Exploratory Data Analysis of fMR Images: Philosophy, Strategies, Tools and Implementation

R.L. Somorjai, M. Jarmasz

Informatics Group, Institute for Biodiagnostics, NRCC, Winnipeg, MB, Canada

INTRODUCTION

Conventional data analysis methods of fMR brain images are based on the assumption that a model of the brain's response to a (designed) stimulus is available (or can be constructed), and that the validity of this model can be tested by statistical methods of inference. This imposes the twin burden of *model arbitrariness*, and the need to confront and deal with the ubiquitous *multiple measurement* problem. However, neuroscientists are designing increasingly complex cognitive/linguistic experiments to probe brain function. This complexity demands that the *model validation* (i.e., hypothesis testing) approach, which is the basis of all currently available popular softwares, such as AFNI or SPM, be preceded by data-driven, bias- and model-free *exploratory data analysis* (EDA) methods that are versatile, flexible, fast and are capable of *generating* models (hypotheses).

METHODS

The above philosophy of rapidly and flexibly generating unbiased, testable models (hypotheses) is the foundation for the analysis strategy that we've implemented in the software EvIdent™. EvIdent clusters the time-courses (TCs) with a much-enhanced variant of Bezdek's Fuzzy C-Means Clustering algorithm (FC)[i]. FC has been chosen over other unsupervised pattern recognition methods such as Kohonen's SOM or some variant of factor analysis, because the former is too computer intensive and the latter too inflexible and limiting. All methods have to deal with the typically low (2-5%) level of activation above the "rest" condition, and with the low SNR. Even more importantly, they must be able to identify and extract the rather small fraction of the total number of voxels that are activated by a particular stimulus.

Because we are interested in the *temporal shapes* or *profiles* of the TCs, the first step of *preprocessing* the data is a type of "normalization" that corrects for the intensity disparity of the voxel TCs. We implemented several options; we found either *subtracting the median* or *dividing by the median* the most useful, depending on whether we want subsequent clustering to differentiate between TCs of the same shape but different *absolute* activation *amplitudes*, or of the same shape but different *relative amplitudes*.

Since we expect that the activated voxels are but a small fraction of the total number, we have devised a flexible approach to *preprocessing* and *preselection* that help identify *potentially interesting* TCs and separate these from the uninteresting ones. In particular, prior to clustering we *exclude* TCs that have significant *trends* and place them in a "trend" cluster for possible further consideration. (Trends may be due to motion artefacts and/or instrumental drift. In EPI data we routinely find trends in as many as 50-60% of the voxels.) The distinction between *exclusion* and *correction* is important. AFNI or SPM correct for, say, linear trends in TCs as part of modelling. (Although EvIdent can also correct both linear and nonlinear trends, we often find that such correction is inadvisable, because it creates artefacts that look like "activations".) After FCA finds "good" clusters, we reassign some of "trendy" TCs to them. The *temporary* exclusion of "trendy" voxels not only speeds up processing; it also removes the contaminating effect of trendy TCs, thus leading to better clustering results. Trend detection is followed by preselection methods that help exclude additional TCs from analysis. We have devised, implemented and tested two *preselection* methods that work well, and can be used either alone or together. One of these[ii] uses *one-scan-shift self-similarity* (a variant of the serial auto-correlation function) of individual voxel TCs, and identifies noise voxels based on standard statistical criteria. The other[iii] is based on a prior sorting of the TCs according to the *number of crossings*, i.e., the number of times a given TC crosses a "deadband" enclosing its median. The many possible combinations of "normalization", trend exclusion and preselection provide the flexibility needed by our EDA software.

RESULTS AND DISCUSSION

One of the essential requirements of any EDA method is that it be fast; we want to generate and test many hypotheses in reasonable computer times. The FC algorithm, as implemented in EvIdent (with the significant enhancement provided by user-controlled *cluster merging*[iv]), satisfies this requirement: we fitted empirically the total execution time t_{exe} (sec) to $N^{\alpha}_{voxel}\, T^{\beta}_{scan}$, where N_{voxel} is the total number of brain voxels and T_{scan} the number of scans. $t_{exe} = A\, N^{\alpha}_{voxel}\, T^{\beta}_{scan}$, $A = 5.02 \times 10^{-5}$, $\alpha = 0.8539$, $\beta = 0.6404$ gives a good fit, with $R^2 = 0.9802$, $\rho_{tY} = 0.9900$, confirming the reliability of EvIdent's algorithms over a wide range of realistic conditions. The ranges (14 data sets) were $0.5 \text{ sec} \le t_{exe} \le 21.3 \text{ sec}$, $4K \le N_{voxel} \le 68K$, $35 \le T_{scan} \le 315$. Computations were carried out on an SGI Origin 2000 CPU (slightly slower than a 400 MHz Pentium II PC).

We emphasize that EDA methods are not meant to, and in fact should not attempt to supplant the model-based, hypothesis-validating methods of statistical inference. Rather, they supplement the latter. Their objective of *rapidly* and *flexibly generating* multiple models of TC behavior *for additional statistical testing* also means that controlling type I errors (false positives) due to repeated measurements is much less critical at this stage and may be left for the follow-up inferential methods. The (possibly "denoised") cluster centroids ought to serve as the data-generated input models that other software, such as SPM can use to construct a design matrix for further analysis[v]. Thus the inevitable arbitrariness of creating *generic* models that may not reflect the peculiarities of the current data can be largely eliminated.

Although we are not ignoring inferential statistical issues, our current emphasis is on cluster validation and improvement; we want to guarantee that the "models" we generate are minimally contaminated by irrelevant TCs.

REFERENCES

[i] Bezdek, J.C. *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, 1981

[ii] Somorjai, R.L., Jarmasz, M., Baumgartner, R. ISMRM 1999, Philadelphia, (submitted)

[iii] Jarmasz, M. & Somorjai, R.L. ISMRM 1999 (submitted).

[iv] Jarmasz, M. & Somorjai, R.L. ISMRM, Sydney 1998, p. 2068

[v] Baumgartner, R. & al. ISMRM 1999 Philadelphia, (submitted).