Available online at www.sciencedirect.com
ScienceDirect
Magnetic Resonance Imaging 27 (2009) 264–278
MAGNETIC RESONANCE IMAGING

Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA

Jing Zhang[a,e,*], Jon R. Anderson[c], Lichen Liang[b], Sujit K. Pulapura[b], Lael Gatewood[a], David A. Rottenberg[c], Stephen C. Strother[a,c,d]

[a]Health Informatics Graduate Program, University of Minnesota, Minneapolis, MN 55455, USA
[b]Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[c]Department of Neurology, University of Minnesota, Minneapolis, MN 55455, USA
[d]Rotman Institute, Baycrest, and Medical Biophysics, University of Toronto, ON, Canada M6A2E1
[e]Neuroscience PET Laboratory, Mt. Sinai School of Medicine, New York, NY 10029, USA

Received 19 November 2007; revised 19 May 2008; accepted 30 May 2008

Abstract

In functional magnetic resonance imaging (fMRI) analysis, although the univariate general linear model (GLM) is currently the dominant approach to brain activation detection, there is growing interest in multivariate approaches such as principal component analysis, canonical variate analysis (CVA), independent component analysis and cluster analysis, which have the potential to reveal neural networks and functional connectivity in the brain. To understand the effect of processing options on performance of multivariate model-based fMRI processing pipelines with real fMRI data, we investigated the impact of commonly used fMRI preprocessing steps and optimized the associated multivariate CVA-based, single-subject processing pipelines with the NPAIRS (nonparametric prediction, activation, influence and reproducibility resampling) performance metrics [prediction accuracy and statistical parametric image (SPI) reproducibility] on the Fiswidgets platform. We also compared the single-subject SPIs of univariate GLM with multivariate CVA-based processing pipelines from SPM, FSL.FEAT, NPAIRS.GLM and NPAIRS.CVA software packages (or modules) using a novel second-level CVA. We found that for the block-design data, (a) slice timing correction and global intensity normalization have little consistent impact on the fMRI processing pipeline, but spatial smoothing, temporal detrending or high-pass filtering, and motion correction significantly improved pipeline performance across all subjects; (b) the combined optimization of spatial smoothing, temporal detrending and CVA model parameters on average improved between-subject reproducibility; and (c) the most important pipeline choices include univariate or multivariate statistical models and spatial smoothing. This study suggests that considering options other than simply using GLM with a fixed spatial filter may be of critical importance in determining activation patterns in BOLD fMRI studies.

Keywords: fMRI; GLM; CVA; fMRI processing pipeline; Pipeline evaluation

1. Introduction

Over the past one and a half decades, functional magnetic resonance imaging (fMRI) has emerged as a powerful neuroimaging tool for the study of brain functions and brain diseases. fMRI, as a noninvasive procedure, has become a critical step in preoperative surgical planning [1–3], and fMRI analysis for individual patient has been used for presurgical mapping. The rapid growth of fMRI applications demands the development of techniques to measure the accuracy and reliability of fMRI analysis software packages such as SPM (Statistical Parametric Mapping), AFNI (Analysis of Functional NeuroImages) and FSL (FMRIB Software Library).

fMRI software diagnostic tools such as Statistical Parametric Mapping Diagnosis (SPMd, http://www.sph.umich.edu/~nichols/SPMd/) have been developed to establish the validity of fMRI models and inferences through diagnosis and exploratory data analysis [4]. However, such tools are specific to certain fMRI software package and not

* Corresponding author. Neuroscience PET Laboratory, Mt. Sinai School of Medicine, New York, NY 10029, USA.
E-mail address: jzhang0000@gmail.com (J. Zhang).

0730-725X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.mri.2008.05.021

easily generalized to others, which makes performance across different packages hard to compare. As a matter of fact, given the same fMRI data, different fMRI analysis software packages with nominally equivalent parameter settings may generate different activation maps or statistical parametric images (SPIs). This variation stems from different approaches in the software implementation for fMRI preprocessing steps and statistical models. In addition, the incompatibility of various fMRI software tools has made it more difficult to compare and evaluate results [5] among numerous fMRI studies [6,7].

In this study, an fMRI processing pipeline (an analysis chain [8] or "functional neuroimaging data chain" [9]) refers to a series of fMRI processing steps after data acquisition and image reconstruction, which include data format conversion, preprocessing, statistical analysis and result visualization. Several fMRI preprocessing steps such as motion correction and spatial smoothing have been found effective in reducing noise and improving signal-to-noise ratios [10–16]. However, the impact of other preprocessing steps is unclear, for example, slice timing correction, global intensity normalization and low-pass filtering [17–22]. In addition, in practice, an SPI is usually obtained with a certain set of preprocessing steps and parameters in fMRI analysis. When preprocessing steps and parameters are changed, the SPI will often change accordingly. Therefore, it is of interest to seek the best combination of such preprocessing steps and parameters in order to better reveal the true activation patterns.

In fMRI statistical analysis, a number of statistical methods or models have been proposed for fMRI activation detection. These models can be classified into two categories: univariate (or model-driven) methods such as the general linear model (GLM) [23–26], which characterizes region-specific responses at each voxel based on assumptions, and multivariate (or data-driven) methods such as principal component analysis (PCA) [27–31], canonical variate analysis (CVA) [32–34], independent component analysis (ICA) [35–37] and cluster analysis [38–41], which are often exploratory and data driven and have the potential to identify activation patterns that may reveal neural networks and functional connectivity [42,43]. To date, univariate GLM is still the dominant approach in fMRI analysis, but there is growing interest in multivariate approaches [43–47]. This may be due to the limitation of univariate analysis (GLM) that it does not take into account network-dependent spatial correlations among voxels, while the advantage of multivariate analysis is that it automatically models the spatial correlations in an fMRI volume, which helps identify large-scale neural networks in the brain. Some multivariate approaches such as CVA, ICA and cluster analysis have been made available in software packages (or modules) such as SPM.MM (Multivariate Methods for fMRI, http://www.madic.org/download/MMTBx/), FSL.MELODIC (Multivariate Exploratory Linear Optimized Decomposition into Independent Components, http://www.fmrib.ox.ac.uk/fsl/melodic/index.html) and EROICA (Exploring regions of interest with cluster analysis, EvIdent).

For univariate GLM-based processing pipelines, much research has been conducted on the impact of different preprocessing or statistical analysis options on fMRI activation detection [6,12,22,48]. In particular, Thirion et al. [49,50] reported that group size, choice of voxel level or cluster level analysis, choice of parametric or nonparametric analysis method and choice of random- or mixed-effects statistics all have significant effects on the reliability of activation maps in fMRI group analysis with GLM. However, there have been few such studies for multivariate fMRI processing pipeline; for example, what pipeline options (including options and parameters for data preprocessing and statistical analysis) are appropriate and can best reveal the true activation patterns? Further, are these pipeline options and parameters similar to, or different from, those of univariate GLM-based fMRI processing pipelines? To address these questions, it is necessary to evaluate and optimize multivariate fMRI processing pipelines.

Receiver operating characteristic (ROC) analysis is the most widely used approach to method evaluation. Since there is no easily measurable ground truth in real fMRI data, ROC analysis requires simulation. Considerable work has been done to evaluate fMRI preprocessing steps and statistical methods with the ROC approach on simulated data [6,21,44,48,51,52]. However, the effectiveness of standard ROC analysis depends on how well the simulated data approximates the real data. To overcome simulation-dependent biases, several modified ROC methods were developed to work with real fMRI data [53–56]. Nevertheless, the fundamental problem of unknown ground truth in real fMRI data ultimately hinders the accuracy of these modified ROC methods.

The NPAIRS (nonparametric prediction, activation, influence and reproducibility resampling) framework has been proposed as an alternative to ROC methods to evaluate and optimize fMRI processing pipelines [9]. The NPAIRS approach is rooted in predictive learning (or machine learning) and works on real fMRI data. The validity of an fMRI processing pipeline can be estimated by measuring its predictive and reproducible performance metrics using "split-half resampling," a combination of twofold cross-validation and delete-d jackknife resampling. In each split, the data are divided into training and test sets (e.g., according to the number of subjects or runs) and model parameters estimated in the training set are used to predict design matrix parameters (e.g., scan state or class) in the test set. Prediction accuracy (p) is measured as the average posterior probability of each fMRI volume's true class membership (i.e., predicted baseline or activation brain state) in the test set based on the training set parameters and Bayes formula [9,57]. For each independent pair of split-half data sets, the resulting SPIs' reproducibility is defined as the correlation (r) between all pairs of spatially aligned voxels in the brain. In general, the average, or median, of the distribution of such correlation

values is obtained from the independent SPIs of many split-half resamplings of the fMRI data. In the NPAIRS approach, reproducible SPIs can be obtained from any data analysis approach on a *Z*-score scale [9]. The NPAIRS approach has been implemented in the NPAIRS software package (Contact S. Strother for software, sstrother@rotman-baycrest.on.ca) which also provides statistical models such as GLM and PCA/CVA for statistical analysis. The utility of the NPAIRS framework has been demonstrated by a number of single-subject and group analyses in functional neuroimaging studies [8,9,58–61].

In this study, to explore pipeline options in fMRI multivariate processing pipelines, we used the NPAIRS approach to evaluate the impact of a series of preprocessing steps (slice timing correction, motion correction, spatial smoothing, temporal filtering and global intensity normalization) and to optimize the associated multivariate CVA-based single-subject processing pipelines. We also compared the univariate GLM (in FSL, SPM2 and NPAIRS) with the multivariate CVA (in NPAIRS) using a novel second-level CVA in order to further identify the most important pipeline options and to reveal how they impact fMRI activation patterns.

2. Methods

2.1. fMRI and MRI data

This study used a BOLD fMRI data set, in which a block-design parametric static force experiment was performed on 16 normal subjects who were scanned on a 1.5-T Siemens scanner. Two fMRI runs per scan session were acquired with an EPI BOLD sequence (TR=3986 ms, TE=60 ms, FA=90°, matrix=64×64, FOV=220×220 mm, number of slices=30, number of time points=135). In each run, there were six baseline periods, which alternated with five activation periods during which a static force was applied to a force transducer held by the subject between the right thumb and forefinger with randomly assigned force levels (200, 400, 600, 800 and 1000 g) monitored via a visual feedback loop. Each baseline and activation epoch lasted for 45 s, and the fundamental frequency of the block design was 0.011 Hz. A detailed description of this data set can be found in Ref. [8].

2.2. Data analysis environment

In order to overcome the incompatibility and lack of interoperability among fMRI analytical software packages, we used Fiswidgets, an interoperable pipeline environment developed by Fissell et al. [5], as the platform to run various processing pipelines. Fiswidgets integrates software packages such as AIR [65,66], AFNI [62] and FSL [63] and provides Java wrappers to encapsulate recently developed software tools.

The software used for preprocessing was part of the Visualization and Analysis Software Tools (VAST), an IDL-based software library developed at the VA Medical Center, Minneapolis, MN. The preprocessing software tools and the IDL-based NPAIRS package were integrated into Fiswidgets through the Java wrappers that Fiswidgets provides. fMRI processing pipelines were built and run on the Fiswidgets GlobalDesktop. Some adaptations were made for models that were not completely incorporated into Fiswidgets such as FSL.FEAT [Image Analysis Group, FMRIB at Oxford (http://www.fmrib.ox.ac.uk/fsl/feat5/index.html)] and SPM2 [Wellcome Department of Imaging Neuroscience, University College London (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/)]. FSL.FEAT was run in batch mode with parameter files through the Unix command widget that Fiswidgets provides. SPM2 was run outside of the Fiswidgets platform because it was not integrated into Fiswidgets and relies on the Matlab environment.

2.3. Preprocessing

Slice timing correction was performed by the FSL. slicetimer. High-pass temporal filtering was realized by applying a band-pass filter to the fMRI time series through multitaper spectrum estimation [64] and setting appropriate cutoffs. Temporal detrending was achieved by specifying a linear combination of cosine basis functions in the GLM design matrix and retaining the residuals and desired effects of the GLM model as the detrended data [58]. Global intensity normalization was performed by dividing the intensities of each scan by its volume mean, for example, proportional scaling [21]. Spatial smoothing was implemented through convolution with a 2D, within-slice Gaussian kernel.

Motion correction was carried out with AIR.alignlinear [65] by applying a six-parameter rigid-body transformation to align each fMRI volume with the first scan of the first run in order to remove head motion. The mean fMRI-to-structural MRI transformation (six parameters) also used AIR.alignlinear. The intrasubject alignment from individual fMRI space to structural MRI space was derived by multiplying the fMRI motion correction transformation and the mean fMRI-to-structural MRI transformation. The fMRI scans were then resampled to the individual MRI space by applying the derived transformation to each fMRI volume and projecting it into the subject's structural MRI space.

Intersubject alignment (or spatial normalization) was also performed for pipeline optimization and evaluation of analytical models. The intersubject alignment transformation was derived by combining the intrasubject alignment transformation with a structural MRI-to-MNI152 (Montréal Neurological Institute brain template) transformation using a seventh-order polynomial warp in AIR5.03 [66]. The fMRI volumes were then aligned to the MNI template brain through intersubject alignment.

2.4. Evaluate the impact of preprocessing steps

The impact of a series of preprocessing steps [slice timing correction, motion correction, spatial smoothing, temporal filtering (high-pass filtering, detrending) and global intensity

Table 1
The pipeline options and parameters for evaluation of preprocessing steps tested

Preprocessing steps		Options and parameters of processing pipelines																	
		1. Slice timing correction			2. Motion correction			3. Spatial smoothing (FWHM in pixels)				4. Temporal filtering						5. Global intensity normalization	
												High pass (×0.001 Hz)			Detrending (cosine)				
1. Slice timing correction		x		F			A			0, 2				x			0, 1, 2		x	
2. Motion correction			x			x		A		0, 2				x			0, 1, 2		x	
3. Spatial smoothing			x				A		0	1, 1.5, 2, 4, 6, 8				x			0, 2		x	
4. Temporal filtering	High pass			x			A			0, 2		0		2, 4, 6, 8			0		x	
	Detrending			x			A			0, 2				x		0	1, 2, 3, 4		x	
5. Global intensity normalization				x			A			0, 2				x			0, 1, 2		x	V

FWHM — full-width-at-half-maximum; F — FSL.slicetimer; A — AIR.alignlinear; x — no operation performed; V — VAST.

normalization] was investigated with the NPAIRS.CVA model in single-subject analysis. These steps were chosen for comparison with published ROC and NPAIRS results.

To avoid testing preprocessing steps in isolation, a set of preprocessing options in a range of settings (Table 1) was set to evaluate each tested preprocessing step. For example, as shown in Table 1, the impact of slice timing correction was tested by turning it on and off in pipelines with motion correction, 0- and 2-pixel spatial smoothing and 0-, 1- and 2-cosine temporal detrending settings. However, tested parameters were dropped from further testing when no significant effect was found to limit the total number of combinations.

After preprocessing, PCA was performed and dimension reduction (denoising) was achieved by keeping a small number (2, 5, 10 or 25) of the principal components (PCs) that accounted for much of the variability in each subject's data. In order to study the primary effect of activation versus baseline, a two-class CVA that categorizes six baseline periods as one class and five activation periods as the other class was performed using the retained PCs. Scans at the baseline-to-activation and activation-to-baseline transitions were dropped to improve the CVA cost function (based on the ratio of between-class to within-class covariance) [8] by avoiding hemodynamic transition-scan effects.

For each pipeline combination tested, the NPAIRS prediction accuracy (calculated as the posterior probability for each scan's true class membership) and reproducibility (obtained by correlating the output SPIs from each of the independent split-half data sets) were used to calculate the Euclidean distance (D) between each prediction accuracy (p) and reproducibility (r) value pair (p, r) to the perfect prediction accuracy and reproducibility pair (1, 1). Equal weighting (1:1) was given to prediction and reproducibility measures in this study to calculate the Euclidean distance between the (p, r) pair of the pipeline tested and the perfect pair (p, r)=(1, 1). These distances were compared for each subject and each preprocessing step (Table 1) with and without the preprocessing step to identify those that significantly changed the metrics (distance D) across the 16 subjects.

Mean distance change (ΔM) was used to measure the impact of each preprocessing step tested, which is defined as the difference of mean distance across all subjects to perfect prediction and reproducibility (1, 1) by subtracting the mean distance with the step from that without the step. It is expressed as:

$$\Delta M = \overline{D_0} - \overline{D} = \frac{1}{N}\left\{\sum_{i=1}^{N}\sqrt{(1-p_{i0})^2+(1-r_{i0})^2} - \sum_{i=1}^{N}\sqrt{(1-p_i)^2+(1-r_i)^2}\right\} \quad (1)$$

where $\overline{D}$ is the mean distance between the (p, r) performance of the pipeline tested and (1, 1) in the prediction and reproducibility plane, p_{i0} and r_{i0} are the prediction accuracy and reproducibility without the preprocessing step tested for the ith subject, p_i and r_i are the ones with the preprocessing step for the ith subject and N is the total number of subjects in the data set.

Table 2
The preprocessing options and parameters for evaluation of nine pipelines

Pipelines/Models	Nonidentical options and parameters
1. NPAIRS.CVA1	2 cosine basis detrending, 5 principal components (#PCs)
2. NPAIRS.CVA2	Optimized #PCs (on reproducibility)
3. NPAIRS.CVA3	Optimized detrending and #PCs (on reproducibility)
4. NPAIRS.CVA4	Optimized smoothing, detrending, #PCs (on prediction and reproducibility)
5. NPAIRS.GLM1	2 cosine basis detrending
6. NPAIRS.GLM2	Optimized detrending options generated from NPAIRS.CVA3
7. NPAIRS.GLM3	Optimized smoothing and detrending options from NPAIRS.CVA4
8. FSL(3.2).FEAT	176 s high-pass filtering cutoff, prewhitening, default Hrf
9. SPM2	176 s high-pass filtering cutoff, prewhitening, default Hrf

A Wilcoxon matched pairs rank sum test (pairing the distances for each subject with and without a particular preprocessing step) was used to evaluate the significance of the steps tested [67]. Relative variation was further computed through dividing the mean distance change (ΔM) by its standard deviation to compare the relative impact of the preprocessing steps tested.

2.5. Optimize the preprocessing steps

Based on the evaluation of preprocessing steps, the processing pipelines per subject were optimized for spatial smoothing, temporal detrending and denoising by reducing the number of PCs for the PCA/CVA model. The parameter choices included spatial smoothing (FWHM=0, 1.5, 2, 4, 6 pixels), temporal detrending (cosine cycle of 0 and ≤1, 1.5, 2, 3, 4) and #PCs (2, 5, 10, 25) for the two-class CVA.

The impact of such optimization on GLM-based pipelines was examined with between-subject reproducibility (BSR). In BSR, the number of activated voxels ($Z>3$) common to each pair of subjects relative to the average number of activated voxels between both subjects was measured, and this procedure was repeated for all possible pairs of subjects to obtain a conjunction matrix. The BSR was based on the average of the conjunction matrix. The BSR measure was applied to SPIs generated from intersubject aligned data alone. For a detailed description of BSR conjunction analysis, see Ref. [59]. Based on the pipeline optimization results for the 16 subjects, an optimized BSR matrix (16×16) was formed. The nonoptimized BSR matrices were calculated using the SPIs generated by the nonoptimized pipelines, and they were ranked by mean BSR across 16 subjects to obtain the best-performing nonoptimized pipeline (or the penultimate pipeline), which was then compared with the optimized BSR matrix to see whether group homogeneity improved after optimization.

A parametric paired t test and nonparametric Wilcoxon matched pairs rank sum test were used to calculate the significance of differences between the optimized BSR matrix and the best-performing nonoptimized matrix.

2.6. Evaluate heterogeneous pipelines

Using both fixed and individually optimized preprocessing pipelines, we compared GLM- and CVA-based analysis models with four analytical modules in FSL3.2, SPM2 and NPAIRS, namely, FSL.FEAT(GLM), SPM2(GLM), NPAIRS.GLM and NPAIRS.CVA. Since a GLM prediction metric has not been implemented in the current NPAIRS package, only between-run SPI reproducibility (i.e., the correlation between the two SPIs generated from each run for each subject) was used as the performance metric. A second-level CVA using NPAIRS.CVA was applied to compare the resulting SPIs and to test for significant spatial pattern differences across preprocessing pipelines, data analysis models and heterogeneous software packages.

To evaluate the analytical methods across FSL3.2, SPM2 and NPAIRS, we used identical initial preprocessing steps: masks were generated by FSL.BET based on the mean fMRI volume of each subject, and spatial smoothing was performed with in-plane Gaussian filtering (FWHM=2 pixels).

The preprocessed data were fed to univariate FSL.FEAT, SPM2 and NPAIRS.GLM and to multivariate NPAIRS.CVA, respectively, for statistical analysis. Since FSL.FEAT and SPM2 convolve the GLM design matrix with the hemodynamic response function and NPAIRS.GLM does not, transition scans were kept for FSL.FEAT and SPM2 but were dropped for NPAIRS.GLM and NPAIRS.CVA. Because of different methods and implementations of temporal filtering in FSL, SPM2 and NPAIRS (FSL uses high-pass filtering while SPM2 and NPAIRS use detrending with a cosine basis function), the parameters of the high-pass filtering and cosine detrending had to be matched in order to remove similar components of low-frequency noise. For the static force data set, a 176-s cutoff was chosen for high-pass filtering in SPM2 and FSL.FEAT, and two cosine cycles were chosen as the corresponding cosine detrending parameter in NPAIRS since after dropping the transition scans, there are, on average, 16 scans for each baseline–activation cycle and there are 5.5 task cycles in a run (corresponding to 352 s).

The number of PCs was fixed at five in order to see the impact of tuning the CVA in the original two-class CVA (CVA1, Table 2). In addition, three other CVA variations were used to monitor changes in between-run reproducibility. They include one with a flexible number of PCs, that is, optimize #PCs for maximum between-run reproducibility, CVA2; one with flexible cosine detrending and #PCs, that is, optimize cosine detrending and #PCs for maximum between-run reproducibility, CVA3; and, finally, one with flexible spatial smoothing, cosine detrending and #PCs, that is, optimize smoothing, detrending and #PCs for the minimum $\overline{D}$ [the mean distance to $(p, r)=(1, 1)$], CVA4. Similarly, for the NPAIRS.GLM model, the original GLM

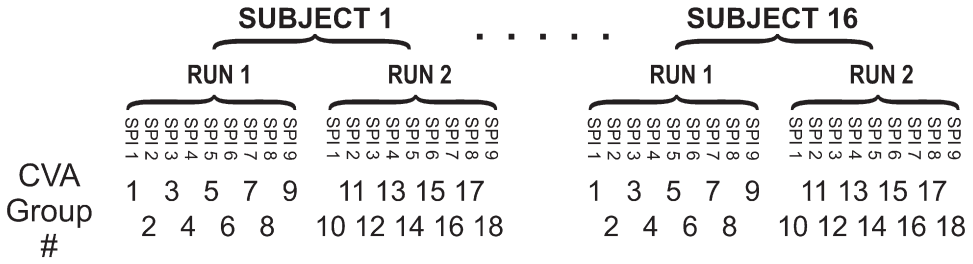

Fig. 1. CVA class structure for SPI comparison. The results of nine pipelines — 288 (16 subjects×2 runs×9 models) SPIs were assigned to 18 classes (or groups) with 16 SPIs in each class.

method was named GLM1. GLM2 was formed by applying CVA3 (flexible cosine detrending) detrending parameters to the input of GLM, and GLM3 was formed by applying CVA4 preprocessing parameters. The seven pipelines together with FSL.FEAT and SPM2 (FSL.FEAT and SPM2 were applied with prewhitening by default) formed the nine pipelines, and their parameter settings are listed in Table 2. Both intrasubject and intersubject alignments were performed.

After running these heterogeneous pipelines, the mean between-run SPI reproducibility across the 16 subjects was calculated. A Wilcoxon matched pairs rank sum test was used to evaluate the significance of selected between-run reproducibility differences.

A second-level NPAIRS-CVA, that is, applying CVA class structure to the SPIs being compared per subject per run, was then applied to the 288 (16 subjects×2 runs×9 pipelines) SPIs generated by the above 9 pipelines with resampled split-half groups of 8 subjects. The class structure of the second-level CVA is presented in Fig. 1. Subject variance and mean run effect were removed by subtracting the mean of each SPI and normalizing it by its spatial variance in order to help CVA capture the variance of these SPIs caused by the processing of pipeline components (i.e., preprocessing steps and statistical analysis).

Prewhitening was performed with FSL.FEAT- and SPM2-based pipelines (see Table 2 for details) but not with NPAIRS.GLM/CVA-based pipelines (since prewhitening was not implemented in the IDL-based VAST and NPAIRS package). In order to test the impact of prewhitening, we added an FSL.FEAT-based pipeline with no prewhitening performed to the above 9 pipelines. A second-level CVA was performed on the SPIs generated by the 10 pipelines to test the relative impact of prewhitening.

In order to see the difference between individually optimized and group-wise spatial smoothing, an SPI comparison was performed by a second-level CVA on the SPIs generated by seven heterogeneous pipelines at 2- and 6-pixel spatial smoothing and 2-cosine detrending. These pipelines were named CVA1 — NPAIRS.CVA with 5 PCs, 2-pixel smoothing; CVA2 — NPAIRS.CVA with optimized #PCs, 2-pixel smoothing; GLM1 — NPAIRS.GLM, 2-pixel smoothing; and FEAT1 — FSL.FEAT with 176 s high-pass filtering, 2-pixel smoothing; CVA3, GLM2 and

Table 3
The impact of preprocessing steps tested with CVA-based pipelines

(P1) The impact of slice timing correction

	0-pixel smoothing				2-pixel smoothing			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
0 cosine	5	0.02	0.219	1.00	5	−0.03	0.196	0.62
2 cosine	5	0.09	0.283	0.07	5	0.02	0.308	0.21

(P2) The impact of motion correction

	0-pixel smoothing				2-pixel smoothing			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
0 cosine	2	0.08	0.206	0.04	2	0.11	0.214	0.06
2 cosine	5	0.16	0.089	0.00	5	0.08	0.094	0.00

(P3) The impact of spatial smoothing

	0-cosine detrending				2-cosine detrending			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
1.5 pixels	5	0.09	0.090	0.00	2	0.12	0.239	0.00
2 pixels	5	0.12	0.095	0.00	5	0.11	0.093	0.00
4 pixels	25	0.12	0.061	0.00	5	0.14	0.141	0.00
6 pixels	25	0.12	0.080	0.00	25	0.10	0.058	0.00

(P4.1) The impact of high-pass filtering

Cutoff (Hz)	0-pixel smoothing				2-pixel smoothing			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
0.002	25	0.04	0.087	0.03	10	0.07	0.103	0.02
0.004	25	0.05	0.083	0.03	10	0.06	0.090	0.02
0.006	25	0.06	0.086	0.02	10	0.10	0.124	0.01
0.008	10	0.02	0.165	0.70	10	0.03	0.123	0.39

(P4.2) The impact of cosine basis detrending

No. of cosine cycles	0-pixel smoothing				2-pixel smoothing			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
1	2	0.09	0.232	0.14	2	0.16	0.250	0.03
2	2	0.15	0.158	0.00	2	0.17	0.190	0.00
3	5	0.09	0.149	0.02	2	0.10	0.125	0.00
4	10	0.04	0.173	0.36	10	0.03	0.130	0.45

(P5) The impact of global intensity normalization

	0-pixel smoothing				2-pixel smoothing			
	#PCs	ΔM	S.D.	Sig.	#PCs	ΔM	S.D.	Sig.
0 cosine	25	0.05	0.164	1.00	5	0.07	0.210	0.37
2 cosine	2	0.03	0.043	0.02	2	0.05	0.162	0.59

Table 4
Summary of the impact of the preprocessing steps tested

Preprocessing steps		ΔM	S.D.	Sig.	Coeff. vari.
1	Slice timing correction	0.07	0.197	0.14	0.355
2	Motion correction	0.08	0.094	0.00	0.851
3	Spatial smoothing (SS)	0.11	0.093	0.00	1.183
4	High-pass filtering (HPF)	0.10	0.124	0.01	0.806
	Temporal detrending (TD)	0.17	0.190	0.03	0.895
5	Global intensity normalization	0.04	0.100	0.13	0.400

Note: ΔM — mean distance change (mean distance without the tested preprocessing step−mean distance with the tested preprocessing step); Sig. — significance tested by Wilcoxon matched pairs rank sum test; Coeff. vari. — coefficient of variation, that is, normalized mean distance change.

FEAT2 have the same pipeline options as CVA1, GLM1 and FEAT1, respectively, except for 6-pixel spatial smoothing instead of 2-pixel spatial smoothing.

3. Results

3.1. Evaluate the impact of preprocessing steps

Table 3 gives the results of preprocessing steps tested in the pipelines as described in Table 1. The mean distance change (ΔM) is a result of turning on and off a preprocessing step; $\Delta M>0$ implies improved performance. Table 3 indicates that slice timing correction and global intensity normalization have little impact. Motion correction, spatial smoothing, temporal detrending and high-pass filtering were found to significantly improve the performance of CVA-based pipelines ($P<.05$).

Table 4 summarizes the evaluation results of all the preprocessing steps tested (in Table 3) at 2-pixel spatial smoothing with optimized detrending parameter that is obtained by calculating the relative variation, that is, the maximum ratio of the mean distance change (ΔM) to the corresponding standard deviation. One can see that for block designs, slice timing correction and global intensity normalization have little impact, while spatial smoothing has the largest positive impact, followed by temporal detrending, motion correction and high-pass filtering on the performance of CVA-based fMRI processing pipelines.

3.2. Optimize the processing pipelines

The optimization results are summarized in Table 5. There is considerable variance among the subjects in the optimized preprocessing parameters and the minimum distance to perfect prediction and reproducibility: it ranges from 0.12 with prediction and reproducibility of (0.90, 0.95) for S4, or (0.96, 0.89) for S15, to 0.53 (0.60, 0.65) for S13. There are differences in optimal processing options and insignificant difference (Min. Dist.) in the minimum distances between intrasubject and intersubject alignment.

The results of the optimization impact on pipeline performance across the 16 subjects are summarized in Table 5B with BSR for all pairs of subjects. In Table 5B, a negative mean difference indicates an improvement in SPIs generated by the optimized preprocessing data. The results of paired t and Wilcoxon matched paired rank sum tests show that the best performing nonoptimized BSR conjunction matrix is significantly better than the optimized BSR conjunction matrix (Mean Diff. > 0), but pipeline optimization on average significantly improves the BSR (Avg. Mean Diff. < 0). This might suggest a minor gain in group homogeneity, namely, the improved common activation detection across all 16 subjects.

3.3. Evaluate heterogeneous pipelines

3.3.1. With one NPAIRS metric — SPI reproducibility

Fig. 2 presents the evaluation results of mean between-run reproducibility for the nine analytical models (listed in Table 2). The range of mean between-run reproducibility is from 50% to 75%. One can see that the lowest mean reproducibility is CVA1 (nonoptimized #PCs), which is significantly lower than that of FEAT and SPM2. However, when the CVA model is optimized on #PCs (with between-run reproducibility metric), its mean reproducibility increases dramatically (indicated as CVA2).

Furthermore, when we allow both #PCs and temporal detrending to be optimized, the reproducibility further increases as shown in Fig. 2 (CVA3). However, when the optimized detrending parameters were applied to the input of GLM, the mean reproducibility stayed the same (GLM2) as that of GLM1. Friedman's test (a nonparametric test for differences of several treatment effects in a randomized complete block design [67]) indicates that there is no significant difference ($P<.05$) in the distribution of mean reproducibility between FEAT, SPM2 and GLM1, between FEAT, SPM2 and CVA2, or between FEAT, SPM2 and CVA3. Nevertheless, when the CVA model is applied to the optimized preprocessing data (where the optimization was based on prediction and between-run reproducibility metrics for the three options of spatial smoothing, temporal detrending and #PCs), it generates even higher mean reproducibility between runs (CVA4). When such optimized preprocessing options were applied to the input of GLM as GLM3 represents, it improved the mean between-run reproducibility for the GLM model. The CVA4 and GLM3 improvements are not unexpected given the overall increase in smoothing kernel size above 2 pixels for most subjects (see Table 5).

Fig. 2 also shows that the mean between-run reproducibility generated from intrasubject aligned data is, in general, higher than that generated from intersubject aligned data, which suggests that some noise was introduced through MRI-to-MNI transformation in intersubject alignment.

3.3.2. SPI comparison through second-level CVA

Fig. 3 illustrates the CVA results of SPI comparisons where the 288 (16 subject×2 runs×9 models) SPIs were generated from the 9 pipeline models (Table 2) on intersubject aligned data. It indicates that the differences

Table 5

(A) The optimal preprocessing options and results for 16 single subjects

Subject	Spatial smoothing (pixel)		Temporal detrending (cosine)		Denoising (#PCs)		Min. Dist.	
	Intra.	Inter.	Intra.	Inter.	Intra.	Inter.	Intra.	Inter.
S1	6	6	1.5	2	10	10	0.38	0.34
S2	6	6	2	2	10	10	0.38	0.42
S3	0	1.5	1.5	2	5	2	0.30	0.29
S4	2	2	2	2	2	2	0.12	0.13
S5	4	6	1	0	5	10	0.29	0.35
S6	4	4	2	2	5	5	0.26	0.31
S7	4	4	1.5	0	2	2	0.17	0.17
S8	4	6	3	1.5	5	5	0.25	0.37
S9	6	4	3	3	2	2	0.40	0.42
S10	6	6	1.5	1.5	2	2	0.18	0.19
S11	4	4	1	1	10	10	0.46	0.47
S12	4	4	2	1	2	2	0.30	0.29
S13	6	6	2	1.5	5	2	0.53	0.51
S14	0	6	1	1.5	10	25	0.46	0.52
S15	2	2	3	3	2	2	0.12	0.13
S16	6	4	2	0	10	10	0.28	0.29

Note: #PCs — number of principal components; Min. Dist. — the minimum distance between the (prediction, reproducibility) pair of the tested pipeline to (1, 1) perfect prediction and reproducibility pair; Intra. — intrasubject alignment; Inter. — intersubject alignment.

(B) The impact of pipeline optimization (with BSR)

	BSR	S.D.	Paired *T* Sig.	W. S. Sig.
Mean Diff.	0.095	0.312	0.00	0.00
Avg. Mean Diff.	−0.106	0.204	0.00	0.00

Note: Mean Diff. BSR — mean difference (best-performing nonoptimized BSR conjunction matrix−optimized BSR conjunction matrix); Avg. Mean Diff. BSR — mean difference (the average of the best-performing and worst-performing nonoptimized BSR conjunction matrix−optimized BSR conjunction matrix). W. S. Sig. — significance of Wilcoxon matched pairs rank sum test.

between the univariate GLM model and the multivariate CVA model account for the largest variance (80.1%) due to activation pattern differences, regardless of the details of the GLM model (e.g., HRF or no HRF), software implementation [SPM2 (blue), FSL.FEAT (red), NPAIRS.GLM (black and green)] and little mismatches in parameter matching (for cutoffs of high-pass filtering in FSL.FEAT and detrending in SPM2 and NPAIRS.GLM). The second largest variance (13.9%) is driven by differences in spatial smoothing across subjects for the individually optimized pipelines — Gaussian FWHM ranges from 0 to 6 pixels (green) while the remaining pipelines have FWHM=2 pixels. There are much smaller variations across the black lines joining the 16-subject means. These are due to preprocessing pipeline differences within a given model class and run. The run-to-run effects are even smaller and occur mainly in the individual subject distributions about the similar means/pipeline (black line segments) across runs. Fig. 4 gives the details of the two sources of variances in canonical eigenimages (CE). Fig. 4A represents the first dimension of CVA variances in SPI comparison, that is, the variance due to activation pattern differences between CVA and GLM where red-yellow regions are significantly higher in CVA SPIs and blue-green regions are significantly higher in GLM SPIs. It shows that the CVA-based models are more sensitive to activations and functional connectivity around the cerebellum (A10) and parietal cortex (A26-27) with some potential artifacts (A17-A20), while the GLM-based models are more sensitive to activations in regions such as the motor and premotor cortex (A26-27) and associate occipital cortex (A15-17) (in blue-green regions). Fig. 4B is the second-dimension CE (or CE2), which illustrates the spatial pattern from variance of individually optimized spatial smoothing.

Fig. 5 provides an overall view of the impact of prewhitening tested with second-level CVA. The CV score plots demonstrate that prewhitening creates little variance (less than 100%−79.6%−14.5%=5.9%) in the SPI comparisons compared to the largest variance (79.6%), which is caused by the model difference between CVA and GLM, which forms canonical variable 1 (CV1, or first-dimension CV), followed by the second largest variance (14.5%) caused by individually optimized spatial smoothing that forms CV2 (or second-dimension CV). Compared with Fig. 2, there was not much change for pipelines with and without prewhitening, which suggests that the overall impact of prewhitening is small relative to other pipeline options such as model selection and spatial smoothing.

For group-wise spatial smoothing, Fig. 6 indicates that the group-mean differences of spatial smoothing between 2

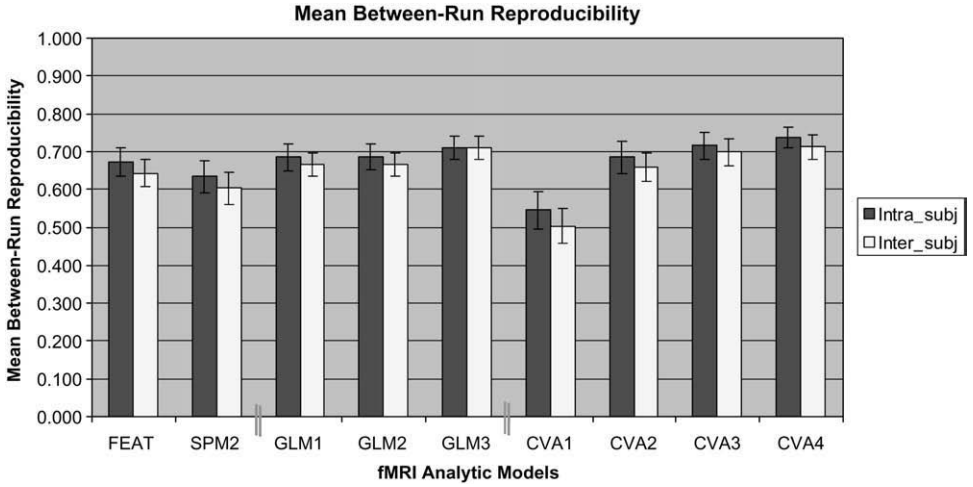

Fig. 2. Mean between-run reproducibility of on-off effect analyzed by nine analytical models (see Table 2 for pipeline details). FEAT, FSL(3.2).FEAT; GLM1, NPAIRS.GLM; GLM2, NPAIRS.GLM2; GLM3, NPAIRS.GLM3; CVA1, NPAIRS.CVA1; CVA2, NPAIRS.CVA2; CVA3, NPAIRS.CVA3; CVA4, NPAIRS.CVA4; Intra_subj. —Intra-subject alignment; Inter_subj — Inter-subject alignment. The error bar is the standard error.

and 6 pixels account for the largest variance (93.1%, CV1) of the second-level CVA, followed by the mean differences between univariate model GLMs and multivariate model CVAs (6.0%, CV2). Fig. 7 illustrates the details of the group-mean differences in the spatial patterns of brain images. Spatial smoothing with 6 pixels (images labeled with A) tends to help identify activations around primary motor cortex, premotor cortex and supplementary motor area (A25-27) while reducing potential artifacts (A17-18), and CV2 (images labeled with B) reflects a pattern similar to CVA vs GLM differences in CE1 of Figure 4. Figs. 6 and 7 provide evidence that spatial smoothing and univariate versus multivariate model selection are the most important pipeline choices in generating SPIs. Also, Figs. 6 and 7 suggest that the order of first-dimension CV (CV1) and second-dimension CV (CV2) in SPI comparison is not fixed, but changes depending on pipeline options — when the pipeline difference in spatial smoothing was larger than the

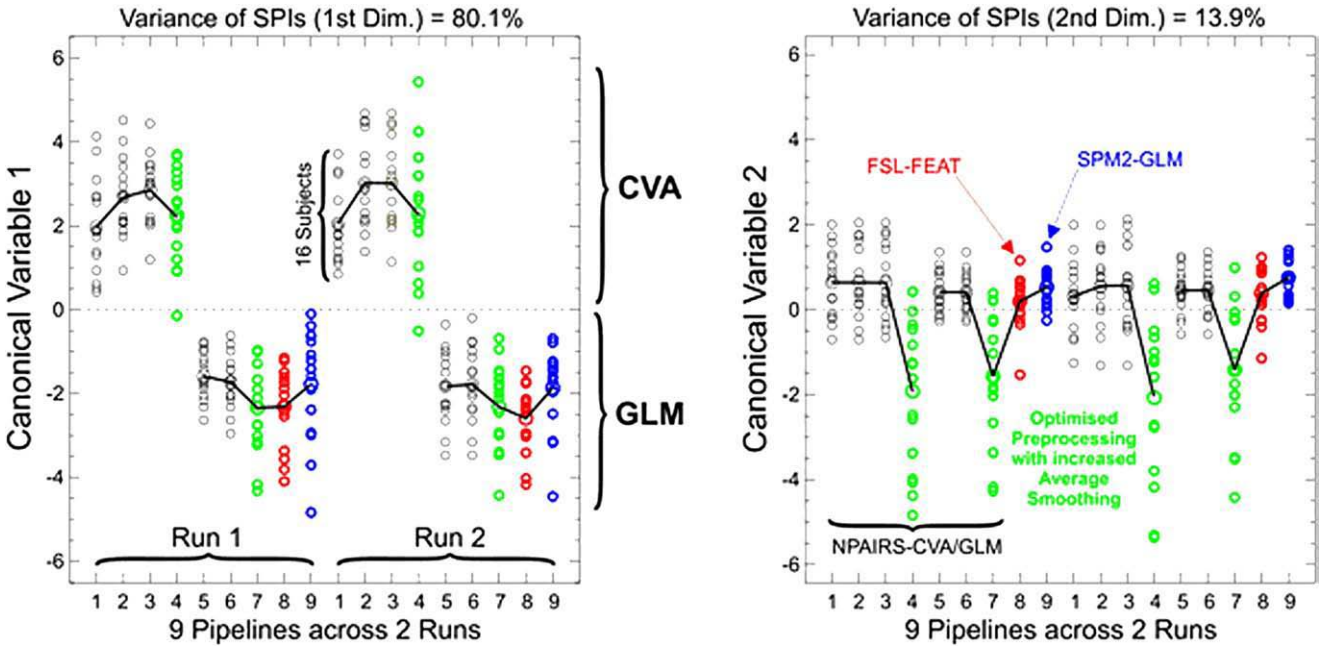

Fig. 3. Canonical score plot for the first and second dimensions of CVA variance analysis nine pipelines for each run (with pipeline details in Table 2): 1. NPAIRS.CVA1; 2. NPAIRS.CVA2; 3. NPAIRS.CVA3; 4. NPAIRS.CVA4 (upper green); 5. NPAIRS.GLM1; 6. NPAIRS.GLM2; 7. NPAIRS.GLM3 (lower green); 8. FSL.FEAT (red); 9. SPM2 (blue).

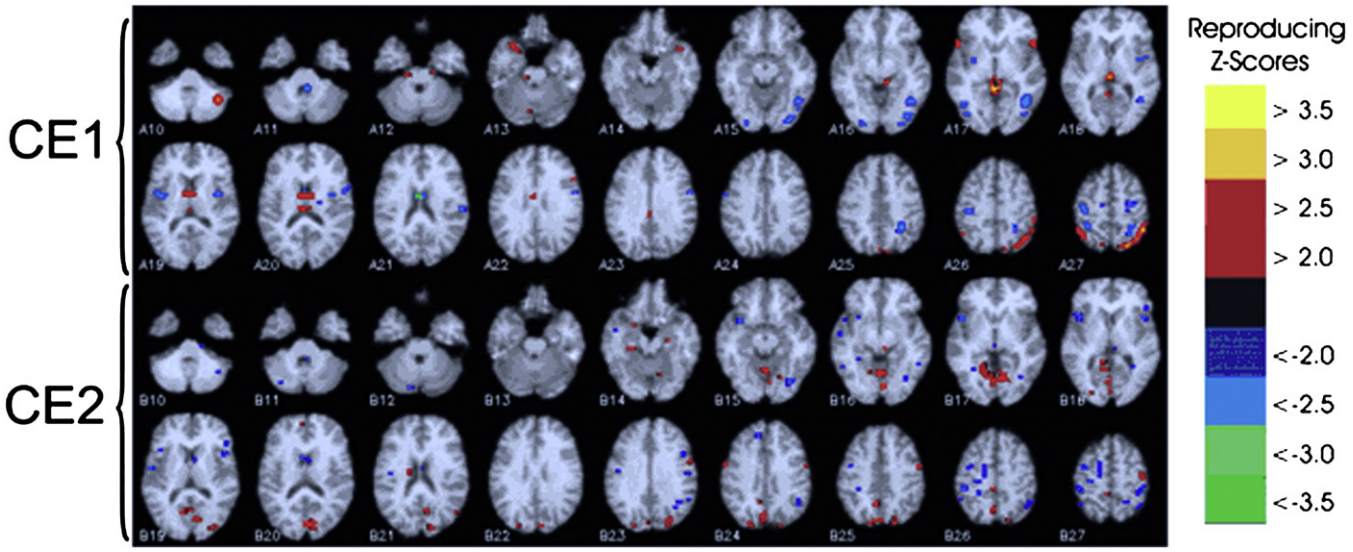

Fig. 4. Spatial patterns (first two dimensions) of CVA variance analysis. (A) First dimension, variance between CVA and GLM; (B) Second dimension, variance of optimized smoothing. CE1: canonical eigenimage of the first dimension; CE2: canonical eigenimage of the second dimension.

model difference, the variance in spatial smoothing became CV1 and model difference became CV2.

4. Discussion

Although univariate GLM is still the dominant approach in fMRI analysis today, there is growing interest in multivariate approaches because they are often data driven and model free and have the potential to directly reveal the functional connectivity of neural networks in the brain. In this study, we applied the NPAIRS approach to the evaluation and optimization of multivariate CVA-based fMRI processing pipelines in single-subject analysis. For preprocessing steps in CVA-based pipelines, both NPAIRS metrics (prediction accuracy and SPI reproducibility) were used as performance measures in this study. To compare with univariate GLM-based pipelines, we used between-run SPI reproducibility alone since a GLM prediction metric has not been implemented in the current NPAIRS package for fMRI software packages such as SPM, which hinders it from evaluating SPM-based pipelines with prediction accuracy metric. In addition, a second-level CVA was applied to the SPIs generated by heterogeneous pipelines to see the relative importance of major fMRI processing pipeline choices.

4.1. Evaluation of the preprocessing steps

The results of the preprocessing step evaluation indicate that, first, slice timing correction in block-design fMRI analysis has little impact on the performance of fMRI processing pipeline, a not unexpected result given that the transition scans most sensitive to the hemodynamic response function had been dropped. However, the remaining block scans reflect a worst-case potential interaction between slice timing and head motion with alternating slice acquisition and an approximately 4-s TR. It is reassuring to find that this does not seem to be a problem. This result supports the current practice that slice timing correction is omitted for block designs (and performed only for event-related fMRI) [18,19]. Another study based on the experiment with the slice timing correction tool in SPM99 concluded that for block designs, not using slice timing correction gave a more robust activation than slice timing correction [68]. While recognizing potential problems with slice timing correction, Smith et al. [63] suggest that slice timing correction should be performed with motion correction in an integral approach.

Second, the results demonstrate that in most cases tested, the impact of within-run global intensity normalization on pipeline performance is not significant, which is consistent with the results using ROC analysis by Skudlarski et al. [6]. This is probably due to nonuniform drift across each scan, and Skudlarski et al. [6] further suggested that it would be unwise to apply global intensity normalization as a general procedure. However, Jezzard et al. [17] pointed out that if global intensity normalization is not carried out, there can be a problem with multisession group analyses and it becomes necessary to perform grand mean normalization.

Third, our results demonstrate that motion correction can significantly improve the performance of CVA-based single-subject fMRI processing pipelines. On the other hand, the results of simulation (with GLM-based pipelines) in Ref. [6] showed that motion correction (performed with the SPM package) did not change the relative efficiency in data analysis. LaConte et al. [8] found that there was little impact with within-subject alignment (which includes motion correction) on NPAIRS performance metrics. However, at a closer look, Jezzard et al. [17] pointed out that when subject motion is uncorrelated with the experimental

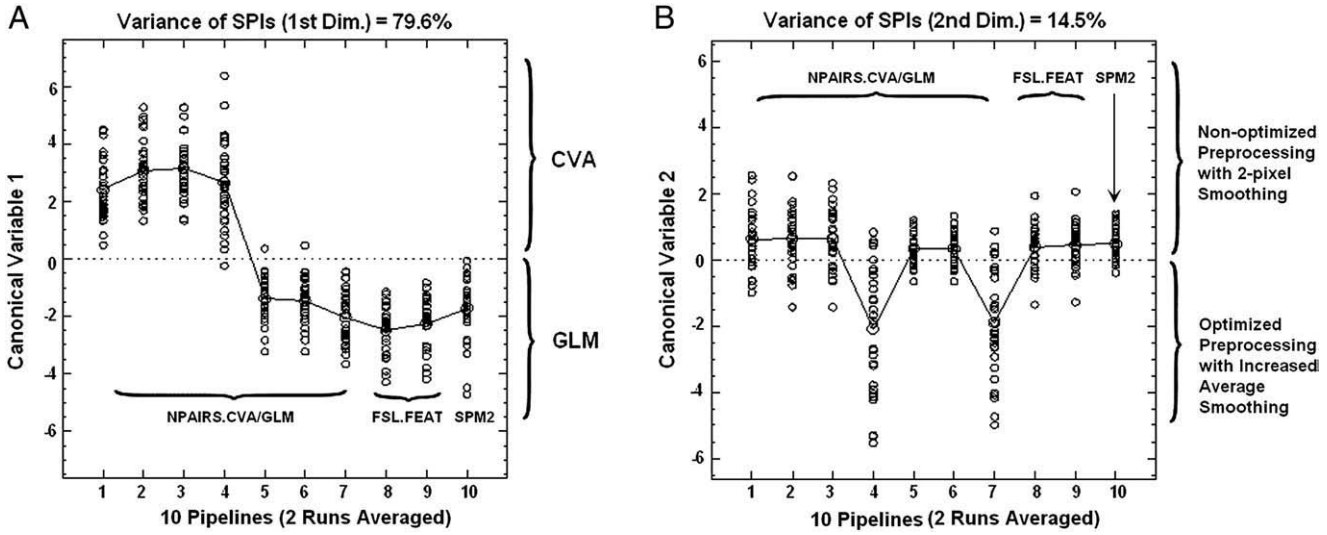

Fig. 5. Canonical variate (CV) scores of 10-pipeline SPI comparisons with and without prewhitening. Ten pipelines for two averaged runs: 1. NPAIRS.CVA1; 2. NPAIRS.CVA2; 3. NPAIRS.CVA3; 4. NPAIRS.CVA4; 5. NPAIRS.GLM1; 6. NPAIRS.GLM2; 7. NPAIRS.GLM3; 8. FSL.FEAT without prewhitening; 9. FSL.FEAT with prewhitening; 10. SPM2 with prewhitening. Pipelines 1–8 are without prewhitening; pipelines 9–10 are with prewhitening. See Table 2 for details of pipelines 1–7, and 9, 10 above, which equal 8, 9 of Table 2, respectively. (A) First dimension, variance between CVA and GLM; (B) Second dimension, variance of optimized smoothing.

paradigm, motion correction makes the activation analysis more sensitive by reducing the standard error and leaving the power of the response unchanged, but when motion is stimulus correlated, motion correction reduces the apparent level of activation. Due to the multivariate nature of CVA, CVA is less influenced by stimulus-correlated motion along each voxel but more influenced by stimulus-uncorrelated motion, which causes the improved signal detection and pipeline performance through motion correction. The difference in results between the LaConte et al. [8] study and this study may stem from the different approaches taken to measure variance (difference of means in Ref. [8] vs. maximum mean difference among all #PCs options in this study) and subject variability.

In addition, we have reported that motion correction had some negative impact on GLM-based processing pipelines [61], which suggests that stimulus-correlated head motion misleads the GLM model (causing lower prediction and reproducibility), but it seemed that it could not mislead the CVA model due to the multivariate nature of CVA. Lund et al. [71] reported that inclusion of motion parameters in the design matrix of GLM analysis

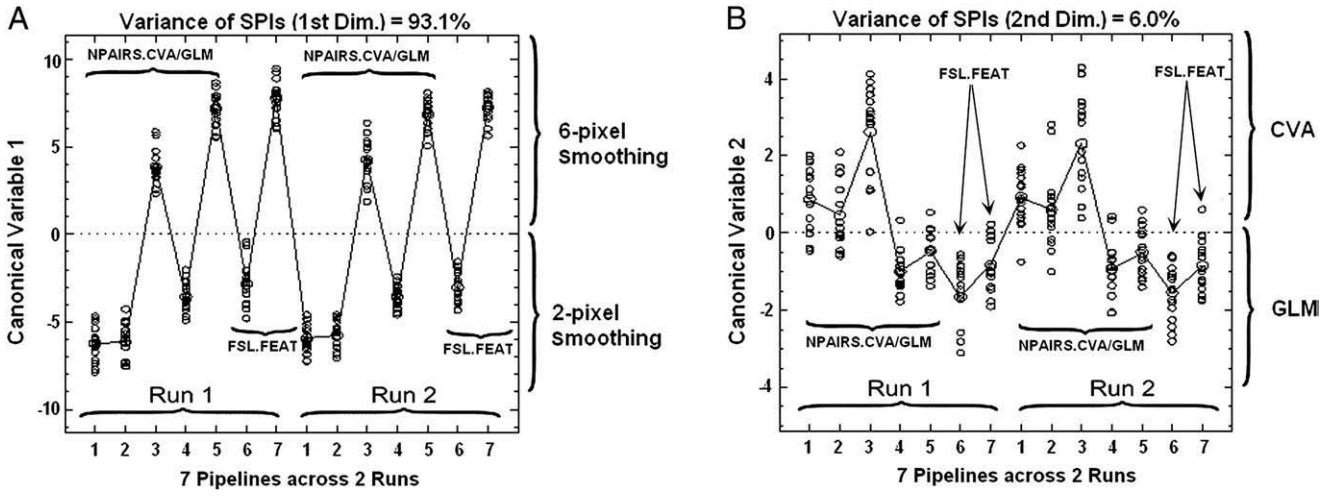

Fig. 6. Canonical variate (CV) score plots of seven-pipeline SPI comparisons through second-level CVA with seven pipelines for each run: (A) First dimension, variance between 6-pixel and 2-pixel spatial smoothing; (B) second dimension, variance between CVA and GLM. 1. CVA1 (2-pixel smoothing, 5 #PCs); 2. CVA2 (2-pixel smoothing, optimized #PCs); 3. CVA3 (6-pixel, optimized #PCs); 4. GLM1 (2-pixel smoothing); 5. GLM2 (6-pixel smoothing); 6. FEAT1 (2-pixel smoothing); 7. FEAT2 (6-pixel smoothing).

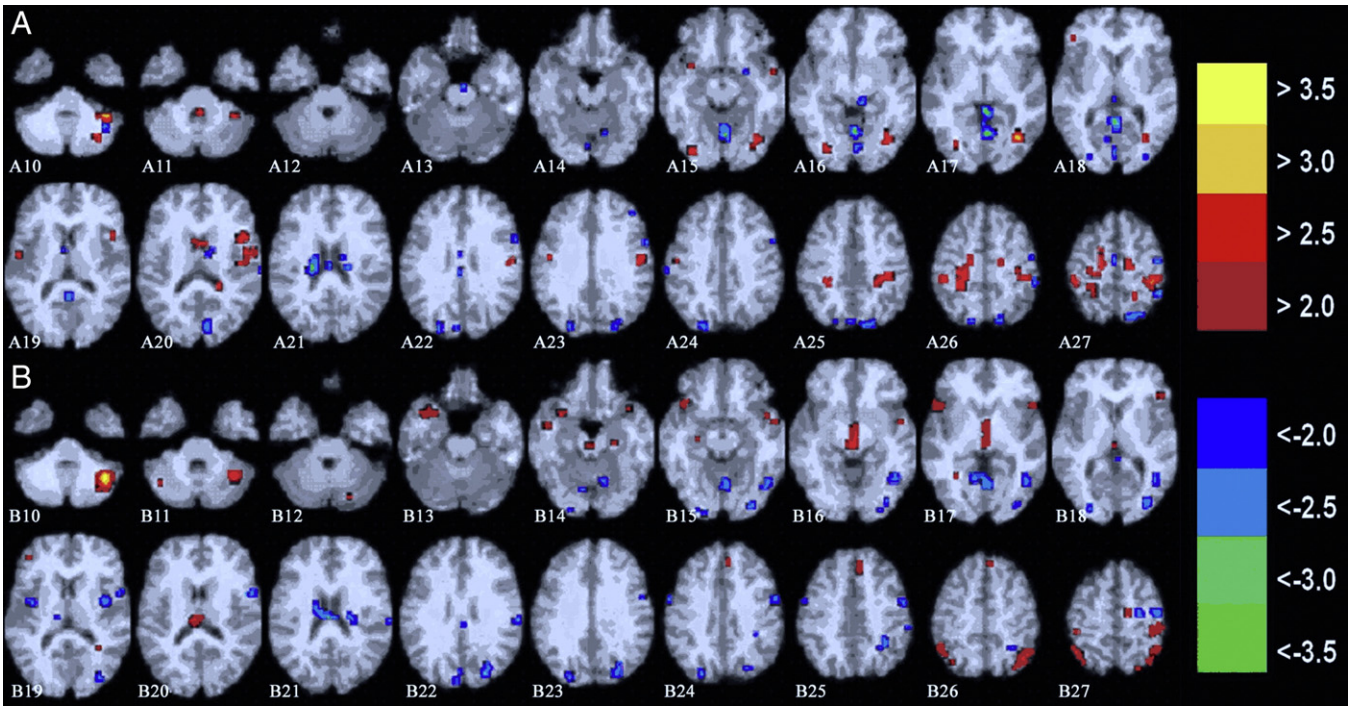

Fig. 7. Spatial patterns of seven-pipeline SPI comparisons through second-level CVA with: (A) First dimension, heavy smoothing (6-pixels, gold-reds)>light smoothing (2-pixels, blue-greens); (B) Second dimension, CVA>GLM (gold-reds), GLM>CVA (blue-greens).

significantly reduced both the intrasubject and the intersubject variance. Several fMRI software packages now allow motion parameters to be included as regressors in the analysis. Morgan et al. [72] compared the sensitivity of four commonly used fMRI software packages such as SPM2 and FSL with simulated data. They found that the most sensitive analysis technique was to perform motion correction and include the realignment parameters as regressors in the GLM, which was most beneficial when stimulus-correlated motion was present.

Fourth, the positive impact of spatial smoothing, high-pass filtering and temporal detrending has been confirmed by ROC analyses [6] (Della-Maggiore et al., 2002) and previous NPAIRS analyses [10,62]. Due to the nonisotropic voxel size (3.4×3.4×5 mm) of the fMRI data set used, 2D in-plane spatial smoothing was performed in this study, while 3D spatial smoothing, which is suitable for data with isotropic voxel size, is more commonly used in fMRI preprocessing nowadays. In general, 3D spatial smoothing may lead to higher sensitivity, but the gain in sensitivity must supersede the loss in specificity induced by the larger number of degrees of freedom that accompanies 3D smoothing. Moreover, Friman et al. [73] reported that compared with 2D spatial smoothing, better performance of 3D smoothing was obtained at the cost of extra computations.

Further, our results show that for block designs, slice timing correction and global intensity normalization have little impact, but spatial smoothing, temporal detrending, motion correction and high-pass filtering significantly improve CVA-based pipeline performance. Interestingly, these results are, in general, similar to the results obtained from GLM-based pipelines [61] where we found that slice timing correction and global intensity normalization have little consistent impact on fMRI processing pipelines and that spatial smoothing and high-pass filtering and/or temporal detrending significantly increase pipeline performance. This suggests that the impact of preprocessing steps on pipeline performance might be somewhat model independent [59].

To allow for the possible interaction among parameter settings of various processing steps along the processing pipeline, we designed our evaluation experiment in a way that each preprocessing step was tested in a variety of parameter settings (e.g., different levels of spatial smoothing and temporal detrending). The evaluation results of the preprocessing steps confirm that the parameter settings of the five preprocessing steps we tested (slice timing correction, motion correction, spatial smoothing, temporal filtering and global intensity normalization) do interact with one another. Therefore, it does not seem reasonable to evaluate and optimize the processing steps in relative isolation [62].

Since this study uses one data set (the static force data set), its results may or may not be applicable to other data sets. However, the majority of our results (i.e., significant positive impact of spatial smoothing, high-pass filtering and temporal detrending as well as little impact of slice timing correction and global intensity normalization) on the

performance of multivariate CVA-based pipelines are consistent with previous ROC findings and/or findings from other studies that are mostly based on univariate GLM [6] (Della-Maggiore et al., 2002). This consistency suggests that the impact of these preprocessing steps tends to be somewhat model independent.

4.2. Optimization of fMRI processing pipelines

In this study, the NPAIRS-based optimization of fMRI single-subject processing pipelines is by no means exhaustive, and only a subset of parameter choices that are of testing importance has been selected. Nevertheless, the results demonstrate considerable variation of optimal options among subjects.

Furthermore, the analysis results based on BSR demonstrate that NPAIRS-based optimization, on average, significantly improves activation detection across all subjects in the group (or group homogeneity), which supports aggregating the individually optimized data in a random-effect group analysis to obtain improved group results [59].

4.3. Evaluation of heterogeneous pipelines

With the available NPAIRS performance metrics in the NPAIRS framework, the evaluation results based on between-run reproducibility and SPI comparison through secondary CVA revealed some valuable information. First, the largest variance in the resulting SPIs (i.e., the output of processing pipelines) is caused by the variance between heterogeneous models GLM and CVA (80.1%) (regardless of the software packages used: FSL.FEAT, SPM2 and NPAIRS). This reveals the fundamental difference between the two kinds of statistical analysis models (univariate GLM and multivariate CVA) in activation detection. The univariate GLM examines one voxel at a time and its variance is based on the mean intensity changes at a single voxel along the time series. The statistical scores generated by the univariate GLM are linearly proportional to the ratio of relative signal to relative noise. On the other hand, multivariate PCA/CVA examines all voxels in a scan simultaneously and characterizes the spatial and temporal response over the entire volume; since the CVA solution is drawn from the eigenstructure of $W^{-1}B$ (W is the within-class covariance and B is the between-class covariance), CVA is sensitive to the mean-difference covariance structure (i.e., B in $W^{-1}B$) and there is a tendency for W^{-1} to act as a pooled noise term compared to the voxel-based variance estimate of GLM. [69]. The relative noise measurements performed by Shaw et al. [70] showed that the CVA was more sensitive to correlated signals while GLM was not. This may explain why CVA highlighted certain brain regions that GLM identifies with lower statistical values. On the other hand, there are other brain regions revealed strongly by univariate GLM that have lower statistical values using multivariate CVA. The reason is that there is a larger mean difference in the voxel intensity between the baseline and activation groups along the single time series (easily identified by the GLM) than the difference in the CVA covariance structure. The fundamental differences between univariate GLM and multivariate CVA suggest that they may necessarily serve as complementary approaches in fMRI analysis.

Second, individually optimized spatial smoothing accounts for 13.9% (Figs. 3 and 4) of the pipeline variance and the group-wise spatial smoothing can account for 93.1% (Figs. 6 and 7) of the pipeline variance. This reveals the important impact of spatial smoothing on the pipeline performance. The important role of spatial smoothing was also observed in Refs. [8,58]. The between-run reproducibility results in Fig. 2 confirm the value of optimizing preprocessing steps (spatial smoothing and temporal detrending) and model parameter (#PCs). When these parameters were optimized, the mean reproducibility between runs for CVA model increased, which indicates the importance and necessity of tuning pipeline parameters.

Third, although there are large differences between the GLM- and CVA-based models, there are also similarities in their SPI reproducibility results. Fig. 2 indicates that both the CVA and GLM model can generate similar mean between-run reproducibility with a range of 60–75% after CVA optimized on #PCs (i.e., the output SPIs are, on average, 60–75% reproducible between runs for both GLM- and CVA-based models). This is consistent with the finding in PET studies (similar activation pattern reproducibility between the GLM and CVA results) [59].

The limitation of this study is that since the GLM prediction metric for SPM is not available in the current NPAIRS package, only between-run reproducibility in the NPAIRS performance metrics is used, which is inadequate by itself for pipeline evaluation [11]. Thus, it is necessary to explore GLM prediction metric for commonly used software packages such as SPM in the future.

In summary, in this study, we investigated the impact of commonly used fMRI preprocessing steps, optimized the associated multivariate CVA-based single-subject fMRI processing pipelines with the NPAIRS approach and compared univariate GLM versus CVA-based processing pipelines with SPI reproducibility and a second-level CVA. The results demonstrated that for the block-design data used, (a) slice timing correction and global intensity normalization have little consistent impact on the fMRI processing pipeline, but spatial smoothing, temporal detrending or high-pass filtering, and motion correction significantly improved pipeline performance across all subjects; (b) the combined optimization of spatial smoothing, temporal detrending and CVA model parameters on average improved between-subject reproducibility; and (c) the most important pipeline choices include univariate or multivariate statistical models and spatial smoothing. The fundamental differences between univariate GLM and multivariate CVA suggest that they may necessarily serve as complementary approaches in fMRI analysis. The

limitations of the current NPAIRS package revealed in this study will motivate further development and improvement of the NPAIRS framework so that it can assess fMRI processing pipelines of both univariate and multivariate models across various fMRI analysis software packages (SPM, AFNI, etc.) and help improve the accuracy and reliability of fMRI software packages to meet the demand of growing fMRI applications.

Acknowledgments

We would like to thank James Ashe, M.D., and Suraj Muley, M.D., for providing the static force data. We are also grateful to Kelly Rehm, Kirt Schaper and Kate Fissell for valuable discussions and technical assistance. This work was partly supported by the NIH Human Brain Project P20 Grant MN EB002013.

References

[1] Stippich C, Rapps N, Dreyhaupt J, Durst A, Kress B, Nennig E, et al. Localizing and lateralizing language in patients with brain tumors: feasibility of routine preoperative functional MR imaging in 81 consecutive patients. Radiology 2007;243(3):828–36.
[2] Bookheimer SY, Strojwas MH, Cohen MS, Saunders AM, Pericak-Vance MA, Mazziotta JC, et al. Patterns of brain activation in people at risk for Alzheimer's disease. N Engl J Med 2000;343(7):450–6.
[3] Tharin S, Golby A. Functional brain mapping and its applications to neurosurgery. Neurosurgery 2007;60(Suppl 2):185–201.
[4] Luo WL, Nichols TE. Diagnosis and exploration of massively univariate neuroimaging models. Neuroimage 2003;19(3):1014–32.
[5] Fissell K, Tseytlin E, Cunningham D, Iyer K, Carter CS, Schneider W, et al. Fiswidgets: a graphical computing environment for neuroimaging analysis. Neuroinformatics 2003;1(1):111–26.
[6] Skudlarski P, Constable RT, Gore JC. ROC analysis of statistical methods used in functional MRI: individual subjects. Neuroimage 1999;9(3):311–29.
[7] Rex DE, Ma JQ, Toga AW. The LONI pipeline processing environment. Neuroimage 2003;19(3):1033–48.
[8] LaConte S, Anderson J, Muley S, Ashe J, Frutiger S, Rehm K, et al. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. Neuroimage 2003;18(1):10–27.
[9] Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, et al. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage 2002;15(4):747–71.
[10] Hopfinger JB, Büchel C, Holmes AP, Friston KJ. A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. Neuroimage 2000;11(4):326–33.
[11] Tanabe J, Miller D, Tregellas J, Freedman R, Meyer FG. Comparison of detrending methods for optimal fMRI preprocessing. Neuroimage 2002;15(4):902–7.
[12] Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 2001;14(6):1370–86.
[13] Caparelli EC, Tomasi D, Arnold S, Chang L, Ernst T. *k*-Space based summary motion detection for functional magnetic resonance imaging. Neuroimage 2003;20(2):1411–8.
[14] Bannister PR, Jenkinson M. Robust affine motion correction in fMRI time series. Neuroimage 2001;13(6):S70.
[15] Bannister PR, Beckmann C, Jenkinson M. Exploratory motion analysis in fMRI using ICA. Neuroimage 2001;13(6):S69.
[16] Erb M, Hidsmann E, Klose U, Thesen S, Grodd W. Brain activation mapping of leg movement using fMRI with prospective motion correction. Neuroimage 2001;13(6):S9.
[17] Jezzard P, Matthews PM, Smith SM. Functional MRI: an introduction to methods. 1st ed. Oxford University Press; 2001.
[18] Calhoun V, Golay X, Pearlson G. Improved fMRI slice timing correction: interpolation errors and wrap around effects. Proceedings, ISMRM, 9th Annual Meeting; 2000. p. 810.
[19] Henson R, Büchel C, Josephs O, Friston K. The slice-timing problem in event-related fMRI. Neuroimage 1999;9(1):125.
[20] Friston KJ, Josephs O, Zarahn E, Holmes AP, Rouquette S, Poline JB. To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. Neuroimage 2000;12:196–208.
[21] Gavrilescu M, Shaw ME, Stuart GW, Eckersley P, Svalbe ID, Egan GF. Simulation of the effects of global normalization procedures in functional MRI. Neuroimage 2002;17(2):532–42.
[22] Macey PM, Macey KE, Kumar R, Harper RM. A method for removal of global effects from fMRI time series. Neuroimage 2004;22(1):360–6.
[23] Friston KJ, Jezzard P, Turner R. Analysis of functional MRI time-series. Hum Brain Mapp 1994;1:153–71.
[24] Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 1995;2:189–210.
[25] Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SCR, Frackowiak RSJ, et al. Analysis of fMRI time series revisited. Neuroimage 1995;2:45–53.
[26] Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited — again. Neuroimage 1995;2:173–81.
[27] Andersen A, Gash DM, Avison MJ. Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. Magn Reson Imaging 1999;17(6):795–815.
[28] Friston KJ, Phillips J, Chawla D, Buchel C. Revealing interactions among brain systems with nonlinear PCA. Hum Brain Mapp 1999;8(2–3):92–7.
[29] Friston KJ, Phillips J, Chawla D, Buchel C. Nonlinear PCA: characterizing interactions between modes of brain activity. Philos Trans R Soc Lond B Biol Sci 2000;355(1393):135–46.
[30] Hansen LK, Larsen J, Nielsen FA, Strother SC, Rostrup E, Savoy R, et al. Generalizable patterns in neuroimaging: how many principal components? Neuroimage 1999;9(5):534–44.
[31] Laiand SH, Fang M. A novel local PCA-based method for detecting activation signals in fMRI. Magn Reson Imaging 1999;17(6):827–36.
[32] Bullmore ET, Rabe-Hesketh S, Morris RG, Williams SC, Gregory L, Gray JA, et al. Functional magnetic resonance image analysis of a large-scale neurocognitive network. Neuroimage 1996;4(1):16–33.
[33] Friston KJ, Frith CD, Frackowiak RSJ, Turner R. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 1995;2(2):166–72.
[34] Worsley KJ, Poline JB, Friston KJ, Evans AC. Characterizing the response of PET and fMRI data using multivariate linear models (MLM). Neuroimage 1997;6:305–19.
[35] McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, et al. Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp 1998;6(3):160–88.
[36] Biswal BB, Ulmer JL. Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. J Comput Assist Tomogr 1999;23(2):265–71.
[37] McKeown MJ. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. Neuroimage 2000;11(1):24–35.
[38] Scarth G, McIntyre M, Wowk B, Somorjai R. Detection of novelty in functional images using fuzzy clustering. Proceedings of the Society of Magnetic Resonance Twelfth Annual Conference, Nice, France, 1; 1995. p. 238.
[39] Baumgartner R, Scarth G, Teichtmeister C, Somorjai R, Moser E. Fuzzy clustering of gradient-echo functional MRI in the human visual

cortex. Part I: reproducibility. J Magn Reson Imaging 1997;7: 1094–101.
[40] Moser E, Diemling M, Baumgartner R. Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part II: quantification. J Magn Reson Imaging 1997;7:1102–8.
[41] Jarmasz M, Somorjai RL. EROICA: exploring regions of interest with cluster analysis in large functional magnetic resonance imaging data sets. Concepts Magn Reson 2003;16A(1):50–62.
[42] Friston KJ, Frith CD, Liddle PF, Frackowiak RS. Functional connectivity: the principal component analysis of large (PET) data sets. J Cereb Blood Flow Metab 1993;13:5–14.
[43] Lin FH, McIntosh AR, Agnew JA, Eden GF, Zeffiro TA, Belliveau JW. Multivariate analysis of neuronal interactions in the generalized partial least squares framework: simulations and empirical studies. Neuroimage 2003;20(2):625–42.
[44] Beckmann CF, Smith SM. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans Med Imaging 2004;23(2):137–52.
[45] LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. Neuroimage 2005;26(2):317–29.
[46] Rowe DB, Hoffmann RG. Multivariate statistical analysis in fMRI. IEEE Eng Med Biol Mag 2006;25(2):60–4.
[47] Lee JH, Lee TW, Jolesz FA, Yoo SS. Multivariate analysis of fMRI group data using independent vector analysis. (independent component analysis and signal separation). 7th International Conference, ICA, London, UK, Proceedings; 2007.
[48] Della-Maggiore V, Chau W, Peres-Neto PR, McIntosh AR. An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data. Neuroimage 2002;17(1):19–28.
[49] Thirion B, Pinel P, Mériaux S, Roche A, Dehaene S, Poline JB. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. Neuroimage 2007;35(1):105–20.
[50] Thirion B, Pinel P, Tucholka A, Roche A, Ciuciu P, Mangin JF, et al. Structural analysis of fMRI data revisited: improving the sensitivity and reliability of fMRI group studies. IEEE Trans Med Imaging 2007;26(9):1256–69.
[51] Lange N, Strother SC, Anderson JR, Nielsen FÅ, Holmes AP, Kolenda T, et al. Plurality and resemblance in fMRI data analysis. Neuroimage 1999;10(3):282–303.
[52] Lukic AS, Wernick MN, Strother SC. An evaluation of methods for detecting brain activations from PET or fMRI images. Artif Intell Med 2002;25:69–88.
[53] Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed 1997;10:160–4.
[54] Genovese CR, Noll DC, Eddy WF. Estimating test–retest reliability in fMRI. I: statistical methodology. Magn Reson Med 1997;38: 497–507.
[55] Maitra R, Roys SR, Gullapalli RP. Test–retest reliability estimation of functional MRI data. Magn Reson Med 2002;48:62–70.
[56] Nandy RR, Cordes D. Novel ROC-type method for testing the efficiency of multivariate statistical methods in fMRI. Magn Reson Med 2003;49(6):1152–62.
[57] Mardia KV, Kent JT, Bibby JM. Multivariate analysis. San Diego: Academic Press; 1979.
[58] Strother SC, Laconte S, Hansen LK, Anderson J, Zhang J, Pulapura S, et al. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics. Neuroimage 2004;23 (Suppl 1):196–207.
[59] Shaw ME, Strother SC, Gavrilescu M, Podzebenko K, Waites A, Watson J, et al. Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. Neuroimage 2003;19(3):988–1001.
[60] Zhang J, Liang L, Anderson J, Gatewood L, Rottenberg D, Strother S. A Java-based fMRI processing pipeline evaluation system for the assessment of GLM and CVA-based heterogeneous pipelines. Neuroinformatics 2008;6(2):123–34.
[61] Zhang J, Liang L, Anderson J, Gatewood L, Rottenberg D, Strother S. Evaluation and comparison of GLM- and CVA-based fMRI processing pipelines with Java-based fMRI processing pipeline evaluation system. Neuroimage 2008;41(4):1242–52.
[62] Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance Neuroimages. Comput Biomed Res 1996;29: 162–73.
[63] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 2004;23 (Suppl 1):208–19.
[64] Mitra PP, Pesaran B. Analysis of dynamic brain imaging data. Biophys J 1999;76(2):691–708.
[65] Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assist Tomogr 1998;22(1):139–52.
[66] Woods RP, Grafton ST, Watson JD, Sicotte NL, Mazziotta JC. Automated image registration: II. Intersubject validation of linear and nonlinear models. J Comput Assist Tomogr 1998;22(1):153–65.
[67] Conover WJ. Practical nonparametric statistics. New York: Wiley; 1999. p. 367–73.
[68] Egolf E. To slice or not to slice (timing). http://www.nrc-iol.org/cores/ccnlab/int_reports/slice_or_not.htm.
[69] Kustra R, Strother SC. Penalized discriminant analysis of [15O]water PET brain images with prediction error selection of smoothing and regularization hyperparameters. IEEE Trans Med Imaging 2001;20: 376–87.
[70] Shaw ME, Strother SC, McFarlane AC, Morris P, Anderson J, Clark CR, et al. Abnormal functional connectivity in posttraumatic stress disorder. Neuroimage 2002;15(3):661–74.
[71] Lund TE, Nørgaard MD, Rostrup E, Rowe JB, Polson OB. Motion or activity: their role in intra- and inter-subject variation in fMRI. Neuroimage 2005;26(3):960–4.
[72] Morgan VL, Dawant BM, Li Y, Pickens DR. Comparison of fMRI statistical software packages and strategies for analysis of images containing random and stimulus-correlated motion. Comput Med Imaging Graph 2007;31(6):436–46.
[73] Friman O, Borga M, Lundberg P, Knutsson H. Adaptive analysis of fMRI data. Neuroimage 2003;19(3):837–45.