Nonparametric and semiparametric regression analysis of group testing samples

Mingyu Li[1] and Minge Xie[2, 3]*

[1]*Celgene Corporation, 110 Allen Road, Basking Ridge, NJ 07920 U.S.A.*

[2]*Department of Statistics and Biostatistics, Rutgers University,*

Hill Center for Mathematical Sciences, Piscataway, NJ 08854 U.S.A.

Abstract

This paper develops a general methodology of nonparametric and semiparametric regression for group testing data, relating group testing responses to covariates at individual level. We fit nonparametric and semiparametric models and obtain estimators of the parameters and the nonparametric regression function by maximizing penalized likelihood function. For implementation, we develop a modified EM algorithm with individual responses as complete data and observed group testing responses as observed data. Numerical results based on simulations and chlamydia data collected in a Nebraska study show that our estimation methods perform well for estimating both the individual probability of positive outcome and the prevalence rate in the population.

***Keywords*:** Group testing, EM algorithm, Smoothing, Generalized linear models, Penalized maximum likelihood.

[1] Email: mli@celgene.com

[2] Correspondence to: Minge Xie, Department of Statistics and Biostatistics, Rutgers University, Hill Center for Mathematical Sciences, Piscataway, NJ 08854, U.S.A. Email: mxie@stat.rutgers.edu; Telephone: 732-445-2690

[3] This research is partly supported by grants NSF-DMS-DMS1107012, NSF-DMS0915139, NSF-SES0851521 and DHS-2008-DN-077-ARI012-03.

I. Introduction

Group testing, or pooled testing, where the samples are tested in pools instead of at individual level, was first introduced by Dorfman [1] to reduce cost and increase efficiency of tests. Under the assumption that the group testing result is positive if at least one individual sample in the corresponding pool is positive and negative if none of the individual samples in that pool are positive, Dorfman [1] showed that his proposed group testing method can significantly reduce the total number of tests compared to individual tests. Since then, the group testing method has been widely used in blood or urine screening, chemical compound screening in drug discovery, infectious disease diagnostic tests and molecular biology screening; see, for example, Stramer et al. [2], Xie et al. [3], Remlinger et al. [4] , Lindan et al. [5] , Rours et al. [6], Du and Hwang [7] and reference therein.

Many papers discussed the issue of optimal designs in attempt to find optimal group size which minimizes the total number of tests under various scenarios; see, for example, Dorfman [1], Yao and Hwang [8], Hughes-Oliver and Swallow [9], Phatarfod and Sudbury [10] and Brookmeyer [11]. But the group testing method can also be used to estimate model parameters such as the overall prevalence of the disease in large population and others. Chen and Swallow [12] showed that group testing can substantially reduce the mean square error of the estimator of prevalence rate and the cost per unit information under certain conditions. Vansteelandt, Goetghebeur and Verstraeten [13] suggested that testing pools can lower false positive and false negative rates in low prevalence cases and yield more precise prevalence estimators. Huang and Tebbs [14] demonstrated that despite the loss in information, pooling by group testing can provide robust and improved parameter estimators in terms of mean-squared error under structural measurement error models. Depending on the purpose of study, the group testing schemes and the information contained in the group testing results may vary. If the objective is to identify all positive individuals, all samples in the pools with positive group testing results may be retested. In this case, the individual sample responses are all available. In many other cases, the individual outcomes can not be implied completely from the group testing results, either by design (see, e.g., Gastwirth and Hammick [15] to protect privacy) or due to testing errors and other considerations (see, e.g., Gastwirth and Hammick [15], Vansteelandt et al. [13], Xie [16], Chen, Tebbs and Bilder [17]).

In many studies, the individual covariate information, such as age, gender, and general health information, is available and it is of interest to explore whether such information is related to the responses or not. Parametric regression methodologies have been developed to analyze the relationship between the group testing responses and the covariate variables. Vansteelandt et al. [13] directly maximized the likelihood function of the group testing responses. Xie [16] considered the individual responses as the complete data and the observed testing results as observed data respectively, and applied an EM algorithm to fit regression models. Furthermore, Chen, Tebbs and Bilder [17] studied heterogeneous populations and included a random effect covariate in regression models. In addition, Huang and Tebbs [14] and Huang [18] considered regression analysis of group testing samples in the situations when covariates have measurement errors. So far, the research on the regression method in group testing has focused on parametric models. Nonparametric or semiparametric regression models have not been considered in analysis of group testing samples, partly due to the difficulty of performing nonparametric regression on missing data especially with potentially correlated testing responses.

In this paper, we extend the parametric regression analysis of Xie [16] to nonparametric and semiparametric regression analyses. We use a penalized maximum likelihood method and a modified EM algorithm to overcome the difficult of missing information and potentially correlated responses. Penalized likelihood contains the likelihood function and a roughness penalty term in which a smoothing parameter controls the trade-off between goodness-of-fit and smoothness. Green and Silverman [19] provided a thorough discussion on the penalized maximum likelihood method for nonparametric and semi-parametric regression and generalized linear models. Green [20] applied the EM algorithm to the penalized maximum likelihood estimator and pointed out that the parameter can represent a smooth function that has been discretized. In our work, we combine the algorithms in Green and Silverman [19] and the methodologies in Green [20], and apply the EM algorithm to the nonparametric and semiparametric regression for group testing data. The results of numerical examples show that our estimation methods perform well in estimating both the individual probability of positive outcome and sample prevalence rate. In the simulation studies, we consider two pooling strategies, `alike' and `random' for comparison, and it turns out that `alike' pooling provides notable improvement of the estimators. This result is consistent with that reported by Bilder and Tebbs [21] in parametric models.

The rest of the paper is organized as follows. In Section 2, we present the models, estimation methodology and algorithms. In Section 3, simulation studies are conducted to illustrate the implementation and to evaluate the performance of the proposed estimation methods for nonparametric and semiparametric models. In Section 4, we provide further discussion and summaries.

2. Estimation method

2.1. Notations and models

We assume, without loss of generality, that samples from N subjects are grouped into, say, n pools. These n pools are tested first, and some individuals or subsets of the n pools are further tested, depending on the group testing scheme and the purpose of the study. Following Xie [16], for $i=1,\cdots,N$, we denote by y_i whether the sample from the i^{th} individual is positive or not; it equals 1 if it is positive and 0 if it is negative. Also, we suppose that m tests in total are performed on m (usually $m \geq n$) sets of individuals, say $g_1, g_2, \cdots, g_m$, where the sets correspond to the pools as well as some subsets of the pools or some regrouped pools or some selected individuals, depending on the group testing scheme used. Corresponding to the m sets of $g_1, g_2, \cdots, g_m$, denote by the m testing results as $\mathbf{t}=\left(t_1,\cdots,t_m\right)$; The testing result t_i equals 1 if it is positive and 0 otherwise. As in Xie [16], we assume that the testing methods may not be perfectly accurate and the notions of sensitivity and specificity are used to specify the accuracy of the testing methods. Here, sensitivity is the probability of a positive sample being tested positive and specificity is the probability of a negative sample being tested negative. Let η and θ denote the sensitivity and specificity respectively, then we have $0<\eta\leq 1$ and $0<\theta\leq 1$. Under this assumption, t_i can be decided by

$$t_i = W_i 1_{\left(\sum_{j\in g_i} y_j > 0\right)} + \left(1-V_i\right) 1_{\left(\sum_{j\in g_i} y_j = 0\right)},$$

where W_i and V_i are independent Bernoulli random variables equal to 1 with probabilities η and θ respectively and $1_{(\cdot)}$ is the indicator function. Note that, depending on the testing scheme used, these observed testing results t_i 's may or may not be correlated; see, Section 3 for an example.

When covariate variables and testing response of each individual subject are available, we can often use generalized linear regression models to model the individual response. If we assume that the covariates are linearly related to the individual response through a link function $h(\cdot)$, we can construct a parametric generalized linear model:

$$h[P(y_i = 1)] = x_i^T \beta . \tag{1}$$

If the relationship is not linear and we'd rather model it by an unknown smooth function, we can use a nonparametric generalized linear model:

$$h[P(y_i = 1)] = f(v_i). \tag{2}$$

If we think that some covariates are linearly related and some are related by an unknown smooth function, we can use a semiparametric generalized linear model:

$$h[P(y_i = 1)] = x_i^T \beta + f(v_i). \tag{3}$$

Here, in Models (1) – (3), x_i is a $p \times 1$ covariate vector for the parametric linear regression and v_i is a covariate variable for the regression of non-parametric component. The link function $h(\cdot)$ is a known monotonic function. The most commonly used link function for binary responses is the logit link function $h(p) = \text{logit}(p) = \log(p/(1-p))$. Model (1) has been discussed in [16]. In this paper, we are interested in estimating the unknown smooth function $f(\cdot)$ in Model (2) and the unknown parameters β and the unknown smooth function $f(\cdot)$ in Model (3), assuming that we only know the observed testing responses $\mathbf{t} = (t_1, \cdots, t_m)$ but not the individual-level responses $\mathbf{y} = (y_1, \cdots, y_N)$.

Sometimes, it is possible to identify all the individual testing results $\mathbf{y} = (y_1, \cdots, y_N)$ from the testing results $\mathbf{t} = (t_1, \cdots, t_m)$. In this case, we can simply fit the aforementioned models directly to the individual testing results $\mathbf{y} = (y_1, \cdots, y_N)$. However, in many other cases, the individual testing results can not be fully determined from the group testing results. Our focus is on the more complicated latter case, and the earlier case can be regarded as a simple special case of the latter one. Note that, for some testing schemes, the explicit formula for the likelihood function of $\mathbf{t} = (t_1, \cdots, t_m)$ (observed likelihood) may be very complicated or even unavailable. Direct maximization of the observed likelihood function could be a tedious task, if not impossible. On the other hand, the log-likelihood function of complete data has simple form for generalized linear models. We develop in the rest of this section a set of EM algorithms for nonparametric and semiparametric generalized linear models.

2.2. Estimation method and EM algorithm for nonparametric model

Under model (2), the joint density function of $(y_1, \cdots, y_N)$ is

$$p(y_1, \cdots, y_N \mid f_1, \cdots, f_N) = e^{\sum_{i=1}^{N}\left[y_i f_i - \log\left(1+e^{f_i}\right)\right]},$$

where $f_i = f(v_i)$. We consider $\mathbf{y} = (y_1, \cdots, y_N)$ as the complete data, which is not completely observed. The observed data are the m testing results $\mathbf{t} = (t_1, \cdots, t_m)$. Depending on the testing schemes, sometimes the joint density of observed samples $\mathbf{t} = (t_1, \cdots, t_m)$ can be very complicated with no explicit formula. We denote by $p(t_1, \cdots, t_m \mid f_1, \cdots, f_N)$ the joint density of observed samples $\mathbf{t} = (t_1, \cdots, t_m)$. We define a penalized observed log-likelihood function as

$$l_p(\mathbf{f}) = \log p(t_1, \cdots, t_m \mid f_1, \cdots, f_N) - \alpha/2 \int f''(v)^2 dv, \tag{4}$$

where $\mathbf{f} = (f_1, ..., f_N)$ and α is the smoothing parameter. Our task is to obtain an estimator of $f(\cdot)$, say $\hat{f}(\cdot)$, by maximizing (4).

Note that, we can write that $\log p(t_1, \cdots, t_m \mid f_1, \cdots, f_N) = \log p(\mathbf{y} \mid f_1, \cdots, f_N) - \log p(\mathbf{y} \mid \mathbf{t}, f_1, \cdots, f_N)$. By taking the conditional expectation on both sides, conditional on $\mathbf{t}$ and $(\tilde{f}_1, \cdots, \tilde{f}_N)$, it follows

$$\begin{aligned}\log p(t_1, \cdots, t_m \mid f_1, \cdots, f_N) = {} & E\left[\log p(\mathbf{y} \mid f_1, \cdots, f_N) \mid \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f}_N\right] \\ & - E\left[\log p(\mathbf{y} \mid \mathbf{t}, f_1, \cdots, f_N) \mid \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f}_N\right],\end{aligned} \tag{5}$$

for any set of given parameters $\tilde{\mathbf{f}} = (\tilde{f}_1, \cdots, \tilde{f}_N)$. Define

$$\begin{aligned}Q(\mathbf{f} \mid \tilde{\mathbf{f}}) &\overset{def}{=} E\left[\log p(\mathbf{y} \mid f_1, \cdots, f_N) \mid \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f}_N\right] - \alpha/2 \int f''(v)^2 \, dv \\ &= \sum_{i=1}^{N} \left[E(y_i \mid \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f}_N) f_i - \log(1 + e^{f_i})\right] - \alpha/2 \int f''(v)^2 \, dv.\end{aligned}$$

By (5) and Jensen's inequality, we can show that

$$l_p(\mathbf{f}) - l_p(\tilde{\mathbf{f}}) \geq Q(\mathbf{f} \mid \tilde{\mathbf{f}}) - Q(\tilde{\mathbf{f}} \mid \tilde{\mathbf{f}}), \tag{6}$$

for any two sets of parameters $\mathbf{f}$ and $\tilde{\mathbf{f}}$. Thus, by (6), each time when we update our parameter values from $\tilde{\mathbf{f}}$ to $\mathbf{f}^* = \text{argmax}_{\{\text{any}\,\mathbf{f}\}} Q(\mathbf{f} \mid \tilde{\mathbf{f}})$, we have $l_p(\mathbf{f}^*) - l_p(\tilde{\mathbf{f}}) \geq 0$ and thus the value of $l_p(\mathbf{f})$ will increase from $l_p(\tilde{\mathbf{f}})$ to $l_p(\mathbf{f}^*)$. Following the reasoning of the standard EM algorithm and also Green [20], we design a modified EM algorithm to obtain an estimate of $\mathbf{f} = (f_1, \cdots, f_N)$: For a selected set of starting values $f^{[0]}(v_i)$ of

$f_i = f(v_i)$ for $i = 1, 2, \cdots, N$, we iterate the following E and M steps until $\left\| \mathbf{f}^{[k+1]} - \mathbf{f}^{[k]} \right\|$ is very small.

- *E-step*: For given $f^{[k]}(v_i)$ for $i = 1, \cdots, N$ at the k^{th} iteration, $k = 0, 1, 2, \cdots$, calculate the conditional expectations

$$c_i^{[k]} = E\left[y_i \mid t_1, \cdots, t_m, f^{[k]}(v_1), \cdots, f^{[k]}(v_N) \right], \qquad i = 1, \cdots, N.$$

- *M-step*: Given $\left(c_1^{[k]}, \cdots, c_N^{[k]}\right)$ for a fixed $k = 0, 1, 2, \cdots$, update the estimates at the $(k+1)^{th}$ iteration, $f_i^{[k+1]} = f^{[k+1]}(v_i)$, for $i = 1, \cdots, N$, by maximizing the following penalized log-likelihood function:

$$\sum_{i=1}^{N} \left[c_i^{[k]} f_i - \log\left(1 + e^{f_i}\right) \right] - \alpha/2 \int f''(v)^2 \, dv . \tag{7}$$

For a fixed α, maximizing [7] can be solved via iterating on the penalized weighted least squares problem (refer to Gu [22])

$$\min \sum_{i=1}^{N} \left[b_i'' \left(z_i - f_i \right)^2 \right] + \alpha \int f''(v)^2 \, dv , \tag{8}$$

where $b_i'' = e^{\tilde{f}_i} \Big/ \left(e^{\tilde{f}_i} + 1\right)^2$, $z_i = \tilde{f}_i + \left(c_i^{[k]} - b_i'\right) \Big/ b_i''$, $b_i' = e^{\tilde{f}_i} \Big/ \left(e^{\tilde{f}_i} + 1\right)$ and $\tilde{f}_i = \tilde{f}(v_i)$ is evaluation of $f(v_i)$ in previous iteration. For notation simplicity and without loss of generality, we re-arrange the indexes i, so that $v_1 \le v_2 \le \cdots \le v_N$ (i.e., the ordered $v_{(i)} = v_i$, for $i = 1, \cdots, N$). By Green and Silverman [19], the solution of problem (8) is a natural cubic spline and the penalty term can be written as

$$\alpha \int f''(v)^2 \, dv = \alpha \mathbf{f}^T K \mathbf{f} ,$$

for the natural cubic spline, where $K = QR^{-1}Q^T$ and $\mathbf{f} = \left(f\left(v_{(1)}\right), \cdots, f\left(v_{(N)}\right)\right)$. Here Q is an $n \times (n-2)$ band matrix and R is an $(n-2)\times(n-2)$ symmetric band matrix and each element of these two matrices is a function of $\left(v_{(1)}, \cdots, v_{(N)}\right)$, which is the ordered values of $\left(v_1, \cdots, v_N\right)$. The matrices, Q and R are given in Appendix 5.1. Denote by W a diagonal matrix with $W_{ii} = b_i''$ and denote working response vector by $\mathbf{z} = \left(z_1, \cdots, z_N\right)$, then the matrix form of problem (8) is

$$\min S(\mathbf{f}) = \left(\mathbf{z}-\mathbf{f}\right)^T W\left(\mathbf{z}-\mathbf{f}\right) + \alpha \mathbf{f}^T K\mathbf{f} \ , \tag{9}$$

and the solution of (9) is

$$\mathbf{f}^{new} = \left(W + \alpha K\right)^{-1} W\mathbf{z}\,. \tag{10}$$

In group testing, the sample size N is usually very large, hence direct use of (10) is very time consuming. We apply the Reinsch algorithm for weighted smoothing (refer to Green and Silverman [19]) to calculate (10). The steps of the algorithm are given in Appendix 5.2 and each step can be performed in $O(N)$ algebraic operations.

Finally, we need to choose the smoothing parameter α for our computing. We apply the following generalized cross-validation (GCV) criterion: Choose the α that minimizes

$$GCV\left(\alpha\right) = \frac{\left\|W^{1/2}\left(\mathbf{z}-\mathbf{f}\right)\right\|^2}{n\left[1-n^{-1}\mathrm{trace}\left\{\left(W+\alpha K\right)^{-1}W^{1/2}\right\}\right]^2} = \frac{n\left\|W^{1/2}\left(\mathbf{z}-\mathbf{f}\right)\right\|^2}{\left\{\mathrm{trace}\left(\alpha W^{-1/2}KW^{-1/2}\right)\right\}^2}\,, \tag{11}$$

where W, $\mathbf{z}$ and $\mathbf{f}$ are all evaluated at the converged estimator, $\hat{f}(\cdot)$. Other criteria, like cross validation and likelihood based cross validation, can also be used.

2.3. Estimation method and EM algorithm for semiparametric model

Semiparametric model can be analyzed by using similar estimation method and algorithm developed in Section 2.2. Under the semiparametric model (3), the joint density function of $\left(y_1, \cdots, y_N\right)$ is again very simple:

$$p\left(y_1, \cdots, y_N \mid \beta,\ f_1, \cdots,\ f_N\right) = \exp\left\{\sum_{i=1}^{N}\left[y_i\left(x_i^T\beta + \ f_i\right) - \log\left(1 + e^{x_i^T\beta + f_i}\right)\right]\right\},$$

although the joint density function of the observed testing results $\left(t_1, \cdots, t_m\right)$, say $p\left(t_1, \cdots, t_m \mid \beta,\ f_1, \cdots,\ f_N\right)$, can sometimes be very complicated. We want to maximize the penalized observed log-likelihood function,

$$l_p\left(\beta, \mathbf{f}\right) = \log p\left(t_1, \cdots, t_m \mid \beta,\ f_1, \cdots,\ f_N\right) - \alpha/2\int f\,''\left(v\right)^2 dv, \tag{12}$$

to obtain the estimators of β and $f(\cdot)$, say $\hat{\beta}$ and $\hat{f}(\cdot)$, where α is a smoothing parameter.

For any set of given parameters $\tilde{\beta}$ and $\tilde{\mathbf{f}} = \left(\tilde{f}_1, \cdots,\ \tilde{f}_N\right)$, we define

$$\begin{aligned} Q\left(\beta, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right) &\overset{def}{=} E\left[\log p\left(\mathbf{y} \mid \beta,\ f_1, \cdots,\ f_N\right) \mid \mathbf{t}, \tilde{\beta},\ \tilde{f}_1, \cdots,\ \tilde{f}_N\right] - \alpha/2\int f\,''\left(v\right)^2 dv \\ &= \sum_{i=1}^{N}\left[E\left(y_i \mid \mathbf{t}, \tilde{\beta},\ \tilde{f}_1, \cdots,\ \tilde{f}_N\right)\left(x_i^T\beta + f_i\right) - \log\left(1 + e^{x_i^T\beta + f_i}\right)\right] - \alpha/2\int f\,''\left(v\right)^2 dv. \end{aligned}$$

Again, as in Section 2.2, we can prove that

$$l_p\left(\beta, \mathbf{f}\right) - l_p\left(\tilde{\beta}, \tilde{\mathbf{f}}\right) \geq Q\left(\beta, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right) - Q\left(\tilde{\beta}, \tilde{\mathbf{f}} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right), \tag{13}$$

for any two sets of given parameters $\left(\beta, \mathbf{f}\right) = \left(\beta,\ f_1, \cdots,\ f_N\right)$ and $\left(\tilde{\beta}, \tilde{\mathbf{f}}\right) = \left(\tilde{\beta},\ \tilde{f}_1, \cdots,\ \tilde{f}_N\right)$. Thus, each time when we update our parameter values from $\left(\tilde{\beta}, \tilde{\mathbf{f}}\right)$ to a new set of values $\left(\beta^*, \mathbf{f}^*\right) = \text{argmax}_{\{\text{any}\,\beta, \mathbf{f}\}} Q\left(\beta, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right)$, the value of the panelized log-likelihood function (12) will be increased. Furthermore, based on (13), we can show that the panelized log-likelihood function (12) will increase, either when we update $\tilde{\beta}$ to $\beta^* = \text{argmax}_{\{\text{any }\beta\}} Q\left(\beta, \tilde{\mathbf{f}} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right)$ while holding $\tilde{\mathbf{f}}$ fixed or when we update $\tilde{\mathbf{f}}$ to $\mathbf{f}^* = \text{argmax}_{\{\text{any }\mathbf{f}\}} Q\left(\tilde{\beta}, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right)$ while holding $\tilde{\beta}$ fixed. Thus, we can also use a backfitting method to update the parametric and nonparametric components β and $\mathbf{f}$ separately.

Note that, similar to the nonparametric model, the second integration (penalty) part in $Q\left(\beta, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right)$ can be written by $\mathbf{f}^T K \mathbf{f}$ for natural cubic splines. By Theorem 5.2 of Green and Silverman [19], the Fisher scoring algorithm for maximizing $Q\left(\beta, \mathbf{f} \mid \tilde{\beta}, \tilde{\mathbf{f}}\right)$ with respect to β and $f(\cdot)$ for a fixed α is given by solving

$$\begin{bmatrix} X^T W X & X^T W \\ W X & W + \alpha K \end{bmatrix} \begin{pmatrix} \beta \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} X^T W \mathbf{z} \\ W \mathbf{z} \end{pmatrix}, \tag{14}$$

where the working response vector $\mathbf{z} = \left(z_1, \cdots, z_N\right)$ has the form

$$z_i = x_i^T \tilde{\beta} + \tilde{f}_i + \left(c_i - b_i^{\prime}\right) \Big/ b_i^{\prime\prime},$$

$c_i = E\left(y_i \mid \mathbf{t}, \tilde{\beta}, \tilde{\mathbf{f}}\right)$, and $b_i^{\prime\prime} = e^{x_i^T \tilde{\beta} + \tilde{f}_i} \Big/ \left(e^{x_i^T \tilde{\beta} + \tilde{f}_i} + 1\right)^2$, $b_i^{\prime} = e^{x_i^T \tilde{\beta} + \tilde{f}_i} \Big/ \left(e^{x_i^T \tilde{\beta} + \tilde{f}_i} + 1\right)$, $X = \left(x_1, \cdots, x_N\right)^T$ and W is a diagonal matrix with $W_{ii} = b_i^{\prime\prime}$. Again in this natural cubic spline formulation, for notation simplicity and without loss of generality, we re-arrange the indexes i, so that $v_1 \le v_2 \le \cdots \le v_N$ (i.e., the ordered $v_{(i)} = v_i$, for $i = 1, \cdots, N$).

Equation (14) forms a system of $p + n$ estimating equations for both the parametric and non-parametric components. It can be written as a pair of simultaneous matrix equations for the parametric and non-parametric components separately (refer to Green and Silverman [19]),

$$X^T W X \beta = X^T W \left(\mathbf{z} - \mathbf{f}\right),$$

$$\left(W + \alpha K\right)\mathbf{f} = W\left(\mathbf{z} - X\beta\right).$$

Based on this separation, we propose a backfitting/EM algorithm which runs iteratively between fitting the parametric component and fitting the nonparametric component while holding the other fixed: For a selected set of starting points $\beta^{[0]}$ and $f_i^{[0]}$ for $i = 1, \cdots, N$, we iterate the following two pairs of E and M steps until both $\left\|\beta^{[k+1]} - \beta^{[k]}\right\|$ and $\left\|\mathbf{f}^{[k+1]} - \mathbf{f}^{[k]}\right\|$ are very small.

- *E-step for parametric part*: For given $\beta^{[k]}$ and $f_i^{[k]}$ for $i = 1, \cdots, N$, for $k = 0, 1, 2, \cdots,$ calculate

  $$c_i^{[k]} = E\left(y_i \mid t_1, \cdots, t_m, \beta^{[k]}, f^{[k]}\left(v_1\right), \cdots, f^{[k]}\left(v_N\right)\right), \qquad i = 1, \cdots, N.$$

- *M-step for parametric part*: Given $\left(c_1^{[k]}, \cdots, c_N^{[k]}\right)$ for a fixed $k = 0, 1, 2, \cdots,$ update the estimator at the $\left(k+1\right)^{th}$ iteration, $\beta^{[k+1]}$, by

  $$\beta^{[k+1]} = \left[X^T W X\right]^{-1} X^T W \left(\mathbf{z} - \mathbf{f}^{[k]}\right).$$

- *E-step for nonparametric part*: For given $\beta^{[k+1]}$ and $f_i^{[k]}$, $i = 1, \cdots, N$ at the $\left[k\right]^{th}$ iteration for $k = 0, 1, 2, \cdots,$ calculate

  $$c_i^{[k]} = E\left(y_i \mid t_1, \cdots, t_m, \beta^{[k+1]}, f_1^{[k]}, \cdots, f_N^{[k]}\right), \qquad i = 1, \cdots, N.$$

- *M-step for nonparametric part*: Given $\left(c_1^{[k]}, \cdots, c_N^{[k]}\right)$ for fixed $k = 0, 1, 2, \cdots$, update the estimator at the $(k+1)^{th}$ iteration, $f_i^{[k+1]}$, for $i = 1, \cdots, N$, by

$$\mathbf{f}^{[k+1]} = \left(W + \alpha K\right)^{-1} W\left(\mathbf{z} - X\beta^{[k+1]}\right).$$

In our computing program, the Reinsch algorithm for weighted smoothing is applied in the M-step for the nonparametric component, and the GCV criterion (11), with $(\mathbf{z} - \mathbf{f})$ replaced by $(\mathbf{z} - X\beta - \mathbf{f})$, is used to choose the smoothing parameter.

3. SIMULATION STUDIES

In this section we conduct simulation studies to evaluate the finite sample performance of the penalized maximum likelihood estimation methodology proposed in Section 2. We use a group testing scheme proposed by Gastwirth-Hammick [15] for illustration.

The Gastwirth and Hammick (GH) group testing scheme is designed for estimating the prevalence of a rare disease, in which to protect the privacy of individuals tests are performed only at the group level but not at the individual level. Under the GH group testing scheme, individual samples are batched into pools first. Then a screening test is performed for each pool. After that, those pools classified as positive are given confirmatory tests. Generally speaking, the screening test is cheap but not quite accurate while the confirmatory test is almost perfect with a much higher cost. Gastwirth and Hammick [15] gave such an example in blood testing practice: for testing HIV positives, the commonly used screening test is the ELISA kit and the standard confirmatory test is the Western blot (WB) analysis.

Without loss of generality and for simplicity, we assume that total $N = nk$ individual samples are grouped into n pools of size k, although the proposed algorithm can be applied to different group sizes. Denote by the screening testing results $t_1^{(s)}, \cdots, t_n^{(s)}$, corresponding to pools $g_1, \cdots, g_n$ respectively. The value of $t_i^{(s)}$ is equal to 1 if the test is positive and equal to 0 otherwise. Suppose there are r positive outcomes $t_{j1}^{(s)}, \cdots, t_{jr}^{(s)}$ of n screening tests, and they correspond to the pools $g_{j1}, \cdots, g_{jr}$. Let $t_{j1}^{(c)}, \cdots, t_{jr}^{(c)}$ denote the r confirmatory testing results. Therefore, we have testing results $\mathbf{t} = \left(t_1^{(s)}, \cdots, t_n^{(s)}, t_{j1}^{(c)}, \cdots, t_{jr}^{(c)}\right) \overset{def}{=} \left(t_1, \cdots, t_m\right)$ from pools $G = \left\{g_1, \cdots, g_n, g_{j1}, \cdots, g_{jr}\right\}$ and the total number of tests is $m = n + r$. Note that, the screening test $t_{js}^{(s)}$ and the confirmatory test $t_{js}^{(c)}$ are correlated.

For the screening tests, the testing results can be written as

$$t_j^{(s)} = W_j^{(s)} 1_{\left(\sum_{i \in g_j} y_i > 0\right)} + \left(1 - V_j^{(s)}\right) 1_{\left(\sum_{i \in g_j} y_i = 0\right)}, \tag{15}$$

where $W_j^{(s)}$ and $V_j^{(s)}$ are independent Bernoulli random variables equal to 1 with probabilities $\eta^{(s)}$ and $\theta^{(s)}$ respectively. For the confirmatory tests, the testing results can be expressed as

$$t_{jl}^{(c)} = W_l^{(c)} 1_{\left(\sum_{i \in g_{jl}} y_i > 0\right)} + \left(1 - V_l^{(c)}\right) 1_{\left(\sum_{i \in g_{jl}} y_i = 0\right)}, \tag{16}$$

where $W_l^{(c)}$ and $V_l^{(c)}$ are independent Bernoulli random variables equal to 1 with probabilities $\eta^{(c)}$ and $\theta^{(c)}$ respectively. Here, $\left(\eta^{(s)}, \theta^{(s)}\right)$ are sensitivity and specificity of screening tests and $\left(\eta^{(c)}, \theta^{(c)}\right)$ are sensitivity and specificity of confirmatory tests.

Suppose that individual i belongs to the group g_j, it is easy to verify under the GH group testing scheme that the conditional expectation of y_i given $\left(t_1, \cdots, t_m\right)$ and $\left(f_1, \cdots, f_N\right)$ under the nonparametric model (2) has an explicit formula,

$$\begin{aligned}
&E\left(y_i \mid f_1, \cdots, f_N, t_1, \cdots, t_m\right) \\
&= \frac{\left(1-\eta^{(s)}\right) p_i}{\left(1-\eta^{(s)}\right)\left[1-\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]+\theta^{(s)}\left[\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]} 1_{\left(t_j^{(s)}=0\right)} \\
&+ \frac{\eta^{(s)}\left(1-\eta^{(c)}\right) p_i}{\eta^{(s)}\left(1-\eta^{(c)}\right)\left[1-\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]+\left(1-\theta^{(s)}\right)\theta^{(c)}\left[\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]} 1_{\left(t_j^{(s)}=1, t_j^{(c)}=0\right)} \qquad (17) \\
&+ \frac{\eta^{(s)}\eta^{(c)} p_i}{\eta^{(s)}\eta^{(c)}\left[1-\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]+\left(1-\theta^{(s)}\right)\left(1-\theta^{(c)}\right)\left[\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]} 1_{\left(t_j^{(s)}=t_j^{(c)}=1\right)},
\end{aligned}$$

where $p_i = \exp\left(f_i\right) / \left(1+\exp\left(f_i\right)\right)$. Under the semiparametric model (3), the conditional expectation of y_i given $\left(t_1, \cdots, t_m\right)$, β and $\left(f_1, \cdots, f_N\right)$, $E\left(y_i \mid \beta, f_1, \cdots, f_N, t_1, \cdots, t_m\right)$ has the same formula as (17), except that $p_i = \exp\left(x_i^T \beta + f_i\right) / \left(1+\exp\left(x_i^T \beta + f_i\right)\right)$.

In the special case with the assumption that $\eta^{(c)} = \theta^{(c)} = 1$, the form of (17) is simpler

$$\begin{aligned}
&E\left(y_i \mid f_1, \cdots, f_N, t_1, \cdots, t_m\right) \\
&= \frac{\left(1-\eta^{(s)}\right) p_i}{\left(1-\eta^{(s)}\right)\left[1-\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]+\theta^{(s)}\left[\prod_{i' \in g_j}\left(1-p_{i'}\right)\right]} 1_{\left(t_j^{(s)}=0\right)} \qquad (18) \\
&+ \frac{p_i}{1-\prod_{i' \in g_j}\left(1-p_{i'}\right)} 1_{\left(t_j^{(s)}=t_j^{(c)}=1\right)}.
\end{aligned}$$

The assumption $\eta^{(c)} = \theta^{(c)} = 1$ means that the confirmatory test is perfectly accurate. It was used in Gastwirth and Hammick [15] in their numerical examples. We also adopt it in our simulation studies.

The next two subsections contain two examples of simulation studies, one for nonparametric model, and the other for semiparametric model. The simulation study for the semiparametric model is based on the chlamydia data of Chen et al. [17] collected by the State of Nebraska as a part of an Infertility Prevention Project.

3.1. Nonparametric model

In the first simulation study, we assume that the true underlying model is

$$\operatorname{logit}\left\{P\left(y_i = 1\right)\right\} = a + b\sin\left(v_i/2\right), \tag{19}$$

where $a = -2.65$ and $b = 0.6$. However, this sin curve function is not known to us. Rather, we use the nonparametric model (2), with the logit link, to study group testing data generated from model (19).

The group testing data from model (19) are generated as follows. First, individual level data $\left(v_i, y_i\right)$, $i = 1, \cdots, N$, are simulated where the covariate v_i is from uniform distribution $U\left(-6.28, 6.28\right)$ and y_i is according to model (19). We use the `random' pooling strategy, in which the N individuals are randomly pooled into $n = N/5$ groups $\left(g_1, \cdots, g_n\right)$ of size 5 regardless of their covariate values. For simplicity, we use the same size for all the pools and take pool size 5 for illustration. After grouping, the results of screening tests are generated according to (15) with $\eta^{(s)} = 0.923$ and $\theta^{(s)} = 0.996$, the same sensitivity and specificity used in the simulation study of Xie [16]. Furthermore, the results of confirmatory tests are generated by (16) assuming $\eta^{(c)} = \theta^{(c)} = 1$ for the pools with positive screening test results. The collection of $\left(v_1, \cdots, v_N\right)$ and $\mathbf{t} = \left(t_1, \cdots, t_n, t_{j1}, \cdots, t_{jr}\right)$ forms a data set under the `random' pooling, and we use them to estimate the nonparametric function $f\left(\cdot\right)$ in model (2).

Bilder and Tebbs [21] studied an `alike' pooling strategy, in which individual samples with similar covariates are grouped together. They reported that the `alike' pooling strategy can improve the estimation of the `random' strategy. We also consider an `alike' pooling strategy in our simulation. In the `alike' pooling strategy, the covariate $v_i$ is first ordered and then grouped in size of 5 according to the order. We proceed the same testing approach as in the `random' pooling strategy, and obtain a data set under the `alike' pooling strategy. This set of group testing data is also analyzed to estimate the same nonparametric function $f(\cdot)$ in model (2).

The above simulation is repeated 2×200 times, 200 times each for sample size $N = 5000$ and 10000 respectively. Since in the simulation study we know the true $f(\cdot)$, we can choose the smoothing parameter $\alpha = \alpha_{MISE}$ by minimizing the mean integrated squared error (MISE) of the estimators $\hat{f}(\cdot)$. We also use generalized cross validation criterion (11) to choose $\alpha = \alpha_{GCV}$ and compare the estimators using α_{GCV} to those using α_{MISE} and true $f(\cdot)$ values. The optimum smoothing parameter α is searched on the grid 0.1(0.05)1.

The simulation results are summarized in Table 1. Table 1 shows the numerical estimates of integrated relative bias (i.e., $\int \left| \left[\hat{f}(v) - f(v) \right] / f(v) \right| dF(v)$), the integrated standard error (i.e., $\int \hat{SE} \left\{ \hat{f}(v) \right\} dF(v)$), the integrated mean integrated square error (MISE) (i.e., $\int \left[\hat{f}(v) - f(v) \right]^2 dF(v)$) and the estimator of prevalence rate (i.e., $\int \exp\left(\hat{f}(v) \right) \Big/ \left[1 + \exp\left(\hat{f}(v) \right) \right] dF(v)$), respectively. In the table, $\alpha = \alpha_{GCV}$ (`random') means the `random' pooling strategy is used and the smoothing parameter α is selected by minimizing GCV score. Similarly, $\alpha = \alpha_{MISE}$ (`alike') means the `alike' pooling strategy is used and the smoothing parameter α is selected by minimizing MISE value and so on.

Table 1: Simulation results for nonparametric model based on 200 replications.

	Relative bias	*Empirical S.E.*	*Empirical MISE*	*prev. (7.08%)*
		α_{GCV} (‘random’)		
$N = 5000$	0.026	0.422	0.0076	6.68
$N = 10000$	0.014	0.301	0.0022	6.86
		α_{MISE} (‘random’)		
$N = 5000$	0.020	0.341	0.0041	6.79
$N = 10000$	0.011	0.247	0.0014	6.93
		α_{GCV} (‘alike’)		
$N = 5000$	0.008	0.237	0.0006	6.99
$N = 10000$	0.005	0.180	0.0002	7.03
		α_{MISE} (‘alike’)		
$N = 5000$	0.005	0.188	0.0002	7.02
$N = 10000$	0.003	0.143	0.0001	7.05

From Table 1, we can see that `alike' method provides better estimators than `random' method for both $\alpha = \alpha_{MISE}$ and $\alpha = \alpha_{GCV}$, which is intuitive. In addition, using $\alpha = \alpha_{MISE}$ gives a little better estimator than using $\alpha = \alpha_{GCV}$ for both pooling strategies, which is also expected. When sample size N is equal to 10000, the relative bias and empirical MISE are reduced about half compared to those when N is equal to 5000 for both $\alpha = \alpha_{MISE}$ and $\alpha = \alpha_{GCV}$. From model (19), when $v_i \sim U(-6.28, 6.28)$, the population prevalence (equivalently, the mean probability that y equal to 1) is 7.08%. Compared to this true prevalence 7.08%, the estimators of the prevalence are very close to true value for all cases.

The point-wise average of the estimated nonparametric curves $\hat{f}(\cdot)$ over 200 replications are displayed in Figure 1. The left panel of Figure 1 shows the estimators using $N = 5000$ and the right panel is for the estimators using $N = 10000$.

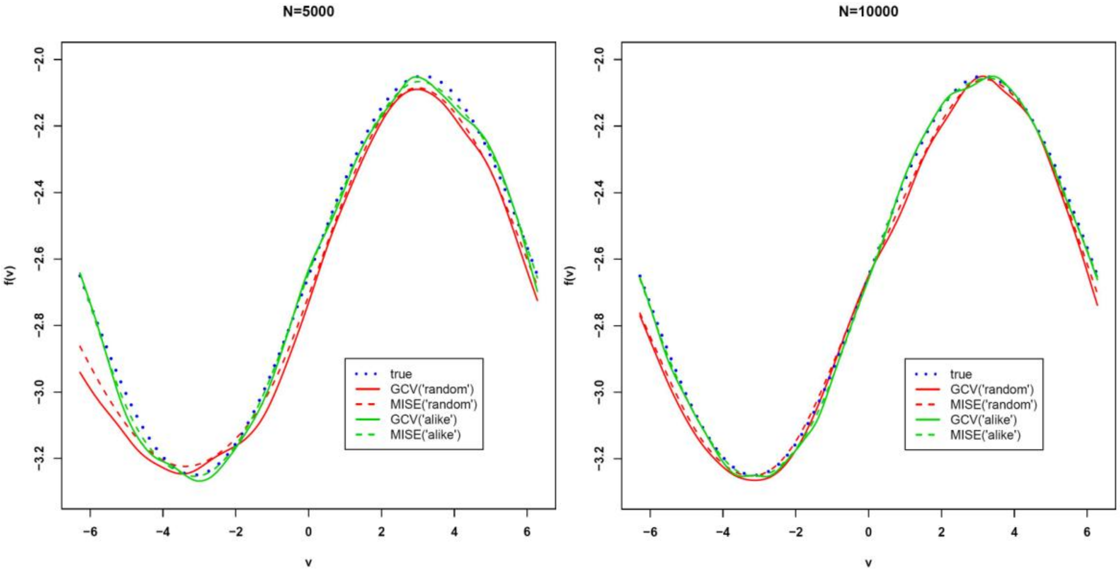

Figure 1: Point-wise average of the estimated nonparametric curve $\hat{f}(\cdot)$ for nonparametric model based on 200 replications. Left panel is for $N = 5000$ and right panel is for $N = 10000$: the blue dotted curve is the true values; the red solid curve is for α_{GCV} and `random' pooling, while the red dashed curve is for $\alpha_{MISE}$ and `random' pooling; the green solid curve is for α_{GCV} and `alike' pooling, while the green dashed curve is for $\alpha_{MISE}$ and `alike' pooling.

In Figure 1, the blue dotted curve represents the true values; the red solid curve is for the estimator using α_{GCV} and `random' pooling, while the red dashed curve represents the estimator using smoothing parameter $\alpha_{MISE}$ and `random' pooling; the green solid curve is for the estimator using α_{GCV} and `alike' pooling, while the green dashed curve represents the estimator using smoothing parameter $\alpha_{MISE}$ and `alike' pooling. First of all, the point-wise average curves of the estimators from 4 methods are all close to the true curve. Second, the estimator using α_{MISE} is closer to the true curve than the one using α_{GCV} given the same pooling strategy, which is expected, however, the difference becomes smaller as the sample size N increases. In addition, the `alike' pooling method has notable improvement compared to the `random' pooling method. When sample size increases, the difference from pooling strategies and smoothing parameter selection criteria becomes smaller and all the estimators are very close to the true curve in the whole support of the covariate.

Figure 2 illustrates the point-wise variances of the estimators $\hat{f}(\cdot)$ over 200 replications, with the left panel for $N = 5000$ and right panel for $N = 10000$. In Figure 2, the red solid curve is for the estimator using α_{GCV} and `random' pooling, while the red dashed curve represents the estimator using smoothing parameter $\alpha_{MISE}$ and `random' pooling; the green solid curve is for the estimator using α_{GCV} and `alike' pooling, while the green dashed curve represents the estimator using smoothing parameter $\alpha_{MISE}$ and `alike' pooling. All the variance curves have a similar trend. They have larger variances in the margin of the support of v and when the corresponding probability of positive, p_i is low. Furthermore, the variances decrease dramatically when the sample size increases, and `alike' method has smaller point-wise variances than `random' method.

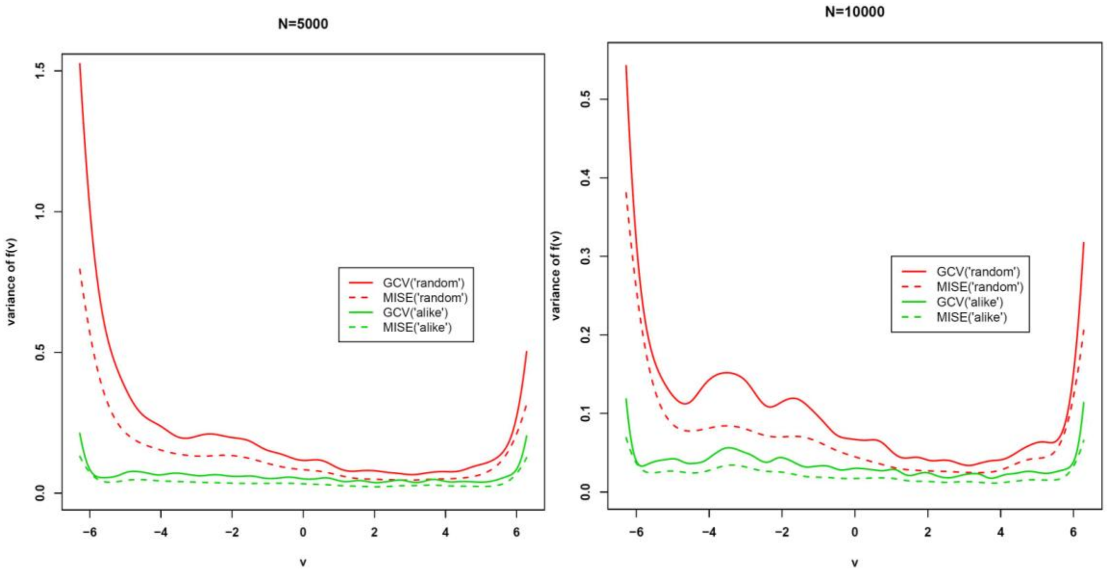

Figure 2: Empirical point-wise variances of the estimated nonparametric curve $\hat{f}(\cdot)$ for nonparametric model based on 200 replications. Left panel is for $N = 5000$ and right panel is for $N = 10000$: the red solid curve is for α_{GCV} and `random' pooling, while the red dashed curve is for $\alpha_{MISE}$ and `random' pooling; the green solid curve is for α_{GCV} and `alike' pooling, while the green dashed curve is for $\alpha_{MISE}$ and `alike' pooling.

In conclusion, the simulation studies demonstrate that our proposed estimation algorithm for nonparametric model performs well and the proposed generalized cross validation criterion can choose proper smoothing parameters in these settings.

3.2. Semiparametric model

In this section we conduct a simulation study based on the chlamydia data example studied in Chen et al. [17]. Chen et al. [17] developed regression method to fit mixed effect models for group testing samples, and applied their method to the chlamydia data collected by the state of Nebraska. The data set consists of chlamydia infection statuses for 6138 subjects, and the risk covariates like age, gender, urethritis status and infection symptoms status. The sample prevalence is 7.8 percent.

In our example, we consider two covariates, Age and some continuous covariate V, and assume that Age is linearly related to the link function, while V has nonparametric relationship with the link function. We fit the semiparametric GLM:

$$\text{logit}\left\{P\left(y_i = 1\right)\right\} = \beta * \text{age}_i + f\left(v_i\right), \tag{20}$$

and estimate β and $f(\cdot)$.

For simplicity, we take the total number of subjects N equal to 6140 and group the samples into 1228 pools with group size 5. Again, the smoothing parameter α is selected by minimizing GCV, and is searched on the grid 0.1(0.05)1. We use both `alike' and `random' pooling strategies. For the `alike' grouping, there are two approaches: `alike' by par-non and `alike' by non, depending on sorting by which covariate first. The `alike' by par-non means that we sort by Age, and then sort by V in the same value of Age; while `alike' by non means that we sort the samples by V (assume that there are no ties in V). For model (20), the covariate Age is generated randomly from {15:45}, V is a continuous random variable from uniform distribution $U(1.57, 7.85)$ and $f(v) = -1.25 + \sin(v)$. Assume that true β is equal to -0.05 and $\eta^{(s)} = 0.95$ and $\theta^{(s)} = 0.98$. Under these settings, the overall positive percentage is about 7.8 percent and only 21.6% tests are needed compared to the individual testing. For the model (20), we estimate β and $f(\cdot)$ by the EM algorithm proposed in Section 2.3.

In this study, we again use 200 replications. Table 2 shows the average and standard error of 200 estimators $\hat{\beta}$, and the integrated relative bias, the integrated S.E. and the integrated MISE of $\hat{f}(\cdot)$. The estimator of prevalence rate is also calculated.

Table 2: Simulation results for semiparametric model based on 200 replications.

	$\hat{\beta}$		$\hat{f}(\cdot)$			
pooling	*Mean*	*S.E.*	*Relative*	*Empirical*	*Empirical*	*prev.*
strategy	(-0.05)		*bias*	*S.E.*	*MISE*	(7.8%)
‘random’	-0.050	0.011	0.052	0.410	0.0018	7.58
‘alike’ by par-non	-0.050	0.006	0.036	0.226	0.0006	7.73
‘alike’ by non	-0.049	0.013	0.031	0.400	0.0006	7.87

Table 2 shows that for the parametric part β, all of the three pooling strategies – `random', `alike' by par-non and `alike' by non, provide good estimators, which are very close to the true value -0.05 with small standard errors. Among them, `random' and `alike' by non have similar S.E.s and `alike' by par-non has smallest S.E. For the nonparametric part, `alike' by non has smallest relative bias and empirical MISE and `alike' by par-non has the smallest empirical S.E. In addition, all the three pooling methods provide the estimators of prevalence rate very close to the true value, 7.8%.

The box-plot of the estimators of β is displayed in Figure 3 for `random', `alike' by par-non and `alike' by non. This plot shows clearly that averages of the estimators of $\beta$ are all very close to the true value -0.05 (the dotted line) for 3 pooling strategies. The standard error of `alike' by par-non is the smallest and `random' method has similar standard error with the `alike' by non approach.

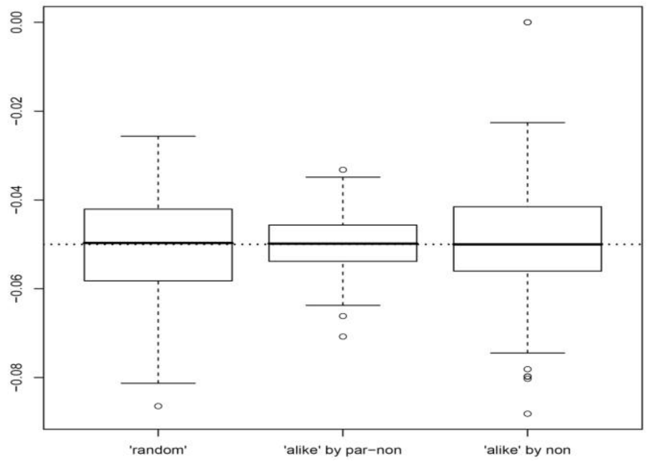

Figure 3: Box-plot of the estimated β for semiparametric model for `random', `alike' by par-non and `alike' by non pooling strategies based on 200 replications. The horizontal dotted line corresponds to the true value of β, -0.05.

The point-wise average (left panel) and point-wise variance (right panel) of the estimators $\hat{f}(\cdot)$ over 200 replications are displayed in Figure 4. In Figure 4, the blue dotted curve represents the true values; the red solid curve is for `random'; and the green solid curve is for `alike' by par-non, while the green dashed curve is for `alike' by non pooling strategy. From left panel of Figure 4, we can notice that the point-wise average curves from two `alike' methods are a little closer to the true curve than `random' method. In the right panel of Figure 4, the `alike' by par-non method gives smallest variances in the whole support of v, and the `random' and `alike' by non methods have similar variance curves.

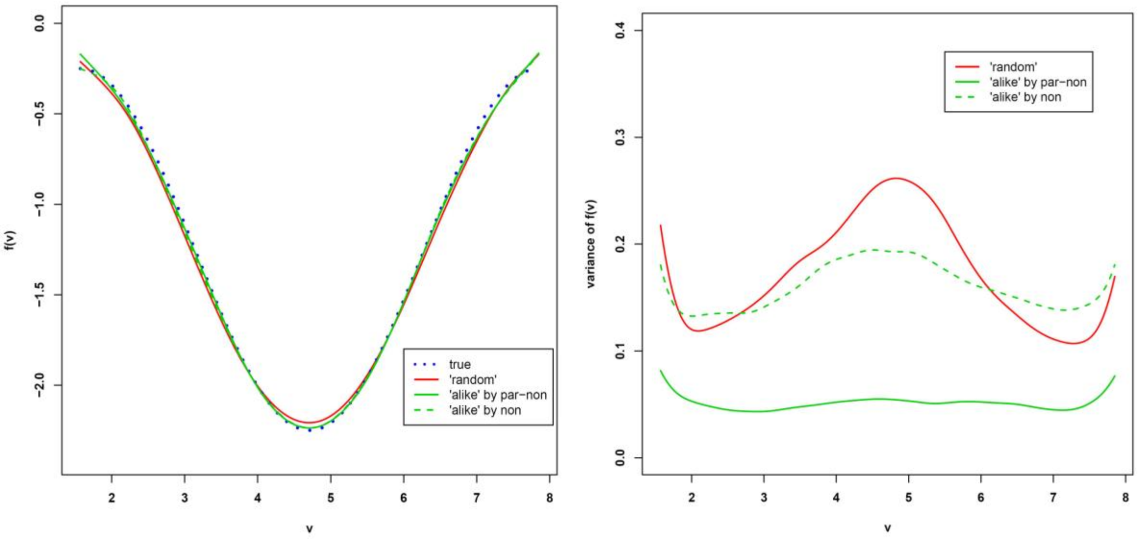

Figure 4: Simulation results of the estimated nonparametric curve $\hat{f}(\cdot)$ for semiparametric model based on 200 replications. Left panel is for point-wise average and right panel is for point-wise variance: the blue dotted curve is for the true values; the red solid curve is for `random'; and the green solid curve is for `alike' by par-non, while the green dashed curve is for `alike' by non.

In conclusion, our estimation methodology gives good estimators for both the parametric component and nonparametric component and the prevalence rate in semiparametric model.

4. DISCUSSION

In this paper, we generalized the parametric model in Xie [16] and fitted nonparametric and semiparametric models for group testing responses using the covariate information. We maximize the penalized likelihood function of group testing results and apply the EM algorithm, considering the group testing as the missing data case. By the information inequality, the EM algorithm can be used in both nonparametric and semiparametric models. For the group testing experiment, since the number of subjects is usually very large, direct use of available software may not be practical. Therefore, the computational aspect has been discussed, and the method of choosing the smoothing parameter has also been considered.

The simulation studies confirm that our proposed estimation methodologies perform very well for both nonparametric and semiparametric models for group testing samples. In simulation studies, we use `random' and `alike' pooling strategies, and the results show that `alike' method improves the estimators significantly, which agrees with the results reported by Bilder and Tebbs [21] and others.

5. APPENDICES

5.1. Q and R matrices

Let $\left(v_{(1)}, \cdots, v_{(N)}\right)$ are ordered values of $\left(v_1, \cdots, v_N\right)$ and assume that there is no tie. Let $h_i = v_{(i+1)} - v_{(i)}$ for $i = 1, \cdots, N-1$. Then Q is an $N \times \left(N-2\right)$ band matrix with entries q_{ij}, for $i = 1, \cdots, N$ and $j = 2, \cdots, N-1$, given by

$$q_{j-1,\, j} = h_{j-1}^{-1},\, q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \text{ and } q_{j+1,\, j} = h_j^{-1}, \text{ for } j = 2, \cdots, N-1,$$

and $q_{ij} = 0$ for $\left|i - j\right| \geq 2$. The columns of Q are numbered starting at $j = 2$, so that the top left element of Q is q_{12}.

$$Q = \begin{pmatrix} q_{12} & q_{13} & \cdots & q_{1,N-1} \\ q_{22} & q_{23} & \cdots & q_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N-1,2} & q_{N-1,3} & \cdots & q_{N-1,N-1} \\ q_{N,2} & q_{N,3} & \cdots & q_{N,N-1} \end{pmatrix}$$

$$= \begin{pmatrix} h_1^{-1} & & & & \\ -h_1^{-1} - h_2^{-1} & h_2^{-1} & & 0 & \\ h_2^{-1} & -h_2^{-1} - h_3^{-1} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & -h_{N-3}^{-1} - h_{N-2}^{-1} & h_{N-2}^{-1} \\ & & & h_{N-2}^{-1} & -h_{N-2}^{-1} - h_{N-1}^{-1} \\ & & & & h_{N-1}^{-1} \end{pmatrix}.$$

The symmetric band matrix R is $(N-2)\times(N-2)$ with elements r_{ij}, for i and j both from 2 to $(N-1)$, given by

$$r_{ii} = \frac{1}{3}\left(h_{i-1} + h_i\right) \text{ for } i = 2, \cdots, N-1,$$

$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6} h_i \text{ for } i = 2, \cdots, N-2,$$

and $r_{ij} = 0$ for $|i-j| \geq 2$.

$$R = \begin{pmatrix} r_{22} & r_{23} & \cdots & r_{2,N-1} \\ r_{32} & r_{33} & \cdots & r_{3,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N-1,2} & r_{N-1,3} & \cdots & r_{N-1,N-1} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & & 0 & \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & & \\ & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \ddots & \\ & & \ddots & \ddots & \frac{h_{N-2}}{6} \\ & 0 & & \frac{h_{N-2}}{6} & \frac{h_{N-2}+h_{N-1}}{3} \end{pmatrix}.$$

5.2. Reinsch algorithm for weighted smoothing

Define the $(N-2)$-vector γ as $\gamma_i = \partial^2 f\left(v_{(i)}\right) / \partial^2 v_{(i)}^2$ for $i = 2, \cdots, N-1$, then we have $Q^T \mathbf{f} = R\gamma$ for natural cubic spline (refer to Green and Silverman [19]).

The solution of (8) satisfies $\mathbf{f} = \left(W + \alpha Q R^{-1} Q^T\right)^{-1} W\mathbf{z}$, which implies

$$W\mathbf{f} = W\mathbf{z} - \alpha Q R^{-1} Q^T \mathbf{f} = W\mathbf{z} - \alpha Q\gamma .$$

Therefore, $\mathbf{f} = \mathbf{z} - \alpha W^{-1} Q\gamma$. Again, by $Q^T \mathbf{f} = R\gamma$,

$$Q^T \mathbf{f} = Q^T \mathbf{z} - \alpha Q^T W^{-1} Q\gamma$$

$$R\gamma = Q^T \mathbf{z} - \alpha Q^T W^{-1} Q\gamma$$

$$\left(R + \alpha Q^T W^{-1} Q\right)\gamma = Q^T \mathbf{z} .$$

The algorithm for weighted spline smoothing is

- Step 1 Evaluate the vector $Q^T \mathbf{z}$.

- Step 2 Find the non-zero diagonals of $R+\alpha Q^T W^{-1} Q$, and its Cholesky decomposition factors L and D.
- Step 3 Solve $LDL^T \gamma = Q^T \mathbf{z}$ for γ by forward and back substitution.
- Step 4 Use $\mathbf{f} = \mathbf{z} - \alpha W^{-1} Q^T \gamma$ to find $\mathbf{f}$.

References

[1] Dorfman R. The detection of defective members of large populations. Ann Math Statist 1943; 14(4):436-440.

[2] Stramer S, Glynn S, Kleinman S, Strong D, Caglioti S, Wright D *et al*. Detection of HIV-1 and HCV infections among antibody-negative blood donors by nucleic acid-amplication testing. N Engl J Med 2004; 351:760-768.

[3] Xie M, Tatsuoka K, Sacks J, Young S. Group testing with Blockers and synergism. J Am Statist Assoc 2001; 96:92-101.

[4] Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. Technometrics 2006; 48:133-144.

[5] Lindan C, Mathur M, Kumta S, Jerajani H, Gogate A, Schachter J *et al*. Utility of pooled urine specimens for detection of Chlamydia trachomatis and Neisseria gonorrhoeae in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. J Clin Microbiol 2005; 43(4):1674-1677.

[6] Rours G, Verkooyen RP, Willemse H, van der Zwaan E, van Belkum A, de Groot R *et al*. Use of pooled urine samples and automated DNA isolation to achieve improved sensitivity and cost-effectiveness of large-scale testing for Chlamydia trachomatis in pregnant women. J Clin Microbiol 2005; 43(9):4684-4690.

[7] Du D, Hwang FK. Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing. World Scientic Publishing Company; 2006.

[8] Yao YC, Hwang FK. On optimal nested group testing algorithms. J Stat Plan Infer 1990; 24(2):167-175.

[9] Hughes-Oliver JM, Swallow WH. A two-stage adaptive group-testing procedure for estimating small proportions. J Am Statist Assoc 1994; 89:982-993.

[10] Phatarfod RM, Sudbury A. The use of a square array scheme in blood testing. Stat Med 1994; 13(22):2337-2343.

[11] Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. Biometrics 1999; 55(2):608-612.

[12] Chen CL, Swallow WH. Using group testing to estimate a proportion, and to test the binomial model. Biometrics 1990; 46:1035-1046.

[13] Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. Biometrics 2000; 56(4):1126-1133.

[14] Huang, X., Tebbs, J. M., On latent-variable model misspecification in structural measurement error models for binary response. Biometrics 2009; 65, 710-718.

[15] Gastwirth JL, Hammick PA. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. J Stat Plan Infer 1989; 22:15-27.

[16] Xie M. Regression analysis of group testing samples. Stat Med 2001; 20(13):1957-1969.

[17] Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. Biometrics 2009; 65(4):1270-1278.

[18] Huang, X. An improved test of latent-variable model misspecification in structural measurement error models for group testing data. Stat Med 2009; 28:3316-3327.

[19] Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models: A roughness penalty approach. Champman & Hall; 1994.

[20] Green PJ. On use of the EM algorithm for penalized likelihood estimation. J R Stat Soc B 1990; 52(3):443-452.

[21] Bilder CR, Tebbs JM. Bias, efficiency, and agreement for group-testing regression models. J Stat Comput Sim 2009; 79(1):67-80.

[22] Gu C. Cross-validating non-gaussian data. J Comput Graph Stat, 1992; 1(2):169-179.