



Single-index regression for pooled biomarker data

Juexin Lin and Dewei Wang 

Department of Statistics, University of South Carolina, Columbia, SC, USA

ABSTRACT

Laboratory assays used to evaluate biomarkers (biological markers) are often prohibitively expensive. As an efficient data collection mechanism to save on testing costs, pooling has become more commonly used in epidemiological research. Useful statistical methods have been proposed to relate pooled biomarker measurements to individual covariate information. However, most of these regression techniques have proceeded under parametric linear assumptions. To relax such assumptions, we propose a semiparametric approach that originates from the context of the single-index model. Unlike with traditional single-index methodologies, we face a challenge in that the observed data are biomarker measurements on pools rather than individual specimens. In this article, we propose a method that addresses this challenge. The asymptotic properties of our estimators are derived. We illustrate the finite sample performance of our estimators through simulation and by applying it to a diabetes data set and a chemokine data set.

ARTICLE HISTORY

Received 14 February 2017
Accepted 24 May 2018

KEYWORDS


Biomarker pooling;
semiparametric regression;
single-index models

1. Introduction

Pooled testing was originally proposed by Dorfman (1943) to detect syphilis among U.S. army recruits during World War II. The general idea is to test pooled specimens formed from combining individual specimens (e.g. blood, urine, plasma, etc.) rather than to test each specimen separately. Since Dorfman's seminal work, pooled testing has been recognised as a cost-effective strategy to perform large-scale screening for rare infectious diseases. In addition to reducing testing costs, pooling specimens can also preserve irreplaceable specimens, tackle the hindrance of detection limits and reduce the impact of potential outliers (Schisterman, Vexler, Yi, and Perkins 2011). Because of these benefits, pooled testing has been employed in a variety of areas, including infectious disease screening (Van et al. 2012), genetics (Gastwirth 2000), blood safety (Dodd, Notari, and Stramer 2002), drug discovery (Remlinger, Hughes-Oliver, Young, and Lam 2006) and animal ecology (Dhand, Johnson, and Toribio 2010).

In addition to being used for case identification, pooling has also been implemented for the purposes of estimation. Thompson (1962) considered estimating the proportion

CONTACT Dewei Wang  deweiwang@stat.sc.edu  Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/10485252.2018.1483501>

of a certain characteristic using binary responses measured on pools. Farrington (1992), Vansteelandt, Goetghebeur, and Verstraeten (2000), Xie (2001) and Wang, McMahan, and Gallagher (2015) used generalised linear models to relate binary pooled responses to individual covariate information. Recently, nonparametric (Delaigle and Meister 2011; Delaigle and Hall 2012; Wang, Zhou, and Kulasekera 2013) and semiparametric (Wang, McMahan, Gallagher, and Kulasekera 2014; Delaigle, Hall, and Wishart 2014) regression methodologies have been proposed for pooled testing data. It is important to note that the articles in this paragraph specifically addressed modelling binary outcomes.

More generally, pooling has also been expanded to assess biomarkers, where continuous assessments (such as biomarker concentration levels) are produced (Weinberg and Umbach 1999; Vexler, Liu, and Schisterman 2006; Whitcomb, Perkins, Zhang, Ye, and Lyles 2012). Early statistical research in this area mainly focused on the evaluation of diagnostic tools by estimating the receiver-operating characteristic curve and the corresponding area under it (see Schisterman, Faraggi, Reiser, and Trevisan 2001; Faraggi, Reiser, and Schisterman 2003; Liu and Schisterman 2003; Mumford, Schisterman, Vexler, and Liu 2006; Bondell, Liu, and Schisterman 2007; Vexler, Schisterman, and Liu 2008). More recently, regression problems that relate continuous biomarker responses measured on pools to individual level covariates have been considered. Ma, Vexler, Schisterman, and Tian (2011) and McMahan, McLain, Gallagher, and Schisterman (2016) used a linear regression model for pooled biomarker responses under the assumption that the individual biomarker levels are conditionally Gaussian given the individual covariates. Under this same assumption, Malinovsky, Albert, and Schisterman (2012) proposed a model that incorporates random effects. Because many biomarkers tend to have right-skewed distributions, Mitchell et al. (2014) explored linear regression analysis under the assumption that individual biomarker levels follow a log-normal distribution. All of these regression techniques were constructed under a parametric linear model and an assumption that the type of the biomarker distribution is known. If the linear model does not hold or the distribution is misspecified, these methods would lead to biased estimates.

In this paper, we propose a semiparametric method to model pooled data with continuous responses and individual covariate information, which overcomes the curse of dimensionality and maintains the important advantage of nonparametric flexibility. Furthermore, our estimation procedure does not require any prior knowledge on the biomarker distribution. The new methodology is proposed in the context of the single-index model. Instead of assuming a linear model, the single-index model assumes the mean of an individual response is related to a linear combination of the covariates through an unknown smooth function. It is a popular semiparametric model to accommodate multi-dimensional covariates while retaining the interpretability of the regression coefficients, see, for example, Ichimura (1993), Härdle, Hall, and Ichimura (1993), Klein and Spady (1993), Xia, Tong, Li, and Zhu (2002), Xia (2006), Zhu and Xue (2006) and Cui, Härdle, and Zhu (2011), who consider responses available on the individual level. In pooled testing, of course, the data are measured on pools. Existing single-index methods in pooled testing were proposed by Wang et al. (2014) and Delaigle et al. (2014) for binary responses. This article presents a new single-index technique to analyse continuous pooled outcomes. We illustrate that when the population size is fixed, pooling could significantly reduce testing costs with only a minor loss in accuracy. On the other hand, when the number of assays

is limited, testing pooled specimens does not compromise the estimation when compared to testing individual specimens.

The rest of this article is organised as follows. In Section 2, we present our semiparametric regression model to analyse biomarkers measured on pooled specimens, and in Section 3, we establish the asymptotic properties of the proposed estimators. We assess the performance of our methods using simulation in Section 4 and apply them to a diabetes data set from the National Health and Nutrition Examination Survey (NHANES) in Section 5.1 and a chemokine data set obtained from the Collaborative Perinatal Project (CPP) in Section 5.2. A discussion is given in Section 6, and regularity conditions are provided in the Appendix. Proofs and additional simulation results are provided in the Web-based Supplementary Materials.

2. Model and methodology

2.1. Assumptions

We consider the situation in which J laboratory assays are taken on pools to measure a continuous biomarker of interest. The j th pool is formed by mixing c_j specimens, each of which is obtained from an individual. The total number of individuals is denoted by $N = \sum_{j=1}^J c_j$. We let Y_{ij} and $\mathbf{X}_{ij} = (X_{ij1}, \dots, X_{ijp})^\top$ denote the continuous biomarker level and the p -dimensional covariate of the i th individual in the j th pool, respectively, where $i = 1, \dots, c_j$ and $j = 1, \dots, J$. Assume throughout that the $(Y_{ij}, \mathbf{X}_{ij})$'s are independent and identical distributed (iid) versions of (Y, \mathbf{X}) , where the mean and variance of Y given $\mathbf{X} = \mathbf{x}$ are

$$E(Y | \mathbf{X} = \mathbf{x}) = \eta(\mathbf{x}^\top \boldsymbol{\beta}) \quad \text{and} \quad V(Y | \mathbf{X} = \mathbf{x}) = \sigma^2,$$

respectively, where $\eta(\cdot)$ is an unknown smooth curve, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is an unknown p -dimensional regression parameter, and $\sigma^2 > 0$. Note that we do not assume the type of the distribution of $Y | \mathbf{X} = \mathbf{x}$ to be known in advance. To ensure identifiability of a single-index model (Lin and Kulasekera 2007), we assume that the support of the covariates, denoted by \mathbb{X} , is a bounded convex set that contains at least one interior point and the parameter space of $\boldsymbol{\beta}$ is $\mathcal{B} = \{\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top : \beta_1 > 0, \|\boldsymbol{\beta}\| = 1\}$, where $\|\boldsymbol{\beta}\| = (\sum_{j=1}^p \beta_j^2)^{1/2}$. If the $(Y_{ij}, \mathbf{X}_{ij})$'s are observed, then traditional single-index modelling techniques can be applied to estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$; e.g. see Ichimura (1993), Xia (2006), Wang, Xue, Zhu, and Chong (2010) and Cui et al. (2011). However, in pooled testing, because assays are not taken on each individual, the Y_{ij} 's are all latent and the responses available to us are on the pooled level.

Denote the biomarker level of the j th pooled specimen by Z_j . In this article, we assume that $Z_j = c_j^{-1} \sum_{i=1}^{c_j} Y_{ij}$; i.e. the observed biomarker response is the arithmetic average of the individual biomarker levels. This is a common assumption in the statistical literature on biomarker pooling (Weinberg and Umbach 1999; Faraggi et al. 2003; Vexler et al. 2008; Malinovsky et al. 2012; Lyles, Van Domelen, Mitchell and Schisterman 2015; McMahan et al. 2016). We view this to be reasonable as long as each individual contributes the same amount to the pool and there is no neutralisation while pooling. The observed data are

$\{(Z_j, \mathbf{X}_{1j}, \dots, \mathbf{X}_{c_jj}) : j = 1, \dots, J\}$, where

$$E(Z_j | \mathbf{X}_{1j} = \mathbf{x}_1, \dots, \mathbf{X}_{c_jj} = \mathbf{x}_{c_j}) = \frac{1}{c_j} \sum_{i=1}^{c_j} \eta(\mathbf{x}_i^\top \boldsymbol{\beta})$$

and $V(Z_j | \mathbf{X}_{1j} = \mathbf{x}_1, \dots, \mathbf{X}_{c_jj} = \mathbf{x}_{c_j}) = c_j^{-1} \sigma^2$. The goal of this work is to estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$ based on the observed data $\{(Z_j, \mathbf{X}_{1j}, \dots, \mathbf{X}_{c_jj}) : j = 1, \dots, J\}$.

2.2. Estimation

In what follows, we propose a method to consistently estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$. If $\eta(\cdot)$ was known, then one could immediately obtain an estimate of $\boldsymbol{\beta}$ by minimising the weighted least squares objective function,

$$S\{\boldsymbol{\beta}, \eta(\cdot)\} = \sum_{j=1}^J c_j \left\{ Z_j - \frac{1}{c_j} \sum_{i=1}^{c_j} \eta(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) \right\}^2,$$

with respect to $\boldsymbol{\beta}$. A primary challenge herein is to account for the dependence between the infinite-dimensional parameter $\eta(\cdot)$ and the finite-dimensional parameter $\boldsymbol{\beta}$. To acknowledge this dependence in our notation, we write $\eta(\cdot)$ as $\eta_{\boldsymbol{\beta}}(\cdot)$; i.e. $\eta_{\boldsymbol{\beta}}(t) = E(Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t)$ for a given $\boldsymbol{\beta}$. If one can find a consistent estimator $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ of $\eta_{\boldsymbol{\beta}}(\cdot)$, then our estimator of $\boldsymbol{\beta}$ can be obtained as $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\operatorname{argmin}} S\{\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}(\cdot)\}$. When each $\mathbf{X}_{ij}^\top \boldsymbol{\beta}$ has its own response Y_{ij}

available, $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ could be obtained as the Nadaraya–Watson or the local–polynomial estimator between the Y_{ij} ’s and the $\mathbf{X}_{ij}\boldsymbol{\beta}$ ’s (see, e.g. Ichimura 1993; Cui et al. 2011). However, in our context, all the Y_{ij} ’s are latent and $\{\mathbf{X}_{ij}\}_{i=1}^{c_j}$ share the same pooled response Z_j for each j . Constructing $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ is not straightforward.

To circumvent this, we simply treat Z_j to be the response for each \mathbf{X}_{ij} and find out what is $E(Z_j | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t)$. Noting that $Z_j = c_j^{-1} \sum_{i=1}^{c_j} Y_{ij}$.

$$E(c_j Z_j | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t) = E\left(\sum_{i=1}^{c_j} Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t\right) = \sum_{l \neq i} E(Y_{lj}) + E(Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t).$$

Because Y_{ij} ’s are iid, we denote by μ the marginal expectation of Y_{ij} , i.e. $\mu = E(Y_{ij})$. Consequently, we have

$$E(c_j Z_j | \mathbf{X}_{ij}^\top \boldsymbol{\beta} = t) = (c_j - 1)\mu + \eta_{\boldsymbol{\beta}}(t). \tag{1}$$

Comparing to the case where Y_{ij} ’s are available, viewing Z_j as the response of \mathbf{X}_{ij} adds one extra intercept term in the form of $(c_j - 1)\mu$. Equation (1) inspires the construction of $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$.

We first estimate μ . Marginally, one could easily recognise that the Z_j ’s are independent variables with mean μ , so a natural estimator of μ is $\hat{\mu} = N^{-1} \sum_{j=1}^J c_j Z_j$. Then, for a given

β and t , we obtain the local linear kernel estimator $\hat{\eta}_\beta(t)$ through minimising the local least squares objective function,

$$\sum_{j=1}^J \sum_{i=1}^{c_j} \{c_j Z_j - (c_j - 1)\hat{\mu} - \eta_\beta(t) - \eta'_\beta(t)(\mathbf{X}_{ij}^\top \beta - t)\}^2 K_h(\mathbf{X}_{ij}^\top \beta - t), \tag{2}$$

with respect to $\eta_\beta(t)$ and $\eta'_\beta(t)$, where $\eta'_\beta(t)$ denotes the derivative of $\eta_\beta(t)$, $K(\cdot)$ is a kernel function, h is a user-selected bandwidth and $K_h(\cdot) = h^{-1}K(\cdot/h)$. The objective function in (2) utilises a local linear approximation that approximates $\eta(\mathbf{X}_{ij}^\top \beta)$ by $\eta(t) + (\mathbf{X}_{ij}^\top \beta - t)\eta'(t)$ at a given t . Because the accuracy of such an approximation depends on the distance between $\mathbf{X}_{ij}^\top \beta$ and t , the kernel term $K_h(\mathbf{X}_{ij}^\top \beta - t)$ weights $\mathbf{X}_{ij}^\top \beta$ more (less) if $\mathbf{X}_{ij}^\top \beta$ is close to (far away from) t . The local linear approximation became a well-accepted nonparametric regression technique due to the seminal work (Fan 1993) where the optimality of local linear smoothers was demonstrated for the nonparametric regression. One could easily extend our method to incorporate a local polynomial (with a higher order) approximation (Fan and Gijbels 1996) if $\eta_\beta(\cdot)$ is smooth enough.

It is worthwhile to point out that the minimiser $\hat{\eta}_\beta(t)$ can be expressed explicitly as

$$\hat{\eta}_\beta(t) = \frac{S_{N2}(t, \beta)\hat{T}_{N0}(t, \beta) - S_{N1}(t, \beta)\hat{T}_{N1}(t, \beta)}{S_{N0}(t, \beta)S_{N2}(t, \beta) - S_{N1}^2(t, \beta)}, \tag{3}$$

where

$$S_{Nl}(t, \beta) = N^{-1}h^{-l} \sum_{j=1}^J \sum_{i=1}^{c_j} K_h(\mathbf{X}_{ij}^\top \beta - t) \left(\mathbf{X}_{ij}^\top \beta - t\right)^l$$

and

$$\hat{T}_{Nl}(t, \beta) = N^{-1}h^{-l} \sum_{j=1}^J \sum_{i=1}^{c_j} \{c_j Z_j - (c_j - 1)\hat{\mu}\} K_h(\mathbf{X}_{ij}^\top \beta - t) \left(\mathbf{X}_{ij}^\top \beta - t\right)^l,$$

for $l \in \{0, 1, 2\}$. Our final estimators are

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} S\{\beta, \hat{\eta}_\beta(\cdot)\} \quad \text{and} \quad \hat{\eta}(\cdot) = \hat{\eta}_{\hat{\beta}}(\cdot). \tag{4}$$

Directly minimising $S\{\beta, \hat{\eta}_\beta(\cdot)\}$ in $\mathcal{B} = \{\beta = (\beta_1, \dots, \beta_p)^\top : \beta_1 > 0, \|\beta\| = 1\}$ might be numerically difficult, because \mathcal{B} is a part of the surface of the unit ball. To reduce such a computational burden, we rewrite β to be $\beta = (\sqrt{1 - \|\beta^{(1)}\|^2}, \beta^{(1)\top})^\top$, where $\beta^{(1)} = (\beta_2, \dots, \beta_p)^\top$. Consequently, the parameter space is transformed from \mathcal{B} to be $\mathcal{B}^{(1)} = \{\beta^{(1)} = (\beta_2, \dots, \beta_p)^\top : \|\beta^{(1)}\| < 1\}$; i.e. the interior of the unit ball in $\mathbb{R}^{(p-1)}$. A numerical search inside a ball of a lower dimension is easier than on the surface of a ball of a higher dimension, even though theoretically they are the same.

3. Asymptotic properties

In this section, we present the asymptotic properties of our proposed estimators. To derive these properties, we assume that the group sizes remain finite as $N \rightarrow \infty$. We view this assumption to be reasonable because, in practice, the characteristics of the assay used often bound the pool size, i.e. larger pool sizes, at a point, could adversely affect an assay's accuracy and therefore would not be employed. For example, in a study on the relationship between chemokine levels and miscarriage, the levels of monocyte chemoattractant protein-1 (MCP1) were measured using pools of size 2 (Whitcomb et al. 2007). In a BioCycle study, the F2-isoprostane level (a biomarker that measures oxidative stress) was measured in pools of size 3 (Malinovsky et al. 2012). To examine whether the pro-inflammatory cytokine interleukin-6 is a good predictor of myocardial infarction, pools of sizes 2 and 4 were used (McMahan et al. 2016). Besides practical concerns, diverging group sizes could also lead to theoretical issues. For instance, if $c_j \rightarrow \infty$ as $N \rightarrow \infty$, we have $E(Z_j | \mathbf{X}_{1j}, \dots, \mathbf{X}_{c_jj}) = c_j^{-1} \sum_{i=1}^{c_j} \eta(\mathbf{X}_{ij}^\top \boldsymbol{\beta})$ converges in probability to μ and $V(Z_j | \mathbf{X}_{1j}, \dots, \mathbf{X}_{c_jj}) = c_j^{-1} \sigma^2$ converges to zero. In other words, when c_j 's are large, all the Z_j 's become nearly the same which makes the estimation of $\eta(\cdot)$ and $\boldsymbol{\beta}$ very challenging. Hence, in this article, we focus on the scenario where c_j 's are all finite.

Because Y_{ij} 's are latent as long as $c_j > 1$, using the method of pooling increases the theoretical complexity when comparing to traditional single-index models. Equation (1) provides an idea to consistently estimate the dependence between $\eta(\cdot)$ and $\boldsymbol{\beta}$ using pooled responses. It treats Z_j as the response for each covariate \mathbf{X}_{ij} in the j th group. As a result, (Z_j, \mathbf{X}_{ij}) 's are not iid observations as $(Y_{ij}, \mathbf{X}_{ij})$'s. Furthermore, one needs to estimate the extra intercept term $(c_j - 1)\mu$ in advance. Despite these theoretical complications caused by pooling, we obtained the asymptotic properties of our estimators $\hat{\eta}(\cdot)$ and $\hat{\boldsymbol{\beta}}$. Before presenting the results, we need to introduce some notation. Because the group sizes (all positive integers) do not change with N , we denote the collection of the values of c_j by $\{c^{(m)} : m = 1, \dots, M\}$, where M is also a fixed value. More explicitly, for each j , there exists an m such that $c_j = c^{(m)}$. For each m , we let J_m denote the number of pools having size $c^{(m)}$. The ratio $J_m c^{(m)} / N$ represents the proportion of individuals that were involved in pools of size $c^{(m)}$. When $N \rightarrow \infty$, we assume that this proportion converges to $\gamma_m \in [0, 1]$, where $\sum_{m=1}^M \gamma_m = 1$. Furthermore, we denote the true parameters by $\eta_0(\cdot)$ and $\boldsymbol{\beta}_0 = (\beta_{01}, \boldsymbol{\beta}_0^{(1)\top})^\top$, where $\boldsymbol{\beta}_0^{(1)} = (\beta_{02}, \dots, \beta_{0p})^\top$. Let \mathcal{J}_0 be the value of $\partial B(\boldsymbol{\beta}^{(1)}) / \partial \boldsymbol{\beta}^{(1)}$ evaluated at $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_0^{(1)}$, where $B(\boldsymbol{\beta}^{(1)}) = (\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\top})^\top$. Moreover, denote by $\boldsymbol{\Omega}_0(\mathbf{X}) = E[\mathbf{X}\mathbf{X}^\top | \mathbf{X}^\top \boldsymbol{\beta}_0] - E[\mathbf{X} | \mathbf{X}^\top \boldsymbol{\beta}_0] E[\mathbf{X}^\top | \mathbf{X}^\top \boldsymbol{\beta}_0]$ and $\boldsymbol{\Omega} = E[\eta_0'(\mathbf{X}^\top \boldsymbol{\beta}_0) \boldsymbol{\Omega}_0(\mathbf{X})]$.

Theorem 3.1 provides the asymptotic properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}(\cdot)$. To obtain these results, we proceed under a few mild regularity conditions, which are provided in the Appendix. The proofs are provided in the Web-based Supplementary Materials.

Theorem 3.1: *Under conditions C1–C6 stated in the Appendix, as $N \rightarrow \infty$,*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma}),$$

where $\Sigma = \sigma^2(\sum_{m=1}^M \gamma_m/c^{(m)})^{-1} \mathcal{J}_0(\mathcal{J}_0^\top \Omega \mathcal{J}_0)^{-1} \mathcal{J}_0^\top$ and \xrightarrow{d} means convergence in distribution. Furthermore,

$$\sup_{\mathbf{x} \in \mathbb{X}} |\hat{\eta}(\mathbf{x}^\top \hat{\boldsymbol{\beta}}) - \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)| = O_p\{(Nh/\log N)^{-1/2}\}.$$

Theorem 3.1 reveals the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the consistency of $\hat{\eta}(\cdot)$. We would like to point out that, when c_j 's are all 1, our pooled biomarker data Z_j 's are exactly the individual-level responses Y_{ij} 's. Thus the proposed estimator is the same as the classical single-index estimator based on all individual-level data, i.e. the asymptotic normality includes $c_j = 1$ as a special case. From the asymptotic variance, we could see some patterns that might help us understand the theoretical impact of pooling. For simplicity, let us consider all the pools to be of the same size, i.e. $c^{(m)} = c$, $\gamma_m = 1$ and $N = cJ$. We see that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is $c\sigma^2 \mathcal{J}_0(\mathcal{J}_0^\top \Omega \mathcal{J}_0)^{-1} \mathcal{J}_0^\top / N$. Consequently, if the number of individuals (N) is fixed in applications, pooling more individuals in a group would lead to a loss of information and yield a larger variability in the resulting estimates of $\boldsymbol{\beta}$. If the number of individuals is not limited but the budget is limited up to J assays, we could rewrite the asymptotic variance to be $\sigma^2 \mathcal{J}_0(\mathcal{J}_0^\top \Omega \mathcal{J}_0)^{-1} \mathcal{J}_0^\top / J$ which does not depend on the pool sizes. Thus, in this case, pooling does not compromise the asymptotic efficiency of $\hat{\boldsymbol{\beta}}$.

One must note that Theorem 3.1 holds when Condition C4 (see Appendix) is satisfied, i.e. as $N \rightarrow \infty$, $h \rightarrow 0$, $Nh^4 \rightarrow \infty$ and $Nh/\log N \rightarrow \infty$. Thus it is important to select a suitable value for the bandwidth h . One could use the traditional cross-validation method. That is leaving one group of data out and fitting the model using the remaining data to predict the response that was left out. After predicting all responses, the bandwidth is chosen to be the one that minimises the sum of the squares of all the prediction errors. In other words, this traditional approach has to numerically search for an estimator of $\boldsymbol{\beta}$ when leaving each group out. When the number of groups J is large, the traditional cross-validation could cause a huge computational burden. In order to make our method more applicable in real applications, we suggest using a revised version of the traditional cross-validation method. This method was originally proposed by Härdle et al. (1993). Denote by $\hat{\eta}_{\boldsymbol{\beta}}^{(-j)}(u)$ the leave-one-out estimator of $\eta_{\boldsymbol{\beta}}(u)$ obtained via the explicit formula (3) without using the data pertaining to the j th pool. Our proposed bandwidth \tilde{h} is chosen so that $(\tilde{\boldsymbol{\beta}}, \tilde{h})$ minimises

$$S_{cv}(\boldsymbol{\beta}, h) = \sum_{j=1}^J c_j \left\{ Z_j - \frac{1}{c_j} \sum_{i=1}^{c_j} \hat{\eta}_{\boldsymbol{\beta}}^{(-j)}(X_{ij}^\top \boldsymbol{\beta}) \right\}^2.$$

Furthermore, we use the value of $\tilde{\boldsymbol{\beta}}$ as a sensible starting point to compute $\hat{\boldsymbol{\beta}}$.

4. Simulation studies

In this section, we illustrate the finite sample performance of our proposed method through simulation. Before presenting our results, we note that, to the best of our knowledge, the literature does not contain any competing methods for simultaneously estimating both $\boldsymbol{\beta}$ and $\eta(\cdot)$ based on continuous pooled assessments. McMahan et al. (2016) proposed a parametric approach to estimate $\boldsymbol{\beta}$ by assuming $\eta(\cdot)$ is linear. Therefore, besides the

main goal of illustrating of the performance of our proposed procedures under a variety of different settings, we have also compared our method with the one proposed by McMahan et al. (2016).

To illustrate that our estimation procedure does not rely on the distribution of biomarker levels, i.e. the distribution of $Y_{ij} \mid \mathbf{X}_{ij}$, we consider the following examples:

$$(D1) \quad Y \mid \mathbf{X} = \mathbf{x} \sim N\{\eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0), \sigma^2\}$$

$$(D2) \quad Y \mid \mathbf{X} = \mathbf{x} \sim \text{Gamma}\{\text{shape} = \eta_0^2(\mathbf{x}^\top \boldsymbol{\beta}_0)/\sigma^2, \text{rate} = \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)/\sigma^2\}$$

$$(D3) \quad Y \mid \mathbf{X} = \mathbf{x} \sim \text{Log-Normal}\{\mu_0(\mathbf{x}^\top \boldsymbol{\beta}_0), g_0(\mathbf{x}^\top \boldsymbol{\beta}_0)\}, \text{ where } \mu_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = \log\{\eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)\} - g_0(\mathbf{x}^\top \boldsymbol{\beta}_0)/2, \text{ and } g_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = \log\{\sigma^2/\eta_0^2(\mathbf{x}^\top \boldsymbol{\beta}_0) + 1\}.$$

The normal distribution is used to simulate symmetric biomarker data, while the other two cases are used to emulate right skewed distributions. These distributions were used in simulating biomarker levels but were not used in the part of estimation. Parameters in these distributions are chosen to satisfy our model assumption that $E(Y \mid \mathbf{X} = \mathbf{x}) = \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)$ and $V(Y \mid \mathbf{X} = \mathbf{x}) = \sigma^2$. We set $\boldsymbol{\beta}_0 = (0.5, 0.5, \sqrt{2}/2)^\top$ and $\sigma = 0.5$. For $\eta_0(\cdot)$, we consider the following four models:

$$(M1) \quad \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = \mathbf{x}^\top \boldsymbol{\beta} + 2$$

$$(M2) \quad \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = (\mathbf{x}^\top \boldsymbol{\beta}_0)^2$$

$$(M3) \quad \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = (\mathbf{x}^\top \boldsymbol{\beta}_0)^2 \exp(\mathbf{x}^\top \boldsymbol{\beta}_0)$$

$$(M4) \quad \eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0) = \sin(a\pi \mathbf{x}^\top \boldsymbol{\beta}_0) + 1, \text{ where } a = 1 \text{ or } 2.$$

One of the most attractive features of a single-index approach is that it does not force any shapes on the regression curve while being able to consistently estimate the regression coefficients. Model (M1) is chosen to be linear purposely. Through this setting, we would like to see the consequences of using our method if ignoring the truly linearity. The regression curves in (M2)–(M4) are nonlinear. They are similar to those discussed in Cui et al. (2011). Models (M2) and (M3) offer smooth curves; in contrast, Model (M4) results in an oscillating curve with the frequency being controlled by a , i.e. larger values of a result in a larger degree of oscillatory behaviour of the curve over the range of $\mathbf{x}^\top \boldsymbol{\beta}_0$. These nonlinear curves could illustrate the benefits of using our method if the regression curve is not linear. The vector of covariates $\mathbf{X} = (X_1, X_2, X_3)^\top$ contains continuous X_1 and X_2 following Uniform $(-1, 1)$ and $N(0, 0.3^2)$ distributions, respectively, and discrete X_3 with $P(X_3 = \pm 0.5) = 0.5$.

To generate pooled observations, we considered two scenarios. In the first, the number of individual specimens to be tested is fixed. Testing the specimens individually is ideal providing full information; however, this may be impractical due to the financial limitations and thus pooling is used. We chose $N \in \{2500, 5000\}$ and specified a common pool size $c_j = c$ for all $j = 1, \dots, J$, where $c \in \{1, 2, 5, 10\}$. Different values of c indicate different levels of savings. For example, $(N, c) = (2500, 5)$ means an 80% reduction in testing cost when compared to $(N, c) = (2500, 1)$ where each individual is tested separately. In this scenario, we are able to evaluate how the reduction of the number of tests would affect the accuracy of estimating $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$. For each combination of (D1)–(D3), (M1)–(M4), and $N \in \{2500, 5000\}$, we randomly generated N samples of (Y, \mathbf{X}) according to the covariate setting, the selected conditional distribution of $Y \mid \mathbf{X}$, and $\eta_0(\cdot)$. Then for

each $c \in \{1, 2, 5, 10\}$, we randomly assigned the N samples into $J = N/c$ pools and labelled them by $(Y_{ij}, \mathbf{X}_{ij})$ where $i = 1, \dots, c, j = 1, \dots, N/c$. The testing response of the j th pooled specimen was determined by $Z_j = c^{-1} \sum_{i=1}^c Y_{ij}$.

In the second scenario, the investigator may have only J assays available due to the limitation of financial resources. The choice is between testing J specimens one-by-one or testing cJ specimens using pools of size c . We considered $J \in \{250, 500\}$ and $c_j = c \in \{1, 2, 5, 10\}$. For example, $(J, c) = (250, 10)$ implies that even though there are only 250 assays, pooling could involve 10 times the number of specimens than testing individual specimens, i.e. $(J, c) = (250, 1)$. Through these settings, we are able to see whether the extra number of individuals could provide more information and thus improve the estimation accuracy. For each combination of (D1)–(D3), (M1)–(M4) and $J \in \{250, 500\}$, we randomly generated $10 \times J$ copies of (Y, \mathbf{X}) to form the specimen bank. Then for each $c \in \{1, 2, 5, 10\}$, we randomly sampled $N = cJ$ individuals from the specimen bank and assigned them to J pools. After labelling them by $(Y_{ij}, \mathbf{X}_{ij})$, where $i = 1, \dots, c, j = 1, \dots, J$, we generated the testing response of the j th pool by $Z_j = c^{-1} \sum_{i=1}^c Y_{ij}$.

Within each configuration in both scenarios, we repeated the data generating process 500 times for all considered pool sizes and applied our methodology to estimate β_0 and $\eta_0(\cdot)$. We specified the kernel function $K(\cdot)$ in (2) to be the probability density function of the standard normal distribution. The bandwidth h was selected via the leave-one-out cross-validation method described at the end of Section 3. In order to reveal the robustness of our method to the shape of a regression curve, we also fitted each data under the parametric linear assumption. The applied parametric method was from McMahan et al. (2016).

Tables 1 and 2 summarise the results for Model (M1) under all the considered distributions (D1)–(D3) when $N \in \{2500, 5000\}$ and when $J \in \{250, 500\}$, respectively. These summary statistics include the sample mean and the standard deviation of the 500 estimates of β_0 . In order to illustrate what role the pool size c plays, we use the average estimation error (AEE), defined by $AEE = \sum_{k=1}^p |\hat{\beta}_{0k} - \beta_{0k}|$, as an overall measure of the estimation accuracy for β_0 and the empirical mean squared error (MSE), calculated by $MSE = N^{-1} \sum_{j=1}^J \sum_{i=1}^c \{\hat{\eta}(\mathbf{X}_{ij}^\top \hat{\beta}) - \eta_0(\mathbf{X}_{ij}^\top \beta_0)\}^2$, to evaluate the accuracy of estimating the entire regression curve $\eta_0(\mathbf{x}^\top \beta_0)$. The sample mean of the 500 AEE's and $MSE \times 10^5$ are also included in the tables. Tables 3 and 4 summarise the same results for estimating Model (M2) under all considered distributions (D1)–(D3) when $N \in \{2500, 5000\}$ and when $J \in \{250, 500\}$, respectively. Results for Models (M3) and (M4) are similar. Hence, we include them in the Web-based Supplementary Materials.

From all the tables, one could see that our estimates of β_0 are generally on target across all models and exhibit little bias. As N or J increases, both the bias and the sample standard deviation of the estimates of β_0 decrease, so do the AEE and MSE. These patterns reinforce the consistency of $\hat{\beta}$ and $\hat{\eta}(\cdot)$, shown in Theorem 3.1. Furthermore, the overall measures (AEE and MSE) are seldom affected by (D1)–(D3). By comparing our results with the ones of McMahan et al. (2016), one could see that from Tables 1 and 2 when the curve is truly linear, both methods yield reasonable estimates of β_0 . The variability of their estimates is smaller than the one of ours. This is expected because our procedure has to estimate the curve η_0 which is given as known to their method. However, when the curve is not linear (Tables 3 and 4), their estimates suffer from a huge bias while ours are still on target. For example, in Tables 3 and 4, almost all estimates of β from the parametric method

Table 1. Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016).

| | <i>N</i> | | Proposed method | | | | Parametric method | | | | |
|------|----------|-----------------------|-----------------------|---------------|---------------|---------------|-------------------|---------------|---------------|---------------|---------------|
| | | | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | |
| (D1) | 2500 | β_{01} | 0.498 (0.028) | 0.497 (0.039) | 0.497 (0.044) | 0.497 (0.065) | 0.500 (0.018) | 0.500 (0.026) | 0.500 (0.042) | 0.500 (0.056) | |
| | | β_{02} | 0.498 (0.033) | 0.494 (0.050) | 0.498 (0.061) | 0.494 (0.088) | 0.499 (0.035) | 0.498 (0.047) | 0.500 (0.072) | 0.499 (0.096) | |
| | | β_{03} | 0.708 (0.025) | 0.710 (0.035) | 0.705 (0.039) | 0.702 (0.058) | 0.706 (0.021) | 0.707 (0.029) | 0.708 (0.045) | 0.710 (0.064) | |
| | | AEE (MSE \times 10) | 0.308 (0.011) | 0.324 (0.019) | 0.340 (0.009) | 0.372 (0.017) | 0.058 (0.004) | 0.080 (0.007) | 0.128 (0.017) | 0.172 (0.030) | |
| | 5000 | β_{01} | 0.501 (0.012) | 0.498 (0.031) | 0.502 (0.029) | 0.499 (0.050) | 0.500 (0.012) | 0.500 (0.018) | 0.499 (0.027) | 0.498 (0.040) | |
| | | β_{02} | 0.501 (0.019) | 0.499 (0.036) | 0.498 (0.042) | 0.497 (0.064) | 0.500 (0.025) | 0.501 (0.033) | 0.503 (0.052) | 0.503 (0.077) | |
| | | β_{03} | 0.706 (0.013) | 0.707 (0.027) | 0.705 (0.026) | 0.704 (0.043) | 0.707 (0.014) | 0.708 (0.020) | 0.708 (0.031) | 0.707 (0.043) | |
| | | AEE (MSE \times 10) | 0.294 (0.002) | 0.305 (0.007) | 0.317 (0.003) | 0.336 (0.008) | 0.041 (0.002) | 0.056 (0.004) | 0.089 (0.008) | 0.128 (0.016) | |
| | (D2) | 2500 | β_{01} | 0.495 (0.040) | 0.500 (0.025) | 0.496 (0.046) | 0.492 (0.069) | 0.499 (0.018) | 0.497 (0.024) | 0.497 (0.039) | 0.498 (0.056) |
| | | | β_{02} | 0.498 (0.044) | 0.500 (0.038) | 0.496 (0.065) | 0.485 (0.099) | 0.501 (0.034) | 0.502 (0.049) | 0.498 (0.074) | 0.500 (0.111) |
| | | | β_{03} | 0.709 (0.033) | 0.706 (0.022) | 0.707 (0.043) | 0.710 (0.054) | 0.708 (0.021) | 0.708 (0.028) | 0.709 (0.043) | 0.711 (0.062) |
| | | | AEE (MSE \times 10) | 0.309 (0.022) | 0.315 (0.005) | 0.347 (0.008) | 0.386 (0.020) | 0.058 (0.004) | 0.080 (0.007) | 0.125 (0.016) | 0.182 (0.032) |
| 5000 | | β_{01} | 0.500 (0.013) | 0.497 (0.032) | 0.500 (0.029) | 0.497 (0.046) | 0.501 (0.013) | 0.500 (0.017) | 0.499 (0.027) | 0.498 (0.037) | |
| | | β_{02} | 0.500 (0.020) | 0.498 (0.038) | 0.499 (0.046) | 0.501 (0.066) | 0.499 (0.023) | 0.499 (0.035) | 0.497 (0.052) | 0.499 (0.074) | |
| | | β_{03} | 0.707 (0.012) | 0.708 (0.027) | 0.705 (0.029) | 0.703 (0.042) | 0.707 (0.015) | 0.708 (0.020) | 0.705 (0.032) | 0.703 (0.044) | |
| | | AEE (MSE \times 10) | 0.296 (0.002) | 0.308 (0.012) | 0.320 (0.004) | 0.339 (0.006) | 0.041 (0.002) | 0.057 (0.004) | 0.090 (0.008) | 0.123 (0.015) | |
| (D3) | | 2500 | β_{01} | 0.498 (0.029) | 0.500 (0.033) | 0.497 (0.048) | 0.496 (0.065) | 0.499 (0.016) | 0.499 (0.024) | 0.499 (0.039) | 0.501 (0.053) |
| | | | β_{02} | 0.498 (0.034) | 0.498 (0.042) | 0.496 (0.069) | 0.492 (0.099) | 0.499 (0.032) | 0.499 (0.046) | 0.504 (0.074) | 0.509 (0.110) |
| | | | β_{03} | 0.708 (0.024) | 0.707 (0.027) | 0.706 (0.047) | 0.702 (0.064) | 0.706 (0.020) | 0.706 (0.029) | 0.706 (0.046) | 0.702 (0.065) |
| | | | AEE (MSE \times 10) | 0.306 (0.009) | 0.316 (0.005) | 0.344 (0.015) | 0.372 (0.019) | 0.056 (0.004) | 0.079 (0.007) | 0.127 (0.016) | 0.181 (0.032) |
| | 5000 | β_{01} | 0.500 (0.012) | 0.499 (0.023) | 0.500 (0.028) | 0.496 (0.053) | 0.500 (0.012) | 0.500 (0.018) | 0.500 (0.028) | 0.500 (0.040) | |
| | | β_{02} | 0.499 (0.018) | 0.500 (0.030) | 0.498 (0.045) | 0.495 (0.071) | 0.499 (0.023) | 0.501 (0.034) | 0.502 (0.053) | 0.502 (0.075) | |
| | | β_{03} | 0.707 (0.012) | 0.707 (0.021) | 0.706 (0.029) | 0.706 (0.047) | 0.707 (0.014) | 0.706 (0.019) | 0.707 (0.031) | 0.707 (0.046) | |
| | | AEE (MSE \times 10) | 0.293 (0.002) | 0.303 (0.007) | 0.323 (0.004) | 0.341 (0.006) | 0.039 (0.002) | 0.056 (0.004) | 0.090 (0.008) | 0.129 (0.016) | |

Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\beta_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE \times 10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.



Table 2. Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016).

| <i>J</i> | Proposed method | | | | Parametric method | | | | | | |
|----------------|-----------------|----------------|---------------|---------------|-------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | | | |
| (D1) | 250 | β_{01} | 0.497 (0.064) | 0.497 (0.064) | 0.497 (0.062) | 0.499 (0.069) | 0.499 (0.054) | 0.501 (0.056) | 0.501 (0.057) | 0.501 (0.054) | |
| | | β_{02} | 0.490 (0.091) | 0.497 (0.087) | 0.498 (0.090) | 0.490 (0.099) | 0.501 (0.111) | 0.504 (0.105) | 0.501 (0.110) | 0.503 (0.110) | |
| | | β_{03} | 0.705 (0.059) | 0.700 (0.058) | 0.699 (0.059) | 0.701 (0.063) | 0.709 (0.065) | 0.710 (0.063) | 0.711 (0.062) | 0.712 (0.063) | |
| | | AEE (MSE × 10) | 0.372 (0.085) | 0.362 (0.033) | 0.366 (0.018) | 0.365 (0.017) | 0.184 (0.041) | 0.179 (0.070) | 0.181 (0.166) | 0.182 (0.316) | |
| | 500 | β_{01} | 0.495 (0.055) | 0.495 (0.050) | 0.494 (0.049) | 0.496 (0.045) | 0.501 (0.037) | 0.501 (0.037) | 0.501 (0.039) | 0.501 (0.038) | |
| | | β_{02} | 0.497 (0.069) | 0.496 (0.067) | 0.497 (0.069) | 0.498 (0.068) | 0.500 (0.076) | 0.498 (0.079) | 0.504 (0.073) | 0.498 (0.077) | |
| | | β_{03} | 0.706 (0.052) | 0.707 (0.046) | 0.706 (0.046) | 0.705 (0.042) | 0.707 (0.044) | 0.705 (0.044) | 0.705 (0.043) | 0.710 (0.044) | |
| | | AEE (MSE × 10) | 0.349 (0.039) | 0.347 (0.029) | 0.345 (0.011) | 0.340 (0.005) | 0.126 (0.019) | 0.128 (0.035) | 0.124 (0.077) | 0.127 (0.156) | |
| | (D2) | 250 | β_{01} | 0.496 (0.066) | 0.498 (0.063) | 0.496 (0.063) | 0.497 (0.065) | 0.495 (0.054) | 0.504 (0.056) | 0.501 (0.056) | 0.502 (0.054) |
| | | | β_{02} | 0.492 (0.096) | 0.498 (0.089) | 0.497 (0.095) | 0.491 (0.092) | 0.495 (0.106) | 0.493 (0.110) | 0.492 (0.111) | 0.505 (0.113) |
| β_{03} | | | 0.703 (0.060) | 0.699 (0.059) | 0.700 (0.060) | 0.704 (0.061) | 0.710 (0.066) | 0.710 (0.064) | 0.704 (0.063) | 0.708 (0.064) | |
| AEE (MSE × 10) | | | 0.367 (0.075) | 0.365 (0.028) | 0.376 (0.019) | 0.370 (0.015) | 0.181 (0.040) | 0.182 (0.072) | 0.183 (0.166) | 0.185 (0.325) | |
| 500 | | β_{01} | 0.499 (0.048) | 0.495 (0.053) | 0.497 (0.047) | 0.498 (0.048) | 0.499 (0.040) | 0.502 (0.038) | 0.496 (0.039) | 0.502 (0.038) | |
| | | β_{02} | 0.491 (0.067) | 0.492 (0.067) | 0.496 (0.071) | 0.494 (0.069) | 0.500 (0.075) | 0.504 (0.077) | 0.503 (0.076) | 0.501 (0.075) | |
| | | β_{03} | 0.708 (0.047) | 0.710 (0.048) | 0.706 (0.046) | 0.707 (0.043) | 0.706 (0.044) | 0.707 (0.047) | 0.706 (0.046) | 0.706 (0.043) | |
| | | AEE (MSE × 10) | 0.347 (0.042) | 0.351 (0.041) | 0.350 (0.041) | 0.347 (0.033) | 0.127 (0.020) | 0.129 (0.036) | 0.127 (0.084) | 0.124 (0.150) | |
| (D3) | | 250 | β_{01} | 0.502 (0.062) | 0.500 (0.061) | 0.493 (0.065) | 0.492 (0.064) | 0.500 (0.054) | 0.499 (0.057) | 0.500 (0.054) | 0.505 (0.054) |
| | | | β_{02} | 0.494 (0.092) | 0.491 (0.095) | 0.488 (0.095) | 0.490 (0.095) | 0.500 (0.102) | 0.502 (0.111) | 0.507 (0.106) | 0.498 (0.105) |
| | β_{03} | | 0.699 (0.056) | 0.702 (0.059) | 0.709 (0.061) | 0.708 (0.061) | 0.706 (0.064) | 0.706 (0.062) | 0.713 (0.059) | 0.707 (0.065) | |
| | AEE (MSE × 10) | | 0.367 (0.072) | 0.374 (0.029) | 0.377 (0.020) | 0.376 (0.015) | 0.174 (0.040) | 0.183 (0.074) | 0.175 (0.152) | 0.181 (0.313) | |
| | 500 | β_{01} | 0.497 (0.052) | 0.499 (0.042) | 0.500 (0.041) | 0.498 (0.049) | 0.499 (0.040) | 0.501 (0.038) | 0.498 (0.040) | 0.500 (0.038) | |
| | | β_{02} | 0.491 (0.066) | 0.501 (0.062) | 0.503 (0.062) | 0.495 (0.068) | 0.499 (0.077) | 0.495 (0.076) | 0.500 (0.071) | 0.501 (0.073) | |
| | | β_{03} | 0.709 (0.047) | 0.702 (0.041) | 0.699 (0.039) | 0.706 (0.044) | 0.706 (0.046) | 0.707 (0.046) | 0.708 (0.045) | 0.706 (0.044) | |
| | | AEE (MSE × 10) | 0.348 (0.041) | 0.335 (0.012) | 0.330 (0.005) | 0.342 (0.008) | 0.131 (0.021) | 0.128 (0.035) | 0.124 (0.079) | 0.123 (0.148) | |

Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\beta_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE × 10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

Table 3. Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016).

| | <i>N</i> | | Proposed method | | | | Parametric method | | | |
|-----------------------|----------|-----------------------|-----------------|---------------|---------------|---------------|-------------------|----------------|----------------|----------------|
| | | | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 |
| (D1) | 2500 | β_{01} | 0.500 (0.041) | 0.502 (0.056) | 0.523 (0.080) | 0.535 (0.107) | 0.000 (0.022) | 0.000 (0.029) | -0.001 (0.046) | 0.000 (0.062) |
| | | β_{02} | 0.496 (0.045) | 0.496 (0.063) | 0.504 (0.085) | 0.502 (0.119) | -0.001 (0.041) | -0.002 (0.057) | 0.001 (0.085) | 0.000 (0.115) |
| | | β_{03} | 0.706 (0.049) | 0.701 (0.063) | 0.673 (0.079) | 0.653 (0.096) | 0.001 (0.022) | 0.002 (0.030) | 0.005 (0.052) | 0.007 (0.071) |
| | | AEE (MSE \times 10) | 0.331 (0.017) | 0.356 (0.021) | 0.363 (0.026) | 0.386 (0.045) | 1.707 (0.672) | 1.707 (0.676) | 1.702 (0.688) | 1.700 (0.705) |
| | 5000 | β_{01} | 0.501 (0.026) | 0.503 (0.039) | 0.506 (0.059) | 0.525 (0.077) | 0.001 (0.016) | 0.000 (0.020) | -0.001 (0.032) | -0.002 (0.043) |
| | | β_{02} | 0.501 (0.030) | 0.501 (0.043) | 0.507 (0.065) | 0.506 (0.088) | 0.000 (0.029) | 0.001 (0.039) | 0.001 (0.059) | -0.003 (0.085) |
| | | β_{03} | 0.704 (0.032) | 0.700 (0.045) | 0.689 (0.066) | 0.669 (0.080) | 0.000 (0.017) | -0.001 (0.023) | -0.001 (0.036) | -0.001 (0.050) |
| AEE (MSE \times 10) | | 0.310 (0.008) | 0.325 (0.010) | 0.343 (0.015) | 0.366 (0.020) | 1.706 (0.671) | 1.707 (0.673) | 1.709 (0.679) | 1.713 (0.688) | |
| (D2) | 2500 | β_{01} | 0.503 (0.031) | 0.504 (0.046) | 0.512 (0.074) | 0.524 (0.097) | -0.001 (0.021) | -0.001 (0.028) | -0.002 (0.044) | -0.002 (0.063) |
| | | β_{02} | 0.503 (0.036) | 0.506 (0.054) | 0.508 (0.083) | 0.517 (0.107) | -0.003 (0.037) | 0.001 (0.050) | 0.000 (0.079) | 0.003 (0.114) |
| | | β_{03} | 0.700 (0.035) | 0.695 (0.053) | 0.679 (0.080) | 0.654 (0.097) | 0.001 (0.024) | 0.000 (0.032) | -0.001 (0.046) | -0.002 (0.071) |
| | | AEE (MSE \times 10) | 0.314 (0.014) | 0.332 (0.017) | 0.359 (0.023) | 0.382 (0.032) | 1.703 (0.671) | 1.708 (0.674) | 1.710 (0.684) | 1.709 (0.704) |
| | 5000 | β_{01} | 0.502 (0.025) | 0.503 (0.036) | 0.511 (0.054) | 0.521 (0.075) | 0.001 (0.019) | 0.002 (0.033) | 0.000 (0.052) | 0.000 (0.054) |
| | | β_{02} | 0.503 (0.028) | 0.502 (0.041) | 0.505 (0.060) | 0.512 (0.084) | 0.001 (0.028) | 0.002 (0.041) | 0.001 (0.062) | 0.004 (0.085) |
| | | β_{03} | 0.702 (0.029) | 0.700 (0.044) | 0.689 (0.057) | 0.669 (0.076) | -0.001 (0.018) | -0.002 (0.022) | 0.000 (0.048) | 0.000 (0.065) |
| AEE (MSE \times 10) | | 0.308 (0.007) | 0.322 (0.009) | 0.336 (0.013) | 0.347 (0.019) | 1.706 (0.672) | 1.705 (0.676) | 1.708 (0.687) | 1.705 (0.696) | |
| (D3) | 2500 | β_{01} | 0.501 (0.040) | 0.504 (0.054) | 0.518 (0.077) | 0.542 (0.103) | 0.001 (0.020) | 0.000 (0.028) | 0.001 (0.045) | 0.000 (0.062) |
| | | β_{02} | 0.498 (0.046) | 0.499 (0.061) | 0.506 (0.087) | 0.507 (0.121) | -0.003 (0.041) | -0.007 (0.055) | -0.005 (0.085) | -0.004 (0.115) |
| | | β_{03} | 0.703 (0.047) | 0.698 (0.061) | 0.675 (0.079) | 0.645 (0.093) | 0.000 (0.022) | 0.000 (0.032) | 0.000 (0.050) | 0.000 (0.070) |
| | | AEE (MSE \times 10) | 0.331 (0.016) | 0.352 (0.019) | 0.360 (0.026) | 0.372 (0.036) | 1.708 (0.673) | 1.713 (0.677) | 1.711 (0.689) | 1.711 (0.707) |
| | 5000 | β_{01} | 0.501 (0.028) | 0.503 (0.041) | 0.507 (0.060) | 0.522 (0.083) | 0.001 (0.015) | 0.001 (0.021) | 0.000 (0.031) | 0.001 (0.044) |
| | | β_{02} | 0.500 (0.032) | 0.500 (0.043) | 0.498 (0.064) | 0.503 (0.089) | 0.001 (0.029) | 0.001 (0.039) | 0.001 (0.059) | -0.001 (0.082) |
| | | β_{03} | 0.704 (0.034) | 0.701 (0.047) | 0.695 (0.066) | 0.673 (0.079) | 0.001 (0.016) | 0.000 (0.022) | 0.002 (0.035) | 0.003 (0.051) |
| AEE (MSE \times 10) | | 0.314 (0.008) | 0.326 (0.010) | 0.347 (0.015) | 0.365 (0.024) | 1.705 (0.671) | 1.705 (0.673) | 1.704 (0.678) | 1.705 (0.688) | |

Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\beta_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE \times 10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.



Table 4. Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016).

| <i>J</i> | Proposed method | | | | Parametric method | | | | | |
|----------|-----------------|----------------|---------------|---------------|-------------------|---------------|----------------|----------------|----------------|----------------|
| | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | <i>c</i> = 1 | <i>c</i> = 2 | <i>c</i> = 5 | <i>c</i> = 10 | | |
| (D1) | 250 | β_{01} | 0.509 (0.069) | 0.508 (0.080) | 0.517 (0.079) | 0.512 (0.077) | -0.001 (0.067) | -0.001 (0.066) | -0.001 (0.065) | -0.002 (0.062) |
| | | β_{02} | 0.500 (0.081) | 0.503 (0.083) | 0.502 (0.085) | 0.510 (0.084) | -0.006 (0.127) | -0.003 (0.125) | 0.007 (0.122) | -0.004 (0.121) |
| | | β_{03} | 0.689 (0.073) | 0.685 (0.080) | 0.679 (0.081) | 0.678 (0.078) | -0.002 (0.069) | -0.004 (0.070) | 0.001 (0.075) | -0.001 (0.071) |
| | 500 | AEE (MSE × 10) | 0.456 (0.067) | 0.468 (0.045) | 0.442 (0.027) | 0.441 (0.021) | 1.716 (0.695) | 1.716 (1.399) | 1.701 (3.535) | 1.714 (7.057) |
| | | β_{01} | 0.513 (0.058) | 0.511 (0.072) | 0.511 (0.076) | 0.512 (0.069) | -0.002 (0.047) | 0.000 (0.045) | -0.003 (0.043) | -0.001 (0.041) |
| | | β_{02} | 0.505 (0.074) | 0.497 (0.077) | 0.500 (0.078) | 0.509 (0.081) | -0.005 (0.087) | 0.000 (0.086) | -0.008 (0.081) | -0.005 (0.088) |
| (D2) | 250 | β_{03} | 0.685 (0.056) | 0.690 (0.067) | 0.687 (0.074) | 0.680 (0.071) | 0.002 (0.051) | 0.000 (0.052) | 0.004 (0.050) | -0.002 (0.051) |
| | | AEE (MSE × 10) | 0.401 (0.059) | 0.422 (0.037) | 0.404 (0.024) | 0.402 (0.017) | 1.712 (0.688) | 1.707 (1.376) | 1.713 (3.436) | 1.716 (6.887) |
| | | β_{01} | 0.517 (0.112) | 0.510 (0.122) | 0.524 (0.126) | 0.518 (0.122) | 0.004 (0.058) | 0.004 (0.061) | 0.000 (0.061) | 0.000 (0.060) |
| | 500 | β_{02} | 0.489 (0.127) | 0.482 (0.128) | 0.498 (0.128) | 0.507 (0.129) | -0.002 (0.129) | -0.007 (0.114) | 0.006 (0.132) | -0.015 (0.113) |
| | | β_{03} | 0.673 (0.109) | 0.680 (0.115) | 0.657 (0.117) | 0.657 (0.111) | -0.004 (0.069) | 0.003 (0.072) | -0.003 (0.071) | 0.001 (0.068) |
| | | AEE (MSE × 10) | 0.600 (0.187) | 0.490 (0.136) | 0.448 (0.103) | 0.422 (0.098) | 1.719 (0.698) | 1.709 (1.414) | 1.712 (3.542) | 1.704 (7.097) |
| (D3) | 250 | β_{01} | 0.505 (0.076) | 0.507 (0.084) | 0.515 (0.082) | 0.528 (0.080) | 0.003 (0.041) | -0.002 (0.044) | 0.000 (0.043) | -0.001 (0.041) |
| | | β_{02} | 0.497 (0.093) | 0.499 (0.087) | 0.500 (0.092) | 0.501 (0.089) | 0.000 (0.084) | -0.006 (0.093) | -0.001 (0.077) | 0.000 (0.084) |
| | | β_{03} | 0.690 (0.085) | 0.687 (0.082) | 0.680 (0.084) | 0.671 (0.079) | -0.001 (0.046) | 0.001 (0.050) | 0.003 (0.052) | 0.001 (0.048) |
| | 500 | AEE (MSE × 10) | 0.546 (0.088) | 0.548 (0.048) | 0.409 (0.040) | 0.400 (0.022) | 1.705 (0.679) | 1.714 (1.372) | 1.705 (3.432) | 1.707 (6.861) |
| | | β_{01} | 0.524 (0.114) | 0.512 (0.119) | 0.515 (0.118) | 0.526 (0.121) | 0.000 (0.065) | 0.001 (0.063) | 0.000 (0.061) | 0.004 (0.063) |
| | | β_{02} | 0.500 (0.120) | 0.497 (0.116) | 0.488 (0.124) | 0.500 (0.130) | -0.008 (0.129) | -0.005 (0.127) | -0.010 (0.122) | 0.000 (0.124) |
| (D3) | 250 | β_{03} | 0.660 (0.108) | 0.672 (0.106) | 0.675 (0.109) | 0.656 (0.107) | -0.004 (0.070) | 0.003 (0.075) | 0.005 (0.070) | -0.001 (0.072) |
| | | AEE (MSE × 10) | 0.436 (0.160) | 0.447 (0.120) | 0.455 (0.097) | 0.436 (0.095) | 1.719 (0.698) | 1.709 (1.414) | 1.712 (3.542) | 1.704 (7.097) |
| | | β_{01} | 0.529 (0.084) | 0.508 (0.094) | 0.515 (0.101) | 0.520 (0.103) | 0.000 (0.047) | -0.002 (0.046) | -0.002 (0.045) | 0.000 (0.043) |
| | 500 | β_{02} | 0.501 (0.099) | 0.504 (0.103) | 0.494 (0.114) | 0.507 (0.114) | 0.003 (0.092) | 0.005 (0.087) | -0.003 (0.091) | 0.002 (0.086) |
| | | β_{03} | 0.669 (0.061) | 0.679 (0.084) | 0.678 (0.091) | 0.664 (0.095) | 0.003 (0.051) | -0.001 (0.054) | -0.002 (0.051) | 0.002 (0.051) |
| | | AEE (MSE × 10) | 0.405 (0.134) | 0.414 (0.074) | 0.406 (0.068) | 0.389 (0.065) | 1.702 (0.681) | 1.705 (1.380) | 1.714 (3.448) | 1.704 (6.891) |

Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\beta_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE × 10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

are centred around 0. If inferences were made based on these estimates, one would incorrectly conclude that all the covariates are insignificant, i.e. a wrongly assumed curve would greatly compromise statistical inferences. However, such concerns do not exist if using our method. To sum up, all the aforementioned observations demonstrate that our estimators are robust to the biomarker distribution and the shape of the regression curve.

Now we look at the impact of pool sizes. When the number of individual N is fixed (Tables 1 and 3), as one might expect, all the standard deviations increase with the pool size c . The loss in estimation efficiency is the price paid for the significant cost reduction realised by pooling. In terms of estimating the entire mean curve $\eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)$, one could see that the MSE's only increase a little when c increases. For example, in Table 3 for (D1) and $N = 2500$, the MSE changes from 0.0017 to 0.0021 when c_j increases from 1 to 2. Note that $c = 2$ represents a 50% saving in cost when comparing to $c = 1$. These results suggest that pooling could provide estimates similar to or not much worse than those obtained from individual testing while conferring a significant cost reduction.

Tables 2 and 4 correspond to the second scenario where the number of assays J is fixed. The results reinforce our findings from Theorem 3.1 which are that the pool size c does not affect the efficiency of estimating $\boldsymbol{\beta}_0$ across different pool sizes when J is fixed. For example, in Table 4 for (D2) and $J = 500$, when $c = 1$, the standard deviations of estimates of $(\beta_{01}, \beta_{02}, \beta_{03})$ are (0.076, 0.094, 0.085) which change to (0.080, 0.089, 0.079) when $c = 10$, respectively. As an overall measure, AEE actually decreases from 0.600 (when $c = 1$) to 0.422 (when $c = 10$). Furthermore, the MSE strictly decreases with c . These patterns indicate that measuring biomarkers on pools will provide nearly the same or even more precise estimates on both $\boldsymbol{\beta}_0$ and $\eta_0(\mathbf{x}^\top \boldsymbol{\beta}_0)$ when compared to individual testing.

Lastly, we consider the case where $V(Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta})$ depends on covariates. We set $V(Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta}) = (0.5\mathbf{X}_{ij}^\top \boldsymbol{\beta})^2$ and repeated the whole simulation study described above. Because the patterns of these results are similar, we present them in the Web-based Supplementary Materials. One conclusion is that our method also performs well if $V(Y_{ij} | \mathbf{X}_{ij}^\top \boldsymbol{\beta})$ changes with covariates.

5. Real data analysis

5.1. NHANES diabetes data

We first illustrate our proposed methodology by applying it to a diabetes data set obtained from NHANES available at https://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx. The data consists of a continuous response variable for each individual, Y , which denotes a patient's 2-hour plasma glucose level concentration (mg/dL), which has been identified as a viable biomarker for detecting diabetes mellitus. In addition, a set of explanatory variables are considered; namely, X_1 gender, X_2 age in month, X_3 the log body mass index (kg/m^2), X_4 systolic blood pressure (mm Hg), X_5 diastolic blood pressure (mm Hg), X_6 the log fasting plasma glucose level (mg/dL), X_7 the log triglycerides level (mg/dL) and X_8 the log HDL-cholesterol level (mg/dL), so that the covariate vector $\mathbf{X} = (X_1, \dots, X_8)^\top$ for each individual. This data set contains $N = 2574$ individual observations with 2318 of them having all of the explanatory variables listed above. In this section, we analyse the diabetes data set of $N = 2318$ individuals with full covariate information. It is important to notice that instead of analysing actual pooled testing data, it is more advantageous to

artificially construct pooled responses using individual level data, because it allows us to investigate the effect that pool size and composition (in terms of the covariates) has on parameter estimation.

The first focus of our analysis is to compare our pool response model to the analogous model in which the individual level data is fully observed. To accomplish this, we randomly assigned individuals to pools of size c , where $c \in \{2, \dots, 10\}$. Note that the sample size $N = 2318$ cannot be divided by some values of c ; in such cases, we pool the remainders as the last group (e.g. when $c = 10$, the pool response data consists of 231 pools of size 10 and 1 of size 8). Pooling responses for the pools were then determined according to $Z_j = c^{-1} \sum_{i=1}^c Y_{ij}$. We repeated the above procedure 500 times and applied our proposed model to each of the resulting pooled data sets. We standardised the continuous covariates so that they had mean 0 and variance 1, while the discrete binary covariates were recoded to be -0.5 or 0.5 , respectively.

Figure 1 presents box plots of the 500 estimates of β obtained from our method across $c \in \{2, 3, \dots, 10\}$. Also included in the figure are quantile plots of the estimates of $\eta_0(t)$ for pool sizes of $c = 1, 2, 5, 10$. For purpose of comparison, we use the $c = 1$ case as a reference by which our estimates can be compared. Note that the reference estimates suggest a nonlinear shape of $\eta_0(\cdot)$ which supports the use of our single-index model. From these results, it is apparent that the estimates of β_0 are largely in agreement with the estimates based on the individual-level data. This can also be said for our estimates of $\eta_0(t)$ across all considered pool sizes. We again observe that the variability in our parameter estimates tends to increase with the pool size, which is expected due to the significant cost reduction. Additionally, one will note that our estimates of $\eta_0(t)$ exhibit evidence of instability toward the upper bound of $X^T \hat{\beta}$ for larger pool size (e.g. $c = 10$). Again this is an expected phenomenon, since the number of observations that occur in that region is relatively small.

The second primary focus is to assess the effect of pooling when the number of assays J is fixed. For this purpose, we set $J = 232$ and consider $c \in \{1, 2, \dots, 10\}$. The pool response data for each c was constructed by randomly sampling cJ specimens from the 2318 individuals and assigning them to pools of size c . Once the pools have been established, we determine the testing response for the j th pool by $Z_j = c^{-1} \sum_{i=1}^c Y_{ij}$. Again, we repeated the procedure 500 times and applied our proposed method to those data sets. Figure 2 presents box plots of the 500 estimates of β_0 's across $c \in \{1, 2, \dots, 10\}$ and quantile plots of the estimates of $\eta_0(t)$ for $c = 1, 2, 5, 10$. It can be seen that the estimates of β_0 and $\eta_0(\cdot)$ generally agree with the reference estimates (obtained when $N = 2318$ and $c = 1$). Furthermore, the box plots are nearly the same across all pool sizes. The variability of estimates of $\eta_0(t)$ when $c > 2$ is smaller than the one when $c = 1$; e.g. the width of the 95% quantile bands when $c = 10$ is apparently smaller than the one when $c = 1$. These results reinforce the main findings of the second scenario in Section 4.

5.2. CPP pooled chemokine data

We now illustrate the proposed methodology using a pooled data. This data was collected from the CPP, a study conducted from 1957 to 1974 to assess various aspects of maternal and child health (e.g. see Whitcomb et al. 2007). In 2007, stored serum samples from CPP participants were measured for levels of many chemokines to study

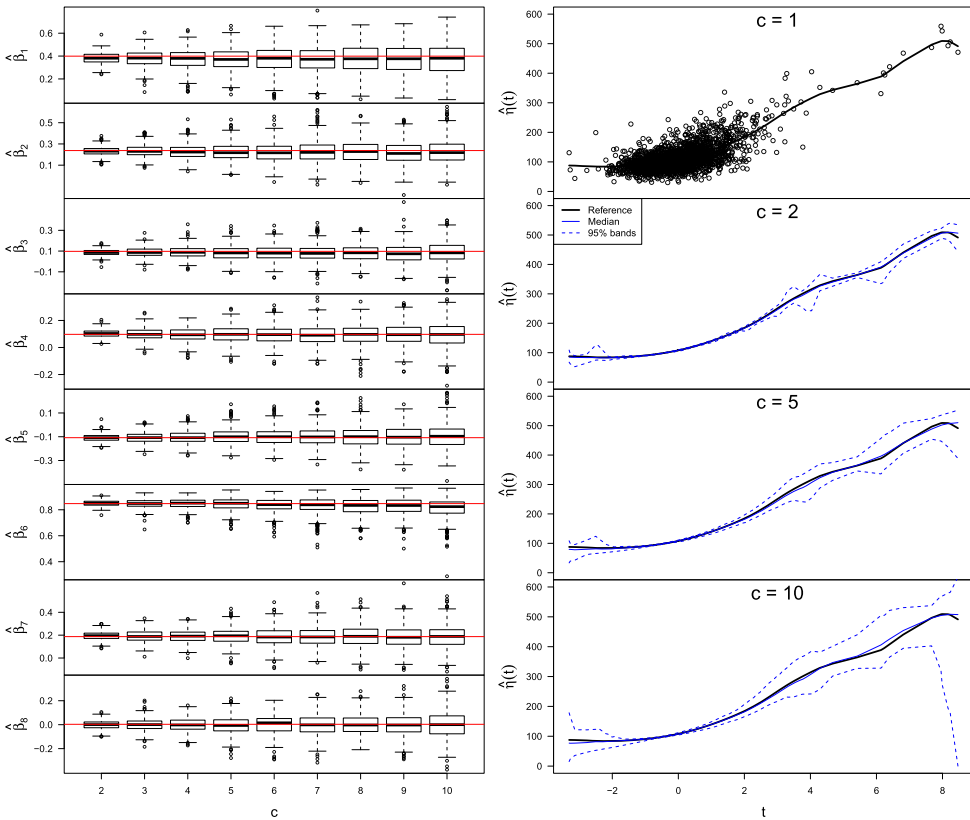


Figure 1. Left: Box plots of the 500 estimates of $\beta = (\beta_{01}, \dots, \beta_{08})^\top$ across $c \in \{2, \dots, 10\}$. Right: The points in the top figure denote the patient’s 2-hour plasma glucose level. The remaining three figures depict the estimate curve of $\eta(t)$ and the quantile plots of the estimates of $\eta(t)$ for $c \in \{2, 5, 10\}$. Specifically, at every value of t we plot the 2.5th, 50th and 97.5th percentiles of the 500 estimates of $\eta(t)$. The solid lines in the figures denote the estimates of $\eta(t)$, when $N = 2318$ and $c = 1$.

whether these biomarkers are related to miscarriage risk. In this article, we focus on the biomarker macrophage inhibitory protein (MIP)-1 α which was measured in pools of size $c = 2$. We consider only the pools with participants whose full covariate information were available. Considered covariates include age (standardised; x_1), race (1 = African-American/0 = otherwise; x_2), and miscarriage status (1 = yes/0 = no; x_3). After removing missing values, the number of pools is $J = 330$. Our goal is to apply our single-index technique to the pooled measurements so that one can estimate the individual-level mean trend of the MIP-1 α given the covariate information.

Applying our methodology yields a bandwidth $h = 0.692$ and estimates of the regression coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^\top = (0.703, 0.700, 0.127)^\top$. The estimated mean curve $\hat{\eta}(t)$ is plotted (the black line) in Figure 3. In order to obtain valid inference, we adopted a bootstrapping method. A general description of this bootstrapping method is presented in the Web-based Supplementary Materials, where a simulation studied is also included to illustrate its performance. We bootstrapped the pooled data for 500 times. On each bootstrap sample, we applied our methodology and obtained 500 bootstrap estimates of $(\beta, \eta(\cdot))$. The

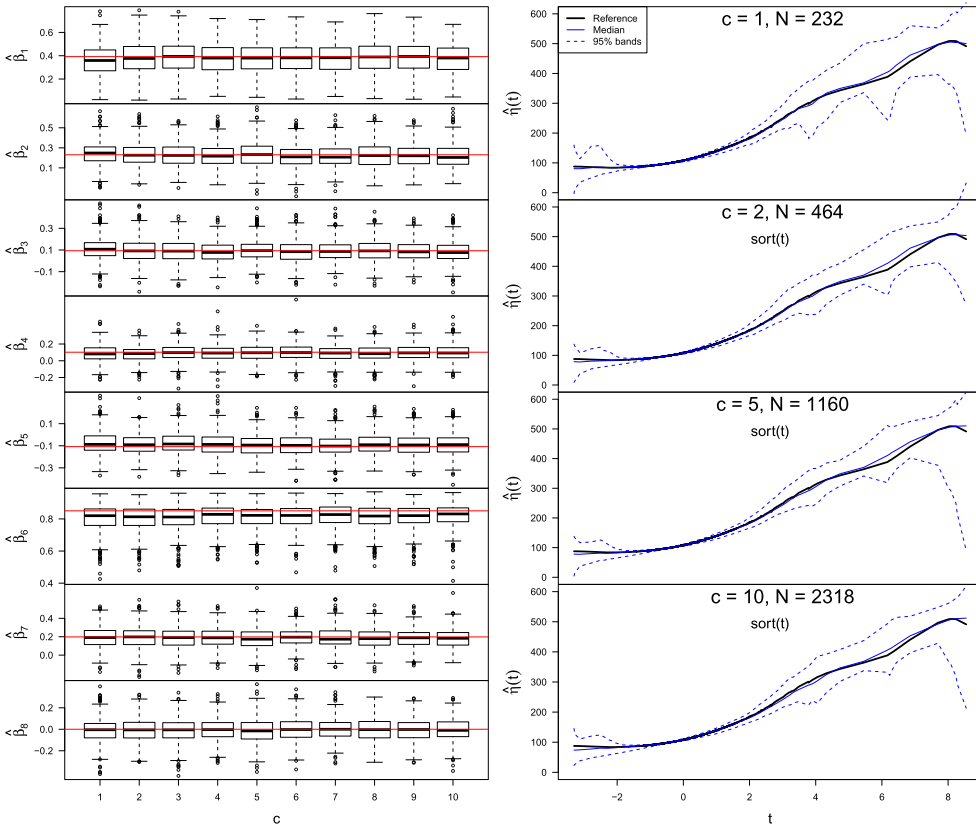


Figure 2. Left: Box plots of the 500 estimates of $\beta = (\beta_{01}, \dots, \beta_{08})^\top$ across $c \in \{1, 2, \dots, 10\}$ when J is fixed to be 232. Right: The fourth figures depict the estimate curve of $\eta(t)$ and the quantile plots of the estimates of $\eta(t)$ for $c \in \{1, 2, 5, 10\}$ when J is fixed to be 232. Specifically, at every value of t we plot the 2.5th, 50th and 97.5th percentiles of the 500 estimates of $\eta(t)$. The solid lines in the figures denote the estimates of $\eta(t)$, when $N = 2318$ and $c = 1$.

standard deviation of these bootstrap estimates of β can be used to estimate the standard error of our point estimates. The resulting estimated standard errors are $SE(\hat{\beta}_1) = 0.178$, $SE(\hat{\beta}_2) = 0.480$ and $SE(\hat{\beta}_1) = 0.368$, which suggest that at least age has a significant impact on the individual’s MIP-1 α mean level. Pointwise quantile plots (2.5th, 50th and 97.5th percentiles) of the 500 bootstrap estimates of $\eta(\cdot)$ are also included in Figure 3. Clearly, one can see a nonlinear mean relationship between the linear predictor ($t = \mathbf{x}^\top \hat{\beta}$) and the MIP-1 α level. This nonlinear relationship further demonstrates the contribution and the flexibility of our proposed single-index methodology.

6. Discussion

In spite of the wide and lasting interest in pooling strategy under restrictive parametric assumptions, nonparametric or semiparametric estimation based on continuous pooled biomarker data received relatively less attention. In this article, we have proposed a general semiparametric framework for modelling pooled biomarker data allowing for the

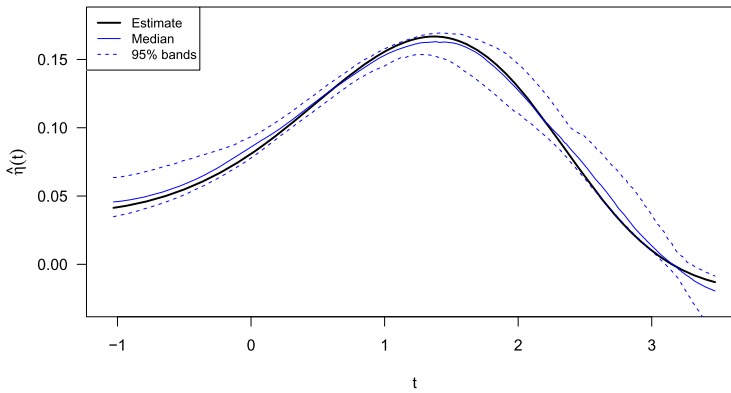


Figure 3. CPP pooled MIP-1 α data: This figure includes the estimate curve of $\eta(t)$ and the quantile plots of the 500 bootstrap estimates of $\eta(t)$ based on bootstrap samples. Specifically, at every value of t we plot the 2.5th, 50th and 97.5th percentiles of the 500 bootstrap estimates of $\eta(t)$.

incorporation of individual covariates. Compared to existing works (Ma et al. 2011; Malinovsky et al. 2012; McMahan et al. 2016), our approach does not force the regression function to be linear nor the type of biomarker distribution to be known. We have shown that our estimates enjoy nice asymptotic properties. To illustrate the performance of our methodology, we have considered two scenarios. In the first, the population size is fixed. Our numerical studies suggest that pooling could reduce the cost substantially with only a minor loss of estimation accuracy. In the second, the number of assays is fixed. We found out that the pooling strategy could be superior providing more information than testing specimens separately. Our estimates performed well under either symmetric or right-skewed biomarker distribution settings.

Because pooling biomarker is now more common in practical applications (see Lyles et al. 2015; Mitchell, Lyles, Manatunga, and Schisterman 2015; Perrier, Giorgis-Allemand, Slama, and Philippat 2016), we believe it would be very beneficial to develop more statistical methods that are flexible to model such data. In this work, we assumed that pools are constructed by randomly mixing individual specimens. One interesting future work is to consider the situation where pooling is performed within stratification of population on the basis of some demographic variables, such as age or gender, which might potentially improve the estimation performance. Caudill (2010) and Mitchell et al. (2014) adopted such grouping criteria to characterise population and analyse biomarker data, respectively. Another interesting extension of our work is to incorporate pooled exposures as a part of the covariate information, which is a more complex problem and received many attention recently (Linton and Whang 2002; Whitcomb et al. 2012; Delaigle and Zhou 2015).

Acknowledgments

We would like to thank an Associate Editor and two anonymous referees for their constructive suggestions that have greatly improved the presentation of this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Dewei Wang  <http://orcid.org/0000-0003-0822-8563>

References

- Bondell, H.D., Liu, A., and Schisterman, E.F. (2007), 'Statistical Inference Based on Pooled Data: A Moment-Based Estimating Equation Approach', *Journal of Applied Statistics*, 34, 129–140.
- Caudill, S.P. (2010), 'Characterizing Populations of Individuals Using Pooled Samples', *Journal of Exposure Science and Environmental Epidemiology*, 20, 29–37.
- Cui, X., Härdle, W., and Zhu, L. (2011), 'The EFM Approach for Single-Index Models', *The Annals of Statistics*, 39, 1658–1688.
- Delaigle, A., and Hall, P. (2012), 'Nonparametric Regression with Homogeneous Group Testing Data', *The Annals of Statistics*, 40, 131–158.
- Delaigle, A., Hall, P., and Wishart, J.R. (2014), 'New Approaches to Nonparametric and Semiparametric Regression for Univariate and Multivariate Group Testing Data', *Biometrika*, 101, 567–585.
- Delaigle, A., and Meister, A. (2011), 'Nonparametric Regression Analysis for Group Testing Data', *Journal of the American Statistical Association*, 106, 640–650.
- Delaigle, A., and Zhou, W.X. (2015), 'Nonparametric and Parametric Estimators of Prevalence from Group Testing Data with Aggregated Covariates', *Journal of the American Statistical Association*, 110, 1785–1796.
- Dhand, N.K., Johnson, W.O., and Toribio, J.L. (2010), 'A Bayesian Approach to Estimate OJD Prevalence from Pooled Fecal Samples of Variable Pool Size', *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 452–473.
- Dodd, R.Y., Notari, E.P., and Stramer, S.L. (2002), 'Current Prevalence and Incidence of Infectious Disease Markers and Estimated Window-Period Risk in the American Red Cross Blood Donor Population', *Transfusion*, 42, 975–979.
- Dorfman, R. (1943), 'The Detection of Defective Members of Large Populations', *The Annals of Mathematical Statistics*, 14, 436–440.
- Fan, J. (1993), 'Local Linear Regression Smoothers and their Minimax Efficiencies', *The Annals of Statistics*, 21, 196–216.
- Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.
- Faraggi, D., Reiser, B., and Schisterman, E.F. (2003), 'ROC Curve Analysis for Biomarkers Based on Pooled Assessments', *Statistics in Medicine*, 22, 2515–2527.
- Farrington, C.P. (1992), 'Estimating Prevalence by Group Testing Using Generalized Linear Models', *Statistics in Medicine*, 11, 1591–1597.
- Gastwirth, J. L. (2000), 'The Efficiency of Pooling in the Detection of Rare Mutations', *American Journal of Human Genetics*, 67, 1036–1039.
- Härdle, W., Hall, P., and Ichimura, H. (1993), 'Optimal Smoothing in Single-Index Models', *The Annals of Statistics*, 21, 157–178.
- Ichimura, H. (1993), 'Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models', *Journal of Econometrics*, 58, 71–120.
- Klein, R.W., and Spady, R.H. (1993), 'An Efficient Semiparametric Estimator for Binary Response Models', *Econometrica*, 61, 387–421.
- Lin, W., and Kulasekera, K.B. (2007), 'Identifiability of Single-Index Models and Additive-Index Models', *Biometrika*, 94, 496–501.
- Linton, O., and Whang, Y.J. (2002), 'Nonparametric Estimation with Aggregated Data', *Econometric Theory*, 18, 420–468.
- Liu, A., and Schisterman, E.F. (2003), 'Comparison of Diagnostic Accuracy of Biomarkers with Pooled Assessments', *Biometrical Journal*, 45, 631–644.
- Lyles, R.H., Van Domelen, D., Mitchell, E.M., and Schisterman, E.F. (2015), 'A Discriminant Function Approach to Adjust for Processing and Measurement Error When a Biomarker Is Assayed

- in Pooled Samples', *International Journal of Environmental Research and Public Health*, 12, 14723–14740.
- Ma, C.X., Vexler, A., Schisterman, E.F., and Tian, L. (2011), 'Cost-Efficient Designs Based on Linearly Associated Biomarkers', *Journal of Applied Statistics*, 38, 2739–2750.
- Malinovsky, Y., Albert, P.S., and Schisterman, E.F. (2012), 'Pooling Designs for Outcomes under a Gaussian Random Effects Model', *Biometrics*, 68, 45–52.
- McMahan, C.S., McLain, A.C., Gallagher, C.M., and Schisterman, E.F. (2016), 'Estimating Covariate-Adjusted Measures of Diagnostic Accuracy Based on Pooled Biomarker Assessments', *Biometrical Journal*, 58, 944–961.
- Mitchell, E.M., Lyles, R.H., Manatunga, A.K., Danaher, M., Perkins, N.J., and Schisterman, E.F. (2014), 'Regression for Skewed Biomarker Outcomes Subject to Pooling', *Biometrics*, 70, 202–211.
- Mitchell, E.M., Lyles, R.H., Manatunga, A.K., and Schisterman, E.F. (2015), 'Semiparametric Regression Models for a Right-Skewed Outcome Subject to Pooling', *American Journal of Epidemiology*, 181, 541–548.
- Mumford, S.L., Schisterman, E.F., Vexler, A., and Liu, A. (2006), 'Pooling Biospecimens and Limits of Detection: Effects on ROC Curve Analysis', *Biostatistics*, 7, 585–598.
- Perrier, F., Giorgis-Allemand, L., Slama, R., and Philippat, C. (2016), 'Within-Subject Pooling of Biological Samples to Reduce Exposure Misclassification in Biomarker-based Studies', *Epidemiology*, 27, 378–388.
- Remlinger, K.S., Hughes-Oliver, J.M., Young, S.S., and Lam, R.L. (2006), 'Statistical Design of Pools Using Optimal Coverage and Minimal Collision', *Technometrics*, 48, 133–143.
- Schisterman, E., Faraggi, D., Reiser, B., and Trevisan, M. (2001), 'Statistical Inference for the Area under the Receiver Operating Characteristic Curve in the Presence of Random Measurement Error', *Annals of Epidemiology*, 154, 174–179.
- Schisterman, E.F., Vexler, A., Yi, A., and Perkins, N.J. (2011), 'A Combined Efficient Design for Biomarker Data Subject to a Limit of Detection Due to Measuring Instrument Sensitivity', *The Annals of Applied Statistics*, 5, 2651–2667.
- Thompson, K.H. (1962), 'Estimation of the Proportion of Vectors in a Natural Population of Insects', *Biometrics*, 18, 568–578.
- Van, T.T., Miller, J., Warshauer, D.M., Reisdorf, E., Jerrigan, D., Humes, R., and Shult, P.A. (2012), 'Pooling Nasopharyngeal/Throat Swab Specimens to Increase Testing Capacity for Influenza Viruses by PCR', *Journal of Clinical Microbiology*, 50, 891–896.
- Vansteelandt, E., Goetghebeur, E., and Verstraeten, T. (2000), 'Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples', *Biometrics*, 56, 1126–1133.
- Vexler, A., Liu, A., and Schisterman, E.F. (2006), 'Efficient Design and Analysis of Biospecimens with Measurements Subject to Detection Limit', *Biometrical Journal*, 48, 780–791.
- Vexler, A., Schisterman, E.F., and Liu, A. (2008), 'Estimation of ROC Curves Based on Stably Distributed Biomarkers Subject to Measurement Error and Pooling Mixtures', *Statistics in Medicine*, 27, 280–296.
- Wang, D., McMahan, C.S., and Gallagher, C.M. (2015), 'A General Parametric Regression Framework for Group Testing Data with Dilution Effects', *Statistics in Medicine*, 34, 3606–3621.
- Wang, D., McMahan, C.S., Gallagher, C.M., and Kulasekera, K.B. (2014), 'Semiparametric Group Testing Regression Models', *Biometrika*, 101, 587–598.
- Wang, J., Xue, L., Zhu, L., and Chong, Y.S. (2010), 'Estimation for a Partial-Linear Single-Index Model', *The Annals of Statistics*, 38, 246–274.
- Wang, D., Zhou, H., and Kulasekera, K.B. (2013), 'A Semi-Local Likelihood Regression Estimator of the Proportion Based on Group Testing Data', *Journal of Nonparametric Statistics*, 25, 209–221.
- Weinberg, C.R., and Umbach, D.M. (1999), 'Using Pooled Exposure Assessment to Improve Efficiency in Case–Control Studies', *Biometrics*, 55, 718–726.
- Whitcomb, B.W., Perkins, N.J., Zhang, Z., Ye, A., and Lyles, R.H. (2012), 'Assessment of Skewed Exposure in Case–Control Studies with Pooling', *Statistics in Medicine*, 31, 2461–2472.
- Whitcomb, B.W., Schisterman, E.F., Klebanoff, M.A., Baumgarten, M., Rhoton-Vlasak, A., Luo, X., and Chegini, N. (2007), 'Circulating Chemokine Levels and Miscarriage', *American Journal of Epidemiology*, 166, 323–331.

- Xia, C. (2006), 'Asymptotic Distributions for Two Estimators of the Single-Index Model', *Econometric Theory*, 22, 1112–1137.
- Xia, Y., Tong, H., Li, W.K., and Zhu, L. (2002), 'An Adaptive Estimation of Dimension Reduction Space', *Journal of the Royal Statistical Society: Series B*, 64, 363–410.
- Xie, M. (2001), 'Regression Analysis of Group Testing Samples', *Statistics in Medicine*, 20, 1957–1969.
- Zhu, L., and Xue, L. (2006), 'Empirical Likelihood Confidence Regions in a Partially Linear Single-Index Model', *Journal of the Royal Statistical Society B*, 68, 549–570.

Appendix. Regularity conditions

We provide the mild regularity conditions under which the theorem in Section 3 holds. These conditions are common in the single-index literature.

- (C1) The curves $d_{\beta}(t) = E(X | X^{\top} \beta = t)$ and $\eta_{\beta}(t)$ have bounded and continuous second-order derivatives.
- (C2) The probability density function of $X^{\top} \beta$ is bounded away from zero and satisfies the Lipschitz condition of order 1 on $\{t = x^{\top} \beta, x \in \mathbb{X}\}$.
- (C3) As $N \rightarrow \infty$, $h \rightarrow 0$, $Nh^4 \rightarrow \infty$ and $Nh / \log N \rightarrow \infty$.
- (C4) $K(\cdot)$ is a bounded and symmetric kernel function with bounded first derivative.
- (C5) Conditional on X , Y has a finite fourth moment.
- (C6) The equation $u^{\top} \Omega u = 0$ has the unique root $u = \beta_0$ in \mathcal{B} .

Conditions C1–C4 are common smoothness assumptions (Xia 2006; Wang et al. 2010; Cui et al. 2011). The Lipschitz condition in C2 allows for the discrete components in the covariates. Condition C5 is similar to the one used in Wang et al. (2010). Condition C6 assures that the matrix $\mathcal{J}_0^{\top} \Omega_a \mathcal{J}_0$ is positive definite.