Computational Statistics (2019) 34:653–682
https://doi.org/10.1007/s00180-018-0838-3

ORIGINAL PAPER

Binary surrogates with stratified samples when weights are unknown

Yu-Min Huang[1]

Received: 9 March 2017 / Accepted: 8 September 2018 / Published online: 19 September 2018

Abstract

In clinical practice, surrogate variables are commonly used as an indirect measure when it is difficult or expensive to measure the primary outcome variable X, based on which the disease status is assessed. In this article, we consider the problem of constructing an optimal binary surrogate Y to substitute such the feature variable X. To retain samples that have rare values in X, the paired sample (X, Y) is usually selected based on stratified sampling, where the strata are constructed using the disjoint intervals with the support of X. For such a sampling design, the stratum proportions are usually unknown such that proportional allocation is infeasible and (X, Y)'s cannot be regarded as an i.i.d. sample between strata. We estimate the unknown cutoff determining higher/lower levels of X that optimally match the variable Y and provide the true positive rates (TPR) adjusted for the disproportionate stratum weights. Our approach is to estimate the underlying distribution of X, then conduct an ad-hoc estimation for the TPR and for the expected prediction errors under zero-one loss function. We develop parametric estimate of the distribution of X under exponential family assumption and a weighted-kernel density estimator when the distribution of X is unspecified. We illustrate our methods on various simulation studies and on a real example where binary surrogates were evaluated for a medical device. The simulation results indicate that our approach performs well.

Keywords Surrogate variable · Biased sampling · Logistic model · Binary classification · Composite likelihood · Kernel density · Optimal cutoff values

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-018-0838-3) contains supplementary material, which is available to authorized users.

✉ Yu-Min Huang
ymhuang0211@gmail.com

[1] Department of Statistics, Tunghai University, Taichung, Taiwan

1 Introduction

In many clinical practices, surrogate variables are commonly used as an indirect measure when it is difficult or expensive to measure the primary outcome variable X, based on which the disease status is assessed. In particular, it is desirable to have a binary surrogate Y, providing positive/negative signs for some disease status. In this article, we consider the problem of constructing an optimal binary surrogate Y, which allow us to predict the disease status. In clinical studies, in order to obtain samples from individuals with rare values of X, the stratified sampling scheme is often used. Specifically, the population is first divided into relevant strata using the disjoint intervals with the support of X, i.e., the support of the real-valued input variable X is partitioned into semi-closed intervals, say $J_s = [b_{s-1}, b_s)$, $b_0 < b_1 \cdots < b_S$, and from each semi-close interval (strata), a paired random sample (X_{si}, Y_{si}), $i = 1, \ldots, n_s$; $s = 1, \ldots, S$ with values restricted in the semi-close interval J_s is randomly drawn. Since the stratum proportions are usually unknown, the proportional allocation is infeasible, i.e., the sample sizes are allowed to be disproportionate to the interval probabilities $P(X \in J_s)$. Thus, under such a sampling scheme, (X_{si}, Y_{si})'s can not be regarded as an i.i.d. sample. Under this disproportional sampling scheme, we shall construct a binary surrogate Y using an estimated cutoff value (threshold) of X. We assume that the surrogate variable Y is monotonically positively associated to the input variable X through a monotone increasing link function $\phi(\cdot)$ such that $P(Y = 1|x) = \phi(x)$. Thus, we are seeking the function value $v(c) = 1$ for $X \geq c$. Accordingly, the optimal relation is subject to finding the optimal c, such that $P(v(c)|X \geq c)$ is optimized. Although such sampling scheme seems to be uncommonly utilized, we have motivating examples and a few more other examples, in that such sampling way may be considered.

Example A: the motivating example for our study

This example has also served as the real example in Sect. 5. The present paper is motivated by a clinical study conducted by a biotech company for developing a rapid test kit, by which, diabetic patients can simply read positive/negative outcomes to indicate whether their glycated haemoglobin (HbA1c or A1c) levels have exceeded a critical point. Normally the measurement of such medical value has to be taken via blood tests with much higher cost and time, therefore the goal is to determine accurate cut-off point serving a threshold value, optimally dividing the +/- revealed on the new test kit. The medical regulation requires the condition that the new test kit should be applicable to both healthy persons and patents with specified precisions for the whole range of A1c values. Due to the cost and insufficient samples sizes for both healthy persons and patients, the medical investigator took random samples from different ranges of the A1c values.

Example B: endpoints example

From literature, we also found an interesting example regarding endpoints in clinical trials. For the evaluation of new treatments in clinical trials, a conventional endpoint

Table 1 Endpoints examples of Table 1 in Buyse et al. (2016)

$Z=0$		$Z=1$				
		Scenario A: Full capture		Scenario Independence		
S	T	S	T	S	T	Prognosis
0	10	0	10	0	5	Poor
0	10	0	10	0	15	
0	10	1	20	1	5	Intermediate
1	20	1	20	1	15	
1	20	1	20	1	40	Good

could be costly or difficult to measure. Thereby, a cheaper and more convenient surrogate endpoint is expected to predict the clinical outcomes. As data abbreviated in Table 1, an example shown in Buyse et al. (2016). The surrogate endpoint S is a binary outcome, with $S = 0$ denoting failure and $S = 1$ success. The true endpoint T is a continuous outcome. The patients are classified by prognosis types: Poor, Intermediate and Good. In such an example, imagine that if the prognosis types are grouped according to a continuous variable (variable reflect the prognosis), by the values on scales, and samples are taken by groups, then we also face the pooled samples which are not IIDs between those groups. According to our simulation studies, matching the continuous endpoint T to the binary surrogate Z with such stratified samples would also encounter the same concerns of biased true positive rates as we had in our original motivation example.

Example C: ecological data

Another potential application may occur in the construction of spatial distribution. As in the paper by Ferrier et al. (2002), a fundamental challenge faced by conservation action is to manage biodiversity. A way to proceed is to use the existence of information on the spatial distribution of biodiversity. To have better information cover the whole range of spatial areas, conducting stratified sample by some different ranges of continuous scores would be applicable and reduce the cost. For instance, for large ecological regions, samples may be drawn by different levels of lapse rates due to the mountain altitudes.

From the above motivation examples, we summarize our study motivation that when specific sample sizes are demanded for different range of a continuous measurements and the distribution is unknown, then such stratification would provide a sampling scheme with an effective cost. We can utilize the first stage of our approach to estimate the distribution of the variable. And for matching the measurements to one binary surrogate, we proposed to further utilize our two-stage method.

One way is to model the optimal cutoff by establishing cutpoint models as proposed by Lausen and Schumacher (1996), Contal and O'Quigley (1999) and da Silva and Klein (2011). However, this method requires modification since (X, Y)'s cannot be

regarded as an i.i.d. sample between strata. On the other hand, in view of the conditional probability $P(v(c)|X \geq c)$, if by fitting the conditional probability with a logistic model, we have biased estimates of TPRs as illustrated in Example 1. The resultant bias is also due to fitting the model while neglecting the unknown stratum weights. Our approach is based on the minimization of the expected *zero-one* loss prediction error (EPE) (Friedman et al. 2001) possessing the following form

$$EPE(v) = E_X E_{Y|X} \left[|Y - v(X)| \, |X\right]. \tag{1}$$

For a given threshold value c, when the expectation is evaluated with a continuous probability density of X, we write it as

$$EPE(v) = P(Y = 0|X \geq c)P(X \geq c) + P(Y = 1|X < c)P(X < c). \tag{2}$$

The classification problem then is equivalent to maximizing the total correct probability (TCP), namely to get

$$\max_c TCP = \max_c \left[P(Y = 1, X \geq c) + P(Y = 0, X < c)\right]. \tag{3}$$

To determine the optimal c based on TCP, we need to estimate the distribution of X, which requires the estimation of the unknown stratum proportion $P(X \in J_s))$, such that the selection bias can be adjusted (Heckman 1979; Zadrozny 2004; Cortes et al. 2008; Richards et al. 2012; Liu and Ziebart 2014). Selection biased problems similar to ours are also investigated as a special case in the work of Vardi (1985), Gill et al. (1988), Gilbert (2000) and Fokianos (2004). In particular, Gill et al. (1988) pointed out that such a biased sampling depending on the partition of the support is not identifiable and there is no way to estimate it using the empirical distribution unless the weights of bias can be extracted from other sources, whereas commonly the biased sampling is referred to sampling the data without the concordance of the original probability weights in the population. They therefore propose to use an enlarged data set by including another extra proportion of the data drawn with the support of X. For inference dealing with similarly biased samples, Wang and Sun (2009) utilized observations from overlapped intervals (i.e., recaptures) from a finite population.

Directly applying non-parametric estimation to TCP, $EPE(v)$, such as simply counting the relative frequencies to estimate these probabilities is not adequate either. We propose to solve the binary classification first by estimating the underlying true distribution for the input variable X, then by estimating the $EPE(v)$ with an ad-hoc approach. The optimal threshold value is then obtained by attaining the smallest $EPE(v)$. We utilize parametric and non-parametric approaches to build the distribution of X, where we assume that the values of X are continuously distributed and have continuous density $f(x)$. Parametrically, let $f(x)$ be specified by $f_{\boldsymbol{\theta}}(x)$, a density up to an unknown parameter $\boldsymbol{\theta}$. We first investigate the distribution in exponential family, well-known to accommodate quite many choices of popular distributions undertaken in practice. For an exponential family, we use composite-type (Varin et al. 2011) likelihood function conditioning on strata (Godambe 1976; Lindsay 1982). Following the results shown by Lindsay (1988) that under unbiased score functions, the

asymptotic properties of estimates solved from the composite likelihood function are asymptotic equivalent to those from a full likelihood function. To handle an unknown density, which may inherently have a multi-modal feature, we build the density using a weighted-type kernel density with the weights adaptively sought by the data. We connect the densities under strata J_s, denoted by $f_{J_s}(x), s = 1, \ldots, S$, side-by-side on the edges of the stratum and apply normalization to obtain the estimation of true density $f(x)$ defined on the whole support. For a such non-parametrically fitting, choices of optimal bandwidth are also under consideration with values searched from minimizing the integrated squared error with cross-validation.

The rest of the article is organized as follows. In Sect. 2, we derive the parametric and non-parametric estimators for the underlying distribution of X when the distribution of X belongs to an exponential family and is unspecified. In Sect. 3, the form for ad-hoc estimation of the expected zero-one loss prediction error is presented and the asymptotic variance of EPE with the fitted parametric distribution of X is described. We present simulation studies in Sect. 4, where we find that the rebuilt distributions are quite close to the truth and the estimated EPE appears reasonable compared with the results generated by proportionately stratified samples, which relatively we set as benchmarks. In Sect. 5, we illustrate the proposed method using a real example.

2 Parametric distribution estimation with stratification

2.1 Parametric distributions from exponential families

We assume that X has a continuous distribution belonging to a p-dimensional exponential family. With respect to some common probability measure μ, the density of X has the form

$$f_{\boldsymbol{\theta}}(x) = \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(x) - A(\eta(\boldsymbol{\theta}))\}h(x), \tag{4}$$

where $\boldsymbol{\eta}$ is the vector of natural parameters and $\mathbf{T}(x) = (T_1(x), \ldots, T_p(x))$ is the vector of corresponding sufficient statistics. For finitely fixed $S \geq 3$, the partition of the sample space is a collection of left-closed intervals $J_s = [b_{s-1}, b_s)$, $b_0 < b_1 \cdots < b_S$. Being subject to the range within each stratum J_s, a random sample of size n_s is collected. We denote the total size as $N = \sum_{s=1}^{S} n_s$.

Let Q_s be the probability that the value of X falls in stratum J_s. While conditioning on stratum J_s, the probability density of one X_{si} is

$$\frac{f_{\boldsymbol{\theta}}(x_{si})}{Q_s} = \frac{\exp\{\boldsymbol{\eta}^T \mathbf{T}(x)\}h(x)}{\int_{J_s} \exp\{\boldsymbol{\eta}^T \mathbf{T}(x)\}h(x)dx} = \frac{\bar{f}_{\boldsymbol{\theta}}(x_{si})}{\bar{Q}_s}, \tag{5}$$

where $\bar{f}_{\boldsymbol{\theta}}(x_{si})$ denotes $\exp\{\boldsymbol{\eta}^T \mathbf{T}(x)\}h(x)$ and $\boldsymbol{\eta}$ simplifies $\boldsymbol{\eta}(\boldsymbol{\theta})$. However, there is no guarantee that the data crossing the stratum are independent, and therefore no particular explicit form to express the full likelihood via a joint probability. But, given on strata, these stratified samples x_{si} are conditionally independent [as the same assumptions

hold in Example 1.3a by Gill et al. (1988) and Section 5.2 by Wu (1997)]. We exploit the composite likelihood function approach that is a likelihood-type object obtained by multiplying a collection of marginal likelihoods (Lindsay 1988; Cox and Reid 2004). Originally, this likelihood-type object was facilitated for spatial data collected over split sites or for the aim of reducing computational burden. While considering the conditioning on stratum $J_s, s = 1, \ldots, S$, we notice that the data is conditionally independent. By multiplying all conditional likelihoods across the strata, the composite likelihood is $L(\boldsymbol{\theta}|\mathbf{x}) = \Pi_{s=1}^{S} \Pi_{i=1}^{n_s} \left[\bar{f_{\boldsymbol{\theta}}}(x_{si}) \Big/ \bar{Q}_s \right]$. For a single data point x_{si} in J_s, the log of likelihood of the parameter $\boldsymbol{\eta}$ and the score function with respective to $\boldsymbol{\eta}$ is

$$l(\boldsymbol{\eta}|x_{si}) = \ln\left(\frac{f_{\boldsymbol{\theta}}(x_{si})}{Q_s}\right) = \boldsymbol{\eta}^T \mathbf{T}(x_{si}) + \ln(h(x_{si})) - \ln\left[\int_{J_s} e^{\boldsymbol{\eta}^T \mathbf{T}(x_{si})} h(x) dx\right], \tag{6}$$

$$\mathbf{u}(\eta_j, x_{si}) = \frac{\partial l(\boldsymbol{\eta}|x_{si})}{\partial \eta_j} = T_j(x_{si}) - E_{J_s}\left[T_j(x_{si})\right], \tag{7}$$

where $\mathbf{u}(\eta_j, x_{si})$ is the score function of the *component log-likelihood* at a data point x_{si}. Given observations in stratum J_s, the conditional log likelihood function of $\boldsymbol{\eta}$ is

$$l(\boldsymbol{\eta}|\mathbf{x}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} l(\boldsymbol{\eta}|x_{si}) = \sum_{s=1}^{S} l^{(s)}(\boldsymbol{\eta}|\mathbf{x}_s), \quad \mathbf{x}_s = (x_{s,1}, \ldots, x_{s,n_s}). \tag{8}$$

After taking differentiation with respective to $\boldsymbol{\eta}$, we have the following *composite score functions*:

$$\frac{l(\boldsymbol{\eta}|\mathbf{x})}{\partial \eta_j} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \mathbf{u}(\eta_j, x_{si}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \frac{\partial l(\boldsymbol{\eta}|x_{si})}{\partial \eta_j} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \left[T_j(x_{si}) - E_{J_s}\left[T_j(x_{si})\right]\right]. \tag{9}$$

We also denote the composite score function of η_j given $\mathbf{x}$ by $\mathbf{u}(\eta_j, \mathbf{x})$ and the composite score function of $\boldsymbol{\eta}$ by $\mathbf{u}(\boldsymbol{\eta}, \mathbf{x})$. If the interest is to make inference about θ, further taking differentiation regarding one component of $\boldsymbol{\theta}$, say θ_k, leads to

$$\frac{l(\boldsymbol{\eta}|\mathbf{x})}{\partial \eta_j} \frac{\partial \eta_j}{\partial \theta_k} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \frac{\partial l(\boldsymbol{\eta}|x_{si})}{\partial \eta_j} \frac{\partial \eta_j}{\partial \theta_k} = \left(\sum_{s=1}^{S} \sum_{i=1}^{n_s} \left[T_j(x_{si}) - E_{J_s}\left[T_j(x_{si})\right]\right]\right) \frac{\partial \eta_j}{\partial \theta_k}. \tag{10}$$

The *composite maximum likelihood estimates* for $\boldsymbol{\eta}$ are solved by equating the composite score functions to zero. For computing the score functions, we need to evaluate the expectation $E_{J_s}\left[T_j(x_{si})\right]$, which is an integral with respect to the probability measure under the stratified interval and therefore sometime intractable. A convenient way is to implement software to obtain the numerical integral, such as using the function `integrate()` in R. When the probability density is not too complicated, usually the numerical integral can be obtained. Here we also propose Monte Carlo likelihood estimation steps as another way to evaluate the composite score functions:

- *1. Monte Carlo E-step* Given the current parametric values $\boldsymbol{\theta}_0$, with Monte Carlo size m_s, simulate $X_{s,1}, \ldots, X_{s,m_s}$ from $f_{\boldsymbol{\theta}}$ with values restricted in disjoint stratified intervals J_s. Compute the Monte Carlo estimate $\hat{E}_{J_s}^{MC}\left[T_j(x_{si})\right]$ by averaging the Monte Carlo samples of $T_j(x_{si})$.
- *2. Maximization step* Use the current estimate $\hat{E}_{J_s}^{MC}\left[T_j(x_{si})\right]$ as the expectation in the composite scores (10). Equate the equations to zero to obtain the updated parameter values $\boldsymbol{\theta}_1$. Then we repeat the iteration for M times.

The computation steps with Monte Carlo can be regarded as an implment of Monte Carlo EM algorithm proposed by Wei and Tanner (1990), and the iterationes achieve convergence to the MLE under certain conditions for $M \to \infty$. Geyer (1991, 1994); Beskos et al. (2009). However, the Monte Carlo EM algorithm does not anymore monotonically increase the likelihood due to the extra randomness introduced by the simulations in the Monte Carlo E-step. Nevertheless, after a long iteration of M times, the solutions will be oscillate within a small range that the changes of likelihood behaves sufficiently small stochatstically (Chan and Ledolter 1995). We stop the iteration when such small changes δ occurs. The Monte Carlo likelihood estimation is its easy implement of integration, however simulation errors may be of concerned.

2.2 Asymptotic properties

Composite likelihoods may be seen as misspecified likelihoods. However, as addressed by Lindsay (1988), under regularity conditions on the component log-densities, i.e. the first properties that only require unbiased component scores and finite covariance matrix, we can conclude with consistent properties for the composite maximum likelihood estimator under our sampling scheme.

Theorem 1 *Consider the stratified sample* $\mathbf{X} = (X_{si}), s = 1, \ldots, S, i = 1, \ldots, n_s$, *conditionally independent on stratum* $s = 1, \ldots, S$ *with the density function (5). Then, the Kullback Leibler information inequality holds for each component log likelihood and hence for the composite log likelihood, i.e.,*

$$E_{\boldsymbol{\eta}_0}\left[\mathbf{u}(\boldsymbol{\eta}, x_{si})\right] \leq E_{\boldsymbol{\eta}_0}\left[\mathbf{u}(\boldsymbol{\eta}_0, x_{si})\right] \Rightarrow \sup_{\boldsymbol{\eta}} E_{\boldsymbol{\eta}_0}\left[L(\boldsymbol{\eta}|x_{si})\right] = E_{\boldsymbol{\eta}_0}\left[L(\boldsymbol{\eta}_0|x_{si})\right], \quad (11)$$

where $\boldsymbol{\eta}_0$ *is the true value of the parameter. Hence the composite MLE is consistent.*

Proof It is straightforward to see that the regularity conditions hold for each component log-density in (6). The asymptotic results follow the first properties addressed by Lindsay (1988). □

Given the strata, the component score functions are conditionally independent across the strata. In the following, we derive weak convergent properties for the composite MLE based on the conditional independent component score functions.

Theorem 2 *Let $(X_{si}), s = 1, \ldots, S, i = 1, \ldots, n_s$ be random samples from stratum J_s respectively. Let $\mathbf{u}(\boldsymbol{\eta}, x_{si})$ denote the component score of $\boldsymbol{\eta}$ at one observation x_{si} obtained at stratum J_s and suppose that each $\mathbf{u}(\boldsymbol{\eta}, x_{si})$ has a $p \times p$ non-singular covariance V_s. There exists $\mathbf{t} = (t_1, \ldots, t_s) \in [0, 1]^s$ that for $n_s = Nt_s$, we have*

$$\frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \mathbf{u}(\boldsymbol{\eta}, x_{si}) \xrightarrow{\mathcal{L}} \sum_{s=1}^{S} V_s^{1/2} \mathbb{W}_s(t_s), \tag{12}$$

where $\mathbb{W}_s(t)$ are independent Wiener processes.

Proof Apply Donsker's theorem on the scores at each stratum, the asymptotic results follow with similar derivations in Theorem 1 of Martsynyuk (2012) by adding up the conditional independent terms together. The proof would be analogous, and so details are omitted. □

Theorem 3 *Let $(X_{si}), s = 1, \ldots, S, i = 1, \ldots, n_s$ be random samples from stratum J_s respectively. Let $\mathbf{u}(\boldsymbol{\eta}, x_{si})$ denote the component score of $\boldsymbol{\eta}$ at one observation x_{si} obtained at stratum J_s, and suppose that each $\mathbf{u}(\boldsymbol{\eta}, x_{si})$ has a $p \times p$ non-singular covariance V_s. Suppose that for each s, $n_s/N \to \lambda_s$ and for some $\lambda_s \in (0, 1)$, the equation is*

$$\frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \mathbf{u}(\boldsymbol{\eta}, x_{si}) \xrightarrow{\mathcal{L}} \sum_{s=1}^{S} \lambda_s V_s^{1/2} \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \tag{13}$$

Theorem 4 *Under the same setting in Theorems 2 or 3, let $\boldsymbol{\Sigma}$ denote asymptotic covariance of $\sum_{s=1}^{S} \sum_{i=1}^{n_s} \mathbf{u}(\boldsymbol{\eta}, x_{si})/\sqrt{N}$ evaluated at the true value $\boldsymbol{\eta}_0$. Let $\widehat{\boldsymbol{\eta}}$ be the composite MLE obtained from the composite score functions (9). Then, as $N \to \infty, n_s \to \infty$, for $s = 1, \ldots, n_s$,*

$$\sqrt{N} \left(\widehat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0 \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{G}(\boldsymbol{\eta}_0)^{-1}), \tag{14}$$

where

$$\mathbf{G}(\boldsymbol{\eta}_0) = \mathbf{H}(\boldsymbol{\eta}_0)^{-1} \boldsymbol{\Sigma} \mathbf{H}(\boldsymbol{\eta}_0) \text{ and } \mathbf{H}(\boldsymbol{\eta}_0) = E_{\boldsymbol{\eta}_0} \left[- \sum_{s=1}^{S} \frac{n_s}{N} \nabla_{\boldsymbol{\eta}_0} \mathbf{u}(\boldsymbol{\eta}_0, x_{si}) \right], \tag{15}$$

with $i \in \{1, \ldots, n_s\}$ and $n_s/N \to \lambda_s$.

Proof Let $\tilde{h}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \mathbf{u}(\boldsymbol{\eta}, x_{si})$. From equation (9), it is obvious that $E\left[\tilde{h}(\boldsymbol{\eta})\right] = 0$ for all $\boldsymbol{\eta}$. From Theorem 2, we have

$$\sqrt{N}\tilde{h}(\boldsymbol{\eta}) \xrightarrow{\mathcal{L}} \sum_{s=1}^{S} \lambda_s V_s^{1/2} \mathcal{N}(\mathbf{0}, \mathbf{I}_p) = \mathcal{N}\left(\mathbf{0}, \sum_{s=1}^{S} \lambda_s^2 V_s\right). \tag{16}$$

Expand the estimating equations $\tilde{h}(\boldsymbol{\eta})$ in a Taylor series with negligible error

$$\mathbf{0} = \tilde{h}(\boldsymbol{\eta}) = \tilde{h}(\boldsymbol{\eta}_0) + \left[\nabla \tilde{h}(\boldsymbol{\eta}_0)\right](\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + o_p(\sqrt{N}), \tag{17}$$

$$-\nabla \tilde{h}(\boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathbf{H}(\boldsymbol{\eta}_0), \tag{18}$$

where

$$\mathbf{H}(\boldsymbol{\eta}_0) = -E\left[\frac{1}{N}\sum_{s=1}^{S} n_s \nabla_{\boldsymbol{\eta}_0} \mathbf{u}(\boldsymbol{\eta}_0, x_s)\right], \quad x_s \in J_s. \tag{19}$$

Equation (17) can be written as

$$\sqrt{N}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = \left[-\nabla \tilde{h}(\boldsymbol{\eta}_0)\right]^{-1}\left[\sqrt{N}\tilde{h}(\boldsymbol{\eta}_0)\right] + o_p(\sqrt{N}), \tag{20}$$

from which by Slustsky's theorem, we get $\sqrt{N}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathbf{H}(\boldsymbol{\eta}_0)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{H}(\boldsymbol{\eta}_0)$. □

- *Remark* Notice that in our case, we have $\mathbf{H}(\boldsymbol{\eta}_0) = \boldsymbol{\Sigma}$.

For IID random samples, with the regularity conditions, the asymptotic results for the MLE are well known. However, for stratified samples such as ours, we encounter the way how to have larger samples. First, we can increase the total sample size N overall managed for the all strata, and with fixed proportions t_s specified at each stratum, $n_s = Nt_s$. Then by Theorem 2, the sum of the score function across the strata asymptotically converges to $\sum_{s=1}^{S} V_s^{1/2} W_s(t_s)$. However, if we simply increase the total sample size N and have the condition that the relative sizes n_s/N are eventually attained at λ_s, then by Theorem 3, we have similar asymptotic normality results. In fact, results on Theorems 2 and 3 are very similar, in that if we have a large sample size n_s for each stratum either the sample is increased with fixed proportions t_s or increases eventually to a proportion λ_s, the scores functions weakly converges to Gaussian weighted by the variance matrix and has variances depending on the proportions t_s or λ_s. Notice however that, in practice, Theorem 2 assures us the asymptotic Gaussian, even we don't increase the sample size for the stratum according to their true proportions behind the population.

By expressing the score function with Taylor expansion up to the second order, the weak convergence of the MLE is presented in Theorem 3. In Theorem 3, we adopt the condition that $n_s/N \to \lambda_s$ and from the proof, we see that the asymptotic variance Σ is asymptotically equivalent to the matrix $\mathbf{H}(\boldsymbol{\eta}_0)$, because the matrix $\mathbf{H}(\boldsymbol{\eta}_0)$ is the Fisher information derived from the score function. In the real example in Sect. 5 of the original manuscript, we show how to implement Theorem 3 numerically to obtain the standard error for the MLEs. Except for the above comments about the asymptotic

properties, which are derived for the stratified samples from disjoint intervals, we have reasonable conjectures that the composite likelihood and the asymptotic properties may be extended to be adopted for the cases when we have stratified samples randomly from overlapped intervals. We will put these extensions in future study.

2.3 Discussions on incorrect distribution assumptions

When the estimation of the distribution of the benchmark variable X is not very accurate, from all of our simulation examples and real examples, we empirically find that the optimal cutoff points would be still very close to the solution under the true distribution. However, incorrect assumptions on the distribution would pass possible bias to the estimated true positive rates. For real application, the true distribution may not always follow what we presume, i.e. the exponential family. We propose here the ways to justify whether the distribution assumptions are suitable and suggest possible modification.

The situations collapse into possible cases that (1) presence of outliers in the data (2) no particular parametric suitable to the data (3) the distribution of the data actually belongs to other parametric family. The first case that the data contain outliers is easier to be handled, because a simple empirical quantile against the fitted quantiles plot may review the extraordinary values. The second and the third case would result in that the fitted parametric distribution in the exponential family would quite differ from the empirical distribution of the data. For the stratified data, we propose to test the goodness of fit, by controlling the familywise error rate with type I error α/S chosen for each stratum and for each stratum we perform the Komogorov–Smirnov test. If it shows that the null hypothesis that the fitted distribution $\hat{F}$ differs from the empirical distribution F_n, we consider the following alternatives.

I. No particular assumption presumed for the stratified data: Under the continuous assumption of the data, we suggest and propose to use the weighted kernel density estimation described in Sect. 2. Then the stratum weights can be estimated.
II. Other family of parametric distributions such as generalized Gamma or non-exponential family: although there are many other parametric distributions not belonging to the exponential family, due to the stratified samples and the profile likelihood is employed to solved for the likelihood estimates, the consistency of the estimates of particular parametric distribution may require some further investigation and we leave this for further study.
III. Box-Cox transformation: if the original observations (stratified) has a nonlinear transformation to normal, we propose the following steps.
   – For unstratified X, the power transformation is utilized as

$$x^{(\lambda_1)} = \begin{cases} \frac{(x^{\lambda_1}+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0 \\ \log(x+\lambda_2), & \lambda_1 = 0 \end{cases}, \quad x > -\lambda_2. \tag{21}$$

– the profile likelihood with stratified power transformed data is

$$\sum_s \frac{1}{Q_s(\boldsymbol{\theta})} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n_s} \exp\left\{-\frac{(\mathbf{x}^{(\lambda_1)}-\boldsymbol{\mu})'(\mathbf{x}^{(\lambda_1)}-\boldsymbol{\mu})}{2\sigma^2}\right\} \mathbf{Jac}(\lambda_1;\mathbf{x}) \tag{22}$$

where the Jacobin is $\mathbf{Jac}(\lambda_1;\mathbf{x}) = \prod_{j=1}^{n_s} \left|\frac{dx_i^{(\lambda_1)}}{dx_i}\right|$

Analogously following the Box-Cox transformation as for unstratified data, we solve for the power λ_1 by maximizing the profile likelihood function.

2.4 Non-parametric distribution with kernel density

Assume that the input variable X has an unknown continuous density function $f(x)$ defined on the support of X that consists of partitioned intervals $J_s = [b_{s-1}, b_s)$, $b_0 < b_1 \cdots < b_S$. With stratified samples, we estimate the underlying density function with the following steps:

Step 1. For each semi-close interval, estimate one single stratum density function $f_{J_s,h_s}(x)$ by the kernel density function

$$\widehat{f}_{J_s,h_s}(x) = \frac{1}{n_s h_s}\sum_{i=1}^{n_s} K\left(\frac{x_{si}-X}{h_s}\right) = \frac{1}{n_s h_s}\sum_{i=1}^{n_s}\frac{1}{2}\mathcal{I}\left(\left|\frac{x_{si}-X}{h_s}\right|<1\right), \tag{23}$$

where $K(u)$ is the uniform kernel defined on $(-1, 1)$.
Step 2. Unify the stratum density functions into a function g:

$$\widehat{f}_{UN}(x) = \sum_{s=1}^{S} \widehat{f}_{J_s,h_s}(x)c(s)I(x\in J_s), \quad \text{for } x\in J_s, \tag{24}$$

where $c(s) > 0$ satisfies the condition that $\widehat{f}_{J_{s-1},h_{s-1}}(b_s^-)c(s-1) = \widehat{f}_{J_s,h_s}(b_s)c(s)$. A particular choice of such $c(s)$ is

$$c(s^*) = \begin{cases} 1, & \text{for } s=1 \\ \prod_{s=2}^{s^*}\left(\widehat{f}_{J_{s-1},h_{s-1}}(b_s^-)\Big/\widehat{f}_{J_s,h_s}(b_s)\right), & \text{for } s^*\geq 2, \end{cases} \tag{25}$$

where $\widehat{f}_{J_{s-1},h_{s-1}}(b_s^-) = \lim_{b\uparrow b_s}\widehat{f}_{J_{s-1},h_{s-1}}(b)$. Notice that the function g is piecewise smooth by stratum.
Step 3. Normalize the function g to a density function over the support $\Omega = J_1 \cup \cdots \cup J_S$ that

$$\widehat{f}_h(x) = \frac{\widehat{f}_{UN}(x)}{\int_\Omega \widehat{f}_{UN}(x)dx} \approx \frac{\widehat{f}_{UN}(x)}{\sum_{s=1}^{S}\sum_{i=1}^{n_s}\widehat{f_{J_s,h_s}}(x_{si})c(s)I(x_{si}\in J_s)\Delta_s} \tag{26}$$

$$= \frac{\sum_{s=1}^{S}\widehat{f}_{J_s,h_s}(x)c(s)I(x\in J_s)}{\sum_{s=1}^{S}\sum_{i=1}^{n_s}\widehat{f_{J_s,h_s}}(x_{si})c(s)I(x_{si}\in J_s)\Delta_s} \tag{27}$$

$$= \sum_{s=1}^{S} w(x,s,h_s)\frac{1}{h_s}\sum_{i=1}^{n_s} K\left(\frac{x_{si}-X}{h_s}\right),\ s=1,\ldots S, \tag{28}$$

where $w(x,s,h_s) = \frac{1}{n_s}\frac{c(s)I(x\in J_s)}{\sum_{s=1}^{S}\sum_{i=1}^{n_s}\widehat{f}_{J_s,h_s}(x_{si})c(s)I(x_{si}\in J_s)\Delta_s}$ and Δ_s is the equal width between bins within each J_s. The fitted $\widehat{f}_h(x)$ is our proposed density to the non-parametric estimate of the underlying density function for X. All the convergence assumptions required in the setting of the kernel density estimation are conventionally assumed to be held under each stratum. We adopt Eq. (28) to express the estimate as a weighted-type kernel density as discussed in Wang and Sun (2009) and Wang and Wang (2007), except that the weights $w(x,s,h_s)$ in (28) are based on our kernel density estimation for each stratum rather than due to any weight that have ever been formerly determined for those data points. The weights $w(x,s,h)$ are driven by the data, and the sum of the total weights is approximately equal to one when $c(s)$ has been obtained. In fact, substituting this result into $w(x,s,h_s)$ gives us the approximation $\sum_{s=1}^{S} w(x_{si},s,h_s) \approx 1$. We may consider $w(x_{si},s,h_s)$ as the weight representing the data points fitting the kernel density.

The bandwidth controls the smoothness of the kernel estimate and is regularly chosen to minimize a measure of the closeness between the estimate $\widehat{f}_h(x)$ and the true density f. Here, universally, we set $h_s = h$. We then measure the closeness by considering the integrated squared error (ISE), that is

$$ISE(\widehat{f}_h) = \int \left(\widehat{f}_h - f\right)^2 dx = \int \left(\widehat{f}_h\right)^2 dx - 2\int \widehat{f}_h f dx + \int f^2 dx. \tag{29}$$

The ISE in (29) can be conveniently approximated by utilizing leave-one-out cross-validation (Bowman 1984; Sheather et al. 2004; Wang and Wang 2007). We apply steps similar to those in the work by Wang and Wang (2007) to the weighted form (28), and denote $w(x_i,s,h)$ as w_i, the first term of ISE is

$$\int \widehat{f}_h(x)^2 dx = \sum_{s=1}^{S}\int \sum_i \frac{w_i}{h}K\left(\frac{x-X_i}{h}\right) \times \sum_j \frac{w_j}{h}K\left(\frac{x-X_j}{h}\right)dx \tag{30}$$

$$\overset{t=x/h}{=} \sum_{s=1}^{S}\frac{1}{h}\sum_i\sum_j w_i w_j \int K\left(\frac{X_i-t}{h}\right)K\left(\frac{t-X_j}{h}\right)dt \tag{31}$$

$$\overset{u=t-(x/h)}{=} \sum_{s=1}^{S}\frac{1}{h}\sum_i\sum_j w_i w_j \int K\left[\frac{X_i}{h}-\frac{X_j}{h}-u\right]K(u)\,du, \tag{32}$$

where $u \sim Unif(-1, 1)$ and the subscripts i, j are denoted inside the stratum J_s. After integrating the convolution part of the kernel, we have

$$\int \widehat{f_h}(x)^2 dx = \sum_{s=1}^{S} \frac{1}{h} \sum_i \sum_j w_i w_j K^*(X_i, X_j), \tag{33}$$

where

$$K^*(X_i, X_j) = \begin{cases} \frac{1}{4}\left[2 - \frac{X_i - X_j}{h}\right], & \text{for } \frac{X_i - X_j}{h} \in (0, 2) \\ \frac{1}{4}\left[2 + \frac{X_i - X_j}{h}\right], & \text{for } \frac{X_i - X_j}{h} \in (-2, 0). \end{cases} \tag{34}$$

Now, the kernel density $\widehat{f}_{-i}$ by leaving the ith case out of computation while the others $\mathbf{x}^{(-i)}$ will receive re-adjusted weights that are the original weight w_j normalized by the the total weights of $\mathbf{x}^{(-i)}$. The leave-one-out estimation of $\widehat{f}_{-i}$ would be

$$\widehat{f}_{-i} = \frac{\sum_{j \neq i} w_j^{(-i)} K_h\left(x - X_j\right)}{\sum_{j \neq i} w_j^{(-i)}}. \tag{35}$$

The second term of ISE can be estimated by $-2/n \sum \widehat{f}_{-i}$. Although, the entire sample is stratified and only conditioning within the same stratum observations are treated independently, the leave-one-out can still be applied because stratification is unchanged. We seek to obtain a bandwidth h to minimize

$$\sum_{s=1}^{S} \frac{1}{h} \sum_i \sum_j w_i w_j K^*(X_i, X_j) - \frac{2}{n} \sum \widehat{f}_{-i}. \tag{36}$$

For a target range of h, we perform a grid search on h and the ISE computed with a corresponding $w(s)$ is minimized. An important issue that requires to be called is the boundary effects for each $\widehat{f}_{J_s, h_s}(x)$. Without further corrections, the estimation bias on the boundaries of each J_s may induce some overall bias for the final estimated density $\widehat{f}(x)$. For our estimation, we use the correction method proposed by Zhang et al. (1999), a combination of methods of pseudo-data, transformation, and reflection, with these advantages that the variance is kept down, and a small bias could only be on the boundary. But Zhang's method is only originally introduced for a density with support on $[0, \infty)$. To apply the correction method to the kernel density estimated on $J_1, \ldots, J_s$, we will need to modify the method to correct boundary effects at a right-hand sided endpoint, and also modify the method that can be utilized for support with two-sided boundaries. We will illustrate these modified corrections in our simulation studies in Sect. 5.

3 Estimation of *EPE*(*v*)

3.1 Ad-hoc estimation

For a cut-off point $c \in J_{\tilde{s}}$ of X, we estimate the $EPE(v)$ with

$$\widehat{EPE}_c(v) = \int_{X<c} \mathbf{1}(Y=1|X)d\widehat{P}_X + \int_{X\geq c} \mathbf{1}(Y=0|X)d\widehat{P}_X, \tag{37}$$

where $\widehat{P}_X$ refers to the fitted distribution obtained by the methods in Sect. 2. Since the distribution of X is assumed to be continuous, this would be equivalent to

$$\widehat{EPE}_c(v) = \sum_{s=1}^{S} \int_{X\in J_s, X<c} \mathbf{1}(Y=1|X)d\widehat{P}_X + \sum_{s=1}^{S} \int_{X\in J_s, X\geq c} \mathbf{1}(Y=0|X)d\widehat{P}_X. \tag{38}$$

While estimating the expectation of indicator functions inside Eq. (38) by the empirical measures that in turn are the relative frequencies of $Y=0$ and $Y=1$ within J_s, we may then evaluate $\widehat{EPE}_c(v)$ by ad-hoc approach as following. Let $Z=1$, if $X \geq c$, otherwise $Z=0$. We may denote the sampled triple as (X_{si}, Y_{si}, Z_{si}), $s = 1,\ldots,S$, $i = 1,\ldots,n_s$. For each stratum J_s, we denote the sample of X as $\{X_{si}\}$. Suppose that c is a value in $J_{s^*} \neq J_S$. Further, write J_s as a union of A_1, A_2:

$$A_1 = \{x \in J_{s^*} : b_{s^*-1} \leq x < c\}, \quad A_2 = \{x \in J_{s^*} : c \leq x \leq b_{s*}\}. \tag{39}$$

Then for a fixed value c, the conditional true positive probability is

$$\begin{aligned} TP(c) &= P(Y=1|Z=1) && (40)\\ &= P(Y=1, Z=1, A_2|Z=1) + \sum_{k=s^*+1}^{S} P(Y=1, J_k|Z=1) && (41)\\ &= \frac{P(Y=1, A_2, Z=1)P(A_2, Z=1)}{P(Z=1)P(A_2, Z=1)} \\ &\quad + \sum_{k=s^*+1}^{S} \frac{P(Y=1|J_k, Z=1)P(J_k, Z=1)}{P(Z=1)} && (42)\\ &= \frac{p_{s*}P(c \leq X < b_{s^*})}{P(X\geq c)} + \sum_{k=s^*+1}^{S} \frac{p_k P(X\in J_k)}{P(X\geq c)}, && (43) \end{aligned}$$

where p_{s^*} denotes $P(Y=1|A_2, Z=1)$ and p_k denotes $P(Y=1|J_k, Z=1)$. With the sample triples (X_{si}, Y_{si}, Z_{si}), we estimate the probabilities $P(c \leq X < b_{s^*})$, $P(X \in J_k)$ and $P(X \geq c)$ using the estimated density from Sect. 2. We then estimate p_{s^*}, p_k with the sample proportions of $Y=1$ observed in strata J_{s^*} and J_k, which are

$$\widehat{p}_{s^*} = \frac{\sum_{i\in A_2} Y_{s^*i}}{m_{s^*}}, \quad \widehat{p}_k = \frac{\sum_{i\in J_k} Y_{ki}}{n_k}, k = s^*+1, \ldots, S, \tag{44}$$

where m_{s^*}, n_k are the number of subjects in A_2 and J_k respectively. The estimated $TP(c)$ given the value c would be

$$\widehat{TP(c)} = \frac{\widehat{p}_{s*}\widehat{P}(c \le X < b_{s^*})}{\widehat{P}(X \ge c)} + \sum_{k=s^*+1}^{S} \frac{\widehat{p}_k \widehat{P}(X \in J_k)}{\widehat{P}(X \ge c)}. \tag{45}$$

Now, if the value of c is given at J_S, then

$$\widehat{TP(c)} = \widehat{p}_S = \frac{\sum_{i\in J_S} Y_{Si}}{n_S}. \tag{46}$$

Analogously, for a fixed value $c \in J_{s^*}, s^* \neq 1$, the estimation of false positive probability is

$$\widehat{FP(c)} = \frac{\widehat{r}_{s*}\widehat{P}(b_{s^*-1} \le X < c)}{\widehat{P}(X < c)} + \sum_{k=1}^{s^*-1} \frac{\widehat{r}_k \widehat{P}(X \in J_k)}{\widehat{P}(X < c)}, \tag{47}$$

where $\widehat{r}_{s^*}$ and $\widehat{r}_k$ are

$$\widehat{r}_{s^*} = \frac{\sum_{i\in A_1} Y_{s^*i}}{m_{s^*}}, \quad \widehat{p}_k = \frac{\sum_{i\in J_k} Y_{ki}}{n_k}, k = 1, \ldots, s^*-1, \tag{48}$$

where m_{s^*}, n_k are the number of subjects in A_1 and J_k respectively. From the above equations, the estimated $EPE(v)$ is

$$\widehat{EPE}(v) = 1 - \left\{ \widehat{P}(X < c) - \left(\widehat{r}_{s*}\widehat{P}(b_{s^*-1} \le X < c) + \sum_{k=1}^{s^*-1} \widehat{r}_k \widehat{P}(X \in J_k) \right) + \widehat{p}_{s*}\widehat{P}(c \le X < b_{s^*}) + \sum_{k=s^*+1}^{S} \widehat{p}_k \widehat{P}(X \in J_k) \right\} \tag{49}$$

and $\widehat{TCP} = 1 - \widehat{EPE}(v)$.

3.2 TPR and the logistic model

Suppose we employ a logistic model $\log(E(Y = 1|\mathbf{1}\{X \ge c\})) = \alpha + \beta\mathbf{1}\{X \ge c\}$, which imposes the constant parameters α and β universally over $X \ge c$. With an

expression of composite likelihood over the strata with the logistic model, we solve the score functions for α and β leading to

$$\sum_{s=s^*}^{S}\left[\sum_{i=1}^{n_s} Y_{si} - \sum_{i=1}^{n_{si}} \frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}\right] P(X \in J_s | X \geq c). \tag{50}$$

The estimated TPR at a cutoff c is a solution sought by a weighted logistic model, that is,

$$\frac{\sum_{s=s^*}^{S}\sum_{i=1}^{n_s} Y_{si}\, P(X \in J_s | X \geq c)}{\sum_{s=s^*}^{S}\sum_{i=1}^{n_s} n_{si}\, P(X \in J_s | X \geq c)}. \tag{51}$$

In contrast to the ad-hoc estimation in Eq. (40), it can be remarked that the ad-hoc estimation does not acquire the universal constant parameters through the strata, but rather fit the expectation *locally* inside the strata. The logistic model turns out to be a special case under our ad-hoc approach.

3.3 Comparisons of EPE and the *q*-statistic

When matching the binary surrogate variable Y by the continuous values of X, we seek a cufoff point dividing the range of X into two divisions, so that the optimal heterogeneity or classification between two groups is attained, and the groups are corresponding to the binary surrogates $Y = 0$ or $Y = 1$. By the concept of spatial stratified heterogeneity, the classified groups are considered as spatial strata or classified groups. We may take further look about comparing EPE with using q-statistic. To avoid using the same term, i.e. "strata" mentioned for the semi-closed intervals in this paper, for our case when discussing the q-statistic, we adopt the term *classified groups* in terms of the spatial strata. The q-statistic, proposed by Wang et al. (2010, 2016), is a measure for presenting the spatial heterogeneity for an interesting variable W:

$$q = 1 - \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(W_{hi} - \overline{W}_h)^2}{\sum_{i=1}^{N}(W_i - \overline{W})^2} = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}, \tag{52}$$

where h is the index of the classified group. The main concept is to measure the dissimilarity between the groups by the difference of 1 and the ratio of the between group variation relative to the total variation. For our study, we set $h = 1, 2$ for the "$Y = 0$" and "$Y = 1$" groups, and we seek to apply the q-statistic to our paired interval stratified data (X, Y) and evaluate the cutoff point with the optimal (maximum) value of q-statistic.

We apply the q-statistic on the binary data of Y. Under a cutoff value c, the q-statistic is

$$1 - \frac{\sum_{s,i,X<c}(Y_{si} - \overline{Y}_1)^2 + \sum_{s,i,X\geq c}(Y_{si} - \overline{Y}_2)^2}{\sum_s \sum_i (Y_{si} - \overline{Y})^2}, \tag{53}$$

where $\sum_{s,i,X<c}(Y_{si}-\overline{Y}_1)^2$ is the total sum of squares variation for the values of Y with $X<c$. The total variation $\sum_s\sum_i(Y_{si}-\overline{Y})^2$ would be the same for any cutoff value c we apply, hence it is not affected by the cutoff. For a choice of cutoff value c, the numerator of the second term in Eq. (53) can be written into

$$\sum_{s,i,X<c}(Y_{si}-\mu_1)^2+\sum_{s,i,X\geq c}(Y_{si}-\mu_2)^2+\sum_{s,X<c}(\overline{Y}_1-\mu_1)^2+\sum_{s,X\geq c}(\overline{Y}_2-\mu_2)^2, \tag{54}$$

where μ_i denotes the true means of the groups based on the classification with such cutoff value c. Recall that if c is a general cutoff value, with stratified data, suppose $c\in J_{s^*}$ for some semi-close interval J_{s^*}, and $w_s^{(1)}, s=1,\ldots,s^*$ are probability weights of J_s conditioning on $X<c$, $w_s^{(2)}, s=s^*,\ldots,S$ are probability weights of J_s conditioning on $X\geq c$:

$$\begin{aligned} &P(Y=0,J_s|X<c)=(1-a_1)w_s^{(1)},\\ &s=1,\ldots,s^*,\ w_1^{(1)}+\cdots+w_{s^*}^{(1)}=1, \end{aligned} \tag{55}$$

$$\begin{aligned} &P(Y=1,J_s|X\geq c)=(1-a_2)w_s^{(2)},\\ &s=s^*,\ldots,S,\ w_{s^*}^{(2)}+\cdots+w_S^{(2)}=1, \end{aligned} \tag{56}$$

for some constants $0<a_1,a_2<1$. Based on this, with a cufoff value c, the true mean μ_1 and μ_2 of Y on the two groups are $\mu_1=a_1=P(Y=1|X<c)$ and $\mu_2=1-a_2=P(Y=1|X\geq c)$. Because Y_{si} are either 0 or 1, conditioning on the classified groups, the marginal expectation over the stratum is

$$\begin{aligned} &E\left[\sum_{s,i,X<c}(Y_{si}-\mu_1)^2+\sum_{s,i,X\geq c}(Y_{si}-\mu_2)^2\right]\\ &=E\left[E\left[\sum_{s,i,X<c}(Y_{si}-\mu_1)^2+\sum_{s,i,X\geq c}(Y_{si}-\mu_2)^2\Big|X\right]\right]\\ &=E\left[\left[(\mathbf{Y}-\boldsymbol{\mu})^T(\mathbf{Y}-\boldsymbol{\mu})\Big|X\right]\right] \end{aligned} \tag{57}$$

where $\mathbf{Y}$ represents the whole binary data Y and $\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)$ with $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2$ vectors of elements μ_1 and μ_2. Equation (57) in fact is also an expected prediction error (EPE) but with square error loss. If we restrict our solutions to those unbiased ones, the forms of solutions of μ_1 and μ_2 are formulated as $\mu_1=E(Y|X<c)=P(Y=1|X<c)$ and $\mu_2=E(Y|X\geq c)=P(Y=1|X\geq c)$. For one binary Y, in order to obtain the unbiased $E(Y|X\geq c)$, accordingly we must have the optimal cutoff c corresponding to the true μ_1 and μ_2. Because under the true mean μ_1 and μ_2, asymptotically $\sum_{s,X<c}(\overline{Y}_1-\mu_1)^2+\sum_{s,X\geq c}(\overline{Y}_2-\mu_2)^2$ goes to zero,[1] we have on

[1] For $X<c$, $\overline{Y}_1=\frac{\sum_{i,s_1}Y_{si}+\cdots+\sum_{i,s^*}Y_{si}}{N_1}\approx\frac{n_1w_1\mu_1+\cdots+n_{s^*}w_{s^*}\mu_1}{N_1}=\mu_1$

average and sufficiently the Eq. (54) would be minimized at the true cutoff value with the optimal solutions on μ_1 and μ_2. Comparing with the EPE in our approach, we use EPE with the zero-one loss function which generates best prediction $v(x) = 0$ when $\max_y P(Y = y|X < c) = P(Y = 0|X < c)$ and $v(x) = 1$ when $\max_y P(Y = y|X \geq c) = P(Y = 0|X \geq c)$. For this binary classification, we can see asymptotically the same cutoff value c minimizing the two versions of EPE. Thus, we conclude that the optimal cufoff based on q-statistic and our approach would be close, but certainly have different values on the statistics. Here we also notice that the unbiased solutions μ_1 and μ_2 are the best unbiased solutions, however may not be the best solutions in minimizing the expected square loss (57), when comparing using some possible other James-Stein type estimators in (57). Such estimators are biased and would not be corresponding to using $\overline{Y}_1$ and $\overline{Y}_2$ in the q-statistic. Empirically, we work on EPE with estimation of the probabilities of $P(X < c)$ and $P(X \geq c)$ hence we encounter some adding prediction errors, while the q-statistic is computed on the summarized sum of squares with raw data but also have some errors due to $\sum_{s,X<c}(\overline{Y}_1-\mu_1)^2+\sum_{s,X\geq c}(\overline{Y}_2-\mu_2)^2$. However, the q-statistic would not report the true positive rates as when of interests of the scientists.

Now, we may want to have some idea about how the value of q-statistic may take. Because the binary data on Y, though independent within a stratum, can not be regarded as independent observations between the strata. With possible minor dependence, we put

$$\sum_{s,i,X<c}(Y_{si} - \overline{Y}_1)^2 \approx (N_1 - 1)s_1^2 + \text{bias}_1,$$

$$\sum_{s,i,X\geq c}(Y_{si} - \overline{Y}_2)^2 \approx (N_2 - 1)s_1^2 + \text{bias}_2, \tag{58}$$

$$\sum_{s}\sum_{i}(Y_{si} - \overline{Y})^2 \approx (N - 1)s_1^2 + \text{bias}_3 \tag{59}$$

With moderate or small dependence, the bias of the above would be small, and consistently the q-statistic has value approximated by $1 - \frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{(N-1)s^2}$. The empirical comparisons are shown in Example 1 in Sect. 4.

We conclude that the optimal cutoff values from our approach are close to the optimal cutoff obtaining from maximizing the q-statistic. However in our problem, our approach in using EPE is more suitable. Because in our case, first, we are also interested in reporting the true positive rates as further results for evaluating the misclassification, and second our classified data is binary, the q-statistic may generate multiple solutions. In fact,the q-statistic may be more suitable for continuous data as according to our experiences.

4 Simulation studies

In this section, we conduct simulation examples to assess the performance of the proposed method. All of our three examples are analyzed on bivariate data (X, Y)

collected with the support of X based on the stratified sampling scheme. We assumed $X \sim F(x)$ is continuous, and Y is a binary outcome monotonically associated with the input variable X by Eq. (61). For each example, we compare the results from our approach with that estimation obtained by the simple, naive non-parametric method (NP), at which, any disproportionate sampling bias is intended to be ignored. Then, by such a naive non-parametric method, for a cutoff point $c \in J_{s^*}$, the non-parametric TCP would only use the relative frequencies for probabilities:

$$TCP_{\mathbf{np}} = \sum_{s=s^*}^{S} \sum_{\{\, i: X_{si} \geq c \,\}} \mathbf{1}\{\, Y_{si} = 1 \,\}/n + \sum_{s=1}^{s^*} \sum_{\{\, i: X_{si} < c \,\}} \mathbf{1}\{\, Y_{si} = 0 \,\}/n. \tag{60}$$

Through the empirical results, we convey the following.

- The estimation of distribution of X, fitted parametrically or non-parametrically by the kernel density, is close to the true distribution.
- The *distribution-adjusted* TCP, where stratum weights were incorporated from our ad-hoc approach with disproportional stratified samples, is close to the TCP that is estimated based on proportional stratified samples if the underlying distribution of X is the same. But *distribution-adjusted* TCPs are quite different from TCP_{np}.
- The cutoff point appears without a significantly large difference to the cutoff point based on $TCP_{\mathbf{np}}$.
- For values $X \geq \widehat{c}$, the true positive probabilities from both methods are close, but these probabilities tend to be much overestimated at $X \geq c$, $c \leq \widehat{c}$ in some cases if $TCP_{\mathbf{np}}$ is used.

Example 1. Gamma distributed input variable

We assume that $X \sim Gamma(15, 3)$ and the variable Y is associated to X according to the following logistic regression

$$P(Y = 1|X = x) = \frac{\exp(-12.5 + 1.92X)}{1 + \exp(-12.5 + 1.92X)} \tag{61}$$

From the relation, one can identify that the optimal cutoff point is 6.5104 at which success probability is 0.5.

(1.a) Bias from using the logistic model

We generate simple random pairs (X, Y) where $X \sim Gamma(15, 3)$ with sample size 500 and generate (X, Y) based on the stratified sampling scheme that with values of X from $J_1 = [0, 6)$, $J_2 = [6, 7)$, $J_3 = (7, \infty)$ and sample sizes 100, 250, 200 respectively. For the two data sets and a sequence of cutoff values $c > 0$, at each c, we fit the following logistic model:

$$P(Y = 1|X \geq c) = \frac{\exp(\alpha + \beta\, \mathbf{1}\{\, X \geq c \,\}}{1 + \exp(\alpha + \beta\, \mathbf{1}\{\, X \geq c \,\}}.$$

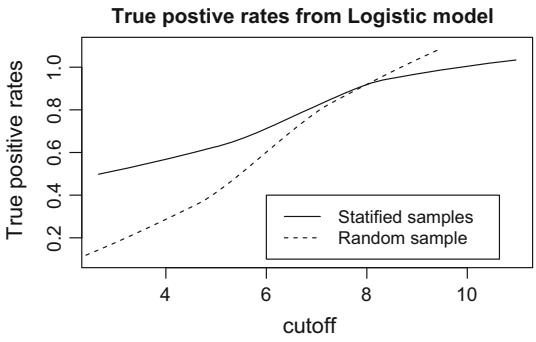

Fig. 1 Estimated TPR at various cutoff points with logistic model

The resultant TPRs for these two simulations are shown in the following chart (Fig. 1).

We found that there is potential bias with stratified samples. The TPR curve for the simple random pairs may be also taken as a benchmark to be compared later with the results from our approach.

(1.b) TCP and TPR

We simulate samples with proportional and disproportionate stratification produced by partitioning the support of X into $J_1 = [0, 6)$, $J_2 = [6, 7)$, $J_3 = [7, \infty)$, and from each J_s, the sample sizes are arranged in 6 cases: (a) $n_1 = 79, n_2 = 7, n_3 = 14$ (b) $n_1 = 395, n_2 = 35, n_3 = 70$ (c) $n_1 = 30, n_2 = 50, n_3 = 20$ (d) $n_1 = 40, n_2 = 150, n_3 = 60$ (e) $n_1 = 100, n_2 = 250, n_3 = 200$ (f) $n_1 = 130, n_2 = 10, n_3 = 100$. The samples under (a) and (b), designed for representing cases of moderate and larger sizes, have relative stratum sample sizes that are proportional to the original probability weights of J_1, J_2, J_3 in $Gamma(15, 3)$. Obviously, the samples under (c), (d) and (e) make the second strata oversampled, and in (f), we undersample the second strata. For each sample size scheme listed above, we run 100 Monte Carlo replications and assess the composite MLEs of the Gamma parameters and the fitted distribution of X. The results are summarized in Table 2, which shows that the composite MLEs of Gamma parameters seem to have potential variances, and may result in a wider range of solutions. However, based on the 100 Monte Carlo replications, the distribution deviations calculated upon $\|\widehat{F}_n - F\|$ only show moderate differences between them, and these differences become relatively much small when sample sizes increase. Further, the TCP plot (Fig. 2) reveals that the distribution-adjusted TCP of all disproportional stratified samples [(c), (d), (e), (f)] are close to the TCP of the proportional stratified samples, and this is also similar to the TCP obtained under simple random pairs. We found that the naive non-parametric TCP (TCP_{np}) has bias and sharply drops when the cutoff point is just slightly away from the cutoff. In addition, the estimate of TPR, however, is bumpy due to the stratification, we present the TPR curve through a Lowess smoothing. From the TPR plot (Fig. 3), we also find that the TPR with adjustment using the correct underlying distribution is similar to the TPR curve with proportionately stratified samples, and so, we consider that it has recovered well under disproportionately sampled data. Nevertheless, in these

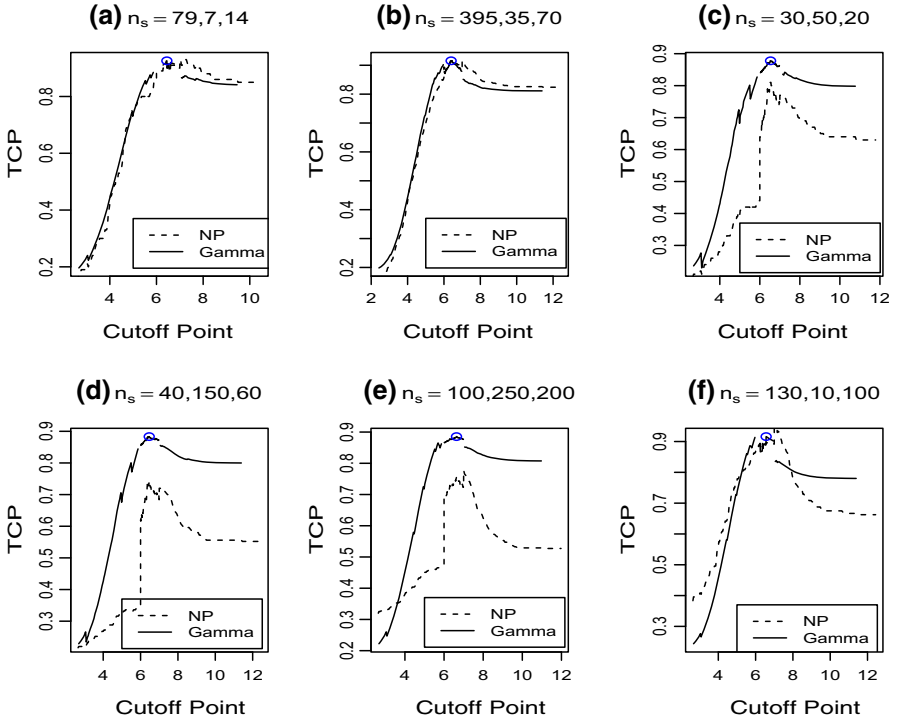

Fig. 2 Estimated TCP at different cutoff points: the solid curves are obtained with adjusting strata probabilities using estimated gamma probabilities, and the dashed curves are obtained using the naive non-parametric estimation

cases, the naive non-parametric TPR may overestimate the TPR if the disproportional sampling weights are ignored while data is oversampled in J_2 or J_3. This suggests that when we give lower cutoff values, the naive nonparametric approach would not rightly explore TPRs under disproportionate stratification.

Empirically, in this example, we compare the cutoff value maximizing the q-statistic with the cufoff values from our two stage with EPE approach based on the simulation Gamma example. Using one simulation data, we found that the cutoff values obtained from the q-statistic are close to the cutoff values from our approach in the sense of 95% confidence intervals as using the Monte Carlo standard error shown in the Table 2. However, because we work on EPE with ad-hoc estimation and the q-statistic is calculated based on the raw data, the q-statistic in our case somehow has bumpy paths. In fact, in the numerical study, we encounter multiple solutions reaching the same maximum of q-statistic, and we take averages over these solutions as the optimal cutoff from the q-statistic. The numbers are summarized in Fig. 4.

Example 2. Kernel density on gamma distributed input variable

In this example, we assume that $X \sim Gamma(\alpha, \beta), \alpha = 15, \beta = 3$ and the variable Y is associated with X according to the logistic function, as per Eq. (61).

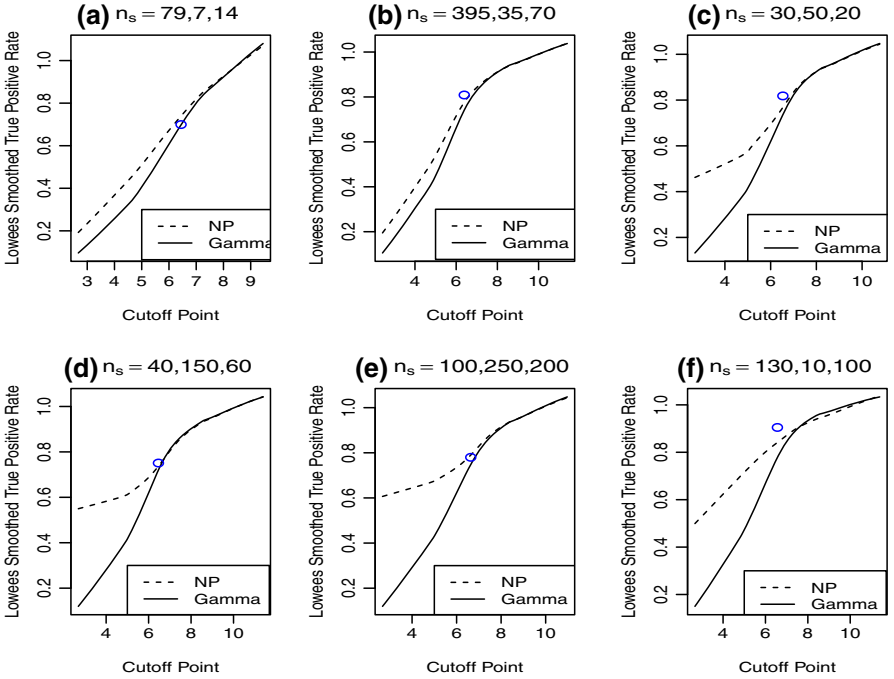

Fig. 3 Smooth true positive probability curve: solid curves are solutions adjusting probabilities using the estimated gamma probabilities, dashed curves are from the naive non-parametric estimation

Table 2 Estimation of gamma distribution with various disproportionate schemes

	Sample size in paritions J_s					
	(79,7,14)	(395,35,70)	(30,50,20)	(40,150,60)	(100,250,200)	(130,10,100)
$\widehat{\alpha}$	13.34 ± 0.54	14.36 ± 0.92	14.02 ± 1.52	13.92 ± 1.43	14.38 ± 1.07	14.11 ± 0.91
$\widehat{\beta}$	2.61 ± 0.16	2.87 ± 0.21	2.70 ± 0.33	2.79 ± 0.25	2.89 ± 0.21	2.81 ± 0.19
Cutoff	6.20 ± 0.42	6.27 ± 0.32	6.25 ± 0.39	6.42 ± 0.31	6.19 ± 0.32	6.18 ± 0.32
$\|\hat{F}_n - F\|$	(0,0,0.1)	(0,0,0.05)	(0,0,0.13)	(0,0,0.09)	(0,0,0.06)	(0,0,0.06)

$\widehat{\alpha}, \widehat{\beta}$ are estimates of shape and scale parameters based on 100 Monte Carlo samples presented with $\pm$ one s.e. The entries of the last row in the Table are the 10th, 50th, 95th percentiles for $\|\widehat{F}_n - F\|$, where $F(x)$ is $Gamma(15, 3)$

The sample is generated with disproportional stratification partitioning the support of X by $J_1 = [0, 6)$, $J_2 = [6, 7)$, $J_3 = [7, \infty)$, and from each J_s, the sample sizes are $n_1 = 100, n_2 = 250, n_3 = 200$. In this study, only one data set is analyzed to be compared to the results from Example 1. The underlying distribution is assumed to be unknown and has continuous density. The kernel density of X is estimated with the method constructed in Sect. 2. Figure 5 summarizes the significant findings of our analysis. The plot of ISE versus bandwidth shows that the lowest ISE is attained at the bandwidth value of around 0.1, but empirically, the locally largest bandwidth usually

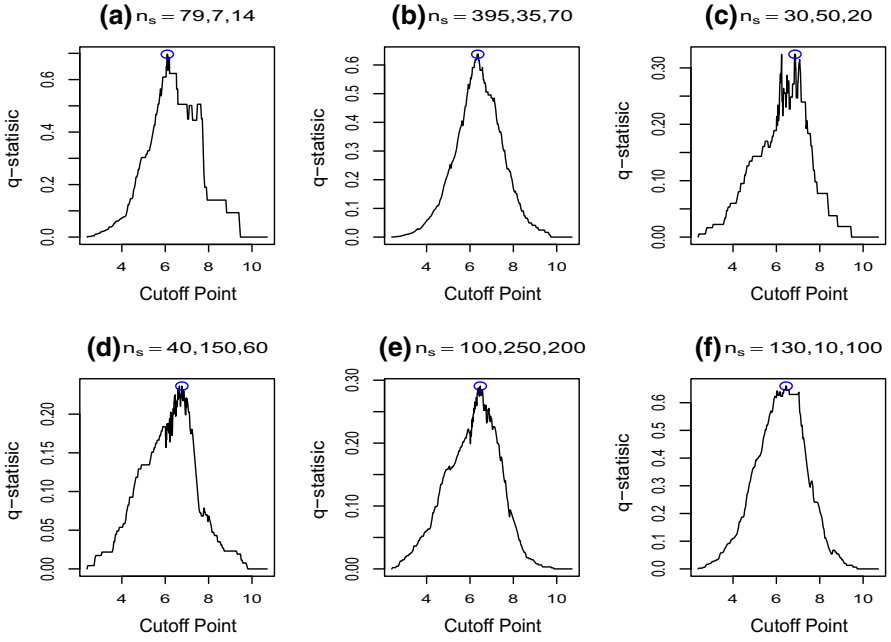

Fig. 4 q-statistics for various cutoff values with the optimal points shown by circles

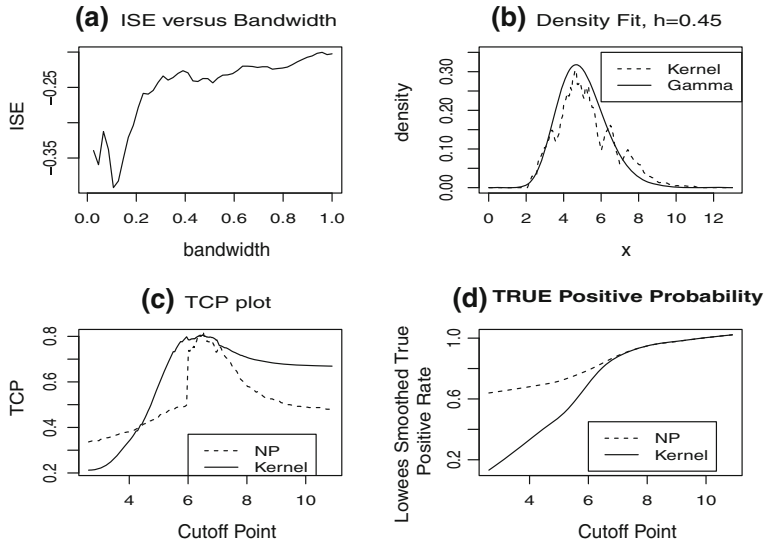

Fig. 5 Estimated TCP, smoothed TPR: solid curves are solutions adjusting probabilities using kernel densities, dashed curves are from the naive non-parametric estimation

provides better performance (Jones et al. 1996). Hence, we choose the bandwidth $h = 0.5$ for this case. We found that the kernel density is close to the true density $Gamma(15, 3)$, but it is not without some bumpy curvatures - particularly around

the boundary of the interval J_2. The TCP seems similar to the TCP from Example 1, though with a lower top value around the cutoff point.

We further conduct boundary correction for the kernel density. The kernel density estimate is constructed with boundary correction on the boundaries of J_1, J_2, J_3 using the improved estimator of the density function at the boundary (Zhang et al. 1999), which is shown to have advantages while keeping the variance down while the bias is also controlled. The improved estimator is formed as a combination of pseudo-data, transformation, and reflection methods using the extended data set consisting of the original data $X_1, \ldots, X_n \in [0, \infty)$ and the pseudo transformed data $-g(X_1), \ldots, -g(X_n)$ pulled into a new kernel density estimator:

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + g(X_i)}{h}\right) \right\}, \tag{62}$$

where $g \in [0, \infty)$ is non-negative, continuous, and monotonically increasing with $g^{-1}(0) = 0, g^{(1)}(0) = 1$. As shown in the work of Zhang et al. (1999), the bias of their estimate is first-orderly trimmed off by letting $g^{(2)}(0) = [2f^{(1)}(0)/f(0)]$. Further, a rationale choice of g is also proposed to be

$$g(x) = x + dx^2 + Ad^2x^3, \tag{63}$$

where $d = f^{(1)}(0)/f(0)$ and $3A > 1$. To implement Zhang's estimate in our example with data presented in partitioned intervals, with the same type of g, we modify the estimate by the followings

- For $J_1 = (0, 6)$: $\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{(6-x)+g(6-X_i)}{h}\right) \right\}$
- For $J_2 = [6, 7)$: $\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + 0.5K\left(\frac{(x-7)+g(X_i-7)}{h}\right) + 0.5K\left(\frac{(6-x)+g(6-X_i)}{h}\right) \right\}$
- For $J_3 = [7, \infty)$: $\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{(x-7)+g(X_i-7)}{h}\right) \right\}$

For J_1 and J_3, the estimates $\widehat{f}(x)$ were only constructed on the data by resetting the data to have the boundary on zero on the left. For J_2, the data is confined to an interval with two endpoints, and there seems to be no direct way of applying Zhang's method. However, if we take the boundary corrections generated by g on the two sides by half them adding to the kernel density, using similar approximation steps of derivation as proved in Zhang's work and according to our computation, we will find that we can still get a certain amount of boundary corrections on the two endpoints although such a correction effect is slightly weaker than the one done in the interval which has only one endpoint. By the choice of kernel density, the same cross-validation steps of choosing the bandwidth were also applied. With these corrections, the connected kernel density appears to be smoother and closer to the true density (Fig. 6). Although, there are other boundary correction methods, such as utilizing beta kernel (Chen 1999) or Probit-transformation on data in (0, 1), while performing the kernel density, either the computation is more complicated, or the correction effects are not quite clearly

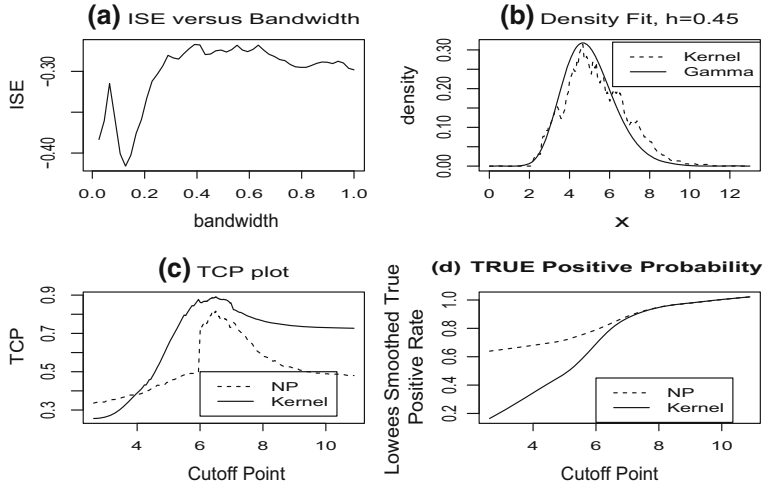

Fig. 6 Estimated TCP, smoothed TPR: solid curves are solutions adjusting probabilities using boundary-corrected kernel densities, and dashed curves are from the naive non-parametric estimation

applicable for our case. Besides, the distribution-adjusted TCP with our modified boundary-corrected kernel density is closer to the distribution-adjusted TCP based on the Gamma distribution that is estimated from the conditional likelihood.

Example 3: Kernel density on mixtures of Normal

In this example, we assume that $X \sim 0.3N(3,1) + 0.4N(6,3) + 0.3N(9,1)$ and the variable Y is associated to X according with the logistic function as per the same Eq. (61). The sample is generated with disproportional stratification partitioning the support of X by $J_1 = [0,6)$, $J_2 = [6,7)$, $J_3 = [7,\infty)$ and from each J_s, the sample sizes are $n_1 = 100, n_2 = 250$ and $n_3 = 200$. To compute the TCP and TPR, as we have previously done for Example 1 and 2, we shall first estimate the density for the distribution. In this case, we estimate $f(x)$ by a kernel density (1) without boundary corrections and (2) with boundary corrections. The results are summarized in the following Figs. 7, 8. In fact, in this example, either with boundary correction or without boundary correction, all the estimates of kernel density, TCP and TPR are quite similar. Similarly, we have gotten better TCP, and TPR that are close to the truth, while using the fitted distribution in calculating the TCP and TPR.

5 Real example

Currently, glycated haemoglobin (A1c) is considered to be the critical marker of metabolic control in diabetic patients, providing a cumulative measurement of the blood glucose concentration over the preceding 2nd-3rd months. However, blood samples should be taken regularly from the patients and the level of A1c has to be obtained

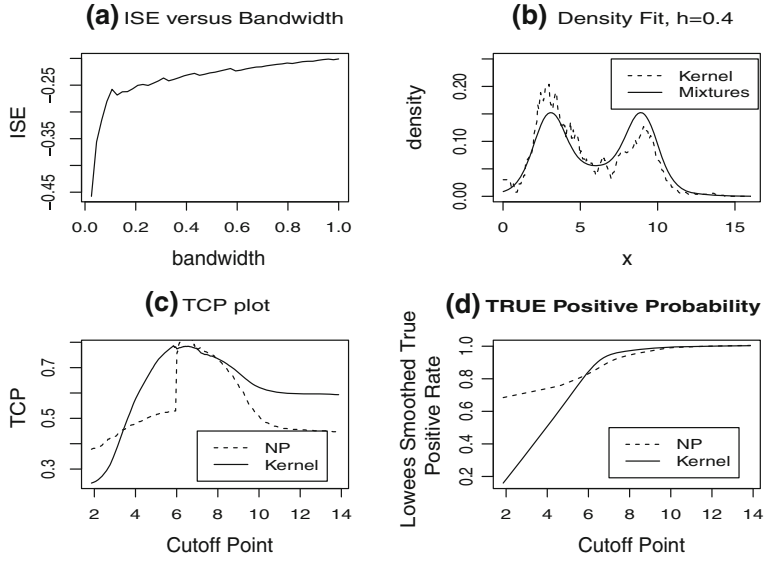

Fig. 7 Estimated TCP, smoothed TPR: solid curves are solutions adjusting probabilities using kernel densities, and dashed curves are from the naive non-parametric estimation. Assume the underlying distribution is a mixture of normal

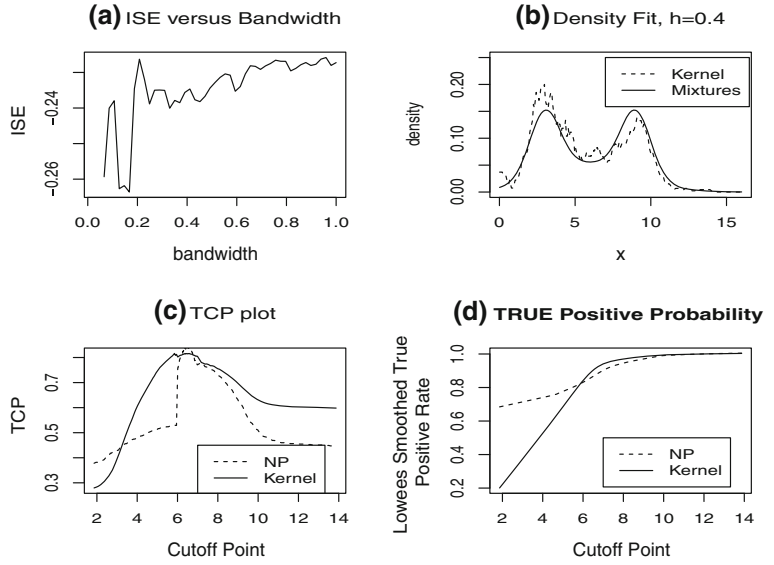

Fig. 8 Estimated TCP, smoothed TPR: solid curves are solutions adjusting probabilities using boundary-corrected kernel densities, and dash curves are from the naive non-parametric estimation. Assume that the underlying distribution is a mixture of normal

from hightech machines in a medical center. In general, one patient is diagnosed as diabetic if his/her A1c level is detected to be higher than 6.5%. A personal rapid test kit is a newly invented device for use at home. With the kit, the patient can simply

read positive/negative outcomes that indicate whether the A1c level has exceeded a critical level. In this clinical study, the goal is to estimate and determine an accurate cut-off point, serving as a threshold value that optimally divides the +/- outcomes on the test kit. This clinical study was conducted at the Chung Shan Medical University Hospital, Taiwan, where a total of 220 of participants, including diabetic patients and healthy individuals, were sampled during their revisits or regular medical examination. We consider all the individuals visiting the hospital during the months as our study population. The sample scheme was set up as following: based on the readings of the machine (rounded up to the first decimal place), subjects with the resultant A1c levels 0–5.9%, 6.0–6.9%, and $\geq 7.0\%$ were randomly sampled subsequently. At the end, we had 94 individuals with A1c levels between 0 and 5.9%, 77 individuals with A1c levels between 6.0 and 6.9% and 49 individuals with A1c levels exceeding 7.0%. The participants drawn at each of the three ranges of A1c levels (0–5.9%, 6.0–6.9%, 7.0% and above) is an independent random sample. Based on literatures on diabetes and primary analysis through preclinical study, we assume that A1c levels of the study population has a Gamma distribution $Gamma(\alpha, \beta)$, where β is the scale parameter. Let $X =$ A1c level and Y be the binary outcome from the test kit for each patient. Expressing Gamma in a density in the exponential family, the sufficient statistics are $T_1 = X$ and $T_2 = \log(X)$ with respect to the parameter $= (\alpha, 1/\beta)$ in its canonical form. To solve the score functions in Eq. (9), we have composite MLE $\widehat{\boldsymbol{\eta}} = (23.9857, 0.3058)$ and $(\widehat{\alpha}, \widehat{\beta}) = (23.9857, 3.2697)$. The asymptotic variance of $\boldsymbol{\eta}$ takes the form

$$\sum_{s=1}^{3} n_s \left\{ \begin{pmatrix} E_{J_s}\left[(\log(X))^2\right] & E_{J_s}\left[X\log(X)\right] \\ E_{J_s}\left[X\log(X)\right] & E_{J_s}\left[X^2\right] \end{pmatrix}^{-1} \right\} \doteq \begin{pmatrix} 0.2653 & 0.0000 \\ 0.0000 & 0.0212 \end{pmatrix} \tag{64}$$

where $\doteq$ represents the right hand side of the equation that is the estimate of variance matrix by plugging in MLEs. By apply the Delta method, the asymptotic variance of $(\widehat{\alpha}, \widehat{\beta})$ is found to be $(0.2653, 0.0002)$. The estimated total correct probability (TCP) and true positive rate at various choices of cutoff points are shown in Fig. 9. From the figure, the optimal cutoff point (circled point) seems quite close in both methods, but the TPR probabilities at different points have large deviations. We also apply the boundary-correct kernel density to calculate distribution-adjusted TCP and TPR, the

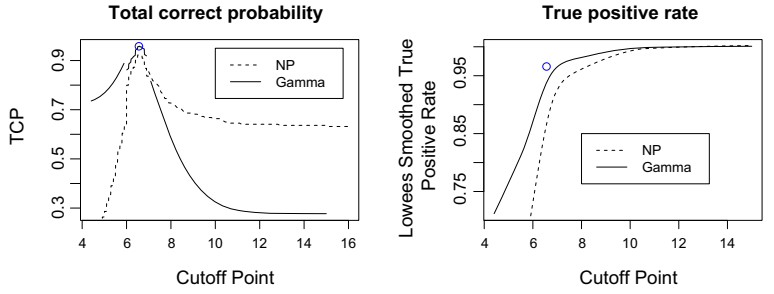

Fig. 9 Estimated TCP, smoothed TPR using the estimated Gamma density of input variable for A1c data

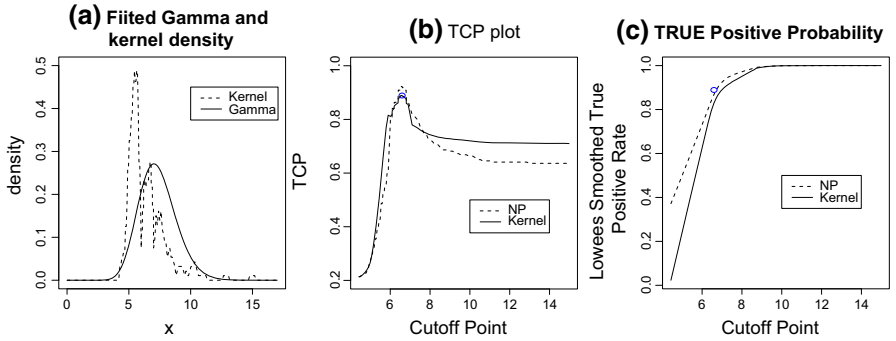

Fig. 10 Estimated TCP, smoothed TPR using the estimated kernel density of input variable for A1c data

results are shown in Fig. 10. The results differ from those obtained in Fig. 9, and the results in Fig. 10 seem to be more rationale.

6 Conclusions

We consider the problem of constructing a binary surrogate, while attempting to use a binary variable Y to substitute a continuous X, with paired data (X, Y) sampled respectively from the partitioned intervals with the support of X. By minimizing the expected prediction error of the data, we estimate the cutoff value c that divides the values of X optimally to match the binary response Y in $(0, 1)$ of interest. We propose a method to estimate the expected prediction error by summing the probability $P(Y = 1|X \geq c)P(X \geq c)$, where we estimate the probability $P(X \geq c)$ with a fitted distribution of X. Our method brings advantages and solutions to the problem. First, it is suspicious that existing methods such as logistic regression, change point method or naive nonparametric method do not take into account possible heterogeneous variances of different strata and the yielding prediction errors or TPRs would be doubtful. Accordingly, we use estimated EPE adjusting for the stratum probability weights. In particular, to circumvent unknown stratum probabilities, we need to estimate distribution of X. However, we found no existing method in the literature to estimate the unknown stratum weights under such a sampling scheme. The estimation of distribution with such stratified samples has ever been discussed by Vardi (1985), Gilbert (1988, 1999) and other recent articles, where non-parametric MLE or kernel densities were proposed, but these estimates are based on an enlarged data set to which another extra portion of data drawn across from the support were added. Our approach provides a novel estimation, requiring no such extra portions. We estimate the distribution of X assumed in the exponential family with composite likelihood where asymptotic results were also presented. Non-parametrically, we establish a kernel density estimates for data that may have other distribution features. From simulation studies and real data analysis, we found that our estimates of the distribution were close to the true one and the optimal cutoff value had adequate true positive rates when compared with the results from the naive non-parametric method. From our proposed methods, further

improvements are possible: (1) for kernel density, use precise boundary corrections on the boundaries between strata (2) Estimate the stratum weights under other parametric families. For either of these two directions of improvements, we need to both further investigate and leave them for future work.

Supplementary material

Supplement to "Binary surrogates with stratified samples when weights are unknown" (gamma-whole-sim-combine-2018Jun.R, gamma-epf-plot.R, gamma-ROC.R, FN-simulation.R, KD-prob.R, KD-boundary). The supplement to this paper contains R codes to perform numerical results in Sects. 4 and 5.

References

Beskos A, Papaspiliopoulos O, Roberts G (2009) Monte carlo maximum likelihood estimation for discretely observed diffusion processes. Ann Stat 37:223–245

Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360

Buyse M, Molenberghs G, Paoletti X, Oba K, Alonso A, der Elst W, Burzykowski T (2016) Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biom J 58:104–132

Chan K, Ledolter J (1995) Monte carlo em estimation for time series models involving counts. J Am Stat Assoc 90:242–252

Chen SX (1999) Beta kernel estimators for density functions. Comput Stat Data Anal 31:131–145

Contal C, O'Quigley J (1999) An application of changepoint methods in studying the effect of age on survival in breast cancer. Comput Stat Data Anal 30:253–270. ISSN 0167-9473. https://doi.org/10.1016/S0167-9473(98)00096-6

Cortes C, Mohri M, Riley M, Rostamizadeh A (2008) Sample selection bias correction theory. In: Algorithmic learning theory, Springer, Berlin, pp 38–53

Cox DR, Reid N (2004) A note on pseudolikelihood constructed from marginal densities. Biometrika 91:729–737

da Silva GT, Klein JP (2011) Cutpoint selection for discretizing a continuous covariate for generalized estimating equations. Comput Stat Data Anal 55:226–235. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2010.02.016

Ferrier S, Watson G, Pearce J, Drielsma M (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. i. Species-level modelling. Biodivers Conserv 11:2275–2307

Fokianos K (2004) Merging information for semiparametric density estimation. J R Stat Soc Ser B (Stat Methodol) 66:941–958

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics Springer, Berlin

Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. Interface Foundation of North America

Geyer CJ (1994) On the convergence of monte carlo maximum likelihood calculations. J R Stat Soc Ser B (Methodol) 56:261–274

Gilbert PB (2000) Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. Ann Stat 28:151–194

Gill RD, Vardi Y, Wellner JA (1988) Large sample theory of empirical distributions in biased sampling models. Ann Stat 16:1069–1112

Godambe V (1976) Conditional likelihood and unconditional optimum estimating equations. Biometrika 63:277–284

Heckman JJ (1979) Sample selection bias as a specification error. Econom J Econom Soc 47:153–161

Jones MC, Marron JS, Sheather SJ (1996) A brief survey of bandwidth selection for density estimation. J Am Stat Assoc 91:401–407

Lausen B, Schumacher M (1996) Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. Comput Stat Data Anal 21:307–326. ISSN 0167-9473. https://doi.org/10.1016/0167-9473(95)00016-X

Lindsay B (1982) Conditional score functions: some optimality results. Biometrika 69:503–512

Lindsay BG (1988) Composite likelihood methods. Contemp Math 80:221–239

Liu, A. and Ziebart, B. (2014). Robust classification under sample selection bias. In: Advances in neural information processing systems, pp 37–45

Martsynyuk YV (2012) Invariance principles for a multivariate student process in the generalized domain of attraction of the multivariate normal law. Stat Probab Lett 82:2270–2277

Richards JW, Starr DL, Brink H, Miller AA, Bloom JS, Butler NR, James JB, Long JP, Rice J (2012) Active learning to overcome sample selection bias: application to photometric variable star classification. Astrophys J 744:192

Sheather SJ (2004) Density estimation. Stat Sci 19:588–597

Vardi Y (1985) Empirical distributions in selection bias models. Ann Stat 13:178–203

Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Stat Sin 21:5–42

Wang B, Sun J (2009) Inferences from biased samples with a memory effect. J Stat Plan Inference 139:441–453

Wang B, Wang X (2007) Bandwidth selection for weighted kernel density estimation. arXiv preprint arXiv:0709.1616

Wang J-F, Li X-H, Christakos G, Liao Y-L, Zhang T, Gu X, Zheng X-Y (2010) Geographical detectors-based health risk assessment and its application in the neural tube defects study of the heshun region, china. Int J Geogr Inf Sci 24:107–127

Wang J-F, Zhang T-L, Fu B-J (2016) A measure of spatial stratified heterogeneity. Ecol Indic 67:250–256

Wei GC, Tanner MA (1990) A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. J Am Stat Assoc 85:699–704

Wu CO (1997) A cross-validation bandwidth choice for kernel density estimates with selection biased data. J Multivar Anal 61:38–60

Zadrozny B (2004) Learning and evaluating classifiers under sample selection bias p 114

Zhang S, Karunamuni R, Jones M (1999) An improved estimator of the density function at the boundary. J Am Stat Assoc 94:1231–1240