Adaptive thresholding for reliable topological inference in single subject fMRI analysis

Krzysztof J. Gorgolewski K.J.GORGOLEWSKI@SMS.ED.AC.UK
Neuroinformatics Doctoral Training Centre, University of Edinburgh, UK

Amos J. Storkey A.STORKEY@ED.AC.UK
Institute for Adaptive and Neural Computation, University of Edinburgh, UK

Mark E. Bastin MARK.BASTIN@ED.AC.UK
Health Sciences (Medical Physics), University of Edinburgh, UK

Cyril Pernet CYRIL.PERNET@ED.AC.UK
Brain Research Imaging Centre, a SINAPSE Collaboration centre, University of Edinburgh, UK

Abstract

Single subject functional Magnetic Resonance Imaging (fMRI) has proved to be a useful tool for mapping functional areas in clinical procedures such as tumour resection. Using fMRI data, clinicians assess the risk, plan and execute such procedures based on thresholded statistical maps. However, because current thresholding methods were developed mainly in the context of cognitive neuroscience group studies, most single subject fMRI maps are thresholded manually to satisfy specific criteria related to single subject analyses. Here, we propose a new adaptive thresholding method which combines Gamma-Gaussian mixture modelling with topological thresholding to improve cluster delineation. In a series of simulations we show that by adapting to the signal and noise properties, the new method performs well in terms of over and underestimation of the true activation border. We also show through simulations and a motor test-retest study on ten volunteer subjects that adaptive thresholding improves reliability, mainly by accounting for the global signal variance. This in turn increases the likelihood that the true activation pattern can be determined.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

1. Introduction

The final outcome from fMRI analyses is a map showing which areas are most likely involved in certain sensori-motor or cognitive skills. After appropriate data pre-processing, a General Linear Model (GLM) is fitted to the measured signal and a T-test looking for differences between conditions or between a given condition versus rest is performed. The result is a 3D volume of T-values. Given these T-values, each voxel is labelled as being active (involved in the task) or not-active (not involved in the task) based on an *ad-hoc* threshold. This procedure has been successfully used in the context of cognitive neuroscience group studies for population inference. However, three major problems need to be addressed in order to improve inference at the subject level when used for clinical decision making, namely: (i) the impact of signal-to-noise ratio (SNR) on thresholding, (ii) the relative importance of Type I versus Type II error rates, and (iii) the spatial accuracy of the thresholded maps. In this paper we investigate how these issues affect statistical maps and describe a new adaptive thresholding method which improves cluster delineation.

The aim of our approach is to perform inference on the cluster level and at the same time provide a good balance between false positive and negative errors in the delineation of activation borders. We therefore propose a Gamma-Gaussian mixture model as a method to account for a distributions of T-values in Statistical Parametric Maps (SPM) (Woolrich et al., 2005) and set a threshold specific to the data at hand (Pendse et al., 2009). A natural way to determine this thresh-

old is to take the crossing point between the Gaussian, the model corresponding to no activation, and the Gamma distribution, the model corresponding to positive activations. This crossing point thus separates signal from noise, and consequently provides a good trade-off between false positive and negative rates. Finally, once this threshold is established, topological inference via False Discovery Rate (FDR) correction over clusters (Chumbley & Friston, 2009) is used to correct for the number of tests performed while accounting for spatial dependencies across voxels, thereby explicitly controlling for Type I cluster rate. This heuristic approach combines advantages of the different methods mentioned above. Specifically it relies on a simple model of the SPM, allows adaptive thresholding, and accounts for multiple comparisons in the context of topological inference.

2. Methods and Materials

2.1. Gamma-Gaussian mixture model

Following Woolrich et al. (2005), the T-value distribution from a SPM is modelled using a Gamma-Gaussian mixture model, with the Gaussian distribution as a model for the null distribution (no activation) and Gamma distributions as models for the negative (deactivation) and positive (activation) distributions. Note that due to high degrees of freedom in a typical fMRI experiment, i.e. the number of time points greatly exceeds number of regressors, a normal distribution is good approximation of Students t-distribution. In practice, three different models are fitted, namely:

$$p(x) = N(x|\mu,\sigma) \tag{1}$$

$$p(x) = \pi_N N(x|\mu,\sigma) + \pi_A Gamma(x+\mu|k,\theta) \tag{2}$$

$$\begin{aligned} p(x) =& \pi_D Gamma(x-\mu|k_D,\theta_D) \\ &+ \pi_N N(x|\mu,\sigma) \\ &+ \pi_A Gamma(x+\mu|k_A,\theta_A) \end{aligned} \tag{3}$$

Model 1 is fitted using maximum likelihood estimator, and Models 2 and 3 are fitted using an expectation-maximization algorithm (Dempster et al., 1977). Gamma components in Models 2 and 3 correspond to activation and deactivation classes. Note that Gamma components are shifted by the estimated mean of the noise components (Gaussian distribution). For each model, Bayesian information criterion (BIC) is calculated and the model with the highest score is selected. Compared to other approaches (e.g. Pendse et al., 2009), the explicit model selection via BIC and the use of Gamma distributions allows the case when no signal is present (Model 1) to be determined, and avoids having to attribute components to noise or activations. In the case that Model 1 is selected it is assumed that the data contains no signal. For Models 2 and 3, each voxel is assigned a label (activation, deactivation and noise) corresponding to the component with the highest posterior probability. In these cases, the highest T-value among voxels belonging to the noise class is chosen as the new cluster forming threshold. Clusters defined this way undergo topoogical FDR procedure (Chumbley & Friston, 2009). Using Random Field Theory each cluster gets assigned a probability value based on it's extent, clusterforming threshold and estimated image smoothness. These values are then corrected for FDR.

2.2. Simulations

To investigate the performance of each method, a total of 2500 time series were simulated with varying activation sizes and signal stength. Each SPM was thresholded using topological FDR with 3 different cluster forming thresholds. Two fixed cluster forming thresholds were used across all 2500 SPMs, specifically a p-values of 0.05 with family wise error (FWE) correction (T-value of 4.47) and 0.001 uncorrected (T-value of 3.19). These thresholds were chosen as they correspond to defaults values used in the SPM software package (`http://www.fil.ion.ucl.ac.uk/spm/`) and we refer to them as fixed thresholds (FT 0.05 FWE and FT 0.001). This contrasts with the cluster forming thresholds obtained with the Gamma-Gaussian mixture model which by nature change with the data. We refer to these thresholds as adaptive thresholds (AT).

2.2.1. Spatial accuracy

For a given true cluster, the degree of underestimation was defined as the number of voxels that were falsely declared as not active, and the degree of overestimation was defined as the number of voxels that were falsely declared as active. Using these definitions, cluster borders can be simultaneously overestimated (voxels declared active that should not be) and underestimated (voxels declared non-active that should not be). Comparisons between AT and FT were performed in a pairwise manner using a percentile bootstrap on the Harrel-Davies estimates of the median differences. Multiple tests correction was applied maintaining FDR at the 0.05 level.

2.3. fMRI data and reliability analyses

2.3.1. Subjects

Eleven healthy volunteers were recruited. One subject had to be discarded due to problems with executing the task. Subjects had to move a body part corresponding to a picture finger, foot or lips. A block design with 3×15 sec activation periods with 15 sec rest periods was used. Four trials were used for training before data acquisition. Four volumes were acquired for signal stabilization before stimulus presentation. There were five repetitions of each activation block.

Scanning was performed using a GE Signa HDx 1.5 T clinical scanner. Each volunteer was scanned twice, two (eight subjects) or three (two subjects) days apart. SPMs were computed for each body part and the resulting maps were thresholded with a cluster threshold of 0.05 FDR corrected but using AT and the two default FT values as in the simulations. For every subject, contrast and thresholding method, Dice similarity (overlap) was calculated between the two sessions. Comparison of thresholding methods was performed using a percentile bootstrap on the differences between Dice coefficients. Finally, to further investigate the impact of AT and FT on reliability, the Dice values were computed using multiple threshold combinations between sessions, and AT and FT located in this space. This allowed an understanding of the underlying behaviour of our reliability metric in relation to different cluster forming thresholds.

3. Results

3.1. Simulations

3.1.1. Spatial accuracy

Due to the fact that the smallest cluster was found by all of the thresholding methods in only a handful of runs it was excluded from further analyses; in other words there were not enough true positives to reliably estimates border accuracy. For the remaining cluster sizes, AT outperformed both default FT values in terms of underestimation of borders, i.e. it showed fewer false negative voxels, but at the same time it performed worst in terms of overestimation with more false positive voxels. However, the difference was such that AT had a better overall spatial accuracy, i.e. trade-off between over and underestimation. AT provided a statistically significant improvement in terms of the border over/under estimation when compared to both of the two FT values. As in the cluster analysis the effect was stronger for lower SNR levels, although in case of the highest tested SNR, 0.16, FT 0.001 performed equally well as AT.

3.1.2. Reliability experiment

For the three evaluated contrasts of the motor task (finger, foot, lips), AT provided improvement in terms of between session Dice overlap over both default FT values. Mapping of the parameter space showed that many combinations of thresholds can lead to high Dice overlap, and that highest values were obtained when different thresholds between sessions were used. The reason behind this phenomenon is that maximum T-values are often shifted between sessions as evidenced by looking at the joint distribution of T-values. Indeed the tail of the joint distribution is off-diagonal, meaning that voxels in the second scan session have higher or lower T-values than the same voxels in the first session. This effect is mostly observed when there is a shift of the overall distribution, i.e. in the context of a global effect (Friston et al., 1990) such as when temporal noise correlates with the stimuli sequence and affects the whole brain. AT attempts to estimate and correct for this effect by allowing the Gaussian component to have non-zero mean and having the activation and non-activation components range fixed to that mean, leading to a choice of a pair of thresholds optimal in terms of Dice overlap (see Figure 1).

4. Discussion

Single subject fMRI analyses have different requirements than group studies mainly because the SNR is often lower, and one wants to reveal specific or expected areas and delineate their spatial extent. For these reasons, a fixed threshold strategy is rarely adopted and each subject's SPM tends to be thresholded differently. Here, we propose a method that thresholds each subject's statistical map differently, but follows an objective criterion rather than a subjective decision. Indeed, we show that our adaptive thresholding method outperforms default fixed thresholds in terms of spatial accuracy. This increase can also explain the reliability results. While validity and reliability can be separated in various conditions, we can infer that, for fMRI, the most valid voxels are the ones detected reliably. Valid and reliable voxels usually correspond to voxels located at the core of a cluster while non-valid and non-reliable voxels are located at the cluster borders. Since AT leads to higher reliability than FT, we can infer that it also improves clusters delineation in real data sets.

A major source of noise in fMRI time series relates to global effects. Because of the shift of the overall T-value distribution below or above 0, a fixed threshold

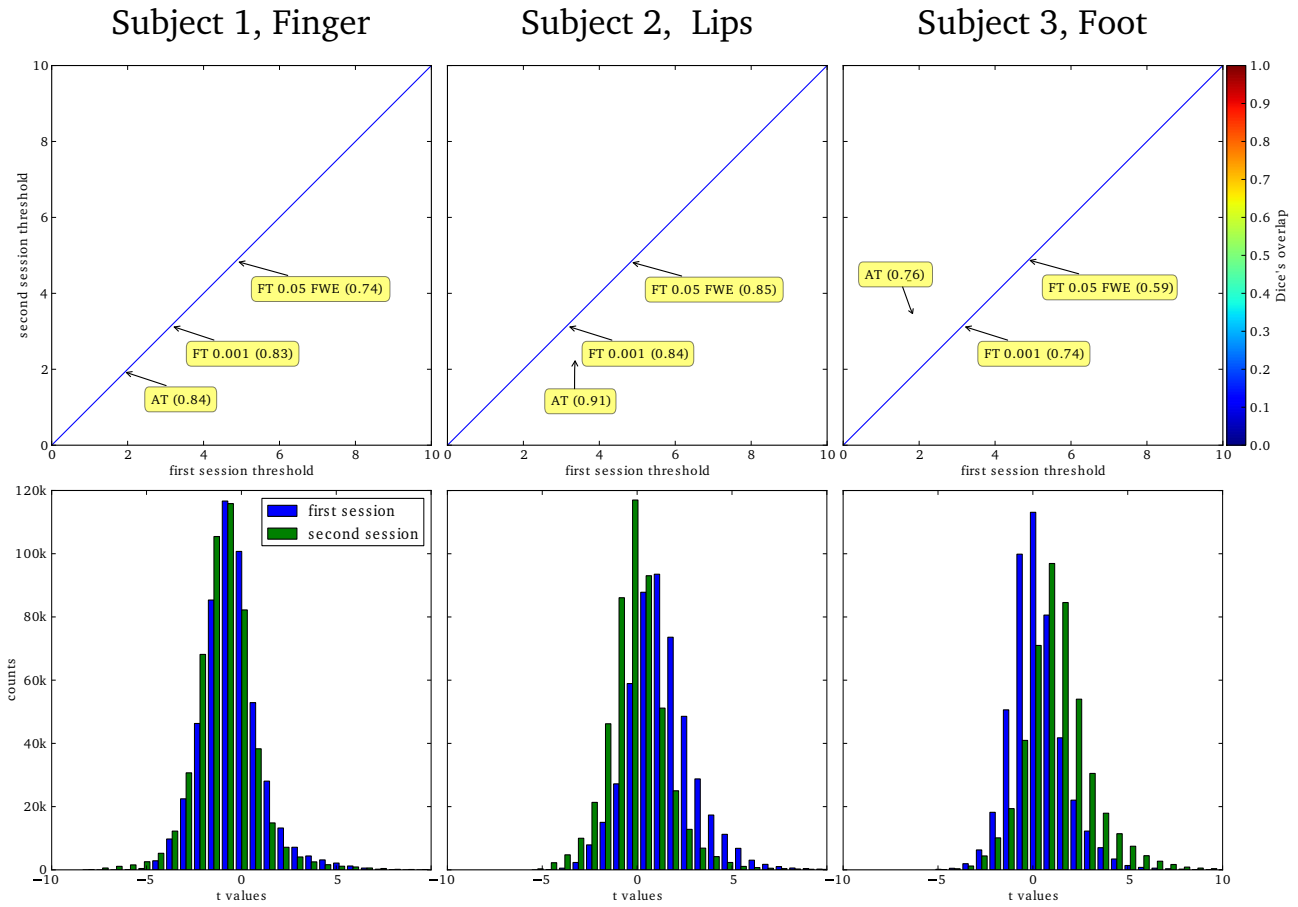

Figure 1. Analysis of the T-map reproducibility of three selected subjects. Heatmasps show between session Dice coefficients for different pairs of cluster forming thresholds. Finger contrast for Subject 1 illustrates the case where choosing the same threshold for both session is the optimal course of action; Lips movement contrast for Subject 2 shows a shift of values between the sessions. This allows AT to choose a lower threshold for the second session and optimize the Dice coefficient value. Foot contrast for Subject 3 presents a shift in the opposite direction.

strategy can lead to the under or overestimation of the true signal. By contrast, we show that AT can correct for global effects by shifting the mean of the Gaussian component in our Gamma-Gaussian mixture model. This ability to adapt to noise translates to improved reliability in a test-retest study on healthy controls. A similar approach has been used before to remove global effect biases in a session variability study by Smith et al. (2005), but not in context of thresholding statistical maps.

Finally, because AT provides a higher spatial accuracy and adapts to noise, it also leads to an increase in reliability. In the context of single subject fMRI analysis, and in particular for data used in clinical procedures such as presurgical planning, it is worth noting that spatial accuracy is essential. Of particular interest here, AT showed much lower underestimation than FT, which may be useful in clinical situations. Increased spatial reliability in healthy controls also means that one can be confident that the method will more often detect valid clusters as suggested by the reduced false negative rate in the simulations. Overall, AT therefore achieves a better balance than FT approaches, and provides a new tool to reliably and objectively threshold multiple single-subject SPMs.

References

Chumbley, Justin R. and Friston, Karl J. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1): 62–70, January 2009.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.

Friston, KJ, Frith, CD, Liddle, PF, Dolan, RJ, Lammertsma, AA, and Frackowiak, RSJ. The relationship between global and local changes in PET scans. *Journal of Cerebral Blood Flow & Metabolism*, 10 (4):458–466, July 1990.

Pendse, G., Borsook, David, and Becerra, Lino. Enhanced false discovery rate using Gaussian mixture models for thresholding fMRI statistical maps. *NeuroImage*, 47(1):231–261, August 2009.

Smith, Stephen M, Beckmann, Christian F, Ramnani, Narender, Woolrich, Mark W, Bannister, Peter R, Jenkinson, Mark, Matthews, Paul M, and McGonigle, David J. Variability in fMRI: a re-examination of inter-session differences. *Human brain mapping*, 24(3):248–57, March 2005.

Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., and Smith, S.M. Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. *IEEE Transactions on Medical Imaging*, 24(1):1–11, January 2005.