Received: 27 January 2019 | Revised: 24 September 2019 | Accepted: 21 November 2019

DOI: 10.1002/sim.8454

RESEARCH ARTICLE

Statistics in Medicine WILEY

Structured sparse logistic regression with application to lung cancer prediction using breath volatile biomarkers

Xiaochen Zhang[1] | **Qingzhao Zhang**[1,2] | **Xiaofeng Wang**[3] | **Shuangge Ma**[4] | **Kuangnan Fang**[1]

[1]Department of Statistics, School of Economics, Xiamen University, China

[2]The Wang Yanan Institute for Studies in Economics, Xiamen University, China

[3]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio

[4]Department of Biostatistics, Yale University, New Haven, Connecticut

Correspondence
Kuangnan Fang, Department of Statistics, School of Economics, Xiamen University, China.
Email: xmufkn@xmu.edu.cn

Funding information
Fundamental Research Funds for the Central Universities, 20720181003,20720171064,20720171095; Humanity and Social Science Youth Foundation of Ministry of Education of China, 19YJC910010; National Natural Science Foundation of China, 71471152

This article is motivated by a study of lung cancer prediction using breath volatile organic compound (VOC) biomarkers, where the challenge is that the predictors include not only high-dimensional time-dependent or functional VOC features but also the time-independent clinical variables. We consider a high-dimensional logistic regression and propose two different penalties: *group spline-penalty* or *group smooth-penalty* to handle the group structures of the time-dependent variables in the model. The new methods have the advantage for the situation where the model coefficients are sparse but change smoothly within the group, compared with other existing methods such as the group lasso and the group bridge approaches. Our methods are easy to implement since they can be turned into a group minimax concave penalty problem after certain transformations. We show that our fitting algorithm possesses the descent property and leads to attractive convergence properties. The simulation studies and the lung cancer application are performed to demonstrate the accuracy and stability of the proposed approaches.

KEYWORDS

group smooth-penalty, group spline-penalty, high-dimensional data, time-dependent variables, variable selection

1 | INTRODUCTION

Group structures in predictors can arise for many reasons in regression modeling. Taking this grouping information into account in the modeling process could improve both the interpretability and the accuracy of the model.[1] Several methods have addressed the group variable selection problem in literature, see for example, the group lasso,[2,3] the method through composite absolute penalties (CAPs),[4] the group minimax concave penalty (gMCP),[5,6] and the group bridge approach.[7]

When features are ordered in some meaningful ways, they are often correlated each other. To take the ordering effects into account, Tibshirani et al[8] proposed the fused lasso to encourage flatness of the coefficients as well as the sparsity of the coefficients. Cao et al[9] introduced a new regularization term, named as the generalized fused group lasso (GFGL), to flexibly model the relationship among the response and predictors based on the prior knowledge. Ciuperca[10] considered the quantile model with grouped explanatory variables, and proposed and studied an adaptive fused group lasso quantile estimator.

In many applications, the features might vary smoothly, rather than being like a step function. The fused lasso-types of approaches cannot work in these situations. Hebiri and Van[11] introduced "smooth-lasso" to deal with sparse problems where successive regression coefficients were known to vary slowly. Huang et al[12] proposed the sparse Laplacian shrinkage (SLS) method for variable selection and estimation that explicitly incorporated the correlation patterns among predictors. Guo et al[13] imposed an L_2 penalty on the discrete version of the second derivatives of coefficients, as well as the sparse penalty on coefficients. However, the above methods only tackled the sparse problems in linear regression without considering the group effects. They are not suitable for the cases where the group structures exist among the predictors.

Motivating from a real clinical study where the features can be divided into groups and the features within each group are ordered in some meaningful way, we propose two penalized logistic regression approaches: group spline-penalty and group smooth-penalty methods in this article. The proposed methods may advance from the existing ones along the following aspects. First, they are the solutions to estimate coefficients which are sparse but change smoothly within a group. Second, our methods are easy to implement since they can be turned into a gMCP problem after certain transformations. It turns out that group spline-MCP (gspMCP) problem and group smooth-MCP (gsmMCP) problem are equivalent to a gMCP-type optimization problem. The simulation studies and real applications show that the estimations and predictions of the proposed method are more accurate than those using the gMCP for logistic regression.

The later sections are organized as follows. We address the background of the lung cancer study in Section 2. In Section 3, we describe the model setting and methodology, as well as the proposed computational algorithm. In Section 4, we present simulation studies under the five different scenarios. We assess the performance of our proposed models by comparing them with the gMCP method in terms of feature selection and prediction error. In Section 5, we apply the proposed methods to the lung cancer study. The article concludes with discussion in Section 6. The detailed computational algorithms and the descent property for our methods are given in Appendix.

2 | THE LUNG CANCER STUDY: A MOTIVATING EXAMPLE

Lung cancer accounts for around 28% of cancer-related deaths worldwide. Recent years have brought forward a new method for the early diagnosis of lung cancer by the analysis of the exhaled breath. The new diagnosis method contains blueprints of gaseous and nongaseous markers that will help in distinguishing the breath of cancer stricken patients from the healthy population.[14] Exhaled breath contains a complex mixture of several hundreds of volatile organic compounds (VOCs). The consumption and production of VOCs are caused by metabolic processes within the cells. After circulating within the blood and transferring to the lungs, these metabolic byproducts are exhaled from the body. Thus, the composition of VOCs in the exhaled breath reflects metabolic activity within the body. It has been shown that the features obtained from VOCs can be used as a noninvasive marker of lung cancer.[15,16] The measures of breath VOCs are often obtained through a colorimetric sensor array (CSA),[17] a disposable cartridge that contents an array of chromogenic reagents. The measurement from the sensor is a change in the colors of its elements every tidal volume (approximately 500 mL) breath air crossed. Sensor arrays do not identify the specific constituents of exhaled breath; rather their output is the result of the entire composition of the breath contents with the sensor.

Previously reported studies suggested that an early version of this sensor system was moderately accurate in identifying the subjects with lung cancer based on their breath profile.[15,16] In this study, the colorimetric sensor array, which was composed of 128 separate colorants, was used to determine whether the color features from CSA are capable of predicting lung cancer. Study subjects were recruited prospectively from outpatient clinics at the Cleveland Clinic, Cleveland, OH, US. All study subjects performed tidal breathing, inhaling unfiltered air through their nose, and exhaling through their mouth into disposable corrugated tubing for a total of 5 min through the CSA electric breath measuring system. Each image of the sensor was converted to numerical values for changes in the red, green, blue, and four ultra-color spectra of each colorant. Then totally 128 (the number of colorants) $\times$ 7 (changes in the red, green, blue, and four ultra-color spectra of each colorant) = 896 groups were obtained for each study subject. In each group, as images of the array were captured at baseline and then at the time that every tidal volume (approximately 500 mL) breath air is crossed throughout breath collection. Thus, VOC features can be divided into 896 groups depending on the color type, and the features are time-dependent and ordered in a meaningful way in each group. Figure 1 shows an example of the observed log-transformed intensities for seven color spectra (red, green, blue, and four ultra-color spectra) of one colorant in one subject. In addition to VOC features, we also collect other four clinical predictors: age, sex, smoking history, and the presence of chronic obstructive pulmonary disease (COPD). Modeling the data is challenging due to the fact that the problem

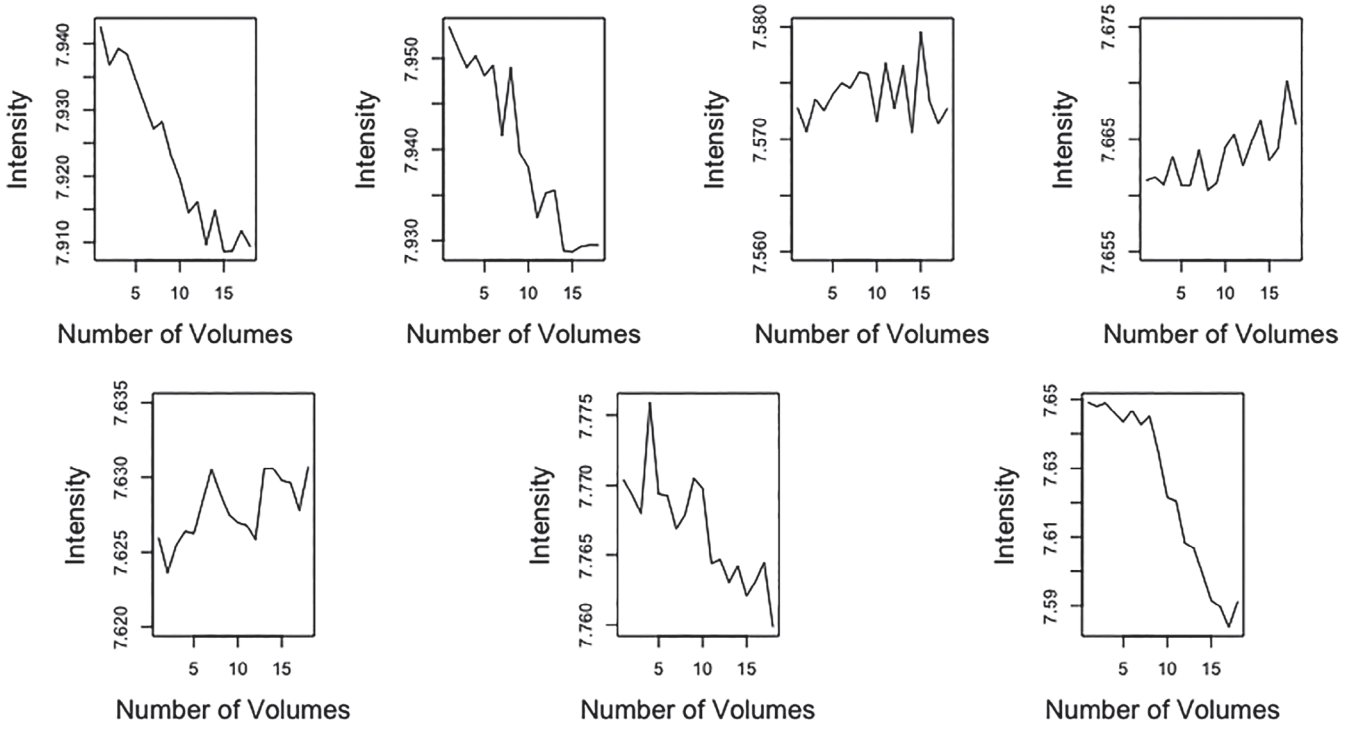

FIGURE 1 An example of the observed log intensities for one subject. Each image of the sensor was converted to numerical values for changes in the red, green, blue, and four ultra-color spectra of each colorant

is a "large-p-small-n" problem with the meaning group effects among the VOC features. Our goal of the study here is to build a prediction model using the VOC features as well as the clinical predictors to predict the presence of lung cancer.

3 | METHODS

3.1 | The model setting and methodology

Suppose that we have independent and identically distributed data from n subjects $(\boldsymbol{z}_i, \boldsymbol{x}_i, y_i; i = 1, \dots, n)$, where $\boldsymbol{z}_i \in \mathbb{R}^{q_0}$ is a q_0-dimensional vector, $\boldsymbol{x}_i \in \mathbb{R}^q$ is a q-dimensional vector, and $y_i \in \{0, 1\}$ is a binary response variable. Here $\boldsymbol{x}_i$ can be divided into J groups, which means there exist grouping structures: $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{\boldsymbol{n}})^\top$ with $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}^\top, \dots, \boldsymbol{x}_{i,J}^\top)^\top$ for $i = 1, \dots, n$. In jth group, the number of the predictors is q_j and $q = \sum_{j=1}^{J} q_j$. The vector $\boldsymbol{z}_i$ includes intercept and other scaler vectors. In the lung study, y_i is the binary outcome that denotes the presence of lung cancer, $\boldsymbol{x}_{i,j}$ is the jth group of VOC features for ith study subject, and the scaler vector $\boldsymbol{z}_i$ includes age, sex, smoking history, and the presence of COPD for the ith subject.

Here we consider a logistic regression model. The conditional probability of the model is given by $\mathbb{P}(y_i = 1|\boldsymbol{x}_i, \boldsymbol{z}_i) = p_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ with $\log\{p_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i)/(1 - p_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i))\} = \eta_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{z}_i^\top \boldsymbol{\beta}_0 + \sum_{j=1}^{J} \boldsymbol{x}_{i,j}^\top \boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_0 \in \mathbb{R}^{q_0}$ is the parameter vector corresponding to $\boldsymbol{z}_i$ and $\boldsymbol{\beta}_j \in \mathbb{R}^{q_j}$ is the parameter vector corresponding to the jth predictor $\boldsymbol{x}_{i,j}$. We denote $\boldsymbol{\beta} \in \mathbb{R}^{q+q_0}$ the whole parameter vector, that is, $\boldsymbol{\beta} = (\boldsymbol{\beta_0}^\top, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_J^\top)^\top$. The log-likelihood function is then

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(p_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i)) + (1 - y_i) \log(1 - p_{\boldsymbol{\beta}}(\boldsymbol{x}_i, \boldsymbol{z}_i))\}. \tag{1}$$

We use the group penalty to select the relevant groups in $\boldsymbol{X} = \{\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_J^\top\}^\top$. Let us consider the penalty

$$P_1(\boldsymbol{\beta}; \lambda_1, a) = \sum_{j=1}^{J} P_{\text{MCP}}\left(\left\|\boldsymbol{\beta}_j\right\|; \sqrt{q_j}\lambda_1, \gamma\right). \tag{2}$$

where $P_{\text{MCP}}(\theta;\lambda,\gamma)=\lambda\int_0^{|\theta|}(1-x/(\lambda\gamma))_+dx$, the notation z_+ stands for the positive part of z: z_+ is z if $z>0$, zero otherwise, $\|\cdot\|$ is the standard Euclidean norm, $\lambda_1\geq 0$ is the penalty parameter, $\sqrt{q_j}$s is the penalty used to adjust for the group sizes, and γ is an additional tuning parameter that is used to modify $P_{\text{MCP}}(\theta,\lambda,\gamma)$.

In this article, we assume that the features in each group are ordered in some meaningful way. Hebiri and Van[11] introduced "smooth-lasso," where the smoothness between $(\beta_{j,1},\ldots,\beta_{j,q_j})$ were encouraged by L_2-penalty $\sum_{m=2}^{q_j}(\beta_{j,m+1}-\beta_{j,m})^2$. Guo et al[13] suggested $\sum_{m=2}^{q_j-1}(\Delta_m^{(2)}\boldsymbol{\beta}_j)^2$, where $\Delta_m\boldsymbol{\beta}_j=:(\beta_{j,m+1}-\beta_{j,m})$, and $\Delta_m^{(2)}\boldsymbol{\beta}_j=:(\Delta_m\beta_j-\Delta_{m-1}\beta_j)=\beta_{j,m+1}-2\beta_{j,m}+\beta_{j,m-1}$. Motivating from the ideas of these two methods to capture the smooth features in a group, we propose two types of penalties to capture smoothing changes in each group, which are named the *group spline-penalty* and *group smooth-penalty*, respectively. These two types of penalties can be written in the following unified formula:

$$\sum_{j=1}^{J}P_2(\boldsymbol{\beta}_j,\lambda_2)=\sum_{j=1}^{J}\lambda_2\boldsymbol{\beta}_j^{\top}L^{j\top}L^j\boldsymbol{\beta}_j. \tag{3}$$

For the group spline-penalty, we set L^j as a $(q_j-2)\times q_j$ matrix with $L^j_{i,i}=L^j_{i,i+2}=1$, $L^j_{i,i+1}=-2$ and $L^j_{i,l}=0$ otherwise; for group smooth-penalty, we set L^j as a $q_j\times q_j$ matrix with $L^j_{i,i}=-1$, $L^j_{i,i-1}=1$ and $L^j_{i,l}=0$ otherwise.

Therefore, our proposed minimization objective function is then

$$Q(\boldsymbol{\beta})=-\frac{1}{n}\ell(\boldsymbol{\beta})+P_1(\boldsymbol{\beta};\lambda_1,\gamma)+\sum_{j=1}^{J}P_2(\boldsymbol{\beta}_j,\lambda_2). \tag{4}$$

In Equation (4), the first penalty is the usual gMCP to select groups of features, and the second penalty could be either the group spline-penalty or the group smooth-penalty. The group spline-penalty mimics the cubic spline, which is penalizing the L_2 norm of a discrete version of the second-order derivatives of the coefficients to encourage smoothness of coefficients. The group smooth-penalty penalizes the difference between adjacent coefficients. A nonzero λ_2 induces smoothness in coefficient estimates within a group. The larger λ_2 is, the smoother the estimates within a group are. The regularization parameters λ_1 and λ_2 can be determined automatically by cross-validation or by using a validation set. The details are shown in Sections 4 and 5.

The proposed methods are easy to implement since it can be turned into a gMCP problem after certain transformations.

3.2 | Computational algorithm

In this subsection, we address the computational algorithm for logistic regression with gspMCP or gsmMCP. Following Liu et al,[18] we resort to a majorization-minimization (MM) approach. The MM algorithm consists of two steps: the majorizing step and the minimizing step. Two steps alternate in turn until convergence is met. After convergence, the final solution achieves a local minimum of the original function. With MM approach, Liu et al[18] majorize the negative log-likelihood function through a quadratic loss:

$$-\frac{1}{n}\tilde{\ell}(\boldsymbol{\beta})\propto\frac{v}{2n}\left(\tilde{\boldsymbol{Y}}-\boldsymbol{Z}^{\top}\boldsymbol{\beta_0}-\sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right)^{\top}\left(\tilde{\boldsymbol{Y}}-\boldsymbol{Z}^{\top}\boldsymbol{\beta_0}-\sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right),$$

where $v=1/4$, $\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{\eta}}+(\boldsymbol{Y}-\boldsymbol{p})/v$ is the pseudo-response vector, and $\boldsymbol{Y}$ is the response vector and $\boldsymbol{p}$ is conditional probability vector. For more technical details, we refer to Breheny and Huang.[1] Given current estimate $\tilde{\boldsymbol{\beta}}$, we then can minimize

$$Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})\propto\frac{v}{2n}\left(\tilde{\boldsymbol{Y}}-\boldsymbol{Z}^{\top}\boldsymbol{\beta_0}-\sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right)^{\top}\left(\tilde{\boldsymbol{Y}}-\boldsymbol{Z}^{\top}\boldsymbol{\beta_0}-\sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right)+P_1(\boldsymbol{\beta};\lambda_1,\gamma)+\sum_{j=1}^{J}P_2(\boldsymbol{\beta}_j,\lambda_2).$$

This is computationally feasible since the L_2 term can be merged with the least squares terms. So, we can transform the problem into another gMCP problem.

Proposition 1. *Given current estimate* $\tilde{\boldsymbol{\beta}}^{(m)}$*, define an artificial dataset* $(\boldsymbol{Y}^*, \boldsymbol{X}^*)$ *by* $\boldsymbol{Y}^* = (\tilde{\boldsymbol{Y}}, \boldsymbol{0})^\top$, $\boldsymbol{X}^* = (\boldsymbol{X}^\top, \boldsymbol{L})^\top$, *where*

$$\boldsymbol{L} = \begin{bmatrix} \sqrt{\lambda_2 n/v}L^1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \sqrt{\lambda_2 n/v}L^2 & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \sqrt{\lambda_2 n/v}L^J \end{bmatrix}.$$

We set L^j *as a* $(q_j - 2) \times q_j$ *matrix with* $L^j_{i,i} = L^j_{i,i+2} = 1$, $L^j_{i,i+1} = -2$, *and* $L^j_{i,l} = 0$ *otherwise in gspMCP and we set* L^j *as a* $q_j \times q_j$ *matrix with* $L^j_{i,i} = -1$, $L^j_{i,i-1} = 1$ *and* $L^j_{i,l} = 0$ *otherwise in gsmMCP. Therefore,*

$$Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{(m)}) \propto \frac{v}{2n}(\boldsymbol{Y}^* - \boldsymbol{Z}^\top\boldsymbol{\beta_0} - \boldsymbol{X}^*\boldsymbol{\beta}^*)^\top(\boldsymbol{Y}^* - \boldsymbol{Z}^\top\boldsymbol{\beta_0} - \boldsymbol{X}^*\boldsymbol{\beta}^*) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma),$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_J^\top)^\top$.

Remark 1. Given current estimate $\tilde{\boldsymbol{\beta}}^{(m)}$, we can majorize $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{(m)})$ by $\frac{v}{2n}(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^\top(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma)$, and minimize $\frac{v}{2n}(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^\top(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma)$ to get $\tilde{\boldsymbol{\beta}}^{(m+1)}$. We perform the majorizing step and the minimizing step in turn until convergence. The theory underlying MM algorithms ensures that this algorithm retains the descent property, which we will show in Appendix C.

Remark 2. We can perform certain transformation on $\boldsymbol{X}_j^*$, the jth group of the artificial dataset $\boldsymbol{X}^*$, using the method described in Appendix B to get orthonormal groups. After the transformation, we can apply the algorithm in Appendix A to get the solution to the orthonormalized problem. One can easily transform back to the solution to the original problem base on the results discussed in Appendix B.

We now present Algorithm 1 to solve the logistic regression model with gspMCP or gsmMCP.

Algorithm 1 Pseudocode for our algorithm

for do $j = 1$ to J
  $\frac{1}{n}\boldsymbol{X}_j^{*\top}\boldsymbol{X}_j^* = \boldsymbol{O}_j\boldsymbol{\Lambda}\boldsymbol{O}_j^\top$ (Singular value decomposition)
  $\widetilde{\boldsymbol{X}}_j = \boldsymbol{X}_j\boldsymbol{O}_j\boldsymbol{\Lambda}_j^{-1/2}$
end for
repeat
  $\boldsymbol{\eta} \leftarrow \widetilde{\boldsymbol{X}}\boldsymbol{\beta}$
  $\boldsymbol{\pi} \leftarrow e^{\eta_i}/(1 + e^{\eta_i})|_{i=1}^n$
  $\boldsymbol{r} = (\boldsymbol{Y} - \pi)/v$
  $\boldsymbol{\beta}_0 \leftarrow (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{r} + \boldsymbol{\beta}_0$
  for $j = 1$ to J **do**
    $\boldsymbol{u}_j = \widetilde{\boldsymbol{X}}_j^\top\boldsymbol{r} + \boldsymbol{\beta}_j$
    $\boldsymbol{\beta}_j' \leftarrow \frac{1}{v}F(v\boldsymbol{u}_j, \sqrt{q_j}\lambda_1, a)$
    $\boldsymbol{r} \leftarrow \boldsymbol{r} - \widetilde{\boldsymbol{X}}_j^\top(\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j)$
    $\boldsymbol{\beta}_j \leftarrow \boldsymbol{\beta}_j'$
  end for
until convergence
for $j = 1$ to J **do**
  $\boldsymbol{\beta}_j \leftarrow \boldsymbol{O}_j\boldsymbol{\Lambda}_j^{-1/2}\boldsymbol{\beta}_j$ (Transform back to the original problem)
end for

The algorithm of gMCP is presented in Appendix A. The details of orthonormalization can be seen in Appendix B. The theory underlying the algorithm ensures that it retains the descent property, which we formally state in Appendix C. As smooth-penalty may not always necessarily maintain sparsity and many small false features may have been left, we use a second-stage "thresholding" procedure when dealing with a gsmMCP problem. We refer to Guo et al[13] for more details.

4 | SIMULATION STUDIES

In this section, we conduct simulation studies to evaluate the performance of the proposed methods. We consider five different scenarios. Within each scenario, our simulated data consist of a training set and a testing set. Models are fitted on the training data, and then tested on the testing set. Following Lee et al,[19] the regularization parameters λ_1 and λ_2 are selected by 5-fold cross-validation base on the training set. We select the regularization parameters that achieve the maximum of the validated log-likelihood. The numbers of observations in the training set and testing set are n and $3n$, respectively. We consider different n in each scenario in order to evaluate the performances of the different methods under different sample sizes. Moreover, we fixed $\gamma = 3$ for logistic gMCP and the proposed approaches. Below are the detail settings of the five scenarios:

Scenario 1 The length of $\boldsymbol{\beta}$ is set to be 400. We divide $\boldsymbol{\beta}$ into 20 groups and $q_j = 20$ for $j = 1, 2, \ldots, 20$. In third, fourth, seventh, and eighth group, the coefficients $\boldsymbol{\beta}_j$s are nonzeros and chosen randomly in sequential order from the sine function $sin(t)$ where $t \in [0, 2\pi]$. $\boldsymbol{X}_j$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$. Two different covariance structures $\boldsymbol{\Sigma}$ are considered: $\Sigma = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$. The numbers of observations in the training set are set as $n = 100$ and 200, respectively.

Scenario 2 The setting of $\boldsymbol{\beta}$ is the same as in Scenario 1. $\boldsymbol{X}$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$. The numbers of observations in training set are set as $n = 100$, 200 and 300, respectively. We consider the cases of weak, moderate, and strong correlations by setting $\rho = 0.2$, 0.5, and 0.8, respectively.

TABLE 1 Summary of performance measures for Scenario 1. Means with standard errors (in parentheses) are obtained from 100 Monte Carlo repetitions

Method	Feature selection		Prediction error	
	Sensitivity	Specificity	AUC	ACC
$\Sigma = I$, n=100				
gMCP	0.403(0.266)	0.990(0.026)	0.663(0.096)	0.654(0.056)
gsmMCP	0.917(0.118)	0.834(0.191)	0.901(0.030)	0.829(0.029)
gspMCP	0.960(0.092)	0.879(0.126)	0.910(0.036)	0.840(0.037)
$\Sigma = I$, n=200				
gMCP	0.833(0.163)	0.998(0.012)	0.827(0.040)	0.754(0.034)
gsmMCP	0.980(0.068)	0.955(0.101)	0.928(0.019)	0.853(0.023)
gspMCP	0.993(0.043)	0.969(0.067)	0.956(0.015)	0.888(0.022)
$\Sigma_{ij} = 0.5^{\|i-j\|}$, n=100				
gMCP	0.430(0.266)	0.989(0.026)	0.679(0.094)	0.665(0.063)
gsmMCP	0.902(0.150)	0.892(0.144)	0.926(0.040)	0.857(0.043)
gspMCP	0.950(0.107)	0.915(0.106)	0.925(0.040)	0.857(0.042)
$\Sigma_{ij} = 0.5^{\|i-j\|}$, n=200				
gMCP	0.820(0.185)	0.999(0.009)	0.839(0.040)	0.765(0.035)
gsmMCP	0.985(0.060)	0.933(0.116)	0.959(0.017)	0.892(0.025)
gspMCP	0.983(0.064)	0.969(0.059)	0.968(0.016)	0.906(0.024)

Abbreviations: ACC, accuracy; AUC, area under curve; gMCP, group minimax concave penalty; gsmMCP, group smooth-MCP; gspMCP, group spline-MCP.

Scenario 3 The number of predictors is set to be 400. We divide the variables into 20 groups. And q_j for $j = 1, 2, \ldots, 20$ are not the same, where $q_1 = q_2 = 10$, $q_3 = q_4 = 15$, $q_5 = 30$, $q_6 = 40$, and $q_j = 20$ for $j = 7, 8, \ldots, 20$. In first, second, and sixth group, the coefficients $\boldsymbol{\beta}_j$s are nonzeros and chosen randomly in sequential order from the sine function $sin(t)$ where $t \in [0, 2\pi]$. $n = 100$ and 200, respectively.

Scenario 4 In this scenario, we increase the number of predictors and set to be 1000. We divide the variables into 50 groups and $q_j = 20$ for $j = 1, 2, \ldots, 50$. $\boldsymbol{\beta}_j$ for $j = 1, 2, \ldots, 20$ are the same as in Scenario 1, and $\boldsymbol{\beta}_j = 0$ for $j = 21, 22, \ldots, 50$. The generation of $\boldsymbol{X}$ is the same as in Scenario 1. $n = 100$ and 200, respectively.

TABLE 2 Summary of performance measures for Scenario 2. Means with standard errors (in parentheses) are obtained from 100 Monte Carlo repetitions

Sample size	Method	Feature selection		Prediction error	
		Sensitivity	Specificity	AUC	ACC
$\rho = 0.2$					
100	gMCP	0.427(0.264)	0.988(0.030)	0.675(0.093)	0.663(0.056)
	gsmMCP	0.958(0.101)	0.813(0.208)	0.924(0.031)	0.854(0.033)
	gspMCP	0.927(0.119)	0.942(0.079)	0.923(0.033)	0.854(0.035)
200	gMCP	0.840(0.157)	0.999(0.009)	0.837(0.040)	0.764(0.034)
	gsmMCP	0.988(0.055)	0.949(0.096)	0.952(0.017)	0.882(0.023)
	gspMCP	0.978(0.072)	0.989(0.030)	0.962(0.016)	0.897(0.023)
300	gMCP	0.910(0.144)	1.000(0.000)	0.880(0.033)	0.801(0.032)
	gsmMCP	0.988(0.055)	1.000(0.000)	0.960(0.013)	0.893(0.019)
	gspMCP	0.988(0.055)	1.000(0.000)	0.970(0.013)	0.909(0.020)
$\rho = 0.5$					
100	gMCP	0.475(0.247)	0.989(0.027)	0.702(0.090)	0.684(0.056)
	gsmMCP	0.973(0.086)	0.839(0.170)	0.945(0.028)	0.879(0.031)
	gspMCP	0.968(0.085)	0.938(0.076)	0.942(0.032)	0.877(0.034)
200	gMCP	0.807(0.166)	0.999(0.006)	0.839(0.044)	0.765(0.039)
	gsmMCP	0.990(0.049)	0.988(0.032)	0.974(0.012)	0.915(0.020)
	gspMCP	0.985(0.060)	0.996(0.018)	0.972(0.013)	0.911(0.020)
300	gMCP	0.877(0.149)	1.000(0.000)	0.884(0.038)	0.804(0.034)
	gsmMCP	0.995(0.035)	1.000(0.000)	0.977(0.009)	0.920(0.015)
	gspMCP	0.998(0.025)	1.000(0.000)	0.980(0.009)	0.926(0.016)
$\rho = 0.8$					
100	gMCP	0.465(0.204)	0.999(0.009)	0.752(0.066)	0.713(0.042)
	gsmMCP	0.940(0.118)	0.929(0.086)	0.947(0.031)	0.881(0.034)
	gspMCP	0.960(0.092)	0.961(0.057)	0.957(0.028)	0.896(0.034)
200	gMCP	0.667(0.188)	1.000(0.000)	0.845(0.047)	0.771(0.041)
	gsmMCP	0.973(0.079)	0.996(0.015)	0.984(0.015)	0.937(0.026)
	gspMCP	0.975(0.075)	0.998(0.011)	0.977(0.013)	0.920(0.022)
300	gMCP	0.750(0.178)	1.000(0.000)	0.882(0.041)	0.804(0.038)
	gsmMCP	0.995(0.035)	1.000(0.000)	0.989(0.008)	0.947(0.016)
	gspMCP	0.990(0.049)	1.000(0.000)	0.983(0.009)	0.933(0.017)

Abbreviations: ACC, accuracy; AUC, area under curve; gMCP, group minimax concave penalty; gsmMCP, group smooth-MCP; gspMCP, group spline-MCP.

Scenario 5 The number of predictors is set to be $400 + n$, where n is the sample size of the training set. We divide the variable into 20 groups. And q_j for $j = 1, 2, \ldots, 20$ are not the same, where $q_1 = q_2 = 10$, $q_3 = q_4 = 15$, $q_5 = 30$, $q_6 = 40 + n$, and $q_j = 20$ for $j = 7, 8, \ldots, 20$. The key difference between this scenario and Scenario 4 is that, in this scenario, the number of variables in the sixth group is larger than the sample size n. In first, second, and sixth group, the coefficients $\boldsymbol{\beta}_j$s are nonzeros and chosen randomly in sequential order from the sine function $sin(t)$ where $t \in [0, 2\pi]$.

The performances of different methods are examined in three aspects: feature selection, group selection, and prediction error. We compare sensitivities and specificities for different methods in terms of feature selection as well as group selection. Sensitivity for feature selection measures the proportion of correctly-identified variables with nonzero coefficient. Specificity for feature selection measures the proportion of correctly-identified variables with zero coefficient. Sensitivity for group selection measures the proportion of correctly-identified groups that content nonzero coefficients. Specificity for group selection measures the proportion of correctly-identified groups that have all zero coefficients. To measure the prediction error, we generate testing sets for each example and use the fitted models from the training set to predict on the testing set. Area under curve (AUC) and accuracy (ACC) of the testing results are reported. Here, ACC is the proportion of true cases (both true positives and true negatives) among the total number of cases examined.

Here we only display the summary tables for Scenarios 1, 2, and 3 (see Tables 1, 2, and 3). The results for other scenarios can be found in Appendix S1. In each scenario, the performances of the three methods all get better as the sample size n increases. The sensitivities of gsmMCP and gspMCP are much higher than those of gMCP, which means that both the models with group spline-penalty and the models with group smooth-penalty can select more true parameters than using gMCP. As for prediction error, AUC and ACC are improved when using group spline-penalty or group smooth-penalty. The proposed methods, gspMCP and gsmMCP, are not affected by the neighboring correlation between groups

TABLE 3 Summary of performance measures for Scenario 3. Means with standard errors (in parentheses) are obtained from 100 Monte Carlo repetitions

	Feature selection		Group selection		Prediction error	
Method	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**AUC**	**ACC**
$\Sigma = I, n = 100$						
gMCP	0.522(0.364)	0.988(0.027)	0.503(0.305)	0.988(0.028)	0.681(0.110)	0.672(0.067)
gsmMCP	0.972(0.071)	0.815(0.200)	0.943(0.143)	0.814(0.198)	0.875(0.022)	0.802(0.024)
gspMCP	0.967(0.082)	0.935(0.076)	0.933(0.164)	0.934(0.078)	0.924(0.028)	0.854(0.031)
$\Sigma = I, n = 200$						
gMCP	0.935(0.094)	1.000(0.004)	0.870(0.189)	0.999(0.006)	0.865(0.026)	0.787(0.025)
gsmMCP	0.972(0.075)	0.991(0.026)	0.943(0.150)	0.991(0.027)	0.885(0.019)	0.806(0.018)
gspMCP	0.990(0.046)	0.998(0.010)	0.980(0.093)	0.998(0.010)	0.954(0.013)	0.886(0.018)
$\Sigma_{ij} = 0.5^{\lvert i-j \rvert}, n = 100$						
gMCP	0.713(0.260)	0.992(0.021)	0.567(0.262)	0.991(0.023)	0.771(0.098)	0.728(0.061)
gsmMCP	0.950(0.087)	0.875(0.135)	0.900(0.174)	0.874(0.136)	0.926(0.021)	0.854(0.024)
gspMCP	0.945(0.092)	0.959(0.065)	0.890(0.184)	0.959(0.064)	0.948(0.027)	0.882(0.034)
$\Sigma_{ij} = 0.5^{\lvert i-j \rvert}, n = 200$						
gMCP	0.908(0.110)	1.000(0.000)	0.817(0.219)	1.000(0.000)	0.894(0.024)	0.814(0.025)
gsmMCP	0.983(0.050)	1.000(0.004)	0.967(0.101)	0.999(0.006)	0.951(0.012)	0.880(0.016)
gspMCP	0.987(0.051)	0.998(0.013)	0.973(0.102)	0.998(0.014)	0.975(0.015)	0.918(0.022)

Abbreviations: ACC, accuracy; AUC, area under curve; gMCP, group minimax concave penalty; gsmMCP, group smooth-MCP; gspMCP, group spline-MCP.

(see Table 2) or the numbers of features in each group (see Tables S1 and S2 in Appendix S1). We can conclude that the logistic gspMCP and logistic gsmMCP both outperform logistic gMCP, while the performances of logistic gspMCP seem equivalent to those of logistic gsmMCP .

5 | APPLICATION TO THE LUNG CANCER STUDY

In this section, we apply the proposed methods to the lung cancer study, which is introduced in Section 2. There are 270 subjects, 92 of which are cancer and the other 178 are control. We analyzed the data using the gMCP, gsmMCP, and gspMCP. The regularization parameters λ_1 and λ_2 are selected by 5-fold cross-validation using the training set. To show the prediction performance of the proposed methods comparing to logistic regression with gMCP, we randomly choose 220 samples from data as the training set, and the remaining samples are used as the testing set. We obtain the final estimates of the regression coefficients based on training test and calculate the sensitivity, specificity, ACC, and AUC on the testing set. The sensitivity are 0.471 (gMCP), 0.647 (gsmMCP), and 0.706 (gspMCP), respectively. The specificity are 0.697 (gMCP), 0.818 (gsmMCP), and 0.818 (gspMCP), respectively. The ACC are 0.620 (gMCP), 0.760 (gsmMCP), and 0.780 (gspMCP), respectively. The AUC are 0.620 (gMCP), 0.800 (gsmMCP), and 0.804 (gspMCP), respectively. Figure 2 shows the ROC curves of three methods for both the training set and testing set. The two proposed methods are much better than the gMCP.

Table 4 shows group selection results for three methods. The numbers of groups identified are 4 for gMCP, 12 for gsmMCP, and 14 for gspMCP, respectively. In Table 4, "S_13_UV2" means the second ultra-color spectrum of thirteenth colorant and other symbols follow the same naming rule. The numbers of selected groups by gsmMCP and gspMCP are larger than gMCP. This result is not surprising since that gsmMCP and gspMCP are designed to capture the smooth changes of features. As shown in simulation, both the models with group spline-penalty and the models with group smooth-penalty can select more true parameters than gMCP.

One problem of the single sample-splitting method is that it is sensitive with respect to the choice of splitting the entire sample: different split sample may lead to different results. To overcome the "randomness" of splitting, we run the sample-splitting method 100 times, and summarize the results. The details of comparison among three methods in terms of sensitivity, specificity, ACC, and AUC are shown in Figure S1 and Table S3 in Appendix S1. There is no much difference in terms of specificity for the three methods, but we see a great improvement in terms of sensitivity, AUC, and ACC by using the proposed methods.

From the result, we can conclude that the gspMCP and gsmMCP both perform better than the gMCP for logistic regression. The performances of the two proposed methods are similar. This is not surprising since the gspMCP and gsmMCP methods can capture smoothing changes in coefficients within groups and select important groups. This leads to a higher detection rate of lung cancer than the gMCP method.

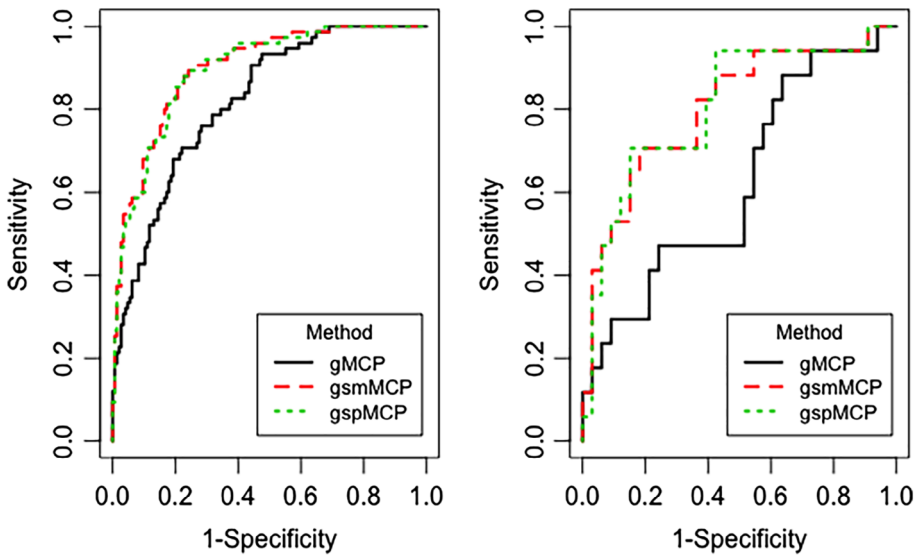

FIGURE 2 ROC curves for three methods: training set (left) and testing set (right) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Group selection results for three methods in application.

Method	gMCP	gsmMCP	gspMCP
Age	√	√	√
Sex		√	√
COPD		√	√
Smoking (current)		√	√
S_13_UV2		√	√
S_33_UV3			√
S_34_UV4		√	√
S_35_UV1			√
S_35_UV3		√	√
S_35_UV4	√	√	√
S_41_UV1		√	√
S_46_UV2		√	√
S_81_UV3	√	√	√
S_85_UV2	√	√	√

Abbreviations: gMCP, group minimax concave penalty; gsmMCP, group smooth-MCP; gspMCP, group spline-MCP.

6 | DISCUSSION

This article is motivated by an analysis of breath VOCs with the inclusion of clinical risk factors for lung cancer diagnosis. The predictors include high-dimensional time-dependent VOC features as well as time-independent clinical variables. Thus, we consider the high-dimensional logistic regression problem and propose two different penalties: gspMCP or gsmMCP to handle the group structures of the time-dependent variables.

The proposed methods may advance from the existing ones along the following aspects. First, the methods are the first attempt to investigate coefficients that are sparse and change smoothly within the group. Second, our methods are easy to implement since they can be turned into a gMCP problem after some transformations. We propose an efficient algorithm to solve the problem, and this algorithm possesses the descent property and leads to attractive convergence properties. The algorithm with the supporting mathematical optimization theory can also apply to other generalized linear models. We apply the new methods on the analysis of breath VOCs and focus on the prediction of lung cancer. The results show that the predictions of the proposed methods are more accurate comparing to that using logistic regression with gMCP. We have concentrated on the group selection by using gMCP as the group penalty. The proposed ideas can be extended for other group selection and bi-level selection methods.

ACKNOWLEDGMENTS

We are grateful to the associate editor and the reviewers for their valuable suggestions which substantially improved the paper. This study was supported by the National Natural Science Foundation of China (71471152), Fundamental Research Funds for the Central Universities (20720181003,20720171064,20720171095), and Humanity and Social Science Youth Foundation of Ministry of Education of China (19YJC910010).

ORCID

Shuangge Ma https://orcid.org/0000-0001-9001-4999
Kuangnan Fang https://orcid.org/0000-0003-0934-7281

REFERENCES

1. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25(2):173-187.
2. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J Royal Stat Soc: Series B*. 2006;68(1):49-67.

3. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J Royal Stat Soc: Series B*. 2008;70(1):53-71.
4. Zhao P, Rocha G, Yu B. Grouped and hierarchical model selection through composite absolute penalties. Paper presented at Department of Statistics, UC Berkeley, Tech. Rep. 2006: 703.
5. Wang L, Li H, Huang JZ. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J Am Stat Assoc*. 2008;103(484):1556-1569.
6. Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Statist Sci*. 2012;27(4):481-499.
7. Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for variable selection. *Biometrika*. 2009;96(2):339-355.
8. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *J Royal Stat Soc: Series B*. 2005;67(1):91-108.
9. Cao P, Liu X, Liu H, et al. Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimers disease. *Comput Methods Prog Biomed*. 2018;162:19-45.
10. Ciuperca G. Adaptive fused lasso in grouped quantile regression. *J Stat Theory Pract*. 2017;11(1):107-125.
11. Hebiri M, Van De Geer S. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron J Stat*. 2011;5:1184-1226.
12. Huang J, Ma S, Li H, Zhang CH. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann Stat*. 2011;39(4):2021-2046.
13. Guo J, Hu J, Jing BY, Zhang Z. Spline-Lasso in high-dimensional linear regression. *J Am Stat Assoc*. 2016;111(513):288-297.
14. Tran VH, Chan HP, Thurston M, et al. Breath analysis of lung cancer patients using an electronic nose detection system. *IEEE Sensors J*. 2010;10(9):1514-1518.
15. Mazzone PJ, Hammel J, Dweik R, et al. Lung cancer diagnosis by the analysis of exhaled breath with a colorimetric sensor array. *Thorax*. 2007;62(7):565-568.
16. Mazzone PJ, Wang XF, Xu Y, et al. Exhaled breath analysis with a colorimetric sensor array for the identification and characterization of lung cancer. *J Thorac Oncol*. 2012;7(1):137-142.
17. Janzen MC, Ponder JB, Bailey DP, Ingison CK, Suslick KS. Colorimetric sensor arrays for volatile organic compounds. *Anal Chem*. 2006;78(11):3591-3600.
18. Liu J, Ma S, Huang J. Integrative analysis of cancer diagnosis studies with composite penalization. *Scand J Stat*. 2014;41(1):87-103.
19. Lee S, Shin H, Lee SH. Label-noise resistant logistic regression for functional data classification with an application to Alzheimer's disease study. *Biometrics*. 2016;72(4):1325-1335.
20. Ortega JM, Rheinboldt WC. *Iterative Solution of Nonlinear Equations in Several Variables, Classics in Applied Mathematics*. 4th ed. Philadeplhia, PA: SIAM; 2000.
21. Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions. *J Comput Graph Stat*. 2000;9(1):1-20.
22. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optimiz Theory Appl*. 2001;109(3):475-494.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang X, Zhang Q, Wang X, Ma S, Fang K. Structured sparse logistic regression with application to lung cancer prediction using breath volatile biomarkers. *Statistics in Medicine*. 2020;39:955–967. https://doi.org/10.1002/sim.8454

APPENDIX A: COMPUTATIONAL ALGORITHM FOR LOGISTIC REGRESSION WITH GMCP

Here we describe the computational algorithm for logistic regression with gMCP. Note that the estimator is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n}\ell(\boldsymbol{\beta}) + \sum_{j=1}^{J} P_{\mathrm{MCP}}\left(\left\|\boldsymbol{\beta}_j\right\|\right); \sqrt{q_j}\lambda, \gamma) \right\},$$

where $\|\cdot\|$ is the standard Euclidean norm, $\lambda \geq 0$ is the penalty parameter, $\sqrt{q_j}$s in the penalty are used to adjust for the group sizes, and γ is an additional tuning parameter that is used to modify $P_{\mathrm{MCP}}(\theta, \lambda, \gamma)$. Under the logistic regression

model, there is no simple, closed-form solution for $\hat{\boldsymbol{\beta}}$. To tackle this problem, Liu et al[18] resort to a MM approach,[20] which majorizes the negative log-likelihood function by a quadratic loss:

$$-\frac{1}{n}\tilde{\ell}(\boldsymbol{\beta}) \propto \frac{v}{2n}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}^{\top}\boldsymbol{\beta_0} - \sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right)^{\top}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}^{\top}\boldsymbol{\beta_0} - \sum_{j=1}^{J}\boldsymbol{X}_j^{\top}\boldsymbol{\beta}_j\right),$$

where $v = 1/4$ and $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{\eta}} + (\boldsymbol{Y} - \boldsymbol{p})/v$ is the pseudo-response vestor. Here we require the groups $\boldsymbol{X}_j$ to be orthonormal (more details about orthonormalization are discussed in Appendix B), then the gradient of $\tilde{\ell}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta_j}$ is given by $\nabla\tilde{\ell}(\boldsymbol{\beta_j}) = -v\boldsymbol{u}_j + v\boldsymbol{\beta}_j$, where $\boldsymbol{u}_j = \boldsymbol{X}_j^{\top}(\tilde{\boldsymbol{Y}} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j})$ is the unpenalized (maximum likelihood) solution for $\boldsymbol{\beta}_j$, $\boldsymbol{X}_{-j}$ is the portion of the design matrix that remains after $\boldsymbol{X}_j$ has been excluded, and $\boldsymbol{\beta}_{-j}$ are its associated regression coefficients.

The updating equation for $j = 1, \ldots, J$ is then:

$$\boldsymbol{\beta}_j \leftarrow \frac{1}{v}F(v\boldsymbol{u}_j, \sqrt{q_j}\lambda, \gamma) = \frac{1}{v}F(v\left\|\boldsymbol{u}_j\right\|, \sqrt{q_j}\lambda, \gamma)\frac{\boldsymbol{u}_j}{\left\|\boldsymbol{u}_j\right\|},$$

where

$$F(u, \lambda, \gamma) = \begin{cases} \frac{S(u,\lambda)}{1-1/\gamma} & if \quad |u| \le \gamma\lambda \\ u & if \quad |u| > \gamma\lambda \end{cases},$$

and

$$S(u, \lambda) = \begin{cases} u - \lambda & if \quad u > \lambda \\ 0 & if \quad |u| \le \lambda \\ u + \lambda & if \quad u < -\lambda \end{cases}.$$

The parameters $\boldsymbol{\beta}_j$ for $j = 1, \ldots, J$ are updated by the above updating equation until convergence.

APPENDIX B: ORTHONORMALIZATION

We address here how one can apply the algorithm for logistic regression with gMCP in Appendix A for the nonorthonormal groups. We perform transformations on nonorthonormal groups following the idea by Breheny and Huang.[1] Taking the singular value decomposition of the Gram matrix of the jth group, we have

$$\frac{1}{n}\boldsymbol{X}_j^{\top}\boldsymbol{X}_j = \boldsymbol{O}_j\boldsymbol{\Lambda}\boldsymbol{O}_j^{\top},$$

where $\boldsymbol{\Lambda}_j$ is a diagonal matrix containing the eigenvalues of $\frac{1}{n}\boldsymbol{X}_j^{\top}\boldsymbol{X}_j$ and $\boldsymbol{O}_j$ is an orthonormal matrix of its eigenvectors.

Now, we may construct a linear transformation $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j\boldsymbol{O}_j\boldsymbol{\Lambda}_j^{-1/2}$ with the following properties:

$$\frac{1}{n}\tilde{\boldsymbol{X}}_j^{\top}\tilde{\boldsymbol{X}}_j = \boldsymbol{I},$$

$$\tilde{\boldsymbol{X}}_j\tilde{\boldsymbol{\beta}}_j = \boldsymbol{X}_j(\boldsymbol{O}_j\boldsymbol{\Lambda}_j^{-1/2}\tilde{\boldsymbol{\beta}}_j),$$

where $\boldsymbol{I}$ is the identity matrix, and $\tilde{\boldsymbol{\beta}}_j$ is the solution on the orthonormalized scale.

This orthonormalization can greatly reduce the computational burden associated with fitting gMCP models. Besides, this orthonormalization can be accomplished without loss of generality since we may easily transform back to the original problem with $\boldsymbol{\beta}_j = \boldsymbol{O}_j\boldsymbol{\Lambda}_j^{-1/2}\tilde{\boldsymbol{\beta}}_j$. It is worthy to mention that these transformations do not change the sparsity of coefficients. Thus, an analyst does not have to worry about issues of orthonormality when applying these algorithms to nonorthonormal group variables.[1] From a computational standpoint, this procedure is not very expensive as the decompositions are being applied only to the groups, not the entire design matrix. Furthermore, because $\boldsymbol{\Lambda}_j$ is diagonal,

the inverses are computed easily. The decompositions only need to be computed once at the initial step, not for every iteration.

APPENDIX C: DESCENT PROPERTY OF ALGORITHM 1

Before proving descent property of Algorithm 1, we establish the group-wise convexity of the objective function. Although the objective function contains nonconvex components and is not necessarily convex overall, the objective function is still convex with respect to the variables in a single group.

Lemma 1. *Let $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ denote the majorizing approximation to $Q(\boldsymbol{\beta})$ at $\tilde{\boldsymbol{\beta}}$. Then $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ is a strictly convex function with respect to $\boldsymbol{\beta}_j$ at $\tilde{\boldsymbol{\beta}}$ for logistic gspMCP and gsmMCP with $\gamma > \frac{4}{5}$.*

Proof. Let $\bigtriangledown^2_{\boldsymbol{d}} Q(\boldsymbol{\beta}_j|\tilde{\boldsymbol{\beta}})$ denote the second derivative of $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ in the direction $\boldsymbol{d}$. Then the strict convexity of $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ follows if $\bigtriangledown^2_{\boldsymbol{d}} Q(\boldsymbol{\beta}_j|\tilde{\boldsymbol{\beta}})$ is positive definite at all $\boldsymbol{\beta}_j$ and for all $\boldsymbol{d}$. Let ξ_* denote the infimum over $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\beta}_j$, and $\boldsymbol{d}$ of the minimum eigenvalue of $\bigtriangledown^2_{\boldsymbol{d}} Q(\boldsymbol{\beta}_j|\tilde{\boldsymbol{\beta}})$, we obtain

$$\xi_* = \frac{1}{4} - \frac{1}{\gamma} + 1,$$

These quantities are positive under the conditions specified in the lemma. ∎

Proposition 2. *Let $\boldsymbol{\beta}^{(m)}$ denote the value of the fitted regression coefficient at the end of iteration m (after transforming back to the original problem with $\boldsymbol{\beta}_j = \boldsymbol{O}_j \boldsymbol{\Lambda}_j^{-1/2} \tilde{\boldsymbol{\beta}}_j$). At every iteration of the proposed group descent algorithm for logistic group spline-penalty or group smooth-penalty,*

$$Q(\boldsymbol{\beta}^{(\boldsymbol{m+1})}) \leq Q(\boldsymbol{\beta}^{(\boldsymbol{m})}),$$

where $Q(\boldsymbol{\beta}) = -\frac{1}{n}\ell(\boldsymbol{\beta}) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma) + \sum_{j=1}^{J_1} P_2(\boldsymbol{\beta}_j, \lambda_2)$.

Furthermore, every limit point of the sequence $\{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(1)}, \ldots\}$ is a stationary point of Q, provided that no elements of $\boldsymbol{\beta}$ tend to $\pm\infty$.

Proof. The proposition makes two claims: descent with every iteration and convergence to stationary points. To establish descent for logistic regression, we note that because ℓ is twice differentiable, for any point $\boldsymbol{\eta}$ there exists a vector $\boldsymbol{\eta}^{**}$ on the line segment joining $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ such that

$$\ell(\boldsymbol{\eta}) = \ell(\boldsymbol{\eta}^*) + (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \bigtriangledown \ell(\boldsymbol{\eta}^*) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \bigtriangledown^2 \ell(\boldsymbol{\eta}^{**})(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \leq \tilde{\ell}(\boldsymbol{\eta}|\boldsymbol{\eta}^*),$$

where the inequality follows from the fact that $v\boldsymbol{I} - \bigtriangledown^2 \ell(\boldsymbol{\eta}^{**})$ is a positive semidefinite matrix. This means

$$-\frac{1}{n}\ell(\boldsymbol{\eta}) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma) + \sum_{j=1}^{J_1} P_2(\boldsymbol{\beta}_j, \lambda_2) \leq -\frac{1}{n}\tilde{\ell}(\boldsymbol{\eta}|\boldsymbol{\eta}^*) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma) + \sum_{j=1}^{J_1} P_2(\boldsymbol{\beta}_j, \lambda_2),$$

$$\propto \frac{v}{2n}(\tilde{\boldsymbol{Y}} - \sum_{j=0}^{J} \boldsymbol{X}_j^\top \boldsymbol{\beta}_j)^\top (\tilde{\boldsymbol{Y}} - \sum_{j=0}^{J} \boldsymbol{X}_j^\top \boldsymbol{\beta}_j) + P_1(\boldsymbol{\beta}; \lambda_1, \gamma) + \sum_{j=1}^{J_1} P_2(\boldsymbol{\beta}_j, \lambda_2).$$

Descent now follows from the descent property of MM algorithms[21] coupled with the fact that each updating step consists of minimizing $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$.

If no elements of $\boldsymbol{\beta}$ tends to $\pm\infty$, then the descent property of the algorithm ensures that the sequence $\boldsymbol{\beta}^{(k)}$ stays within a compact set and therefore possesses a limit point $\tilde{\boldsymbol{\beta}}$. Lemma 1 provides sufficient conditions to apply Theorem 4.1 of Tseng[22] and conclude that $\tilde{\boldsymbol{\beta}}$ must be a stationary point of $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$. Furthermore, because $Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ is tangent to $Q(\boldsymbol{\beta})$ at $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ must also be a stationary point of $Q(\boldsymbol{\beta})$. ∎