Adjusting the Neuroimaging Statistical Inferences for Nonstationarity

Gholamreza Salimi-Khorshidi[1], Stephen M. Smith[1], and Thomas E. Nichols[1,2]

[1] Centre for Functional MRI of the Brain (FMRIB),
University of Oxford, Oxford, UK
reza@fmrib.ox.ac.uk
[2] GlaxoSmithKline Clinical Imaging Centre, London, UK

Abstract. In neuroimaging cluster-based inference has generally been found to be more powerful than voxel-wise inference [1]. However standard cluster-based methods assume stationarity (constant smoothness), while under nonstationarity clusters are larger in smooth regions just by chance, making false positive risk spatially variant. Hayasaka et al. [2] proposed a Random Field Theory (RFT) based nonstationarity adjustment for cluster inference and validated the method in terms of controlling the overall family-wise false positive rate. The RFT-based methods, however, have never been directly assessed in terms of homogeneity of local false positive risk. In this work we propose a new cluster size adjustment that accounts for local smoothness, based on local empirical cluster size distributions and a two-pass permutation method. We also propose a new approach to measure homogeneity of local false positive risk, and use this method to compare the RFT-based and our new empirical adjustment methods. We apply these techniques to both cluster-based and a related inference, threshold-free cluster enhancement (TFCE). Using simulated and real data we confirm the expected heterogeneity in false positive risk with unadjusted cluster inference but find that RFT-based adjustment does not fully eliminate heterogeneity; we also observe that our proposed empirical adjustment dramatically increases the homogeneity and TFCE inference is generally quite robust to nonstationarity.

1 Introduction

When detecting changes in functional or structural brain image data, it is necessary to have powerful inference methods that offer precise control of false positive risk. To assess the evidence of a change at each voxel of a statistic image, the two most common approaches are voxel- and cluster-based inferences. While voxel-wise methods use a single threshold to classify signals as *real*, cluster-based inference defines clusters as contiguous voxels whose intensity exceeds a predefined cluster-forming threshold u_c, and then detects signals based on the spatial extent of a cluster. Cluster-based inference is known to have a higher sensitivity compared to voxel-intensity-based tests when the signal is spatially extended.

Cluster-size tests have been widely used under different implementations [1,3], including simulation-based tests [4,5], random field theory (RFT-based) tests [6],

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part I, LNCS 5761, pp. 992–999, 2009.

and permutation tests [7,8]. However most of these proceedures are based on a stationarity assumption, that the spatial autocorrelation function is shift-invariant. When stationarity assumption is violated, the sensitivity and specificity of the test depend on local smoothness of the image, which justifies the use of adjusted cluster size which measures cluster size relative to local smoothness [6,2]. Independent of the stationarity issue, cluster-based inference also is limited by the arbitrariness of its important u_c parameter and the amount of pre-smoothing. To address these problems, threshold-free cluster enhancement (TFCE) was introduced [9], which in essence integrates out the u_c parameter while produces an image of local cluster-like evidence of a signal and was shown to generally have better detection power while being less sensitive to the amount of smoothing used [9,10].

In this work we propose a new adjustment for nonstationarity based on the local empirical distribution of cluster size in a two-pass permutation method. We evaluate this new approach, in the context of both standard cluster-based and TFCE inferences. We compare the impact of using no adjustment, RFT-based adjustment, and our proposed empirical adjustment under various simulated and real data with spatially-varying smoothness.

2 Method

A fitted general linear model at a voxel i has residuals

$$\hat{\epsilon}_i = Y_i - X\hat{\beta}_i \tag{1}$$

where Y_i is the observed intensity (M×1), $\hat{\beta}_v$ is the parameter vector (P×1), X is the design matrix (M×P), and $\hat{\epsilon}_v$ is the residual error (M×1), the estimated residuals of the linear model fit ($\hat{\epsilon}$) can be used to yield a smoothness estimation. This estimator explains the spatial correlation structure of the SPM (statistical parameter map) by assuming that it can be modeled as being due to a convolution of the signal with a Gaussian filter with an unknown, but determinable, width (σ), which is used to estimate the effective resolution of the data.

2.1 Kiebel et al.'s Method

Using the RFT concepts [11], an unbiased estimator for the covariance of the partial derivatives (at direction j) at voxel i in a D-dimensional Gaussian random field is calculated as

$$\lambda_{i,j} = \frac{\nu - 2}{\nu - 1} \cdot \frac{1}{M} \sum_{t=1}^{M} \left(\frac{\partial S_{it}}{\partial x_j} \right)^2 \tag{2}$$

where S is the standardized error, ν is the number of degrees of freedom, and M is the number of observations (time points or subjects), which yields a voxel-wise estimate of smoothness as

$$RESEL_i = \prod_{j=1}^{D} (8 \cdot \ln(2))^{1/2} \, \sigma_{i,j} \tag{3}$$

where $\sigma_{i,j} = (2\lambda_{i,j})^{-1/2}$, and $RESEL_i$ is the volume of a resolution element (resel) at voxel i (note that $1/RESEL$ is resels per voxel or RPV). As an alternative, a more robust estimate can be defined using just the control group observations or using all the observations after excluding any outlier (in terms of their smoothness estimates) observations.

2.2 Jenkinson et al.'s Method

As the smoothness extent decreases Kiebel's estimator becomes increasingly inaccurate. An alternative estimate can use the autocorrelation of the standard error at voxel i (S_i^2) and its cross-correlation with neighboring voxels (i.e., $SS_{i,j}$ for the next voxel in direction j) with a Gaussian autocorrelation function assumption [12,13]. A voxel-wise smoothness can be estimated using the σ from

$$\sigma_{i,j}^2 = \left(4 \cdot \ln\left(\frac{S_i^2}{SS_{i,j}}\right)\right)^{-1} \tag{4}$$

with the rest of the operation as before (which can also result in a robust estimate as in Kiebel's).

2.3 Empirical Cluster-Size Normalization

In FSL's *randomise* analysis[1], after N_P permutations, let $N_v \leq N_P$ be the number of permutations that clusters with sizes $S_1(v)$, $S_2(v)$, ... $S_{N_v}(v)$ hit the voxel v. The empirical cluster size per voxel (ECSPV) for this voxel is calculated as

$$ECSPV(v) = \left(\frac{\sum_{i=1}^{N_v} S_i(v)^E}{N_v}\right)^{1/E} \tag{5}$$

where E is the cluster size histogram's normalization parameter. Having ECSPV from the first run, the adjustment in the second run can either be voxel-wise

$$S_C^{vn} = \sum_{v \in C} \frac{1}{ECSPV(v)} \tag{6}$$

or cluster-wise

$$S_C^{cn} = \frac{S_C}{\sum_{v \in C} ECSPV(v)}. \tag{7}$$

Cluster-based inference using these normalized statistics (S_C^{vn} or S_C^{cn}) is expected to be adjusted for nonstationarity (estimating the smoothness/roughness of each area by using the cluster sizes hitting each voxel at different permutations under the null hypothesis).

[1] see http://www.fmrib.ox.ac.uk/fsl/randomise/

2.4 Empirical TFCE Normalization

TFCE is a tool developed based on the idea of cluster size accumulation on a range of possible cluster-forming thresholds [9]. In order to adjust the TFCE statistic for nonstationarity, either its corresponding cluster sizes can be adjusted or TFCE scores at each voxel can be empirically normalized by

$$\begin{aligned} ETPV(v) &= \frac{\sum_{i=1}^{N_v} TFCE(v)}{N_v} \\ TFCE^N(v) &= \frac{TFCE(v)}{ETPV(v)} \end{aligned} \tag{8}$$

where N_v is the number of permutations a voxel has a nonzero TFCE score and ETPV is the empirical TFCE per voxel.

2.5 Nonstationarity Assessment

Using null data, spatial variation in cluster-related inference's false positive rate (P-value) is used to assess different methods' performance in correction for nonstationarity. In case of using a stationary null data for a statistical inference, the output P-value volume should follow a uniform distribution ($U(0,1)$ with the mean of 0.5) at each voxel. Since $-\ln$ P-values are easier to visualize, note that if X has a uniform distribution, $-\ln(X)$ has an exponential distribution with parameter $\lambda = 1$, so $E[-\ln(X)] = \frac{1}{\lambda} = 1$, and $Var[-\ln(X)] = \frac{1}{\lambda^2} = 1$, and hence $E[-\log_{10}(X)] = \frac{1}{\ln(10)} = 0.4343$, and $Var[-\log_{10}(X)] = 0.4343$.

Thus, the deviation of an inference from this expected distribution can be a good indicator of the existence of nonstationarity in the image. Three statistical indicators (mean, standard deviation or SD, and coefficient of variation or CV) are employed to assess this deviation for each adjustment technique. To implement this idea, a group of permutations in the *first run* result in a distribution of cluster-related statistic. Then, for the same group of permutations in the *second run*, the resulting cluster-related statistic image is converted to a P-value volume with respect to the distribution from the first run. This P-value volume is then converted to $-\log_{10}(P)$ volume. Averaging these -$\log_{10}(P)$ volumes across all the permutations results in a single -$\log_{10}(P)$ volume whose mean, SD and CV indicates the extent of nonstationarity in the data. Note that, as a result of the averaging of $-\log_{10}$(P) images, the mean can be compared with the expected mean (0.4343), however the SD should be smaller than the expected SD (0.4343).

2.6 Data

To assess different methods' performance, both simulated and real data are tested. For the stationary null data simulation, two groups of 150×150×150 Gaussian noise ($\sim N(0,1)$) images (with 20 images in each group) are generated and smoothed with a Gaussian smoothing kernel (with σ=2, 3, 4 and 5 voxels). To avoid the nonstationarity at the edge, the outer 30 voxels are excluded. To simulate the nonstationary data, two groups of 150×150×150 white noise images (with 20 images in each group) are first smoothed with three different 3D Gaussian kernels (representing the low, medium, and high smoothness extents). These images are combined in a way that an outer layer smoothed with σ_1 encloses a middle layer smoothed with σ_2, which encircles a core smoothed with σ_3 (referred to as $\sigma_1\sigma_2\sigma_3$). The center core is 30×30×30 voxels, centered within a 60×60×60 voxel middle layer, which itself was centered in a 90×90×90 volume. The combined image is smoothed again with a 3D Gaussian filter with σ =1.5 voxels (to eliminate discontinuities at the borders), and just as in the stationary case, to avoid the nonstationarity at the edges, outer 30 voxel are excluded.

To assess each method's performance on real data, null fMRI and VBM datasets are also fed into the analysis. The fMRI dataset is a pain study with 16 healthy subjects. The first-level analysis of the data using FSL includes motion correction and spatial smoothing (FWHM (full width half maximum) = 2, 4, 5, 7 and 10mm) prior to temporal model fitting (including autocorrelation correction), and a two-step registration to the MNI152 standard brain space. The null VBM dataset includes structural gray matter images of 35 healthy control subjects smoothed with different Gaussian kernel sizes (σ =2, 3, 4 and 5mm). Dividing these subjects into two groups results in a null data analysis (with no expected difference). The final assessment of the methods' performance is on real VBM data with three groups of subjects: 46 controls, 50 Alzheimer's disease (AD) and 57 mild cognitive impairment (MCI).

3 Results

Fig. 1 illustrates statistical indicators (mean, SD and CV of $-\log_{10}(P)$ volumes) of the previously mentioned analyses on both simulated and real data. Kiebel's and Jenkinson's method have very similar results, which is why only Jenkinson's results (referred to as RPV1) are shown (RPV0 refers to no adjustment). Also, results for $E = 1$ and 2 are not shown as $E = 2/3$ shows a better adjustment; *vw* and *cw* refer to voxel-wise and cluster-wise normalization. Fig. 2 and 3 illustrate the effect of different adjustment techniques in localizing the differences between two groups of subjects in real data, by assessing the change in P-value as a function of clusters' unadjusted P-value and average FWHM (see figure captions for details).

4 Discussion

Using cluster-size adjustment techniques, cluster sizes depend on both their connected component's size and the local roughness of their region. The results show that there is a substantial spatial variation in cluster-based inference's FPR that cannot be completely corrected. Using the uniformity indices, adjustments improve the homogeneity of the result image. According to the results, empirical adjustment causes a reduction of sensitivity, which may also be an indication of increased power of the inference. The optimal empirical adjustment is seen at

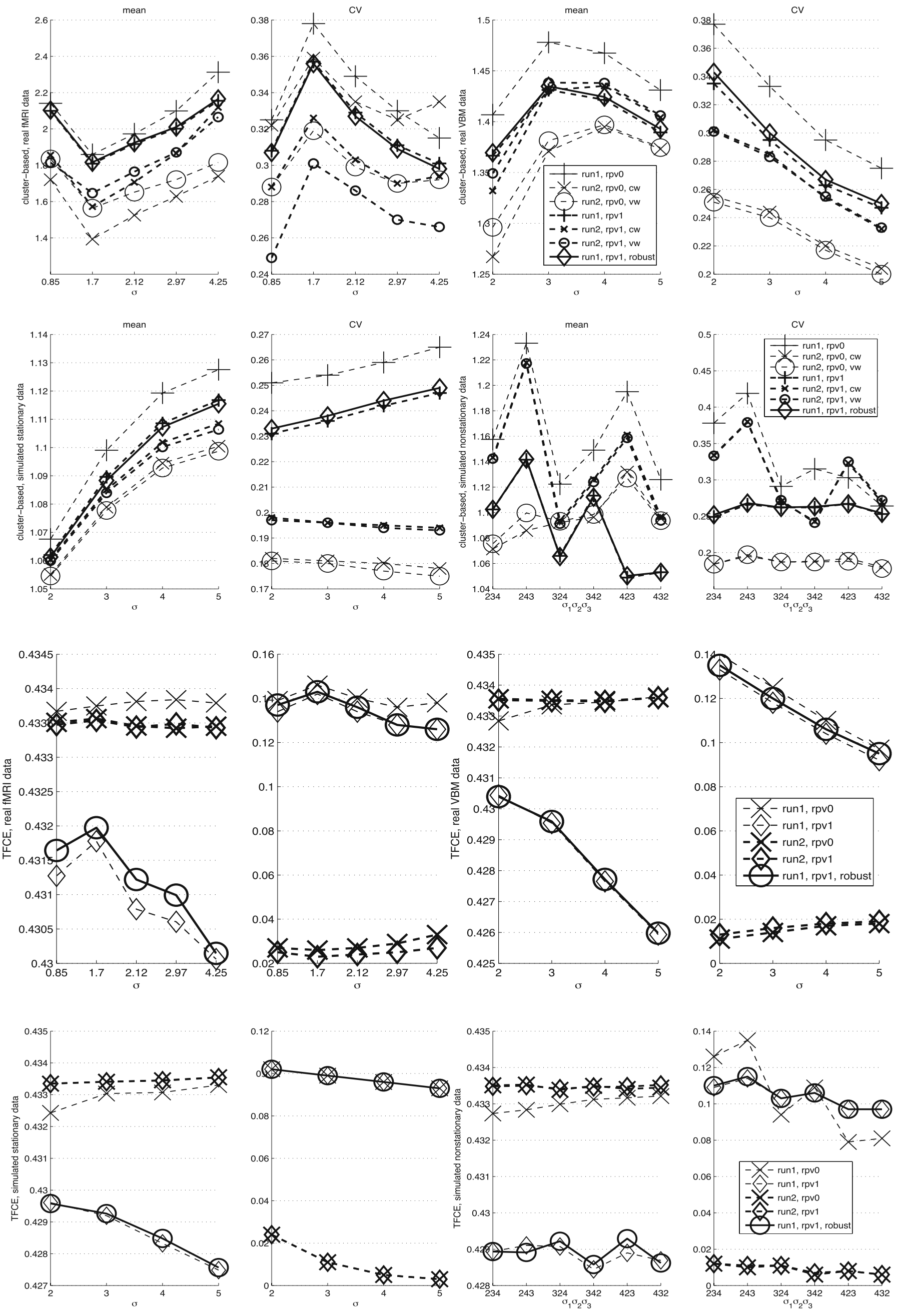

Fig. 1. The spatial mean and CV of -$\log_{10}(p)$ as a function of smoothing extent. These results show how adjustment can improve the stationarity of the cluster-related inferences and TFCE's robustness to nonstationarity in the data. Note that the legend in the first (third) row can be used for the third (fourth) row.

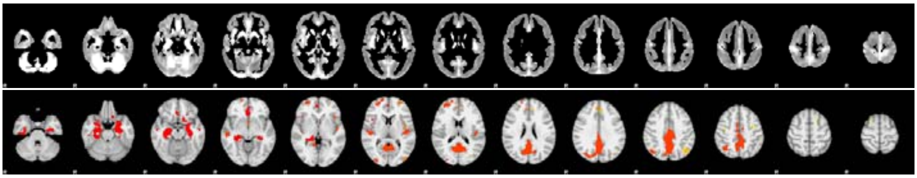

Fig. 2. Smoothness map corresponding the three-group VBM data (top image) and the resulting clusters after thresholding the T-statistic map at $T = 4$ (for the MCI-AD contrast). Smoothness estimates and clusters formed in this images are the basis of the analyses and the results sin Fig. 3. Note that the displayed slices are selected from z=-32mm to z=64mm including every eighth millimeters (in MNI coordinates).

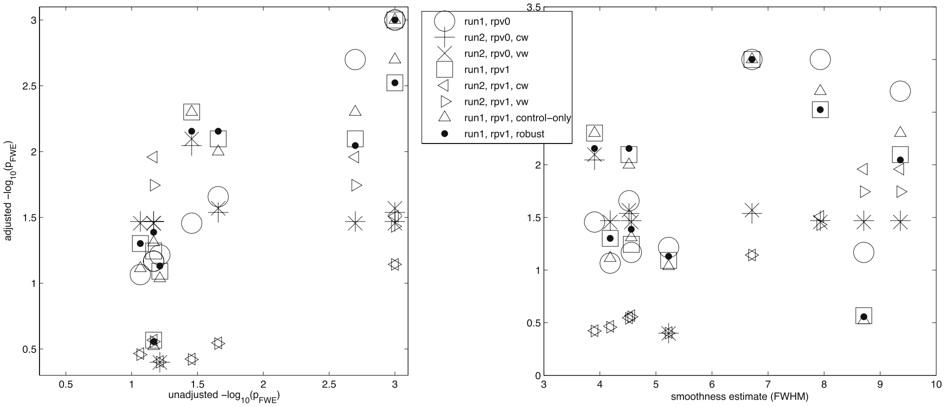

Fig. 3. The effect of the adjustment on significance level of clusters as a function of its FWHM (right column) and unadjusted P-value (left column). This figure illustrates the increase in sensitivity (specificity) in rough (smooth) regions after appropriate adjustments. Note that the clusters in this figure are shown in Fig. 2, which will remain the same for all the adjustment analyses. According to the right column, the expected variation of the significance (increase/decrease in rough/smooth regions) can be observed in empirical adjustment when no RFT-based adjustment is present, and RFT-based adjustment without empirical adjustment. Also, according to the left column, adjusting the cluster-based inference for nonstationarity seems to reduce the significance-level of the clusters, which agrees with the result in the top row to some extent (as more clusters are to be formed in smooth regions, where adjustment is expected to result in a significance decrease.

$E = 2/3$ and voxel-wise normalization in ECSPV calculation. The empirical adjustment is not recommended for adjusted cluster-sizes as the second correction applied to a unified field, may be similar to using $CS = \sum_{v \in C} 1$, which is the classic unadjusted cluster size. Note that, the use of cluster sizes to extract a voxel-wise characteristic (i.e., ECSPV) can be an imprecise estimate, because of the censoring (the P-value corresponding to a cluster, is not a precise voxel-wise

P-value). On the other hand, TFCE inference is very robust with respect to spatial variations in image smoothness. In both adjusted and unadjusted TFCE inferences, the summary measure of performance, perfectly matches the expected measure of a uniform image at all of the tested smoothing extents.

References

1. Friston, K., Holmes, A., Poline, J., Price, C., Frith, C.: Detecting activations in pet and fmri: levels of inference and power. NeuroImage 4, 223–235 (1996)
2. Hayasaka, S., Phan, K., Liberzon, I., Worsley, K., Nichols, T.: Nonstationary cluster-size inference with random field and permutation methods. Neuroimage 22(2), 676–687 (2004)
3. Poline, J., Worsley, K., Evans, A., Friston, K.: Combining spatial extent and peak intensity to test for activations in functional imaging. Neuroimage 5(2), 83–96 (1997)
4. Forman, S., Cohen, J., Fitzgerald, J., Eddy, W., Mintun, M., Noll, D.: Improved assessment of significant activation in functional magnetic resonance imaging (fmri): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647 (1995)
5. Ledberg, A., Åkerman, S., Roland, P.: Estimation of the probability of 3d clusters in functional brain images. Neuroimage 8, 113–128 (1998)
6. Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., Evans, A.C.: A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp 4, 58–73 (1996)
7. Holmes, A., Blair, R., Watson, J., Ford, I.: Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16(1), 7–22 (1996)
8. Nichols, T., Holmes, A.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain. Mapp. 15(1), 1–25 (2002)
9. Smith, S.M., Nichols, T.E.: Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44(1), 83–98 (2009)
10. Smith, S., Douaud, G., Salimi-Khorshidi, G., Webster, M., Mackay, C., Groves, A., Nichols, T.: Threshold-free cluster enhancement: Practical examples. In: 14th Anuual Meeting of the Organization for Human Brain Mapping, Melbourne, Australia (2008)
11. Kiebel, S., Poline, J., Friston, K., Holmes, A., Worsley, K.: Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. Neuroimage 10(6), 756–766 (1999)
12. Flitney, D., Jenkinson, M.: Cluster analysis revisited. Technical report, FMRIB Centre, University of Oxford (2000)
13. Nichols, T.: Cluster analysis revisited - again: Implementing nonstationary cluster size inference. Technical report, FMRIB Centre, University of Oxford (2008)