Classification on Brain Functional Magnetic Resonance Imaging: Dimensionality, Sample Size, Subject Variability and Noise

Jean Honorio
CSAIL, MIT
Cambridge, MA 02139, USA
jhonorio@csail.mit.edu

Abstract

Our aim in this chapter is to study the conditions for the reasonably good performance of classifiers on brain functional magnetic resonance imaging. We propose a synthetic model for the systematic study of aspects such as dimensionality, sample size, subject variability and noise. Our simulations highlight the key factors that affect generalization accuracy.

1 Introduction

Functional magnetic resonance imaging (fMRI) has become one of the methods of choice for looking at the activity of the human brain. Neural activity is followed by an increase in blood flow and oxigenation in the local vasculature. This phenomenon is called *hemodynamic response* and it is used by fMRI in order to undirectly measure neural activity.

Classification on brain fMRI faces various challenges. fMRI data is three-dimensional and thus, datasets are very high dimensional. Typical datasets contain tens of thousands of voxels. Due to the cost and time needed in order to capture fMRI along with other imaging modalities and clinical data, the number of available subjects is small. Usually, datasets have only a few tens of subjects. Many datasets also show high subject variability, depending upon the nature of the neuropsychological process and task. Additionally, it is well known that fMRI signal is noisy.

Classification has been used on brain fMRI for two goals: the *prediction of cognitive states* and *group classification*. In the *prediction of cognitive states* [1–15], the goal is to infer the experimental condition (e.g. calculation vs. reading) that a given subject was undergoing at a given time. In *group classification* [16–23], the goal is to infer the group membership (e.g. cocaine addicted vs. control) of a given subject.

Besides using of all voxels as features [5, 8, 10, 14], classifiers are also trained with features extracted from the original data. Feature extraction

methods can be categorized into two groups. Some of the methods extract features that are a weighted combination from all the available voxels by using different schemes, such as principal component analysis [6, 13, 20, 21], independent component analysis [16, 17, 19] and a coarse resolution of the original image [3]. The remaining methods select a subset of voxels by following different criteria, such as most discriminative voxels [12], most active voxels [1,2,9,12,15,17,18], searchlight accuracy [24], mutual information [11], threshold-split region [22, 23] and recursive feature elimination [4, 7].

Several classifiers commonly encountered in the machine learning literature have been proposed for brain fMRI data, such as Gaussian naïve Bayes [12], k-nearest neighbors [15], Fisher linear discriminant [18–20], logistic regression [14], linear support vector machines [1,2,4,5,7,8,11–13,15,21], Gaussian support vector machines [3,6,9], Adaboost [10], random forests [16], neural networks [17] and majority voting [22, 23].

Some of the proposed classification frameworks use a predefined set of regions of interest (ROIs) [2, 5, 10, 15, 17, 19]. As noted in [23], a possible drawback in these frameworks is that the practitioner needs either prior knowledge of the underlying neuropsychological process or an additional independent dataset in order to find the set of ROIs. If one selects the set of ROIs from the same dataset (*double dipping*) the significance of the cross-validation results is compromised [25].

Our goal in this chapter is to study the conditions for the reasonably good performance of linear support vector machines (SVMs) on brain fMRI data. To this end, we propose a synthetic model for the systematic study of aspects such as dimensionality, sample size, subject variability and noise. We believe this study is a first step in understanding the key factors that affect generalization accuracy in brain fMRI data.

We chose to focus on linear SVMs since it has been largely used for group classification as well as the prediction of cognitive states [1, 2, 4, 5, 7, 8, 11–13, 15, 21]. Other linear classifiers such as the Fisher linear discriminant and logistic regression have also been proposed in the literature [14, 18–20]. Besides using all voxels as features [5, 8, 10, 14], we chose a method that extracts features that are a weighted combination from all the available voxels, namely principal component analysis [6, 13, 20, 21]. We also chose to evaluate a feature extraction method that selects a subset of voxels, namely the most discriminative voxels [12].

2 Materials and Methods

In this section, we present our synthetic model, its parameters and the procedure for data generation. We also describe the feature extraction methods used in our experiments, as well as the techniques used for estimating the generalization accuracy of the classifiers.

Our synthetic model is based on a number of Gaussian-distributed regions with Gaussian spatially-correlated noise. Despite our simplifying assumptions, we believe our model introduces many intuitions from neuro-

science. First, we assume that a number of brain regions are involved in a specific activity, such that their activation level differs between two classes. (Class refers to a group in *group classification* or a experimental condition in the *prediction of cognitive states*.) Second, it is well known that brain fMRI data contains spatially-correlated noise and that there is high subject variability, and thus we introduce these elements to our model. Finally, our experimental setting assumes high dimensionality of the feature vector and a small number of samples.

For simplicity of presentation, we consider a *one-dimensional* brain. Note that most classification algorithms (including linear SVMs) do not take into account the three-dimensional structure of the data. In fact, voxels are treated as one-dimensional. Our synthetic model has the following parameters:

- the number of original features: F,
- the number of samples per class: S,
- the number of involved brain regions: R,
- the distance between the means of the classes: μ_{signal},
- the variance per class: σ^2_{signal} (both parameters μ_{signal} and σ^2_{signal} allow for modeling subject variability),
- the radius of the involved brain regions: r_{signal},
- the noise variance: σ^2_{noise} and
- the radius of the spatially-correlated noise: r_{noise}.

We generate a complete synthetic dataset with F features and $2S$ samples (each class contains S samples) by using the following procedure:

1. We select R features from $\{1, \ldots, F\}$ uniformly at random.
2. For each of the S samples in class $\in \{-1, +1\}$:
   (a) We create an F-dimensional "signal vector" as follows: Each of the R selected features is independently sampled from a Gaussian distribution with mean $+\mu_{\text{signal}}/2$ for class $+1$ ($-\mu_{\text{signal}}/2$ for class -1) and variance σ^2_{signal}. The remaining $F-R$ unselected features are set to zero. After this vector has been created, we smooth it with a Gaussian filter of radius r_{signal} and standard deviation $r_{\text{signal}}/2$. (We normalize the Gaussian filter so that its center has weight 1 and thus, each of the R selected features in the signal vector retains its variance.)
   (b) We create an F-dimensional "noise vector" as follows: Each of the F features is independently sampled from a Gaussian distribution with mean zero and variance σ^2_{noise}. After this vector has been created, we smooth it with a Gaussian filter of radius r_{noise} and standard deviation $r_{\text{noise}}/2$. (We normalize the Gaussian filter so that it has unit ℓ_2-norm and thus, each of the F features in the noise vector retains its variance.)
   (c) The generated sample is the summation of the "signal vector" from step (2a) the "noise vector" from step (2b).

We used LIBLINEAR [26] in order to train linear SVMs with ℓ_2^2-regularization, hinge loss and soft-margin parameter $C = 1$. Additionally, we use the following feature extraction methods:

- *Original features.* We use the F generated features from the above-described procedure.
- *Principal component analysis (PCA) features.* [6,13,20,21] We perform singular value decomposition of the F original features and use all available components.
- *Most discriminative features.* [12] We first rank each of the F original features independently with a Gaussian classifier and then select the top 100 performing features.

In order to estimate the generalization accuracy of the classifiers, we rely on three different methods:

- *k-fold cross-validation.* We hold out S/k samples in turn while training on the other $S(k-1)/k$ samples. The held out samples are used for measuring the classification accuracy. (We chose $k = 5$.)
- *.632 Bootstrapping.* For each of B independent repetitions, we perform the following procedure. From the S samples in the dataset, we pick S random samples with replacement as training set. (The training set has approximately $0.632\,S$ unique samples.) After training, we measure the classification accuracy for the samples not in the training set. The final estimator is an average between the above quantity and the classification accuracy by using the whole dataset of S samples for training and testing. We refer the interested reader to [27]. (We chose $B = 5$.)
- *Independent set.* After training in the whole dataset of S samples, we measure the classification accuracy in an independent set of 1000 samples per class.

The latter method is an unbiased estimator which we use only for assessing the quality of the former two biased methods, which are used in practice. Indeed, the independent-set-of-samples method is impractical for most brain fMRI problems.

3 Results

We perform several synthetic experiments in order to analyze the different aspects of brain fMRI data. For each experiment, we change one parameter while keeping the rest of the parameters fixed to a default value. In order to obtain a measure of confidence of the results, we perform 20 repetitions and report error bars at 90% significance level. Next, we show the set of values for each parameter (the default values are in parentheses): $F \in \{500,\ 1000,\ 2000,\ 5000,\ 10000,\ 20000,\ (50000)\}$, $S \in \{10,\ (20),\ 50,\ 100,\ 200,\ 500,\ 1000\}$, $R \in \{1, 2, 3, (4), 5, 6, 7\}$, $\mu_{\text{signal}} \in \{0.5, 1, 1.5, (2), 2.5, 3, 3.5\}$, $\sigma^2_{\text{signal}} \in \{0.05,\ 0.1,\ 0.2,\ 0.5,\ (1),\ 2,\ 5\}$, $r_{\text{signal}} \in \{1,\ 2,\ 3,\ 4,\ 5,\ (6),\ 7\}$, $\sigma^2_{\text{noise}} \in \{0.05,\ 0.1,$

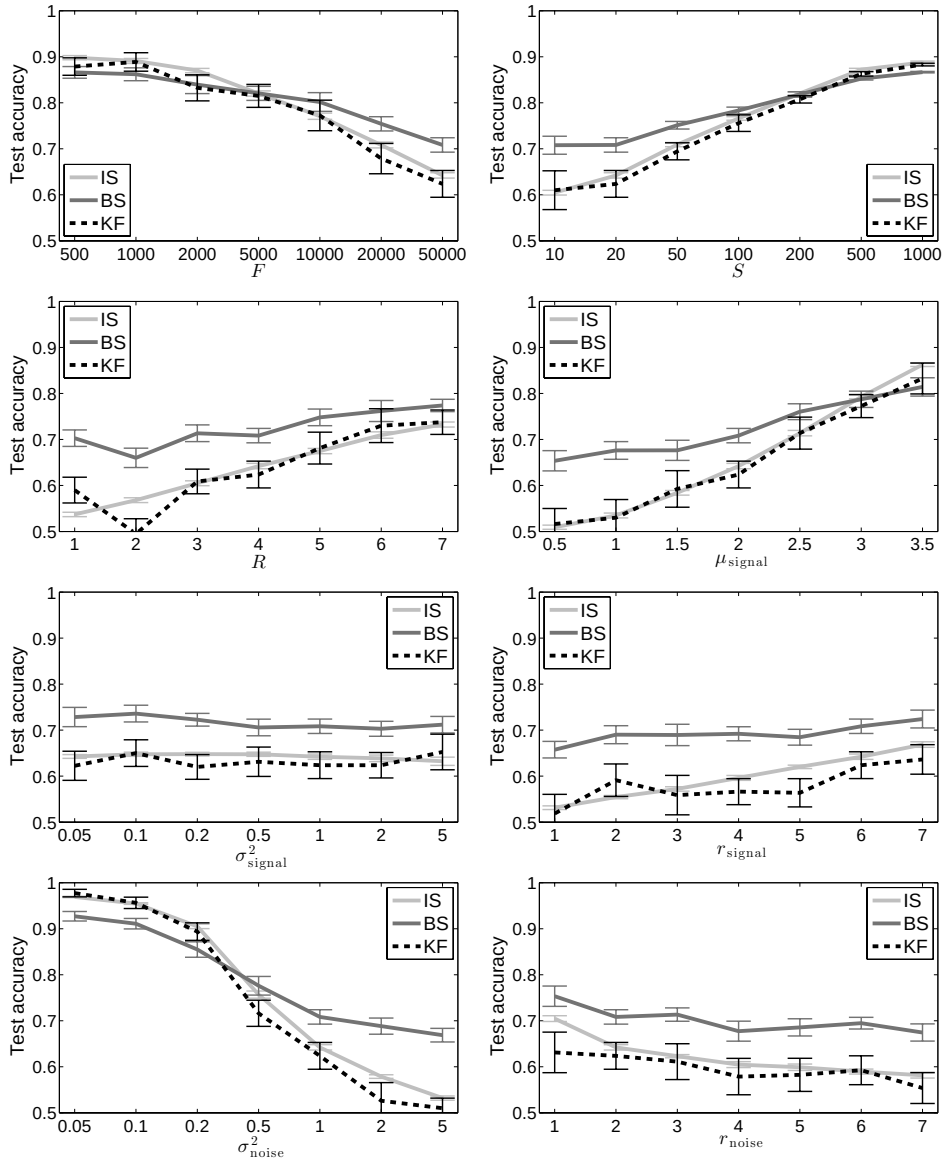

Figure 1: Generalization accuracy for linear SVMs with the original features, for a different: number of original features (F), number of samples per class (S), number of involved brain regions (R), distance between means of the classes (μ_{signal}), variance per class (σ^2_{signal}), radius of the involved brain regions (r_{signal}), noise variance (σ^2_{noise}) and radius of the spatially correlated noise (r_{noise}). Default values were set to $F = 50000$, $S = 20$, $R = 4$, $\mu_{\text{signal}} = 2$, $\sigma^2_{\text{signal}} = 1$, $r_{\text{signal}} = 6$, $\sigma^2_{\text{noise}} = 1$ and $r_{\text{noise}} = 2$. We also include error bars at 90% significance level. Note that k-fold cross-validation (KS) is a better estimator than .632 bootstrapping (BS), since it is always closer to the unbiased independent-set-of-samples method (IS). Note that generalization accuracy is increasing with respect to S, R, μ_{signal} and r_{signal}. It is also decreasing with respect to F, σ^2_{noise} and r_{noise}. Generalization accuracy does not significantly change with respect to σ^2_{signal}.

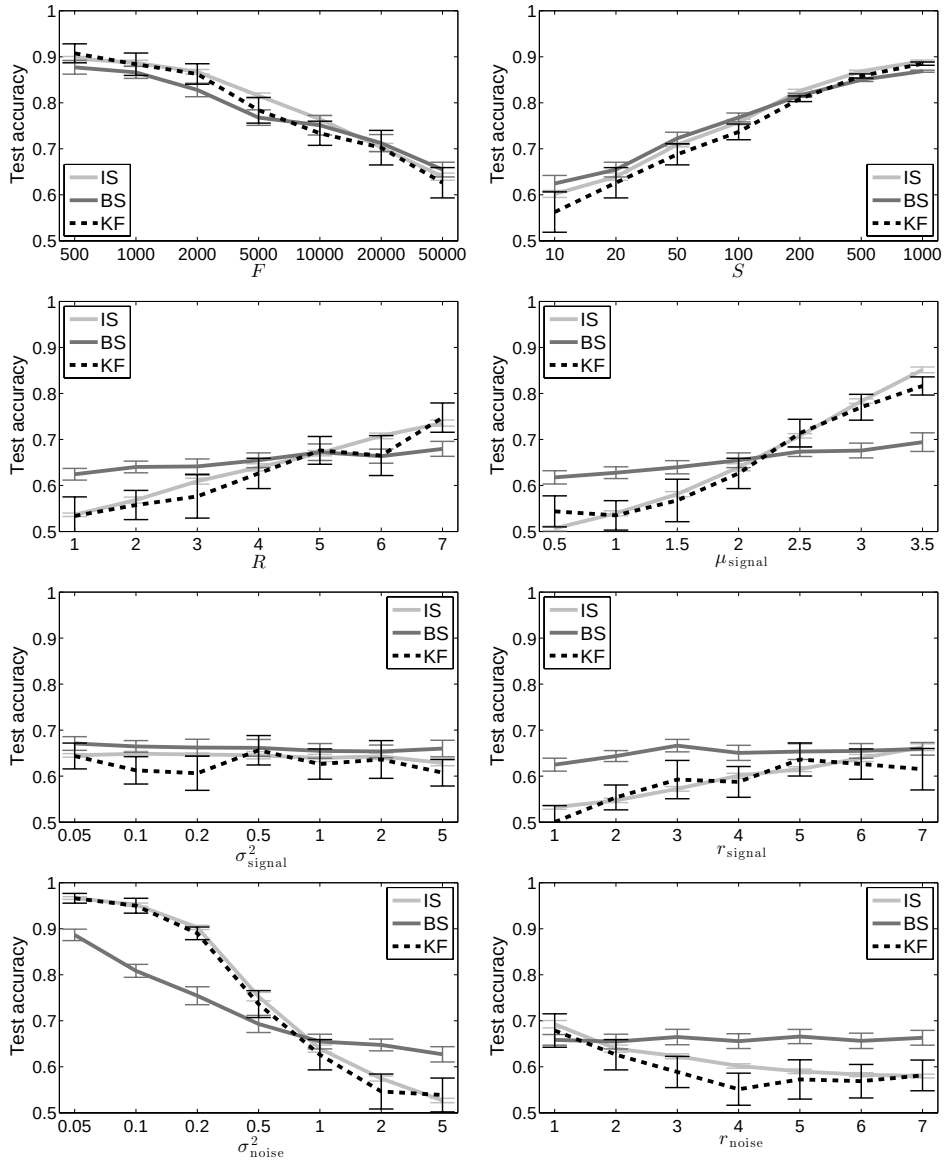

Figure 2: Generalization accuracy for linear SVMs with PCA features, for a different: number of original features (F), number of samples per class (S), number of involved brain regions (R), distance between means of the classes (μ_{signal}), variance per class (σ^2_{signal}), radius of the involved brain regions (r_{signal}), noise variance (σ^2_{noise}) and radius of the spatially correlated noise (r_{noise}). Default values were set to $F = 50000$, $S = 20$, $R = 4$, $\mu_{\text{signal}} = 2$, $\sigma^2_{\text{signal}} = 1$, $r_{\text{signal}} = 6$, $\sigma^2_{\text{noise}} = 1$ and $r_{\text{noise}} = 2$. We also include error bars at 90% significance level. Note that k-fold cross-validation (KS) is a better estimator than .632 bootstrapping (BS), since it is always closer to the unbiased independent-set-of-samples method (IS). Note that generalization accuracy is increasing with respect to S, R, μ_{signal} and r_{signal}. It is also decreasing with respect to F, σ^2_{noise} and r_{noise}. Generalization accuracy does not significantly change with respect to σ^2_{signal}.

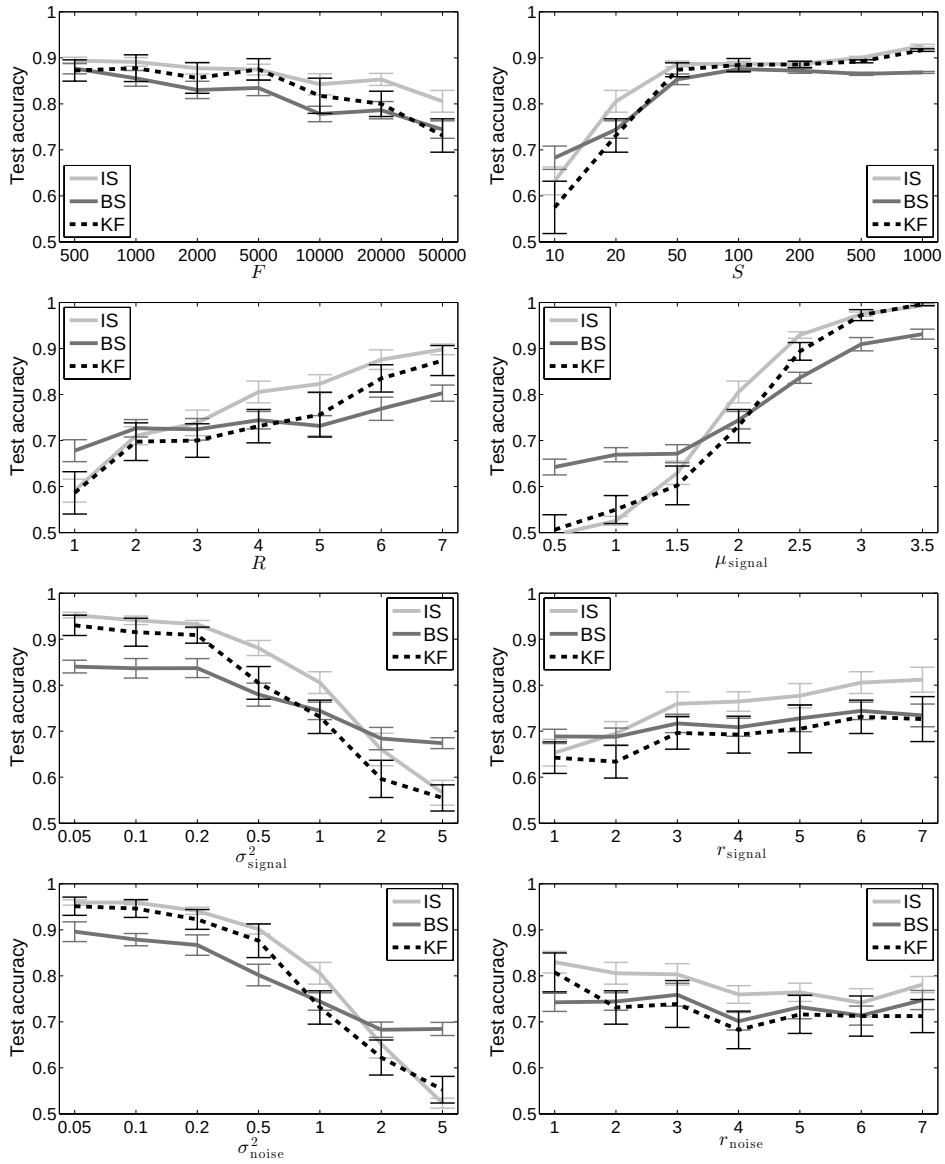

Figure 3: Generalization accuracy for linear SVMs with the most discriminative features, for a different: number of original features (F), number of samples per class (S), number of involved brain regions (R), distance between means of the classes (μ_{signal}), variance per class (σ^2_{signal}), radius of the involved brain regions (r_{signal}), noise variance (σ^2_{noise}) and radius of the spatially correlated noise (r_{noise}). Default values were set to $F = 50000$, $S = 20$, $R = 4$, $\mu_{\text{signal}} = 2$, $\sigma^2_{\text{signal}} = 1$, $r_{\text{signal}} = 6$, $\sigma^2_{\text{noise}} = 1$ and $r_{\text{noise}} = 2$. We also include error bars at 90% significance level. Note that k-fold cross-validation (KS) is a better estimator than .632 bootstrapping (BS), since it is always closer to the unbiased independent-set-of-samples method (IS). Note that generalization accuracy is increasing with respect to S, R, μ_{signal} and r_{signal}. It is also decreasing with respect to F, σ^2_{signal}, σ^2_{noise} and r_{noise}.

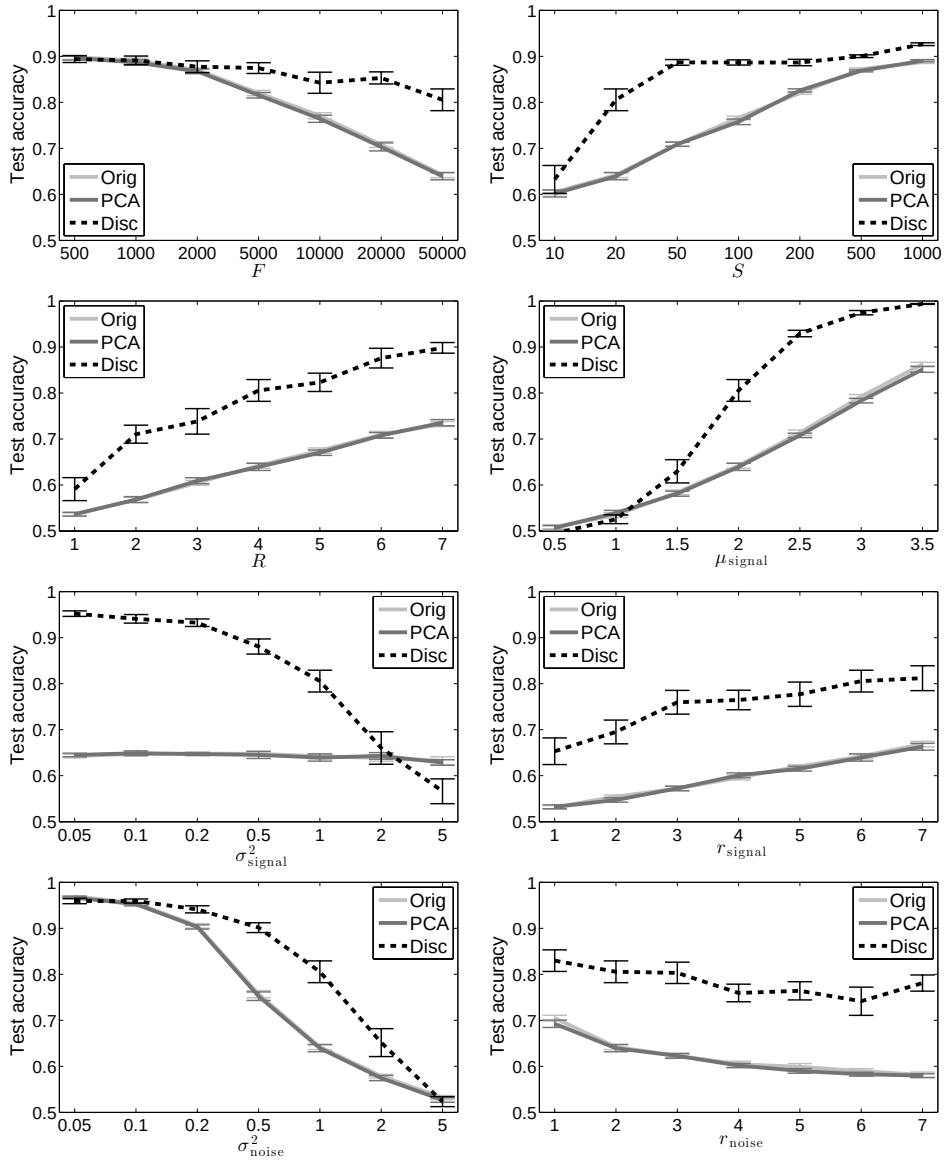

Figure 4: Generalization accuracy for linear SVMs with the most discriminative features, for a different: number of original features (F), number of samples per class (S), number of involved brain regions (R), distance between means of the classes (μ_{signal}), variance per class (σ^2_{signal}), radius of the involved brain regions (r_{signal}), noise variance (σ^2_{noise}) and radius of the spatially correlated noise (r_{noise}). Default values were set to $F = 50000$, $S = 20$, $R = 4$, $\mu_{\text{signal}} = 2$, $\sigma^2_{\text{signal}} = 1$, $r_{\text{signal}} = 6$, $\sigma^2_{\text{noise}} = 1$ and $r_{\text{noise}} = 2$. We report the unbiased independent-set-of-samples method. We also include error bars at 90% significance level. The results with the original features (Orig) and PCA features (PCA) are almost equal. In general, the results with the most discriminative features (Disc) are significantly better than Orig and PCA.

0.2, 0.5, (1), 2, 5} and $r_{\text{noise}} \in \{1, (2), 3, 4, 5, 6, 7\}$. We believe that some of these default values are typically encountered in brain fMRI problems, specifically the number of original features (F), the number of samples per class (S) and the signal-to-noise ratio $\sigma^2_{\text{signal}}/\sigma^2_{\text{noise}}$.

We report the generalization accuracy of linear SVMs with the original features in Figure 1, with PCA features in Figure 2 and with the most discriminative features in Figure 3. Note that k-fold cross-validation is a better estimator than .632 bootstrapping, since it is always closer to the unbiased independent-set-of-samples method. For the three feature extraction methods, generalization accuracy is increasing with respect to the number of samples per class (S), the number of involved brain regions (R), the distance between means of the classes (μ_{signal}) and the radius of the involved brain regions (r_{signal}). Generalization accuracy is also decreasing with respect to the number of original features (F), the variance per class (σ^2_{signal}), the noise variance (σ^2_{noise}) and the radius of the spatially correlated noise (r_{noise}). Although, for the original features as well as PCA features, generalization accuracy does not significantly change with respect to the variance per class (σ^2_{signal}). The behavior with the most discriminative voxels is more pronounced than with the other two feature extraction methods.

Figure 4 shows a comparison of the three feature extraction methods: the original features, PCA features and the most discriminative features. The results with the original features and PCA features are almost equal. In general, the results with the most discriminative features are significantly better than the other two methods.

4 Discussion

As in many classification tasks, having a small number of discriminative features allows for obtaining good generalization accuracy. Thus the goal of a practitioner is to decrease the number of features while retaining discriminability. In brain fMRI, we recommend to use methods such as the most discriminative voxels [12], which obtained significantly better results than using all voxels as features [5, 8, 10, 14] or PCA features [6, 13, 20, 21]. Note that the latter fact is not surprising since the objective of PCA is not to find discriminative features but to best explain the variance in the whole dataset.

In what follows, our observations are mainly with respect to the best performing feature extraction method, i.e. the most discriminative features.

Our experiments suggest that it is recommendable to collect data from at least $S = 50$ samples per class, which allows for obtaining good classification accuracy ($\sim 90\%$). Having more than $S = 50$ samples per class seems to increase generalization accuracy only marginally.

Some aspects of brain fMRI data cannot be controlled since they depend on the nature of the neuropsychological process and task. Among these aspects, we have the number and radius of the involved brain regions (R and r_{signal}) and the subject variability parameters (μ_{signal} and σ^2_{signal}). When there are few involved brain regions or when the brain regions are small, we

obtain poor classification accuracies ($60 - 65\%$). When there are several involved brain regions or when the brain regions are large, we obtain good classification accuracies ($80 - 90\%$). Additionally, in a regime of *low subject variability*, when the distance between means of the classes (μ_{signal}) is large or when the variance per class (σ^2_{signal}) is small, we obtain almost perfect classification ($96 - 100\%$). In a regime of *high subject variability*, when the distance between means of the classes (μ_{signal}) is small or when the variance per class (σ^2_{signal}) is high, we obtain very poor classification accuracies ($50 - 56\%$).

Other aspects of brain fMRI data can be controlled up to some extent. Controlling for the noise variance (σ^2_{noise}) seems far more important than controlling for the amount of spatial-correlatedness of the noise (r_{noise}). Indeed, when the noise variance is small, all the feature extraction methods obtained remarkably good generalization accuracy ($\sim 96\%$). In this sense, we recommend to take into account the reduction of noise variance when designing neuropsychological tasks as well as when devising proper signal processing methods for the captured data.

5 Concluding Remarks

We chose a reasonable model based on a number of Gaussian-distributed regions with Gaussian spatially-correlated noise, although other more complex synthetic models could have been chosen. Note that it is not possible to know the *true* probabilistic model that generated real-world data, unless we work under the unsatisfiable assumption of having access to infinite data. In practice, only a finite number of samples is available and *objectively* assessing the *level of realism* of a model is not possible. Despite the simplifying assumptions made in our model, we believe it introduces many intuitions from neuroscience. Having said that, we believe that more complex synthetic models that better introduce other neuropsychological aspects will be very beneficial.

We did not include the *leave-one-out* method, where we hold out each of the samples in turn while training on the other $S - 1$ samples. The main reason for excluding this method is that our experimental setting includes training sets of up to $S = 1000$ samples, where leave-one-out is computationally demanding. We preferred to include k-fold cross-validation, since includes leave-one-out as an specific instance ($k = S$). Moreover, in our experiments, k-fold cross-validation was a good estimator of the generalization accuracy, since it was always close to the unbiased (but impractical) independent-set-of-samples method. Experiments with leave-one-out cross-validation will be of importance, given its use in many studies.

Note that in most experimental settings, the parameters (e.g. number of PCA components, number of most discriminative features, soft-margin parameter C) are selected for each training set by using either a validation set or a nested cross-validation procedure. We chose to keep the parameters fixed for computational reasons. The experimental study of parameter

selection will be also beneficial.

While we focused exclusively on generalization accuracy, it is also important to analyze other aspects of feature selection and classification. For instance, it would be interesting to analyze whether the most discriminative features include the R ground-truth involved brain regions, or whether linear SVMs with the original features produce higher weights for the R ground-truth involved brain regions.

References

[1] S. Balci, M. Sabuncu, J. Yoo, S. Ghosh, S. Whitfield-Gabrieli, J. Gabrieli, and P. Golland. Prediction of successful memory encoding from fMRI data. *MICCAI Workshop on Analysis of Functional Medical Images*, pages 97–104, 2008.

[2] D. Cox and R. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19:261–70, 2003.

[3] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. Langlebenb. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28:663–8, 2005.

[4] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43:44–58, 2008.

[5] J. Etzel, V. Gazzola, and C. Keysers. An introduction to anatomical ROI-based fMRI classification analysis. *Brain Research*, 1282:114–25, 2009.

[6] Y. Fan, D. Shen, and C. Davatzikos. Detecting cognitive states from fMRI images by machine learning and multivariate classification. *IEEE CVPR Workshop on Mathematical Methods in Biomedical Image Analysis*, 2006.

[7] S. Hanson and Y. Halchenko. Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Computation*, 20:486–503, 2008.

[8] S. La Conte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 26:317–29, 2005.

[9] J. Lee, M. Marzelli, F. Jolesz, and S. Yoo. Automated classification of fMRI data employing trial-based imagery tasks. *Medical Image Analysis*, 13:392–404, 2009.

[10] M. Martínez-Ramón, V. Koltchinskii, G. Heileman, and S. Posse. fMRI pattern classification using neuroanatomically constrained boosting. *NeuroImage*, 31:1129–41, 2006.

[11] V. Michel, C. Damon, and B. Thirion. Mutual information-based feature selection enhances fMRI brain activity classification. *IEEE International Symposium on Biomedical Imaging*, pages 592–5, 2008.

[12] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57:145–75, 2004.

[13] J. Mourão-Miranda, A. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage*, 28:980–95, 2005.

[14] S. Ryali, K. Supekar, D. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51:752–64, 2010.

[15] X. Wang, R. Hutchinson, and T. Mitchell. Training fMRI classifiers to discriminate cognitive states across multiple subjects. *Neural Information Processing Systems*, 16:709–16, 2003.

[16] A. Anderson, I. Dinov, J. Sherin, J. Quintana, A. Yuille, and M. Cohen. Classification of spatially unaligned fMRI scans. *NeuroImage*, 49:2509–19, 2010.

[17] J. Arribas, V. Calhoun, and T. Adali. Automatic Bayesian classification of healthy controls, bipolar disorder and schizophrenia using intrinsic connectivity maps from fMRI data. *IEEE Transactions on Biomedical Engineering*, 57(12):2850–60, 2010.

[18] C. Damon, P. Pinel, M. Perrot, V. Michel, E. Duchesnay, J. Poline, and B. Thirion. Discriminating between populations of subjects based on fMRI data using sparse feature selection and SRDA classifier. *MICCAI Workshop on Analysis of Functional Medical Images*, pages 25–32, 2008.

[19] O. Demirci, V. Clark, and V. Calhoun. A projection pursuit algorithm to classify individuals using fMRI data: Application to Schizophrenia. *NeuroImage*, 39:1774–82, 2008.

[20] J. Ford, H. Farid, F. Makedon, L. Flashman, T. Mc Allister, V. Megalooikonomou, and A. Saykin. Patient classification of fMRI activation maps. *Medical Image Computing and Computer Assisted Intervention*, 2879:58–65, 2003.

[21] C. Fu, J. Mourão-Miranda, S. Costafreda, A. Khanna, A. Marquand, S. Williams, and M. Brammer. Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. *Biological Psychiatry*, 63:656–62, 2008.

[22] J. Honorio, D. Samaras, D. Tomasi, and R. Goldstein. Simple fully automated group classification on brain fMRI. *IEEE International Symposium on Biomedical Imaging*, pages 1145–48, 2010.

[23] J. Honorio, D. Tomasi, R. Goldstein, H.C. Leung, and D. Samaras. Can a single brain region predict a disorder? *IEEE Transactions on Medical Imaging*, 31(11):2062–72, 2012.

[24] F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45:S199–S209, 2009.

[25] N. Kriegeskorte, W. Simmons, P. Bellgowan, and C. Baker. Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12:535–40, 2009.

[26] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–4, 2008.

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.