Journal of Data Science 23 (4), 607–623
October 2025
DOI: 10.6339/25-JDS1188
Statistical Data Science

Exploring Massive Risk Factors of Categorical Outcomes via Supervised Dimension Reduction

Yan Li[1], Kangni Alemdjrodo[2], Yanzhu Lin[3], Min Zhang[1], and Dabao Zhang[1,*]

[1]*Department of Epidemiology and Biostatistics, University of California, Irvine, CA 92617, United States*
[2]*Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, United States*
[3]*Eli Lilly and Company, Indianapolis, IN 46285, United States*

Abstract

We propose to explore high-dimensional data with categorical outcomes by generalizing the penalized orthogonal-components regression method (POCRE), a supervised dimension reduction method initially proposed for high-dimensional linear regression. This generalized POCRE, i.e., gPOCRE, sequentially builds up orthogonal components by selecting predictors which maximally explain the variation of the response variables. Therefore, gPOCRE simultaneously selects significant predictors and reduces dimensions by constructing linear components of these selected predictors for a high-dimensional generalized linear model. For multiple categorical outcomes, gPOCRE can also construct common components shared by all outcomes to improve the power of selecting variables shared by multiple outcomes. Both simulation studies and real data analysis are carried out to illustrate the performance of gPOCRE.

Keywords *gPOCRE; latent model; logistic regression; multinomial regression; orthogonal components*

1 Introduction

High-dimensional data with categorical outcomes, such as data from genome-wide association studies of single or multiple related diseases, challenge statistical inference as we are usually interested in building models to distinguish the different groups and identifying risk factors that cause such classification (Tam et al., 2019). Many tools have been developed to address these two issues in analyzing classical categorical data with few risk factors of interest, for example, logistic regression, linear discriminant analysis (LDA) (Fisher, 1936), and classification tree (CT) (Loh, 2011). Since high-dimensional categorical data usually come with a massive number of features but a relatively small sample size, direct application of these classical methods is either computationally infeasible or methodologically inappropriate (Fan et al., 2009; Xie et al., 2020).

Classification based on generalized linear regression models, such as logistic or probit regression, is challenged by multicollinearity and perfect separation due to the available massive features (Shen and Gao, 2008). Extending LASSO (Tibshirani, 1996) to generalized linear regression models still faces such challenges (Van de Geer, 2008). LDA aims to discriminate cases and controls in a one-dimensional space and avoid such challenges (McLachlan, 2005). However,

*Corresponding author. Email: dabao.zhang@uci.edu.

Received September 29, 2024; Accepted May 2, 2025

such a one-dimensional space may involve all features and is still insufficient to discriminate cases and controls. The classification tree instead builds a decision tree, leading to a conclusion on case or control based on selected features (Freeman et al., 2013). It models the nonlinear relationship between features and responses. However, evaluating the significance of individual features is not straightforward.

Supervised dimension reduction plays a vital role in exploring high-dimensional data in linear regression (Massy, 1965). While it privileges over the relation between responses and features to propose effective data analysis (Velliangiri et al., 2019), supervised dimension reduction has been studied for partial least squares (PLS) regression (Wold, 1966, 1975; Hoskuldsson, 1988, 1992; De Jong, 1993; Boulesteix and Strimmer, 2006), and further for sparse partial least squares (SPLS) (Lê Cao et al., 2008; Chun and Keleş, 2010) enabled with variable selection. Zhang et al. (2009) developed a penalized orthogonal-components regression (POCRE) to fit high-dimensional linear regression models. POCRE sequentially constructs orthogonal components for massive features to maximize, upon standardization, their correlations with the response variables. A penalization framework was implemented to select sparse features for each component. POCRE is computationally efficient owing to its sequential construction of leading sparse principal components.

Here we will extend the idea of POCRE to generalized linear models, especially multinomial logistic regression models, by developing a supervised dimension reduction method allowing for selecting sparse features for categorical outcomes. The challenge here lies in the fact that fitting a generalized linear model usually relies on iteratively regressing different sets of working responses against predictors. As the regression at each iteration also presents dynamic weights for observations, Chung and Keles (2010) proposed the sparse generalized least squares (SGPLS) by taking each iteration as an independent task of supervised dimension reduction and constructing its own set of components for the underlying working responses. Therefore, these components do not converge and indeed different sets may have different numbers of components.

Lin et al. (2015) proposed a generalized orthogonal-components regression (GOCRE) to address the challenge by fixing the weights, which may be initialized via classical approaches, and targeting to sequentially construct a set of orthogonal components to maximally account for the variation in the categorical outcome. GOCRE addresses well the multicollinearity and perfect separation issues. However, GOCRE lacks the variable selection ability as it constructs each component with all available predictors. Here we will impose the variable selection ability on such a supervised dimension reduction method to build up low-dimensional linear combinations of sparse features and provide a valuable tool for exploratory analysis of high-dimensional data with categorical outcomes.

In the following sections, we first briefly review GOCRE within the context of generalized linear regression. Then, we propose gPOCRE by imposing a penalty function into the GOCRE framework. The algorithms are provided to implement gPOCRE. Section 3 and Section 4 present simulation studies and real data analysis using gene expression data. We conclude with a discussion in Section 5.

2 Generalized POCRE

For a set of high-dimensional data with multivariate response $\{(\mathbf{y}_i, \mathbf{x}_i), i = 1, 2, \ldots, n\}$, we have each response $\mathbf{y}_i$ a q-dimensional row vector and each predictor $\mathbf{x}_i$ a p-dimensional row vector.

The multivariate generalized linear model can be defined as

$$G(E[\mathbf{y}_i|\mathbf{x}_i]) = (G_1(E[\mathbf{y}_i|\mathbf{x}_i]), \ldots, G_q(E[\mathbf{y}_i|\mathbf{x}_i])) = \boldsymbol{\mu} + \mathbf{x}_i\mathbf{B}, \tag{1}$$

where μ is a q-dimensional row vector, $\mathbf{B}$ is a $p \times q$ dimension coefficient matrix, and $G(\cdot)$ is a vector of q link functions with each modeling one component of the response. In general, this model allows integration of diverse link functions.

Without loss of generality, we will focus on the case of multiple categorical outcomes modeled via multinomial logistic regression, specifically the baseline categorical model, for easy illustration. Suppose that each $\mathbf{y}_i$ is a multinomial response indicating $q+1$ categories. That is, the q-dimensional $\mathbf{y}_i$ includes q dummy variables with binary values, summing to at most one. Thus, we have the multinomial logistic regression specified with the multilogit link function,

$$G(E[\mathbf{y}_i|\mathbf{x}_i]) = \left(\log \frac{E[\mathbf{y}_{i1}|\mathbf{x}_i]}{1-\sum_{j=1}^{q} E[\mathbf{y}_{ij}|\mathbf{x}_i]}, \ldots, \log \frac{E[\mathbf{y}_{iq}|\mathbf{x}_i]}{1-\sum_{j=1}^{q} E[\mathbf{y}_{ij}|\mathbf{x}_i]} \right). \tag{2}$$

When $q = 1$, the above model reduces to a logistic regression.

Denote

$$\begin{aligned} \mathbf{X} &= (\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t)^t = (\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p}), \\ \mathbf{Y} &= (\mathbf{y}_1^t, \ldots, \mathbf{y}_n^t)^t = (\mathbf{y}_{\cdot 1}, \ldots, \mathbf{y}_{\cdot q}). \end{aligned}$$

Assume that a variance-related weight w_i has been appropriately defined, and denote $\mathbf{W} = diag(w_1, \ldots, w_n)$. Further assume that each $\mathbf{x}_{\cdot j}$ have been centered such that $\mathbf{1}_n^t\mathbf{W}\mathbf{X} = \mathbf{0}_p^t$, where $\mathbf{1}_n$ is an n-dimensional column vector with all components as one and $\mathbf{0}_p$ is an p-dimensional column vector with all components as zero.

We want to explore the model (1) in a low-dimensional space through building sparse orthogonal components $\mathbf{x}_i\varpi_j$, $j = 1, 2, \ldots$, to account for the variation of the nominal multinomial outcomes, that is,

$$G(E[\mathbf{y}_i|\mathbf{x}_i]) = \mu + \sum_j \vartheta_j(\mathbf{x}_i\varpi_j), \tag{3}$$

where each ϑ_j is a p-dimensional column vector. In addition, the components are orthogonal in a space with inner product $< z_1, z_2 >= E[z_1^t\mathbf{W}z_2]$.

We first review the GOCRE in the following and then introduce the idea of sparsifying these components to construct sparse orthogonal components sequentially.

2.1 Generalized Orthogonal-Component Regression

The orthogonal components are sequentially constructed based on the prespecified weights $\mathbf{W}$ and accordingly centralized $\mathbf{X}$.

First, let $\mathbf{X}_1 = \mathbf{X}$ and the $n \times q$ matrix η is initialized at, e.g., $\eta = \eta^{(0)}$. For convenience, we also denote

$$\eta = (\eta_1^t, \ldots, \eta_n^t)^t = (\eta_{\cdot 1}, \ldots, \eta_{\cdot q}),$$

which leads to the calculation of

$$\mathbf{Z}(\eta) = \left(\mathbf{Z}_{\cdot 1}^t(\eta), \ldots, \mathbf{Z}_{\cdot q}^t(\eta)\right)^t,$$

where

$$\mathbf{Z}_{\cdot j}(\eta) = \eta_{\cdot j} + \left(\nabla G^{-1}(\eta_{\cdot j})\right)^{-1} \left(\mathbf{y}_{\cdot j} - G^{-1}(\eta_{\cdot j})\right). \tag{4}$$

Here the function $G^{-1}(\eta)$ is the inverse function of the multilogit link function $G(\eta)$ with

$$G^{-1}(\eta) = (G^{-1}(\eta_1)^t, \ldots, G^{-1}(\eta_n)^t)^t = (G^{-1}(\eta_{\cdot 1}), \ldots, G^{-1}(\eta_{\cdot q}))$$

where, for each η_i,

$$G^{-1}(\eta_i) = \left(\frac{e^{\eta_{i1}}}{1 + \sum_{j=1}^{q} e^{\eta_{ij}}}, \ldots, \frac{e^{\eta_{iq}}}{1 + \sum_{j=1}^{q} e^{\eta_{ij}}}\right). \tag{5}$$

A component $\mathbf{X}_1\alpha(\eta)$ can be constructed with $\alpha = \alpha(\eta)$ maximizing

$$\|\mathbf{Z}(\eta)^t \mathbf{W} \mathbf{X}_1 \alpha\|^2 = \alpha^t \mathbf{X}_1^t \mathbf{W} \mathbf{Z}(\eta) \mathbf{Z}(\eta)^t \mathbf{W} \mathbf{X}_1 \alpha,$$

under the condition $\|\alpha(\eta)\| = 1$. Then regressing $\mathbf{Z} = \mathbf{Z}(\eta)$ against $\mathbf{X}_1\alpha$ with $\alpha = \alpha(\eta)$ leads to an update of η,

$$\eta(\alpha) = \mathbf{W}\mathbf{Z}/(\mathbf{1}_n^t \mathbf{W} \mathbf{1}_n) + \mathbf{X}_1 \alpha \gamma_1, \tag{6}$$

where

$$\gamma_1 = \alpha^t \mathbf{X}_1^t \mathbf{W} \mathbf{Z} / \alpha^t \mathbf{X}_1^t \mathbf{W} \mathbf{X}_1 \alpha.$$

Alternatively update $\alpha(\eta)$ and $\eta(\alpha)$ until $\alpha(\eta)$ converges to α_1, which leads to the construction of the first component $\mathbf{X}_1\alpha_1$.

After constructing the $(j-1)$-st component $\mathbf{X}_{j-1}\alpha_{j-1}$, we obtain $\mathbf{X}_j = \mathbf{X}_{j-1} - \mathbf{X}_{j-1}\alpha_{j-1}\theta_j$ by removing $\mathbf{X}_{j-1}\alpha_{j-1}$ from $\mathbf{X}_{j-1}$ such that

$$\mathbf{X}_j^t \mathbf{W} \mathbf{X}_{j-1} \alpha_{j-1} = 0,$$

i.e., $\mathbf{X}_j$ is orthogonal to $\mathbf{X}_{j-1}\alpha_{j-1}$, leading to

$$\theta_{j-1} = \alpha_{j-1}^t \mathbf{X}_{j-1}^t \mathbf{W} \mathbf{X}_{j-1} / \alpha_{j-1}^t \mathbf{X}_{j-1} \mathbf{W} \mathbf{X}_{j-1} \alpha_{j-1}. \tag{7}$$

With the estimate of η from constructing the first $j-1$ orthogonal components, we can update $\mathbf{Z}(\eta)$ following (4). Then the component $\mathbf{X}_j\alpha(\eta)$ can be constructed with

$$\alpha(\eta) \quad = \quad \arg \max_{\alpha : \|\alpha\| = 1} \{\|\mathbf{Z}(\eta)^t \mathbf{W} \mathbf{X}_j \alpha\|^2\}, \tag{8}$$

where $\alpha(\eta)$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{X}_j^t \mathbf{W} \mathbf{Z}(\eta) \mathbf{Z}(\eta)^t \mathbf{W} \mathbf{X}_j$.

Regressing $\mathbf{Z} = \mathbf{Z}(\eta)$ against $\mathbf{X}_j\alpha(\eta)$ as well as the other j components updates η as,

$$\eta(\alpha) = \mathbf{W}\mathbf{Z}/(\mathbf{1}_n^t \mathbf{W} \mathbf{1}_n) + \sum_{k=1}^{j-1} \mathbf{X}_k \alpha_k \gamma_k + \mathbf{X}_j \alpha \gamma, \tag{9}$$

where, for $k = 1, \ldots, j-1$,

$$\gamma_k = \alpha_k^t \mathbf{X}_k^t \mathbf{W} \mathbf{Z} / \alpha_k^t \mathbf{X}_k^t \mathbf{W} \mathbf{X}_k \alpha_k, \qquad \gamma = \alpha^t \mathbf{X}_j^t \mathbf{W} \mathbf{Z} / \alpha^t \mathbf{X}_j^t \mathbf{W} \mathbf{X}_j \alpha.$$

Alternatively update $\alpha(\eta)$ and $\eta(\alpha)$ until $\alpha(\eta)$ converges to α_j, leading to the construction of the j-th component $\mathbf{X}_j\alpha_j$. Thus, the j-th component $\mathbf{X}_j\alpha_j$ maximizes its correlation with the working response $\mathbf{Z}$ and maximally explains the variation of the response variables.

Such construction stops whenever $\mathbf{W}^{1/2}\mathbf{Z}(\eta)$ is uncorrelated to $\mathbf{W}^{1/2}\mathbf{X}_j$. Because

$$\mathbf{X}_j\alpha_j = \mathbf{X}_{j-1}(I - \alpha_{j-1}\theta_{j-1})\alpha_j = \cdots = \mathbf{X}\left\{\prod_{l=1}^{j-1}(I - \alpha_{j-l}\theta_{j-l})\right\}\alpha_j,$$

we can denote each component $\mathbf{X}_j\alpha_j = \mathbf{X}\varpi_j$. Upon completion of the construction, we have the generalized orthogonal-components regression model with orthogonal components $\mathbf{X}\varpi_1$, $\mathbf{X}\varpi_2$, $\mathbf{X}\varpi_3$, $\cdots$, because $\mathbf{W}^{1/2}\mathbf{X}\varpi_1$, $\mathbf{W}^{1/2}\mathbf{X}\varpi_2$, $\mathbf{W}^{1/2}\mathbf{X}\varpi_3$, $\cdots$, are uncorrelated.

Theorem 1. *Each component $\mathbf{X}_j\alpha_j$ can be rewritten as $\mathbf{X}\varpi_j$ 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 before, the components $\mathbf{X}\varpi_1, \mathbf{X}\varpi_2, \ldots$, are orthogonal.

2.2 Sparsifying the Components

As shown in the above, we construct the j-th component $\mathbf{X}_j\alpha_j$ by maximizing $\|\mathbf{Z}(\eta)^t\mathbf{W}\mathbf{X}_j\alpha_j\|^2$ under the condition that $\|\alpha_j\| = 1$. That is, it seeks a sequence of loading vectors that not only relate $\mathbf{X}$ to $\mathbf{Z}(\eta)$ but also capture the variation in $\mathbf{X}$, and each loading vector α_j turns out to be the leading eigenvector of $\mathbf{X}_j^t\mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t\mathbf{W}\mathbf{X}_j$. The solution to this optimization problem is not a sparse vector, which leads to the fact that each component is a linear combination of all features.

Here we intend to enforce the variable selection function in constructing orthogonal components and build each component with selected important features. Specifically, to get the sparse loading vectors, we follow Zhang et al. (2009) and consider the following optimization problem to obtain $\alpha(\eta) = \alpha/\|\alpha\|$,

$$(\alpha, \vartheta) = \arg\min_{\alpha,\vartheta:\|\vartheta\|=1}\{-2\alpha^t\mathbf{X}_j^t\mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t\mathbf{W}\mathbf{X}_j\vartheta + \|\alpha\|^2 + p_\lambda(\alpha)\}, \tag{10}$$

where $p_\lambda(\alpha)$ is a penalty function with tuning parameter λ. Different penalty functions will be considered to obtain sparse $\alpha(\eta)$ to sparsify the constructed components, with $\alpha(\eta) = 0$ implying uncorrelated $\mathbf{W}^{1/2}\mathbf{Z}(\eta)$ and $\mathbf{W}^{1/2}\mathbf{X}_j$.

When $q = 1$, $\mathbf{Z}(\eta)$ is a column vector. Applying the method of Lagrange multipliers to (10), we can get

$$\vartheta = \mathbf{X}_j^t\mathbf{W}\mathbf{Z}(\eta)/\|\mathbf{X}_j^t\mathbf{W}\mathbf{Z}(\eta)\|,$$

which implies the following results.

Theorem 2. *In the case of logistic regression with $q = 1$, the optimization problem in (10) can be simplified and rewritten as, with $\tau = \|\mathbf{X}_j^t W\mathbf{Z}(\eta)\|$,*

$$\alpha = \arg\min_{\alpha}\{\left\|\alpha - \tau\mathbf{X}_j^t\mathbf{W}\mathbf{Z}(\eta)\right\|^2 + p_\lambda(\alpha)\}. \tag{11}$$

2.2.1 Sparsifying the Components via Empirical Bayes Thresholding

When $p_\lambda(\cdot)$ is specified by the logarithm of a prior density function, the optimal α in (11) will be a Bayesian shrinkage of $\mathbf{X}_j^t\mathbf{WZ}(\eta)$. In consideration of the sparsity of α, we can employ the empirical Bayes thresholding (EBT) proposed by Johnstone and Silverman (2004) to obtain a sparsified $\alpha(\eta)$ from $\mathbf{X}_j^t\mathbf{WZ}(\eta)$.

In general, we can solve (10) by iterating alternatively between optimal ϑ and α. For a fixed α, we have,

$$\vartheta = \mathbf{X}_j^t\mathbf{WZ}(\eta)\mathbf{Z}(\eta)^t\mathbf{WX}_j\alpha/\|\mathbf{X}_j^t\mathbf{WZ}(\eta)\mathbf{Z}(\eta)^t\mathbf{WX}_j\alpha\|.$$

For a fixed ϑ, we denote

$$\xi = \mathbf{X}_j^t\mathbf{WZ}(\eta)\mathbf{Z}(\eta)^t\mathbf{WX}_j\vartheta.$$

Then we have

$$\alpha = \arg\min_\alpha\{-2\alpha^t\xi + \|\alpha\|^2 + p_\lambda(\alpha)\} = \arg\min_\alpha\{\|\xi - \alpha\|^2 + p_\lambda(\alpha)\}. \tag{12}$$

Thus the optimal α is an estimate of the mean of ξ under the prior distribution specified by $p_\lambda(\alpha)$, which will be selected to obtain a sparse α.

Since each $\xi_i/\|\xi\|$ is an estimate of the certain conditional correlation coefficient, we can take a Fisher's z-transformation,

$$z_i = \frac{1}{2}\log\frac{1 - \xi_i/\|\xi\|}{1 + \xi_i/\|\xi\|},$$

and further assume,

$$z_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \frac{\lambda^2}{p-3}\right),$$

where $\mu_i = \frac{1}{2}\log\{(1-\alpha_i)/(1+\alpha_i)\}$, and λ partially accounts for possible under-dispersion or over-dispersion due to dependent data.

To obtain sparse μ and thus sparse α, we assume a mixture prior with a point mass at zero and a quasi-Cauchy distribution for each μ_i, i.e.,

$$\pi(\mu_i) = (1-w)\delta_0(\mu_i) + w\frac{1}{\sqrt{2\pi}}\left\{1 - \frac{|\mu_i|\Phi(-|\mu_i|)}{\phi(\mu_i)}\right\},$$

where $\delta_0(\cdot)$ is Dirac's delta function. An estimate of w, say $\hat{w}$, can be calculated by maximizing the marginal likelihood, and μ_i can be estimated by the posterior median, i.e.,

$$\hat{\mu}_i = \hat{\mu}(\xi_i) = median(\mu_i|z_i, \hat{w}),$$

leading to

$$\hat{\alpha}_i = \frac{1 - e^{2\hat{\mu}_i}}{1 + e^{2\hat{\mu}_i}}\|\xi\|.$$

Note that, as $\hat{w}$ provides a data-driven estimate of the parameter sparsity, the resultant estimate is adaptive to the sparsity of the underlying parameter and can reach the overall risk bounds.

2.2.2 Sparsifying the Components via Parametric Penalties

Many parametric penalty functions have been proposed to combine variable selection and parameter estimation, for example, the L_1 penalty function by Tibshirani (1996), smoothly clipped absolute deviation (SCAD) by Fan and Li (2001), and minimax concave penalty (MCP) by Zhang (2010), among others. As shown below, such parametric penalty functions result in explicit solutions to (12).

With sparse α in (12), we can calculate

$$\sigma = \text{median}_{1\leqslant j\leqslant p}\{|\xi_j|\}/\Phi^{-1}(0.75),$$

and rewrite (12) as

$$\alpha = \arg\min_{\alpha}\{\|\xi - \alpha\|^2/\sigma^2 + p_\lambda(\alpha/\sigma)\}, \tag{13}$$

that is, optimal α can be obtained from estimating the mean α/σ from standardized data ξ/σ under different penalty functions, such as L_1, SCAD, and MCP. When ξ contains pure noise, α can be zero, which implies no further construction of components.

3 Simulation Studies

We consider different cases of large p small n data to compare the performance of gPOCRE with other approaches, *i.e.*, sparse generalized partial least squares (SGPLS) by Chung and Keles (2010) and generalized linear model with lasso penalty (LASSO) by Friedman et al. (2010). For gPOCRE, we use four different methods to enable the variable selection, that is, empirical Bayes thresholding, L_1 penalty, SCAD, and MCP, and the corresponding algorithms are denoted as gPOCRE_{EB}, gPOCRE_{L_1}, gPOCRE_{MCP}, and gPOCRE_{SCAD}, respectively.

We present six case studies here. The first two consider highly and mildly correlated predictors. The third one has clustered predictors, the fourth one demonstrates a measurement-error model, and the fifth and sixth ones study the multinomial response case. Within each case, the simulated data consists of a training set with sample size 200 or 500 and a test set with sample size fixed at 200, with the number of predictors fixed at $p = 1000$. For each simulated data set, a five-fold cross-validation method is used to select the optimal tuning parameters based on the training data. The sample misclassification rate (MR), number of detected true predictors (NTP), and number of false predictors (NFP) are calculated based on the test data set.

3.1 Predictors with High or Mild Correlations

To study the effects of correlated predictors, we simulate data from the following model,

$$\text{logit}(E(y_i|\mathbf{x}_i)) = 2\sum_{j=1}^{10} x_{ij} + \sum_{j=101}^{110} x_{ij}, \quad i = 1, \ldots, n,$$

where each $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ consists of ten independently distributed blocks, with each block $\{x_{i,k+1}, \ldots, x_{i,k+100}\}$ simulated from an $AR(1)$ process with the correlation coefficient ρ. We consider $\rho = 0.9$ and 0.5 for high and mild correlations, respectively.

We summarize the simulation results in Table 1. First, we note that as the ratio of the number of observations to the number of predictors increases, the prediction ability of each

Table 1: Results from simulation studies of predictors with high/mild correlation. Reported are the median values across 100 simulation runs. Numbers in the parentheses are the corresponding standard errors.

ρ		n	LASSO	SGPLS	gPOCRE (EB)	gPOCRE (L_1)	gPOCRE (SCAD)	gPOCRE (MCP)
0.9	MR	200	.0700	.0600	.0500	.0500	.0500	.0500
			(.0021)	(.0022)	(.0023)	(.0024)	(.0024)	(.0024)
		500	.0500	.0425	.0400	.0450	.0475	.0450
			(.0016)	(.0016)	(.0016)	(.0017)	(.0016)	(.0017)
	NTP	200	11	19	20	20	20	20
			(.1941)	(.2183)	(.1218)	(.1058)	(.0946)	(0.1058)
		500	16	20	20	20	20	20
			(.1608)	(.0834)	(.0403)	(.0273)	(.0239)	(.0273)
	NFP	200	5	1	3.5	6	6	6
			(.6488)	(.2788)	(.2624)	(.5747)	(.3657)	(.5586)
		500	5.5	1	3	6	6	6
			(.4856)	(.2156)	(.1457)	(.3111)	(.3215)	(.4564)
0.5	MR	200	.1400	.1425	.1400	.1550	.1550	.1550
			(.0029)	(.0038)	(.0038)	(.0036)	(.0034)	(.0036)
		500	.0800	.0800	.0800	.0850	.0900	.0085
			(.0020)	(.0023)	(.0025)	(.0026)	(.0023)	(.0027)
	NTP	200	15	13	14	15	15	15
			(.2301)	(.3044)	(.2672)	(.2793)	(.2864)	(.2793)
		500	19	18	18	19	19	19
			(.0832)	(.1311)	(.1458)	(.0815)	(.0928)	(.0815)
	NFP	200	14	1	2	4.5	3	4.5
			(1.4567)	(3.1936)	(1.9000)	(2.8678)	(2.7846)	(2.8678)
		500	18.5	0	0	1.5	1	1.5
			(1.8809)	(.1423)	(.1114)	(.3427)	(.2796)	(.3292)

method increases as expected. For example, MR of gPOCRE_{L_1} decreases from 0.05 to 0.045, and MR of SGPLS decreases from 0.06 to 0.043, as the ratio increases from 1/5 to 1/2 in the case of high correlation. In terms of MR, simulation results suggest that gPOCRE_{EB} performs better than the other methods in most cases. In terms of NTP and NFP, all methods improve the ability to identify the true predictors as the number of observations increases, and gPOCRE_{EB} slightly outperforms the other methods. We noticed that when the correlation between predictors

increases, all methods except LASSO improve the performance to identify the true predictors.

3.2 Clustered Predictors

We consider a latent variable model as follows,

$$\text{logit}(E(y_i|\mathbf{x}_i)) = 1.5\sum_{j=1}^{30} x_{ij}, \quad i = 1, \ldots, n,$$

where, for $j = 1, \ldots, p$,

$$x_{ij} = z_{i1}1_{\{j\leqslant 10\}} + z_{i2}1_{\{11\leqslant j\leqslant 20\}} + z_{i3}1_{\{21\leqslant j\leqslant 30\}} + \xi_{ij},$$

with the three latent variables $z_{i1}, z_{i2}, z_{i3} \overset{iid}{\sim} N(0, 1)$ and $\xi_{ij} \overset{iid}{\sim} N(0, 0.1^2)$.

The simulation results are summarized in Table 2. We noticed that different versions of gPOCRE have similar performance as SGPLS, and both of them outperform LASSO in terms of MR, NTP, and NFP. With much smaller standard errors reported, different versions of gPOCRE also demonstrate robust performance than SGPLS and LASSO across different criteria.

3.3 Predictors Observed with Errors

With the common concern of errors in predictors, we consider the following model, including predictors observed with errors,

$$\text{logit}(E(y_i|\mathbf{x}_i)) = z_{i1} + 2z_{i2} + z_{i3}, \quad i = 1, \ldots, n,$$

Table 2: Results from simulation studies of clustered predictors. Reported are the median values across 100 simulation runs. Numbers in the parentheses are the corresponding standard errors.

	n	LASSO	SGPLS	gPOCRE (EB)	gPOCRE (L_1)	gPOCRE (SCAD)	gPOCRE (MCP)
MR	200	.0500	.0400	.0500	.0500	.0500	.0500
		(.0019)	(.030)	(.0021)	(.0030)	(.0030)	(.0030)
	500	.0300	.0300	.0300	.0300	.0325	.0300
		(.0014)	(.0015)	(.0017)	(.0020)	(.0017)	(.0020)
NTP	200	7	30	30	30	30	30
		(.1739)	(.1960)	(.0141)	(.0100)	(.0100)	(.0100)
	500	11	30	30	30	30	30
		(.2068)	(.0539)	(.0100)	(.0000)	(.0000)	(.0000)
NFP	200	2	0	0	0	0	0
		(.9085)	(.2031)	(.8663)	(.0656)	(.1149)	(.06557)
	500	1	0	0	0	0	0
		(.8560)	(.1444)	(.0141)	(.0389)	(.0321)	(.0389)

Table 3: Results from simulation studies of predictors observed with errors. Reported are the median values across 100 simulation runs. Numbers in the parentheses are the corresponding standard errors.

	n	LASSO	SGPLS	gPOCRE (EB)	gPOCRE (L_1)	gPOCRE (SCAD)	gPOCRE (MCP)
MR	200	.2600	.2450	.2300	.2450	.2500	.2450
		(.0032)	(.0034)	(.0033)	(.0036)	(.0040)	(.0036)
	500	.2350	.2200	.2150	.2200	.2200	.2200
		(.0031)	(.0030)	(.0028)	(.0029)	(.0030)	(.0030)
NTP	200	13	22	21.5	21	21.5	21
		(.3025)	(.7015)	(.5434)	(.5472)	(.5902)	(.5472)
	500	20	29	26	29	29	29
		(.2918)	(.3074)	(.3102)	(.2419)	(.2418)	(.2399)
NFP	200	5.5	2	.5	0	0	0
		(1.7845)	(17.9083)	(.6991)	(.6642)	(1.0049)	(.6642)
	500	4	0	0	0	0	0
		(2.2903)	(7.2884)	(.0609)	(.1785)	(.2286)	(.2270)

where, for $j = 1, \ldots, 1000$,

$$x_{ij} = \text{sign}(5.5 - j)z_{i1}1_{\{j \leqslant 10\}} + \text{sign}(15.5 - j)z_{i2}1_{\{11 \leqslant j \leqslant 20\}} + z_{i3}1_{\{21 \leqslant j \leqslant 30\}} + \xi_{ij},$$

with latent variables $z_{i1}, z_{i2}, z_{i3} \overset{iid}{\sim} N(0, 1)$ and $\xi_{ij} \overset{iid}{\sim} N(0, 1)$.

We summarize the simulation results in Table 3. In terms of MR and NTP, the four versions of gPOCRE are comparable to each other and SGPLS. While LASSO is outperformed by all other methods in terms of MR and NTP, it can identify much less number of true predictors than others. In terms of NFP, the four versions of gPOCRE perform similarly but outperm both SGPLS and LASSO with LASSO the worst.

3.4 Multinomial Models

Here we simulate data from the following multinomial model,

$$\begin{cases} \log(E[y_{i1}|\mathbf{x}_i]/E[y_{i3}|\mathbf{x}_i]) = \theta * \sum_{j=1}^{10} x_{ij} + \theta * \sum_{j=101}^{110} x_{ij}, \\ \log(E[y_{i2}|\mathbf{x}_i]/E[y_{i3}|\mathbf{x}_i]) = \frac{1-\sqrt{3}}{2} * \theta * \sum_{j=1}^{10} x_{ij} + \frac{1+\sqrt{3}}{2}\theta * \sum_{j=101}^{110} x_{ij}, \end{cases} \tag{14}$$

where, for $i = 1, 2, \ldots, n$, each $\mathbf{x}_i$ is simulated from an $AR(1)$ process with $\rho = 0.5$.

We summarized the simulation results in Table 4, which shows that SGPLS outperforms all other methods in terms of MR. However, the four versions of gPOCRE reports higher NTP, especially that gPOCRE_{L_1}, gPOCRE_{SCAD}, and gPOCRE_{MCP} tend to identify all true predictors. On other hand, all methods, except gPOCRE_{EB}, may report a large number of false predictors.

Table 4: Simulation results of the multinomial model. Reported are the median values across 100 simulation runs. Numbers in the parentheses are the corresponding standard errors.

	θ	n	LASSO	SGPLS	gPOCRE (EB)	gPOCRE (L_1)	gPOCRE (SCAD)	gPOCRE (MCP)
MR	2	200	0.19	0.1	0.18	0.2	0.21	0.2
			(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)
		500	0.11	0.07	0.1	0.12	0.14	0.12
			(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	4	200	0.18	0.09	0.16	0.2	0.2	0.2
			(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)
		500	0.085	0.06	0.09	0.12	0.12	0.12
			(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
NTP	2	200	31	36	36	40	40	40
			(3.03)	(5.19)	(2.5)	(2.06)	(2.15)	(2.04)
		500	36	38	40	40	40	40
			(1.4)	(2.83)	(0.69)	(0.34)	(0.59)	(0.34)
	4	200	31	36	36	40	40	40
			(2.92)	(4.59)	(2.3)	(1.99)	(1.97)	(1.99)
		500	37	38	40	40	40	40
			(1.34)	(2.86)	(0.73)	(0.28)	(0.34)	(0.28)
NFP	2	200	19.5	2	4	31	29	35
			(30.64)	(11.53)	(50.69)	(66.8)	(65.62)	(66.56)
		500	21.5	20	0	4	2	4
			(36.67)	(27.13)	(2.15)	(8.53)	(9.08)	(10.16)
	4	200	24	2	4	24	29	25
			(28.79)	(10.46)	(46.02)	(65.22)	(64.16)	(64.98)
		500	17	17	0	4	4	4
			(28.32)	(43.25)	(1.64)	(8.66)	(8.52)	(8.66)

3.5 Running Time Analysis

In high-dimensional data analysis, computational time is an important factor when comparing the performance of different methods. Table 5 summarizes the time used to analyze one data set from the case with highly correlated predictors with sample size fixed at 200. All methods used 5-fold cross-validation to select tuning parameters. We noted that LASSO uses the shortest time to select the tuning parameter and fit the model. Among the rest, gPOCRE$_{EB}$ takes much less

Table 5: Time consumption (in seconds) in analyzing one dataset.

Model	LASSO	SGPLS	gPOCRE (EB)	gPOCRE (L_1)	gPOCRE (SCAD)	gPOCRE (MCP)
High Correlation	0.36	31	16	27	27	27
Mild Correlation	0.41	35	16.6	25.6	26	26
Clustered Predictors	0.46	32	18	30	37	30
Predictors with Errors	0.46	30	19	31	35	32
Multinomial Model	0.88	63	43	66	80	66

time than the others, while SGPLS is comparable to gPOCRE$_{L_1}$ and gPOCRE$_{MCP}$. This result was obtained on a *MacBook Pro with 2.5 GHz Intel Core i7.*

4 Real Data Analysis

4.1 Logistic Regression for Isolated Letter Speech Recognition

Here we analyze a subset of the Isolated Letter Speech Recognition (ISOLET) data (Fanty and Cole, 1990), collected from 150 subjects, each speaking the first two letters of the alphabet twice. A total of 617 waveform features are available for predicting the spoken letter. We randomly chose 80% of the data as training data and the remaining 20% as the test data, with data stratified between the letters. We applied each of gPOCRE, SGPLS, and LASSO to the training data to build a logistic model, using a five-fold CV to optimize the tuning parameters. The MR values were calculated based on the test data. We repeated this procedure 50 times, and reported all MR and numbers of non-zero coefficients in Figure 1.

All methods performed well in terms of MR on the test sets, with median values at zero. While all three models reported MR values predominantly at 0 and 0.008, gPOCRE and LASSO occasionally reported higher MR values of 0.017. SGPLS, in contrast, produced one notable outlier with an MR of 0.034. Additionally, SGPLS exhibited considerably variability in feature selection, identifying between tens and over 600 features across the 50 models. As shown in Table 6, only 15 features were consistently selected by SGPLS in every model. LASSO, despite producing models with similarly small numbers of features (median at 40), consistently selected only four features across all 50 models. In contrast, gPOCRE not only stably selected larger numbers of features (median at 192), but also demonstrated strong consistency by selecting the same 39 features across all 50 models, with 91 features appearing in at least 80% of them.

4.2 Multinomial Regression for Breast Cancer Data

We employed a subset of the breast cancer data from The Cancer Genome Atlas (TCGA) (Hutter and Zenklusen, 2018). We focused our analysis on the three breast cancer subtypes, i.e., Basal, HER2, and Luminal, with sample sizes of 66, 44, and 110, respectively. Each subject included 384 predictors representing the gene expressions of 184 miRNAs and 200 mRNAs. We allocated the data into training and test sets using an 80%-20% stratified random split, and applied gPOCRE, SGPLS, and LASSO to predict the three subtypes of breast cancer. For each split, the tuning parameters for each method were obtained through five-fold cross-validation. The process was repeated 50 times, and the misclassification rates of the three methods are reported in Table 7.

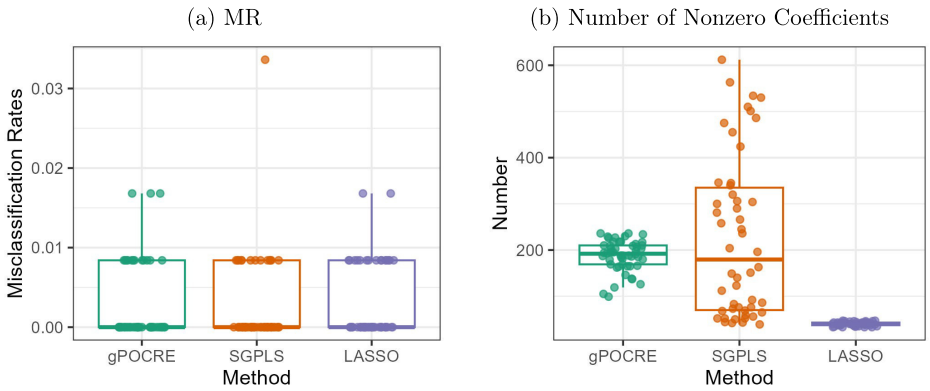

Figure 1: Boxplot of results from analyzing ISOLET data across 50 random splits.

Table 6: The number of variables consistently identified by different methods.

Identified Times	gPOCRE	SGPLS	LASSO
50	39	15	4
45∼49	35	13	5
40∼44	17	13	4

Table 7: Summary for the breast cancer data analysis. Reported are the median values across 50 random splits. The numbers in the parentheses are the corresponding standard errors.

Methods	$\#(\beta \neq 0)$	MR
gPOCRE_{EB}	74.5 (1.30)	0.07 (0.005)
SGPLS	301 (10.79)	0.09 (0.006)
LASSO	60 (0.86)	0.09 (0.006)

Among the methods, gPOCRE outperformed both SGPLS and LASSO in terms of misclassification rate (MR). LASSO has similar MR to SGPLS, but it has selected the fewest genes. In contrast, gPOCRE selected more genes than LASSO but fewer than SGPLS. Notably, SGPLS demonstrated instability, exhibiting a large standard error in the number of selected genes, while gPOCRE achieved a more stable selection with the lowest MR.

5 Discussion

In this work, we introduced gPOCRE, a regression-based model tailored for high-dimensional generalized linear models (GLMs). By constructing sparse components, gPOCRE addresses the challenges posed by high dimensionality and multicollinearity, offering a solution to ill-posed problems while identifying important features. Both simulation studies and real data examples highlight the superior performance of gPOCRE and its comparability to other existing methods.

Several algorithms have been proposed to implement sparse partial least squares (PLS) for high-dimensional GLMs. Chung and Keles (2010) introduced SPLSDA and SGPLS, which extend sparse partial least squares (Chung and Keles (2010)) to classification problems. SPLSDA

is a two-stage procedure that replaces PLS with a sparse PLS method in the first stage of the classification framework developed by Nguyen and Rocke (2002a,b). However, SGPLS extends SPLS by, within each iteration of IRWLS, employing weighted SPLS instead of weighted least squares (LS). SGPLS uses cross-validation (CV) to determine the optimal tuning parameters and the number of components. Simulations and real data examples show that SGPLS is significantly more time-consuming than gPOCRE$_{EB}$.

gPOCRE applies penalization through four strategies: empirical Bayes thresholding, L_1 penalty, SCAD penalty, and MCP. Each of these strategies enables gPOCRE to select sparse variables effectively within the large p, small n paradigm. Empirical Bayes thresholding can take a threshold adaptive to parameter sparsity, thus providing data-driven sparsification of components in gPOCRE. Our simulation studies have verified its advantage over other penalty methods. On the other hand, gPOCRE$_{EB}$ also outperforms LASSO and SGPLS in practical scenarios such as predictors observed with errors and multiple outcomes. As shown in Table 3 and Table 4, SGPLS may unstably report a large number of false predictors, which resonates with the observed large standard errors of SGPLS when analyzing both sets of real data.

Supplementary Material

The MATLAB code for gPOCRE is available on the journal's website. The ISOLET data by Fanty and Cole (1990) can be downloaded from https://www.openml.org/search?type=data&sort=version&status=any&order=asc&exact_name=isolet&id=41966, and the breast cancer data can be found in the R package mixOmics (https://mixomics.org/).

A Proof of Theorem 2

We can apply Lagrange multipliers to (10),

$$L(\alpha, \vartheta, \gamma) = -2\alpha^t \mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t \mathbf{W}\mathbf{X}_j \vartheta + \|\alpha\|^2 + p_\lambda(\alpha) + \gamma(\|\vartheta\|^2 - 1).$$

Taking the partial derivative of ϑ, we have

$$\frac{\partial L(\alpha, \vartheta, \gamma)}{\partial \vartheta} = -2\alpha^t \mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t \mathbf{W}\mathbf{X}_j + 2\gamma\vartheta^t = 0,$$

which leads to

$$\vartheta^t = \frac{1}{\gamma}\alpha^t \mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t \mathbf{W}\mathbf{X}_j.$$

Taking the partial derivative of γ, we have

$$\frac{\partial L(\alpha, \vartheta, \gamma)}{\partial \gamma} = \|\vartheta\|^2 - 1 = 0,$$

which implies $\gamma = \|\alpha^t \mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)\mathbf{Z}(\eta)^t \mathbf{W}\mathbf{X}_j\|$. Because $q = 1$, both γ and $\alpha^t \mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)$ are scalars. Thus we have

$$\vartheta = \frac{\mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)}{\|\mathbf{X}_j^t \mathbf{W}\mathbf{Z}(\eta)\|}.$$

Plugging ϑ into (10), we can get (11).

Acknowledgement

The authors thank the editor, associate editor, and referees for their constructive comments which has led to significant improvement of this paper. The results of breast cancer data are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding

This research was partially supported by NSF CAREER award IIS-0844945, NIH grants R01GM131491, R01GM131491-02S1, R01GM131491-02S2, R01AG080917, and R01AG080917-02S1, NCI grant P30CA062203, and UCI Anti-Cancer Challenge funds from the UC Irvine Comprehensive Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Chao Family Comprehensive Cancer Center.

References

Boulesteix AL, Strimmer K (2006). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioformatics*, 8: 32–44. https://doi.org/10.1093/bib/bbl016

Chun H, Keleş S (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1): 3–25. https://doi.org/10.1111/j.1467-9868.2009.00723.x

Chung D, Keles S (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 9. Article 17.

De Jong S (1993). Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18: 251–263. https://doi.org/10.1016/0169-7439(93)85002-X

Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360. https://doi.org/10.1198/016214501753382273

Fan J, Samworth R, Wu Y (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10: 2013–2038.

Fanty M, Cole R (1990). Spoken letter recognition. *Proceedings of the International Conference on Neural Information Processing Systems*, 4: 220–226.

Fisher RA (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Freeman C, Kulić D, Basir O (2013). Feature-selected tree-based classification. *IEEE Transactions on Cybernetics*, 43(6): 1990–2004. https://doi.org/10.1109/TSMCB.2012.2237394

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1. https://doi.org/10.18637/jss.v033.i01

Hoskuldsson A (1988). PLS regression methods. *Journal of Chemometrics*, 2: 211–228. https://doi.org/10.1002/cem.1180020306

Hoskuldsson A (1992). The h-principle in modelling with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 14: 139–153. https://doi.org/10.1016/0169-7439(92)80099-P

Hutter C, Zenklusen JC (2018). The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell*, 173(2): 283–285. https://doi.org/10.1016/j.cell.2018.03.042

Johnstone IM, Silverman BW (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4): 1594–1649. https://doi.org/10.1214/009053604000000030

Lê Cao KA, Rossouw D, Robert-Granié C, Besse P (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. 7(1): Article 35.

Lin Y, Zhang M, Zhang D (2015). Generalized orthogonal components regression for high dimensional generalized linear models. *Computational Statistics & Data Analysis*, 88: 119–127. https://doi.org/10.1016/j.csda.2015.02.006

Loh WY (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1): 14–23.

Massy WF (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309): 234–256. https://doi.org/10.1080/01621459.1965.10480787

McLachlan GJ (2005). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons.

Nguyen DV, Rocke DM (2002a). *Classification of Acute Leukemia Based on DNA Microarray Gene Expressions Using Partial Least Squares*. Springer.

Nguyen DV, Rocke DM (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18: 39–50. https://doi.org/10.1093/bioinformatics/18.1.39

Shen J, Gao S (2008). A solution to separation and multicollinearity in multiple logistic regression. *Journal of Data Science*, 6(4): 515. https://doi.org/10.6339/JDS.2008.06(4).395

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews. Genetics*, 20(8): 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van de Geer SA (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2): 614–645.

Velliangiri S, Alagumuthukrishnan S, et al. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, 165: 104–111. https://doi.org/10.1016/j.procs.2020.01.079

Wold H (1966). Estimation of principal components and related models by iterative least squares. In PR Krishnajad (Ed.), *Multivariate Analysis*, 391–420. New York: Academic Press.

Wold H (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1): 117–142. https://doi.org/10.1017/S0021900200047604

Xie J, Lin Y, Yan X, Tang N (2020). Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data. *Journal of the American Statistical Association*, 115(530): 747–760. https://doi.org/10.1080/01621459.2019.1573734

Zhang C (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894–942.

Zhang D, Lin Y, Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781–796.