The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines

Joset A. Etzel, Nikola Valchev, Christian Keysers

ABSTRACT

Multivoxel pattern analysis of functional magnetic resonance imaging (fMRI) data is continuing to increase in popularity. Like all fMRI analyses, these analyses require extensive data processing and methodological choices, but the impact of these decisions on the final results is not always known. This study explores the impact of four methodological choices on analysis outcomes and introduces the technique of partitioning on random runs for characterizing temporal dependencies and evaluating partitioning methods. The analyses were performed on two fMRI data sets, which were repeatedly analyzed with support vector machines, varying the method of temporal compression, smoothing, voxel-wise detrending, and partitioning into training and testing sets. Smoothing sometimes slightly increased classification accuracy. Partitioning other than on the runs increased classification accuracy, and the random runs technique allowed us to attribute this improvement to the increased amount of training data, rather than to bias. The impact of the temporal compression and detrending methods varied so strongly with data set that general recommendations could not be drawn. These interactions suggest that, rather than searching for a universally superior set of methodological choices, researchers must carefully consider each choice in the context of each experiment.

Introduction

Recently, classification analyses of functional magnetic resonance imaging (fMRI) data have become relatively common. These techniques, also called multivoxel pattern analysis (MVPA) or multivariate analysis, use machine learning algorithms or multivariate statistics to identify patterns across multiple voxels corresponding to mental states (Haynes and Rees, 2006; Norman et al., 2006; O’Toole et al., 2007). A few examples of the wide range of possible applications include reconstruction of viewed images (Miyawaki et al., 2008), identification of the category and identity of viewed objects (Shinkareva et al., 2008), and comparing distributed representations in subjects with and without schizophrenia (Yoon et al., 2008). Reflecting the increasing use and wide array of techniques, several introductions to these methods have appeared recently (Etzel et al., 2009; Mur et al., 2009; Pereira et al., 2009).

The multiple papers describing how to perform classification analyses reflect that, like all fMRI data analyses, these methods are highly complex and involve many decisions, from experiment design and scanning parameters through to data preprocessing and cross-validation methods. This complexity creates a situation in which it is impossible to fully evaluate the impact of each choice on every analysis, so the experimenter must rely on theory, prior experience, and educated guesses. Some recommended practices arise from firm foundations, such as the need to keep the training and testing data independent; or from experimental hypotheses, such as whether to perform the analysis within-subjects or between-subjects. In other cases, however, such as the recommendation that images not be smoothed, the common practice appears to derive primarily from intuition and convention. The impact of other choices are simply unknown, such as whether more accurate results will derive from performing temporal compression (creating a single image from the multiple images corresponding to an experimental event) by fitting a hemodynamic response function or by averaging a time-shifted set of images.

Many previous investigations of the influence of processing choices on analysis results have been carried out. One area of focus is the evaluation of different classification algorithms and multivariate statistics on fMRI data sets, in an attempt to determine classes of well-performing algorithms (Chu, 2009; Ku et al., 2008; Mitchell et al., 2004; Mourao-Miranda et al., 2005; Schmah et al., 2009). Another area explores dimensionality reduction: methods of choosing voxels to include in an analysis, or to transform the voxels to a lower-dimensional space (e.g. De Martino et al., 2008; Mourao-Miranda et al., 2006; Pereira, 2007). Yet other work investigates image processing decisions fundamental to all fMRI analyses, such as the influence of...
alignment and slice timing methods (e.g. LaConte et al., 2003; Strother, 2006; Tanabe et al., 2002). Mourao-Miranda et al. (2006) compared the accuracies obtained when classifying individual volumes against temporally compressing the volumes within each block. LaConte et al. (2005) concentrated on fMRI data analysis with support vector machines (SVMs), comparing (among other factors) the results obtained with different kernels and algorithm parameter values, as well as two processing factors: smoothing and detrending. Running the analyses on a single data set (classifying whether the image was from a rest or task period; the task was an isometric grasp), they concluded that the performance was relatively insensitive to smoothing, but decreased with detrending.

This paper evaluates on the impact of four data processing decisions on SVM analyses of fMRI data by comparing the results of different choices on the outcome of analysis of two fMRI data sets. The first processing step considered was temporal compression method: was the data for each event temporally compressed using averaging or by fitting a model (creating parameter estimate images, PEIs)? The second decision was smoothing: were the images smoothed or left unsmoothed? Third was detrending method: was the data uncorrected, mean-subtracted, or linear-detrended? Finally, the partitioning method used for SVM cross-validation was varied: splitting the data on the runs or by leaving out two examples of each class. Additionally, this paper introduces the technique of partitioning on random runs, used to evaluate the presence of temporal dependencies in the partitioning structure.

Materials and methods

Two data sets were analyzed for these comparisons. The data sets were drawn from two separate unpublished fMRI experiments on human motion perception and control. Feature selection in both data sets was done by anatomical regions of interest (ROIs), including all voxels in each ROI. From previous experience with these types of tasks it was expected that classification will be possible in the chosen anatomical ROIs in both data sets. Different subjects, stimuli, and scanning parameters were used in each study, although both were carried out on the same scanner (3T Phillips). The relevant aspects of each data set will be described briefly. Notably, the first data set employs sparse sampling and a block design, while the second continuous sampling and an event-related design.

First data set

The first data set is composed of images collected while subjects viewed movies of actors performing simple hand actions (e.g. putting a CD into a player) and control movies in which the actors kept their hands flat on the table beside the props (e.g. with the CD and player sitting on the table). The analysis here consists of classifying the type of movie the subject was viewing (action or control), using five ROIs: primary motor cortex (M1), premotor cortex (preM), primary somatosensory cortex (S1), secondary somatosensory cortex (S2), and visual cortex (visual), all in the left hemisphere. Sixteen healthy males ranging in age from 19 to 33 years (mean = 22.6, sd = 3.9) participated in the experiment, which was approved by the Ethical Review Committee of the University Medical Center Groningen and conducted in accordance with the Declaration of Helsinki. One subject was excluded from analysis due to excessive movement during scanning.

Five different movies of each type (action or control) were used, presented in blocks of three stimuli of the same type. The stimuli were randomly presented in a sparse block design with three runs and eight blocks of each movie type in each run; consecutive blocks were separated by three TR (about 16 s). A sparse sampling acquisition was used because blocks of sound stimuli were also included, randomly mixed with the movie stimuli blocks within each run, though the sound stimuli will not be discussed here. Movies were presented in the 4-s silent interval between the acquisition of volumes (acquisition time 1.5 s), using a T2* weighted acquisition at 3 T (TE = 30 ms, TR = 5.3 s, TA = 1.3 s, 25 axial slices, 4.5 mm thick, 3.5 x 3.5 mm in-plane resolution). 24 blocks of each type were presented in total. The subjects performed an odd-ball detection task, so blocks containing an odd-ball or button push (i.e. erroneous odd-ball detection) were omitted from analysis and the number of usable blocks per subject ranged from 21 to 24.

Second data set

The second data set is composed of images collected while subjects performed object-directed actions with the right hand, and here we will classify the action being performed: squeezing or rotating. A continuous sampling acquisition was used for this study (TE = 28 ms, TR = 1 s, 15 slices, 4.0 mm thick, 2.29 x 2.29 mm in-plane resolution). Partial brain volumes focused on the left hemisphere premotor cortex were acquired; due to this the visual cortex anatomical ROI included in the previous data set is not included here; the other four (M1, preM, S1, S2) are the same. Seven healthy undergraduates (5 females, 2 males) ranging in age from 20 to 25 years (mean = 21.9, sd = 1.6) participated in the experiment, which was approved by the Ethical Review Committee of the University Medical Center Groningen and conducted in accordance with the Declaration of Helsinki.

During each trial the subject watched a visual cue (grey circle) which enlarged and turned pink to indicate the subject should grasp the object and begin the rotating action, or blue to indicate the subject should grasp the object and begin the squeezing action. The circle shrank to indicate the remaining time left in the trial, returning to grey to cue the subject to stop the action and release the object. The actions were carried out in individual events (action period of 4 s, followed by 10 s of rest), separated into two runs (squeezing and rotating randomly alternated within each run). The event timing was temporally aligned (“time-locked”) to the volume acquisition. The two runs were separated by at least two runs of other experimental conditions (different actions and objects). Each action was repeated 8 times in each run, for 16 repetitions total, with trials which were not performed correctly (e.g. the object was dropped) omitted from analysis, leaving 13 to 16 usable examples of each action per subject.

Image preprocessing

There were differences in the image preprocessing between the data sets, some of which derived from the different sampling (sparse vs. continuous), design (block vs. event-related), and volume acquisition (full vs. partial, voxel sizes) used in the two data sets. This paper is concerned with characterizing the effects of certain processing choices on the outcome of classification analyses, so differences between the data sets are valuable for probing the sensitivity of the outcomes to data set characteristics. However, since only two data sets are included in the current work it is not possible to fully explore the influence of all factors or confidently assign differences in outcome between the data sets to specific data set characteristics.

Preprocessing of both data sets was performed using SPM (Wellcome Department of Imaging Neuroscience, London, UK), version 2 for the first data set and version 5 for the second. In both cases the volumes were first realigned. The images from the first data set were normalized; normalization allows the same voxels to be compared in all subjects, necessary for between-subjects analyses, but potentially distorting activation patterns. The second data set was not normalized, instead the anatomical ROI masks were reverse-normalized into subject space and analyses were performed in subject space. Specifically, high resolution T1 images were coregistered with the mean EPI image and segmented. Normalization parameters were then
estimated to transform the gray matter segment onto the MNI gray-matter segment. These parameters were then applied to all EPI images (first data set) or the inverse transformation was applied to the ROI masks (second data set).

In the first data set the voxels were resized to $4 \times 4 \times 4 \text{ mm}$ during normalization (acquired at $3.5 \times 3.5 \times 4.5 \text{ mm}$), while in the second data set the voxels were not resized, but kept at the acquired size of $2.29 \times 2.29 \times 4.0 \text{ mm}$. The smoothed images were created in SPM by Gaussian smoothing at $10 \times 10 \times 10 \text{ mm}$ FWHM (first data set) or $8 \times 8 \times 8 \text{ mm}$ FWHM (second data set); different FWHM were used because of the difference in voxel size in the two data sets. This amount of smoothing is somewhat large, but typical for these voxel sizes. As is common in fMRI analyses smoothing was performed before anatomical ROI identification, so border voxels will contain some information from voxels in adjacent areas. It would perhaps be considered desirable for these comparisons to have been performed on data from experiments using identical scanning parameters, since identical normalization, voxel sizes, and smoothing could have been applied (voxel sizes could be matched with the current data, but the amount of interpolation required would necessarily vary, thus not eliminating preprocessing differences). However, since one of the motivations for this work was to examine the stability of the observed interactions over different data sets, common variations in image characteristics are not necessarily detrimental.

Rather than consider individual volumes, temporal compression was used to create one summary volume per stimulus block (first data set) or event (second data set). Two common methods of temporal compression were considered here: averaging and the creation of parameter estimate images (PEIs). The averaged images were created by averaging the volumes judged, based on a canonical HRF and the scanning parameters, to contain the most signal. For the first experiment, the volumes collected immediately following the presentation of each 4-s movie were averaged (three volumes per block, due to the sparse sampling). In the second experiment the volumes collected starting 3 s and ending at 9 s after the beginning of each 4-s event were averaged. The PEIs were created in SPM by fitting a boxcar function convolved with the canonical hemodynamic response function using a GLM for each block (first data set) or event (second data set). During PEI creation a high-pass filter in the temporal domain was applied, using a frequency cutoff calculated for each subject as 10% more than the maximum time between two events of the same type for that subject. Temporal compression by PEI creation thus incorporates steps to remove linear trends from the images, which are not included during temporal compression by averaging.

The PEIs and averaged images were further processed (in R, version 2.10.1, R Development Core Team, 2009) to remove voxels with zero variance in any run. Some voxels were removed in one set of images (e.g. the smoothed PEIs) but not the others. To be able to compare the results across image type (smoothed or not, PEIs or averaged images) an intersected data set was created for each preprocessing combination, containing only those voxels present in all four types of images. Since the first data set was normalized, zero-variance voxels in any subject were removed from all subjects, allowing analyses to be performed using the same voxels in all subjects.

The maximum probability maps based on the probabilistic cytoarchitectonic maps from the SPM Anatomy Toolbox (Eickhoff et al., 2006; Eickhoff et al., 2005) were used to create the anatomical ROIs. The SPM Anatomy Toolbox areas used for each ROI are listed in Table 1, as well as the final number of voxels in each. All voxels in each ROI were included in the classifications; no further feature selection was performed.

Two voxel-wise detrending methods were compared, on both the averaged images and the PEIs: linear detrending and mean-subtraction. The mean-subtraction was performed by subtracting the mean for each run from each voxel, while linear detrending kept the residuals from fitting a line to the values for each voxel over each run. The images before detrending were also analyzed ("none"). Note that this detrending was performed after averaging or PEI creation; the temporal filtering and model fitting during PEI creation will already have detrended these images to some degree. Accordingly, in the "mean-subtraction" or "linear detrending" conditions the PEIs will have been subjected to three detrending stages (temporal filtering, model fitting, mean-subtraction or linear detrending), while the averaged images only one (mean-subtraction or linear detrending). Since the PEI creation procedure was not varied, conclusions can only be made about the influence of the last detrending stage.

### Classification

Analyses were performed within-subjects, fitting a support vector machine classifier to each subject individually and averaging the accuracies across subjects. The classifications were performed in R using the svm command in the e1071 package with a linear kernel, the cost parameter fixed at 1, and default scaling (both voxels and examples standardized to zero mean and unit variance; standardization parameters determined on training data only), common choices for fMRI data. Additionally, between-subjects analyses were performed on the first data set, and appear in the Supplementary Information.

Partitioning the data into training and testing sets is a necessary step in classification analyses, but requires particular attention when working with fMRI data, due to the numerous unavoidable temporal dependencies and typical organization into runs and events (Etzel et al., 2009; Mitchell et al., 2004; Mur et al., 2009; Pereira et al., 2009). Temporal dependencies arise from (among other causes) the slow speed of the BOLD response, scanner drift, and subject movement. If these factors were ignored large distortions in classifier performance could occur. For example, if the individual images making up a block were considered separately, and some images from a block were put into the training set, but others into the testing set, we would expect that images from within the same block would be more similar than images from other blocks. The classifier could then learn this similarity, inflate accuracy, or the similarity could mask true patterns in the images. Temporal dependencies can be reduced by careful experimental design and temporal compression, but it is not easy to determine if these methods were successful.

Two common methods of partitioning the data were used, with an additional method added to evaluate possible distortions from temporal dependencies. The first method was partitioning on the runs, in which each run is used as the testing set in turn, training on the remaining runs (i.e. “leave one run out”). This is considered the “baseline” method: the method with the least risk of temporal dependencies inflating accuracy, but also the most conservative.

The second partitioning method is leave-two-out, in which two examples of each class are randomly selected to serve as the test set, with the remaining examples the training data, ignoring run labels. Note that individual examples (temporally compressed images) are

<table>
<thead>
<tr>
<th>ROI</th>
<th>Anatomy Toolbox areas</th>
<th>Number of voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>BA 4a, BA 4p</td>
<td>156</td>
</tr>
<tr>
<td>preM</td>
<td>BA 4A, BA 6</td>
<td>449</td>
</tr>
<tr>
<td>S1</td>
<td>BA 1, BA 2, BA 3a, BA 3b</td>
<td>291</td>
</tr>
<tr>
<td>S2</td>
<td>OP1, OP2, OP3, P4</td>
<td>165</td>
</tr>
<tr>
<td>Visual</td>
<td>BA17, BA18, HOC5</td>
<td>376</td>
</tr>
</tbody>
</table>

Since the second data set was not normalized the number of voxels in each ROI varied for each subject. The minimum and maximum voxel count of the subjects is therefore reported.

### Table 1

Number of voxels and the SPM Anatomy Toolbox areas used to create each ROI.
the unit for partitioning in this method, not entire runs: examples from a single run will appear in both the training and testing sets. In many cases the number of examples is low or the experiment design otherwise makes it undesirable to partition on the runs. This method creates the largest training sets, and so is often considered when the number of examples is small. However, the increase in training set size comes with an increased risk of distortion from temporal dependencies (from ignoring the run structure). Also, if the test sets are too small (e.g. one or two examples of each class) there is the risk that the significance will be reduced because of increased variance (Golland and Fischl, 2003; Mukherjee et al., 2003). Leave-two-out partitioning was chosen since more than ten examples of each class were available in both data sets; the tests sets could also have been made with one or three examples.

The final partitioning method is “random runs,” a new method designed to evaluate the extent to which temporal dependencies remain in the data, such that including examples collected near in time in the training and testing sets will affect performance. Random runs partitioning ignores the run structure (like leave-two-out) but keeps the same number of examples in the training and testing sets as partitioning on the runs, thus providing diagnostic information. The random runs partitioning is done by a stratified permutation of the run labels: the run labels are permuted within each class. In this case the random runs partitioning method is similar to splitting the data into thirds (for the first data set) or halves (for the second data set) at random (while maintaining class structure). If the accuracy with random runs partitioning is substantially greater than partitioning on the true runs it can be assumed that temporal dependencies (or other distortions) remain in the data. In this situation the image preprocessing needs to be reassessed and partitioning methods should consider the run structure. However, if the accuracies are similar a method which ignores the run structure, such as leave-two-out, can be considered.

The training and testing data sets were balanced when necessary by removing examples at random (i.e. to ensure equal numbers of each class). If removing examples was required the analysis was repeated ten times and the results averaged; the partitioning methods relying on randomness (random runs and leave-two-out) were also repeated ten times and the results averaged.

Statistical testing

Mixed models (linear mixed-effects models) were used for the statistical testing. Given the complexity of these analyses and comparisons, the testing will be described in detail; additional descriptions are also provided in the Supplementary Information. The statistical testing is somewhat unusual because of the nature of the comparisons being made: the treatments (processing techniques) were not applied at the subject level but rather at the data level; the data for each subject was smoothed, partitioned, detrended, and temporally compressed, and it is performance at the data level (the accuracy) which we wish to compare. The experiments most similar to the structure and hypotheses of the current work come from the machine learning literature. The field of machine learning is concerned with (among other topics) finding the most efficient and accurate classification algorithms, which requires methods to evaluate algorithms. Statistical difficulties can arise because the observations are not independent when multiple classification algorithms are evaluated on the same data (Dietterich, 1998; Salzberg, 1997), and creating the testing and training sets introduces an additional source of variance (Nadeau and Bengio, 2003).

Demšar (2006) discusses the evaluation of multiple classifiers on multiple sets of data; a situation similar to the present structure and hypotheses (see also García et al., 2010; García and Herrera, 2008). The strategy he recommends, and which is used in the current work, is to treat the datasets as the experimental unit: a single measure of performance (e.g. accuracy) is obtained with each algorithm on each set, and statistics are performed on these accuracies. This avoids some troublesome sources of shared variance, assuming the datasets themselves are independent. In this case the datasets (in the sense of (Demšar, 2006)) are the voxels making up the ROIs within each subject, because the treatments (processing techniques) were applied at this level. While different subjects are independent, the ROIs within each subject are not. We therefore have a nested data structure: ROIs within subject, subjects within data set (in the sense of this paper: the fMRI experiment).

Mixed models were used to perform the statistical tests, a technique which allows the data grouping structure and heteroscedasticity to be specified. All mixed models were calculated using the nlme package (version 3.1-96) in R (Pinheiro and Bates, 2000). The data were summarized such that a single accuracy was listed for each ROI, ROI, and combination of processing factors (smoothing, temporal compression, detrending, and partitioning). The ROIs were recoded as nine different factors (five from the first data set, four from the second), instead of keeping the true ROI names, since four of the ROIs were present in both data sets but it was not expected that the ROI effect would be similar or meaningful across both experiments; differences in accuracy between ROIs was not of interest. Reported $p$ and $F$ values were calculated by Wald tests for the terms in the model, where a marginal sum of squares (each factor against the whole model) was calculated. Significance was taken as $p < 0.05$ within each model. 95% confidence intervals of the estimates were calculated using the confint command in the R multcomp package (version 1.1-6), and adjusted $p$-values are reported and shown when simultaneous intervals were made.

The analytical strategy was to first construct an overall model containing all the factors, followed by tests on subsets of the data to characterize the relevant effects, due to the numerous significant interactions. The overall models are discussed and presented in the Supplementary Information, as are full results of the analyses described below.

Results

The average classification accuracy obtained in each ROI for each manipulation appears graphically in Fig. 1 and as numerical values in Tables S2–S4 in the Supplementary Information. The results of statistical testing are summarized here, but shown and described in greater detail in the Supplementary Information. The results were somewhat robust to the processing techniques assessed here in that the experimental interpretations would be similar: in the first data set accuracy was highest in S1 and visual, middle in preM and M1, and lowest in S2, while in the second data set the accuracy was more similar in the ROIs, though often highest in preM and S1. The different manipulations had an impact, however, and will be described for each data processing treatment in turn.

Temporal compression method: by averaged images or PEIs?

The impact of temporal compression method can be assessed by comparing the first six and last six columns of each table in Fig. 1. In the models incorporating all factors temporal compression was present in significant interactions with partitioning, detrending, and data set, as well as smoothing in the first data set. Mixed models were constructed to characterize the effect of temporal compression within each combination of interacting factors. The results appear in Fig. 2, Tables S12–S15, and Figs. S1–S2.

Unsurprisingly, the largest effects, with PEIs significantly more accurate than averaged images, occurred in both data sets for non-detrended images partitioned on the runs (note that this “non-detrended” in the sense of the detrending applied here; some linear trends were removed during PEI creation). This is a special situation in
which averaged images are expected to perform quite poorly, and so is not a helpful case for determining which form of temporal compression produced the best results in typical conditions. Disregarding the non-detrended images partitioned on the runs, the most striking effect is the interaction with data set: PEIs were either significantly more accurate than averaged images or indistinguishable on the first data set, but significantly less accurate than averaged images on the second.

Unusual comparisons are being made here (e.g. mean-subtracted PEIs against mean-subtracted averaged images); most often averaged images are detrended, while PEIs are not (other than the detrending and/or temporal filtering employed during PEI creation). Models were therefore constructed to compare the accuracies of linearly-detrended averaged images and non-detrended PEIs, both unsmoothed, for partitioning on the runs and by leaving out two of each class (Fig. 2). Consistent with the previous conclusion, there was a strong interaction with data set: these PEIs were more accurate than these averaged images in the first data set (sparse sampling block design, normalized images), but not the second (continuous sampling event-related design), with the strongest effect in leave-two-out partitioning.

**Smoothing: were the data smoothed during preprocessing?**

The impact of smoothing can be assessed by comparing the first group of three columns to the second group of three columns (for averaged images) and the third group of three columns to the fourth (for PEIs) of each table in Fig. 1. In the models incorporating all factors smoothing was present in significant interactions with temporal compression, detrending, and data set, but not partitioning, so mixed models were constructed to characterize the effect of smoothing within each combination of detrending, temporal compression, and data set. The results appear in Tables S16 and S17, and the direction and magnitude of the effects can be seen in Fig. 3 and Fig. S3.

**Fig. 1.** Average classification accuracy for each manipulation; see Tables S2 to S4 in the Supplementary Information for numerical values. Each pane has the results for a different partitioning method: (a) partitioning on the runs; (b) partitioning on random runs; c. partitioning by leaving out two examples of each class. All figures use the same color scale to indicate mean accuracy (proportion correctly classified). The data in each figure are grouped by image type, smoothing, and detrending (columns), with the ROIs within each data set in the rows. See the text for a full description of these manipulations.

**Fig. 2.** Temporal compression manipulation. Results of the models contrasting non-detrended PEIs against linear-detrended averaged images (unsmoothed images), separately for each data set, for partitioning on the runs and by leaving two out. 95% confidence intervals for the contrast PEI-averaged images = 0 are shown; an estimate above zero indicates that the PEIs produced higher accuracies than the averaged images, below zero, that the averaged images produced higher accuracies. See the Supplementary Information for the other treatment combinations, numerical results, and additional discussion.
As expected from the many significant interactions between smoothing, data set, and temporal compression, a simple description of the impact of smoothing on accuracy cannot be made. It can be observed that the magnitude of the effect is modest here in all cases: the significant effects are estimated with a mean impact on accuracy of $+0.05$ to $-0.06$. The effect of smoothing is consistent in the averaged images and PEIs for the first data set: smoothing produced a significant increase in accuracy for the mean-subtracted and linear-detrended images, but had no significant effect on the non-detrended images. The detrended averaged images also showed a significant increase in accuracy with smoothing in the second data set, but the detrended PEIs a significant decrease.

Considering only the most common processing combinations (non-detrended PEIs, mean-subtracted or linear-detrended averaged images; see Fig. 3), a conclusion can be made: smoothing either led to a small increase in accuracy or had no significant effect, even though these are within-subjects analyses. However, the interactions with data set and temporal compression method suggest that additional factors likely also influence the effect of smoothing.

**Detrending method: are the data uncorrected, mean-subtracted, or linear-detrended?**

The impact of the detrending method can be seen by comparing the columns of the tables in Fig. 1 within each image and smoothing type. In the models incorporating all factors detrending was found in significant interactions with temporal compression, partitioning, and data set, but not smoothing. Mixed models accordingly were constructed to characterize the effect of detrending within each combination of temporal compression, partitioning, and data set, the results of which appear in Fig. 4, Tables S18–S19, and Fig. S4.

Focusing first on the averaged images partitioned on the runs, it was expected that detrending would have a large impact on accuracy, given the recognized problem of shifting baselines between (and sometimes during) runs, which temporal compression by averaging does not eliminate. This was found in both data sets: the largest improvements in accuracy of any treatment combination were seen here. While there was a trend towards linear detrending producing higher accuracy than mean subtraction in both data sets, this difference was not significant. The results are slightly less consistent between the data sets for the averaged images when partitioned on random runs or by leaving out two of each class. Mean-subtraction always resulted in significantly higher accuracies than no detrending, as did linear detrending in all but the first data set when leaving out two of each class. Linear detrending produced significantly higher accuracies than mean-subtraction in the second data set, but significantly lower in the first. Based on these results we can confirm that detrending is required when temporal compression is performed by averaging, and perhaps lean towards preferring linear detrending to mean subtraction, with the caveat that linear detrending is not always superior.
Detrending beyond that performed when creating PEIs (i.e. GLM estimation and/or temporal filtering) is not often used, but was included here. If all linear run-related trends were removed from the PEIs during creation we would expect the additional detrending to have no effect on accuracy, or perhaps reduce accuracy (if the detrending added noise). This was not found: additional detrending produced a significant increase in accuracy in a few cases. Focusing first on partitioning on the runs, mean-subtraction produced a small significant increase in accuracy over both no detrending and linear detrending in the first data set. No differences were significant in the second data set when partitioning on the runs, although the mean estimates point in the direction of a small increase in accuracy with detrending. Contradictory impacts of detrending the PEIs were found between the first and second data sets when partitioning on random runs or by leaving out two of each class: detrending produced lower accuracies in the first, but higher in the second. The reason for this discrepancy cannot be determined from the current study, but may be related to the different designs of the two experiments (e.g. block vs. event-related, normalized or not). These results show that there is sometimes an advantage to performing additional detrending when temporal compression is performed by creating PEIs, but the factors modulating this advantage are not clear.

**Partitioning method: was the analysis performed by partitioning on the runs, on random runs, or by leaving out two of each class?**

The classification results are shown separated by partitioning method: partitioned on the runs in Fig. 1a, partitioned on random runs in Fig. 1b; and partitioned by leaving out two examples of each class in Fig. 1c. In the models incorporating all factors partitioning was present in significant interactions with temporal compression and detrending, so mixed models were constructed to characterize the effect of partitioning within each combination of detrending and temporal compression; see Fig. 5, Table S20, and Fig. S5.

As previously discussed, the non-detrended averaged images were not expected to classify well when partitioned on the runs, since differences between the images from different runs which are known to be present were not removed. Interestingly, the large increase in accuracy when non-detrended averaged images were partitioned on random runs or by leave-two-out suggests that the classifiers were able to overcome the presence of these differences when examples from all runs were present in the training data. Effects of this size were not present between partitioning methods when the averaged images were detrended, or with the PEIs. The effect of partitioning on non-detrended averaged images will not be considered further, as this processing combination is not generally used in actual analyses.

Of interest is whether partitioning on random runs is significantly more accurate than partitioning on the runs, and whether partitioning by leaving out two of each class is significantly more accurate than partitioning on random runs, since these contrasts give an indication of the amount of run-related information (temporal dependencies) remaining in the images, and whether the increase in accuracy with partitioning by leaving out two of each class is due to the increase in the amount of training data or by mixing examples from within runs. As can be seen in Figs. 5 and S5, partitioning on random runs was not significantly more accurate than partitioning on the runs for PEIs under any type of detrending, nor for averaged images with linear detrending. There is a significant increase in accuracy with partitioning on random runs over partitioning on the runs for mean-subtracted averaged images, but the magnitude of the increase is small, and partitioning by leaving out two of each class was significantly more accurate still. This pattern of results is consistent with the idea that in this case the increase in accuracy with partitioning by leaving two out is due to the increase in training data, rather than due to mixing examples across runs.

**Fig. 5.** Partitioning manipulation: 95% confidence intervals for the estimates of the contrasts for the effect of partitioning within detrending and temporal compression. Only the results of the detrended averaged images (first two columns) and non-detrended PEIs (third column) are shown here; see the Supplementary Information for the full results. Three different contrasts (linear hypotheses) of partitioning methods appear in each graph pane: random runs—true runs=0 (rdm-run); leave-two-out—runs=0 (l2o-run); and leave-two-out-random runs=0 (l2o-rdm). In all cases an estimate above zero indicates that the first partitioning method of the particular contrast produced higher accuracies, below, that the second produced higher accuracies.

**Discussion**

This paper evaluates the impact of four data processing decisions on the outcome of SVM analyses of fMRI data: the temporal compression method (by averaging or by fitting a hemodynamic response model to create PEIs), smoothing (smoothed or not), detrending method (uncorrected, mean-subtracted, or linear-detrended), and partitioning method (on the runs, on random runs, or by leaving out two examples of each class). Unfortunately, definitive guidelines for all of these factors cannot be drawn from the current work. Instead, these results underscore the complexity of these decisions, and the need to carefully evaluate each step in every situation. The results of each manipulation will be discussed in turn.

It was expected that PEIs would produce markedly better results than averaged images, particularly in the continuously sampled data (second data set), since PEI creation considers the shape of the hemodynamic response. This expected advantage did not occur. Instead, PEIs resulted in lower accuracy levels than (detrended) averaged images on the second data set, but higher accuracy levels than averaged images in the first (sparse sampling, normalized images). This striking difference between the data sets suggests that neither temporal compression method is best in general: PEI creation sometimes produces worse results than averaging. Further research on this topic is needed, particularly considering that whether, and how, to conduct temporal compression is a critical step in multivariate analyses of fMRI data. Whether averaging or PEI creation produces the best result is likely dependent upon the characteristics of the individual experiment; important factors probably include stimulus
length, sampling rate, and jitter in image acquisition timing. The optimum time periods to use for averaging are also unknown.

The lack of significantly decreased performance with smoothed data was initially surprising, as smoothing is generally advised against for multivariate analyses, particularly for within-subjects analyses, as it was thought to blur the patterns detected by multivoxel methods (e.g. Haynes and Rees, 2006; Mur et al., 2009). We are aware of several other reports of smoothing improving (or not degrading) classification performance (Chu, 2009; LaConte et al., 2005; Op de Beeck, 2010a, b; Op de Beeck et al., 2008), suggesting that this pattern may be seen in many multivariate analyses. The current work contributes an additional smoothing comparison: smoothing was applied with a FWHM of 2.5 or 3 voxels, within-subjects analyses were used, and the classification likely relied on relatively subtle activation patterns (as two types of movies or hand actions were distinguished).

In the current analyses smoothing likely reduced the amount of high spatial frequency information (patterns which change in a short distance, such as over adjacent voxels) available to the classifiers, but accuracy was not significantly reduced. Invertibility of the smoothing, as described by Kamitani and Sawahata (2010), is not likely to be an explanation, since linear support vector machines were used for classification and smoothing was applied prior to anatomical ROI selection. A recent theoretical proposal by Kriegeskorte et al. (2010) may provide an explanation: sub-voxel scale information may be present in the low spatial frequencies surviving smoothing if the voxel acts as a complex spatiotemporal filter. We conclude that smoothing should be considered when preprocessing data for a multivariate analysis. Further investigations are needed to determine for which experiments smoothing might be beneficial and the optimum amount of smoothing for particular analyses and designs, as different algorithms and dimension reduction strategies will likely influence the optimum amount of smoothing.

It is widely recognized that detrending is required when conducting analyses of fMRI data, given the typical signal drift both within and between runs. Some detrending is performed when temporal compression is done by creating PEIs, both by the model fitting itself and by the temporal filtering often included in the image preprocessing. Temporal compression by averaging (or other methods, such as analyzing single images) does not include detrending, so detrending is performed on the compressed images. Two methods of detrending were compared here: mean-subtraction and linear detrending, on both the averaged images and the PEIs.

As expected, detrending produced significant improvements in accuracy for the averaged images. The amount of improvement and which method was superior varied with data set and partitioning method, but either method is superior to no detrending. This conclusion is consistent with previous investigations of the influence of detrending: LaConte et al. (2005) found that svm analysis results were fairly robust to different levels of detrending, and Tanabe et al. (2002) (in the context of mass-univariate analysis) found that while correcting for mean signal differences between runs was essential (i.e. mean-subtraction), no other detrending procedure was clearly superior in general.

Additional detrending is not usually applied when temporal compression is performed by creating PEIs, but produced a significant increase in accuracy in some cases in the current comparisons, interacting with data set and partitioning method. When partitioning by leaving two out, linear detrending led to more than a 0.1 accuracy increase compared to no additional detrending in the second data set, but a 0.05 decrease in the first. In contrast, the results were similar between the data sets when partitioning on the runs: a small increase in accuracy with detrending. Like the choice of temporal compression method itself, these interactions between data set suggest that whether additional detrending should be applied to PEIs is still an open question.

Partitioning on the runs is sometimes preferred for classification-based fMRI analyses, given the well-founded concern that having examples from the same run in both the training and testing sets may inflate the performance, given the changes over time unavoidable in fMRI studies (e.g. from scanner drift and subject movement). However, partitioning on the runs is not always practical, such as when few runs or examples are present. Partitioning on random runs, as presented here, allows characterization of the amount of inflation that occurs from mixing examples in each experiment, and so provides an indication of when partitioning other than on the runs is valid. Specifically, if partitioning on random runs leads to a large increase in performance compared with partitioning on the runs, then it is likely that temporal bias is present in the data. In this case the data processing steps should be reexamined, to determine if other methods of temporal compression or detrending can reduce the temporal dependencies. However, if partitioning on random runs does not lead to an increase in performance, partitioning methods which ignore the run structure (such as leaving out two examples of each class) can be utilized, and resultant increases in performance be attributed to the increase in the number of training examples.

At the present time we are unfortunately not able to determine with complete certainty which set of processing decisions are best for a particular experiment before starting an analysis. Multivariate analyses of fMRI data require many decisions that were not compared here, just a few of which are the standardization method (e.g. voxels within runs only, within examples only, none), classification algorithm, and dimension reduction method. The included factors can also be performed in additional ways: smoothing at a higher or lower FWHM, temporal compression by averaging can be done on different time windows, PEIs can be made with different hemodynamic response functions or fitting methods, individual volumes could be analyzed.

It can be tempting to try a large number of processing choices (e.g. detrending method, classification algorithm, standardization method, feature selection method) on a single data set, in the hope of maximizing classification accuracy. This strategy should be viewed with great caution, since as the number of treatments and analyses increases the potential for spurious results also increases. The processing decisions should be made based on experiment characteristics, and restricted to the fewest number of factors possible. For example, if an experiment is interested in events shorter than a TR with jittered timing, temporal compression by PEIs (or a similar weighting method) should be chosen, as identifying individual volumes with equivalent signal would be impossible. If a partitioning scheme ignoring run structure is planned, partitioning on the runs and on random runs should be performed first, if possible, to evaluate temporal dependencies. However, some decisions, such as detrending method, cannot always be determined a priori. In these cases the identified different methods should be tried on a subset of the data (rather than the entire data set) whenever possible, to avoid overfitting the data and producing incorrect results.

The current work, evaluating the impact of several methodological decisions on multivariate analyses of fMRI data, unfortunately does not provide definitive answers, even on all of the factors considered. Instead, it suggests that the impact of some factors (such as the temporal compression method) may interact so strongly with the characteristics of each particular data set that a generally superior method does not exist. The interaction of processing choices and data set challenges scientists performing these analyses to characterize the methods more fully, to be able to understand the data set attributes driving the interactions. New (and existing) analytical methods should be applied to as many data sets as possible, and an attempt made to understand the underlying dynamics. At the current time, when planning a multivariate analysis we recommend that the methodological strategies most suitable for the particular design and goals of each experiment be carefully considered. Exploration of the impact of methodological choices on the analysis outcome may be necessary when principled a priori decisions are impossible.
Acknowledgments

The research was supported by a Nederlandse Organisatie voor Wetenschappelijk Onderzoek Vidi grant and a Marie Curie Excellence grant to Christian Keysers. Nikola Valchek was supported by the Foundation for Science and Technology of the Portuguese Ministry for Higher Education and Technology (SFRH/BD/47576/2008). Computer time was provided by the Centre for High-Performance Computing and Visualisation at the University of Groningen (The Netherlands) on the HPC Cluster.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2010.08.050.

References