Exploration of high dimensional biomedical datasets with low-distortion embeddings [*]

François Meyer [†] Xilin Shen [‡]

Abstract

Different physical phenomena contribute to the dynamical changes in the functional magnetic resonance imaging (fMRI) signal: task-related hemodynamic response, non-task-related physiological rhythms, machine and motion artifacts, etc. The contribution of this work is a novel method to chart functional maps, that are build globally based on the geometry of the fMRI time series. This method relies on nonlinear mapping that embed the dataset into a low dimensional space, and minimizes the local distortion. After embedding, low dimensional structures emerge that can be interpreted as task-related hemodynamic responses, or non-task-related physiological rhythms.

1 Introduction

Functional brain imaging utilizes the coupling between local electrical activity and regional changes in blood flow and blood oxygenation level. FMRI datasets consist of time series of three-dimensional volumes acquired while a subject is submitted to sensory or cognitive stimulations inside an MRI scanner. The goal of the analysis is to detect among the N voxels the "activated" voxels i where the dynamic changes in the fMRI signal $\mathbf{x}_i(t), t = 0, \cdots, \mathsf{T}-1$ can be considered to be triggered by the stimulus. In the absence of any detailed substantive understanding of the mechanism of the fMRI response, exploratory methods have been proposed for the analysis of fMRI datasets. Principal and independent component analysis (PCA and ICA) have been used to identify non-trivial components, which help to understand the nature of the confounds and generate new hypotheses [7]. Of course, the orthogonality imposed by PCA and the statistical independence required by ICA are very strict constraints with no physiological interpretation. Interestingly, empirical evidence from fMRI data analysis using both PCA and ICA has shown that only the top few components provided by the decomposition are useful for the analysis.

In this work we view each time series as a point in $\mathbb{R}^{\mathsf{T}}$, where T is the number of time samples. We will take advantage of the fact that there are subsets of fMRI time series that have a much smaller intrinsic dimensionality than T. Instead of projecting the time series on a linear subspace we propose to analyze the nonlinear structure formed by the set of fMRI time series. The geometry of the dataset is defined in terms of how the different time series organize themselves with respect to one another in $\mathbb{R}^{\mathsf{T}}$. For fMRI data analysis, we are particularly interested in the organization of the activated time series into a connected region of $\mathbb{R}^{\mathsf{T}}$, and the relative position of this region with respect to the rest of the dataset. We will show in this work that the set of activated time series in a single fMRI dataset forms a distinct low dimensional structure, which can be easily separated from the rest of the time series.

The contribution of this paper is a novel method to chart functional maps, that are build globally based on the "functional" geometry of the fMRI dataset. This method relies on nonlinear mapping techniques to build a new parametrization of an fMRI dataset, and performs a clustering of the low dimensional representation. The new parametrization results in a clear separation of the time series into: (1) response to stimulus, (2) coherent physiological signals, (3) artifacts, and (4) background time series. We have performed a detailed evaluation and comparison to linear and nonlinear techniques using *in vivo* datasets.

2 Overview of the approach

At a microscopic level, a large number of internal variables associated with various physical and physiological phenomena contribute to the dynamic changes in the fMRI signal inside an activated region. Because the fMRI signal is a large scale (as compared to the scale of the neurons spike trains) measurement of neuronal activity, we expect that many of these variables will be coupled resulting in a low dimensional configuration of the activated fMRI signal. We assume therefore that

[*] This paper appeared in the Data Mining for Biomedical Informatics Workshop which was held in conjunction with the 7th SIAM International Conference on Data Mining on April 28, 2007 in Minneapolis, MN.

[†] Center for the Study of Brain, Mind and Behavior ; Program in Applied and Computational Mathematics, Princeton University, http://www.princeton.edu/~fmeyer, fmeyer@princeton.edu

[‡] Electrical Engineering, University of Colorado at Boulder

the activated fMRI time series vary smoothly as a function of a small number of "hidden" variables, and form a well defined low dimensional structure. This hypothesis is completely consistent with the fact that parametric models, with a small number of parameters have been successful in modeling the hemodynamic response [8]. In addition, we also assume that there exist other well-defined physical or physiological processes that give rise to low-dimensional subsets of time series.

In this paper, we call *source signals* the time series that are generated by physical or physiological processes that have an intrinsic low dimension. Example of *source signals* include task-related hemodynamic responses, non-task-related physiological rhythms (breathing and heart beating motion). We expect that only a small fraction of time series in an fMRI dataset will be *source signals*. We call the time series that are not associated with any *source signals* background time series. As shown in our experiments with *in vivo* data, background time series cannot be approximated with a small number of parameters, and do not form low dimensional subsets of $\mathbb{R}^T$.

We can take advantage of the low dimensionality of the subsets of *source signals* to construct a new low dimensional parametrization of the dataset. This new parametrization will allow us to cluster the dataset into the various *source signals*. The curse of dimensionality forces us to perform the clustering using a low-dimensional parametrization of the dataset. As we construct a new parametrization of the dataset, it is critical to preserve local distances between time series: time series that are functionally close should remain close in the new parametrization. Formally, the construction of the new parametrization amounts to finding a function Φ such

- $\Phi(\mathbf{x}_i)$ has less degrees of freedom than $\mathbf{x}_i$,

$$\mathbf{x}_i \in \mathbb{R}^T,\ \Phi(\mathbf{x}_i) \in \mathbb{R}^K, \quad \text{with} \quad K \ll T;$$

- Φ minimizes the local distortion introduced by the mapping,

$$\text{if} \quad \|\mathbf{x}_i - \mathbf{x}_j\| \quad \text{is small, then} \\ \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\| \approx \|\mathbf{x}_i - \mathbf{x}_j\|.$$

Most existing methods for reducing the dimensionality of fMRI datasets are linear: the time series $\mathbf{x}_i$ are projected onto a set of (fixed or adaptively chosen) basis functions. The resulting coefficients serve as the new coordinates in the low dimensional representation. However, in the presence of nonlinearity, a linear mapping may fail to preserve local proximity, distorts the geometry of the dataset, and make the clustering difficult. For this reason we propose to use nonlinear maps to reduce the dimensionality of the dataset. Our experiments with *in vivo* data confirm that the subsets formed by the *source signals* have a nonlinear geometry that cannot be captured by standard linear methods.

In summary, our proposed approach relies on the following three steps:

1. Construction of a new parametrization.

2. Clustering of the dataset in the low dimensional space formed by the new parametrization.

3. Identification of the set of activated time series.

3 Random walk on the dataset

Given a $T \times N$ fMRI dataset $\mathbf{X} = \{\mathbf{x}_i(t)\}$, we define a random walk on the dataset. First, we replace the dataset by a graph G: the time series $\mathbf{x}_i$ originating from the voxel i becomes the vertex i of the graph (we slightly abuse notation here: i is the spatial index of the voxel, as well as the vertex index). Edges are defined by the n nearest neighbors (according to $\|\mathbf{x}_i - \mathbf{x}_j\|$). The weight $w_{i,j}$ on the edge $\{i, j\}$ quantifies the functional proximity between i and j,

$$w_{i,j} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2},$$

if $\mathbf{x}_i$ is connected to $\mathbf{x}_j$, and $w(\mathbf{x}_i, \mathbf{x}_j) = 0$ otherwise.

We consider a random walk Z_n on the graph with transition probability $\mathbf{P}$, defined by $P_{i,j} = w_{i,j}/d_i$, where $d_i = D_{i,i} = \sum_j w_{i,j}$ is the degree of the vertex i. The transition probability of the Markov random walk in s steps is $\mathbf{P}^s$. As s goes to infinity, $\mathbf{P}^s$ converges to the stationary distribution $\boldsymbol{\pi}$.

We define a similarity measure between any two vertices i and j. The similarity should be able to distinguish between strongly connected vertices (the two voxels belong to the same functional region) and weakly connected vertices (the two voxels have similar time series but belong to different functional regions). For this purpose, we consider the average commute time, $\kappa(i,j) = H(j,i) + H(i,j)$, which is a a symmetric version of the average hitting time from i to j,

$$H(i,j) = E_i[T_j] \quad \text{with} \quad T_j = \min\{n \geqslant 0; Z_n = j\}.$$

The notation E_i indicates that the random walk is started at i, $Z_0 = i$. The average commute time quantifies the expected path length between i and j.

We can show that the commute time defines a distance on the graph. This distance can be compared to the standard distance δ on the graph.

Proposition 3.1. *If i and j are at a distance $\delta(i,j)$ on the graph, then*

$$2\delta(i,j) \leqslant \kappa(i,j) \leqslant C\delta(i,j),$$

where $C = \max_{i,j}, Q(i,j)^{-1}$ *and* $Q(i,j) = \pi_i P_{i,j}$.

We note that the Markov chain defined on the graph is reversible, so $Q(i,j) = Q(j,i)$. The lower bound is obvious. The upper bound is a direct consequence of the following result.

Lemma 3.1. *If i and j are two adjacent vertices, then*

$$\kappa(i,j) \leqslant Q(i,j)^{-1}.$$

Proof. From the strong Markov property [3], we have

$$\kappa(i,j) = \frac{1}{\pi_i P_i(T_i < T_j)}.$$

But if i and j are adjacent, the probability that, starting at i, j is visited before returning to i, $P(T_j < T_i)$, is greater than the probability that, starting from i, the walk visits j at the next instant. So,

$$\kappa(i,j) = \frac{1}{\pi_i P_i(T_i < T_j)} \leqslant \frac{1}{\pi_i P_{i,j}}.$$

Unfortunately, the constant C can be quite large. For instance, if we choose uniform weights, $w_{i,j} = 1$, then $P_{i,j} = 1/d_i$, and $\pi_i = d_i/(2|\mathcal{E}|)$, where $|\mathcal{E}|$ is the cardinality of the set of edges $\mathcal{E}$. In this case, we have $Q_{i,j} = 1/(2|\mathcal{E}|)$, and $C = 2|\mathcal{E}|$.

The commute time can be conveniently computed from the eigenfunctions $\phi_1, \cdots, \phi_N$ of

$$\mathbf{N} = \mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}, \tag{3.1}$$

with the eigenvalues $-1 \leqslant \lambda_N \cdots \leqslant \lambda_2 < \lambda_1 = 1$. Indeed, we have

$$\kappa(i,j) = \sum_{k=2}^{N} \frac{1}{1-\lambda_k}\left(\frac{\phi_k(i)}{\sqrt{d_i}} - \frac{\phi_k(j)}{\sqrt{d_j}}\right)^2.$$

As proposed in [2, 5], we define an embedding

$$i \mapsto I_k(i) = \frac{1}{1-\lambda_k}\frac{\phi_k(i)}{\sqrt{d_i}}, \quad k = 2, \cdots, N \tag{3.2}$$

Because each ϕ_k is also an eigenfunction of the Laplacian [1, 11], it minimizes the "distortion"

$$\min_{\phi,\|\phi\|=1} \sum_{i,} \sum_{j \neq i} (\phi(i) - \phi(j))^2$$

with ϕ_k orthogonal to $\{\phi_0, \phi_1, \cdots, \phi_{k-1}\}$. In practice, the reduction of dimensionality is achieved by using only the first K eigenfunctions. The construction of the embedding is summarized in Fig. 1.

```
Algorithm 1: Construction of the embedding
  Input:
    – $x_i(t), t = 0, \cdots, T-1, i = 1, \cdots, N$,
    – $\sigma$ (see (3)); $n$: number of nearest neighbors.
    – $K$: number of eigenfunctions.
  Algorithm:
    1. construct the graph defined by the $n$ nearest
       (according to $\|x_i - x_j\|$) neighbors of each $x_i$
    2. compute $\mathbf{P}$, and $\mathbf{N}$ defined by (3.1).
    3. find the first $K$ eigenfunctions, $\phi_k$, of $\mathbf{N}$
  • Output: For all $x_i$
    – new co-ordinates of $x_i$ are given by (3.2).
```

Figure 1: Construction of the embedding

The spatial co-ordinates of the voxels from which the time series originate are not used in the computation of $w_{i,j}$. We know that spatial information can be useful: truly activated voxels tend to be spatially clustered. However, because of the complexity of the cortical surface, neighboring voxels may belong to different functional units, thus produce entirely different responses.

3.1 Determining the dimensionality In our problem, we do not intend to control the residual variance for the entire dataset, however we have found in the experiments that the first few eigenfunctions ϕ_k largely contribute to describing the variation within subsets of low dimensionality, while the subsequent ϕ_k are dedicated to modeling the variation of the background time series. At time t, we reconstruct an image of N voxels from K projections and compute the residual energy at voxel i, $\varepsilon_i(K) = \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2/|\mathbf{x}_i\|^2$. We can average the residual energy over a region $\mathcal{R}$, and compute $\varepsilon_{\mathcal{R}}(K) = \sum_{i \in \mathcal{R}} \varepsilon_i(K)$. We expect $\varepsilon_{\mathcal{R}}(K)$ to decay fast with K when the region $\mathcal{R}$ is well approximated by $\phi_1, \cdots, \phi_K$. For instance, in Fig. 4 the red and blue clusters (activated time series) are well approximated with $K = 4$ eigenfunctions, while the green cluster (background time series) is poorly approximated.

3.2 k-Means clustering We use the k-Means clustering algorithm for partitioning the low dimensional parametrization of the dataset. The k-Means algorithm does not estimate the number of clusters. Because of the existence of various *source signals* other than the task-related hemodynamic response, it is impractical to specify the number of clusters in advance. We solve this problem heuristically by first over-partitioning the dataset, and progressively reducing the number of clusters, until a stable partition emerges.

4 Experiments

We describe below the results of an experiment with an event-related fMRI dataset. More experiments with synthetic and *in-vivo* data can be found in [9]. We evaluate our approach using two different criteria. First, we compare the parametrization provided by our method with the parametrization generated by ISOMAP and PCA. We look for well defined structures in the new parametrization. We also compare the result of activation detection obtained using our approach with the one obtained by the General Linear Model (GLM).

In [4] the authors study age-related changes in functional anatomy. Subjects were instructed to press a key with their right index finger upon the visual stimulus onset. The stimulus lasted for 1.5 second. Functional images were collected using a Siemens 1.5-T Vision System with an asymmetric spin-echo sequence sensitive to BOLD contrast (volume TR $=$ 2.68 sec, 3.75 $\times$ 3.75mm in-plane resolution; T2* evolution time $=$ 50 msec, alpha $=$ 90°). Sixteen contiguous axial slices were acquired. For each slice, 128 sequential images were obtained consisting of one run of functional imaging. The image resolution is 64 $\times$ 64. For every 8 images, the subjects were presented with one of the two trial conditions. The "one-trial" condition involves one isolated stimulus. The "two-trial" condition has two consecutive stimuli with an inter-stimulus interval of 5.36 sec. The one-trial conditions and two-trial conditions were mixed in a pseudo random fashion. There were 15 trials per run. We use one dataset from a single subject consisting of a single run of experiment. For the data analysis, the first and last four images are discarded. So the total length of time series is 120. Each run of experiment consists of 15 trials, and each trial consists of 8 temporal samples. According to the ordering information, 8 of the 15 trials are of one-trial condition and the other 7 are of two-trial condition. Time series from one-trial and two-trial conditions are averaged separately. Therefore, each voxel is associated with two averaged time series, and $\mathbf{X}$ contains two columns of $\mathrm{T} = 8$ samples for each voxel i. The linear trend is removed from all averaged time series. The results published in [4] show activation in the visual cortex, motor cortex, and cerebellum. We focus our analysis to a volume extracted from the posterior region of the brain which encompasses the visual cortex. The volume includes 4 axial slices, with the original index 7, 8, 9 and 10. There are altogether 1025 intra-cranial voxels within the selected region. And the number N of time series included for the analysis is 2050 (twice the number of voxels). $\mathbf{X}$ is a 8×2050 matrix.

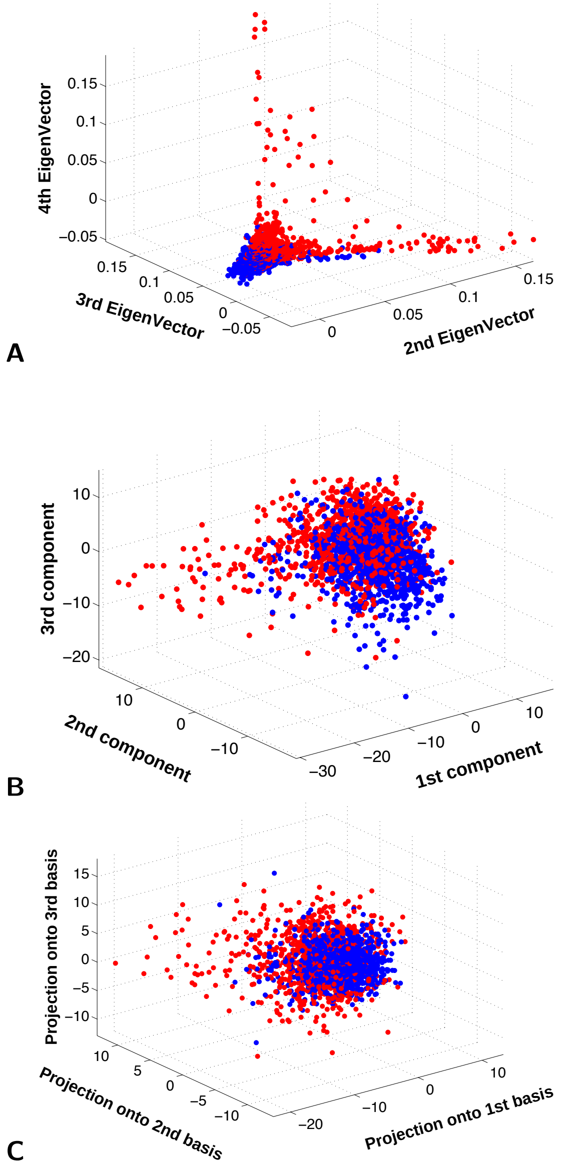

Figure 2: Parametrization given by the embedding (3.2) in **A**, ISOMAP in **B**, and PCA in **C**. Blue points correspond to time series from one-trial condition. Red points correspond to time series from two-trial condition.

Fig. 2-**A** shows the embedding provided by our approach. We note two distinctive "branches" forming a "V" shape. Both "branches" are nearly one dimensional. Because we know if each time series was acquired in the one-trial or two-trial conditions, we color-code the two conditions: blue = one-trial condition ; red = two-trial condition. For comparison purposes, we show the embedding generated by ISOMAP [10] (Fig. 2-**B**), and the embedding provided by PCA using the first three components (Fig. 2-**C**). The representation given by PCA and ISOMAP are less conspicuous than the representation obtained by our approach. In Fig. 2-**B** and 2-**C**, points from the two-trial are distributed as outliers to the main point cloud. We cannot see any low dimensional structures in these two plots.

The outcome of the clustering is shown in Fig. 3. The plot of the residual energy (Fig. 4) indicates that eigenfunctions ϕ_2 and ϕ_4 create the largest drop in the residual energy, and should be the most useful.

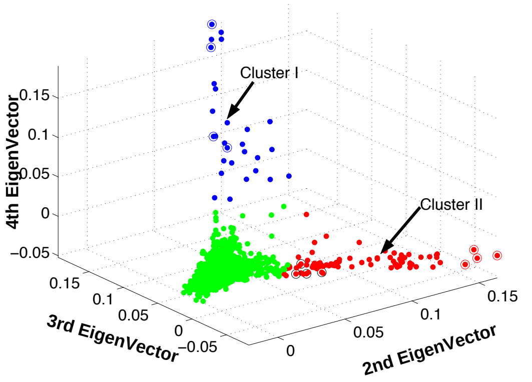

Figure 3: Clustering into 3 classes: cluster I (blue), II (red), and background (green). Time series associated with points marked by a circle are shown in Fig. 5.

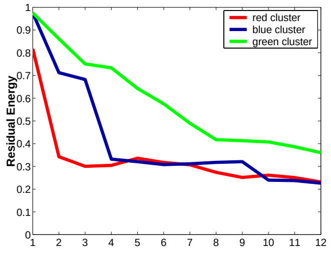

Figure 4: Average residual energy for each cluster (red=II, blue=I, green=background) as a function of K.

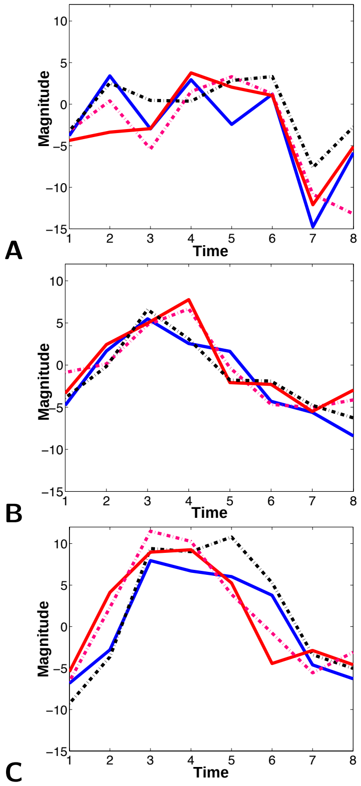

Figure 5: Four representative time series from cluster I (**A**) , cluster II with one-trial condition (**B**), and from cluster II with two-trial condition (**C**).

We have selected three groups of four time series (identified by circle in the scatter plot in Fig. 3) and have plotted them in Fig. 5. There are two groups of time series selected from cluster II. In Fig. 3, red points with red circles are from the two-trial condition and are located at the tip of the branch. Red points with black circles are from the one-trial condition, and are located at the border with the background cluster. Time series from cluster I all have an abrupt dip at $t = 7$. The corresponding voxels are located along the border of the brain (see Fig. 7). It is most likely that time series from cluster I all suffer from a motion artifact.

Time series from cluster II clearly have a shape similar to an hemodynamic response, and the corresponding voxels are located in the visual cortex (see Fig. 7). Therefore, we assume that cluster II contains times series triggered by the stimulus. Moreover, the embedding has sorted the time series along the branch of cluster II from two-trial condition (strong response) at the tip, to one-trial condition (weak response) at the stem (close to the background time series). The embedding has preserved the intrinsic organization of the activated time series, and has organized the time series according to the strength of the activation. This is a remarkable result since no information about the stimulus, or the type of trial was provided to the algorithm.

Each eigenfunction ϕ_k can be viewed as a function defined on the vertices of the graph. Therefore we can plot each of the eigenfunctions as an image on the original dataset. For instance, Fig. 6 shows the (color-coded) graph of ϕ_2 for the region of analysis in the four different slices. The majority of the voxels take negative value (blue). However, a few voxels take positive values (red and orange). The nodal lines where ϕ_2 changes sign are localized around the area of activation. Of course, we can check in Fig. 3 that the activated time series have a positive ϕ_2 co-ordinate. In fact, ϕ_2 is known as the Fiedler vector and is used for optimally splitting a dataset into two parts.

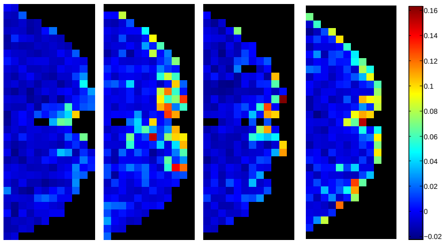

Figure 6: ϕ_2, as a function on the dataset.

Fig. 7 shows the location of the voxels (corresponding to the time series) of cluster I (blue) and II (red). We also computed the activation map obtained using the GLM. The averaged time series from the two-trial condition are used for the regression analysis. We use the hemodynamic response function defined in [6] for the regression. We threshold the p-value at 0.005, and the activation map is shown in Fig. 8. The two activation maps are completely consistent with one another.

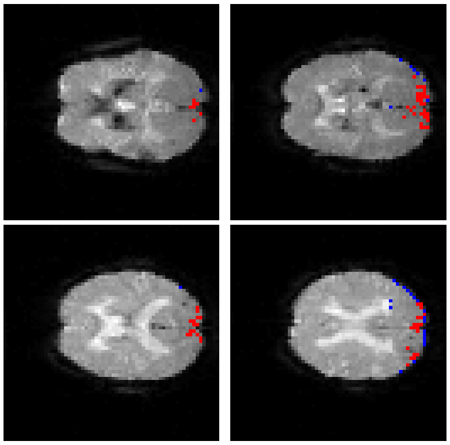

Figure 7: Activation maps: voxels of cluster I (blue); voxels of cluster II (red).

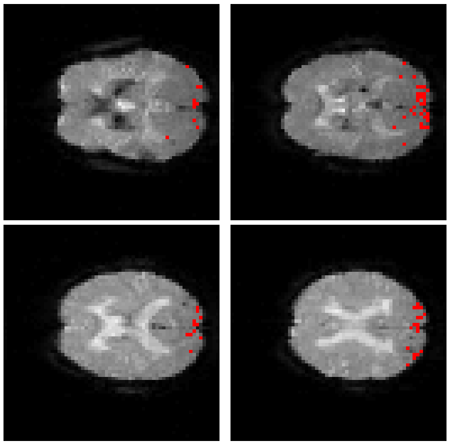

Figure 8: Activation maps obtained using the GLM (p-value = 0.005).

References

[1] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computations **15** (2003), 1373–1396.

[2] P. Bérard, G. Besson, and S. Gallot, *Embeddings Riemannian manifolds by their heat kernel*, Geometric and Functional Analysis **4(4)** (1994), 373–398.

[3] P. Bremaud, *Markov chains*, Springer Verlag, 1999.

[4] R.L. Buckner, A.Z. Snyder, A.L. Sanders, M.E. Raichle, and J.C. Morris, *Functional brain imaging of young, nondemented, and demented older adults*, Journal of Cognitive Neuroscience **12** (2000), 24–34.

[5] R.R. Coifman and S. Lafon, *Diffusion maps*, Applied and Computational Harmonic Analysis **21** (2006), 5–30.

[6] A. M. Dale and R. L. Buckner, *Selective averaging of rapidly presented individual trials using fMRI*, Human Brain Mapping **5** (1997), 329–340.

[7] K. M. Petersson, T.E. Nichols, J-B Poline, and A.P. Holmes, *Statistical limitations in functional neuroimaging I. Non-inferential methods and statistical models*, Phil. Trans. R. Soc. Lond. B (1999), no. 354, 1240–60.

[8] ———, *Statistical limitations in functional neuroimaging II. Signal detection and statistical inference*, Phil. Trans. R. Soc. Lond. B (1999), no. 354, 1261–81.

[9] X. Shen and F.G. Meyer, *Low dimensional embedding of fMRI datasets*, Submitted, available online at `ece.colorado.edu/~fmeyer`, 2007.

[10] J. Tenenbaum, V. de Silva, and J. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), 2319–2322.

[11] B. Thirion and O. Faugeras, *Nonlinear dimension reduction of fMRI data: the Laplacian embedding approach*, IEEE International Symposium on Biomedical Imaging, 2004, pp. 372–375.