Chemometrics and Intelligent Laboratory Systems 142 (2015) 70–77

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Short Communication

Classification of spectral data using fused lasso logistic regression

Donghyeon Yu [a,1], Seul Ji Lee [c,1], Won Jun Lee [c], Sang Cheol Kim [d], Johan Lim [b,*], Sung Won Kwon [c,*]

[a] Department of Statistics, Keimyung University, Daegu, Republic of Korea
[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea
[c] College of Pharmacy, Seoul National University, Seoul, Republic of Korea
[d] Samsung Genome Institute, Samsung Medical Center, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 10 April 2014
Received in revised form 31 December 2014
Accepted 13 January 2015
Available online 22 January 2015

Keywords:
Classification
Fused lasso regression
Mass spectral data
ℓ_1-regularization
Penalized logistic regression

ABSTRACT

Spectral data contain powerful information that can be used to identify unknown compounds and their chemical structures. In this paper, we study fused lasso logistic regression (FLLR) to classify the spectral data into two groups. We show that the FLLR has a grouping property on regression coefficients, which simultaneously selects a group of highly correlated variables together. Both the sparsity and the grouping property of the FLLR provide great advantages in the analysis of the spectral data. In particular, it resolves the well-known peak misalignment problem of the spectral data by providing data dependent binning, and provides a better interpretable classifier than other ℓ_1-regularization methods. We also analyze the gas chromatography/mass spectrometry data to classify the origin of herbal medicines, and illustrate the advantages of the FLLR over other existing ℓ_1-regularized methods.

1. Introduction

This paper aims to "jointly" find differently expressed peaks between two groups and build an efficient classifier of high-dimensional spectral data using the fused lasso logistic regression (FLLR). The high dimensional spectral data – near infrared spectral (NIR) data, nuclear magnetic resonance (NMR) data, liquid chromatography mass spectral (LC/MS) data, and gas chromatography mass spectral (GC/MS) data – are widely used in various biological and medical disciplines. For example, the NMR, which observes magnetic properties from the energy absorbed and re-emitted via atomic nucleus, is used to identify compounds in a given sample mixture [1,2]. On the other hand, mass spectrometry (MS), which ionizes chemical compounds and measures mass-to-charge ratio of charged particles (of ion fragments), is popular in many bio-analytical sectors [3,4].

Despite its usefulness in various scientific fields, the analysis of spectral data (or mass spectral data) comes with some difficulties. Mass spectrometry data usually have unwanted local or global shifts of peaks (misalignment of peaks) due to instrumental instability or small differences in experimental conditions. The misalignment of peaks over samples weakens the strength of major signals and introduces loss of efficiency in statistical analyses. This requires us to align spectrum before analysis or to use a more complex model to explain it [5–8]. Second, the mass spectral data are typical examples of high dimensional and low sample size (HDLSS) data, which introduces ill-posedness of the problem. To resolve this difficulty, many advanced statistical procedures are proposed and the ℓ_1-regularized regression is one of the most popular recently [9–14].

This paper shows the fused lasso regression, a variation of the ℓ_1-regularized regression, can resolve the difficulties addressed above. The classical lasso regression by Tibshirani [9] penalizes the ℓ_1-norm of the coefficient vector (the sum of absolute values of coefficients) of ordinary least square, and provides a sparse solution that estimates many of coefficients as 0. The sparse estimate of the coefficient vector additionally allows us to select variables of the model, and this becomes the most attractive feature that makes the lasso regression be widely used in various applications. Many variations of the lasso regression are proposed in the literature. In particular, we find several modifications of the lasso regression for strongly correlated covariates. The classical lasso regression randomly chooses one of them as non-zero, if the model has a group of strongly correlated covariates (with non-zero coefficients). Unlike the classical lasso regression, the elastic-net regression by Zou and Hastie [11] has a penalty on a convex combination of the ℓ_1-norm and the square of the ℓ_2-norm of the coefficient vector. It has the grouping property that selects or removes strongly correlated covariates simultaneously. The fused lasso regression of this paper, firstly proposed by Tibshirani et al. [12] and studied much recently [15–20], assumes that the covariates are observed in order (for example, in time order) and penalizes a convex combination of the ℓ_1-norm of differences of adjacent coefficients and the ℓ_1-norm of the coefficient vector itself. As in the elastic-net, the fused lasso regression also simultaneously

* Corresponding authors.
E-mail addresses: johanlim@snu.ac.kr (J. Lim), swkwon@snu.ac.kr (S.W. Kwon).
[1] Donghyeon Yu and Seul Ji Lee contributed equally to this paper.

http://dx.doi.org/10.1016/j.chemolab.2015.01.006
0169-7439/

selects a group of strongly correlated adjacent covariates, and further makes their estimates be equal to each other.

The spectral data we study in this paper have three interesting features. (i) the dimension of covariates is much larger than the number of samples (high-dimensionality), (ii) many covariates are zero or close to zero. They are simply noise and their coefficients are zero in the model (sparsity), and (iii) covariates are observed in order, e.g., in the order of mass-to-charge ratio or retention time in mass spectral data. For analyzing the spectral data, the fused lasso regression has several advantages compared to other ℓ_1-regularization methods. The sparsity and the grouping property of the fused lasso regression are well fitted to the spectral data. In addition, the grouping property of the fused lasso regression naturally provides data dependent binning of covariates and resolves the difficulty from misalignment of samples.

This paper is organized as follows. In Section 2, the FLLR is introduced and several of its properties are studied. In particular, we show the grouping property of the FLLR, and propose a logistic modification of the split Bregman (SB) algorithm to solve the FLLR. Here, the main interest of this paper is to find a classifier of spectral data, and we focus our discussion on the logistic regression rather than the classical linear regression. Section 3 analyzes the GC/MS data, where we aim to jointly find differentially expressed peaks and classify the origin of oriental herbal medicines. Section 4 concludes the paper with a brief discussion of the FLLR for the two-dimensional spectral data.

2. Fused lasso logistic regression

Let p denote the number of discrete retention times (t) or mass-to-charge ratios (m/z). Let $X \in \mathbb{R}^p$ be the intensities of ions in a sample. We consider two-class problem, whose classes are labeled by 0 and 1. In our example, the class corresponds to the origin of an herbal medicine; 0 (respectively, 1) indicates its origin is China (respectively, Korea). Our goal is to build a linear classifier $f(\mathrm{X}) = \mathrm{X}\beta = \sum_{j=1}^{p} X^j\beta_j$ to predict the origin of a new sample. To be specific, if $f(\mathrm{X}) \geq 0$, we classify the sample into class 1 ($Y = 1$); otherwise classify it into class 0 ($Y = 0$).

Suppose we have n (training) samples $\{(\mathrm{x}_i, y_i), i = 1, 2, \ldots, n\}$. To build the classifier, we often minimize the empirical risk based on the training sample as

$$\mathrm{R}_{\mathrm{emp}}(\beta_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + \mathrm{x}_i\beta), \tag{1}$$

where $y_i \in \{0, 1\}$, $\mathrm{x}_i = (x_i^1, x_i^2, \ldots, x_i^p)$ is a $1 \times p$ covariate vector of the i-th subject, $\beta_0 \in \mathbb{R}$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p$. This paper studies the logistic loss that is

$$l(y_i, \beta_0 + \mathrm{x}_i\beta) = -y_i(\beta_0 + \mathrm{x}_i\beta) + log\{1 + exp(\beta_0 + \mathrm{x}_i\beta)\}. \tag{2}$$

The spectral data is of HDLSS. The sparse regression using the ℓ_1-regularization is one of the most popular tool for the HDLSS data. The FLLR, which is a main theme of this paper, proposes to minimize

$$\sum_{i=1}^{n}[-y_i(\beta_0 + \mathrm{x}_i\beta) + log\{1 + exp(\beta_0 + \mathrm{x}_i\beta)\}] + \lambda\Omega(\beta), \tag{3}$$

where

$$\Omega(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\|\mathrm{D}\beta\|_1 \tag{4}$$

with

$$\mathrm{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}. \tag{5}$$

In the above, $\lambda > 0$ and $\alpha \in (0, 1)$ are tuning parameters to decide the magnitudes of the shrinkage and fusion of the estimates of β, respectively.

The FLLR has the grouping property for a set of highly correlated variables as claimed in Theorem 1, which is an analogy of Theorem 1 in Bondell and Reich [14]. Theorem 1 shows that the estimates of coefficients of any two highly correlated adjacent variables become one of the following two cases. They are equal to each other, and are simultaneously selected or removed from the model. If not, their estimates simply bridge the monotone trend of their neighboring coefficients. Let $\mathrm{y} = (y_1, y_2, \ldots, y_n)^T$ and

$$\mathrm{X} = \begin{pmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_n \end{pmatrix} = \left(\mathrm{x}^1; \mathrm{x}^2; \cdots; \mathrm{x}^p\right) = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^p \end{pmatrix}.$$

Theorem 1. *Let λ and α be the two tuning parameters in the FLLR. Given data (y, X) with a binary response y and standardized covariates $\mathrm{X} = (\mathrm{x}^1, \ldots, \mathrm{x}^p)$, let $\hat{\beta}(\lambda)$ be the regression estimate using the tuning parameters λ for a fixed constant $\alpha \in (0, 1)$. Assume that the predictors are signed such that $\hat{\beta}_j(\lambda) \geq 0$ for all j. Let $\rho_j = (\mathrm{x}^j)^T\mathrm{x}^{j+1}$ be the sample correlation between standardized covariates x^j and x^{j+1}, for $j = 2, \ldots, p-2$. For each j, if $\lambda > \sqrt{2n\left(1-\rho_j\right)}\Big/(1-\alpha)$, then either (i) $\hat{\beta}_j(\lambda) = \hat{\beta}_{j+1}(\lambda)$ or (ii) $\mathrm{sign}\left(\hat{\beta}_{j+1}(\lambda) - \hat{\beta}_j(\lambda)\right) = \mathrm{sign}\left(\hat{\beta}_j(\lambda) - \hat{\beta}_{j-1}(\lambda)\right) = \mathrm{sign}\left(\hat{\beta}_{j+2}(\lambda) - \hat{\beta}_{j+1}(\lambda)\right) \neq 0$.*

Proof. The proof is similar to that of Theorem 1 in Bondell and Reich [14]. The sub-differential of the fusion penalty function is:

$$\partial_j(\|\mathrm{D}\beta\|_1) = \begin{cases} \mathrm{sgn}(\beta_1 - \beta_2) & \text{if } j = 1, \\ \mathrm{sgn}\left(\beta_p - \beta_{p-1}\right) & \text{if } j = p, \\ \mathrm{sgn}\left(\beta_j - \beta_{j-1}\right) - \mathrm{sgn}\left(\beta_{j+1} - \beta_j\right) & \text{otherwise}, \end{cases}$$

where $\mathrm{sgn}(x) = \mathrm{sign}(x)$ if $x \neq 0$ and $\mathrm{sgn}(x) \in [-1, 1]$ if $x = 0$.

Suppose that $\hat{\beta}_j(\lambda) \neq \hat{\beta}_{j+1}(\lambda)$, then, the differentiation of (3) with respect to β_j becomes

$$-\left(\mathrm{x}^j\right)^T\left\{\mathrm{y} - \mu\left(\hat{\beta}_0 + \sum_k \mathrm{x}^k\hat{\beta}_k\right)\right\} + \lambda\alpha + \lambda(1-\alpha) \times \left(\mathrm{sign}\left(\hat{\beta}_j - \hat{\beta}_{j-1}\right) - \mathrm{sign}\left(\hat{\beta}_{j+1} - \hat{\beta}_j\right)\right) = 0, \tag{6}$$

where $\mu\left(\hat{\beta}_0 + \sum_k \mathrm{x}^k\hat{\beta}_k\right) = \left(\mu_1\left(\hat{\beta}_0 + \mathrm{x}_1\hat{\beta}\right), \mu_2\left(\hat{\beta}_0 + \mathrm{x}_2\hat{\beta}\right), \cdots, \mu_n\left(\hat{\beta}_0 + \mathrm{x}_n\hat{\beta}\right)\right)$ with

$$\mu_i\left(\hat{\beta}_0 + \mathrm{x}_i\hat{\beta}\right) = \frac{exp\left\{\hat{\beta}_0 + \sum_{j=1}^{p} x_i^j\hat{\beta}_j\right\}}{1 + exp\left\{\hat{\beta}_0 + \sum_{j=1}^{p} x_i^j\hat{\beta}_j\right\}}.$$

By differentiating (3) with respect to β_{j+1}, we also have

$$-\left(\mathrm{x}^{j+1}\right)^T\left\{\mathrm{y} - \mu\left(\hat{\beta}_0 + \sum_k \mathrm{x}^k\hat{\beta}_k\right)\right\} + \lambda\alpha + \lambda(1-\alpha) \times \left(\mathrm{sign}\left(\hat{\beta}_{j+1} - \hat{\beta}_j\right) - \mathrm{sign}\left(\hat{\beta}_{j+2} - \hat{\beta}_{j+1}\right)\right) = 0. \tag{7}$$

Subtracting (6) from (7) gives

$$-\left(\mathrm{x}^{j+1} - \mathrm{x}^j\right)^T\left\{\mathrm{y} - \mu\left(\hat{\beta}_0 + \sum_k \mathrm{x}^k\hat{\beta}_k\right)\right\} + \lambda(1-\alpha)\left(\kappa_{j+1} - \kappa_j\right) = 0,$$