Received: 13 January 2017 | Revised: 9 April 2018 | Accepted: 6 August 2018

DOI: 10.1002/qre.2376

RESEARCH ARTICLE

WILEY

# Phase I outlier detection in profiles with binary data based on penalized likelihood

Zhen Li[1] | Yanfen Shang[2] | Zhen He[2]

[1]Tianjin University, College of Management and Economics, Tianjin, Tianjin, China

[2]Industrial Engineering, College of Management and Economics, Tianjin University, Tianjin, Tianjin, China

**Correspondence**
Yanfen Shang, Industrial Engineering, College of Management and Economics, Tianjin University, Nankai District, Tianjin, Tianjin 300072, China.
Email: syf8110@gmail.com

**Funding information**
National Natural Science Foundation of China, Grant/Award Numbers: 71672122, 71532008, 71401123, 71402118 and 71472132

**Abstract**

Profile monitoring has been proven to be important among statistical process control problems. In some specific applications, the response variable of a profile can be categorical data, and numerous methods have been proposed for monitoring this type of profiles. Nevertheless, outlier detection for profiles with categorical data still attracted insufficient attention in the literature. To this end, this paper focuses on binary profiles and develops two schemes for outlier detection, from the viewpoint of penalized likelihood, based on the group LASSO method and directional information, respectively, in which the profiles of interest are treated as an integrated high-dimensional vector. Simulation study shows that the proposed group-type scheme usually performs better than the existing $T_I^2$ chart in terms of detecting outliers correctly and alleviating the masking effect. Finally, a real example is used for illustrating the implementation of the proposed scheme.

**KEYWORDS**

binary profile, group LASSO, masking effect, outlier detection, penalized likelihood

## 1 | INTRODUCTION

It has been well demonstrated that statistical process control (SPC) is a remarkable tool for monitoring process or product quality in both manufacturing and service industries. A variety of schemes have been developed to tackle univariate and multivariate SPC problems. However, in some industrial applications, the quality of a process or product is characterized preferably by a relationship between the response variable and one or more explanatory variables, which is typically referred to as a profile. In the past several decades, profile monitoring has been studied extensively in the literature. Some of the previous works focused on Phase I (cf. Mahmoud and Woodall,[1] Mahmoud et al,[2] Zhang and Albin,[3] Zou et al[4]), while some other works focused on Phase II (cf. Kang and Albin,[5] Kim et al,[6] Zou et al,[7] Noorossana et al[8]). In this paper, we are mainly concerned with the Phase I problem.

Focusing on Phase I monitoring of linear profiles, Mestek et al[9] proposed a $T^2$ control chart based on successive response vectors. Mahmoud and Woodall[1] pointed out that the $T^2$ method introduced by Mestek et al[9] is inapplicable when the explanatory variables are random and thus they developed a $F$ test approach which combined the indicator variables (dummy variables) with a univariate control chart. Mahmoud et al[2] proposed a change point method to analyze Phase I profile data by using a likelihood ratio test (LRT) in a segmented simple linear regression model. To monitor simple linear profiles with individual observations, Yeh et al[10] developed a Phase I control chart based on the conventional assumption that the error term follows a normal distribution. For Phase I monitoring of multiple linear profiles, Mahmoud[11] suggested monitoring the intercept, the slope, and the variance of a simple linear profile by treating the average fitted response vector as the explanatory variable.

  

As for nonlinear profiles, Williams et al[12] introduced four $T^2$ methods based on nonlinear regression. Considering both within-profile and between-profile variations, Paynabar and Jin[13] developed a wavelet-based mixed-effect model for complex nonlinear profiles. Zhang and Albin[3] proposed a $\chi^2$ control chart by treating profiles as high-dimensional vectors and argued that the $\chi^2$ chart is of great help for detecting outliers for nonlinear profiles especially when the profiles are very complex. Zou et al[4] indicated that the $\chi^2$ control chart does outperform its counterparts in the case of complex profiles, but it suffers from a certain masking effect. For this reason, they proposed a new procedure for outlier detection, based on penalized regression, by treating profiles as vectors following Zhang and Albin.[3] In certain cases, a single profile may require quite a long time to generate, which is a challenging problem for conventional SPC methods. Motivated by an ingot growth process in semiconductor manufacturing, Dai et al[14] proposed a method for monitoring the dynamically growing profile trajectory to detect unexpected changes during the long processing cycle. In a recent paper, Paynabar et al[15] focused on Phase I analysis of multichannel nonlinear profiles and proposed a new framework, in which the monitoring statistics is constructed by incorporating the multi-dimensional functional principal component analysis into change-point models.

All the aforementioned works rely on a basic assumption that the response variable of the profile is continuous. However, it is very likely that only categorical data are available in practical applications, as a result of which this assumption can be violated. To this end, Yeh et al[16] investigated a profile problem whose response variable is binary. In their study, the logistic regression model was used to express the relationship between the response variable and the explanatory variables and five $T^2$ control charts were proposed and compared in terms of the signal probability in various scenarios. Integrating the EWMA scheme and the LRT, Shang et al[17] studied Phase II monitoring and proposed a novel control chart for binary profiles with random covariates. Paynabar et al[18] developed a Phase I chart for surgical-operation improvement, in which the surgical outcomes are binary. The authors constructed the control chart via a likelihood-ratio test derived from a change-point model ($LRT_{CP}$) based on the risk-adjustment logistic regression. Recently, Shadman et al[19] proposed a unified framework combining a change point model with a generalized linear model, which can be used to develop Phase I control charts for profiles with continuous, count, or categorical data.

In a typical Phase I profile problem, one of the most significant steps is to identify the outlying profiles among the reference dataset and remove them so that a reliable functional curve can be established for use in Phase II.[20] However, to the best of our knowledge, there are limited previous works specializing in outlier detection for profile processes with categorical data. Although the $T_I^2$ control chart proposed by Yeh et al[16] can be used to detect outliers, its performance is only studied in terms of probability of signal, which is incapable of determining the masking effect (see Zou et al[4] for details). Therefore, in this study, we aim to develop new schemes for detecting outliers efficiently and analyze their detecting performance more thoroughly. The differences between our work and the existing research are the following: first of all, our study focuses on diagnosing the outliers in Phase I profiles with binary data. Therefore, it is different from those existing schemes which are about change-point detection, Phase II profile monitoring, and/or profiles with numerical data. Moreover, our proposed schemes can be extended to other profiles with categorical data, and the extension is presented in the following section. Secondly, the proposed schemes are quite different from the control-chart-based methods, such as the $T_I^2$ chart in Yeh et al.[16] To use this kind of scheme, the control limit is necessary. So, if it cannot be calculated, and then it is usually obtained by simulation by utilizing more data and/or time. However, our schemes do not need the control limit and then can save data and time. Moreover, the superiority in detecting performance is investigated in the following by comparing with the $T_I^2$ chart.

Note that the outliers in a given dataset generally possess the sparsity property, which means only a small portion of profiles are outliers. Based on this property, there have been several methods for detecting outliers, such as cluster-based methods, classification methods, machine learning methods, statistical model-based methods, and so on. The detailed review can be found in Hodge and Austin[21] and Jobe and Pokojovy.[22] Considering that variable selection can be used for selecting sparsity model, so we detect outliers based on it in this paper. In fact, variable selection methods have been widely used in the SPC literature. Zou and Qiu[23] adapted the LASSO to the multivariate SPC problem and proposed a Phase II control chart by integrating the LASSO-based test statistic with the EWMA procedure to determine the shift direction based on observed data. Combining Bayesian Information Criterion (BIC) with the adaptive LASSO, Shang et al[24] developed a diagnosis scheme for multistage processes with binary outputs. Therefore, inspired by Zhang and Albin,[3] in this article, we transform the historical profiles into a high-dimensional vector, and then develop an outlier detection scheme based on the group LASSO method which can be found in Yuan and Lin[25] and Meier et al.[26] Furthermore, the priori directional information

has been utilized to tackle various SPC problems and was recently applied to binary profiles by Shang et al.[27] Hence, we propose another scheme based on the directional information.

The rest of this paper is organized as follows. In Section 2, the model setup is discussed briefly. In Section 3, we propose two schemes for outlier detection based on penalized likelihood and discuss their extensions. In Section 4, the detecting performance of our proposed schemes is studied via simulations. A real example is given in Section 5. Concluding remarks are presented in Section 6.

## 2 | MODEL SETUP

Suppose we have an off-line dataset that contains $m$ binary profiles. In each profile, there are $n$ fixed independent experimental settings. The $p$-dimensional vector of the explanatory variables corresponding to a specific experimental setting is denoted by

$$\mathbf{x}_j = \left(x_{j0}\cdots, x_{j(p-1)}\right)^T,\ \ j = 1, \cdots n.$$

$\pi_{ij}$ denotes the probability that the quality characteristic of interest fails in the jth experimental setting of the ith profile. In the context of the logistic regression model, $\pi_{ij}$ is supposed to be a function of $\mathbf{x}_j$ and the link function of the ith profile can be expressed as follows,

$$\begin{aligned}\text{logit}\left(\pi_{ij}\right) &= \ln\frac{\pi_{ij}}{1-\pi_{ij}} = \mathbf{x}_j^T\boldsymbol{\beta_i},\ \ i = 1, \cdots, m; j \\ &= 1, \cdots n,\end{aligned} \tag{1}$$

where $\boldsymbol{\beta_i} = (\beta_{i0}, \cdots, \beta_{i(p\ -\ 1)})^T$ is the parameter vector of the logistic regression model for the ith profile. Moreover, it is customary to set $x_{j0} \equiv 1$ so that $\beta_{i0}$ will be the intercept of the model.

Assume that $\boldsymbol{\beta_0} = (\beta_{00}, \cdots, \beta_{0(p\ -\ 1)})^T$ is the in-control parameter vector of the process of interest. Apparently, the ith profile will be considered as an outlier if $\boldsymbol{\beta_i} \neq \boldsymbol{\beta_0}$. Note that $\boldsymbol{\beta_0}$ is usually unavailable in Phase I. We replace $\boldsymbol{\beta_0}$ by its robust estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ which is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathbf{0}} = \text{median}\left\{\widehat{\boldsymbol{\beta}}_{\mathbf{1}}, \cdots, \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}\right\},$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is the maximized likelihood estimator obtained by solving iterative weighted least square (IWLS, see McCullagh and Nelder[28]).

For each profile, we define a shift as $\boldsymbol{\delta_i} = \boldsymbol{\beta_i} - \widehat{\boldsymbol{\beta}}_{\mathbf{0}}$, and then the integrated shift vector $\boldsymbol{\delta} = \left(\boldsymbol{\delta}_1^T\cdots\boldsymbol{\delta}_m^T\right)^T$. Consequently, the problem can be further expressed as follows,

$$\boldsymbol{\delta_i} = \begin{cases} \mathbf{0}, & \text{if } i \notin \Omega \\ \boldsymbol{\delta}_\Omega, & \text{if } i \in \Omega \end{cases},$$

where $\Omega$ denotes the set of outliers which is a subset of $\{1, \cdots, m\}$ and $\boldsymbol{\delta}_\Omega$ denotes a nonzero vector corresponding to some assignable cause. Then, we aim to detect those nonzero $\boldsymbol{\delta_i}$s.

## 3 | METHODOLOGY

Suppose that there are $N_{ij}$ Bernoulli observations in the jth experimental setting of the ith profile, and thus the observation $y_{ij}$ follows a binomial distribution, ie,

$$y_{ij} \sim \text{BIN}\left(N_{ij}, \pi_{ij}\right).$$

Assuming that the observations are independent, we can obtain the following joint log-likelihood function,

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\ln\binom{N_{ij}}{y_{ij}} + y_{ij}\ln\pi_{ij} + \left(N_{ij} - y_{ij}\right)\ln\left(1-\pi_{ij}\right)\right).$$

By substituting $\left(\boldsymbol{\delta_i} + \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right)$ for $\boldsymbol{\beta_i}$, $\pi_{ij}$ can be expressed as follows,

$$\pi_{ij} = \frac{\exp\left(\mathbf{x}_j^T\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\delta_i}\right)\right)}{1 + \exp\left(\mathbf{x}_j^T\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\delta_i}\right)\right)}.$$

Thus, the log-likelihood function becomes the following form,

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\begin{array}{c}\ln\binom{N_{ij}}{y_{ij}} + y_{ij}\mathbf{x}_j^T\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\delta_i}\right) - N_{ij}\ln \\ \left(1 + \exp\left(\mathbf{x}_j^T\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\delta_i}\right)\right)\right)\end{array}\right). \tag{2}$$

### 3.1 | The detection scheme based on group-type variable selection

As is known, an outlier is defined as an observation that deviates so significantly from others that it may be generated from a changed process. [29] Besides, we usually assume that only a small percentage of profiles are outliers (ie, the so-called sparsity characteristic) in Phase I study. [20] Consequently, the problem of detecting outliers here is similar to a variable selection issue. It is noteworthy that the variable selection here is not a typical one, because we want to select nonzero subvectors instead of nonzero components of $\boldsymbol{\delta}$. That is, we need to select

variable at the factor level (ie, subvector level). Naturally, we can get an original estimator $\widehat{\boldsymbol{\delta}}^{(o)}$ by maximizing $\ell(\boldsymbol{\delta})$ via the procedure in Appendix. However, most subvectors of $\widehat{\boldsymbol{\delta}}^{(o)}$ remain nonzero, even though some of them are close to $\mathbf{0}$, which makes this estimator improper.

The BIC proposed by Schwarz,[30] which is shown especially effective in selecting the true sparsity model, can be applied here. However, if $m$ is too large, it is unpractical to calculate the BIC value for each possible $\Omega$ (see Zou and Qiu[22] and the references therein). Therefore, we utilize the following penalized loss function to reduce the complexity of calculation,

$$PL(\boldsymbol{\delta}) = -\ell(\boldsymbol{\delta}) + P_\lambda,$$

where $P_\lambda$ is the penalty function with a tuning parameter $\lambda$.

To select variable at the factor level, we apply the group-type penalty that can encourage sparsity at the factor level, motivated by Yuan and Lin[25] and Meier et al.[26] Then, the penalized loss function becomes

$$PL(\boldsymbol{\delta}) = -\ell(\boldsymbol{\delta}) + \lambda \sum_{i=1}^{m} \|\boldsymbol{\delta_i}\|_{\boldsymbol{K_i}},$$

where $\|\boldsymbol{\delta_i}\|_{\boldsymbol{K_i}} = \left(\boldsymbol{\delta_i}^T \boldsymbol{K_i} \boldsymbol{\delta_i}\right)^{1/2}$ and the kernel matrix $\boldsymbol{K_i} = p\boldsymbol{I}_p$.

Given a specified value of $\lambda$, we need the minimizer

$$\widehat{\boldsymbol{\delta}}_\lambda = \arg\min_{\delta} PL(\boldsymbol{\delta}).$$

To obtain $\widehat{\boldsymbol{\delta}}_\lambda$, we get the following score function,

$$\frac{\partial PL(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = -\boldsymbol{X}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}) + \boldsymbol{\Lambda}\boldsymbol{\delta} = \mathbf{0}, \tag{3}$$

where $\boldsymbol{X}^* = \text{diag}\{\boldsymbol{X}, \cdots, \boldsymbol{X}\}$ is a block diagonal matrix with $m$ identical blocks and $\boldsymbol{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$; $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_m^T\right)^T$ and $\boldsymbol{y_i} = (y_{i1}, \cdots, y_{in})^T$; $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^T, \cdots, \boldsymbol{\mu}_m^T\right)^T$ and $\boldsymbol{\mu_i} = (N_{i1}\pi_{i1}, \cdots, N_{in}\pi_{in})^T$; $\boldsymbol{\Lambda}$ is an $mp \times mp$ diagonal matrix that equals

$$\boldsymbol{\Lambda} = \lambda\sqrt{p} \cdot \text{diag}\left\{\frac{1}{\|\boldsymbol{\delta_1}\|}\boldsymbol{I}_p, \cdots, \frac{1}{\|\boldsymbol{\delta_m}\|}\boldsymbol{I}_p\right\}.$$

Based on Equation (3), we have the IWLS equation as follows,

$$\left(\boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{X}^* + \boldsymbol{\Lambda}\right)\boldsymbol{\delta} = \boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{q}. \tag{4}$$

where $\widehat{\boldsymbol{W}} = \text{diag}\left\{\widehat{\boldsymbol{W}}_1, \cdots, \widehat{\boldsymbol{W}}_m\right\}$, $\widehat{\boldsymbol{W}}_i = \text{diag}\{N_{i1}\widehat{\pi}_{i1}(1-\widehat{\pi}_{i1}), \cdots, N_{in}\widehat{\pi}_{in}(1-\widehat{\pi}_{in})\}$ and $\boldsymbol{q}$ denotes the adjusted dependent variate,

$$\boldsymbol{q} = \text{logit}(\boldsymbol{\pi}) - \boldsymbol{X}^*\widehat{\boldsymbol{\beta}}_0^* + \widehat{\boldsymbol{W}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{q} = \left(\boldsymbol{q}_1^T, \cdots, \boldsymbol{q}_m^T\right)^T$, $\boldsymbol{q_i} = (q_{i1}, \cdots, q_{in})^T$, $\boldsymbol{\pi} = \left(\boldsymbol{\pi}_1^T, \cdots, \boldsymbol{\pi}_m^T\right)^T$, $\boldsymbol{\pi_i} = (\pi_{i1}, \cdots, \pi_{in})^T$; $\widehat{\boldsymbol{\beta}}_0^* = \left(\widehat{\boldsymbol{\beta}}_0^T, \cdots, \widehat{\boldsymbol{\beta}}_0^T\right)^T$ is an $m \times p$-dimensional vector with each subvector being $\widehat{\boldsymbol{\beta}}_0$.

We can obtain $\widehat{\boldsymbol{\delta}}_\lambda$ by solving Equation (4) and then calculate the BIC value with $\widehat{\boldsymbol{\delta}}_\lambda$. Based on previous study (Tibshirani,[31] Yuan and Lin,[25] and Meier et al[26]), the value of BIC can be obtained by the following equation,

$$\text{BIC}_\lambda = -\ell\left(\widehat{\boldsymbol{\delta}}_\lambda\right) + \frac{1}{2}d_\lambda \ln(n), \tag{5}$$

where $d_\lambda$ denotes the degree of the model,

$$d_\lambda = \sum_i I\left(\left\|\widehat{\boldsymbol{\delta}}_{\lambda \boldsymbol{i}}\right\| > 0\right) + \sum_i \frac{\left\|\widehat{\boldsymbol{\delta}}_{\lambda \boldsymbol{i}}\right\|}{\left\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}^{(o)}\right\|}(p-1),$$

where $\widehat{\boldsymbol{\delta}}_{\lambda \boldsymbol{i}}$ and $\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}^{(\mathbf{o})}$ are the ith subvectors of $\widehat{\boldsymbol{\delta}}_\lambda$ and $\widehat{\boldsymbol{\delta}}^{(\mathbf{o})}$, respectively.

Then, set a series of values for $\lambda$ appropriately and utilize BIC to determine the optimal value defined as follows,

$$\lambda^{(*)} = \arg\min_{\lambda} \text{BIC}_\lambda.$$

Finally, the estimation $\widehat{\boldsymbol{\delta}}_{\lambda^{(*)}}$ with sparsity property can be a measurement of the outliers in the dataset and the set of outliers is $\widehat{\Omega}^* = \left\{i : \widehat{\boldsymbol{\delta}}_{\lambda^{(*)}\boldsymbol{i}} \neq \mathbf{0}\right\}$.

The above steps are further integrated into the following procedure that is called Group-type Penalized Outlier Detection (GPOD).

Step 1. Given an off-line dataset, obtain the maximized likelihood estimator for each profile, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ ($i$ = 1, ⋯, $m$). Let $\widehat{\boldsymbol{\beta}}_0$ be the median of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$.

Step 2. Define the high-dimensional shift vector $\boldsymbol{\delta}$ and obtain its original estimation $\widehat{\boldsymbol{\delta}}^{(o)}$ based on the procedure in appendix.

Step 3. Choose an appropriate range and set $T$ levels for the tuning parameter $\lambda$. For each $\lambda^{(t)}$ ($t$ = 1, ⋯, $T$), obtain the $\widehat{\boldsymbol{\delta}}_{\lambda^{(t)}}$ by using the following IWLS procedure:

(a) let $\widehat{\boldsymbol{\delta}}^{(o)}$ be the starting point $\widehat{\boldsymbol{\delta}}^{(0)} = \widehat{\boldsymbol{\delta}}^{(o)}$;

(b) at the $l$th iteration ($l \geq 0$), calculate the $\boldsymbol{q}^{(l)}$, $\widehat{\boldsymbol{W}}^{(l)}$, $\boldsymbol{\Lambda}^{(l)}$ based on $\widehat{\boldsymbol{\delta}}^{(l)}$, and then update the estimation of $\boldsymbol{\delta}$ via the following equation,

$$\widehat{\boldsymbol{\delta}}^{(l+1)} = \left(\boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}^{(l)}\boldsymbol{X}^* + \boldsymbol{\Lambda}^{(l)}\right)^{-1}\boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}^{(l)}\boldsymbol{q}^{(l)}.$$

(c) repeat step (b) until the following convergence criterion is met,

$$\left\|\widehat{\boldsymbol{\delta}}^{(l)}-\widehat{\boldsymbol{\delta}}^{(l-1)}\right\|_1 / \left\|\widehat{\boldsymbol{\delta}}^{(l)}\right\|_1 \leq \varepsilon,$$

where $\widehat{\boldsymbol{\delta}}^{(l)}$ is the estimation of $\boldsymbol{\delta}$ at the $l$th iteration. $\varepsilon$ is a small positive constant (eg, $\varepsilon = 10^{-4}$). $\|\mathbf{v}\|_1$ denotes the sum of absolute values of all components of $\mathbf{v}$.

Step 4. Use BIC to determine $\lambda^{(*)}$.

Step 5. Based on the final estimation $\widehat{\boldsymbol{\delta}}_{\lambda^{(*)}}$, obtain the set of outliers $\widehat{\Omega}^{*} = \left\{i : \widehat{\boldsymbol{\delta}}_{\lambda^{(*)}\boldsymbol{i}} \neq \mathbf{0}\right\}$.

## 3.2 | The detection scheme based on directional information

In this section, we consider the directional information assuming that the shift of each parameter vector only occurs in a presupposed component $\beta_{ik}$ ($k \in \{0, 1, ..., (p-1)\}$) and shifts in other components are zero. To simplify the discussion, we just present the directional scheme which assumes the shift only occurs in $\beta_{i0}$. Moreover, this directional scheme can be modified without any extra difficulties when the assumption is the shift occurs in any other $\beta_{ik}$. To be distinguishable from the GPOD, the shift vector here is denoted by

$$\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}} = \left(\underset{\sim}{\delta}_{01}, \cdots, \underset{\sim}{\delta}_{0m}\right)^T.$$

where $\underset{\sim}{\delta}_{0i} = \beta_{i0} - \widehat{\beta}_{00}$.

Due to the changing form of the shift vector, $\pi_{ij}$ can be expressed as follows,

$$\pi_{ij} = \frac{\exp\left(\underset{\sim}{\delta}_{0i} x_{j0} + \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right)}{1 + \exp\left(\underset{\sim}{\delta}_{0i} x_{j0} + \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right)}.$$

Consequently, we have the following joint log-likelihood function,

$$\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\begin{array}{c} \ln\binom{N_{ij}}{y_{ij}} + y_{ij}\left(\underset{\sim}{\delta}_{0i} x_{j0} + \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right) - N_{ij} \ln \\ \left(1 + \exp\left(\underset{\sim}{\delta}_{0i} x_{j0} + \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right)\right) \end{array}\right). \tag{6}$$

Define the maximizer of $\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right)$ as the original estimator $\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}}^{(o)}$, which can be obtained via the procedure in Appendix. Similarly, the penalized technique is combined with Equation (6) leading to the following penalized loss function,

$$PL\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) = -\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) + \sum_{i=1}^{m} P_{\lambda i}\left(\left|\underset{\sim}{\delta}_{0i}\right|\right).$$

The penalty function applied here is supposed to encourage sparsity at the component level. Moreover, adaptive weights are utilized for penalizing different coefficients in the $\ell_1$ penalty (see the adaptive lasso in Zou[32]). The penalized loss function becomes the following form,

$$PL\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) = -\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) + \lambda \sum_{i=1}^{m} \widehat{w}_i \left|\underset{\sim}{\delta}_{0i}\right|,$$

where $\widehat{w}_i$ is the ith component of $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \cdots, \widehat{w}_m)^T$ referred to as weight vector. Based on suggestions in Zou,[32] $\widehat{\boldsymbol{w}}$ should be data dependent, and here is defined as follows,

$$\widehat{\boldsymbol{w}} = \frac{1}{\left|\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}}^{(o)}\right|}.$$

Thus, the penalized loss function can be further expressed as follows,

$$PL\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) = -\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) + \lambda \sum_{i=1}^{m} \frac{\left|\underset{\sim}{\delta}_{0i}\right|}{\left|\widehat{\underset{\sim}{\delta}}_{0i}^{(o)}\right|}. \tag{7}$$

Given a specified value of $\lambda$, to obtain the estimator $\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}\lambda}$, the above penalized loss function needs to be differentiated. However, the penalty term in Equation (7) is singular at the origin. Based on Fan and Li,[33] we replace the penalty function by its local quadratic approximation. Define $\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}(\mathbf{0})}$ as the current estimation of $\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}$. If $\underset{\sim}{\delta}_{0(0)i}$ is quite close to 0, then set $\underset{\sim}{\delta}_{0i}$ as 0 exactly; otherwise, the penalty function can be locally approximated by the following quadratic approximation,

$$P_{\lambda i}\left(\left|\underset{\sim}{\delta}_{0i}\right|\right) \approx P_{\lambda i}\left(\left|\underset{\sim}{\delta}_{0(0)i}\right|\right) + \frac{1}{2}\frac{P'_{\lambda} i\left(\left|\underset{\sim}{\delta}_{0(0)i}\right|\right)}{\left|\underset{\sim}{\delta}_{0(0)i}\right|}\left(\underset{\sim}{\delta}_{0i}^2 - \underset{\sim}{\delta}_{0(0)i}^2\right).$$

Therefore, Equation (7) can be approximated as follows,

$$PL\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) = -\ell\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right) + \frac{1}{2}\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}^T \boldsymbol{\Theta} \underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}} + C, \tag{8}$$

where $C$ is a constant corresponding to $\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}(\mathbf{0})}$ and $\boldsymbol{\Theta}$ is an $m \times m$ diagonal matrix defined as follows,

$$\boldsymbol{\Theta} = \lambda \cdot \operatorname{diag}\left\{\frac{1}{\left|\widehat{\underset{\sim}{\delta}}_{01}^{(o)}\right| \cdot \left|\underset{\sim}{\delta}_{0(0)1}\right|}, \cdots, \frac{1}{\left|\widehat{\underset{\sim}{\delta}}_{0m}^{(o)}\right| \cdot \left|\underset{\sim}{\delta}_{0(0)m}\right|}\right\}.$$

Differentiate Equation (8) with respect to $\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}$ and the score function is as follows,

$$\frac{\partial PL\left(\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}\right)}{\partial \underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}}} = -\boldsymbol{I}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}) + \boldsymbol{\Theta}\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}} = \mathbf{0},$$

where $\boldsymbol{I}^{*} = \mathrm{diag}\{\widetilde{\mathbf{1}}, \cdots, \widetilde{\mathbf{1}}\}$ is a block diagonal matrix with $m$ identical blocks and $\widetilde{\mathbf{1}} = (1, \cdots, 1)^T$ is an $n$-dimensional vector; $\boldsymbol{y}$ and $\boldsymbol{\mu}$ have the same definitions as that in Equation (3).

Based on the score function, the following IWLS equation can be used to obtain $\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}\lambda}$ for a given $\lambda$,

$$\left(\boldsymbol{I}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{I}^{*} + \boldsymbol{\Theta}\right)\underset{\sim}{\boldsymbol{\delta}}_{\mathbf{0}} = \boldsymbol{I}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{q}.$$

where $\widehat{\boldsymbol{W}}$ and $\boldsymbol{q}$ have the same definitions as that in Equation (4).

After getting $\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}\lambda}$, the BIC value can be calculated based on Equation (5). It is worth noting that the definition of $d_\lambda$ here has been different from that in Equation (5), although the equation form of BIC is the same. According to Zou et al,[34] $d_\lambda$ denotes the number of nonzero components of $\widehat{\underset{\sim}{\boldsymbol{\delta}}}_{\mathbf{0}\lambda}$ here.

The above steps can be integrated into one procedure called Directional Penalized Outlier Detection (DPOD). The ways of calculating the relative parameters are different between the DPOD and the GPOD. Despite that, the notations and implementation steps of the two procedures are quite similar. Therefore, the steps of implementing the DPOD are omitted intentionally here.

### 3.3 | Extensions

The schemes in subsections 3.1 and 3.2 are proposed to detect outliers for profiles with binary response data. Additionally, the proposed schemes are actually not limited to binary profiles and can be extended to generalized linear profiles modeled by generalized linear model. The unified form of the link function is as follows,

$$g\left(\mu_{ij}\right) = \mathbf{x}_j^T\boldsymbol{\beta}_{\boldsymbol{i}}, i = 1, \cdots, m; j = 1, \cdots n, \tag{9}$$

where $g(\cdot)$ is called canonical link function, $\mu_{ij}$=E($y_{ij}$), and $y_{ij}$follows the exponential family of distributions with a canonical form. Specifically,

$$f\left(y_{ij}, \theta_{ij}\right) = \exp\left(y_{ij}b(\theta_{ij}) - \gamma(\theta_{ij}) + d\left(y_{ij}\right)\right), \tag{10}$$

where $b(\cdot)$, $\gamma(\cdot)$ and $d(\cdot)$ are known functions, $\theta_{ij}^{'}s$ are the canonical parameters of the exponential family of distributions, and $\mu_{ij} = \gamma^{'}(\theta_{ij})/b^{'}(\theta_{ij})$. Based on the function (9) and (10), the log-likelihood function for the ith profile could be obtained, and then similarly, the penalized loss function and the corresponding detecting schemes can be written and derived according to the steps in 3.1 and 3.2.

For example, if $y_{ij}$ follows Poisson distribution, the link function becomes

$$\ln\theta_{ij} = \mathbf{x}_j^T\boldsymbol{\beta}_{\boldsymbol{i}}, i = 1, \cdots, m; j = 1, \cdots n.$$

and $b(\theta_{ij}) = \ln\theta_{ij}$, $\gamma(\theta_{ij}) = \theta_{ij}$, $d(y_{ij}) = -\ln(y_{ij}!)$. Therefore, the log-likelihood function for the ith profile is the following

$$l_i = \sum_{j=1}^{n} y_{ij}\ln\theta_{ij} - \theta_{ij} - \ln\left(y_{ij}!\right)$$

Recall the associated notation in the previous subsections, the log-likelihood function for m profiles is written as

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij}\mathbf{x}_j^T\left(\boldsymbol{\delta}_{\boldsymbol{i}} + \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right) - \exp\left(\mathbf{x}_j^T\left(\boldsymbol{\delta}_{\boldsymbol{i}} + \widehat{\boldsymbol{\beta}}_{\mathbf{0}}\right)\right) - \ln\left(y_{ij}!\right)\right).$$

Except for the log-likelihood function, the penalty function and the estimation steps of $\boldsymbol{\delta}$ are same as that of binary profiles, and the corresponding detecting schemes for profiles with Poisson response data can be similarly obtained based on the procedures in previous subsections.

## 4 | SIMULATION STUDY

In this section, we study the detecting performance for binary profiles of the proposed GPOD and DPOD schemes in various scenarios via Monte Carlo simulations. As we mentioned in Section 1, limited literature specialized in outlier detection for binary profiles. Yeh et al[16] first studied such a problem and proposed five $T^2$ control charts, among which the $T_I^2$ chart performs best in detecting outliers. Besides, Shadman et al[19] proposed a change-point (CP) approach chart, but their simulation results show that the CP approach chart is less effective than the $T_I^2$ chart in detecting the presence of outlying observations. Therefore, the $T_I^2$ chart will be used as the benchmark in the following simulation study. Without loss of generality, we assume $p = 2$ and the in-control parameter vector $\boldsymbol{\beta}_{\mathbf{0}} = (1, -1)^T$. The number of experiment settings of each profile is set to be $n = 20$ ranging from 1 to 20. Note that $x$ is usually replaced by ln($x$) in actual modeling. The design matrix $\boldsymbol{X}$ is set as follows,

$$\boldsymbol{X}=\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \ln 1 & \ln 2 & \cdots & \ln 20 \end{pmatrix}^{T}.$$

The parameter in binomial distribution corresponding to the jth setting of the ith profile is assumed to be fixed $N_{ij} = 100$, and the generated dataset consists of $m = 20$ binary profiles. Among them, $m_o$ outlying profiles, $m_o = 1$, 2, 3, 4, 5, and 6, respectively, are considered in different scenarios. To apply the $T_I^2$ chart, the upper control limit (UCL) should be obtained based on large quantities of simulations, and here the UCL is approximated by simulated data generated based on the parameter estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$. Given a profile dataset, one needs to calculate the value of $T_I^2$ for each profile. A profile will be viewed as an outlier if its $T_I^2$ value exceeds the UCL. As for the GPOD and the DPOD, one needs to implement the schemes as described in subsections 3.1 and 3.2 respectively, and then the outliers can be indexed by $\widehat{\Omega}^*$ directly.

Our simulation results show that the two proposed schemes generally have satisfactory performance in terms of the signal probability. However, we do not recommend it for assessing the performance in identifying outliers because it can only reflect limited information about the outliers. In a typical outlier detection issue, evaluating the masking effect and the swamping effect is usually of great significance.[4] Inspired by Shang et al[24] and Zou et al,[35] we adopt the probability that a scheme succeeds in identifying all the outliers correctly, which is denoted by "Cf." Besides, "Uf" presents the probability that a scheme only identifies some (at least one) of the outliers correctly but fails to identify all of them; "Of" presents the probability that a scheme detects all the outliers correctly and meanwhile identifies some normal profiles as outliers; "Rf" presents the probability of the remaining cases. Apparently, "Uf" and "Of" can, respectively, reflect the information of masking and swamping effects.

Three scenarios are considered in the simulation study: (1) various shifts in $\beta_{i0}$; (2) various shifts in $\beta_{i1}$; and (3) various shifts in $\beta_{i0}$ and $\beta_{i1}$ simultaneously. The simulation results of the three scenarios are tabulated in Tables 1–3, respectively.

According to "Cf" in Table 1, we can find out that these three schemes usually performs quite well when the ratio of outliers is small ($m_o/m \leq 0.1$), and their performance degrades as the ratio of outliers increases. However, the GPOD performs better than the $T_I^2$ chart when the ratio becomes large ($m_o/m > 0.1$). Besides, The DPOD outperforms the $T_I^2$ chart as the number of outlying profiles $m_o$ is large, especially for the case when there are small shifts in $\beta_{i0}$. And when the shift gets larger, the performance of these three schemes becomes better.

Based on the results in columns "Uf" and "Of" of Table 1, we can see that the $T_I^2$ chart tends to detect fewer true outliers than the GPOD and DPOD schemes because of the masking effect, which means outliers are masked and undetected. In comparison, the DPOD is hardly prone to masking, and the $T_I^2$ chart has more tendency than the GPOD in terms of masking. That is to say, our

**TABLE 1** Performance comparison among the $T_I^2$ chart, GPOD and DPOD with various shifts in $\boldsymbol{\beta_{i0}}$ for $\boldsymbol{m = 20}$

| | | Cf | | | Uf | | | Of | | | Rf | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_\Omega{}^T$ | $m_o$ | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD |
| (0.2, 0) | 1 | 0.58 | 0.63 | 0.38 | - | - | - | 0 | 0 | 0 | 0.42 | 0.37 | 0.62 |
| | 2 | 0.26 | 0.41 | 0.23 | 0.47 | 0.36 | 0.08 | 0 | 0 | 0 | 0.27 | 0.23 | 0.69 |
| | 3 | 0.08 | 0.24 | 0.16 | 0.69 | 0.54 | 0.08 | 0 | 0 | 0 | 0.23 | 0.22 | 0.76 |
| | 4 | 0.02 | 0.14 | 0.09 | 0.73 | 0.61 | 0.07 | 0 | 0 | 0 | 0.25 | 0.25 | 0.84 |
| | 5 | 0 | 0.06 | 0.06 | 0.70 | 0.63 | 0.06 | 0 | 0 | 0 | 0.30 | 0.31 | 0.88 |
| | 6 | 0 | 0.03 | 0.03 | 0.61 | 0.60 | 0.05 | 0 | 0 | 0 | 0.39 | 0.37 | 0.92 |
| (0.3, 0) | 1 | 0.92 | 0.87 | 0.60 | - | - | - | 0.06 | 0.12 | 0.38 | 0.02 | 0.01 | 0.02 |
| | 2 | 0.84 | 0.83 | 0.44 | 0.06 | 0.03 | 0 | 0.10 | 0.13 | 0.56 | 0 | 0.01 | 0 |
| | 3 | 0.68 | 0.76 | 0.29 | 0.15 | 0.05 | 0 | 0.15 | 0.15 | 0.70 | 0.02 | 0.04 | 0.01 |
| | 4 | 0.45 | 0.64 | 0.20 | 0.26 | 0.09 | 0 | 0.21 | 0.19 | 0.78 | 0.08 | 0.08 | 0.02 |
| | 5 | 0.20 | 0.52 | 0.13 | 0.34 | 0.12 | 0 | 0.21 | 0.19 | 0.85 | 0.25 | 0.17 | 0.02 |
| | 6 | 0.04 | 0.36 | 0.07 | 0.31 | 0.15 | 0 | 0.13 | 0.19 | 0.87 | 0.52 | 0.30 | 0.06 |
| (0.4, 0) | 1 | 0.92 | 0.89 | 0.69 | - | - | - | 0.08 | 0.11 | 0.31 | 0 | 0 | 0 |
| | 2 | 0.85 | 0.85 | 0.55 | 0 | 0 | 0 | 0.15 | 0.15 | 0.45 | 0 | 0 | 0 |
| | 3 | 0.68 | 0.78 | 0.43 | 0 | 0 | 0 | 0.32 | 0.22 | 0.57 | 0 | 0 | 0 |
| | 4 | 0.44 | 0.70 | 0.30 | 0.01 | 0 | 0 | 0.55 | 0.29 | 0.70 | 0 | 0.01 | 0 |
| | 5 | 0.18 | 0.60 | 0.18 | 0.01 | 0 | 0 | 0.78 | 0.39 | 0.82 | 0.03 | 0.01 | 0 |
| | 6 | 0.03 | 0.47 | 0.10 | 0.01 | 0 | 0 | 0.87 | 0.50 | 0.90 | 0.09 | 0.03 | 0 |

**TABLE 2** Performance comparison among the $T_I^2$ chart, GPOD, and DPOD with various shifts in $\beta_{i1}$ for $m = 20$

| | | Cf | | | Uf | | | Of | | | Rf | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_\Omega{}^T$ | $m_o$ | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD |
| (0, 0.1) | 1 | 0.63 | 0.74 | 0.37 | - | - | - | 0.04 | 0.09 | 0.37 | 0.33 | 0.17 | 0.26 |
| | 2 | 0.33 | 0.56 | 0.24 | 0.46 | 0.26 | 0.08 | 0.03 | 0.07 | 0.50 | 0.18 | 0.11 | 0.18 |
| | 3 | 0.13 | 0.39 | 0.16 | 0.67 | 0.42 | 0.09 | 0.02 | 0.05 | 0.53 | 0.18 | 0.14 | 0.22 |
| | 4 | 0.03 | 0.26 | 0.09 | 0.75 | 0.51 | 0.08 | 0.01 | 0.03 | 0.54 | 0.21 | 0.20 | 0.29 |
| | 5 | 0 | 0.14 | 0.06 | 0.70 | 0.55 | 0.07 | 0 | 0.01 | 0.46 | 0.30 | 0.30 | 0.41 |
| | 6 | 0 | 0.07 | 0.04 | 0.62 | 0.53 | 0.05 | 0 | 0.01 | 0.37 | 0.38 | 0.39 | 0.54 |
| (0, 0.2) | 1 | 0.91 | 0.88 | 0.71 | - | - | - | 0.09 | 0.12 | 0.29 | 0 | 0 | 0 |
| | 2 | 0.82 | 0.85 | 0.57 | 0 | 0 | 0 | 0.18 | 0.15 | 0.43 | 0 | 0 | 0 |
| | 3 | 0.63 | 0.78 | 0.44 | 0 | 0 | 0 | 0.37 | 0.22 | 0.56 | 0 | 0 | 0 |
| | 4 | 0.35 | 0.70 | 0.32 | 0 | 0 | 0 | 0.65 | 0.30 | 0.68 | 0 | 0 | 0 |
| | 5 | 0.11 | 0.58 | 0.20 | 0 | 0 | 0 | 0.89 | 0.42 | 0.80 | 0 | 0 | 0 |
| | 6 | 0.02 | 0.43 | 0.11 | 0 | 0 | 0 | 0.95 | 0.56 | 0.89 | 0.03 | 0.01 | 0 |
| (0, 0.3) | 1 | 0.87 | 0.89 | 0.77 | - | - | - | 0.13 | 0.11 | 0.23 | 0 | 0 | 0 |
| | 2 | 0.62 | 0.85 | 0.69 | 0 | 0 | 0 | 0.38 | 0.15 | 0.31 | 0 | 0 | 0 |
| | 3 | 0.20 | 0.80 | 0.60 | 0 | 0 | 0 | 0.80 | 0.20 | 0.40 | 0 | 0 | 0 |
| | 4 | 0 | 0.70 | 0.47 | 0 | 0 | 0 | 1 | 0.30 | 0.53 | 0 | 0 | 0 |
| | 5 | 0 | 0.57 | 0.34 | 0 | 0 | 0 | 1 | 0.43 | 0.66 | 0 | 0 | 0 |
| | 6 | 0 | 0.42 | 0.22 | 0 | 0 | 0 | 1 | 0.58 | 0.78 | 0 | 0 | 0 |

**TABLE 3** Performance comparison among the $T_I^2$ chart, GPOD, and DPOD with various shifts in both $\beta_{i0}$ and $\beta_{i1}$ for $m = 20$

| | | Cf | | | Uf | | | Of | | | Rf | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_\Omega{}^T$ | $m_o$ | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD | $T_I^2$ | GPOD | DPOD |
| (0.1, 0.1) | 1 | 0.93 | 0.88 | 0.59 | - | - | - | 0.06 | 0.12 | 0.39 | 0.01 | 0 | 0.02 |
| | 2 | 0.85 | 0.81 | 0.41 | 0.04 | 0.01 | 0 | 0.10 | 0.17 | 0.58 | 0.01 | 0.01 | 0.01 |
| | 3 | 0.71 | 0.71 | 0.25 | 0.10 | 0.03 | 0 | 0.17 | 0.23 | 0.74 | 0.02 | 0.03 | 0.01 |
| | 4 | 0.47 | 0.58 | 0.14 | 0.20 | 0.04 | 0 | 0.24 | 0.30 | 0.84 | 0.09 | 0.08 | 0.02 |
| | 5 | 0.21 | 0.42 | 0.08 | 0.28 | 0.06 | 0 | 0.28 | 0.33 | 0.88 | 0.23 | 0.19 | 0.04 |
| | 6 | 0.05 | 0.30 | 0.04 | 0.25 | 0.07 | 0 | 0.18 | 0.31 | 0.88 | 0.52 | 0.32 | 0.08 |
| (0.1, 0.2) | 1 | 0.91 | 0.87 | 0.73 | - | - | - | 0.09 | 0.13 | 0.27 | 0 | 0 | 0 |
| | 2 | 0.75 | 0.80 | 0.60 | 0 | 0 | 0 | 0.25 | 0.20 | 0.40 | 0 | 0 | 0 |
| | 3 | 0.43 | 0.69 | 0.42 | 0 | 0 | 0 | 0.57 | 0.31 | 0.58 | 0 | 0 | 0 |
| | 4 | 0.12 | 0.53 | 0.25 | 0 | 0 | 0 | 0.88 | 0.47 | 0.75 | 0 | 0 | 0 |
| | 5 | 0.01 | 0.37 | 0.15 | 0 | 0 | 0 | 0.99 | 0.63 | 0.85 | 0 | 0 | 0 |
| | 6 | 0 | 0.22 | 0.06 | 0 | 0 | 0 | 1 | 0.78 | 0.94 | 0 | 0 | 0 |
| (0.2, 0.2) | 1 | 0.88 | 0.87 | 0.75 | - | - | - | 0.12 | 0.13 | 0.25 | 0 | 0 | 0 |
| | 2 | 0.64 | 0.76 | 0.60 | 0 | 0 | 0 | 0.36 | 0.24 | 0.40 | 0 | 0 | 0 |
| | 3 | 0.23 | 0.60 | 0.41 | 0 | 0 | 0 | 0.77 | 0.40 | 0.59 | 0 | 0 | 0 |
| | 4 | 0.01 | 0.41 | 0.22 | 0 | 0 | 0 | 0.99 | 0.59 | 0.78 | 0 | 0 | 0 |
| | 5 | 0 | 0.22 | 0.09 | 0 | 0 | 0 | 1 | 0.78 | 0.91 | 0 | 0 | 0 |
| | 6 | 0 | 0.10 | 0.03 | 0 | 0 | 0 | 1 | 0.90 | 0.97 | 0 | 0 | 0 |

proposed penalized-type schemes can alleviate masking effects to a certain extent. This finding is similar to that in Zou et al.[4]

As described in Zou et al,[4] swamping can lead to normal profiles being considered as outliers, and in outlier detection, masking is usually worse than swamping because it can result in gross distortions. Therefore, the DPOD could be an effective method if masking effect is very serious in reality. Moreover, the GPOD is the most recommendable method in various situations due to its consistently remarkable performance in detecting true outlying profiles.

As shown in Table 2, the similar conclusions are obtained. The GPOD outperforms the $T_I^2$ chart and DPOD as there are multiple outlying profiles, and the superiority of the proposed DPOD to $T_I^2$ is more obvious when the

shift is small. It is noteworthy that the $T_I^2$ chart does not perform better as the magnitude of shift increases. We think the reason may be that the large shift can influence the accuracy of the computation of $T_I^2$ values.

Table 3 shows the simulation results when shifts are imposed on the model parameters $\beta_{i0}$ and $\beta_{i1}$ simultaneously, the finding of which is same as above. The DPOD performs best among these three schemes in terms of avoiding masking effect. Besides, the GPOD outperforms the other two methods in detecting true outliers across almost all situations.

We also investigate the detecting performance of the methods for $m = 10$ and $m = 30$. Likewise, the largest ratio of outliers is set to be $m_o/m = 0.3$. For $m = 10$, $m_o$ =1, 2, and 3, and for $m = 30$, $m_o$ takes nine values ranging from 1 to 9. For the sake of brevity, only the probability of diagnosing all the true outliers correctly is shown in Table 4 and Figure 1, but the original data is available from the authors.

According to the performance comparison in Table 4 and Figure 1, the GPOD is further proven to be superior to the $T_I^2$ and the DPOD in most cases, and the DPOD outperforms the $T_I^2$ as the number of outlying profiles is larger. Moreover, it can be found out that the $T_I^2$ remains incapable of identifying outliers correctly in some cases when the ratio of the outliers in the dataset is slightly large.

Besides, in order to further study the detection performance of the proposed methods, we consider another logistic regression model, of which the in-control parameter vector is $\boldsymbol{\beta_0} = (1, -1, 0.5)^T$ (ie, $p = 3$) and the design matrix is as follows,

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \ln 1 & \ln 2 & \cdots & \ln 20 \\ 0.1 & 0.2 & \ldots & 2 \end{pmatrix}^T.$$

The simulated results are summarized in Table 5. Only the "Cf" for $m = 10$ is presented here, because the findings for $p = 3$ are quite similar with that for $p = 2$. As shown in Table 5, the GPOD performs best in terms of "Cf" in most cases, and the DPOD also performs better than the $T_I^2$ when the ratio of outliers is large.

**TABLE 4** Performance comparison of the three methods in various scenarios for $m = 10$

| | | Cf | | | | Cf | | | | Cf | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_o$ | $\delta_\Omega{}^T$ | $T_I^2$ | GPOD | DPOD | $\delta_\Omega{}^T$ | $T_I^2$ | GPOD | DPOD | $\delta_\Omega{}^T$ | $T_I^2$ | GPOD | DPOD |
| 1 | (0.2, 0) | 0.61 | 0.60 | 0.45 | (0, 0.1) | 0.67 | 0.71 | 0.46 | (0.1, 0.1) | 0.88 | 0.88 | 0.65 |
| 2 | | 0.21 | 0.32 | 0.28 | | 0.26 | 0.45 | 0.28 | | 0.64 | 0.70 | 0.38 |
| 3 | | 0.02 | 0.13 | 0.16 | | 0.03 | 0.23 | 0.16 | | 0.22 | 0.47 | 0.17 |
| 1 | (0.3, 0) | 0.89 | 0.89 | 0.69 | (0, 0.2) | 0.83 | 0.91 | 0.76 | (0.1, 0.2) | 0.77 | 0.87 | 0.76 |
| 2 | | 0.65 | 0.77 | 0.46 | | 0.42 | 0.80 | 0.56 | | 0.16 | 0.67 | 0.50 |
| 3 | | 0.22 | 0.55 | 0.24 | | 0.04 | 0.59 | 0.34 | | 0 | 0.40 | 0.23 |
| 1 | (0.4, 0) | 0.86 | 0.90 | 0.75 | (0, 0.3) | 0.66 | 0.90 | 0.81 | (0.2, 0.2) | 0.67 | 0.85 | 0.76 |
| 2 | | 0.49 | 0.79 | 0.54 | | 0.02 | 0.78 | 0.65 | | 0.03 | 0.57 | 0.43 |
| 3 | | 0.07 | 0.60 | 0.30 | | 0 | 0.57 | 0.45 | | 0 | 0.25 | 0.15 |

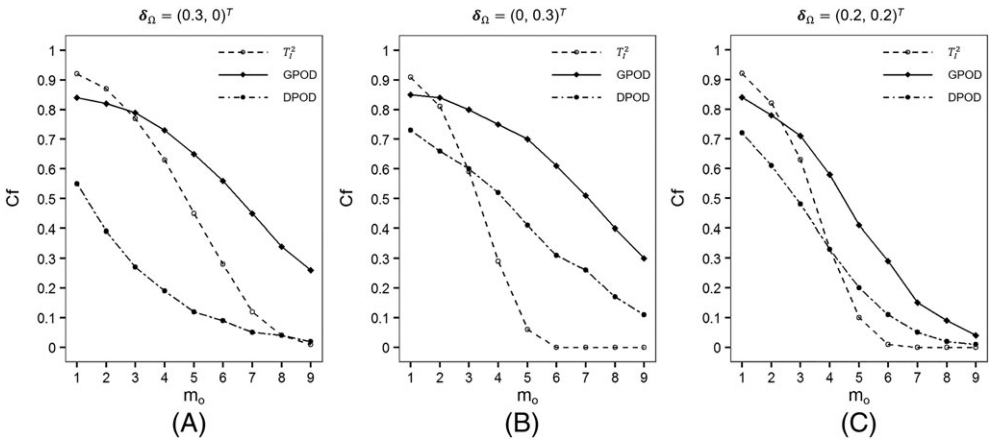


**FIGURE 1** Performance comparison among the $T_I^2$, GPOD, and DPOD in three typical scenarios for $m = 30$

**TABLE 5** "Cf" comparison among the $T_I^2$ chart, GPOD, and DPOD for $m = 10$ with various shifts in $\boldsymbol{\beta}$ when $p = 3$

| | | Cf | | | | Cf | | |
|---|---|---|---|---|---|---|---|---|
| $m_o$ | $\delta_\Omega{}^T$ | $T_I^2$ | GPOD | DPOD | $\delta_\Omega{}^T$ | $T_I^2$ | GPOD | DPOD |
| 1 | (0.2, 0, 0) | 0.68 | 0.45 | 0.37 | (0, 0.1, 0) | 0.81 | 0.67 | 0.3 |
| 2 | | 0.26 | 0.32 | 0.23 | | 0.41 | 0.29 | 0.25 |
| 3 | | 0.04 | 0.16 | 0.21 | | 0.1 | 0.21 | 0.18 |
| 1 | (0.3, 0, 0) | 0.91 | 0.75 | 0.6 | (0, 0.2, 0) | 0.84 | 0.81 | 0.55 |
| 2 | | 0.69 | 0.66 | 0.39 | | 0.34 | 0.59 | 0.34 |
| 3 | | 0.26 | 0.52 | 0.19 | | 0.02 | 0.48 | 0.19 |
| 1 | (0.4, 0, 0) | 0.87 | 0.87 | 0.58 | (0, 0.3, 0) | 0.63 | 0.81 | 0.61 |
| 2 | | 0.46 | 0.74 | 0.45 | | 0 | 0.55 | 0.33 |
| 3 | | 0.06 | 0.66 | 0.23 | | 0 | 0.32 | 0.12 |
| 1 | (0, 0, 0.1) | 0.15 | 0.19 | 0.15 | (0.1, 0.1, 0.1) | 0.85 | 0.78 | 0.48 |
| 2 | | 0.01 | 0.02 | 0.07 | | 0.4 | 0.65 | 0.28 |
| 3 | | 0 | 0 | 0.03 | | 0.02 | 0.36 | 0.15 |
| 1 | (0, 0, 0.2) | 0.82 | 0.65 | 0.32 | (0.2, 0.2, 0.1) | 0.5 | 0.62 | 0.5 |
| 2 | | 0.49 | 0.44 | 0.22 | | 0 | 0.43 | 0.27 |
| 3 | | 0.1 | 0.24 | 0.22 | | 0 | 0.18 | 0.1 |
| 1 | (0, 0, 0.3) | 0.89 | 0.79 | 0.4 | (0.2, 0.2, 0.2) | 0.34 | 0.68 | 0.48 |
| 2 | | 0.61 | 0.68 | 0.21 | | 0 | 0.36 | 0.21 |
| 3 | | 0.17 | 0.51 | 0.27 | | 0 | 0.13 | 0.07 |

To sum up, from the above tables and figure, we conclude that the GPOD scheme can be a favorable detecting method in identifying outlying profiles with binary data. To alleviate masking effect, the DPOD scheme can also be the alternative effective method.

## 5 | A REAL EXAMPLE APPLICATION

In this section, we illustrate the proposed GPOD scheme using the warranty claims data of automobiles as the example. Nowadays, most of the durable products are sold with warranty. In the warranty period, sellers or manufacturers are obligated to repair or replace the failed products for customers, and the failures result in the warranty claims. Therefore, to access the quality and reliability performance of the sold products, the warranty claims data are usually collected for analysis. In Table 6, some warranty claims data from an automobile manufacturer of China are presented.

As shown in Table 6, the column labeled "$N_i$" presents the number of automobiles that are made in $MOP(i)$ (the ith month of production) and have been sold. Besides, the number of warranty claims corresponding to the early failures of the automobiles for each MOP, $y_{ij}$, is recorded against each MIS (month in service). That is, after $MIS(j)$, $y_{ij}$ out of $N_i$ automobiles failed. So $y_{ij}$ follows the binomial distribution, $y_{ij}\sim\text{BIN}(N_i, \pi_{ij})$ where $\pi_{ij}$ is the failure rate of the automotive sample $MOP(i)$ at time point $MIS(j)$. In this way, the data in Table 6 can be viewed as a dataset consisting of 10 binary profiles, where the response variable is $y_{ij}$ and the explanatory variable is $MIS(j)$ denoted by $x_j$ here. The design matrix is as follows,

$$\boldsymbol{X}=\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 12 \end{pmatrix}^T.$$

We assume that all the above profiles are collected from an in-control process and model them by Equation (1). To demonstrate the implementation of the proposed scheme, we change the first profile sample into an outlier by imposing a shift, $\boldsymbol{\delta}_\Omega = (0.02, 0.03)^T$, on its estimated parameter vector, and the modified data is shown in Table 7.

Given the dataset that contains 10 profiles with the first profile being an outlier, the steps of detecting the outlier by the GPOD scheme are described as follows:

Step 1. Estimate the parameter vector of each profile by solving the widely used IWLS equation and denote it by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ ($i = 1, \cdots, 10$). Then, the in-control parameter of the process is assumed to be $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} = \text{median}\left\{\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_{10}\right\}$.

Step 2. Define the high-dimensional shift vector $\boldsymbol{\delta}$ and obtain its original estimator $\widehat{\boldsymbol{\delta}}^{(0)}$ based on equations in appendix.

Step 3. Set an appropriate range and grid it into $T$ levels for the penalty tuning parameter $\lambda$. For each

**TABLE 6** Warranty claims in different MIS for various MOP

| | | MIS($j$) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOP($i$) | $N_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 9451 | 26 | 43 | 66 | 84 | 100 | 115 | 126 | 137 | 158 | 170 | 178 | 203 |
| 2 | 10 167 | 32 | 56 | 73 | 92 | 110 | 126 | 149 | 168 | 186 | 191 | 210 | 248 |
| 3 | 4602 | 10 | 15 | 22 | 29 | 37 | 48 | 56 | 64 | 66 | 81 | 96 | 108 |
| 4 | 8572 | 20 | 41 | 61 | 72 | 86 | 101 | 117 | 135 | 139 | 173 | 194 | 218 |
| 5 | 9923 | 25 | 45 | 64 | 80 | 109 | 125 | 142 | 146 | 184 | 202 | 238 | 254 |
| 6 | 9431 | 29 | 47 | 70 | 89 | 112 | 128 | 139 | 148 | 179 | 205 | 246 | 266 |
| 7 | 8907 | 25 | 45 | 59 | 83 | 102 | 119 | 121 | 136 | 165 | 185 | 207 | 220 |
| 8 | 9176 | 21 | 35 | 46 | 60 | 77 | 87 | 94 | 132 | 160 | 177 | 192 | 204 |
| 9 | 8064 | 16 | 30 | 45 | 61 | 71 | 73 | 92 | 115 | 135 | 150 | 157 | 166 |
| 10 | 8667 | 24 | 39 | 53 | 60 | 72 | 80 | 142 | 182 | 209 | 222 | 225 | 236 |

**TABLE 7** The modified warranty claims data for MOP(1)

| | | MIS($j$) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOP($i$) | $N_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 9451 | 52 | 61 | 73 | 86 | 101 | 120 | 141 | 167 | 197 | 232 | 273 | 322 |

$\lambda^{(t)}$ $(t = 1, \cdots, T)$, obtain the $\widehat{\boldsymbol{\delta}}_{\lambda^{(t)}}$ and calculate the value of $\mathrm{BIC}_{\lambda^{(t)}}$ based on the procedure in Section 3.1.

Step 4. Find $\lambda^{(*)}$ corresponding to the minimizer of $\mathrm{BIC}_{\lambda^{(t)}}$, and $\widehat{\boldsymbol{\delta}}_{\lambda^{(*)}}$ will be the optimal estimation.

In this application, only the first subvector of $\widehat{\boldsymbol{\delta}}_{\lambda^{(*)}}$ is a nonzero vector, $\widehat{\boldsymbol{\delta}}_{\lambda^{(*)}1} = (0.200, 0.002)^T$. Then, we can conclude that the first profile is the only outlier among the 10 profiles. That is, the GPOD scheme identifies the outliers correctly.

## 6 | CONCLUDING REMARKS

Detecting outliers for profiles with binary response still remains a challenging problem and attracts insufficient attention in the literature. In this article, we propose two diagnosing schemes under the framework of penalized likelihood, based on the group LASSO method and directional information, respectively, to detect outliers among profiles with binary data. As shown in the simulations, the proposed GPOD scheme can detect the true outliers with large probabilities and usually outperforms the existing $T_I^2$ chart especially when the ratio of outliers in the dataset is slightly large. Moreover, the GPOD scheme suffers from less masking effects than the $T_I^2$ chart. The DPOD scheme, however, tends to identify excess outliers in many cases, which reduces its ability to detect outliers correctly but meanwhile alleviates masking effects to quite a large extent. Therefore, if the masking effect is more serious, the DPOD can be the alternative of the GPOD. Furthermore, not requiring large quantities of simulations to obtain the UCL, the proposed penalized-type schemes are less computationally expensive than the $T_I^2$ chart in real-world applications.

In this paper, the outliers are supposed to follow the linear model structure. It is worthwhile to extend our schemes to the cases when the model structure of outliers is nonlinear. In some applications, the quality of a process or a product is better characterized by two or more correlated profiles, which is referred to as multivariate profiles. How to develop penalized schemes for such profiles can be an interesting future research topic. Both schemes we proposed in this paper are based on BIC. That is, the optimal tuning parameter is purely determined by minimizing the BIC value, leading to the fact that the Type-I error cannot be controlled effectively. The outlier detection schemes based on controlling the Type-I error can be studied in the future research.

## ACKNOWLEDGEMENT

The authors would like to thank the Editor and two anonymous referees for their many helpful comments

that have resulted in significant improvements in the article. This research was supported by the National Natural Science Foundation of China (Nos. 71672122, 71532008, 71472132, 71402118 and 71401123).

## REFERENCES

1. Mahmoud MA, Woodall WH. Phase I analysis of linear profiles with calibration applications. *Dent Tech*. 2004;46:380-391.
2. Mahmoud MA, Parker PA, Woodall WH, Hawkins DM. A change point method for linear profile data. *Qual Reliab Eng Int*. 2007;23(2):247-268.
3. Zhang H, Albin S. Detecting outliers in complex profiles using a $\chi^2$ control chart method. *IIE Trans*. 2009;41:335-345.
4. Zou C, Tseng ST, Wang Z. Outlier detection in general profiles using penalized regression method. *IIE Trans*. 2014; 46(2):106-117.
5. Kang L, Albin S. On-line monitoring when the process yields a linear. *J Qual Technol*. 2000;32(4):418-426.
6. Kim K, Mahmoud MA, Woodall WH. On the monitoring of linear profiles. *J Qual Technol*. 2003;35(3):317-328.
7. Zou C, Zhang Y, Wang Z. A control chart based on a change-point model for monitoring linear profiles. *IIE Trans*. 2006;38(12):1093-1103.
8. Noorossana R, Eyvazian M, Amiri A, Mahmoud MA. Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application. *Qual Reliab Eng Int*. 2010;26(3):291-303.
9. Mestek O, Pavlík J, Suchánek M. Multivariate control charts: control charts for calibration curves. *Fresenius J Anal Chem*. 1994;350(6):344-351.
10. Yeh A, Zerehsaz Y. Phase I control of simple linear profiles with individual observations. *Qual Reliab Eng Int*. 2013;29(6):829-840.
11. Mahmoud MA. Phase I analysis of multiple linear regression profiles. *Commun Stat Simul Comput*. 2008;37(10):2106-2130.
12. Williams JD, Woodall WH, Birch JB. Statistical monitoring of nonlinear product and process quality profiles. *Qual Reliab Eng Int*. 2007;23(8):925-941.
13. Paynabar K, Jin J. Characterization of non-linear profiles variations using mixed-effect models and wavelets. *IIE Trans*. 2011;43(4):275-290.
14. Dai C, Wang K, Jin R. Monitoring profile trajectories with dynamic time warping alignment. *Qual Reliab Eng Int*. 2014; 30(6):815-827.
15. Paynabar K, Zou C, Qiu P. A change-point approach for phase-I analysis in multivariate profile monitoring and diagnosis. *Dent Tech*. 2016;58:191-204.
16. Yeh AB, Huwang L, Li YM. Profile monitoring for a binary response. *IIE Trans*. 2009;41(11):931-941.
17. Shang Y, Tsung F, Zou C. Profile monitoring with binary data and random predictors. *J Qual Technol*. 2011;43(3):196-208.
18. Paynabar K, Jin J, Yeh AB. Phase I risk-adjusted control charts for monitoring surgical performance by considering categorical covariates. *J Qual Technol*. 2012;44(1):39-53.
19. Shadman A, Mahlooji H, Yeh AB, Zou C. A change point method for monitoring generalized linear profiles in phase I. *Qual Reliab Eng Int*. 2015;31(8):1367-1381.
20. Qiu P, Zou C, Wang Z. Nonparametric profile monitoring by mixed effects modeling. *Dent Tech*. 2010;52:265-277.
21. Hodge V, Austin J. A survey of outlier detection methodologies. *Artif Intell Rev*. 2004;22(2):85-126.
22. Jobe JM, Pokojovy M. A cluster-based outlier detection scheme for multivariate data. *J Am Stat Assoc*. 2015;110(512):1543-1551.
23. Zou C, Qiu P. Multivariate statistical process control using LASSO. *J Am Stat Assoc*. 2009;104(488):1586-1596.
24. Shang Y, Zi X, Tsung F, He Z. LASSO-based diagnosis scheme for multistage processes with binary data. *Comput Ind Eng*. 2014;72:198-205.
25. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodology*. 2006;68(1):49-67.
26. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J R Stat Soc Series B Stat Methodology*. 2008;70(1): 53-71.
27. Shang Y, Man J, He Z, Ren H. Change-point detection in phase I for profiles with binary data and random predictors. *Qual Reliab Eng Int*. 2016;32(7):2549-2558.
28. McCullagh P, Nelder JA. *Generalized Linear Models*. 37 CRC Press; 1989.
29. Hawkins DM. *Identification of Outliers*. 11 London: Chapman and Hall; 1980.
30. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
31. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol*. 1996;58:267-288.
32. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
33. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001; 96(456):1348-1360.
34. Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the lasso. *Ann Stat*. 2007;35(5):2173-2192.
35. Zou C, Ning X, Tsung F. LASSO-based multivariate linear profile monitoring. *Ann Oper Res*. 2012;192(1):3-19.

**Zhen Li** is a graduate student of the College of Management and Economics at Tianjin University, China.

**Yanfen Shang** is an associate professor of the College of Management and Economics at Tianjin University, China. She received her BS and MS degrees from Tianjin University, Tianjin, People's Republic of China, and PhD degree from Hong Kong University of Science and Technology (HKUST), Hong Kong. Her research interests include quality management and statistical process control.

**Zhen He** is a professor in the College of Management and Economics, Tianjin University. He received his PhD in Management Science and Engineering from Tianjin University, China, in 2001. He is the recipient of Outstanding Research Young Scholar Award of the National Natural Science Foundation of China. His research interests focus on statistical quality control, DOE, and Six Sigma management.

**How to cite this article:** Li Z, Shang Y, He Z. Phase I outlier detection in profiles with binary data based on penalized likelihood. *Qual Reliab Engng Int*. 2019;35:1–13. https://doi.org/10.1002/qre.2376

## APPENDIX: OBTAIN THE ORIGINAL ESTIMATORS OF THE SCHEMES

The original estimator $\widehat{\boldsymbol{\delta}}^{(o)}$ can be obtained by solving the IWLS equation. Take the derivative of Equation (2) with respect to $\boldsymbol{\delta}$ and obtain the following score function,

$$\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \boldsymbol{X}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{A1}$$

where $\boldsymbol{X}^{*} = \text{diag}\{\boldsymbol{X},\cdots,\boldsymbol{X}\}$ is a block diagonal matrix with $m$ identical blocks and each block is $\boldsymbol{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)^T$, $\boldsymbol{y}=\left(\boldsymbol{y}_1^T,\ \cdots,\ \boldsymbol{y}_m^T\right)^T$ and $\boldsymbol{y}_i = (y_{i1},\cdots,y_{in})^T$, $\boldsymbol{\mu}=\left(\boldsymbol{\mu}_1^T,\ \cdots,\ \boldsymbol{\mu}_m^T\right)^T$ and $\boldsymbol{\mu}_i = (N_{i1}\pi_{i1},\cdots,N_{in}\pi_{in})^T$.

Assuming $\boldsymbol{X}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}) = \mathbf{0}$, we have the following IWLS equation,

$$\boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{X}^{*}\boldsymbol{\delta} = \boldsymbol{X}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{q}, \tag{A2}$$

where $\widehat{\boldsymbol{W}} = \text{diag}\left\{\widehat{\boldsymbol{W}}_1,\cdots,\widehat{\boldsymbol{W}}_m\right\}$, $\widehat{\boldsymbol{W}}_i = \text{diag}\{N_{i1}\widehat{\pi}_{i1}(1-\widehat{\pi}_{i1}),\cdots,N_{in}\widehat{\pi}_{in}(1-\widehat{\pi}_{in})\}$ and $\widehat{\pi}_{ij}$ needs to be updated based on the current $\widehat{\boldsymbol{\delta}}$ for each step of iteration; $\boldsymbol{q}$ denotes the adjusted dependent variate,

$$\boldsymbol{q} = \text{logit}(\boldsymbol{\pi}) - \boldsymbol{X}^{*}\widehat{\boldsymbol{\beta}}_{\mathbf{0}}^{*} + \widehat{\boldsymbol{W}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{q}=\left(\boldsymbol{q}_1^T,\cdots,\boldsymbol{q}_m^T\right)^T$, $\boldsymbol{q}_i = (q_{i1},\cdots,q_{in})^T$, $\boldsymbol{\pi}=\left(\boldsymbol{\pi}_1^T,\cdots,\boldsymbol{\pi}_m^T\right)^T$, $\boldsymbol{\pi}_i = (\pi_{i1},\cdots,\pi_{in})^T$, $i = 1,\ \cdots,\ m$; $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}^{*}=\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}}^T,\cdots,\widehat{\boldsymbol{\beta}}_{\mathbf{0}}^T\right)^T$ is an $m \times p$-dimensional vector with each subvector being $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$. Accordingly, the $\widehat{\boldsymbol{\delta}}^{(o)}$ can be obtained by solving Equation A. (A 2).

Similarly, the $\widehat{\boldsymbol{\delta}}_{\underset{\sim}{\mathbf{0}}^{(o)}}$ also needs to be obtained by solving IWLS equation. Take the derivative of Equation (6) with respect to $\boldsymbol{\delta}_{\underset{\sim}{\mathbf{0}}}$ and obtain the following score function,

$$\frac{\ell\left(\boldsymbol{\delta}_{\underset{\sim}{\mathbf{0}}}\right)}{\partial \boldsymbol{\delta}_{\underset{\sim}{\mathbf{0}}}} = \boldsymbol{I}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{I}^{*} = \text{diag}\{\widetilde{\mathbf{1}},\cdots,\widetilde{\mathbf{1}}\}$ is a block diagonal matrix with $m$ identical blocks, and $\widetilde{\mathbf{1}} = (1,\cdots,1)^T$ is an $n$-dimensional vector, $\boldsymbol{y}$ and $\boldsymbol{\mu}$ have the same definitions as that in Equation A. (A 1).

Assume $\boldsymbol{I}^{*T}(\boldsymbol{y}-\boldsymbol{\mu}) = \mathbf{0}$, and we will have the IWLS equation as follows,

$$\boldsymbol{I}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{I}^{*}\boldsymbol{\delta}_{\underset{\sim}{\mathbf{0}}} = \boldsymbol{I}^{*T}\widehat{\boldsymbol{W}}\boldsymbol{q}, \tag{A3}$$

where $\widehat{\boldsymbol{W}}$ and $\boldsymbol{q}$ have the same definitions as that in Equation A. (A 2). Then, the $\widehat{\boldsymbol{\delta}}_{\underset{\sim}{\mathbf{0}}^{(o)}}$ can be obtained by solving Equation A. (A 3).