A SEMI-SUPERVISED METHOD FOR MULTI-SUBJECT FMRI FUNCTIONAL ALIGNMENT

Javier S. Turek, Theodore L. Willke

Parallel Computing Lab
Intel Labs
Hillsboro, Oregon, USA

Po-Hsuan Chen, Peter J. Ramadge

Department of Electrical Engineering
Princeton University
Princeton, New Jersey, USA

ABSTRACT

Practical limitations on the duration of individual fMRI scans have led neuroscientist to consider the aggregation of data from multiple subjects. Differences in anatomical structures and functional topographies of brains require aligning data across subjects. Existing functional alignment methods serve as a preprocessing step that allows subsequent statistical methods to learn from the aggregated multi-subject data. Despite their success, current alignment methods do not leverage the labeled data used in the subsequent methods. In this work we propose a semi-supervised scheme that simultaneously learns the alignment and performs the analysis. We derive a specific instance of the scheme using the Shared Response Model for alignment and Multinomial Logistic Regression for classification. In our experiments this method improves the average classification accuracy from 65.5% to 68.5%, and from 5.3% to 6.1% over the independently-trained methods. Furthermore, our method achieves similar prediction with almost half the samples used for alignment.

Index Terms— fMRI, functional alignment, shared response model, semi-supervised method

1. INTRODUCTION

Functional Magnetic Resonance Imaging, or fMRI, is a non-invasive imaging technique that allows neuroscientists to learn how the human brain works. A neuroscience experiment proceeds by having a subject inside the scanner doing some tasks, while three-dimensional volumes are acquired. The volumetric information shows the activity of the subject's brain. Physical and practical factors limit the duration of an experiment, and because fMRI has high-spatial-low-temporal resolution, the number of volumes (samples) is smaller than the number of voxels (features) per volume. The imbalance between features and samples lessens the statistical power of the methods used for classification, regression, and other types of analysis on the data.

Because neuroscience experiments require scanning several subjects, a natural means to compensate the limitation on the number of samples is to aggregate data from multiple subjects. Combining samples from multiple subjects requires to account for their brain differences. Anatomical alignment methods aim at solving the anatomical structure problem using specific anatomical features for alignment [1, 2, 3]. Yet, they fail to align functional topographies satisfactorily [2, 4].

More recently, methods that align functional topographies have emerged. These methods leverage the smoothness properties of fMRI data [5, 6] to extract shared aspects across subjects using factor models. Techniques based on Principal Component Analysis [7], Independent Component Analysis [8], Dictionary Learning [9], and Canonical Correlation Analysis [10] have been proposed. Another example is the Hyperalignment method [11] that assumes that subjects were presented with a time-synchronized stimulus and extracts shared features across subjects with orthogonal projections. The Shared Response Model (SRM) [12] follows a similar idea but boosts the predictive power of the subsequent analysis methods.

Despite their success in aggregating data from multiple subjects, functional alignment methods have disadvantages. Some of these methods require additional alignment data for optimal performance, hence reducing the available time for the actual experiment. Moreover, these methods are usually followed by an analysis stage consisting of supervised learning (e.g., classification, regression [11, 12]), which requires labeled data. Functional alignment methods are typically unsupervised and fail to use these training labels to learn the shared characteristics of the multi-subject data.

Consequently, we propose a semi-supervised scheme that enforces simultaneous training of a functional alignment model and a supervised model. We suggest a general approach to computing any instance of this scheme. Next, we derive a specific instance of the scheme with the Shared Response Model [12] for alignment and a Multinomial Logistic Regression (MLR) for a classification analysis. We demonstrate its superiority by achieving better classification performance or using less samples for alignment to obtain the same performance as unsupervised methods.

2. THE SHARED RESPONSE MODEL

Recently, Chen et al. [12] devised a multi-subject functional alignment method, the Shared Response Model (SRM) achieving state-of-the-art prediction results. SRM assumes that all the subjects receive the same stimuli during scanning,

and hence, there is an underlying low-dimensional shared representation across subjects for all samples. Let $\mathbf{X}_i \in \mathbb{R}^{v \times t}$ be t column-vectorized fMRI volumes for subject i with v voxels each. The model proposes that the data samples $\mathbf{X}_i$ for subject i are obtained by transforming the k-dimensional shared responses $\mathbf{S} \in \mathbb{R}^{k \times t}$ to the voxel space of the subject with a subject-specific mapping $\mathbf{W}_i$. Namely, the model is described as $\mathbf{X}_i = \mathbf{W}_i\mathbf{S} + \mathbf{E}_i$, where $\mathbf{E}_i \in \mathbb{R}^{v \times t}$ is the i^{th} subject representation error. Furthermore, this model assumes that the mapping $\mathbf{W}_i$ is in the Stiefel manifold $\mathcal{V}_{v,k}$ of orthonormal matrices, i.e., $\mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}$. The parameter k describes the dimension or the number of "features" of the shared response subspace, which typically is much smaller than the number of voxels v and samples t.

The authors propose in [12] to solve a constrained optimization problem to compute the shared response $\hat{\mathbf{S}}$ and the mappings $\hat{\mathbf{W}}_i$. The constraints are due to the fact that the mappings $\mathbf{W}_i$ are in the Stiefel manifold. The optimization task is given by

$$\hat{\mathbf{S}}, \left\{\hat{\mathbf{W}}_i\right\}_{i=1}^{N} = \arg\min_{\mathbf{W}_i,\mathbf{S}} \frac{1}{2}\sum_i \|\mathbf{X}_i - \mathbf{W}_i\mathbf{S}\|_F^2 \quad (1)$$
$$\text{s.t.} \quad \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I} \qquad i = 1 \dots N,$$

where N is the number of subjects. The Frobenius norm aims at reducing the representation error $\mathbf{E}_i$. Problem (1) is non-convex and finding the global optimum is hard. Therefore, the authors follow a block coordinate descent approach, fixing all but one matrix, which leads to a set of convex optimization problems. In particular, computing the shared responses $\mathbf{S}$ when fixing the mappings $\mathbf{W}_i$, has a closed-form solution of the form

$$\mathbf{S} = \frac{1}{N}\sum_i \mathbf{W}_i^T\mathbf{X}_i. \quad (2)$$

Additionally, fixing $\mathbf{S}$ and all but one mapping $\mathbf{W}_i$, N separate subproblems are required to be solved. These problems have a closed-form solution [13] of the form $\mathbf{W}_i = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices of the Singular Value Decomposition of the matrix $\mathbf{X}_i\mathbf{S}^T$. A local solution to Problem (1) is computed by iteratively updating $\mathbf{S}$ and all the mappings $\mathbf{W}_i$.

3. A SEMI-SUPERVISED APPROACH FOR FUNCTIONAL ALIGNMENT

To obtain a semi-supervised scheme, we add to the typical loss function $\mathcal{L}_{Align}$ that fits a multi-subject functional alignment model, a loss term $\mathcal{L}_{Sup}$ that simultaneously requires to learn a supervised task by fitting the model to extra labeled data samples. The proposed semi-supervised scheme involves solving the general optimization problem[1]

$$\min_{\psi,\theta} (1-\alpha)\,\mathcal{L}_{Align}(\psi) + \alpha\mathcal{L}_{Sup}(\theta;\psi) + R(\theta), \quad (3)$$

where ψ and θ are the semi-supervised model parameters, $R(\theta)$ is a regularization term for the supervised task, and $\alpha \in [0,1]$ is a scalar value that controls the bias between the functional alignment term and the supervised term.

Instead of finding a solution to the functional alignment problem first and then fitting a supervised model to the labeled data, Problem (3) computes the alignment parameters and fits the supervised task at once. The difference between these approaches is clearer when the objective function is differentiated to obtain the gradient. In the former case, there is no feedback from the supervised task to the functional alignment model, whereas in Problem (3) the gradient for updating the functional alignment term parameters depends also on the supervised term $\mathcal{L}_{Sup}(\theta;\psi)$. Parameter α controls the mixture between the functional alignment model and a pure supervised task solution. When $\alpha = 0$ only $\mathcal{L}_{Align}$ is minimized, and with $\alpha = 1$ only $\mathcal{L}_{Sup}$ is minimized.

3.1. Estimating the Semi-Supervised Model

The general semi-supervised optimization problem in Equation (3) is likely to inherit properties from any of the penalty terms. For example, using the objective function from SRM in (1) for $\mathcal{L}_{Align}$ would make the semi-supervised problem non-convex. Therefore, it is difficult to make assumptions for the general semi-supervised model. However, there are two typical methods that can compute a solution in many cases. One option is the gradient descent method. This method is very useful, but it usually needs a lot of iterations to converge. An alternative option is a Block-Coordinate Descent (BCD) approach [14]. In each step it solves a subproblem for a group of variables while fixing the values of the remaining ones. BCD methods have been shown to perform well empirically in various problems [15, 14, 16]. Furthermore, this method allows for more complex ideas like using a second order method for a block.

A BCD approach to (3) seems natural, because the semi-supervised scheme can be divided into two blocks: one for ψ and one for θ. In particular, solving for θ leads to exactly the same problem as minimizing the supervised task with $\mathcal{L}_{Sup}$ alone. For example, when $\mathcal{L}_{Sup}$ represents a classifier or a regression objective, a well-known algorithm for that problem can be leveraged to obtain a solver for such a semi-supervised method. An additional benefit is that the functional alignment parameters ψ are updated with the latest θ values instead of those from the previous iteration like in the gradient descent approach. Nevertheless, it is worth mentioning that many methods could be developed for more specific versions of $\mathcal{L}_{Align}$ and $\mathcal{L}_{Sup}$.

3.2. Semi-Supervised Shared Response Model

To derive a specific version of the semi-supervised scheme in (3), we use SRM as the functional alignment method and an MLR [17] classifier to exploit the labeled data. Although MLR allows for multi-class classification with a simple function to optimize, but other alternatives such as using hinge

[1]For simplicity, we limit the notation in Equation (3) to only the parameters of the model and the tasks being solved, leaving out the data dependencies.

loss for multi-class SVM [18, 19] can be considered as well. Recall the SRM objective term in equation (1):

$$\mathcal{L}_{SRM}\left(\{\mathbf{W}_i\}_i, \mathbf{S}; \{\mathbf{X}_i\}_i\right) = \frac{1}{2t}\sum_i \|\mathbf{X}_i - \mathbf{W}_i\mathbf{S}\|_F^2,$$

where the $\mathbf{W}_i$s follow the orthogonality constraint as in SRM, and t is a normalization factor representing the number of alignment data samples in each $\mathbf{X}_i$.

For the supervised term $\mathcal{L}_{Sup}$, we choose an MLR classifier. In our semi-supervised model there are various ways to feed the resulting functional alignment model to the classifier, and there are several options to apply the classifier as well. Each such decision would yield a different update to the model parameters and its usefulness would depend on the application. We consider the case where all labeled samples are used to train a single classifier that fits all subjects, having one set of parameters θ for all subjects. Training a classifier per subject does not exploit the functional alignment across brains, missing an opportunity to boost prediction.

Let $\mathbf{Z}_i \in \mathbb{R}^{v\times q}$ be q column-vectorized samples of subject i for the supervised task, and let $\mathbf{y}_i \in [1,\ldots,C]^q$ be the corresponding labels[2] for the samples in $\mathbf{Z}_i$. Training the MLR with the labeled data samples $\mathbf{Z}_i$ directly maintains the MLR term unpaired to the SRM term and does not exploit the alignment benefits. The resulting alignment SRM can be applied to the data $\mathbf{Z}_i$ such that the classifier is trained in the original voxel space of each subject, in the voxel space of a specific subject, or in the shared response subspace. The first case is obtained by applying the subject's specific mapping to project the samples into the shared response subspace and back from it, i.e., $\hat{\mathbf{Z}}_i = \mathbf{W}_i\mathbf{W}_i^T\mathbf{Z}_i$. This functions as a denoising step for the data before classification. This requires that all data samples for all subjects have the same number of voxels. However, the voxel space of each subject could be different enough to limit the classifier predictive power. The second case gets around the difference in subject voxel spaces by projecting the shared responses back from the shared response subspace to a subject-specific voxel space, i.e., $\hat{\mathbf{Z}}_i = \mathbf{W}_j\mathbf{W}_i^T\mathbf{Z}_i$, where j is the same for all i. In the third case, the labeled samples $\mathbf{Z}_i$ are projected once using the mappings $\mathbf{W}_i$, i.e., $\hat{\mathbf{U}}_i = \mathbf{W}_i^T\mathbf{Z}_i$. This functions as a dimensionality reduction method by reducing the number of features for the classifier (with $k \ll v$), and allows for a simpler algorithmic derivation for updating the subject mappings as they depend linearly on all the mappings. We continue from here applying MLR on the shared response subspace.

We define the MLR classifier with parameters $\mathbf{\Theta} \in \mathbb{R}^{k\times C}$ and bias terms $\mathbf{b} \in \mathbb{R}^C$, where C is the number of classes. The MLR penalty function $\mathcal{L}_i\left(\mathbf{\Theta}, \mathbf{b}; \mathbf{Z}_i, \mathbf{y}_i, \mathbf{W}_i\right)$ for subject i is defined as

$$\mathcal{L}_i = -\frac{1}{2q}\sum_j \log\left(\mathrm{softmax}_{(\mathbf{y}_i)_j}\left(\mathbf{\Theta}^T(\mathbf{W}_i^T\left(\mathbf{Z}_i\right)_j) + \mathbf{b}\right)\right),$$

[2]In the case of having a regression problem within $\mathcal{L}_{Sup}$, the labels would be described as real values.

where the notation $(\mathbf{y}_i)_j$ refers to the element j in $\mathbf{y}_i$, $(\mathbf{Z}_i)_j$ is the sample (column) j of matrix $\mathbf{Z}_i$, and the *softmax* function is defined as $\mathrm{softmax}_k(\mathbf{v}) = \exp(\mathbf{v}_k)/\sum_l \exp(\mathbf{v}_l)$. Eventually, the MLR penalty function including all subjects is given by their sums:

$$\mathcal{L}_{MLR}\left(\mathbf{\Theta}, \mathbf{b}; \{\mathbf{Z}_i, \mathbf{y}_i, \mathbf{W}_i\}_i\right) = \frac{1}{\gamma}\sum_i \mathcal{L}_i, \tag{4}$$

where $\gamma > 0$ is a scalar value that controls the influence of the regularization term $R(\mathbf{\Theta})$.

To estimate the semi-supervised model based on SRM with MLR, we derive the BCD approach for each parameter in the method: $\mathbf{S}$, each $\mathbf{W}_i$, and the pair $(\mathbf{\Theta}, \mathbf{b})$. The method initializes $\mathbf{W}_i$s with random orthonormal matrices and updates each variable at a time. Because $\mathcal{L}_{MLR}$ is independent of $\mathbf{S}$, the optimal shared response is obtained with Equation (2) like in SRM. The parameters $(\mathbf{\Theta}, \mathbf{b})$ are updated by minimizing the MLR objective function (4) with the selected regularization term $R(\mathbf{\Theta})$. In our case, we use an ℓ_2-regularization term and a Conjugate Gradients (CG) solver. On the other hand, updating a subject's mapping $\mathbf{W}_i$ requires computing the gradient of the objective function in (3) w.r.t. $\mathbf{W}_i$. We note that only one term of $\mathcal{L}_{SRM}$ and $\mathcal{L}_i$ depends on $\mathbf{W}_i$. Therefore, the gradient is given by

$$\nabla_{\mathbf{W}_i}\mathcal{L}_{SRM}\left(\mathbf{W}_i, \mathbf{S}\right) + \nabla_{\mathbf{W}_i}\mathcal{L}_{MLR}\left(\mathbf{\Theta}, \mathbf{b}; \mathbf{W}_i\right) = \tag{5}$$

$$\frac{\alpha}{t}\mathbf{W}_i\mathbf{S}\mathbf{S}^T - \frac{\alpha}{t}\mathbf{X}_i\mathbf{S}^T + \frac{1-\alpha}{\gamma q}\sum_j\left[\sum_{l=1}^C \left(\mathbf{Y}_i\right)_{j,l}\left(\mathbf{Z}_i\right)_j\left(\mathbf{\Theta}\right)_l^T - \frac{\sum_{l=1}^C \exp\left\{\left(\mathbf{b}\right)_l + \left(\mathbf{\Theta}\right)_l^T\mathbf{W}_i\left(\mathbf{Z}_i\right)_j\right\}\left(\mathbf{Z}_i\right)_j\left(\mathbf{\Theta}\right)_l^T}{\sum_{l=1}^C \exp\left\{\left(\mathbf{b}\right)_l + \left(\mathbf{\Theta}\right)_l^T\mathbf{W}_i\left(\mathbf{Z}_i\right)_j\right\}}\right],$$

where $\mathbf{Y}_i$ is an indicator matrix for vector $\mathbf{y}_i$ with $\left(\mathbf{Y}_i\right)_{j,l} = 1$ if $\left(\mathbf{y}_i\right)_j = l$, and 0 otherwise. As the mapping $\mathbf{W}_i$ is orthonormal and resides in Stiefel manifold $\mathcal{V}_{v,k}$, any optimization method for updating the mappings should be constrained to this manifold [20, 21]. Note how Equation (5) depicts the influence of the classifier when updating the mapping $\mathbf{W}_i$, in contrast to the mapping updates for SRM in Section 2.

4. EXPERIMENTS AND RESULTS

In this section we demonstrate our semi-supervised scheme instantiated with SRM and MLR (SS-SRM). We use two datasets for the experiments. The *raider* dataset [11] includes samples acquired while the subjects watched the movie "Raiders of the lost ark" for 110 minutes followed by 8 sets presenting still images from 7 different categories. The region of interest was limited to the ventral temporal cortex [22] and preprocessed as described in [12, 11]. The final dataset contains 1000 voxels (500 voxels per hemisphere), 2203 volumes for the movie stimulus and 56 volumes for the still image stimuli. Overall 10 subjects were scanned. In the *sherlock* dataset [23], 17 subjects were scanned while viewing a portion of an episode of the "Sherlock" BBC series,

Dataset	Experiment	MLR	SRM	SS-SRM
raider	Image category	56.25%	65.53%	68.57%
sherlock	Scene recall	4.28%	5.31%	6.12%

Table 1. Comparison of average accuracy for brain decoding experiments.

yielding 1976 volumes. Then, the subjects were requested to verbally recall the episode as best as they could. The recall section was divided in 50 scenes and all the volumes describing one scene were averaged and labeled. The number of scenes varies per subject between 24 and 44. The posterior medial cortex region of interest was used to extract 813 voxels per volume.

We implement the SS-SRM in Python with the pyManOpt package [24] for updating the mappings with the Conjugate Gradient in the Stiefel manifold. The SRM and SS-SRM code is available in the Brain Imaging Analysis Kit[3]. We fixed the number of iteration to 15 for both SRM and SS-SRM.

In the first experiment we evaluate the average classification accuracy of the methods by classifying the categories of the volumes with image stimuli from the *raider* dataset. We consider different numbers of data samples for the functional alignment part. We vary the length of the movie stimuli taking less volumes and use these for the functional alignment task. Next, we partition the volumes of the still image stimuli in 8 folds (each fold is a set) for all subjects and run an 8-fold cross-validation. We run three methods: a plain MLR classifier without functional alignment applied, SRM followed by an MLR classifier, and SS-SRM. All the methods used an ℓ_2-regularization with the γ parameter selected for best performance and to avoid overfitting. The regularizer parameter for the MLR classifier was $\gamma = 0.001$. SRM was trained with $k = 50$ features for the shared response and the regularizer value for MLR was $\gamma = 0.001$. The SS-SRM method used also $k = 50$ features and parameters $\alpha = 0.2$ and $\gamma = 1.0$. The average accuracy performance of the methods and their standard errors are presented in Figure 1. Also, the resulting predicting performance including the entire movie stimuli is also shown in Table 1. The improvement obtained with the functional alignment methods over a plain classifier is notable, while SS-SRM obtains better accuracy over SRM for any movie length. SS-SRM needs about half the samples for the functional alignment to achieve the same results as SRM with the entire movie samples. This experiment shows the potential of a semi-supervised method to achieve the same accuracy levels when using less fMRI volumes for alignment. Moreover, with more unlabeled data, it also has the potential to improve prediction results on labeled datasets. This is crucially important because it is much easier to collect large unlabeled datasets than large labeled datasets.

Next, we compare the prediction capabilities of the methods in the recall experiment using the *sherlock* dataset. Here,

[3]http://www.brainiak.org

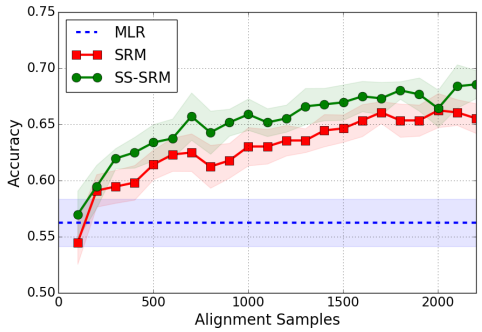

Fig. 1. Average accuracy as a function of the number of alignment samples.

we compare the performance when classifying unseen scenes from a single subject. We apply cross-validation by leaving out 7 scenes from a specific subject in every fold. The scenes are not necessarily contiguous in time, and as the number of scenes per subject vary, a few were not left-out. The same three methods as in the previous experiment were tested. The MLR classifier used $\gamma = 0.01$, SRM with MLR ran with $\gamma = 10$, whereas SS-SRM had $\gamma = 0.1$. Both, SRM and SS-SRM used $k = 25$ features for the shared response dimension. Table 1 shows the accuracy for these methods. The apparently low results are due to the high number of classes, although all of them are above the chance level of 2.12%[4]. Nevertheless, SS-SRM leverages the classification data better and achieves a higher accuracy than the other methods.

5. CONCLUSIONS

We presented a semi-supervised scheme that achieves simultaneous multi-subject fMRI functional alignment and supervised task learning with labeled data. Our scheme is applicable to combinations of functional alignment methods defined as optimization tasks and subsequent analyses defined as supervised learning tasks. The scheme generates an inherent feedback loop from the supervised task to the functional alignment method. We derived and implemented SS-SRM, which aligns functional topographies using the Shared Response Model and trains a Multinomial Logistic Regression classifier. Our proposed implementation achieves better prediction accuracy than its unsupervised SRM counterpart combined with an independently-trained MLR classifier. We found that the same accuracy can be achieved with less input brain volumes in our setup, leaving more time for additional neuroscientific experimentation. The semi-supervised scheme does not preserve spatial locality, which is of crucial importance for interpretation of results in neuroscience. Therefore, it would be beneficial to extend the semi-supervised scheme to spatial preserving methods, like a searchlight analysis.

[4]Three scenes were recalled only by less than 3 subjects and they were removed from the data, leaving 47 different scenes for the experiment.

6. REFERENCES

[1] J. Talairach and P. Tournoux, *Co-planar Stereotaxic Atlas of the Human Brain: 3-dimensional Proportional System : an Approach to Cerebral Imaging*, Thieme classics. G. Thieme, 1988.

[2] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. L. Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer, "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm)," *Phil. Trans. of the Royal Society of London B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[3] B. Fischl, M. I. Sereno, R. B.H. Tootell, and A. M. Dale, "High-resolution intersubject averaging and a coordinate system for the cortical surface," *Human Brain Mapping*, vol. 8, no. 4, pp. 272–284, 1999.

[4] M. Brett, I. Johnsrude, and A. Owen, "The problem of functional localization in the human brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 243–249, 2002.

[5] H. Op de Beeck, "Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fmri analyses?," *NeuroImage*, vol. 49, no. 3, pp. 1943–1948, 2010.

[6] S.A. Huettel, A.W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*, Freeman, 2009.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[8] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar, "A method for making group inferences from functional mri data using independent component analysis," *Human Brain Mapping*, vol. 14, no. 3, pp. 140–151, 2001.

[9] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *IPMI 22*, 2011, pp. 562–573.

[10] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional mri using canonical correlation analysis," *Magnetic Resonance in Medicine*, vol. 45, no. 2, pp. 323–330, 2001.

[11] J. Haxby, J. Swaroop Guntupalli, A. Connolly, Y. Halchenko, B. Conroy, M. Gobbini, M. Hanke, and P. Ramadge, "A common, high-dimensional model of the representational space in human ventral temporal cortex," *Neuron*, vol. 72, no. 2, pp. 404–416, 2011.

[12] P.-H. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, and P. Ramadge, "A reduced-dimension fmri shared response model," in *NIPS 28*, pp. 460–468. 2015.

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis:*, Cambridge University Press, Cambridge, 2 edition, 10 2012.

[14] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, June 2001.

[15] E. Treister and J. S. Turek, "A block-coordinate descent approach for large-scale sparse inverse covariance estimation," in *NIPS 27*, pp. 927–935. 2014.

[16] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, no. 3, pp. 346–367, 2007.

[17] J. Engel, "Polytomous logistic regression," *Statistica Neerlandica*, vol. 42, no. 4, pp. 233–252, 1988.

[18] Z. Jiang, "Support vector machines for multi-class pattern recognition based on improved voting strategy," in *2010 Chinese Control and Decision Conference*, May 2010, pp. 517–520.

[19] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *7th European Symp. on Artificial Neural Networks*, 1999, pp. 219–224.

[20] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[21] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.

[22] 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.

[23] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson, "Shared memories reveal shared structure in neural activity across individuals," *Nature Neuroscience*, vol. 20, no. 1, pp. 115–125, 2017.

[24] J. Townsend, N. Koep, and S. Weichwald, "Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation," *arXiv preprint arXiv:1603.03236*, 2016.