Estimation of the Intrinsic Dimensionality of fMRI Data

D. Cordes[1], R. Nandy[2]

[1]Radiology, University of Washington, Seattle, WA, United States, [2]Psychology, UCLA, Los Angeles, CA, United States

Introduction

A difficult problem in fMRI is to estimate the true dimensionality of the data to determine of what is *essential* in the data[1,2,3]. There are several popular methods based on information-theoretic criteria to estimate this dimension from the eigenspectrum, in particular Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Minimum Description Length (MDL), and a Laplace approximation based on model evidence (PPCA). Except AIC (which is asymptotically incorrect for a large number of samples), all these methods perform virtually identical when the noise model is purely white. The PPCA method appears to work reasonably well for fMRI data and is implemented in MELODIC, part of the FSL analysis package (FMRIB, Oxford, UK). However, using simulated data, it can be shown that none of these methods offer very accurate estimates of the dimension *when the noise is colored*. For example, in real resting-state fMRI data, a common characteristic of each of these methods is to provide dimensionality estimates that grow about linearly with the number of time-frames used in the data. Although the true dimension may increase when more data is acquired in time, it is unlikely that the dimension will be significantly different by 50% or more. In this research we investigate AIC, MDL, and PPCA for simulated and real resting-state fMRI data, and present a novel empirical method to identify more robustly the intrinsic dimension in fMRI data.

Theory and Methods

The noisy linear mixing model for mean removed and variance normalized fMRI data can be described according to $\boldsymbol{x}_i = \mathbf{A}\boldsymbol{s}_i + \boldsymbol{\eta}_i$, $i=1,\ldots,N$. where $\boldsymbol{x}_i$ is the observed time course at voxel i with T time points, $\mathbf{A}$ is the T x p dimensional mixing matrix, $\boldsymbol{s}_i$ is the signal vector (zero mean) with p components at voxel i, $\boldsymbol{\eta}_i$ is the noise that follows a multivariate Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, and N is the number of voxels. Due to the presence of noise, the observation matrix $\mathbf{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N]$ is always of full rank T, irrespective of the number of components of the signal. If T>p, which is reasonable to assume for fMRI data, the dimension of the full dataset is larger than the number of "true" biological components. Since fMRI data have autocorrelations in the noise subspace, we rely on an autoregressive AR(1) noise model, and compute the eigenvalues of the sample covariance matrix for different autoregressive coefficients by simulation. For typical fMRI parameters (N=20000, T=160), the k-th eigenvalue can be accurately approximated by $\lambda(k)=a\exp(-b\,k)$, $k\in[1,T]$ with coefficients $a=a(\phi)$ and $b=b(\phi)$, where $\phi \in [0,0.3]$ is the autoregressive coefficient (see Figure 1 left). Similarly, $a(\phi)$ and $b(\phi)$ can be accurately parameterized by $a(\phi)=a_1\exp(-b_1\,\phi)$ and $b(\phi)=a_2\exp(-b_2\,\phi)$.

Our new method to obtain the dimensionality of the data relies on the following steps: First, the AR(1) coefficient ϕ for the noise subspace is estimated using the tail eigenvalues of the sample covariance matrix. Since the functions $a(\phi)$ and $b(\phi)$ are determined with high accuracy by simulation, a best fit of the tail eigenvalue spectrum to the exponential model will determine ϕ. Then, confidence intervals for the noise eigenvalues can be calculated, if necessary, by simulation, and the dimension of the signal space can be determined by the number of eigenvalues that are significantly larger than the eigenvalue of pure noise.

To make this method practical, the following steps are carried out. First, we assume that $a(\phi)$ and $b(\phi)$ have been determined by simulation using only Gaussian noise with AR(1) covariance structure (for variable ϕ). For example, we found that $a=0.9664\exp(2.868\,\phi)$ and $b=0.002562\exp(4.66\,\phi)$ provide an excellent fit ($R^2>0.98$) for $\phi \in [0,0.3]$ and $\gamma=T/N=160/20000$. Then, to determine the particular value of ϕ for real data, the tail eigenvalue spectrum of the real data is fitted to $\lambda(k)=a_r\exp(-b_r\,k)$, and the coefficients a_r and b_r are obtained. The value ϕ is determined from the known parameterization $b=b(\phi)$. Because of the variance normalization of each time series, the tail eigenvalues for real data will always be smaller by a shift Δ than the corresponding simulated Gaussian noise eigenvalues due to the fact that real data contains both signal and noise, and the variance of the sum of these quantities is normalized to one. This shift Δ can be determined from the tail spectrum, and thus the noise spectrum can be properly adjusted. An equivalent AR(1) noise eigenvalue spectrum is then obtained by simulation, adjusted, and compared to the eigenvalue spectrum of the real data. The particular eigenvalue of the real data that is larger than the corresponding adjusted simulated noise eigenvalue then defines the dimension of the signal subspace.

Results and Conclusion

The estimated dimension of simulated data (50 Gaussian sources, SNR=2, ϕ=0.2) as a function of the temporal dimension is shown in Figure 1 (middle). On the right, the results for real resting-state fMRI data (1.5T, EPI, TR 2sec, 20 slices, TE 40ms, 64x64 resolution) from 10 subjects is shown. Both figures show that the PPCA and MDL estimate with larger T (ex. T=320) is significantly larger than the estimate with smaller T (ex. T=160), which is inconsistent with the eigenvalues spectra of the data. Our method showed exact results for simulated data. For real data, we obtained a far smaller difference in the dimensionality estimate when comparing data with different temporal sizes. Although the intrinsic dimensionality is unknown for real data, our method incorporating the estimation of the AR(1) coefficient appears to provide a more consistent estimate. All real resting-state fMRI data yielded values of ϕ in the range [0.14, 0.24]. AIC failed to produce any local minima as shown by the high estimates identical to the temporal size of the data.

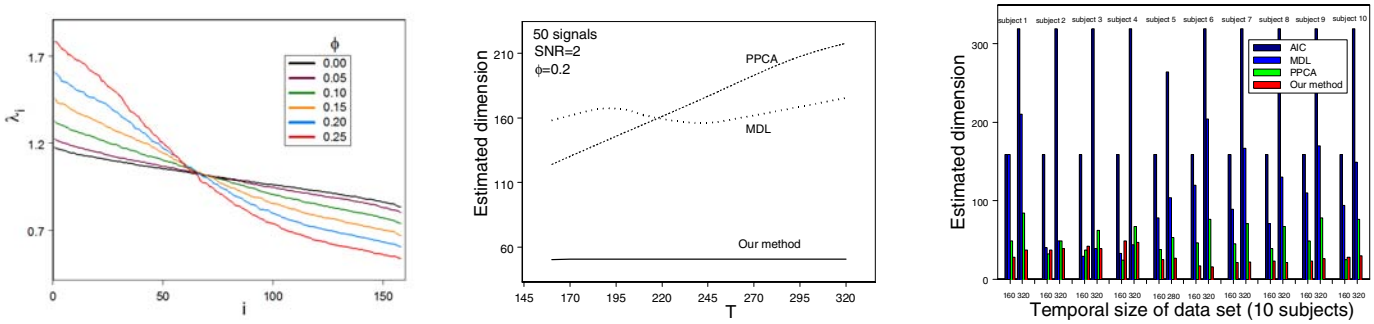

Figure 1. (**Left**) Exponential behavior of the eigenvalues for simulated correlated Gaussian noise corresponding to an AR(1) model (the AR(1) coefficient is ϕ). **Middle:** Estimated dimension for simulated data with 50 Gaussian signals and correlated noise corresponding to an AR(1) model as a function of the temporal size of the data. Note, the inaccuracy of predicting the dimension by MDL and PPCA. **Right:** Estimated dimension of real resting-state echoplanar data for 10 different subjects. Each data set was analyzed using the full temporal dimension (T=320 for 9 subjects, T=280 for 1 subject) and also using only the first half of the data (T=160 for all subjects). Note that both MDL and PPCA show a significant increase in dimensionality from T=160 to T=320 (280) contrary to our method.

References

1. Beckmann, C.F., Smith, S.M. IEEE Trans. Med. Imag. **23**:137-152 (2004).
2. Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J. Hum. Brain Mapp. **14**:140-151 (2001).
3. McKeown, M.J. Neuroimage **11**: 24-35 (2000).