Clinical Risk Prediction with Multilinear Sparse Logistic Regression

Fei Wang[1], Ping Zhang[1], Buyue Qian[1], Xiang Wang[1], Ian Davidson[2]
[1]IBM T. J. Watson Research Center. Yorktown Heights, NY
[2]Department of Computer Science. University of California, Davis.
{fwang,pzhang,bqian,wangxi}@us.ibm.com, davidson@cs.ucdavis.edu

ABSTRACT

Logistic regression is one core predictive modeling technique that has been used extensively in health and biomedical problems. Recently a lot of research has been focusing on enforcing sparsity on the learned model to enhance its effectiveness and interpretability, which results in sparse logistic regression model. However, no matter the original or sparse logistic regression, they require the inputs to be in vector form. This limits the applicability of logistic regression in the problems when the data cannot be naturally represented vectors (e.g., functional magnetic resonance imaging and electroencephalography signals). To handle the cases when the data are in the form of multi-dimensional arrays, we propose *MulSLR*: Multilinear Sparse Logistic Regression. *MulSLR* can be viewed as a high order extension of sparse logistic regression. Instead of solving one classification vector as in conventional logistic regression, we solve for K classification vectors in *MulSLR* (K is the number of modes in the data). We propose a block proximal descent approach to solve the problem and prove its convergence. The convergence rate of the proposed algorithm is also analyzed. Finally we validate the efficiency and effectiveness of *MulSLR* on predicting the onset risk of patients with Alzheimer's disease and heart failure.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; G.3 [**Probability and Statistics**]: Correlation and Regression Analysis

Keywords

Logistic Regression; Multilinear; Proximal Gradient; Healthcare

1. INTRODUCTION

Clinical risk prediction, such as predicting the onset [27] or hospitalization [19] risk of patients with chronic diseases, is an important problem in health informatics. Accurate risk prediction can greatly help reduce the unnecessary costs in hospitals as well as provide the right service at point-of-care.

There has been quite a few existing works in both data mining and health informatics domains on clinical risk prediction. For example, Sun *et al.* [27] developed a LASSO type of method for identifying important risk factors for predicting the onset risk of heart failure patients. Xiang *et al.* [30] proposed a multi-source learning approach for predicting the risk measured by cognitive score of patients with Alzheimer's disease. Miravitlles *et al.* [18] analyzed the factors associated with increased risk of exacerbation and hospital admission for patients with Chronic Obstructive Pulmonary Disease (COPD). In most of those works, logistic regression is at the heart of the predictive modeling process. Because of the large number of potentially related factors in different scenarios, a sparsity constraint is usually added on the learned model coefficients. The resultant model is referred to as sparse logistic regression, which can do both prediction and feature selection simultaneously. Depending on the different sparsity structures the model wants to explore, we can construct different sparsity-induced regularization terms. By adding them to the objective of conventional logistic regression we can get different types of sparse logistic regression models [13][17][26].

Until now most of the existing sparse logistic regression type of approaches assume their inputs are a set of data vectors. This means that we need to have a vector based representation for each patient if we want to adopt logistic regression type of methods to evaluate their risk. However, many patient medical data are not naturally in vector form. For example, X-Ray images are two-dimensional; Electroencephalography (EEG) is two dimensional if you stack all signals captured from different poles; functional Magnetic Resonance Imaging (fMRI) is three-dimensional. Even in patient Electronic Health Records (EHRs), there could be multiple diagnosis/symptoms accompanied with several drugs on the same claim. Therefore it is natural to represent a patient with a diagnosis by drug co-occurrence matrix if we want to consider the correlation between diagnosis and drugs when predicting the patient risk. We can also represent the patients with high order tensors if we want to consider more than two factors that are inter-correlated with each other. In these cases, if we still want to apply logistic regression one straightforward way is to stretch those matrices and tensors into vectors as people did in image processing, but this will lose the correlation information among different dimensions. Moreover, after stretching the dimensionality of

KDD'14, August 24–27, 2014, New York, New York, USA.
Copyright © 2014 ACM 978-1-4503-2956-9/14/08…$15.00.
http://dx.doi.org/10.1145/2623330.2623754 .

the data objects will become very high, which will make traditional logistic regression inefficient.

Based on the above considerations, many researches on extending traditional vector based approaches to two (matrix based) or high order (tensor based) approaches have been becoming more and more popular. For example, two-dimensional Principal Component Analysis (PCA) [32] and Linear Discriminant Analysis [33] have been found to be more effective on face recognition task compared to traditional vector based PCA and LDA. Cai *et al.* [4] also extend Support Vector Machine to multidimensional data and got better results on document classification. Recently, Huang and Wang [9] developed a matrix variate logistic regression model and applied it in electroencephalography data analysis. Tan *et al.* [28] further extended logistic regression to tensor inputs and achieved good performance in a video classification task.

In this paper, we propose *MulSLR*, a multi-linear sparse logistic regression method that can directly take matrices or tensors as inputs and perform prediction. Because of the added L1 sparsity regularization terms, we developed a *Block Proximal Gradient* (BPG) method to solve the problem iteratively. We theoretically prove the convergence of the proposed algorithm and analyze the convergence rate based on the Kurdyka–Lojasiewicz inequality [3]. Finally we validate the effectiveness of our algorithm on both synthetic and real world data sets.

It is worthwhile to highlight the following aspects of *MulSLR*.

- *MulSLR* is able to directly take matrices and tensors as inputs, thus the resultant model can encode the correlation structure of the different types of data features.
- *MulSLR* does not need to stretch the data into vectors for model training, thus it avoids the curse of dimensionality and as a result the model can be more efficiently trained.
- The BPG method we developed to train *MulSLR* is theoretically guaranteed to converge. We also analyze the converge rate of *MulSLR*.
- We apply *MulSLR* to evaluate the onset risk of Alzheimer's disease and congestive heart failure on real world data sets and show some interesting results.

The rest of this paper is organized as follows. Section 2 reviews some related works. The details along with the convergence analysis of *MulSLR* is introduced in Section 3. Section 4 presents the experimental results, followed by the conclusions in Section 5.

2. RELATED WORK

Logistic regression [8] is a statistical classification method that has widely been used in many application areas, such as computer vision [23], information retrieval [6] and health informatics [18][30]. Suppose we have a training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ $(1 \leqslant i \leqslant n)$ is the i-th training data vector. Associated with each $\mathbf{x}_i$ we also have its corresponding label $y_i \in \{0, 1\}$. The goal of logistic regression is to train a linear classification function $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$ to discriminate the data in class 1 from the data in class 0 by minimizing the following logistic loss

$$\ell_{org}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top\mathbf{x}_i + b))\right] \qquad (1)$$

where $\mathbf{w} \in \mathbb{R}^d$ is the classification vector and b is the bias. They can be learned with gradient descent type of approaches.

In many real world applications, the data vectors $\{\mathbf{x}_i\}_{i=1}^n$ are usually sparse and high-dimensional (e.g., each patient could be a tens of thousands dimensional vector with bag-of-feature representation [27]). To enhance the interpretability of the model in these scenarios, we can add a sparsity regularization term on $\mathbf{w}$ and minimize the following ℓ_1-regularized logistic loss

$$\ell_{sp}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top\mathbf{x}_i + b))\right] + \lambda\|\mathbf{w}\|_1 \qquad (2)$$

where $\|\cdot\|_1$ is the vector $\ell 1$ norm and $\lambda > 0$ is a factor trading off the prediction accuracy and model sparsity. The resultant model is usually referred to as sparse logistic regression model [12]. Compared with the conventional logistic regression model obtained by minimizing $\mathcal{J}_{org}$, the $\mathbf{w}$ obtained by minimizing $\mathcal{J}_{sp}$ is sparse. In this way, we can not only get a predictor, but also know what are the feature dimensions that are important to the prediction, and these are the features with nonzero classification coefficients. Zou and Hastie [34] pointed out that there are some limitations if we only adopt 1 norm regularization, and they proposed a regularization term called *elastic net*, which is a mixture of 1 and 2 norm regularizers.

Sparse logistic regression has widely been used in health informatics because the applications in this field usually not only want a good prediction performance but also need the reason why. Basically what are the key factors that will affect the prediction performance. For example, sparse logistic regression has been used in the prediction of Leukemia [15], Alzheimer's disease [22] and cancers [10]. In recent years people also designed different regularization terms [13][17][26] to a enforce more complex sparsity patterns on the learned model. However, all these works require a vector based representation of the data. Fig.1 provides a graphical illustration on the difference of traditional vector based logistic regression and multilinear logistic regression when working on multi-dimensional data. Those traditional approaches need to stretch the data into an ultra-high dimensional vector first before they can be applied. This may suffer from the curse of dimensionality.

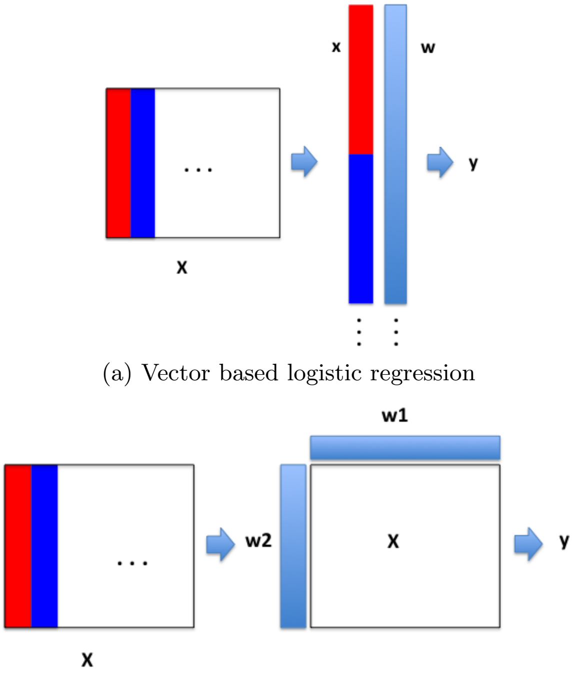
3. METHODOLOGY

We introduce the details of *MulSLR* in this section. First we will formally define the problem.

3.1 Problem Statement

Without the loss of generality, we assume each observation is a tensor $\mathcal{X}^i \in \mathbb{R}^{d_1\times d_2\times\cdots d_K}$, suppose its corresponding response is $y^i \in \{0, 1\}$, then *MulSLR* assumes

$$y^i \leftarrow \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \qquad (3)$$



(a) Vector based logistic regression

(b) Multilinear logistic regression

Figure 1: Traditional vector based logistic regression and multilinear logistic regression work on multi-dimensional data.

where $\times_k$ is the mode-k product, and $\mathbf{w}^k \in \mathbb{R}^{d_k \times 1}$ is the prediction coefficients on the k-th dimension. Then

$$\begin{aligned} & \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K \\ & = \sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_2}\cdots\sum_{i_K=1}^{d_K} w^1_{i_1} w^2_{i_2} \cdots w^K_{i_K} X^i_{i_1 i_2 \cdots i_K} \qquad (4) \end{aligned}$$

Let $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\}$ be the set of prediction coefficient vectors. The loss we want to minimize is

$$\begin{aligned} \ell(\mathcal{W}, b) &= \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{X}_i, y_i, \mathcal{W}) && (5) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_{(\mathcal{W},b)}(\mathcal{X}^i)) \end{aligned}$$

where for notational convince, we denote

$$f_{(\mathcal{W},b)}(\mathcal{X}^i) = \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \qquad (6)$$

The loss function we considered in this paper is *Logistic Loss*:

$$\ell_l(\mathcal{W}, b) = \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))] \qquad (7)$$

We adopt the following elastic net regularization term [34]

$$\begin{aligned} \mathcal{R}(\mathcal{W}) &= \mathcal{R}_1(\mathcal{W}) + \mathcal{R}_2(\mathcal{W}) && (8) \\ &= \sum_{k=1}^{K} \lambda_k \|\mathbf{w}^k\|_1 + \frac{1}{2}\sum_{k=1}^{K} \mu_k \|\mathbf{w}^k\|_2^2 && (9) \end{aligned}$$

Then the optimization problem we want to solve is

$$\min_{\mathcal{W}} \mathcal{J}(\mathcal{W}, b) = \ell(\mathcal{W}, b) + \mathcal{R}(\mathcal{W}) \qquad (10)$$

We adopt a *Block Coordinate Descent* (BCD) procedure to solve the problem. Starting from some initialization $(\mathcal{W}_{(0)}, b_{(0)})$, at the i-th step of the t-th round of updates, we update $(\mathbf{w}^k_{(t)}, b_{(t)})$ by

$$\begin{aligned} & (\mathbf{w}^k_{(t)}, b_{(t)}) && (11) \\ & = \arg\min_{(\mathbf{w},b)} \left[\ell(\mathcal{W}^{1\sim(k-1)}_{(t)}, \mathbf{w}, \mathcal{W}^{(k+1)\sim K}_{(t-1)}, b) + \lambda_k \|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\right] \end{aligned}$$

where

$$\begin{aligned} \mathcal{W}^{1\sim(k-1)}_{(t)} &= \left\{\mathbf{w}^1_{(t)}, \mathbf{w}^2_{(t)}, \cdots, \mathbf{w}^{k-1}_{(t)}\right\} \\ \mathcal{W}^{(k+1)\sim K}_{(t-1)} &= \left\{\mathbf{w}^{k+1}_{(t-1)}, \mathbf{w}^{k+2}_{(t-1)}, \cdots, \mathbf{w}^{K}_{(t-1)}\right\} \end{aligned}$$

Algorithm 1 Block Coordinate Descent Procedure

```
Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$
Initialization: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 0$
while Not Converge do
    for $k = 1 : K$ do
        Update $(\mathbf{w}^k_{(t)}, b_{(t,k)})$ by solving problem (11)
        $t = t + 1$
    end for
end while
```

3.2 Proximal Gradient Descent

The proximal descent procedure for updating $(\mathbf{w}^i_{(t)}, b_{(t,k)})$ is

$$\begin{aligned} & \mathbf{w}^k_{(t)} = && (12) \\ & \arg\min_{\mathbf{w}} \Big[(\mathbf{w} - \widetilde{\mathbf{w}}^k_{(t)})^\top \nabla_{\mathbf{w}^k} \ell(\mathcal{W}^{1\sim(k-1)}_{(t)}, \widetilde{\mathbf{w}}^k_{(t)} \mathcal{W}^{(k+1)\sim K}_{(t-1)}, b_{(t,k-1)}) \\ & + \frac{\tau^k_{(t)}}{2}\|\mathbf{w} - \widetilde{\mathbf{w}}^k_{(t)}\|_2^2 + \lambda_k \|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\Big] \end{aligned}$$

where

$$\begin{aligned} & \nabla_{\mathbf{w}^k} \ell(\mathcal{W}, b) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}^k} \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))] \\ & = \frac{1}{n}\sum_{i=1}^{n} -y_i \frac{\exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))}{1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))} \nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i) \\ & = -\frac{1}{n}\sum_{i=1}^{n} \left[1 + \exp(y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))\right]^{-1} y_i \nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i) \quad (13) \end{aligned}$$

and

$$\begin{aligned} & \nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i) && (14) \\ & = \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_{(k-1)} \mathbf{w}^{(k-1)} \times_{(k+1)} \mathbf{w}^{(k+1)} \cdots \times_K \mathbf{w}^K \end{aligned}$$

$\widetilde{\mathbf{w}}^k_{(t)}$ is an extrapolated point defined as

$$\widetilde{\mathbf{w}}^k_{(t)} = \mathbf{w}^k_{(t-1)} + \omega^k_{(t)}(\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t-2)}) \qquad (15)$$

The optimal solution to problem (12) can be obtained as

$$\mathbf{w}^k_{(t)} = \mathcal{S}_{\alpha^k_{(t)}}\left(\frac{\tau^k_{(t)}\widetilde{\mathbf{w}}^k_{(t)} - \nabla_{\mathbf{w}^k}\ell^k_{(t)}(\widetilde{\mathbf{w}}^k_{(t)}, b_{(t,k-1)})}{\tau^k_{(t)} + \mu_k}\right) \qquad (16)$$

where $\alpha^k_{(t)} = \lambda_k/(\tau^k_{(t)} + \mu_k)$ and $\mathcal{S}_{\alpha^k_{(t)}}$ is the component-wise shrinkage operator defined as

$$\left(\mathcal{S}_{\alpha^k_{(t)}}(\mathbf{v})\right)_i = \begin{cases} v_i - \alpha^k_{(t)}, & \text{if } v_i > \alpha^k_{(t)} \\ v_i + \alpha^k_{(t)}, & \text{if } v_i < -\alpha^k_{(t)} \\ 0, & \text{if } |v_i| \leqslant |\alpha^k_{(t)}| \end{cases} \qquad (17)$$

Similarly we can update $b_{(t,k)}$ as

$$\begin{aligned} b_{(t,k)} \quad = \quad & \arg\min_b \Big[(b - \widetilde{b}_{(t,k)})^\top \nabla_b \ell(\mathcal{W}^{1\sim k}_{(t)}, \mathcal{W}^{(k+1)\sim K}_{(t-1)}, \widetilde{b}_{(t,k)}) \\ & + \frac{\tau^k_{(t)}}{2}|b - \widetilde{b}_{(t,k)}|_2^2\Big] \end{aligned} \qquad (18)$$

and

$$\begin{aligned} \nabla_b \ell(\mathcal{W}, b) &= \frac{1}{n}\sum_{i=1}^n \nabla_b \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))] \\ &= -\frac{1}{n}\sum_{i=1}^n \Big[1 + \exp(y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))\Big]^{-1} y_i \end{aligned} \qquad (19)$$

$\widetilde{b}_{(t,k)}$ is the extrapolated point defined as

$$\widetilde{b}_{(t,k)} = b_{(t,k-1)} + \omega^k_{(t)}(b_{(t,k-1)} - b_{(t,k-2)}) \qquad (20)$$

The optimal solution of problem (18) can be obtained by simple gradient descent as

$$b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau^k_{(t)}}\nabla_b \ell^k_{(t)}(\mathbf{w}^k_{(t)}, \widetilde{b}_{(t,k)}) \qquad (21)$$

Putting Eq.(14) and Eq.(19) together, we have the following theorem

THEOREM 3.1. *The partial gradient $\nabla_{(\mathbf{w}^k,b)}\ell(\mathcal{W}, b)$ is Lipschitz continuous with constant*

$$\tau^k = \frac{\sqrt{2}}{n}\sum_{i=1}^n \Big(\|\nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i)\|_2 + 1\Big)^2 \qquad (22)$$

PROOF. See Appendix I. □

$\tau^k_{(t)}$ in Eq.(16) and Eq.(21) can be set as the Lipschitz constant according to Theorem 3.1. For notational convenience, we first introduce

$$\begin{aligned} & \nabla^{(t,k)}_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i) \\ & = \mathcal{X}^i \times_1 \mathbf{w}^1_{(t)} \times_2 \mathbf{w}^2_{(t)} \cdots \times_{(k-1)} \mathbf{w}^{(k-1)}_{(t)} \\ & \quad \times_{(k+1)} \mathbf{w}^{(k+1)}_{(t-1)} \cdots \times_K \mathbf{w}^K_{(t-1)} \end{aligned} \qquad (23)$$

Then $\tau^k_{(t)}$ can be set as

$$\tau^k_{(t)} = \frac{\sqrt{2}}{n}\sum_{i=1}^n \Big(\Big\|\nabla^{(t,k)}_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i)\Big\|_2 + 1\Big)^2 \qquad (24)$$

$\omega^k_{(t)}$ can be set as

$$\omega^k_{(t)} = \min\left(\omega_{(t)}, \delta_\omega\sqrt{\frac{\tau^k_{(t-1)}}{\tau^k_{(t)}}}\right) \qquad (25)$$

where $\omega_{(t)} = (\eta_{(t-1)} - 1)/\eta_{(t)}$ with $\eta_{(0)} = 1$ and $\eta_{(t)} = \frac{1}{2}\left(1 + \sqrt{1 + 4\eta^2_{(t-1)}}\right)$.

Algorithm 2 summarized the whole algorithmic flow of our algorithm. At each iteration the most time consuming part is evaluating the gradient, which takes $O(n\prod_{i=1}^K d_i)$ time, that is linear with respect to data set size and data dimension.

Algorithm 2 Block Proximal Gradient Descent for Multilinear Sparse Logistic Regression

```
Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$, $r_0 = 1$, $\delta_\omega < 1$
Initialization: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 1$
while Not Converge do
    for $k = 1 : K$ do
        Compute $\tau^k_{(t)}, \omega^k_{(t)}$ with Eq.(24) and Eq.(25)
        Compute $\widetilde{\mathbf{w}}^k_{(t)}, \mathbf{w}^k_{(t)}$ with Eq.(15) and Eq.(16)
        Update $\widetilde{b}_{(t,k)}, b_{(t,k)}$ by and Eq.(20) and Eq.(21)
    end for
    if $\ell(\mathcal{W}_{(t-1)}, b_{(t-1,K)}) \leqslant \ell(\mathcal{W}_{(t)}, b_{(t,K)})$ then
        Reupdate $\mathbf{w}^k_{(t)}$ and $b_{(t,k)}$ using Eq.(16) and Eq.(21),
        with $\widetilde{\mathbf{w}}^k_{(t)} = \mathbf{w}^k_{(t-1)}$ and $\widetilde{b}_{(t,k)} = b_{(t,k-1)}$
    end if
    $t = t + 1$
end while
```

3.3 Convergence Analysis

THEOREM 3.2. *Let $\mathcal{W}_{(t)}$ be the sequence generated by Algorithm 1 with $0 \leqslant \omega^k_{(t)} \leqslant \delta_\omega\sqrt{\tau^k_{(t-1)}/\tau^k_{(t)}}$ for $\delta_\omega < 1$. Then when $t \to \infty$, the sequence of $(\mathcal{W}_{(t)}, b_{(t,k)})$ will converge to some point $(\bar{\mathcal{W}}, \bar{b})$.*

PROOF. See Appendix II. □

With Theorem 3.2, it is not hard to see that $(\bar{\mathcal{W}}, \bar{b})$ is also a stationary point. This is because when $\mathcal{W}_{(t)} \to \bar{\mathcal{W}}$, $b_{(t,k)} \to \bar{b}$, according to Eq.(24), $\tau^k_{(t)} \to \bar{\tau}^k$. Therefore

$$\begin{aligned} & \bar{\mathbf{w}}^k = \\ & \arg\min_{\mathbf{w}} \Big[(\mathbf{w} - \bar{\mathbf{w}}^k)^\top \nabla_{\mathbf{w}^k}\ell(\bar{\mathcal{W}}^{1\sim(k-1)}, \bar{\mathbf{w}}^k\bar{\mathcal{W}}^{(k+1)\sim K}, \bar{b}) \\ & + \frac{\bar{\tau}^k}{2}\|\mathbf{w} - \bar{\mathbf{w}}^k\|_2^2 + \lambda_k\|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\Big] \end{aligned} \qquad (26)$$

Let $r(\mathbf{w}^k) = \lambda_k\|\mathbf{w}^k\|_1 + \frac{\mu_k}{2}\|\mathbf{w}^k\|_2^2$, then Eq.(26) suggests $\mathbf{0} \in \nabla_{\mathbf{w}^k}\ell(\bar{\mathcal{W}}, \bar{b}) + \partial r(\mathbf{w}^k)$ $(1 \leqslant k \leqslant K)$. Similarly, we also have $\nabla_b\ell(\bar{\mathcal{W}}, \bar{b}) = 0$. Therefore $(\bar{\mathcal{W}}, \bar{b})$ is a stationary point. To establish the convergence rate estimation result, we first introduce the following K-L inequality.

DEFINITION 3.3. **(Kurdyka-Lojasiewicz (K-L) Inequality)** *[11] [14] A function f is said to satisfy the Kurdyka-Lojasiewica inequality at point $\bar{\mathbf{\Psi}}$, if there exists $\theta \in [0, 1)$ such that*

$$\frac{|f(\mathbf{\Psi}) - f(\bar{\mathbf{\Psi}})|^\theta}{dist(\mathbf{0}, \partial f(\mathbf{\Psi}))} \qquad (27)$$

is bounded for any $\mathbf{\Psi}$ near $\bar{\mathbf{\Psi}}$, where $\partial f(\mathbf{\Psi})$ is the limiting subdifferential of f at $\mathbf{\Psi}$ [24], and

$$dist(\mathbf{0}, \partial f(\mathbf{\Psi})) \equiv \min\{\|\mathbf{\Phi}\|_F : \mathbf{\Phi} \in \partial f(\mathbf{\Psi})\} \qquad (28)$$

The K-L inequality was first introduced by Lojasiewicz [14] on real analytic functions, for which Eq.(27) is bounded around any stationary point $\bar{\mathbf{\Psi}}$ for $\theta \in [\frac{1}{2}, 1)$. Kurdyka later extended this property to the functions on the o-minimal structure, and recenttly Bolte *et al.* [3] further extended it to non smooth sub analytic functions, and loss function $\mathcal{J}(\mathcal{W}, b)$ is one function of such type. Then it satisfies the K-L inequality. According to [31], we have the following theorem stating the convergence rate of our BPG method.

THEOREM 3.4. *Let $(\bar{\mathcal{W}}, \bar{b})$ be a stationary point of the $(\mathcal{W}_{(t)}, b_{(t)})$ sequence, then depending on the θ in Eq.(27), we have the following convergence rate.*

- *If $\theta = 0$, then $(\mathcal{W}_{(t)}, b_{(t)})$ converges to $(\bar{\mathcal{W}}, \bar{b})$ in finite iterations.*
- *If $\theta \in (0, \frac{1}{2}]$, then $(\mathcal{W}_{(t)}, b_{(t)})$ converges to $(\bar{\mathcal{W}}, \bar{b})$ at least linearly, i.e., $\|(\mathcal{W}_{(t)}, b_{(t)}) - (\bar{\mathcal{W}}, \bar{b})\|_F \leqslant Cr^t$ for some constant C and $r < 1$.*
- *If $\theta \in (\frac{1}{2}, 1)$, then $(\mathcal{W}_{(t)}, b_{(t)})$ converges to $(\bar{\mathcal{W}}, \bar{b})$ at least sublinearly, i.e., $\|(\mathcal{W}_{(t)}, b_{(t)}) - (\bar{\mathcal{W}}, \bar{b})\|_F \leqslant Ct^{-\frac{1-\theta}{2\theta-1}}$ for some constant C.*

PROOF. The proof can easily be derived from the proof of Theorem 2.9 in [31], thus we neglect the details here. □

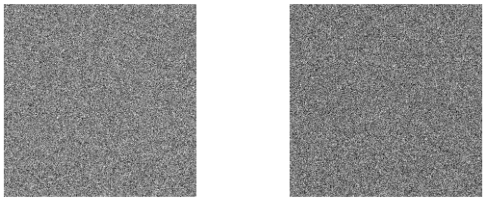
4. EXPERIMENTS

In this section we will present the results of a set of experiments we conducted to test the effectiveness and efficiency of our method, including both synthetic examples and real world examples.

4.1 Synthetic Examples

We constructed some synthetic data sets to investigate two types of questions:

1. Whether *MulSLR* can effectively discover the latent data structure or not?
2. What is the scalability behavior of *MulSLR* when it is implemented on data sets with different scales?



(a) Sample from class 1 (b) Sample from class 0

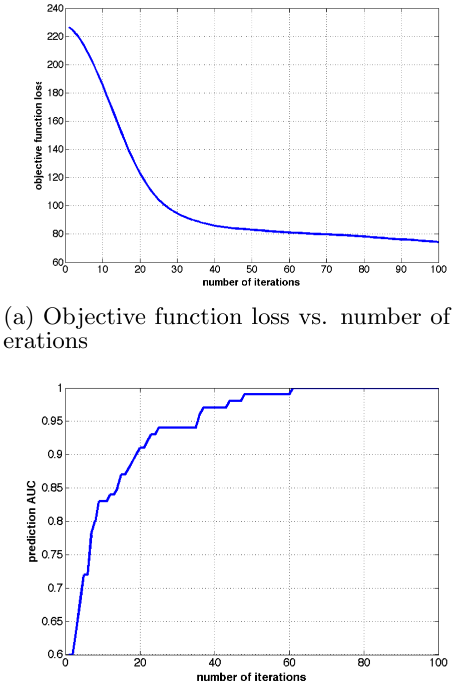
Figure 2: Two samples from the synthetic data set we generated. (a) is from class 1, (b) is from class 0. The intensities of the pixels indicates the values of the corresponding entries, where dark means small values and bright means large values.

In order to answer question 1, we constructed a data set with two different patterns. Each data object is a square matrix with size 100 × 100, where the elements in the data matrices are generated independently from $\mathcal{N}(0, 1)$, i.e., univariate Gaussian distribution with zero mean and unit variance. The upper-left 20 × 20 block was different for the data matrices in class 1 and class 0 in the following sense. We generate two vectors $\mathbf{w}_1 \in \mathbb{R}^{20}$ and $\mathbf{w}_2 \in \mathbb{R}^{20}$ whose elements are generated independently from uniform distribution between 0 and 1. For any data matrix $\mathbf{X}$ from class 1, we have $\mathbf{w}_1^\top \widehat{\mathbf{X}} \mathbf{w}_2 + 1 \geqslant 0.5$, where $\widehat{\mathbf{X}}$ is the upper-left 20×20 block of $\mathbf{X}$. For any data matrix $\mathbf{Y}$ from class 2, we have $\mathbf{w}_1^\top \widehat{\mathbf{X}} \mathbf{w}_2 + 1 \leqslant -0.5$, where $\widehat{\mathbf{Y}}$ is the upper-left 20×20 block of $\mathbf{Y}$. Therefore there is a special correlation structure on the two dimensions of those data matrices. Basically the data from those two classes can only be identified from a bilinear combination on their upper-left 20×20 blocks. We provide two sample data matrices on Fig.2, one from each class. From the figure we cannot judge whether there is any differences between them. We generated 1000 samples for each class and thus the entire data set has 2000 samples.

In our implementation, we initialize $\mathbf{w}_1 \in \mathbb{R}^{100}, \mathbf{w}_2 \in \mathbb{R}^{100}$ as uniform vectors, and we iterate the *MulSLR* until a certain termination condition is satisfied. Such termination condition could be either a maximum number of iteration steps or the absolute difference of objective function value between two consecutive steps is less than a certain tolerance value. For those free parameters we set $\lambda_1 = \lambda_2 = 0.01$, $\mu_1 = \mu_2 = 0.0001$, $\delta_\omega = 0.99$. We set the maximum number of iterations to be 100. We randomly select 80% of the data for training (the data in class 1 and 0 are evenly sampled), and the rest 20% data for testing. The objective function value convergence plot is shown in Fig.3 (a), from which we can see that with the iterations going on, the objective function value decreases very fast during the first 30 steps, and decreases slowly from 30 to 60 steps, and becomes almost stable from then on. We also evaluated the prediction performance on the testing data set in terms of *Area Under the receiver operating characteristic Curve* (AUC) value using $\mathbf{w}_1$ and $\mathbf{w}_2$ obtained from each iteration. The prediction AUC versus number of iterations plot is illustrated in Fig.3 (b), which shows that the prediction performance also increases very fast during the first 30 steps, and slowly increase to 1 from step 30 to step 60.

As illustrations, we also plotted the matrix of $\mathbf{W} = \mathbf{w}_1 \mathbf{w}_2^\top$ after 100 iterations with *MulSLR* in Fig.4 (b). This is interesting because *MulSLR* makes decisions with a linear function, and in two dimension case the weight, or importance of the (i, j)-th entry is W_{ij} (as $\mathbf{w}_1^\top \mathbf{X} \mathbf{w}_2 = vec(\mathbf{W})^\top vec(\mathbf{X})$, where vec is a function vectorizing a matrix). From the we can clearly observe a block structure on the upper left corner. This complies with the latent data structure and explains the reason why we can achieve a 1 AUC. For comparison purpose, we also plotted the matrix of $\mathbf{w}_1 \mathbf{w}_2^\top$ after 100 iterations with *MulSLR* with $\lambda_1 = \lambda_2 = 0$ in Fig.4 (a), in which case the sparsity ($\ell 1$) regularizations on $\mathbf{w}_1$ and $\mathbf{w}_2$ do not take effect. We can observe that in this case the matrix is dense, which is because $\mathbf{w}_1$ and $\mathbf{w}_2$ are dense vectors. We also checked the predicted AUC value this dense $\mathbf{w}_1$ and $\mathbf{w}_2$ can get, which is only 0.6525. This validates the superiority of sparse multilinear logistic regression over plain multilinear logistic regression in this case.

To answer the second question, we conducted two sets of experiments on Mac OS 10.7 with 2.2GHz CPU and 12GB main memory. In the first set of experiments, we randomly generated a set of 100×100 data matrices, and we record the



(a) Objective function loss vs. number of iterations

(b) Prediction AUC vs. number of iterations

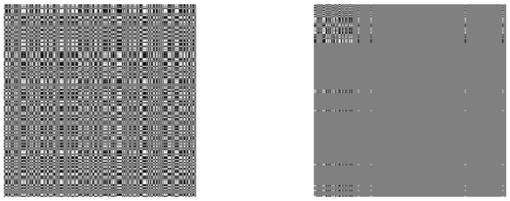
Figure 3: Convergence plots on the synthetic data. (a) shows how the objective function value with respect to the number of iterations when training with 80% of the data. (b) shows how the corresponding testing AUC goes with number of iterations on the rest 20% data. From the figure we can see that in this case *MulSLR* converges in about 60 iterations

averaged running time per iteration for *MulSLR* with respect to different data set size. The result is shown in Fig.5(a), which shows a clear linear trend between the running time and data scale. In the second set of experiments, we fixed the data set size to be 100, but varying the data dimensionality from 50 × 50 to 2000 × 2000. The result is provided in Fig.5(b), which shows that the trend of that curve is slightly quadratic (as the entire data dimensionality is the square of the horizontal axis values). This is in accordance with our complexity analysis of Algorithm 2.

4.2 Experiments on fMRI Data

Functional magnetic resonance imaging or functional MRI (fMRI) is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting associated changes in blood flow[1]. fMRI is an effective approach to investigate alterations in brain function related to the earliest symptoms of Alzheimer's disease, possibly before development of significant irreversible structural damage.

The raw fMRI scans used in our experiments were collected from real clinic cases of 1,005 patients [21], whose cognitive function scores (semantic, episodic, executive and spatial - ranges between -2.8258 and 2.5123) were also acquired at the same time using a cognitive function test.



(a) No $\ell 1$ regularizations (b) With $\ell 1$ regularizations

Figure 4: The matrix of $\hat{\mathbf{w}}_1\hat{\mathbf{w}}_2^\top$, where $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2$ are the converged solution over 100 iterations. (a) shows the result with $\lambda_1 = \lambda_2 = 0$, i.e., no $\ell 1$ regularizations. (b) shows the result of *MulSLR* with $\lambda_1 = \lambda_2 = 0.01$.

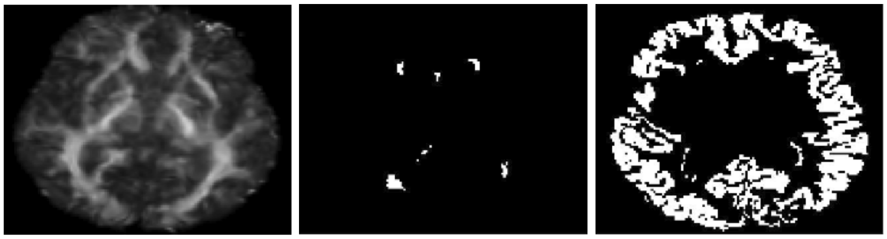
There are three types of MRI scans that were collected from the subjects: (1) FA, the fractional anisotropy MRI gives information about the shape of the diffusion tensor at each voxel, which reflects the differences between an isotropic diffusion and a linear diffusion; (2) FLAIR, Fluid attenuated inversion recovery is a pulse sequence used in MRI, which uncovers the white matter hyperintensity of the brain; (3) GRAY, gray MRI images revealing the gray matter of the brain. In the raw scans, each voxel has a value from 0 to 1, where 1 indicates that the structural integrity of the axon tracts at that location is perfect, while 0 implies either there are no axon tracts or they are shot (not working). The raw scans are preprocessed (including normalization, denoising and alignment) and then restructured to 3D tensors with a size of 134 × 102 × 134. Fig.6 demonstrate a sample image for each of thee three types of scans. Another information we have for this data set is associated with each sample we have a label, which could be either *normal*, *Mild Cognitive Impairment* (MCI) or *demented*.

Because this is a three-class problem and logistic regression is for binary classification, we constructed three prediction tasks with one-versus-rest strategy, i.e., *normal* vs. *MCI* and *demented*, *MCI* vs. *normal* and *demented*, *demented* vs. *normal* and *MCI*. For *MulSLR*, we set the ℓ_1 term regularization parameters $\lambda_1 = \lambda_2 = \lambda_3$ and tune it from the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ with five fold cross validation. The ℓ_2 term regularization parameters are set to $\mu_1 = \mu_2 = \mu_3 = 10^{-4}$. For comparison purpose, we also implemented the following baseline algorithms:

- **Nearest Neighbor** (NN). This is the one nearest neighbor classifier with standard Euclidean distance.
- **Support Vector Machine** (SVM). This is the regular vector based SVM method.
- **Logistic Regression** (LR). This is the traditional vector based logistic regression method.
- **Sparse Logistic Regression** (SLR). This is the vector based sparse logistic regression.
- **Multilinear Logistic Regression** (MLR). This is equivalent to *MulSLR* with all $\ell 1$ regularization parameters setting to 0.

We use *LIBLINEAR* [7] for the implementation of LR and SLR, and *LIBSVM* [5] for the implementation of SVM. Note

[1] `http://en.wikipedia.org/wiki/Functional_magnetic_resonance_imaging`



(a) FA (b) FLAIR (c) GRAY

Figure 6: Sample images of the three different types of scans.

Table 1: Prediction AUC Comparison for Different Methods and Different Tasks

method	*normal* vs. *MCI* and *demented*	*MCI* vs. *normal* and *demented*	*demented* vs. *normal* and *MCI*
NN	0.5052 ± 0.0555	0.5085 ± 0.0473	0.5106 ± 0.0487
SVM	0.6491 ± 0.1355	0.5230 ± 0.0567	0.5409 ± 0.0473
LR	0.6591 ± 0.1421	0.5860 ± 0.1007	0.6203 ± 0.1253
SLR	0.6595 ± 0.1396	0.5877 ± 0.0993	0.6220 ± 0.1243
MLR	0.6754 ± 0.0896	0.5994 ± 0.0627	0.6382 ± 0.1201
MulSLR	$\mathbf{0.6923 \pm 0.0706}$	$\mathbf{0.6015 \pm 0.0520}$	$\mathbf{0.6731 \pm 0.1054}$

that in order to test those vector based approaches, we need to stretch those fMRI tensors into very long vectors (with dimensionality 1,831,512).Table 1 summarized the average and standard deviation over 5-fold cross validation in terms of Areas Under the receiver operating characteristics Curve (AUC) values. The data we used are the FLAIR images. From the table we can observe that:

- The multilinear methods work better than traditional vector based approaches. One possible reason is because there is a clear spatial structure on fMRI images which those multilinear methods can take advantage of.
- The sparse methods work better than their non sparse counterparts. This can also be understood because the FLAIR images are sparse in nature.
- Discriminating MCI from normal and demented is more difficult compared with the other two tasks. This is because MCI is an intermediate state during the progression of Alzheimer's disease from normal to demented.

4.3 Experiments on CHF Onset Prediction

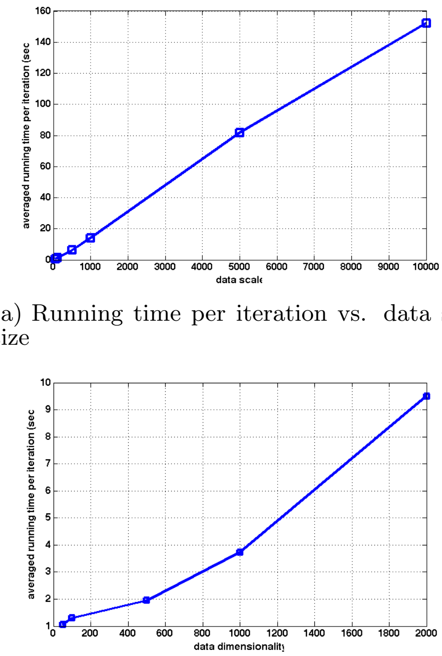
Congestive heart failure (CHF), which refers to a condition where the heart cannot pump enough blood to meet the body's needs, is a major chronic illness in the U.S., affecting more than five million patients. It is estimated CHF costs the nation an estimated \$32 billion each year[2]. Effective prediction of the onset risk of potential CHF patients would help identify the patient at risk in time. Thus the decision makers can provide the proper treatment to the right patients. this can also help save unnecessary costs.

Electronic Health Records (EHR) are systematic collection of patient health information including diagnosis, medication, lab, procedure, demographics, etc. It has now been becoming one of the major information source for conducting healthcare analytics research. The data set we use in our experiments is from a real world EMR data warehouse including the longitudinal EHR of 319,650 patients over 4 years. On this data set, we identified 1,000 CHF case patients, i. e., the patients who are confirmed with CHF with the identification criterion in [29]. Then we did a group match according to patient demographics, comorbidities and primary care physicians similar as in [29] identifying 2,000 control patients. We use the medication orders of those patients within two years from their operational criteria date (for case patients, their operational criteria dates are just their CHF confirmation date. For control patients that date is just the date of their last records in the database). On each medication order we use the corresponding pharmacy class and the primary diagnosis in terms of Hierarchical Condition Category (HCC) codes [20] for the medication prescription. In total there are 92 unique pharmacy classes and 195 distinct HCC codes. Therefore each patient can be represented as a 92×195 matrix, where the (i, j)-th entry indicates the frequency that the i-th drug was prescribed during the two years with the j-th diagnosis code as primary diagnosis.

The parameters for *MulSLR* are set in the same manner as the experiments in last subsection. For comparison purpose, we also implemented NN, SVM, LR, SLR, MLR and reported the averaged AUC value over 5-fold cross validation along with their standard deviations on Fig.7. From the figure we can clearly observe that multilinear methods still work better than vector based methods (in order to implement vector based methods, we still need to stretch the patient matrices into vectors). *MulSLR* performs better than MLR, which is the regular ridged bilinear logistic regression, because the patient matrices formed are typically very sparse. However, in this case the vector based SLR works slightly worse than LR, this could be due to the sparsity structures of the patient matrices get lost when stretching them into vectors.

One interesting thing to check is the product of $\mathbf{w}_{med}$, which is the classification vector on the medication side, and $\mathbf{w}_{diag}^{\top}$, which is the classification vector on the diagnosis side. Just like what we examined on the toy data. In this way we

[2] http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/docs/fs_heart_failure.pdf



(a) Running time per iteration vs. data set size

(b) Running time per iteration vs. data dimensionality

Figure 5: Averaged running time per iteration of *MulSLR*. (a) shows the running time per iteration vs. data set size plot, where the data dimensionality is fixed to 100. (b) shows the running time per iteration vs. one-side data dimensionality plot, where the data set size is fixed to 500.

can find some strongly correlated medications and diagnosis that could be highly predictive for CHF onset risk. To achieve this goal, we use all the data to train the $\mathbf{w}_{med}$ and $\mathbf{w}_{diag}$ and plot their outer-product as in Fig.8 (where we just show a submatrix due to space limitation), where warm color indicate high values and cold color indicates small values. From the image we can observe that:

- **Cardiac disease and their corresponding treatment drugs are highly correlated and predictive**. For example, *Antihypertensive, Antihyperlipidemic, Calcium Blocker, Beta Blockers, Cardiotonic* drugs with *CHF* (HCC080), *Acute Myocardial Infarction* (HCC081), *Hypertensive Heart Disease* (HCC090), *Hypertension* (HCC091) diagnosis.
- **CHF commorbidities and their corresponding treatment drugs are predictive**. For example, *Chronic Obstructive Pulmonary Disease* (COPD) (HCC108) [25] with *Corticosteroids*, and also *Chronic Kidney Disease* (CKD) (HCC436 and HCC439) [1].
- **CHF related symptoms and their corresponding treatment drugs are predictive**, such as Gout, which is a well-known Framingham symptom [16] for CHF patients.

Actually all those findings are also clinically validated in those references we cited. We also provide the detailed description of all HCC codes at `https://www.dropbox.com/s/6e1qbjf1ce7yi6x/hcc_codes.pdf`.

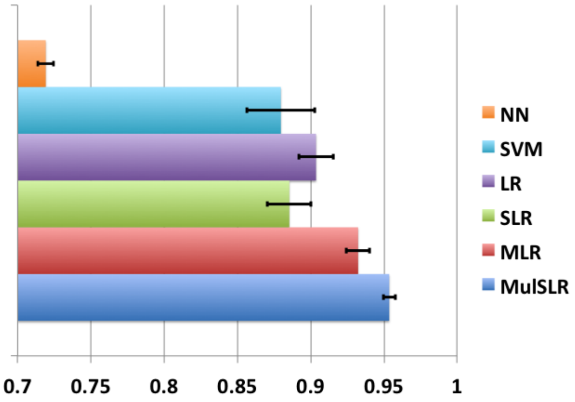


Figure 7: Prediction performance for different methods on the CHF onset prediction task in terms of averaged AUC value with 5-fold cross validation along with their standard deviations.

5. CONCLUSIONS

We propose a multilinear sparse logistic regression method called *MulSLR* in this paper, which can directly take data matrices or tensors as inputs and do prediction on that. *MulSLR* is formulated as an optimization problem and we propose an effective BCD strategy to solve it. We proved the convergence and analyzed the convergence rate theoretically. Finally we validate the effectiveness and efficiency of *MulSLR* on both synthetic and real world data sets. We demonstrate that *MulSLR* can not only achieve good performance, but also discover interesting predictive patterns.

Appendix I: Proof of Theorem 3.1

Let

$$\mathcal{W}_{\setminus k} = (\mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^{(k-1)}, \mathbf{w}^{(k+1)}, \cdots, \mathbf{w}^K) \tag{29}$$

Then for any $(\mathcal{W}_{\setminus k}, \mathbf{w}^k, b)$ and $(\mathcal{W}_{\setminus k}, \widehat{\mathbf{w}}^k, \widehat{b})$, we have

$$\begin{aligned}
& \left\|\nabla_{(\mathbf{w}^k,b)}\ell(\mathcal{W}_{\setminus k}, \mathbf{w}^k, b) - \nabla_{(\mathbf{w}^k,b)}\ell(\mathcal{W}_{\setminus k}, \widehat{\mathbf{w}}^k, \widehat{b})\right\|_2 \\
\leqslant\; & \frac{1}{n}\sum_{i=1}^{n}\left|\left[1+\exp\left(y_i f_{(\mathcal{W}_{\setminus k},\mathbf{w}^k,b)}(\mathcal{X}^i)\right)\right]^{-1}\right. \\
& \left. - \left[1+\exp\left(y_i f_{(\mathcal{W}_{\setminus k},\widehat{\mathbf{w}}^k,\widehat{b})}(\mathcal{X}^i)\right)\right]^{-1}\right|\left(\|\nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i)\|_2 + 1\right) \\
\leqslant\; & \frac{1}{n}\sum_{i=1}^{n}\left(\|\nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i)\|_2 + 1\right)^2\left(\left\|\mathbf{w}^k - \widehat{\mathbf{w}}^k\right\|_2 + |b - \widehat{b}|\right) \\
\leqslant\; & \frac{\sqrt{2}}{n}\sum_{i=1}^{n}\left(\|\nabla_{\mathbf{w}^k} f_{(\mathcal{W},b)}(\mathcal{X}^i)\|_2 + 1\right)^2\left\|\left(\mathbf{w}^k, b\right) - \left(\widehat{\mathbf{w}}^k, \widehat{b}\right)\right\|_2
\end{aligned}$$

This completes the proof.

Appendix II: Proof of Theorem 3.2

For notational convenience, we define

$$\ell^k_{(t)}(\mathbf{w}^k, b) = \ell(\mathcal{W}^{1\sim(k-1)}_{(t)}, \mathbf{w}^k, \mathcal{W}^{(k+1)\sim K}_{(t-1)}, b) \tag{30}$$

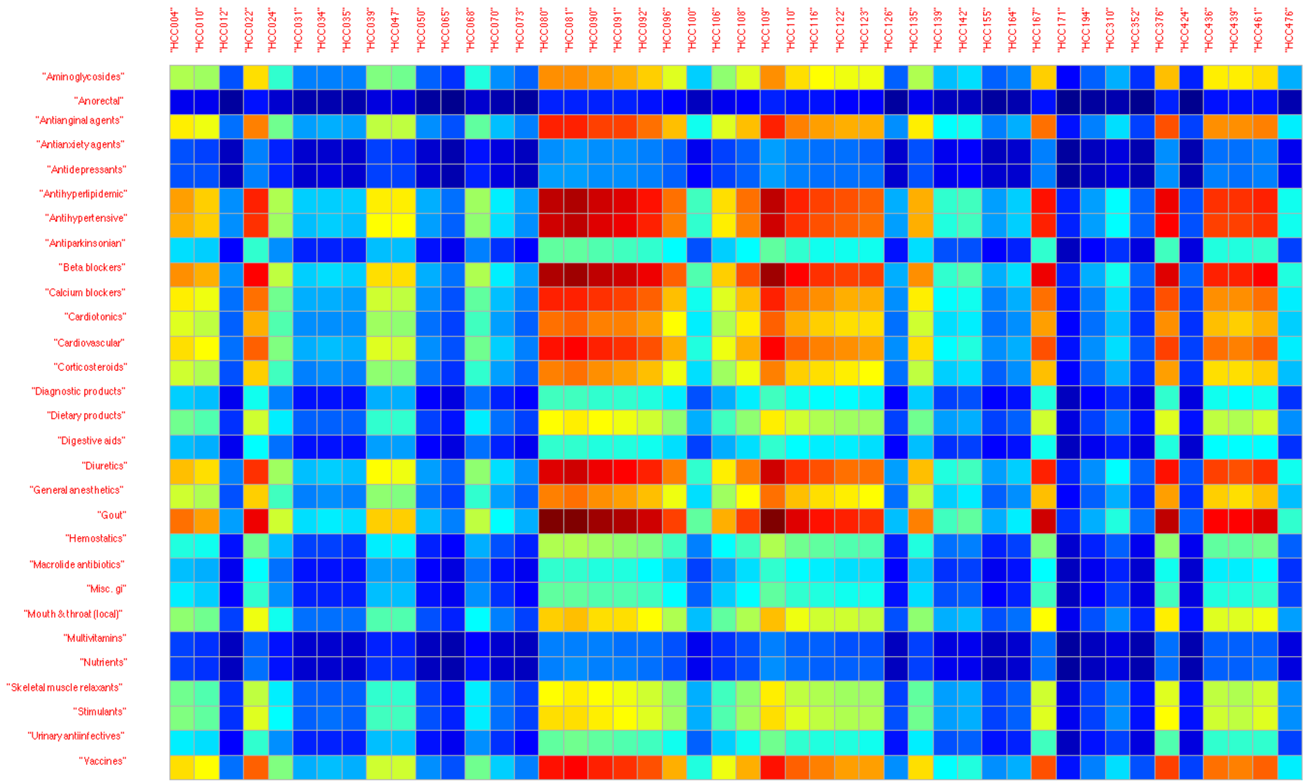


Figure 8: A sub matrix from the matrix of $\mathbf{w}_{med}\mathbf{w}_{diag}^{\top}$. Warm color indicates large value, and cold color indicates small value. Thus warm color indicates *MulSLR* gives larger weights to the corresponding entry when making a decision. There are some interesting red blocks in the image. For example, *Antihypertensive, Antihyperlipidemic, Calcium Blocker, Beta Blockers, Cardiotonic* drugs with *CHF* (HCC080), *Acute Myocardial Infarction* (HCC081), *Hypertensive Heart Disease* (HCC090), *Hypertension* (HCC091) diagnosis. Those are all cardiac disease and their corresponding treatment drugs. The row of drug *Corticosteroids* is also warm because it is used for treating pulmonary disease such as COPD, which is a common commorbidity of CHF. For similar reason, the column of diagnosis COPD (HCC108) is warm.

With Lemma 2.3 in [2], we have

$$\begin{aligned}
&\ell^k_{(t)}(\mathbf{w}^k_{(t-1)}, b_{(t,k-1)}) - \ell^k_{(t)}(\mathbf{w}^k_{(t)}, b_{(t,k)}) \qquad (31)\\
&\geqslant \frac{\tau^k_{t-1}}{2}\|\widetilde{\mathbf{w}}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + \frac{\tau^k_{t-1}}{2}|\widetilde{b}_{(t,k-1)} - b_{(t,k)}|^2\\
&\quad + \tau^k_{(t-1)}(\widetilde{\mathbf{w}}^k_{(t-1)} - \mathbf{w}^k_{(t-1)})^\top(\mathbf{w}^k_{(t)} - \widetilde{\mathbf{w}}^k_{(t-1)})\\
&\quad + \tau^k_{(t-1)}(\widetilde{b}_{(t,k-1)} - b_{(t,k-1)}) \cdot (b_{(t,k)} - \widetilde{b}_{(t,k-1)})\\
&= \frac{\tau^k_{t-1}}{2}\|\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + \frac{\tau^k_{t-1}}{2}|b_{(t,k-1)} - b_{(t,k)}|^2\\
&\quad - \frac{\tau^k_{t-1}}{2}(\omega^k_{(t-1)})^2[\|\mathbf{w}^k_{(t-2)} - \mathbf{w}^k_{(t-1)}\|^2 + |b_{(t,k-2)} - b_{(t,k-1)}|^2]\\
&\geqslant \frac{\tau^k_{t-1}}{2}\|\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + \frac{\tau^k_{t-1}}{2}|b_{(t,k-1)} - b_{(t,k)}|^2\\
&\quad - \frac{\tau^k_{t-2}}{2}(\delta_\omega)^2[\|\mathbf{w}^k_{(t-2)} - \mathbf{w}^k_{(t-1)}\|^2 + |b_{(t,k-2)} - b_{(t,k-1)}|^2]
\end{aligned}$$

Then

$$\begin{aligned}
&\ell(\mathcal{W}_{(t-1)}, b_{(t-1,K)}) - \ell(\mathcal{W}_{(t)}, b_{(t,K)}) \qquad (32)\\
&= \textstyle\sum_{k=1}^K \left(\ell^k_{(t)}\left(\mathbf{w}^k_{(t-1)}, b_{(t,k-1)}\right) - \ell^k_{(t)}\left(\mathbf{w}^k_{(t)}, b_{(t,k)}\right)\right)\\
&\geqslant \sum_{k=1}^K \left(\frac{\tau^k_{t-1}}{2}\|\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + \frac{\tau^k_{t-1}}{2}|b_{(t,k-1)} - b_{(t,k)}|^2\right.\\
&\quad \left. - \frac{\tau^k_{t-2}}{2}(\delta_\omega)^2\|\mathbf{w}^k_{(t-2)} - \mathbf{w}^k_{(t-1)}\|^2 - \frac{\tau^k_{t-2}}{2}(\delta_\omega)^2|b_{(t,k-2)} - b_{(t,k-1)}|^2\right)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\ell(\mathcal{W}_{(0)}, b_{(0,0)}) - \ell(\mathcal{W}_{(t)}, b_{(t,K)}) \qquad (33)\\
&\geqslant \sum_{k=1}^K \sum_{t=1}^T \left(\frac{\tau^k_{t-1}}{2}\|\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + \frac{\tau^k_{t-1}}{2}|b_{(t,k-1)} - b_{(t,k)}|^2\right.\\
&\quad \left. - \frac{\tau^k_{t-2}}{2}(\delta_\omega)^2\|\mathbf{w}^k_{(t-2)} - \mathbf{w}^k_{(t-1)}\|^2 \; \frac{\tau^k_{t-2}}{2}(\delta_\omega)^2|b_{(t,k-2)} - b_{(t,k-1)}|^2\right)\\
&\geqslant \sum_{k=1}^K \sum_{t=1}^T \left(\frac{\tau^k_{t-1}(1-\delta^2_\omega)}{2}[\|\mathbf{w}^k_{(t-1)} - \mathbf{w}^k_{(t)}\|_2^2 + |b_{(t,k-1)} - b_{(t,k)}|^2]\right)
\end{aligned}$$

As $\ell(\mathcal{W}, b)$ is lower bounded, when $T \to \infty$, we have

$$\sum_{t=0}^{\infty} \|(\mathcal{W}_{(t)}, b_{(t,k)}) - (\mathcal{W}_{(t+1)}, b_{(t+1,k)})\|_2^2 < \infty$$

and $\|(\mathcal{W}_{(t-1)}, b_{(t-1,k)}) - (\mathcal{W}_{(t)}, b_{(t,k)})\|_F^2 \to 0$. This completes the proof.

6. REFERENCES

[1] A. Ahmed, M. W. Rich, P. W. Sanders, G. J. Perry, G. L. Bakris, M. R. Zile, T. E. Love, I. B. Aban, and M. G. Shlipak. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *The American journal of cardiology*, 99(3):393–398, 2007.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] J. Bolte, A. Daniilidis, and A. Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[4] D. Cai, X. He, J.-R. Wen, J. Han, and W.-Y. Ma. Support tensor machines for text categorization. *Department of Computer Science Technical Report No. 2714, University of Illinois at Urbana-Champaign (UIUCDCS-R-2006-2714)*, 2006.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] W. S. Cooper, F. C. Gey, and D. P. Dabney. Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 198–210. ACM, 1992.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[9] H. Hung and C.-C. Wang. Matrix variate logistic regression model with application to eeg data. *Biostatistics*, 14(1):189–202, 2013.

[10] Y. Kim, S. Kwon, and S. Heun Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3):1643–1655, 2006.

[11] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783. Institut Fourier, 1998.

[12] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[13] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S. J. Meltzer, and M. Tan. Sparse logistic regression with lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[14] S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595. Institut Fourier, 1993.

[15] T. Manninen, H. Huttunen, P. Ruusuvuori, and M. Nykter. Leukemia prediction using sparse logistic regression. *PloS one*, 8(8), 2013.

[16] P. A. McKee, W. P. Castelli, P. M. McNamara, and W. B. Kannel. The natural history of congestive heart failure: the framingham study. *New England Journal of Medicine*, 285(26):1441–1446, 1971.

[17] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[18] M. Miravitlles, T. Guerrero, C. Mayordomo, L. Sánchez-Agudo, F. Nicolau, and J. L. Segú. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory copd patients: a multiple logistic regression analysis. *Respiration*, 67(5):495–501, 2000.

[19] E. F. Philbin and T. G. DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

[20] G. C. Pope, R. P. Ellis, A. S. Ash, J. Ayanian, D. Bates, H. Burstin, L. Iezzoni, E. Marcantonio, and B. Wu. Diagnostic cost group hierarchical condition category models for medicare risk adjustment. *Health Economics Research, Inc. Waltham, MA*, 2000.

[21] B. Qian, X. Wang, F. Wang, H. Li, J. Ye, and I. Davidson. Active learning from relative queries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1614–1620. AAAI Press, 2013.

[22] A. Rao, Y. Lee, A. Gass, and A. Monsch. 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*, pages 4499–4502. IEEE, 2011.

[23] X. Ren and J. Malik. Learning a classification model for segmentation. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 10–17. IEEE, 2003.

[24] R. T. Rockafellar, R. J.-B. Wets, and M. Wets. *Variational analysis*, volume 317. Springer, 1998.

[25] F. H. Rutten, M.-J. M. Cramer, J.-W. J. Lammers, D. E. Grobbee, and A. W. Hoes. Heart failure and chronic obstructive pulmonary disease: an ignored combination? *European journal of heart failure*, 8(7):706–711, 2006.

[26] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, 51(2):752–764, 2010.

[27] J. Sun, J. Hu, D. Luo, M. Markatou, F. Wang, S. Edabollahi, S. E. Steinhubl, Z. Daar, and W. F. Stewart. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2012, page 901. American Medical Informatics Association, 2012.

[28] X. Tan, Y. Zhang, S. Tang, J. Shao, F. Wu, and Y. Zhuang. Logistic tensor regression for classification. In *Intelligent Science and Intelligent Data Engineering*, pages 573–581. Springer, 2013.

[29] J. Wu, J. Roy, and W. F. Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[30] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye. Multi-source learning with block-wise missing data for alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193. ACM, 2013.

[31] Y. Xu and W. Yin. A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3), 2013.

[32] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(1):131–137, 2004.

[33] J. Ye, R. Janardan, Q. Li, et al. Two-dimensional linear discriminant analysis. In *NIPS*, volume 4, page 4, 2004.

[34] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.