◆ Human Brain Mapping 35:4499–4517 (2014) ◆

Comparing Within-Subject Classification and Regularization Methods in fMRI for Large and Small Sample sizes

Nathan W. Churchill,[1,2]* Grigori Yourganov,[1,3] and Stephen C. Strother[1,2,3]

[1]*Rotman Research Institute, Baycrest Hospital, Toronto, Ontario, Canada*
[2]*Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada*
[3]*Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada*

Abstract: In recent years, a variety of multivariate classifier models have been applied to fMRI, with different modeling assumptions. When classifying high-dimensional fMRI data, we must also regularize to improve model stability, and the interactions between classifier and regularization techniques are still being investigated. Classifiers are usually compared on large, multisubject fMRI datasets. However, it is unclear how classifier/regularizer models perform for within-subject analyses, as a function of signal strength and sample size. We compare four standard classifiers: Linear and Quadratic Discriminants, Logistic Regression and Support Vector Machines. Classification was performed on data in the linear kernel (covariance) feature space, and classifiers are tuned with four commonly-used regularizers: Principal Component and Independent Component Analysis, and penalization of kernel features using L_1 and L_2 norms. We evaluated prediction accuracy (P) and spatial reproducibility (R) of all classifier/regularizer combinations on single-subject analyses, over a range of three different block task contrasts and sample sizes for a BOLD fMRI experiment. We show that the classifier model has a small impact on signal detection, compared to the choice of regularizer. PCA maximizes reproducibility and global SNR, whereas L_p-norms tend to maximize prediction. ICA produces low reproducibility, and prediction accuracy is classifier-dependent. However, trade-offs in (P,R) depend partly on the optimization criterion, and PCA-based models are able to explore the widest range of (P,R) values. These trends are consistent across task contrasts and data sizes (training samples range from 6 to 96 scans). In addition, the trends in classifier performance are consistent for ROI-based classifier analyses. *Hum Brain Mapp 35:4499–4517, 2014.* © 2014 Wiley Periodicals, Inc.

Key words: BOLD fMRI; preprocessing; model optimization; data-driven metrics; head motion; physiological noise; multivariate analysis

Contract grant sponsor: CIHR; Contract grant number: MOP84483; Contract grant sponsor: Bridging CIHR; Contract grant number: IAO123872; Contract grant sponsor: Brain, Mind and Behaviour Grant from the James S. McDonnell Foundation and Heart and Stroke Foundation of Ontario, through the Centre for Stroke Recovery (to S.C.S.)

*Correspondence to: Nathan W. Churchill, Rotman Research Institute, Baycrest 3560 Bathurst Street, Toronto, ON, Canada.
E-mail: nchurchill@research.baycrest.org

Received for publication 13 August 2013; Revised 3 December 2013; Accepted 30 January 2014.

DOI 10.1002/hbm.22490

Published online 17 March 2014 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Blood-oxygenation level dependent functional MRI (BOLD fMRI) is an imaging technique that is sensitive to the changes in blood oxygenation that accompany neuronal activity. This is used to localize stimulus-evoked changes in brain activity, and make predictions about brain regions implicated in experimental stimuli, in order to characterize the relationship between brain function and behavior. However, fMRI is characterized by a weak, complex BOLD response, and significant noise confounds. Furthermore, identifying discriminative brain regions in whole-brain analysis is an ill-conditioned problem, with far more variables (brain voxels) than observations (time points). There is therefore a need for sophisticated analysis models, to reliably and accurately measure the predictors of brain-state.

Multivariate classification and brain mapping techniques have long been applied to functional neuroimaging data [e.g. Hansen et al., 1999; Haxby et al., 2001; Kustra and Strother, 2001; Morch et al., 1997; Tegeler et al., 1999], and have become increasingly popular in fMRI in recent years, due to their sensitivity to spatially distributed, functionally connected brain networks. They have proven invaluable in numerous studies of sensory, motor and cognitive neuroscience [Berman et al., 2013; Broderson et al., 2012; Carlson et al., 2003; Davatzikos et al., 2005; LaConte et al., 2005; Mitchell et al., 2008; Raizada et al., 2010; Shaw et al., 2003; Shinkareva et al., 2007], neurofeedback and brain-computer interface [LaConte et al., 2005; Sitaram et al., 2010; Yoo et al., 2004] and predicting disease state and recovery in clinical populations [Fu et al., 2008; Menzies et al., 2007; Modinos et al., 2012; Pagani et al., 2009; Schmah et al., 2010; Wang et al., 2006; Yoon et al., 2008]. Multivariate classification is well-suited to fMRI studies in which the activation patterns are not known *a priori*, and distributed brain networks are potentially recruited (e.g. novel experimental designs, complex integrative tasks, groups with different behavioral profiles, and the complex effects of disease state); these models allow researchers to simultaneously localize activated brain areas and measure their prediction (generalization) to independent test data.

Research in these domains is typically concerned with both the generalizability of multivariate classifier models, and the confidence with which we can interpret discriminant brain maps. Prediction accuracy (P) measures the generalizability of the classifier model, based on how well we can predict brain-state for new, independent data. Conversely, reproducibility (R) measures the stability of the discriminating brain maps (or. "sensitivity maps") across repeated analyses. That is, how confident we can be in the identified brain regions that contain discriminative information. Both P and R are important measures of model performance, and their relationship characterizes bias-variance tradeoffs in the classifier models [Rasmussen et al., 2012; Strother et al., 2002; Yourganov et al., 2011].

However, the researcher who wishes to employ multivariate classification must choose from among a wide variety of available models, with different underlying assumptions, and without clear information on the (P, R) tradeoffs between models. For example, generative models construct a probability distribution from the data prior to classification, whereas non-probabilistic classifiers directly estimate an optimal decision rule on the data. In addition, the classifier's decision surface may be a linear function in the feature space (a hyperplane), or a nonlinear surface. Despite the proliferation of models, there have been only a few studies on the impact of classifier choices on fMRI results, including [Hansen et al., 2001; LaConte et al., 2005; Lange et al., 1999; Misaki et al., 2010; Rasmussen et al., 2012; Schmah et al., 2010].

Multivariate classifiers for fMRI must also be regularized, which involves imposing constraints that control model complexity, to reduce over-fitting, and optimize the generalizability of the classifier model. Part of model training is thus to determine the optimal regularization parameters. Many regularization strategies have been applied to fMRI analysis, typically by directly penalizing features of the fMRI data (e.g. L_p-norm penalties). Other strategies include component decompositions (e.g. Principal Component and Independent Component Analysis); although they are not often thought of as regularizers, they can be defined as special cases of L_p penalization [Kustra and Strother, 2001]. Alternatively, many studies perform a feature selection step that extracts a subset of brain voxels during model training (e.g. Haxby et al., 2001; Mitchell et al., 2004; Pereira et al., 2009; Misaki et al., 2010]. However, this approach of feature selection driven by task prediction may decrease model sensitivity and generalizability [Rasmussen et al., 2012]. The effect of regularizer choice is even less explored than classifier models, despite evidence that it significantly impacts classifier performance [Churchill et al. 2012; LaConte et al., 2005; Rasmussen et al., 2012; Strother et al., 2004].

There are other, practical considerations when selecting classifiers; namely, how stable are these models as a function of signal strength and sample size? The majority of classifier-based studies use large, multisubject experimental datasets with hundreds of scans, and conservative preprocessing strategies. This stabilizes classifiers, and minimizes the influence of noise and artifact. It is not clear how performance generalizes to within-subject analysis and small datasets, where only a few scans may be available per stimulus type. These issues arise when analyzing brief, single-subject datasets and measuring cognitive changes over short time scales. They are also relevant for clinical studies, in which patients may only tolerate a few minutes of scanning, and a significant portion of the dataset may be corrupted with head motion and other noise [Carter et al., 2008]. An additional consideration is the number of input voxels in the analysis. In general, classifier studies focus exclusively on either whole-brain analysis [e.g. LaConte et al., 2005; Rasmussen et al., 2012; Schmah et al., 2010;] or focus on smaller, predefined regions of interest (ROIs) [e.g. Haxby et al., 2001; Misaki

et al., 2010]. However, it is unclear if the findings of whole-brain studies are relevant to ROI-driven analysis, and vice versa.

In this article, we examined the performance of different classifier/regularizer choices for the analysis of individual subject datasets, using task battery data collected from young, healthy subjects. We examined four standard classifiers: Linear Discriminant (LD), Quadratic Discriminant (QD), Logistic Regression (LR), and Support Vector Machines (SVM), all of which have been previously applied to fMRI data. We also examined the impact of four different regularization methods: L_2 and L_1 penalties, along with two methods of subspace estimation, Principal Component Analysis (PCA) and Independent Component Analysis (ICA). All classification and model regularization was applied to data in the linear kernel space (i.e. the untransformed covariance matrix between image volumes) because in our experience, sparse regularizers applied in the voxel space tend to produce overly sparse activation maps [Rasmussen et al., 2012]. We compared model performance under a range of different task contrasts, and as a function of sample size, by sub-sampling from within-subject datasets. We evaluated the classifier/regularizer models using measures of (P,R) and global Signal-to-Noise Ratio (gSNR), estimated in the NPAIRS cross-validation framework [Strother et al., 2002].

We also measured the relative similarity of brain activation patterns across the different classifier models, using DISTATIS, a multidimensional scaling technique that computes bootstrapped confidence estimates of pattern similarity (Abdi et al., 2009]. Finally, we tested whether the trends in classifier performance for whole-brain analysis generalized to ROI-based analyses. We do not address the issue here, but for data on the effects of nonlinear kernels, see Schmah et al. [2010] and Rasmussen et al. [2012].

METHODS

We begin by reviewing the different classifier models and regularization methods that are evaluated in this article. Afterwards, we describe the experimental data that is used to test the different models, along with data preprocessing steps. Finally, we discuss the resampling framework and provide details of the metrics used to evaluate model performance: prediction (P), reproducibility (R), global Signal-to-Noise Ratio (gSNR) and the DISTATIS measure of brain map similarity.

Classifier Models

We examined four binary classifiers that are used in fMRI analyses: Linear Discriminant (LD), Quadratic Discriminant (QD), Logistic Regression (LR), and Support Vector Machines (SVM). In fMRI analysis, classifiers are typically used to predict the behavioral task (the "class") that was performed during acquisition of a brain image. This is achieved in a cross-validation framework, with a labeled training set and an unlabeled test set [Pereira et al., 2009]. We acquired a set of N image volumes, containing V brain voxels, which are represented as V-dimensional column vectors of $\boldsymbol{X}$. For ill-conditioned $\boldsymbol{X}$ ($V >> N$), we greatly reduce the number of features while preserving data dimensionality, by classifying on the NxN linear kernel $\boldsymbol{\Phi}=\boldsymbol{X}^T\boldsymbol{X}$; this is the covariance matrix of $\boldsymbol{X}$, if the mean is subtracted from each voxel. We emphasize that all classifier and regularizer combinations (jointly termed "models") were applied in the linear kernel space $\boldsymbol{\Phi}$; this is in contrast to much of the fMRI classifier literature, which directly regularizes the voxel features of $\boldsymbol{X}$, imposing spatial sparsity to achieve a well-posed solution. We chose our alternate approach, as prior research has demonstrated that voxel-space regularization tends to produce overly-sparse representations, particularly for weaker task contrasts [Rasmussen et al., 2012] whereas the effects of linear kernels are less understood.

The input data includes $\boldsymbol{\Phi}$, which is treated as a column matrix $\boldsymbol{\Phi}=[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,...,\boldsymbol{\phi}_N]$ where each $\boldsymbol{\phi}_n$ is an Nx1 feature vector; it also includes class labels $\boldsymbol{y}=[y_1 \ldots y_N]$ with $y_n \in \{-1,1\}$, denoting scans acquired during class 1 or 2 stimuli, respectively. Each classifier uses training data to estimate a decision boundary $D(\boldsymbol{\phi}_n)$ between the classes, where $D(\boldsymbol{\phi}_n)<0$ assigns feature vector $\boldsymbol{\phi}_n$ to class 1, and $D(\boldsymbol{\phi}_n)>0$ assigns $\boldsymbol{\phi}_n$ to class 2. For all models except SVMs, we assign each data vector $\boldsymbol{\phi}_n$ to the class that maximizes posterior probability $p(\text{class } i|\boldsymbol{\phi}_n)$. The decision boundary is given by the log-ratio of posterior probabilities:

$$D(\boldsymbol{\phi}_n)=\log \frac{p(\boldsymbol{\phi}_n|\text{class } 2)p(\text{class } 2)}{p(\boldsymbol{\phi}_n|\text{class } 1)p(\text{class } 1)} \quad (1)$$

For current results, we have equal sample sizes in each class, and assume equal prior probabilities $p(\text{class } 1)=p(\text{class } 2)$. This requires that we only estimate the conditional distribution $p(\boldsymbol{\phi}_n|\text{class } i)$ for each class. The classifier models are defined as follows; see Table I for a summary of their properties.

Quadratic and linear discriminants

These are generative classifier models, which assume that the data from each class has a multivariate Gaussian posterior distribution, with mean vector μ_i and covariance matrix Σ_i:

$$p(\boldsymbol{\phi}_n|\text{class } i)=(2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2}e^{-\frac{1}{2}(\boldsymbol{\phi}_n-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\phi}_n-\boldsymbol{\mu}_i)} \quad (2)$$

During model training, the maximum-likelihood Gaussian parameters are estimated, including within-class means $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, and covariances $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. For QD, the decision function $D(\boldsymbol{\phi}_n)$ of Eq. (1) is the ratio of Gaussian log-likelihoods. This reduces to a quadratic (that is, nonlinear) function of $\boldsymbol{\phi}_n$:

TABLE I. Properties of the evaluated classifier models

	Decision surface	Model type	Loss function
Linear discriminant (LD)	Linear	Probabilistic	Gaussian
Quadratic discriminant (QD)	Nonlinear	Probabilistic	Gaussian
Logistic regression (LR)	Linear	Probabilistic	Logistic
Support vector machines (SVM)	Linear	Nonprobabilistic	Hinge-loss

$$D(\boldsymbol{\phi}_n)=\frac{1}{2}\log\frac{|\boldsymbol{S}_1|}{|\boldsymbol{S}_2|}-\frac{1}{2}\{(\boldsymbol{\phi}_n-\boldsymbol{m}_2)^T\boldsymbol{S}_2^{-1}(\boldsymbol{\phi}_n-\boldsymbol{m}_2) -(\boldsymbol{\phi}_n-\boldsymbol{m}_1)^T\boldsymbol{S}_1^{-1}(\boldsymbol{\phi}_n-\boldsymbol{m}_1)\} \quad (3)$$

The LD classifier makes the simplifying assumption that the population covariances of the two classes are equal; in place of class-specific sample covariances $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, LD uses the pooled covariance estimate $\boldsymbol{S}=(\boldsymbol{S}_1+\boldsymbol{S}_2)/_2$. This reduces Eq. (3) to a linear function of $\boldsymbol{\phi}_n$:

$$D(\boldsymbol{\phi}_n)=\left(\boldsymbol{S}^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1)\right)^T\boldsymbol{\phi}_n-\frac{1}{2}(\boldsymbol{m}_2+\boldsymbol{m}_1)^T\boldsymbol{S}^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1) \quad (4)$$

This is a decision hyperplane, defined by normal vector $\boldsymbol{w}=\boldsymbol{S}^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1)$ and a bias term $\boldsymbol{b}=-^1/_2(\boldsymbol{m}_2+\boldsymbol{m}_1)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1)$, which can be expressed as

$$D(\boldsymbol{\phi}_n)=\boldsymbol{w}^T\boldsymbol{\phi}_n+\boldsymbol{b} \quad (5)$$

Note that the LD model produces the same solution as fitting a least-squares regression onto target variables $\boldsymbol{y}$, which minimizes the Gaussian loss function [Hastie et al., 2005]:

$$L=\frac{1}{2}\sum_n\left(\boldsymbol{y}_n-\boldsymbol{w}\cdot\boldsymbol{\phi}_n-\boldsymbol{b}\right)^2 \quad (6)$$

Both QD and LD classifiers therefore assign data points to the class of highest posterior probability, under Gaussian model assumptions.

Logistic regression

This classifier maps the unbounded features of $\boldsymbol{\phi}_n$ onto the range [0, 1] using the logistic sigmoid function $\sigma\{a\}=1/(1+e^{-a})$. This is a probabilistic model, derived by expressing the posterior probability of Class 1 as a function of the log-odds ratio from Eq. (1):

$$p(\text{class 1})=\frac{1}{1+\exp\left\{-\log\frac{p(\boldsymbol{\phi}_n|\text{class 1})}{p(\boldsymbol{\phi}_n|\text{class 2})}\right\}} \quad (7)$$

Assuming the log-odds decision boundary is a linear function of the features, we produce a decision hyperplane, defined by normal vector $\boldsymbol{w}$ and bias term $\boldsymbol{b}$ (as in Eq. (5)), with a nonlinear mapping function $\sigma\{D(\boldsymbol{\phi}_n)\}$. This solution requires that we minimize the logistic loss function, which may be compactly expressed as:

$$L=\sum_n\log\left(1+\exp\left\{y_n(\boldsymbol{w}\cdot\boldsymbol{\phi}_n-\boldsymbol{b})\right\}\right) \quad (8)$$

There is no analytic solution to minimize this function, although a number of iterative methods have been developed; we employed the Iteratively Reweighted Least Squares (IRLS) procedure (Bishop, 2006; code developed in-house), which can be readily adapted to L_1 (Lasso) and L_2 (ridge regression) regularized models (applied with Matlab's Statistics toolbox (R2007b); see Model Regularization for more details).

Support vector machines

These classifiers estimate a hyperplane (defined by Eq. (7)) in which the distance from $D(\boldsymbol{\phi}_n)$ to the nearest data point (defined as the "margin") is maximized. If we penalize misclassification errors using "slack variable" weights ξ_n for individual data points, this is solved for weight parameter C and slack variable parameter p by the condition:

$$\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{b}}\left\{\frac{1}{2}||\boldsymbol{w}||^2+C\sum_n|\xi_n|^p\right\} \quad (9)$$

under the constraint

$$y_n(\boldsymbol{w}\cdot\boldsymbol{\phi}_n-b)\geq 1-\xi_n,\xi_n\geq 0 \quad (10)$$

That is, scans must be projected onto the correct side of the decision surface, and also projected beyond the (slack variable-adjusted) margins of this decision surface. We use a non-standard SVM approach. Conventionally, Eq. (10) is defined for voxel-space data (replacing $\boldsymbol{\phi}_n$ with $\boldsymbol{x}_n$) and the SVM is solved in "dual-space" by re-formulating it as a function of its kernel $\boldsymbol{\Phi}$; in standard SVM, vector $\boldsymbol{w}$ represents voxel weights (e.g. a sensitivity map; referred to as the "primal weight vector"). We instead transform the data into kernel space $\boldsymbol{\Phi}$, and the SVM solves for optimal weights $\boldsymbol{w}$ on the kernel features themselves. See *Constructing Voxel Sensitivity Maps* below for more details on how $\boldsymbol{w}$ is transformed into voxel-space. This approach was chosen to maintain consistency with the other classifier models, which all estimate the decision boundary on the kernel-space features. The SVM estimates $\boldsymbol{w}$ as a linear combination of input vectors:

$$w=\sum_n \alpha_n y_n \phi_n \quad (11)$$

In general, the solution admits nonzero weights $\alpha_n > 0$ for a few ϕ_n along the margins of the hyperplane [Rasmussen et al., 2012]. These ϕ_n are the "support vectors" which drive discrimination; thus ϕ_n which are most difficult to classify have the most influence on the decision surface. The misclassification penalty is controlled by parameter C, where a small C allows the margin to include more support vectors contributing $\alpha_n > 0$ to the discriminant function. SVM was the only tested classifier that has no probabilistic interpretation for $D(\phi_n)$. This model was implemented using *svmtrain* and *svmclassify* functions from MatLab's Bioinformatics Toolbox R [2007b].

Constructing voxel sensitivity maps

An important part of classification is to derive a "sensitivity map" s_{vox}, which weights brain voxels based on their importance in constructing the decision boundary $D(\phi_n)$. For all models, we used the approach defined in [Kjems et al. 2002], which estimates s_{vox} based on the partial derivative $s_{vox} = \partial D(x)/\partial x$, a function of training inputs x. For LD, LR and SVM, this is given by the linear feature weighting vector w, that minimizes loss functions in Eqs. ((6), (8), and (10)), respectively [Kjems et al., 2002; LaConte et al., 2005; Rasmussen et al., 2012]. Because w is solved in the kernel feature space, we reconstruct the voxel image via the transformation $s_{vox} = Xw$. For example in LD, $s_{vox} = X\left(S^{-1}(m_2 - m_1)\right)$, where m_1, m_2 and are the kernel-space class means and S is the within-class covariance. For QD, the voxel-space sensitivity map is defined [see Yourganov et al., 2010, for details]:

$$s_{vox} = X\left(S_2^{-1}(x-m_2) - S_1^{-1}(x-m_1)\right)$$

Model Regularization

For this article, we classify on the linear kernel $\Phi_{(NxN)}$, which greatly reduces data dimensionality relative to the original $X_{(VxN)}$ data space. However, further regularization is important, given the noisy, often collinear nature of fMRI data. In this article, we consider "regularization" as any mathematical constraint that limits classifier model complexity, in order to stabilize the decision surface. A variety of such regularization methods are available, including directly penalizing the kernel-space features, and projecting data into a lower-dimensional basis space. For the current article, we examined L_1 and L_2-norm penalization on the features of Φ (i.e. rows of the untransformed NxN covariance matrix), as well as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) subspace selection methods, all of which are commonly used in the fMRI literature.

L_1 and L_2 penalization

The L_p-norm methods constrain the decision surface by directly penalizing the kernel basis weights of Φ. For LD and LR models, which have linear decision boundaries defined by a projection vector w, their loss functions Eqs. (6) (LD) and (8) (LR) have an added penalty term $\frac{\lambda}{2}\sum_i |w_i|^p$, which forces the maximum-likelihood solution of w to have a smaller norm. For example, the penalized LD loss function is:

$$L = \frac{1}{2}\sum_n (y_n - w \cdot \phi_n)^2 + \frac{\lambda}{2}\sum_i |w_i|^p \quad (12)$$

Parameter λ dictates the amount of regularization, where a larger λ increases the influence of the penalty function on our solution.

A linear model with $p = 2$ (known as "ridge regression") tends to assign uniformly high/low discriminant weights w_i to groups of correlated features. For LD, it is equivalent to estimating a decision boundary on the regularized covariance $S^* = (S + \lambda I)$ (13), which produces an increasingly diagonal covariance structure [Kustra and Strother, 2001]. In contrast, the $p = 1$ penalty (known as "lasso") tends to select only a single feature from among a correlated group of features. This is conceptually similar to step-wise regression [Efron et al., 2004], and is of interest because it is able to define a parsimonious subset of predictors in a regression model. The L_1 penalty produces a non-differentiable objective function, and so its decision function $D(\phi_n)$ does not have an analytic solution. The LD–L_2 model was solved using standard ridge regression, and LD-L_1 was solved using a modified version of the computationally efficient LARS procedure [Efron, 2004] (code from www2.imm.dtu.dk/pubdb/views/publication_details.p hp?id=3897). For LR, we used the Iterative Reweighted Least-Squares model, which regresses on a set of adjusted target variables; this model directly performs ridge/Lasso regression for each algorithm iteration (software developed in-house).

For QD, the quadratic decision surface requires alternative methods to implement L_1 and L_2-penalties. For L_2, in an analogous extension of the LD model [Eq. (12)], we directly regularized the within-class covariance matrices using a fixed λ value:

$$S_1* = (S_1 + \lambda I), S_2* = (S_2 + \lambda I) \quad (13)$$

For L_1 regularization, we penalized the inverse covariance matrices of Eq. (3) (i.e. precision matrices), using the "graphical lasso" model [Friedman et al., 2008] (code from www-stat.stanford.edu/~tibs/glasso). This technique sequentially fits an L_1 regression to each variable $1 \ldots N$ of the inverse covariance matrix, treating all other variables as predictors. It gives zero weight to elements of multiple highly collinear (i.e. redundant) variables, producing a sparse inverse covariance matrix. This is a computationally

efficient method of applying L_1 regularization to the QD classifier model.

For SVM, we enforce an L_p-norm penalty on the support vectors, via slack variable parameter p in Eq. (9). The L_2-constrained model is solved using quadratic programming methods; the L_1-constrained model is solved using the Sequential Minimization Optimization algorithm [Cristianini et al., 2000; both provided in the MatLab Bioinformatics Toolbox; R (ver. 2007b; The MathWorks, Inc., Natick, MA).

PCA and ICA decomposition

Both of these models transform the feature-space data $\mathbf{\Phi}$ into a linear subspace of independent and/or orthogonal components, with the component basis estimated in different ways. For all classifiers, we project the data onto a subset of components, and then perform classification in the new coordinate space.

Principal Component Analysis projects $\mathbf{\Phi}$ onto a set of K orthonormal Principal Components (PCs; for $K<N$), where the k^{th} PC explains the greatest possible data variance that is orthogonal to all preceding PCs $1\ldots(k-1)$. This is expressed as the unique decomposition of the symmetric, positive definite matrix $\mathbf{\Phi}$:

$$svd(\mathbf{\Phi}) = \boldsymbol{V}\Lambda^2\boldsymbol{V}^T \tag{14}$$

$\boldsymbol{V}$ = orthonormal basis vectors, and Λ = diagonal matrix of associated singular values. The data are represented as PC-space coordinates $\boldsymbol{Q} = \boldsymbol{V}^{\mathrm{T}}\Lambda$. We perform classification on a subspace of 1 to k PCs, where we vary $(1 < k < K)$. Lower k corresponds to a smaller, more regularized classifier subspace. Strother et al. [in press] have demonstrated that PCA is an important regularization technique, because it provides a highly efficient feature extraction that reflects the first-order, linear covarying network structure of the brain, as previously shown in Positron Emission Tomography studies [Kustra and Strother, 2001; Moeller and Strother, 1991; Rottenberg et al., 1987].

ICA refers to a broad class of models that linearly decompose $\mathbf{\Phi}$ into a set of statistically independent, non-Gaussian source signals $\boldsymbol{V_{IC}}$, by estimating the linear "unmixing" transformation matrix $\boldsymbol{W}$ that produces $\boldsymbol{V_{IC}} = \boldsymbol{W}\mathbf{\Phi}$. Unlike PCs, ICs are not constrained to orthonormality, and ICA models may enforce independence in the spatial or temporal domain, producing non-orthogonal temporal or spatial components, respectively. These models do not have an analytic solution, requiring iterative estimation; they are usually estimated on an initial reduced PCA subspace, which stabilizes the solution and reduces computational burden. In principle, ICA models allow for flexible modeling of signal components, but they are limited by issues of dimensionality, and solutions tend to be unstable [Yourganov et al., 2011]. In this article, we used the FastICA package [Hyvärinen et al., 2000) (research.ics.tkk.fi/ica/fastica, ver. 2.5), which performs ICA following an initial PCA decomposition. We employed the "tanh" (hyperbolic tangent) nonlinearity measure, which produced optimal results out of the available set of constraints, and used the deflation approach to sequentially estimate the ICs. The ICA components are sensitive to the input PC dimensionality. Since it has previously been shown that theoretical estimates of the intrinsic dimensionality are suboptimal, we empirically identified the optimal subspace, by varying both the input PC subspace and number of ICs used to perform classification (see next paragraph).

On the range of regularization

For classification in a PCA subspace, we varied the subspace dimensionality of the within-subject analyses from 1 to $K_{\max}$, where $K_{\max}$ is the maximum non-rank deficient dimensionality $\leq N$. Similarly, we varied L_1 regularization to admit $1-N$ nonzero basis weights. For ICA, we tested all values of 1 to K ICs, for each of $K = 10\%$, 25% and 50% of $K_{\max}$ input PC dimensionality (with ICs ordered by variance). For L_2 regularization, we identified the highest and lowest values of λ at which an increment produced no statistical change in the discriminant maps, and subdivided this range into $K_{\max}$ steps, on the log scale. This was done to ensure that the tested range was comparable to the other regularization methods.

Experimental Data

In order to evaluate the different classifier/regularizer models, we analyzed a set of experimental fMRI task contrasts. The fMRI dataset, originally presented in [Grady et al., 2010], included 19 young, healthy adults (mean age 25 ± 3 years, range 20–30, eight women), screened using a health questionnaire to exclude health issues and/or medications that might affect cognitive function and brain activity. Images were acquired with a Siemens Trio 3-T magnet. T_2 functional images (TE/TR = 30/2,000 ms, FA = 70°, FOV = 200 mm) were obtained using echo planar acquisition. Each image volume consisted of 28 5-mm thick axial slices with 3.125×3.125 mm^2 pixels. T1-weighted anatomical volumes were also obtained using SPGR (TE/TR = 2.6/2,000 ms, FOV = 256 mm, slice thickness = 1 mm). Subject brain volumes were nonlinearly co-registered to a common anatomical template, created from a group-average anatomical reference; this was obtained from an iterative optimization procedure that is robust to inter-subject anatomical variability [Kovacevic et al., 2005]. Functional data were corrected for slice-timing offsets using AFNI (afni.nimh.nih.gov/afni) and motion corrected using AIR (bishopw.loni.ucla.edu/AIR5). Additional preprocessing consisted of spatial smoothing with a FWHM=7 mm Gaussian kernel and regressing out white matter time signal, using an averaged white matter time series as a regressor for each voxel/subject separately; we also performed linear detrending to control for residual motion and scanner drift effects.

Subjects performed four task runs (total duration of 300 TR = 600 s, for each run), in which blocks of four different

stimuli were presented twice in each run, interleaved with fixation blocks; for full details, see [Grady et al., 2010]. The visual stimuli consisted of bandpass filtered white noise patches with different center frequencies; for all tasks, when subjects detected a "target" they pressed 1 of 3 buttons to indicate where the stimulus appeared. The stimulus types include:

Fixation (FX): the participant observed a dot presented on the middle of the screen.

Reaction Time (RT): when the participant detected a target image on a screen, they pressed the corresponding button as quickly as possible.

Delayed Match to sample (DM): test of working memory. A target stimulus was presented and then removed from the screen, followed by a 2.5 s blank-screen delay. Three stimuli were presented and the participant pressed the button corresponding to the target.

Attentional Cueing (AT): the participant pressed the button corresponding to the target image, following a brief directional cue.

Perceptual Matching (PM): the participant was presented a target in the upper portion of the screen, and three stimuli on the lower portion; they pressed the button corresponding to the one of three stimuli on the lower portion that matched the target.

During each of the four experimental runs, the FX condition was presented in 8× (10 TR) blocks, producing 32 blocks in total. The other three conditions were presented in 2× (~20 TR) blocks per run for each condition although there is some variability due to variance in stimulus timing; this produced eight blocks in total. For all analyses, we discarded the first two TRs of each block, producing 8-TR blocks in the FX condition, and task conditions blocks that ranged from 16 to 18 TR (except a single block of 14TR in AT). In order to maintain consistency in block sizes, we therefore trimmed all task blocks to 16TR, discarding scans from the end of the block.

We examined three contrasts, ordered by increasing difficulty of classification: (1) FXvDM, (2) ATvPM, (3) PMvDM. We consider (1) to be a "strong" contrast, as we attempt to classify between relatively large changes in cognitive state, compared to the "weaker" cognitive state differences in (2) and (3). This can be confirmed by the systematically lowered prediction and gSNR in going from (1) to (3), shown in Figs 2 to 4. We selected these contrasts, in order to test whether results are consistent across different strengths of task contrast, as there is evidence that this has an impact on optimal image processing and analysis methods [Churchill et al., 2012; Rasmussen et al., 2012].

Resampling, Performance Metrics, and Regularizer Optimization

We evaluated model performance for the single-subject analyses, in a cross-validation framework. The task blocks constitute our resampling units, as scans within a block are temporally correlated. The overall structure of the data consists of four task runs, with two condition blocks per run (we concatenated pairs of successive FX runs to match the 16-TR task blocks). Given the four scanning runs, we performed fourfold cross-validation (CV) on the data. The (two blocks) × (two conditions) from a single run were designated independent "test" data, and the remaining (six blocks) × (two conditions) were used to perform model training and validation (i.e. optimize model regularization). This procedure allowed us to maximize the independence of "test" data, relative to the training and validation stage.

We performed model training and validation in the NPAIRS split-half cross-validation framework [Strother et al., 2002, 2010], in which we randomly split the training and validation data into two pseudo-independent sets of (three blocks) × (two conditions). We trained the classifier model and generated a discriminant brain map, independently for each data split. We used split-1 data to estimate decision surface $D_1(\boldsymbol{\phi}_n)$ and voxel sensitivity map $\boldsymbol{s}_{vox,1}$, and used split-2 data to estimate $D_2(\boldsymbol{\phi}_n)$ and $\boldsymbol{s}_{vox,2}$. Prediction and Reproducibility were used to quantify model performance.

Prediction (P)

P measures how well the decision surface $D_i(\boldsymbol{\phi}_n)$ predicts the experimental condition of data that was not used to build the decision surface. To estimate prediction, we use $D_1(\boldsymbol{\phi}_n)$ to assign each data-point from split-2 to class 1 or 2. Prediction is measured as the fraction of scans in split-2 that are assigned to the correct class. We then use $D_2(\boldsymbol{\phi}_n)$ to classify split-1 data, and take the average prediction from both splits, denoted P_{valid}. This provides a measure of classifier generalizability that allows us to compare probabilistic QD, LD and LR models with the non-probabilistic SVM. We also tested an alternative prediction of Bayes' posterior probability for all models except SVM, but found no significant differences in model performance compared with classification accuracy.

Reproducibility (R)

R measures the stability of discriminating brain regions in $\boldsymbol{s}_{vox,i}$, across (pseudo)independent data splits. This was estimated as the Pearson correlation of the pairwise voxel values between $\boldsymbol{s}_{vox,1}$ and $\boldsymbol{s}_{vox,2}$. Under Gaussian signal and noise assumptions, this R metric has the advantage of being directly related to the global Signal-to-Noise Ratio (gSNR) of the brain map, by the relation $g\text{SNR} = \sqrt{2R/(1-R)}$. We also used split-half results to generate a reproducible Z-scored Statistical Parametric Map (rSPMZ) of brain activation, using the bivariate scatter plot of $\boldsymbol{s}_{vox,1}$ versus $\boldsymbol{s}_{vox,2}$ voxel values. We project voxel values onto the first PC of the scatterplot (the signal axis) to compute reproducible signal, and normalize by the standard deviation of the second PC (noise axis), producing a Z-scored SPM [Strother et al., 2002].

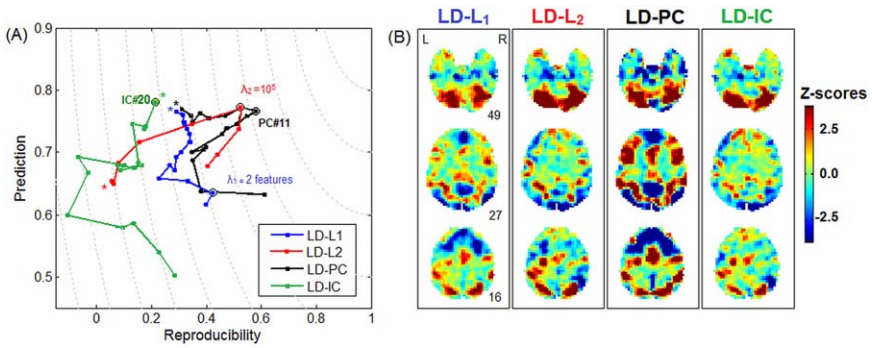

Figure 1.

The impact of regularization choice for a representative subject in the ATvFX contrast, with the linear discriminant (LD) classifier, for a restricted range of 20 regularizer values. (**A**) Plots showing (P_{valid} = validation prediction, R = reproducibility) curves for regularization via Principal Components (PC), L_1 and L_2 penalties, and Independent Components (IC). Each curve traces out (P_{valid}, R) as a function of degree of regularization. A "*" indicates the point of weakest regularization; PC and IC with 20 components (both ordered by variance), L_1 with nonzero weights on 20 of the linear kernel features (rows of covariance matrix $\mathbf{\Phi}$), L_2 with $\lambda_2 = 10^{-8}$. The optimal point that minimizes Euclidean distance from ($P_{valid} = 1$, $R = 1$) is circled for each regularizer. (**B**) The Z-scored subject SPM produced by each optimized regularizer. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For a given analysis, we computed (P_{valid}, R) across the range of regularization parameter values, and used these metrics to select the optimal regularization parameter (e.g. λ for L_1 or L_2; subspace size k for PCA and ICA). The (P_{valid}, R) metrics reflect important tradeoffs in fMRI classification in a pseudo-ROC curve [LaConte et al., 2003; Strother et al., in press]. In general, greater model regularization leads to decreased model prediction, and large, complex changes in reproducibility (see Fig. 1). An optimal classifier should jointly optimize P_{valid} and R; we chose to minimize the Euclidean distance $D=\sqrt{(1-P_{valid})^2+(1-R)^2}$, for which smaller D indicates better model performance. This metric has been previously implemented in model and preprocessing optimization [e.g. Churchill et al., 2012; LaConte et al., 2003; Shaw et al., 2003; Zhang et al., 2009]. Furthermore, Rasmussen et al. [2012] demonstrated that the D metric is an effective compromise between optimizing exclusively on R or P metrics: it improved the stability of brain SPMs across different classifier models, while improving detection of the less predictive nodes of the brain network that underpin motor performance. After selecting the level of regularization that minimizes D, we used the optimized model to classify data in a held-out "test" run. This produced P_{test}, an independent, unbiased measure of prediction accuracy for comparison with the biased P_{valid}.

In this resampling framework, we designated each of the four runs the "test" dataset in sequence (i.e. fourfold CV), and we measured average P_{test} across the four folds. For each fold, we performed 25 training/validation resamples (of $\frac{1}{2}\binom{6}{3}\times\frac{1}{2}\binom{6}{3}=100$ possible ways of splitting the two task conditions) and computed median (P_{valid}, R) values, in order to optimize regularization. When reporting final results, we computed median P, R, gSNR and rSPMZs across the 4x25 =100 resampling splits, for each subject. Note that all figures showing prediction refer to P_{test}, excepting Figure 1, which demonstrates the P_{valid} versus R optimization tradeoff as a function of model regularization.

Model Performance and Sample Size

We analyzed each of the 19 subjects in the fourfold CV framework, for all 16 possible classifier/regularizer combinations (LD, QD, LR or SVM)x(L_1, L_2, PCA, or ICA). This produced median estimates of P_{test}, R, gSNR, and a set of 19 rSPMZs for each subject.

In order to examine the relationship between model performance and sample size, we sampled backward from the end of each task block, retaining only a subset of N_{block} volumes from every block. We tested a range of $N_{block} = \{1, 2, 4, 8, 16\}$, which gives effective training sample sizes per split, for two blocks from each of 3 runs, of {6, 12, 24, 48, 96}, which was used to build the decision surface $D(\boldsymbol{\phi}_n)$. For every block size N_{block}, we then generated all of our performance metrics, and we plot performance as a function of N_{block} for every classifier/regularizer model combination.

We then tested for significant trends in model performance using nonparametric testing. We used a Friedman test to identify models that are significantly different based on ranking. This was depicted using a critical-difference diagram, with significant differences in models assessed using the Nemenyi permutation test at $\alpha = 0.05$ [see Conover et al., 1999 for a review of these procedures]. These

results were computed separately for each of the three different task contrasts (FXvDM, ATvPM, PMvDM). We show ranking results for a fixed N_{block}=16, although model rankings were generally consistent for smaller N_{block} values.

Additionally, we examined the relative bias of the validation prediction estimates P_{valid} (which are used to optimize model regularization), compared to independent unbiased test prediction P_{test}. We plot median P_{valid} and P_{test} for each classifier model, and test whether P_{valid} is significantly higher than P_{test}, using nonparametric paired Wilcoxon tests. A significant difference indicates a consistently biased classifier model. This was evaluated for the minimum and maximum sample size limits, of N_{block}=1 and N_{block}=16, respectively.

The Dynamic Range of Classifier Models

For the previous sections, we chose model regularization to optimize the metric $D(P_{valid}, R)$ of Euclidean distance from perfect ($P_{valid} = 1$, $R = 1$). This provides a compromise that demonstrates how the different models perform, under equal weighting of the two metrics. However, this does not provide information on how models perform at extremes of regularization, when maximizing either P_{valid} or R alone. That is, what is the maximum *attainable* Prediction and Reproducibility for each model? We define this as the effective "dynamic range" of each model. To examine this, we optimized each classifier/regularizer model by (A) maximized P_{valid}, (B) minimized $D(P_{valid}, R)$ and (C) maximized R, for each individual subject analysis. We then plotted median (P_{test}, R) values of the different models under these optimization criteria. The results are shown for a representative LD classifier model, and ATvPM task contrast, but the findings are consistent for all classifiers and contrasts. There is evidence that the brain expresses a range of different task-related activation patterns as a function of increased regularization, ranging from the highly reproducible and spatially extensive but less predictive of the task (e.g. default-mode networks), to the highly task-predictive and more spatially sparse (e.g. task-positive networks) [Rasmussen et al., 2012; Strother et al., in press]. We suggest that an ideal classifier model should have a wide "dynamic range", allowing the user to tune the model to identify this spectrum of network structures.

Comparing SPM Patterns

Although prediction and reproducibility are valuable measures of model performance, they provide no information on the spatial structure of the SPM patterns themselves. In experimental data, two models can provide similar P_{test} and R values, and yet produce significantly different spatial patterns [Churchill et al., 2012]. To address this issue, we used the bootstrapped multidimensional scaling method of DISTATIS [Abdi et al., 2007]. This model identifies consistent trends in the within-subject correlation of model SPMs, and does not directly compare between subject SPMs. The following steps are performed to compare 16 different classifier/regularizer models:

1. For the kth subject ($1 \leq k \leq 19$), compute the doubly-centered 16 × 16 correlation matrix $\boldsymbol{S}_k$, between SPMs generated by the different classifier/regularizer models.
2. Form the 19 × 19 matrix $\boldsymbol{C}$ of pairwise similarity between the $\boldsymbol{S}_k$ matrices, measured using the R_V coefficient of matrix similarity.
3. Perform PCA on matrix $\boldsymbol{C}$. The first eigenvector gives the set of coefficient weights α_k, which indicate how similar each $\boldsymbol{S}_k$ (and thus the k_{th} subject) is to the strongest common $\boldsymbol{S}$-matrix pattern. Then compute the weighted compromise matrix $\boldsymbol{S}_+ = \sum_k \alpha_k \boldsymbol{S}_k$; this is the most common 16x16 cross-correlation between model SPMs across subjects.
4. Perform PCA on $\boldsymbol{S}_+$, and project $\boldsymbol{S}_+$ into the new PCA basis space.

The models' rSPM(z)s are expressed as points in 15-dimensional PCA space, typically reduced to the first two PC dimensions for visualization. The distance between any two models' points represents the Euclidean "distance" between rSPM(Z)s, with points that are closer together indicating more similar activation maps. We produced confidence estimates on the models, by Bootstrap resampling of the subject correlation matrices (1,000 iterations). The bootstrapped matrices were projected into the same PC space, and the resultant 95% confidence ellipses were drawn around the centroids.

We examined model SPMs for clustering, defined as overlap in the 95% confidence ellipses, which indicates that the model SPMs are not significantly distinguishable. This provides a representation of the relative similarity between classifier/regularizer model SPMs. We applied DISTATIS to data at $N_{block} = 16$ sample size, although trends are consistent for smaller samples.

The Effects of Voxel ROI Selection

As a final test, we examined whether results of whole-brain analyses generalize to smaller numbers of input voxels. It is common in many fMRI studies to predefine a voxel ROI for analysis, particularly if there is an a priori hypothesis about a specific subset of brain regions. It remains unclear whether differences between classifiers are preserved when the number of input voxels (V) are much less than the number of timepoints (N). We re-ran the analyses in Model Performance and Sample Size, for the 16 combinations of classifier/regularizer and a full sample size $N_{block} = 16$, comparing (1) whole-brain analysis (21,535 voxels; $V >> N$), (2) a "liberal" ROI mask of

dorsal task-positive regions, including parietal lobes and dorsolateral prefrontal cortex (276 voxels; $V \approx N$), and (3) a "conservative" ROI mask of the bilateral parietal lobes (29 voxels; $V << N$). These regions were chosen as areas typically implicated in overt attention and visual search, with ROIs predefined using an AAL anatomical atlas.

The analyses were run for the intermediate ATvPM contrast. We plotted median P_{test}, R, and gSNR as a function of the number of input voxels. We also plotted SPMs from "conservative" ROI results, to demonstrate the importance of spatial reliability, even in a relatively small subset of voxels.

RESULTS

Model Performance and Sample Size

As a demonstration of the importance of model choice, Figure 1 plots sample (P_{valid}, R) curves as a function of regularization, for a single representative subject (ATvFX contrast, with an LD classifier, for a maximum sample size of $N_{block} = 16$). As shown in Figure 1(A), the regularizers produce different curves in (P_{valid}, R) space, with different optimization points. For all models except LD-ICA, if we trace the curve as a function of increasing regularization, we approach a point of minimum distance from ($P_{valid} = 1, R = 1$), after which overall model performance declines. Figure 1(B) further demonstrates that the resulting activation patterns may be quite different, even under a fixed classifier model. Although there are activations common to all regularizers (e.g. upper cerebellum; slice 49), there are also regions of extreme heterogeneity (e.g. superior frontal, strong signal Z-scores for PCA and L_1 only; slice 16).

Figure 2 plots the median prediction accuracy P_{test} across subjects for each classifier/regularizer model, as a function of sample size. In general, prediction distributions are overlapped, indicating relatively large inter-subject variability in classifier performance. Only the strong FXvDM contrast shows consistent separation between different models. However, there are a few reliable trends; the significant differences are displayed as critical difference diagrams in Figure 4. Comparing regularizers, L_1 and L_2 maximize median prediction, and PCA generally performs worse. The effects of ICA are the most classifier-dependent. For LD and LR, ICA is often comparable to L_p-norm regularizers; for SVM it only weakly out-performs PCA, and for QD it is the least predictive model. Comparing different classifiers, trends in performance depend on the chosen regularizer: for L_p-norms, SVM > LR,LD > QD, whereas for PCA we generally observe LD > QD > LR > SVM. These trends are ambiguous for the weaker task contrasts. Interestingly, the prediction power curves do not increase monotonically with sample size, but plateau at $N_{block} = 4$ to 8 for many models. This is evidence of within-run heterogeneity: as scans are added to the block, they increase the variability of the data distribution, making it more difficult to build a generalizable, predictive model.

Figure 3 plots median spatial reproducibility and gSNR for each classifier/regularizer model, as a function of sample size. We observe much stronger trends in the effect of regularizer choice compared to prediction, with non-overlapped inter-quartile error bars. In general, PCA > L_p > ICA, and PCA shows greater relative benefit in the weaker task contrasts. We also observe a "cross-over" effect between L_1 and L_2: for weaker contrasts, L_1 has greater reproducibility for smaller sample sizes (typically $N_{block} < 4$), but for larger sample sizes it becomes worse than L_2. Therefore, the relative stability of the L_p-norm regularizers is sample-size dependent. The ICA model generally gives poor reproducibility and gSNR, and for QD and SVM classifiers, and gets worse with increased sample size, indicating that it is prone to over-fitting these more complex decision functions. Comparing across classifiers, the choice in classifier model has a much smaller impact on reproducibility and gSNR than regularizer choice, as classifiers with the same regularization tend to have similar performance. Finally, with the exception of ICA, the reproducibility/gSNR curves increase monotonically with sample size and as a function of condition contrast, whereas a much weaker trend with contrast is observed in the prediction plots (Fig. 2). For example, SVM-PC prediction plateaus at $N_{block} = 4$ to 8 for all contrasts, but reproducibility monotonically rises with increasing sample size, showing continued improvement in spatial SNR that may be uncoupled from prediction performance.

Figure 4 provides statistical significance tests on the relative performance of the different classifier models. For prediction P_{test} (left), many models are not statistically distinguishable, with increasing model overlap for weaker task contrasts. In general, models regularized with L_1 and L_2 perform better than PCA, and ICA is highly variable depending on classifier and task contrast. The SVM-L_1, SVM-L_2, and LD-ICA models have consistently optimal prediction. For reproducibility and gSNR (right), a greater number of models are significantly different. PC-based models are not significantly different, and consistently outperform all other models. L_1 and L_2 models offer moderate performance, and become more similar for weaker contrasts, while IC gives lowest performance. The SVM-L_1 and SVM-L_2 models are unique, as they have moderate R and gSNR for strong task contrasts, but consistently decrease in rank for weaker contrasts.

Tables II and III compare P_{valid} versus P_{test} for the 16 different models as a function of task contrast, for sample sizes of $N_{block} = 1$ and 16, respectively. At the small-sample limit (Table II), there is a strong contrast and regularizer-dependent effect. For FXvDM, all models are significantly biased, for ATvPM only PC-based models (and QD-L_2) are not significantly biased, and for PMvDM, no models are significantly biased. At the large-sample limit (Table III), there is an interaction between classifiers and task contrast. For the strongest contrast (FXvDM), LD models are consistently biased,

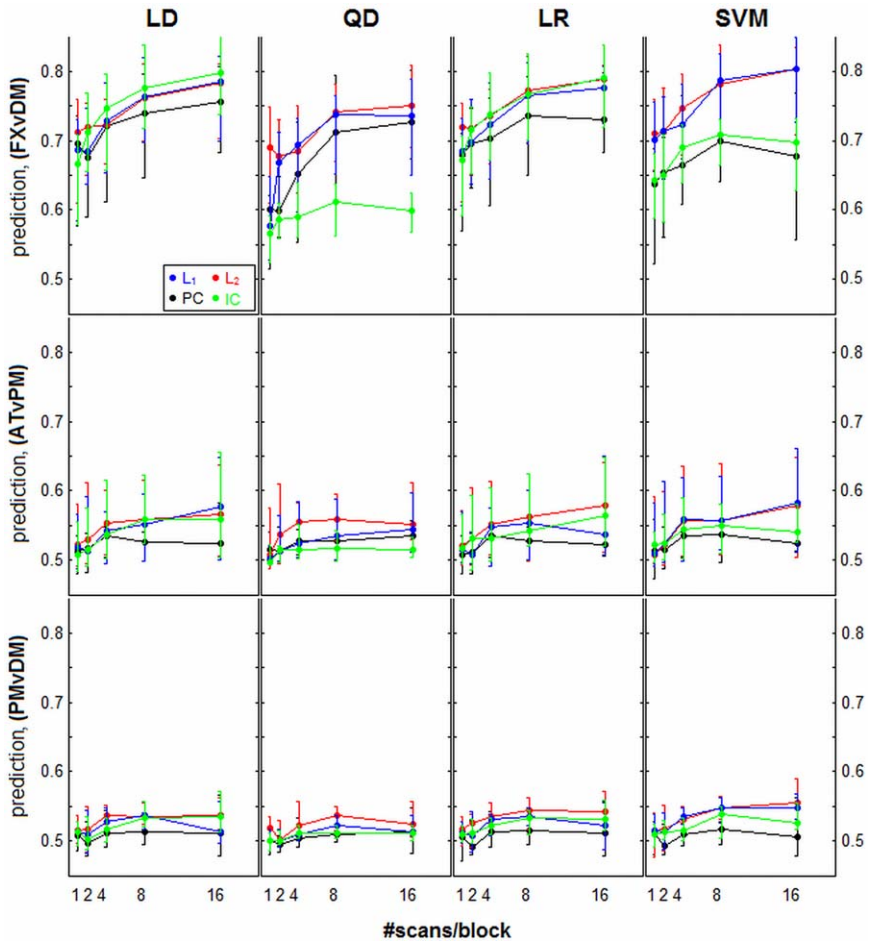

Figure 2.

Prediction accuracy curves as a function of sample size, for different regularizer/classifier model combinations. Panels represent different classifiers of linear discriminant (LD), quadratic discriminant (QD), logistic regression (LR), and support vector machines (SVM). Each curve represents median prediction across subjects (with interquartile error bars), for a different regularization method. Results are shown for three different task contrasts, going from strongest (top row; FXvDM) to weakest (bottom row; PMvDM). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

whereas for the weakest contrast (PMvDM), SVM models are all significantly biased. However, the magnitude of upward prediction biases tends to be relatively small; in the most extreme case of FXvDM and $N_{\text{block}} = 1$, the average magnitude of $(P_{\text{valid}} - P_{\text{test}})$ is 0.06 ± 0.02 (mean $\pm$ standard deviation, over all subjects and models; a mean change of 8%).

The Dynamic Range of Classifier Models

Figure 5 plots the median (P_{test}, R) values across subjects for a representative LD classifier and ATvPM task contrast, for the four different regularization methods. Results are shown for three different optimization criteria: (A) maximized P_{valid}, (B) minimized $D(P_{\text{valid}}, R)$ and (C) maximized R. PCA regularization consistently produces the greatest range in (P_{test}, R) values as a function of optimization criterion. If optimized with P_{valid} (Fig. 5A), it is both the most predictive and reproducible model; if optimized with $D(P_{\text{valid}}, R)$ (Fig. 5B), it dramatically increases reproducibility at the expense of decreased prediction; if optimized with R (Fig. 5C),we see a further small mean increase in R and decrease in P_{test}. In contrast, all of the other regularizers (L_1, L_2, and ICA) are highly consistent between the three panels of Figure 5, indicating that choice of optimization metric has little effect on model performance for these regularizers, which all have relatively low gSNR/R values. These trends were consistently observed for all task contrasts and classifier models.

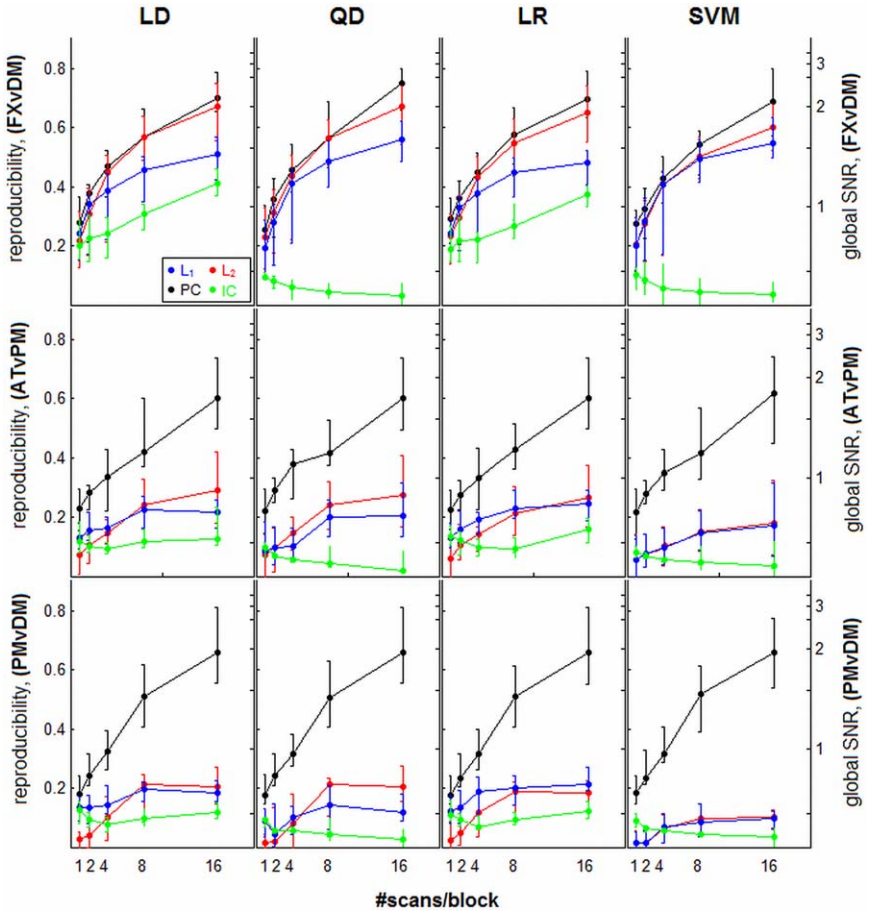

Figure 3.

Spatial reproducibility and global signal-to-noise (gSNR) are both plotted as a function of sample size, for different regularizer/classifier model combinations; gSNR is a monotonic function of reproducibility. Panels represent different classifiers of linear discriminant (LD), quadratic discriminant (QD), logistic regression (LR), and support vector machines (SVM). Each curve represents median reproducibility across subjects (with interquartile error bars), for a different regularization method. Results are shown for three different task contrasts, going from strongest (top row; FXvDM) to weakest (bottom row; PMvDM). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Spatial Pattern Effects

Figure 6 shows the DISTATIS-space plots; these are two-dimensional representations of the relative similarity between SPMs of the 16 different models. For all task contrasts, we observe a "U-shaped" trend. Dimension 1 (horizontal axis; axis of greatest variance) shows that the largest difference in SPMs is between PCA and ICA-regularized models. The second dimension of effect (vertical axis) expresses a contrast between component-regularized (PCA, ICA) versus L_p-norm models. The primary difference between models is again driven by regularizers, and the ellipses of classifier models for a fixed regularizer are significantly overlapped. The patterns of L_1 and L_2-regularized models also tend to be highly similar. The only outlier in this trend is the SVM-L_p models, which tend to be more similar to ICA-based models than the other classifiers.

Figure 7 plots the average Z-scored SPMs for three representative models that are most different in DISTATIS space. For a fixed LD classifier, we show SPMs for L_1, PCA and ICA regularization. PCA consistently provides the most extensive and highest magnitude activation Z-scores. There is a contrast-dependent effect: the mean FXvDM patterns are relatively similar across models, although the LD-PC model is more sensitive to decreases in mean activation for DM (increases for FX) that reflect default network-like regions. For weaker contrasts, LD-PC consistently produces high, spatially extensive Z-scores, again with strong activation decreases that reflect default-network like regions in the states of lower task

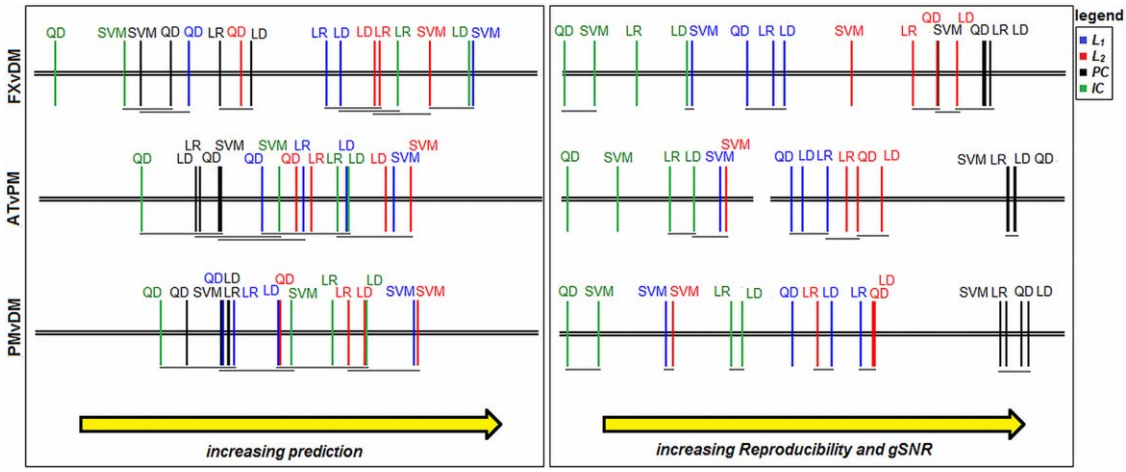

Figure 4.

Critical difference diagrams. Each plot shows average ranking of the 16 different classifier/regularizer models for metrics of Prediction, Reproducibility and gSNR (note that Reproducibility and gSNR have the same rankings, as they are monotonically related; see Resampling, Performance Metrics, and Regularizer Optimization), computed across subjects. All rankings are significant ($P < 0.01$, Friedman test), and models that are not significantly different are connected by horizontal grey bars ($\alpha = 0.05$, Nemenyi test). Results are shown for the maximum number of datapoints ($N_{block} = 16$), for each of the three task contrasts. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

engagement (AT in ATvPM; PM in PMvDM), but these activations are much weaker for other models.

The Effects of Voxel ROI Selection

Figure 8 plots median prediction accuracy P_{test} (Fig. 8A) and median reproducibility (Fig. 8B) across subjects for each classifier/regularizer model, as a function number of input voxels. As the number of voxels decreases, median P_{test} also decreases, and the difference between classifier/regularizer choices is reduced. However, L_p-norm regularizers still have generally higher median performance than PC and IC. In contrast, R is higher for smaller voxel ROIs, and the differences between regularizers remain consistent across ROI size, with PC > L_p > IC. Figure 8C demonstrates the importance of spatial reproducibility for ROI-based analysis. We show sample SPMs from 3 representative subjects for a "conservative" ROI mask. The reproducible LD-PC model produces clear regions of focal activation, with high Z-scored reliability, showing differential lateralization across subjects. The LD-IC model patterns are highly diffuse, with no statistical confidence (very low Z-scores) in the pattern of BOLD activity.

DISCUSSION

We compared a set of commonly-used classification and regularization techniques, for single-subject analyses over a range of different experimental task contrasts, with model performance quantified via metrics of prediction accuracy (P), spatial reproducibility (R), and global Signal-to-Noise Ratio (gSNR). For reproducibility and gSNR, the choice of classifier had a relatively small impact, whereas the regularizer choice had large effects that generalized across all tested datasets and classifier models. Conversely, prediction showed complex interactions with classifiers and regularizers, but the choice of model tended to have a smaller effect on prediction accuracy, with greater statistical overlap between models. An important finding in this article is that researchers should consider their regularizer choice at least as carefully as their choice of classifier model, when performing model selection.

Regularizer choices involve a tradeoff between spatial reproducibility versus prediction accuracy, when these metrics are given equal weight during regularization (i.e. optimizing the D metric). A PCA-based model provides a consistently higher reproducibility and gSNR than all other regularizers, irrespective of the classifier model, but generally lower prediction accuracy. The possible advantages of PCA are demonstrated in Figure 7 where such models are better able to detect the more variable, but spatially stable, reduced activity in the states of lower relative task engagement (FX in FXvDM; AT ATvPM; PM in PMvDM), which are heavily overlapped with default network areas. In comparison, L_p-norm regularized models had lower reproducibility and do not show default-mode areas well for the weaker contrasts, but were more predictive. For LD and LR (linear, probabilistic models), ICA was most extreme, with the lowest model reproducibility, but relatively high prediction for LD and LR. This result was unstable for QD and SVM classifiers, where prediction was also low; this is

TABLE II. Comparison of classifier prediction accuracy on validation data P_{valid}, which is used to optimize model regularization, versus independent test data P_{test}

	FXvDM			ATvPM			MvDM		
	P_{valid}	P_{test}	Signif.	P_{valid}	P_{test}	Signif.	P_{valid}	P_{test}	Signif.
LD-L1	0.77	0.68	0.01	0.56	0.51	0.07	0.50	0.51	0.50
LD-L2	0.77	0.72	0.01	0.58	0.53	0.01	0.51	0.52	0.70
LD-PC	0.77	0.68	0.00	0.55	0.52	0.25	0.47	0.50	0.88
LD-IC	0.76	0.71	0.01	0.59	0.52	0.01	0.51	0.50	0.25
QD-L1	0.72	0.67	0.00	0.57	0.51	0.01	0.51	0.50	0.20
QD-L2	0.74	0.68	0.00	0.57	0.54	0.23	0.50	0.50	0.42
QD-PC	0.67	0.60	0.00	0.52	0.51	0.35	0.51	0.50	0.47
QD-IC	0.63	0.59	0.00	0.53	0.51	0.09	0.51	0.50	0.08
LR-L1	0.77	0.70	0.00	0.57	0.51	0.05	0.49	0.51	0.72
LR-L2	0.77	0.72	0.00	0.59	0.53	0.00	0.52	0.53	0.46
LR-PC	0.79	0.70	0.00	0.53	0.51	0.15	0.47	0.49	0.68
LR-IC	0.77	0.72	0.00	0.59	0.53	0.02	0.50	0.51	0.45
SVM-L1	0.77	0.71	0.01	0.60	0.52	0.05	0.51	0.51	0.47
SVM-L2	0.78	0.71	0.02	0.60	0.52	0.07	0.53	0.52	0.40
SVM-PC	0.72	0.65	0.00	0.53	0.52	0.34	0.47	0.49	1.00
SVM-IC	0.72	0.65	0.00	0.57	0.52	0.00	0.52	0.51	0.06

We plot median prediction across subject, for each classifier/regularizer model combination. The models are significantly biased if P_{valid} is consistently higher than P_{test} (significance assessed using non-parametric paired Wilcoxon test; significant tests shaded in grey). Results are shown for the minimum sample size ($N_{block} = 1$), for the three task contrasts.

potentially due to interactions with the complex quadratic- and hinge-loss functions (respectively); regularization with ICA is most sensitive to classifier choice.

The high reproducibility of PCA is consistent with the preprocessing optimization studies of Churchill et al. [2012], which showed that LD-PCA optimization using the *D* metric is more sensitive to reproducibility than prediction. This suggests that the variance-driven PCA subspace is generally robust between data split-halves. This strong spatial stability also generalizes down to relatively small-

TABLE III. comparison of classifier prediction accuracy on validation data P_{valid}, which is used to optimize model regularization, versus independent test data P_{test}

	FXDM			ATvPM			PMvDM		
	P_{valid}	P_{test}	Signif.	P_{valid}	P_{test}	Signif.	P_{valid}	P_{test}	Signif.
LD-L1	0.77	0.73	0.02	0.55	0.52	0.17	0.53	0.51	0.34
LD-L2	0.76	0.73	0.04	0.54	0.53	0.95	0.52	0.51	0.27
LD-PC	0.77	0.76	0.03	0.55	0.52	0.11	0.53	0.51	0.32
LD-IC	0.69	0.68	0.06	0.55	0.52	0.79	0.52	0.50	0.17
QD-L1	0.79	0.79	0.39	0.60	0.58	0.18	0.55	0.54	0.29
QD-L2	0.76	0.75	0.14	0.56	0.55	0.53	0.54	0.52	0.26
QD-PC	0.79	0.78	0.28	0.58	0.57	0.07	0.55	0.54	0.20
QD-IC	0.79	0.80	0.50	0.58	0.58	0.53	0.54	0.55	0.62
LR-L1	0.79	0.78	0.11	0.57	0.54	0.14	0.54	0.52	0.27
LR-L2	0.74	0.74	0.02	0.55	0.54	0.68	0.53	0.51	0.01
LR-PC	0.80	0.79	0.40	0.57	0.58	0.59	0.54	0.53	0.17
LR-IC	0.80	0.80	0.33	0.58	0.58	0.45	0.54	0.55	0.70
SVM-L1	0.79	0.79	0.81	0.56	0.56	0.30	0.56	0.53	0.07
SVM-L2	0.61	0.60	0.12	0.53	0.51	0.05	0.54	0.51	0.02
SVM-PC	0.81	0.80	0.49	0.57	0.56	0.09	0.55	0.53	0.05
SVM-IC	0.70	0.70	0.29	0.54	0.54	0.16	0.54	0.53	0.03

We plot median prediction across subject, for each classifier/regularizer model combination. The models are significantly biased if P_{valid} is consistently higher than P_{test} (significance assessed using nonparametric paired Wilcoxon test; significant tests shaded in grey). Results are shown for the maximum sample size (N_{block}=16), for the three task contrasts.

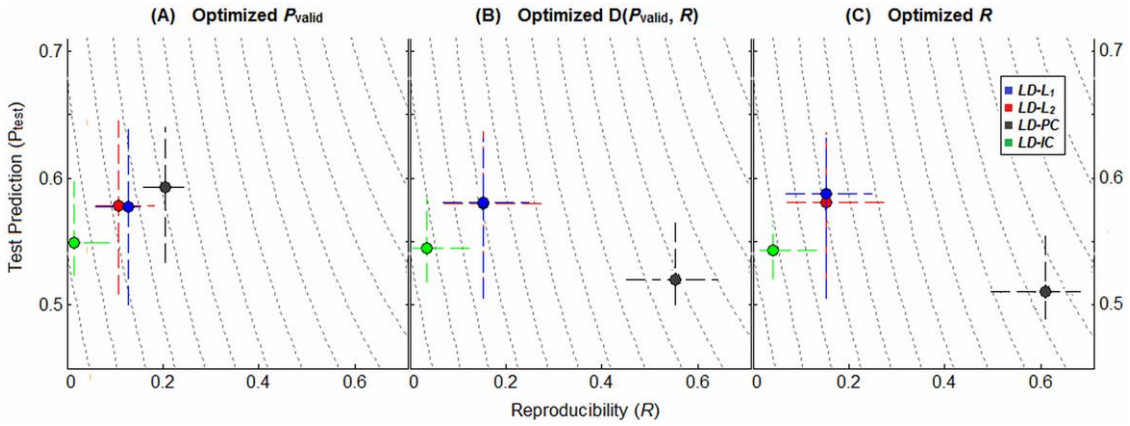

Figure 5.

Plots illustrating the dynamic range of different regularizers in (Prediction, Reproducibility) space, for representative Linear Discriminant (LD) classifier, and the intermediate ATvFX contrast. Points represent the median (P_{test}, R) values for each model, with upper and lower quartile error bars. We show median performance metrics under three different optimization criteria: (**A**) maximized prediction P_{valid}, (**B**) minimization of the $D(P_{valid}, R)$ metric which equally weights P and R (this is the metric used for all previous results), and (**C**) maximized reproducibility R. Results are shown for the maximum sample size ($N_{block} = 16$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

sample SPMs. However, the experimental data of this study has been extensively preprocessed, including ICA-based artifact removal, white matter and CSF regression, and temporal detrending [Grady et al., 2010]. The PCA model is sensitive to high-variance, spatially distributed signal changes, such as motion artifact; therefore, careful preprocessing may be critical to ensure good performance when classifying on a PCA basis sensitive to signal variance [Garrett et al., 2011].

The L_p-norm models had consistently higher prediction accuracy, across a range of task contrasts and sample sizes. The increased prediction at the expense of reproducibility is expected, as L_p-norm analyses operate directly on the linear kernel. The kernel features are non-orthogonal, allowing models to select the optimal predictors from among a redundant set of features, but at the expense of stability in the chosen features. For consistency with PCA/ICA, we applied L_p penalties to the linear kernel, but most other studies apply L_1 and L_2 penalties directly on voxel features. Interestingly, many such studies have shown that L_1 produces excessively sparse brain maps [Carroll et al., 2009; Rasmussen et al., 2012; Ryali et al., 2010]. However,

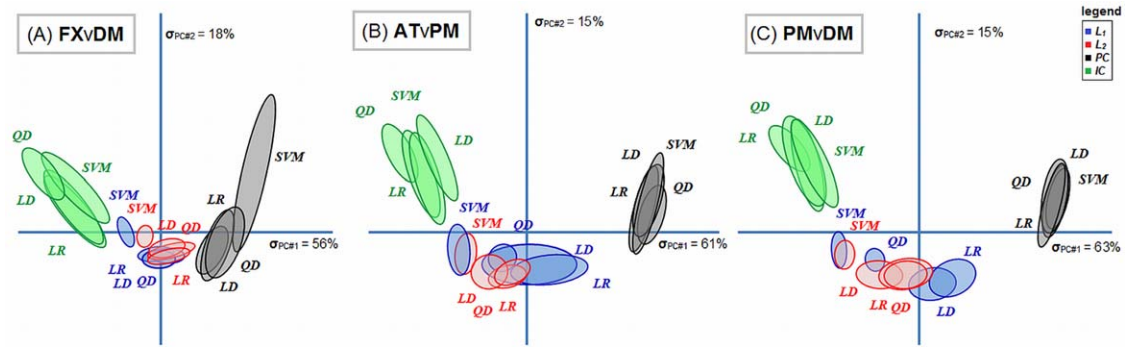

Figure 6.

DISTATIS multidimensional scaling plot, depicting the relative similarity of different classifier/regularizer model SPMs. The ellipses denote the 95% Bootstrap confidence bounds for each model. Ellipses that are closer in DISTATIS space indicate more similar SPM patterns, and models are not significantly different if the ellipses overlap. Colors denote different regularizers: L_1 = blue, L_2 = red, PCA = black, ICA = green. Results are shown for the maximum sample size ($N_{block} = 16$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

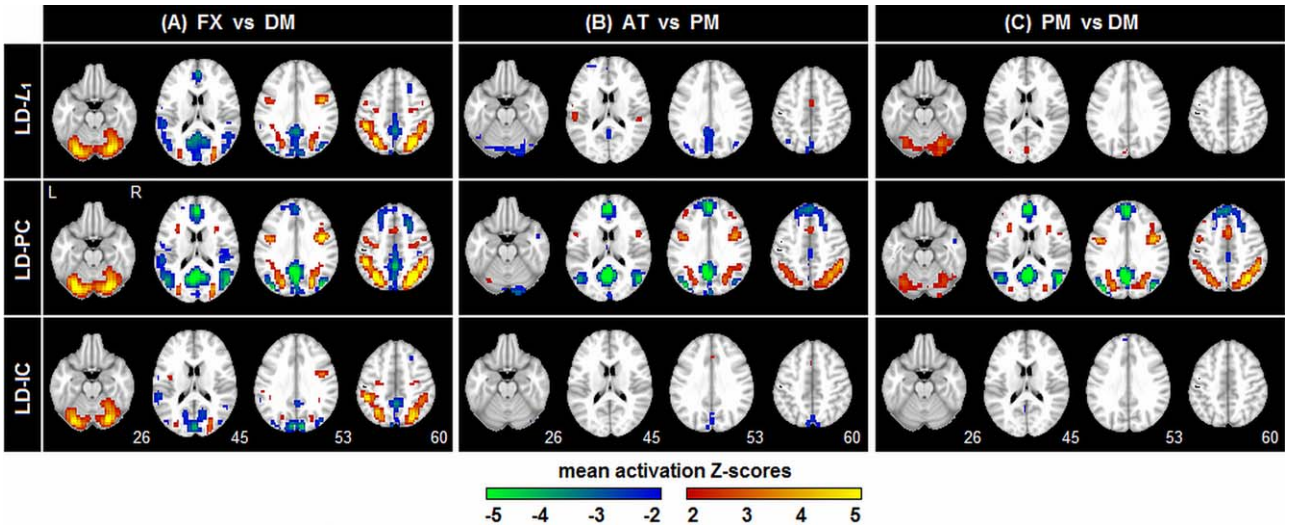

Figure 7.

Average *Z*-scored SPMs, computed across subjects, for three representative classifier/regularizer models, seen in the DISTATIS plot of Figure 6. The SPMs are plotted for each of the three different task contrasts, at the maximum sample size ($N_{block} = 16$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

we found that L_1 penalization on the kernel features only had worse gSNR than L_2 for the strongest task contrast and large sample sizes. It appears that applying L_1 to the kernel features is therefore a more stable implementation of this regularizer.

We also tested ICA, implemented in the FastICA package, and found overall poor reproducibility and gSNR for the resulting classifiers, with high prediction only for LD and LR. The poor reproducibility is not unexpected, given that (a) there is risk of model overfitting in ICA, as it estimates higher-order statistical moments, and (b) its components tend to be unstable (e.g. performing ICA twice may not give the same subspace), often requiring additional computation methods to stabilize the subspace. Since ICA refers to a variety of different model implementations (e.g. enforcing independence on spatial vs. temporal domains, the choice of non-Gaussianity metric, the convergence algorithm, etc.), these results are not definitive. Nonetheless, we have searched over a range of different initial PCA subspaces (retaining up to 50% of a maximum of training 96 scans), and our results are consistent with prior evaluations of FastICA in the MELODIC package, which is a sub-optimal estimator of the signal subspace in both simulated [Yourganov et al., 2011] and experimental data [Churchill et al., 2012].

Our finding that different classifiers offer relatively similar performance for a fixed regularizer is supported in prior literature, for less extensive model comparisons. The LD and SVM classifiers on a linear kernel have comparable classification accuracy for many single-subject analyses [Misaki et al., 2010; Ng and Abugharbieh, 2011; Schmah et al., 2010], although LaConte et al. [2005] showed that this may be a strong function of preprocessing choices. Similarly, Rasmussen et al. [2012] showed for group-level analyses that LD, SVM, and LR on a linear kernel basis produced similar (P, R) values across a broad range of regularization values.

A number of studies also showed significant differences between classifiers, particularly linear versus nonlinear models. For example, [Misaki et al.; 2010] showed better performance for linear models in standard task analyses, and [Schmah et al., 2010] demonstrated significantly better prediction for non-linear models for a subset of contrasts. However, these results were based on optimizing model prediction. Our results, which show non-linear QD to be comparable to other models for most regularizers, suggest that some of the apparent differences may be due to choice in regularizer and optimization criteria, rather than the classifier itself. In the case of Schmah et al. [2010], who examined longitudinal and clinical datasets, it is also possible that there are fundamental differences in classifier performance that are simply not evident for some standard task analyses [Carter et al., 2008].

Along with classifier and regularizer choice, the criterion used to optimize regularization has a strong impact on model performance. The choice of whether to optimize R, P, or D metrics, may produce entirely different measures of model performance (Fig. 7). This has previously been observed in group analysis [Rasmussen et al., 2012], where L_2-regularized classifiers become increasingly similar if the D metric is chosen as the optimization criterion. We show that maximizing P_{valid} produces similar (P_{test}, R) for different regularizers, but including R as an optimization criterion increases the difference between models. Furthermore, PCA-based regularization

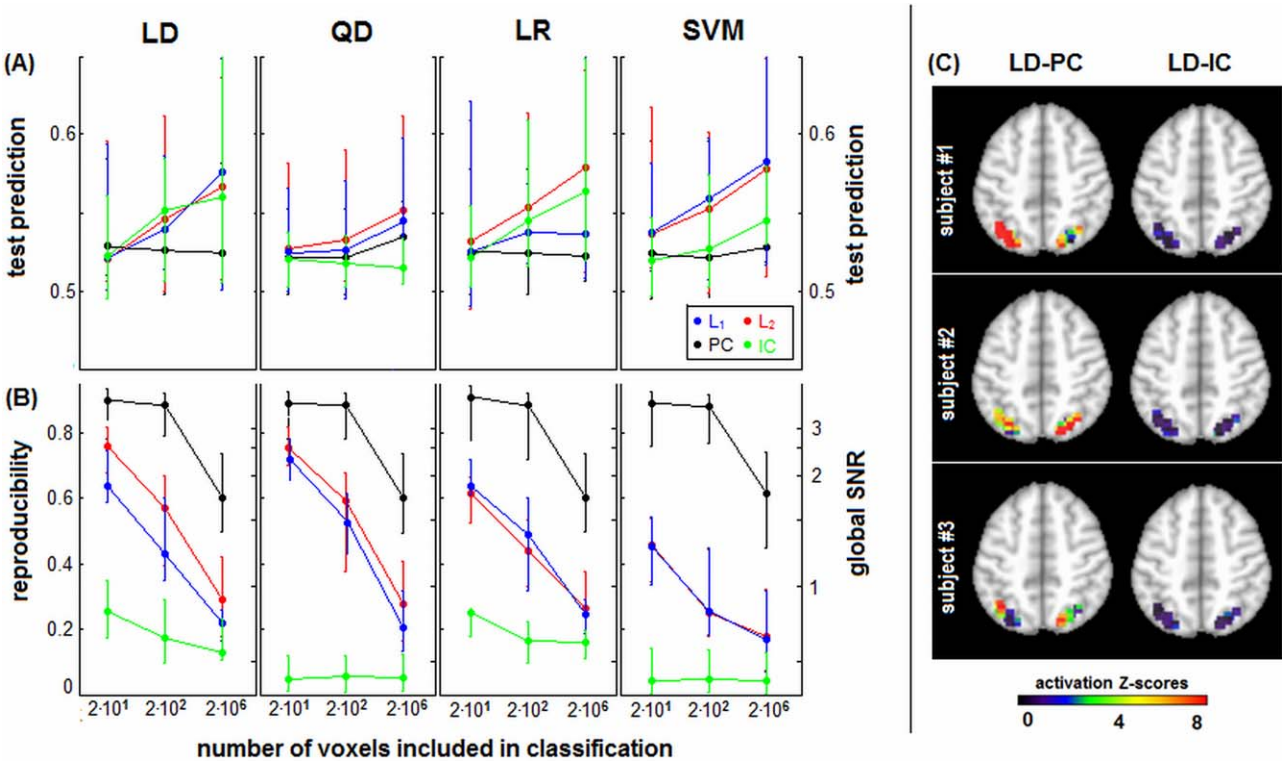

Figure 8.

(**A**) Prediction accuracy and (**B**) spatial reproducibility curves as a function of number of voxels, for different regularizer/classifier model combinations. We compared ROI masks of 24 and 276 voxels, against whole-brain classification (21,535 voxels). Panels represent different classifiers of linear discriminant (LD), quadratic discriminant (QD), logistic regression (LR), and support vector machines (SVM). Each curve represents median prediction across subjects (with interquartile error bars), for a different regularizer. Results are shown for the representative ATvPM contrast, and maximum sample size of $N_{block} = 16$. (**C**) Reproducible *Z*-scored activation maps for three example subjects, showing results of classification analysis for the 29-voxel ROI mask of the parietal lobes. Results are shown for LD-PC and LD-IC classifiers, which have comparable prediction accuracy but the greatest difference in spatial reproducibility. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

expresses the widest "dynamic range" of (P, R) values, indicating that it may be able to explore a greater range of brain states and potential network patterns, ranging from highly reproducible to highly predictive. Moreover, it is possible to extend PCA-based models to explore even wider ranges of (P, R) values, using multistage PCA decompositions that exploit data heterogeneity [Strother et al., in press].

These findings have implications for multiple research domains where the user is performing multivariate classification, including experimental neuroscience, neurofeedback, and clinical fMRI studies. Moreover, we tested a range of different task contrast strengths, sample sizes, and both whole-brain and ROI-based classification, showing that the results are highly generalizable. We conclude that when researchers are choosing a classifier model, they should focus on (a) the type of regularizer and (b) the optimization criterion. For example, users who are primarily interested in classification accuracy and not underlying spatial patterns (e.g. neurofeedback with brain-computer interfaces, many clinical biomarker studies) should either chose an L_p-norm classifier, which has high prediction regardless of optimization criterion, or PC subspace optimized on P_{valid}. If the user is interested in identifying stable, coherent brain networks that may support behavioral performance without requiring optimal classification accuracy on experimental conditions (e.g. neuroscience studies of the temporal dynamics of default-mode networks), a PC subspace optimized on $D(P_{valid}, R)$ is an effective choice. Finally, in studies where both spatial localization and prediction are important (e.g. clinical neuroscience, such as some biomarker studies, pre-surgical mapping), we suggest L_p-norm regularization for results that are insensitive to the optimization criteria but may have a relatively low gSNR, or a carefully tuned PC-space optimization such as available with the NPAIRS software [https://code.google.com/p/plsnpairs/) [e.g. Strother et al., in press].

ACKNOWLEDGMENTS

The authors thank Dr. Cheryl Grady for providing the experimental fMRI data.

REFERENCES

Abdi H, Valentin D, Chollet S, Chrea C (2007): Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. Food Quality Prefer 18:627–640.

Berman MG, Yourganov G, Askren MK, Ayduk O, Casey BJ, Gotlib IH, Kross E, McIntosh AR, Strother S, Wilson NL, Zayas V, Mischel W, Shoda Y, Jonides J (2013): Dimensionality of brain networks linked to life-long individual differences in self-control. Nat Commun 4:1373.

Bishop CM (2006): In: Jordan M, Kleinberg J, Scholkopf B, editors. Pattern Recognition and Machine Learning. Chapter 4: Linear Models for Classification p. 207–208, New York, NY: Springer.

Brodersen KH, Wiech K, Lomakina EI, Lin C, Buhmann JM, Bingel U, Ploner M, Stephan KE, Tracey I (2012): Decoding the perception of pain from fMRI using multivariate pattern analysis. Neuroimage 63:1162–1170.

Carlson TA, Schrater P, He S (2003): Patterns of activity in the categorical representations of objects. J Cog Neurosci 15:704–717.

Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR (2009): Prediction and interpretation of distributed neural activity with sparse models. Neuroimage. 44:112–122.

Carter CS, Heckers S, Nichols T, Pine DS, Strother S (2008): Optimizing the design and analysis of clinical functional magnetic resonance imaging research studies. Biol Psychiatry 64:842–849.

Churchill NW, Yourganov G, Oder A, Tam F, Graham SJ, Strother SC (2012): Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. PLoS One 7:e31147.

Conover WJ (1999): Practical Nonparametric Statistics, 3rd ed. Weinheim: Wiley.

Cristianini N, Shawe-Taylor J (2000): An Introduction to Support Vector Machines. Cambridge MA: Cambridge University Press.

Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, Gur RC, Langleben DD (2005): Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. Neuroimage 28:663–668.

Demirci O, Clark VP, Calhoun VD (2008): A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. Neuroimage 4:1774–1782.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004): Least angle regression. The Annals of statistics, 32(2):407–499.

Fleisher AS, Sherzai A, Taylor C, Langbaum JBS, Chen K, Buxton RB (2009): Resting-state BOLD networks versus task-associated functional MRI for distinguishing Alzheimer's disease risk groups. Neuroimage 47:1678–1690.

Friedman J, Hastie T, Tibshirani R (2008): Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441.

Fu CH, Mourao-Miranda J, Costafreda SH, Khanna A, Marquand AF, Williams SCR, Brammer MJ (2008): Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. Biol Psychiatry 2008;63:656–662.

Garrett DD, Kovacevic N, McIntosh AR, Grady CL (2011): The importance of being variable. J Neurosci 31:4496–4503.

Grady CL, Protzner AB, Kovacevic N., Strother SC, Afshin-Pour B, Wojtowicz M, Anderson JAE, Churchill N, McIntosh AR (2010): A multivariate analysis of age-related differences in default mode and task-positive networks across multiple cognitive domains. Cereb Cortex 20:1432–1447.

Greicius MD, Srivastava G, Reiss AL, Menon V (2004): Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. Proc Natl Acad Sci USA 101:4637–4642.

Hastie T, Buja A, Tibshirani R (1995): Penalized discriminant analysis. Ann Stat 23:73–102.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001): Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.

Hyvärinen A, Oja E (2000): Independent component analysis: Algorithms and applications. Neural Netw 13:411–430.

Kjems U, Hansen LK, Anderson J, Frutiger S, Muley S, Sidtis J, Rottenberg D, Strother SC (2002): The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. Neuroimage 15:772–786.

Kovacevic N, Henderson JT, Chan E, Lifshitz N, Bishop J, Evans AC, Henkelman RM, Chen XJ (2005): A 3D atlas of the mouse brain with estimates of the average and variability. Cereb Cortex 15:639–−645.

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

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

Lukic AS, Wernick MN, Strother SC (2002) An evaluation of methods for detecting brain activations from functional neuroimages. Artif Intell Med 25:69–88.

Menzies L, Achard S, Chamberlain SR, Fineberg F, Chen CH Campo ND, Sahakian BJ, Robbins TW, Bullmore Ed (2007): Neurocognitive endophenotypes of obsessive-compulsive disorder. Brain 130:3223–3236.

Misaki M, Kim Y, Bandettini P, Kriegeskorte N.(2010): Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53:103–118.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004): Learning to decode cognitive states from brain images. Machine Learn 57:145–175.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA (2008): Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195.

Modinos G, Pettersson-Yeo W, Allen P, McGuire PK, Aleman A, Mechelli A (2012): Multivariate pattern classification reveals differential brain activation during emotional processing in individuals with psychosis proneness. Neuroimage 59:3033–3041.

Moeller JR, Strother SC (1991): A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. J Cereb Blood Flow Metab 11:121–135.

Mørch N, Hansen LK, Strother SC, Svarer C, Rottenberg DA, Lautrup B, Savoy R, Paulson OB (1997): Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. Lecture Notes Comput Sci 1230:259–270.

Ng B, Abugharbieh R (2011): Generalized sparse regularization with application to fMRI brain decoding. Lecture Notes Comput Sci 6801:612–623.

Ogawa S, Menon RS, Kim SG, Ugurbil K (1998): On the characteristics of functional magnetic resonance imaging of the brain. Annu Rev Biomol Struct 27:447–474.

Pagani M, Salmaso D, Rodriguez G, Nardo D, Nobili F. (2009): Principal component analysis in mild and moderate Alzheimer's disease—A novel approach to clinical diagnosis. Psychiatry Res 173:8–14.

Pereira F, Mitchell T, Botvinick M. (2009): Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45:S199–S209.

Raizada RDS, Tsao FM, Liu HM, Holloway ID, Ansari D (2010): Linking brain-wide multivoxel activation patterns to behaviour: Examples from language and math. Neuroimage 51:462–471.

Rasmussen, PM, Hansen LK, Madsen KH, Churchill NW, Strother SC (2012): Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognit 45(6): 2085–2100.

Rasmussen PM, Abrahamsen TJ, Madsen KH, Hansen LK (2012): Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation. Neuroimage 60:1807–1818.

Rottenberg DA, Moeller JR, Strother SC, Sidtis JJ, Navia BA, Dhawan V, Ginos JZ, Price RW (1987): The metabolic pathology of the AIDS dementia complex. Ann Neurol 22:700–706.

Ryali S, Supekar K, Abrams DA, Menon V (2010): Sparse logistic regression for whole brain classification of fMRI data. Neuroimage 51:752–764.

Saur D. Ronneberger O, Kummerer D, Mader I, Weiller C, Kloppel S (2010): Early functional magnetic resonance imaging activations predict language outcome after stroke. Brain 133:1252–1264.

Schmah T, Yourganov G, Zemel RS, Hinton GR. (2010): Comparing classification methods for longitudinal fMRI studies. Neural Comput 22:2729–2762.

Shaw ME, Strother SC, 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.

Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA (2007): Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS ONE 3: e1394.

Sitaram R, Lee S, Ruiz S, Rana M, Veit R, Birbaumer N (2010): Real-time support vector classification and feedback of multiple emotional brain states. Neuroimage 56:753–765.

Strother SC, Anderson J, Hansen LK, Kjems U (2002): The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. Neuroimage 15:747–771.

Strother SC, Rasmussen P, Churchill N, Hansen LK (in press): In: Rish I, Cecchi G, Lozano A, Niculescu-Mizil A, editors. Stability and Reproducibility in fMRI Analysis in Practical Applications of Sparse Modeling. p. 111–136, Cambridge MA: MIT Press.

Wang K, Liang M, Wang L, Tian L, Zhang X, Li K, Jiang T (2006): Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study. Hum Brain Mapp 28:976–978.

Yoo SS, Fairneny T, Chen NK, Choo SE, Panych LP, Lee S-Y, Jolesz FA (2004): Brain-computer interface using fMRI: spatial navigation by thoughts, Neuroreport 15:1591–1595.

Yoon JH, Tamir D, Minzenberg MJ, Ragland JD, Ursu S, Carter CS (2008): Multivariate pattern analysis of functional magnetic resonance imaging data reveals deficits in distributed representations in schizophrenia. Biol Psychiatry 64:1035–1041.

Yourganov G, Schmah T, Small SL, Rasmussen PM, Strother SC (2010): Functional connectivity metrics during stroke recovery. Arch Ital Biol 148:259–270.

Yourganov, G, Chen, X, Lukic, AS, Grady, CL, Small, SL, Wernick MN, Strother SC (2011): Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data. Neuroimage 56:531–543.

Zhang J, Anderson J, Liang L, Pulapura S, Gatewood L, Rottenberg DA, Strother SC (2009): Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. Magn Reson Imag 27:264–278.

Zou H, Hastie T (2005): Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301–320.