Generalized Sparse Regularization with Application to fMRI Brain Decoding

Bernard Ng and Rafeef Abugharbieh

Biomedical Signal and Image Computing Lab, UBC, Canada
bernardyng@gmail.com

Abstract. Many current medical image analysis problems involve learning thousands or even millions of model parameters from extremely few samples. Employing sparse models provides an effective means for handling the curse of dimensionality, but other propitious properties beyond sparsity are typically not modeled. In this paper, we propose a simple approach, generalized sparse regularization (GSR), for incorporating domain-specific knowledge into a wide range of sparse linear models, such as the LASSO and group LASSO regression models. We demonstrate the power of GSR by building anatomically-informed sparse classifiers that additionally model the intrinsic spatiotemporal characteristics of brain activity for fMRI classification. We validate on real data and show how prior-informed sparse classifiers outperform standard classifiers, such as SVM and a number of sparse linear classifiers, both in terms of prediction accuracy and result interpretability. Our results illustrate the added-value in facilitating flexible integration of prior knowledge beyond sparsity in large-scale model learning problems.

Keywords: brain decoding, fMRI classification, prior-informed learning, sparse optimization, spatiotemporal regularization

1 Introduction

Recent years witnessed a surging interest in exploiting sparsity [1-9] to handle the ever-increasing scale and complexity of current medical image analysis problems [10,11]. Oftentimes, one is faced with exceedingly more predictors than samples. Under such ill-conditioned settings, incorporating sparsity into model learning proved to be of enormous benefits. In particular, enforcing sparsity enables model parameters associated with irrelevant predictors to be implicitly removed, i.e. shrunk to exactly zero [1], which reduces overfitting thus enhances model generalizability. Learning parsimonious models by imposing sparsity also simplifies result interpretation [1], which is of utmost importance in most medical studies.

Since the advent of the least absolute shrinkage and selection operator (LASSO) regression model [1], where Tibshirani showed that penalizing the l_1 norm induces sparsity in the regression coefficients, numerous powerful variants were subsequently proposed [2-9]. Zou and Hastie, for instance, proposed the elastic net penalty [2], which retains the sparse property of LASSO but additionally encourages correlated

predictors to be jointly selected in a data-driven manner. For applications where a natural grouping of the predictors exists, Yuan and Lin proposed the group LASSO penalty [3], which sparsely selects subsets of predefined groups with non-zero weights assigned to all predictors within the selected groups. To re-enable predictor-level sparsity back into group LASSO, Sprechmann et al. proposed combining the LASSO and group LASSO penalties under the name, hierarchical LASSO [4], also known as sparse group LASSO [5]. Other extensions include collaborative hierarchical LASSO [4] and overlapped group LASSO [6] among many others. All these models provide effective means for imposing structural constraints on the model parameters [4], but lack the flexibility for incorporating potentially advantageous problem-specific properties beyond sparsity [7]. For example, adjacent pixels in an image are typically correlated. Model parameters associated with these pixels should thus be assigned similar magnitudes to reflect the underlying correlations. However, merely enforcing sparsity does not model such associations between predictors.

To encourage smoothness in model parameters, in addition to sparsity, Tibshirani et al. proposed the fused LASSO model [8], which combines the LASSO penalty with a term that penalizes the l_1 norm of the differences between model parameters of adjacent predictors. To extend beyond smoothness, Tibshirani and Taylor recently proposed the Generalized LASSO model [7], which penalizes the l_1 norm of a weighted combination of the model parameters. By varying the choice of weights, Generalized LASSO facilitates numerous applications, such as trend filtering, wavelet smoothing, and outlier detection [7]. In a previous work [9], we proposed an extension of LASSO that also enables such modeling flexibility but is much simpler to optimize. Specifically, we proposed adding a generalized ridge penalty (l_2 norm of a weighted combination of the model parameters) to the LASSO regression model and showed that the resulting optimization problem can be efficiently minimized with existing LASSO solvers [12-15].

In this paper, we propose a simple yet effective approach for incorporating prior knowledge into a wide collection of sparse linear models, as motivated by our previous work [9]. We refer to our approach as generalized sparse regularization (GSR). In contrast to [7] and [9], we show that GSR is applicable to a much broader set of sparse linear models than just the LASSO regression model. The adaptability of GSR to such a wide range of models stems from how any l_2 norm penalty can be merged into an l_2 data fitting loss through a simple augmentation trick. Thus, adding a generalized ridge penalty to any sparse linear models with an l_2 data fitting loss, as commonly used in regression models [1-9], enables problem-specific properties to be easily integrated while preserving the functional form of the original optimization problem that these models entail. This desirable property of GSR facilitates the direct deployment of a wealth of existing sparse optimizers [12-15].

To demonstrate the power of GSR, we apply GSR to a large-scale functional magnetic resonance imaging (fMRI) classification problem, where only tens of samples are available for training a classifier with several tens of thousands of coefficients. Recent research in this area of fMRI analysis has mainly focused on exploitation of sparsity-enforcing techniques, such as sparse logistic regression [16,17], elastic net [18], and group LASSO [19] among others [20], to mitigate the curse of dimensionality. However, merely enforcing sparsity does not promote spatial smoothness in classifier weight patterns [9], which deviates from how spatially

proximal voxels tend to display similar level of brain activity [21]. We previously proposed remedying this limitation by modeling spatial correlations using a generalized ridge penalty [9]. Recently, van Gerven et al. proposed a Bayesian formulation for incorporating a spatiotemporal prior, where the authors opted to model uncertainty by estimating the posterior probabilities of the classifier weights as opposed to obtaining sparse weightings through finding the maximum a posterior solution. [22]. In this work, we model other characteristics of brain activity, in addition to spatial correlations, by exploiting the flexibility of GSR. In particular, we apply GSR to build anatomically-informed sparse classifiers that simultaneously model the intrinsic spatiotemporal structure in fMRI data, and explore whether incorporating such additional prior knowledge can further enhance prediction accuracy and result interpretation.

2 Methods

2.1 Overview of Sparse Linear Models

In this section, we focus on the problem of regression, since most sparse linear models [1-9] are inherently designed for such application, and defer discussion of how the sparse linear models described in this section can be employed for classifier learning in Section 2.3. Consider the standard least square regression problem:

$$\hat{a} = \min_{a} \left\| y - Xa \right\|_2^2 , \tag{1}$$

where y is an $N\times1$ response vector, X is an $N\times M$ predictor matrix, a is an $M\times1$ coefficient vector, N is the number of samples, and M is the number of predictors. The closed-form solution of (1) is given by:

$$\hat{a} = \left(X^T X \right)^{-1} X^T y . \tag{2}$$

When $N << M$, which is typical in many medical imaging problems [10,11], $(X^TX)^{-1}$ is ill-conditioned. Thus, direct estimation of $\hat{a}$ using (2) usually results in overfitting. To obtain a more generalizable estimate of $\hat{a}$, a common strategy is to employ regularization. In particular, Tibshirani proposed enforcing sparsity on a to achieve the dual objective of reducing overfitting and enhancing interpretability [1]:

$$\hat{a} = \min_{a} \left\| y - Xa \right\|_2^2 + \alpha \left\| a \right\|_1 , \tag{3}$$

where $\alpha \geq 0$. The model (3) is commonly referred to as the LASSO regression model, where Tibshirani showed that penalizing the l_1 norm induces sparsity on a [1].

The success of the LASSO regression model in numerous applications resulted in an explosion of research on sparse linear models [14]. Although many LASSO variants involve only a simple addition of other penalty terms to (3), the modeling power that these extensions facilitate proved substantial. For example, Zou and Hastie proposed adding a ridge penalty to (3), which is known as the elastic net model [2]:

$$\hat{a} = \min_a \left\| y - Xa \right\|_2^2 + \alpha \left\| a \right\|_1 + \beta \left\| a \right\|_2^2, \tag{4}$$

where $\beta \geq 0$. This model has two key advantages over LASSO. First, in contrast to LASSO, the number of non-zero elements in a is no longer limited by the number of samples [2]. This property is especially important in medical imaging applications, since the number of samples is often much smaller than the number of predictors. Second, in cases where the predictors are correlated, elastic net tends to jointly select the correlated predictors, whereas LASSO would arbitrarily select only one predictor among each correlated set [2].

Another widely-used extension of LASSO is the group LASSO, proposed by Yuan and Lin for applications where a natural grouping of the predictors is present [3]:

$$\hat{a} = \min_a \left\| y - Xa \right\|_2^2 + \alpha \left\| a_g \right\|_{2,1}, \tag{5}$$

$$\left\| a_g \right\|_{2,1} = \sum_{h=1}^{H} \left\| a_{g_h} \right\|_2, \tag{6}$$

where a_{gh} are the coefficients associated with predictors belonging to the predefined group g_h, and $h \in \{1, \ldots, H\}$ where H is the number of predefined groups [3]. As evident from (6), $\|a_g\|_{2,1}$ is exactly the l_1 norm of $\|a_{gh}\|_2$. Thus, minimizing $\|a_g\|_{2,1}$ encourages sparse subsets of *groups* to be selected with non-zero coefficients assigned to all predictors within each selected group. However, since not all predictors within a group are necessarily relevant, Sprechmann et al. [4] proposed the hierarchical LASSO model, also known as sparse group LASSO [5]:

$$\hat{a} = \min_a \left\| y - Xa \right\|_2^2 + \alpha \left\| a_g \right\|_{2,1} + \beta \left\| a \right\|_1, \tag{7}$$

which is essentially group LASSO combined with a LASSO penalty. Reintroducing the LASSO penalty encourages internal sparsity such that only a sparse subset of predictors within each selected group is chosen [4]. This property of (7) is particularly useful in cases where one is uncertain about the exact group assignment.

To extend (7) to scenarios where multiple sets of samples, y^s, are generated from different processes, e.g. when observations are collected from multiple subjects [11], but associated with the same set of predictors, e.g. using the same set of semantic features to predict brain activation patterns of different subjects [11], Sprechmann et al. [4] proposed the collaborative hierarchical LASSO model:

$$\hat{a} = \min_a \left\| Y - XA \right\|_F^2 + \alpha \sum_{k=1}^{K} \left\| A_{g_h} \right\|_F + \beta \sum_{s=1}^{S} \left\| a^s \right\|_1, \tag{8}$$

where S is the number of processes, $Y = (y^1, \ldots, y^S)$, $A = (a^1, \ldots, a^S)$, A_{gh} are the rows of A belonging to group g_h, and a^s is the coefficient vector of process s. $\| \cdot \|_F$ denotes the Frobenius norm. Minimizing (8) promotes the same groups to be selected across processes with the chosen sparse subset of predictors within each selected group being potentially different [4]. This model thus enables information across processes

to be shared while providing flexibility in handling inter-process variability by allowing sparsity patterns within the selected groups to vary across processes.

All of the LASSO extensions above, as well as others not discussed in this paper due to space limitations, provide effective means for modeling data structure through promoting groups of predictors to be jointly selected. However, the associations between predictors are ignored in these models, i.e. associated predictors may be jointly selected but the coefficients assigned to these predictors can greatly vary. Since the associations between predictors can be an important source of information for further constraining the ill-conditioned problem (1), we propose a simple approach to incorporate predictor associations into all of the models above, as discussed next.

2.2 Generalized Sparse Regularization

To integrate properties beyond sparsity into sparse linear models as those described in Section 2.1, we propose complementing these models with the following penalty:

$$J_{GSR}(a) = \left\|\Gamma a\right\|_2^2, \tag{9}$$

where $\Gamma = (\gamma_1, \ldots, \gamma_R)^{\mathrm{T}}$ is an $R \times M$ penalty matrix for modeling the associations between predictors, γ_r is an $M \times 1$ vector encoding the penalty for each association r, and R is the number of associations being modeled. For instance, if predictors x_p and x_q are associated with each other, say by virtue of being spatially adjacent voxels, then a_p and a_q should presumably be assigned similar values to reflect this intrinsic association. This can be accomplished by, e.g. setting γ_{rp} to 1 and γ_{rq} to -1 such that differences in a_p and a_q are penalized. Many other context-specific properties can be analogously modeled by varying Γ as discussed in Section 2.4 and [7].

One may envisage other penalties to facilitate similar modeling flexibility. However, not all penalties added to the models in Section 2.1 will result in a practical optimization problem. The critical question is thus whether the optimization problem resulting from integrating (9) into models in Section 2.1 can be efficiently minimized. To address this question, we draw upon a basic property of the Euclidean norm:

$$\sum_{\omega=1}^{\Omega} \left\|z_\omega\right\|_2^2 = \left\|z\right\|_2^2, \quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_\Omega \end{pmatrix}, \tag{10}$$

where z_ω is a vector of arbitrary length. Since models (3), (4), (5) and (7) in combination with (10), can all be written in the form:

$$\hat{a} = \min_a \left\|y - Xa\right\|_2^2 + \lambda\left\|\Gamma a\right\|_2^2 + J(a), \tag{11}$$

where $\lambda \geq 0$ and $J(a)$ is the sparse penalty term of the respective models, invoking (10) on (11) results in the following optimization problem:

$$\hat{a} = \sqrt{1+\lambda}\, \min_{\tilde{a}} \left\|\tilde{y} - \tilde{X}\tilde{a}\right\|_2^2 + J(\tilde{a}), \tag{12}$$

$$\tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \tilde{X} = (1+\lambda)^{-\frac{1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda}\Gamma \end{pmatrix}, \tag{13}$$

which has the exact same functional form as the optimization problems of the original sparse linear models. Thus, existing solvers of the respective sparse linear models [12-15] can be directly employed to efficiently minimize (11).

For the multi-process model (8), if the associations between predictors are similar across processes, then the same matrix augmentation trick can be applied with the zero vector in (13) replaced by an $R \times S$ zero matrix. If associations between predictors vary across processes, a minor modification is needed:

$$\tilde{y} = \begin{pmatrix} vector(Y) \\ 0 \end{pmatrix}, \quad \tilde{X} = (1+\lambda)^{-\frac{1}{2}} \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & X \\ & & \sqrt{\lambda}\Gamma & \end{pmatrix}, \tag{14}$$

where vector(·) is an operator that stacks the columns of its argument into a vector, and a in the first term of (11) is now vector(A). We highlight that, in addition to enabling different predictor associations to be modeled for each process, (14) permits modeling of predictor associations *across* processes. This modification of the matrix augmentation trick thus builds even greater flexibility into (8).

2.3 GSR for Classification

The models described in Section 2.1 are inherently designed for regression problems with y being a continuous variable. To extend the GSR-augmented sparse regression models to the classification setting without altering the functional form of the associated optimization problems, we employ the spectral regression technique [23]:

Step 1. Learn the constraint-free optimal projection of the training data X, e.g. using graph embedding (GE) [24]:

$$Wy = \lambda Dy\,, \tag{15}$$

where y is the projection of X on the subspace defined by W [24]. W_{ij} models the intrinsic relationships between samples i and j of X, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. We note that the key advantage of GE is that it enables various subspace learning algorithms to be used as classifiers by simply varying W [24].

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

$$\hat{a} = \min_a \left\| y - Xa \right\|_2^2 + Q(a)\,. \tag{16}$$

Clearly, setting $Q(a)$ to $\lambda\|\Gamma a\|_2^2 + J(a)$ results in our proposed GSR model (11). Hence, exploiting spectral regression in combination with GSR facilitates sparsity and predictor associations to be jointly integrated into classifier learning.

2.4 Application to fMRI Spatiotemporal Classification

Given N M×1 feature vectors, x_i, forming the rows of an N×M predictor matrix, X, the general classification problem involves finding the corresponding N×1 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 $g_h \in G = \{g_1, \ldots, g_H\}$. In the context of spatiotemporal fMRI classification, we treat the signal intensity of each brain voxel p at time t_k within a trial as a feature, and all brain volumes within a trial of the same experimental condition as a sample x_i, as illustrated in Fig. 1.

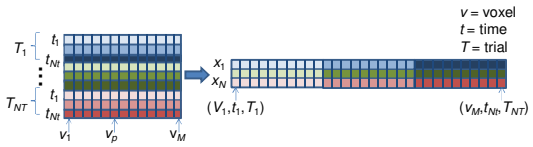

Fig. 1. Predictor matrix. Brain volumes within the same trial are concatenated and treated a single sample x_i. N_t is the number of volumes within a trial and N_T is the number of trials for a particular experimental condition.

Our goal is thus to determine to which condition, l_i, does each concatenated set of brain volumes x_i belongs. Since the brain is known to be functionally organized into specialized neural regions [25], this provides a natural grouping of the voxels.

To model this modular property of the brain and the spatiotemporal correlations in brain activity, we build an anatomically-informed spatiotemporally smooth sparse linear discriminant analysis (ASTSLDA) classifier by first solving for y in (15) with:

$$W_{ij} = \begin{cases} 1/m_c, & l_i = l_j = c \\ 0, & otherwise \end{cases}, \tag{17}$$

where m_c is the number of samples in class c [23]. We then optimize (11) with $J(a)$ set to the group LASSO penalty (6), and the signal intensity of the voxels within each brain region of interest (ROI) at each time point t_k treated as a group (Section 3). Γ is set as the spatiotemporal Laplacian operator:

$$\Gamma_{p_k q_s} = \begin{cases} -1, q_s \in N_{p_k} \\ 0, otherwise \end{cases}, \quad \Gamma_{p_k p_k} = -\sum_{q_s \neq p_k} \Gamma_{p_k q_s}, \tag{18}$$

where N_{pk} is the spatiotemporal neighborhood of voxel p at time t_k. Specifically, 6-connected spatial neighbors and signal intensity of voxel p itself at adjacent time points are defined as the spatiotemporal neighbors.

We can easily build other sparse LDA variants in an analogous manner. To build an anatomically-informed spatially smooth sparse LDA (ASSLDA) classifier, we employ (18) but with N_{pk} being the 6-connected spatial neighborhood of voxel p. To build a sparse LDA classifier that is only anatomically-informed (ASLDA), we set λ in (11) to 0 with $J(a)$ being the group LASSO penalty (6). Voxel-level sparse LDA classifiers that incorporates a spatiotemporal prior (STSLDA), a spatial prior (SSLDA), and no prior (SLDA) can be built in a similar manner as their anatomically-informed counterparts, but with $J(a)$ in (11) being the LASSO penalty (3). LDA classifier with elastic net penalty (ENLDA) can also be built from (11) by setting Γ to identity and $J(a)$ as the LASSO penalty (3). Our proposed model (11) thus encompasses many of the widely-used sparse linear models.

3 Materials

The publicly available StarPlus database [26] was used for validation. We provide here a brief description of the data. Details can be found in [26,27]. In the StarPlus experiment, all subjects performed 40 trials of a sentence/picture matching task. In each trial, subjects were required to look at a picture (or sentence) followed by a sentence (or picture), and then 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. fMRI brain volumes were acquired from 13 normal subjects at a TR of 500 ms, but only 6 of the subjects' data are available online [26]. Each subject's dataset comprised voxel time courses within 25 ROIs that were chosen by neuroscience experts, resulting 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. No spatial normalization was performed.

4 Results and Discussion

In this work, we explored the implications of incorporating prior knowledge into model learning using GSR on a large-scale fMRI classification problem. We treated the signal intensity of each voxel at each time point within a trial as a feature, and all brain volumes within the same trial of each experimental condition as a sample. The classification task was thus to discriminate sets of brain volumes associated with a picture from those associated with a sentence. To account for the delay in the hemodynamic response (HDR) [28], we only used the 8 brain volumes collected 4 s after stimulus onset within each trial. Each sample thus consisted of approximately 40,000 features (i.e. roughly 5000 voxels × 8 volumes, see Section 3) with 40 samples per class (i.e. 40 trials per class) for each subject. For comparison, we contrasted ASTSLDA against LDA, linear SVM, SLDA, ENLDA, SSLDA [9], ASLDA, and ASSLDA. We employed the spectral projected gradient technique (SPG) [13] to

minimize the optimization problem of the respective classifier models. SPG was chosen for its ability to efficiently handle large-scale problems (classifiers with ~40,000 coefficients in this study). All contrasted classifiers could be learned within a few seconds using SPG for a fixed set of parameter values λ and α (β was not involved in the classifier models contrasted in this work). Ten-fold nested cross validation was used [27] to select λ and α, and estimate the prediction accuracy. The average prediction accuracies over subjects are shown in Fig. 2. An axial slice of the classifier weight patterns corresponding to 5.5 s and 6 s after stimulus onset (i.e. when HDR typically peaks) was also plotted (Fig. 3) for result interpretations. Only 3 exemplar subjects were included due to space limitations.

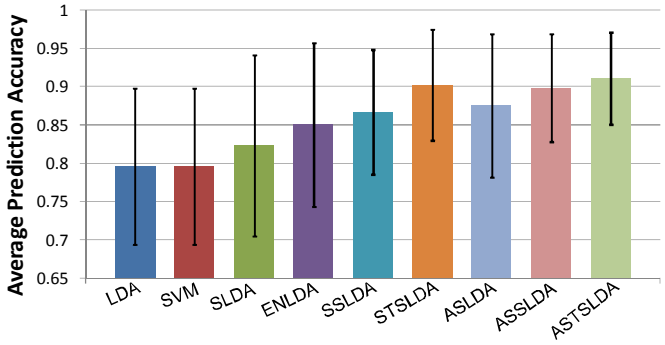

Fig. 2. Prediction accuracy comparisons. ASTSLDA outperformed all other contrasted methods with an average accuracy of 91%.

LDA resulted in the worst average prediction accuracy, which was likely due to overfitting. The classifier weight patterns also appeared spatially-random. Using SVM led to similar predictive performance and randomly-distributed weight patterns. Reducing overfitting using SLDA increased accuracy, but the weight patterns seemed overly-sparse. The reason was likely due to LASSO's constraint on the number of non-zero elements in the classifier weight vector a, i.e. restricted by the sample size [2]. In fact, since we treated the signal intensity of the brain volumes within a trial as a single feature vector, weights would be spread across time, resulting in even sparser spatial patterns than what would be obtained in the conventional case where each brain volume is taken as a sample. Also, we observed little consistency in the SLDA weight patterns across subjects and time.

Alleviating LASSO's constraint on the number of non-zero elements in a using ENLDA [2] improved prediction accuracy over SLDA. However, the weight patterns remained overly-sparse, which demonstrate the highly-overlooked fact that higher predictive accuracy does not necessarily translate to more neurologically-sensible weight patterns. In contrast, modeling spatial correlations using SSLDA resulted in higher prediction accuracy and spatially smoother weight patterns that better conform to how spatially proximal voxels tend to be correlated [21]. Our results thus illustrate the added-value of incorporating prior knowledge, both in terms of predictive performance and result interpretability. These benefits of exploiting prior knowledge

were further exemplified using STSLDA, which obtained similar spatially smooth patterns but a further increase in prediction accuracy. This increase likely arose from how STSLDA pooled information across brain volumes through penalizing discrepancies in classifier weights between spatiotemporal neighbors.

Accounting for the modular structure of the human brain using ASLDA improved prediction accuracy over the non-anatomically-informed classifiers except STSLDA, which again demonstrate the importance of constraining the model learning problem with prior knowledge (i.e. through grouping of associated features) to handle the curse of dimensionality. However, the weight patterns displayed little consistency between subjects and across time points. Modeling spatial correlations using ASSLDA resulted in higher accuracy than ASLDA, but it too obtained (slightly) lower accuracy than STSLDA. These results further support that exploiting the commonality between adjacent brain volumes, in addition to modeling spatial correlations within each brain volume, can greatly improve predictive performance.

Modeling both the modular property of the brain and the spatiotemporal characteristics of brain activity using ASTSLDA resulted in the best overall predictive performance with an average accuracy of 91%, which is an 11% increase over LDA and SVM. ASTSLDA also achieved the lowest variability in prediction accuracy, thus demonstrating stable performance despite the considerable inter-subject variability often seen in fMRI studies [21]. Moreover, higher consistency in weight patterns was observed across both subjects and time, with weights correctly assigned to the dorsolateral prefrontal cortex, which is responsible for verification task (e.g. decide if a sentence matches a picture) [29], as well as the visual cortex along the calcarine fissure and the temporal lobe, which pertain to picture/sentence discrimination [30].

5 Conclusions

We proposed GSR, a general approach for enabling properties beyond sparsity to be incorporated as an integral part of sparse model learning. By exploiting a basic property of the Euclidean norm, we showed that GSR can be directly applied to many widely-used sparse linear models without altering the functional form of their respective optimization problems. GSR hence facilitates greater modeling flexibility without the need for devising new complex optimization routines. We validated GSR on a large-scale classification problem and demonstrated how jointly modeling the modular nature of the human brain and the intrinsic spatiotemporal structure of brain activity can substantially improve prediction accuracy over standard techniques, such as LDA and SVM. An important extension of our current work will be to employ hierarchical LASSO in combination with GSR to model how some voxels within an ROI may be irrelevant for task discrimination. Also, exploiting the commonality between subjects within a group by integrating GSR into the collaborative hierarchical model may further improve prediction, which we intend to explore.

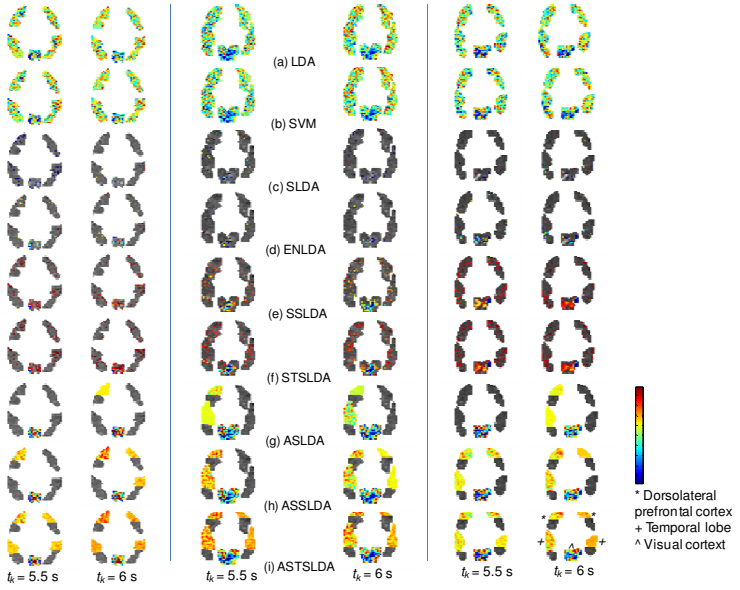

Fig. 3. Classifier weight patterns. Weight patterns 5.5 s and 6 s after stimulus onset for 3 exemplar subjects. (a) LDA and (b) SVM resulted in randomly-distributed weight patterns. (c) SLDA generated overly-sparse patterns, which was mildly improved with (d) EN-LDA. (e) SSLDA and (f) STSLDA produced spatially smoother patterns. (g) ASLDA and (h) ASSLDA weight patterns displayed little consistency across time and subjects. (i) ASTSLDA patterns were more consistent across subjects and time compared to the contrasted classifiers.

References

1. Tibshirani, R.: Regression Shrinkage and Selection via the LASSO. J. Royal Stat. Soc. Series B 58, 267--288 (1996)
2. Zou, H., Hastie, T.: Regularization and Variable Selection via the Elastic Net. J. Royal Stat. Soc. Series B 67, 301--320 (2005)
3. Yuan, M., Lin, Y.: Model Selection and Estimation in Regression with Grouped Variables. J. Royal Stat. Soc. Series B 68, 49--67 (2006)
4. Sprechmann, P., Ramirez, I., Sapiro, G.: Collaborative Hierarchical Sparse Modeling. Technical report, arXiv:1003.0400v1 (2010)
5. Friedman, J., Hastie, T., Tibshirani, R.: A Note on the Group LASSO and a Sparse Group LASSO. Technical report, arXiv:1001.0736v1 (2010)
6. Jacob, L., Obozinski, G., Vert., J.P.: Group Lasso with overlaps and graph Lasso. In: Proc. Int. Conf. Mach. Learn., pp. 433--440 (2009)
7. Tibshirani, R., Taylor, J.: The Solution Path of the Generalized Lasso. Ann. Stat. (in press)
8. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and Smoothness via the Fused Lasso. J. Royal Stat. Soc. Series B 67, 91--108 (2005)
9. Ng, B., Vahdat, A., Hamarneh, G., Abugharbieh, R.: Generalized Sparse Classifiers for Decoding Cognitive States in fMRI. In: Proc. MICCAI Workshop on Mach. Learn. Med. Imaging 6357, 108--115 (2010)

10. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531--536 (1999)
11. Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V. L., Mason, R. A., Just, M. A.: Predicting Human Brain Activity Associated with the Meanings of Nouns. Science 320, 1191--1195 (2008)
12. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. Ann. Stat. 32, 407--499 (2004)
13. van den Berg, E., Friedlander, M.P.: Probing the Pareto Frontier for Basis Pursuit Solutions. SIAM J. Sci. Comput. 31, 890--912 (2008)
14. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software 33, 1--22 (2010)
15. Schmidt, M., Fung, G., Rosales, R.: Optimization Methods for L1-Regularization. Technical report, the University of British Columbia (2009)
16. Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse Estimation Automatically Selects Voxels Relevant for the Decoding of fMRI Activity Patterns. NeuroImage 42, 1414--1429 (2008)
17. Ryali, S., Supekar, K., Abrams, D.A., Menon, V.: Sparse Logistic Regression for Whole-brain Classification of fMRI Data. NeuroImage 51, 752--764 (2010)
18. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and Interpretation of Distributed Neural Activity with Sparse Models. NeuroImage 44, 112--122 (2009)
19. van Gerven, M., Takashima, A., Heskes, T.: 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)
20. Michel, V., Eger, E., Keribin, C., Thirion, B.: Multi-Class Sparse Bayesian Regression for Neuroimaging Data Analysis. In: Wang, F., Yan, P., Suzuki, K., and Shen, D. (eds.) MLMI 2010. LNCS, vol. 6357, pp. 50--57. Springer-Verlag Berlin, Heidelberg (2010)
21. Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B.: Dealing with the Shortcomings of Spatial Normalization: Multi-subject Parcellation of fMRI Datasets. Hum. Brain Mapp. 27, 678--693 (2006)
22. van Gerven, M., Cseke, B., de Lange, F.P., Heskes, T.: Efficient Bayesian Multivariate fMRI Analysis Using a Sparsifying Spatio-temporal Prior. NeuroImage 50, 150--161 (2010)
23. Cai, D., He, X., Han, J.: Spectral Regression: A Unified Approach for Sparse Subspace Learning. In: Proc. IEEE Int. Conf. Data Mining, pp. 73--82 (2007)
24. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph Embedding and Extension: A General Framework for Dimensionality Reduction. IEEE Trans. Pat. Ana. Machine Intell. 29, 40--50 (2007)
25. J.A. Fodor. The Modularity of the Mind. MIT. 2--47 (1983)
26. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/.
27. Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to Decode Cognitive States from Brain Images. Mach. Learn. 57, 145--175 (2004)
28. Liao, C.H., Worsley, K.J., Poline, J.B., Aston, A.D., Duncan, G.H., Evans, A.C.: Estimating the Delay of the fMRI Response. NeuroImage 16, 593--606 (2002)
29. Manentiab, R., Cappab, S.F., Rossiniac, P.M., Miniussiad, C.: The Role of the Prefrontal Cortex in Sentence Comprehension: An rTMS Study. Cortex 44, 337--244 (2008)
30. Vandenberghe, R., Price, C., Wise, R., Josephs, O., Frackowiak, R.S.J.: Functional Anatomy of a Common Semantic System for Words and Pictures. Nature 383, 254--256 (1996)