Group analysis in functional neuro-imaging : selecting subjects using global spatial and temporal similarity between subjects

Ferath Kherif[1,2,4], Jean-Baptiste Poline[1,2],
Sébastien Mériaux [1,2], Habib Benali[2,3], Guillaume Flandin[5,1] Matthew Brett[4]

1: Service Hospitalier Frédéric Joliot, CEA, Orsay, France

2: IFR 49, Institut d'Imagerie Neurofonctionnelle, Paris, France.

3: INSERM U 494, CHU Pitié-Salpêtrière, Paris, France.

4: MRC Cognition and Brain Unit, Cambridge, United Kingdom.

5: INRIA, Epidaure Project, Sophia Antipolis, France.

Short Title Multivariate Model Specification For fMRI Data

Keywords Group analysis, multivariate analysis, statistical analysis, fMRI, brain imaging method.

Address for correspondence

Jean-Baptiste Poline, Service Hospitalier Frédéric Joliot, CEA/DRM,
4, Place du Général Leclerc, 91401 ORSAY Cedex, France

Tel +33 1 69 86 78 37

Fax +33 1 69 86 77 86

Email poline@shfj.cea.fr

Abstract

Standard group analyses of fMRI data rely on spatial and temporal averaging of individuals. This averaging operation is only sensible when the mean is a good representation of the group. This is not the case if subjects are not homogeneous, and it is therefore a major concern in fMRI studies to assess this group homogeneity. We present a method that provides relevant distances or similarity measures between temporal series of brain functional images belonging to different subjects. The method allows a multivariate comparison between data sets of several subjects in the time or in the space domain. These analyses assess the global inter-subject variability before averaging subjects and drawing conclusions across subjects, at the population level. We adapt the *RV-coefficient* to measure meaningful spatial or temporal similarities and use multidimensional scaling to give a visual representation of each subject's position with respect to other subjects in the group. We also provide a measure to detect subjects that may be outliers. Results show that the method is a powerful tool to detect subjects with specific temporal or spatial patterns, and that, despite the apparent loss of information, restricting the analysis to an homogeneous sub-group of subjects does not reduce the statistical sensitivity of standard group fMRI analyses.

1 Introduction

During a typical fMRI experiment, several data sets are collected from different subjects scanned during a common experimental paradigm. At the data analysis level, a decision must be taken on how to extract relevant information from this vast amount of data. The choice of the methods for statistical analysis depends mainly on the sort of inference required. If the results are to be restricted to the subjects studied, then fixed effect analyses can be used ; random effects analyses are needed for inference to a parent population from which the subjects are drawn (Holmes and Friston, 1998; Friston et al., 1999a,b; Petersson et al., 1999). Generally, in the cognitive neuroscience community, one wishes to extend the results to a parent population (Mandzia and Black, 2001; Detre and Floyd, 2001; D'Esposito, 2000; Cabeza and Nyberg, 2000; Posner and Raichle, 1998).

This introduces two important concepts : the *population* and the *sample*. The population consists of the entire group which we wish to investigate. The sample represents a subset of this population which may well be selected for its homogeneous characteristics in relation to the aim of the study (left or right hand preference, age, ability for languages, and so on). If the individuals in the sample have been *drawn randomly* from the population, the characteristics of the sample are likely to reflect the characteristics of the entire population (Cochran, 1977; Govindarajulu, 1999) and statistical results obtained in the sample can be generalised to the population of interest. The sampling step is therefore critical in determining to what extent findings can be generalised. Only representative samples will ensure the validity of the inference.

In neuro-imaging, random effect models that average the sample are the most popular techniques used to make these inferences that take into account the inter-subject variability. Random effect group analyses (and most other group analysis techniques) depend on the assumption that the subjects are *independently identically distributed* and have been drawn from the *same homogeneous population* (Petersson et al., 1999). In particular the subjects are expected to have the same common mean which will be very close to the true mean of the population. The main drawback is that the mean is sensitive to outliers (Brammer et al., 1997), and moreover, the mean is not a very useful statistic for representing multimodal distributions or non-homogeneous

populations. Thus, if there is an important *known factor* that is not taken into account (say right or left hand preference), the pooling operation would be invalid.

The problem of pooling across sub-groups is best illustrated by example. Imagine that 50% of the population use an area in the right frontal cortex for a memory task, while deactivating the same area on the left. The other half of the population have the opposite pattern, activating the left but deactivating the right. If we average activation across a randomly selected group of subjects, we are likely to conclude that frontal lobe activation does not change for this memory task, even though this is not true for any subject in the population. So, if one could demonstrate that there are two different sub-groups in the population, it would clearly be preferable to conduct an analysis for each group.

These problems are not limited to the spatial domain but can also arise in the temporal domain where a common behaviour is also most often assumed. For example, Burbaud et al. (2000) shows that different activation patterns for mental calculation are obtained depending on the strategy (verbal or visual), while Aguirre et al. (1998), White et al. (2001) present results where subjects show variability in the time domain for a common experimental paradigm.

There are several unknown or uncontrolled factors that can contribute to the fact that the *mean* sampled group will not be representative of the parent population. One is anatomical variability (Brett et al., 2002): extreme examples of brain anatomy can lead to unusual patterns of functional activation. However, even with similar anatomy, differences in subject strategies or mental states may result in different activation with regard to the paradigm. Lastly, any undetected scanning problem can produce outliers in terms of signal to noise ratio or other artefacts.

The goal of this work is to provide a method which can investigate the homogeneity of the sample subjects' datasets. To do so, a measure of similarity between the subjects has to be defined. This measure should indicate which subjects are similar to each other : if the subjects' data can be clustered into separate sub-groups, this means that there probably is not a single underlying population. The measure should also indicate which subjects should be considered as outliers.

Finding a comparison measure for subjects fMRI data sets is a challenging task. Firstly, this

comparison may have to be done with respect to the experimental paradigm under study. Secondly, the datasets are huge (at least several hundred 3D volumes per subject) and complex, because they lie in a high dimensional space and exhibit both spatial and temporal correlations. The measure chosen should therefore be based on the multivariate nature of the data. Lastly, to be practical, the method has to be computationally efficient and the measure should be easy to interpret – for example it should be apparent if the distance between two datasets is due to temporal or spatial differences.

There are many previous papers that partially address these issues using multivariate methods : *Fisher's linear discriminant analysis* (Tegeler et al., 1999), *Principal component analysis* ((Moeller et al., 1987; Strother et al., 1995a,b; Hansen et al., 1999; Strother et al., 2002)), *Partial least square* (McIntosh et al., 1996), *Independent Component analysis* (Calhoun et al., 2001; Nybakken et al., 2002) or *Multidimensional scaling* (Welchew et al., 2002). Several univariate methods have also been proposed for example *test-retest methods* Maitra et al. (2002); Genovese et al. (1997) or *Meta-analysis* procedure (Fox et al., 1999; Xiong et al., 2000). The related question on how to combine inter-subject information is treated by (Lazar et al., 2002; Brammer et al., 1997). However, to our knowledge, no previous method has been specifically designed to investigate group homogeneity. The purpose of this work is to propose such a method, that is applicable in the spatial and temporal domains.

The rest of the paper is organised as follows: in the methods section, we present our chosen measure and describe how the measure can be applied in the spatial and temporal domains. We discuss the use of MultiDimensional Scaling (MDS) for displaying the results for a group of subjects. Next, we apply the method to an experimental dataset of 10 subjects who performed a mental computation task. These subjects are analysed for their homogeneity in the spatial and temporal domain. We present methods for deciding which subjects may be outliers using a graphical tool and the Cook distance measure (Seber, 1977; Christensen, 1987), and compare the original group results with those obtained on a more homogeneous group after an outlier has been removed. The discussion section describes possible extensions and applications of these techniques.

2 Methods

2.1 The *RV-coefficient* as a similarity measure

A large number of different similarity measures have been proposed according to the type of data studied. Our approach falls in the domain of multidimensional statistics that try to answer the following question : if n observations are described by p variables on the one hand and q other variables on the other hand, how does one compare these observations? The method that we suggest was described by Escoufier (Robert and Escoufier, 1976) and is based on the calculation of a coefficient of similarity called the *RV-coefficient*. In this section, we describe this coefficient and present its adaptation to brain imaging data.

Let us imagine that we have n observations of p variables on one individual, and n observations of q variables on another. These result in two matrices Y_1 $(p \times n)$ and Y_2 $(q \times n)$ corresponding to subject 1 and subject 2. The *RV-coefficient* (Robert and Escoufier, 1976) is a measure of the closeness between the two configurations (or patterns) of points in Y_1 and Y_2 that can be viewed as an extension of the correlation coefficient for multivariate data. This measure is based on a comparison of summary matrices representing the individual patterns. The summary matrices Z_{11} and Z_{22} are derived from Y_1 and Y_2 by computing respectively $Y_1'Y_1$ and $Y_2'Y_2$ (Y' is the transposed matrix of Y). If we assume that the raw data are first mean centered then Z_{11} and Z_{22} simply correspond to the sampled covariance matrices (to a multiplication factor of $1/n$). The *RV-coefficient* is defined by:

$$RV(Y_1, Y_2) = \frac{trace(Z_{11}Z_{22})}{\sqrt{trace(Z_{11}Z_{11})}\sqrt{trace(Z_{22}Z_{22})}} \tag{1}$$

It measures the similarity of the two covariance matrices Z_{11} and Z_{22}. Note that Z_{11} and Z_{22} are translation and rotation independent and can be associated with an operator for which we can derive an inner product (and a distance metric) based on the Hilbert-Schmidt norm. If we denote the norm as $|A|_2 = \sqrt{trace(A'A)}$ and the scalar product as $< A, B >= trace(A'B)$, for the square matrices A and B, the *RV-coefficient* can be written as :

$$RV(Y_1, Y_2) = \frac{< Z_{11}, Z_{22} >}{|Z_{11}|_2 |Z_{22}|_2} \tag{2}$$

In this context, the *RV-coefficient* can be seen as the square of the cosine of the angle between Z_{11} and Z_{22}, thus it can be considered as a multivariate extension of the classical Pearson correlation coefficient. This coefficient is bounded between 0 and 1. A value close to one indicates a high degree of similarity between two datasets Y_1 and Y_2. If RV is null, the two sets are independent if normally distributed[1]. When RV is exactly one, then the eigen-components of data set Y_1 can be derived from Y_2 through an homothetic transformation (for a proof, see Lavit, 1984). Note, however, that this coefficient only considers linear relationships between two data sets. This may seem a strong assumption but the linear part is likely to be predominant when comparing two fMRI datasets.

The RV similarity measure can be transformed to a distance measure. For example, when comparing the configuration of n points, the distance D(Y_1,Y_2) can be defined as :

$$D(Y_1, Y_2) = ||\frac{Z_{11}}{\sqrt{trace(Z_{11}^2)}} - \frac{Z_{22}}{\sqrt{trace(Z_{22}^2)}}|| \tag{3}$$

$$D(Y_1, Y_2) = \sqrt{2}\ \sqrt{1 - \frac{trace(Z_{11} Z_{22})}{\sqrt{trace(Z_{11}^2)}\sqrt{trace(Z_{22}^2)}}} \tag{4}$$

$$D(Y_1, Y_2) = \sqrt{2}\ \sqrt{1 - RV_{Y_1, Y_2}} \tag{5}$$

The computation of the RV coefficient described above allows the comparison of the n observations in each data set. Such methods that look at the relationship between observations across variables are called *Q-mode* analysis in the multivariate terminology. However, the RV can be computed as long as one dimension is common to the two datasets Y_1 and Y_2. When the number of variables are equal in the two data sets ($p = q$) a similarity measure can be computed comparing variables across observations. This can simply be done by considering Y_1' and Y_2', the transpose matrices of Y_1 and Y_2 and then computing $RV(Y_1', Y_2')$. By comparison, we call

[1] If the data are not normally distributed, RV can be null and the data not independent.

this *R-mode* analysis. In this case, the RV coefficient measures the similarity of the $(p \times p)$ covariance matrices of the variables of two datasets, Z_{11} and Z_{22}, obtained respectively from the cross-product matrices $Y_1 Y_1'$ and $Y_2 Y_2'$. Thus, the RV coefficient allows two modes of analysis: *Q-mode*, which investigates the interactions of observations between datasets, and *R-mode*, which reveals the interaction among the variables.

The coefficient has other nice features. Escoufier demonstrated that several multivariate methods can be seen as processes that maximise the *RV-coefficient* of various matrices. For example, PCA analysis of a matrix Y can be viewed as finding an orthogonal transformation matrix L such that RV(Y,LY) is maximum and LY is orthogonal. The maximisation of the RV coefficient RV($L_1 Y_1$,$L_2 Y_2$) is equivalent to a Canonical Correlation Analysis (CCA) of Y_1 and Y_2. L_1 and L_2 are the canonical components, linear combinations of the initial variables in each dataset. When Y_2 is replaced by a design matrix X, this maximisation corresponds to PLS. It easy to see that Canonical variate analysis (CVA) and (Multivariate Linear Models) MLM, which are closely related to PLS, can also be described in terms of maximisation of the RV coefficient between the data and the model. All these approaches, which differ in their noise assumptions and the use of prior information (Kherif et al., 2002), can be embedded in the same framework, which is the maximisation of the RV coefficient. The interpretation of the *RV-coefficient* does not fundamentally change with the data distribution since it represents in all cases the closeness of two datasets after a linear transformation.

The *RV-coefficient* has also been used for model selection by Tanaka and Mori (1997) and for principal component sensitivity by Castaño-Tostado and Tanaka (1990).

2.2 Adapting the *RV-coefficient* to fMRI data

This section describes the calculation of the *RV-coefficient* for comparing fMRI data from two subjects. The fMRI data for a subject can be formatted in a matrix Y of size $t \times n$, with t the number of scans and n the number of voxels. In this instance, Y_1 and Y_2 represent the data for subject numbers 1 and 2. If both dimensions are equal for the two subjects (the same number of scans t and same number of voxels n), RV can be computed using time (*Q-mode*) or space

(*R-mode*) as the common dimension.

2.2.1 Temporal and spatial similarity

Spatial similarities

Generally, we can find a common voxel space for all subjects by stereotactic standardisation methods (Brett et al., 2002), such that all subjects have the same number n of voxels to be analysed.

We saw previously that the *RV-coefficient* can be calculated in the space represented by the configuration of n points (*Q-mode*). If these n points represent the voxels, RV(Y_1,Y_2) represents a measure of the spatial disparity between two subjects. A small value of RV(Y_1,Y_2) would be interpreted as showing that the two subjects had activity located in different brain regions regardless of their temporal dynamics.

Temporal similarities

Temporal similarities are computed in a similar way, whenever the data have a common time dimension. This will be the case when the experimental paradigm has been repeated across subjects. We can compute the corresponding *R-mode* RV-coefficient. A small value of RV(Y_1',Y_2') would reflect different temporal patterns whatever the spatial distribution. Note that this allows the computation of temporal distance in the original voxel space, such that there is no need for stereotactic transformation.

2.2.2 Model based *RV-coefficient*

The comparison of raw data across different subjects might not be relevant because of the existence of confounding factors that are not of interest. For instance, low frequency drifts, if not removed from the data, will randomly perturb the similarity measure. More generally, a good similarity measure should focus on the particular aspect of the data that the experimenter defines to be of interest. Here we use the classical framework of the general linear model and represent

the various effects thought to influence the fMRI data by a set of r regressors of dimension t grouped in the matrix X. We therefore extend the *RV- coefficient* to take into account the experimental paradigm X. Thus we replace Y_1 by $X'Y_1$ and Y_2 by $X'Y_2$ in order to address the data projected into the space defined by the model. We can therefore use the corresponding model based temporal $((X'Y_1)(X'Y_2)')$ and spatial $((X'Y_1)'(X'Y_2))$ covariance matrices to calculate the *RV-coefficient*. Often the interest may only be in a sub-space G of the model X – for instance the sub-space defined by the difference between two experimental conditions. In this case, both model and data can be projected onto this sub-space, in a way similar to the PLS method. The model X and the data Y become respectively X_G and Y_G, leading to an *RV-coefficient* tuned to the specific question represented by G.

We also need to deal with the fact that fMRI data are not independent in space or time. We therefore introduce two metrics $\mathcal{M}$ and $\mathcal{N}$ to handle temporal ($\mathcal{M}$) and spatial ($\mathcal{N}$) correlations. $\mathcal{M}$ deals with correlation in the columns of the preceding matrices, and $\mathcal{N}$ with correlation in the rows.

The matrix $\mathcal{M}$ is defined by:

$$\mathcal{M}^{-\frac{1}{2}} = (X_G^{'} V X_G)^{-1/2} \tag{6}$$

$$\tag{7}$$

$\mathcal{M}$ corrects for scaling differences in the model regressors and takes into account the temporal correlation of the data, represented by the time by time matrix V. Our method, and more generally fMRI data analysis, requires knowledge of the matrix V. The correct specification of this matrix is a major subject of research and several methods have been proposed. One solution is first to convolve the time series with a known kernel and derive the autocorrelation matrix from the applied convolution matrix (Friston et al., 2000). More sophisticated methods (Purdon and Weisskoff, 1998; Worsley et al., 2002; Marchini and Smith, 2003) estimate the temporal correlation from the data by assuming some model of the structure in the residuals (generally

an AR model).

Similarly, the $\mathcal{N}$ metric in the *RV-coefficient* corrects for differences in voxel variances. The determination of spatial correlation is complex. It is well known that voxels are spatially correlated, but few methods have been proposed to address this problem. One reason for this is that most of the analysis procedures are univariate. Another is that the inverse of the $n \times n$ spatial covariance matrix, $n >> t$, cannot be estimated accurately from the data. Here, following Worsley et al. (1997), we suppose that the spatial correlation can be represented by a diagonal spatial covariance matrix:

$$\mathcal{N}^{-\frac{1}{2}} \quad = \quad \text{diag}\{\hat{\sigma}_1^{-1}, \hat{\sigma}_2^{-1}, \cdots \hat{\sigma}_n^{-1}\} \tag{8}$$

The values used for the diagonal elements $\hat{\sigma}_1..\hat{\sigma}_n$ are estimates of the residual variance at each voxel after regression against a model X. It is clear that scaling by $\mathcal{N}^{-1/2}$ cannot be seen as a whitening process, as it does not attempt to address correlation between voxels. However, the use of a diagonal matrix makes the method more robust than the estimation of the full matrix. Note that this scaling gives all the voxels the same importance; without scaling, it is possible for a few noisy voxels with high magnitude to dominate the results.

It is important to note that, as long as the various datasets have similar spatial or temporal structures, the exact specification of those covariances has a limited impact on the relative distances between datasets.

In summary, we use the following matrices :

$$\begin{cases} Y_1^* = \mathcal{M}^{-1/2} X_G' Y_{1_G} \mathcal{N}_1^{-1/2} & \text{for the first dataset,} \\ Y_2^* = \mathcal{M}^{-1/2} X_G' Y_{2_G} \mathcal{N}_2^{-1/2} & \text{for the second.} \end{cases} \tag{9}$$

Then, we compute the RV coefficient for each dimension :

$$\begin{cases} RV(Y_1^*, Y_2^*) & \text{for the spatial RV } (Q\text{-}mode), \\ RV(Y_1^{*'}, Y_2^{*'}) & \text{for the temporal RV } (R\text{-}mode). \end{cases} \tag{10}$$

Lastly, these coefficients are transformed into distances using equation 3.

The method based on the *RV-coefficient* involves the computation of the trace of matrices. Due to the large amount of data in an fMRI experiment, computation and data storage can be very cumbersome. Our implementation is designed to avoid direct computation of the products of large matrices. The trace of the product of two matrices is computed as the overall sum of the the Hadamard (elementwise) product of those matrices. Only one pass through the data performed simultaneously for all subjects is needed. Whenever possible, computations are done in the model parameter space, reducing computational cost considerably.

2.3 Distances for several data sets

2.3.1 MultiDimensional Scaling to represent spatial and temporal similarities

For more than two subjects, the *RV-coefficient* is computed in pairwise fashion and transformed to a distance as in equation (3). A symmetric matrix made of the distances between subjects is then obtained. To visualise this distance matrix in a two dimensional space, several procedures can be employed, for example clustering or MultiDimensional Scaling (MDS) (see Gower, 1984). The purpose of multidimensional scaling (MDS) is to provide a visual representation of the distance matrix. MDS plots the subjects as points on a map such that the distances on this map are similar to the original distances. MDS is used to provide an optimal configuration of points in (for example) 2-dimensional space by minimising the mismatch between the distances between the points in the MDS map and the original distances (Torgerson, 1952). However, it is possible for this configuration in two dimensions to be a poor representation of the data. This is measured by the percentage of variance captured by the MDS plot.

If a plane is not able to capture most of the variability of the data, the representation can be

completed with other dimensions. Note that MDS has been used previously in the context of brain imaging in Friston et al. (1996). It is a simple visual representation tool that can show if there is a specific pattern to the distances between subjects, such as the presence of outliers or sub-groups.

The temporal or spatial patterns that lead to the overall observed distances cannot easily be represented (for K subjects there are $K(K-1)/2$ distances). Canonical analysis of the whole data set would be a useful method but is computationally expensive. Here, we chose to show each subject's main pattern by using the first component of an MLM analysis (temporal) or the individual's SPM (spatial).

2.3.2 Outlier detection

The small number of subjects participating in brain imaging studies (usually around 15) does not in general allow the use of classical clustering methods. However, the distance matrix computed above can be used more quantitatively to develop a diagnostic measures for detecting subjects that may be outliers.

The definition of an outlier is an observation that lies far from the rest of data. Recall that, for K subjects, the distance matrix is $K \times K$, where each row (and each column) contains the distances of one subject from each of the other subjects. Hence, the mean of each row (excluding the diagonal element) represents the average distance of this subject from the rest of the group. This mean can be used to judge if this subject may be an outlier.

If $m_1, m_2, ..., m_K$ are these mean distance measures for the K subjects, then detecting an outlier reduces to finding if one or several values $m_1, m_2, ..., m_K$ are large compared to the rest. If the group is homogeneous, we would expect that these values will be approximately equal and centered around the same overall mean. These K values can be ordered from highest to the lowest and plotted, in a similar way to eigenvalue plots in PCA analyses.

As a complement of this plot, statistical tests for outlier detection can also be performed. For more details about the following tests see Hawkins (1980) and Barnett and Lewis (1994). The

so-called discordancy tests represent one class of outlier tests. For example, Grubb's or Dixon tests can be used to detect unusual extreme values. These tests generally assume a normal distribution of the data. Grubb's method provides critical values for testing if the maximum values are likely to occur by chance for K>3. For the data we describe in the following section, where K=10, the critical value for testing the maximum is $G >= 3.5$ with $G = \frac{m_{(K)} - m_{(1)}}{s}$, where $m_{(K)}$ and $m_{(1)}$ are respectively the maximum and minimum values in $m_1, ..., m_K$ and s the data standard deviation.

Another approach is to make use of regression diagnostics – i.e. estimate a regression model (here a simple intercept model) and then check the residuals. More sophisticated diagnostic measures based on the case-deletion approach have been proposed, among them the most popular are Cook distances, DFFITS or DFBETA. These measures are also called influence diagnostics because they consist in measuring the effect on certain parameters of the model when one observation is omitted; here the parameter is the overall mean. We have preferred these diagnostic methods because they are more exploratory and allow us to check the distance values for all subjects and not only the maximum. In this paper, we have chosen to use the well-known Cook distance measure (Cook and Weisberg, 1982) to look for atypical subjects. Cook distances are known to perform better than DFFITS for highlighting influential observations. Because we have a single parameter, Cook distances and DFBETA will provide the same information.

The Cook distance measures the influence of each value in $m_1, ..., m_K$ on the mean over all subjects, denoted μ. Let μ_i the mean over all subjects without subject i; then the Cook distance for each subject i is given by:

$$C_i = \frac{(\mu - \mu_i)^2}{s^2} \tag{11}$$

where s^2 is the standard deviation around the global mean m. The size of C_i is assessed by comparing it to percentiles of the F distribution: $F(1, K-1)$. There are no well established critical thresholds for the Cook distance and in general an observation is said to have a large

influence if C_i is greater than the 50th percentile. One uses the heuristic condition of $C_i > 1$ to detect outliers. This value is close to the median of the F distribution for a large range of degrees of freedom. However, in our case, with only one single parameter and low degrees of freedom, the value of 0.5 is a more appropriate approximation.

Influence diagnostics can be generalized to deal with more than one outlier. One method is to iterate the detection procedure several times; in each iteration the sample is reduced by removing the detected outlier. It is also possible to check for several outliers at the same time, by looking at the joint influence of a set of observations. This is important because, if there are several outliers, detection methods can be subject to masking and swamping effects – outliers may not be detected or "good" observations may be declared as outliers.

2.3.3 Homogeneous sub-group analysis

The purpose of this analysis is not to remove data from the group analysis in order to obtain more favorable statistical results, but to warn about potential problems in the sample and to detect subjects for which the data may not be representative.

We present below the results of analyses conducted both with and without the suspected outlier subjects. To make the comparison easier, we have also analysed the data choosing a more heterogenous sub-group by discarding a less influential observation. Hence this *disparate group* and the *homogeneous group* (without the outlier) have the same degree of freedom (see the results section). If the results differ greatly, then interpretation of the results should be done with caution. The data from subjects identified as outliers may need to be checked further.

In general, before removing an outlier from a set of subjects, several cases need to be considered. First, the data should be checked for artefacts (movements,..) using either multivariate methods (Kherif et al., 2001) or other methods such as in Luo and Nichols (2003). If possible, data can be corrected and the analysis performed again. Another possibility is that the model used to perform the analysis (at the subject level) is not correct (see Kherif et al. (2002) for a method of guiding model specification). If subjects seem to be clustered in several sub-groups, one should

look for the factors that could explain the apparent heterogeneity such as behavioral data. If the data cannot be corrected, or if there are reasons to believe that the outliers belong to another population then the group analysis should be performed with an homogeneous sub-group.

3 Application

This section presents the results of an application of the method to fMRI data. The data consists of ten subjects who underwent two acquisition series.

3.1 Description of the fMRI dataset

We tested our method on a cognitive paradigm that investigates the brain network involved in a mental calculation task. The experiment was conducted on 10 subjects (Simon et al., 2002). During fMRI scanning, six blocks were presented of 13 trials each (26 sec). Each block was preceded by a 4-sec instruction period, and there were alternating blocks of the computation-task and the control-task. Each subject performed two of such sequences. TR was 2 seconds giving a total of 186 scans per subject. The (linear) model used for analysing the data consisted of 3 regressors per condition (computation and control), which were the block regressor convolved with the standard hemodynamic response, and its temporal and spatial derivatives. Within this model, a sub-space of interest X_G was formed using an F-contrast to highlight activations induced by the calculation task relative to the control task. Data were corrected for low frequency drift using a cutoff of 1/120 Hz.

Each subject's data were corrected for movement and stereotactically normalized to the Montreal Neurological Institute (MNI) template. We also apply a spatial smoothing with different kernel values (FWHM = 5, 8 and 12 mm). The smoothing parameters have an impact on the RV coefficient, so that, in general, the similarity between subjects increases with higher smoothing. However, the relative distances between subjects remain similar and conclusions are unchanged. We here present the results with a smoothing kernel value set to 5mm (the same value used in the original paper (Simon et al., 2002)).

The data were also temporally filtered (gaussian FWHM = 4s) and the corresponding convolution matrix used to estimate the temporal correlation matrix V.

3.2 Results

This section presents the results of temporal and spatial comparisons for the data from ten subjects. Here we use the adapted model-based *RV-coefficient* to investigate inter-subject distances with respect to the comparison between activation and control. For this purpose, we use equation (9) and (10), with a sub-space X_G that represents the expected activation.

3.2.1 Temporal distances

Figure 1 (center panel) shows a 2D MDS representation of the temporal distance between subjects. The figure shows that although subjects cannot be easily divided into more than one group, subjects 4 and possibly 3 lie a long way from the group centre of mass, indicating a different temporal behaviour. The representation of the distances between subjects in a space with two dimensions captured approximately 85% of the variance.

In order to illustrate the temporal difference between subjects we applied a Multivariate Linear Model (MLM) analysis (Worsley et al., 1997; Kherif et al., 2002) to the dataset for each subject. MLM can summarize in a small number of components, the the predominant temporal dynamics observed in the whole brain. These temporal differences between subjects are observed in Figure 1. The time series shown in this figure are the first components of the output of the MLM analyses. These components are clearly seen to be similar for two subjects (8,9) that are close on the MDS plot. In contrast, the patterns for subjects 8 and 9 are clearly different from those for subject 4 and subject 3. These two subjects are far away from subjects 8 and 9 on the MDS plot. The temporal behaviour of subjects 3 and 4 may be due to a quadratic-time-by-experiment interaction.

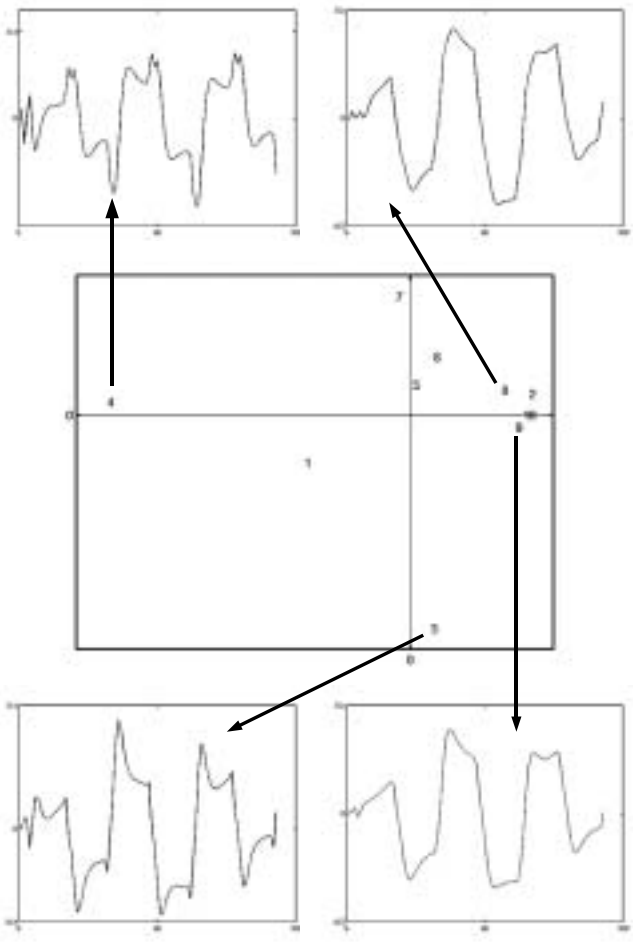

Figure 1: Inter-subject variability in terms of temporal distances (MDS plot in center panel) and illustration of this variability using the first temporal MLM eigencomponents.

3.2.2 Spatial distances

Figure 2 (center panel) shows a 2D MDS representation of the spatial distance between subjects. In this plot, some of the inhomogeneity found in the temporal domain is observed again. In particular, subject numbers 1, 3, and 4 are found to be the farthest from the group center. The spatial distance is illustrated in Figure 2 using a thresholded statistical parametric map ($p < 0.05$ corrected) showing the activation effect: one coronal slice (MNI coordinates $y = 64$) and one axial slice (MNI coordinates $z = 44$) for each subject. Distances between subjects 3 and 4 and subjects 8 and 9 are mainly reflected in a greater bilateral activity in the parietal lobes for the latter. Theses differences can also be observed in the pre-motor areas. The plot only captures 25% of the variance of the distances and should therefore be interpreted with care. Alternatively, extra dimensions can easily be added to construct a 3D plot, or two 2D plots. Note that this representation is only a visual aid and is not used to exclude or classify subjects.

3.3 Outlier identification using Cook test

While the previous plots do not show a clear clustering between two or more groups of subjects, it is not completely clear whether there are subjects that should be considered as outliers. To quantify this, we plot the ordered mean distances and the Cook distance plot (with 0.5 as an heuristic cut-off value, see method section).

In the temporal domain, the figure 3 (*left*) shows that subject number 4 has the highest mean distance from the other subjects. The Cook distance for this subject (Figure 3 *right*) is larger than the cut-off value suggesting an important influence over the mean. Subject 4 should most probably be considered as an outlier in the temporal domain.

In the spatial domain, the mean ordered distance exhibits very little variation across subjects (Figure 4 *left*). Although subject 9 can not be considered as an outlier, figure 4 *right* shows that this subject has an unusual influence over the sampled group population mean because of its small distances from the other subjects. In other words, it is a particularly representative subject.

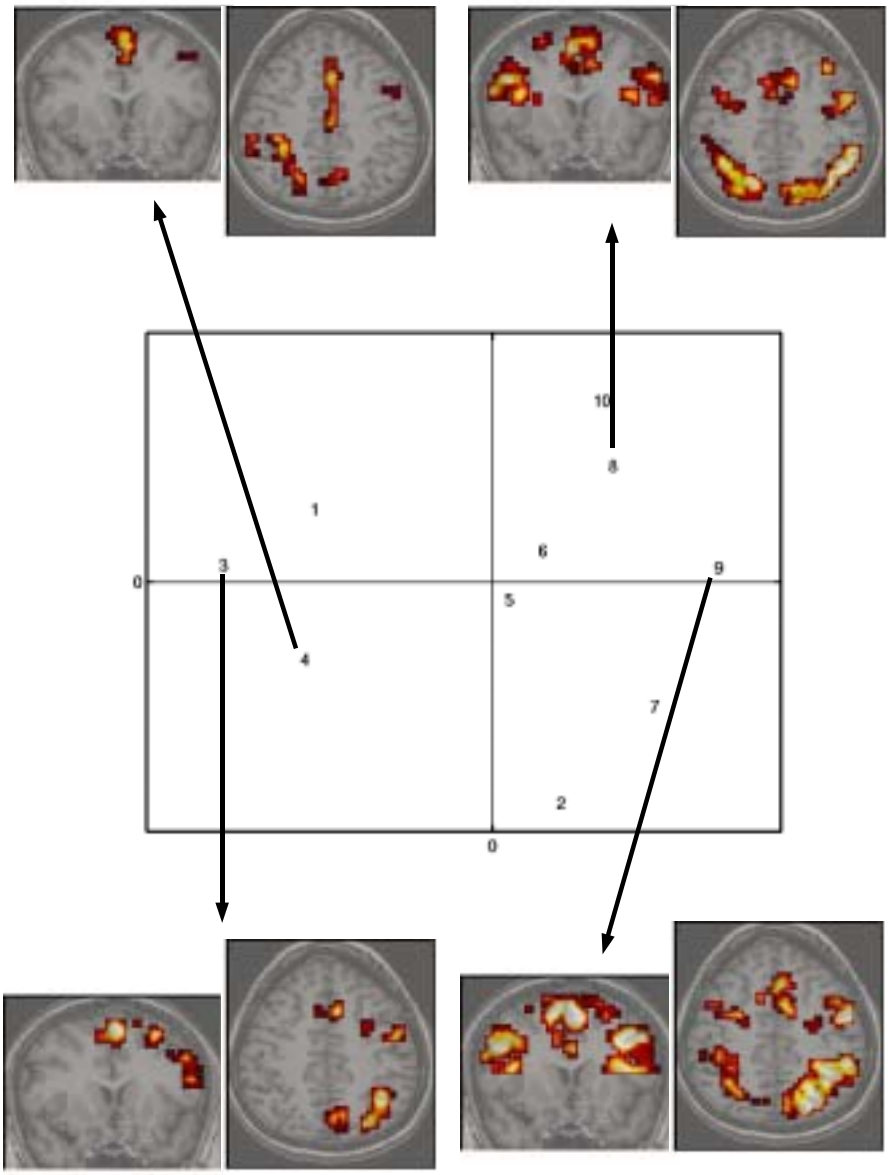

Figure 2: Inter-subject variability in terms of spatial distances (MDS plot in center panel) and illustration of this variability on two slices of a thresholded statistical parametric map ($p < 0.05$) : one coronal slice ($y = 64$ *left*) and one axial slice ($z = 44$ *right*).

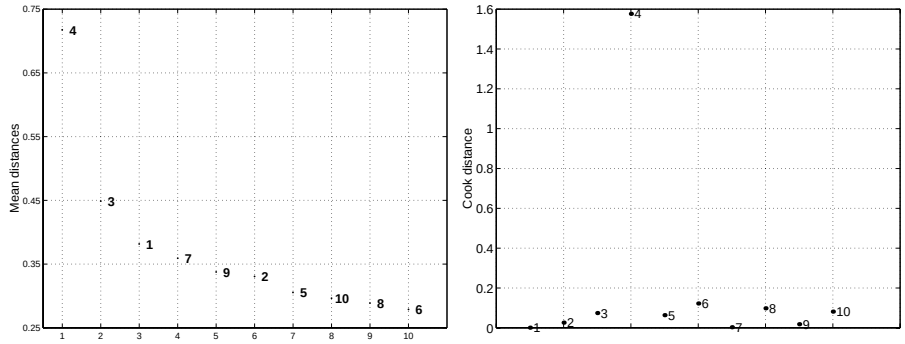

Figure 3: **Temporal similarity** *Left*: Ordered mean temporal distances (similar to a scree plot): subject 4 has the highest mean distance from the other subjects; *Right*: Cook distances: subject 4 is the most influential observation on the group distances.

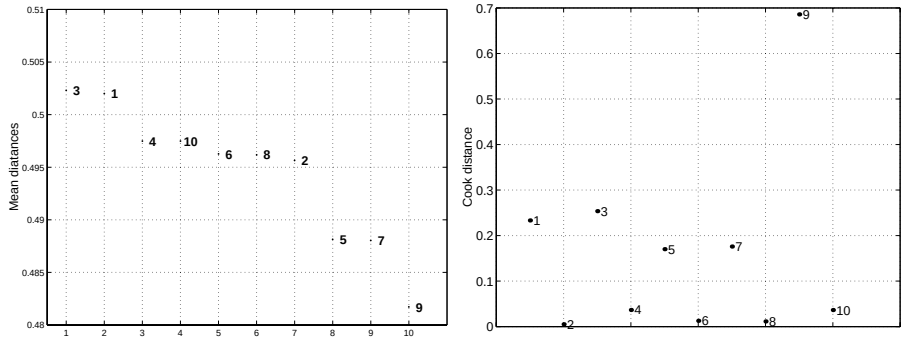

Figure 4: **Spatial similarity** *Left*: Ordered mean spatial distances; *Right*: Cook distances. Subject 9 has the most influence on the group because of its small average distances to the other subjects (see the plot on the *left*).

Comparing the Cook distances in the temporal and spatial domain, we observed that subjects with large influence in one domain are likely to have a large influence in the other dimension.

3.3.1 Homogenous sub-group analysis

Here we follow the method described in section 2.3.3. To demonstrate the potential influence of atypical subjects on the results for the group, figure 5 shows random effect analyses for three different groups: the whole sampled group (*center*); an homogeneous sub-group (*left*), where subject 4 (the outlier subject) is excluded, and an inhomogeneous sub-group (*right*), where subject 7 (one of the less influential subjects in the time domain) is excluded. Results are

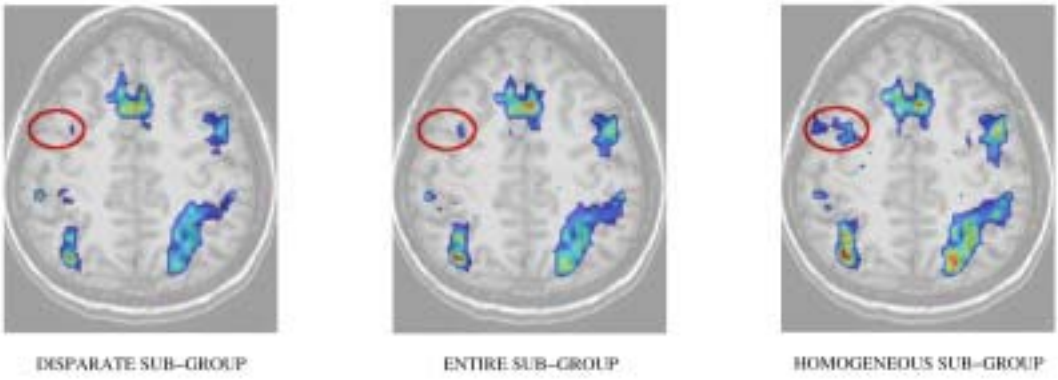

Figure 5: Random effect analyses on three different groups (threshold $p < 0.05$, axial slice $z = 44$) : *left* disparate sub-group formed by all subjects except subject 7, *center*: all subjects, *right*: homogeneous sub-group formed by all subjects except subject 4.

thresholded at a level corresponding to a p-value of 5% (corrected). Clearly, the statistical sensitivity is reduced when the non-homogeneous sub-group is used (see activations in the pre-motor areas in red circle on figure 5). Interestingly, it appears that results obtained with all subjects are also less sensitive than those for the homogeneous sub-group. However, note that it is not the actual sensitivity of the analysis that is at stake but the interpretability of the results when represented by the mean.

4 Discussion and conclusion

The RV coefficient is a multivariate extension of the simple correlation coefficient. We have shown that the RV coefficient can provide a meaningful summary of the spatial or temporal resemblance between the fMRI datasets of two different subjects. We have used this measure to address the difficult problem of assessing the homogeneity of the subjects sampled. If the subject group is not homogenous, then group analyses that aim at giving a representation of the behaviour of the population could be misleading or erroneous. Specifically, we have been able to detect temporal (or spatial) outliers in the sampled sub-population and have showed that exclusion of outliers can lead to qualitatively different results. This allows a better interpretability of the results when represented by the mean. Although the t-test is known to be robust against non-normality, this

statistic is based on the sample mean and variance which are sensitive to outliers when assessed using small samples. In some cases, detection and removal of outliers can yield more sensitive results for a subset of the population.

The technique proposed can be adapted in two main ways. First, the set of voxels chosen to compute either the spatial or temporal distance can vary from application to application. In studying language representation in the brain, one may for instance choose to compare the between-subject spatial distances in the temporal lobes only. Note that when *apriori* knowledge is poor, it may be necessary to use methods that reduce data dimensionality, such as PCA. Second, the choice of temporal representation of functional activity, represented by the linear model X, is also left to the experimenter. This is the case for most neuro-imaging data analysis methods and it is well known that the choice made can seriously influence the results obtained. Multivariate techniques can be used to find optimal models for the data without requiring the input of the experimenter (Kherif et al., 2002).

It is likely that this technique will be useful for other applications. The modified RV- coefficient is general in the sense that it is an association measure between any two data matrices, but it can be easily tuned to specific questions using the extension proposed. This combination of properties makes the RV coefficient useful in many different situations. Below we list some possible applications of this measure in the field of brain imaging.

Comparing several groups of subjects. It would clearly be interesting to apply the technique when there is a specific known clustering of the sampled subjects – for example, based on performance, phenotypic or genotypic information. First, one could check the relevance of the distance measure with respect to the original subject classification, and second, one could use the technique to find the set of relevant subject parameters that would predict the subjects distances and clustering. This can be seen as the dual problem of finding the temporal and spatial patterns best predicting a group classification. A group classification could be qualitative such as 0 or 1, or quantitative on a given scale, or indeed multivariate, for instance made up of several physiological parameters. This is related to discriminant function analysis techniques (Kustra and Strother, 2001). These

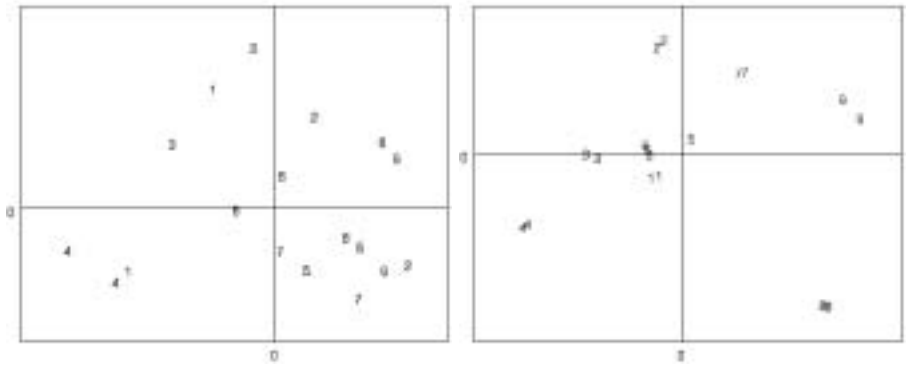

Figure 6: *Left* : MDS plot based on temporal distances for the two sessions of each subject. *Right* : MDS plot based on spatial distances for the two sessions of each subject. Subjects are identified with their number.

kind of studies are useful for characterising the difference between a population of patients and normal subjects (Meyer-Lindenberg et al., 2001). It is worth noting that a test can be easily developed to verify the association between a pre-specified group membership and the RV distance measures. Subjects are assigned to a sub-group (represented by a simple binary matrix constituted of 0's and 1's). A non-parametric test based on the permutation of this matrix, analogous to "AnoSim" method (Clarke, 1993), allows us to check whether the subjects that belong to the same group are similar or not.

Comparing several sessions within or across subjects. One important issue in functional neuroimaging is the reproducibility of the results within subjects but between sessions that replicate the same (or similar) experimental paradigms. A few papers have addressed the issue (see for instance (McGonigle et al., 2000)). However, the *RV-coefficient* can provide a global measure of the relative distance in time or space of the replication of sessions with respect to the replication of subjects; to our knowledge, this has not previously been shown. We present such results in figure 6 and it is clear that two sessions from the same subject are *spatially* close from each other, while the *temporal* distance is much higher and comparable to the inter-subject variance. This will have implications for the modelling of different sessions within subjects.

Comparing different regions of interest for the same subject in the time domain. In learning or adaptation experiments, the most important information lies in the temporal

behaviour of brain regions. Often, we can define a network of areas that may be involved in the experiment. The RV coefficient can be used to calculate the relative distances between brain regions in the temporal domain. This can be seen as an index of their functional connectivity. Thus, the RV coefficient could be used as a more complete measure of the link between regions, links that are usually assessed using the correlation of the mean region activity. This would be at the expense of more difficult inferences. However, non-parametric permutation tests could be used to asses the significance of these links.

Comparing several tasks within and across subjects. Brain imaging paradigms are increasingly likely to involve several tasks or control conditions that allow for a better interpretation of the condition or task at the heart of the experiment. Experiments can sometimes involve 5 or more different tasks or conditions. These extra conditions can be thought of as functional localisers and can be used to define regions in which more specific functions are studied. Alternatively, different paradigms can be designed to study the spatial correspondence of several similar cognitive functions (Simon et al., 2002). In both cases, it will be important to characterise the distances between the different tasks within subject or across subjects. This may be particularly useful in the spatial domain if the main aim of the study is to understand the relative mapping of the functions involved to brain areas, and whether this mapping is stable across the population.

The method can also be extended to characterise individual differences across a series of tasks. Individual differences are now the topic of much current research.

In this paper we have addressed the problem of the using an average across subjects to represent the temporal or spatial characteristics of a group. We have provided a general tool to measure distances between subjects in the temporal or spatial domain. We believe that routine assessment of group homogeneity may have important consequences, because this is rarely addressed before using averaged results to report group data. Future work might address the problem of population sampling given a first sample of subjects. If the homogeneity analysis suggests the population may be clustered, we suggest separate group analyses for each sub-group. It is possible that the sampled sub-groups do not contain enough subjects to perform group analysis. In this case one

needs to repeat the sampling stage. Adaptive Sampling (Thompson and Seber, 1996) is one such method, that uses the detected characteristics of the subjects already sampled to guide future sampling.

References

Aguirre G.K., Zarahn E., and D'esposito M. The variability of human, BOLD hemodynamic responses. *Neuroimage*, (8(4)):360–9, Nov 1998.

Barnett V. and Lewis T. *Outliers in Statistical Data*. 1994.

Brammer M.J., Bullmore E.T., Simmons A., Williams S.C., Grasby P.M., Howard R.J., Woodruff P.W., and Rabe-Hesketh S. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. *Magn Reson Imaging*, (15(7)):763–70, 1997.

Brett M., Johnsrude I.S., and Owen A.M. The problem of functional localization in the human brain. *Nat Rev Neurosci*, (3(3)):243–9, Mar 2002.

Burbaud P., Camus O., Guehl D., Bioulac B., Caille J., and Allard M. Influence of cognitive strategies on the pattern of cortical activation during mental subtraction. a functional imaging study in human subjects. *Neurosci Lett*, 16(287(1)):76–80, Jun 2000.

Cabeza R. and Nyberg L. Imaging cognition II: An empirical review of 275 PET and fMRI studies. *J Cogn Neurosci*, (12(1)):1–47, Jan 2000.

Calhoun V.D., Adali T., Pearlson G.D., and Pekar J.J. A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp*, (14(3)): 140–51, Nov 2001.

Castaño-Tostado E. and Tanaka Y. Some comments in Escoufier's RV - coefficient as a sensitivity measure in principal component analysis. *Communications in Satistics*, A(19)12:4619–26, 1990.

Christensen R. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer-Verlag, 1987.

Clarke K.R. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143, 1993.

Cochran W.G. *Sampling Techniques*. John Wiley, 1977.

Cook R.D. and Weisberg S. *Residual and Influence in Regression*. 1982.

D'Esposito M. Functional neuroimaging of cognition. *Semin Neurol*, (20(4)):487–98, 2000.

Detre J.A. and Floyd T.F. Functional MRI and its applications to the clinical neurosciences. *Neuroscientist*, (7(1)):64–79, Feb 2001.

Fox P.T., Huang A.Y., Parsons L.M., Xiong J.H., Rainey L., and Lancaster J.L. Functional volumes modeling: scaling for group size in averaged images. *Hum Brain Mapp*, (8(2-3)): 143–50, 1999.

Friston K.J., Frith C.D., Fletcher P., Liddle P.F., and Frackowiak R.S. Functional topography: multidimensional scaling and functional connectivity in the brain. *Cereb Cortex*, (6(2)):156–64, Mar-Apr 1996.

Friston K.J., Holmes A.P., Price C.J., Buchel C., and Worsley K.J. Multisubject fMRIstudies and conjunction analyses. *Neuroimage*, 4(10):385–96, Oct 1999a.

Friston K.J., Holmes A.P., and Worsley K.J. How many subjects constitute a study? *Neuroimage*, (10(1)):1–5, Jul 1999b.

Friston K.J., Josephs O., Zarahn E., Holmes A.P., Rouquette S., and Poline J. To smooth or not to smooth? bias and efficiency in fMRI time-series analysis. *Neuroimage*, (12(2)):196–208, Aug 2000.

Genovese C.R., Noll D.C., and Eddy W.F. Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. *Magn Reson Med*, (38(3)):497–507, Sep 1997.

Govindarajulu Z. *Elements of Sampling Theory and Methods*. Prentice Hall, 1999.

Gower C. Multidimensional scaling displays. In New York: Praeger., editor, *Research methods for multimode data analysis*. Law,H.G., 1984.

Hansen L.K., Larsen J., Nielsen F.A., Strother S.C., Rostrup E., Savoy R., Lange N., Sidtis J., Svarer C., and Paulson O.B. Generalizable patterns in neuroimaging: how many principal components? *Neuroimage*, (9(5)):534–44, May 1999.

Hawkins D.M. *Identification of Outliers.* 1980.

Holmes A.P. and Friston K.J. Generalisability, random effects and population inference. *Neuroimage*, 7:S754, 1998.

Kherif F., Andrade A., Benali H., Le-Bihan D., and Poline J.-B. Multivariate analysis for fMRI data investigation and model checking. *Neuroimage*, 13(6):S171, 2001.

Kherif F., Poline J.-B., Flandin G., Benali H., Simon O., Dehaene S., and Worsley K. Multivariate Model Specification for fMRI Data. *Neuroimage*, 16(4):1068–83, Aug 2002.

Kustra R. and Strother S. Penalized discriminant analysis of [15o]-water PET brain images with prediction error selection of smoothness and regularization hyperparameters. *IEEE Trans Med Imaging*, 20(2):376–87, 2001.

Lavit C. *Analyse conjointe de tableaux quantitatifs.* Masson, 1984.

Lazar N.A., Luna B., Sweeney J.A., and Eddy W.F. Combining brains: A Survey of Methods for Statistical Pooling of Information. *Neuroimage*, 16(2):538–50, Aug 2002.

Luo W. and Nichols T. Diagnosis and Exploration of Massively Univariate 4D Spatiotemporal Models. *submitted*, 2003.

Maitra R., Roys S.R., and Gullapalli R.P. Test-retest reliability estimation of functional MRI data. *Magn Reson Med*, (48(1)):62–70, Jul 2002.

Mandzia J. and Black S.E. Neuroimaging and behavior: probing brain behavior relationships in the 21st century. *Curr Neurol Neurosci Rep*, (1(6)):553–61, Nov 2001.

Marchini J.L. and Smith S.M. On bias in the estimation of autocorrelations for fMRI voxel time-series analysis. *Neuroimage*, (18(1)):83–90, Jan 2003.

McGonigle D.J., Howseman A.M., Athwal B.S., Friston K.J., Frackowiak R.S., and Holmes A.P. Variability in fMRI: an examination of intersession differences. *Neuroimage*, (11(6 Pt 1)): 708–34, Jun 2000.

McIntosh A.R., Bookstein F.L., Haxby J.V., and Grady C.L. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, (3(3 Pt 1)):143–57, Jun 1996.

Meyer-Lindenberg A., Poline J.-B., Kohn P.D., Holt J.L., Egan M.F., Weinberger D.R., and Berman K.F. Evidence for abnormal cortical functional connectivity during working memory in schizophrenia. *Am J Psychiatry*, 158(11):1809–17, Jun 2001.

Moeller J.R., Strother S.C., Sidtis J.J., and Rottenberg D.A. Scaled subprofile model: a statistical approach to the analysis of functional patterns in positron emission tomographic data. *J Cereb Blood Flow Metab*, (7(5)):649–58, Oct 1987.

Nybakken G.E., Quigley M.A., Moritz C.H., Cordes D., Haughton V.M., and Meyerand M.E. Test-retest precision of functional magnetic resonance imaging processed with independent component analysis. *Neuroradiology*, (44(5)):403–6, May 2002.

Petersson K.M., Nichols T.E., Poline J.B., and Holmes A.P. Statistical limitations in functional neuroimaging. ii. Signal detection and statistical inference. *Philos Trans R Soc Lond B Biol Sci*, 354(1387):1261–81, Jul 1999.

Posner M.I. and Raichle M.E. The neuroimaging of human brain function. *Proc Natl Acad Sci U S A*, 3(95(3)):763–4, Feb 1998.

Purdon P.L. and Weisskoff R.M. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fmri. *Hum Brain Mapp*, (6(4)): 239–49, 1998.

Robert P. and Escoufier Y. A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics*, 25:257–265, 1976.

Seber G.A.F. *Linear Regression Analysis*. John Wiley & Sons, 1977.

Simon O., Mangin J.F., Cohen L., Le Bihan D., and Dehaene S. Topographical layout of hands, eye, calculation, and language-related areas in the human parietal lobe. *Neuron*, 31(33(3)): 475–87, Jan 2002.

Strother S.C., Anderson J., Hansen L.K., Kjems U., Kustra R., Sidtis J., Frutiger S., Muley S., LaConte S., and Rottenberg D. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage*, (15(4)):747–71, Apr 2002.

Strother S.C., Anderson J.R., Schaper K.A., Sidtis J.J., Liow J.S., Woods R.P., and Rottenberg D.A. Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping: I. "functional connectivity" of the human motor system studied with [15o]water PET. *J Cereb Blood Flow Metab*, (15(5)):738–53, Sep 1995a.

Strother S.C., Kanno I., and Rottenberg D.A. Commentary and opinion: I. principal component analysis, variance partitioning, and "functional connectivity". *J Cereb Blood Flow Metab*, (15(3)):353–60, May 1995b.

Tanaka Y and Mori Y. Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematics and Management Sciences*, 17:61–89, 1997.

Tegeler C., Strother S.C., Anderson J.R., and Kim S.G. Reproducibility of BOLD-based functional MRI obtained at 4 T. *Hum Brain Mapp*, (7(4)):267–83, 1999.

Thompson S.K. and Seber G.A.F. *Adaptive Sampling*. John Wiley, 1996.

Torgerson W.S. Multidimensional scaling. I. Theory and method. *Psychometrika*, 17:401–419, 1952.

Welchew D.E., Honey G.D., Sharma T., Robbins T.W., and Bullmore E.T. Multidimensional scaling of integrated neurocognitive function and schizophrenia as a disconnexion disorder. *Neuroimage*, 17(3):1227–39, Aug 2002.

White T., O'Leary D., Magnotta V., Arndt S., Flaum M., and Andreasen N.C. Anatomic and functional variability: the effects of filter size in group fmri data analysis. *Neuroimage*, (13(4)): 577–88, Apr 2001.

Worsley K.J., Liao C.H., Aston J., Petre V., Duncan G.H., Morales F., and Evans A.C. A General Statistical Analysis for fMRI Data. *NeuroImage*, 15(1):1–15, 2002.

Worsley K.J., Poline J.B., Friston K.J., and Evans A.C. Characterizing the response of PET and fMRIdata using multivariate linear models. *Neuroimage*, 6(4):305–19, Nov 1997.

Xiong J., Rao S., Jerabek P., Zamarripa F., Woldorff M., Lancaster J., and Fox P.T. Intersubject variability in cortical activations during a complex language task. *Neuroimage*, (12(3)):326–39, Sep 2000.