THE PAIRWISE ELASTIC NET SUPPORT VECTOR MACHINE FOR AUTOMATIC FMRI FEATURE SELECTION

Alexander Lorbert, Peter J. Ramadge

Dept. of Electrical Engineering, Princeton University, Princeton NJ

ABSTRACT

A support vector machine (SVM) regularized with the Pairwise Elastic Net (PEN) penalty is used to automatically select a sparse set of brain voxel clusters based on the fMRI responses to two stimuli classes. This requires solving the PEN-SVM quadratic program. We show how to design the PEN regularization to encode, in a graph-based fashion, the pairwise similarity structure of the voxel fMRI responses and how to control the spatial locality of the encoding using a voxel searchlight. The voxel similarity encoding is reflected in the sparse structure of the weights of trained PEN-SVM and these weights automatically select a sparse set of voxel clusters. We empirically demonstrate the effectiveness of the approach using a real-world, multi-subject fMRI dataset.

***Index Terms*—** Support Vector Machine, Pairwise Elastic Net, fMRI, Sparsity, Feature Selection

1. INTRODUCTION

The fMRI voxel selection problem requires selecting the subset of brain voxels, based on measured fMRI responses, that jointly discriminate between two stimuli. Traditionally this has been done by spatial smoothing and mass thresholding of a univariate statistic across voxels [1]. Alternatives include hypothesis testings on voxel clusters [2], and thresholding a statistic in a transformed domain (e.g. wavelet) [3, 4]. However, these methods sacrifice spatial resolution since averaging or clustering voxels hides fine patterns in the data.

Multivariate analysis holds the promise of more sophisticated selection mechanisms since it allows, for example, distributed patterns of activation to be captured that might be otherwise missed by univariate tests [5, 6, 7, 8, 9, 10]. A novel, recent approach selects voxels using tree-based spatial regularization of a univariate statistic [11, 12]. This achieves spatial precision and smoothness but uses a complex regularization method.

An alternative approach uses labeled training data to tune a pattern classifier. The optimized classifier weights are then used to select the informative voxels [9, 13]. However, this is not without pitfalls. First, the weights may require thresholding and hence the selection of a threshold parameter. Consequently, voxel selection is not "automatic". Second, the weights may not reflect expected properties of the informative voxels, e.g., spatial smoothness, spatial clustering, etc., since standard machine learning methods are not necessarily informed by these desired characteristics. The first issue leads to the idea that classifier weights should be sparse, thus ensuring automatic feature selection (no threshold). This, however, exacerbates the second problem as sparsity is achieved by selecting a weight pattern that is sufficient for classification but not reflective of the overall pattern of informative voxels [10].

This has led to a variety of recent work addressing these issues. To ensure spatial locality of the selected voxels one can use a two phase approach. First, run a multivariate analysis on a set of searchlights (spherical masked regions) [14], to test if the searchlight contains informative data. Then train a classifier on the preselected searchlight voxels. A variant of this two-stage framework is to train classifiers on several predefined masks, and then aggregate the classifiers using boosting [15, 16]. This is faster but assumes detailed prior knowledge to select the predefined masks. An alternative is a one step approach in which one constrains or regularizes the classifier during training to attain desired weight characteristics. For example, [17] used AdaBoost to train classifiers with "rich features" (features involving the values of several adjacent voxels) to capture spatial structure in the data. This yielded superior performance but selected "rich features" rather than discriminating voxels. The method of [18], also based on boosting, uses voxels as base classifiers and favors adding base classifiers (voxels) that are spatially contiguous to voxels already selected in the boosted classifier.

Recently, there has been interest in using support vector machine (SVM) [19] methods to capture multivariate relationships in MRI and fMRI data [9, 13, 20]. However, without regularization the SVM weights need not exhibit desired spatial characteristics across voxels. To address this [20] considers various forms of quadratic regularization based on Laplacian operators that encode spatial and anatomical consistency, e.g., voxel-to-voxel proximities.

The focus of this paper is on using a trained regularized SVM to automatically select brain voxels (features) based on measured fMRI responses to two classes of stimuli. What is distinctive about our approach is that we regularize the SVM with the Pairwise Elastic Net (PEN) regularization penalty. This leads to what we call the PEN-SVM quadratic program.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

The PEN regularization encodes, in a graph-based fashion, the pairwise similarity structure of the voxel fMRI responses. The spatial locality of this encoding can be controlled using a searchlight. This principled similarity encoding is then reflected in the structure of the trained PEN-SVM. Since the PEN penalty seeks similarity modulated weight sparsity, the trained PEN-SVM weights automatically select a sparse set of voxel clusters.

The remainder of the paper is organized thus: the PEN-SVM is presented in §3 and principled, graph-based construction of the penalty matrix $\mathbf{P}$ is detailed in §3.1. Experimental results on fMRI data are given in §4 and we conclude in §5.

2. PRELIMINARIES

Given a set of m example-label pairs, $\{\mathbf{x}_i, y_i\}_{i=1}^m$, with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{\pm 1\}$, an SVM seeks $(\mathbf{w}, b)$ to minimize the regularized hinge-loss [19]:

$$\arg\min_{(\mathbf{w},b)} \textstyle\sum_{i=1}^m \left[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right]_+ + J(\mathbf{w}), \tag{1}$$

where $[z]_+ \triangleq \max\{z, 0\}$ is the positive part of $z \in \mathbb{R}$ and for the standard SVM $J(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$. We will call this ridge-SVM. In additional we consider:

$$J(\mathbf{w}) = \begin{cases} \lambda\|\mathbf{w}\|_1 & \text{lasso-SVM} \\ \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2 & \text{EN-SVM} \\ \lambda_1\|\mathbf{w}\|_1 + \lambda_2\mathbf{w}^T\mathbf{\Lambda}\mathbf{w} & \text{SEN-SVM} \\ \lambda|\mathbf{w}|^T\mathbf{P}|\mathbf{w}| & \text{PEN-SVM} \end{cases} . \tag{2}$$

Here $|\mathbf{w}| \in \mathbb{R}^n$ with $|\mathbf{w}|_i = |w_i|$, and $\mathbf{\Lambda}, \mathbf{P} \in \mathbb{R}^{n\times n}$ are symmetric, positive semidefinite. The lasso-SVM is a linear program (LP) and the other SVMs are quadratic programs (QP). In all cases, the parameter λ is selected using cross validation to maximize training data classification accuracy.

The lasso-SVM was considered in [21, 22, 23]: the ℓ_1 penalty is responsible for producing a sparse weight vector, and is therefore appropriate for automatic feature selection. The elastic net SVM (EN-SVM) was first proposed in [24] where it was called the doubly regularized support vector machine (DrSVM). It attempts to balance the clustering property of ridge with the sparsity of lasso. The structured elastic net SVM (SEN-SVM), [25], is also a form of double regularization where the $\mathbf{\Lambda}$ matrix encodes graph structure. The PEN penalty was first introduced for linear regression [26], where it was shown to encompass ridge, lasso, and elastic net. Moreover, PEN allows one to customize the sparsity relationship between any two features.

3. PEN-SVM

In [26], the authors proved that the PEN penalty $|\mathbf{w}|^T\mathbf{P}|\mathbf{w}|$ is convex if and only the matrix $\mathbf{P}$ is psd with nonnegative-valued entries. Additionally, various constructions for $\mathbf{P}$ were offered ranging from correlation-based to group-based. The key to designing an "appropriate" $\mathbf{P}$ is [normalized] similarity: when features i and j are similar, we want p_{ij} to be small, and vice versa. The extreme cases are

$$|\mathbf{w}|^T\mathbf{I}|\mathbf{w}| = \|\mathbf{w}\|_2^2 \tag{3}$$

$$|\mathbf{w}|^T(\mathbf{1}\mathbf{1}^T)|\mathbf{w}| = \|\mathbf{w}\|_1^2 , \tag{4}$$

which lead to a ridge penalty (3) and a squared lasso penalty (4). A ridge approach implicitly groups all features as indicated by the small (i.e., zero) off-diagonal elements in (3). In contrast, a lasso penalty is intended to accomplish sparse feature selection. It hence assumes that all features are pairwise dissimilar and sets the off-diagonal elements in (4) equal to 1. Note that the PEN penalty forms a ℓ_1-squared/ℓ_2 tradeoff rather than an ℓ_1/ℓ_2 tradeoff.

Assuming $\mathbf{P}$ is psd and nonnegative-valued, the PEN-SVM is a convex quadratic program. To see this, note that $|\mathbf{w}|^T\mathbf{P}|\mathbf{w}| = \min_{\mathbf{v}\succeq|\mathbf{w}|} \mathbf{v}^T\mathbf{P}\mathbf{v}$. Thus, we attain the PEN-SVM by solving the QP:

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T\mathbf{u} + \mathbf{v}^T\mathbf{P}\mathbf{v} \\ \text{subject to} & u_i \geq 1 - \mathbf{w}^T(y_i \cdot \mathbf{x}_i) - y_i b \\ & u_i \geq 0 \\ & v_j \geq w_j \\ & v_j \geq -w_j \quad . \end{array} \tag{5}$$

The unknowns are $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^n$, and $b \in \mathbb{R}$. The QP (5) can be solved, for example, using `cvx` [27, 28].

3.1. Graph-based regularization

We now outline a new automatic construction of the matrix $\mathbf{P}$ using a graph-based approach. Let $\mathbf{A} = [a_{ij}] = [\,\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n\,] \in \{0,1\}^{n\times n}$ be a symmetric adjacency matrix of an n-node graph. Hence $a_{ij} = a_{ji} = 1$ if nodes i and j are linked, and 0 otherwise. We also impose $a_{ii} = 0$, i.e., no self-links. The degree of node i is $d_i = \sum_j a_{ij}$. It follows that $d_i = \mathbf{1}^T\mathbf{a}_i = \mathbf{a}_i^T\mathbf{a}_i \in \{0, 1, \ldots, n-1\}$ and $\mathbf{d} = \mathbf{A}\mathbf{1}$. Lastly, we set $\mathbf{D} = \mathbf{diag}(\mathbf{d})$ and let $\mathbf{L} = \mathbf{D} - \mathbf{A}$ be the graph Laplacian [29]. Necessarily, $\mathbf{1}$ is an eigenvector of $\mathbf{L}$ with corresponding eigenvalue of zero.

We now construct $\mathbf{P} = [p_{ij}]$ as follows. Let $\mathbf{B} = \mathbf{1}\mathbf{1}^T - \mathbf{A} + \mathbf{D} = \mathbf{1}\mathbf{1}^T + \mathbf{L}$. Then set $\mathbf{C} = \mathbf{B}^\alpha$, some $\alpha \in \{1, 2, \ldots\}$. Finally let $p_{ij} = c_{ij}/\sqrt{c_{ii}c_{jj}}$.

Note that $\mathbf{B}$ is a nonnegative-valued psd matrix so $\mathbf{C}$ is also nonnegative-valued psd. The third step is a diagonal normalization to have $\mathbf{P}$ fit the mold of an auto-correlation matrix [30]. Thus, $\mathbf{P}$ is psd and nonnegative-valued, and so the resulting penalty function will be convex.

Since $\mathbf{L}\mathbf{1} = \mathbf{0}$, it follows that

$$\mathbf{B}^\alpha = (\mathbf{1}\mathbf{1}^T)^\alpha + \mathbf{L}^\alpha = n^{\alpha-1}\mathbf{1}\mathbf{1}^T + \mathbf{L}^\alpha \quad . \tag{6}$$

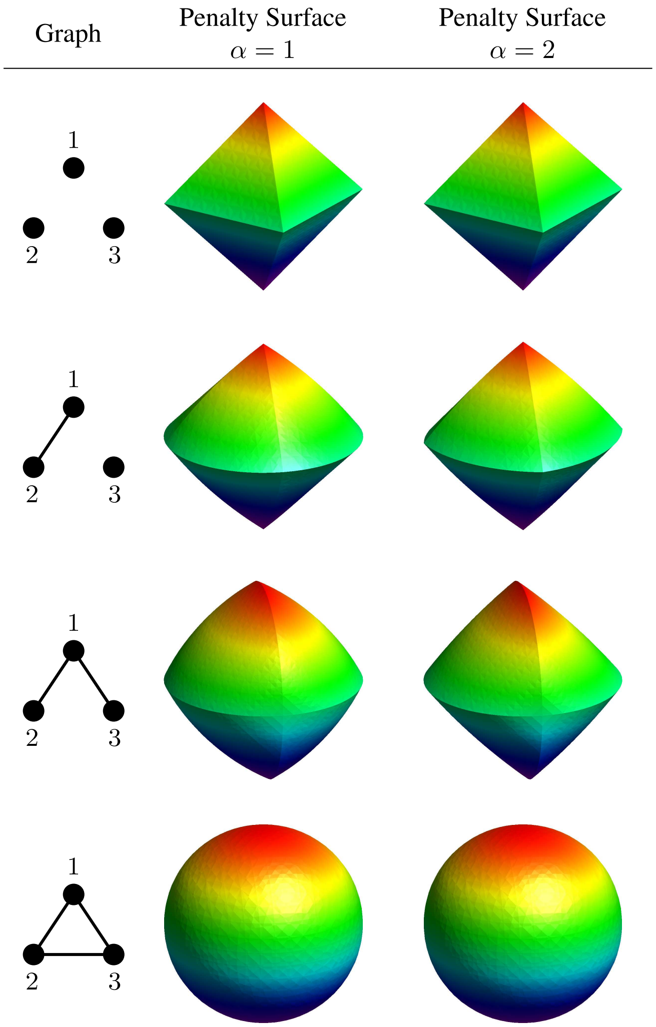

Fig. 1. Penalty surfaces resulting from 3-node graphs using construction in §3.1 (best seen in color).

For $\alpha = 1$, we have

$$c_{ij}^{\alpha=1} = \begin{cases} 1 + d_i & i = j \\ 1 - a_{ij} & i \neq j \end{cases} . \tag{7}$$

So the off-diagonal elements are 0 when nodes i and j are connected, and 1 when disconnected. For $\alpha = 2$:

$$c_{ij}^{\alpha=2} = \begin{cases} n + d_i + d_i^2 & i = j \\ n - (d_i + d_j)a_{ij} + \mathbf{a}_i^T \mathbf{a}_j & i \neq j \end{cases} . \tag{8}$$

The quantity $\mathbf{a}_i^T \mathbf{a}_j$ is a count of how many nodes are connected to both node i and node j. With a_{ij} indicating whether nodes i and j are connected, we see that there is a balancing act taking place within c_{ij} ($i \neq j$). Suppose that $d_i + d_j > 0$. It follows that c_{ij} is largest when nodes i and j are not connected but share many one-link connections. Similarly, when nodes i and j are connected and do not share many one-link connections, c_{ij} is smallest. Thus, c_{ij} possesses limited propagation information (one-hop and two-hop) connecting nodes i and j. Certainly, if nodes i and j do not share common connections and are not themselves connected, then $c_{ij} = n$. On the other hand, when nodes i and j are maximally connected in the one-hop and two-hop sense, i.e., $d_i = d_j = n - 1 = \mathbf{a}_i^T \mathbf{a}_j + 1$, then $c_{ij} = 0$.

We can summarize the above analysis as follows: when nodes i and j are connected, or "similar", we expect c_{ij} to be small. Conversely, when there is a lack of connection, we expect c_{ij} to be large. Recall that small [large] off-diagonal elements lead to a ridge [lasso] penalty. Therefore, we expect $\mathbf{P}$ to possess the encoded ℓ_1^2/ℓ_2 tradeoff. The parameter α encodes the propagation (α hops).

We can also ask what $\mathbf{P}$ will look like as $\alpha \to \infty$. First we note that the largest magnitude eigenvalue of $\mathbf{B}$ is n. This is because (i) $\mathbf{B}$ is a nonnegative matrix and (ii) $\mathbf{1}$, an eigenvector with eigenvalue n, is nonnegative-valued [31]. Note that $\mathbf{P}$ is invariant to a scaling of $\mathbf{C}$, so we can consider $n^{-\alpha}\mathbf{C} = n^{-\alpha}\mathbf{B}^\alpha = n^{-1}\mathbf{1}\mathbf{1}^T + n^{-\alpha}\mathbf{L}^\alpha$. If the largest eigenvalue of $\mathbf{L}$, λ, is less than n, then $n^{-\alpha}\mathbf{C} \to n^{-1}\mathbf{1}\mathbf{1}^T$ and $\mathbf{P} \to \mathbf{1}\mathbf{1}^T$ because $\lambda^\alpha n^{-\alpha} \to 0$. The only other scenario is that $\mathbf{B}$ has $k+1$ eigenvectors $\mathbf{V} = [\frac{1}{\sqrt{n}}\mathbf{1} \mid \mathbf{v}_1 \mid \cdots \mid \mathbf{v}_k]$ with eigenvalue n,[1] whereby $n^{-\alpha}\mathbf{C} \to \mathbf{V}\mathbf{V}^T$. We then normalize to obtain $\mathbf{P}$. As a final note, when the maximum degree of the graph is less than $n/2$, then the maximum eigenvalue of $\mathbf{L}$ will be less than n (Gershgorin Circle Theorem), implying $\mathbf{P} \to \mathbf{1}\mathbf{1}^T$ as $\alpha \to \infty$.

Figure 1 exhibits all possible 3-node graphs and the resulting penalty surfaces using the construction presented in §3.1. When there are no links present, a lasso surface is generated (ℓ_1 ball). Similarly, a clique yields a ridge surface (ℓ_2 ball). This will always be the case, i.e., $\mathbf{A} = \mathbf{0} \Rightarrow \ell_1$-ball and $\mathbf{A} = \mathbf{1}\mathbf{1}^T - \mathbf{I} \Rightarrow \ell_2$-ball. When there is a single connection ($1 \leftrightarrow 2$), the 12-plane has a ridge-like cross section while the 13- and 23-planes have a lasso-like cross section. Lastly, with two links present ($1 \leftrightarrow 2$ and $1 \leftrightarrow 3$), the 23-plane has a lasso-like cross section while the 12- and 13-planes have a lasso-like cross section.

4. EXPERIMENTAL RESULTS

Functional MRI data were collected from 10 subjects participating a block-design visualization experiment [32]. In each run, a subject was shown images belonging to a certain class—each class appearing once—for 16 TRs followed by 10 TRs of rest. The different image categories were (1) female-face, (2) male-face, (3) monkey, (4) house, (5) chair, (6) shoe, and (7) dog. To create an example, we took the time average of each voxel response over a 16 TR window, offset by 6 seconds to account for hemodynamic response. With 8 runs,

[1] This occurs, for example, when two or more nodes have maximal degree.

SVM	Sparsity	Training Acc.	Test Acc.
ridge	5963	80/80	74/80
lasso	29	80/80	71/80
EN	243	80/80	75/80
SEN	194	80/80	75/80
PEN	49	80/80	74/80

Table 1. Face-versus-house results from training on the first 4 runs and testing on the remaining 4 runs.

this yielded 560 labeled examples (10 subjects · 8 runs/subject · 7 examples/run). We then extracted the female-face and house examples from Talairach-aligned Ventral Temporal cortex (VT). This provided 160 examples (80 face + 80 house) of dimension 5994 (2997 voxels per hemisphere of VT). Then, we trained binary SVM classifiers on the first four runs and tested on the last four runs. Regularization parameters were selected to optimize training accuracy and then sparsity.

For both SEN-SVM and PEN-SVM we first constructed a 1/0 adjacency matrix by setting $a_{ij} = 1$ if (i) voxels i and j reside in the same hemisphere and (ii) voxel i is in voxel j's searchlight [14]. For a given voxel, we considered spherical searchlight of radius 3, leading to a 1.56%-sparse symmetric, adjacency matrix. We used $\alpha = 2$ for PEN-SVM.

Table 1 features the sparsity[2] and accuracy results. Additionally, the weight vectors are graphically overlaid in Figure 2. As only sign information is used to classify, each $(\mathbf{w}, b)$ pair was renormalized to have $\|\mathbf{w}\|_2 = 1$.

In terms of classification accuracy, there is no compelling reason to deviate from the standard ridge-SVM. This is not surprising given the small TR-count and large voxel-count [33]. The drawback of ridge-SVM, however, is poor feature selection. With the inclusion of an ℓ_1 (ℓ_1^2) penalty, the number of non-zero weights decreases to nearly 4.1% for EN-SVM, 3.2% for SEN-SVM, and less than 1% for both lasso-SVM and PEN-SVM. The distribution of the non-zero weights for EN-SVM are sparse but "speckled". SEN-SVM possesses a sparsity advantage over EN-SVM and is also less speckled. Whereas lasso-SVM provides the sparsest array of weights, there is no local clustering—only single voxels unto themselves. Like SEN-SVM, PEN-SVM provides a spatially-grouped and competitively-sparse weight vector.

5. CONCLUSION

We introduced the Pairwise Elastic Net Support Vector Machine. This adds PEN regularization to the hinge loss resulting in a quadratic program. The PEN penalty requires a matrix $\mathbf{P}$ with $O(n^2)$ degrees of freedom. We proposed a principled construction for $\mathbf{P}$ based on voxel response similarity that can be spatially modulated by a searchlight window. We also indicated how to incorporate a priori graph structure. Several questions remain. First, for a large number of features, employing a general purpose QP-solver is inefficient. An efficient direct method for solving the PEN-SVM would be preferred. Other avenues of investigation include alternative constructions for the PEN penalty matrix and the regularization path of the PEN-SVM [34].

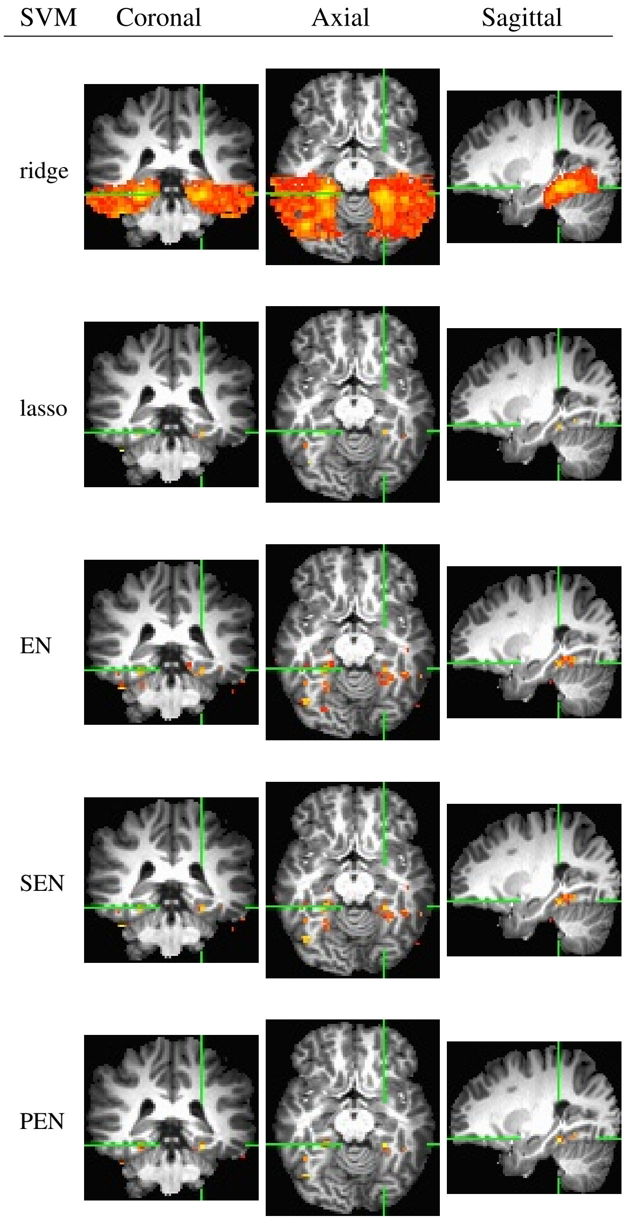

Fig. 2. Face-versus-house weight vectors for various SVMs (yellow [red] indicates a larger [smaller] magnitude).

[2] Given the size of the problem and the precision for which a solution was deemed acceptable, we took sparsity to mean the number of weights with magnitude less than 10^{-6}.

6. REFERENCES

[1] K.J. Friston, J. Ashburner, J. Heather, et al., "Statistical parametric mapping," *Neuroscience Databases: A Practical Guide*, p. 237, 2003.

[2] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini, "Cluster-based analysis of FMRI data," *NeuroImage*, vol. 33, no. 2, pp. 599–608, 2006.

[3] D. Van De Ville, T. Blu, and M. Unser, "Integrated wavelet processing and spatial statistical testing of fMRI data," *NeuroImage*, vol. 23, no. 4, pp. 1472–1485, 2004.

[4] D. Van De Ville, M.L. Seghier, F. Lazeyras, T. Blu, and M. Unser, "WSPM: Wavelet-based statistical parametric mapping," *NeuroImage*, vol. 37, no. 4, pp. 1205–1217, 2007.

[5] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[6] K.A. Norman, S.M. Polyn, G.J. Detre, and J.V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[7] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," *NeuroImage*, pp. 46–57, 2004.

[8] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "Compare: Classification of morphological patterns using adaptive regional elements," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 1, pp. 93 –105, jan. 2007.

[9] S. Kloppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Automatic classification of mr scans in alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[10] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fmri: A tutorial overview," *NeuroImage*, vol. 45, no. 1, Supplement 1, pp. S199 – S209, 2009.

[11] Z. Harmany, R. Willett, A. Singh, and R. Nowak, "Controlling the error in fmri: Hypothesis testing or set estimation?," in *Biomedical Imaging, 5th IEEE International Symposium on*, 2008, pp. 552–555.

[12] R.M. Willett and R.D. Nowak, "Minimax optimal level-set estimation," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2965–2979, 2007.

[13] P. Vemuri, J. L. Gunter, M. L. Senjem, Whitwell J. L., K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack Jr., "Alzheimer's disease diagnosis in individual subjects using structural mr images: Validation studies," *NeuroImage*, vol. 39, no. 3, pp. 1186 – 1197, 2008.

[14] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *PNAS*, vol. 103, no. 10, pp. 3863–3868, 2006.

[15] V. Koltchinskii, M. Martınez-Ramon, and S. Posse, "Optimal aggregation of classifiers and boosting maps in functional magnetic resonance imaging," *Advances in Neural Information Processing Systems*, vol. 17, pp. 705–712, 2005.

[16] M. Martínez-Ramón, V. Koltchinskii, G.L. Heileman, and S. Posse, "fMRI pattern classification using neuroanatomically constrained boosting," *NeuroImage*, vol. 31, no. 3, pp. 1129–1141, 2006.

[17] Melissa K. Carroll, Kenneth A. Norman, James V. Haxby, and Robert E. Schapire, "Exploiting spatial information to improve fmri pattern classification," in *12th Annual Meeting of the Organization for Human Brain Mapping, Florence, Italy*, 2006.

[18] Z.J. Xiang, Y.T. Xi, U. Hasson, and P.J. Ramadge, "Boosting with spatial regularization," in *Advances in Neural Information Processing Systems*, 2009.

[19] V.N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, 2000.

[20] R. Cuingnet, M. Chupin, H. Benali, and O. Colliot, "Spatial and anatomical regularization of SVM for brain image analysis," *Advances in Neural Information Processing Systems 23*, pp. 1–9, 2010.

[21] P.S. Bradley and O.L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *ICML*, 1998, pp. 82–90.

[22] M. Song, C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, and N. Tugcu, "Prediction of protein retention times in anion–exchange chromatography systems using support vector regression," *Journal of chemical information and computer sciences*, vol. 42, no. 6, pp. 1347–1357, 2002.

[23] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1–norm support vector machines," *NIPS*, vol. 16, no. 1, pp. 49–56, 2004.

[24] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, 2006.

[25] M. Slawski, "The structured elastic net for quantile regression and support vector classification," *Statistics and Computing*, pp. 1–16, 2012.

[26] A. Lorbert, D. Eis, V. Kostina, D.M. Blei, and P.J. Ramadge, "Exploiting covariate similarity in sparse regression via the pairwise elastic net," in *AISTATS*, 2010, vol. 9, pp. 477–484.

[27] Inc. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," `http://cvxr.com/cvx`, Sept. 2012.

[28] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds., Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.

[29] F.R.K. Chung, *Spectral graph theory*, vol. 92, American Mathematical Society, 1997.

[30] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[31] C. Meyer, *Matrix analysis and applied linear algebra book and solutions manual*, Number 71. Society for Industrial Mathematics, 2000.

[32] M.R. Sabuncu, B.D. Singer, B. Conroy, R.E. Bryan, P.J. Ramadge, and J.V. Haxby, "Function based inter-subject alignment of human cortical anatomy," *Cerebral Cortex*, 2009.

[33] P. Hall, J.S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *J. R. Statist. Soc. B*, vol. 67, no. 3, pp. 427–444, 2005.

[34] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, no. 2, pp. 1391, 2005.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.