Generalized Group Sparse Classifiers with Application in fMRI Brain Decoding

Bernard Ng
The University of British Columbia
2366 Main Mall, Vancouver, BC, V6T 1Z4
bernardn@ece.ubc.ca
http://bisicl.ece.ubc.ca

Rafeef Abugharbieh
The University of British Columbia
2366 Main Mall, Vancouver, BC, V6T 1Z4
rafeef@ece.ubc.ca
http://bisicl.ece.ubc.ca

Abstract

The perplexing effects of noise and high feature dimensionality greatly complicate functional magnetic resonance imaging (fMRI) classification. In this paper, we present a novel formulation for constructing "Generalized Group Sparse Classifiers" (GSSC) to alleviate these problems. In particular, we propose an extension of group LASSO that permits associations between features within (predefined) groups to be modeled. Integrating this new penalty into classifier learning enables incorporation of additional prior information beyond group structure. In the context of fMRI, GGSC provides a flexible means for modeling how the brain is functionally organized into specialized modules (i.e. groups of voxels) with spatially proximal voxels often displaying similar level of brain activity (i.e. feature associations). Applying GSSC to real fMRI data improved predictive performance over standard classifiers, while providing more neurologically interpretable classifier weight patterns. Our results thus demonstrate the importance of incorporating prior knowledge into classification problems.

1. Introduction

The curse of dimensionality presents major challenges to pattern classification [1]. In many real world problems, the number of features often far exceeds the number of samples [2]. Thus, direct application of standard classifiers will likely result in overfitting [1]. To reduce overfitting, various dimension reduction techniques have been put forth, which can be broadly divided into two categories: feature extraction and feature selection [1]. Extracting features by applying subspace learning algorithms, such as principal component analysis (PCA), linear discriminant analysis (LDA), and Laplacian eigenmap [1,3] among many others, provide an effective means of reducing feature dimensionality while preserving certain intrinsic relationships between the samples. However, the new features created by combining the original features lack intuitive physical meanings, which hinders result interpretation [1].

The alternative dimension reduction strategy is to select the subset of features with highest discriminability. This approach simplifies result interpretation since the physical meanings of the features are retained. Conventional techniques under this category include sequential forward selection and sequential backward selection [1]. More advanced methods, such as support vector machine with recursive feature selection (SVM-RFE) [4] and random forest [5], that enable multivariate feature selection have also been explored. More recently, a surge of sparsity-enforcing techniques, such as relevance vector machine [6], informative vector machines [7], joint classifier and feature optimization [8], sparse multinomial logistic regression [9], sparse SVM [10], and spectral regression [11], have been proposed. Many of these techniques perform feature selection by exploiting the sparse property of the least absolute shrinkage and selection operator (LASSO) penalty [12] to shrink classifier weights associated with irrelevant features to exactly zero. However, merely enforcing sparsity without accounting for domain-specific information may result in spurious classifier weight patterns [13]. For instance, in the context of image classification, adjacent pixels are typically correlated. Thus, these pixels should desirably be jointly selected to reflect their correlations [14].

To extend the sparse property of LASSO to the group level so that related features (e.g. adjacent voxels) would be jointly selected, group LASSO was introduced [15,16]. Given that the input features can be grouped based on some common "factors" [16], the group LASSO penalty sparsely selects a subset of the groups and assigns non-zero weights to all features within the selected groups. Identification of relevant key factors is thus simplified. However, the group LASSO penalty does not model the associations between features within a group. For instance, adjacent pixels predefined to be in the same group would be jointly selected, but the classifier weight patterns might not be spatially smooth.

Another widely-used technique that generates similar effects as group LASSO is elastic net [17]. This technique combines a ridge penalty with the LASSO penalty, which has been shown to jointly select sets of correlated features. However, the ridge penalty alone does not provide the flexibility to model other properties beyond correlations.

In many applications, jointly modeling group structure and the associations between features within groups can be beneficial. In particular, in the presence of strong noise, the amount of discernable information can be limiting. In such cases, prior knowledge, such as feature associations and group structure, can serve as a valuable source of information. An area that can enormously benefit from incorporation of prior knowledge is functional magnetic resonance imaging (fMRI) analysis, which suffers from exceedingly low samples-to-features ratio coupled with severely noisy data. In a typical fMRI study, multiple subjects are recruited and presented with a set of stimuli as three dimensional (3D) volumes of their brains are acquired at regular time intervals. Brain regions pertinent to the experimental conditions will consume oxygen in response to stimulus, resulting in changes in signal intensity, which are used for inferring brain activation. However, the complex confluence of artifacts, such as scanner noise, head motions, and oxygenation changes due to cardiac and respiratory cycles, greatly hinders accurate fMRI analysis. Thus, additional information is critically needed for disambiguating signal from noise.

The standard approach for analyzing fMRI data compares the intensity time course of each brain voxel independently against an expected response to estimate the likelihood of activation [18]. The multivariate nature of fMRI data is thus completely ignored. To remedy this limitation, classification approaches that enable all brain voxels to be jointly analyzed have been explored [19,20]. Under the classification framework, the signal intensity of each voxel is usually taken as a feature with each brain volume treated as a sample [19,20]. The classification task is thus to determine the associating experimental condition of each brain volume based on its signal patterns. Typical fMRI datasets consist of considerably more voxels (~tens of thousands) than brain volumes (~hundreds). Hence, direct application of standard classifiers, such as support vector machines [21] and LDA [22], with all brain voxels used as features will likely result in overfitting [19,20]. To reduce the dimensionality of the feature space, a common strategy is to apply univariate analysis [18] to restrict the feature set to only those voxels displaying significant brain activation or discriminability [21]. Alternatively, principal component analysis (PCA) may be applied to reduce feature dimension [23]. However, both of these strategies discard the collective discriminant information encoded by the voxel patterns, which may lead to suboptimal feature selection [13,24-26].

Recently, there is an increasing interest in applying sparsity-enforcing techniques to fMRI classification [13,24-27]. However, naively enforcing sparsity may result in spurious classifier weight patterns [13,27]. In particular, classifier weights derived from enforcing sparsity alone often deviate from how brain activity is observed to be distributed in localized spatial clusters [13]. To encourage spatial contiguity, spatial SVM [28] has been proposed. However, this technique assigns weights to all input voxels, which renders identification of relevant voxels nontrivial. Group sparse penalties have also been explored to jointly select spatially proximal voxels as well as voxels within brain regions of interest (ROIs) [29]. The resulting weight patterns, however, may not be spatially smooth, since group sparse penalties do not model the relationships between features within each group. Elastic net [13] has also been employed but suffers from similar limitations [27].

In this paper, we present a novel formulation for jointly integrating prior knowledge on group structure and feature associations into classifier learning. Our main contribution is an extension of group LASSO that provides the flexibility to model associations between features within (predefined) groups. We refer to classifiers built from our formulation as "Generalized Group Sparse Classifiers" (GGSC), which can be viewed as a group level extension of our recently proposed "Generalized Sparse Classifiers" [27]. The motivation behind GGSC is rooted from how the human brain is functionally organized into specialized modules (i.e. in groups of voxels) [30], where spatially proximal voxels within each module tend to display similar level of brain activity (i.e. feature associations) [31]. Thus, in addition to handling noise and the curse of dimensionality, GGSC facilitates more explicit modeling of prior knowledge. The implications in modeling spatial voxel correlations, while enforcing group sparsity are investigated.

2. Proposed Method

2.1. Problem Formulation

In a standard classification framework, given N $M\times1$ feature vectors, x_i, $i = 1\ldots N$, forming the rows of an $N\times M$ predictor matrix, X, the overall objective is to find the corresponding $N\times1$ label vector, l, containing the class label l_i of x_i. We assume here that each feature x_{ip} can be naturally assigned to a group $g_k \in G = \{g_1, \ldots, g_K\}$. In the context of fMRI, we treat signal intensity of each brain voxel p as a feature x_{ip}, and each brain volume as a sample x_i. Our goal is thus to determine to which experimental condition l_i does each brain volume x_i belong. Since the human brain is known to be functionally organized into specialized neural regions [30], this provides a natural grouping of the voxels. Also, the observation that spatially proximal voxels tend to jointly activate [31] can serve as prior knowledge. Our classifier learning formulation for incorporating this prior information on the functional organization of the brain is presented next.

2.2. Generalized Group Sparse Classifiers

During the past decade, an affluence of sparsity-enforcing techniques has been proposed. Beginning with the introduction of LASSO [12], numerous variants subsequently followed, including the elastic net [17], group LASSO [15,16], and sparse group LASSO [32,33] among many others. Although the extension from LASSO to its variants often involved only a simple addition of other penalties, the effects are quite substantial. For example, by adding a ridge penalty to LASSO, Zou et al. [17] showed that a "grouping effect" that promotes joint selection of correlated features was enabled. Extending the elastic net, we proposed combining a "generalized" ridge penalty with LASSO to facilitate modeling of properties beyond correlations for classifier learning [27]. This penalty has also been proposed for penalizing differences in successive weights of ordered features in the context of regression under the name, "Smooth LASSO" [34]. Other examples include hierarchical LASSO, which combines LASSO and group LASSO to reintroduce feature-level sparsity to group LASSO [32,33]. In our work here, we propose augmenting the group LASSO with the generalized ridge penalty, so that feature associations within groups can be modeled:

$$J_{GGSC}(a) = \alpha\left\|\Gamma a\right\|_2^2 + \beta\left\|a_g\right\|_{2,1}, \quad (1)$$

where in the context of classification, a is the classifier weight vector we are estimating, $\left\|a_g\right\|_{2,1} = \sum_{k=1}^{K}\left\|a_{g_k}\right\|_2$ is the original group LASSO penalty, a_{gk} are weights associated with features belonging to group g_k, and α and β control the amount of regularization. Γ is a penalty matrix for modeling the associations between features (Section 2.3). We note that (1) can be viewed as a generalization of a number of widely-used penalties. Specifically, if we define each input feature as a group, (1) reduces to the penalty used for constructing "Generalized Sparse Classifiers" [27]. If we additionally set Γ to the identity matrix, the elastic nets penalty [17] is obtained. Furthermore, setting α in (1) to zero results in the group LASSO penalty [15,16].

To model group structure and feature associations during classifier learning, we propose the following formulation which consists of the typical misclassification cost combined with our proposed penalty (1):

$$\hat{a} = \min_a \left\|l - f(Xa)\right\|_0 + \alpha\left\|\Gamma a\right\|_2^2 + \beta\left\|a_g\right\|_{2,1}, \quad (2)$$

where $f(\cdot)$ maps Xa to the label space. We have restricted our attention to linear classifiers, i.e. $y = Xa$, so that the relative contribution of each feature x_{ip} can be directly discerned from a_p, which is critical for result interpretation in fMRI studies [19]. Our strategy for optimizing (2) is discussed in Section 2.4.

2.3. Feature Association Modeling

To model the associations between features within each group g_k, we set the elements of Γ such that discrepancies in classifier weights between features that are associated with each other are penalized. For example, if feature x_{ip} is associated with features x_{iq1} and x_{iq2}, we can encourage the corresponding classifier weights, a_p, a_{q1}, a_{q2}, to have similar magnitudes by setting Γ to:

$$\begin{array}{l}\Gamma_{p,:} = \begin{pmatrix} 2 & \cdots & -1 & -1 & \cdots \end{pmatrix} \\ \qquad\quad \uparrow \qquad\quad \uparrow \quad \uparrow \qquad , \\ \qquad\quad p \quad \cdots \quad q_1 \quad q_2 \quad \cdots \end{array} \quad (3)$$

assuming all features are equally important. Nevertheless, the relative importance of the features can be straightforwardly encoded into (2) by adjusting the penalties assigned to different classifier weights. In the context of fMRI, we define feature associations based on spatial locality. Specifically, we assume that each voxel p is associated with its 6-connected spatial neighbors within the same brain region g_k. This definition of feature association models how spatially proximal voxels tend to display similar level of brain activity [31].

2.4. GGSC Optimization

To efficiently minimize (1), we employ a two-step optimization strategy originally proposed in the seminal paper by Zou et al. for enforcing sparsity on PCA [35] and subsequently extended by Cai et al. to various standard classifiers under the name "spectral regression" [11]:

Step 1. Learn the constraint-free optimal projection of the training data, X, e.g. using graph embedding (GE) [3]:

$$Wy = \lambda Dy, \quad (4)$$

where y is the projection of X in the subspace defined by W [3]. W_{ij} models the intrinsic relationships between samples i and j of the training data, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. The advantage of GE is that it enables various subspace learning algorithms to be used as classifiers by simply varying W [3].

Step 2. Determine the classifier weights a such that y and Xa are as similar as possible under the desired constraints:

$$\hat{a} = \min_a \left\|y - Xa\right\|_2^2 + \alpha\left\|\Gamma a\right\|_2^2 + \beta\left\|a_g\right\|_{2,1}. \quad (5)$$

Using the above strategy converts the classification problem (2) into a regularized regression problem (5). If we then transform (5) by augmenting X and y as follows:

$$\tilde{X} = (1+\alpha)^{-\frac{1}{2}}\begin{pmatrix} X \\ \sqrt{\alpha}\Gamma \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad (6)$$

we obtain the group LASSO regression problem [15,16]:

$$\hat{a} = \sqrt{1+\alpha}\min_{\tilde{a}}\left\|\tilde{y} - \tilde{X}\tilde{a}\right\|_2^2 + \beta\left\|\tilde{a}_g\right\|_{2,1}, \quad (7)$$

which can be solved using various existing optimization

algorithms [15,16,36-38]. The spectral projected gradient technique [38] was employed in our work with parameters α and β selected using nested cross-validation [21].

2.5. Implementation Details

Applying (1) for classifier learning necessitates that the input features be meaningfully assigned into groups. In the context of fMRI, voxel grouping can be performed by dividing the brain into ROIs. One way is to warp a labeled atlas to each subject's brain images to extract the ROIs. However, the typically large anatomical variability renders this approach prone to mis-registration errors [31]. The alternative is to have experts manually segment the ROIs [21], which is the approach used in this work.

Another issue specific to fMRI studies is that temporally adjacent brain volumes, which are used as samples, typically display high correlations due to the slow hemodynamic response. Thus, assigning adjacent brain volumes to both the training and test sets will inflate the accuracy estimates. We have thus ensured that brain volumes belonging to the same experiment trial were not assigned to both the training and test sets in our parameter selection and prediction accuracy estimation.

3. Materials

The publicly available StarPlus database [39] was used for validation. We provide here a brief description of the data and the experimental setup for convenience. Further details can be found in [21,39].

3.1. fMRI Experiment

In the StarPlus experiment, all subjects performed 40 trials of a sentence/picture matching task. In each trial, subjects looked at a picture (or sentence) followed by a sentence (or picture), and to decide whether the sentence (picture) correctly described the picture (sentence). The first stimulus was presented for 4 s followed by a blank screen for 4 s. The second stimulus was then presented for up to 4 s followed by a 15 s rest period. In half of the trials, the picture preceded the sentence, and vice versa.

3.2. Imaging Data

fMRI brain volumes were acquired from 13 normal subjects at a TR of 500 ms [21]. 6 of the subjects' data were made available online [39]. Each subject's dataset comprised voxel time courses within 25 ROIs that were chosen by neuroscience experts. ROIs included the calcarine fissure, supplementary motor areas, left inferior frontal gyrus, bilateral dorsolateral prefrontal cortex, frontal eye fields, inferior parietal lobule, intraparietal sulcus, inferior temporal lobule, opercularis, posterior precentral sulcus, supramarginal gyrus, superior parietal lobule, temporal lobe, and triangularis, which in all resulted in approximately 5000 voxels per subject. Inter-subject differences in the number of voxels were due to anatomical variability. Motion-correction and temporal detrending were applied on the voxel time courses to account for head motions and low frequency signal drifts [21]. No spatial normalization was performed.

4. Results and Discussion

In this work, we considered the classification task of discriminating brain volumes associated with a sentence from those associated with a picture. We treated the signal intensity of each voxel as a feature, and each brain volume as a sample. Each sample (feature vector) thus consisted of approximately 5000 features (i.e. roughly 5000 voxels within the 25 ROIs, Section 3.2). To account for the delay in the hemodynamic response [18], we only used the 8 brain volumes collected 4 s after stimulus onset as samples. This results in 320 samples per class, i.e. 8 brain volumes by 40 trials, for each subject.

To validate GGSC, we constructed a spatially smooth sparse region LDA (SSRLDA) classifier by first solving for y in (4) with:

$$W_{ij} = \begin{cases} 1/m_t, & l_i = l_j = t \\ 0, & otherwise \end{cases}, \tag{8}$$

where m_t is the number of samples in class t, and $D = I$ [11]. We then optimized (2) with Γ set in a manner analogous to (3) to estimate SSRLDA's classifier weights. For comparison, we also tested LDA [22], linear SVM [21], LDA with a LASSO penalty (SLDA), LDA with an elastic net penalty (ENLDA), LDA with a group LASSO penalty to promote sparsity at the region-level (SRLDA), and spatially smooth sparse LDA (SSLDA) [27] on the same StarPlus data. Ten-fold cross validation was used to estimate prediction accuracy [21].

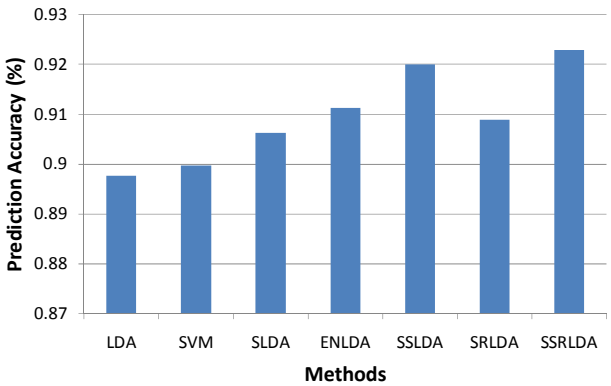

Figure. 1. Predictive accuracy comparisons. Our proposed SSRLDA resulted in the best overall predictive performance over all other contrasted methods. This demonstrates the importance of modeling the modularity property of the brain and its inherent spatial characteristics in fMRI classification.

LDA resulted in the lowest average prediction accuracy over subjects (Fig. 1), which was likely due to overfitting. Using SVM led to a slight improvement in accuracy. Controlling overfitting using SLDA also only improved accuracy mildly, which might be due to LASSO's limitation on the number of features (i.e. constrained by the sample size) [17]. Alleviating this constraint using ENLDA further improved predictive performance. Grouping voxels into ROIs and sparsely selecting the more discriminant ROIs using SRLDA also improved accuracy over SLDA, but the increase was miniscule. In contrast, explicitly modeling local spatial correlations by using SSLDA led to a more substantial accuracy increase. Exploiting both group structure and spatial correlations using our proposed SSRLDA achieved the best overall prediction performance over all contrasted methods with an average accuracy of 92.3%. Our results thus suggest added benefits in modeling both the modularity property of the human brain and its local spatial characteristics.

In addition to improving predictive performance, SSRLDA also provides classifier weight patterns that simplify interpretation. A representative axial slice is shown in Fig 2. Only an exemplar subject is included due to space limitations. Classifier weight patterns of both LDA and SVM appeared to be spatially dispersed in a random fashion, which substantially deviates from the widely-accepted conjecture of how brain activity is distributed in local spatial clusters [31]. These results illustrate an important yet highly overlooked point in that achieving higher predictive accuracy (as obtained with SVM compared to LDA) does not necessarily translate into more interpretable weight patterns [13].

Using SLDA resulted in overly-sparse weight patterns, which was partially overcome with ENLDA. SSLDA provided spatially smoother weight patterns than ENLDA, but isolation of discriminant brain regions is nontrivial since most brain regions contained some voxels with non-zero weights. Enforcing region-level sparsity using SRLDA provided a more clear-cut indication of which brain regions were most relevant for the sentence/picture classification task. However, the resulting weight patterns do not possess the spatially smooth property that SSLDA provides. Using SSRLDA generated smoother classifier weight patterns than SRLDA, while providing the region-level sparsity needed for simple identification of discriminant ROIs. In particular, the prefrontal cortex, which is responsible for verification task (e.g. decide if a sentence matches a picture) [40], as well as the visual cortex along the calcarine fissure and the temporal lobe, which pertain to picture/sentence discrimination [41], were correctly selected by SSRLDA. We note that the anatomical ROIs likely comprise multiple functionally-homogeneous sub-regions, where only a subset of these sub-regions may be relevant for class discrimination. Thus, functionally parcellating the brain into finer regions [31] would provide us a more precise localization of the discriminant areas while still allowing us to take advantage of our prior knowledge on the modular property of the brain [30].

To quantify the increase in local spatial smoothness using SSRLDA as compared to SRLDA, we employed the spatial distribution metric (SDM) proposed in [13]:

$$SDM = H / H_0,\ H = -\sum_{b=1}^{B} p_b \log p_b,\ p_b = Q^{-1} \sum_{q \in b} \left| a_q \right|, \quad (9)$$

where we divided each subject's brain into B 3×3×3 bins [13]. $Q = \|a\|_1$ and $H_0 = \log \|a\|_0$. SDM $\in [0,1]$, where 0 implies that a_q are concentrated within one bin and 1 indicates that a_q are evenly distributed across the bins. SSRLDA's SDM (0.5677) was found to be lower than that of SRLDA (0.5726), thus quantitatively demonstrating that modeling feature associations using SSRLDA provided spatially smoother weight patterns than SRLDA.

5. Conclusions

We proposed a novel classifier learning formulation that jointly exploits the group structure of the data as well as the associations between features. Our work amounts to augmenting group LASSO with a generalized ridge penalty that allows for incorporation of general properties, such as the commonly important spatial smoothness constraint. Applying SSRLDA, built from our GGSC formulation, resulted in enhanced predictive performance over state-of-the-art classifiers. In addition, more neurologically interpretable classifier weight patterns were obtained. From a general classification standpoint, having the flexibility to more elaborately model prior knowledge can be highly beneficial as demonstrated by our results. We believe that GGSC is a significant step forward in providing such modeling flexibility.

References

[1] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. IEEE Trans. Pat. Ana. Machine Intell., 22:4-37, 2000.

[2] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, and M. Caligiuri. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286:531-536, 1999.

[3] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extension: A General Framework for Dimensionality Reduction. IEEE Trans. Pat. Ana. Machine Intell., 29:40-50, 2007.

[4] S. Hanson and Y. Halchenko. Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There is no Face Identification Area. Neural Comput., 20:486-503, 2008.

[5] G. Langs, B.H. Menze, D. Lashkari, and P. Golland, Detecting Stable Distributed Patterns of Brain Activation Using Gini Contrast. NeuroImage, 2011. (in press)

[6] M.E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. J. Machine Learning Research, 1:211-244, 2001.
[7] N. Lawrence, M. Seeger, and R. Herbrich. Fast Sparse Gaussian Process Methods: The Informative Vector Machine. In: Proc. NIPS, 15:609-616, 2003.
[8] B. Krishnapuram, L. Carin, and A.J. Hartemink. Joint Classifier and Feature Optimization for Comprehensive Cancer Diagnosis Using Gene Expression Data. J. Comput. Biol., 11:227-242, 2004.
[9] B. Krishnqpuram, L. Carin, M.A.T. figueiredo, and A.J. Hartemink. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. IEEE Pat. Ana. Machine Intell., 27:957-968, 2005.
[10] J.B. Bi, K.P. Bennett, M. Embrechts, C.M. Breneman, and M. Song. Dimensionality Reduction via Sparse Support Vector Machines. J. Machine Learning Research, 3:1229-1243, 2003.
[11] D. Cai, X. He, and J. Han. Spectral Regression: A Unified Approach for Sparse Subspace Learning. In: Proc. IEEE Int. Conf. Data Mining, 73-82, 2007.
[12] R. Tibshirani. Regression Shrinkage and Selection via the LASSO. J. Royal Stat. Soc. Series B, 58:267:288, 1996.
[13] M.K. Carroll, G.A. Cecchi, I. Rish, R. Garg, and A.R. Rao. Prediction and Interpretation of Distributed Neural Activity with Sparse Models. NeuroImage, 44:112-122, 2009.
[14] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a Spatially Smooth Subspace for Face Recognition. In: Proc. CVPR, 1-7, 2007.
[15] M. Yuan and Y. Lin. Model Selection and Estimation in Regression with Grouped Variables. J. Royal Stat. Soc. Series B, 68:49-67, 2006.
[16] L. Meier, S. van de Geer, and P. Buhlman. The Group LASSO for logistic regression. J. Royal Stat. Soc. Series B, 70:53-71, 2008.
[17] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. J. Royal Stat. Soc. Series B, 67:301-320, 2005.
[18] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C.D. Frith, and R.S.J. Frackowiak. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Hum. Brain Mapp., 2:189-210, 1995.
[19] K.A. Norman, S.M. Polyn, G.J. Detre, and J.V. Haxby. Beyond Mindreading: Multi-voxel Pattern Analysis of fMRI Data. Trends Cogn. Sci., 10:424-430, 2006.
[20] J.D. Haynes and G. Rees. Decoding Mental States from Brain Activity in Humans. Nat Rev. Neurosci., 7:523-534, 2006.
[21] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to Decode Cognitive States from Brain Images. Mach. Learn., 57:145-175, 2004.
[22] J.D. Haynes and G. Rees. Predicting the Orientation of Invisible Stimuli from Activity in Human Primary Visual Cortex. Nat. Neurosci., 8:686-691, 2005.
[23] T.A. Carlson, P. Schrater, and S. He. Patterns of Activity in the Categorical Representations of Objects. J. Cogn. Neurosci., 15:704-717, 2003.
[24] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse Estimation Automatically Selects Voxels Relevant for the Decoding of fMRI Activity Patterns. NeuroImage 42:1414-1429, 2008.
[25] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon. Sparse Logistic Regression for Whole-brain Classification of fMRI Data. NeuroImage 51:752-764, 2010.
[26] V. Michel, E. Eger, C. Keribin, and B. Thirion. Multi-Class Sparse Bayesian Regression for Neuroimaging Data Analysis. In: MICCAI Workshop on Mach. Learn. Med. Imaging, 50-57, 2010.
[27] B. Ng, A. Vahdat, G. Hamarneh, and R. Abugharbieh. Generalized Sparse Classifiers for Decoding Cognitive States in fMRI. In: Proc. MICCAI Workshop on Machine Learning Med. Imaging, 6357:108-115, 2010.
[28] L. Liang, V. Cherkassky, and D.A. Rottenberg. Spatial SVM for Feature Selection and fMRI Activation Detection. In: Int. Joint Conf. on Neural Networks, 1463-1469, 2006.
[29] M. Van Gerven, A. Takashima, and T. Heskes. Selecting and Identifying Regions of Interest Using Groupwise Regularization. In: NIPS Workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis (2008).
[30] J.A. Fodor. The Modularity of the Mind. The Massachusetts Institute of Technology. 2-47, 1983.
[31] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.B. Poline. Dealing with the Shortcomings of Spatial Normalization: Multi-subject Parcellation of fMRI Datasets. Hum. Brain Mapp., 27:678-693, 2006.
[32] P. Sprechmann, I. Ramirez, and G. Sapiro. Collaborative Hierarchical Sparse Modeling. In: CISS, 2010.
[33] J. Friedman, T. Hastie, and R. Tibshirani. A Note on the Group LASSO and a Sparse Group LASSO. Preprint, 2010.
[34] M. Hebiri, "Regularization with the Smooth-Lasso Procedure," preprint, 2008.
[35] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. J. Comp. Graph. Stat., 15:265-286, 2006.
[36] H. Liu, M. Palatucci, and J. Zhang. Blockwise Coordinate Descent Procedures for the Multi-task LASSO, with Application to Neural Semantic Basis Discovery. In: Proc. Int. Conf. Mach. Learning, 2009.
[37] J. Liu, S. Ji, and J. Ye. Multi-task Feature Learning via Efficient $l_{2,1}$-norm minimization. In: Proc. Conf. Uncertainty in Artificial Intell., 2009.
[38] E. van den Berg and M.P. Friedlander. Probing the Pareto Frontier for Basis Pursuit Solutions. SIAM J. Sci. Comput., 31:890-912, 2008.
[39] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/.
[40] R. Manentiab, S.F. Cappab, P.M. Rossiniac, and C. Miniussiad. The Role of the Prefrontal Cortex in Sentence Comprehension: An rTMS Study. Cortex, 44:337-244, 2008.
[41] R. Vandenberghe, C. Price, R. Wise, O. Josephs, and R.S.J. Frackowiak. Functional Anatomy of a Common Semantic System for Words and Pictures. Nature, 383:254-256, 1996.

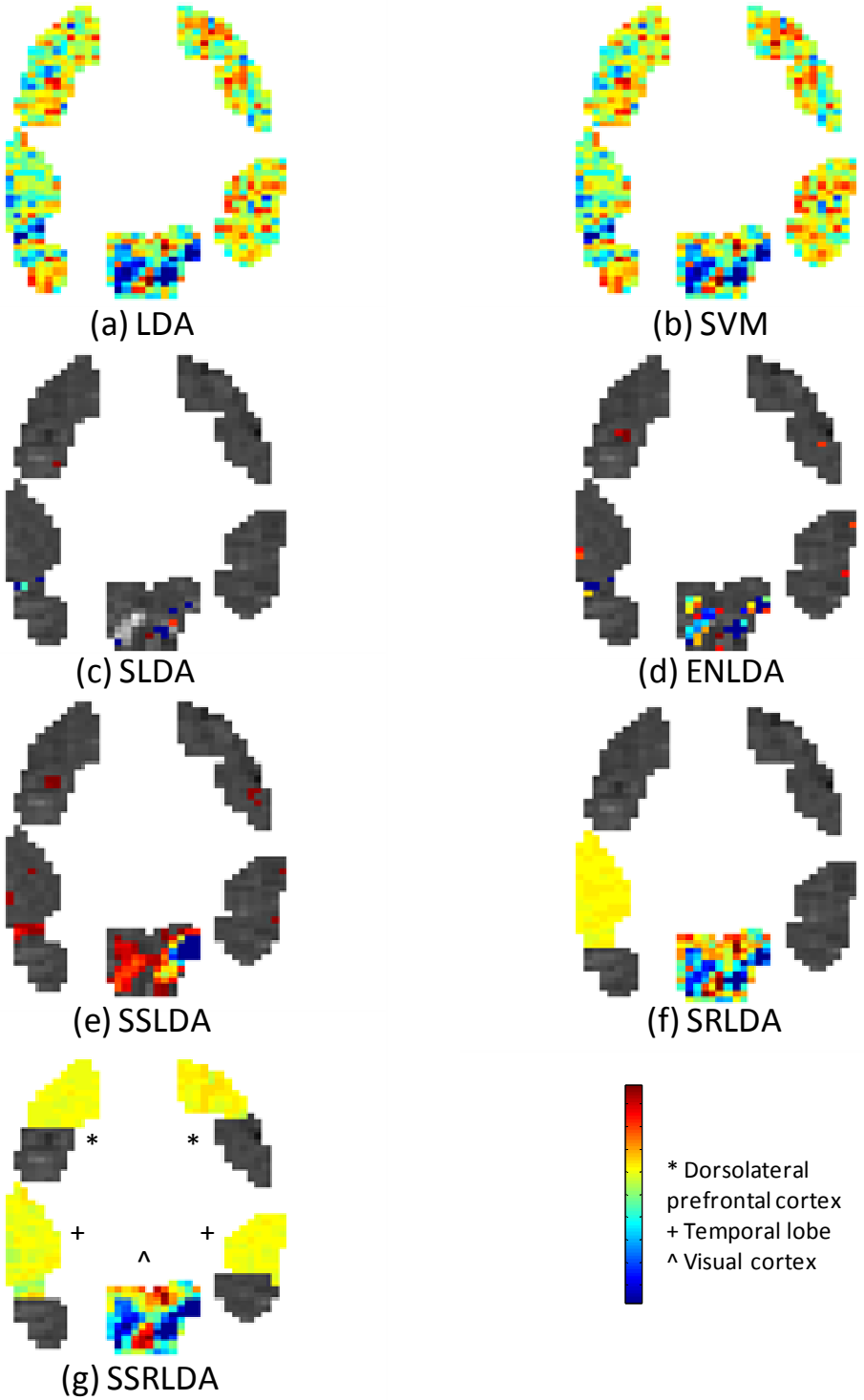

Figure. 2. Classifier weight patterns of contrasted methods. (a) LDA and (b) linear SVM resulted in randomly-distributed weight patterns. (c) SLDA generated overly-sparse weight patterns, which was partially alleviated with (d) EN-LDA. (e) SSLDA produced spatially smooth patterns, but the patterns do not facilitate clear-cut isolation of discriminant brain regions. (f) SRLDA provided region-level sparse weightings, but the weight patterns do not possess the spatial smoothness that SSLDA offers. Modeling both group structure and feature associations using (g) SSRLDA enabled brain regions pertinent to the picture/sentence classification task to be more easily identified. Smoother weight patterns were also obtained as compared to SRLDA, which we quantified using SDM [13].