Classification of MRI under the Presence of Disease Heterogeneity using Multi-Task Learning: Application to Bipolar Disorder

Xiangyang Wang[1,2], Tianhao Zhang[1,⋆], Tiffany M. Chaim[3], Marcus V. Zanetti[3], and Christos Davatzikos[1]

[1] Center for Biomedical Image Computing and Analytics, and Department of Radiology, University of Pennsylvania, Philadelphia PA 19104, United States
[2] School of Communication and Information Engineering, Shanghai University, 200444, Shanghai, China
[3] Laboratory of Psychiatric Neuroimaging (LIM-21), Department and Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil
tianhao.zhang@uphs.upenn.edu

Abstract. Heterogeneity in psychiatric and neurological disorders has undermined our ability to understand the pathophysiology underlying their clinical manifestations. In an effort to better distinguish clinical subtypes, many disorders, such as Bipolar Disorder, have been further sub-categorized into subgroups, albeit with criteria that are not very clear, reproducible and objective. Imaging, along with pattern analysis and classification methods, offers promise for developing objective and quantitative ways for disease subtype categorization. Herein, we develop such a method using learning multiple tasks, assuming that each task corresponds to a disease subtype but that subtypes share some common imaging characteristics, along with having distinct features. In particular, we extend the original SVM method by incorporating the sparsity and the group sparsity techniques to allow simultaneous joint learning for all diagnostic tasks. Experiments on Multi-Task Bipolar Disorder classification demonstrate the advantages of our proposed methods compared to other state-of-art pattern analysis approaches.

1 Introduction

Most neurodegenerative and neuropsychiatric disorders are very heterogeneous, both from an imaging and from a clinical perspective, likely reflecting underlying complex genetic and environmental factors. Heterogeneity is further complicated by the fact that oftentimes different pathologies co-exist in the same individual, thereby confounding the structural and the clinical phenotypes. In the past decade, we have witnessed a great deal of progress in the use of advanced pattern analysis and machine learning methods for the classification of individuals, which is important for diagnostic and predictive purposes, and ultimately

⋆ Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 125–132, 2015.
DOI: 10.1007/978-3-319-24553-9_16

for individualized medicine. To date, however, most attempts for multivariate pattern analysis (MVPA) methods, such as Support Vector Machines (SVM), especially linear formulations, are primarily focused on the problem of finding a single direction separating two groups, and not on capturing multiple directions in heterogeneous populations.

For example, Bipolar Disorder (BD) mainly consists of BD type I and type II [1]. Multiple tasks of classifications including the whole patients (BD) vs normal controls (NC), each subtype of BD vs NC (i.e., BD I vs NC and BD II vs NC), and the distinguishing between different subtypes of BD (BD I vs BD II), are thereby more necessarily implemented rather than simple binary categorization of BD vs NC for computerized MRI diagnosis.

Multi-task learning [2] is a relatively recent development in the field of machine learning, and might be better suited for classification under phenotypic heterogeneity, as it simultaneously solves multiple classification tasks. Herein, we develop a novel multi-task SVM method, named Multi-Task $l_{2,1} + l_1$-norm SVM (mtSVM$L21L1$), which can work in the context of multiple classification tasks, by solving the multi-task hinge loss with sparsity [3] and group sparsity [4] regularization minimization problem. The learned weight coefficients W which defining the hyperplane are endowed with group sparsity property across multiple tasks while allow different patterns between tasks. This, therefore, can facilitate us to select a subset of features from the original input variables, which are meaningful for all the tasks. Our method is different from the multi-task feature learning methods [5][6][7][8][9][10][11] which are based on the least square (LS) loss technique. Actually, hinge loss based SVM (adopted in our method) has been validated [12][13] to have better performance than LS based methods for feature selection and classification. To the best of our knowledge, this is the first multi-task pattern classification method invented to identify the individual-level biomarkers for diagnosis of the heterogeneous neuropsychiatric data.

2 Multi-Task $l_{2,1} + l_1$-norm Support Vector Machine

2.1 Formulation

Assuming that we have t supervised learning tasks, let $X_i = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d \times n}$ as the training data matrix on i^{th} task, $i = 1, \cdots, t$, where d is the feature dimension, n is the number of input data samples, and let $Y_i = [y_1, y_2, \cdots, y_n] \in \mathbb{R}^n$ as the corresponding labels from these training samples for task i, where $y_j \in \{+1, -1\}$ is the binary label for each task. Let $W = [w_1, w_2, \cdots, w_t] \in \mathbb{R}^{d \times t}$ be the weight coefficient matrix for all t tasks, whose column $w_i \in \mathbb{R}^d$ parameterizes the linear discriminant function and whose row $w^k \in \mathbb{R}^t$ is the vector of coefficients associated with the k^{th} feature across different tasks. Then the hinge loss based multi-task model, i.e., Multi-Task $l_{2,1}+l_1$-norm SVM (mtSVM$L21L1$) can be defined by the following minimization problem:

$$\min_W \sum_{i=1}^{t} f(w_i^T X_i, Y_i) + \alpha \|W\|_{2,1} + \beta \|W\|_1 \tag{1}$$

where f is the hinge loss function as used in standard SVM [14] and defined as:

$$f(w_i^T X_i, Y_i) = \sum_{i=1}^{n} (1 - y_{ji}(w_i^T x_{ji} + b_i))_+ \tag{2}$$

where $(a)_+ = \max(0, a)$, b is the bias term. In the second term of (1), $\|W\|_{2,1} = \sum_{k=1}^{d} \|w^k\|_2$ is the structural sparsity, i.e., $l_{2,1}$-norm regularization [4], which encourages the weight coefficient matrix with many near-zero rows, while endows the coefficients that are significant to all the tasks to have larger weights. It will make sense if all classification tasks more or less share some common features. This may be true in our BD problem, because some brain regions might be abnormal in all subgroups (BD I, BD II) here. However, on the other hand, each task may have its specific features that are important for this task while unimportant for some others. So the l_1-norm regularization term $\|W\|_1$ is included in (1) in order to induce sparsity among tasks. This idea can be illustrated by Fig. 1: Fig. 1A is the standard sparsity pattern, and the models for different tasks are built independently; Fig. 1B is the pattern learned by the model with only $l_{2,1}$-norm, which enforces all models from different tasks to select a common set of features; Fig. 1C shows the learned pattern with $l_{2,1} + l_1$-norm, which makes sparsity weight coefficients that are similar, but not identical, across tasks.

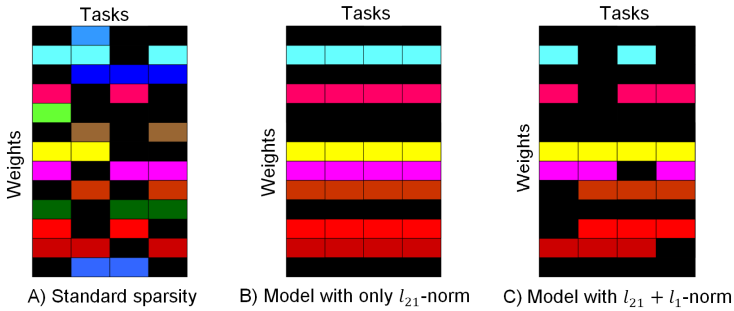

Fig. 1. Illustrations of sparsity effects. Different colors indicate different weight coefficients. A) Standard sparsity; B) Model with only $l_{2,1}$-norm; C)Model with $l_{2,1}+l_1$-norm.

2.2 Solution

We use the Optimal Stochastic Alternating Direction Method of Multipliers (SADMM) method [15] to solve our $l_{2,1}+l_1$-norm SVM problem. We first convert (1) to the following equivalent problem:

$$\min_W \sum_{i=1}^{t} \sum_{j=1}^{n} \max(0, 1 - y_{ji} w_i^T x_{ji}) + \alpha \|Z\|_{2,1} + \beta \|Z\|_1 \quad \text{s.t. } Z = W \tag{3}$$

This is a non-smooth but strongly convex problem. Let $f(w,\xi) = \max(0, 1 - yw^T x)$, where $\xi = \{x, y\}$ is a feature-label pair, and $h(Z) = \alpha\|Z\|_{2,1} + \beta\|Z\|_1$, the augmented Lagrangian will be:

$$L_\mu^k(W, Z, \lambda) = f(W_k) + < g_k, W > + \frac{1}{2\eta_k}\|W - W_k\|_2^2 + h(Z) \\ - < \lambda, Z - W > + \frac{\mu}{2}\|Z - W\|_2^2 \tag{4}$$

where $g_k = f^{'}(W_k, \xi_{k+1})$ is a stochastic sub-gradient of $f(W_k)$ at the current search point W_k of the k^{th} iteration, λ is the Lagrangian multipliers, $\mu > 0$ is a penalty parameter, $< A, B >= \text{trace}(A^T B)$, η_k is the step size and is set as $\eta_k = 2/\gamma(k+2)$ as well as in [15]. Applying SADMM to problem (4) produces closed-form updating rules as follows:

$$\begin{aligned} W_{k+1} &= \arg\min_W W^T f^{'}(W_k, \xi_{k+1}) + \frac{\mu}{2}\|Z_k - W - \frac{\lambda_k}{\mu}\|_2^2 + \frac{1}{2\eta_k}\|W - W_k\|_2^2 \\ Z_{k+1} &= \arg\min_Z \alpha\|Z\|_{2,1} + \beta\|Z\|_1 + \frac{\mu}{2}\|Z - W_{k+1} - \frac{\lambda_k}{\mu}\|_2^2 \\ \lambda_{k+1} &= \lambda_k - \mu\,(Z_{k+1} - W_{k+1}) \end{aligned} \tag{5}$$

Let $L_\mu^k(W) = W^T f^{'}(W_k, \xi_{k+1}) + \frac{\mu}{2}\|Z_k - W - \frac{\lambda_k}{\mu}\|_2^2 + \frac{1}{2\eta_k}\|W - W_k\|_2^2$, have $\partial L_\mu^k(W)/\partial W = 0$, and then we get the updating rule:

$$W_{k+1} = (\frac{1}{\eta_k} + \mu)^{-1}\left[\mu Z_k - \lambda_k + \frac{1}{\eta_k}W_k - f^{'}(W_k, \xi_{k+1})\right] \tag{6}$$

where $f^{'}(w, \xi) = -yx$, if $yw^T x < 1$; otherwise 0.

The implementation of the method can be summarized in **Algorithm 1**. Note that the Step 4 is solved by utilizing the decomposition property [7]. For further details see **Supplementary Material**[1], in which we also provided the proof on the convergence property of the algorithm.

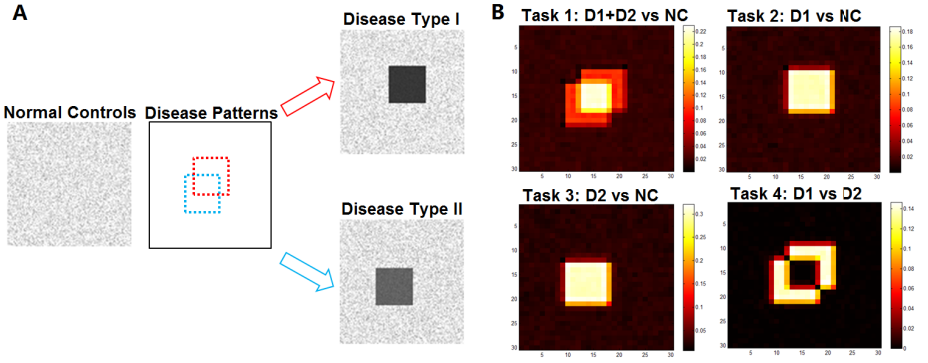

Fig. 2. Simulated data and results. A) Data generation; B) Learned weight coefficients.

[1] www.cbica.upenn.edu/sbia/Tianhao.Zhang/MICCAI2015.html

Algorithm 1. Multi-Task $l_{2,1} + l_1$-norm SVM (mtSVM$L21L1$)

Input: data matrix X, labels Y, and parameters $\alpha,\ \beta$
Initialize: $W_0 = Z_0 = \lambda_0 = 0$, $\mu = 10^{-6}$, $\mu_{max} = 10^{10}$, $\rho_0 = 1.1$, $\epsilon = 10^{-8}$, $\gamma = 2$, maxIter$= 10^3$, $k = 0$.
Output: W
while not converge, $k <$maxIter, **do**
1 $\eta_k = 2/\gamma(k+2)$
2 Obtain stochastic gradient g_k; build L_μ^k via (4)
3 Fix the others and update W by (6)
4 Fix the others and update Z by:
$$Z_{k+1} = \arg\min_Z \alpha\|Z\|_{2,1} + \beta\|Z\|_1 + \frac{\mu}{2}\|Z - W_{k+1} - \frac{\lambda_k}{\mu}\|_2^2$$
5 Update the multiplier λ by: $\lambda_{k+1} = \lambda_k - \mu\,(Z_{k+1} - W_{k+1})$
6 Update the parameter μ by: $\mu = \min(\rho_0\mu, \mu_{max})$
7 Check the convergence conditions: $\|Z_{k+1} - W_{k+1}\|_\infty < \epsilon$
8 $k = k + 1$
end while

3 Results

3.1 Multi-Task Feature Learning on the Simulated Data

The Data: We generated three groups of images: 1) disease type I (D1) data, 2) disease type II (D2) data, and 3) normal control data (NC). Each group had 30 samples, resulting in a total of 90 samples. All images are of size 100×100. The data are generated as follows. For each of the normal data, the mean is in [0.8, 0.95] with some Gaussian noise. In D1 and D2 images, there is an area of size 30×30, in which the values are decreased to [0.1, 0.6] with some Gaussian noise. The locations of such patches in D1 and D2 are not identical, but they have an overlapping area of size 20×20. The generation is illustrated in Fig. 2A.

The Results: The results obtained by mtSVM$L21L1$ are shown in Fig. 2B. Both disease patterns are identified in the comparison of D1+D2 vs NC (Task 1), with their overlapping area being more highlighted. In Task 2, it's shown that the abnormal patch in D1 is identified, while in Task 3, the abnormal patch in D2 is well marked. In Task 4, we can see that only the differences between D1 and D2 are highlighted. Taken together, the simulation results show our proposed method works effectively and correctly for multi-task feature learning.

3.2 Multi-Task Classification on the Bipolar Disorder Data

The Data: We evaluated the proposed methods using the structural brain MRIs on Bipolar Disorder (BD), a typical heterogeneous neuropsychiatric illness. From the total of 71 subjects, 44 were treatment-naive patients of BD and 27 were age and gender matched normal controls (NC). According to the DSM-IV criteria, each patient was assigned into BD I (22 subjects) or BD II (22 subjects) subgroups. Details on demographic characteristics, and image acquisition and preprocessing can be found in [16]. T1-weighted images were preprocessed according to a number of steps [16], including 1) AC-PC plane alignment; 2) Skull

removal; 3) Tissue segmentation into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF); and 4) High-dimensional image warping to a standard MNI space, resulting in the mass-preserved tissue density maps.

Experimental Design: Based on the voxel-wise tissue density values of GM, we performed the multiple classification tasks, including 1) Task 1: BD vs NC, 2) Task 2: BD I vs NC, 3) Task 3: BD II vs NC, and 4) Task 4: BD I vs BD II. According to the absolute values of weight coefficients W, we select the respective K top-ranked features [6][12] for each task, with which the linear SVM is used in the final step for the binary classification for each task. Other than the proposed mtSVM$L21L1$ method, some comparative methods are also carried out as below: 1) stLS$L1$: Single Task l_1-norm Least Square (LS) loss function feature selection; 2) stSVM$L1$: Single Task l_1-norm SVM feature selection; 3) mtLS$L21$: Multi-Task $l_{2,1}$-norm LS loss function feature selection [6][7]; 4) mtSVM$L21$: Multi-Task $l_{2,1}$-norm SVM, i.e., the case that only $l_{2,1}$-norm term is included in Equation (1); 5) mtLS$L21L1$: Multi-Task $l_{2,1} + l_1$-norm LS loss function [17] feature selection. mtLS$L21L1$ is built upon mtLS$L21$ by adding the l_1-norm, and solved by using the Accelerated Proximal Gradient (APG) [7] method. To compare all methods, we used 5-fold cross-validation: four random subsets for training and the remaining one subset for testing.

Parameters Tuning: The above methods can be classified into three groups according to regularization terms: 1)l_1-norm: stLS$L1$ and stSVM$L1$; 2)$l_{2,1}$-norm: mtLS$L21$ and mtSVM$L21$; 3)$l_{2,1} + l_1$-norm: mtLS$L21L1$ and mtSVM$L21L1$. They are related with two parameters, α or/and β which regulate the effects of the $l_{2,1}$ or/and l_1 terms respectively. We searched them in the range of $\alpha, \beta \in \left[10^{-5}, \cdots, 10^{-1}, 0.5, 1, 10^{1}, \cdots, 10^{5}\right]$. Another important parameter is K, i.e., the number of features which are selected from tasks. In our experiments, this number is in the area of [5, 5500].

Table 1. The ACCs (%) and the AUCs of the competing methods, calculated from four different tasks, respectively. The right columns list the average values.

Methods	Task 1		Task 2		Task 3		Task 4		ACC (avg)	AUC (avg)
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC		
stLS$L1$	59.23	0.49	51.78	0.54	48.79	0.44	46.80	0.42	51.65	0.47
stSVM$L1$	56.48	0.47	51.79	0.54	53.52	0.43	50.81	0.45	53.15	0.48
mtLS$L21$	67.52	0.61	66.75	0.62	58.52	0.56	68.52	0.61	65.33	0.60
mtSVM$L21$	70.38	0.66	74.78	0.73	72.43	0.73	71.42	0.61	72.25	0.68
mtLS$L21L1$	76.00	0.64	74.83	0.64	74.28	0.61	72.34	0.63	74.36	0.63
mtSVM$L21L1$	**78.95**	**0.67**	**84.23**	**0.76**	**84.18**	**0.77**	**78.35**	**0.71**	**81.42**	**0.72**

Classification Results: The optimal classification accuracy (ACC) and the area under curve (AUC) measures of all methods are listed in Table 1. As shown, Multi-Task methods performed better than Single Task ones. Among all the methods, mtSVM$L21L1$ has the best performances. The fact that mtSVM$L21L1$ outperformed mtSVM$L21$ reveals the benefit of characterizing specific patterns

related to different tasks. The heterogeneity in the BD group resulted in inferior performance of BD vs NC than BD I/II vs NC. In addition, we find that BD I vs BD II is the most difficult task, and likely requires a much larger training set.

Feature Interpretations: We overlaid the output weight coefficients obtained by mtSVM$L21L1$ onto the standard template for visual inspection. The representative sections are displayed in Fig. 3. We can see that BD I and BD II share similar patterns of GM abnormalities around the Frontal Pole (Fig. 3A) and the Precuneus (Fig. 3B) which are present in the results of Tasks 1, 2, and 3 (namely, BD vs NC, BD I vs NC, BD II vs NC), but not in Task 4 (BD I vs BD II). Relative to BD I vs NC (Task 2), BD II vs NC (Task 3) demonstrated more widely spread patterns including not only the Frontal Pole and the Precuneus but also more signals around the Cerebellum (Fig. 3C), and the Middle Frontal Gyrus (Fig. 3D) which were further confirmed by the direct comparison between BD I and BD II, that is, Task 4.

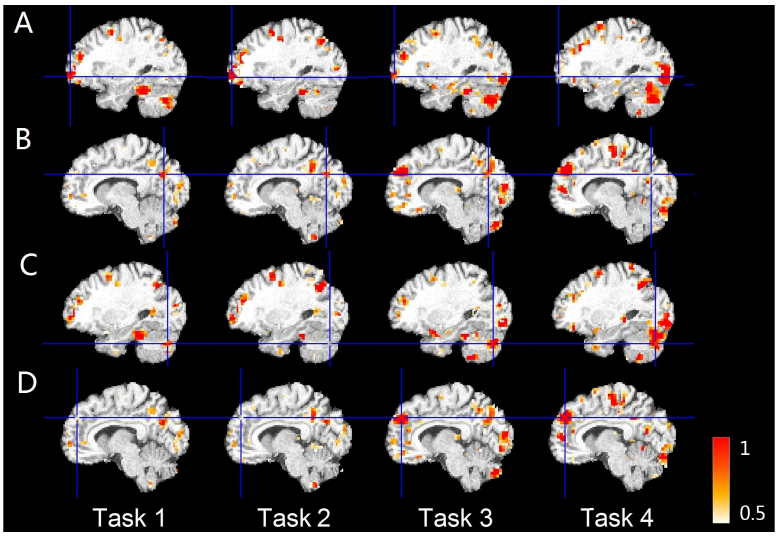

Fig. 3. Representative slices of regions, including A) Frontal Pole, B) Precuneus, C) Cerebellum, and D) Middle Frontal Gyrus, obtained from all four tasks. The scale indicates the absolute values of weights.

4 Conclusions

In this paper, we propose a novel method named Multi-Task $l_{2,1} + l_1$-norm Support Vector Machine (mtSVM$L21L1$) for classifying Bipolar Disorder (BD) disease under the presence of phenotypic heterogeneity. We adopt the framework of multi-task hinge loss with sparsity regularization terms to jointly learn features that are commonly shared among all the tasks and which are characterized with specific patterns in each task. Experimental results have shown that, compared with other state-of-the-art methods, our proposed method can achieve the best

performances for multi-tasks, also yielding better results than previous works on MRI-based classification in BD [16]. Furthermore, the features learned by the proposed method reveals the heterogeneous patterns of structural abnormalities from different tasks. Taken together, the proposed methods have deepened our insight into the neurobiological basis of the disorder′s clinical heterogeneity and helped us make progress on individual-level patient stratification.

Acknowledgement. This was supported in part by NIH R01AG14971, CNPq-Brazil & NARSAD (for clinical data), and FAPESP 13/03905-4 (to M.V.Z).

References

1. Dunner, D.L., Gershon, E.S., Goodwin, F.K.: Heritable factors in the severity of affective illness. Biological Psychiatry 11(1), 31–42 (1976)
2. Caruana, R.: Multitask learning. Machine Learning 28(1), 41–75 (1997)
3. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58(1), 267–288 (1996)
4. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B 68(1), 49–67 (2006)
5. Evgeniou, A., Pontil, M.: Multi-task feature learning. In: NIPS, pp. 41–48 (2007)
6. Wang, H., Nie, F., Huang, H., Risacher, S., et al.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: ICCV, pp. 557–562 (2011)
7. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via multi-task learning. NeuroImage 78, 233–248 (2013)
8. Rao, N., Cox, C., Nowak, R., Rogers, T.T.: Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. In: NIPS, pp. 2202–2210 (2013)
9. Jie, B., Zhang, D., Cheng, B., Shen, D.: Manifold regularized multitask feature learning for multimodality disease classification. Human Brain Mapping 36(2), 489–507 (2015)
10. Metsis, V., Makedon, F., Shen, D., Huang, H.: DNA copy number selection using robust structured sparsity-inducing norms. IEEE/ACM Transactions on Computational Biology and Bioinformatics 11(1), 168–181 (2014)
11. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $l2$, 1-norms minimization. In: NIPS, pp. 1813–1821 (2010)
12. Cai, X., Nie, F., Huang, H., Ding, C.: Multi-class l2, 1-norm support vector machine. In: ICDM, pp. 91–100 (2011)
13. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 28–55 (2011)
14. Scholkopf, B., Smola, A.J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press (2002)
15. Azadi, S., Sra, S.: Towards an optimal stochastic alternating direction method of multipliers. In: ICML, pp. 620–628 (2014)
16. Serpa, M.H., Ou, Y., Schaufelberger, M.S., Doshi, J., et al.: Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. BioMed Research International (2014)
17. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. Journal of Computational and Graphical Statistics 22(2), 231–245 (2013)