EMPIRICALLY INVESTIGATING THE STATISTICAL VALIDITY OF SPM, FSL AND AFNI FOR SINGLE SUBJECT FMRI ANALYSIS

Anders Eklund^{a,b,c}, Thomas Nichols^d, Mats Andersson^{a,c}, Hans Knutsson^{a,c}

^aDepartment of Biomedical Engineering, Linköping University, Sweden

^bDepartment of Computer and Information Science, Linköping University, Sweden

^cCenter for Medical Image Science and Visualization (CMIV), Linköping University, Sweden

^dDepartment of Statistics, University of Warwick, Coventry, United Kingdom

ABSTRACT

The software packages SPM, FSL and AFNI are the most widely used packages for the analysis of functional magnetic resonance imaging (fMRI) data. Despite this fact, the validity of the statistical methods has only been tested using simulated data. By analyzing resting state fMRI data (which should not contain specific forms of brain activity) from 396 healthy controls, we here show that all three software packages give inflated false positive rates (4%-96% compared to the expected 5%). We isolate the sources of these problems and find that SPM mainly suffers from a too simple noise model, while FSL underestimates the spatial smoothness. These results highlight the need of validating the statistical methods being used for fMRI.

1. INTRODUCTION

Analysis of functional magnetic resonance imaging (fMRI) data is typically performed using parametric statistical methods, which are based on several assumptions. The assumptions are often only tested using simulated data [1], but it is extremely hard to simulate a brain and an MR scanner. In our previous work [2], real fMRI data were used to show that single subject fMRI analysis using the SPM [3, 4] software package can lead to inflated false positive rates (e.g. 70% compared to the expected 5%). The main reason for the high degree of false positives was found to be that the SPM software uses a rather simple model of the errors in the general linear model (GLM). As other fMRI software packages, e.g. FSL [5] and AFNI [6], use more advanced noise models, here we extend our investigation to see if these software packages give false positive rates that are more accurate.

2. METHODS

Resting state fMRI data from 396 healthy controls were downloaded from the homepage of the 1000 functional con-

nectomes project [7] (<http://fcon.1000.projects.nitrc.org/fcpClassic/FcpTable.html>), and are summarized in Table 2. Each rest dataset was analyzed with several different activity paradigms (see Table 1), as it should be impossible to find significant activity in rest data. For 100 rest (null) datasets and a familywise error corrected significance threshold of 5%, significant activity is on average expected in 5 of the datasets. The number of significant activations, divided by the number of analyzed subjects, can thus be seen as an estimate of the familywise false positive rate. In this paper the focus has been on cluster level inference [8], as it is more commonly used than voxel level inference. Two common cluster defining thresholds [9, 10] were tested; $p = 0.01$ ($z = 2.3$) and $p = 0.001$ ($z = 3.1$).

Table 1. Length of activity and rest periods for the activity paradigms, R stands for randomized.

Paradigm	Activity periods (s)	Rest periods (s)
B1	10	10
B2	30	30
E1	2	6
E2	1-4 (R)	3-6 (R)

The parameters used for each software package are given in Table 3. The ambition was to use the default settings for each software package, with the exception that motion regressors have been used for all packages (motion regressors are default in AFNI, but not in SPM or FSL). All the processing scripts are available at <https://github.com/wanderine/ParametricSinglesubjectfMRI> to facilitate further investigations.

2.1. SPM

According to [9], SPM2 is the most common SPM version used since 2007, but SPM8 (with the latest updates) was used

Table 2. fMRI data used for estimating false positive rates for SPM, FSL and AFNI. For each subject there is one anatomical T_1 -weighted volume and one resting state scan.

Study	# Subjects	Age (years)	Voxel size (mm)	Repetition time (s)	# Timepoints
Beijing	198	18-26 (21.2 ± 1.8)	$3.13 \times 3.13 \times 3.6$	2.0	225
Cambridge	198	18-30 (21.0 ± 2.3)	$3.0 \times 3.0 \times 3.0$	3.0	119

Table 3. Settings used for the different software packages.

Parameter / Software	SPM 8	FSL 5	AFNI
Slice timing correction	None	None	None
Motion correction	Realign	mcfliirt	3dvolreg
Normalization	'Segment', default parameters	Affine, 12 parameters	Affine, 12 parameters
Normalization template	MNI	MNI 152, 2 mm	Talairach
fMRI-T1 registration	Linear	Linear	Linear
Smoothing	4 - 10 mm FWHM	4 - 10 mm FWHM	4 - 10 mm FWHM
Motion regressors	Yes, 6	Yes, 6	Yes, 6
High-pass filter cutoff	128 s	B1 20 s, B2 60 s, E1, E2 100 s	Detrending
Noise modelling	Global AR(1)	FILM prewhitening [11]	Voxel-specific ARMA(1,1)
Cluster defining threshold	$p = 0.01$, $p = 0.001$	$p = 0.01$, $p = 0.001$	$p = 0.01$, $p = 0.001$

for this study, as it is more likely to be used in future studies. Analyses were performed using a Matlab batch script.

2.2. FSL

Analyses in FSL 5.0.7 were performed by using a processing configuration (design.fsf) generated by the FEAT GUI. The default settings were used, except for adding motion parameters to the design matrix. The cutoff of the highpass filter was automatically changed by the GUI from the default 100 seconds, to 20 seconds for the fast block based design (B1) and to 60 seconds for the slow block based design (B2).

2.3. AFNI

Analyses in AFNI were performed using the standardized python script *afni_proc.py*, setup through the master script *uber_subject.py*. The default settings were used, except for running the statistical analysis with 3dREMLfit which uses a voxel-specific ARMA(1,1) model for the GLM residuals (the default function 3dDeconvolve assumes that the residuals are independent).

3. RESULTS

Smoothness estimates (in the x-direction) for the Beijing data sets are given in Figure 1, as cluster based thresholding strongly depends on these estimates. Estimated familywise error rates are given in Figure 2. Estimated power spectra of the whitened GLM residuals for the Beijing data sets are given in Figure 3, to see how well the different noise models

perform. Prior to estimating power spectra, each residual time series was standardized to have a variance of 1 (only time series within the brain were used). Probability distributions of the estimated t-values are given in Figure 4. Proportions of estimated t-values being larger than different cluster defining thresholds are given in Table 4. See the github repository for details about the calculations.

Fig. 1. Smoothness estimates for the Beijing datasets, for SPM, FSL and AFNI. A Gaussian smoothing kernel of size 6 mm FWHM (full width at half maximum) was applied during the preprocessing of each dataset.

Fig. 2. Estimated familywise error rates for 4-10 mm of smoothing and four different activity paradigms (B1, B2, E1, E2), for SPM, FSL and AFNI. Each activity map was first thresholded using a voxel-wise threshold ($p = 0.01$ or $p = 0.001$, uncorrected for multiple comparisons) and then the surviving clusters (if any) were compared to a cluster extent threshold corresponding to $p = 0.05$ (corrected for multiple comparisons). The estimated familywise error rates are simply the number subjects with significant activation, divided by the number of analyzed subjects. **Top:** results for Cambridge data, **Bottom:** results for Beijing data, **Left:** results for a cluster defining threshold of $p = 0.01$ ($z = 2.3$), **Right:** results for a cluster defining threshold of $p = 0.001$ ($z = 3.1$). Note that the default amount of smoothing is 8 mm in SPM, 5 mm in FSL and 4 mm in AFNI.

Fig. 3. Power spectra of the GLM residuals (averaged over all brain voxels for the 198 Beijing data sets), for SPM, FSL and AFNI. The dip at 0.05 Hz is from the activity paradigm B1 (a square wave with a period of 20 seconds). If the residuals are uncorrelated, the power spectra will be flat.

4. DISCUSSION

The fMRI software packages SPM, FSL and AFNI have been tested in terms of statistical validity for single subject analysis. It is clear that all three software packages give unreliable results, and that the more advanced noise models in FSL and AFNI are not sufficient to lower the false positive rates to expected values. While SPM mainly suffers from a too simple noise model (Figure 3), FSL instead seems to underestimate the spatial smoothness (Figure 1).

It may seem strange that AFNI also gives unreliable results, since AFNI does not rely on Gaussian random field theory (GRFT) [12, 8] to correct for multiple comparisons (like SPM and FSL do). Instead, AFNI uses the function 3dClustSim to perform a Monte Carlo simulation of how likely it is to find clusters of a certain size in white noise. The main assumptions of 3dClustSim are, however, the same as for GRFT. Both methods assume perfect Gaussian data and a constant smoothness (which needs to be estimated from the data). 3dClustSim can give better results for low amounts of smoothness, as GRFT also requires that the activity map is sufficiently smooth. As Table 4 shows that the t-values from AFNI closely follow a true t-distribution, it is likely that a non-uniform spatial smoothness is the main problem. A non-uniform smoothness is more problematic for lower cluster defining thresholds, which is reflected in the estimated false positive rates and elsewhere [10]. Non-uniform smoothness has previously been shown to be a problem in voxel-based morphometry [13, 14].

It is disappointing that the standard software packages still give unreliable statistical results, more than 20 years after the initial fMRI experiments. Considering that non-parametric

Fig. 4. Log probability distributions of the estimated t-values (pooled over all brain voxels for the 198 Beijing data sets analyzed with paradigm B1), for SPM, FSL and AFNI. All three software packages generate t-values with a heavier tail compared to a true t-distribution. Note that the distributions for SPM and AFNI are below the true t-distribution for t-values smaller than 2.5 (not possible to see in this graph).

methods have become more flexible [15], that multivariate statistical tests have more complicated null distributions than univariate tests [16, 17, 18] and that more computing power is available [19, 20], it may be time for the fMRI community to consider non-parametric statistical methods. A permutation test, for example, does not assume Gaussian data, a constant noise variance or a constant smoothness (and the smoothness does not need to be estimated from the data). Permutation tests are easy to use for group analyses, but single subject fMRI data contain temporal auto correlation, making it harder to permute the volumes. Nevertheless, permutation tests show promising results also for single subject fMRI analysis [2, 21].

Table 4. Proportion of t-values being larger than different cluster defining thresholds (for Beijing data sets analyzed with paradigm B1). The expected proportions are 1%, 0.1% and 0.0001%, respectively.

t-value	SPM	FSL	AFNI
2.34	1.08%	2.12%	0.99%
3.13	0.16%	0.37%	0.11%
4.89	0.00126%	0.00282%	0.00053%

5. ACKNOWLEDGEMENT

This work was supported by the Neuroeconomic research initiative at Linköping University. The authors would like to thank the NITRC for sharing the fMRI data.

6. REFERENCES

- [1] M. Welsaert and Y. Rosseel, "A review of fMRI simulation studies," *PLoS ONE*, vol. 9, pp. e101953, 2014.
- [2] A. Eklund, M. Andersson, C. Josephson, M. Johanneson, and H. Knutsson, "Does parametric fMRI analysis with SPM yield valid results? - An empirical study of 1484 rest datasets," *NeuroImage*, vol. 61, pp. 565–578, 2012.
- [3] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, *Statistical Parametric Mapping: the Analysis of Functional Brain Images*, Elsevier/Academic Press, 2007.
- [4] J. Ashburner, "SPM: a history," *NeuroImage*, vol. 62, pp. 791–800, 2012.
- [5] M. Jenkinson, C. Beckmann, T. Behrens, M. Woolrich, and S. Smith, "FSL," *NeuroImage*, vol. 62, pp. 782–790, 2012.
- [6] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, vol. 29, pp. 162–173, 1996.
- [7] B. Biswal, M. Mennes, X. Zuo ..., and M. Milham, "Toward discovery science of human brain function," *PNAS*, vol. 107, pp. 4734–4739, 2010.
- [8] K. J. Friston, K. J. Worsley, R. S. J. Frackowiak, J. C. Mazziotta, and A. C. Evans, "Assessing the significance of focal activations using their spatial extent," *Human Brain Mapping*, vol. 1, pp. 210–220, 1994.
- [9] J. Carp, "The secret lives of experiments: Methods reporting in the fMRI literature," *NeuroImage*, vol. 63, pp. 289–300, 2012.
- [10] C. Woo, A. Krishnan, and T. Wager, "Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations," *NeuroImage*, vol. 91, pp. 412 – 419, 2014.
- [11] M. Woolrich, B. Ripley, M. Brady, and S. Smith, "Temporal autocorrelation in univariate linear modeling of FMRI data," *NeuroImage*, vol. 14, pp. 1370–1386, 2001.
- [12] K. Worsley, A. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for CBF activation studies in human brain," *Journal of cerebral blood flow and metabolism*, vol. 12, pp. 900–918, 1992.
- [13] M. Silver, G. Montana, and T. Nichols, "False positives in neuroimaging genetics using voxel-based morphometry data," *NeuroImage*, vol. 54, pp. 992–1000, 2011.
- [14] C. Scarpazza, G. Sartori, M. de Simone, and A. Mechelli, "When the single matters more than the group: very high false positive rates in single case voxel based morphometry," *NeuroImage*, vol. 70, pp. 175–188, 2013.
- [15] A. Winkler, G. Ridgway, M. Webster, S. Smith, and T. Nichols, "Permutation inference for the general linear model," *NeuroImage*, vol. 92, pp. 381–397, 2014.
- [16] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Adaptive analysis of fMRI data," *NeuroImage*, vol. 19, pp. 837–845, 2003.
- [17] M. Jin, R. Nandy, T. Curran, and D. Cordes, "Extending local canonical correlation analysis to handle general linear contrasts for fMRI data," *International journal of biomedical imaging, Article ID 574971*, vol. 2012, 2012.
- [18] J. Stelzer, Y. Chen, and R. Turner, "Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control," *NeuroImage*, vol. 65, pp. 69 – 82, 2013.
- [19] A. Eklund, P. Dufort, D. Forsberg, and S. LaConte, "Medical image processing on the GPU - Past, present and future," *Medical Image Analysis*, vol. 17, pp. 1073–1094, 2013.
- [20] A. Eklund, P. Dufort, M. Villani, and S. LaConte, "BROCCOLI: Software for fast fMRI analysis on many-core CPUs and GPUs," *Frontiers in Neuroinformatics*, vol. 8:24, 2014.
- [21] D. Adolf, S. Weston, S. Baecke, M. Luchtman, J. Bernarding, and S. Kropf, "Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method," *Frontiers in Neuroinformatics*, vol. 8:72, 2014.