ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 26 (2005) 317 – 329

Support vector machines for temporal classification of block design fMRI data

Stephen LaConte,[a] Stephen Strother,[b] Vladimir Cherkassky,[c] Jon Anderson,[b] and Xiaoping Hu[a,*]

[a]*Biomedical Engineering, Georgia Institute of Technology, Emory University, 531 Asbury Circle, Suite N305, Atlanta, GA 30322, USA*
[b]*Biomedical Engineering, University of Minnesota, Minneapolis, MN 55455, USA*
[c]*Electrical Engineering, University of Minnesota, Minneapolis, MN 55455, USA*

Received 23 December 2003; revised 10 January 2005; accepted 31 January 2005
Available online 24 March 2005

This paper treats support vector machine (SVM) classification applied to block design fMRI, extending our previous work with linear discriminant analysis [LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., Strother, S., 2003a. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. NeuroImage 18, 10–27; Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Siditis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15, 747–771]. We compare SVM to canonical variates analysis (CVA) by examining the relative sensitivity of each method to ten combinations of preprocessing choices consisting of spatial smoothing, temporal detrending, and motion correction. Important to the discussion are the issues of classification performance, model interpretation, and validation in the context of fMRI. As the SVM has many unique properties, we examine the interpretation of support vector models with respect to neuroimaging data. We propose four methods for extracting activation maps from SVM models, and we examine one of these in detail. For both CVA and SVM, we have classified individual time samples of whole brain data, with TRs of roughly 4 s, thirty slices, and nearly 30,000 brain voxels, with no averaging of scans or prior feature selection.

Keywords: Support vector machine; Functional magnetic resonance imaging; Canonical variates analysis

Introduction

The advent of functional magnetic resonance imaging (fMRI) in the early 1990s has provided a revolutionary means for non-invasively probing spatiotemporal variations of brain function. For whole brain studies, the number of acquired brain voxels can be in the tens of thousands, which are sampled in time to acquire hundreds of measurements. The flexibility of this technique in terms of experimental design and data analysis is virtually limitless. As has been stated in a variety of ways by several researchers, neuroimaging data are extremely rich in signal information and poorly characterized in terms of signal and noise structure (Cox and Savoy, 2003; Hansen et al., 2001; LaConte et al., 2003a; Lange et al., 1999; Skudlarski et al., 1999; Strother et al., 2002).

During this same period of fMRI development, advances in the interrelated fields of machine learning, data mining, and statistics have enhanced our capabilities to extract and characterize subtle features in data sets from a wide variety of scientific fields (Cherkassky and Mulier, 1998; Hastie et al., 2001; Mjolsness and DeCoste, 2001). Among these developments, support vector machines (SVMs) have been an active area of research and have been applied to a broad range of problems. SVMs arise from the Statistical Learning Theory of Vapnik (Vapnik, 1995) and possess several unique properties appropriate for real world applications, including fMRI. Among these is the fact that the formulation of SVMs was motivated to deal with small sample sizes and high dimensional inputs, which match the situation involved for temporally predictive modeling of fMRI data.

There are several reasons to consider temporal predictive modeling[1] for fMRI data analysis. First, as argued by Strother and Hansen (Strother et al., 2002) and mentioned in Morch et al. (1997), from a Bayesian perspective, there is no obvious mathematical advantage for choosing to estimate a spatial summary map from our knowledge of the experiment (e.g. general linear model approaches (Friston et al., 1995)) over trying to estimate these experimental parameters from our input patterns. Second, as demonstrated in LaConte et al. (2003a), Shaw et al. (2003), and Strother et al. (2002), prediction accuracy along with other model performance metrics such as spatial pattern reproducibility can be used as a data-dependent means of methodological

* Corresponding author. Fax: +1 404 727 9873.
E-mail address: xhu@bme.emory.edu (X. Hu).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.01.048

[1] Predictive modeling in this article generically refers to classification (detection) or regression (estimation) models. In most cases, however, we focus on the classification problem.

validation. Currently, the most common tool for such validation is the receiver operating characteristic (ROC) analysis (Constable et al., 1995; Hansen et al., 2001; Le and Hu, 1997; Metz, 1978; Skudlarski et al., 1999; Xiong et al., 1996), measuring a method's accuracy by comparing the true positive fraction of activated pixels against the false positive fraction varied over some modeling parameter. Unlike standard ROC analysis in neuroimaging, the approach of LaConte et al. (2003a), Shaw et al. (2003), and Strother et al. (2002) need not rely on simulations. Third, predictive modeling explicitly uses the assumption that we have more reliable knowledge about the temporal aspects of the data than the spatial activation patterns. This is the same assumption implicitly used for generating SPMs, interpreting "data-driven" results, and modeling the hemodynamic response, which follows from the fact that we designed the temporal nature of the experiment and/or simultaneously measure behavioral performance. Finally, temporal predictive modeling is a much more natural way to examine the recent interest in using fMRI for brain computer interface (BCI) and biofeedback studies (LaConte et al., 2004).

The work in predictive modeling has primarily been developed by Strother and Hansen (Hansen et al., 1999; Kjems et al., 2002; Kustra and Strother, 2001; LaConte et al., 2003a; Lautrup et al., 1994; Morch et al., 1997; Shaw et al., 2003; Strother et al., 2002) with recent explicit testing of distributed brain systems by Haxby et al. (Haxby et al., 2001) and Cox and Savoy (Cox and Savoy, 2003). The implication of the classification setting is that fMRI can be used for predicting brain states to enhance our understanding of brain systems, rather than the standard emphasis on spatial mapping.[2] Recently, Strother has introduced a formal framework in which predictive modeling plays a prominent role. This framework, termed NPAIRS for Nonparametric Prediction Activation Influence and Reproduciblity reSampling (Strother et al., 2002), provides a disciplined approach for exploring multivariate signal and noise spaces and the impact of various factors such as experimental and data analysis parameters as well as the influence of outliers on these subspaces.

The use of SVM has recently been reported in the fMRI literature (Cox and Savoy, 2003; LaConte et al., 2003b). In LaConte et al. (2003b), we dealt with the efficacy of SVM compared to CVA and discussed SVM model interpretation. Here, we greatly expand that initial study. The work of Cox and Savoy (Cox and Savoy, 2003) differs from our approach on several points. First, we are classifying individual scans, with TRs of roughly 4 s, rather than 20-s blocks. Whereas Cox and Savoy used ten classes of visual objects, we focus on the two-class problem to illustrate the important issue of visualization and interpretation of SVM models applied to fMRI. Finally, we build our SVM models based on whole brain data, without selectively choosing voxels through an initial statistical parametric mapping.

In the context of BCI-type studies such as LaConte et al. (2004) or analyses of distributed representations of sensory information (Cox and Savoy, 2003), predictive modeling may be the ultimate goal. A major impetus for performing MRI-based experiments in the first place, however, is to obtain spatially localized information. With this experimental data, one advantage of predictive modeling is that it allows for spatially distributed patterns of activation while also incorporating the temporal structure of the experiment. In other words, we are dealing with multivariate approaches. These spatial summary maps provide aid in model interpretation as well as a tangible means of comparing different models (e.g. (Hansen et al., 2001)). For the SVM, generation of these summary maps requires special consideration, and we outline four methods for doing this, demonstrating one as an example.

This article has the explicit aim of formally describing SVM classification in the application domain of fMRI analysis and to propose interpretation (mapping) strategies for this application. We give a careful description of SVM classifiers and examine more closely the selection and tuning of SVM model parameters. We then illustrate SVM classification through comparisons to CVA, a previously published technique. We do not attempt a definitive verdict on CVA vs. SVM, but rather highlight some currently known merits of both approaches within the fMRI application domain.

Theory

Here, we summarize only the salient concepts for SVM-based classification that are essential for describing its application to fMRI. For a more general treatment, please refer to Burges (1998), Cherkassky and Mulier (1998), and Muller et al. (2001). See Kjems et al. (2002), LaConte et al. (2003a), and Strother et al. (2002) for a description of CVA. In particular, we discuss the classification setting and its relevance to fMRI data analysis, the use of SVM classifiers, and interpretation of SVM results in the context of fMRI. Fig. 1 provides a graphical depiction of these topics using simulated data. The primary new contribution of this section is the discussion of SVM interpretation and activation map generation in the context of fMRI.

The classification problem

The classification problem is one of determining a scalar class label, y_t, from a measurement vector, $\vec{x}_t$ For temporal classification of fMRI data, each $\vec{x}_t$ represents all brain voxels at a given time, t of the total scan time T $(1 \leq t \leq T)$, and y_t is the experimental design value for that time (see Figs. 1A and B). For example, at each sampled repetition time (TR), we know the subject was performing task A (class = 1) or performing a control task (class = −1). Thus, an fMRI experiment consists of a series of brain images being collected while class labels are changed. In this situation, we have labeled data. For simplicity, we examine the binary classification problem ($y_t = \pm 1$).

In the general SVM formulation, the input vectors are mapped to a high dimensional feature space, $\vec{z}$, via a non-linear transformation function, $g(\cdot)$: $\vec{z} = g(\vec{x})$. In practice, g is often expressed in its dual form, termed the inner product kernel $H(\vec{x}, \vec{x}^T) = g(\vec{x})g(\vec{x}^T)$. For linear SVM, the feature space is the original input space $(\vec{z} = \vec{x})$. The SVM algorithm attempts to find a linear decision boundary (separating hyperplane) in the feature space, formalized by the decision function

$$D(\vec{z}) = (\vec{w} \cdot \vec{z}) + w_0, \tag{1}$$

where $\vec{w}$ defines the linear decision boundaries. Since the kernel can be non-linear, the decision boundaries in the original data space may also be non-linear (Fig. 1C). The solution of $\vec{w}$ satisfies

[2] Here, we alternatively refer to statistical parametric maps (SPMs), activation maps, and summary images as the spatial pattern that summarizes the interaction between the fMRI experiment, data acquisition, and data analysis.

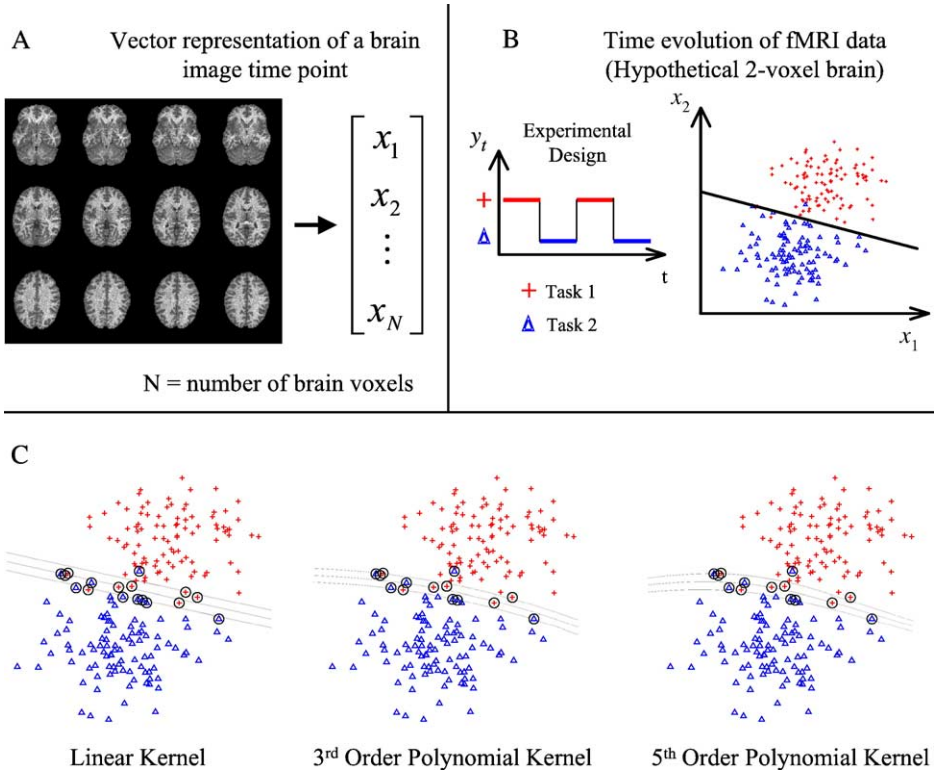

Fig. 1. Representation of fMRI data for predictive modeling: As shown in panel (A), each set of measured time images is represented as an N-dimensional vector, where N is the number of brain voxels. Panel (B) depicts the time evolution of an experiment in the first two dimensions of the input space (or for a brain consisting solely of two voxels). Panel (C) illustrates the effect of polynomial kernels of 1st, 3rd, and 5th order in the original input space. Circled data time points are support vectors. The center boundary represents a decision value $D = 0$, while the two outer lines are the margin of $D = +/- 1$.

$y_t\left[\left(\vec{w}\cdot\vec{z_t}\right)+w_0\right]\geq 1$ and is optimal when $\|\vec{w}\|^2$ is minimized under this constraint. It is also possible to introduce a *soft margin* formulation (see also next paragraph) to allow for data that cannot be separated without error and/or to permit a degree of training error to improve generalization (classification accuracy on test data that is independent of the training data) (Cortes and Vapnik, 1995). For a soft margin, slack variables, ξ, are the distance from the correct margin boundary for data points within the margin or on the wrong side of the margin. That is, their values are the error (subtraction) between the true class labels and the decision function in Eq. (1). In this case, the hyperplane is defined by $y_t\left[\left(\vec{w}\cdot\vec{z_t}\right)+w_0\right]\geq 1-\xi_t$ and is optimal under this constraint when $\frac{C}{T}\sum_{t=1}^{T}\xi_t+\frac{1}{2}\|\vec{w}\|^2$ is minimized. The free parameter, C, affects the tradeoff between complexity and number of non-separable samples. It does this by controlling the degree of compromise between having zero training error (which is only possible in the separable case) and increased flexibility permitted by the slack variables. An important result (see (Cherkassky and Mulier, 1998)) is that the solution of the optimal $\vec{w}$ is a linear combination of a subset of the training vectors, $\vec{x_t}$, termed support vectors. Moreover, a dual form of Eq. (1) exists (Cherkassky and Mulier, 1998; Vapnik, 1995):

$$D(\vec{z})=\sum_{t=1}^{T}\alpha_t y_t(\vec{z}\cdot\vec{z_t})+w_0=\sum_{t=1}^{T}\alpha_t y_t H(\vec{x},\vec{x_t})+w_0 \qquad (2)$$

where those $\alpha_t > 0$ specify this linear combination of the training (support) vectors. Thus, the trained model consists of the predefined non-linear function $g(\cdot)$, the weighted support vectors, and the support vector class labels.

The minimization of $\|\vec{w}\|^2$ translates to maximization of what is termed the *margin*, whose boundaries are defined as $\left(\vec{w}\cdot\vec{z_t}\right)+w_0=1$ and $\left(\vec{w}\cdot\vec{z_t}\right)+w_0=-1$. The minimization of $\frac{C}{T}\sum_{t=1}^{T}\xi_t+\frac{1}{2}\|\vec{w}\|^2$ permits a larger margin by allowing (penalized) errors, leading to margin boundaries $\left(\vec{w}\cdot\vec{z_t}\right)+w_0=(1-\xi_t)$ and $\left(\vec{w}\cdot\vec{z_t}\right)+w_0=-1(1-\xi_t)$. In general, maximization of the margin is desirable for better accuracy on test data, but the soft margin allows for a tradeoff between complexity and the number of training errors as described above. For the soft margin case, observations with $\xi > 0$ fall on the wrong side of the margin boundary. Any data points falling within the margin or on the wrong side of the margin are the support vectors and correspond to an $\alpha > 0$ in Eq. (2). See Fig. 1C for simulated examples of SVM models using polynomial kernels of different order with support vectors circled.

Model interpretation

Even though the development of the SVM was motivated purely by the predictive learning problem, general work in interpretation has been done by several researchers, including Smola (Smola et al., 1998) and Kwok (Kwok, 1999; Kwok, 2000) to integrate SVM into a Bayesian framework. In addition, Scholkopf has studied the general relationship between arbitrary locations in the feature space and the original input space. One important point from that work is that mapping from a general

point in the feature space to the original input space is an ill-posed problem (Scholkopf et al., 1999).

For fMRI, Fig. 1 summarizes the idea of representing a single time volume as a point in a vector space and the distribution of the time volumes for an experiment in that space. For convenience of illustration, we use a toy example and display only two dimensions (which would correspond to voxels) of the input space in Figs. 1B and C. For our application, it is important to remember that each data point in the input space (and therefore the feature space) represents a spatial pattern (the fMRI image at a certain time). Also important is that the SVM model we wish to interpret is the subset of data that defines the margin. In fact, removing non-support vector images from the training exercise would result in an identical model. From an interpretation perspective, the support vectors, including those data with $\xi > 0$, are the observations that are the most difficult to classify as they are least distinct in the feature space from members of the other class (they are at the boundaries of the classes).

With these considerations in mind, we propose four methods for obtaining summary maps from fMRI data using SVMs. The first is the direct visualization of the SVM training weight vector, $\vec{w}$, directly. This is possible when a linear kernel is used. In this case, the dimensionality of $\vec{w}$ corresponds to that of the original input data, $\vec{x}$.

The second is the use of sensitivity maps as proposed by Kjems et al. (2002). Defined as

$$s(i) = \left\langle \left[\frac{\partial p(y \mid \vec{x})}{\partial x(i)} \right]^2 \right\rangle_{p(x,y)} \tag{3}$$

and approximated over the finite number of time observations, T, as

$$s(i) \approx \frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial p(y_t \mid \vec{x}_t)}{\partial x(i)} \right]^2, \tag{4}$$

where $s(i)$ is estimated at each pixel location i. Crucial to estimating these sensitivities is an estimate of the conditional distribution. The work of Kwok (Kwok, 1999) provides the approximation

$$p_\theta(y_t \mid \vec{x}_t) \approx \exp(-\xi_t), \tag{5}$$

where θ represents the model parameterization, and the slack variables are the ξ_ts. For multiple slice data sets, sensitivity maps become computationally expensive to generate. In addition, actual sensitivity depends on the accuracy of the density estimation and its partial derivative. As stated by Kjems, however, sensitivity maps have an advantage in the fact that the approach is general enough to be applied to any model, allowing for direct comparison across methods.

We term the third method for visualizing SVM results "feature space weighting" (FSW). This approach derives from the intuition of distance from the margin being related to ease of discrimination. The simplest approach (and the one that we will demonstrate—see Methods and Results) is to perform a correlation test for every pixel time series with the reference experimental design as proposed by Bandettini (Bandettini et al., 1993). Instead of using every time point in the reference and image set, however, we exclude those times corresponding to support vectors. More generally, this approach could be adapted to use a distance measure with respect to the margin, leading to a weighted average contrast function. That is, rather than using a reference waveform with just two on/off levels, the values at each time could be weighted by some distance measure from the margin and this continuous valued waveform used for pixel-wise correlation.

We also propose a fourth approach for obtaining summary images, called decision weighting. Similar to FSW, a refined contrast function is obtained for the validation data based on model results. Variations of this approach would be to use $D(\vec{z}_t)$ from Eq. (1) directly, or fix positive and negative values to +1 and −1, respectively.

The tradeoffs and subtle differences in comparing these methods are beyond the scope of the current work. It is also quite likely that these approaches are not exhaustive. We limit our exploration of this topic to the simplified FSW compared with standard cross correlation analysis.

Methods

The SVM implementation used was SVMlight (Joachims, 1999). For I/O speed considerations, we modified this C-based software to read binary image files. CVA along with NPAIRS was implemented in IDL (RSI, Boulder, CO) and is part of the VAST software library (http://neurovia.umn.edu/incweb/npairs_info.html) at the VA Medical Center, Minneapolis, Minnesota. Visualization of SVM models was accomplished with Matlab (MathWorks, Natick, MA) and AFNI (Cox, 1996; Cox and Hyde, 1997).

Comparison of SVM with CVA

Data and CVA results reported in LaConte et al. (2003a) were used for comparison with the SVM. Sixteen right-handed volunteers performed two repeated runs of a static force task alternating between six rest and five force periods/run (45 s/period; {200, 400, 600, 800, 1000} g randomized forces with thumb and forefinger). Data were collected on a Siemens 1.5 T scanner (EPI BOLD: TR/TE = 3986/60 ms, slices = 30, voxel = 3.44 × 3.44 × 5 mm).

In LaConte et al. (2003a), we examined modeling performance with respect to ten preprocessing combinations. These were generated by (1) aligning each fMRI volume and resampling it into a Talairach reference space (Talairach and Tournoux, 1988), (2) smoothing each axial slice with a 2D Gaussian kernel at one of three levels {0, 1.5, 6.0 pixels full-width at half-maximum}, and (3) detrending and removing confounds by performing volume mean normalization and then removing temporal trends and experimental block effects within a GLM framework as suggested by Holmes et al. (Holmes et al., 1997); constant terms and cosine terms at one of three levels {0, 0.5, 2.0 cycles} constituted the covariates within a design matrix and the residuals of the GLM model were retained as the detrended data. In all detrending cases, the run mean was also subtracted from each time course. For both CVA and SVM, only scans acquired entirely within the 45-s control and 45-s force states were considered. Scans occurring at transitions between force or rest periods were excluded from the modeling exercise (LaConte et al., 2003a). From this approach, the ten preprocessing combinations studied included no preprocessing (run mean subtraction, no smoothing, no alignment—unaligned mean volumes for each run were used to obtain a resampling transformation matrix) and the nine (aligned) combi-

nations derived from the three detrending and three smoothing levels. See Table 1 for the preprocessing abbreviations used with many of the figures.

For CVA/PCA, model complexity can be controlled through the number of principal components used. For SVM, the choice of kernel and the parameter C control model complexity. We used resampling techniques for model selection (Cherkassky and Mulier, 1998; Efron and Tibshirani, 1993; Hastie et al., 2001; Ripley, 1998). This is done by dividing a given data set into two disjoint samples—a learning set and a validation set. For comparison with LaConte et al. (2003a), we use two repeated fMRI experimental runs as our independent splits, resulting in only one possible split, generating two prediction accuracy estimates. That is, training with run 1 and estimating prediction accuracy on run 2 and vice versa. For each preprocessing combination, CVA/PCA was performed for several levels of model complexity (using 10, 25, 50, 75, and 100 components). These same data were analyzed with SVM using a polynomial kernel of the form $H(\vec{x}, \vec{x}^T) = \left[(\vec{x} \cdot \vec{x}^T) + 1\right]^q$ (using degree 1, 2, and 3) to vary complexity for SVMlight's default C value $\left(\frac{1}{T}\sum_{t=1}^{T} \| \vec{x}_t \| \right)^{-1}$. With this approach, we found very little benefit in using the more flexible 2nd and 3rd degree kernels (see Fig. 2). We therefore focused our model tuning resampling on the C parameter, which controls the tradeoff between training error and the margin. To investigate the sensitivity of this parameter, we used several values around $C = 1$, which is close to the default for our data, and also included extreme high and low values. Specifically, we used the following C values: (0.0001, 0.4, 0.6, 0.8, 1, 1.2 1.4, 1.6, and 10000). For both CVA and SVM, two-class (force-baseline) classification was compared, using run 1 for training and run 2 for validation (and vice versa) to generate two sets of classification results per subject. Considering each subject, with two functional runs and ten preprocessing levels at each level of model complexity, 1600 CVA models ($16 \times 2 \times 10 \times 5$) and 2880 SVM models ($16 \times 2 \times 10 \times 9$) were generated. For each model, percent prediction accuracy was calculated on the validation set, calculating $\frac{[\text{number correctly classified scans}]}{[\text{total number of scans}]} \times 100$. The best CVA and SVM model for each preprocessing combination and each subject were selected by choosing the most accurate test result for each level of complexity and each run. As noted by Cherkassky and Mulier (1998) and Friedman (1994), this approach has limitations for comparing prediction accuracy estimates since a single resampling for both complexity control (the model selection we have just described) and methodological comparisons (CVA vs. SVM and even one preprocessing to another) results in optimistic prediction accuracy estimates that may not be consistent across classification methods. We have not followed the prescribed double resampling here because of constraints in the amount of data and because our aim is to demonstrate the ability to evaluate relative performance for different methodologies rather than focus on the true predictive ability of our models.

Table 1
Preprocessing abbreviations

Abbreviation	Description
hh	High detrending, high smoothing
hl	High detrending, low smoothing
hn	High detrending, no smoothing
lh	Low detrending, high smoothing
ll	Low detrending, low smoothing
ln	Low detrending, no smoothing
nh	No detrending, high smoothing
nl	No detrending, low smoothing
nn	No detrending, no smoothing
xx	No detrending, no smoothing, (no alignment)

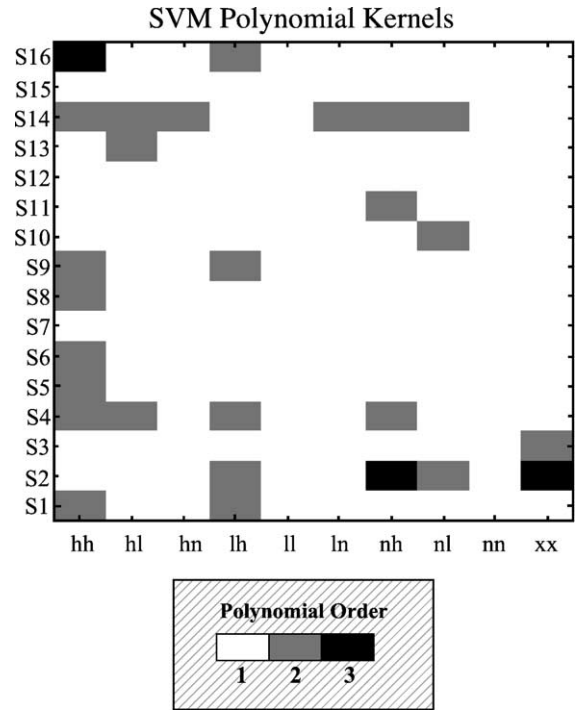

Fig. 2. Kernel parameter selection: for each subject and preprocessing, the kernel parameter (polynomial order) was varied (C was fixed to SVMlight's default). The winning model, based on prediction accuracy estimates, is reported here as one of three gray values representing 1st, 2nd, and 3rd order polynomials.

FSW maps were generated by excluding support vector time points from a pixel-wise correlation test. These were compared with conventional correlation maps (using the original, on/off, experimental reference function). Beyond graphical descriptions, we consider the minimum and maximum correlation coefficients obtained between the two methods. Intuitively, it is more likely to obtain larger correlation values from FSW, which has fewer observations. To interpret the significance of these differences (see Press et al., 1992), we convert the extreme correlation values to z scores using Fisher's z transformation,

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{6}$$

where r is Pearson's correlation coefficient. Significance between two correlation values (e.g. max reference vs. max FSW) is obtained by

$$s = \operatorname{erfc}\left(\frac{|z_1 - z_2|}{\sqrt{2}\sqrt{\frac{1}{T_1 - 3} + \frac{1}{T_2 - 3}}}\right), \tag{7}$$

where T_1 and T_2 are not NMR relaxation times, but the number of time observations used to obtain the corresponding correlation coefficients. The function erfc is the complementary error function.

Results

Our initial exploration of polynomial kernels convinced us to focus on linear kernels (where the input space is equivalent to the feature space). Fig. 2 shows our justification for this. It also represents the first of many result "images," where gray scale represents the value of a result (in this case polynomial order), with one row per subject and one column per preprocessing. Readers are again referred to Table 1 for the preprocessing abbreviations used for these figures.

From our resampling of C, we observed several things. One was that, across a set of C values (fixing subject and preprocessing), one run tended to dominate. That is, using one run as training data would consistently lead to better prediction results for the other run. Interestingly, for CVA, this effect was consistent while for SVM no obvious pattern emerged across subject and preprocessing (Fig. 3). In both cases, the performance gap between training with run 1 and training with run 2 as well as the mean performance from both runs varied across both subject and preprocessing, again with no obvious pattern. Focusing on SVM models, we observed that prediction accuracy was typically only degraded with very small C values (Fig. 4). The subset of the data used in the winning SVM model varied from roughly 30% to 100% of the total training data, varying mostly with subject, but also tending to do better (fewer SVs) with high detrending, and to some extent with high smoothing (Fig. 5).

In Fig. 6, we show the SVM %PA for the winning model of both runs and nine C-value models for each subject and preprocessing. For comparison, we also show our previous CVA results across five model complexity levels (LaConte et al., 2003a) in the same format. Figs. 6A and B are summary images, showing

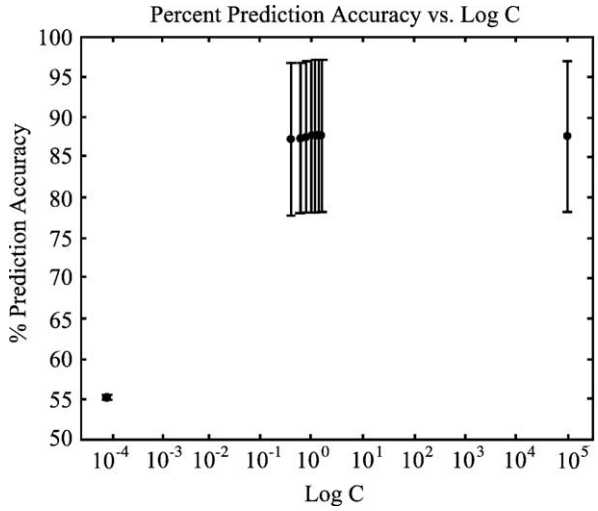

Fig. 4. Percent prediction accuracy vs. C for all subjects, preprocessing, and runs: Here, log(C) is plotted for the values chosen (.0001, 0.4, 0.6, 0.8, 1.0, 0.2, 0.4, 0.6, 100,000). Error bars represent +/− 1 standard deviation.

%PA of the winning models. Figs. 6C and D allow a side-by-side comparison of Figs. 6A and B for both preprocessing and subject. We again must point out that we do not have enough data for a robust estimate of true prediction accuracy. There is, however, much information in Fig. 6 in terms of the number of subjects and preprocessing combinations. Perhaps the most striking observation is that SVM appears to be less sensitive to preprocessing decisions, while CVA prediction tends to be affected by detrending, as was also noted in LaConte et al. (2003a). From Fig. 6C, it appears that

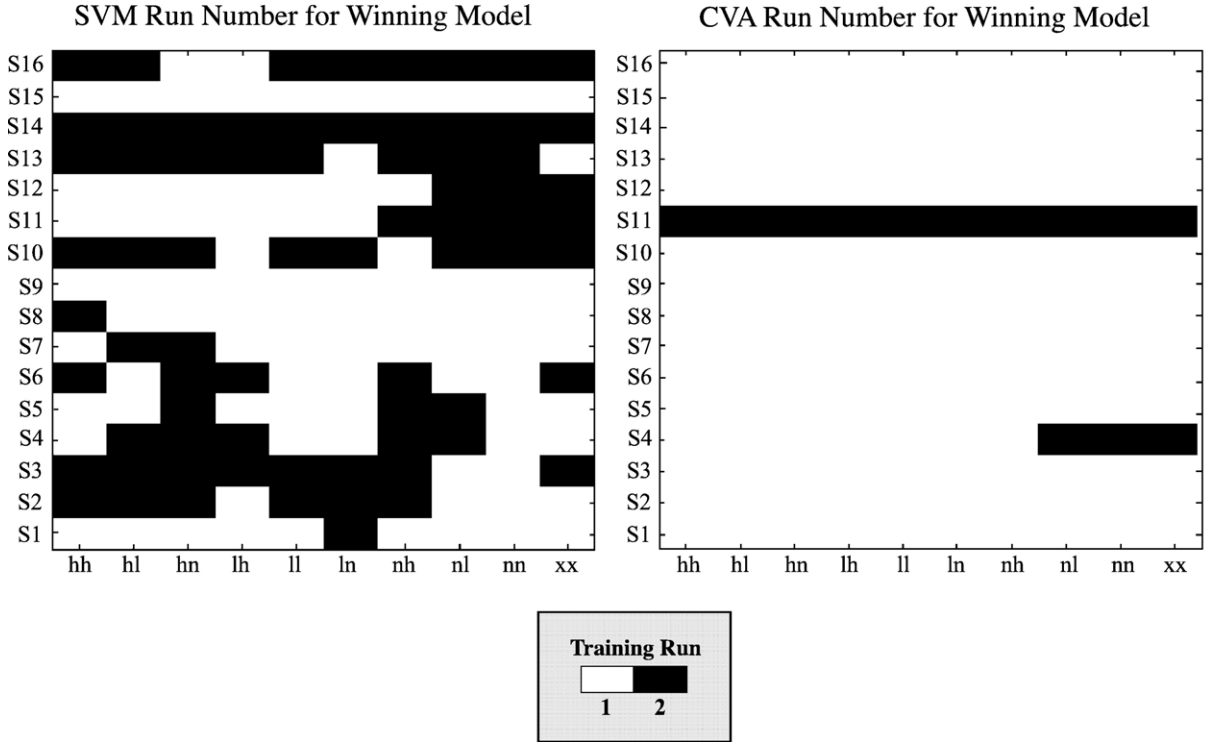

Fig. 3. Training run providing highest prediction accuracy for the pair of splits: Our resampling exercise provided two prediction accuracy estimates for each subject, preprocessing, and model complexity value (C for SVM and number of PCs for CVA). That is, the model from run 1 predicted labels of run 2 and vice versa. Here, we show the run that provided the best model for predicting the opposite run for the winning complexity values of Fig. 6.

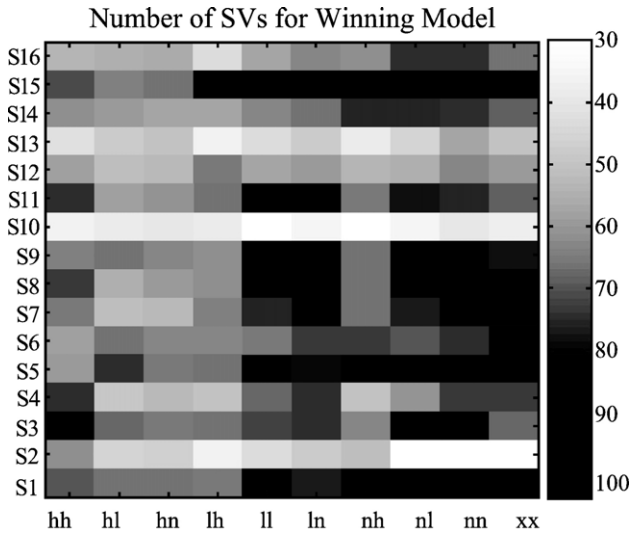

Fig. 5. Number of support vectors used in winning models: The subset of the data used in the winning SVM model varied from roughly 30% to 100% of the total training data.

SVM is also primarily affected by detrending. Looking at the mean and spread of results, low detrending may be optimal for SVM, while CVA appears to improve with increased detrending. Fig. 6D supports our impression from A and B that, for a given subject, SVM classification is less variable.

In terms of SVM activation maps, FSW is compared to conventional cross-correlation for subjects 4 and 14 in Figs. 7 and 8 respectively, and quantified in Table 2. In this case, FSW was implemented by simply discarding observations corresponding to support vector images (weight = 0) and retaining non-support vector times (weight = 1). From Fig. 6A, subject 4 provided consistently good models, while subject 14 performed well except for cases of low detrending. For these two subjects, we show FSW results for training models using the first experimental run, linear kernel and $C = 1.0$ with preprocessing combinations (A) dc detrending, no smoothing, no alignment; (B) low detrending, low smoothing, and alignment; and (C) high detrending, high smoothing, and alignment.

The scatter plots to the left of the summary maps in Figs. 7 and 8 are individual pixel results for conventional cross correlation vs. FSW for all brain voxels. We chose the top 10% from each test to choose the activation threshold. Thus, pixels above the horizontal threshold were the top 10% from the conventional cross-correlation test (blue and orange). Similarly, pixels to the right of the vertical threshold line were the top 10% for the FSW method (magenta and orange). Orange, then, represents the agreement of both methods for the chosen thresholds. As demonstrated by both the summary maps and the scatter plots (see also Table 2), there is a high level of agreement between the two methods. Correlations between the two methods' activation maps were larger in subject 4 than subject 14. For both subjects, the correlation of the activation maps was consistent with the amount of preprocessing (correlation values were highest for high detrending and high smoothing, and lowest for no preprocessing). This is interesting in that such a ranking is not evident based solely on prediction accuracy. Since the maps from the two methods largely agree, it is difficult to draw strong contrasts between them. It does appear, however, that blue pixels tend to be more isolated and spurious. It is also important to note that (based on the number of support vectors required) the FSW maps are derived from only roughly half of the data from the experimental run.

Table 2 quantifies some of the relationships between the conventional reference correlation and feature space weighting maps. The agreement between the maps for both methods is presented in terms of linear, pixel-wise correlations under the third column (Map Corr.). Of the 105 time points per run, the number of support vectors used in the trained model (and thus discarded from the FSW maps) is SVs. We demonstrate the significance levels of the differences in correlation values obtained with the FSW method using the conservative two-sided test described in Eqs. (6) and (7).

Figs. 9 and 10 give an additional perspective on the models from Table 2 as well as Figs. 7 and 8. Regardless of the kernel used, it is always possible to examine the relative role of each time measurement in the training data through the α_ts. Shown as dots are $\alpha_t y_t$. Non-zero α_ts are support vectors and are emphasized with red circles. The square waves idealize the experimental paradigm (representing the training class labels), with x marks indicating discarded transition scans that were not included in the training model. As indicated in Table 2, the number of SVs required tends to decrease with degree of preprocessing. As shown in Figs. 9 and 10, however, the amplitudes of the α_ts tend to increase dramatically. They are bounded by C (in this case $C = 1$). For these plots, it is difficult to draw generalizations concerning temporal distributions of the SVs. It does seem, however, that the final baseline period is important in both subjects. In other data sets (unpublished results), we have noted that transition scans tend to become upper-bound support vectors.

Discussion

Focusing primarily on linear (soft margin) SVM classifiers, we have described many of the issues important to block design fMRI. To assess performance, prediction accuracy results of tuned SVM models were compared with our recent CVA work across sixteen subjects and ten preprocessing combinations. We also discussed various aspects of interpretation and visualization of SVM models in the context of fMRI.

Three important issues relevant to this work require further elaboration. These are classification performance, model interpretation, and validation. A good model, from a classification perspective, generalizes well (as measured by prediction accuracy), implying that it has characterized an important aspect (or feature, or statistical property) of the training data that is consistent for independent data arising from the same system (sampling distribution). Typically, good generalization relates not only to the appropriateness of the chosen model, but also to the selection of appropriate model parameters and preprocessing operations (which may also be viewed as model hyper-parameters). Thus, training a model requires expertise in terms of model selection, while interpreting the model requires understanding of the model assumptions as well as the origin and context of the data itself. For fMRI, we most often assume our a priori knowledge exists in the temporal characteristics of the data, even though most current research in this field focuses on spatial detection. One difficulty with spatial detection in this setting is the issue of validation. This is one strong argument for multivariate techniques, which simultaneously consider both temporal and spatial aspects of this inherently spatiotemporal data. Here, we have only treated model

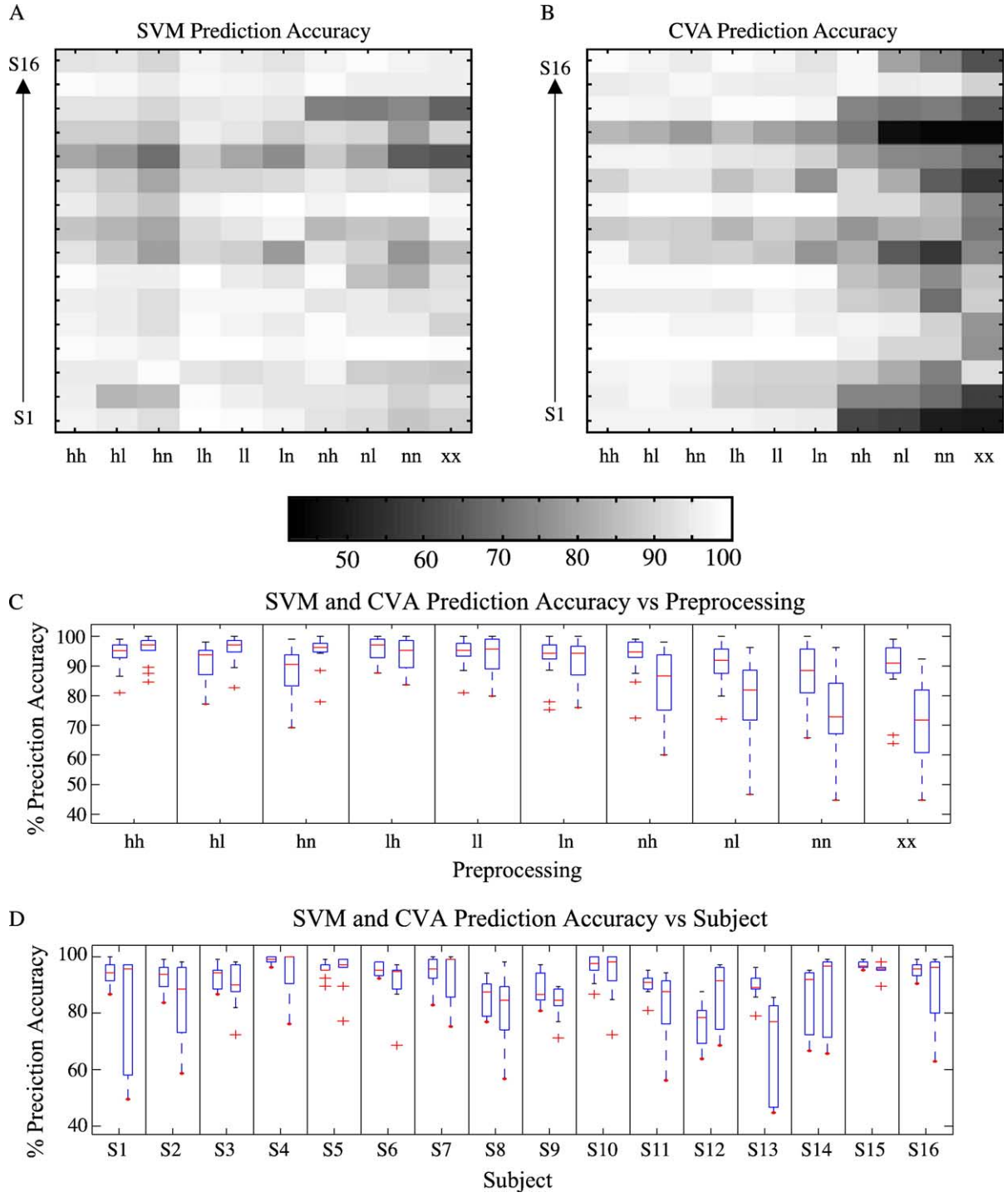

Fig. 6. Winning prediction accuracy results for SVM and CVA models for each subject and preprocessing combination: Results in panel (A) are for SVM while results in panel (B) are for CVA. Both panels (A) and (B) use a common color map. Panels (C) and (D) provide a side-by-side comparison of the results in panels (A) and (B). Panel (C) is each preprocessing across subjects (columns of panels (A) and (B)) and panel (D) is each subject across preprocessings (rows of panels (A) and (B)). For panels (C) and (D), the box plots on the left are for SVM results and on the right are CVA.

selection based on prediction accuracy. Our work on prediction vs. reproducibility plots, however, not only gives a means of validating results (by optimizing both prediction and reproducibility), but also adds flexibility to the model selection process. As we have noted, this is a unique approach to model selection, but one that we argue makes sense for neuroimaging (LaConte et al., 2003a). In essence, this approach gives an investigator a formal means for balancing his or her confidence in known temporal information of the data with the priority of obtaining interpretable (and reproducible) summary images.

We also observed a preference towards the linear kernel and an insensitivity of the SVM regarding the parameter C for sufficiently large values. The kernel often acts to increase the dimensionality of the feature space, however, our data are already of very high

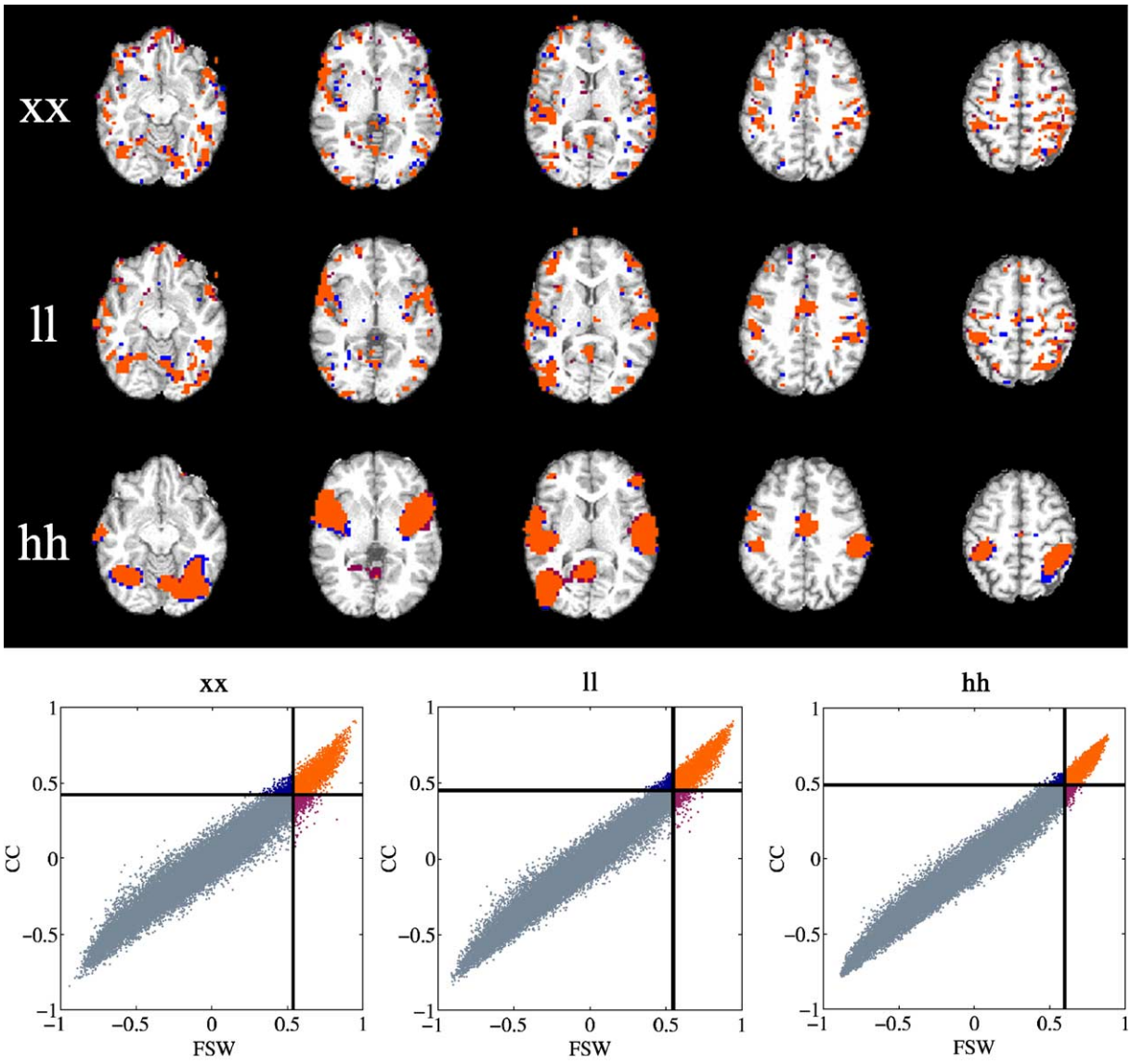

Fig. 7. Comparison of feature space weighting (FSW) to conventional cross-correlation for subject 4 for three levels of preprocessing: the scatter plots to the left of the summary maps are individual pixel results for conventional cross-correlation vs. FSW for all brain voxels. We chose the top 10% from each test to choose the activation threshold. Thus, pixels above the horizontal threshold were the top 10% from the conventional cross-correlation test (blue), pixels to the right of the vertical threshold line were the top 10% for the FSW method (magenta), and pixels satisfying both thresholds represent the agreement of both methods (orange).

dimensionality (note that this is a dramatically different situation from the Cox and Savoy study, where the classification exercise was predicated by GLM-based selection of voxels). That small C values consistently led to poor SVM models indicates that such values do not adequately penalize data points within the margin or on the wrong side of the margin. With C large enough, prediction accuracy was relatively consistent. In practice, then, complexity resampling may not be crucial if this and the kernel selection results hold in future studies.

In Fig. 3, we noted sensitivity to experimental run with CVA that was not apparent with SVM. In that figure, the sampling took place at each grid coordinate for each C value. Thus, the results for CVA are an interesting observation and are clearly systematic. For all but subject 11 (and for subject 4 with low preprocessing), training with run 1 was preferred. Several explanations for this may be possible. One is that for the majority of subjects, an assumption of stationarity (in time) is being violated, which CVA seems to be more sensitive to than SVM. From this, it is possible that the observed time effect (learning, fatigue, scanner performance) is changing the within and/or between covariance of force vs. rest but preserving the margin or interface between these two states.

Although we believe that classification performance, per se, is a relevant and important scientific question for fMRI, we have narrowed our focus to the sensitivity of CVA and SVM to preprocessing combinations frequently applied to fMRI data. Results in Fig. 6 suggest that SVM is less sensitive to preprocessing than CVA. Notably, CVA prediction tends to be affected by low frequency characteristics of the data (altered by detrending). One possible argument against kernel SVM is that (in our current approach) fMRI data already occupy high dimensional feature spaces. We can think of preprocessing as changing the coordinates of each time observation in this feature space. Dramatic changes in

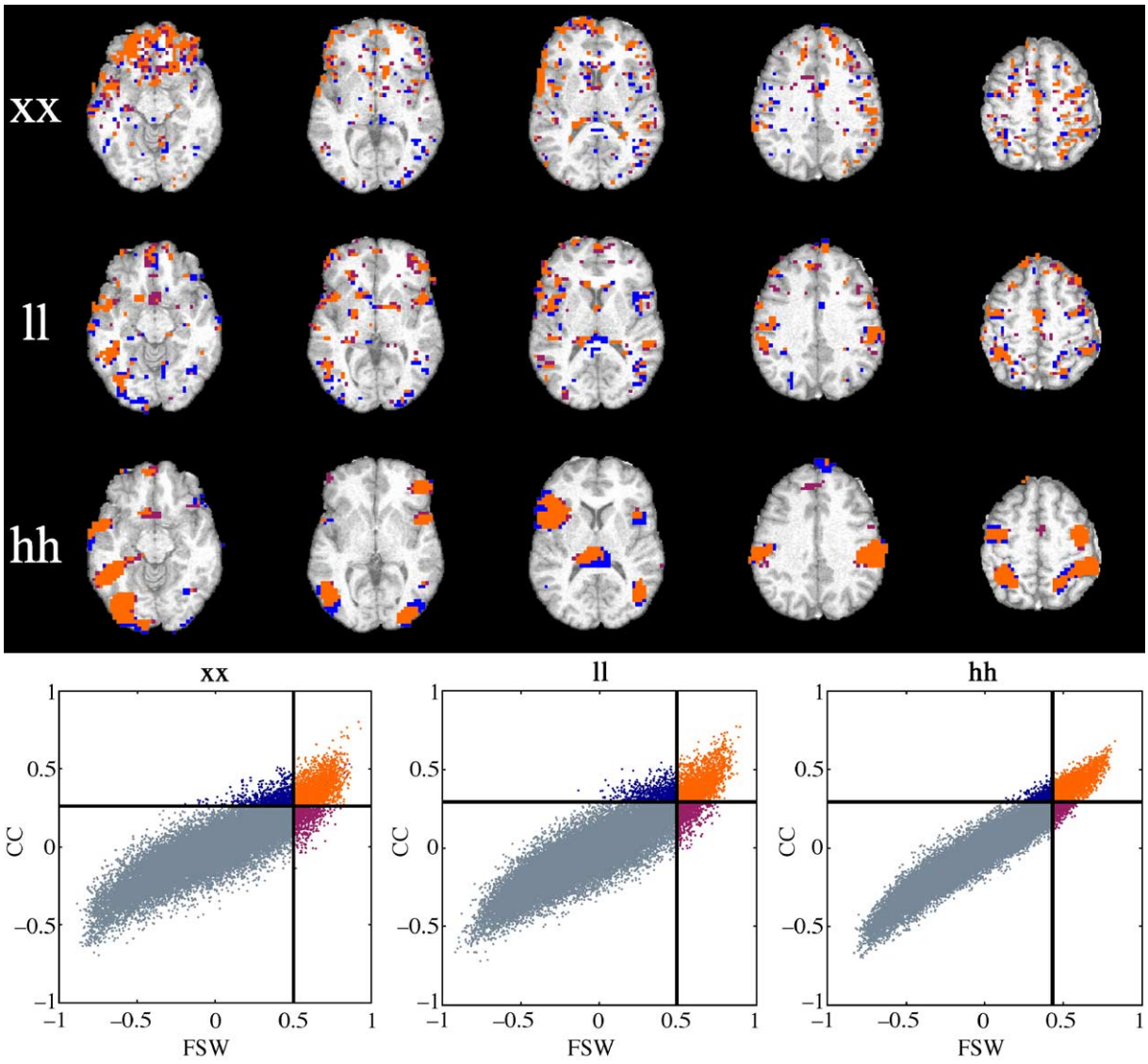

Fig. 8. Comparison of feature space weighting (FSW) to conventional cross-correlation for subject 14 for three levels of preprocessing. The presentation here is exactly analogous to Fig. 7.

an SVM model, however, require that preprocessing change the relative locations of those observations closest to the margin. Depending on the scientific emphasis, we believe that both CVA and SVM have appropriate uses. CVA is formulated based on variance/covariance concepts, and the fact that it does seem sensitive to common preprocessing strategies suggest that it is

Table 2
Comparison of conventional cross-correlation and feature space weighting (FSW) activation maps

Subject	pp	Map Corr.	SVs	Conventional reference		Feature space weighting		Significance	
				Minimum value: r (z)	Maximum value: r (z)	Minimum value: r (z)	Maximum value: r (z)	Minimum values	Maximum values
4	xx	0.97	50	−0.84 (−1.2)	0.91 (1.5)	−0.94 (−1.8)	0.95 (1.8)	0.002	0.05
	ll	0.98	44	−0.83 (−1.2)	0.90 (1.5)	−0.91 (−1.5)	0.94 (1.8)	0.04	0.1
	hh	0.99	41	−0.78 (−1.1)	0.82 (1.2)	−0.89 (−1.4)	0.89 (1.4)	0.03	0.1
	xx	0.87	79	−0.70 (−0.86)	0.80 (1.1)	−0.88 (−1.4)	0.93 (1.7)	0.02	0.02
14	ll	0.90	69	−0.72 (−0.91)	0.77 (1.0)	−0.92 (−1.6)	0.90 (1.4)	0.001	0.04
	hh	0.96	62	−0.70 (−0.87)	0.68 (0.83)	−0.83 (−1.2)	0.83 (1.2)	0.09	0.06

This table quantifies information presented in Figs. 7 and 8. The first two columns define subject number and preprocessing combination. From 105 time points per run, SVs are the number of support vectors used in the trained model (and thus discarded from the FSW maps). The next columns list minimum and maximum correlation coefficients (r) and their converted z scores for both methods. The significance of the differences between these minimum and maximum values is given in the last two columns.

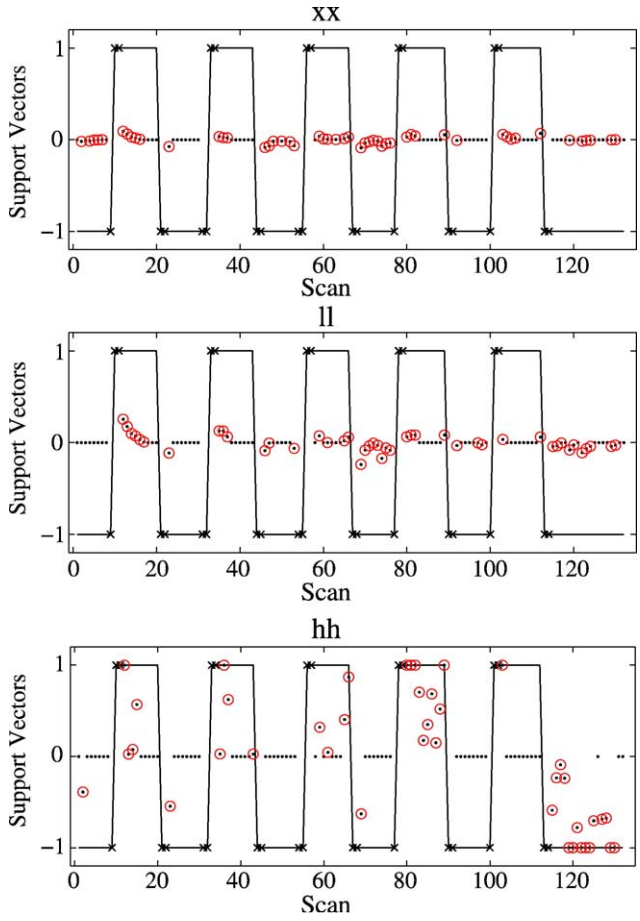

Fig. 9. Temporal occurrence of support vectors for subject 4: Dots in this figure represent each $\alpha_t y_t$ (see Eq. (2)) for the models in Fig. 7 and Table 2. Non-zero α_ts are support vectors and are emphasized with red circles. The square waves idealize the experimental paradigm (representing the training class labels), with x marks indicating discarded transition scans that were not included in the training model. The magnitude of α is bounded by C (in this case $C = 1$).

well suited for examining the signal and noise questions we are currently pursuing (LaConte et al., 2003a; Strother et al., 2002). On the other hand, the unique formulation of the SVMs necessitates further study and provides a unique window for studying the complex structure of fMRI data.

In this paper, we have also begun the important examination of model interpretation of SVMs in the context of fMRI. As aptly stated by Mjolsness and DeCoste (2001),

> Discriminative models make no attempt to explicitly capture the true underlying physics of the phenomena. Nevertheless, as many recent successful applications of methods such as SVMs have shown, such classifiers can provide strong insights into the nature of the phenomena, including such aspects as which input dimensions are most useful, which examples are most likely to be outliers, and what new observations might be most worthwhile to gather.

We have proposed four methods for generating activation maps from SVMs. Weight vector maps, sensitivity maps, and FSW maps are based on the training data, while decision weighting is based on the validation data. The proposal of these four techniques illustrates that there are multiple methods for extracting interpretable information from the SVM framework. In general, training-based interpretation makes use of the model generation step and validation-based interpretation addresses issues of model generalization. Weight vector maps are directly related to the support vectors. For this reason, they may highlight data features that tend to be ambiguous between the classes. In poor models, where all data are support vectors, and all α_ts are equal, weight vector maps reduce to subtraction images. Though it may be possible to develop meaningful intuition from nonlinear kernel models, direct visualization of $\vec{w}$ is most straightforward with a linear kernel. Sensitivity maps elucidate the importance of a particular subspace or set of features (a reduced set of voxels) on SVM training. FSW looks at the distance of each observation from the separating hyperplane, using the intuition that observations closest to this boundary are most difficult to classify. Decision weighting, through $D(\vec{z}_t)$, also considers distances in the feature space, but on new data, independent of the training data. Further work is needed to establish the relative strengths and weaknesses for these model visualization techniques. In our initial work with FSW (LaConte et al., 2003b), we looked at the t score distribution of all brain voxels and observed an enhanced t score distribution, with a slightly taller peak and slightly longer tails. This subtle effect is consistent with our significance results in Table 2.

To demonstrate concepts, we have focused on two-class analysis for runs within a single scanning session, but this is not a limitation

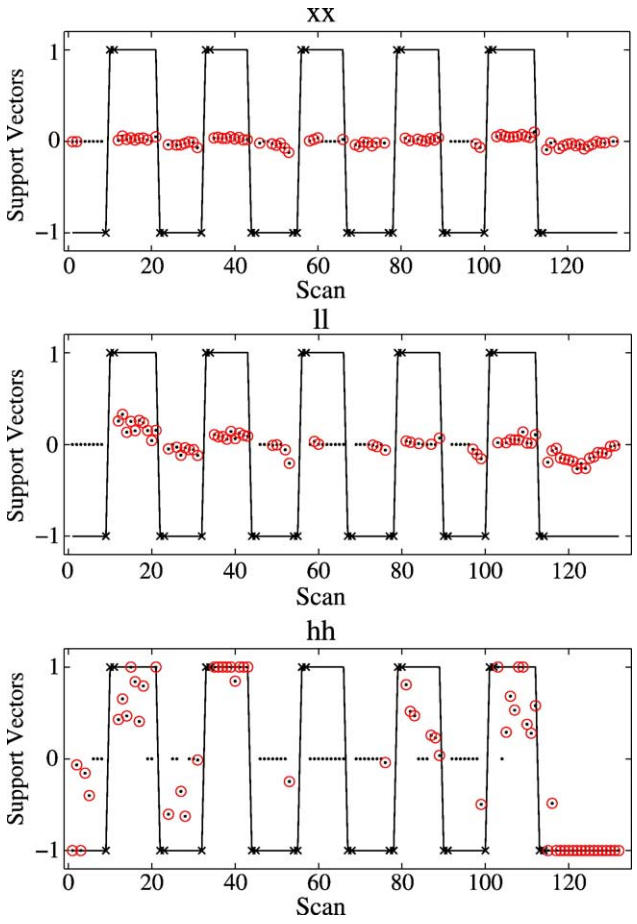

Fig. 10. Temporal occurrence of support vectors for subject 14: As in Fig. 9, the $\alpha_t y_t$s give the importance of a given scan for the models in Fig. 8 and Table 2.

of SVM. Indeed Cox and Savoy demonstrated ten categories with success in data acquired more than 1 week apart (Cox and Savoy, 2003). As SVMs are new to fMRI, we have carefully outlined the essential concepts for SVM classification. Our hope is that this description provides a useful reference for other investigators. We have argued that understanding of both the model and the application is essential to interpretation and visualization of results. As further evidence of the capabilities of SVM classification, we have classified individual time samples of whole brain data, with TRs of roughly 4 s, thirty slices, and nearly 30,000 brain voxels, with no averaging of scans or prior voxel selection.

Looking to the future, our work as well as recent work by Cox and Savoy (2003) illustrate the feasibility and potential of SVM for fMRI data. We look forward to additional evidence concerning linear vs. non-linear kernel selection. Further investigations evaluating map generating strategies and other insights into model interpretation will greatly enhance the use of SVM classification in fMRI research. A final issue is that of experimental design. We have only dealt with block design data, but extensions targeting predictive modeling of event-related fMRI will be a major contribution to the field. Going further, there is a growing interest in real-time fMRI and fMRI-based feedback. In many cases, multivariate models are a much more natural approach for these types of studies compared to univariate tests since, in the multivariate case, feedback can occur for each measurement. Temporally predictive models provide the capacity for adaptive feedback of the stimulus paradigm to the subject based on classified brain state, constituting additional flexibility for experimental design (LaConte et al., 2004).

Acknowledgments

Many people have helped with various aspects of this project. We especially wish to acknowledge Dr. Jihong Chen, Dr. Yasser Kadah, Dr. Scott Peltier, Dr. Shing-Chung Ngan, Mr. Kirt Schaper, and Dr. Kelly Rehm.

References

Bandettini, P.A., Jesmanowicz, A., Wong, E.C., Hyde, J.S., 1993. Processing strategies for time-course data sets in functional MRI of the human brain. Magn. Reson. Med. 30, 161–173.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Discov. 2, 121–167.

Cherkassky, V., Mulier, F., 1998. Learning from Data: Concepts, Theory, and Methods. John Wiley and Sons, Inc., New York.

Constable, T.R., Skudlarski, P., Gore, J.C., 1995. An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. Magn. Res. Med. 34, 57–64.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

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

Cox, R.W., Hyde, J.S., 1997. Software tools for analysis and visualization of FMRI data. NMR Biomed. 10, 171–178.

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

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Academic Press, San Diego.

Friedman, J.H., 1994. An overview of predictive learning and function approximation. In: Cherkassky, V., Friedman, J.H., Wechsler, H. (Eds.), From Statistics to Neural Networks: Theory and Pattern Recognition Applications. Springer-Verlag, Berlin, pp. 1–61.

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

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

Hansen, L.K., Nielsen, F.A., Strother, S.C., Lange, N., 2001. Consenus inference in neuroimaging. NeuroImage 13, 2001.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining Inference, and Prediction. Springer-Verlag, New York.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Holmes, A.P., Josephs, O., Buchel, C., Friston, K.J., 1997. Statistical modelling of low-frequency confounds in fMRI. Neuroimage 5, S480.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Scholkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge, MA.

Kjems, U., Hansen, L.K., Strother, S.C., 2002. The quantitative evaluation of functional neuroimaging experiments: generalization error and learning curves. NeuroImage 15, 772–786.

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

Kwok, J.T.Y., 1999. Moderating the outputs of support vector machine classifiers. IEEE Trans. Neural Netw. 10, 1018–1031.

Kwok, J.T.Y., 2000. The evidence framework applied to support vector machines. IEEE Trans. Neural Netw. 11, 1162–1173.

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

LaConte, S., Strother, S., Cherkassky, V., Hu, X., 2003. Predicting motor tasks in fMRI data with support vector machines. Proceedings of the 11th Scientific Meeting of the International Society for Magnetic Resonance in Medicine, July 10–16. Toronto, Canada p. 494.

LaConte, S., Peltier, S., Hu, X., 2004. Real-time prediction of brain states using fMRI. Proceedings of the 12th Scientific Meeting of the International Society for Magnetic Resonance in Medicine, May 15–21. Kyoto, Japan, p. 2551.

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

Lautrup, B., Hansen, L.K., Law, I., Morch, N., Svarer, C., Strother, S.C., 1994. Massive weight-sharing: a cure for extremely ill-posed problems. In: Hermann, H.J., Wolf, D.E., Poeppel, E. (Eds.), Proceedings of the Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific, Ulich, Germany.

Le, T.H., Hu, X., 1997. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed. 10, 160–164.

Metz, C.E., 1978. Basic principles of ROC analysis. Semin. Nucl. Med. 8, 283–298.

Mjolsness, E., DeCoste, D., 2001. Machine learning for science: state of the art and future prospects. Science 293, 2051–2055.

Morch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B., 1997. Nonlinear versus linear models in functional neuroimaging: learning curves and generalization crossover. In: Duncan, J., Gindi, G. (Eds.), Lecture Notes in Computer Science 1230: Information Processing in Medical Imaging. Springer-Verlag, pp. 259–270.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. 12, 181–201.

Press, W.H., Vetterling, W.T., Flannery, B.P., Teukolsky, S.A., 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univ. Press, New York.

Ripley, B.D., 1998. Statistical theories of model fitting. In: Bishop, C.M. (Ed.), Neural Networks and Machine Learning. Springer-Verlag, Berlin, pp. 3–25.

Scholkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Muller, K.R., Ratsch, G., Smola, A.J., 1999. Input space versus feature space in kernel-based methods. IEEE Trans. Neural Netw. 10, 1000–1017.

Shaw, M.E., Strother, S.C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., Egan, G., 2003. Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. NeuroImage 19, 988–1001.

Skudlarski, P., Constable, R.T., Gore, J.C., 1999. ROC analysis of statistical methods used in functional MRI: individual subjects. NeuroImage 9, 311–329.

Smola, A., Scholkopf, B., Muller, K.R., 1998. The connection between regularization operators and support vector kernels. Neural Netw. 11, 637–649.

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

Talairach, P., Tournoux, J., 1988. A Stereotactic Coplanar Atlas of the Human Brain. Thieme, Stuttgart.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.

Xiong, J., Gao, J.-H., Lancaster, J.L., Fox, P.T., 1996. Assessment and optimization of functional MRI analysis. Hum. Brain Mapp. 4, 153–167.