An empirical diagnostic tool to estimate the true dimension of fMRI data

R. R. Nandy[1], D. Cordes[1]

[1]Radiology, University of Washington, Seattle, WA, United States

Introduction

An interesting problem in fMRI is to estimate the true dimensionality of the data which is an important step for popular fMRI post-processing tools like PCA and ICA. Typical fMRI data has inherent noise which always makes the data full rank. The true dimension of the noise-free part of the data is usually estimated by using the eigenvalues of the covariance matrix of the data. The plot of eigenvalues in ascending order has a sharp ascent at the point where the eigenvalues have contribution from components other than noise. It is difficult to identify the discontinuity since the SNR for the true components are not always strong and the noise is strongly correlated. There are several methods to estimate the dimension from the eigenspectrum including BIC, MDL and a Laplace approximation based on model evidence [1]. The last method appears to work best for correlated noise and is implemented in MELODIC, part of FSL fMRI data analysis package [2]. However, using simulated data, it can be easily shown that none of these methods offer very accurate estimates of the dimension. Even in real resting-state data, a common characteristic of each of these methods is to provide estimates that grow linearly with the number of time-frames used in the data. Although the true dimension may increase when larger time-frames are used, it is unlikely that the dimension will grow linearly. In light of these facts, we present a simple empirical diagnostic tool to identify the point of discontinuity, which is far more stable in estimating the true dimension as the number of time-frames increases compared to the other procedures mentioned above.

Methods

The noisy mixing model for mean removed fMRI data is $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{s}_i + \boldsymbol{\eta}_i,\ \boldsymbol{\eta}_i \sim N(\boldsymbol{0},\boldsymbol{\Sigma}), i = 1,...,q,$ where q is the number voxels, $\boldsymbol{x}_i$ is the observed time-course at voxel i with T time points, $\boldsymbol{A}$ is the $T \times p$ dimensional mixing matrix, $\boldsymbol{s}_i$ is the source vector with p components at voxel i and $\boldsymbol{\eta}_i$ is the noise that follows a Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$. Let $(\lambda_1,....,\lambda_T)$ be the eigenvalues of the data covariance matrix in ascending order. It can be shown that the first $T-p$ eigenvalues have no contribution from the sources and the discontinuity in the slopes of the eigenspectrum is precisely at this point. Our proposed diagnostic tool to identify the point exploits the fact that the plot of ordered eigenvalues exhibit locally exponential behavior (except at the very beginning and at the end which are not relevant in the estimation). So we plot $\Delta\lambda/\lambda$ (which is scale invariant) against the index of the eigenvalue to identify the discontinuity point. It is expected that the plot will have a strong ascent at the point of discontinuity. A detailed theoretical justification of this method is beyond the scope of this abstract and will be presented in future in a full manuscript. The data is variance normalize. Furthermore, since our method depends on analyzing differences (which are not very robust), we apply a moving average filter of length 7 to the observed eigenvalues.

Results

We first consider a simulated data with 50 sources from Laplace distribution and a noise covariance matrix estimated by scanning a phantom with 299 time frames. In figure 1 (top), we have plotted $\Delta\lambda/\lambda$ by considering all 299 time frames, 250 time frames, 200 time frames and 150 time frames (the last 25 points are dropped in the plot as they are too large and has no role to play in the estimation process). In the plot of $\Delta\lambda/\lambda$, there is always an initial peak before it stabilizes. There is again a sharp ascent once the true components start contributing to the eigenvalues. The initial peak serves as a good threshold to decide the cutoff point to estimate the dimension. The index on x-axis at which the curve crosses the threshold is the estimated number of eigenvalues from pure noise. Subtracting it from the time dimension gives the estimate for the true dimension of the data. For each of the 4 chosen lengths of time frames, the estimated number of dimensions is approximately 75, which overestimates the true dimension. But the estimates are very stable relative to the number of chosen time frames and are far superior to all the other methods which are not only larger than our estimate but also grew significantly with the number of chosen time frames. The estimate using Laplace approximation based on model evidence for 160 time frames was 112. In Figure 1 (bottom), we have used real resting-state data with 350 time frames. As before, we estimated the dimension by using 4 different lengths of time frames (350, 300, 250 & 200). As with the phantom, the estimates are remarkably stable at about 70. All the other methods provided much larger estimates and grow significantly with the number of chosen time frames.

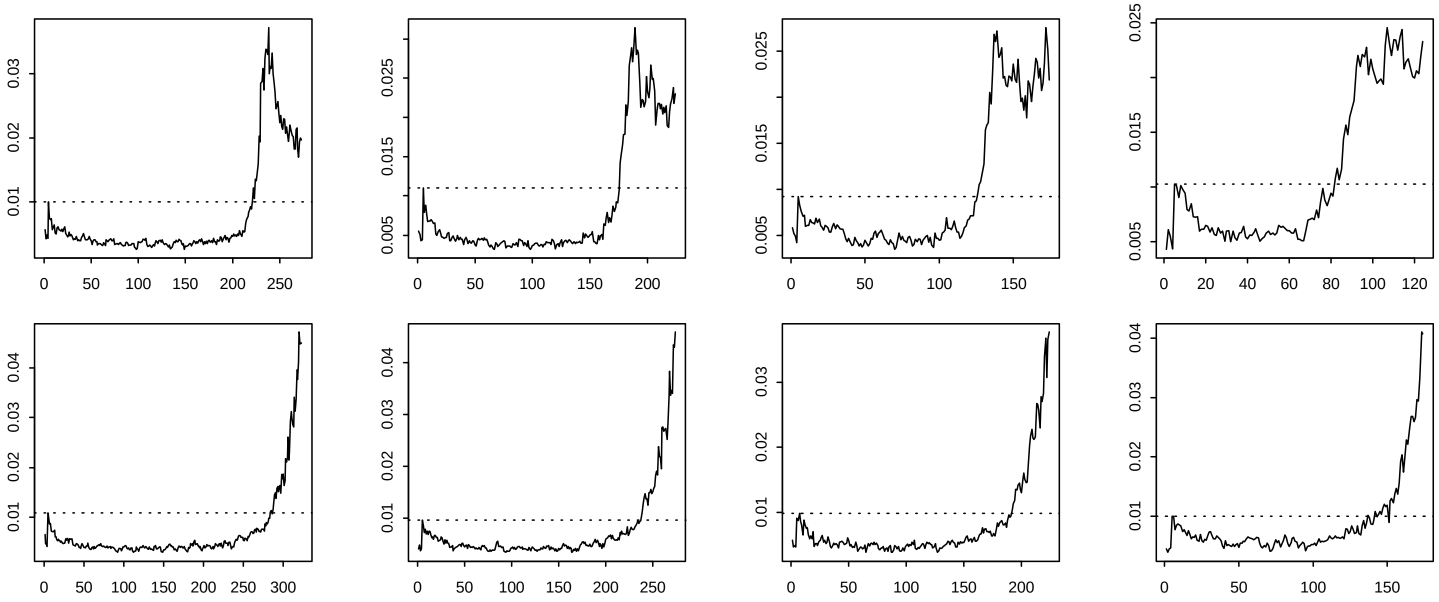

Figure 1. Plot of $\Delta\lambda/\lambda$ for indices of ordered eigenvalues with simulated (top) and real data(bottom). The last 25 values are too large and not shown in the plot.

References

1. Minka, T. Massachusetts Inst.Technol., Cambridge, Tech. Rep. 514, 2000.
2. Beckmann, C.F. and Smith, S.M. (2004). IEEE Trans. on Medical Imaging; 23(2):137--152.