Automated detection of fMRI artefacts from Shannon entropy distributions

John McGonigle[1]
mcgonigle@cs.bris.ac.uk
Majid Mirmehdi[1]
majid@cs.bris.ac.uk
Andrea L. Malizia[2]
andrea.l.malizia@bristol.ac.uk

[1] Department of Computer Science
University of Bristol, UK

[2] Psychopharmacology Unit
University of Bristol, UK

Abstract

As the number of subjects in modern fMRI experiments increases, the use of automated analysis pipelines is becoming more popular, leading to less manual inspection of the data. Here we promote the use of Shannon entropy distributions to discover those datasets in large studies suffering from various artefacts. Entropy distributions of 1444 resting state fMRI datasets from the 1000 Functional Connectomes Project are examined and mean distributions found after each of several different preprocessing steps. Empirically derived envelopes are generated so that significantly outlying datasets may be identified. This process of outlier detection may be automated such that those datasets with characteristic shifts in entropy caused by specific artefacts may be flagged for further manual examination or removed from further analysis. We conclude this technique will be a useful quality control method when dealing with data from large studies.

1 Introduction

The number of subjects in modern functional Magnetic Resonance Imaging (fMRI) experiments is increasing as researchers seek to examine effects across larger populations and more groups pool data. With this increase has come the more common use of automated analysis pipelines, especially at the preprocessing stage. As such, it becomes difficult for an individual to manually check for sometimes subtle artefacts which may have a detrimental effect on further analysis. This is especially true when there may be hundreds of subjects, and the artefacts transient in nature. Techniques such as Independent Component Analysis (ICA) are often used to guide these manual checks [4].

The recent public release of more than 1000 resting state fMRI (R-fMRI) datasets as part of the *1000 Functional Connectomes Project* [1] provides the neuroimaging community with the opportunity to apply and test analysis techniques on a much larger number of subjects than is usually available locally. With the potential to examine data from many sources comes the issue of how the characteristics of this data vary between sites, and also between studies at the same site. It also allows for typical distributions of various summary statistics to be found, and their dependence on scan parameters determined.

Here we explore the use of Shannon entropy [3] distributions in order to automatically identify outlier datasets. Removal of these is prudent before carrying out further data-driven methods as are often used in the analysis of R-fMRI. Specifically, we show that shifts in these distributions can be characteristic of certain types of artefact.

2 Methodology

2.1 Shannon entropy for outlier detection

The entropy, H, of a discrete random variable, X, is the average minimum amount of information that is necessary to encode a string of symbols, and may be found as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i) \tag{1}$$

where $p(x_i)$ is the probability mass function which may be determined from a histogram of the original data [3]. In the case of fMRI data, each voxel time course (expressed as percentage signal change) may be divided into a number of discrete signal levels, and these levels used as symbols in an entropy calculation [2]. In this paper 20 signal levels are used for each voxel. As the entropy calculation is carried out at every voxel in each dataset there will be tens to hundreds of thousands of entropy values for each individual. Distributions of these values may then be compared to others.

By collating and scaling the entropy distributions of many individual subjects a mean distribution may be found together with the 10th and 90th empirical percentiles. Those distributions where more than 25% of their voxels are found to lie outside of this "envelope" of the empirical percentiles may be deemed to be outliers and flagged for further examination and possible removal (these values have been found empirically).

2.2 Resting state data

The data explored here is freely available from the 1000 Functional Connectomes Project (www.nitrc.org/projects/fcon_1000/) and represents 1444 sessions, involving more than 1300 individual subjects, collected independently during 31 studies at 24 sites. Full details of the geographic distribution and exact parameters for each site may be found on the project website and in [1]. The mean number of volumes for each session is 174, with a mean repetition time (TR) of 2.14s, with the mean scanning time being just over 6 minutes. The variance from the slightly differing resting state protocols (e.g., eyes open or closed) does not affect the summary distributions examined here.

2.2.1 Preprocessing

To explore the effect of various steps in a conventional fMRI preprocessing pipeline on entropy distributions the following were carried out in several combinations. All data was motion corrected, spatially normalised to the MNI152 template (with only gray matter voxels examined further) and smoothed with a Gaussian kernel of full width at half maximum (FWHM) of 5mm, all using SPM8 (www.fil.ion.ucl.ac.uk/spm/). A detrending was also carried out, incorporating the global mean and motion parameters to remove the effects of global signal and residual motion as would be implicit in a conventional general linear model

approach to analysis (where they would be used as unwanted effects regressors). All values were converted to percentage change differences from the mean of timepoints 2-25.

3 Results

Figure 1 shows those datasets identified as outlier after only motion correction (the most important stage to find artefacts before they are smeared throughout the data by the interpolation steps of spatial normalisation and further corrections). It should be noted that only 3% of all the datasets shown here are marked as outlier (48 out of the 1444 examined). In practice one might wish to only examine the furthest outlying distributions manually.

The banding artefact (the alternating light and dark regions) shown in Figure 2 is due to a possible signal dropout. It was identified as an outlier dataset through its Shannon entropy distribution being shifted lower and flatter (a peak of 2.2 nats compared to a mean peak of 2.7 nats). Since this is a transient artefact not affecting all scans in a session its presence is easy to miss in a manual data quality check, but can have a detrimental effect on further analysis. The severe susceptibility distortion artefact shown in Figure 3 is likely due to microscopic pieces of metal near the eye. This type of artefact is almost constant for each timepoint but the effect of slight motion will change the distortion, leading to abnormal signal. In this case the entropy distribution is shifted higher (with a peak of 2.79 nats). The datasets highlighted here do not appear as significant outliers when looking at standard deviation distributions alone.

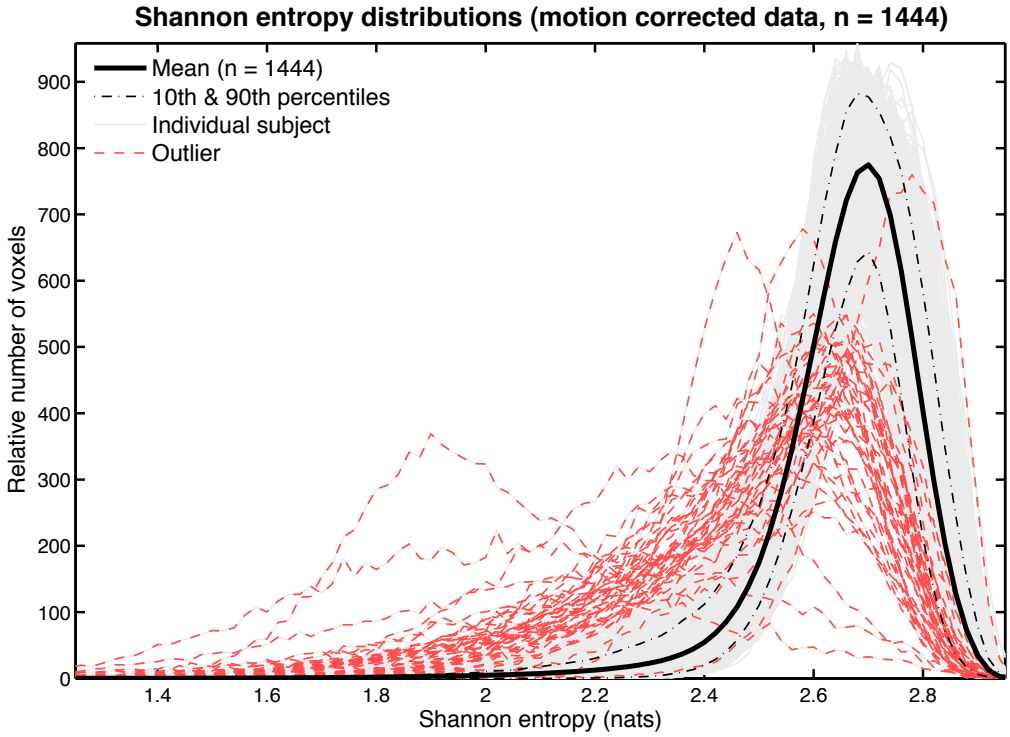


Figure 1: The Shannon entropy distributions for all 1444 datasets with those having 25% or more of their voxels lying outside the 10th and 90th percentile envelope classed as outliers and highlighted in red. 48 datasets (just over 3%) are flagged here. These could then be investigated in detail as in Figures 2 and 3.

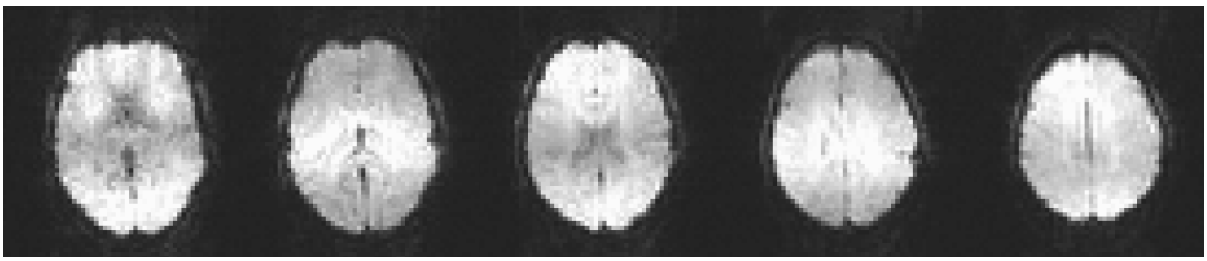


Figure 2: Banding artefact due to possible signal dropout from dataset New Haven (a): subject 13647b. Its Shannon entropy distribution was flagged as an outlier having a peak shifted to lower entropy. Image resolution is typical for whole brain fMRI.

It would appear that many of the distributions with lower shifted peaks are due to susceptibility artefact around the skull, higher shifted peaks caused by frontal lobe susceptibility distortions, with flattened distributions caused by signal dropout.

The mean entropy distributions after several combinations of preprocessing steps can be seen in Figure 4(a). It is interesting to note that only performing a spatial normalisation to a standard space does not have a significant effect on the entropy distribution. As one might expect there is an overall decrease in the amount of information necessary to represent the data, on average, when the contributions from drift and residual motion have been removed.

For comparison, the effects of the preprocessing steps on standard deviation distributions are shown in Figure 4(b), where it is evident their relationships do not follow directly from the entropy distributions.

4 Discussion

With the growing availability of large numbers of datasets for analysis there is the potential for including many containing artefacts which would ideally be excluded from further processing. Individually checking all datasets in detail is often not practical and is prone to human error due to the subtlety and transient nature of many artefacts. Thus, finding ways to exclude these before the mass scripted analysis which is common in large studies will be important, especially those which might not be apparent without detailed study. Furthermore, R-fMRI is commonly analysed using either ICA or a seed-region based analysis strategy. These are affected by data-quality issues more than the hypothesis based approaches used in conventional fMRI.

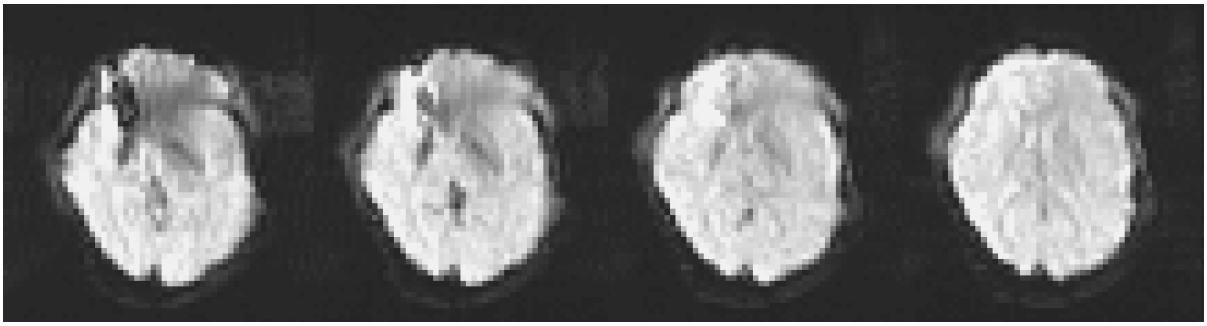
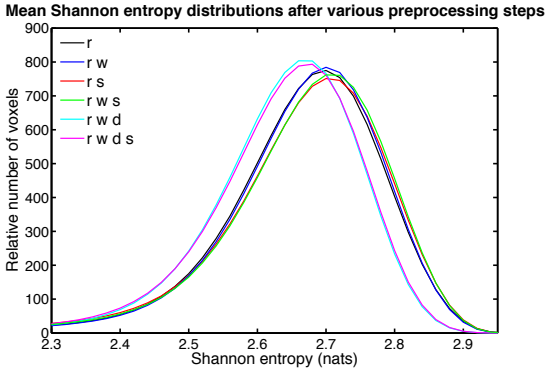
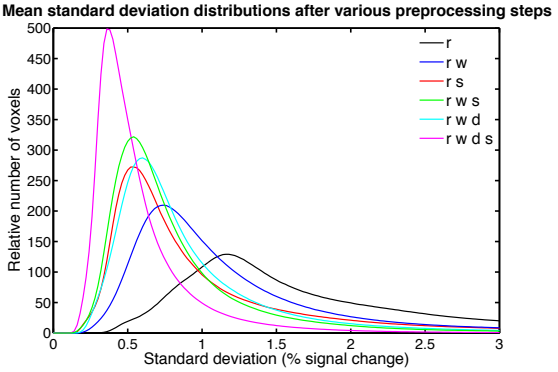


Figure 3: Susceptibility distortion artefact from dataset Taipei (a): subject 03537. Its Shannon entropy distribution was flagged as an outlier having a peak shifted to higher entropy.



(a) The mean Shannon entropy distributions for 1444 datasets after combinations of motion correction (r), spatial normalisation to a standard space (w), smoothing with FWHM of 5mm (s) and motion aware detrending (d). Note that spatial normaisation alone does not significantly change the entropy distributions.



(b) The mean of standard deviation distributions for 1444 datasets after combinations of motion correction (r), spatial normalisation to a standard space (w), smoothing with FWHM of 5mm (s) and motion aware detrending (d).

Figure 4: Comparing the effect of preprocessing steps on the mean distributions

We have proposed that fMRI datasets containing certain forms of artefact may initially be recognised through distributions of their Shannon entropy and flagged for examination and possible rejection or correction. Future work will examine how these distributions manifest themselves both spatially in the brain, and between different scanners and centers.

References

[1] Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M. Smith, Christian F. Beckmann, Jonathan S. Adelstein, Randy L. Buckner, Stan Colcombe, Anne-Marie Dogonowski, Monique Ernst, Damien Fair, and others. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010. URL http://dx.doi.org/10.1073/pnas.0911855107.

[2] D.B. de Araujo, W. Tedeschi, A.C. Santos, J. Elias Jr., U.P.C. Neves, and O. Baffa. Shannon entropy applied to the analysis of event-related fMRI time series. *NeuroImage*, 20 (1):311–317, 2003. URL http://dx.doi.org/10.1016/S1053-8119(03)00306-9.

[3] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[4] J. Sui, T. Adali, G.D. Pearlson, and V.D. Calhoun. An automatic artifact removal method for independent components derived from second-level fMRI analysis. *NeuroImage*, 47(Supplement 1):S81 – S81, 2009. URL http://dx.doi.org/10.1016/S1053-8119(09)70576-2.