# Solving Logistic Regression with Group Cardinality Constraints for Time Series Analysis

Yong Zhang[1] and Kilian M. Pohl[1,2]

[1] Department of Psychiatry and Behavioral Sciences, Stanford University, USA
[2] Center for Health Sciences, SRI International, USA

**Abstract.** We propose an algorithm to distinguish 3D+t images of healthy from diseased subjects by solving logistic regression based on cardinality constrained, group sparsity. This method reduces the risk of overfitting by providing an elegant solution to identifying anatomical regions most impacted by disease. It also ensures that consistent identification across the time series by grouping each image feature across time and counting the number of non-zero groupings. While popular in medical imaging, group cardinality constrained problems are generally solved by relaxing counting with summing over the groupings. We instead solve the original problem by generalizing a penalty decomposition algorithm, which alternates between minimizing a logistic regression function with a regularizer based on the Frobenius norm and enforcing sparsity. Applied to 86 cine MRIs of healthy cases and subjects with Tetralogy of Fallot (TOF), our method correctly identifies regions impacted by TOF and obtains a statistically significant higher classification accuracy than logistic regression without and relaxed grouped sparsity constraint.

## 1 Introduction

Identifying image phenotypes by automatically classifying 3D+t MRIs is popular in medical image analysis, such as for cardiac motion [1,2] and brain development [3,4]. Automatic classification is difficult due to the complex information captured by MRIs coupled with small sample sizes of most MRI studies. Besides avoiding overfitting, the image phenotypes extracted by the classifiers need to be meaningful to clinicians. To address both issues, group sparsity constrained logistic regression models [5,6] first reduce the dense image information to a small number of features by grouping image features via the $l_2$-norm and counting the number of non-zero groupings ($l_0$ "norm"). The identified groupings then often can be directly linked to specific anatomy or function. Motivated by [5,7], these methods generally find a solution by relaxing the $l_0$ "norm" to the sum over the $l_2$-norm values. However, this solution relates to the original sparse problem only in specific conditions, e.g. compressed sensing [8], which generally do not hold for medical image applications. We now propose an algorithm for solving the logistic regression problem with the group cardinality constraint and show on a data set of cine MRIs that it produces statistically significant better results than the corresponding approach based on the relaxed norm .

© Springer International Publishing Switzerland 2015
N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 459–466, 2015.
DOI: 10.1007/978-3-319-24574-4_55

Specifically, we model consistency constraints of disease patterns across the series of images via group sparsity, i.e. we group the weights associated with each image feature across time and then apply the $l_0$ "norm" to those groupings. We find a solution within this model by generalizing Penalty Decomposition (PD) [9] from solving sparse logistic regression problem with group size one to more than one. In detail, we decouple the logistic regression function from enforcing the group sparsity constraint by introducing a penalty term based on the Forbenius norm. We minimize the smooth and convex logistic regression with the penalty term via gradient descent and derive a closed form solution for the modified group sparsity problem. We then find a local minimum of the original problem by iteratively solving the two subproblems via block coordinate descent.

We apply our method to cine MRI of 40 patients with reconstructive surgery of Tetralogy of Fallot (TOF) and 46 healthy volunteers, whose diagnosis we view as ground truth. As the residual effects of TOF mostly impact the shape of the RV [10], the regions impacted by TOF should not change across time, i.e. an ideal test bed for our logistic regression approach. We encode each cine MRI by first non-rigidly registering the image series to a 3D+t template, which divides the left and right ventricle into subregions. For each of those regions, we then record the average of the Jacobian determinant of the deformation map.

We measure the accuracy of our model via five-fold cross-validation. During training, we automatically set all important parameters of our approach by first creating a separate classifier for each setting of the parameter space. We then reduce the risk of overfitting by combining those classifiers into a single *ensemble of classifiers* [11], i.e. we compute a weighted average across the outputs of the individual classifiers. As we will show later, the ensemble of classifiers correctly favors subregions of the ventricles most likely impacted by TOF. Furthermore, the accuracy of the ensemble of classifiers is statistically significant higher than the outcomes obtained by relaxing and omitting group cardinality constraint, i.e. the classical logistic regression model that keeps all regions for disease detection.

## 2 Solving Sparse Group Logistic Regression

Our aim is to train a classifier to correctly assign subjects to cohorts based on their corresponding image data. To do so, we first describe the logistic regression model with group cardinality constraint. We then find a solution within that model by generalizing the PD approach.

**Model:** We assume that our training set consists of $N$ subjects that are represented by their corresponding diagnosis $\{b_1, \ldots, b_N\}$ and the encoding of 3D+t medical images $\{Z^1, \ldots, Z^N\}$ with $T$ time points. The value of $b_s \in \{-1, +1\}$ depends on the subject '$s$' being healthy $(+1)$ or diseased $(-1)$. The corresponding feature matrix $Z^s := [z_1^s \; z_2^s \ldots z_T^s] \in \mathbb{R}^{M \times T}$ of that subject is composed of vectors $z_t^s \in \mathbb{R}^M$ encoding the $t^{th}$ frame with $M$ features.

Next, we train the classifier on that data by solving the *logistic regression* problem confined by group sparsity. To define the problem, we introduce the variable $A^s := b_s \cdot Z^s$ for $s = 1, \ldots, N$, the weight matrix $W \in \mathbb{R}^{M \times T}$ defining

the importance of each feature in correctly classifying 3D+t images, the trace of a matrix $\mathrm{Tr}(\cdot)$, the logistic function $\theta(y) := \log(1 + \exp(-y))$, and the *average logistic loss* function with respect to the label weight $v \in \mathbb{R}$

$$l_{\mathrm{avg}}(v, W) := \frac{1}{N} \sum_{s=1}^{N} \theta\left(\mathrm{Tr}(W^\top A^s) + v \cdot b_s\right). \tag{1}$$

Ignoring the time point associated with each feature, we now assume that the disease is characterized by $r \in \mathbb{N}$ features, i.e. '$r$' rows of the feature matrix $Z^s$ subject to (s.t.) $r \leq M$. Then the logistic regression problem with group sparsity constraint is defined as

$$(\hat{v}, \widehat{W}) := \underset{v \in \mathbb{R}, W \in \mathbb{R}^{M \times T}}{\arg\min} \; l_{\mathrm{avg}}(v, W) \;\; \text{s.t.} \;\; \|\widetilde{W}\|_0 \leq r, \tag{2}$$

where $\widetilde{W} := (\|W^1\|_2, \ldots, \|W^M\|_2)^\top$ groups the weight vectors over time as $W^i := (W_1^i, \ldots, W_T^i)$ is the $i^{th}$ row of matrix $W$. Note, $\|\widetilde{W}\|_0$ equals the number of nonzero components of $\widetilde{W}$, i.e. the model accounts for the relationship of the same features across all the time points. Intuitively, if the model chooses a feature in one frame, the corresponding features in other frames should also be chosen since the importance of a feature should be similar across time. Replacing $\|\widetilde{W}\|_0$ with $\|W\|_0$ (or in the case of $T = 1$) results in the more common *sparse logistic regression* problem, which, in contrast, chooses individual features of $W$ ignoring any temporal dependency, which is not desired for our application.

**Algorithm:** We find a local minimum to Eq. (2) by decoupling the minimization of $l_{\mathrm{avg}}(\cdot, \cdot)$ from the sparsity constraint

$$\mathcal{X} := \{W \in \mathbb{R}^{M \times T} : \widetilde{W} := (\|W^1\|_2, \ldots, \|W^M\|_2)^\top \text{ and } \|\widetilde{W}\|_0 \leq r\}.$$

Specifically, we approximate the sparse weights $W \in \mathcal{X}$ via the unconstrained variable $Y \in \mathbb{R}^{M \times T}$ so that Eq. (2) changes to

$$(\hat{v}, \widehat{Y}, \widehat{W}) := \underset{v \in \mathbb{R}, Y \in \mathbb{R}^{M \times T}, W \in \mathcal{X}}{\arg\min} \; l_{\mathrm{avg}}(v, Y) \;\; \text{s.t. } W - Y = 0. \tag{3}$$

Denoting with $\varrho > 0$ a penalty parameter and the matrix Frobenius norm with $\|\cdot\|_F$, we solve Eq. (3) by generalizing PD proposed in [9] from enforcing the cardinality constraint for group size $T = 1$ to $T \geq 1$, i.e. we obtain a local solution $(\hat{v}, \widehat{Y}, \widehat{W})$ of the following nonconvex problem

$$\min_{v \in \mathbb{R}, Y \in \mathbb{R}^{M \times T}, W \in \mathcal{X}} f(v, Y, W) = l_{\mathrm{avg}}(v, Y) + \frac{\varrho}{2} \|W - Y\|_F^2. \tag{4}$$

Ascending $\varrho$ at each iteration, we use Block Coordinate Descent (BCD) to determine the solution by alternating between minimizing Eq. (4) with fixed $W$ and by fixing $v$ and $Y$. When $W$ is set to $W'$, finding the minimum with respect to $v$ and $Y$ turns into a smooth and convex problem

$$(v', Y') \leftarrow \underset{v \in \mathbb{R}, Y \in \mathbb{R}^{M \times T}}{\arg\min} \left\{ l_{\mathrm{avg}}(v, Y) + \frac{\varrho}{2} \|W' - Y\|_F^2 \right\}, \tag{5}$$

which can be solved via a gradient descent. In turn, minimizing the objective function just with respect to $W$, i.e.

$$W' \leftarrow \arg\min_{W \in \mathcal{X}} \|W - Y'\|_F^2, \tag{6}$$

can now be solved in closed form. We derive the closed form solution by assuming (without loss of generality) that the rows of $Y'$ are nonzero and listed in descending order according to their $l_2$-norm, i.e. let $(Y')^i$ be the $i^{th}$ row of $Y'$ for $i = 1, \ldots, M$ then $\|(Y')^1\|_2 \geq \|(Y')^2\|_2 \geq \ldots \geq \|(Y')^M\|_2 > 0$. It can then be easily shown (see [1]) that $W'$ is defined by the first '$r$' rows of $Y'$, i.e.

$$(W')^i = \begin{cases} (Y')^i, \text{ if } i \leq r; \\ 0, \quad \text{otherwise,} \end{cases} \quad \text{for } i = 1, \ldots, M. \tag{7}$$

In theory, the global solution derived above is not unique for Eq. (6), which we have not experienced in practice. One can also prove (similar to [9]) that the method converges with respect to $\varrho \to +\infty$ to a local minimum of Eq. (2).

## 3 Characterizing TOF Based on Cine MRIs

We apply our approach to cine MRIs of 40 TOF cases and 46 healthy controls. We choose this dataset due to the ground-truth diagnosis, i.e. each subject received reconstructive surgery for TOF during infancy or not. Furthermore, refining the quantitative analysis of these scans could lead to improved timing for follow-up surgeries in TOF patients. Finally, the assumption of our group sparsity model, i.e. the features extracted from each time point of the image series are sample descriptions of the same phenomena, fits well to this dataset. As we describe next, the features of this experiment are deformation-based shape encodings of the heart. As the residual effects of TOF reconstructive surgery mostly impact the shape of the right ventricle [10], our encoding should capture differences between the two groups in the same heart regions across time. We not only show that this is the case but our approach defined by Eq. (2) achieves significantly higher accuracy than logistic regression with relaxed sparsity constraints, i.e. given the sparse regularizing parameter $\lambda$ solving

$$\min_{v \in \mathbb{R}, W \in \mathbb{R}^{M \times T}} l_{\text{avg}}(v, W) + \lambda \sum_{i=1}^{M} \|W^i\|_2, \tag{8}$$

and without sparsity constraints, *i.e.*, $\lambda = 0$. Note we omit comparison to implementations replacing the $l_2$-norm in Eq. (8) with the $l_1$-norm since the resulting regularizor ignores consistency of the features across time. In other words, the regions identified by those approaches most likely have no relevance with respect to TOF, whose residual effects should be consistent across time.

**Extract Image Feature:** All 86 cine MR scans are defined by 30 time points. Each time point of a cine MRI is represented by a 3D image and a semi-automatically segmented label map of the right ventricular blood pool (RV)

[1] `http://www.stanford.edu/~yzhang83/MICCAI2015_supplement.pdf`

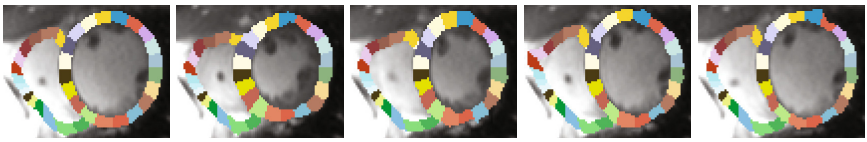



**Fig. 1.** Example of partitioning the template at different time points (T1 to T25) of the heart cycle. Each section of the RV is subtended by 9° and by 18° for the LV.

and the myocardium of the left ventricle (LV). Based on those segmentations, we omitted non cardiac structures from the scans and corrected each scan for image inhomogeneity and slice motion [12]. Next, we randomly choose the first time point of a healthy subject as a template and rigidly registered each case to the template. We encode the shape of the RV and LV by non-rigidly registering (via Demon's approach [13]) each time point of the remaining 85 scans to the template. We then compute the Jacobian determinant of the deformation maps [14]. We confine the determinant maps to the LV segmentations and for the RV reduce the maps to a 7mm band along its boundary, which is similar to the width of the LV myocardium. We also parcellate the RV and LV into smaller regions by first computing its corresponding mass center and then sectoring it into region of interests by the subtended angles as shown in Fig. (1). We finalize our shape descriptor by mapping the refined Jacobians into the template space and computing their average value with respect to individual sections. The size of the sections are defined with respect to degrees.

**Measure Accuracy:** We measure the accuracy of our approach via five-fold cross validation. We define the parameter space of the smoothness parameter $p = \{0, 0.5, \ldots, 5.0\}$ of the Demon's algorithm, the subtended angle $d = \{45°, 36°, 30°, 24°, 20°, 18°, 15°, 12°, 10°, 8°, 6°, 5°, 4°\}$ of each section of the LV and RV, and the sparsity constraint $r \in \{5, 10, \ldots, 50\}$, i.e. the number of average regional values chosen by our approach. Note, that we choose such a broad parameter space to allow for a fair comparison with other logistic regression models. For each parameter setting, we specify a regressor by computing the optimal weights $\widehat{W}$ with respect to training data. We do so by initializing the penalty parameter $\rho$ at 0.1 and increase $\rho$ by a factor of $\sqrt{10}$ at each PD iteration until convergence, *i.e.*, $\|\widehat{W} - \widehat{Y}\|_F^2 \leq 10^{-3} \cdot f(v, \widehat{Y}, \widehat{W})$. We also compute its training accuracy with respect to correctly assigning cine MRIs to each patient group. We account for the imbalance in cohort size by computing the normalized accuracy (nAcc), i.e. we separately compute the accuracy for each cohort and then average their values. On our dataset, parameter exploration on the training set resulted in multiple settings with 100% training accuracy. We therefore cross-validate a classifier based on training a single ensemble of classifiers [11]. We do so by treating the training nAcc score of a regressor as its weight in the decision of the ensemble of classifiers. In other words, the final label is then determined by the weighted average across the set of regressors, all of whom have different parameter settings. We also refer to this ensemble of classifier as $\boldsymbol{l_0}$**-Grp**. Fig. (2) shows the weight of each region according to the $l_0$-Grp across time. We notice

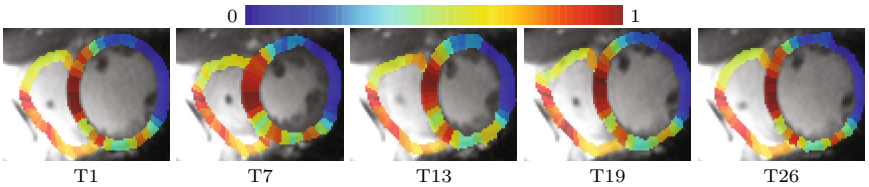



**Fig. 2.** Regions weighted by importance according to the proposed $l_0$-Grp approach. The sparsity constraints correctly identified the RV (left side) and specifically the right ventricular outflow tract (the regions where RV and LV meet) as important markers for identifying the residual TOF effects.

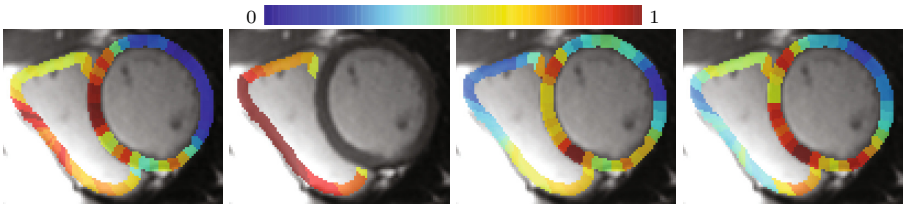
that the weights are fairly consistent across time due to the group sparsity term. In addition, the maps correctly point out (in red and yellow) the importance of the RV (left side) and more specifically the right ventricular outflow tract (the area where RV meets the LV) in identifying TOF.

**Alternative Models:** For comparison, we also generated an ensemble of logistic regression classifiers omitting the sparsity constraint (called **No-Grp**) and one by relaxing the group cardinality constraint (called **Rlx-Grp**) as defined by Eq. (8) [15]. These ensembles of classifiers used the same parameter space for the smoothness parameter $p$ and the subtended angle $d$ of each section. For Rlx-Grp, the sparsity parameter $\lambda$ of Eq. (8) was automatically set so that the number of chosen group features were similar to the group cardinality, i.e. $\{5, 10, \ldots, 50\}$. We applied each ensemble implementation to the dataset of just the LV, RV, and both ventricles.

**Classification Results:** The accuracy scores in % of all three ensembles are listed in Table 1 with respect to the encoding of the LV, RV, and both ventricles (LV&RV). Scores in bold represent results, which are significantly better (p-value $< 0.05$) than those of the other two methods. We compute the p-value based on the DeLongs Test [16] of the methods' ROC curves.

All three methods have similar accuracy scores for the LV (around 74.5%). As expected, the accuracy scores of all methods improve (by at least 9.9%) when the classification is based on the RV. The scores further improve for No-Grp and $l_0$-Grp by including both ventricles in the analysis. The increase in the score of $l_0$-Grp is explained by looking at the weight map of LV&RV in Fig. (3) (a), which, compared to the map of RV only in Fig. (3) (b), also highlights part of the LV. The impact of TOF on the LV was also recently echoed by [10].

Not only does the proposed $l_0$-Grp achieve the overall best score with 94% but the scores involving the RV are significantly better than those of the other two methods (p-values $\leq 0.036$). While these results further validate the proposed group cardinality constraint model for logistic regression, the relatively poor accuracy scores by Rlx-Grp is somewhat surprising (84.7%). Comparing the regional weight maps of the two approaches (see Fig. (3) (a) + (d)), the map of Rlx-Grp mostly ignores the RV failing to identify regions impacted by TOF. It not only explains the relatively low score but further highlights the superiority of the group cardinality over the $l_2$-norm relaxation with respect to this experiment.




(a) $l_0$-Grp RV&LV (b) $l_0$-Grp RV only (c) $l_0$-Grp $1^{st}$ TP only (d) Rlx-Grp RV&LV

**Fig. 3.** Regions of the first time point weighted by importance according to different implementations: $l_0$-Grp based on RV&LV (a), just based on the RV (b), based only on the first time point (c), and $l_0$-Grp based on RV&LV (d). Note, implementations putting more emphasis on the RV (a+b) are those with the higher accuracy scores.

**Table 1.** Accuracy scores (nAcc) in % of No-Grp, Rlx-Grp, and our proposed model ($l_0$-Grp). As expected, all methods achieve higher accuracy scores based on RV encoding vs. LV encoding. Including both ventricles in the analysis, leads to higher accuracy scores for No-Grp and the proposed $l_0$-Grp. The bold accuracy scores of $l_0$-Grp are significantly higher (p-value $< 0.05$) than the other two implementations, which was the case for the two experiments including the RV.

| | LV | RV | LV&RV |
|---|---|---|---|
| No-Grp | 74.5 | 85.1 | 86.2 |
| Rlx-Grp | 74.8 | 84.7 | 83.8 |
| $l_0$-Grp | 74.1 | **91.8** | **94.0** |

To study the importance of the group-sparsity over just sparsity, we confined $l_0$-Grp to the first time point, i.e. reducing the sparsity term to the cardinality constraint without group constraint. The accuracy of $l_0$-Grp dropped for all three experiments by a minimum of 4.5%. The weight map in Fig. (3) (c) reveals the reason for the drop in accuracy as it puts less emphasis on the RV. Furthermore, its scores are now similar to those of No-Grp (less than 2.5% difference). We therefore conclude that the $l_0$ group sparsity takes proper advantage of the repeat samples of the same disease patterns provided by the image time series.

## 4 Conclusion

We proposed an algorithm for solving logistic regression based on the group cardinality constraint. Unlike existing approaches, our algorithm did not relax the $l_0$-norm regularizer but instead used penalty decomposition to solve the original problem. Applied to 86 cine MRIs of healthy cases and subjects with TOF, our method not only correctly identified regions impacted by TOF but also obtains statistically significant higher classification accuracy than logistic regression without and with relaxed group sparsity constraints.

**Acknowledgement.** We would like to thank DongHye Ye for his help on generating the cardiac 515 dataset. This project was supported in part by the NIH grant R01HL127661. It was also supported by the Creative and Novel Ideas in HIV Research Program through a supplement to NIH P30 AI027763. This funding was made possible by collaborative efforts of the Office of AIDS Research, the National Institutes of Allergies and Infectious Diseases, and the International AIDS Society.

## References

1. Yu, Y., Zhang, S., Li, K., Metaxas, D., Axel, L.: Deformable models with sparsity constraints for cardiac motion analysis. Med. Image Anal. 18(6), 927–937 (2014)
2. van Assen, H.C., Danilouchkine, M.G., Frangi, A.F., Ordás, S., Westenberg, J.J.M., Reiber, J.H.C., Lelieveldt, B.: SPASM: Segmentation of sparse and arbitrarily oriented cardiac MRI data using a 3D-ASM. In: Frangi, A.F., Radeva, P., Santos, A., Hernandez, M. (eds.) FIMH 2005. LNCS, vol. 3504, pp. 33–43. Springer, Heidelberg (2005)
3. Serag, A., Gousias, I.S., Makropoulos, A., Aljabar, P., Hajnal, J.V., Boardman, J.P., Counsell, S.J., Rueckert, D.: Unsupervised learning of shape complexity: Application to brain development. In: Durrleman, S., Fletcher, T., Gerig, G., Niethammer, M. (eds.) STIA 2012. LNCS, vol. 7570, pp. 88–99. Springer, Heidelberg (2012)
4. Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R.: Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81, 358–370 (2013)
5. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. Roy. Soc. Ser. B 70(1), 53–71 (2008)
6. Wu, F., Yuan, Y., Zhuang, Y.: Heterogeneous feature selection by group lasso with logistic regression. In: ACM-MM, pp. 983–986 (2010)
7. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736 (2010)
8. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory 52(2), 489–509 (2006)
9. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM J. Optim. 23(4), 2448–2478 (2013)
10. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Syst. 16(6), 345–379 (2010)
11. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. 33(1-2), 1–39 (2010)
12. Ye, D., Desjardins, B., Hamm, J., Litt, H., Pohl, K.: Regional manifold learning for disease classification. IEEE T. Med. Imaging 33(6), 1236–1247 (2014)
13. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45(1), S61–S72 (2009)
14. Shen, D., Davatzikos, C.: Very high-resolution morphometry using mass-preserving deformations and hammer elastic registration. NeuroImage 18(1), 28–41 (2003)
15. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Arizona State University (2009)
16. DeLong, E., DeLong, D., Clarke-Pearson, D.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3), 837–845 (1988)