Computational Statistics and Data Analysis 88 (2015) 119–127

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Generalized orthogonal components regression for high dimensional generalized linear models☆

Yanzhu Lin [a], Min Zhang [b], Dabao Zhang [b,*]

[a] Laboratory of Systems Genetics, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, United States
[b] Department of Statistics, Purdue University, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 17 December 2014
Accepted 9 February 2015
Available online 17 February 2015

Keywords:
Categorical data
Classification
Collinear
Dimension reduction multicollinear

ABSTRACT

The algorithm, generalized orthogonal components regression (GOCRE), is proposed to explore the relationship between a categorical outcome and a set of massive variables. A set of orthogonal components are sequentially constructed to account for the variation of the categorical outcome, and together build up a generalized linear model (GLM). This algorithm can be considered as an extension of the partial least squares (PLS) for GLMs, but overcomes several issues of existing extensions based on iteratively reweighted least squares (IRLS). First, existing extensions construct a different set of components at each iteration and thus cannot provide a convergent set of components. Second, existing extensions are computationally intensive because of repetitively constructing a full set of components. Third, although they pursue the convergence of regression coefficients, the resultant regression coefficients may still diverge especially when building logistic regression models. GOCRE instead sequentially builds up each orthogonal component upon convergent construction, and simultaneously regresses against these orthogonal components to fit the GLM. The performance of the new method is demonstrated by both simulation studies and a real data example.

1. Introduction

Available high-throughput biotechnologies have made it possible to genotype thousands of genetic markers, meanwhile, they bring challenges to statistical analyses of these data. Such data are characterized by a large number of variables (p) observed from a relatively small number of subjects (n), and create the well-known large p small n problems. To deal with this issue, an important strategy is to reduce the high dimensionality of the predictors before fitting models. As a supervised dimension-reduction method, partial least squares (PLS) by Wold (1975) has drawn considerable attention, see Vinzi et al. (2010). PLS constructs orthogonal components such that these components capture information of original predictors predicting response variables, and linear models are built on the base of these components instead of the original predictors. It is computationally fast and able to take collinear or multicollinear predictors.

Success of PLS in fitting linear models motivates extensions to generalized linear models (GLMs). With the iteratively reweighted least squares (IRLS) algorithm commonly used for building regular GLMs (Green, 1984), Marx (1996) proposed

☆ This work was partially supported by NSF CAREER award IIS-0844945 and the Cancer Care Engineering project at the Oncological Science Center of Purdue University.

* Corresponding author.
E-mail address: zhangdb@stat.purdue.edu (D. Zhang).

http://dx.doi.org/10.1016/j.csda.2015.02.006
0167-9473/

an extension, i.e., the iteratively reweighted partial least squares (IRPLS) algorithm, which replaces the least squares estimates with PLS estimates at each iteration. It is a natural extension of PLS, however, a different set of orthogonal components are constructed at each iteration and thus the convergence of original regression coefficients is pursued. As a result, the loadings of orthogonal components never converge, and even regression coefficients especially for logistic regressions rarely converge. A full set of distinct components at each iteration not only make it difficult to interpret, but also demand intensive computation.

Much effort has been devoted to solving the non-convergence issue of IRPLS. Ding and Gentleman (2004) applied the bias reduction procedure proposed by Firth (1993) to IRPLS, specifically for the classification problems. Firth (1993) modified the score function to remove the first order term of the asymptotic bias of maximum likelihood estimators for GLMs. Heinze and Schemper (2002) showed that this bias reduction procedure may also avoid the common infinite estimate problem of logistic regressions. However, the non-convergence issue still exists in IRPLS by Ding and Gentleman (2004), possibly due to varying components at each iteration. Alternatively, Fort and Lambert-Lacroix (2005) proposed to build up continuous pseudo-responses via ridge regression and then apply PLS to regress these pseudo-responses against the predictors; Nguyen and Rocke (2002) instead proposed to first apply PLS by treating the responses as continuous, and then fit a regular GLM using the resultant orthogonal components instead of the original predictors.

Here we propose a different strategy, namely, the generalized orthogonal components regression (GOCRE), to extend the supervised dimension reduction idea in PLS and fit high dimensional GLMs. While IRPLS repetitively constructs a different set of components at each iteration and targets a convergent set of regression coefficients, GOCRE sequentially constructs orthogonal components which maximally account for the remaining variation in the categorical outcome. The bias correction procedure by Firth (1993) is also applied. The proposed method enjoys computational privilege over IRPLS since IRPLS needs to rebuild all orthogonal components at each iteration. The construction of orthogonal components is also different from the methods by Fort and Lambert-Lacroix (2005) and Nguyen and Rocke (2002), both directly maximizing correlation between categorical responses and components.

This paper is organized as follows. The next section introduces our proposed method in details. Simulation studies are shown in Section 3, and an application of the proposed method to a real data set is presented in Section 4. We close the paper with a brief discussion.

2. The method

2.1. High dimensional generalized linear model

Suppose the distribution of response Y is a member of the exponential family distribution,

$$f(y|\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1}$$

where θ is the canonical parameter and ϕ is the known dispersion parameter. A link function $g(\cdot)$ further relates the mean of response Y to the p predictors in X, i.e.,

$$g(E[Y|X]) = \mu + X\beta, \tag{2}$$

where μ is the intercept and β is a p-dimensional column vector containing all regression coefficients of the predictors. The inverse function of $g(\cdot)$ is denoted as $g^{-1}(\cdot)$.

With a size n sample $\{(y_i, \mathbf{x}_i),\ i = 1, 2, \ldots, n\}$, a common issue is how to provide a legitimate estimate of β in (2) when $p \gg n$. Denote $\mathbf{X} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t)^t$, an $n \times p$ matrix with rank $r_x \le min(n, p)$. The classical maximum likelihood estimators (MLEs) of β form a space with dimension at least $p - r_x$. Suppose an $n \times r_x$ matrix $\mathbf{X}_{\mathcal{S}}$ is constructed by a subset of columns of $\mathbf{X}$, and further assume that there is a unique maximum likelihood estimator of $\beta_{\mathcal{S}}$ for the following model,

$$g(E[Y|X_{\mathcal{S}}]) = \mu + X_{\mathcal{S}}\beta_{\mathcal{S}}. \tag{3}$$

Correspondingly there exists a unique MLE of β, namely $\hat{\beta}$, in model (2), satisfying the following assumption:

Assumption 1. $\hat{\beta}^t\psi = 0$ whenever $\mathbf{X}\psi = \mathbf{0}_{n\times 1}$.

In the case that $\mathbf{X}$ is of rank r_x, the above assumption equivalently puts $p - r_x$ constraints on MLE $\hat{\beta}$ to make model (2) identifiable. This assumption makes practical sense in solving the collinearity or multicollinearity issue. For example, if the jth predictor consistently doubles the value of the k-th predictor, we have $\hat{\beta}_j = 2\hat{\beta}_k$. Therefore, the scale of the predictor, if preserved, may indicate its importance. On the other hand, when the predictors are identical, the corresponding regression coefficients will also be identical.

Due to the aforementioned multicollinearity issue, we can focus on building model (2) with β satisfying the following assumption, a population version of Assumption 1.

Assumption 2. $\beta^t\psi = 0$ whenever $X\psi = 0$, a.s.