

Functional connectivity metrics during stroke recovery

G. YOURGANOV^{1,2}, T. SCHMAH³, S.L. SMALL⁴,
P.M. RASMUSSEN⁵, S.C. STROTHER^{1,2,6}

¹ Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Canada;

² Institute of Medical Science, University of Toronto, Canada;

³ Department of Computer Science, University of Toronto, Canada;

⁴ Department of Neurology, University of Chicago, USA;

⁵ DTU Informatics, Technical University of Denmark and The Danish National Research Foundation's Center for Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark;

⁶ Department of Medical Biophysics, University of Toronto, Canada

ABSTRACT

We explore functional connectivity in nine subjects measured with 1.5T fMRI-BOLD in a longitudinal study of recovery from unilateral stroke affecting the motor area (Small et al., 2002). We found that several measures of complexity of covariance matrices show strong correlations with behavioral measures of recovery. In Schmah et al. (2010), we applied Linear and Quadratic Discriminants (LD and QD) computed on a principal components (PC) subspace to classify the fMRI volumes into “early” and “late” sessions. We demonstrated excellent classification accuracy with QD but not LD, indicating that potentially important differences in functional connectivity exist between the early and late sessions. Motivated by McIntosh et al. (2008), who showed that EEG brain-signal variability and behavioral performance both increased with age during development, we investigated complexity of the covariance matrix for this longitudinal stroke recovery data set. We used three complexity measures: the sphericity index described by Abdi (2010); “unsupervised dimensionality”, which is the number of PCs that minimizes unsupervised generalization error of a covariance matrix (Hansen et al., 1999); and “QD dimensionality”, which is the number of PCs that minimizes the classification accuracy of QD. Although these approaches measure different kinds of complexity, all showed strong correlations with one or more behavioral tests: nine-hole peg test, hand grip test and pinch test. We could not demonstrate that either sphericity or unsupervised dimensionality were significantly different for the “early” and “late” sessions using a paired Wilcoxon test. However, the amount of relative behavioral improvement was correlated with sphericity of the overall covariance matrix (pooled across all sessions), as well as with the divergence of the eigenspectra between the “early” and “late” covariance matrices. Complexity measures that use the number of PCs (which optimize QD classification or unsupervised generalization) were correlated with the behavioral performance of the final session, but not with the relative improvement. These are suggestive, but limited, results given the sample size, restricted behavioral measurements and older 1.5T BOLD data sets. Nevertheless, they indicate one potentially fruitful direction for future data-driven fMRI studies of stroke recovery in larger, better-characterized longitudinal stroke data sets recorded at higher field strength. Finally, we produced sensitivity maps (Kjems et al., 2002) corresponding to both linear and quadratic discriminants for the “early” vs. “late” classification. These maps measure the influence of each voxel on the class assignments for a given classifier. Differences between the scaled sensitivity maps for the linear and quadratic discriminants indicate brain regions involved in changes in functional connectivity. These regions are highly variable across subjects, but include the cerebellum and the motor area contralateral to the lesion.

Key words

fMRI • Functional connectivity • Complexity • Stroke recovery •
Principal Component Analysis (PCA) • Sensitivity maps

1. Introduction

We studied functional connectivity metrics during stroke recovery, using 1.5T fMRI-BOLD data from a longitudinal study of the recovery of nine subjects from unilateral stroke affecting the motor area (Small et al., 2002). The present study was motivated by several previous results. First, the performance of quadratic discriminant (QD) was far superior to linear discriminant (LD) in “early” vs. “late” within-subject classification on the same dataset, as shown in Schmah et al. (2010). Both LD and QD were computed using a principal components basis obtained from the spectral decomposition of the data matrix; the number of PCs was chosen to optimize classification accuracy. Each fMRI volume was classified as “early” if it was recorded in the first or second session (1 or 2 months post-stroke), or “late” if it was recorded in the third or fourth session (3 or 6 months post-stroke). Quadratic discriminant, performed on a PC subspace, achieved high within-subject, out-of-sample classification accuracy (QD median accuracy > 0.99 for every subject; LD median accuracy 0.61–0.90). The superior performance of QD relative to LD is likely due to the inequality of the within-class covariance matrices, and indicates that potentially important differences in functional connectivity exist between the early and late sessions. Such differences likely reflect the functional re-organization that is known to be involved in recovery from stroke (see for example James et al., 2009). We suggest that the hierarchy of LD and QD prediction performance is particularly useful for simultaneously probing longitudinal changes that are expressed in both the mean signal and/or in functional connectivity measures in BOLD fMRI.

Our second source for motivation of this study was McIntosh et al. (2008, 2010) who showed EEG brain-signal variability (i.e. complexity) increasing with age during development, with higher variability associated with higher performance. Our working hypothesis was that behavior and complexity would improve together with recovery from stroke. In this work, functional connectivity between voxels is captured by covariance matrices. To measure the complexity of covariance matrices, in this work we used the sphericity index of Abdi (2010) and the unsupervised dimensionality of Hansen et al. (1999). The first measure indicates complexity as

the rate at which the eigenvalues of the covariance matrix decrease when we go from largest to smallest eigenvalue, i.e. it measures the curvature of the eigenvalue plot or “spectrum”. The second measure indicates complexity as the number of degrees of freedom of the unsupervised Principal Component Analysis (PCA) model that has optimal predictive power. We computed these individually for the “early” and “late” within-class covariance matrices, and also for the overall data covariance matrix that includes all sessions. In addition, for the overall covariance matrix, we used a third complexity measure, the QD dimension that records the optimal dimensionality (the number of PCs) for distinguishing early from late volumes using the quadratic discriminant. Finally we also measured the difference in eigenspectra of the two within-class covariance matrices. We investigated correlations between these measures and behavioral measures of recovery from stroke.

The measures discussed so far give overall indications of the complexity of functional connectivity in our study. To investigate which brain regions are involved in the associated changes in mean BOLD and functional connectivity covariance patterns during stroke recovery, we produced sensitivity maps corresponding to both linear and quadratic discriminants for the early vs. late classification, using a method based on Zurada et al. (1997) and Kjems et al. (2002). Sensitivity maps measure the influence of each voxel on the class assignments for a given classifier. Differences between the scaled sensitivity maps for the linear and quadratic discriminants indicate brain regions that are involved in changes in functional connectivity.

2. Methods

2.1. Data and preprocessing

The original study of Small et al. (2002) analyzed mean volumes of BOLD activation over four regions of interest in each hemisphere, in twelve subjects. We have previously (Schmah et al., 2010) used this data set to compare ten linear and nonlinear classification algorithms; please refer to that paper for a more detailed description of data collection and preprocessing. In the present paper, as in Schmah et al. (2010), we studied only nine of the subjects. The

subjects were studied in four sessions, at 1 and 2 (early), and 3 and 6 (late) months post-stroke. Each session consisted of an fMRI imaging session and a series of behavioral tests.

2.1.1. fMRI data

Each of the four imaging sessions consisted of four continuous recording runs. During each run, the subject alternated two kinds of hand movement: tapping finger and thumb together, or wrist flexion/extension; with rest breaks in between. The movement was paced auditorily at 1 Hz. Within each run, the experimental design was: (12 seconds rest, 24 seconds finger tap, 12 seconds rest, 24 seconds wrist flexion), repeated 8 times. Whole-brain fMRI volumes, each consisting of 24 axial slices, were collected at the University of Maryland in a 1.5T scanner (slice thickness: 6 mm; pixel size: 1.88 mm x 1.88 mm; FOV: 240 x 240 x 144; FA = 60; TE = 35 ms; TR 4000 ms). In terms of the TR, the design of each run may be summarized as: (3 TR rest, 6 TR finger tap, 3 TR rest, 6 TR wrist flexion), repeated 8 times.

The data for all 9 subjects were co-registered and motion-corrected via 12-parameter affine transformations found using the Automated Image Registration (AIR) package (Woods et al., 1998). After this, for computational ease, we retained only 7 axial fMRI slices, namely slices 2, 3, 4, 5, 21, 22 and 23, where 24 is the top of the head. These slices were chosen to ensure coverage of the cerebellum and sensorimotor cortex. This results in 10499 voxels. Next, for every “active” volume, i.e., those recorded during finger or wrist movement, we divided the intensity of each voxel by the mean of the corresponding voxel intensities in the previous two rest volumes, as proposed in McIntosh and Lobaugh (2004), to get rid of low-frequency temporal drifts. Following this normalization, the rest volumes were discarded. The first volume in each active block was also discarded to reduce the haemodynamic response transients intrinsic to fMRI imaging.

The data now consist of 256 “condition groups” of 5 consecutively recorded normalized volumes, each corresponding to either finger or wrist movement, resulting in 1280 volumes overall. Within each condition group, the 5 volumes were recorded consecutively, in the same run of the same session, under the same experimental conditions (i.e. the

subject was making the same movement). The final preprocessing step is to scale the data, separately for each subject using all 1280 volumes. We scaled each volume by subtracting the mean volume and dividing each voxel by the mean standard deviation of the voxel intensity (pooled across all voxels).

2.1.2. Behavioral data

Of the several behavioral tests in the original study, we used the results of three tests (each was performed on both the healthy and the impaired hand in each of the 4 sessions):

1. Strength of the *hand grip*, measured with a dynamometer.
2. Strength of the *pinch* between thumb and index finger, measured the same way.
3. Performance on the *nine-hole peg test*, defined as $1/(\text{time to complete})$.

Each test produced two behavioral measures of recovery: *improvement of performance* and *final performance*. Improvement was calculated as the difference between the performance of the impaired hand on the first and last session, divided by the mean (across all 4 sessions) performance of the healthy hand. The final performance was computed as the ratio of the performance of the impaired hand on the fourth session only, again divided by the mean performance of the healthy hand.

2.2. Classification methods: adaptive linear and quadratic discriminants

We used two probabilistic binary classification methods, both of which can be described in terms of a decision function, D , defined by

$$D(\mathbf{x}) = \log \left(\frac{P(\text{class} = 1 | \mathbf{x})}{P(\text{class} = 0 | \mathbf{x})} \right) \quad (2.1)$$

Note that if $D(\mathbf{x}) < 0$ for a certain volume \mathbf{x} , then that volume is predicted to be in class 0, while if $D(\mathbf{x}) > 0$, then it is predicted to be in class 1. If we assume that, *a priori*, membership in each of the two classes is equally probable, that is, $P(\text{class} = 1) = P(\text{class} = 0) = 0.5$, then it is easily shown, using Bayes’ theorem, that

$$D(\mathbf{x}) = \log \left(\frac{P(\mathbf{x} | \text{class} = 1)}{P(\mathbf{x} | \text{class} = 0)} \right) \quad (2.2)$$

To obtain sample maximum-likelihood classification, both linear and quadratic discriminant models assume that, for each class, the volumes are independent samples from multivariate Gaussian distributions. (For the fMRI volumes in this study, this assumption is not strictly valid, so we therefore resample using more nearly independent groups of scans - see the next subsection.) For the quadratic discriminant (QD), no further assumptions are made. Substituting Gaussian distributions into equation 2.2 leads to the following decision function (the discriminant):

$$\lambda_{QD}(\mathbf{x}) = \frac{1}{2} \log \frac{|\mathbf{S}_0|}{|\mathbf{S}_1|} - \frac{1}{2}(\mathbf{x} - \mathbf{m}_1)^T \mathbf{S}_1^{-1} (\mathbf{x} - \mathbf{m}_1) - \frac{1}{2}(\mathbf{x} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{x} - \mathbf{m}_0) \quad (2.3)$$

Here, \mathbf{m}_c and \mathbf{S}_c are means and covariance matrices for class c , which are taken to be the sample means and covariances of the training data for class c . (See below for the case where one or both of the sample covariance matrices is not invertible.) Linear discriminant (LD) is similar to quadratic discriminant, but makes the extra assumption that the covariance matrices of the two classes are equal. In place of \mathbf{S}_0 and \mathbf{S}_1 , we used \mathbf{S} , the pooled sample covariance matrix (the mean of \mathbf{S}_0 and \mathbf{S}_1). The decision function simplifies to:

$$D_{LD}(\mathbf{x}) = - \left(\mathbf{x} - \frac{1}{2}(\mathbf{m}_0 + \mathbf{m}_1) \right)^T \mathbf{S}^{-1} (\mathbf{m}_0 - \mathbf{m}_1) \quad (2.4)$$

In our application, the number of voxels was greater than the number of observations (the number of volumes), so the sample covariance matrices were always rank-deficient and cannot be inverted. To avert this problem, we computed the first K principal components of the mean-centered data, which were the K eigenvectors of the covariance matrix corresponding to the highest eigenvalues. This was done efficiently using the singular value decomposition (see for example Mardia et al., 1979). We then applied LD and QD to the orthogonal projection of the data onto the first K principal components. Following Strother et al. (2002) and LaConte et al. (2003), the value of K was chosen to optimize the

classification performance on a validation set. In our implementation, the K values tried were 1, 11, 21, ... 471, with the maximum value determined by the number of volumes per class in the training set, which was 480. We found that the optimal values of K ranged from 1 to 471. We call the combined method we have described, i.e., PCA followed by the application of LD (or QD) to an optimal-dimension PC subspace, ‘‘adaptive LD’’ (or QD).

2.3. Cross-validation procedures

The classification methods described above were applied as follows. Data for each subject were treated separately. For each subject, the preprocessed and mean-centered data from all sessions were pooled, resulting in 1280 volumes. We compared the methods in 20 trials, with each trial run on a different random splitting of the pooled data into two subsets: 75% training (960 volumes), 12.5% validation (160 volumes), and 12.5% test (160 volumes). The splits of the data were created using a resampling method, with the sampling units being condition groups, defined by groups of 5 normalized volumes recorded consecutively under the same experimental conditions (see Section 2.1). By assigning all volumes in the same condition group to the same subset (training, validation or test), we reduced the dependence between volumes in different subsets. The sampling method was chosen to create training, validation and test sets that were balanced with respect to all experimental conditions, i.e., for each of the three sets in a split, the set contained the same number of volumes corresponding to finger movement as to wrist movement, and equal numbers of volumes from each run. See Schmah et al. (2010) for details of the resampling method.

The purpose of the validation set was to tune hyperparameters before final testing of a method: the training set was used to build a family of classifiers that are identical except for certain hyperparameter values, and these classifiers were tested on the validation set in order to determine the optimal hyperparameter value(s). The classifier with the optimal hyperparameter value(s) was then tested on the test set. In the present paper, the final classification accuracy is not reported (see Schmah et al., 2010), as it is the optimal hyperparameter values themselves that are of primary interest.

2.4. fMRI-based measures

We computed several measures that reflect the complexity of our fMRI data as well as the development of functional connectivity across sessions. N denotes the number of volumes in the data set.

- *Sphericity index*: This measure was calculated on the covariance matrix with the data pooled across all sessions. The sphericity index reflects the curvature of the plot of this matrix’s eigenvalues. An index of 1 corresponds to a perfectly spherical matrix, in which the eigenvalues are all identical. The smaller this index, the sharper is the drop in the eigenvalues when we go from largest to smallest. We define this index using the Greenhouse-Geisser correction to the Box criterion (Abdi, 2010):

$$\varepsilon = \frac{1}{N-1} \frac{\left(\sum_{i=1}^N \lambda_i\right)^2}{\sum_{i=1}^N \lambda_i^2} \quad (2.5)$$

- *Divergence of eigenspectra of “early” and “late” covariance matrices*: Here, instead of pooling the data across all sessions, we consider the two within-class covariance matrices (corresponding to early and late sessions) separately, and measure the degree of their divergence. The measure is defined as the Euclidean distance between the eigenvalues of the two matrices:

$$d = \sqrt{\sum_{i=1}^N \left(\lambda_i^{(E)} - \lambda_i^{(L)}\right)^2} \quad (2.6)$$

- *QD dimensionality*: The number of principal components that optimize the classification accuracy of a Quadratic Discriminant when classifying the volumes into “early” and “late” sessions, as described in Section 2.2.
- *Unsupervised dimensionality*: The number of principal components that optimize the unsupervised generalization error. Following Hansen et al. (1999), we measure this error using cross-validation. The data are split into two sets of equal size: the training set and the test set. The

mean training volume is subtracted from both training and test sets. We perform singular-value decomposition on the training set, and assume that the signal is contained in the first K principal components, and the remaining $N-K-1$ components contain noise. The variance of the noise is estimated as the mean of these $N-K-1$ eigenvalues. After that, we approximate the covariance matrix of the training set as a sum of the *signal* and *noise* covariance matrices: the first is constructed with the K principal components of the original covariance matrix, and the second is the diagonal matrix with the estimate of noise variance on the diagonal. Then, we can see how well this approximation generalizes to the test set, i.e. calculate the posterior probability of the test set using a multivariate Gaussian distribution with the population covariance matrix approximated using the K PCs and noise estimate.

2.5. Sensitivity maps

Given any probabilistic binary classifier defined by a decision function $D(\mathbf{x})$, we define the *sensitivity* of the classifier to voxel i as

$$s_i = \frac{1}{N} \sum_j \left(\left. \frac{\partial}{\partial x_i} D(\mathbf{x}) \right|_{\mathbf{x}^{(j)}} \right)^2 \quad (2.7)$$

where $\mathbf{x}^{(j)}$ is the j^{th} volume. This is similar to the definition in Zurada et al. (1997) and Kjems et al. (2002), except that we have used the decision function instead of the probability of class membership. To explain this choice, note that the decision function (defined in Equation 2.1) is equivalent to

$$P(\text{class} = 1 | \mathbf{x}) = \frac{1}{1 + \exp(-D(\mathbf{x}))} = \sigma(D(\mathbf{x})) \quad (2.8)$$

where $\sigma(z) = 1 / (1 + \exp(-z))$. Thus

$$\frac{\partial}{\partial x_i} P(\text{class} = 1 | \mathbf{x}) = \sigma'(D(\mathbf{x})) \frac{\partial}{\partial x_i} D(\mathbf{x}) \quad (2.9)$$

where $\sigma'(z) = \exp(-z) / (1 + \exp(-z))^2 = \sigma(z) (1 - \sigma(z))$. The term $\sigma'(D(\mathbf{x}))$ is close to zero whenever \mathbf{x} is far from the decision boundary. This may be considered

a desirable property theoretically, but in our experiments, only a small number of training examples were close enough to the decision boundary for this term to be numerically nonzero, leading to a non-robust dependence on a small number of training examples. The *sensitivity map* of a classifier, for a given set of training examples, is the vector \mathbf{s} of sensitivities for all voxels, computed according to Equation 2.7. The corresponding *scaled sensitivity map* is $\mathbf{s}/\|\mathbf{s}\|_1$, in which each sensitivity has been scaled by the same global intensity factor $\|\mathbf{s}\|_1 = \sum_i(s_i)$ (recall that sensitivities are always positive). We found that global normalization was important in comparing sensitivity maps between subjects and even between different splits of the data for a single subject, since otherwise their global intensities varied by several orders of magnitude. Linear discriminant (LD) and quadratic discriminant (QD) are defined by the decision functions in Equations 2.4 and 2.3 respectively. The corresponding voxel sensitivities are:

$$s_i^{LD} = (\mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_0))_i^2 \quad (2.10)$$

and

$$s_i^{QD} = (-\mathbf{S}_1^{-1}(\mathbf{x}^{(j)} - \mathbf{m}_1) + \mathbf{S}_0^{-1}(\mathbf{x}^{(j)} - \mathbf{m}_0))_i^2$$

For fMRI, both LD and QD must be applied in a PC subspace of small enough dimension that the covariance matrices can be inverted. For every subject and every split of the data, we used the PC dimension that was found to be optimal for prediction. The sample covariance matrices were computed in this subspace. The new decision functions, expressed in the original voxel space, take the same form as above except that \mathbf{S}^{-1} is replaced by $\mathbf{US}^{-1}\mathbf{U}^T$ (and similarly for \mathbf{S}_0 and \mathbf{S}_1), where \mathbf{U} is the orthogonal matrix of eigenimages obtained from the singular value decomposition (each column of \mathbf{U} is an eigenimage). Thus the voxel sensitivities are:

$$s_i^{LD} = (\mathbf{US}^{-1}\mathbf{U}^T(\mathbf{m}_1 - \mathbf{m}_0))_i^2 \quad (2.11)$$

and

$$s_i^{QD} = (-\mathbf{US}_1^{-1}\mathbf{U}^T(\mathbf{x}^{(j)} - \mathbf{m}_1) + \mathbf{US}_0^{-1}\mathbf{U}^T(\mathbf{x}^{(j)} - \mathbf{m}_0))_i^2$$

The mean of the corresponding scaled sensitivity maps (see above), across all different splits of the data for a given subject, is the final sensitivity map for that subject.

3. Results

3.1. Correlations between fMRI-based measures and behavioral measures

We examined correlations between the four fMRI-based measures defined in Section 2.4, and the six behavioral measures of stroke recovery described in Section 2.1.2. Recall that for each of the behavioral measures, larger values correspond to better recovery. For each combination of fMRI-based and behavioral measures, we computed the Spearman correlation coefficient. Table I reports these coefficients and associated p -values. Fig. 1 shows scatter plots for selected pairs of measures, as detailed below.

The correlations that are significant for at least one metric and behavioral measure at the 0.05 level (uncorrected) are:

1. Sphericity is negatively correlated with improvement in both the peg test and pinch strength (there is no evidence for a significant relationship with hand strength performance). This indicates that subjects showing a highly non-spherical covariance matrix (i.e. with a sharp drop in the eigenspectrum) show the greatest improvement in recovery measures. A scatter plot of sphericity versus pinch strength improvement is shown in Fig. 1A.
2. Divergence of the eigenspectra of the “early” and “late” covariance matrices is positively correlated with pinch strength improvement. This is expected, because the difference between these matrices points to the re-organization of the brain, which leads to effective recovery. A scatter plot of these variables appears in Fig. 1B.
3. QD dimensionality is negatively correlated with final peg test performance. In general, we can

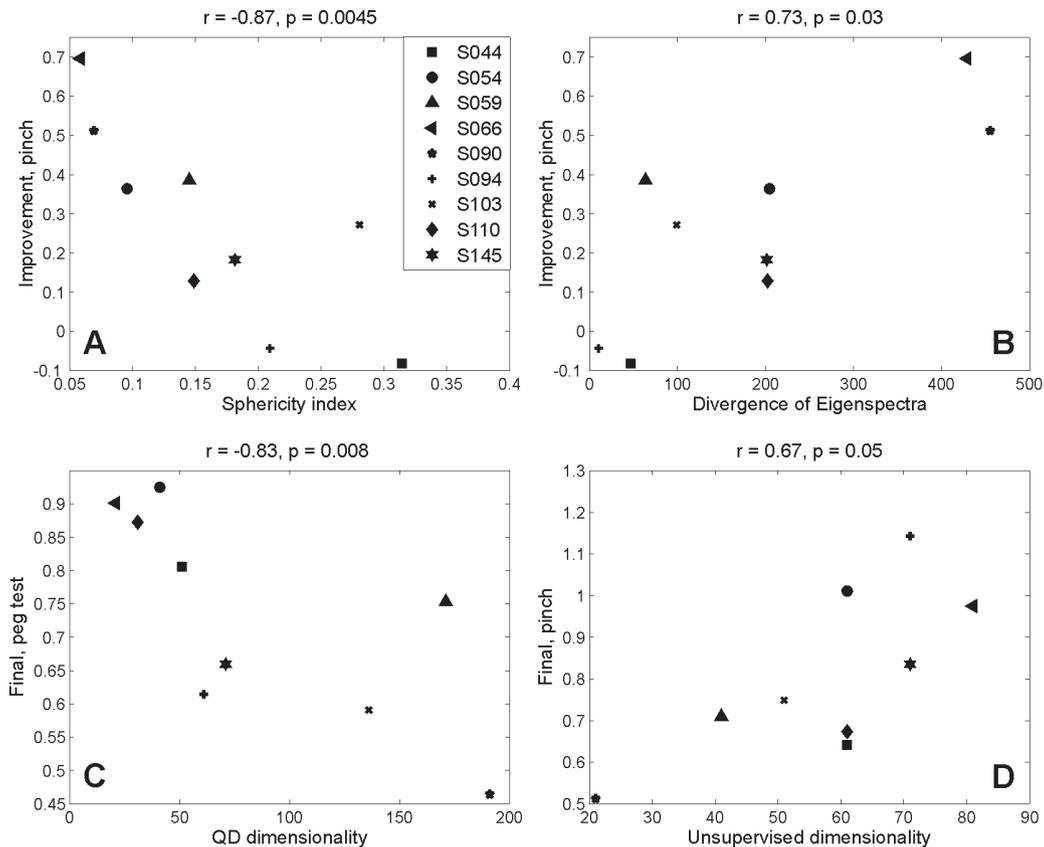


Fig. 1. - Scatter plots for four combinations of fMRI-based measures and behavioral measures: sphericity index versus pinch test improvement (A); divergence of eigenspectra of covariance matrices versus pinch test improvement (B); QD dimensionality versus final peg test performance (C); Unsupervised dimensionality versus final pinch test performance (D).

say that subjects with the best final performance require a small number of PCs in QD classification. To classify the bad performers, we need more PCs to capture the difference between the early and late covariance matrices. A scatter plot of these variables appears in Fig. 1C.

- Unsupervised dimensionality is positively correlated with both final pinch strength and final grip strength. This could indicate that the overall covariance matrix is *more* complex in good final performers, because their brain at the last session is more complex due to the post-stroke re-organization. A scatter plot of unsupervised dimensionality versus final pinch strength appears in Fig. 1D.

In all of these cases, the absolute value of the correlation coefficient is at least 0.67.

Note from Fig. 1 that Subject S059 does not follow the general trend in both the upper right and lower left scatter plots, corresponding to the correlations in points 2 and 3 above. Subjects S066 and S090 look very similar in the upper left and upper right plots, with both showing high pinch test improvement, low

sphericity and high divergence of eigenspectra. In contrast on the same plots subjects S044 and S103 show none to low pinch test improvement, higher sphericity and approximately zero divergence of eigenspectra. These results indicate that nonsphericity is being at least partly driven by changes in the eigenspectrum between the early and late matrices. However, a paired Wilcoxon test failed to reveal a significant change in sphericity computed separately on “early” and “late” sessions ($p = 0.91$). Unsupervised dimensionality computed separately for “early” and “late” sessions also did not differ significantly ($p = 0.86$). This lack of significance is perhaps an indicator of inefficiency of the paired Wilcoxon test and/or the large heterogeneity in our subject set. Finally, note that while S066 and S090 are very similar in the upper left and right plots, they sharply diverge to either ends of the dimensionality scales in the lower left and lower right plots in Fig. 1. To further analyze the correlations between fMRI-based measures and behavioral test results, we performed a Partial Least Squares (PLS) analysis

(McIntosh and Lobaugh., 2004; Krishnan et al., 2010). The matrix of correlations \mathbf{R} (displayed in Table I) was decomposed using singular-value decomposition: $\mathbf{R} = \mathbf{USV}^T$. The first two principal components explain 43% and 39.5%, respectively, of the total correlation (this is computed by dividing each singular value in \mathbf{S} by the total sum of singular values); taken together, they explain 82.5% of the total correlation. These two principal components can be seen as the two orthogonal directions that capture (in the optimal least-squares sense) the associations between fMRI-based and behavioral measures. Fig. 2 shows the scatter plot of the first two saliences that correspond to columns of \mathbf{U} for fMRI-based measures, and to columns of \mathbf{V} for behavioral measures. We can see that the first (horizontal) direction captures the fMRI-based measures that are computed on the eigenvalues (sphericity and eigenspectra divergence) and the behavioral measures based on the improvement in performance. The second (vertical) direction captures the measures based on the PC dimensionality that optimizes prediction (unsupervised or supervised) and the behavioral measures based on the performance at the final session.

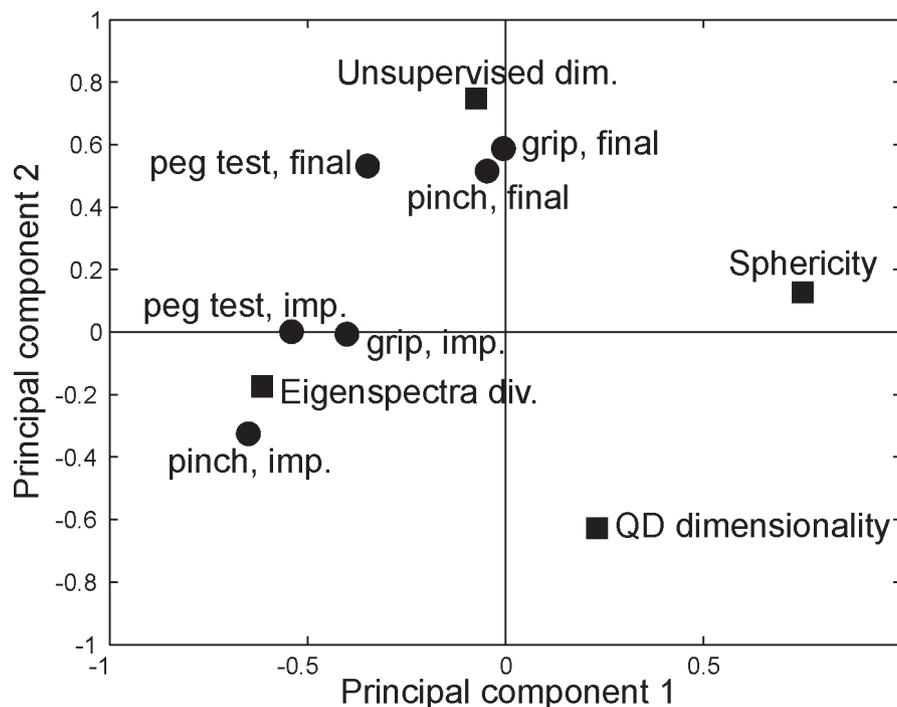


Fig. 2. - Scatter plot for the first vs. the second principal component produced by PLS analysis of the correlation matrix given in Table I. Squares and circles denote fMRI-based and behavioral measures, respectively.

3.2. Sensitivity maps

For each subject, the mean scaled sensitivity maps for both linear discriminant and quadratic discriminant, for the “early” versus “late” task, were calculated as described in Section 2.5. Both sensitivity maps for Subject S090 are shown in Fig. 3. Fig. 4 shows, for each subject, the difference between the LD and QD maps, with differences in sensitivity thresholded at $p < 0.05$ using paired t -tests with Bonferroni correction for multiple comparisons.

The sensitivity maps and the differences between them vary greatly between subjects. The most salient patterns of sensitivity are the following:

- For Subject S054, quadratic (but not linear) discriminant is sensitive to a large bilateral area of prefrontal cortex, while linear (but not quadratic) discriminant is sensitive to parts of the right motor and premotor areas (contralateral to the lesion).
- For Subject S059, quadratic (but not linear) discriminant is sensitive to part of the left premotor area (ipsilateral to the lesion) and the right motor area (contralateral to the lesion).
- For Subject S066, quadratic (but not linear) discriminant is sensitive to several areas on both sides of the cerebellum.

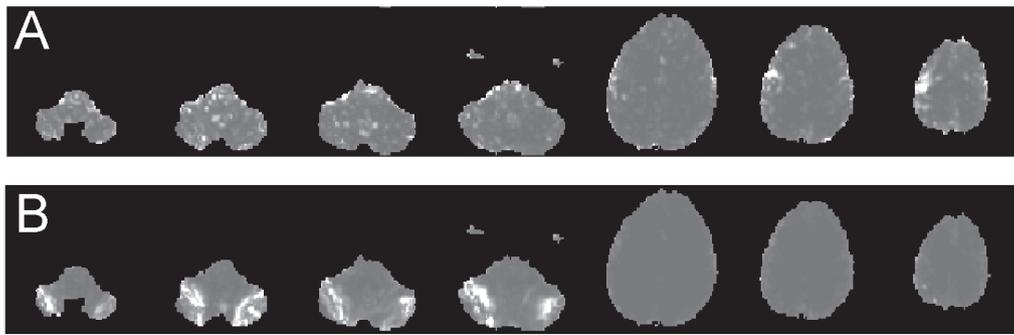


Fig. 3. - Mean scaled sensitivity maps for LD (A) and QD (B) for Subject S090, on the early vs. late task.

- For Subject S090 (see Fig. 3), linear discriminant is sensitive to a large number of voxels in the left premotor area, which is contralateral to the subject's lesion. It is interesting that quadratic discriminant is not sensitive to these voxels, but instead is sensitive to a large number of voxels in both sides of the cerebellum.
- For Subject S103, linear discriminant is sensitive to voxels in the left motor area (contralateral to the lesions), but quadratic discriminant is not sensitive to this area.
- For Subject S145, quadratic (but not linear) discriminant is sensitive to several areas of the frontal and parietal lobes.

Note that for the two subjects who showed greatest performance improvement on the pinch test, namely S066 and S090 (see Fig. 1), quadratic (but not linear) discriminant is sensitive to several areas on both sides of the cerebellum.

4. Discussion

We have presented some interesting associations between measures derived from fMRI data and results of behavioral tests of recovery. Based on sphericity of the data covariance matrix, subjects showing a highly non-spherical covariance matrix across the four sessions (i.e. with a sharp drop in the eigenspectrum) show the greatest improvement in recovery measures. The divergence between the covariance matrices reflects similar but inverse behavioral relationships (see Table I and Figs. 1 and 2). Taken together these tests suggest that significant changes in the eigenspectra across sessions reflect underlying changes in the BOLD functional connectivity and the re-organization

of the brain, which leads to effective recovery. In particular the largest relative behavioral improvements are seen in those with the largest changes in their eigenspectra across longitudinal scanning sessions.

Our two other complexity measures (i.e., number of PCs that optimize QD performance, and unsupervised PCA dimensionality) show strong correlations with behavioral tests that are measured on the last session (six months post-stroke). PLS analysis of the data in Table I shows that these measures of complexity based on PC dimensionality tend to correlate with the behavioral measures of final recovery, and the measures computed on eigenspectra of the fMRI covariance matrix tend to correlate with behavioral measures of relative improvement in performance. These two directions of correlation are mutually orthogonal (Fig. 2), showing that they capture two relatively separate behavioral recovery processes: the absolute level of performance reached over 6 months recovery, and the change from the initial damaged brain required to reach this final performance. Fig. 1 shows that the subjects who have the best final performance (i.e., recovery) on two out of three behavioral tests require the *smallest* number of PCs to optimize QD classification, but the *largest* number of PCs to minimize the unsupervised generalization error.

One possible explanation for these results is consistent with our original motivating hypothesis. Subjects who come closest to recovering normal motor behavior come closest to reinstating the relationship between complexity, or unsupervised dimensionality, and behavior proposed by McIntosh et al. (2008, 2010), i.e., the better the behavior the larger the supporting brain complexity and associated dimensionality. This relationship is reflected in Fig. 1D, the positive correlation between final

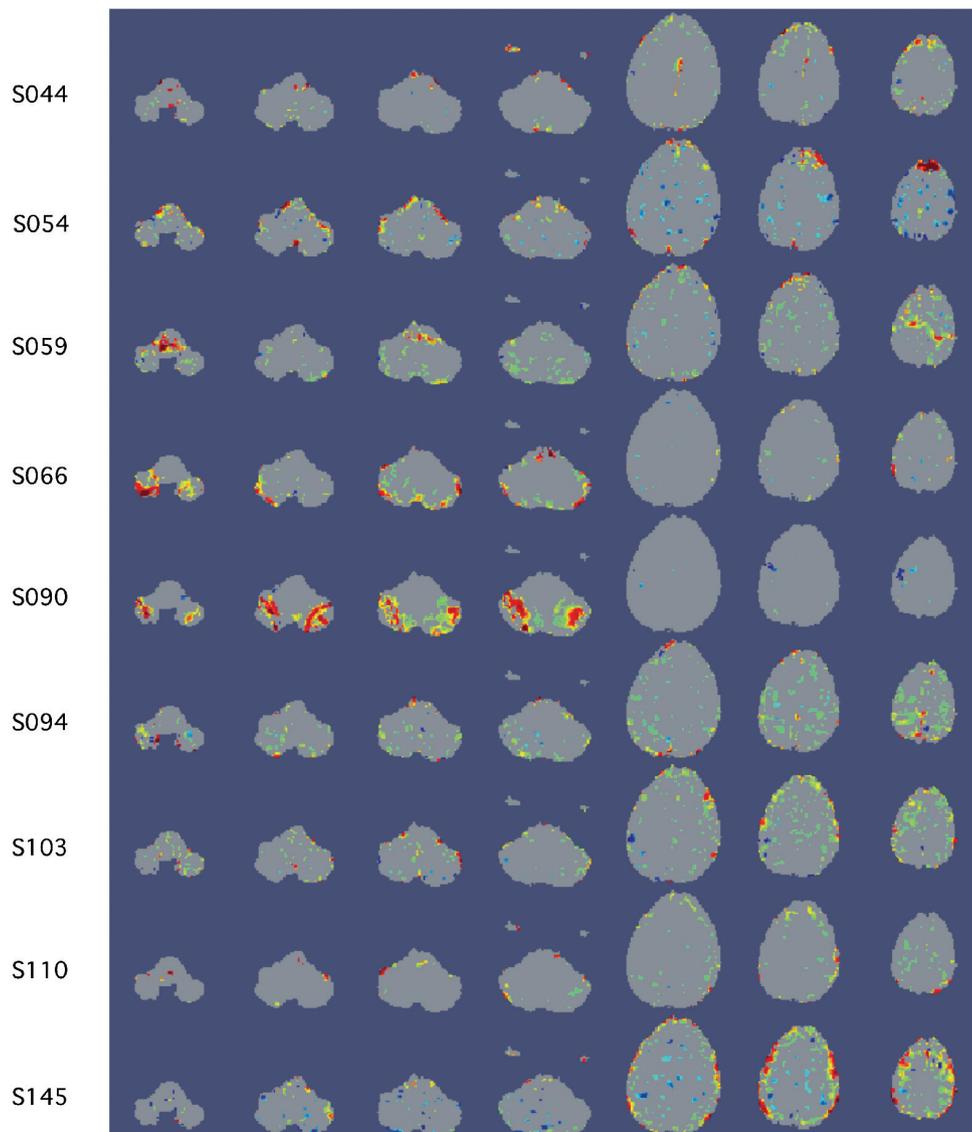


Fig. 4. - Mean differences between the scaled sensitivity maps for QD and LD on the early vs. late task. Each row corresponds to a different subject. Red areas are where QD has higher scaled sensitivity than LD. Blue areas are where LD has higher scaled sensitivity than QD. The differences in sensitivity have been thresholded at $p < 0.05$ using paired t -tests with a Bonferroni correction.

grip strength and unsupervised dimensionality in Table I, and the second principal component of Fig. 2. Subjects may show a relatively large behavioral improvement (upper plots, particularly S066 and S090), but may (S066) or may not (S090) come close to recovering normal function in 6 months. This large relative improvement while failing to recover near normal function (S090) is associated with intermediate to low brain complexity possibly reflecting the damaged and still somewhat disconnected nature of their motor networks (James et al., 2009). Curiously QD dimensionality (Fig. 1C) reflects an opposite association between dimension-

ality for maximum prediction and final behavior, which is highest for the poorest final behavioral performance. This may reflect yet a third influence on our results related to recent work on subspace phase transitions required to support optimal discrimination on a PCA basis set (Yourganov et al., 2010; Strother et al., 2010); the weaker and less coherent the signal the larger the dimensionality required to maximize discriminant performance.

The measures discussed so far give overall indications of complexity of functional connectivity or, in the case of the divergence of eigenspectra measure, a partial indication of the degree of change in

Table I.- Correlations between fMRI-based measures and behavioral measures. Each table entry shows Spearman's correlation coefficient, with the corresponding p -value in parentheses. The p -values are uncorrected for multiple comparisons, and therefore reflect trends in this limited data set.

	Peg test		Pinch strength		Grip strength	
	Improvement	Final	Improvement	Final	Improvement	Final
Sphericity index	-0.68 (0.05)	-0.30 (0.44)	-0.87 (0.0045)	-0.07 (0.88)	-0.43 (0.25)	0.17 (0.68)
Eigenspectra divergence	0.48 (0.19)	0.18 (0.64)	0.73 (0.03)	-0.18 (0.64)	0.50 (0.18)	-0.03 (0.95)
QD dimensionality	-0.28 (0.46)	-0.83 (0.008)	0.15 (0.71)	-0.38 (0.31)	-0.02 (0.98)	-0.42 (0.27)
Unsupervised dimensionality	-0.01 (0.99)	0.45 (0.22)	-0.21 (0.6)	0.67 (0.05)	0.16 (0.68)	0.78 (0.015)

functional connectivity during the course of stroke recovery. Sensitivity maps indicate that at least some of the brain regions are involved in changes in functional connectivity over time. If a voxel has a high sensitivity for QD but not for LD, for “early” versus “late” discrimination, this implies that functional connectivity involving that voxel has changed over time. We observed clusters of such voxels in various areas, but especially in the cerebellum, motor and premotor areas and in prefrontal cortex. Of particular interest is the fact that, for the two subjects who showed greatest performance improvement on the pinch test, namely S066 and S090 (see Fig. 1A), quadratic (but not linear) discriminant is sensitive to several areas on both sides the cerebellum despite the very different PCA subspaces required to support the near perfect prediction performance, i.e., 20 versus 190 PCs, respectively (see Fig. 1C). This suggests that changes in functional connectivity involving the cerebellum play an important role in stroke recovery.

Given the limited number and heterogeneity of subjects and old 1.5T BOLD fMRI scans available for this study we are encouraged by the relatively strong associations found between functional connectivity metrics and measures of behavioral recovery. We believe they strongly support use of such metrics to study recovery in longitudinal stroke studies with state-of-the-art BOLD fMRI acquisition at 3.0T. In addition, such studies help to shed light on optimal performance conditions for predictive classifiers and clearly show the importance of model regularization (i.e., dimensionality selection) and potential difficulties in generalizing such results beyond the current data set, and even across individual subjects when studying brain recovery.

Acknowledgments

This work was partly supported by a James S. McDonnell Foundation, USA (Bridging Brain, Mind, and Behavior) grant to Dr. McIntosh. SCS gratefully acknowledges the support of the Heart and Stroke Foundation of Ontario through the Centre for Stroke Recovery. We thank Natasa Kovacevic for co-registering and motion-correcting the fMRI data used in this study, and Hervé Abdi for his comments on applying the sphericity index and PLS analysis to fMRI data.

References

- Abdi H. The Greenhouse-Geisser correction. pp. 544-548. In: Salkind N. (Ed.) *Encyclopedia of Research Design*. Sage Publications, 2010.
- Hansen L.K., Larsen J., Nielsen F.-A., Strother S.C., Rostrup E., Savoy R., Lange N., Sidtis J., Svarer C., Paulson O.B. Generalizable patterns in neuroimaging: how many principal components? *NeuroImage*, **9**: 534-544, 1999.
- James A., Lu Z., Vanmeter J., Sathian K., Hu X., Butler A. Changes in resting state effective connectivity in the motor network following rehabilitation of upper extremity poststroke paresis. *Top Stroke Rehabil.*, **16** (4): 270-281, 2009.
- Kjems U., Hansen L.K., Anderson J., Frutiger S., Muley S., Sidtis J., Rottenberg D., Strother S.C. The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. *NeuroImage*, **15**: 772-786, 2002.
- Krishnan G., Williams L., McIntosh A., Abdi H. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage*, in press, 2010.
- LaConte S., Anderson J., Muley S., Ashe J., Frutiger S., Rehm K., Hansen L., Yacoub E., Hu X.,

- Rottenberg D., Strother S.C. The evaluation of pre-processing choices in single-subject BOLD fMRI using NPAIRS performance metrics. *NeuroImage*, **18**: 10-27, 2003.
- Mardia K., Kent J., Bibby J. *Multivariate Analysis*. Academic Press, San Diego, 1979.
- McIntosh A.R. and Lobaugh N.J. Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage*, **23** (S1): S250-263, 2004.
- McIntosh A., Kovacevic N., Itier R. Increased brain signal variability accompanies lower behavioral variability in development. *PLoS Comput. Biol.*, **4** (7), 2008.
- McIntosh A.R., Kovacevic N., Lippe S., Garrett D., Grady C., Jirsa V. The development of a noisy brain. *Arch. Ital. Biol.*, 2010 (this issue).
- Schmah T., Yourganov G., Zemel R.S., Hinton G.E., Small S.L., Strother S.C. A comparison of classification methods for longitudinal fMRI studies. *Neural Computation*, in press, 2010.
- Small S.L., Hlustik P., Noll D.C., Genovese C., Solodkin A. Cerebellar hemispheric activation ipsilateral to the paretic hand correlates with functional recovery after stroke. *Brain*, **125** (7): 1544-1557, 2002.
- Strother S.C., Anderson J., Hansen L. K., Kjems U., Kustra R., Sidtis J., Frutiger S., Muley S., LaConte S., Rottenberg D. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, **15**: 747-771, 2002.
- Strother S.C., Oder A., Spring R., Grady C. The NPAIRS computational statistics framework for data analysis in neuroimaging. pp. 111-120. In: Lechevallier Y. and Saporta G. (Eds.), *19th International Conference on Computational Statistics*. Physica-Verlag, Springer. 2010.
- Woods R.P., Grafton S.T., Holmes C.J., Cherry S.R., Mazziotta J.C. Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, **22**: 139-152, 1998.
- Yourganov G., Xu C., Lukic A., Grady C., Small S., Wernick M., Strother S.C. Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data. In press, *NeuroImage*, 2010.
- Zurada J.M., Malinowski A., Usui S. Perturbation method for deleting redundant inputs of perceptron networks. *Neurocomputing*, **14** (2): 177-193, 1997.