PCA-Preprocessing of fMRI Data Adversely Affects the Results of ICA

Christopher G. Green, Rajesh R. Nandy, Dietmar Cordes
Department of Radiology, University of Washington, Seattle, WA, USA

Abstract

Independent Component Analysis (ICA) is a new technique for analyzing fMRI data. Unfortunately, the size of fMRI datasets sometimes renders this technique computationally intractable, and certain compromises must be made to perform the analysis. One such compromise is to project the dataset onto a lower-dimensional subspace using a Principal Components Analysis (PCA). This subspace, which in some sense captures the essence of the data, is then used as the input to ICA. It is demonstrated herein, however, that an ICA analysis of a PCA-preprocessed dataset will tend to favor Gaussian and near-Gaussian distributions and can miss task-related activation components.

Introduction

Independent Component Analysis (ICA) is a means of recovering source signals **S** from their observed mixture **X**. In its most common form, ICA assumes that there exists a matrix **A** such that **X = AS** and that the sources **S** are mutually independent. By constructing a suitable measure of independence among the components of **S**, we may estimate **S** by optimizing this measure.

The use of ICA for analyzing fMRI data was first proposed by McKeown [1]. Here the measured MRI signal is a mixture of the signals arising from various biological processes. It is desired to recover activations maps and time courses corresponding to the functional task in question.

The sheer size of the dataset often causes many problems for analysts, however, leading them to employ Principal Component Analysis (PCA) to reduce the dimension of the input to ICA. The assumption behind employing PCA to project the data onto a lower-dimensional subspace is that PCA will capture "what is important about the data". The validity of this assumption, however, has been challenged by Porrill and Stone [2]. In this study we demonstrate that the interesting signals in fMRI data are easily missed by even a modest PCA reduction in dimension.

Methods

Data Acquisition: Three normal volunteers in their twenties participated in the experiment. Each subject was given a bilateral finger-tapping exercise consisting of four 64 second blocks of periodic opposition of thumb and other fingers for 32 seconds followed by 32 seconds of rest. All scans were done on 1.5 T GE Signa Horizon MRI scanners equipped with high-speed gradients and standard birdcage head coils. Twenty slices (7 mm thick/2 mm gap) were obtained in the coronal direction using single-shot EPI with a 64 x 64 matrix and the following parameters: FA, 90°; TE, 40 ms; TR, 2000 ms; FOV, 24 cm; BW, ± 62.5 kHz.

For the computer-simulated portion of this experiment, three sources were used: the function exp(-r), $r = (x^2 + y^2)^{1/2}$, which models the point-source activations typically encountered in fMRI; a 4096-element random vector with entries drawn from N(0,1); and a (portion of a) natural image (obtained from [3]). The original images are shown in Figure 1.

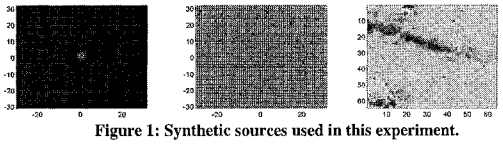

Figure 1: Synthetic sources used in this experiment.

These three sources were mixed using the linear transformation $\mathbf{S} \rightarrow \mathbf{AS}$, where **A** is the matrix

$$A = \begin{bmatrix} 0.33 & 0.34 & 0.33 \\ 0.8 & 0.15 & 0.05 \\ 0.4 & 0.5 & 0.1 \end{bmatrix},$$

to form the observed synthetic data set.

Data Analysis: ICA was performed using two freely available packages [4,5]. The authors wrote additional processing code in MATLAB (The MathWorks, MA).

For the real datasets, ICA was performed with and without PCA. All components accounting for 99.99% percent of the observed variance were kept in the PCA trial. All independent components were then ranked according to the correlation of their time-series and periodograms with those of a square wave modeling our paradigm and examined visually in AFNI (Robert Cox, NIH) for physiological significance.

For the synthetic dataset, the two "most significant" principal components were kept. FastICA was then performed on the reduced dataset. These results were compared with those obtained by using FastICA on the full dataset to extract only the first two independent components. All components generated were visually compared with the input dataset.

Results

In the simulation, ICA successfully found the Gaussian noise and the natural image from the two principal components (Figure 2) but missed the point-source activation entirely. From the full dataset, however, FastICA found the point-source activation and the natural image (Figure 3).

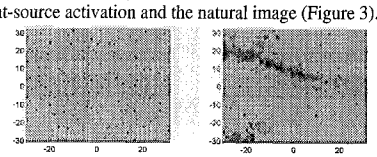

Figure 2: Sources found by ICA in PCA-preprocessed data.

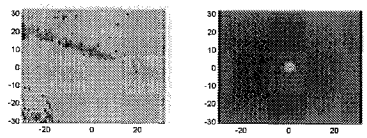

Figure 3: Sources found by ICA in unprocessed data.

The results from the real data were similar. PCA-preprocessing extracted approximately 40 components, only one of which correlated strongly with the reference function (r = 0.71) and with its power spectrum (r = 0.993). This component shows activation in the motor cortex (Figure 4a). Both ICA algorithms, without any preprocessing, found a similar component in each subject that correlated highly with the reference function (r > 0.6 for time-series, r > 0.9 for power spectra) and demonstrated activation in the motor cortex. The unreduced ICA, however, yielded additional components in each subject that correlated highly with the reference function (one is shown in Figure 4b); these components were not found in the PCA-reduced case.

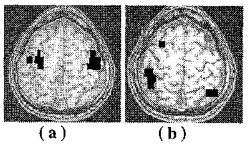

Figure 4: Axial views of activations found by ICA. (a) Activation found by ICA in both processed and unprocessed data. (b) Activation found by ICA in unprocessed data but not processed data.

Discussion

These results show that preprocessing via PCA can affect significantly the output of the ICA algorithm. Indeed, in both the live data analysis and the simulation, PCA-preprocessed ICA failed to detect all the activations associated with the paradigm when the dimension was reduced too aggressively. Commonly used thresholds in PCA may not capture enough of the observed variance to find components corresponding to highly localized brain activity, as such components only explain a small fraction of the total observed variance across the brain. Furthermore, the results of the synthetic data experiment seem to suggest that PCA is biased towards Gaussian components, whereas in fMRI most components of interest are supergaussian. This agrees with the theoretical calculations of Diaconis and Freedman [6], which show that low-dimensional projections of high-dimensional datasets tend to have Gaussian distributions.

References

1. McKeown MJ et al. *Human Brain Mapping* **6**:160-188 (1998).
2. Porrill J, Stone JV. Undercomplete ICA for signal separation and dimension reduction. Tech Report, Dept Psychology, Sheffield Univ, England (1998).
3. Sinkkonen J. Natural Image Collection for ICA experiments. ICA group, Lab Comp Info Sci, Helsinki Univ Tech.
4. Makeig S et al. MATLAB Toolbox for ICA, Ver 3.52. Comp Neurobiology Lab, Salk Inst Bio Studies (1997)
5. Hyvärinen A, Oja E. The FAST-ICA MATLAB package, Ver 2.1. ICA Group, Lab Comp Info Sci, Helsinki Univ Tech (2001).
6. Diaconis P, Freedman D. *Annals Stat* **12**(3):793-815 (1984).

© Proc. Intl. Soc. Mag. Reson. Med. 10 (2002)