Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-021-00444-9

REGULAR ARTICLE

Estimating the class prior for positive and unlabelled data via logistic regression

Małgorzata Łazęcka[1,2] · Jan Mielniczuk[1,2] · Paweł Teisseyre[1,2]

Received: 23 July 2020 / Revised: 3 February 2021 / Accepted: 17 May 2021

Abstract
In the paper, we revisit the problem of class prior probability estimation with positive and unlabelled data gathered in a single-sample scenario. The task is important as it is known that in positive unlabelled setting, a classifier can be successfully learned if the class prior is available. We show that without additional assumptions, class prior probability is not identifiable and thus the existing non-parametric estimators are necessarily biased in general if extra assumptions are not imposed. The magnitude of their bias is also investigated. The problem becomes identifiable when the probabilistic structure satisfies mild semi-parametric assumptions. Consequently, we propose a method based on a logistic fit and a concave minorization of its (non-concave) log-likelihood. The experiments conducted on artificial and benchmark datasets as well as on a large clinical database MIMIC indicate that the estimation errors for the proposed method are usually lower than for its competitors and that it is robust against departures from logistic settings.

Keywords Positive unlabelled learning · Class prior estimation · Logistic regression · Non-convex optimisation · Minorization-maximization algorithm

Mathematics Subject Classification 62H30 · 62J12

✉ Paweł Teisseyre
Pawel.Teisseyre@ipipan.waw.pl

Małgorzata Łazęcka
Malgorzata.Lazecka@ipipan.waw.pl

Jan Mielniczuk
Jan.Mielniczuk@ipipan.waw.pl

1 Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

2 Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland

Published online: 03 June 2021

1 Introduction

Positive and unlabelled (PU) learning focuses on the setting where the data contains labelled positive examples and unlabelled ones. Unlabelled examples can be either positive or negative. In this setting, the true class label $Y \in \{0, 1\}$ is not observed directly. We only observe the surrogate variable $S \in \{0, 1\}$, which indicates whether an example is labelled (and thus positive; $S = 1$) or unlabelled ($S = 0$). This type of data naturally arises in many applications (see Bekker and Davis 2020; Jaskie and Spanias 2019 for reviews). Below we describe some illustrative examples. As a first example, consider the problem of predicting disease based on patient characteristics. Medical databases usually list only diagnosed diseases. However, many diseases, such as hypertension or diabetes, are often undiagnosed (Walley 2018). Therefore, the absence of the diagnosis does not mean that the patient does not have the disease in question. Consequently, we can distinguish three groups of patients: patients with the diagnosed disease ($S = 1$ and thus $Y = 1$); patients without diagnosed disease who have the disease ($S = 0$ and $Y = 1$) and finally patients without diagnosed disease who really do not have the disease ($S = 0, Y = 0$). Importantly, it is not possible to make a distinction between the second and the third group using observed data. Secondly, PU data occur frequently in text classification problems (Liu et al. 2003; Fung et al. 2006; Li and Liu 2003). For example, when classifying web page preferences, some web pages can be bookmarked as positive ($S = 1$) by the user whereas all other pages are treated as unlabelled ($S = 0$). Among unlabelled pages ($S = 0$), one can find both positive and negative pages. Thirdly, PU data stems from under-reporting (Sechidis et al. 2017) which frequently happens in survey data, and it refers to the situation when some respondents fail to answer a question truthfully. Under-reporting may occur, e.g. when the question concerns dangerous or unlawful behaviour such as taking illicit drugs (Bahorik et al. 2014; Chen et al. 2006). So in surveys, one group of respondents may admit such behaviours truthfully ($S = 1 \implies Y = 1$) and the other group do not ($S = 0$). The second group consists of respondents who have engaged in dangerous behaviours but do not report them ($Y = 1, S = 0$) and those who have nothing to report ($Y = 0, S = 0$). Other examples include modelling wildlife habitat selection (Ward et al. 2009; Pearce and Boyce 2006), detection of causative genes for various human diseases (Yang et al. 2014) and predicting drug-target interactions (Lan et al. 2016).

There are many partial observability schemes related to PU learning (we refer to Section 8 in Bekker and Davis (2020) for detailed discussion, see also Menon et al. (2015)). Semi-supervised learning is a general related scenario in which the goal is to learn from labelled and unlabelled data but, in contrast to PU learning, labeled examples from both classes are assumed to be present in the data (Chapelle et al. 2010). PU setting can be also seen as a special case of more general problem of learning from noisy labels (Natarajan et al. 2013; Frenay and Verleysen 2014) when labels are incorrectly assigned. In such general scenario, value of the true class variable Y can be flipped with some probability, i.e. instead Y we observe $S = 1 - Y$. Probabilities of incorrect assignment are: $\omega_1 := P(S = 0|Y = 1)$ and $\omega_2 := P(S = 1|Y = 0)$. Obviously, this problem reduces to PU setting for $\omega_2 = 0$. Even more general scenario is learning from two contaminated distributions (Scott et al. 2013). Finally, PU learning

can be also seen as a special case of 'coarse data' analysis (Heitjan and Rubin 1991; Couso et al. 2017), which covers situations where one does not have access to the exact value of the class variable Y, but only to some subset of the possible values of Y that contains it.

It is known that in PU learning, a classifier can be successfully learned if the class prior $\alpha = P(Y = 1)$ is available (Bekker and Davis 2020; Elkan and Noto 2008). More precisely, knowledge of the class prior can be used in three ways. The first approach is represented by so-called post-processing methods which first train classifier using S as a surrogate variable for Y (which is called non-traditional classification or naive classification) and then modify output probabilities using the class prior (Elkan and Noto 2008). The second approach are pre-processing methods that weigh the examples using the class prior (Steinberg and Cardell 1992; Lancaster and Imbens 1996; Kiryo et al. 2017). We refer to Bekker and Davis (2020) (Section 5.3.2) for a description of a general empirical risk minimization framework in which the weights of observations depending on α are determined for any loss function. In the third approach, class prior is incorporated into learning algorithms. A representative algorithm from this group is POSC4.5 (Denis et al. 2005), which is PU tree learning method.

The class prior is usually not known (except from situations when, for example, disease prevalence is known or can be learnt from other studies) and therefore the problem of its estimation from PU data has attracted significant attention (Elkan and Noto 2008; Jain et al. 2016; Plessis et al. 2017; Bekker and Davis 2018). There are three key contributions of the present paper concerning this problem. First, we formally analyse the problem of prior estimation; we show that in general this problem is ill-defined in non-parametric setting, i.e. class prior is not identifiable without some assumptions on conditional distribution of Y given X. We show however that the class prior becomes identifiable when we impose mild semi-parametric model assumptions on conditional distribution of Y given X. Secondly, we analyse in detail the most popular existing methods: the classical method EN proposed by Elkan and Noto (2008), TIcE algorithm (Bekker and Davis 2018), Partial Matching (Plessis et al. 2017), KM estimators (Ramaswamy et al. 2016) and MLR (Jaskie et al. 2020). We formally show that in some situations, some of the above methods underestimate label frequency $c = P(S = 1|Y = 1)$ and thus overestimate class prior. Finally, we consider the method based on logistic regression which allows to estimate label frequency and parameters of the logistic model simultaneously. The method (called JOINT method) was proposed in recent work (Teisseyre et al. 2020), in the more general context of estimation of the posterior for PU data. Its main limitation lies in necessity of the optimization of the non-concave log-likelihood function. Teisseyre et al. (2020) used simple gradient method, which may fail in some situations. Here we propose a novel procedure, called CD+MM, that combines cyclic coordinate descent (CD) and minimization-majorization (MM) algorithms and allows for more stable and efficient optimization. The method is based on a simple but consequential fact that the log-likelihood treated as function of logistic parameters is bounded from below by a concave function. Indeed, the experiments indicate that CD+MM outperforms other methods, including JOINT method, with respect to estimation error of the class prior.

This paper is organized as follows. In Sect. 2 we introduce notation and basic assumptions. In Sect. 3 we discuss identifiability of class prior; in Sect. 4 we analyse

the existing methods; Sect. 5.1 introduces the novel method, Sect. 5.2 compares it with MLR method (Jaskie et al. 2020), Sect. 6 summarizes the results of numerical experiments and Sect. 7 concludes the paper.

2 Notation and assumptions

We first introduce basic notations. Let $X \in \mathcal{X}$ be a feature vector, $Y \in \{0, 1\}$ be a true class label and $S \in \{0, 1\}$ an indicator of whether an example is labelled ($S = 1$) or not ($S = 0$). We assume that there is some unknown distribution $P(Y, X, S)$ such that (y_i, x_i, s_i), $i = 1, \ldots, n$ is i.i.d. sample drawn from it and data (x_i, s_i), $i = 1, \ldots, n$, is observed. Only positive examples ($Y = 1$) can be labelled, i.e. $P(S = 1|X, Y = 0) = 0$. Thus we know that $Y = 1$ when $S = 1$ but when $S = 0$, Y can be either 1 or 0 . As the aim is to learn the distribution of (X, Y) and we only observe samples from distribution of (X, S), where $S = Y$ with a certain probability, this is a partial observability scenario similar e.g. to a right censoring scheme when Y is observed provided its value is smaller than value of a censoring variable. In this work we also adopt a commonly used assumption called SCAR (Selected Completely At Random) which states that labelled examples are selected randomly from a set of positives examples, independently from X, i.e.

$$P(S = 1|Y = 1, X) = P(S = 1|Y = 1). \tag{1}$$

Parameter $c := P(S = 1|Y = 1)$ is called the label frequency and plays an important role in PU learning. In particular it is closely related to the class prior $\alpha = P(Y = 1)$ Elkan and Noto (2008), i.e. we have

$$\alpha = P(Y = 1) = P(S = 1)/c. \tag{2}$$

The probability $P(S = 1)$ can be directly estimated as the fraction of labelled examples. In the proposed method we first estimate c and then use (2) to estimate α, the similar approach was used e.g. in Bekker and Davis (2018). Under SCAR assumption we have the following property

$$P(S = 1|X) = cP(Y = 1|X), \tag{3}$$

which will be directly used in the proposed method.

In the paper we also take advantage of a representation of variable (X, S) when SCAR assumption is valid, introduced recently in Teisseyre et al. (2020). Namely, we have shown that S can be represented as

$$S = Y \cdot \varepsilon, \text{ where } \varepsilon \perp (X, Y) \text{ and } \varepsilon \sim Bern(1, p) \tag{4}$$

(with $\perp$ denoting independence), for a certain $0 < p < 1$ and where $Bern(1, p)$ stands for Bernoulli distribution. Indeed, we have $S = Y\varepsilon \perp X$ given Y, as $\varepsilon \perp (X, Y)$

implies that $\varepsilon \perp X$ given Y. Moreover,

$$P(S = 1|Y = 1) = P(Y\varepsilon = 1|Y = 1) = P(\varepsilon = 1) = p.$$

Thus probability of success $P(\varepsilon = 1)$ coincides with c. Moreover, it easily follows from the fact that $\varepsilon \perp X$ given $Y = 1$ that $P(X|Y = 1) = P(X|S = 1)$ and thus conditional distribution of X given $Y = 1$ is identifiable.
We stress that the scenario we consider here, namely that i.i.d. sample (x_i, s_i) is observed, called single sample scenario should be distinguished from case-control scenario when one sample is drawn from distribution $P(X|Y = 1)$ and the second one is drawn independently from $P(X)$. The differences between these two schemes are discussed in (Bekker and Davis 2020).

3 Identifiability of class prior

We start by stating an intuitive fact that neither c or α is identifiable for a PU single sample scenario given a full knowledge of the distribution of (X, S), if no assumptions on distribution of (X, Y) are imposed. This was already noticed for PU case control scenario in Ward et al. (2009), see also Scott (2015). In the present case of single sample scenario it follows from noting that distribution of X is mixture of two distributions

$$p(x) = \alpha c p_l(x) + (1 - \alpha c) p_u(x),$$

where $p_l(\cdot)$ and $p_u(\cdot)$ are densities or distribution mass functions of conditional distributions of X given $S = 1$ and $S = 0$, respectively. Note that $\alpha c = P(S = 1)$. We have that $p_l(\cdot)$ coincides with density of X given $Y = 1$ (see Sect. 2), whereas p_u equals

$$p_u(x) = \frac{p(x) - \alpha c p_l(x)}{1 - \alpha c} = \frac{p(x) - \alpha c p(x|Y = 1)}{1 - \alpha c}. \tag{5}$$

Thus mixing proportion is αc and distributions $P(X|S = i)$ of labelled and unlabelled mixture components depend solely on distributions $P(X)$, $P(X|Y = i)$ and αc. This means that changing α and c in such a way that their product is constant we obtain the same distribution (X, S), however the situations when α is small and c is large or otherwise are very different. Thus neither α or c is identifiable.
The situation changes dramatically when we impose semi-parametric model assumptions on conditional distribution of Y given X. We consider so called single index model (see e.g. Ichimura 1993) and prove directly that the parameters of the corresponding model for aposteriori probability of Y are identifiable. This in turn implies Fisher consistency of the likelihood method (part (ii)) which is the theoretical underpinning of empirical likelihood optimisation considered in this paper. In Remark 1 below we discuss that the considered model satisfies the positive function condition which is one of the identifiability conditions considered in Bekker and Davis (2020).

Theorem 1 *(i) Let $f(s)$ be an arbitrary strictly monotone response function with values in $[0, 1]$ such that $P(Y = 1|X = x) = f(\beta^T x)$, for some $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. In addition assume that there exists $i \neq 0$, such that $\beta_i \neq 0$ (i.e. Y and X are not independent). If*

$$cf(\beta^T x) = \tilde{c} f(\tilde{\beta}^T x)$$

for all $x \in R^{p+1}$ where $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p)$, then $c = \tilde{c}$ and $\beta = \tilde{\beta}$.
(ii) Assume that (X, Y) is as in (i), $c = P(S = 1|Y = 1)$ and let $p_{\tilde{c},\tilde{\beta}}(x, 1) = P(X = x, S = 1) = \tilde{c} f(\tilde{\beta}^T x) p(x)$, $p_{\tilde{c},\tilde{\beta}}(x, 0) = P(X = x, S = 0) = 1 - p_{\tilde{c},\tilde{\beta}}(x, 1)$. Then the maximiser of the expected loglikelihood

$$\mathbb{E}_{X,S}\{\log p_{\tilde{c},\tilde{\beta}}(X, S)\}$$

with respect to $(\tilde{c}, \tilde{\beta})$ is unique and equals (c, β).

Part (i) claims that when $P(Y = 1|x)$ is logistic then parameters determining $P(S = 1|x)$, that is c and β are uniquely determined. This is proved by assuming the contrary, checking the stated equality for the specific values of x and showing that it forces two possibly different sets of parameters to coincide. In part (ii) it is shown that when the expected likelihood is maximised, the maximisers correspond to the true parameters of generating mechanism, which are uniquely determined in view of (i).

Proof (i) Note that β_0 is the intercept corresponding to the first coordinate of x which is equal 1. We first consider a situation when there exists an index $i \neq 0$ such that $\beta_i \neq \tilde{\beta}_i$. Plugging the value $x := (1, 0, \ldots, 0, (\tilde{\beta}_0 - \beta_0)/(\beta_i - \tilde{\beta}_i), 0, \ldots, 0)$ (with $(i+1)$th coordinate equal $(\tilde{\beta}_0 - \beta_0)/(\beta_i - \tilde{\beta}_i)$) into the assumed functional equality we obtain $c = \tilde{c}$ as $f(\beta^T x) = f(\tilde{\beta}^T x) \neq 0$. Taking $x := (1, 0, \ldots, 0)$ yields $\beta_0 = \tilde{\beta}_0$. Considering $x = (1, 0, \ldots, 1, 0, \ldots, 0)$, with the first and the $(i + 1)$th coordinate equal to 1, we obtain $\beta_i = \tilde{\beta}_i$, a contradiction. Thus it is enough to consider the equality

$$cf(\beta_0 + \beta_{-0}^T x_{-0}) = \tilde{c} f(\tilde{\beta}_0 + \beta_{-0}^T x_{-0}),$$

where $\beta = (\beta_0, \beta_{-0}^T)^T$. It follows from assumption that we can take coordinate x_i such that $\beta_i \neq 0$ and $i > 0$. Now considering the sequence $x^{(n)} = (1, 0, \ldots, 0, sign(\beta_i) \times n, 0, \ldots, 0)$, with $sign(\beta_i) \times n$ at the $(i + 1)$th place, we obtain $\tilde{c} f(\infty) = cf(\infty)$ and thus $\tilde{c} = c$ again. Taking $x = (1, 0, \ldots, 0)$ yields now $\beta_0 = \tilde{\beta}_0$ and thus in a view of the first part of the proof $\beta = \tilde{\beta}$.
Part (ii) follows from properties of Kullback–Leibler divergence which imply that for any $X = x$ the conditional expected value $\mathbb{E}_{S|X=x}\{\log p_{\tilde{c},\tilde{\beta}}(x, S)\}$ is maximised by success probability equal $cf(\beta^T x)$ in view of Information Inequality (see Cover and Thomas (2006), Theorem 2.6.3). Then it follows from (i) that c and β are uniquely defined. □

Some remarks are in order.

Remark 1 The setting where $P(Y = 1|X = x) = f(\beta^T x)$ and f is some unknown monotone response function is very flexible semi-parametric model. However, its assumptions imply that the positive function property is valid which ensures identifiability (see Bekker and Davis 2020, Section 3.3 for taxonomy of identifiability assumptions and (Ramaswamy et al. 2016), Definition 9 for the precise statement of the positive function property called there separability condition with margin α). Namely, letting $A := \{x : \beta^T x > t\}$ where $t \in R$ is chosen such that $f(t) \geq 1/2 + \varepsilon$ for some $\varepsilon > 0$ and $h(x)$ being the indicator function of A, it is easy to see that

$$\mathbb{E}_{X|Y=1}h(X) \geq \Big(\frac{1}{2} + \varepsilon\Big)\frac{P(X \in A)}{P(Y = 1)} \geq \Big(\frac{1}{2} - \varepsilon\Big)\frac{P(X \in A)}{P(Y = 1)} \geq \mathbb{E}_{X|Y=0}h(X).$$

Thus separability condition with margin α holds with $\alpha = \varepsilon P(X \in A)/P(Y = 1)$ and $\beta = P(X \in A)/(2P(Y = 1))$.

Remark 2 Observe that in the above Theorem we assume that Y and X are not independent. Note that this is obviously necessary. Indeed, when Y is independent from X, it is not possible to infer anything about Y (and S) using X. In such situation the class prior cannot be identified using solely knowledge of S. Part (ii) states that when the model for (X, Y) is correctly specified then log-likelihood method based on (X, S) is Fisher consistent (Li and Duan 1989). In particular, the above result holds true for the logistic response function and also for f being the cumulative distribution function of the standard normal distribution which corresponds to the probit model. It is known that fitting logistic model is robust to misspecification of the response function under parametric assumptions on distribution of X (Li and Duan 1989; Mielniczuk and Teisseyre 2016) what suggests that estimation of c will also enjoy this property. This will be investigated for artificially generated data in Sect. 6. In Sect. 5.1 we focus on the case when the conditional distribution of Y given X is logistic.

4 Existing methods of class prior estimation: theoretical analysis

In this section we review and discuss all available (best to our knowledge) non-parametric methods of class prior estimation except methods based on parametric models including the new proposal which are discussed in the next section.

4.1 Elkan-Noto estimator

We first consider Elkan–Noto estimator of c (cf. Elkan and Noto (2008)) denoted by e_1 on p. 214 of their paper. It is introduced under tacit separability assumption stating that supports of conditional distributions of X given $Y = 1$ and $Y = 0$ are disjoint. The estimator is defined as follows

$$\hat{c}_{EN} = \frac{1}{|\mathcal{A}|}\sum_{i \in \mathcal{A}} \hat{P}(S = 1|X_i), \tag{6}$$

where $\mathcal{A}$ is a labelled part of a test sample $\mathcal{T}$, namely $\mathcal{A} = \{i : (X_i, S_i) \in \mathcal{T} : S_i = 1\}$ and estimator $\hat{P}$ of P is based on training sample $\mathcal{U}$ independent of $\mathcal{T}$. We show below that limit of $\hat{c}_{EN}$ is c under the separability condition but differs from it if the condition is not satisfied.

Lemma 1 *Assume that estimator $\hat{P}(S = 1|X = \cdot)$ is uniformly consistent i.e.*

$$\sup_x |\hat{P}(S = 1|X = x) - P(S = 1|X = x)| \to 0 \tag{7}$$

in probability then

$$\hat{c}_{EN} \to \frac{\mathbb{E}_X P^2(S = 1|X)}{P(S = 1)} \tag{8}$$

in probability when sample size $n = |\mathcal{T}|$ tends to infinity. When the separability assumption holds, the limit in (8) equals c, i.e. $\hat{c}_{EN}$ is consistent.

Intuitively, the proof exploits the idea that since $\hat{c}_{EN}$ is given as an average of random variables, Law of Large Numbers allows us to study its limit. Before we prove the result, we discuss its consequences. First, we note that if Y is independent of X and thus S is independent of X, we have that the limit in (8) equals $P(S = 1) \leq P(S = 1)/P(Y = 1) = c$. Moreover, considering representation $S = Y\varepsilon$, discussed in Sect. 2, it is easy to see that $c = P(Y\varepsilon = 1|Y = 1) = P(\varepsilon = 1)$ whereas

$$\frac{\mathbb{E}_X P^2(S = 1|X)}{P(S = 1)} = \frac{P^2(\varepsilon = 1)\mathbb{E}_X P^2(Y = 1|X)}{P(Y = 1)P(\varepsilon = 1)} = P(\varepsilon = 1)\frac{\mathbb{E}_X P^2(Y = 1|X)}{P(Y = 1)}.$$

Thus the estimator is not consistent in general and multiplicative bias is $\mathbb{E}P^2(Y = 1|X)/P(Y = 1)$. Note that it holds

$$\frac{\mathbb{E}_X P^2(Y = 1|X)}{P(Y = 1)} \leq \frac{\mathbb{E}_X P(Y = 1|X)}{P(Y = 1)} = \frac{P(Y = 1)}{P(Y = 1)} = 1$$

with the inequality above being strict in discrete case if for some x there is $0 < P(Y = 1|X = x) < 1$. In the separability case, when we have $P(Y = 0|X = x) = 1$ or $P(Y = 1|X = x) = 1$ for any x, inequality above becomes equality and the limit in (8) is c. When the separability condition does not hold, the bias of $\hat{c}_{EN}$ is always negative and increases with c. In particular, it is easily checked that in the case of Example 1 discussed below the multiplicative bias equals 0.68.

We also note that consistency assumption (7) in the Lemma is satisfied for discrete finite X when $|\mathcal{U}| \to \infty$ in view of weak Law of Large Numbers and in general case, as S is binary, from uniform consistency of nonparametric regression estimators (cf. e.g. Bierens 1983, for uniform consistency of kernel estimators).

Proof From consistency assumption (7) it easily follows that the limit of $\hat{c}_{EN}$ is the same as the limit

$$\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} P(S=1|X_i) = \frac{n}{n_1}\frac{1}{n}\sum_{i\in\mathcal{T}} P(S=1|X_i)I\{S_i=1\} =: I_1 \times I_2, \tag{9}$$

where $n_1 = |\mathcal{A}|$. Obviously $I_1 = n/n_1 \to P(S=1)^{-1}$ in probability and in view of the weak Law of Large Numbers

$$I_2 \to \mathbb{E}_{(X,S)}\Big(P(S=1|X)I\{S=1\}\Big).$$

We compute the above limit for discrete X, in the general case proof is similar.

$$\begin{aligned}\mathbb{E}_{(X,S)}\Big(P(S=1|X)I\{S=1\}\Big) &= \sum_{x\in\mathcal{X}} \frac{P(S=1,X=x)}{P(X=x)}P(S=1,X=x)\\ &= \sum_{x\in\mathcal{X}} \frac{P^2(S=1,X=x)}{P^2(X=x)}P(X=x)\\ &= \mathbb{E}P^2(S=1|X).\end{aligned} \tag{10}$$

From convergences of I_1 and I_2 the result readily follows. □

4.2 TIcE estimator

We now discuss recent estimator of c introduced in Bekker and Davis (2018). It relies on the observation that SCAR assumption i.e. conditional independence of X and S given Y implies that for any $A \subseteq \mathcal{X}$

$$c = P(S=1|Y=1) = P(S=1|X\in A, Y=1) \tag{11}$$

and the right hand side above is equal

$$\frac{P(S=1,X\in A,Y=1)}{P(X\in A,Y=1)} = \frac{P(S=1|X\in A)}{P(Y=1|X\in A)}. \tag{12}$$

Thus if A is such a set that $P(Y=1|X\in A)\approx 1$ (known as positive subdomain assumption and A is called an anchor set in Bekker and Davis (2020)) then c may be estimated as a fraction of observations with $S=1$ such that their concomitant X falls into A. This is an essence of the proposed method in which A is found using induction tree built on the training sample and $P(S=1|X\in A)$ is estimated using testing sample. Denote by $\hat{c}_{TIcE}$ the resulting estimator (TIcE standing for 'Tree Induction for c Estimation'). Such an approach will not yield satisfactory results if $P(Y=1|X\in A)$ is bounded away from 1 for any set A. Consider the following simple example.

Example 1 Let (X, Y) be such that both X and Y take values 0, 1 with probability 1/2 and

$$P(Y = 1|X = 1) = P(Y = 0|X = 0) = 0.8.$$

Thus the anchor set A in this case (taken as a set maximising $P(Y = 1|X \in A)$) equals $A = \{X = 1\}$ and for any S such that $S \perp X|Y$ the corresponding value of $c = c(S)$ will satisfy

$$c = \frac{P(S = 1|X = 1)}{P(Y = 1|X = 1)} = \frac{10}{8} P(S = 1|X = 1).$$

Divide data into training (i.e. 'tree' using Authors' terminology) and testing (i.e. 'estimation') data as in Bekker and Davis (2020) using (default) proportion 1:4 and let n be the total number of observations, $T_{est} = (4/5)n$ be the size of estimation data. Moreover, using notation from Bekker and Davis (2018), let $T_{est}(X = 1)$ be the expected number of observations from estimation data with $X = 1$ and

$$L_{est}(X = 1) = P(S = 1|X = 1)T_{est}(X = 1) = \frac{8}{10} \times \frac{1}{2} \times \frac{4}{5} \times n \times c$$

the expected number of observations from estimation data with $X = 1$ and $S = 1$. Using value of δ defined by equation (8) on p. 2714 in Bekker and Davis (2018), we can obtain 'ideal' value of $\hat{c}_{TIcE}$ from the equation

$$\hat{c} := \frac{L_{est}(X = 1)}{T_{est}(X = 1)} - \varepsilon, \tag{13}$$

where the correction (error) term equals to $\varepsilon = (\hat{c}(1-\hat{c})(1-\delta)/\delta T_{est}(X = 1))^{1/2}$ and is derived from the one-sided Chebyschev inequality. The 'ideal' value is the output value of the algorithm (cf. definition of c_{low} in Algorithm 1 of Bekker and Davis 2018) when an estimated anchor set $\{S_e : a^* = v\}$ in the algorithm is replaced by the true anchor set $\{X = 1\}$. Note that value of δ is given by (8) in Bekker and Davis 2018). Figure 1 shows values of 'ideal' $\hat{c}_{TIcE}$ obtained from (13) (red dots) and actual values of $\hat{c}_{TIcE}$ (light red dots) for 100 trials and for c ranging from 0 to 1 and $n = 10^3$ and $n = 10^4$. The performances of other methods discussed in the paper is also shown for comparison. KM2 estimator is discussed in Sect. 4.4 whereas MLR, JOINT and CD+MM methods are introduced in Sect. 5. It is seen that $\hat{c}_{TIcE}$ underestimates true c with pronounced bias for larger c and that bias is not negligible even for $n = 10^4$. The situation does not improve even with 'ideal' $\hat{c}_{TIcE}$ when the anchor set is assumed known. The reason for this behaviour is, that for any A, probability $P(Y = 1|X \in A)$ is significantly smaller than 1. The Elkan–Noto estimator performs similarly, on the other hand KM2 defined below has the smaller bias and the CD+MM method proposed in this paper is approximately unbiased in this case.

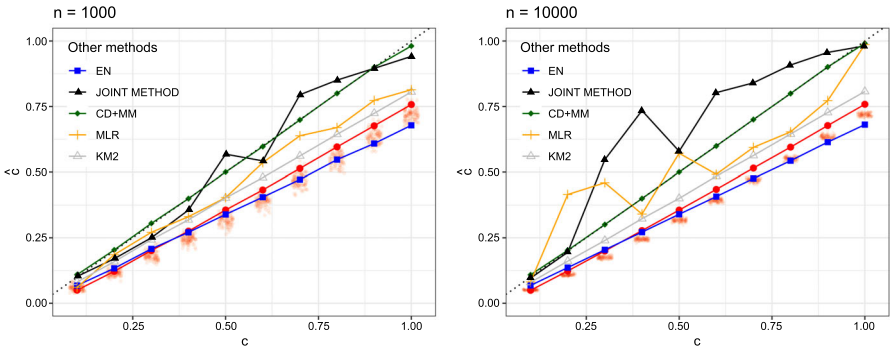

Fig. 1 Example 1. Values of $\hat{c}_{TIcE}$ (light red dots) for 100 trials and 'ideal' $\hat{c}_{TIcE}$ (red dots) for $n = 10^3, 10^4$ (colour figure online)

4.3 Partial matching

PE estimator proposed in du Plessis and Sugiyama (2014) is based on a partial matching in terms of the Pearson divergence defined as $PE(p(x), q(x)) = \int (p(x) - q(x))^2 / p(x)\, dx$ and is approximately unbiased under separability assumption. Namely, it is proposed to find the minimiser of $PE(a \times f(x|Y = 1), f(x))$ over $a > 0$ and then consider its estimator as an estimator of α. Note that $\alpha f(x|Y = 1) \approx f(x)$ only for such x that $(1 - \alpha) f(x|Y = 0) \approx 0$. It is easily derivable that the minimiser equals

$$\left(\int \frac{P^2(x|Y = 1)}{p(x)}\, dx \right)^{-1} = \frac{P(Y = 1)}{1 - (1 - P(Y = 1))A}, \tag{14}$$

where $A = \int p(x|Y = 1) p(x|Y = 0) / p(x)\, dx$ and the equality follows from (5). This quantity is estimable as both $p(x)$ and $p(x|Y = 1) = p(x|S = 1)$ are observable. The estimator is approximately unbiased under the separability assumption i.e. when the class-conditional densities have disjoint supports, as then $A = 0$. Otherwise, it is positively biased and thus suffers from intrinsic bias problem. We also note that the minimiser equals

$$\left(\frac{\int p(x) P^2(Y = 1|X = x)}{P^2(Y = 1)}\, dx \right)^{-1} = \frac{P^2(Y = 1)}{\mathbb{E} P^2(Y = 1|X)} = \alpha \left\{ \frac{\mathbb{E} P^2(Y = 1|X)}{P(Y = 1)} \right\}^{-1}$$

which shows that estimation of the minimiser will lead to biased estimator of α. Note also that the multiplicative constant above is the reciprocal of the multiplicative bias of $\hat{c}_{EN}$. In view of the correspondence between α and c (cf. (2)) this suggests close correspondence between PE estimator and Elkan-Noto estimator.

4.4 KM estimators (Ramaswamy et al. (2016))

Denote by F and H the distribution functions of X given $S = 0$ and $S = 1$, respectively and note that (5) implies that F can be represented as a mixture

$$F(x) = \frac{\alpha - \alpha c}{1 - \alpha c} H(x) + \frac{1 - \alpha}{1 - \alpha c} G(x),$$

where G is distribution of X given $Y = 0$. Following (Ramaswamy et al. 2016), denote the mixing proportion by $\kappa^* = (\alpha - \alpha c)/(1 - \alpha c)$ and note that estimator of α may be obtained from the equality $\alpha = \kappa^*(1 - P(S = 1)) + P(S = 1)$ once κ^* is estimated, as S is observable. The problem of estimating κ^* is approached by transforming PU data into Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ by appropriate function ϕ and solving this problem in $\mathcal{H}$. More specifically, it is shown that κ^* can be recovered by truncating either the distance function $d(\lambda) = \inf_{w \in \mathcal{C}} ||\lambda \phi(\hat{F}) + (1-\lambda)\phi(\hat{H}) - w||_{\mathcal{H}}$, where the set $\mathcal{C}$ consists of convex combinations of transformed data points and $||\cdot||_{\mathcal{H}}$ denotes the norm of $\mathcal{H}$, or the gradient of $d(\lambda)$. In this way two estimators KM1 and KM2 are obtained, which are shown to be consistent under appropriate conditions (Theorems 12 and 13 in Ramaswamy et al. 2016). The advantage of the results is that they hold provided the positive function condition is valid, which is weaker than the anchor set condition (see Ramaswamy et al. 2016, Section 4). Disadvantage is that the formal consistency result holds only when permissible truncation thresholds for d or its gradient depend on the unknown κ^* which we want to estimate. Thus the studied versions of estimators are not necessarily consistent. Theoretical comparison of KM1 and KM2 with other proposals seems out of reach at the moment. The relevant numerical analysis is provided in the following section. We stress that KM1 and KM2 are based on fully nonparametric approach whereas the method proposed here relies on the assumption that distribution of Y given X is logistic.

5 Estimating the class prior via logistic regression

5.1 CD+MM algorithm: description and its properties

In the following we introduce CD+MM algorithm which attempts to optimize log-likelihood function of PU data. We compare it with JOINT method introduced in Teisseyre et al. (2020) and discuss why the new method is beneficial. For standard supervised scenario, the logistic regression involves optimizing log-likelihood function

$$\sum_{i=1}^{n} [y_i \log(\sigma(x_i^T b)) + (1 - y_i) \log(1 - \sigma(x_i^T b))], \tag{15}$$

with respect to $b = (b_0, \ldots, b_p)$, where $\sigma(t) := \exp(t)/(1 + \exp(t))$ is sigmoid logistic function and posterior probability $P(Y = 1|X = x)$ is assumed equal to $\sigma(x^T b^*)$ for a certain $b^* \in R^{p+1}$. Obviously, for PU data, this approach is not feasible

as we do not observe Y and the loss function has to be based on S. To tackle the above problem, we can use equality $P(S = 1|X) = cP(Y = 1|X)$. In this context we mention (Jaskie et al. 2020) who used modified logistic function to approximate $P(S = 1|X)$. We write the observed log-likelihood function

$$L(c, b) = \sum_{i=1}^{n} [s_i \log(c\sigma(x_i^T b)) + (1 - s_i) \log(1 - c\sigma(x_i^T b))]. \tag{16}$$

The main idea of the proposed approach is to maximise the above function simultaneously with respect to b and c. This has been already investigated in Teisseyre et al. (2020) as JOINT Method which relied on simple gradient algorithm to maximise $L(c, b)$ where it is shown that such approach works on par or better than other competitors for estimating aposteriori probability $P(Y = 1|X = x)$. However, it is known that $L(c, b)$ is not a concave function jointly in both arguments (see Song and Raskutti 2020, p. 5). This can be seen by noting that the sum in (16) over $s_i = 0$ is not concave as a function of b and can dominate the remaining sum over $s_i = 1$ (which is concave) to that effect that $L(c, \cdot)$ will not be concave.[1] Thus, despite established good performance of JOINT Method, it may fail to find a global maximum of $L(c, b)$ by using gradient search. Below we introduce a different method of searching for maximizer of $L(c, b)$ and show this gives improvement for the problem of estimating class prior. It is based on Minorization–Maximisation (MM) algorithm (see e.g. Lange 2010) in which a (non-concave) criterion function is bounded from below by a concave function at each step of iteration procedure. Under mild conditions it is shown that the maximizers of the lower bounds in the consecutive iterations yield a non-decreasing sequence of criterion function values.

Define function $L_b(c) := L(c, b)$, to be profile log-likelihood function i.e. the log-likelihood treated as a function of c for fixed b. Below we prove its concavity by showing that its second derivative is non-positive.

Lemma 2 *Function $L_b(c)$ is concave with respect to c.*

Proof The proof follows from a simple calculation showing that

$$\frac{\partial^2}{\partial^2 c} L_b(c) = -\sum_{i=1}^{n} \left(\frac{s_i}{c^2} + \frac{(1 - s_i)\sigma^2(x_i^T b)}{(1 - c\sigma(x_i^T b))^2} \right) \leq 0$$

which implies that $L_b(\cdot)$ is concave. □

Define function $L_c(b) := L(c, b)$, i.e. profile log-likelihood function for fixed c. Moreover, let X be a $n \times p$ matrix of features, whose ith row is x_i^T; $v(x_i^T b) := \sigma(x_i^T b)(1 - \sigma(x_i^T b))$, Σ is $n \times n$ diagonal matrix with

$$\frac{v(x_i^T b)}{\sigma(x_i^T b)(1 - c\sigma(x_1^T b))} = \frac{1 - \sigma(x_i^T b)}{1 - c\sigma(x_i^T b)},$$

[1] We thank Wojciech Rejchel for pointing out this reasoning to us.

$i = 1, \ldots, n$ on the diagonal; $\mathbf{s} = (s_1, \ldots, s_n)^T$ and $\mathbf{p} = (c\sigma(x_1^T b), \ldots, c\sigma(x_n^T b))^T$. Gradient of $L_c(b)$ with respect to b is of the form

$$\nabla L_c(b) = X^T \Sigma(\mathbf{s} - \mathbf{p}). \tag{17}$$

We consider function

$$\Psi_c(b, b^0) := L_c(b^0) + (b - b^0)^T \nabla L_c(b^0) - \frac{1}{8}(b - b^0)^T X^T X(b - b^0).$$

Note that $\Psi_c(\cdot, b^0)$ is concave. The following Lemma gives the lower bound for function $L_c(b)$. The lower bound will serve as a proxy to be maximised in MM algorithm and it is obtained by bounding from below the second term in Taylor expansion of $L_c(b)$.

Lemma 3 *The following inequality holds: $L_c(b) \geq \Psi(b, b^0)$ for any vector b^0.*

Proof We denote by $L_{i,c}(b)$ the ith summand of (16) treated as a function of b. We first calculate a form of a second derivative of $L_{i,c}(b)$ using (17). Namely, noting that $\sigma'(t) = \sigma(t)(1 - \sigma(t)) = v(t)$, we have for $s_i = 1$ and $x_i = (x_{i1}, \ldots, x_{ip})^T$

$$\frac{\partial^2}{\partial b_j \partial b_k} L_{i,c}(b) = -x_{ij} x_{ik} v(x_i^T b)$$

and for $s_i = 0$, using the fact that $v'(t) = v(t)(1 - 2\sigma(t))$ we have

$$\begin{aligned}\frac{\partial^2}{\partial b_j \partial b_k} L_{i,c}(b) &= -c x_{ij} x_{ik} \Big[\frac{v(x_i^T b)(1 - 2\sigma(x_i^T b))(1 - c\sigma(x_i^T b)) + cv^2(x_i^T b)}{(1 - c\sigma^2(x_i^T b))^2} \Big] \\ &= -c x_{ij} x_{ik} v(x_i^T b) \frac{(c\sigma^2(x_i^T b) - 2\sigma(x_i^T b) + 1)}{(1 - c\sigma^2(x_i^T b))^2}.\end{aligned}$$

Observe that

$$\frac{c(c\sigma^2(x_i^T b) - 2\sigma(x_i^T b) + 1)}{(1 - c\sigma^2(x_i^T b))^2} = \frac{(1 - c\sigma(x_i^T b))^2 + c - 1}{(1 - c\sigma^2(x_i^T b))^2} \leq 1,$$

as $c \leq 1$ and $0 \leq \sigma(s) \leq 1$. Thus denoting by $H(b) = \nabla^2 L_c(b) = \Big(\frac{\partial^2}{\partial b_j \partial b_k} L_c(b) \Big)_{j,k}$ Hessian of $L_c(b)$ with respect to b and taking into account the inequality $v(t) \leq 1/4$ we have that

$$H(b) \geq -\frac{1}{4} X^T X, \tag{18}$$

where $' \geq'$ above denotes matrix ordering ($A \geq B$ when $A - B$ is a positive semi-definite matrix). For (18) we additionally used $X^T \Delta X \leq X^T X$ when Δ is a diagonal

matrix with all elements on the diagonal not larger then 1. Taylor expanding $L_c(b)$ around b_0 we have

$$L_c(b) = L_c(b^0) + (b - b^0)^T \nabla L_c(b^0) + \frac{1}{2}(b - b^0)^T H(\tilde{b}^*)(b - b^0)$$

where $\tilde{b}^*$ belongs to the interval $[b^0, b]$. Using inequality (18) for the Hessian with $H(b)$ replaced by $H(b^*)$ it follows that

$$(b - b^0)^T H(\tilde{b}^*)(b - b^0) \geq -\frac{1}{4}(b - b^0)^T X^T X (b - b^0),$$

what combined with the previous inequality gives the proof of the Lemma. □

Algorithm 1: Minorization-Maximization (MM) for optimizing $L_{c^0}(b)$

Input : Observed data (x_i, s_i), number of iterations T, convergence threshold ϵ, c^0.
Initialize: $\hat{b}^0 = (0, \ldots, 0)^T$,
for *t=1,2,...,T* **do**
  $\hat{b}^t = \arg\max_b \Psi_{c^0}(b, \hat{b}^{t-1})$
  if $\max_j |\hat{b}^{t-1}_j - \hat{b}^t_j| < \epsilon$ **then**
    break loop
Output : $\hat{b}^t$

The result below shows that in the consecutive iterations $t = 1, 2, \ldots$ the values of $L_{c^0}(\hat{b}^t)$ form a nondecreasing sequence. This in practical terms means its convergence to the local minimum.

Theorem 2 *Let* $\hat{b}^0 \in R^p$ *and* $\hat{b}^t = \arg\max_b \Psi_{c^0}(b, \hat{b}^{t-1})$ *for* $t \geq 1$. *Then*

$$L_{c^0}(\hat{b}^{t-1}) \leq L_{c^0}(\hat{b}^t).$$

Proof Proof of the Theorem follows from properties of Minorization-Maximization (MM) algorithm (see e.g. inequality (5.69) in Hastie et al. (2015) applied to the negative loglikelihood). Indeed, in view of Lemma 3 we have $L_c(b) \geq \Psi(b, b^0)$, moreover $L_c(b_0) = \Psi(b^0, b^0)$ from the definition of $\Psi(b, b^0)$ and $\Psi(\cdot, b^0)$ is concave. □

Theorem 2 justifies Algorithm 1 which for a given value of c^0 and $\hat{b}^{t-1}$ yields $\hat{b}^t$ such that $L_{c^0}(\hat{b}^{t-1}) \leq L_{c^0}(\hat{b}^t)$ by maximizing $\Psi_{c^0}(\cdot, \hat{b}^{t-1})$. Note here that maximization of $\Psi_{c^0}(\cdot, \hat{b}^{t-1})$ is fast as it is simple maximization of the quadratic function. This provides very plausible justification for applying MM algorithm in this case.
We now describe a novel algorithm CD+MM which combines MM algorithm with Cyclic Coordinate Descent (CD), see Algorithm 2 for the pseudo-code. The algorithm works as follows. We cyclically iterate through b and c, one at a time, maximizing the objective function with respect to each coordinate at a time. After $(t-1)^{\text{th}}$ iteration

in which $\hat{b}^{t-1}$ is obtained, $\hat{c}^t$ is sought by maximising function $L_{\hat{b}^{t-1}}(c)$ with respect to c (note that in the view of Lemma 2 this a concave function of c) and then $\hat{b}^t$ is obtained by maximising $L_{\hat{c}^t}(b)$, which is done using Algorithm 1.

Algorithm 2: Cyclic coordinate descent + Minorization-Maximization (CD+MM)

Input : Observed data (x_i, s_i)
Initialize: $\hat{b}^0 = (0, \ldots, 0)^T$, $\hat{c}^0 = 0.5$
for *t=1,2,...* **do**
  $\hat{c}^t = \arg\max_c L_{\hat{b}^{t-1}}(c)$
  $\hat{b}^t = \arg\max_b L_{\hat{c}^t}(b)$ # use MM Algorithm 1
Output : $\hat{c}^t$

5.2 MLR estimator and its comparision with JOINT method

The idea of MLR estimator (Jaskie et al. 2020) is as follows. First note that $c \leq \max_x P(S = 1|x)$ (and equality holds when $\max_x P(Y = 1|x) = 1$)) and thus c can be estimated as $\hat{c} = \max_x \hat{P}(S = 1|x)$, where $\hat{P}(S = 1|x)$ is some estimator of posterior probability. In MLR method, the following parametric model is used to estimate $P(S = 1|x)$

$$g_{MLR}(x, b, \gamma) = \frac{1}{1 + b^2 + \exp(-\gamma^T x)}, \tag{19}$$

where $b > 0$ and $\gamma \in R^{p+1}$. Estimator $(\hat{b}, \hat{\gamma})$ is obtained using gradient-based algorithms. Then noting that $\max_x g_{MLR}(x, b, \gamma) = 1/(1 + b^2)$ one considers $\hat{c} = 1/(1 + \hat{b}^2)$. Assume now, similarly to JOINT method, that aposteriori probability $P(Y = 1|x)$ is logistic, i.e. $P(Y = 1|x) = \sigma(\beta^T x)$. The following Lemma clarifies the relation between (c, β) and (b, γ).

Lemma 4 *Assume that* $P(Y = 1|x) = \sigma(\beta^T x)$, $\beta_i \neq 0$ *for at least one* $i \geq 1$ *and for certain* $(b, \gamma^T)^T$ *and all* $x \in R^{p+1}$ *it holds that*

$$P(S = 1|x) = c\sigma(\beta^T x) = g_{MLR}(x, b, \gamma). \tag{20}$$

Then we have

$$\gamma_0 = \ln c + \beta_0, \quad \gamma_{-0} = \beta_{-0}, \quad c = \frac{1}{1 + b^2}, \tag{21}$$

where $\beta^T = (\beta_0, \beta_{-0}^T)$ *and* $\gamma^T = (\gamma_0, \gamma_{-0}^T)$.

Proof Similarly to the Proof of Theorem 1, by choosing appropriate values of x we deduce the relations between parameters of the two competing models.

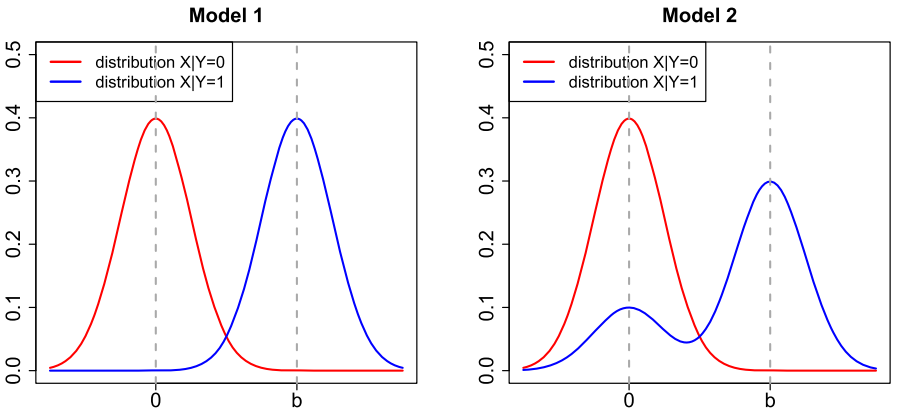

Fig. 2 Density functions of distributions $X|Y = 0$ and $X|Y = 1$ for simulation models 1 and 2

The proof easily follows from (20) after noting that for any $i \geq 1$ such that $\beta_i \neq 0$ (existence of such i is guaranteed by assumptions) and taking $x_n = (0, \ldots, -(\beta_i) \times n, 0 \ldots, 0)^T$, where non-zero element is placed at $(i + 1)$th coordinate, it implies that $P(S = 1|x_n)$ is equivalent to $c \exp(\beta_0 - |\beta_i| \times n)$ when n tends to infinity. Comparison with the rate of convergence of RHS of (21) for x_n yields that $\beta_i = \gamma_i$ and $c \exp(\beta_0) = \exp(\gamma_0)$. If $\beta_i = 0$ by similar reasoning we obtain $\gamma_i = 0$. Thus we obtain the first two desired equalities. The last one follows immediately. □

We note that although (21) suggests that $g_{MLR}(x, b, \gamma)$ yields equivalent parametrisation of $P(S = 1|X)$ to that given by (c, β), this is not true. This is due to subtle but crucial difference. Namely, in contrast to c and β which are independent parameters, b and γ are not algebraically independent in view of equality $\gamma_0 = -\ln(1 + b^2) + \beta_0$, where β_0 is an unknown constant. This entails that b and γ may not be treated as independent parameters while performing gradient-based optimization. In particular, the derivative of $g_{MLR}(x, b, \gamma)$ with respect to b is not $-2b/(1 + b^2 + \exp(-\gamma^T x))^2$. Inadequate estimation may be expected in particular for small c when the absolute value of γ_0 in the view of (21) becomes large. This is indeed confirmed by our numerical analysis in the next section.

6 Experiments

In the experiments we compare the performance of the proposed method CD+MM with the following methods: JOINT method (Teisseyre et al. 2020), TIcE (Bekker and Davis 2018), EN (Elkan and Noto 2008), MLR Jaskie et al. (2020), KM1, KM2 (Ramaswamy et al. 2016). We do not show the results for Partial Matching method (du Plessis and Sugiyama 2014) due to similarity with EN estimator discussed above. The source code of our method is available at https://github.com/teisseyrep/PU_class_prior.

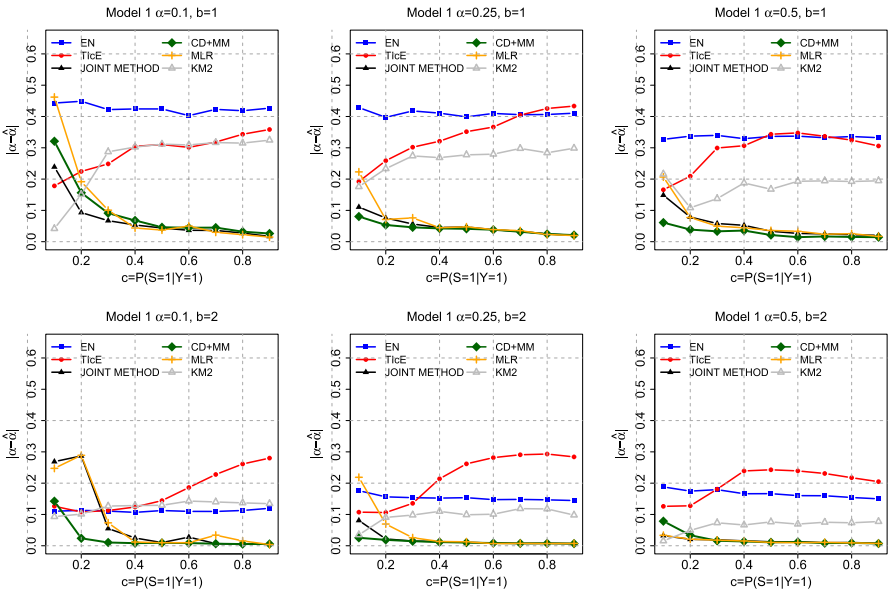

Fig. 3 Estimation error $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c for simulation model 1. Parameters: $b = 1, 2$, $\alpha = 0.1, 0.25, 0.5$ and $n = 5000$

6.1 Simulation models

We generate artificial data as follows. First Y is drawn from Bernoulli distribution with $\alpha = P(Y = 1)$. We consider $\alpha = 0.1, 0.25, 0.5$. Observed binary variable S is generated in such a way that $P(S = 1|Y = 1) = c$, where c is treated as a parameter ranging from 0.1 to 0.9 and $P(S = 1|Y = 0) = 0$. Then X_1 is generated using conditional distributions described below. We consider 2 scenarios:

- *Simulation model 1* X_1 is generated using conditional distributions $X_1|Y = 0 \sim N(0, 1)$ and $X_1|Y = 1 \sim N(b, 1)$, where b is a parameter.
- *Simulation model 2* X_1 is generated using conditional distributions $X_1|Y = 0 \sim N(0, 1)$ and $X_1|Y = 1 \sim 0.25N(0, 1) + 0.75N(b, 1)$.

Figure 2 shows density functions corresponding to conditional distributions of $X_1|Y = 0$ and $X_1|Y = 1$ for the simulation models. Parameter b controls the dependence strength between X_1 and Y. We consider two values $b = 1$ and $b = 2$. Larger value of b corresponds to stronger dependence between X_1 and Y. In order to make a task more challenging, we also add spurious noise variables $X_2, \ldots, X_{10}$, generated from $N(0, 1)$, independently from Y and let $X = (X_1, \ldots, X_{10})$. Note that model 1 corresponds to logistic regression model, whereas in model 2 the dependence between Y and X_1 cannot be described by logistic model. Thus the first model favours the proposed method which is based on logistic model, whereas model 2 allows to analyse the robustness of the proposed method.

Figures 3 and 4 show estimation errors $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c for simulation models (for better presentation we do not show the curves for KM1

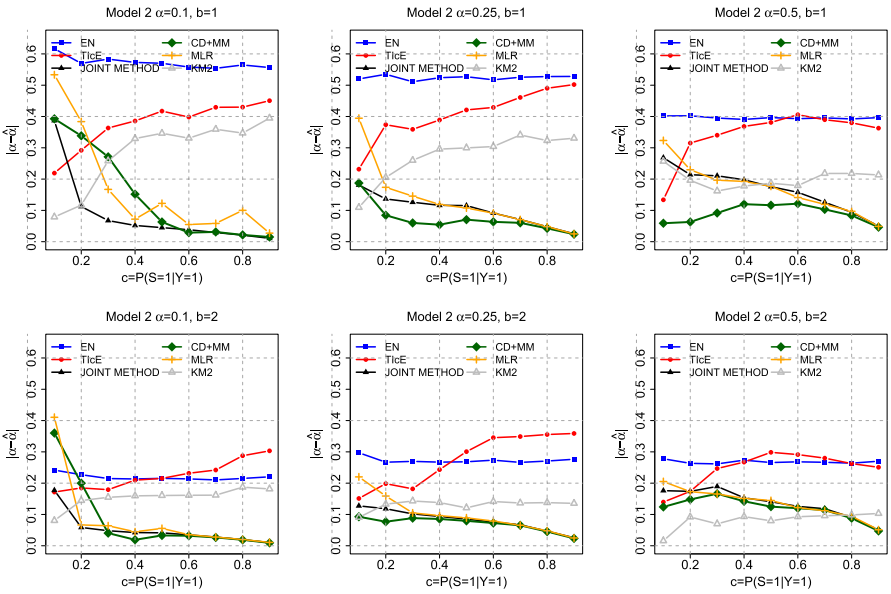

Fig. 4 Estimation error $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c for simulation model 2. Parameters: $b = 1, 2$, $\alpha = 0.1, 0.25, 0.5$ and $n = 5000$

as it worked systematically much worse than KM2). Observe that CD+MM achieves small averaged errors for both models and almost all parameter settings. The JOINT method and MLR usually work worse than CD+MM, whereas TIcE, EN and KM2 work poorly in most cases. The averaged estimation error for CD+MM decreases with c. We observed the largest estimation errors for small α and small c, which is due to the fact that for this setting we observe very few labelled observations. Indeed, for sample size $n = 5000$, $\alpha = 0.1$ and $c = 0.1$ we have on average only 50 labelled examples. For simulation model 1, the estimation error of CD+MM increases when b decreases. This suggests that, when the dependence between Y and X_1 is weak,s then estimation of α is more challenging. In extreme case $b = 0$, Y and X_1 are independent and in such case α is not identifiable, see the discussion in Sect. 3. Advantage of CD+MM and JOINT methods over competitors is larger for Model 1 than for Model 2. This is understandable since, in the first case the logistic model for which these methods are designed, is well specified. Good performance of CD+MM and JOINT in the case of Model 2 indicates that the methods are robust against departures from the logistic model.

Figures 5 and 6 show empirical distributions of $\hat{\alpha}$ in form of boxplots against the true parameter c for simulation models 1 and 2. First, it is clearly seen that EN, TIcE, KM1 and KM2 overestimate α, which is a result of their underestimation of c and agrees with theoretical analysis of Sect. 4. Secondly, we observe large variance for JOINT method, especially for small α and small b. This suggests that simple gradient optimization used in JOINT method may be insufficient; the algorithm probably is getting stuck in local minima. The variance of CD+MM is much lower, when compared to JOINT method, which indicates that the proposed optimization procedure based

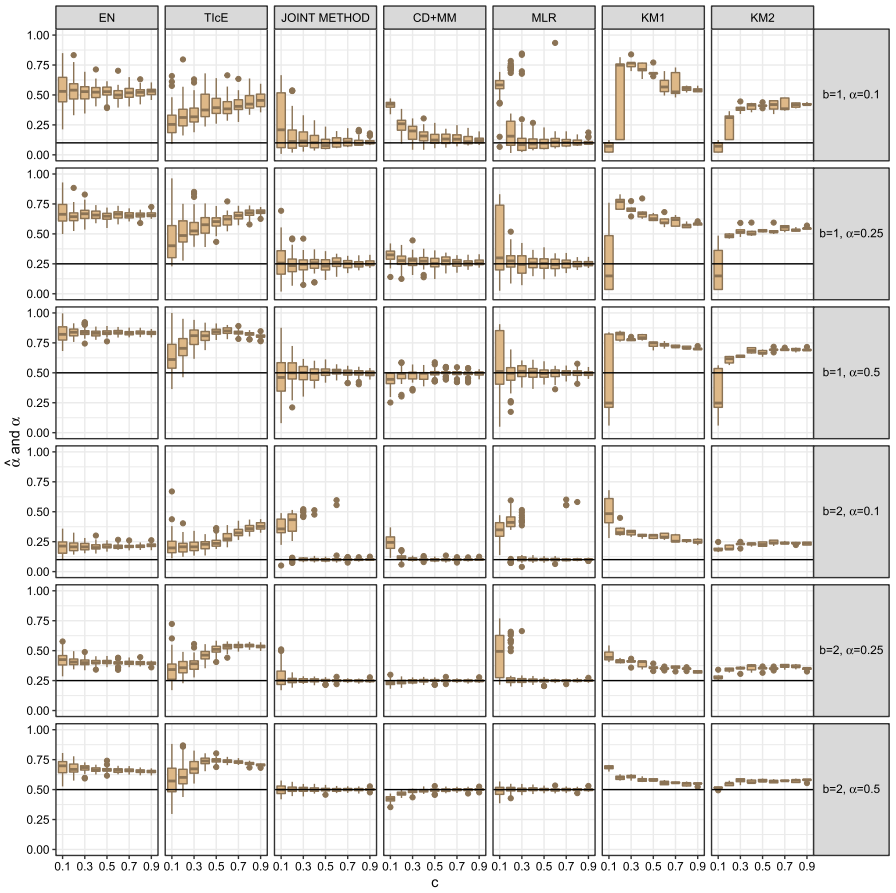

Fig. 5 Distribution of $\hat{\alpha}$ wrt the true parameter c, for simulation model 1. The horizontal line corresponds to the true α

on MM-algorithm allows to find optimal solutions more frequently which results in more stable estimation. We also stress that KM methods are the most computationally expensive among studied methods, especially for larger samples sizes. For example, when dimension of feature vector X is 10 and sample size is $n = 2000$, KM works about 2 times slower than CD+MM, whereas for sample size $n = 5000$, KM works about 30 times slower than CD+MM (assuming that for CD+MM, the maximal numbers of iterations are 300 and 50, for CD and MM algorithms, respectively; for KM we used default settings).

Finally, we performed convergence analysis of the three methods based on parametric modelling: JOINT method, MLR and the proposed CD+MM. Recall, that in CD+MM, in each step of cyclic coordinate descent algorithm 2 we perform iterative MM algorithm 1. To make a comparison between CD+MM and two remaining methods fair, for CD+MM algorithm, we analyze the total number of iterations, i.e. the number of iterations in cyclic coordinate descent Algorithm 2 multiplied by the num-

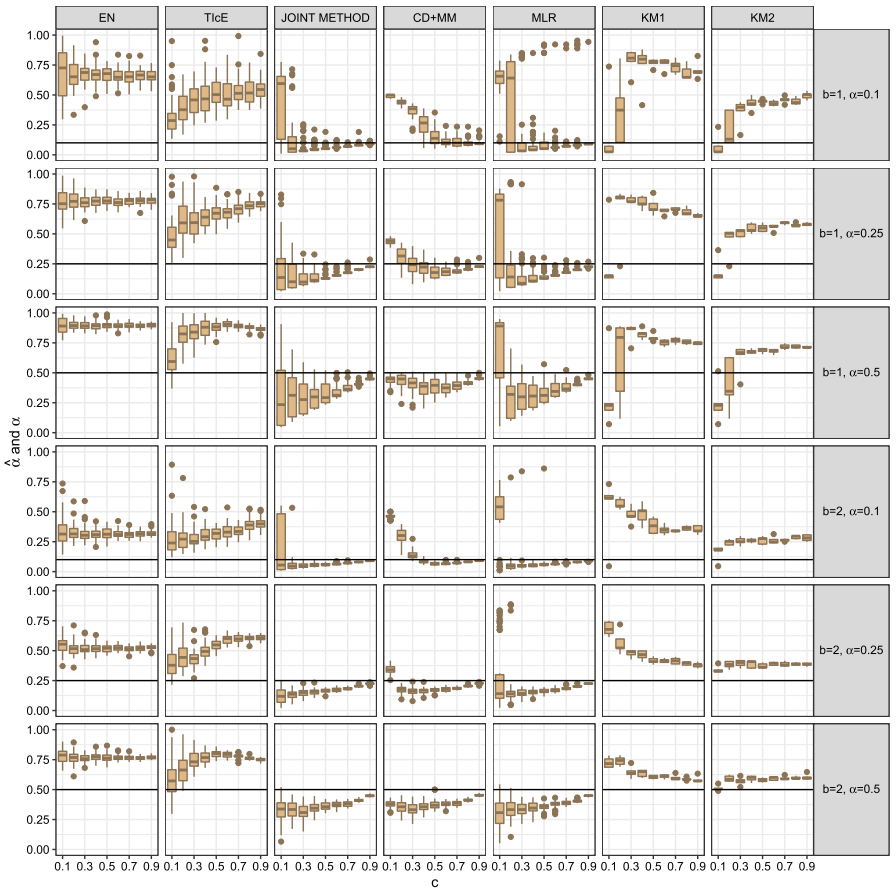

Fig. 6 Distribution of $\hat{\alpha}$ wrt the true parameter c, for simulation model 2. The horizontal line corresponds to the true α

ber of iterations of the MM algorithm. In this experiment, we take only 10 iterations of MM algorithm for each step. Figure 7 shows how the value of log-likelihood changes with the number of iterations for simulation models 1 and 2 for $c = 0.3, 0.5, 0.7$. In all considered cases, CD+MM achieves larger value of the loglikelihood than two remaining methods, within 100 first iterations. Interestingly, the curves for JOINT method and MLR are similar. The plateau of the curves corresponding to JOINT method and MLR may indicate the problem of non-convergence to the global optimum discussed above.

6.2 Benchmark datasets

We use 8 popular benchmark datasets from UCI Machine Learning Repository and one that was used for the IJCNN 2001 neural network competition (Prokhorov 2001). A short summary of each dataset can be found in Table 1. They were chosen to represent

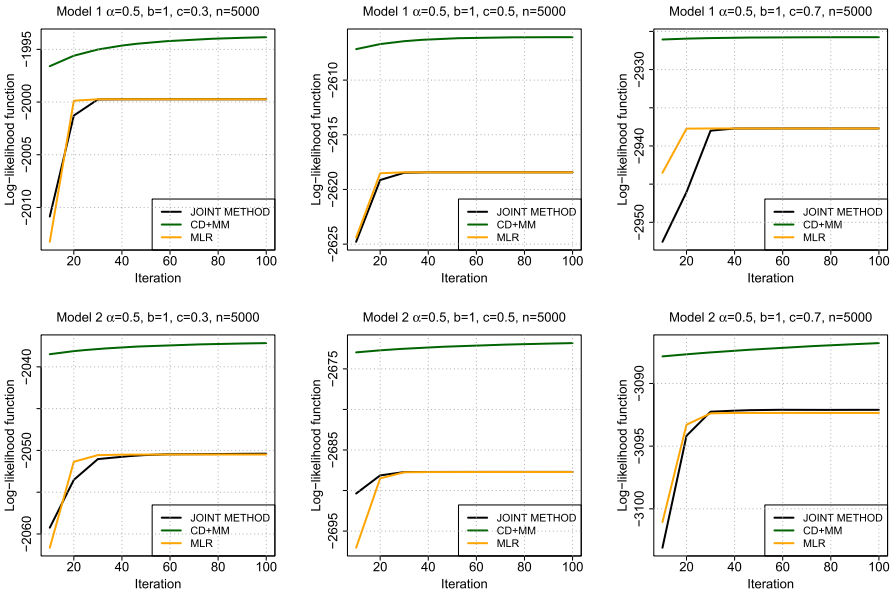

Fig. 7 Convergence analysis. Log-likelihood function with respect to the number of iterations for logistic regression-based methods: JOINT METHOD, CD+MM and MLR

various characteristics of data (number of observations, number of features and the fraction of positive examples). To adjust these data to our problem, we created PU datasets from the labelled datasets, the positive examples were selected to be labelled with label frequencies $c = 0.1, 0.2, \ldots, 0.9$. For each label frequency c, we generated 100 PU datasets labelling randomly elements having $Y = 1$ with probability c and then averaged the results over 100 repetitions. The true class prior for each dataset was estimated as the number of positive examples divided by the number of examples. All numerical features were scaled between 0 and 1 with the standard transformation $(x - \min(x))/(\max(x) - \min(x))$. Such transformation was recommended for TIcE algorithm (Bekker and Davis 2018). Due to computational cost of KM1 and KM2, for these methods, as in Ramaswamy et al. (2016) and Bekker and Davis (2018), we subsampled two largest datasets (mushroom and ijcnn2001) choosing $n = 2000$ observations and averaged the results of 5 such trials for each experiment. Figure 8 shows averaged values of $|\hat{\alpha} - \alpha|$ and Fig. 9 empirical distributions of $\hat{\alpha}$ against c. In terms of an error $|\alpha - \hat{\alpha}|$ the method CD+MM achieves the most accurate results for five datasets (credit-g, diabetes, heart-c, mushroom, spambase) for all or almost all c values with the JOINT or MLR method being the second best (see Fig. 8). In two cases the KM2 works best (vote and wdbc). The MLR method works well on average except for small values of α and two data sets: credit-g and ijcnn2001 (in the latter case it behaved very erratically and due to this it has been removed from the respective plot). It can be seen from Fig. 9 that for both CD+MM and JOINT method $\hat{\alpha}$ is usually less variable than EN and TIcE and the estimation error decreases with c. For EN and TIcE underestimation of c results in overestimated α for BreastCancer,

Table 1 Summary statistics of benchmark datasets

Dataset	n	p	α
Mushroom	8124	21	0.48
ijcnn2001	35000	22	0.10
wdbc	569	31	0.37
Vote	435	32	0.39
Spambase	4601	57	0.39
Heart-c	303	19	0.46
Diabetes	768	8	0.35
Credit-g	1000	48	0.30
BreastCancer	683	9	0.35

credit-g, diabetes, heart-c, spambase and mushroom datasets. On the other hand, the CD+MM and MLR methods tend to underestimate α for small values of c.

6.3 Experiment on clinical dataset MIMIC

We performed an experiment on large clinical database MIMIC III (Johnson et al. 2016). The database contains information on 33166 patients treated in intensive care units (ICU) who are diagnosed according to the coding scheme ICD-9. Patients are diagnosed with various diseases, among which we consider 5 diseases: hypertension, kidney and liver disease, diabetes and chronic pulmonary obstructive disease (copd). The above families of diseases were already investigated in previous studies (Zufferey et al. 2015; Teisseyre et al. 2019; Teisseyre 2020). Similarly as for benchmark datasets, the true prevalence α is computed as a fraction of patients in the database with the given disease. Table 2 shows the values of α for the considered diseases. Note that these values do not match the prevalences of the diseases in the population, which is due to the fact that the considered database is related to the ICU patients and thus it cannot be treated as a representative sample from the population. The original dataset consists of 308 features which correspond to certain blood and diagnostic tests (e.g. Glucose, Sodium, etc.), administrative information (e.g. sex, age, marital status) and medical scores used to track a person's status during the stay in an intensive care unit (e.g. Braden score used to assess a risk of developing a pressure ulcer). The list of all features can be found at https://home.ipipan.waw.pl/p.teisseyre/PUBLICATIONS/parcc/parcc_supplement.pdf. For each disease, we select 30 features using a simple filter based on mutual information, i.e. we first calculate the mutual information between the given disease and the features and then select 30 features corresponding to largest values of mutual information.

To create PU datasets from the completely labelled datasets, the positive examples are selected to be labelled with label frequencies $c = 0.1, 0.2, \ldots, 0.9$. For each label frequency c we generated 100 PU datasets labelling randomly elements having $Y = 1$ with probability c and then averaged the results over 100 repetitions. The above scheme corresponds to the situation when actually occurring disease is diagnosed with probability c.

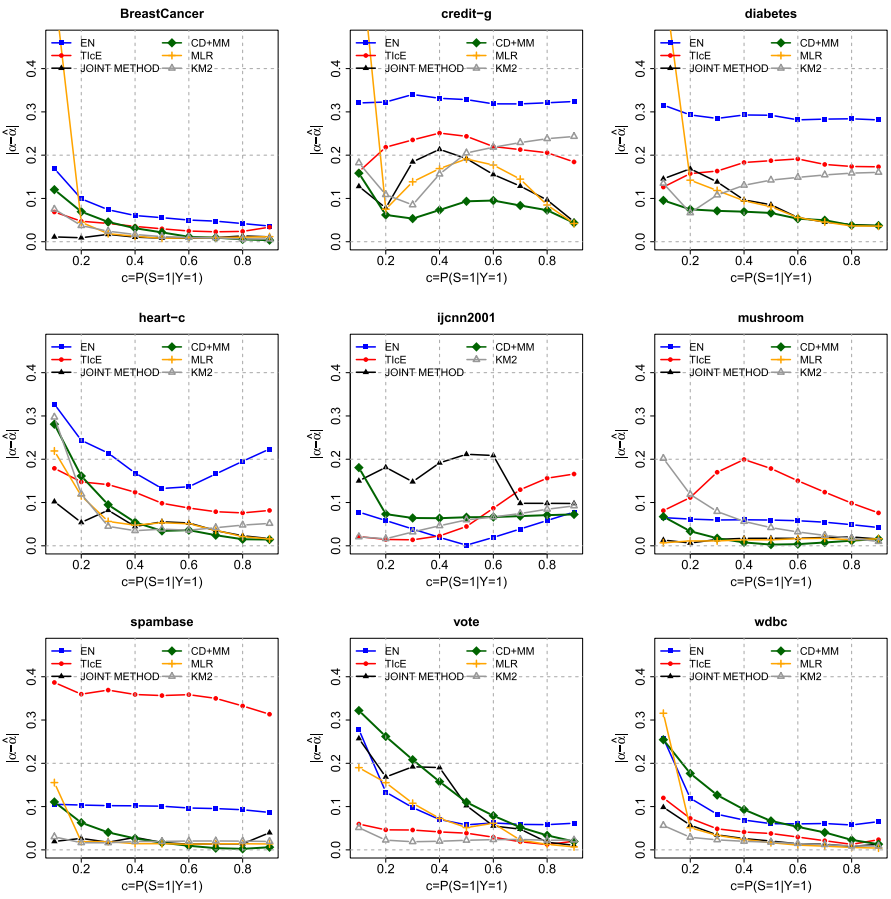

Fig. 8 Estimation error $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c for benchmark datasets

Figure 10 shows distributions of $\hat{c}$ wrt the true parameter c. Observe that EN and TIcE underestimate label frequency c, for all 5 considered datasets. The same is true for KM1 and KM2, though they work better than EN and TIcE. The proposed methods work much better for all datasets except liver, for which they show the same tendency to underestimate c as EN and TIcE but to lesser degree. Figure 11 shows distributions of $\hat{\alpha}$ wrt the true parameter c. Underestimation of c results in considerable overestimation of α in the case of both EN and TIcE. The proposed method CD+MM gives fully satisfactory results for diabetes and kidney. Although, CD+MM slightly overestimates α for copd and liver, its error is still much lower than for EN and TIcE. For hypertension, CD+MM underestimates α for small c and overestimates α for large c. Finally, for CD+MM we observe smaller errors than for JOINT method, which indicates that the proposed optimization procedure based on MM-algorithm allows to improve the estimation accuracy. Figure 12 shows estimation error $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c. We observe the smallest averaged estimation errors for CD+MM for all cases but two with $c = 0.1$. CD+MM outperforms MLR,

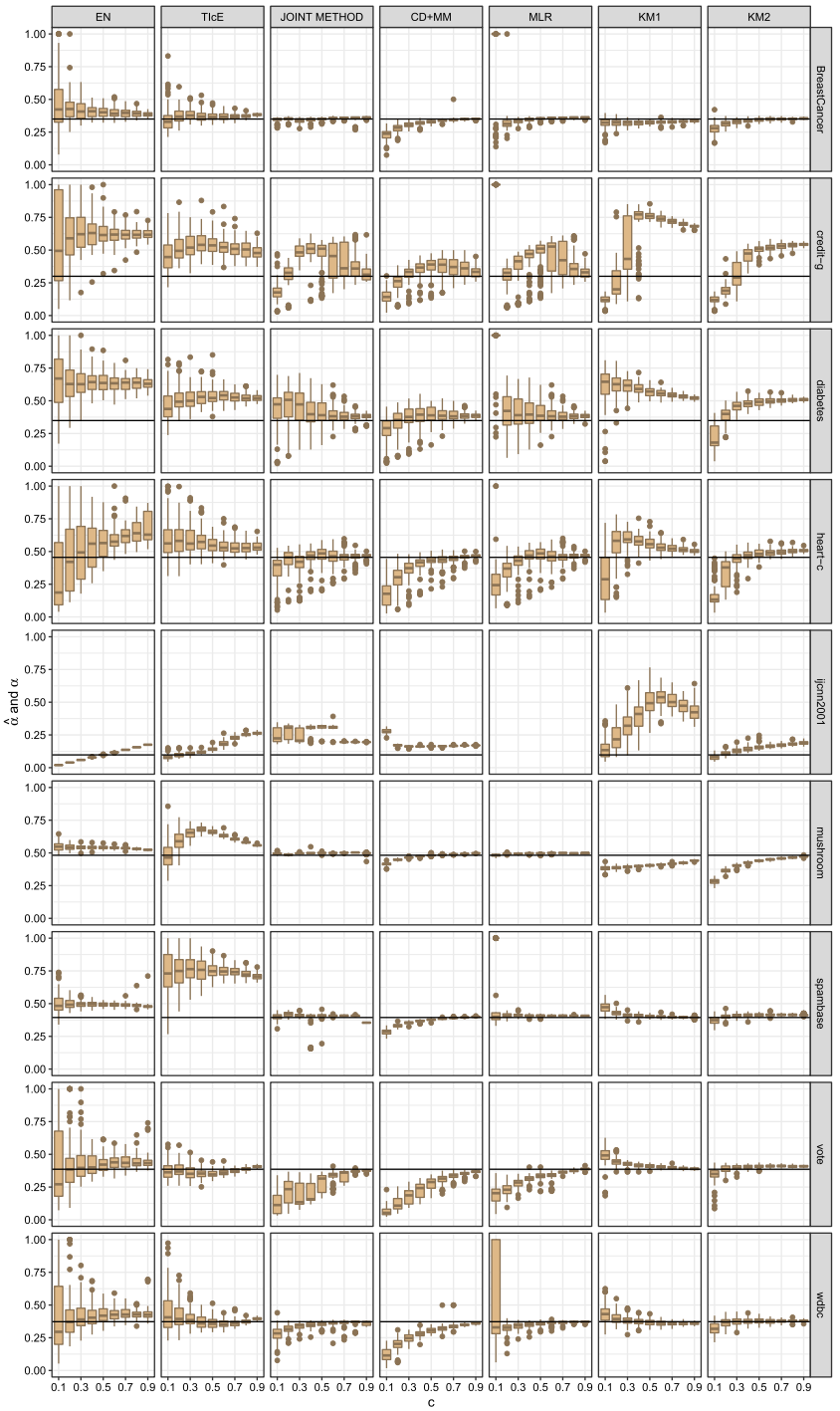

Fig. 9 Distribution of $\hat{\alpha}$ wrt the true parameter c, for benchmark datasets. The horizontal line corresponds to the true α. The plot for MLR on ijcnn2001 data is omitted to its erratic behaviour

Table 2 Summary statistics of MIMIC III database

Number of observations (patients)	33166
Number of features	308
% of patients with hypertension disease	$\alpha = 66.7\%$
% of patients with kidney disease	$\alpha = 34.98\%$
% of patients with liver disease	$\alpha = 6.71\%$
% of patients with diabetes disease	$\alpha = 31.82\%$
% of patients with copd disease	$\alpha = 23.24\%$

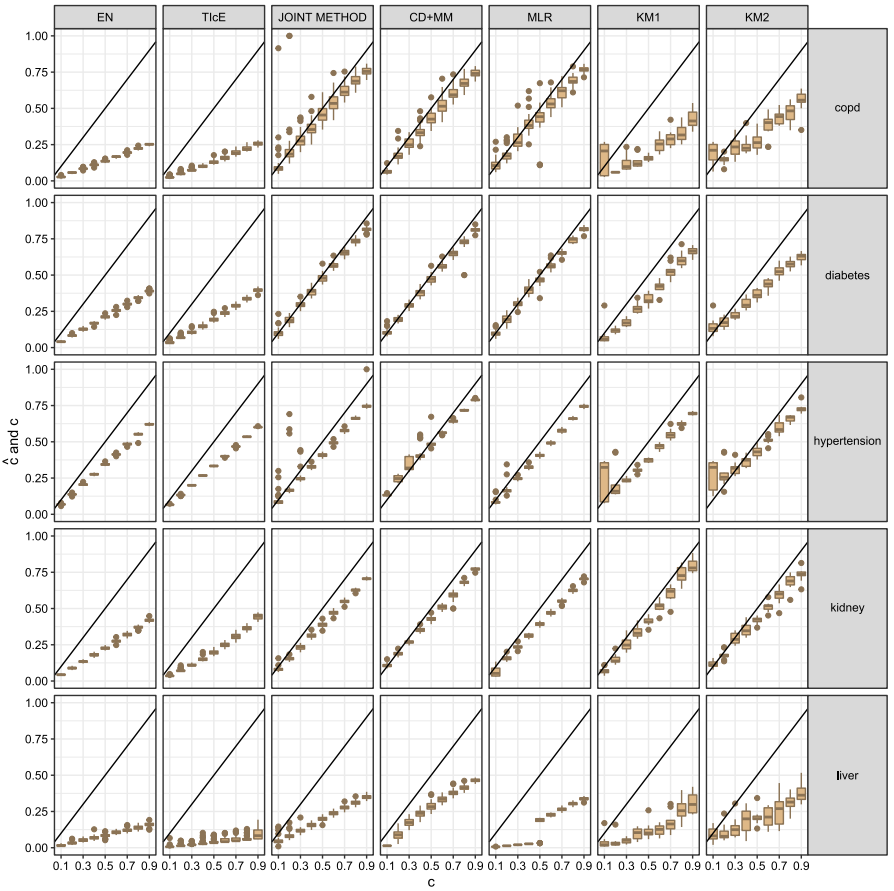

Fig. 10 Distribution of $\hat{c}$ wrt the true parameter c, for MIMIC-III datasets

which usually works on par with JOINT method, but in some situations is unstable (for example liver disease and small c).

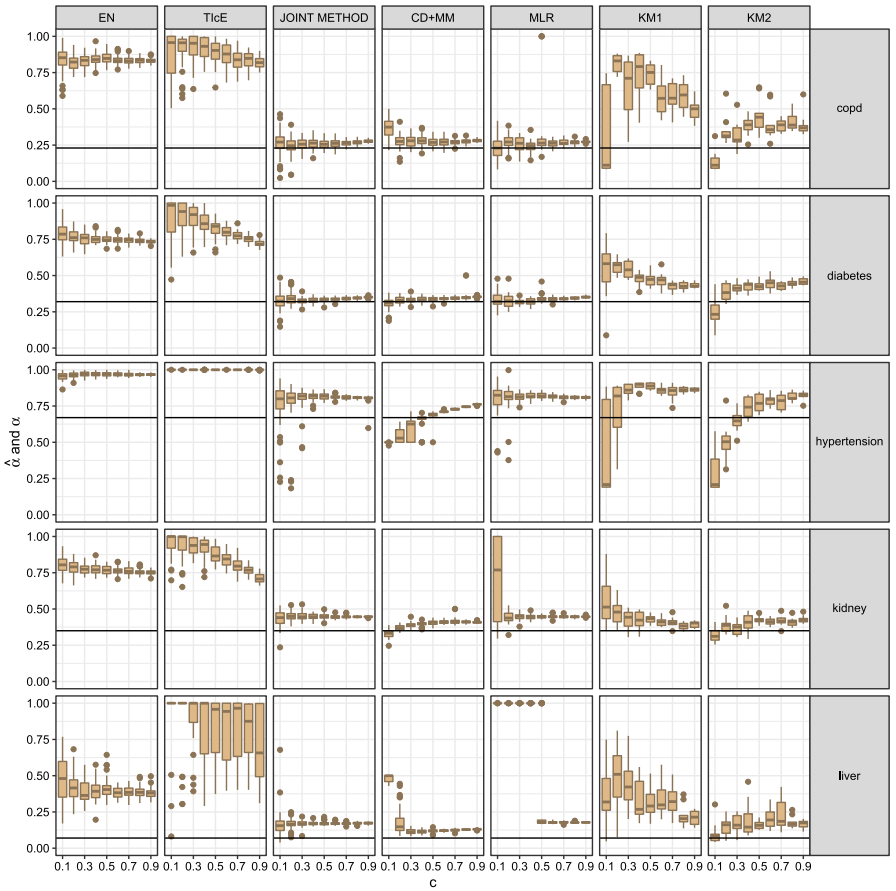

Fig. 11 Distribution of $\hat{\alpha}$ wrt the true parameter c, for MIMIC-III datasets. The horizontal line corresponds to the true α

7 Conclusions and future work

In this paper we analysed different methods of class prior estimation in positive unlabelled learning. We showed that class prior probability is not identifiable for a PU single sample scenario given a full knowledge of distribution of (X, S) if no assumptions on distribution of (X, Y) are imposed. The class prior becomes identifiable when we impose mild semi-parametric model assumptions on conditional distribution of Y given X. We formally show that in some situations, some of the existing algorithms tend to underestimate label frequency c and overestimate class prior probability. This property is confirmed by the numerical experiments. The proposed approach, based on logistic regression, involves simultaneous estimation of label frequency c and model parameters. In order to account for the non-concavity of the likelihood function, a novel optimization procedure, called CD+MM, is proposed in this paper, which is a

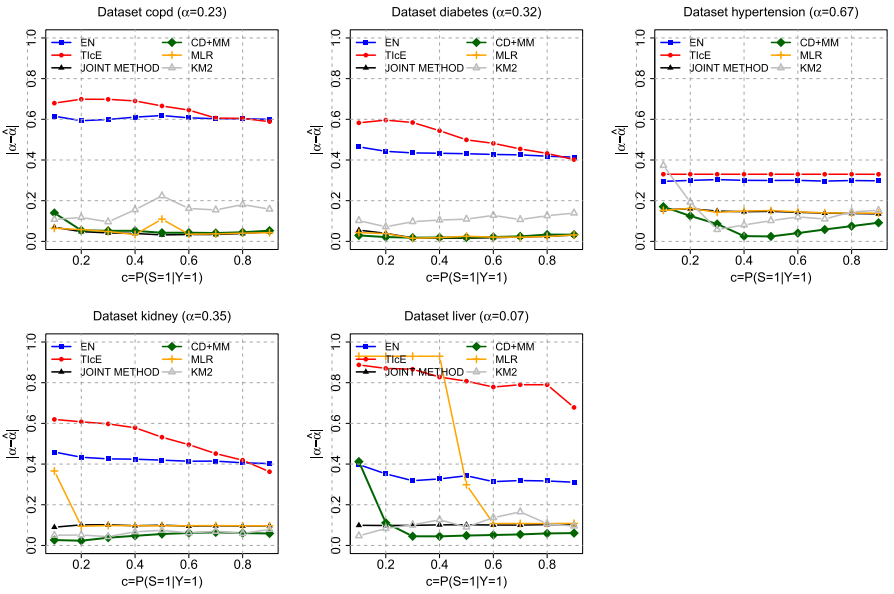

Fig. 12 Estimation error $|\alpha - \hat{\alpha}|$ (averaged over 100 simulations) wrt c for MIMIC III datasets

combination of cyclic coordinate descent and Minorization-Maximization algorithms. In the method, we iteratively optimize profile log-likelihood functions. The experiments, performed on artificial and benchmark datasets as well as on large clinical database MIMIC, indicate that the proposed method CD+MM achieves lower estimation errors of the class prior than other considered methods, for most of the datasets and parameter settings. Indirectly, this indicates that CD+MM is robust to departures from parametric setting of a logistic model from which it has been derived. For the benchmark datasets, KM2 was very competitive, however it was computationally very costly. Moreover, CD+MM is significantly less variable than related JOINT method which uses simple gradient optimization of the likelihood function. It follows from experiments, that class prior estimation becomes more challenging when label frequency is small and the dependence between class variable Y and feature vector X is weak.

Since the performance of the methods based on logistic regression (JOINT method and CD+MM) seems promising, we believe that this approach is worth pursuing. It would be of interest to develop modifications of the method for number of predictors p larger than sample size n using e.g. regularised version of logistic regression. Such version would be particularly useful to deal with high-dimensional data. Moreover, finding concave lower bound of $L(c, b)$ both in c and b would lead to another potentially interesting modification of the proposed method.

Acknowledgements Insightful remarks of two reviewers which substantially influenced the final form of the manuscript are gratefully acknowledged.

References

Bahorik AL, Newhill CE, Queen CC, Eack SM (2014) Under-reporting of drug use among individuals with schizophrenia: prevalence and predictors. Psychol Med 44(1):61–69

Bekker J, Davis J (2018) Estimating the class prior in positive and unlabeled data through decision tree induction. In: Proceedings of the 32th AAAI conference on artificial intelligence

Bekker J, Davis J (2020) Learning from positive and unlabeled data: a survey. Mach Learn 109:719–760. https://doi.org/10.1007/s10994-020-05877-5

Bierens H (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. J Am Stat Assoc 78:699–707

Chapelle O, Schölkopf B, Zien A (2010) Semi-supervised learning. The MIT Press, Cambridge

Chen WJ, Fang CC, Shyu RS, Lin KC (2006) Underreporting of illicit drug use by patients at emergency departments as revealed by two-tiered urinalysis. Addict Behav 31(12):2304–2308

Couso Inés, DD, Hüllermeier E (2017) Maximum likelihood estimation and coarse data. In: Proceedings of the international conference on scalable uncertainty management, volume 10564 of SUM 2017, pp 3–16. Springer

Cover TM, Thomas JA (2006) Elements of information theory (Wiley Series in Telecommunications and Signal Processing). Wiley, New York

Steinberg D, Cardell NS (1992) Estimating logistic regression models when the dependent variable has no variance. Commun Stat Theory Methods 21(2):423–450

Denis F, Gilleron R, Letouzey F (2005) Learning from positive and unlabeled examples. Theor Comput Sci 348(1):70–83

du Plessis M, Sugiyama M (2014) Class prior estimation for positive and unlabeled data. IEICE Trans Inf Syst E–97D(5):1358–1372

Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08, pp 213–220

Frenay B, Verleysen M (2014) Classification in the presence of label noise: A survey. IEEE Trans Neural Netw Learn Syst 25(5):845–869

Fung GPC, Yu JX, Lu H, Yu PS (2006) Text classification without negative examples revisit. IEEE Trans Knowl Data Eng 18(1):6–20

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton

Heitjan DF, Rubin DB (1991) Ignorability and coarse data. Ann Stat 19:2244–2253

Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single index models. J Econom 58(1):71–120

Jain S, White M, Radivojac P (2016) Estimating the class prior and posterior from noisy positives and unlabeled data. In: Proceedings of the 30th international conference on neural information processing systems, pp 2693–2701

Jaskie K, Elkan C, Spanias A (2020) A modified logistc regression for positive and unlabeld learning. In: 53rd Asilomar conference on signals, systems, and computers, pp 2007–2011

Jaskie K, Spanias A (2019) Learning algorithms and applications : a survey. In: IEEE IISA, Patras, Greece,Jul. 2019., pp 1–8

Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony CL, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:1–9

Kiryo R, Niu G, du Plessis MC, Sugiyama M (2017) Positive-unlabeled learning with non-negative risk estimator. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17, pp 1674–1684

Lan W, Wang J, Li M, Liu J, Li Y, Wu F, Pan Y (2016) Predicting drug-target interaction using positive-unlabeled learning. Neurocomputing 206:50–57

Lancaster T, Imbens G (1996) Case-control studies with contaminated controls. J Econom 71(1):145–160

Lange K (2010) Numerical analysis for statisticians. Springer Verlag, New-York

Li K, Duan N (1989) Regression analysis under link violation. Ann Stat 17(3):1009–1052

Li X, Liu B (2003) Learning to classify texts using positive and unlabeled data. In: Proceedings of the 18th international joint conference on artificial intelligence, pp 587–592

Liu B, Dai Y, Li X, Lee WS, Yu PS (2003) Building text classifiers using positive and unlabeled examples. In: Proceedings of the third IEEE international conference on data mining, ICDM '03, pp 179–

Menon A, Rooyen B, Ong C, Williamson R (2015) Learning from corrupted binary labels via class-probability estimation. In: Proceedings of the 32nd international conference on machine learning, pp 1–10

Mielniczuk J, Teisseyre P (2016) What do we choose when we err? Model selection and testing for misspecified logistic regression revisited. Studies in Computational Intelligence, vol 605. Springer, Berlin, pp 271–296

Natarajan N, Dhillon IS, Ravikumar P, Tewari A (2013) Learning with noisy labels. In: Proceedings of the 26th international conference on neural information processing systems, NIPS'13, pp. 1196–1204, Red Hook, NY, USA. Curran Associates Inc

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl Ecol 43(3):405–412

Plessis MC, Niu G, Sugiyama M (2017) Class-prior estimation for learning from positive and unlabeled data. Mach Learn 106(4):463–492

Prokhorov D (2001) IJCNN 2001 neural network competition. Slide presentation in ijcnn'01, Ford Research Laboratory

Ramaswamy H, Scott C, Tewari A (2016) Mixture proportion estimation via kernel embeddings of distributions. In: Proceedings of The 33rd international conference on machine learning, vol 48, pp 2052–2060

Scott C (2015) A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: Proceedings of the eighteenth international conference on artificial intelligence and statistics. PMLR, vol 38, pp 838–846

Scott C, Blanchard G, Handy G (2013) Classification with asymmetric label noise: Consistency and maximal denoising. In: Conference on learning theory (COLT), volume 30 of JMLR proceedings, pp 489–511

Sechidis K, Sperrin M, Petherick ES, Luján M, Brown G (2017) Dealing with under-reported variables: An information theoretic solution. Int J Approx Reason 85:159–177

Song H, Raskutti G (2020) PUlasso: High-dimensional variable selection with presence-only data. J Am Stat Assoc 115(529):334–347

Teisseyre P (2020) Learning classifier chains using matrix regularization: application to multimorbidity prediction. In: Proceedings of the european conference on artificial intelligence, ECAI'20

Teisseyre P, Mielniczuk J, Łazecka M (2020) Different strategies of fitting logistic regression for positive and unlabelled data. In: Proceedings of the international conference on computational science, ICCS'20

Teisseyre P, Zufferey D, Słomka M (2019) Cost-sensitive classifier chains: Selecting low-cost features in multi-label classification. Pattern Recogn 86:290–319

Walley NM et al (2018) Characteristics of undiagnosed diseases network applicants: implications for referring providers. BMC Health Serv Res 18(1):1–8

Ward G, Hastie T, Barry S, Elith J, Leathwick J (2009) Presence-only data and the EM algorithm. Biometrics 65:554–563

Yang P, Li X, Chua HN, Kwoh CK, Ng SK (2014) Ensemble positive unlabeled learning for disease gene identification. PLoS ONE 9(5):1–11

Zufferey D, Hofer T, Hennebert J, Schumacher M, Ingold R, Bromuri S (2015) Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. Comput Biol Med 65:34–43