lower SNR, and vice versa. This smoothing effect introduced by different transformations mainly affects the intra-subject error term ε_s , not the inter-subject term ε_g , because the smoothing takes place within a subject, not across subjects. This effect will be reflected in the variance of ε_s , estimated from intra-subject time series, and taken account by the mixed effect model. If the variance of the intra-subject term is already much smaller than the inter-subject term, then its influence is limited. This phenomenon essentially is rooted in image voxelization. In continuous cases, warping images involves neither interpolation nor smoothing, and consequently dismisses this problem.

Among the many aforementioned factors, the mixed-effect model issue is intrinsic because it still stands even in continuous cases. Errors derived from discretization, such as interpolation errors, inaccurate composition of transformations, and the smoothing effect raised from individual warping of subject images, are not essential, because they do not exist given continuous images and continuous transformations. The nature of a method is determined by its properties in the continuous situation where its mathematical foundation lies. Our experiments, all carried out with voxelized images, evidence our significant improvement of consistency under spatial transformations.

We draw the following conclusions. The FDR defined in image spaces yields different q -values under nonlinear spatial transformations. With regional expansion or shrinkage, detection sensitivity may increase or decrease. The FDR defined in the signal domain produces q -values invariant under diffeomorphic transformations, as theoretically formulated in Section II-D. When the proposed method is applied to discrete grids, experiments with synthetic, semi-synthetic and real images show that it significantly reduces inconsistency across testing spaces. This novel definition of the FDR can be applied to many FDR variants [12, 15, 18, 20–22], and equips them with transformation invariance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Data collection and sharing in Section III-E was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work is supported by grants P41EB015922, U54EB020406, R01MH094343, and K01EB013633 from the National Institutes of Health (NIH).

We acknowledge the MR images distributed by the project "Neurodevelopmental Genomics: Trajectories of Complex Phenotypes". Support for the collection of the datasets was provided by grants RCMH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. All subjects in the project were recruited through the Center for Applied Genomics at the Children's Hospital in Philadelphia.

We also acknowledge data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

References

1. Rosenthal R. The file drawer problem and tolerance for null results. *Psychological Bulletin*. 1979; 86(3):638–641.
2. Bennett C, Miller M, Wolford G. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. *NeuroImage*. Jul.2009 47:S125.
3. Hochberg Y, Tamhane AC. Multiple comparison procedures. *Wiley Series in Probability and Statistics*. Sep.1987
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995; 57(1): 289–300.
5. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 2002; 64(3):479–498.
6. Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT. Automated Talairach atlas labels for functional brain mapping. *Hum Brain Mapp*. 2000; 10(3):120–131. [PubMed: 10912591]
7. Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*. Feb; 2008 39(3):1064–1080. [PubMed: 18037310]
8. Fonov V, Evans A, McKinstry R, Almlí C, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*. Jul.2009 47:S102.
9. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*. Apr; 2002 15(4):870–878. [PubMed: 11906227]
10. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*. 2001; 29(4):1165–1188.

11. Srikanth R, Casanova R, Laurienti PJ, Peiffer AM, Maldjian JA. Estimation of false discovery rates for wavelet-denoised statistical parametric maps. *Neuroimage*. Oct; 2006 33(1):72–84. [PubMed: 16919480]
12. Chen S, Wang C, Eberly LE, Caffo BS, Schwartz BS. Adaptive control of the false discovery rate in voxel-based morphometry. *Hum Brain Mapp*. Jul; 2009 30(7):2304–2311. [PubMed: 19034901]
13. Reiss PT, Schwartzman A, Lu F, Huang L, Proal E. Paradoxical results of adaptive false discovery rate procedures in neuroimaging studies. *NeuroImage*. Dec; 2012 63(4):1833–1840. [PubMed: 22842214]
14. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. Sep; 2006 93(3):491–507.
15. Schwartzman A, Dougherty RF, Lee J, Ghahremani D, Taylor JE. Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage*. Jan; 2009 44(1):71–82. [PubMed: 18547821]
16. Efron B. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*. Mar; 2004 99(465):96–104.
17. Pendse G, Borsook D, Becerra L. Enhanced false discovery rate using Gaussian mixture models for thresholding fMRI statistical maps. *NeuroImage*. Aug; 2009 47(1):231–261. [PubMed: 19269334]
18. Lawyer G, Ferkingstad E, Nesvag R, Varnas K, Agartz I. Local and covariate-modulated false discovery rates applied in neuroimaging. *NeuroImage*. Aug; 2009 47(1):213–219. [PubMed: 19341802]
19. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*. Dec; 2001 96(456):1151–1160.
20. Langers DRM, Jansen JFA, Backes WH. Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate. *Neuroimage*. Oct; 2007 38(1):43–56. [PubMed: 17825583]
21. Zhang C, Fan J, Yu T. Multiple testing via FDR for large-scale imaging data. *The Annals of Statistics*. Feb; 2011 39(1):613–642. [PubMed: 21643445]
22. Nguyen HD, McLachlan GJ, Cherbuin N, Janke AL. False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Trans Med Imaging*. Aug; 2014 33(8):1735–1748. [PubMed: 24816549]
23. Perone Pacifico M, Genovese C, Verdinelli I, Wasserman L. False discovery control for random fields. *Journal of the American Statistical Association*. Dec; 2004 99(468):1002–1014.
24. Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. Cluster-based analysis of FMRI data. *NeuroImage*. Nov; 2006 33(2):599–608. [PubMed: 16952467]
25. Benjamini Y, Heller R. False discovery rates for spatial signals. *Journal of the American Statistical Association*. Dec; 2007 102(480):1272–1281.
26. Chumbley JR, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage*. Jan; 2009 44(1):62–70. [PubMed: 18603449]
27. Chumbley J, Worsley K, Flandin G, Friston K. Topological FDR for neuroimaging. *Neuroimage*. Feb; 2010 49(4):3057–3064. [PubMed: 19944173]
28. Statistical parametric mapping (spm). [Online]. Available: <http://www.fil.ion.ucl.ac.uk/spm/>
29. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005 Aug; 2(8):e124. [PubMed: 16060722]
30. Carp J. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*. 2012; 63(1):289–300. [PubMed: 22796459]
31. Li J, Shi Y, Toga AW. Controlling false discovery rate in signal space for transformation-invariant thresholding of statistical maps. *Information Processing in Medical Imaging*. 2015:125–136. [PubMed: 26213450]
32. Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*. Sep; 1997 24(3):407–418.
33. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage*. 2011; 55(3):954–967. [PubMed: 21216294]

34. Cheverud JM. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*. Jul; 2001 87(1):52–58. [PubMed: 11678987]
35. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*. Aug; 2005 95(3):221–227. [PubMed: 16077740]
36. Kahan W. Pracniques: Further remarks on reducing truncation errors. *Commun ACM*. 1965; 8(1): 40.
37. Satterthwaite TD, Elliott MA, Ruparel K, Loughhead J, Prabhakaran K, Calkins ME, Hopson R, Jackson C, Keefe J, Riley M, et al. Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage*. Feb.2014 86:544–553. [PubMed: 23921101]
38. ANTS: <http://stnava.github.io/ants/>.
39. Giedd JN, Blumenthal J, Jeffries NO, Castellanos FX, Liu H, Zijdenbos A, Paus T, Evans AC, Rapoport JL. Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience*. Oct; 1999 2(10):861–863. [PubMed: 10491603]
40. Gogtay N, Giedd JN, Lusk L, Hayashi KM, Greenstein D, Vaituzis AC, Nugent TF, Herman DH, Clasen LS, Toga AW, et al. Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*. May; 2004 101(21): 8174–8179.
41. Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*. Nov; 1992 12(6):900–918. [PubMed: 1400644]
42. Worsley K, Andermann M, Koulis T, MacDonald D, Evans A. Detecting changes in nonisotropic images. *Hum Brain Mapp*. 1999; 8(2–3):98–101. [PubMed: 10524599]
43. Taylor JE, Worsley KJ. Random fields of multivariate test statistics, with applications to shape analysis. *Ann Statist*. Feb; 2008 36(1):1–27.
44. Friston K, Stephan K, Lund T, Morcom A, Kiebel S. Mixed-effects and fMRI studies. *NeuroImage*. Jan; 2005 24(1):244–252. [PubMed: 15588616]
45. Mumford J, Nichols T. Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*. Mar; 2006 25(2):42–51. [PubMed: 16568936]

Figure 1.

Problem of Interest: Transformation Invariance. Correction for multiple testing applied after a spatial transformation and then followed by the inverse transformation is unnecessarily equivalent to directly thresholding in the original space. Transformation-variant methods suffer from the bias introduced by the image space where multiple testing takes place.

Figure 2.

Transformation Invariance in Signal Space. (a) When an image is spatially transformed, some voxels will expand and some will shrink, so does the volumetric measure of positive and false positive regions. As a result, the FDR defined in the image space changes under spatial transformation. On the other hand, the mapped areas in the signal space do not change under diffeomorphic spatial transformation, so the FDR defined in the signal space is transformation invariant. (b) Green curves represent signals at four voxels a, b, c, and d; red curves on the purple sphere represent mapped area of the neighborhood (a,b,c,d) in the signal space. In the signal space, neighboring voxels with higher correlation will be placed closer than those with lower correlation, reflecting their lower degree of independence. The areas in the signal space reflect the “effective volume of randomness” held in a region. The more randomness a region experiences, and the more trustful positive results in it are, and so higher weights shall be assigned to the region.

Figure 3.

Numerical implementation. (a) Immediate neighbors of a voxel are denoted with 1-, 1+, 2-, 2+, 3- and 3+, respectively. $d-/d+$ (where $d \in \{1, 2, 3\}$) stands for the neighboring voxel by decreasing/increasing one index along the d -th dimension. Colors red, green and blue code the 1st, 2nd and 3rd dimensions respectively. (b) The normalized residual vector $u^\dagger$ (where $\dagger \in \{d-, d+ | d = 1, 2, 3\}$) of a neighboring voxel is mapped to the tangent plane at the central one u via spherical logarithmic mapping. The mapped tangent vector $v^\dagger$ has the same length as the great arc connecting u and $u^\dagger$, its direction is the initial direction from u to $u^\dagger$ along the great arc.

Figure 4.

Data of the experiment in Section III-A. FDR correction is performed in two spaces, A and B , as shown in subfigures (a) and (b) respectively. Subfigure (c) shows the signals, as presented in space A . The signals have the shape of a Gaussian distribution with radius 10.0, with the peak value normalized to 1.0. White noises with amplitude 1.0 are generated on discrete grids spaced by 1mm. The white noises are then smoothed with anisotropic diffusion to produce spatial inhomogeneity. The diffusion conductivity is shown in subfigure (d), and the diffusion time is 0.5. Subfigure (e) shows the noise amplitude after the anisotropic diffusion. Subfigure (f) shows average correlation coefficients with nearest neighbors.

Figure 5.

Results of the experiment in Section III-A, as shown in space A . The upper row and the lower row show results of the volFDR method and the sigFDR method, respectively. The left panel shows the difference between q -values obtained in the two image spaces, in the logarithmic scale. The right panel shows the $-\ln(q)$ map at the resolution of 0.25mm. Though the image resolution becomes finer and finer, the volFDR method does not decrease its inconsistency between the two image spaces. In contrast, the sigFDR method diminishes the inconsistency to almost zero. This clearly demonstrates that the sigFDR method as formulated with the continuous equations in Section II-D, is transformation invariant in its nature.

Figure 6. Ground truth images for the experiment in Section III-B. Signals with SNR ranges from 0.1 to 0.2 are added to the red highlighted regions.

Figure 7.

Image-wise performances of the experiment in Section III-B, evaluated in the fixed image space. Green and red curves respectively show the performance of the sigFDR and the volFDR methods. In the FDR and the Power plots, solid curves are for detection directly in the fixed image space, and dashed curves (except the black one) are for detection in the moving image space later warped back to the fixed image space. The dashed black curve is the target FDR value 5%. In the Dice plot, curves show the consistency between the detection taking place in the two image spaces.

Figure 8.

Voxel-wise performances of the experiment in Section III-B, at SNR=0.16. For definitions of the presented indices, please refer to Section III-B1. The power images of the volFDR method show considerable difference between the two image spaces, while the sigFDR results are much more consistent. The Dice images show clear evidence of the sigFDR method's improved consistency.

Figure 9.

Ground truth images for the experiment in Section III-C. Signals with SNR ranges from 0.1 to 0.2 are added to the red highlighted regions.

Figure 10.

Image-wise performances of the experiment in Section III-C, evaluated in the fixed image space. Green and red curves respectively show the performance of the sigFDR and the volFDR methods. In the FDR and the Power plots, solid curves are for detection directly in the fixed image space, and dashed curves (except the black one) are for detection in the moving image space later warped back to the fixed image space. The dashed black curve is the target FDR value 5%. In the Dice plot, curves show the consistency between the detection taking place in the two image spaces.

Figure 11.

Voxel-wise performances of the experiment in Section III-C, at SNR=0.15. For definitions of the presented indices, please refer to Section III-B1. The detection power of the volFDR decreases considerably in the moving image space, while the sigFDR method does not show significant changes. The sigFDR also achieves high voxel-wise consistency than the volFDR does.

Figure 12.

Experiment design with real 3D images. (1) Randomly sample a subset of data images in the template space Ω . (2) Warp the sampled data images to subjects' image spaces. (3) Perform FDR control in the subjects' image spaces and calculate q -values according to Eq. (3c). (4) Inverse-warp statistics of multiple testing back to space Ω . (5) Check consistency of the $-\ln(q)$ values inverse-warped back from difference spaces.

Figure 13.

Performance statistics on PNC real 3D images. All statistics are calculated across different spatial transformations and then averaged over random samples of signals. Subfigure (a) for the volFDR and subfigure (b) for the sigFDR show standard deviations s of $-\ln(q)$, with degrees of freedom equal 23. Subfigure (c) shows in logarithms the F-statistics in comparison of s_{volFDR} and s_{sigFDR} . Subfigure (d) shows in negative logarithms p -values of the F-statistics. Subfigure (e) for volFDR and subfigure (f) for sigFDR show expected values g of $-\ln(q)$. Subfigure (g) shows the contrast of g_{volFDR} and g_{sigFDR} . (h) shows the t -statistics of correlation between Jacobian determinants and age, with gender as the cofactor.

Figure 14.

Performance statistics on ADNI real 3D images. All statistics are calculated across different spatial transformations and then averaged over random samples of signals. (a) for the volFDR and (b) for the sigFDR show standard deviations s of $-\ln(q)$, with degrees of freedom equal 24. (c) shows in logarithms the F-statistics in comparison of s_{volFDR} and s_{sigFDR} . (d) shows in negative logarithms p -values of the F-statistics. (e) for volFDR and (f) for sigFDR show expected values g of $-\ln(q)$. (g) shows the contrast of g_{volFDR} and g_{sigFDR} . (h) shows the t -statistics about brain atrophy rate contrasting Alzheimer's Disease patients over normal elders, with age and gender as the cofactors.