NeuroImage 56 (2011) 531–543

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data

Grigori Yourganov [a,b,*], Xu Chen [c], Ana S. Lukic [d], Cheryl L. Grady [b,e], Steven L. Small [f], Miles N. Wernick [g,d], Stephen C. Strother [a,b,d]

[a] Institute of Medical Science, University of Toronto, Toronto, ON, Canada
[b] Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, ON, Canada
[c] Case Center for Imaging Research, Case Western Reserve University, Cleveland, OH, USA
[d] Predictek, Inc., Chicago, IL, USA
[e] Department of Psychology, University of Toronto, Toronto, ON, Canada
[f] Department of Neurology, University of Chicago, Chicago, IL, USA
[g] Medical Imaging Research Center, Illinois Institute of Technology, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 27 November 2009
Revised 10 September 2010
Accepted 14 September 2010
Available online 19 September 2010

Keywords:
Dimensionality estimation
Model order selection
Linear discriminant Analysis (LDA)
Principal component analysis (PCA)
Signal detection
fMRI

ABSTRACT

Estimation of the intrinsic dimensionality of fMRI data is an important part of data analysis that helps to separate the signal of interest from noise. We have studied multiple methods of dimensionality estimation proposed in the literature and used these estimates to select a subset of principal components that was subsequently processed by linear discriminant analysis (LDA). Using simulated multivariate Gaussian data, we show that the dimensionality that optimizes signal detection (in terms of the receiver operating characteristic (ROC) metric) goes through a transition from many dimensions to a single dimension as a function of the signal-to-noise ratio. This transition happens when the loci of activation are organized into a spatial network and the variance of the networked, task-related signals is high enough for the signal to be easily detected in the data. We show that reproducibility of activation maps is a metric that captures this switch in intrinsic dimensionality. Except for reproducibility, all of the methods of dimensionality estimation we considered failed to capture this transition: optimization of Bayesian evidence, minimum description length, supervised and unsupervised LDA prediction, and Stein's unbiased risk estimator. This failure results in sub-optimal ROC performance of LDA in the presence of a spatially distributed network, and may have caused LDA to underperform in many of the reported comparisons in the literature. Using real fMRI data sets, including multi-subject group and within-subject longitudinal analysis we demonstrate the existence of these dimensionality transitions in real data.

Introduction

Typical fMRI data sets consist of a relatively small number of temporally correlated observations recorded for a large number of spatial locations (voxels). When trying to create activation maps with multivariate analysis and treating each fMRI volume as a vector, one is faced with an ill-posed problem, because the number of variables (voxels) greatly exceeds the number of observations (fMRI volumes). In addition, not all dimensions of the vector space are useful, as many are dominated by noise. A well-known strategy to reduce the dimensionality and noise prior to analysis is to discard a subset of the components in a principal component analysis (PCA) (Friston et al., 1995a; Strother et al., 1997; Tegeler et al., 1999; Hansen et al., 1999; Laconte et al., 2003). This requires estimation of the dimensionality of the signal-carrying subspace, a task for which many solutions have been proposed in the literature.

Intrinsic PC dimensionality of the data is defined as the number of principal components (PCs) that contain signal and should be retained for further analysis. Some early methods (or, more accurately, rules-of-thumb) of intrinsic dimensionality estimation are described in Mardia et al. (1979). For example, one can retain the PCs that, taken together, explain 90% of the variance in the data; or one can look for the "knee" in a "scree plot" (the point at which the eigenvalue spectrum of the covariance matrix flattens out, which in a white-noise model indicates noise-dominated components). Both of these methods are subjective, because the threshold of 90% is an arbitrary choice, and the scree-plot method involves visual inspection.

Beckmann and Smith (2004) discuss (alongside other methods of dimensionality estimation) a more-sophisticated technique based on scree plots, which works as follows. When the data are Gaussian-distributed, the estimated covariance matrix has a Wishart distribution, and one can calculate the expected distribution matrix's eigenvalues

* Corresponding author. Rotman Research Institute, Baycrest Centre for Geriatric Care, 3560 Bathurst St., Toronto, ON, Canada M6A 2E1. Fax: +1 416 785 2862.
E-mail address: gyourganov@rotman-baycrest.on.ca (G. Yourganov).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.09.034

analytically. PCs corresponding to eigenvalues that are not significantly different from the expected eigenvalues of a Wishart matrix are discarded. However, this method assumes a well-posed problem (the number of samples should be at least equal to the number of voxels) making it impractical for fMRI data analysis without voxel-based feature selection. Peres-Neto et al. (2005) gives a review of several other methods of dimensionality estimation that work for well-posed problems, and compared their performance when the data are not Gaussian-distributed. The best-performing methods were based on permutation tests, where independent observations were randomly shuffled 999 times.

Several techniques for estimating the number of signal-carrying PCs have been developed specifically for ill-posed problems. These methods can be classified into two categories: analytic and empirical. Analytic methods are based on information-theoretic criteria to estimate the optimal number of principal components. We have studied several analytic methods, such as minimal description length (Calhoun et al., 2001; Li et al., 2007), Stein's unbiased risk estimator (Ulfarsson and Solo, 2008), and Laplace approximation to Bayesian evidence (Minka, 2000), and examined their utility in situations usual to fMRI research, where sample size is small and signal-to-noise ratio is low.

Empirical methods of dimensionality estimation select the number of principal components to optimize some metric of performance calculated with resampling techniques; therefore, these methods are typically more computationally expensive than analytic methods. Early examples of empirical dimensionality estimation are Wold (1978) and Eastment and Krzanowski (1982); see also Krzanowski and Kline (1995). They proposed a method when the matrix is approximated with a subset of principal components, and dimensionality is estimated by the number of principal components that results in approximation with an optimal predicted residual sum of squares (PRESS) statistic. Each element of the matrix is compared with its corresponding element in a PC approximation. However, we need to make sure that the information about the element was not used in the approximation of this element (i.e. to ensure the independence of training and test data). For this purpose, the PC approximation of an element is computed on a matrix from which the row and the column containing this element have been removed. This leads to a rather cumbersome and very computationally expensive procedure of numerous PC decompositions on the matrices from which a particular row and column of the original data matrix were removed.

More recently, Hansen et al. (1999) proposed another method based on cross-validation, in which the data were separated into independent training and test sets and the number of PC components that minimize generalization error of the test set was used as an intrinsic dimensionality estimate. Strother et al. (1997, 2002) introduced the split-half resampling framework, where the data are split into two independent sets of roughly the same size. The number of PCs to be retained is selected to optimize the reproducibility of activation maps (calculated separately on the two half-sets), or classification accuracy (when one half serves as a training set, and the other as a test set). There is usually a tradeoff between reproducibility and classification accuracy, and one might wish to optimize a combination of these two metrics (LaConte et al., 2003; Strother et al., 2004; Jacobsen et al., 2008).

Several authors have compared the efficacy of various dimensionality estimation methods. Minka (2000) has shown that his method of optimizing Bayesian evidence is significantly more accurate than 5-fold cross-validation if the sample size and number of voxels in simulated data are both small (<15). His method was also better than cross-validation when the data were non-Gaussian, which was the reason to use his method in probabilistic independent component analysis (Beckmann and Smith, 2004), in which the signal sources are assumed to be non-Gaussian. When the number of observations was larger (>60, in both well-posed and ill-posed situations), optimization of Bayesian evidence was found to be slightly, but not significantly, better than cross-validation in estimating the dimensionality of simulated Gaussian data. Cordes and Nandy (2006) have shown that estimates of dimensionality calculated with analytic methods are strongly influenced by sample size (when the number of observations grows, so do the dimensionality estimates, although the underlying intrinsic dimensionality stays the same). Li et al. (2007) have addressed this problem by subsampling the data, when a large portion of observations is discarded prior to dimensionality estimation so the remaining observations are independent and identically distributed. Ulfarsson and Solo (2008) have shown their method to be more accurate than both Minka's method and minimum description length in simulated Gaussian data. In their simulations, the ratio of the number of variables to the number of observations did not exceed 5/2.

Using the simulation framework introduced in Lukic et al. (2002), we tested a wide selection of dimensionality estimation methods, and compared their performance in situations when the task-related signal was organized into a spatial network of functionally connected loci. In the simulations, we sampled the signal from a multivariate Gaussian distribution, and embedded it in additive Gaussian noise. We used linear discriminant analysis (LDA) on a PC subspace as our basic method of analysis. When the loci are not correlated, signal detection is optimal when the number of PCs in our analysis is roughly equal to the number of loci. However, as the correlations increase and the spatial network starts to become apparent in the data, intrinsic dimensionality goes through a transition from many dimensions to a single dimension, and signal detection is optimized when we use just one PC. This transition is captured when we estimate the intrinsic dimensionality by optimizing the reproducibility of activation maps; all other methods of estimation fail to capture the transition.

Using a metric of global signal-to-noise ratio (gSNR) based on the reproducibility of independent spatial activation maps (Strother et al., 2002, 2010), we demonstrate an asymptotic relationship between optimal dimensionality and gSNR: when gSNR is high enough, the network can be captured with a single dimension, but as gSNR drops, the optimal dimensionality starts to rise sharply. We have shown this asymptotic relationship in two sets of real data: analysis of multi-subject groups from a study of cognitive impairment associated with aging (Grady et al., 2006), and within-subject analysis of a longitudinal stroke study (Small et al., 2002).

This result has a parallel in statistical physics: in order to capture the structure of the data, the ratio of number of observations to the number of dimensions has to reach a certain critical level. A phase transition happens at this point, and, if this level has not been reached, it is impossible to identify the signal-carrying components in the noisy data although good signal detection is still possible under some circumstances (Watkin and Nadal, 1994; see also Results and Discussion below).

Materials and methods

Simulated data

In this study we used computer-generated data to simulate a block-design experiment with two conditions: *activation* and *baseline* (refer to Lukic et al. (2002) for details of the simulation). All images contained the same simplified single-slice "brain-like" background structure with additive Gaussian noise. An elliptical background structure contained in a 60×60 pixel image consisted of "grey matter" in the center and on the rim of the phantom, and "white matter" in between; the amplitude of the background signal in the "grey matter" was 4 times higher than in the "white matter". Gaussian noise was spatially smoothed using a Gaussian filter with full-width-at-half-maximum (FWHM) of 2 pixels. After smoothing, the standard deviation of the noise was 5% of the background signal. Images in the "activation" condition contained 16 Gaussian-shaped signal "blobs" distributed over the image (12 in the "grey matter" and 4 in the "white matter") and added to the smoothed noisy background

image. Fig. 1 shows examples of baseline and activation images (noise is not displayed; although activation signal could be negative as well as positive, we only show positive signal for the sake of clarity). The FWHM of the activation blobs varied between 2 and 4 pixels. Simulated experimental data sets were composed of N baseline and N activation images per set ($N=100$), so the total number of observations in a set is $2N=200$. We used a mask with $J=2072$ pixels covering the "brain" to exclude locations outside of the phantom from analysis.

Images were arranged into 10 "epochs" of 20 images each to simulate a block design with epochs of 10 "baseline" images followed by 10 "activation" images. To simulate the hemodynamic response, each pixel's time course was convolved with a hemodynamic response function (HRF) defined by the sum of two Gamma functions (Glover, 1999). Parameters of the HRF model were taken from Worsley (2001): $a_1=6$, $a_2=12$, $b_1=b_2=0.9$ s, $c=0.35$, TR = 2 s.

Amplitudes of the Gaussian activation signal blobs were sampled from a multivariate Gaussian distribution. The mean amplitude of each activation was specified proportionally to the local value of the background signal:

$$E[a_k] = Mb_k, \tag{1}$$

where a_k is the amplitude of kth activation, $E[a_k]$ is its expected value, b_k is the value of noise-free baseline image at the center of the kth activation, and M is the proportionality constant. To study the effect of M on dimensionality estimation, M was set to different levels (0.01, 0.03 and 0.05) in different realizations of our simulated experiment. These levels of M corresponded to contrast-to-noise ratios (CNRs) of 0.2, 0.6 and 1.0. HRF convolution changed these values to empirical measurements of 0.3, 1.0 and 1.6, respectively. The variance of the amplitude of the Gaussian activation signal in our multivariate Gaussian distribution, denoted by σ_k^2, was defined proportionally to the variance of the independent background Gaussian noise added to each voxel, v_k^2:

$$\sigma^2 = Vv_k^2, \tag{2}$$

where the proportionality constant V was varied from 0.1 to 1.6 in different realizations of the experiment. In this paper, we refer to V as the "relative signal variance", which may be thought of as a form of physiological variation of the activation signal. The third parameter of our multivariate Gaussian model was the correlation coefficient, ρ, which defined the covariance between Gaussian activation signal amplitudes at kth and lth locations ($k \neq l$):

$$cov(a_k, a_l) = \rho\sigma_k\sigma_l. \tag{3}$$

The value of ρ was set to 0, and 0.5 and 0.99 to define a simple distributed spatial network (Lukic et al., 2002).

The amplitudes of the multivariate Gaussian signal in the "active" state are defined by the three parameters: CNR (or M), V and ρ. These values are the same for all volumes in a simulated experimental set, but may differ across sets. All CNR values are those measured empirically after convolution by the HRF.

This Gaussian signal simulation incorporates the three ideas of (1) a mean signal level, (2) physiological variation of signal levels about their mean across successive scans, and (3) a partition of this signal variation between physiological noise and network variation defined by the chosen value of the correlation coefficient coupling distributed Gaussian blobs.

We have also tested the algorithms of dimensionality estimation on large-sample data sets. Twenty artificial data sets were created using $N=1500$ images per class, so each data set consisted of $2N=3{,}000$ images, versus the 200 images data sets described above. The parameters of multivariate Gaussian signal for these sets were set to CNR = 1, $V=1.1$, $\rho=0.5$. Data were convolved with the HRF model as described above.

Real data

In addition to analyzing simulated data, we analyzed two real fMRI data sets to investigate whether similar results would be found.

Group aging study

The first was acquired by Grady et al. (2006). The subjects belonged to three different age groups: young (20 to 30 years of age, 10 subjects), middle-aged (40 to 60 years, 12 subjects) and old (65 to 78 years, 11 subjects). They were presented with black line drawings of nameable objects, and words corresponding to names of objects. The experiment consisted of two "shallow" memory-encoding tasks, two "deep" encoding tasks, and two recognition tasks. During the two shallow-encoding tasks, the subjects were asked to report whether the pictures were large or small, and whether the words were printed in upper or lower case. During the two deep-encoding tasks, the subjects were asked to determine whether the pictures (or words) corresponded to living or non-living entities. During the two recognition tasks, the subjects reported whether or not they had seen the presented stimuli (pictures or words) previously.

The BOLD signal was measured by using a 1.5T magnetic-resonance imaging (MRI) scanner; 26 slices were acquired (5 mm thick; TR = 2.5 s, TE = 40 ms, flip angle 80°). Motion correction was performed with the AFNI software package (Cox, 1996), and between-subject alignment was performed using the FSL package (Smith et al., 2004). The data were smoothed with a Gaussian kernel (FWHM = 6 mm). For each subject, one run was collected for every

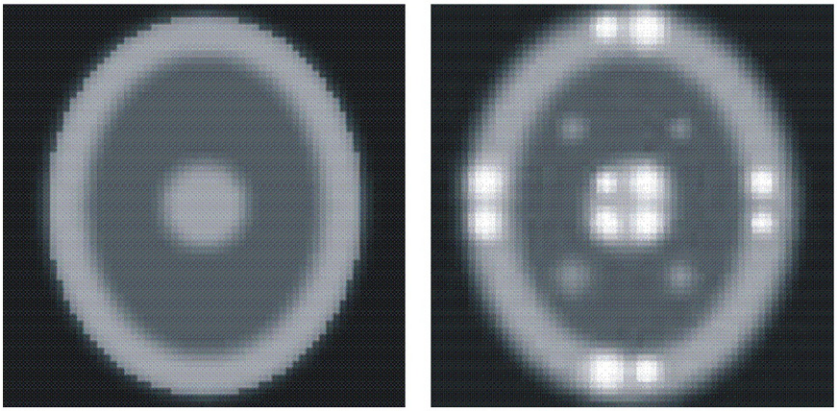

Fig. 1. The phantom in baseline (left) and activation (right) states. Noise is not displayed.

task (89 volumes for encoding tasks, and 166 volumes for recognition tasks). Voxels outside the brain were excluded by masking, leaving 50,308 voxels for further analysis. For further experimental details see Grady et al. (2006).

Stroke recovery study

We also analyzed data collected by Small et al. (2002), in a longitudinal block-design study of stroke recovery, in which 9 stroke patients were scanned in 4 different sessions at 1, 2, 3 and 6 months after the stroke. During each session, subjects were instructed to perform a motor task (finger tapping alternating with wrist flexion) alternating with blocks of rest. Two runs were recorded for subjects performing the task with their healthy hand, and 2 runs with the hand impaired by the stroke. Whole-brain fMRI data were acquired using 1.5T MRI scanner (6 mm thick; TR = 4 s, TE = 35 ms, flip angle 60°). To ease the computational load for more complex nonlinear classifiers in a separate study, we used 7 horizontal slices out of the available 24; we selected the slices that corresponded to parts of the brain involved in finger and wrist motion (4 slices containing the cerebellum and 3 slices containing the motor areas). The number of voxels selected for analysis was 10,499.

The data for all 9 subjects were co-registered spatially and corrected for motion. The volumes within each motor-task block were divided by the average volume of the last two scans of the preceding rest block, for intensity normalization and to filter out low temporal frequencies (McIntosh and Lobaugh, 2004). Rest blocks were discarded after that. By pooling the data across runs and sessions, we obtained 1280 volumes per individual subject; this large number of volumes per subject allowed us to carry out within-subject analysis easily (by contrast, in the aging study the data were analyzed for each group pooled across subjects). This analysis is described in detail in Schmah et al. (2010).

ROC methodology

Detection of the activation signal in the simulated data was evaluated using receiver operating characteristic (ROC) methodology utilizing the knowledge of the "ground truth" in the simulation. The ROC methodology evaluates the efficacy of the algorithm in terms of its capability to make Type I (false positive) and Type II (false negative) detection errors at known signal locations. Error frequency was calculated by generating data under two hypotheses, H_1 (activation present) and H_0 (activation absent), with 500 simulated experimental data sets for each hypothesis:

- H_1: each data set consisted of N activation images and N baseline images ($N = 100$);
- H_0: each data set consisted of $2N$ baseline images.

Dimensionality was estimated on each H_1 set separately, and each set was processed by linear discriminant analysis using that estimate. H_0 sets were processed with the dimensionality estimated on the corresponding H_1 set to reflect real-world processing constraints in which H_0 data are typically unavailable. Values of the test statistic were measured at each of the center voxels of the 16 Gaussian activation loci. For each location, a ROC curve was calculated using LABROC1 software (Metz et al., 1998), and the partial area under the curve (corresponding to the frequency of false positives between 0 and 0.1) was used as a metric of algorithm performance (see Fig. 2).

Linear discriminant analysis

We chose Fisher's Linear Discriminant Analysis (LDA) as the primary method of data analysis for this study based on excellent results obtained in earlier work when the PC subspace is tuned to maximize signal detection (Schmah et al., 2010; LaConte et al., 2003; Strother et al., 2004, 2010; Zhang et al., 2008). The discriminant

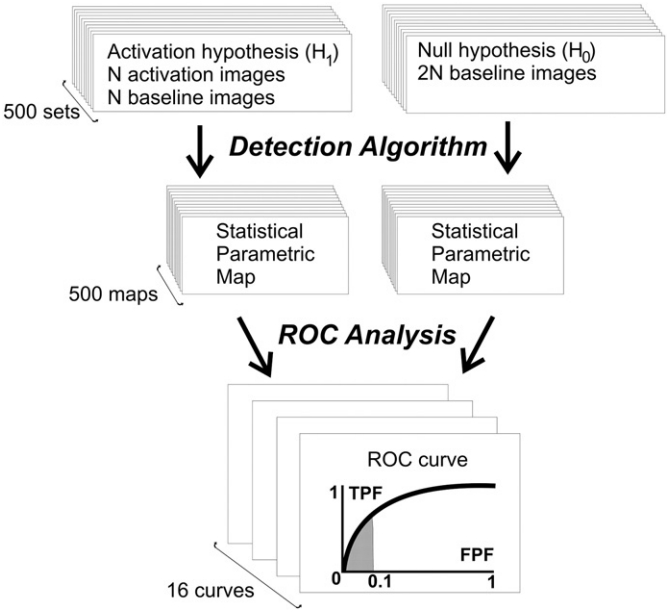

Fig. 2. Scheme of the ROC calculation process.

vectors are the eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$ are within- and between-class covariance matrices (Mardia et al., 1979). In our case, we have two classes of images ("active" and "baseline" states), and $\mathbf{W}^{-1}\mathbf{B}$ has only one eigenvector, given by $\mathbf{W}^{-1}\boldsymbol{\Delta}\mathbf{m}$, where $\boldsymbol{\Delta}\mathbf{m}$ is the difference between the mean vectors of the two classes. This eigenvector defines the hyperplane that separates the two classes. When the data are multinormal and homoscedastic (i.e. all classes share the same within-class covariance matrix), this separation is optimal in a least-squares sense. When the assumption of homoscedasticity is not satisfied, the performance of LDA declines; however, this decline is found to be very slight if the results are regularized by correctly estimating the intrinsic dimensionality (see below).

Typically, the dimensionality of fMRI data (the number of voxels) greatly exceeds the sample size (the number of time points), and, therefore, the within-class covariance matrix $\mathbf{W}$ is rank-deficient and cannot be inverted. To avert this problem, we estimate the linear discriminant in a principal components subspace, approximating data matrix $\mathbf{X}$ by its first K components:

$$\mathbf{X} = \mathbf{USV}^T \approx \mathbf{U}_K\mathbf{S}_K\mathbf{V}_K^T = \mathbf{U}_K\mathbf{Z}_K \tag{4}$$

which allows us to compute the within-class covariance matrix $\mathbf{W}$ on the K-dimensional columns of the $\mathbf{Z}_K$ matrix. To project the discriminant $\mathbf{W}^{-1}\boldsymbol{\Delta}\mathbf{m}$ into the image space, we pre-multiply it by $\mathbf{U}_K$. Selection of dimensionality K both of the intrinsic underlying network signal and for optimization of LDA signal detection is the central issue of this study.

Dimensionality estimation

In our implementation, the performance of the linear discriminant critically depends on the choice of K, the number of principal components that define the subspace on which the LDA operates. Ideally, these K components contain most of the signal of interest, while the remaining $2N$-K components are dominated by noise and can be discarded. In practice, we must select a metric that we can use to separate the potential signal-carrying components from those containing predominantly noise, and find K that optimizes this metric. One goal of this paper is to test and compare the wide range of such

metrics that are discussed below. Note that in all data sets analyzed, both simulated and real, we subtract the mean spatial pattern from the data matrix before performing PCA and estimating the dimension of the subspace to be used with LDA.

ROC optimization

To optimize signal detection, we seek the value of K that maximizes the area under the ROC curve for a useful range of false-positive probabilities. Because one is typically interested in low numbers of false positives, we maximize the area under the portion of the ROC curve that corresponds to frequencies of false positives not exceeding 0.1. However useful this metric is when applied to simulated data, it is unsuitable for real fMRI data since it requires the knowledge of "ground truth," i.e. precise location of truly activated loci.

Reproducibility of spatial maps

Another metric, which is applicable to both real and synthetic data, is based on the reproducibility of spatial maps using a split-half resampling procedure (Strother et al., 1997, 2002, 2010). The data set is randomly split into two halves containing independent subsamples (e.g., in our simulations splitting on epochs), and spatial maps are computed using any model of analysis for each half; in our case using LDA on K principal components. Then we calculate Pearson's correlation coefficient (r) between these two maps (treating each map as a vector). Note that this is equivalent to some recent uses of an intra-class correlation coefficient (Raemaekers et al., 2007). We perform 20 such splits, and the median value of r serves us as a measure of reproducibility. The value of K that maximizes this measure is our estimate of dimensionality.

Since the two split halves should be independent of one another, we split the simulated data so that all the images from the same epoch remain in the same half-set. Also, we exclude the first two images of each set of baseline or activation images to avoid transient effects of the HRF that are not captured in simple Fisher's LDA. In the simulated data, this leaves 160 of the original 200 images for analysis with each split-half subsample containing 80 images. To make sure that the discriminant activation maps from the two split-halves have the same sign (i.e. to avoid the potential ambiguity of the sign of eigenimages in PC decomposition), we match the split-half activation maps to a reference activation map computed on the full data set using exactly the same estimation parameters, i.e., K. If the correlation coefficient between the split-half map and the reference map is negative, we flip the sign of the split-half map. Reproducibility in the real data sets is estimated similarly as described in Strother et al. (2010).

Classification accuracy

Classification accuracy can also serve as a metric of efficacy of our dimensionality estimation. In our simulations we have used leave-one-epoch-out, 10-fold cross-validation, where LDA was performed on the training set with K components used to classify the test set (volumes from the left-out epoch) into "baseline" and "activation" classes. Splitting of the data into training and test sets was performed 10 times, and dimensionality was estimated as the value of K that optimizes median classification accuracy across 10 splits. For the real data sets classification was estimated using the 20 split-half sets described for reproducibility above, with each pair of split-half subsamples defining a training and test set as described in Strother et al. (2010).

Unsupervised prediction

Hansen et al. (1999) have proposed another method of dimensionality estimation that uses resampling. The data are randomly split into training and test sets, and the unsupervised generalization error of the test set is minimized to estimate the dimensionality. In our implementation, we use split-half resampling as described above so the training and test sets are approximately equal in size. Generalization error is defined as

$$G(K) = \frac{1}{2}\log\left|2\pi\mathbf{S}_K^{(train)}\right| + \frac{1}{2}\mathrm{trace}\left[\mathbf{S}_K^{(train)-1}\mathbf{S}^{(test)}\right] \quad (5)$$

Here, $\mathbf{S}_K^{(train)}$ and $\mathbf{S}^{(test)}$ are the sample covariance matrices of the training and the test sets, respectively. We assume that signal is contained in the first K principal components of the training set and the remaining components contain white Gaussian noise, so we can write $\mathbf{S}_K^{(train)}$ as $\mathbf{S}_K^{(signal)} + \sigma^2\mathbf{I}$, where as $\mathbf{S}_K^{(signal)}$ is constructed using the first K principal components of training-set covariance matrix, σ^2 is the estimated noise variance and $\mathbf{I}$ is the identity matrix. We need to find the value of K that results in minimization of $G(K)$ for each of the 20 splits described above.

Minimum description length

Separation of signal from the noise can be achieved with metrics from information theory, such as minimum description length (MDL) (Wax and Kailath, 1985; Li et al., 2007; Calhoun et al., 2001; note that the latter paper contains an error in the estimation of degrees of freedom of the model). In simplified form, this metric is given by

$$MDL(K) = -J\left(\sum_{i=K+1}^{2N}\log l_i - \prod_{i=K+1}^{2N}\log l_i\right) + \frac{1}{2}\tilde{k}\log(2N) \quad (6)$$

Here J is number of voxels, N is the number of images per class (so $2N$ is the total number of images), K is the number of principal components assumed to contain signal, $\tilde{k}$ is the penalty term:

$$\tilde{k} = K(2N + 0.5) - K^2/2 \quad (7)$$

and l_i is the ith eigenvalue of the sample covariance matrix of the data. We have used the implementation of MDL given by the MATLAB function icatb_estimate_dimension from the GIFT toolbox (version 2.0c, downloaded from http://icatb.sourceforge.net).

Bayesian evidence

Minka (2000) has proposed another way of estimating dimensionality based on information theory. He uses the probabilistic principal component analysis (Tipping and Bishop, 1999) framework, in which the data vectors are represented as a weighted sum of basis vectors $\{\mathbf{h}_i\}$, plus the mean vector $\mathbf{m}$, and the error vector $\mathbf{e}$:

$$\mathbf{x} = \sum_{i=1}^{K} w_i\mathbf{h}_i + \mathbf{m} + \mathbf{e} = \mathbf{Hw} + \mathbf{m} + \mathbf{e}. \quad (8)$$

The error term $\mathbf{e}$ is Gaussian white noise: $p(\mathbf{e}) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We aim to find the dimensionality K that optimizes the Bayesian evidence of the data, i.e. probability of the data given K:

$$p(\mathbf{X}|K) = \int_\theta p(\mathbf{X}|\theta)p(\theta|K)d\theta, \quad (9)$$

where $\theta = \{\mathbf{H}, \mathbf{m}, \sigma\}$ contains the parameters of the model. To simplify the integration, Laplace interpolation (Bishop, 2006) is used. We have used the implementation provided by MELODIC software (version 3.05) from the FSL software package (Beckmann and Smith, 2004).

Stein's unbiased risk estimator (SURE)

Ulfarsson and Solo (2008) introduced an analytic method of estimating PC dimensionality of the data. The objective function is the unbiased estimate of the L2-norm risk (i.e. mean squared error),

proposed by Stein. The signal is represented by the probabilistic PCA model:

$$\mathbf{x} = \mathbf{Hw} + \mathbf{m} + \mathbf{e} = \boldsymbol{\mu} + \mathbf{e}, \tag{10}$$

and the maximum-likelihood estimate of $\boldsymbol{\mu}$ is calculated using the first K components of the data. The objective function is the unbiased estimate of the L2-norm risk (i.e. mean squared error), proposed by Stein:

$$R(K) = \frac{1}{2N}\sum||\mathbf{x}_t - \hat{\boldsymbol{\mu}}||^2 + \frac{\sigma^2}{N}\sum \text{trace}\frac{\partial\hat{\boldsymbol{\mu}}}{\partial\mathbf{x}_t} - J\sigma^2. \tag{11}$$

The variance of the noise σ^2 was estimated using methods from random matrix theory, by approximating the median eigenvalue of noisy components with the theoretical estimate using the Marcenko–Pastur distribution.

Other methods of analysis: GLM, PCA, PDA

To compare the efficacy of PC-based LDA to other methods of analysis, we have analyzed the simulated data with several methods that are well-established in the community: general linear model (GLM), principal component analysis (PCA) and penalized discriminant analysis (PDA) with a ridge penalty. Univariate GLM (Friston et al., 1995b) creates activation maps by computing the regression of each voxel's time course with the reference time course (the box-car function convolved with model HRF). A t statistic is obtained by dividing the regression coefficient by the standard deviation of the residuals. PCA was evaluated by using the first eigenimage of the data matrix as our activation map. Penalized discriminant analysis (Kustra and Strother, 2001) regularizes LDA with ridge regression: since we cannot compute $\mathbf{W}^{-1}\Delta\mathbf{m}$ (because $\mathbf{W}$ is not invertable), we compute penalized discriminant $(\mathbf{W} + \lambda\mathbf{I})^{-1}\Delta\mathbf{m}$. In our implementation, the penalty term λ is tuned to optimize reproducibility of activation maps using split-half resampling with a common penalty term applied to both split-half PDA models. Reproducibility is computed as Pearsons correlation coefficient between the activation maps, analogously to optimization of reproducibility for dimensionality estimation, described above. For computational efficiency, initial SVD was performed on the data set, penalized discriminant was computed in the PC space (all PCs were kept) and then projected into the image space.

Global Signal-to-noise Ratio

Ideally, the spatial map obtained from one group of subjects should be identical to that obtained from a disjoint set of similar subjects. Of course, in reality, this ideal is not realized owing to imaging noise and physiological variability. However, reproducibility can be used as a measure of the degree to which this ideal is achieved, and can serve as an effective measure of global signal-to-noise ratio (Strother et al., 2002, 2010).

For a given data set, when the data are divided using split-half resampling, reproducibility can be defined as the Pearson correlation coefficient r between the voxel values in spatial maps obtained from the two halves of the data set. If one standardizes both spatial maps (to have zero mean and unit variance), then the scatter plot of voxel values from one spatial map versus those in the other map will appear as an ellipse-shaped distribution for which the major axis is the line of identity (at 45 degrees to the axes of this two-dimensional space). To the extent that the two spatial patterns contain common, reproducible information, spatial-map components in the direction of the major axis must, by definition, contain a mixture of signal and noise, while the minor-axis direction contains only noise components that are uncorrelated with the signal. By construction, the variance along the signal-bearing (major) axis is, $e_1 = 1 + r$, and the variance along the noise axis (the minor axis) is, $e_2 = 1 - r$. Thus, the signal variance can be computed as $e_1\text{-}e_2 = (1 + r) - (1 - r) = 2r$, and we can define a measure of global signal-to-noise based on $(e_1\text{-}e_2)/e_2$ using reproducibility as:

$$gSNR = \sqrt{\frac{\text{signal variance}}{\text{noise variance}}} = \sqrt{\frac{2r}{1-r}}. \tag{12}$$

When the signal contains no reproducible signal structure ($r = 0$), the major and minor axes are equal and the scatter plot becomes a circle. In this case, $gSNR = 0$. As the reproducibility grows ($r \to 1$), the scatter plot becomes more elongated along the identity line, and $gSNR$ grows to infinity.

Results

Simulated data

Figs. 3A and B show the dimensionality of simulated data estimated by the methods described above. The contrast-to-noise ratio was fixed at 0.3 (Fig. 3A) and at 1 (Fig. 3B). The three panels from left to right in A and B correspond to levels of long-range spatial correlation in the network ($\rho = 0$, $\rho = 0.5$, $\rho = 0.99$), and the horizontal axis shows the relative signal variance, V, for each panel. The plots record the median estimate of dimensionality, measured across 500 data sets, and the error bars show the 25%-75% percentile range of the estimates. Error bars are displayed only for the smallest (0.1) and the largest (1.6) levels of V, and there are no error bars for ROC optimization curves as all 500 data sets are used to generate a single ROC curve for each plot value.

When $\rho = 0$ (left panel), the variation in signal amplitude is independent across the active loci, and the intrinsic dimensionality of our data appears to be approximately equal to the number of active Gaussian blobs, which is 16 in our simulation, i.e., the number of independent eigenvalues along the diagonal of the data covariance matrix. When $\rho > 0$ (middle and right panels), the loci become coupled and while there are still 16 partly independent Gaussian blob amplitudes the intrinsic dimensionality switches from approximately 16 dimensions to one dimension when the covariance structure is dominated by the coupled network with increasing V and ρ. ROC optimization captures this transition in dimensionality when the relative signal variance is sufficiently large for the network structure to be reflected in the signal ($V > 1$ for $\rho = 0.5$, $V > 0.5$ for $\rho = 0.99$; see 1st eigenvalue levels in Fig. 5).

Optimization of reproducibility yields dimensionality estimates that follow ROC optimization quite well, although the estimation is not very robust at $\rho = 0$ and at lower levels of V when $\rho > 0$. When a one-dimensional network structure is strongly manifest in the data with large V and ρ, optimization of reproducibility identifies it very robustly. The estimates obtained with optimization of classification accuracy only start to show the same decreasing trend as V grows for CNR = 1.0. In addition, they are much less robust compared to optimization of reproducibility. The variation in estimates increases with increasing V, presumably because the optimality of linear discrimination relies on the assumption of homoscedasticity, which is violated when the activation images reflect a clear correlated spatial network that does not exist in the baseline images. In addition, for CNR = 0.3 the classification dimensionality estimates are quite skewed with a long tail towards lower dimensionality for large ρ and V indicating an inconsistent response to the dimensionality phase transition.

Estimation of dimensionality with optimization of unsupervised prediction identifies the data as one-dimensional (and, on rare occasions, as two-dimensional) irrespective of the spatial correlation and relative variance of the signal amplitudes. Interestingly, if we ignore the independently varying amplitudes and instead just

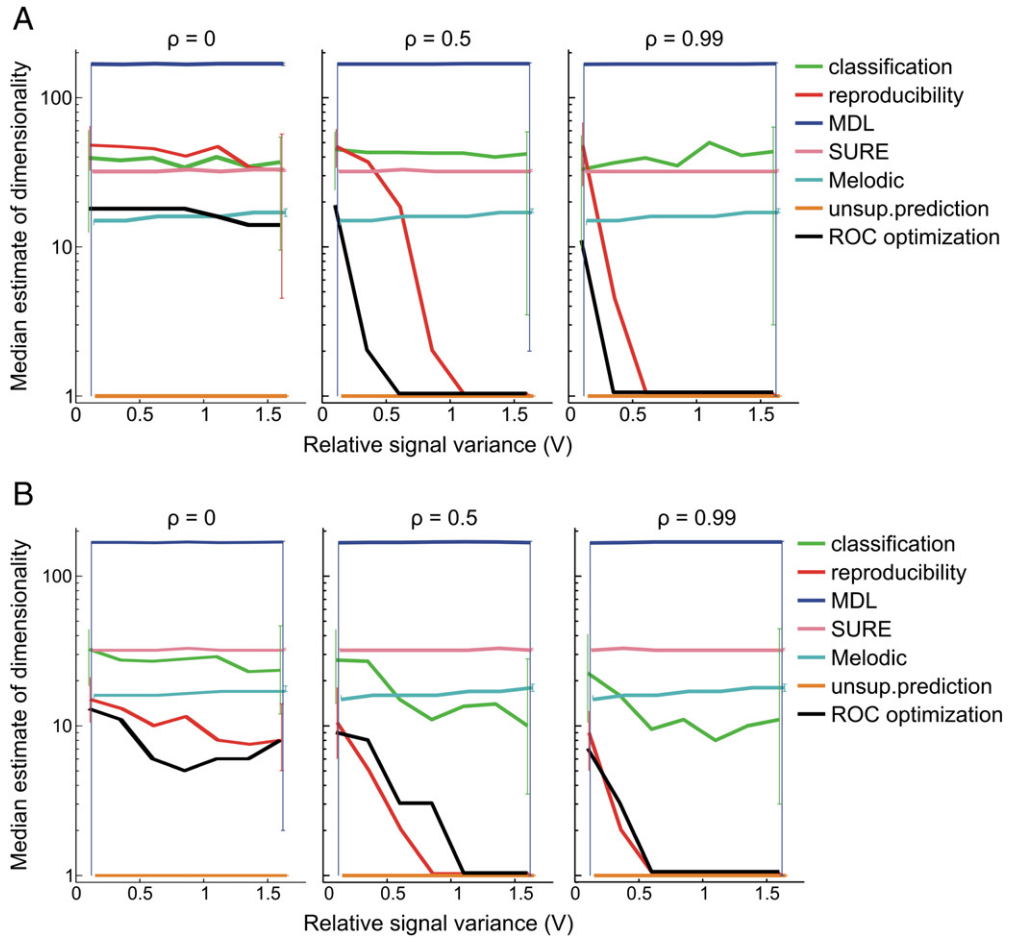

Fig. 3. Median dimensionality estimates in simulations, as calculated by various methods (see legend and text), shown as a function of the relative signal variance, *V*, defined as the variance of the amplitude of the Gaussian activation blobs relative to the variance of the independent background Gaussian noise added to each voxel. CNR is set to 0.3 for Fig. 3A and to 1.0 for Fig. 3B. The three panels from left to right in A and B show three levels of correlation, ρ, between Gaussian activation blob amplitudes. Range bars on the first ($V=0.1$) and last ($V=1.6$) data points reflect the 25%–75% interquartile distribution range across 500 simulation estimates.

consider the common mean Gaussian amplitude, the signal would be considered one-dimensional and this is what this measure seems to reflect. The medians of the analytic methods of dimensionality estimation are largely insensitive to changes in V and ρ. Minka's method (optimization of Bayes evidence) seems to identify the number of active loci correctly and therefore is a valid method when $\rho=0$, but fails to capture the transition to a single dimension when $\rho>0$. The estimates obtained with SURE are consistently somewhat larger (around 30).

Estimation of dimensionality with MDL (as implemented in the GIFT toolbox) produces variable results: in about 60% of the data sets, intrinsic dimensionality was estimated to be about 170, and in the remaining data sets it was estimated as 1 (hence the large error bars for MDL on Figs. 3A and B). This behaviour was consistent across V and ρ and did not depend on the strength of the networked signal.

We should note that when large dimensionality estimates are used in LDA, signal detection suffers as our model becomes unstable (see Appendix A). In particular, there is an upper bound on dimensionality, based on the rule that the number of observations should ideally not exceed the number of degrees of freedom in the LDA model. In our case of 200 observations, this upper bound is approximately equal to 18 dimensions. Consequently LDA built on a PC subspace chosen using MDL in the GIFT toolbox will often, but not always perform poorly. Our own implementation of MDL (data not shown) gave similarly high median dimensionality estimates, but with less variability.

Figs. 4A and B show the performance (as measured by partial ROC area) of LDA based on different methods of dimensionality estimation (reported in Fig. 3), as well as a simple comparison with other methods of fMRI data analysis: GLM, PCA and PDA. As in Fig. 3, CNR was fixed at 0.3 (Fig. 4A) and 1 (Fig. 4B), the three panels represent three levels of ρ, and the horizontal axes show of V varying from 0.1 to 1.6. The ROC curves were calculated separately for the 16 voxels at the center of each Gaussian activation blob, and the plot shows the mean partial area under the curve (with error bars representing the standard deviation across the 16 loci). Partial ROC area would be 0.1 in the best case (perfect detector) and 0.005 in the worst case (random guessing).

Here we see that LDA is often a very effective method of signal detection when the dimensionality of the PC subspace is selected carefully despite the strong heteroscedascity of our simulated data with increasing ρ and V. When Minka's method is used, LDA is sometimes the best performer in all plots, particularly when $\rho=0$, and/or V is small where it is approximately equal to GLM. Optimization of reproducibility is less robust, and the signal detection is reduced compared to using Minka's method particularly when $\rho=0$, or when $\rho>0$ and V and CNR are low (see Fig. 4A). However, when optimization of reproducibility correctly identifies the one-dimensional network in the signal, the performance of LDA rises to near-perfect levels for CNR = 1 and is second only to the signal in the first PCA component for CNR = 0.3. The methods that fail to identify the network structure when $\rho>0$ result in large drops in performance of LDA as V grows, unless they are already performing poorly such as LDA with MDL. In particular, when MDL is used to estimate the dimensionality (either from the GIFT toolbox or our own

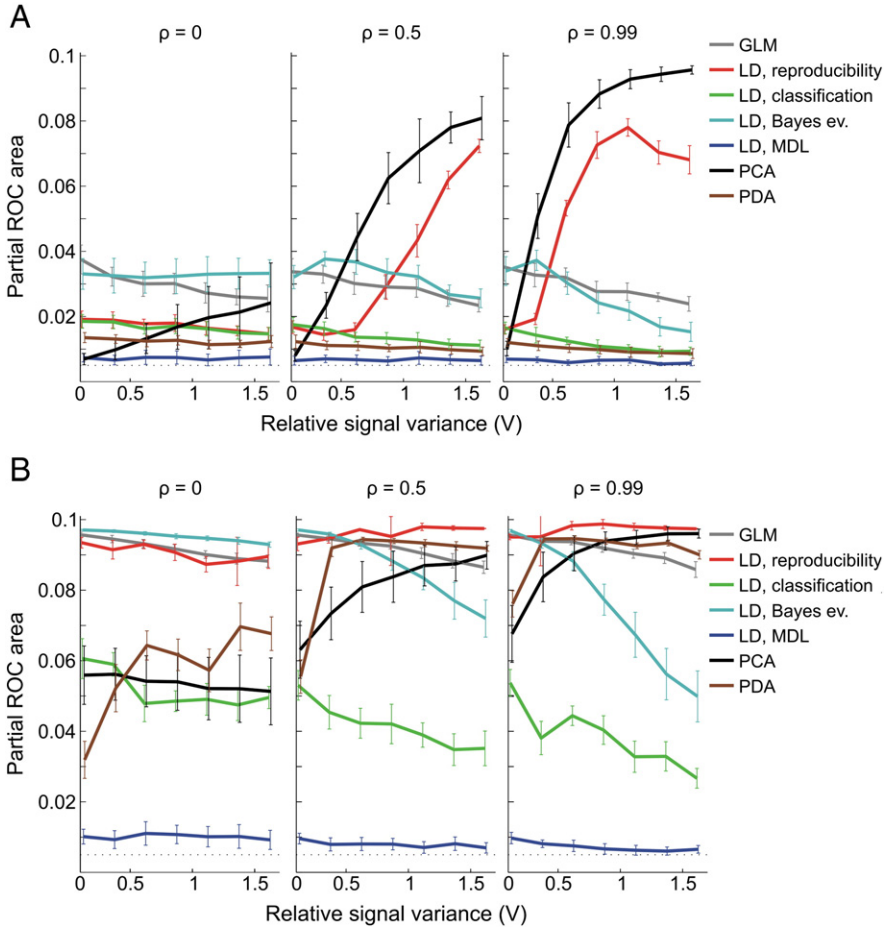

Fig. 4. Partial ROC area (corresponding to false positive frequency [0, 0.1]) as a function of the relative signal variance, V, calculated for different methods of analysis: linear discriminants (LD, on the principal component subspace, with subspace size selected by various methods reported in Fig. 3), univariate general linear model (GLM), penalized discriminant analysis (PDA) with a ridge penalty, and the first component from principal component analysis (PCA). CNR is set to 0.3 for Fig. 4A and to 1 for Fig. 4B. The three panels from left to right in A and B show three levels of correlation, ρ, between Gaussian activation blob amplitudes. Error bars reflect the 16 partial ROC areas across the centre voxels of the sixteen Gaussian activation blobs (see Fig. 1).

implementation), the performance of LDA is only slightly better than chance for all signal structures we tested.

For CNR = 0.3 PDA with a ridge penalty also performs poorly, but is only slightly better than LDA with MDL-defined dimensionality, and a little worse than LDA using dimensionality based on classification accuracy; for most combinations of ρ and V PDA is outperformed by a simple PCA. For CNR = 1.0 the situation is reversed and for most combinations of ρ and V PDA performs better than dimensionality based on classification accuracy or simple PCA. In particular for $\rho = 0.5$ and 0.99 for $V > 0.1$ PDA performs only slightly worse than LDA with dimensionality chosen using reproducibility and much better than LDA with dimensionality chosen using classification accuracy. Overall the soft threshold inherent in the ridge penalty of PDA produces a poorer signal detection algorithm than the hard dimensionality threshold based on reproducibility.

Univariate GLM performs well at high CNR = 1 (Fig. 4B), but rather poorly at CNR = 0.3 when clear network structure emerges for $\rho > 0$ and larger V (Fig. 4A). It also becomes less effective as V grows, because the theoretical assumption of equal variance of the two classes ("activation" and "baseline") for inferential optimality of the GLM t-test is violated by the increasing signal amplitude variance with constant baseline variance. As expected, signal detection of univariate GLM is unaffected by introducing network structure. PCA, on the other hand, works better as V grows provided $\rho > 0$. PCA is especially good when $\rho = 0.99$, becoming the best (CNR = 0.3) or second-best (CNR = 1.0) performer when $V > 1$, a not unexpected result as it exactly matches our simulation conditions. Here, the activation network can be captured sufficiently well by the first principal component when the first eigenvalue clearly stands out in the eigenspectrum of the covariance matrix.

Fig. 5 shows how the first eigenvalue starts to stand out in the eigenspectrum as V increases. This happens when $V \geq 1.1$ for

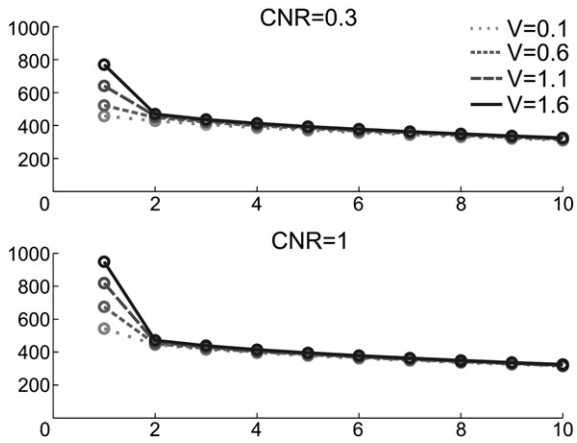

Fig. 5. Plot of the first 10 eigenvalues of the covariance matrix of a single data set, for CNR = 0.3 (top) and CNR = 1 (bottom). ρ is set to 0.5, and V varies from 0.1 to 1.6. Eigenvalues are averaged across 500 simulated data sets.

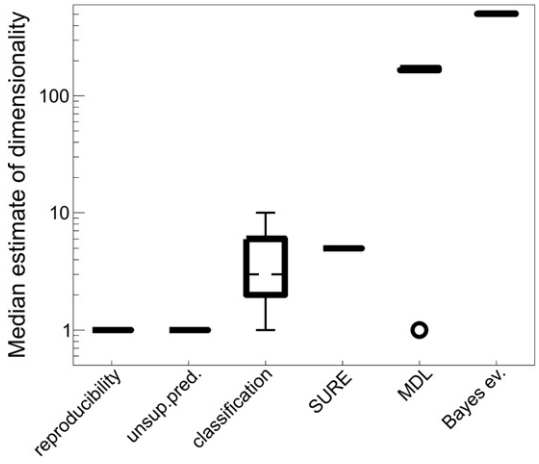

Fig. 6. Box plot of median dimensionality estimates estimated for 20 simulated large data sets ($N=1500$). CNR is 1, ρ is to 0.5, and V is 1.1.

CNR = 0.3, and for $V \geq 0.6$ for CNR = 1, reflecting the ROC dimensionality results for $\rho = 0.99$ in Fig. 3.

Fig. 6 presents the results obtained by running our algorithms on 20 large-sample data sets ($N=1{,}500$) with a moderately strong network structure (CNR = 1.0, $\rho = 0.5$, $V = 1.1$). Given $2N = 3{,}000$ images per set with $J = 2{,}072$ voxels per image, the sample covariance matrix is no longer rank-deficient. As in small-sample ($N = 100$) data sets, reproducibility and unsupervised prediction are optimized with a single PC dimension. Compared with these two methods, classification accuracy is maximized with slightly larger PC dimensionality (median value across all sets is 3 PCs); also, the estimates are much more robust relative to small-sample sets. This suggests that we need rather large samples in order to use classification accuracy as our metric for dimensionality estimation. Using SURE also results in small estimates (5 PCs, very robust). However, algorithms based on MDL and Bayesian evidence produce rather inflated dimensionality estimates (median values are 170 and 523, respectively). MDL estimates single-PC dimensionality in 4 sets out of 20, and a large number ($\approx$170) in the remaining 16 sets; this is consistent with performance on small-sample sets. On the other hand, estimates given by optimization of Bayesian evidence are much larger on large-sample sets compared with small-sample sets. It seems that this method is quite sensitive to sample size, and estimated PC dimensionality grows with the number of observations. This was also reported by Cordes and Nandy (2006). They have also reported similar behaviour for MDL, which does not agree with our results, perhaps because we used the MDL estimates from the GIFT toolbox, which includes subsampling of the data as described in Li et al. (2007).

Fig. 7 shows how the classification accuracy and reproducibility of LDA, computed on a subset of principal components, are influenced by the size of this subset (K). This is illustrated using the prediction–reproducibility (P–R) plots introduced by Strother et al. (2002) and LaConte et al. (2003). We start with computing LDA on the first principal component (corresponding to the largest eigenvalue), then on the first two principal components, then first three, etc., increasing the PC subspace size K. For each value of K, we calculate the reproducibility of spatial maps using split-half resampling, and classification accuracy (i.e. supervised prediction) using leave-one-epoch-out cross-validation. The P–R plot shows the trajectories in the (P–R) space as K increases. Fig. 7 displays these trajectories, averaged across 500 simulated data sets. The three panels correspond to three levels of network correlation ($\rho = 0$, 0.5 and 0.99), and trajectories are shown for four levels of relative signal variance V (0.1, 0.6 and 1.1); CNR was fixed at 1. The left panel shows typical P–R trajectories when the signal is not organized into a covarying spatial network: reproducibility and prediction slowly get better as K increases to about 10 components.

In the middle and right panels, we see that a different pattern emerges as V reaches a certain level: the point corresponding to the first component starts to become separate from the rest of the P–R trajectory. Reproducibility is optimal at this point of $K = 1$. Prediction is maximized for a small value of K, around 2 or 3;. this happens when V is 1.1 at $\rho = 0.5$, and when $V \geq 0.6$ when $\rho = 0.99$. This is the point at which the intrinsic dimensionality of the data goes through a transition from many components to a single component, and when the first eigenvalue separates from the remaining spectrum of the covariance matrix (see Fig. 5). This dimensionality transition can be illustrated using the global signal-to-noise (gSNR) metric. Similar results appear common with LDA in neuroimaging, where

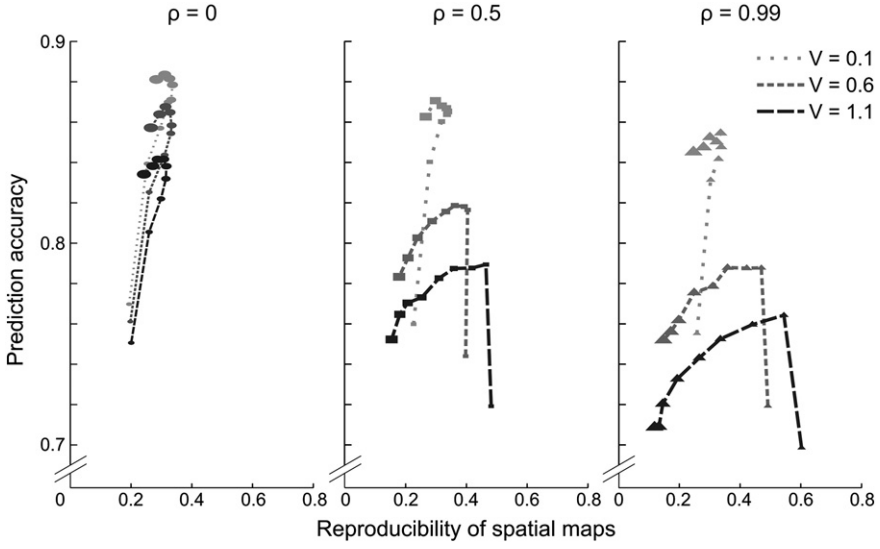

Fig. 7. Scatter plots of prediction accuracy vs. reproducibility of spatial maps, as calculated for a linear discriminant on the principal components subspace of the simulated data. Size of the subspace was varied from $K = 1$ (smallest symbol) to $K = 40$ (largest symbol) components; here we show results for K = 1, 2, 4, 7, 10, 15, 22, 30, 40. Different trajectories correspond to different levels of relative signal variance, V, and the three panels correspond to three levels of the coupling between activation blob amplitudes, ρ. CNR was fixed at 1.

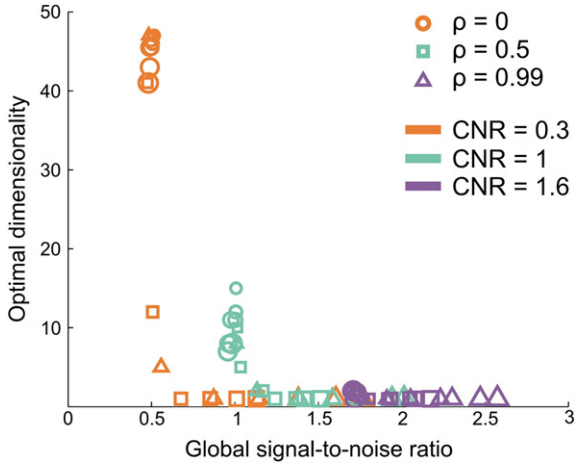

Fig. 8. Asymptotic relationship between global signal-to-noise ratio (gSNR) and optimal dimensionality. Marker size indicates relative signal variance, V, from 0.1 (small) to 1.6 (large). Three colours encode three different levels of CNR, and spatial correlation is encoded by markers.

optimization of prediction requires a larger number of components than optimization of reproducibility (LaConte et al., 2003; also, see below).

Fig. 8 shows the dependence between gSNR and dimensionality that optimizes reproducibility. Unlike previous figures, here we vary the contrast-to-noise ratio of the task-related signal as well as V and ρ; for the sake of simplicity, we use the same marker with increasing size for increasing levels of V. For lower levels of CNR (0.3 and 1) there is an asymptotic relationship between gSNR and optimal dimensionality: when gSNR is large enough, using a single principal component becomes an optimal choice, but, as gSNR decreases, optimal dimensionality becomes larger and larger. The level of gSNR corresponding to the transition in intrinsic dimensionality depends on the mean strength of activation: it is around 0.5 for CNR = 0.3, around 1 for CNR = 1, and for CNR = 1.6 the critical level seems to be around 1.7, although optimal dimensionality never rises above 2.

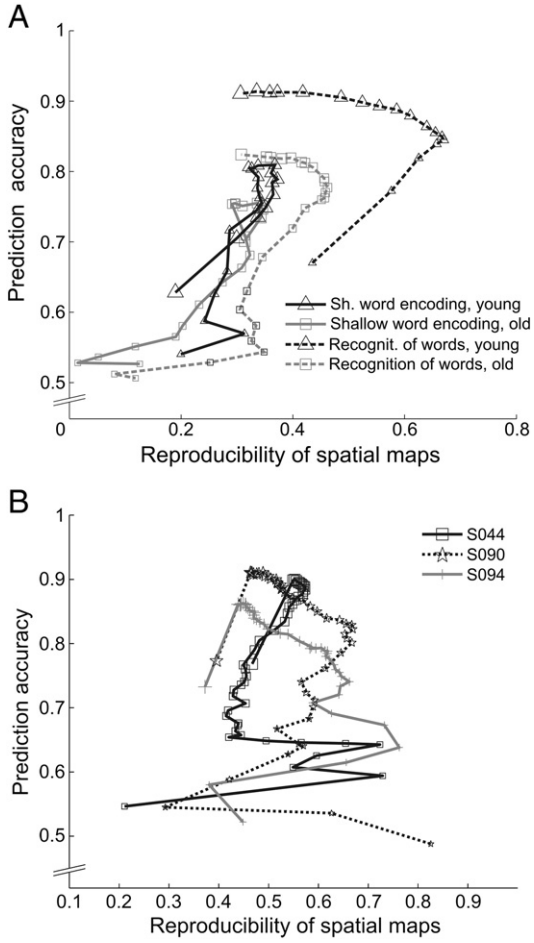

Fig. 9. Prediction–reproducibility plots, calculated on real data: the aging study (A; 2 age groups and 2 behavioral tasks), and the stroke recovery study (B; 3 stroke patients). Marker size indicates K, size of the PC subspace, with smaller markers corresponding to lower K. The range of K was from 4 to 140 (A) and from 1 to 300 (B).

Real data

Figs. 9A and B show examples of P–R plots for real data: the aging study and the stroke recovery study, respectively. Fig. 9A shows results from "young" and "old" age groups performing two tasks of the six tasks: shallow word encoding and word recognition. The second task had twice as many volumes as the first. Data were processed with linear discriminant analysis using the NPAIRS package (for software see: http://code.google.com/p/plsnpairs/). Subspace size was varied from 4 to approximately 140 principal components; reproducibility and prediction were measured for each subspace size using split-half resampling with split-half subsamples of 5 subjects each, repeated 50 times. Fig. 9A shows that for young subjects fewer PCs are needed to achieve maximum reproducibility of activation maps than are needed for older subjects; this holds true for both tasks. Optimal prediction and reproducibility are higher in the recognition task than in the encoding task; possibly because the number of volumes was twice as high in the recognition task as in the encoding task.

Fig. 9B displays the P–R plots for three (out of 9) subjects in the stroke recovery study. The two classes of images in LDA corresponded to epochs of finger tapping and wrist flexion (rather than to activation and rest blocks, as in the aging study). LDA was quite accurate in separating these two classes (classification accuracy was at least 90% for all subjects; see Schmah et al. for a detailed description of results). Unlike the aging study (when the subjects were pooled across age groups) in this study nine subjects were analyzed separately. The resulting within-subject P–R trajectories show much more fluctuation, especially when the PC subspace size is small. Sometimes we observe that points on the P–R trajectory obtained with small K have higher reproducibility but relatively low prediction compared with larger intermediate values of K: subject S090 is a good example (a similar example is recognition of words in the old subject group, Fig. 9A). These points in the P–R trajectory typically correspond to highly reproducible artifacts in the fMRI signal, which have no clear relationship to the task (Strother et al., 2004). Even if reproducibility seems to be maximized here, these points should be ignored. For further analysis, we have included only points with at least 60% prediction accuracy in our estimation of dimensionality.

In the simulated data, we described the demonstrated asymptotic relationship between optimal reproducibility and a measure of global signal-to-noise ratio. This relationship exists in real data as well, as shown in Figs. 10A and B. Fig. 10A plots the optimal dimensionality (in the sense of reproducibility given prediction accuracy of at least 60%) versus gSNR, measured for a large set of subjects belonging to three age groups (young, middle-aged and old), performing a range of six

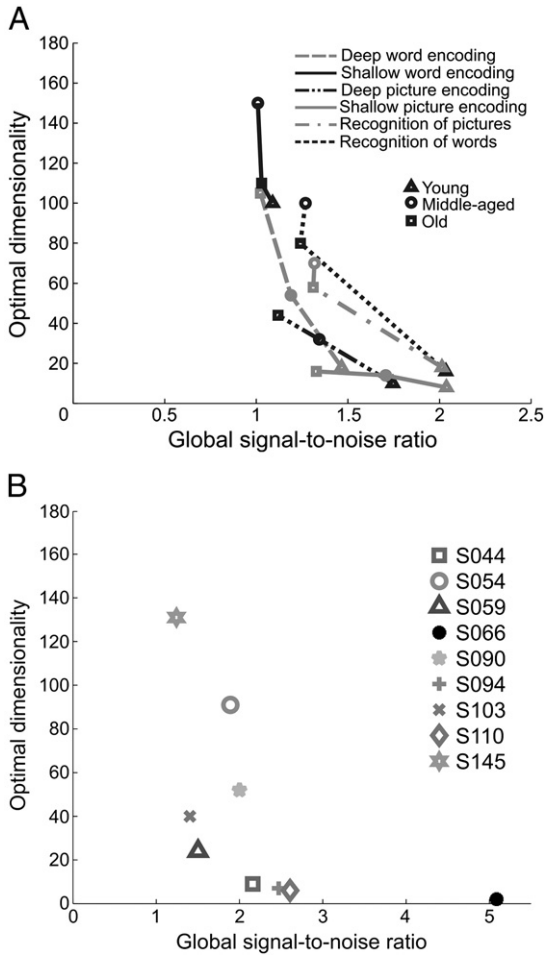

Fig. 10. Optimal dimensionality and gSNR in real data: the aging study (A) and the stroke recovery study (B).

memory tasks. The figure shows that optimal dimensionality is small for data sets with high gSNR, and dimensionality increases greatly as gSNR decreases to 1. Young subjects' data sets have larger gSNR and smaller dimensionality than middle-aged and old subject groups. This might reflect the difference in cortical networks recruited during memory tasks in young and older subjects (Grady et al., 2006; Grady, 2008).

Fig. 10B shows that the asymptotic relationship between estimated dimensionality (selected to optimize reproducibility) and gSNR also holds for within-subject analysis, as performed in a stroke recovery study. Here we also see that low dimensionality is optimal for analyzing the data sets with high gSNR (and, correspondingly, high reproducibility); this dimensionality increases greatly for the subjects whose data has low gSNR. The subjects could be separated into two groups with respect to classification accuracy, representing two possible asymptotic curves. The mean classification accuracy of LDA on subjects S059, S103 and S145 was lower (90%) than on subjects S054, S090 and S044 (96%); perhaps the lower accuracy and gSNR values correspond to a lower CNR in the underlying BOLD signal, which explains why the asymptotic curve is shifted to the left for subjects that are harder to classify. Interestingly, gSNR asymptotes reflecting the underlying average signal structure in these real data sets are all larger than 1.0 and fall between 1 and 2. Both of these studies were performed on 1.5T scanners so that we can expect even larger gSNR asymptotes (and, therefore, smaller dimensionality estimates) from subjects performing similar tasks in scanners with magnetic field strengths $\geq$3T.

Discussion

Choice of dimensionality for optimal signal detection involves a trade-off between the number of independently varying spatial activation loci, and the number of brain networks that are sufficiently strongly coupled and have a sufficient signal variance to be detected in the data covariance matrix as single components (Fig. 3). Selection of dimensionality based on optimal reproducibility of spatial maps is sensitive to the emergence of a functionally connected spatial network from noise, and correctly captures the point at which the network signal can be characterized with a single principal component. Other methods of estimation that we tested failed to reflect this change in the network structure relative to noise. For example, the tendency for Bayes evidence optimization, as implemented in MELODIC software (Beckmann and Smith, 2004), to report large numbers of dimensions probably reflects lack of sensitivity to actual underlying network structures, and instead may reflect an approximate count of the number of spatially distinct non-zero signals in the data set. This is expected to rise with increasing amounts of data as smaller and smaller mean changes and local variance differences can be detected leading to increasing numbers of localized mean activations becoming apparent in the data set. This would be consistent with the tendency for MELODIC dimensionality estimates to increase with increasing sample size seen in our results, and by Cordes and Nandy (2006). A caveat to our dimensionality selection results is the case where CNR=0, which we have not tested, as LDA is then degenerate and not an appropriate analysis technique.

This task of identifying signal-carrying principal components has been analyzed in statistical physics, where the phenomenon of "retarded classification" was described (Watkin and Nadal, 1994). Let α be the ratio of the number of examples to the number of variables. If there is only one signal-carrying principal component, its successful identification is only possible when α reaches a certain critical level. Using simulated two-class Gaussian data, Biehl and Mietzner (1994) describe how this critical level is influenced by various parameters, such as the amount of overlap in the two classes and the separation of their means. When the signal is embedded in more than one component, there are several such critical levels; when α passes through each critical level, there is a phase transition and one more signal-carrying component becomes possible to identify (Hoyle and Rattray, 2007).

In our simulations, we observe a similar effect of a transition from many to one principal component. This transition occurs when the task-related network becomes visible amidst the noise, or, alternatively, when the first eigenvalue (which corresponds to the component with task-related signal) starts to "stand out" in the eigenspectrum of the covariance matrix. There are many factors that influence the occurrence of the switch; we have shown how it is influenced by the contrast-to-noise ratio and relative variance of the simulated signal, and by the spatial correlation of the functionally connected activation loci. In real data, we have also observed that data sets with high reproducibility (e.g., young subjects) require considerably fewer components than data sets with low reproducibility (such as elderly subjects) to optimize the performance of LDA. Based on our results in data sets from scanners at 1.5T we believe the PC dimensionality for optimal signal detection is likely in general to decrease further for 3.0T data sets, because the signal-to-noise ratio of BOLD signal increases with field strength. This will make techniques, such as reproducibility, for accurately estimating the low dimensionality of the fMRI signal subspace even more important.

We have tested many methods of dimensionality estimation, empirical as well as analytic. Of these methods, selecting the PC subspace that maximizes reproducibility of LDA spatial maps appears to often work the best in the signal-detection sense: it provides estimates that are consistent with optimization of the area under the early portion of the ROC curve when LDA is used for analysis. This method captures the switch from high to low dimensionality in the narrow signal range of this phase transition. In recent literature, in addition to the NPAIRS framework, selecting the components of maximal reproducibility was proposed as a promising method of dimensionality estimation for independent component analysis (Yang et al., 2008; Wang and Peterson, 2008). On the other hand, analytical methods (such as minimum description length and maximizing Bayesian evidence) tend to overestimate the number of components (this is especially so in the case of MDL), both of the underlying intrinsic dimensionality and in terms of the optimal subspace for LDA. This is consistent with the finding of Cordes and Nandy (2006) who also show that these methods do not scale well as they select higher dimensionality when the sample size grows. This problem may be reduced by running the analytical methods on a small sub-sample of the data that is constructed so the observations are independent and identically distributed (i.i.d.; Li et al., 2007). However, reducing the total sample size is likely to reduce the accuracy of the resulting estimates making this technique applicable only when there is a strong, well-defined signal with many more than the minimal number of observations required to detect that signal. For the technique described by Li and implemented in GIFT we did not reach this stable point using 3000 simulated scans with a moderately strong connected network structure.

A number of recent papers in the literature have compared LDA with various other classifiers (i.e., support vector machines: Cox and Savoy, 2003; Mourão-Miranda et al., 2005; LaConte et al., 2005; artificial neural networks: Mørch et al., 1997; Hanson et al., 2004). As outlined by LaConte et al. (2005), there are many variables that must be carefully considered and optimized in such comparisons, including the PCA subspace if ill-posed data sets are to be used in LDA. In many of the published comparisons to date it is either unclear how critical optimization variables were treated, or clearly non-optimal choices were made for the LDA models used, particularly with respect to PC dimensionality. For example, Mourão-Miranda et al. (2005) retain all PCs on which to build their discriminant, which, based on our results reported above, is likely to be a non-optimal choice for signal detection of the resulting LDA spatial patterns. A complication for such optimal subspace selection is the result reported here, and in our earlier work, that it may not be possible to optimize prediction and reproducibility simultaneously for LDA applied to real fMRI data sets (LaConte et al., 2003; Strother et al., 2004) or in more general nonlinear models (Jacobsen et al., 2008). A further complication is the issue of "crossed learning curves" that show that a simple linear model may outperform a more highly parameterized, potentially nonlinear model unless large enough sample sizes are available (Mørch et al., 1997). Many of these issues have yet to be carefully compared in relative performance comparisons of multivariate classifiers in the literature and as a result we believe that the eventual utility of LDA, and related multivariate Gaussian models in neuroimaging remains unresolved.

Overall, following Lange et al. (1999) we propose a pluralistic analysis strategy allowing for both weak and strong activation signals with varying levels of network connectivity. If PCA dimensionality selected by optimizing split-half reproducibility of spatial maps is low, there is evidence of a strong spatial network in the BOLD signal and a linear discriminant can provide near optimal signal detection in an ill-posed data set. If the evidence for a network seems to be weak (dimensionality for optimal split-half reproducibility is high), a univariate method such as GLM might be used although our results represent upper bounds on GLM performance with exactly Gaussian noise and perfect HRF models. Practically Minka's Bayes Evidence dimensionality estimate with LDA is likely to be at least an equivalent signal detector in this case, although the robustness of SURE estimates with increasing data set may makes it a more attractive choice in this situation. Selecting between these two approaches requires further study in low SNR, weak-network imaging situations. Analysis of P–R plots, calculated within the NPAIRS framework, can be used to obtain estimates of the strength and extracted global signal-to-noise of the underlying cortical networks in order to guide such analysis choices.

Acknowledgments

The authors gratefully acknowledge helpful conversations with Lars Kai Hansen who brought the work in statistical physics to their attention. This work was partially supported by CIHR grants MOP84483 and MOP14036, and a Bridging Brain, Mind and Behaviour grant from the James S. McDonnell Foundation. GY is partly supported by Ydessa Hendeles scholarship from Baycrest Centre and University of Toronto, and SCS gratefully acknowledges support of the Heart & Stroke Foundation of Ontario through the Centre for Stroke Recovery. Participation of MNW and AL was supported by NIH/NIMH grant 073204.

Appendix A. Stability of Linear Discriminant in High-Dimensional Space

A.1. Degrees of freedom

When the linear discriminant is calculated as a vector $\mathbf{W}^{-1}\Delta\mathbf{m}$ in K-dimensional subspace, it has $K+K(K+1)/2$ degrees of freedom (we need to calculate the K-dimensional vector $\Delta\mathbf{m}$ and symmetric K-by-K matrix $\mathbf{W}$). Accurate estimation of the linear discriminant requires the number of images, $2N$, to be greater or equal to the number of degrees of freedom. This leads to an upper bound on the dimensionality K:

$$K \leq \lfloor \sqrt{4N + 2.25} - 1.5 \rfloor \tag{13}$$

In our case of 200 images (100 images per class), accurate estimation of the linear discriminant and a correspondingly small condition number (see below) requires that the size of the subspace does not exceed $K = 18$.

A.2. Condition number

Condition number of a linear equation $\mathbf{y} = \mathbf{A}\mathbf{x}$ measures the sensitivity of $\mathbf{y}$ to perturbations of $\mathbf{x}$ (Golub and van Loan, 1996). For linear discriminant, it reflects how much influence the noise in $\Delta\mathbf{m}$ has on the calculation of $\mathbf{W}^{-1}\Delta\mathbf{m}$. Using the L2 norm, condition number is defined as the ratio of the largest and smallest eigenvalues of $\mathbf{W}$. When the K used to calculate LDA rises above 100 components, condition number becomes extremely high (especially when the network is present). This leads to large variability in the calculation of LDA in our simulation results (Table 1). Eigendecompositions were performed using Matlab 7.4 and routine svd.

Table 1
Condition number of the estimated within-class covariance matrix W calculated on the principal component subspace of size K. Three different levels of spatial correlation of the signal (ρ) are listed. Dynamic range was fixed at 1.6.

K	$\rho = 0$	$\rho = 0.5$	$\rho = 0.99$
1	1	1	1
10	1.78	2.54	3.78
25	3.12	4.44	6.36
50	8.03	13.33	19.24
100	72.24	104.74	180.18
150	820.20	1228.30	1850.20

References

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional medical imaging. IEEE Trans. Med. Imaging 23 (2), 137–152.

Biehl, M., Mietzner, A., 1994. Statistical mechanics of unsupervised structure recognition. J. Phys. A 27, 1885–1897.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp. 14, 140–151.

Cordes, D., Nandy, R., 2006. Estimation of intrinsic dimensionality of fMRI data. Neuroimage 29, 145–154.

Cox, R.W., 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19, 261–270.

Eastment, H.T., Krzanowski, W.J., 1982. Cross-validatory choice of the number of components from a principal component analysis. Technometrics 24 (1), 73–77.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J., Turner, R., 1995a. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2, 166–172.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1995b. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 9, 416–429.

Golub, G.H., van Loan, C.F., 1996. Matrix Computations, 3rd ed. The Johns Hopkins University Press.

Grady, C.L., 2008. Cognitive Neuroscience of Aging. In: Kingstone, A., Miller, M.B. (Eds.), The Year in Cognitive Neuroscience. Blackwell Publishing, Oxford, UK, pp. 127–144.

Grady, C.L., Springer, M.V., Hongwanishkul, D., McIntosh, A.R., Winocur, G., 2006. Age-related changes in brain activity across the adult lifespan. J. Cogn. Neurosci. 18 (2), 227–241.

Hansen, L.K., Larsen, J., Nielsen, F.A., Strother, S.C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J.J., Svarer, C., Paulson, O.B., 1999. Generalizable patterns in neuroimaging: How many principal components? Neuroimage 9, 534–544.

Hanson, S.J., Matsuka, T., Haxby, J.V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? Neuroimage 23 (1), 156–166.

Hoyle, D.C., Rattray, M., 2007. Statistical mechanics of learning multiple orthogonal signals: asymptotic theory and fluctuation effects. Phys. Rev. E 75 (1 Pt 2), 016101.

Jacobsen, D.J., Hansen, L.K., Madsen, K.H., 2008. Bayesian model comparison in nonlinear BOLD fMRI hemodynamics. Neural Comput. 20 (3), 738–755.

Krzanowski, W.J., Kline, P., 1995. Cross-Validation for Choosing the Number of Principal Components in Principal Component Analysis. Multivariate Behav. Res. 30 (2), 149–165.

Kustra, R., Strother, S.C., 2001. Penalized discriminant analysis of 15O-water PET brain images with prediction error selection of smoothing and regularization hyper-parameters. IEEE Trans. Med. Imaging 20, 376–387.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., Strother, S.C., 2003. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. Neuroimage 18, 10–27.

LaConte, S., Strother, S.C., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. Neuroimage 26, 317–329.

Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.A., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis. Neuroimage 10, 282–303.

Li, Y.O., Adali, T., Calhoun, V.D., 2007. Estimating the number of independent components for functional magnetic resonance imaging data. Hum. Brain Mapp. 28, 1251–1266.

Lukic, A.S., Wernick, M.N., Strother, S.C., 2002. An evaluation of methods for detecting brain activations from functional neuroimages. Artif. Intell. Med. 25, 69–88.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, San Diego.

McIntosh, A.R., Lobaugh, N.J., 2004. Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23, S250–S263.

Metz, C.E., Herman, B.A., Shen, J.H., 1998. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat. Med. 17 (9), 1033–1053.

Minka, T.P., 2000. Automatic choice of dimensionality for PCA. Technical Report 514, MIT Media Laboratory, Perceptual Computing Section.

Mørch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B., 1997. Nonlinear versus linear models in functional neuroimaging: learning curves and generalization crossover. Lect. Notes Comput. Sci. 1230, 259–270.

Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage 28, 980–995.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput. Stat. Data Anal. 49, 974–997.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532–542.

Schmah, T., Yourganov, G., Zemel, R.S., Hinton, G.E., Small, S.L., Strother, S.C., 2010. Comparing classification methods for longitudinal fMRI studies. Journal of Neural Computation 22 (11), doi:10.1162/NECO_a_00024.

Small, S.L., Hlustik, P., Noll, D.C., Genovese, C., Solodkin, A., 2002. Cerebellar hemispheric activation ipsilateral to the paretic hand correlates with functional recovery after stroke. Brain 125 (7), 1544–1557.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (S1), 208–219.

Strother, S.C., Lange, N., Anderson, J.R., Schaper, K.A., Rehm, K., Hansen, L.K., Rottenberg, D.A., 1997. Activation pattern reproducibility: Measuring the effects of group size and data analysis models. Hum. Brain Mapp. 5, 312–316.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutinger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage 15, 747–771.

Strother, S.C., LaConte, S., Hansen, L.K., Anderson, J., Zhang, J., Pulapura, S., Rottenberg, D., 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. Neuroimage 23, S196–S207.

Strother, S.C., Oder, A., Spring, R., Grady, C., 2010. The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging. In: Lechevallier, Y., Saporta, G. (Eds.), 19th International Conference on Computational Statistics: Keynote, Invited and Contributed Papers. Physica-Verlag, Springer, Paris, France, pp. 111–120.

Tegeler, C., Strother, S.C., Anderson, J.R., Kim, S.G., 1999. Reproducibility of BOLD-based functional MRI obtained at 4T. Hum. Brain Mapp. 7, 267–283.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. R. Stat. Soc. B 61 (3), 611–622.

Ulfarsson, M.O., Solo, V., 2008. Dimension estimation in noisy PCA with SURE and random matrix theory. IEEE Trans. Signal Process. 56 (12), 5804–5816.

Wang, Z., Peterson, B.S., 2008. Partner-matching for the automated identification of reproducible ICA components from fMRI datasets: Algorithm and validation. Neuroimage 29, 875–893.

Watkin, T.L.H., Nadal, J.P., 1994. Optimal unsupervised learning. J. Phys. A 27, 1899–1915.

Wax, M., Kailath, T., 1985. Detection of signals by information theoretic criteria. IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (2), 387–392.

Wold, S., 1978. Cross-validatory estimation of the number of principal components in factor and principal components analysis. Technometrics 20 (4), 397–405.

Worsley, K.J., 2001. Statistical analysis of activation images. In: Jezzard, P., Matthews, P.M., Smith, S.M. (Eds.), Functional MRI: An Introduction to Methods. Oxford University Press, NY, pp. 251–270.

Yang, Z., LaConte, S., Weng, X., Hu, X., 2008. Ranking and averaging independent component analysis by reproducibility (RAICAR). Hum. Brain Mapp. 29, 711–725.

Zhang, J., Liang, L., Anderson, J.R., Gatewood, L., Rottenberg, D.A., Strother, S.C., 2008. A Java-based fMRI processing pipeline evaluation system for assessment of univariate general linear model and multivariate canonical variate analysis-based pipelines. Neuroinformatics 6 (2), 123–134.