HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

Med Image Anal. 2017 January ; 35: 58–69. doi:10.1016/j.media.2016.05.011.

Computing Group Cardinality Constraint Solutions for Logistic Regression Problems

Yong Zhang[a], **Dongjin Kwon**[a,b], and **Kilian M. Pohl**[a,b]

[a]Department of Psychiatry & Behavioral Sciences, Stanford University, Palo Alto, CA 94304 USA

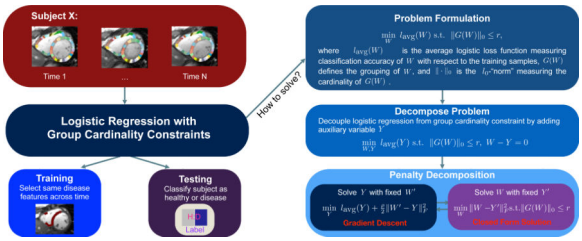
[b]Center for Health Sciences, SRI International, Menlo Park, CA, 94025 USA

Abstract

We derive an algorithm to directly solve logistic regression based on cardinality constraint, group sparsity and use it to classify intra-subject MRI sequences (e.g. cine MRIs) of healthy from diseased subjects. Group cardinality constraint models are often applied to medical images in order to avoid overfitting of the classifier to the training data. Solutions within these models are generally determined by relaxing the cardinality constraint to a weighted feature selection scheme. However, these solutions relate to the original sparse problem only under specific assumptions, which generally do not hold for medical image applications. In addition, inferring clinical meaning from features weighted by a classifier is an ongoing topic of discussion. Avoiding weighing features, we propose to directly solve the group cardinality constraint logistic regression problem by generalizing the Penalty Decomposition method. To do so, we assume that an intra-subject series of images represents repeated samples of the same disease patterns. We model this assumption by combining series of measurements created by a feature across time into a single group. Our algorithm then derives a solution within that model by decoupling the minimization of the logistic regression function from enforcing the group sparsity constraint. The minimum to the smooth and convex logistic regression problem is determined via gradient descent while we derive a closed form solution for finding a sparse approximation of that minimum. We apply our method to cine MRI of 38 healthy controls and 44 adult patients that received reconstructive surgery of Tetralogy of Fallot (TOF) during infancy. Our method correctly identifies regions impacted by TOF and generally obtains statistically significant higher classification accuracy than alternative solutions to this model, *i.e.*, ones relaxing group cardinality constraints.

Graphical abstract

[2]By convexity, see Proposition B.3 of Bertsekas (1999).



1. Introduction

An important topic in medical image analysis is to identify image phenotypes by automatically classifying time series of 3D Magnetic Resonance Images (MRIs). For example, intra-subject MRI sequences are used to analyze cardiac motion (Osman et al., 1999; Sermesant et al., 2003; Chandrashekara et al., 2004; Huang et al., 2005; Besbes et al., 2007; Sundar et al., 2009; Zhang et al., 2010a; Margeta et al., 2012; Wang et al., 2012; Yu et al., 2014), and brain development (Chetelat et al., 2005; Zhang et al., 2010b; Aljabar et al., 2011; Serag et al., 2012; Toews et al., 2012; Bernal-Rusiel et al., 2013; Schellen et al., 2015). However, the automatic classification of medical images is generally challenging. First, the number of features extracted from medical images is usually much larger than the number of samples. This generally results in overfitting of the method to the data, *i.e.*, much higher classification accuracy during training than on test data (Ryali et al., 2010; Marques et al., 2012; Deshpande et al., 2014). In addition, the image phenotypes identified by automatic classifiers are often difficult to relate to the medical literature (Qu et al., 2003). In this article, we propose an algorithm that addresses both issues by directly solving the so called logistic regression problem with group sparsity constraints.

Classifiers based on sparse models reduce the dense image data to a small number of features by counting the number of selected features via the l_0-"norm" and are configured so that the count is below a predefined threshold (Yamashita et al., 2008; Carroll et al., 2009; Rao et al., 2011; Liu et al., 2012; Lv et al., 2015). A generalization of that concept are group sparsity models, which first group image features based on predefined rules and then count the number of non-zero groupings (Ng et al., 2010; Wu et al., 2010; Ryali et al., 2010). To solve the underlying minimization problem, however, these methods relax the feature selection process from (group) cardinality constraints to weighting feature by, for example, replacing the l_0-"norm" with the l_2-norm (Meier et al., 2008; Friedman et al., 2010; Ryali et al., 2010; Li et al., 2012). The solution of those methods relates to the original sparse problem only under specific assumptions, e.g., the data entry matrix needs to satisfy the restricted isometry property in compressed sensing problem (Candes and Tao, 2005; Candès et al., 2006). However, matrices generally do not satisfy this property, such as those of the appendix of (Lu and Zhang, 2013), and most data matrices of medical image applications, e.g., matrices defined by the regional volume scores of subjects. Thus, with the exception of sparse models applied to compressed sensing, the solution obtained with respect to the relaxed norm generally does not recover the one of the original sparse model defined by the l_0-"norm". In addition, the number of measures selected by the classifier depends now on the training data due to the soft selection scheme. One can select a predefined number by

choosing measures whose weight is above a certain threshold. However, in the case of sparse logistic regression the corresponding classifier depends on the measures below the threshold and the relevance of those weights with respect to the disease under study is an ongoing topic of discussion (Haufe et al., 2014; Sabuncu, 2014). Alternatively, the upper bound associated with the sparse constraint is set so that the classifier returns the wanted number of measures for a given training data set (Vounou et al., 2012; Zhang and Shen, 2012; Ma and Huang, 2008; Zhang et al., 2012). The tuning is now data dependent, *i.e.*, each training set is generally associated with a different upper bound so that selected number of scores is constant across training sets. Even comparing the patterns of different subsets of the same data set, *i.e.*, folds, is none trivial as each pattern is the solution to a minimization problem, whose sparsity constraint is unique to a fold. Avoiding soft feature selection and thus these issues, our algorithm solves the original group sparsity constrained, logistic classification problem defined by the l_0-"norm" by extending the Penalty Decomposition (**PD**) method (Lu and Zhang, 2013). By doing so, our method uses a single model to not only classify samples but also directly select patterns (without thresholding or changing upper bounds) that potentially are image phenotypes meaningful to medical community.

To further investigate its potential, we now generalize PD from solving sparse logistic regression problems with group size one to more than one. Specifically, we assume that an intra-subject series of images represents repeated samples of the same disease patterns. In other words, selecting an image feature for disease identification needs to account for the entire series of measurements created by that feature across time. We model this assumption by combining each "feature series" into a single group. The proposed PD algorithm then derives a solution within that model by decoupling the minimization of the logistic regression function from enforcing the group sparsity constraint. Applying Block Coordinate Descent (**BCD**), the minimum to the smooth and convex logistic regression problem is determined via gradient descent while we derive a closed form solution for finding a sparse approximation of that minimum.

We apply our method to cine MRI of 38 healthy adults and 44 adult patients, that received reconstructive surgery of Tetralogy of Fallot (**TOF**) during infancy. The data sets fulfill the assumption of the group sparsity model as the residual effects of TOF mostly impact the shape of the right ventricle (Atrey et al., 2010; Bailliard and Anderson, 2009) so that the regions impacted by TOF should not change across the time series captured by a cine MRI. During training, we automatically set all important parameters of our approach by first training a separate regressor for each setting of the parameter space. We then reduce the risk of over-fitting by combining those classifiers into a single *ensemble of classifiers* (Rokach, 2010). This ensemble of classifiers correctly favors subregions of the ventricles most likely impacted by TOF. For most experiments, it also produces statistically significant higher accuracy scores than ensemble of classifiers that relax the group cardinality constraint.

We first proposed to generalize PD to group sparsity constraints at MICCAI 2015 (Zhang and Pohl, 2015). This article provides a more in-depth view of this idea. Specifically, we expand PD to guarantee convergence of the sparse approximation to a local minimum of the group-sparsity confined, logistic regression problem, which is the primary contribution of this manuscript. We also modify the experiments by replacing the morphometric encodings

of heart regions based on the average of the Jacobian determinants with simple volumetric scores. This simplifies preprocessing as alignment of each cine MRI to a template is unnecessary. It also reduces the size of the parameter search space, which now omits the smoothing parameters associated with the alignment process. Moreover, we not only record a single accuracy score for each implementation but instead generate distributions of scores by modifying the number of training samples. For each training size, we apply the method to 10 different training and testing sets. Finally, we distinguish the ventricular septum from the left ventricle to refine our findings from the previous publication (Zhang and Pohl, 2015) and support those findings with new plots that visualize the selection of regions across the entire heart.

Beyond our MICCAI publication, a possible alternative regression approach for simultaneous classification and pattern extraction is the random forest method (Lempitsky et al., 2009). However, it is unclear how to expand this technology to group-wise selection schemes that enforce temporal consistency in selecting regions, *i.e.*, the same regions are picked across all time points. Due to these difficulties most machine learning approaches applied to cine MRI just focus on disease classification, such as (McLeod et al., 2013; Afshin et al., 2014; Bai et al., 2015). They often improve results by manually selecting regions thought to be impacted by the disease before performing classification (Wald et al., 2009). An exception are (Qian et al., 2011; Ye et al., 2014; Bhatia et al., 2014), which seperately perform disease classification and weigh individual regions possibly impacted by disease. The disconnect between the two steps and the weighing of individual regions makes clinical interpretation of the findings more difficult as, in addition to the earlier mentioned issues associated with the interpretation of weights, it increases the risk of false positive findings compared to directly identifying patterns of regions. Our experimental results echo these issues, where logistic regression with relaxed sparsity constraints was generally significantly less accurate than our proposed solution to the original sparsity constraint. We conclude that our proposed approach is the first to solve a single optimization problem for simultaneous disease classification and group-based pattern identification based on segmentation of cine MRIs.

The rest of this paper is organized as follows. Section 2 provides an in-depth description of PD algorithm and its convergence properties. Section 3 summarizes the experiments on the TOF dataset and Section 4 concludes the paper with final remarks.

2. Solving Sparse Group Logistic Regression

We first describe the logistic regression model with group cardinality constraint, which accurately assigns subjects to cohorts based on features extracted from intra-subject image sequences. We then generalize the PD approach to find a solution within that model. We end the section deriving convergence properties of the resulting algorithm.

2.1. The Model

The input to our model are N subjects, their diagnosis $\{b_1,...,b_N\}$ and features $\{Z^1,..., Z^N\}$ extracted from 3D+t medical images with T time points. The diagnosis $b_s \in \{-1, +1\}$ is +1

if subject 's' is healthy and −1 otherwise. The feature matrix $Z^s := \left[z_1^s \quad z_2^s \dots z_T^s \right] \in {}^{M \times T}$ of subject s is composed of vectors $z_t^s \in {}^M$ encoding the t^{th} time point through M features. Our goal is now to accurately model the relationship between the labels $\{b_1,...,b_N\}$ and the features $\{Z^1,...,Z^N\}$.

Given the large number of features and the relatively small number of samples, we make the model tractable by assuming that the disease is best characterized by the same 'r' features (subject to (s.t.) $r \leq M$) at each time point, which means the same 'r' rows of each feature matrix Z^s. One way of modeling this relationship is via *group-sparsity constraint* solutions to a *logistic regression* problem. Weighing features according to their importance in separating the healthy from the disease group, the problem of logistic regression is to find the configuration that most accurately infers the diagnosis of each sample. Sparsity constraints simply confine the search space to those configurations that only select a subset of features, *i.e*, the weight of non-selected features is zero. Our model aims to identify 'r' rows of a feature matrix, which group-sparsity constraint does by first defining the features of a row as a group before enforcing the sparsity constraint on those groupings. In other words, if the model chooses a feature in one time point, the corresponding features in other time points should also be chosen since the importance of a feature should be similar across time.

To formally define this model, we now introduce

- the diagnosis-weighted feature matrix $A^s := b_s \cdot Z^s$ for $s = 1,...,N$,
- the weight matrix $W \in {}^{M \times T}$ defining the importance of each feature in correctly classifying subjects,
- $W^i := \left(W_1^i, \dots, W_T^i \right)$ being the i^{th} row of matrix W,
- 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 $v \in$

$$l_{avg}(v, W) = \frac{1}{N} \sum_{s=1}^{N} \theta \left(Tr \left(W^{\mathsf{T}} A^s \right) + v \cdot b_s \right). \quad (1)$$

The logistic regression problem with group sparsity constraint is then defined as

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

where $\widetilde{W} := \left(\|W^1\|_2, \dots, \|W^M\|_2 \right)^{\mathsf{T}}$ groups the weight vectors over time by computing the l_2-norm of the rows. Thus, $\|\widetilde{W}\|_0$ equals the number of nonzero components of $\widetilde{W}$, *i.e.*, the non-zero rows of W. The sparsity constraint search space is then formally defined as

$$\mathcal{X} := \left\{ W \in {}^{M\times T} : \|\widetilde{W}\|_0 \leq r \right\},$$

so that Eq. (2) shortens to

$$(v^*, W^*) := \underset{v\in\mathbb{R}, W\in\mathcal{X}}{arg\,min}\, l_{avg}(v, W). \quad (3)$$

Note, that in the case of $T = 1$ or replacing $\|\widetilde{W}\|_0$ with $\|W\|_0$ then Eq. (3) changes to the more common *sparse logistic regression* problem, which, in contrast, chooses individual features of W ignoring any temporal dependency. While the accuracy of this model might be similar to the proposed group-sparsity model, the selected features have limited meaning for diseases such as the residual effects of TOF. TOF impacts the morphometry and thus leads to changes in local shape patterns that are consistent across the cardiac cycle.

2.2. Approximating the Solution to the Group Sparsity Constraint, Minimization Problem

PD of (Lu and Zhang, 2013) estimates the sparse solution ($T = 1$) to logistic regression problems by decoupling the minimization of the logistic regression $l_{\text{avg}}(\cdot,\cdot)$ from finding a solution within sparsity constraint space $\mathcal{X}$. It does so by defining a penalty function consisting of two components: (1) the logistic regression $l_{avg}(\cdot,\cdot)$ dependent on v and an auxiliary variable $Y \in {}^{M\times T}$, and (2) a similarity measure $S(Y, W)$ between the non-sparse solution Y and the approximated sparse solution W. In other words, the penalty function is defined as

$$q_\varrho(v, Y, W) := l_{avg}(v, Y) + \varrho S(W, Y),$$

where the penalty parameter $\varrho < 0$ weighs the importance of $S(\cdot,\cdot)$. At each iteration, PD increase ϱ and then determines the $\left(\hat{Y}, \hat{W}\right)$ minimizing $q_\varrho(\cdot,\cdot,\cdot)$. Thus, as ϱ increases, the difference reduces between $\hat{Y}$, the solution of the regularized logistic problem, and its sparse approximation $\hat{W}$. Once the algorithm converges, $\hat{W}$ is the approximated solution of the original sparse problem. In the remainder of this subsection, we generalized PD to approximate the group sparsity constraint solution ($T \geq 1$) of the logistic regression problem defined by Eq. (3).

To adapt PD to our model, we first assume that our algorithm is initialized qith ϱ_i and (v_i, W_i), where $W_i \in \mathcal{X}$. Each iteration 'p' of our algorithm is then composed of three steps: define penalty function, minimize penalty function, and update and check convergences of results.

Step 1 - Define penalty function—Finding (v^*, W^*) of Eq. (3) is equivalent to solving

$$(v^*, Y^*, W^*) := \underset{v\in\mathbb{R}, Y\in\mathbb{R}^{M\times T}, W\in\mathscr{X}}{arg\,min} l_{avg}(v, Y) \quad \text{s.t.} \quad W - Y = 0. \tag{4}$$

Introducing the matrix Frobenius norm $\|\cdot\|_F$, the above equation is equivalent to

$$(v^*, Y^*, W^*) := \underset{v\in\mathbb{R}, Y\in\mathbb{R}^{M\times T}, W\in\mathscr{X}}{arg\,min} l_{avg}(v, Y) \quad \text{s.t.} \quad \|W - Y\|_F^2 = 0 \tag{5}$$

so that $S(W, Y) := \frac{1}{2}\|W - Y\|_F^2$ is a natural metric to measure the similarity between W and Y for our PD algorithm. For the current penalty parameter ϱ_p, we define the PD characteristic penalty function as

$$q_{\varrho_p}(v, Y, W) := l_{avg}(v, Y) + \frac{\varrho_p}{2}\|W - Y\|_F^2. \tag{6}$$

Thus, the solution to the following non-convex and non-continuous minimization problem approximates the original sparse solution of Eq. (3):

$$\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right) := \underset{v\in\mathbb{R}, Y\in\mathbb{R}^{M\times T}, W\in\mathscr{X}}{arg\,min} q_{\varrho_p}(v, Y, W). \tag{7}$$

Step 2 - Determine local minimum point of Eq. (7) via BCD—To find a local minimum point of Eq. (7), BCD alternates between minimizing the penalty function (Eq. (6)) with respect to the non-sparse terms (v, Y) and updating the sparse term W. Specifically, let (v_{b-1}, Y_{b-1}, W_{b-1}) be the current estimate of BCD. The b^{th} iteration of BCD then determines (v_b, Y_b) by solving the smooth and convex problem

$$(v_b, Y_b) \leftarrow \underset{v\in\mathbb{R}, Y\in\mathbb{R}^{M\times T}}{arg\,min} \left\{ l_{avg}(v, Y) + \frac{\varrho_p}{2}\|W_{b-1} - Y\|_F^2 \right\}, \tag{8}$$

which can be done via a gradient descent. To update W_b, minimizing the penality function with respect to W, *i.e.*,

$$W_b \leftarrow \underset{W\in\mathscr{X}}{arg\,min} \|W - Y_b\|_F^2, \tag{9}$$

can now be solved in closed form. We derive the closed form solution by assuming (without loss of generality) that the rows of Y_b are nonzero and listed in descending order according to their l_2-norm, *i.e.*, let Y_b^j be the j^{th} row of Y_b for $j = 1,..., M$ then $\|Y_b^1\|_2 \geq \|Y_b^2\|_2 \geq \ldots \geq \|Y_b^M\|_2 > 0$. Lemma A.1 (see Appendix for this and any proceeding

lemmas and theorems) then shows that the closed form solution W_b is defined by the first 'r' rows of Y_b,

$$W_b^j = \begin{cases} Y_b^j, & \text{if} \quad j \leq r; \\ 0, & \text{otherwise}, \end{cases} \quad \text{for} \quad i=1,\ldots,M. \tag{10}$$

In theory, multiple global solutions W_b exist in case $\|Y_b^r\|_2 = \|Y_b^{r+1}\|_2$. In practice, we have not come across that scenario.

Denoting with $\|\cdot\|_{\max}$ the max norm of a matrix, *i.e.*,, $\|A\|_{max} := max_{i,j}\{\|a_{ij}\|\}$, BCD stops updating the results when the relative change of each variable is smaller than a benchmark value ϵ_{BCD}, *i.e.*,

$$max\left\{\frac{|v_b - v_b|}{max\left(|v_b|,1\right)}, \frac{\|Y_b - Y_b\|_{max}}{max\left(\|Y_b\|_{max},1\right)}, \frac{W_b - W_b\|_{max}}{max\left(\|W_b\|_{max},1\right)}\right\} \leq \epsilon_{BCD}. \tag{11}$$

We choose this criteria over the absolute change of the sequence (v_b, Y_b, W_b), *i.e.*,

$$max\left\{|v_{b-1} - v_b|, \|Y_{b-1} - Y_b\|_{max}, \|W_{b-1} - W_b\|_{max}\right\} \leq \epsilon_{BCD},$$

as it is more robust when variables have large values, *i.e.*,

$$max\left\{|v_b|, \|Y_b\|_F, \|W_b\|_F\right\}$$

is large.

Step 3 - Update results, penalty parameter, and check the stopping criterion— Let BCD stop at the b'^{th} iteration, $\hat{Y}_p$ is then set to Y_b' and $\hat{v}_p$ to v_b'. To update $\hat{W}_p$, we first define an upper bound $\Gamma = l_{\text{avg}}(v_i, W_i)$ with respect to the initialization, and then check whether

$$\min_{v\in\mathbb{R}, Y\in\mathbb{R}^{M\times T}} q_{\varrho_{p+1}}\left(v, Y, W_{b'}\right) \leq \Gamma. \tag{12}$$

In case the condition Eq. (12) holds, $\hat{W}_p$ is set to W_b' and otherwise to W_i. According to Lemma A.5, this check guarantees that in case PD converges $\hat{W}_p$ also converges to $\hat{Y}_p$, *i.e.*, $\lim_{p\to\infty} \hat{W}_p - \hat{Y}_p = 0$. Finally, PD stops updating the results when $\hat{Y}_p$ and $\hat{W}_p$ are similar enough according to the similarity parameter ϵ_{PD}, *i.e.*,

$$\|\hat{W}_p - \hat{Y}_p\|_{max} \leq \epsilon_{PD}. \quad (13)$$

Note, Algorithm 1 is the pseudocode of our PD approach (Step 1-3), whose corresponding software implementation used for this publication can be dowloaded via https://dx.doi.org/10.6084/m9.figshare.3398332 or the current version via https://github.com/sibis-platform/PDLG.

2.3. Convergence Properties of Penalty Decomposition

For the interested reader, we now briefly derive the properties guaranteeing that the converged solution of our PD approach is a local minimum point of Eq. (3). Focusing just on one iteration of PD, we first show that if the BCD approach of Step 2 converges then the corresponding accumulation point is also a local minimum point of Eq. (7). Across iterations of PD, these local minima define another sequence, which is defined by Step 3. We then show if the sequence converges to an accumulation point with exact r nonzero rows then this point is a local minimum point of the original sparse problem defined by Eq. (3). Note, deriving these properties of our algorithm is non-trivial as the PD penalty function Eq. (6) is non-convex and the sparse space $\mathscr{X}$ is non-continuous. The Appendix contains the complete proofs of the properties of our method described below.

Convergence property of Step 2—In the p^{th} iteration of PD, let $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ be an accumulation point of the converged sequence (v_1, Y_1, W_1), (v_2, Y_2, W_2),... produced by BCD. To show that $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is a local minimum point of Eq. (7), the triple needs to be the minimum point of $q_{\varrho_p}(\cdot,\cdot,\cdot)$ with respect to a neighborhood of this triple. We confine the neighborhood to those triples (v, Y, W), where the sign of non-zero components of $\hat{W}_p$ equals those of W. We formally express this constraint by introducing the set of indices corresponding to non-zero components of $\hat{W}_p$

$$\Omega_{\hat{W}_p} := \left\{ i \in \{1, \ldots, M\} : \|\left(\hat{W}_p\right)^i\|_2 \neq 0 \right\},$$

so that the neighborhood is defined as

$$\mathscr{N} := \left\{ \left(v, Y, \hat{W}_p + H\right) \in \left(, ^{M\times T}, \mathscr{X}\right) : \|(H)^i\|_2 < \|\left(\hat{W}_p\right)^i\|_2, \quad \forall i \in \Omega_{\hat{W}} \right\}. \quad (14)$$

An interesting characteristic of that neighborhood is that for any triple $\left(v, Y, \hat{W}_p + H\right) \in \mathscr{N}$ the following relation holds among H, $\hat{Y}_p$ and $\hat{W}_p$ (see also Lemma A.3)

$$H\left(\hat{Y}_p - \hat{W}_p\right)^T = \mathbf{0}, \quad (15)$$

where **0** is a matrix whose entries are all zero. We make use of this property in Theorem A.4, where we derive the following lower bound for $q_{\varrho_p}(\cdot,\cdot,\cdot)$ with respect to $\left(v, Y, \hat{W}_p + H\right) \in \mathcal{N}$

$$q_\varrho\left(v, Y, \hat{W}_p + H\right) \geq q_\varrho\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right) + Tr\left(H\left(\hat{Y}_p - \hat{W}_p\right)^T\right)$$

so that applying Eq. (15) results in

$$= q_\varrho\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right).$$

In other words, $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is a local minimum point of Eq. (7).

Convergence property of Step 3—Let (v^*, Y^*, W^*) be the accumulation point of the converging sequence $\left(\hat{v}_1, \hat{Y}_1, \hat{W}_1\right)$, $\left(\hat{v}_2, \hat{Y}_2, \hat{W}_2\right)$,... produced by Step 3. Furthermore, assume that W^* has exact 'r' nonzero rows, *i.e.*, $\|\widetilde{W^*}\|_0 = r$, which has always been the case in our experiments. Lemma A.7 then states that (v^*, W^*) is a local minimum point of Eq. (3) if there exists a matrix $\Lambda_{\mathcal{K}} \in {}^{M\times T}$ so that the following holds:

$$\begin{aligned} \frac{\partial l_{avg}(v,W)}{\partial W}\Big|_{v=v^*, W=W^*} + \Lambda_{\mathcal{K}} &= 0, \\ \frac{\partial l_{avg}(v,W)}{\partial v}\Big|_{v=v^*, W=W^*} &= 0, \\ (\Lambda_{\mathcal{K}})^{\mathcal{K}} &= 0, \end{aligned} \quad (16)$$

where $\mathcal{K}$ is a set of 'r' indices for which $(W^*)^i \neq 0$ and $f(x)\,|x=x'$ is the value of $f(\cdot)$ at x'. To determine, $\Lambda_{\mathcal{K}}$, we note that $Y^* = W^*$ (see also Lemma A.5). Theorem A.8 then states that the sequence

$$Z_p := \varrho_p\left(\hat{Y}_p - \hat{W}_p\right) \quad (17)$$

is a bounded and converges to Z^*. The theorem furthermore notes that (v^*, W^*) fulfills the condition of Eq. (16) when $\Lambda_{\mathcal{K}} = Z^*$. In summary, if PD converges then it converges to a local minimum point of the original sparse problem is defined by Eq. (3).

3. Testing Algorithms on Correctly Classifying TOF

To better understand the strength and weakness of our proposed Algorithm 1, we compare the accuracy of our approach to alternative solver of sparsity constraint logistic regression

problems on a data set consisting of regional volume scores extracted from cine MRIs of 44 TOF cases and 38 healthy controls. The dataset provides an ideal test bed for such a comparison as it contains 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 monitoring of TOF patients, *i.e.*, timing for follow-up surgeries. Finally, the residual effects of TOF reconstructive surgery mostly impact the morphometry, *i.e.*, the shape of the right ventricle (Bailliard and Anderson, 2009; Atrey et al., 2010), so that the regional volume scores extracted from each time point of the image series are sample descriptions of the same phenomena, a core assumption of the group sparsity model. We not only show that our PD approach (Algorithm 1) reflects this manifestation of the disease by mostly weighing its decision based on regions within the right ventricle but also achieves significantly higher accuracy than alternative solutions to this model, such as solving logistic regression with relaxed sparsity constraints (Ryali et al., 2010) [1].

3.1. Experimental Setup

Extracting Regional Volume Scores—Each sample of the data set consists of a segmentation of each time point of a motion-corrected cine MRI, *i.e.*, we corrected for slice misalignment due to breathing motion by detecting the center of the left ventricle via Hough transform (Duda and Hart, 1972) and then stacking the slices so that the center of the left ventricle aligns across the slices. Covering the basal, mid-cavity, and apical part with 8 slices, the segmentation outlines the right ventricular blood pool, the wall of the Left Ventricle (**LV**), and the Ventricular Septum (**VS**), which was done at end-diastole according to the semi-automatic procedure described in (Ye et al., 2014) and then propagated from end-diastole to the other time points via non-rigid registration (Avants et al., 2008). For the right ventricular blood pool, we reduce the maps to a 7mm band along its boundary, which is similar to the width of the wall of the other two structures, and name it **RV** (see Fig. (1)). For each time point and image slice, we then parcellate the three structures into smaller sections based on a predefined subtended angles from the center of mass (of the RV or LV&VS). More specifically, RV and LV are divided into the same number of sections, while the VS is divided into $\frac{1}{3}$ of that number reflecting its relative size to RV and LV. For example in Fig. (2), the VS is divided into six sections with respect to each slice and time point of the scan while the RV and LV are divided into 18 sections each. Finally, the input to our proposed solver is the volume of each section.

Measure Classification Accuracy—We measure the accuracy of our approach with respect to correctly classifying samples just based on the sectional volumes scores of the RV alone, the LV alone, and the VS alone as well as using the scores of the whole heart (RV, LV&VS). We train our algorithm with different numbers of training samples, which are defined by the percentage {5%, 10%, . . . , 75%} of cases captured by the entire data set, *i.e.*, 82 cases. For each training sample size, we run ten experiments to estimate a distribution of accuracy scores.

[1]We use the SLEP package to solve the relaxed sparsity constraint model, see (Liu et al., 2009) for details.

For each experiment, we randomly select the training samples from both groups and label the remaining cases as test subjects. For a fair comparison with other sparse logistic regression solvers, we initialize our algorithm according to (Lu and Zhang, 2013), *i.e.*, $\varrho = 0.1$, $\sigma = \sqrt{10}$, $\epsilon_{BCD} = 10^{-4}$, $\epsilon_{PD} = 10^{-3}$, $W_i = \mathbf{0}$ and $v_i = 1$. For each training set, we then determine the optimal setting of our algorithm with respect to the broad parameter space of the remaining two parameters:

- the number of sections $s \in \{2, 3, \ldots, 9\}$ that each slice of the VS is divided into (so that each subject is represented by 3360 to 15120 volume scores)
- and the maximum percentage of regional values chosen by our approach $\omega \in \{5\%, 10\%, \ldots, 50\%\}$. In other words, the sparsity constraint is defined as $r := ceil(M \cdot s \cdot \omega)$ where $M \cdot s$ is the total number of section that the heart cycle is divided into.

We determine the optimal setting by first experimenting with parameter exploration. Specifically, for each of the 80 unique parameter settings we define a regressor by computing the optimal weights W^* of Algorithm 1 with respect to training data. In all experiments, PD then converged within 5 iterations and BCD converged within 500 iterations for each penalty parameter ϱ_p. After convergence, we compute the accuracy of the regressor with respect to the training set via the normalized accuracy (nAcc), *i.e.*, we separately compute the accuracy for each cohort and then average their values to account for the imbalance in cohort size. The entire process of training and measuring the accuracy of the 80 regressors took less than 10 seconds on a single PC (Intel(R) Xeon(R) CPU E5-2603 v2 @1.80GHz and 32G memory). It also resulted in multiple settings, *i.e.*, regressors, with 100% classification accuracy. In case of parameter exploration failing, a common solution (and the one we chose) is to train an ensemble of classifiers (Rokach, 2010). The final label of the ensemble is then defined by the weighted average across the set of regressors, where the weight of the regressor in the decision of the ensemble of classifiers is simply its training accuracy. Once trained, we then measure the accuracy of the resulting ensemble on correctly assigning test samples to the patient groups using the nAcc score. In the remainder of this section, we refer to this ensemble of classifier as **L0-Grp**.

Alternative Models—We compare our solver to alternative algorithm using the same mechanism as above, *i.e.*, we create ensemble of classifiers with respect to the same parameter space and training data sets, and measure their accuracy on the same test data sets. Specifically, we use the algorithm by (Liu et al., 2009) to solve the logistic regression with the sparsity constraints relaxed via the l_2-norm, *i.e.*,

$$\min_{v \in \mathbb{R}, W \in \mathbb{R}^{M \times T}} l_{avg}(v, W) + \lambda \sum_{i=1}^{M} \|W^i\|_2 \quad (18)$$

with λ being the sparse regularizing parameter. We refer to the corresponding ensemble as **Rlx-Grp**. Note, relaxing the sparsity constraint via the l_1-norm results in a optimization problem ignoring temporal consistency, which violates our initial assumption of the model.

In addition, we investigate the accuracy of PD and (Liu et al., 2009) when only applied to a single time point ($T = 1$). When $T = 1$, then Eq. (3) simplifies to

$$\min_{v\in\mathbb{R},W\in\mathbb{R}^M} l_{avg}(v,W) \quad \text{s.t.} \quad \|W\|_0 \leq r, \tag{19}$$

i.e., the sparsity constrained problem solved by PD in (Lu and Zhang, 2013). We refer to the corresponding ensemble as **L0-nGrp**. Furthermore, for $T = 1$, Eq. (18) is equivalent to

$$\min_{v\in\mathbb{R},W\in\mathbb{R}^{M\times T}} l_{avg}(v,W) + \lambda\|W\|_1, \tag{20}$$

whose corresponding ensemble of classifier we refer to **Rlx-nGrp**. Finally, we note that the sparsity parameter λ of Rlx-Grp and Rlx-nGrp is not directly related to the number of selected sections. For a fair comparison to the sparsity constrained methods, we therefore automatically tune the sparsity parameter λ of Eq. (18) and Eq. (20) so that the number of chosen sections $\bar{r}$ were similar to those defined in Eq. (3), *i.e.*, $|r - \bar{r}| \leq 1$.

3.2. Experimental results

The box plots of Fig. (3) summarize the distribution of the accuracy scores associated with each implementation and structure by the average, the first quin-tile, and the fourth quintile of the nAcc scores across the 10 test data sets run for each training sample size. For all four implementations, the average nAcc scores generally increase with the number of training samples. For the RV and the whole heart (RV,LV&VS), the proposed L0-Grp implementation (red boxplots) generally achieves a higher average accuracy, first quintile, and fourth quintile scores than then other three approaches. The difference is especially large for smaller number of training samples, *i.e.*, 5% to 35%. For large training samples, *i.e.*, 70% and 75%, L0-Grp is the only method with a fourth quintile score of 100%. To follow up these observations, we computed the p-value of the paired-sample t-test (McDonald, 2009) between L0-Grp and other three implementations. Table 1 summarizes those computations. For the whole heart, L0-Grp is significantly better ($p < 0.05$) than Rlx-nGrp and L0-nGrp with respect to 13 out of the 15 training sample sizes, and 10 out of the 15 training sets with respect to the Rlx-Grp. For the RV, the counts for statically significant differences increase with respect to the implementations with relaxed constraints. However, this is not true for L0-nGrp implementation, in which case L0-Grp is significantly better in 8 experiments. In this experiment, we conclude that the impact of reducing the number of time points on the accuracy scores is less than relaxing the sparsity constraint. An explanation for this observation might lie in the fact that solution generated by solvers relaxing the sparsity constraint, *i.e.*, Rlx-nGrp and Rlx-Grp, is only accurate with respect to the original problem Eq. (3) under certain conditions (Candès et al., 2006), which are not satisfied here and in medical image analysis in general.

With respect to the LV and VS, Fig. (3) reports insignificant differences between the accuracy of all four methods. The average nAcc scores start at around 55% and generally increase with the number of training samples. While the average scores of Rlx-Grp almost

match those of L0-Grp, L0-Grp achieves the highest average sore with 69% for the LV and 74% for the VS indicating that the VS is slightly more impacted by TOF than the LV. This observation is also in alignment with the literature (Bailliard and Anderson, 2009; Atrey et al., 2010) reporting the residual effects of TOF impacting the RV, which the VS is attached to. Increasing the number of training samples also impacts the spread between the first and fourth quintile. This issue is partly due to the fact that the larger the training data set, the smaller the test set. From a statistical perspective, e.g. recording the outcome of flipping a biased coin several times, one expects this outcome as the first and fourth quintile scores deviate further from the average score compared to experiments with larger test sets. Interestingly enough, only the average scores of the L0-nGrp are not improving with large training sets as it peaks at a training size of 50% for the LV and gets unstable starting at 50% for the VS. In other words, adding more samples to the training is not more informative than adding more information of each individual sample by, for example, increasing the number of time points or including measurements from the RV, the structure most impacted by TOF.

To gain a deeper understanding of the experimental results, Fig. (4) plots the importance of heart sections in distinguishing TOF from healthy controls based on the regional volume scores and with respect to the type of solver and percentage associated with the training sample size. The incomplete circle on the left represents the importance of sections of the RV and on the right the importance of sections of the LV and VS. Each ring of those (incomplete) circles represents a slice of the cine MRI with the outer circle representing the base of the heart. As it is common in the cardiac literature, we overlay the bullseye plot over the LV & VS maps. For each type and percentage, we infer the importance of a section from their average importance across the corresponding ensembles of classifiers, *i.e.*, the number of times a section was selected by a sparse solver multiplied by the solvers' training accuracy. Sections in white were never selected, those in blue had very low and those in red very high impact on the final classification. The ensembles were indifferent to sections labeled turquoise, green, yellow, and orange. We first note that across solvers the number of indifferent section reduces and the number of ignored section increases with larger number of training samples. This indicates a higher confidence of the ensembles in the importance of sections. Furthermore, the number of selected RV sections (left) increases, which is inline with the increase in testing accuracy. Of all methods, L0-Grp relies least on LV sections (right). This could explain its significantly higher accuracy scores compared to those other method in the majority of whole heart experiments. As noted in Section 1, one has to be careful when relating the weights of classifiers to biomedical landmarks. We are in this experiment as the importance maps of Fig. (4) are based on the number of times regions were selected by solvers (not their soft weights) and those of the L0-Grp are in accordance with the medical literature stating that residual effects of TOF mostly impact the RV.

4. Conclusion

We generalized the PD approach to directly solve group cardinality constraint logistic regression, *i.e.*, simultaneously performing disease classification and temporal-consistent pattern identification. To do so, we assumed that an intra-subject series of images represents repeated samples of the same disease patterns. We modeled this assumption by combining series of measurements created by a feature across time into a single group. Unlike existing

approaches, our algorithm then derived a solution within that model by decoupling the minimization of the logistic regression function from enforcing the group sparsity constraint. The minimum to the smooth and convex logistic regression problem was determined via gradient descent while we derived a closed form solution for finding a sparse approximation of that minimum. We applied our method to cine MRI of 38 healthy controls and 44 adult patients that received reconstructive surgery of Tetralogy of Fallot (TOF) during infancy. Our method correctly identified the RV to be most impacted by TOF and generally obtained statistically significant higher classification accuracy than alternative solutions to this model, *i.e.*, ones relaxing group cardinality constraints or ones only applied to a single time point.

While the experiments were confined to regional volumes scores extracted from cine MRIs, the method could be applied to any features computed from intra-subject sequence of images. One only has to ensure that the assumption holds that series of images represent repeated samples of the same disease patterns. Furthermore, one has to be careful when relating the selected features to disease patterns. We did so on our experiments as the selcted features agreed with the medical literature.

Acknowledgment

We would like to thank Drs. Benoit Desjardins and DongHye Ye for their help on generating the cardiac dataset. This research was supported by NIH grants (R01 HL127661, K05 AA017168) and the Creative and Novel Ideas in HIV Research (CNIHR) Program through a supplement to the University of Alabama at Birmingham (UAB) Center For AIDS Research funding (P30 AI027767). This funding was made possible by collaborative efforts of the Office of AIDS Research, the National Institute of Allergy and Infectious Diseases, and the International AIDS Society.

Appendix A

Lemma A. 1

Given a matrix $A \in {}^{M\times T}$, *let* A^j *be the* j^{th} *row of A for j* = 1,..., *M. Without loss of the generality, we assume that*

$$\|A^1\|_2 \geq \|A^2\|_2 \geq \ldots \geq \|A^M\|_2 > 0.$$

Then the solution for the minimization problem

$$W^* := \underset{W\in\mathscr{X}}{arg\,min}\|W - A\|_F^2, \quad \text{(A.1)}$$

within $\mathscr{X} := \left\{W \in {}^{m\times T} : \widetilde{W} := \left(\|W^1\|_2, \ldots, \|W^m\|_2\right)^{\mathsf{T}} \quad and \quad \|\widetilde{W}\|_0 \leq r\right\}$ *are the first r rows of A, i.e.,*

$$(W^*)^j = \begin{cases} A^j, & if \quad j \leq r; \\ 0, & otherwise, \end{cases} \quad for \quad j=1,\ldots,M. \quad \text{(A.2)}$$

Proof

Suppose there is a solution $\overline{W}$ to Eq. (A.1) which is different from W^* defined in Eq. (A.2). We now show that the $\|\overline{W} - A\|_F^2 \geq \|W^* - A\|_F^2$.

We know that for all $\|\overline{W}^j\| \neq 0$, the following is true $\overline{W}^j = A^j$ as otherwise $\overline{W}$ can not be an optimal solution for Eq. (A.1). To see that, simply replace $\overline{W}^j$ with A^j for any j such that $\|\overline{W}^j\| \neq 0$, which results in returns a smaller value for $\|W - A\|_F^2$. Since $\overline{W}$ is different from W^*, there must exists $j_1 > r$ such that $\overline{W}^{j_1} = A^{j_1}$. Given that $\overline{W} \in \mathcal{X}$, then there must exists a row $\overline{W}^{j_2}$ such that $\|\overline{W}^{j_2}\|_2 = 0$ for $j_2 \leq r$. By using the definition of W^*, we also have $\|(W^*)^{j_2} - A^{j_2}\|_2 = 0$ and $(W^*)^{j_1} = 0$. In addition, $\|A^{j_2}\|_2 \geq \|A^{j_1}\|_2$ according to the assumption that $\|A^1\|_2 \geq \|A^2\|_2 \geq \ldots \geq \|A^M\|_2$. Then we have

$$\begin{aligned} \|A^{j_1} - \overline{W}^{j_1}\|_2^2 + \|A^{j_2} - \overline{W}^{j_2}\|_2^2 &= \|A^{j_2}\|_2^2 \\ &\geq \|A^{j_1}\|_2^2 \\ &= \|A^{j_1} - (W^*)^{j_1}\|_2^2 + \|A^{j_2} - (W^*)^{j_2}\|_2^2, \end{aligned}$$

which implies that if we define

$$\left(W'\right)^j = \begin{cases} \overline{W}^j, & \text{if} \quad j \neq j_1, j_2; \\ (W^*)^j, & \text{otherwise} \end{cases} \quad \text{for} \quad j = 1, \ldots, M,$$

then $\|\overline{W} - A\|_F^2 \geq \|W' - A\|_F^2$. Continuing this procedure of replacing values results at some point $W' = W^*$ and thus $\|\overline{W} - A\|_F^2 \geq \|W^* - A\|_F^2$. Thus W^* is an optimal solution of problem Eq. (A.1).

Lemma A. 2

Suppose that $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ *is an accumulation point of the sequence* $\{(v_b, Y_b, W_b)\}$ generated by BCD described in Section 2. Then it is also the block coordinate minimum point of Eq. (7), i.e.,

$$\begin{aligned} \left(\hat{v}_p, \hat{Y}_p\right) &:= \underset{v \in \mathbb{R}, Y \in \mathbb{R}^{M \times T}}{arg\,min}\, q_{\varrho_p}\left(v, Y, \hat{W}_p\right), \\ \hat{W}_p &:= \underset{W \in \mathcal{X}}{arg\,min} q_{\varrho_p}\left(\hat{v}_p, \hat{Y}_p, W\right). \end{aligned} \tag{A.3}$$

Proof

First, note that

$$\begin{aligned} q_{\varrho_p}(v_{b+1}, Y_{b+1}, W_b) &\leq q_{\varrho_p}(v, Y, W_b) \quad \forall v \in , Y \in {}^{M\times T}, \\ q_{\varrho_p}(v_{b+1}, Y_{b+1}, W_{b+1}) &\leq q_{\varrho_p}(v_{b+1}, Y_{b+1}, W) \quad \forall W \in \mathscr{X}. \end{aligned} \tag{A.4}$$

It then follows that

$$q_{\varrho_p}(v_{b+1}, Y_{b+1}, W_{b+1}) \leq q_{\varrho_p}(v_{b+1}, Y_{b+1}, W_b) \leq q_{\varrho_p}(v_b, Y_b, W_b) \quad \forall b \geq 1. \tag{A.5}$$

Hence, the sequence $\{q_{\varrho_p}(v_b, Y_b, W_b)\}$ is non-increasing. Since $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is an accumulation point of $\{(v_b, Y_b, W_b)\}$, there exists a subsequence L such that

$$\lim_{b\in L\to\infty} \{(v_b, Y_b, W_b)\} = \left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right).$$

We then observe that $\{q_{\varrho_p}(v_b, Y_b, W_b)\}_{b\in L}$ is bounded, which together with the monotonicity of $\{q_{\varrho_p}(v_b, Y_b, W_b)\}$ implies that $\{q_{\varrho_p}(v_b, Y_b, W_b)\}$ is bounded below and hence $\lim_{b\to\infty} q_{\varrho_p}(v_b, Y_b, W_b)$ exists. This observation, Eq. (A.5), and the continuity of $q_{\varrho_p}(\cdot,\cdot,\cdot)$ yield

$$\begin{aligned} lim_{b\to\infty} q_{\varrho_p}(v_{b+1}, Y_{b+1}, W_b) &= lim_{b\to\infty} q_{\varrho_p}(v_b, Y_b, W_b) \\ &= lim_{b\in L\to\infty} q_{\varrho_p}(v_b, Y_b, W_b) \\ &= q_{\varrho_p}\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right). \end{aligned}$$

Given that $q_{\varrho_p}(\cdot,\cdot,\cdot)$ is continuous, then taking limits on both sides of Eq. (A.4) with respect to $b \in L \to \infty$ results in

$$\begin{aligned} q_{\varrho_p}\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right) &\leq q_{\varrho_p}\left(v, Y, \hat{W}_p\right) \quad \forall v \in , Y \in {}^{M\times T}, \\ q_{\varrho_p}\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right) &\leq q_{\varrho_p}\left(\hat{v}_p, \hat{Y}_p, W\right) \quad \forall W \in \mathscr{X}. \end{aligned} \tag{A.6}$$

Note that $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is the accumulation point of $\{(v_b, Y_b, W_b)\}_{b\in L\to\infty}$. Then according to the definition of $\mathscr{X}$, we have $\|\widetilde{W}_b\|_0 \leq r$ which immediately implies $\|\tilde{\hat{W}}_p\|_0 \leq r$. Thus, $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is a block coordinate minimum point of Eq. (7).

Lemma A. 3

Let $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ *be a block coordinate minimum of Eq. (7) and* $H \in {}^{M\times T}$ *define a small feasible step of* $\hat{W}_p$, *i.e.,* $\hat{W}_p + H \in \mathscr{X}$. *Then*

$$H^i\left(\hat{Y}_p^i - \hat{W}_p^i\right)^T = 0, \quad for \quad i \in \{1, \ldots, M\}. \tag{A.7}$$

Proof

We note that if $\|\tilde{\hat{Y}}_p\|_0 \leq r$ where $\tilde{\hat{Y}}_p := \left(\|\hat{Y}^1\|_2, \ldots, \|\hat{Y}^m\|_2\right)^{\mathsf{T}}$, then according to Eq. (A.2), $\hat{W}_p = \hat{Y}_p$. Thus, Eq. (A.7) is true. For $\|\tilde{\hat{Y}}_p\|_0 > r$, we observe from Eq. (A.2) that $\left(\hat{W}_p\right)^i = \left(\hat{Y}_p\right)^i$ for all $i \in \Omega_{\hat{W}_p}$ where $\Omega_{\hat{W}_p} := \left\{i \in \mathscr{I} : \|\left(\hat{W}_p\right)^i\|_2 \neq 0\right\}$. Thus $(H)^i\left(\hat{Y}_p^i - \hat{W}_p^i\right)^T = 0$ for $i \in \Omega_{\hat{W}_p}$. On the other hand, $\|(H)^i\|_2 < \|\left(\hat{W}_p\right)^i\|_2$ for all $i \in \Omega_{\hat{W}_p}$, so that $\left(\hat{W}_p + H\right)^i \neq 0$ for all $i \in \Omega_{\hat{W}_p}$. From $\hat{W}_p + H \in \mathscr{X}$, it follows that $(H)^i = 0$ for all $i \notin \Omega_{\hat{W}_p}$ and hence $(H)^i\left(\hat{Y}_p^i - \hat{W}_p^i\right)^T = 0$ for $i \notin \Omega_{\hat{W}_p}$.

Theorem A. 4

Suppose that $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ *is an accumulation point of the sequence* $\{(v_b, Y_b, W_b)\}$ *generated by BCD described in Section 2. Then,* $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ *is a local minimum point of Eq. (7).*

Proof

According to Lemma A.2, we have $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is a block coordinate minimum point of Eq. (7). Next we show that $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)$ is also a local minimum point of Eq. (7).

Since $l_{\text{avg}}(\cdot,\cdot)$ is a convex function, we know that $q_{\varrho p}(\cdot,\cdot,\cdot)$ is also convex. It then follows from the first relation of Eq. (A.3), the partial derivative of $q_{\varrho p}(\cdot,\cdot\cdot)$ 1.1.1 in Bertsekas (1999)), that

$$\begin{aligned}
&\left.\frac{\partial q_{\varrho p}(v,Y,W)}{\partial Y}\right|_{v=\hat{v}_p, Y=\hat{Y}_p, W=\hat{W}_p} = 0, \\
&\left.\frac{\partial q_{\varrho p}(v,Y,W)}{\partial v}\right|_{v=\hat{v}_p, Y=\hat{Y}_p, W=\hat{W}_p} = 0, \\
&\left.\frac{\partial q_{\varrho p}(v,Y,W)}{\partial W}\right|_{v=\hat{v}_p, Y=\hat{Y}_p, W=\hat{W}_p} = \varrho_p\left(\hat{Y}_p - \hat{W}_p\right).
\end{aligned} \tag{A.8}$$

Let H be a "small" feasible step of $\hat{W}_p$, $h \in$ and $G \in {}^{M \times T}$. Then using Lemma 3, Eq. (A.8) and the convexity of $q_{\varrho p}$, we have

$$q_{\varrho_p}\left(\hat{v}_p+h,\hat{Y}_p+G,\hat{W}_p+H\right) \overset{2}{\geq} q_{\varrho_p}\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)+h\left[\left.\frac{\partial q_{\varrho_p}(v,Y,W)}{\partial v}\right|_{v=\hat{v}_p,Y=\hat{Y}_p,W=\hat{W}_p}\right]+Tr\left(G\left[\left.\frac{\partial q_{\varrho_p}(v,Y,W)}{\partial Y}\right|_{v=\hat{v}_p,Y=\hat{Y}_p,W=\hat{W}_p}\right]^T\right)+$$
$$\overset{Eq.(A.8)}{=} q_{\varrho_p}\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)+\varrho_p Tr\left(H\left(\hat{Y}_p-\hat{W}_p\right)^T\right)$$
$$\overset{\text{Lemma A. 3}}{\geq} q_{\varrho_p}\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right),$$

which together with the above choice of *h*, *G* and *H* implies that $\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)$ is local minimum point of Eq. (7).

Lemma A.5

Let $\left\{\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)\right\}$ *be the sequence generated by* ***PD***, $\rho_p=\rho\cdot\sigma^{p-1}$ *and* $\Gamma=l_{avg}\left(v_i,W_i\right)$. *Suppose* (*v**, *Y**, *W**) *is an accumulation point of* $\left\{\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)\right\}$. *Then*

$$Y^*=W^*.$$

Proof

Since (*v*$_i$, *W*$_i$) defined in Algorithm 1 is a feasible point of Eq. (3), we have

$$\min_{v,Y} q_{\varrho_{p+1}}\left(v,Y,W_i\right)\leq q_{\varrho_{p+1}}\left(v_i,W_i,W_i\right)=l_{avg}\left(v_i,W_i\right)=\Gamma.$$

By the specification of W_0 according to line 17-20 of Algorithm 1, we have

$$\min_{v,Y} q_{\varrho_{p+1}}\left(v,Y,W_0\right)\leq\Gamma.$$

From Eq. (A.4), we know that the sequence of $q_{\varrho_p}(v_b, Y_b, W_b)$ is non-increasing. Thus in view of Eq. (7) and the choice of $\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)$ that is specified in line 21 of Algorithm 1, we observe that $\forall p$

$$l_{avg}\left(\hat{v}_p,\hat{Y}_p\right)+\frac{\varrho_p}{2}\|\hat{W}_p-\hat{Y}_p\|_F^2=q_{\varrho_p}\left(\hat{v}_p,\hat{Y}_p,\hat{W}_p\right)\leq\Gamma.$$

By the definition of $l_{avg}\left(\hat{v}_p,\hat{Y}_p\right)$, we have $l_{avg}\left(\hat{v}_p,\hat{Y}_p\right)>0$. Then we obtain that

$$\|\hat{W}_p-\hat{Y}_p\|_F^2\leq 2\left[\Gamma-l_{avg}\left(\hat{v}_p,\hat{Y}_p\right)\right]/\varrho_p\leq 2\Gamma/\varrho_p. \quad \text{(A.9)}$$

Since (v^*, Y^*, W^*) is an accumulation point of $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}$, there exists a subsequence $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}_{k\in\mathscr{S}} \to (v^*, Y^*, W^*)$. Then taking limits on both sides of Eq. (A.9) as $p \in \mathscr{S} \to \infty$, and using $\varrho_p \to \infty$ as $p \to \infty$, we see that $\|W^* - Y^*\|_F^2 = 0$. Thus, the conclusion holds immediately.

Lemma A. 6

Let $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}$ *be the sequence generated by PD,* $\mathscr{K}_{\hat{W}_p} = \{i_1^p, \ldots, i_r^p\}$ *be a set of r district indices in* $\mathscr{I} := \{1, \ldots, M\}$ *such that* $\left(\hat{W}_p\right)^i = 0$ *for any* $i \notin \mathscr{K}_{\hat{W}_p}$. *Suppose* ($v^*$, Y^*, W^*) *is an accumulation point of* $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}$, *then when k* ∈ *S is sufficiently large,*

$$\mathscr{K}_{\hat{W}_p} = \mathscr{K}_{W^*}$$

for some index set $\mathscr{K}_{W^*} \subseteq \mathscr{I}$.

Proof

Since (v^*, Y^*, W^*) is an accumulation point of $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}$, there exists a subsequence $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}_{k\in\mathscr{S}} \to (v^*, Y^*, W^*)$. Since $\mathscr{K}_{\hat{W}_p}$ is an index set, $\{(i_1^p, \ldots, i_r^p)\}_{p\in\mathscr{S}}$ is bounded for all *k*. Thus, there exists a subsequence $S \subseteq \mathscr{S}$ such that $\{(i_1^p, \ldots, i_r^p)\}_{p\in S} \to (i_1^*, \ldots, i_r^*)$ for some *r* distinct indices $i_1^*, \ldots, i_r^*$. Since $i_1^k, \ldots, i_r^k$ are *r* distinct integers, one can easily conclude that $(i_1^p, \ldots, i_r^p) = (i_1^*, \ldots, i_r^*)$ for sufficiently large $p \in S$. Let $\mathscr{K}_{W^*} = \{i_1^*, \ldots, i_r^*\}$. It then follows that $\mathscr{K}_{\hat{W}_p} = \mathscr{K}_{W^*}$ when $k \in S$ is sifficiently large, and moreover, $\left\{\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p\right)\right\}_{k\in S} \to (v^*, Y^*, W^*)$.

Lemma A. 7

Let (v^*, W^*) ∈ $(, \mathscr{X})$ *and the family of subsets of* $\mathscr{I}$ with size ‘r’ and the sets’ complement only consisting of the zero rows of W* be defined as

$$\mathscr{I}_{W^*} := \left\{\mathscr{K} \subseteq \mathscr{I} : |\mathscr{K}| = r, (W^*)^{\overline{\mathscr{K}}} = 0\right\},$$

where $\overline{\mathscr{K}} := \mathscr{I} \setminus \mathscr{K}$. ($v^*$, W^*) is then a local minimum point of Eq. (3) if for each $\mathscr{K} \in \mathscr{I}_{W^*}$ *there exists a matrix* $\Lambda_{\mathscr{K}} \in {}^{M\times T}$ *so that the following holds*:

$$\begin{aligned} \frac{\partial l_{avg}(v,W)}{\partial W}|_{v=v^*,W=W^*}+\Lambda_{\mathscr{K}} &=0, \\ \frac{\partial l_{avg}(v,W)}{\partial v}|_{v=v^*,W=W^*} &=0, \\ (\Lambda_{\mathscr{K}})^{\overline{\mathscr{K}}} &=0 \end{aligned} \quad \text{(A.10)}$$

with $f(x)|_{x=x'}$ *being the value of f(·) at x′.*

Proof

The proof first determines the necessary condition for minima of Eq. (3) and then shows that any (v^*, W^*) fulfilling the necessary condition also fulfils the sufficient condition. To derive the necessary condition, let us assume that (v^*, W^*) is a local minimum point of Eq. (3). Now, we know that $\mathscr{I}_{W^*}$ may contain more than one component due to the inequal cardinality constraints in Eq. (3), *i.e.,* $\|\widetilde{W^*}\|_0 \leq r$. Then for any $\mathscr{K} \in \mathscr{I}_{W^*}$, we observe that ($v^*$, W^*) also minimizes the following problem

$$\min_{v\in\mathbb{R},W\in\mathbb{R}^{M\times T}} l_{avg}(v,W) \quad s.t. \quad W^{\overline{\mathscr{K}}}=0. \quad \text{(A.11)}$$

Then according to Proposition 3.1.1 of (Bertsekas, 1999), for any (v^*, W^*) being the solution of Eq. (A.11), it is necessary that there exists a matrix $\Lambda_{\mathscr{K}} \in {}^{M\times T}$ so that the following holds:

$$\begin{aligned} \frac{\partial l_{avg}(v,W)}{\partial W}|_{v=v^*,W=W^*}+\Lambda_{\mathscr{K}} &=0, \\ \frac{\partial l_{avg}(v,W)}{\partial v}|_{v=v^*,W=W^*} &=0, \\ (\Lambda_{\mathscr{K}})^{\overline{\mathscr{K}}} &=0. \end{aligned} \quad \text{(A.12)}$$

Now any (v^*, W^*) that is a local minimum point of Eq. (3) also has to be a local minimum point of Eq. (A.11) for all $\mathscr{K} \in \mathscr{I}_{W^*}$. Thus it has to fulfill Eq. (A.12) for all $\mathscr{K} \in \mathscr{I}_{W^*}$ so that Eq. (A.12) is a first-order necessary condition of Eq. (3).

From now on, we call any (v^*, W^*) fulfilling Eq. (A.12) for all $\mathscr{K} \in \mathscr{I}_{W^*}$ a stationary point. We now show that a first-order sufficient condition of Eq. (3) is the existence of a stationary point, *i.e.,* any point fulfilling the necessary condition is also a local minimum point of Eq. (3). From Proposition 3.4.1 of (Bertsekas, 1999), we know that (v^*, W^*) is the global minimum point of Eq. (A.11) for all

$$\overline{\mathscr{K}} \in \left\{\mathscr{I}\backslash\mathscr{K}_{W^*} : \mathscr{K}_{W^*} \in \mathscr{I}_{W^*}\right\}.$$

Hence, there exists $\epsilon > 0$ and neighborhoods of (v^*, W^*) for different $\mathscr{K} \in \mathscr{I}_{W^*}$, *i.e.,*

$$\mathcal{N}_{\mathcal{K}}(v^*, W^*;\epsilon) := \left\{ v \in , W \in {}^{M\times T} : W^{\overline{\mathcal{K}}} = 0 \quad \text{and} \quad \sqrt{(v-v^*)^2 + \|W - W^*\|_F^2} < \epsilon \right\}$$

with $\epsilon > 0$, such that for all members of the union of neighborhoods, *i.e.*,

$$\forall (v, W) \in \cup_{\mathcal{K}\in\mathcal{I}_{W^*}} \mathcal{N}_{\mathcal{K}}(v^*, W^*;\epsilon),$$

the following is true

$$l_{avg}(v, W) \geq l_{avg}(v^*, W^*).$$

Now we define the neighborhood of (v^*, W^*) with respect to the sparse space as $\mathcal{X}$ as

$$\mathcal{N}(v^*, W^*;\epsilon) := \left\{ v \in , W \in \mathcal{X} : \sqrt{(v-v^*)^2 + \|W - W^*\|_F^2} < \epsilon \right\}.$$

Since $\cup_{\mathcal{K}\in\mathcal{I}_{W^*}} \mathcal{N}_{\mathcal{K}}(v^*, W^*;\epsilon) = \mathcal{N}(v^*, W^*;\epsilon)$, then for any $(v, W) \in \mathcal{N}(v^*, W^*;\epsilon)$ there exists $\mathcal{K} \in \mathcal{I}_{W^*}$ so that according to Eq. (A.11) $l_{avg}(v, W) \geq l_{avg}(v^*, W^*)$. Hence

$$l_{avg}(v, W) \geq l_{avg}(v^*, W^*), \quad \forall (v, W) \in \mathcal{N}(v^*, W^*;\epsilon).$$

Thus, any stationary point (v^*, W^*) is a local minimum point of Eq. (3).

Theorem A. 8

Let (v^*, Y^*, W^*) *be an accumulation point of the sequence* $\left\{ \left(\hat{v}_p, \hat{Y}_p, \hat{W}_p \right) \right\}$ *generated by PD. Assume that the solution* $\left(\hat{v}_p, \hat{Y}_p, \hat{W}_p \right)$ *obtained by BCD satisfies*

$$\left\| \frac{\partial q_{\varrho_p}(v, Y, W)}{\partial v} \Big|_{v=\hat{v}_p, Y=\hat{Y}_p, W=\hat{W}_p} \right\|_F < \epsilon_p \tag{A.13}$$

for $\epsilon_p \to 0$. *If* $\|\widetilde{W^*}\|_0 = r$, *then* ($v^*$, W^*) *is a local minimum point of problem Eq. (3).*

Proof

We now prove the statement by first showing $Z_p := \varrho_p \left(\hat{Y}_p - \hat{W}_p \right)$ is bounded, then that converges to $\Lambda_{\mathcal{K}}$ for some $\mathcal{K} \in \mathcal{I}_{W^*}$ (the matrix defined in Eq. (A.12)), and finally the (v^*, W^*) have to be a minimum point of Eq. (3) when $\|\widetilde{W^*}\|_0 = r$. Now let us assume that Z_p is not bounded. To contradict this results, let

$$P_p := -\frac{\partial q_{\varrho_p}(v, Y, W)}{\partial v}|_{v=\hat{v}_p, Y=\hat{Y}_p, W=\hat{W}_p}.$$

It then follows from Eq. (A.13) that $\|P_p\|_F \leq \epsilon_p$ for all p. Thus for $lim_{p\to\infty}\epsilon_p=0$ follows $lim_{p\to\infty}P_p=0$. Applying Eq. (17) and the definition of q_{ϱ_p} in Eq. (7), we have

$$-\frac{\partial l_{avg}(v, Y)}{\partial Y}|_{v=\hat{v}_p, Y=\hat{Y}_p} - Z_p - P_p=0. \quad \text{(A.14)}$$

Now $\left\{\|Z_p\|_F\right\}_{k\in S} \to \infty$. Let $\overline{Z}_p=Z_p/\|Z_p\|_F$. Obviously, the sequence $\left\{\overline{Z}_p\right\}$ is bounded. Then by using Bolzano-Weierstrass Theorem, there must exists an accumulation point $\overline{Z}$ such that $\left\{\overline{Z}_p\right\}_{p\in K\subseteq S} \to \overline{Z}$. Clearly $\|\overline{Z}\|_F=1$. Dividing both sides Eq. (A.14) by $\|Z_p\|_F$, taking the limits with respect to $p \in K \to \infty$, and using the relation $lim_{p\in S\to\infty}P_p=0$, we obtain that

$$\bar{Z} =0, \quad \text{(A.15)}$$

which contradicts $\|\overline{Z}\|_F=1$. Therefore, the subsequence $\{Z_p\}_{p\in S}$ is bounded. By applying Bolzano-Weierstrass Theorem Bartle and Sherbert (1982) and the boundness of $\{Z_p\}_{p\in S}$, there must exists an accumulation point Z^* such that $\{Z_p\}_{p\in K\subseteq S} \to Z^*$. Taking limits on both sides of Eq. (A.14) as $p \in K \subseteq S \to \infty$, and using the relations $lim_{k\in S\to\infty}P_p=0$, we see that the first two relations of Eq. (A.12) hold with $Z^* = \Lambda^*$. From Eq. (10) and the definitions of $\mathscr{I}_{\hat{W}_p}$, we have $\left(\hat{W}_p\right)_p^{\mathscr{K}} = \left(\hat{Y}_p\right)^{\mathscr{K}}$ for and hence $(Z_p)^{\mathscr{K}_{\hat{W}_p}}=0$. In addition, we know from Lemma 6 that $\mathscr{K}_{\hat{W}_p}=\mathscr{K}_{W^*}$ when $k \in S$ is sufficiently large. Hence $(Z^*)^{\mathscr{K}_{W^*}}=0$. This together with the definitions of $\mathscr{K}_{W^*}$ and $\overline{\mathscr{K}_{W^*}}$ implies that Z^* satisfies

$$(Z^*)^i = \begin{cases} 0 & \text{if} \quad i \in \mathscr{K}_{W^*}, \\ \left(\Lambda_{\mathscr{K}_{W^*}}\right)_i & \text{if} \quad i \in \overline{\mathscr{K}_{W^*}}. \end{cases}$$

Now $\|\widetilde{W^*}\|_0=r$ so that

$$\mathscr{I}_{W^*} = \left\{\mathscr{K}_{W^*} \subseteq \mathscr{I} : |\mathscr{K}_{W^*}|=r, (W^*)^i=0, \quad \forall i \notin \mathscr{K}_{W^*}\right\}$$

has only one unique component. Hence, Z^* together with (v^*, W^*) satisfies Eq. (A.12) so that according to Lemma A.7 (v^*, W^*) is a local minimum point of Eq. (3).

References

Afshin M, Ben Ayed I, Punithakumar K, Law M, Islam A, Goela A, Peters TM, Li S. Regional assessment of cardiac left ventricular myocardial function via mri statistical features. IEEE Transactions on Medical Imaging. 2014; 33:481–494. [PubMed: 24184708]

Aljabar P, Wolz R, Srinivasan L, Counsell SJ, Rutherford MA, Edwards AD, Hajnal JV, Rueckert D. A combined manifold learning analysis of shape and appearance to characterize neonatal brain development. IEEE Transactions on Medical Imaging. 2011; 30:2072–2086. [PubMed: 21788184]

Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: a survey. Multimedia System. 2010; 16:345–379.

Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis. 2008; 12:26–41. [PubMed: 17659998]

Bai, W.; Peressutti, D.; Oktay, O.; Shi, W.; O'Regan, DP.; King, AP.; Rueckert, D. Functional Imaging and Modeling of the Heart. Springer; 2015. Learning a global descriptor of cardiac motion from a large cohort of 1000+ normal subjects; p. 3-11.

Bailliard F, Anderson RH. Tetralogy of Fallot. Orphanet Journal of Rare Diseases. 2009; 4:1–10. [PubMed: 19133130]

Bartle, R.; Sherbert, D. Matemáticas (Limusa). Wiley; 1982. Introduction to real analysis..

Bernal-Rusiel JL, Reuter M, Greve DN, Fischl B, Sabuncu MR. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage. 2013; 81:358–370. [PubMed: 23702413]

Bertsekas, D. Athena Scientific optimization and computation series. Athena Scientific; 1999. Nonlinear programming..

Besbes, A.; Komodakis, N.; Glocker, B.; Tziritas, G.; Paragios, N. Advances in Visual Computing. Springer; 2007. 4D ventricular segmentation and wall motion estimation using efficient discrete optimization; p. 189-198.

Bhatia KK, Rao A, Price AN, Wolz R, Hajnal JV, Rueckert D. Hierarchical manifold learning for regional image analysis. IEEE Transactions on Medical Imaging. 2014; 33:444–461. [PubMed: 24235274]

Candès EJ, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory. 2006; 52:489–509.

Candes EJ, Tao T. Decoding by linear programming. IEEE Transactions on Information Theory. 2005; 51:4203–4215.

Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage. 2009; 44:112–122. [PubMed: 18793733]

Chandrashekara R, Mohiaddin RH, Rueckert D. Analysis of 3-D myocardial motion in tagged MR images using nonrigid image registration. IEEE Transactions on Medical Imaging. 2004; 23:1245–1250. [PubMed: 15493692]

Chetelat G, Landeau B, Eustache F, Mezenge F, Viader F, de La Sayette V, Desgranges B, Baron JC. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. NeuroImage. 2005; 27:934–946. [PubMed: 15979341]

Deshpande, H.; Maurel, P.; Barillot, C. 2nd International Workshop on Sparsity Techniques in Medical Imaging (STMI). MICCAI 2014; 2014. Detection of multiple sclerosis lesions using sparse representations and dictionary learning.

Duda RO, Hart PE. Use of the hough transformation to detect lines and curves in pictures. Communications of the ACM. 1972; 15:11–15.

Friedman J, Hastie T, Tibshirani R. A note on the group Lasso and a sparse group Lasso. arXiv. 2010:1001.0736.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage. 2014; 87:96–110. [PubMed: 24239590]

Huang H, Shen L, Zhang R, Makedon F, Hettleman B, Pearlman J. A prediction framework for cardiac resynchronization therapy via 4D cardiac motion analysis. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005. volume 3749 of Lecture Notes in Computer Science. 2005:704–711.

Lempitsky, V.; Verhoek, M.; Noble, JA.; Blake, A. Functional Imaging and Modeling of the Heart. Springer; 2009. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography; p. 447-456.

Li S, Yin H, Fang L. Group-sparse representation with dictionary learning for medical image denoising and fusion. IEEE Transactions on Biomedical Engineering. 2012; 59:3450–3459. [PubMed: 22968202]

Liu, J.; Ji, S.; Ye, J. SLEP: Sparse Learning with Efficient Projections. Arizona State University; 2009. URL: http://www.public.asu.edu/~jye02/Software/SLEP

Liu M, Zhang D, Shen D. Ensemble sparse classification of Alzheimer's disease. NeuroImage. 2012; 60:1106–1116. [PubMed: 22270352]

Lu Z, Zhang Y. Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization. 2013; 23:2448–2478.

Lv J, Jiang X, Li X, Zhu D, Chen H, Zhang T, Zhang S, Hu X, Han J, Huang H, et al. Sparse representation of whole-brain fMRI signals for identification of functional networks. Medical Image Analysis. 2015; 20:112–134. [PubMed: 25476415]

Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Briefings in Bioinformatics. 2008; 9:392–403. [PubMed: 18562478]

Margeta J, Geremia E, Criminisi A, Ayache N. Layered spatio-temporal forests for left ventricle segmentation from 4D cardiac MRI data. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. volume 7085 of Lecture Notes in Computer Science. 2012:109–119.

Marques, J.; Clemmensen, LKH.; Dam, E. 1st International Workshop on Sparsity Techniques in Medical Imaging (STMI). MICCAI 2012; 2012. Diagnosis and prognosis of ostheoarthritis by texture analysis using sparse linear models.

McDonald JH. Handbook of biological statistics. volume 2. Sparky House Publishing Baltimore. 2009

McLeod, K.; Mansi, T.; Sermesant, M.; Pongiglione, G.; Pennec, X. Modeling in Computational Biology and Biomedicine. Springer; 2013. Statistical shape analysis of surfaces in medical images applied to the tetralogy of Fallot heart; p. 165-191.

Meier L, Van De Geer S, Bühlmann P. The group Lasso for logistic regression. Journal of the Royal Society. Series B. 2008; 70:53–71.

Ng B, Vahdat A, Hamarneh G, Abugharbieh R. Generalized sparse classifiers for decoding cognitive states in fMRI. Machine Learning in Medical Imaging. volume 6357 of Lecture Notes in Computer Science. 2010:108–115.

Osman NF, Kerwin WS, McVeigh ER, Prince JL. Cardiac motion tracking using CINE harmonic phase (HARP) magnetic resonance imaging. Magnetic Resonance in Medicine. 1999; 42:1048–1060. [PubMed: 10571926]

Qian Z, Liu Q, Metaxas DN, Axel L. Identifying regional cardiac abnormalities from myocardial strains using nontracking-based strain estimation and spatio-temporal tensor analysis. IEEE Transactions on Medical Imaging. 2011; 30:2017–2029. [PubMed: 21606022]

Qu Y, Adam B.l. Thornquist M, Potter JD, Thompson ML, Yasui Y, Davis J, Schellhammer PF, Cazares L, Clements M, Wright GL, Feng Z. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. Biometrics. 2003; 59:143–151. [PubMed: 12762451]

Rao A, Lee Y, Gass A, Monsch A. Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. 2011:4499–4502.

Rokach L. Ensemble-based classifiers. Artificial Intelligence Review. 2010; 33:1–39.

Ryali S, Supekar K, Abrams DA, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. NeuroImage. 2010; 51:752–764. [PubMed: 20188193]

Sabuncu M. A universal and efficient method to compute maps from image-based prediction models. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. volume 8675 of Lecture Notes in Computer Science. 2014:353–360.

Schellen C, Ernst S, Gruber GM, Mlczoch E, Weber M, Brugger PC, Ulm B, Langs G, Salzer-Muhar U, Prayer D, Kasprian G. Fetal MRI detects early alterations of brain development in Tetralogy of Fallot. American Journal of Obstetrics and Gynecology. 2015; 213:392.e1–392.e7. [PubMed: 26008177]

Serag A, Gousias I, Makropoulos A, Aljabar P, Hajnal J, Boardman J, Counsell S, Rueckert D. Unsupervised learning of shape complexity: Application to brain development. Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data. volume 7570 of Lecture Notes in Computer Science. 2012:88–99.

Sermesant M, Forest C, Pennec X, Delingette H, Ayache N. Deformable biomechanical models: Application to 4D cardiac image analysis. Medical Image Analysis. 2003; 7:475–488. [PubMed: 14561552]

Sundar H, Litt H, Shen D. Estimating myocardial motion by 4D image warping. Pattern Recognition. 2009; 42:2514–2526. [PubMed: 20379351]

Toews M, Wells, William M, I. Zöllei L. A feature-based developmental model of the infant brain in structural MRI. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. volume 7511 of Lecture Notes in Computer Science. 2012:204–211.

Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, Montana G. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage. 2012; 60:700–716. [PubMed: 22209813]

Wald RM, Haber I, Wald R, Valente AM, Powell AJ, Geva T. Effects of regional dysfunction and late gadolinium enhancement on global right ventricular function and exercise capacity in patients with repaired tetralogy of Fallot. Circulation. 2009; 119:1370–1377. [PubMed: 19255342]

Wang H, Amini A, et al. Cardiac motion and deformation recovery from MRI: a review. IEEE Transactions on Medical Imaging. 2012; 31:487–503. [PubMed: 21997253]

Wu, F.; Yuan, Y.; Zhuang, Y. Proceedings of the International Conference on Multimedia. ACM; New York, NY, USA: 2010. Heterogeneous feature selection by group Lasso with logistic regression; p. 983-986.

Yamashita O, Sato M, Yoshioka T, Tong F, Kamitani Y. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage. 2008; 42:1414–1429. [PubMed: 18598768]

Ye DH, Desjardins B, Hamm J, Litt H, Pohl KM. Regional manifold learning for disease classification. IEEE Transactions on Medical Imaging. 2014; 33:1236–1247. [PubMed: 24893254]

Yu Y, Zhang S, Li K, Metaxas D, Axel L. Deformable models with sparsity constraints for cardiac motion analysis. Medical Image Analysis. 2014; 18:927–937. [PubMed: 24721617]

Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage. 2012; 59:895–907. [PubMed: 21992749]

Zhang H, Wahle A, Johnson RK, Scholz TD, Sonka M. 4-D cardiac MR image analysis: left and right ventricular morphology and function. IEEE Transactions on Medical Imaging. 2010a; 29:350–364. [PubMed: 19709962]

Zhang J, Peng Q, Li Q, Jahanshad N, Hou Z, Jiang M, Masuda N, Lang-behn DR, Miller MI, Mori S, et al. Longitudinal characterization of brain atrophy of a huntington's disease mouse model by automated morphological analyses of magnetic resonance images. NeuroImage. 2010b; 49:2340–2351. [PubMed: 19850133]

Zhang S, Zhan Y, Metaxas DN. Deformable segmentation via sparse representation and dictionary learning. Medical Image Analysis. 2012; 16:1385–1396. [PubMed: 22959839]

Zhang Y, Pohl KM. Solving logistic regression with group cardinality constraints for time series analysis. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. volume 9351 of Lecture Notes in Computer Science. 2015:459–466.

Highlights

- Model concurrent disease classification and temporal-consistent pattern selection
- Minimize model by directly solving logistic regression confined by group cardinality
- Correctly identify ROIs differentiating the cine MRs of 44 TOF from 38 controls
- Generally significantly more accurate than approaches relaxing group sparsity

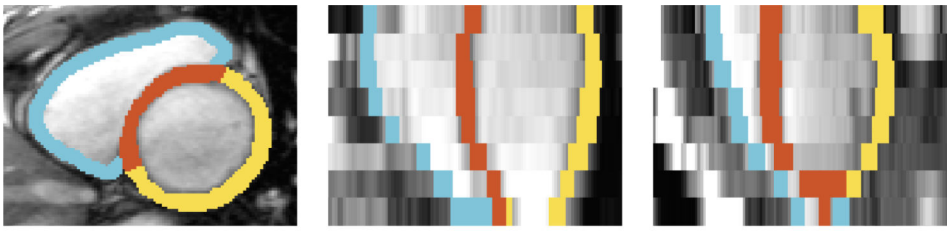


Figure 1.
Example segmentation of the RV (blue), VS (red), and LV (yellow)

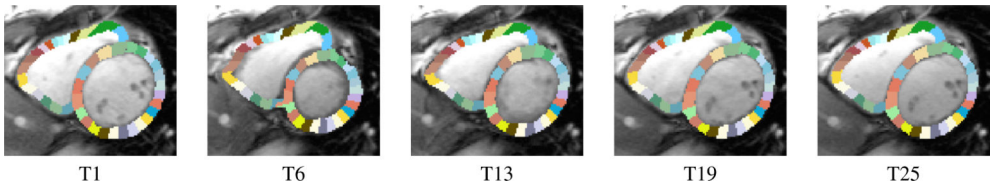


Figure 2.

An example slice of the partitioning of the RV and LV into 18 sections and the VS into six sections at time point 1, 6, 13, 19, and 25.

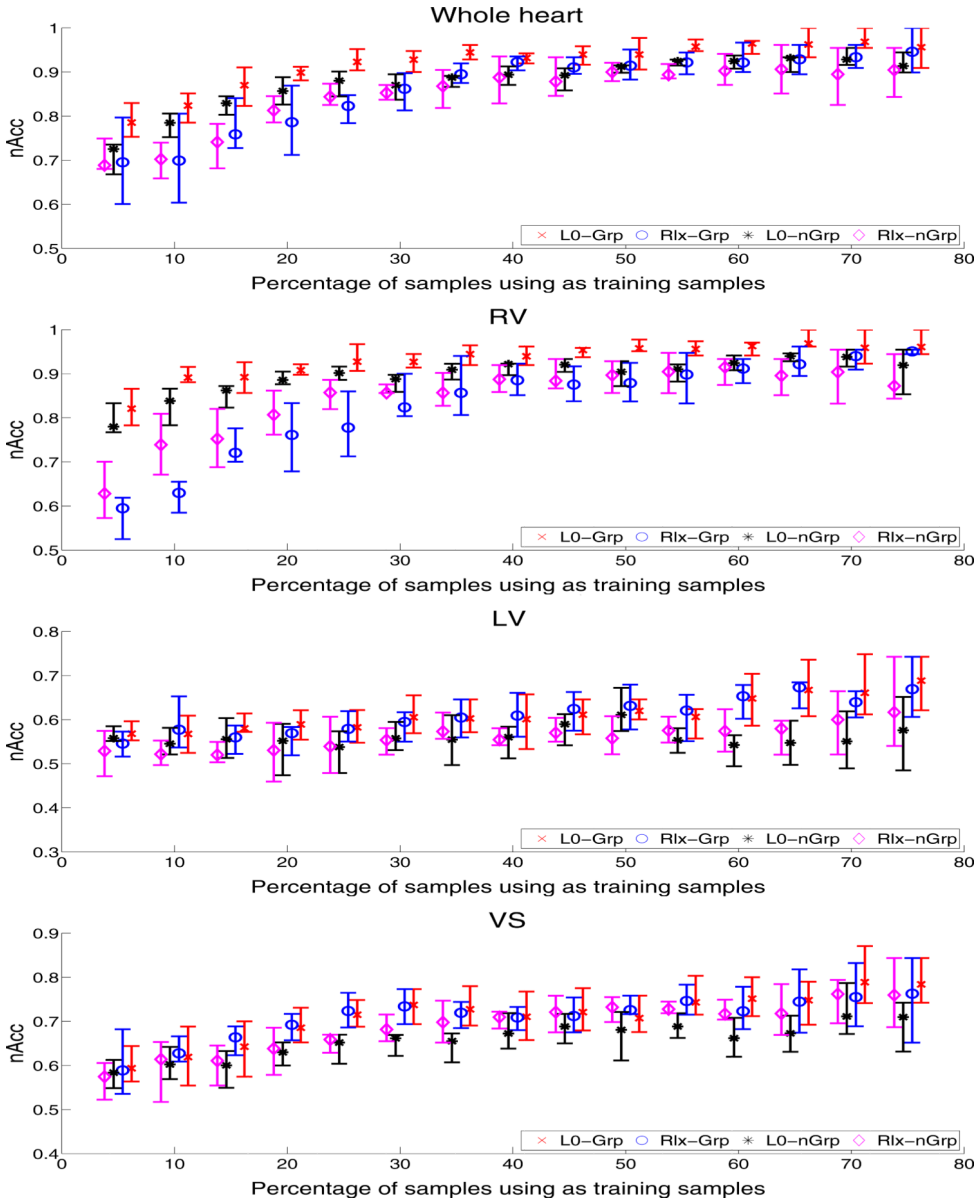


Figure 3.

The boxplots of the average, first quintile, and fourth quintile nAcc scores for all four ensembles with respect to the percentage of the whole data set used for training and the encoding of the LV, RV, VS and the whole heart (RV,LV&VS). For the RV and whole heart, the proposed L0-Grp implementation (red box plots) generally achieves a higher average, first quintile, and fourth quintile scores than then other three approaches. With respect to the LV and VS, all four methods perform similarly with the average nAcc scores starting at around 55% and generally increasing with the number of training samples.

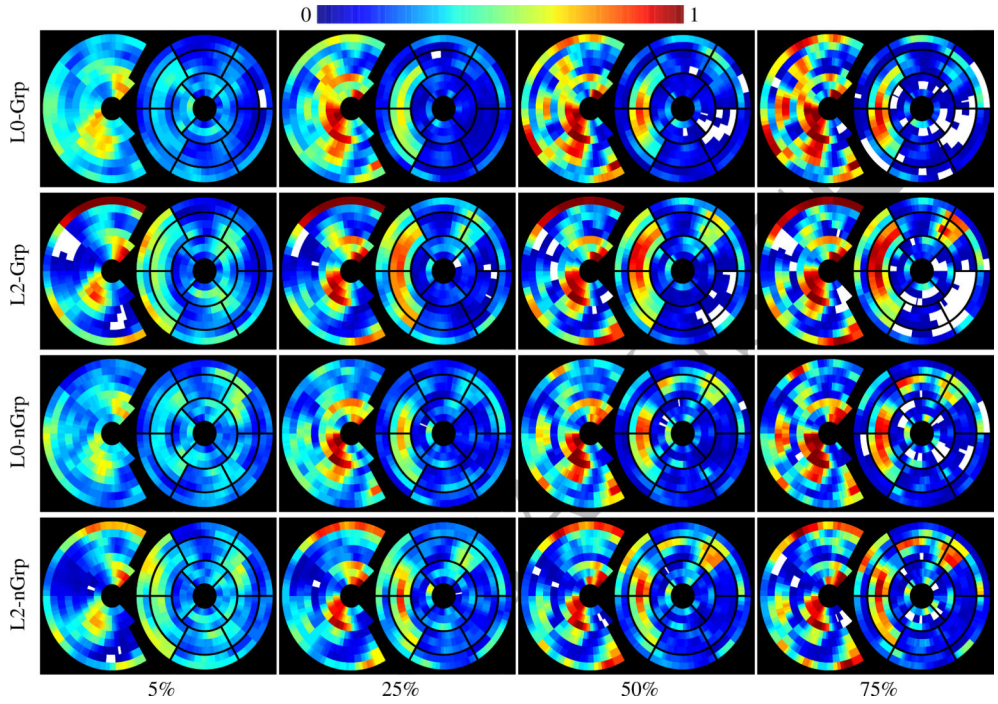


Figure 4.

Importance of heart sections in distinguishing TOF from healthy controls with respect to the type of solver and percentage associated with the training sample size. The incomplete circle on the left represents the importance of sections of the RV and on the right the importance of sections of the LV and VS. Each ring of those (incomplete) circles represents a slice of the cine MRI. As it is common in the cardiac literature, we overlay the bullseye plot over the LV & VS maps. For each type and percentage, the importance of a section is inferred from its average importance across the corresponding ensembles of classifiers, which is based on the number of times a section was selected by a sparse solver. The number of white regions (never selected), blue (weight close to 0) and red (weight close to 1) is generally increasing with the number of training samples indicating that the confidence of the ensemble increases in the selection of the sections. Furthermore, all methods correctly emphasize more sections of the RV than the LV&VS. L0-Grp ignores the LV sections the most of all solvers. This could explain its significantly higher accuracy in most experiments of the whole heart compared to those other three methods.

Table 1

Significant: $p \leq 0.05$; Trend: $0.05 < p \leq 0.1$; Indifferent: $p > 0.1$. Frequency of p-value of the paired-sample t-test between L0-Grp and the other three methods. In most tests, the results obtained by L0-Grp are significantly more accurate than the other three methods with respect to the paired-sample t-test (McDonald, 2009).

	RV			RV, LV & VS		
	Significant	**Trend**	**Indifferent**	**Significant**	**Trend**	**Indifferent**
Rlx-Grp	13	0	2	10	2	3
L0-nGrp	8	4	3	13	2	0
Rlx-nGrp	15	0	0	13	2	0

Algorithm 1

Penalty Decomposition (PD) Applied to Eq. (3)

Initialization: Choose a sparsity parameter $r \in \mathbb{N}$, scalar weight $v_i \in \mathbb{R}$ and a feasible sparsity constrained weights $W_i \in \mathcal{X}$ where

$$\mathcal{X} := \left\{ W \in \mathbb{R}^{M \times T} : \left\| \widetilde{W} \right\|_0 \leq r \right\}.$$

Furthermore, set the following parameters

- $W_0 \leftarrow W_i$ (initialize weight)
- $\varrho_0 > 0$ (initial penalty)
- $\sigma > 1$ (penalty updating factor)
- $\epsilon_{BCD} > 0$ (upper bound for convergence of BCD)
- $\epsilon_{PD} > 0$ (upper bound for convergence of PD)
- $p \leftarrow 0$ (PD index)
- $\Gamma = l_{\text{avg}}(v_i, W_i)$ (upper bound for $q_{\varrho_p}(v, Y, W)$)

repeat (PD Loop)

% Step 1: Define the penalty function

$$q_{\varrho_p}(v, Y, W) := l_{\text{avg}}(v, Y) + \frac{\varrho_p}{2} \left\| W - Y \right\|_F^2$$

% Step 2: Determine the local minimum point of q_{ϱ_p} via BCD

$b \leftarrow 0$ (BCD index)

repeat (BCD Loop)

$b \leftarrow b + 1$

% Solve the following via Gradient Descent

$$\left(v_b, Y_b \right) \leftarrow \underset{v \in \mathbb{R},\, Y \in \mathbb{R}^{M \times T}}{\arg\min} \left\{ l_{\text{avg}}(v, Y) + \frac{\varrho_p}{2} \left\| W_{b-1} - Y \right\|_F^2 \right\}$$

$$W_b^j = \begin{cases} Y_b^j, & \text{if } j \leq r; \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } l = 1, \ldots, M.$$

until $\max \left\{ \dfrac{| v_b - v_{b-1} |}{\max\left(| v_b |, 1 \right)}, \dfrac{\left\| Y_b - Y_{b-1} \right\|_{\max}}{\max\left(\left\| Y_b \right\|_{\max}, 1 \right)}, \dfrac{\left\| W_b - W_{b-1} \right\|_{\max}}{\max\left(\left\| W_b \right\|_{\max}, 1 \right)} \right\} \leq \epsilon_{BCD}$

% Step 3: Update results, penalty parameter, and check the stopping criterion

$\varrho_{p+1} \leftarrow \sigma \cdot \varrho_p$ (increase penalty parameter)

if $\min_{v, Y} q_{\varrho_{p+1}}(v, Y, W_b) \leq \Gamma$ **then**

$W_0 \leftarrow W_b$

else

$W_0 \leftarrow W_i$

$p \leftarrow p + 1$, $\hat{v}_p \leftarrow v_b$, $\hat{Y}_p \leftarrow Y_b$ and $\widehat{W}_p \leftarrow W_0$

until $\left\| \widehat{W}_p - \hat{Y}_p \right\|_{\max} \leq \epsilon_{PD}$