

UC Irvine

UC Irvine Previously Published Works

Title

Generalized orthogonal components regression for high dimensional generalized linear models

Permalink

<https://escholarship.org/uc/item/2198t9xk>

Journal

Computational Statistics & Data Analysis, 88

ISSN

0167-9473

Authors

Lin, Yanzhu

Zhang, Min

Zhang, Dabao

Publication Date

2015-08-01

DOI

10.1016/j.csda.2015.02.006

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at

<https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed



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.

© 2015 Elsevier B.V. All rights reserved.

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).

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,

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

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.,

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

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 $\{(y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$, a common issue is how to provide a legitimate estimate of β in (2) when $p \gg n$. Denote $\mathbf{X} = (\mathbf{x}_1^t, \dots, \mathbf{x}_n^t)^t$, an $n \times p$ matrix with rank $r_x \leq \min(n, p)$. 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 $\mathbf{X}_s$ is constructed by a subset of columns of $\mathbf{X}$, and further assume that there is a unique maximum likelihood estimator of β_s for the following model,

$$g(E[Y|X_s]) = \mu + X_s \beta_s. \quad (3)$$

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 j th 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.

In the next section, we consider the construction of the GOCRE model for any random pair (Y, X) from the population. GOCRE sequentially builds orthogonal components $X\varpi_j$, $j = 1, 2, \dots$, to account for the variation of the categorical outcomes. For the same reason mentioned above, each ϖ_j satisfies the following assumption, leading to β satisfying [Assumption 2](#) when a full set of components are used to build model [\(2\)](#).

Assumption 3. $\varpi_j^t \psi = 0$ whenever $X\psi = 0$, a.s.

2.2. Generalized orthogonal components regression (GOCRE)

The orthogonal components will be sequentially constructed with a prespecified weight w for each pair of Y and X in the whole population such that $E[w]$ is finite. We further assume that $E[wX] = \mathbf{0}_p^t$, where $\mathbf{0}_p$ is a p -dimensional column vector with all components as zero. Note that such a weighted centralization of the random vector X plays an important role in carrying out GOCRE. Since GOCRE constructs orthogonal components relying on a linear regression model whose response value changes at each iteration, the intercept has to be updated at each iteration (unlike PLS which removes intercept from the regression model). This weighted centralization allows separate calculation of the intercept and orthogonal components. For convenience, we denote $\nabla g^{-1}(\eta) = dg^{-1}(\eta)/d\eta$ in the following.

First, let $X_1 = X$ and for a specific η , i.e., $\eta = \eta^{(0)}$, we calculate

$$Z(\eta) = \eta + \{Y - g^{-1}(\eta)\} / \nabla g^{-1}(\eta). \quad (4)$$

A component $X_1\alpha(\eta)$ can be constructed with $\alpha = \alpha(\eta)$ maximizing $\|E[Z(\eta)wX_1\alpha]\|^2$ under the condition $\|\alpha\| = 1$. With a scalar variable $Z(\eta)$, we indeed have

$$\alpha(\eta) = E[X_1^t wZ(\eta)] / \|E[X_1^t wZ(\eta)]\|.$$

Then regressing $Z = Z(\eta)$ against $X_1\alpha$ with $\alpha = \alpha(\eta)$ leads to an update of η ,

$$\eta(\alpha) = E[wZ]/E[w] + X_1\alpha\gamma_1, \quad (5)$$

where $\gamma_1 = E[\alpha^t X_1^t wZ] / E[\alpha^t X_1^t wX_1\alpha]$. Alternatively update $\alpha(\eta)$ and $\eta(\alpha)$ until $\alpha(\eta)$ converges to α_1 , which leads to the construction of the first component $X_1\alpha_1$.

After constructing the j th component $X_j\alpha_j$, we remove $X_j\alpha_j$ from X_j such that $X_{j+1} = X_j - X_j\alpha_j\theta_j$ is orthogonal to $X_j\alpha_j$, i.e.,

$$E[X_{j+1}^t wX_j\alpha_j] = 0 \implies \theta_j = E[\alpha_j^t X_j^t wX_j] / E[\alpha_j^t X_j^t wX_j\alpha_j]. \quad (6)$$

Since

$$X_j\alpha_j = X_{j-1}(I - \alpha_{j-1}\theta_{j-1})\alpha_j = \dots = X \left\{ \prod_{l=1}^{j-1} (I - \alpha_{j-l}\theta_{j-l}) \right\} \alpha_j,$$

we have the following proposition.

Proposition 1. Each component $X_j\alpha_j$ can be rewritten as $X\varpi_j$ where

$$\varpi_j = \left\{ \prod_{l=1}^{j-1} (I - \alpha_{j-l}\theta_{j-l}) \right\} \alpha_j.$$

Furthermore, with the inner product defined as $\langle x, y \rangle = E[xwy]$, the components $X\varpi_1, X\varpi_2, \dots$, are orthogonal.

Through these first j orthogonal components, we can obtain an estimate of η , say η_j . Taking $\eta = \eta_j$, we calculate $Z(\eta)$ following [\(4\)](#). A component $X_{j+1}\alpha(\eta)$ is then constructed with

$$\alpha(\eta) = \arg \max_{\alpha: \|\alpha\|=1} \{\|E[Z(\eta)wX_{j+1}\alpha]\|^2\} = E[X_{j+1}^t wZ(\eta)] / \|E[X_{j+1}^t wZ(\eta)]\|. \quad (7)$$

Regressing $Z = Z(\eta)$ against $X_{j+1}\alpha(\eta)$ as well as the first j components leads to an update of η ,

$$\eta(\alpha) = E[wZ]/E[w] + \sum_{k=1}^j X_k\alpha_k\gamma_k + X_{j+1}\alpha\gamma, \quad (8)$$

where $\gamma = E[\alpha^t X_{j+1}^t wZ] / E[\alpha^t X_{j+1}^t wX_{j+1}\alpha]$, and $\gamma_k = E[\alpha_k^t X_k^t wZ] / E[\alpha_k^t X_k^t wX_k\alpha_k]$ for $k = 1, \dots, j$. Alternatively update $\alpha(\eta)$ and $\eta(\alpha)$ until $\alpha(\eta)$ converges to α_{j+1} , which leads to the construction of the $(j+1)$ -st component $X_{j+1}\alpha_{j+1}$.

Such construction stops whenever $w^{1/2}Z(\eta)$ is uncorrelated to $w^{1/2}X_{j+1}$. Upon completion of the construction, $w^{1/2}X\varpi_1, w^{1/2}X\varpi_2, w^{1/2}X\varpi_3, \dots$, are uncorrelated, which lead to the generalized orthogonal-components regression model with orthogonal components $X\varpi_1, X\varpi_2, X\varpi_3, \dots$.

Proposition 2. Upon completion of the construction, we can build up the generalized orthogonal-components regression model,

$$g(E[Y|X]) = \mu + \sum_j \vartheta_j (X\varpi_j), \quad (9)$$

where ϖ_j , $j = 1, 2, \dots$, are as specified in Proposition 1, and ϑ_j , $j = 1, 2, \dots$, are the regression coefficients of the corresponding orthogonal components. Furthermore each ϖ_j satisfies Assumption 3, and $\beta = \sum_j \vartheta_j \varpi_j$ satisfies Assumption 2.

Proof. When $X_1\psi = X\psi = 0$, a.s., $\alpha_1^t\psi = 0$ following (7). It leads to $X_2\psi = 0$, a.s. Iteratively we have $X_j\psi = 0$, a.s., and $\alpha_j^t\psi = 0$. Hence $\varpi_j^t\psi = 0$, $j = 1, 2, \dots$, which leads to $\beta^t\psi = 0$.

2.3. The algorithm

With observed data $\mathbf{Y} = (y_1, \dots, y_n)^t$ and $\mathbf{X} = (\mathbf{x}_1^t, \dots, \mathbf{x}_n^t)^t$, we can follow the above idea to sequentially construct orthogonal components accounting for the variation in $\mathbf{Y}$, and also provide an estimate of β satisfying Assumption 1. The construction proceeds on the basis of prespecified weight w_i for the i th observation. We denote $W = \text{diag}\{w_1, \dots, w_n\}$. Without loss of generality, we further assume that $\mathbf{X}_1 = \mathbf{X}$ has been column-wisely centralized, i.e., $\mathbf{X}^t W \mathbf{1}_n = \mathbf{0}_p$, where $\mathbf{1}_n$ is an n -dimensional column vector with all components as one.

Suppose that components $\mathbf{X}_1\alpha_1, \dots, \mathbf{X}_{j-1}\alpha_{j-1}$ have been constructed, η_{j-1} is output from the construction of the $(j-1)$ -st component $\mathbf{X}_{j-1}\alpha_{j-1}$, and $\mathbf{X}_j$ is also constructed. We can therefore proceed to construct the j th component $\mathbf{X}_j\alpha_j$, η_j , and $\mathbf{X}_{j+1}$ as follows:

1. Initialize $\eta_j = \eta_{j-1}$;
2. Update $\mathbf{Z} = \eta_j + H^{-1}\{\mathbf{Y} - g^{-1}(\eta_j)\}$, with $H = \text{diag}\{\nabla g^{-1}(\eta_{j1}), \dots, \nabla g^{-1}(\eta_{jn})\}$;
3. Update $\mu = \mathbf{1}_n^t W \mathbf{Z} / \{\mathbf{1}_n^t W \mathbf{1}_n\}$;
4. Update $\alpha_j = \mathbf{X}_j^t W \mathbf{Z} / \|\mathbf{X}_j^t W \mathbf{Z}\|$;
5. Update $\gamma_k = \alpha_k^t \mathbf{X}_k^t W \mathbf{Z} / \{\alpha_k^t \mathbf{X}_k^t W \mathbf{X}_k \alpha_k\}$ for $k = 1, \dots, j$;
6. Update $\eta_j = \mu \mathbf{1}_n + \sum_{k=1}^j \mathbf{X}_k \alpha_k \gamma_k$;
7. Iterate between 2 and 6 until α_j converges;
8. Calculate $P_j = \alpha_j^t \mathbf{X}_j^t W \mathbf{X}_j / \{\alpha_j^t \mathbf{X}_j^t W \mathbf{X}_j \alpha_j\}$, and $\mathbf{X}_{j+1} = \mathbf{X}_j - \mathbf{X}_j \alpha_j P_j$.

Note that $\eta_j = (\eta_{j1}, \dots, \eta_{jn})^t$. In Step 2, we also abuse the notations by defining $g^{-1}(\eta_j) = (g^{-1}(\eta_{j1}), \dots, g^{-1}(\eta_{jn}))^t$.

Remark 1. For each k , $\mathbf{X}_k = \mathbf{X}_{k-1}(I_p - \alpha_{k-1}P_{k-1}) = \dots = \mathbf{X} \times \prod_{l=1}^{k-1} (I_p - \alpha_l P_l)$, therefore $\mathbf{X}_k^t W \mathbf{1}_n = \mathbf{0}_p$ following $\mathbf{X}^t W \mathbf{1}_n = \mathbf{0}_p$. The weighted least squares estimation equation $\mathbf{1}_n^t W \mathbf{Z} = \mathbf{1}_n^t W (\mu \mathbf{1}_n + \sum_{k=1}^j \gamma_k \mathbf{X}_k \alpha_k)$ leads to Step 3.

Remark 2. Calculation of P_j in Step 8 implies that $\mathbf{X}_{j+1}^t W \mathbf{X}_j \alpha_j = 0$. Iteratively, it leads to $\mathbf{X}_{j+1}^t W \mathbf{X}_k \alpha_k = 0$ for $k = j, j-1, \dots, 1$. That is, the components $\mathbf{X}_1\alpha_1, \mathbf{X}_2\alpha_2, \mathbf{X}_3\alpha_3, \dots$, are orthogonal when the inner product is defined as $\langle x, y \rangle = x^t W y$.

Remark 3. Step 5 follows the application of $\mathbf{X}_k^t W \mathbf{1}_n = \mathbf{0}_p$ and $\alpha_k^t \mathbf{X}_k^t W \mathbf{X}_l \alpha_l = 0$, $l \neq k$, to the weighted least squares estimation equation $\alpha_k^t \mathbf{X}_k^t W \mathbf{Z} = \alpha_k^t \mathbf{X}_k^t W (\mu \mathbf{1}_n + \sum_{l=1}^j \gamma_l \mathbf{X}_l \alpha_l)$.

Remark 4. From the construction of the last component, say $\mathbf{X}_m \alpha_m$, we have the estimate

$$\hat{\beta} = \sum_{j=1}^m \left\{ \prod_{k=1}^{j-1} (I_p - \alpha_k P_k) \right\} \alpha_j \gamma_j,$$

which can be sequentially updated and satisfies Assumption 1. The parameter μ in the model (2) can be estimated by μ upon constructing the last component.

Compared to the original model (2), the GOCRE model (9) not only present the unique MLE satisfying Assumption 1, but also calculate the MLE without computing inverse of any matrix. While both features are desirable in analyzing $p \gg n$ data, the latter one particularly speeds up the calculation.

2.4. Selection of weights

Note that, for a specific η , $Z(\eta)$ in (4) has the variance

$$\text{var}(Z(\eta)) = b''(\theta) a(\phi) / \{\nabla g^{-1}(\eta)\}^2.$$

Therefore, it is preferred to have a dynamic weight $w(\eta) \propto 1/\text{var}(Z(\eta))$. However, such a dynamic weight makes it impossible to construct orthogonal components in any specific inner product space.

One strategy is to take dynamic weights when the first component is being iteratively constructed with the identity matrix as the initial value. Once the first component is constructed, we have a converged weight matrix, and therefore use this weight matrix for constructing all subsequent components.

Another strategy is to run the aforementioned algorithm twice. The first run of the algorithm may take an identity weight matrix or use the above strategy to construct a weight matrix. The second run can construct a weight matrix based on the η value from the last step of the previous run.

Our simulation study demonstrated that the first strategy usually performs well and there is negligible gain in taking a second run of the algorithm (results not shown).

2.5. Convergence failures and bias correction

For some GLMs, especially the logistic regression model for binary responses, MLE may not exist due to complete separation, or quasicomplete separation of different categories (Albert and Anderson, 1984), and it usually results in non-convergence of the corresponding algorithm. Heinze and Schemper (2002) proposed that the penalized likelihood method by Firth (1993) can solve the convergence problem due to the aforementioned separation issue.

Suppose the model in (2) has log-likelihood $\ell(\mu, \beta)$ and information matrix $I(\mu, \beta)$. Instead of directly maximizing the log-likelihood function, Firth (1993) proposed to maximize the penalized log-likelihood

$$\ell^*(\mu, \beta) = \ell(\mu, \beta) + \frac{1}{2} \log\{|I(\mu, \beta)|\},$$

where the penalty corresponds to the Jeffreys invariant prior (Jeffreys, 1946). Firth (1993) initially took this modification to reduce the bias of maximum likelihood estimates, and showed that the first order bias can be removed.

For logistic regression, we will modify our algorithm using the same idea to reduce the bias and solve the non-convergence issue. Note that only Step 2 of the algorithm in Section 2.3 needs to be modified. Define

$$\begin{cases} \Delta = (\delta_{kl})_{n \times n} \triangleq W^{\frac{1}{2}} \mathbf{X}(\mathbf{X}^t W \mathbf{X})^+ \mathbf{X}^t W^{\frac{1}{2}}, \\ \zeta \triangleq (\delta_{11}, \dots, \delta_{nn})^t, \\ \Lambda \triangleq \text{diag}\{\delta_{11}, \dots, \delta_{nn}\}, \end{cases}$$

where $(\mathbf{X}^t W \mathbf{X})^+$ is a Moore–Penrose pseudo-inverse. We then replace Step 2 with the following steps,

- 2a. Calculate $H = \text{diag}\{(1 + \delta_{11})\nabla g^{-1}(\eta_{j1}), \dots, (1 + \delta_{nn})\nabla g^{-1}(\eta_{jn})\}$;
- 2b. Update $\mathbf{Z} = \eta_j + H^{-1}\{\mathbf{Y} + \frac{1}{2}\zeta - (\mathbf{I}_n + \Lambda)g^{-1}(\eta_j)\}$.

Note that, once the weight matrix W is fixed, we then have fixed values of δ_{kk} , $k = 1, \dots, n$. Therefore, unlike the IRPLS method modified by Ding and Gentleman (2004) which requires re-calculation of the weight matrix at each iteration, we do not need this re-calculation when constructing all components other than the first component, because of the fixed weight matrix W . Since high-dimensional data imply large matrices involved in calculation of the weight matrix, our algorithm can be more efficient in terms of computational cost.

We use GOCRE₀ to refer to the above implementation for bias correction. As shown in the following, the calculation of Δ can be simplified through a singular value decomposition, which will essentially speed up the computation of GOCRE₀.

For large n , it is still computationally intensive to calculate and maintain the matrix Δ . Instead, Chung and Keles (2010) approximate each diagonal component of Δ with $\text{trace}\{\Delta\}/n$. A similar strategy has been utilized in constructing generalized cross-validation (Golub et al., 1979). Observing that Δ is usually of full rank and therefore $\text{trace}\{\Delta\}/n = 1$ when $p \gg n$, Chung and Keles (2010) always take $\zeta = \mathbf{1}_n$ and $\Lambda = \mathbf{I}_n$. Indeed, as shown below, $\Delta = \mathbf{I}_n$ when $p \geq n$ and $\mathbf{X}$ is of full rank.

Proposition 3. Assume $\text{rank}(\mathbf{X}) = k$ and therefore the singular value decomposition $W^{1/2}\mathbf{X} = U\Omega V^t$ where U is $n \times k$ with $U^t U = \mathbf{I}_k$, Ω is a $k \times k$ diagonal matrix with positive diagonal elements, and V is $p \times k$ with $V^t V = \mathbf{I}_k$. Then $\Delta = UU^t$. Furthermore, if $k = n$, then $\Delta = \mathbf{I}_n$; if $k = n - 1$ and $\mathbf{1}_n^t W \mathbf{X} = \mathbf{0}_p^t$, then $\Delta = \mathbf{I}_n - W^{1/2} \mathbf{1}_n \mathbf{1}_n^t W^{1/2} / \|\mathbf{1}_n\|_1$, where $\|\mathbf{1}_n\|_1 = \sum_{i=1}^n w_i$.

As we preprocess $\mathbf{X}$ such that $\mathbf{1}_n^t W \mathbf{X} = \mathbf{0}_p^t$, Δ will usually have the rank of $n - 1$ instead of full rank when $p \gg n$. The above proposition implies that $\zeta = \mathbf{1}_n - W \mathbf{1}_n / \|\mathbf{1}_n\|_1$ and $\Lambda = \mathbf{I}_n - W / \|\mathbf{1}_n\|_1$. Hereafter, we will use GOCRE to refer to such an implementation, that is, taking $\delta_{kk} = 1 - w_k / \sum_{i=1}^n w_i$, $k = 1, 2, \dots, n$.

3. Simulation studies

We simulated large p small n data to evaluate the performance of GOCRE and compared it with IRPLS implemented by Marx (1996) and Ding and Gentleman (2004), which are hereafter denoted by IRPLS-M and IRPLS-DG respectively. The underlying models take the logit link function in (2) with $\mu = 0$ and $p = 1000$. The predictors were divided into ten blocks, where each block was simulated from an AR(1) process with the correlation ρ prespecified at $\rho = 0, 0.3, 0.5$, and 0.7 respectively. The regression coefficients $\{\beta_j, 1 \leq j \leq p\}$, were generated from a Laplace distribution with location parameter two and scale parameter one.

Table 1

Convergence frequencies of different methods in analyzing simulated data.

Methods	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
IRPLS – M	0%	0%	0%	0%
IRPLS – DG	79%	82%	77%	94%
GOCRE ₀	100%	100%	100%	100%
GOCRE	100%	100%	100%	100%

Table 2

Performance comparison in analyzing simulated data. Reported are the median MR and PRESS across 100 simulated data sets, with standard errors presented in the parentheses.

Criterion	Model	IRPLS-M	IRPLS-DG	GOCRE ₀	GOCRE
MR	$\rho = 0.0$	0.4350(0.0313)	0.4250(0.0349)	0.4250 (0.0330)	0.4275(0.0332)
	$\rho = 0.3$	0.3900(0.0350)	0.3950(0.0364)	0.3825 (0.0365)	0.3850 (0.0365)
	$\rho = 0.5$	0.3525(0.0378)	0.3450(0.0337)	0.3350 (0.0336)	0.3350 (0.0336)
	$\rho = 0.7$	0.3050(0.0337)	0.2900(0.0310)	0.2850 (0.0347)	0.2850 (0.0346)
PRESS	$\rho = 0.0$	0.2671(0.0169)	0.2414(0.0057)	0.2405 (0.0058)	0.2405 (0.0058)
	$\rho = 0.3$	0.2475(0.0217)	0.2330(0.0064)	0.2313 (0.0066)	0.2312 (0.0067)
	$\rho = 0.5$	0.2290(0.0195)	0.2223(0.0069)	0.2208 (0.0072)	0.2207 (0.0072)
	$\rho = 0.7$	0.2004(0.0185)	0.2034(0.0073)	0.2034(0.0084)	0.2033(0.0085)

For each different ρ , we conducted 100 simulation studies. In each simulation study, we simulated three independent sets of data: a training set with 100 observations, a validation set with 100 observations and a test set with 200 observations. Each method was used to fit the models using the training data, and the optimal number of components was chosen using the validation data. The maximum number of components is ten for all methods. The performance was evaluated based on the misclassification rate (MR) and sum of squares of the prediction residuals (PRESS) calculated from the test data.

Shown in Table 1 are the frequencies of each method which has converged in analyzing 100 simulated data sets with each specific ρ . It is well known that there is a divergence problem for both IRPLS implementations (Ding and Gentleman, 2004; Fort and Lambert-Lacroix, 2005; Boulesteix and Strimmer, 2006; Chung and Keles, 2010). Indeed, the IRPLS-M did not converge in analyzing any of the simulated data sets. IRPLS-DG partially solved this problem through Firth's procedure. However, it still did not converge in analyzing, for example, 23% of the data sets with $\rho = 0.5$. On the other hand, both GOCRE₀ and GOCRE converged in all data analyses.

The MR and PRESS in analyzing different models are shown in Table 2 for all methods. We observe that, for each method, the higher the correlation among the predictors, the lower MR and PRESS. For either GOCRE₀ or GOCRE, bold MR and PRESS values indicate better performance than IRPLS-M as well as IRPLS-DG. Interestingly, IRPLS-DG reported smaller MR than IRPLS-M except for the case of $\rho = 0.3$, while IRPLS-DG always reported smaller PRESS than IRPLS-M except for the case of $\rho = 0.7$. In all cases, GOCRE₀ performed better than both IRPLS methods in terms of either criterion, except that IRPLS-M reported the smallest PRESS when $\rho = 0.7$. Indeed, GOCRE reported larger MR than IRPLS-DG only in the case of $\rho = 0.0$, but performed better than both IRPLS methods in all other cases. In addition to their competitive performance and solving the convergence issue, GOCRE₀ and GOCRE also enjoy advantage over the other two methods in computing time which is a critical issue in analyzing high-dimensional data. As shown in the next section, both GOCRE₀ and GOCRE can dramatically reduce the computing time.

As shown in Proposition 3, GOCRE₀ and GOCRE should report exactly the same results since we preprocessed $\mathbf{X}$ in simulated data such that $\mathbf{1}_n' \mathbf{W} \mathbf{X} = \mathbf{0}_p'$. However, $\mathbf{1}_n' \mathbf{W} \mathbf{X} = \mathbf{0}_p'$ could not be computationally obtained due to the computer precision. For example, our computation in MATLAB would return centralized $\mathbf{X}$ with column means in the scale of 10^{-16} instead of exact zero, which resulted in the slight difference between GOCRE₀ and GOCRE.

4. Application to gene expression profiling

Here we use the lung cancer data set (Gustafson et al., 2010) from Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) to illustrate the performance of GOCRE when compared to the two different IRPLS implementations. In this data set, a total of 187 arrays were used to monitor the expression levels of 22,215 genes from 97 cancer patients and 90 healthy individuals. For each gene, a p -value was obtained from the Wilcoxon rank-sum test. Ranking the genes ascendingly on the basis of their p -values, we constructed four data sets by including the top 1000, 2000, 5000, and all genes respectively.

For each data set, we randomly selected one-quarter of the samples to form the test data (with 24 cancer patients and 23 normal persons) and used the rest as training data. We applied each method to build up the models with different number of components (up to 20 components) using the training data, and then calculated MR and PRESS of each model based on the test data. The results are shown in Figs. 1 and 2.

As shown in Fig. 1, IRPLS-M did not converge in constructing almost every component. Although IRPLS-DG converged in constructing majority of the components, it did not converge in constructing some components in analyzing each of the four data sets. As expected, both GOCRE₀ and GOCRE performed similarly, and converged in constructing each component.

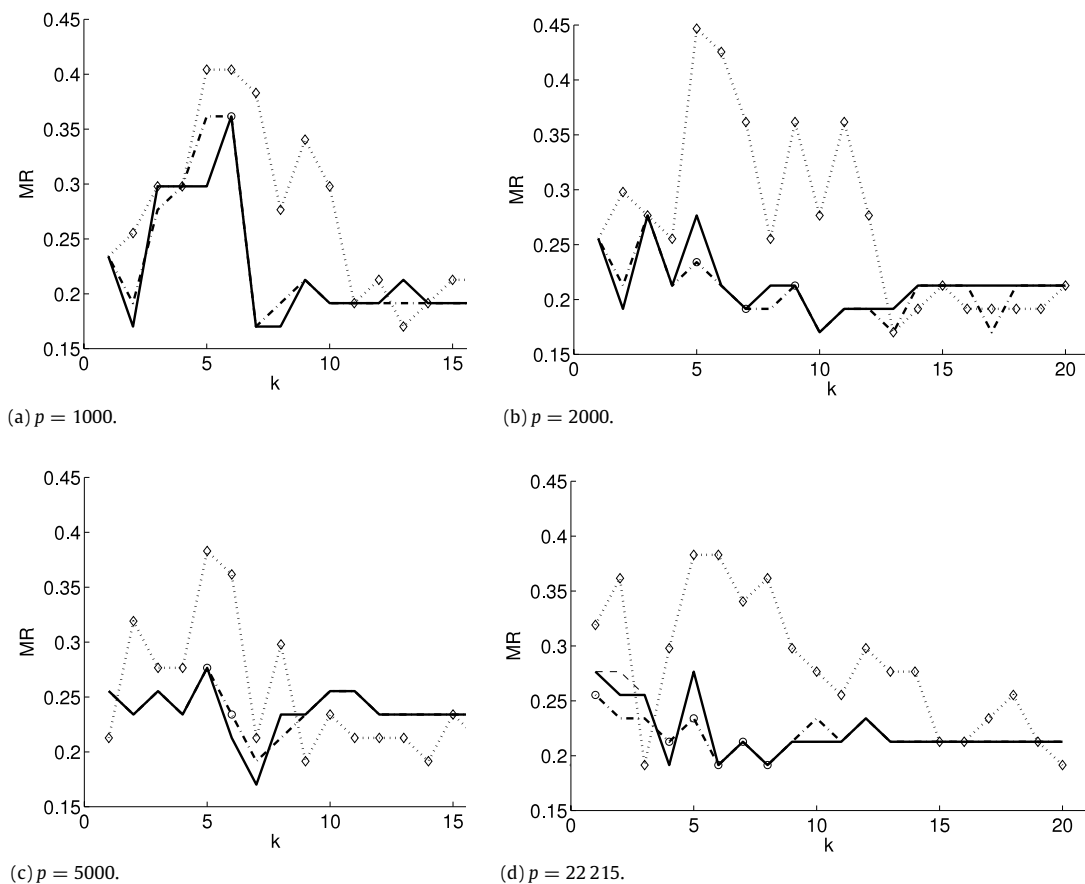


Fig. 1. Misclassification rates (MR) in analyzing the lung cancer data. The results were obtained using IRPLS-M (dotted lines), IRPLS-DG (dashed dotted lines), GOCRE₀ (dashed lines), and GOCRE (solid lines) respectively to build up models with different number of components (k). Non-converged methods were marked by diamonds for IRPLS-M, and circles for IRPLS-DG.

Except for the case with $p = 5000$ where GOCRE₀ and GOCRE might obtain smaller MR than other methods, all methods obtained the same smallest MR in other cases.

Similar to the performance in terms of MR in Fig. 1, IRPLS-M presented rather wildly varying PRESS when different number of components were considered. The other three methods instead presented very stable and similar PRESS while the PRESS of IRPLS-DG are slightly smaller than those of GOCRE₀ and GOCRE. Note that the slight advantage of IRPLS-DG in PRESS did not imply improvement in MR. Indeed, IRPLS-DG and the proposed methods perform quite similarly across different models.

We conducted all the data analyses on a UNIX server with eight CPUs (Intel® Xeon® Processor E5410), a total of 32 GB RAM, and MATLAB® 7.12. Not surprisingly, both GOCRE₀ and GOCRE were much faster than the other two methods because the IRPLS implementations need to reconstruct a set of components at each iteration, but GOCRE₀ and GOCRE sequentially construct all components. As shown in Table 3, they took much less computing time than the other two methods when analyzing the lung cancer data using GLMs with different number of components (up to 20 components). Indeed, for each GLM model with k components, both IRPLS implementations need to update the predictor matrix k times within each iteration. For large p small n data set which produces a high dimensional predictor matrix, it is time consuming to update the predictor matrix.

5. Discussion

Zhang et al. (2009) proposed an orthogonal-component regression (OCRE) for supervised construction of principal components which account for the variation of continuous responses. OCRE can be considered as an alternative implementation of PLS. We here propose GOCRE which extends OCRE for GLMs, focusing on binary outcomes. Such an extension makes it feasible to extend POCRE in Zhang et al. (2009) for GLMs, allowing to select variables from a large amount of candidates. However, the need of iterated procedures like IRLS for fitting classical GLMs challenges the extension. One challenge follows the use of weighted linear models to construct the components. Available components would suggest a weighted linear

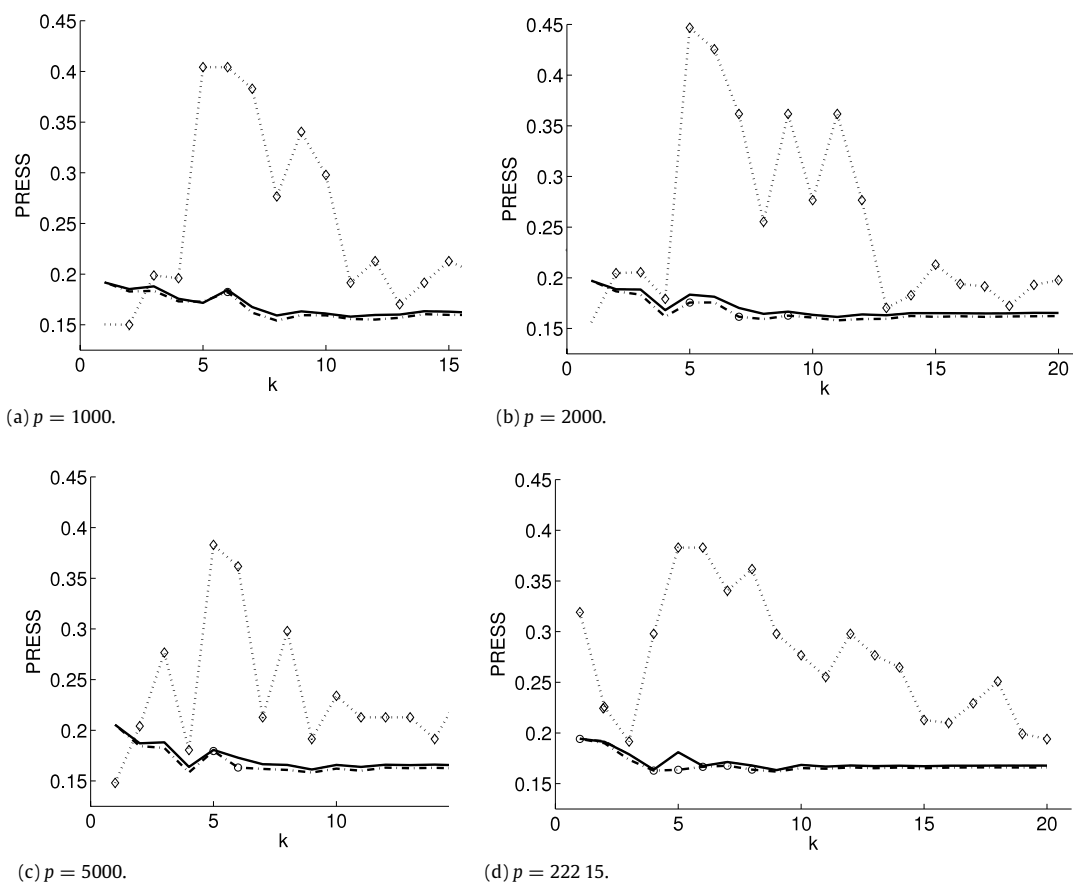


Fig. 2. Sum of squares of the prediction residuals (PRESS) in analyzing the lung cancer data. The results were obtained using IRPLS-M (dotted lines), IRPLS-DG (dashed dotted lines), GOCRE₀ (dashed lines), and GOCRE (solid lines) respectively. Non-converged methods were marked by diamonds for IRPLS-M, and circles for IRPLS-DG.

Table 3

Computation time (seconds) in analyzing the lung cancer data.

p	IRPLS-M	IRPLS-DG	GOCRE ₀	GOCRE
1 000	2,384	263	8	7
2 000	5,334	609	15	12
5 000	11,750	819	35	28
22 215	48,531	2972	443	370

model with updated weights when constructing a new component. Indeed, any update of the same component in each iteration would suggest a new set of weights. Here we suggest to fix the weights upon the construction of the first component, which usually provides a satisfactory set of orthogonal components. A fixed set of weights can also be suggested using some other preliminary analysis of the data.

GOCRE is proposed to sequentially construct a set of orthogonal components and it allows to investigate the relationship between a categorical outcome and a set of variables of interest. Such sequential construction leads to efficient computation in GOCRE while GOCRE performs competitive to IRPLS-DG. As shown in Figs. 1 and 2, GOCRE performs robust over different numbers of components. This is a very important property of supervised dimension reduction as it is important to predetermine the number of components and such robustness buffers less-optimized choice of the number of components.

IRPLS-DG improves IRPLS-M over the non-convergence of regression coefficients. However, it still repetitively constructs a different set of components at each iteration. The convergence of constructed components could be very important in some applications. For example, next generation sequencing data always present a large number of rare genetic variants (Shendure and Ji, 2008). Therefore, GOCRE can play an important role to consolidate rare genetic variants, and reveal genetic architecture underlying phenotypes of interest (Wang et al., 2011).

Appendix. Proof of Proposition 3

Expand V to $\tilde{V} = (V \ V_c)$ such that $\tilde{V}^t \tilde{V} = I_p$, and further expand U to $\tilde{U}$ such that $\tilde{U}^t \tilde{U} = I_p$ in the following way:

$$\tilde{U} = \begin{cases} (U \ U_c), & \text{if } p \leq n; \\ \begin{pmatrix} U & U_c \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, & \text{if } p > n. \end{cases}$$

Accordingly, we expand Λ to a $p \times p$ matrix $\tilde{\Lambda}$ as follows:

$$\tilde{\Lambda} = \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Let $\mathbf{X}_w = W^{1/2} \mathbf{X}$ and $\tilde{\mathbf{X}}_w = \tilde{U} \tilde{\Lambda} \tilde{V}^t$, then $\mathbf{X}_w^t \mathbf{X}_w = \tilde{\mathbf{X}}_w^t \tilde{\mathbf{X}}_w = \tilde{V} \tilde{\Lambda}^2 \tilde{V}^t$, which implies that $(\mathbf{X}_w^t \mathbf{X}_w)^+ = \tilde{V} \tilde{\Lambda}^{+2} \tilde{V}^t$, and

$$\tilde{\Lambda}^+ = \begin{pmatrix} \Lambda^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

It follows that $\mathbf{X}_w (\mathbf{X}_w^t \mathbf{X}_w)^+ \mathbf{X}_w^t = U \Lambda V^t \tilde{V} \tilde{\Lambda}^{+2} \tilde{V}^t V \Lambda U^t = U U^t$. That is, $\Delta = U U^t$.

When $k = n$, U is an orthonormal matrix which implies that $\Delta = U U^t = I_n$.

When $k = n - 1$ and $\mathbf{1}_n^t W \mathbf{X} = (W^{1/2} \mathbf{1}_n)^t \mathbf{X}_w = \mathbf{0}_p^t$, then $\tilde{U} = (U \ W^{1/2} \mathbf{1}_n / \sqrt{\|W\|_1})$ is an orthonormal matrix. That is, $I_n = \tilde{U} \tilde{U}^t = U U^t + W^{1/2} \mathbf{1}_n \mathbf{1}_n^t W^{1/2} / \|W\|_1$, which implies $\Delta = U U^t = I_n - W^{1/2} \mathbf{1}_n \mathbf{1}_n^t W^{1/2} / \|W\|_1$.

References

- Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71, 1–10.
- Boulesteix, A.-L., Strimmer, K., 2006. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* 8, 32–44.
- Chung, D., Keles, S., 2010. Sparse partial least squares classification for high dimensional data. *Stat. Appl. Genet. Mol. Biol.* 9 (1), Article 17.
- Ding, B., Gentleman, R., 2004. Classification using generalized partial least squares. *J. Comput. Graph. Statist.* 14, 280–298.
- Firth, D., 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38.
- Fort, G., Lambert-Lacroix, S., 2005. Classification using partial least squares with penalized logistic regression. *Bioinformatics* 21, 1104–1111.
- Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215–223.
- Green, P.J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Stat. Soc. Ser. B* 46, 149–192.
- Gustafson, A.M., Soldi, R., Anderlind, C., Scholand, M.B., Qian, J., Zhang, X., Cooper, K., Walker, D., McWilliams, A., Liu, G., Szabo, E., Brody, J., Massion, P.P., Lenburg, M.E., Lam, S., Bild, A.H., Spira, A., 2010. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transl. Med.* 2, 26ra25.
- Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. *Stat. Med.* 21, 2409–2419.
- Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* 186, 453–461.
- Marx, B., 1996. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* 38, 374–381.
- Nguyen, D.V., Rocke, D.M., 2002. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18, 39–50.
- Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnol.* 26, 1135–1145.
- Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H., 2010. *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Springer, Berlin.
- Wang, L., Pungpapont, V., Lin, Y., Zhang, M., Zhang, D., 2011. Genome-wide case-control study in GAW17 using coalesced rare variants. *BMC Proc.* 5 (Suppl 9), S110.
- Wold, H., 1975. Soft modelling by latent variables: the nonlinear iterative partial least squares approach. In: Gani, J. (Ed.), *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*. Academic Press, London.
- Zhang, D., Lin, Y., Zhang, M., 2009. Penalized orthogonal-components regression for large p small n data. *Electron. J. Stat.* 3, 781–796.