Published in final edited form as:
Neuroimage. 2008 October 1; 42(4): 1414–1429. doi:10.1016/j.neuroimage.2008.05.050.

Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns

Okito Yamashita[a,*], **Masa-aki Sato**[a], **Taku Yoshioka**[a,b], **Frank Tong**[c], and **Yukiyasu Kamitani**[a,b]

[a]ATR Computational Neuroscience Laboratories, Japan

[b]National Institute of Information Technology, Japan

[c]Vanderbilt University, Psychology Department, USA

Abstract

Recent studies have used pattern classification algorithms to predict or decode task parameters from individual fMRI activity patterns. For fMRI decoding, it is important to choose an appropriate set of voxels (or features) as inputs to the decoder, since the presence of many irrelevant voxels could lead to poor generalization performance, a problem known as overfitting. Although individual voxels could be chosen based on univariate statistics, the resulting set of voxels could be suboptimal if correlations among voxels carry important information. Here, we propose a novel linear classification algorithm, called sparse logistic regression (SLR), that automatically selects relevant voxels while estimating their weight parameters for classification. Using simulation data, we confirmed that SLR can automatically remove irrelevant voxels and thereby attain higher classification performance than other methods in the presence of many irrelevant voxels. SLR also proved effective with real fMRI data obtained from two visual experiments, successfully identifying voxels in corresponding locations of visual cortex. SLR-selected voxels often led to better performance than those selected based on univariate statistics, by exploiting correlated noise among voxels to allow for better pattern separation. We conclude that SLR provides a robust method for fMRI decoding and can also serve as a stand-alone tool for voxel selection.

Introduction

Conventional fMRI data analysis has primarily focused on voxel-by-voxel functional mapping using the general linear model, in which stimuli or behavioral parameters are used as regressors to account for the BOLD response (Friston et al., 1995; Worsely et. al., 2002). Recently, much attention has been paid to pattern classification, or decoding, as an alternative approach to conventional functional mapping. In this approach, fMRI activation patterns of many voxels can be used to characterize subtle differences between different stimuli or subjects' behavioral/mental states. The pioneering work by Haxby et al. (2001) has demonstrated that broadly distributed fMRI activity patterns can discriminate pictures of visual objects, which cannot be easily distinguished by the conventional functional mapping (see also Strother et al., 2002; Spiridon and Kanwisher, 2002; Cox and Savoy, 2003; Carlson et al., 2003; Mitchell et. al., 2004; Laconte et. al., 2005; O'Toole et al., 2005 for other examples). Furthermore, the decoding approach has proved useful in extracting information about fine-scale cortical representations, which has been thought to lie beyond the resolution

*Corresponding author. oyamashi@atr.jp (O. Yamashita).

of fMRI. Kamitani and Tong (2005, 2006) showed that low-level visual features, such as orientation and motion direction, can be reliably decoded by pooling weakly selective signals in individual voxels. Since cortical columns representing orientation or motion direction are thought to be much smaller than standard fMRI voxels, the signal in each voxel may arise from voxel sampling with biases due to variability in the distribution of cortical feature columns or their vascular supply. Decoding analysis can exploit such subtle information, available in individual voxels, to obtain robust selectivity from the ensemble activity pattern of many voxels (‘ensemble feature selectivity’). For comprehensive reviews, see Haynes and Rees (2006) and Norman et al. (2006).

For fMRI decoding, selecting an appropriate set of voxels as the input for classification analysis is important for several reasons. First, voxel selection could improve decoding performance. fMRI decoding analysis takes a form of supervised learning (classification or regression), in which voxel values are the input variables or ‘features’, and a stimulus/task parameter is the output variable or ‘labelled’ category. In supervised learning, too many features can sometimes lead to poor generalization performance, a problem called overfitting. With many adjustable model parameters associated with the features, the learning model may fit to the noise present in the training data, and generalize poorly to novel test data. In a typical fMRI experiment, only tens or perhaps hundreds of samples (task blocks or volumes) are obtained, while the whole brain can contain as many as a hundred thousand voxels or features. Thus, fMRI decoding can easily lead to overfitting if all available voxels are used as input features. Support vector machines (SVM), one of the most popular classifiers in the fMRI decoding literature, avoids this problem by simultaneously minimizing the empirical classification error and maximizing the margin (Boser et al., 1992; Vapnik, 1998). However, generalization performance of SVM will still be degraded if too many irrelevant features are included.

Second, voxel selection is also useful for understanding neural information coding. Voxels can be selected based on separate anatomical or functional knowledge, so that decoding performance for one set of voxels can be compared with that of another. The higher the performance is, the more likely it is that the voxels represent information relevant to the task. Although careful examination is necessary to determine whether the voxels represent the decoded task parameter or some other correlated variable (Kamitani and Tong, 2005), comparisons of decoding performance for different brain areas can provide a powerful method for mapping the information available in local regions (see also Kriegeskorte et al., 2006).

In most previous studies, voxels have been selected based on anatomical landmarks, functional localizers (e.g., retinotopic mapping), or a voxel-by-voxel univariate statistical analysis obtained from training data or data from a separate experiment. The selected voxels were then used as input features for decoding analysis. An alternative to voxel selection for reducing dimensionality is to project the original feature space into a subspace of fewer dimensions using principal component analysis (PCA) (Carlson et al., 2003) or independent component analysis (ICA). The new dimensions can then be used as input features for decoding analysis. Such two-step methods for feature selection and decoding analysis have proven effective. But they could be suboptimal because the voxel/feature selection step does not take into consideration the discriminability of multi-voxel patterns.

In this paper, we introduce novel linear classification algorithms for binary and multi-class classification, which we will refer to as sparse logistic regression (SLR) and sparse multinomial logistic regression (SMLR), respectively. (Note that the term SLR will be used to refer to both binary and multi-class classifiers if the distinction is not critical). SLR is a Bayesian extension of logistic regression, which simultaneously performs feature (voxel)