Predicting cross-task behavioral variables from fMRI data using the *k*-support norm

Michail Misyrlis[1,3], Anna B. Konova[2,3], Matthew B. Blaschko[4,5], Jean Honorio[6], Nelly Alia-Klein[3], Rita Z. Goldstein[3], and Dimitris Samaras[1]

[1] Department of Computer Science, Stony Brook University, Stony Brook, NY, USA
[2] Department of Psychology, Stony Brook University, Stony Brook, NY, USA
[3]Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Center for Visual Computing, École Centrale Paris, France
[5]Équipe Galen, INRIA Saclay, Île-de-France, France
[6]CSAIL, MIT, Cambridge, MA, USA

Abstract. Sparsity regularization allows handling the curse of dimensionality, a problem commonly found in fMRI data. In this paper, we compare LASSO (ℓ_1 regularization) and the recently introduced k-support norm on their ability to predict real valued variables from brain fMRI data for cocaine addiction, in a principled model selection setting. Furthermore, in the context of those two regularization methods, we compare two loss functions: squared loss and absolute loss. With the squared loss function, k-support norm outperforms LASSO in predicting real valued behavioral variables measured in an inhibitory control task given fMRI data from a different task, designed to capture emotionally-salient reward. The absolute loss function leads to significantly better predictive performance for both methods in almost all cases and the k-support norm leads to more interpretable and more stable solutions often by an order of magnitude. Our results support the use of the k-support norm for fMRI analysis and the generalizability of the I-RISA model of cocaine addiction.

Keywords: Functional magnetic resonance Imaging (fMRI), Regularization, Sparse representations

1 Introduction

Functional magnetic resonance imaging (fMRI) is a widely used modality within the field of neuroimaging, that measures brain activity by detecting associated changes in blood oxygenation. One of the goals of fMRI data analysis is to detect correlations between brain activation and a task the subject performs during the scan.

The main challenges in statistical fMRI data analysis [1–4] are (i) the curse of dimensionality (ii) a small number of samples, due to the high cost of fMRI acquisition, and (iii) high levels of noise, such as system noise and random neural activity.

Sparsity regularizers are key statistical methods for improving predictive performance in the event that the number of observations is substantially smaller than the dimensionality of the data, as is the case in fMRI analysis. In this paper we compare the most frequently applied sparsity regularizer developed in the statistics literature, LASSO [5], with the k-support norm [6], a recently introduced method which is less biased towards sparse solutions.

The k-support norm can be viewed as a generalization of LASSO when $k = 1$ and ridge regression when $k = d$, where d is the dimensionality of the data. The k-support norm has previously been used in [6] for classification. It was first used for fMRI data modelling in [7] with a specific choice k parameter of the norm, that is, the method was not tested in a model selection setting. In both cases the k-support norm was used with the squared loss function.

We focus on comparing LASSO with the k-support norm in order to establish the latter regularizer's superiority in analyzing fMRI data. We use two loss functions, namely the squared error and the absolute error. The advantage of the absolute error loss is that it is more robust, in that it penalizes outliers less than squared loss, while still retaining convexity which guarantees finding the global optimum. We compare the methods not only in their predictive accuracy but also in the interpretability and stability of their results.

Our contribution in this paper is threefold. First, we introduce a novel method, the k-support norm with absolute error. Second, this is the first attempt to compare LASSO with the k-support norm in a principled model selection setting. Finally, to the best of our knowledge this is the first application of the k-support norm to a real valued response variable in a challenging clinical setting where the fMRI signal collected during *one task* is used to predict behavioral responses collected at a *different time* during a *second task*.

The neuroscientific motivation for our experiments is the exploration of human drug addiction. Basic studies have led to a theoretical model of human drug addiction, characterized by Impaired Response Inhibition (RI) and Salience Attribution (SA) (hence, I-RISA) [8]. According to the model, the skew in SA is predictive of impaired RI, together contributing to excessive drug use and relapse, core clinical symptoms of cocaine addiction. We use the fMRI data from a SA task (drug Stroop) in order to predict behavioral data in a RI task (color-word Stroop) collected at a different time, hence providing further evidence to support the I-RISA model.

2 Methods

We denote by $X \in \mathbb{R}^{n \times d}$ the design matrix of n samples each with d dimensions; we denote by $y \in \mathbb{R}^n$ the vector of targets.

A basis of statistical inference is the application of regularized risk, in which a loss function is evaluated over a sample of data and is linearly combined with a regularizer that penalizes some norm of the prediction function as in (Eq. (1)),

where the first term is the loss function and the second is the penalty term:

$$\min_{\beta} f(\beta, X, y) + \lambda J(\beta). \tag{1}$$

The scalar parameter $\lambda > 0$ controls the degree of regularization and J is a scalar valued function monotonic in a norm of $\beta \in \mathbb{R}^n$. Sparsity regularization is a key family of priors over linear functions that prevents overfitting and aids interpretability of the resulting models [5, 6]. Key to the mathematical understanding of sparsity regularizers is their interpretation as convex relaxations to quantities involving the ℓ_0 norm, which simply counts the number of non-zero elements of a vector. One of the most important sparsity regularizers is the LASSO [5], where $\lambda J(\beta) = \lambda \|\beta\|_1$. In many learning problems of interest, LASSO has been observed to shrink too many of the β variables to zero. In the presence of a group of highly correlated variables, LASSO may prefer a sparse solution. However including all correlated variables in the model could potentially lead to higher predictive accuracy [6] and the k-support norm provides a way of calibrating the cardinality of the regression vector β so as to include more variables.

The k-support norm can be computed as

$$\|\beta\|_k^{sp} = \left(\sum_{i=1}^{k-r-1} (|\beta|_i^{\downarrow})^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}} \tag{2}$$

where $|\beta|_i^{\downarrow}$ is the ith largest element of the vector and r is the unique integer in $\{0, \ldots, k-1\}$ satisfying

$$|\beta|_{k-r-1}^{\downarrow} > \frac{1}{r+1} \sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \geq |\beta|_{k-r}^{\downarrow}. \tag{3}$$

In this paper, we consider LASSO and the k-support norm with two loss functions: the squared error $f(\beta, X, y) = \|y - X\beta\|_2^2$ and the absolute error $f(\beta, X, y) = \|y - X\beta\|_1$.

In practice, we approximate the absolute error with a Huber type smoothing around zero to ensure differentiability.

3 Experimental Set-up

In this section we present our experiments and the data sets used in them. Our experiments aim at providing empirical evidence for the support of the I-RISA model.

We use the fMRI drug-word task described in [9, 10]. The neuropsychological experiment for cocaine addiction data set has a block design, which includes eight sessions, with each of them having different conditions. The two varying conditions are the monetary reward (50¢, 25¢, 1¢ and 0¢) and the cue shown (drug words, neutral words). The session consists of an initial screen displaying

the monetary reward and then presenting a sequence of forty words in four different colors (yellow, blue, red or green). The subject was instructed to press one of four buttons matching the color of the word they had just read. The subjects were rewarded for correct performance depending on the monetary condition.

We use the behavioral responses of the same subjects in a color-word task [11], a classic task of inhibitory control. In this task the subjects pressed for ink color of color words printed in either their congruent or incongruent colors. Four colors and words (red, blue, yellow and green) were used in all possible combinations. Both congruent and incongruent stimuli were presented randomly. The subjects performed four consecutive runs of this task. As there were 12 incongruent events in each run of 200 events, each subject's data contained up to 48 incongruent events.

For 38 control subjects and 74 cocaine abusers, we use the fMRI data from the drug-word task, to predict color-word behavioral variables.

Our experimental setting consists of 500 trials with an 85% / 15% random split between training and test sets. We perform model selection on the training set. That is, for each combination of parameters ($\lambda \in \{10^i : i = -2, \cdots, 8\}$ for LASSO, $\lambda \in \{10^i : i = -2, \cdots, 8\}$, $k \in \{1, 2, 3, 6, 12, 100, 200, 300, 600\}$ for k-support norm), we do a leave-one-subject-out cross validation on the samples that constitute the training set. We measure the correlation between the predicted and the true response variables on the training set. The parameter setting that leads to the highest correlation is used on the whole training set in order to learn a set of weights for each method, which are then applied on the test set. Finally, we measure the correlation between the predicted and the true response variables on the test set. We report the mean correlation on the holdout test samples and its standard error across the 500 random permutations in Sec. 4. We note that the same sample randomization is used for both LASSO and k-support norm.

In **experiment 1** we use the fMRI contrast drug > neutral words, averaged over monetary reward condition, to predict the conflict effect in the subjects' reaction time on the color-word task, defined as the difference in time between correctly performing the task for congruent and incongruent events. We use the Insula, Hippocampus Complex, Amygdala and ACC, part of the brain's limbic (emotion) circuit, as regions of interest (ROIs) for this experiment. These regions are chosen on the basis of previous studies on independent datasets that showed limbic system modulation by drug-related cues, eg. drug words [12].

In **experiment 2** we use the fMRI contrast 50¢ > 0¢, averaged over word type condition, in order to predict the subjects' responses on the color-word task, defined as the difference in percent accuracy between performing the task for congruent and incongruent events. We use the Basal Ganglia and Thalamus, part of the brain's reward circuit, as ROIs for this experiment. We chose these ROIs on the basis of previous studies on independent datasets that showed reward system modulation by primary and secondary reinforcers, including money [13].

4 Results

We compare the performance of the two methods in Table 1 for the first experiment and Table 2 for the second experiment.

Mean Correlation, D>N, Conflict effect on Reaction Time

Control Subjects

Norm / Loss	Squared	Absolute	p
LASSO	0.16 (0.02)	0.27 (0.02)	<0.01
k-support	0.22 (0.02)	0.24 (0.02)	<0.05
p	<0.001	0.21	

Cocaine Subjects

Norm / Loss	Squared	Absolute	p
LASSO	0.27 (0.01)	0.37 (0.01)	<0.001
k-support	0.33 (0.01)	0.36 (0.02)	<0.001
p	<0.001	0.96	

Table 1. Mean (SE) correlation over 500 random permutations of the samples between the predicted and the actual conflict effect on the reaction times for drug > neutral using the limbic ROI, for all combinations of regularizers and loss functions. The p-values were computed with a Wilcoxon signed rank test between the 500 correlation values for the two combinations of regularizer and loss function in the preceding rows or columns. Based on the p-values, there is a statistically significant difference between absolute loss predictions and squared loss predictions and between LASSO and k-support norm with the squared loss function in both cocaine and control subjects.

With the squared loss function, the k-support norm outperforms LASSO for almost all cases, while when combined with the absolute loss function, the regularizers do not significantly differ in their predictive performance. The absolute loss function, for both regularizers, leads to correlations that are significantly higher than those with the squared loss function in almost all cases.

We report the fraction of non-zero weights that were selected by each method for over 50% of the 500 trials in Tables 3 and 4 for the first and the second experiment respectively.

We average the weights assigned to the voxels over the 500 permutations and then compute the cumulative distribution function (CDF) for those weights. We threshold the CDF at 0.9 and visualize the weights of the voxels up to that threshold[1] in Fig. 1. The overly sparse solutions of the LASSO (Fig. 1(b), 1(d)) lead to models that cannot be interpreted as easily as the solutions of the k-support norm method (Fig. 1(a), 1(c)).

In the presence of correlated features, the degree of sparsity of the solution can be tuned with the k-support norm in order to include several highly correlated features. In contrast, LASSO tends to pick one representative feature with

[1] Due to space constraints we include one representative example out of two for each experiment. The omitted results are qualitatively similar.

Mean Correlation, 50¢>0¢, Conflict effect on Accuracy

Control Subjects

Norm / Loss	Squared	Absolute	p
LASSO	0.25 (0.02)	0.09 (0.02)	<0.001
k-support	0.26 (0.02)	0.09 (0.02)	<0.001
p	0.42	0.78	

Cocaine Subjects

Norm / Loss	Squared	Absolute	p
LASSO	0.22 (0.02)	0.42 (0.02)	<0.001
k-support	0.27 (0.01)	0.41 (0.02)	<0.001
p	<0.001	0.78	

Table 2. Mean (SE) correlation over 500 random permutations of the samples between the predicted and the actual response variables for 50¢> 0¢ using the Basal Ganglia, Thalamus ROI, for all combinations of regularizers and loss functions. The p-values were computed with a Wilcoxon signed rank test between the 500 correlation values for the two combinations of regularizer and loss function in the preceding rows or columns. Based on the p-values there is a statistically significant difference between absolute loss predictions and squared loss predictions and between k-support and LASSO with the squared loss in cocaine subjects only.

Voxel Selection Stability, D>N, Conflict effect on Reaction Time

Control Subjects

Norm / Loss	Squared	Absolute
LASSO	0.0004	0.0007
k-support	0.0029	0.0018

Cocaine Subjects

Norm / Loss	Squared	Absolute
LASSO	0	0.0023
k-support	0.0058	0.0734

Table 3. Voxel Selection stability over 500 random permutations of the samples for drug > neutral using the limbic ROI, for all combinations of regularizers and loss functions. The fraction of voxels which are selected for more than 50% of the 500 trials are presented. The higher values reported for k-support norm indicate that it makes more stable voxel selection than LASSO over different training sets.

no guarantee of consistency in feature selection across different splits of the data samples into training and test sets. In all cases the fraction of non-zero weights selected by the k-support norm is higher than that of LASSO, indicating that the k-support norm method leads to more stable solutions as compared to those obtained with LASSO.

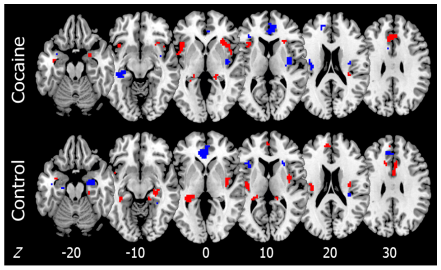
Voxel Selection Stability, 50¢>0¢, Conflict effect on Accuracy		
Control Subjects		
Norm / Loss	Squared	Absolute
LASSO	0.0004	0.0050
k-support	0.0037	0.0083
Cocaine Subjects		
Norm / Loss	Squared	Absolute
LASSO	0.0008	0.0013
k-support	0.0223	0.0122

Table 4. Voxel Selection stability over 500 random permutations of the samples for 50¢> 0¢ using the Basal Ganglia, Thalamus ROI, for all combinations of regularizers and loss functions. The fraction of voxels which are selected for more than 50% of the 500 trials are presented. The higher values reported for k-support norm indicate that it makes more stable voxel selection than LASSO over different training sets.

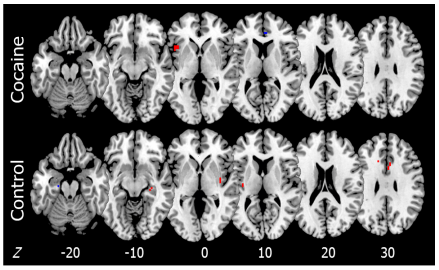
5 Discussion

In our experiments, in almost all cases, the k-support norm outperforms LASSO in predicting the behavioral measures given fMRI data when combined with squared loss, while when combined with the absolute loss, the predictive accuracy of the two regularizers does not differ significantly. The absolute loss led to higher predictions than squared loss for both regularizers for almost all cases. The LASSO leads to sparse solutions, since it tends to pick one feature per group of correlated features. On the other hand, the k-support norm allows calibrating the cardinality of the solutions and thus can select more interpretable groupings of correlated features and also leads to more stable results across different training sets. Thus, our results support the further exploration of the k-support norm for fMRI analysis.

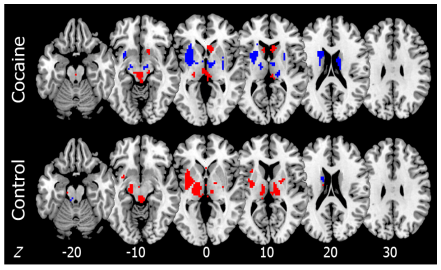
We also provide further evidence to support the I-RISA model of drug addiction, whereby the skew in SA in cocaine abusers, as indexed by fMRI response to drug words and monetary rewards, two motivationally salient stimuli, is predictive of RI, as indexed by response slowing and accuracy on a task requiring inhibitory control (the color-word Stroop). Specifically, we show that in cocaine users, response to drug words in voxels located in limbic brain regions, such as the anterior insula and ACC implicated in emotion processing and emotion regulation, was predictive of slower responses on the RI task (Exp. 1), while



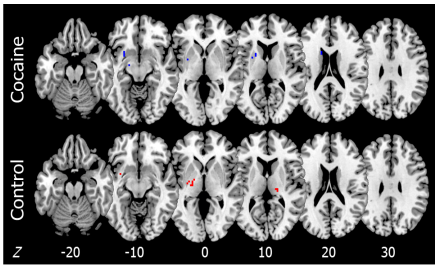
(a) Most predictive voxels in Exp. 1 using the k-support norm with the Absolute Loss



(b) Most predictive voxels in Exp. 1 using the LASSO with the Absolute Loss



(c) Most predictive voxels in Exp. 2 using the k-support norm with the Squared Loss



(d) Most predictive voxels in Exp. 2 using the LASSO with the Squared Loss

Fig. 1. Visualization of the most predictive voxels in Exp. 1 (top row) and Exp. 2 (bottom row) over the 500 permutations. The degree of sparsity of the solution can be tuned with the k-support norm, thus leading to models ((a), (c)) that are easier to interpret than those of LASSO ((b), (d)). (Best viewed in color)

response to money in voxels located in reward-related brain regions, such as the putamen implicated in habits, was predictive of lower accuracy on the RI task (Exp. 2).

References

1. J Honorio, D Tomasi, R Goldstein, H Leung, and D Samaras, "Can a single brain region predict a disorder?," *IEEE Transactions on Medical Imaging*, 2012.
2. S Song, Z Zhan, Z Long, J Zhang, and L Yao, "Comparative study of svm methods combined with voxel selection for object category classification on fmri data," *PloS one*, vol. 6, no. 2, pp. e17191, 2011.
3. Andreas Bartels and Semir Zeki, "Brain dynamics during natural viewing conditionsa new guide for mapping connectivity in vivo," *Neuroimage*, vol. 24, no. 2, pp. 339–349, 2005.
4. D R Hardoon, J Mourao-Miranda, M Brammer, and J Shawe-Taylor, "Unsupervised analysis of fmri data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

5. R Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
6. A Argyriou, R Foygel, and N Srebro, "Sparse prediction with the k-support norm," *Neural Information Processing Systems*, 2012.
7. K Gkirtzou, J Honorio, D Samaras, RZ Goldstein, M Blaschko, et al., "fmri analysis of cocaine addiction using k-support sparsity," in *International Symposium on Biomedical Imaging*, 2013.
8. RZ Goldstein and ND Volkow, "Drug addiction and its underlying neurobiological basis: neuroimaging evidence for the involvement of the frontal cortex," *The American journal of psychiatry*, vol. 159, no. 10, 2002.
9. RZ Goldstein, D Tomasi, S Rajaram, LA Cottone, L Zhang, T Maloney, F Telang, N Alia-Klein, and ND Volkow, "Role of the anterior cingulate and medial orbitofrontal cortex in processing drug cues in cocaine addiction," *Neuroscience*, vol. 144, no. 4, pp. 1153–1159, 2007.
10. AB Konova, SJ Moeller, D Tomasi, MA Parvaz, N Alia-Klein, ND Volkow, and RZ Goldstein, "Structural and behavioral correlates of abnormal encoding of money value in the sensorimotor striatum in cocaine addiction," *European Journal of Neuroscience*, vol. 36, no. 7, 2012.
11. SJ Moeller, D Tomasi, J Honorio, ND Volkow, and RZ Goldstein, "Dopaminergic involvement during mental fatigue in health and cocaine addiction," *Translational psychiatry*, vol. 2, no. 10, pp. e176, 2012.
12. HW Chase, SB Eickhoff, AR Laird, and L Hogarth, "The neural basis of drug stimulus processing and craving: an activation likelihood estimation meta-analysis," *Biological psychiatry*, vol. 70, no. 8, pp. 785–793, 2011.
13. X Liu, J Hairston, M Schrier, and J Fan, "Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 5, pp. 1219–1236, 2011.