ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230342

Double Logistic Regression Approach to Biased Positive-Unlabeled Data

Konrad Furmańczyk[a], Jan Mielniczuk[b,d], Wojciech Rejchel[c] and Paweł Teisseyre[b,d;*]

[a]University of Life Sciences, Warsaw, Poland
[b]Polish Academy of Sciences, Warsaw, Poland
[c]Nicolaus Copernicus University, Toruń, Poland.
[d]Warsaw University of Technology, Warsaw, Poland
ORCiD ID: Konrad Furmańczyk https://orcid.org/0000-0002-7683-4787

Abstract. Positive and unlabelled learning is an important non-standard inference problem which arises naturally in many applications. The significant limitation of almost all existing methods addressing it lies in assuming that the propensity score function is constant and does not depend on features (Selected Completely at Random assumption), which is unrealistic in many practical situations. Avoiding this assumption, we consider parametric approach to the problem of joint estimation of posterior probability and propensity score functions. We show that if both these functions are logistic with different parameters (double logistic model) then the corresponding parameters are identifiable. Motivated by this, we propose two approaches to their estimation: a joint maximum likelihood method and the second approach based on an alternating maximization of two Fisher consistent approximations. Our experimental results show that the proposed methods perform on par or better than the existing methods based on Expectation-Maximisation scheme.

1 Introduction

Positive unlabelled (PU) inference is based on data sets containing labelled observations ($S = 1$) which are all positive ($Y = 1$), and unlabelled ones ($S = 0$) which may either belong to a positive or a negative class (Y is either 1 or 0). Examples of such experimental setup abound in medicine [36, 22, 6, 38], text and image analysis [9, 27, 26, 15], ecology [37, 29] and survey data [33]. For example, medical databases may contain only information about diagnosed patients who have a certain disease ($S = 1$) whereas un-diagnosed patients ($S = 0$) may have it or not. In survey sampling, asking a sensitive question (e.g. on use of illicit drugs) may lead to under-reporting, as beside the positive respondents ($Y = 1$) who answer the question truthfully, there are respondents who engage in this activity and do not admit it ($Y = 1, S = 0$). Their answers are merged together with those of people who abstain from such behaviour and answer the question negatively ($Y = 0, S = 0$) [1]. PU data occur frequently in text classification problems. 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. The other important example is associated with detecting unlawful content in social networks. In this case, certain content has been marked as unlawful (e.g. some images or posts), however unlawful content may still exist among the unmarked profiles.

In the seminal paper [8] an influential approach to this problem which is proposed based on assumption that probability of labelling of positive elements is not instance dependent, i.e. $P(S = 1|Y = 1, x) = P(S = 1|Y = 1)$ (Selected Completely at Random, or SCAR, assumption), where x is a feature vector and constant $c = P(S = 1|Y = 1)$ is called label frequency. For a review of the developments, almost exclusively based on SCAR, we refer to [4]. The SCAR assumption facilitates inference significantly, as in this case, aposteriori probability $P(Y = 1|x)$, which is often of the main interest, can be written as $P(Y = 1|x) = c^{-1}P(S = 1|x)$, where $P(S = 1|x)$ can be estimated using the observed PU data. In view of this, estimation of c becomes a crucial problem [8, 31, 16, 30, 3, 17, 23]. The common approach here is to treat first the unlabelled observations as coming from negative ($Y = 0$) class and then detect among them those observations, which, due to their covariates' closeness to the labelled data, should be assigned to the positive class (see e.g. [32], [39]). Other important approaches are based on suitable modification of the risk function using weighting to account for unobservability of negative examples (see [7] and [21]).

However, the SCAR assumption fails in many practical situations. For example, an age is an important factor in screening for many diseases (such as a prostate cancer) [24] which may lead to a positive diagnosis. Moreover, the occurrence of other diseases (e.g. obesity), may play a role in undertaking an in-depth scrutiny for other potential illnessess (e.g. diabetes) [35]. In surveys, the criminal background of the interviewee is a strong indication that obtaining an untrustworthy answer is likely. In a general case, the situation is much more complicated than under SCAR, as $P(S = 1|x)$ may be small even if $P(Y = 1|x)$ is large. Importantly, ignoring the fact that the probability of labelling depends on features will lead to biased estimation of the posterior probability. An accurate estimation of this probability enables a precise estimation of the posterior probability which, in its turn, leads to an accurate prediction. The problem draws more attention recently [5, 11, 10]. Besides approaches based on EM algorithms (see below), other methods to tackle this problem are based on the concept of probabilistic gap ([13]), assumptions that ordering of posterior and propensity score with respect to x coincide ([18])

* Corresponding Author. Email: teisseyrep@ipipan.waw.pl

or on application of deep learning techniques ([28]). In a broader context, the situation when elements of underlying sample were chosen taking the values of their covariates into account is frequently termed labelling (selection) bias or covariate shift, and its importance is recognised by many authors ([14], [19]).

The restrictiveness of SCAR assumption calls for functional modelling of the probability of being labelled which corresponds to propensity score in causal inference. The important steps in this direction has been taken recently in [5] and [11], where variants of Expectation-Maximisation (EM) algorithm have been considered, see Section 5 for the detailed discussion of the methods. The present paper also addresses this issue and contains the following new developments: firstly, we consider parametric models for propensity score $e(x) = P(S = 1|Y = 1, x)$ and $P(Y = 1|x)$ and show that their parameters are identifiable given values of $P(S = 1|x)$ only. In particular, the both functions can be modeled as logistic functions with different parameters; this setup will be called double logistic model in the following. This naturally leads to an introduction of a joint maximum likelihood (ML) estimators of these parameters and establishing their consistency when the model is well specified. We note that the analysis of this parametric approach is hindered by the fact that even in the SCAR case log-likelihood *is not necessarily a concave function* of the underlying parameters. Secondly, we introduce a method (called 'Two MODELS' method, 'TM' in brief) consisting in alternate maximising *concave* empirical surrogates for expected likelihoods of posterior probability of $Y = 1$ and the propensity score. We prove that the method is consistent under some simplifying assumptions. Moreover, we numerically investigate the behaviour of our proposals and show that the TM method consistently exhibits superior or comparable behaviour to the best of existing methods [5, 11].

The rest of the paper is structured as follows. In Section 2 we formally describe the PU learning problem and define basic quantities. In Section 3 we discuss the problem of joint estimation of posterior probability and propensity score functions and state main theoretical results. The algorithms (including the proposed ones) are described in Section 4, whereas the two existing most related methods (EM and LBE) are discussed in Section 5. In Section 6 we describe the results of experiments and in Section 7 we conclude our work. The Appendix available at https://github.com/teisseyrep/putm contains the proofs and some additional numerical results.

2 Background

We first introduce basic notations. Let X be a random variable corresponding to feature vector, $Y \in \{0, 1\}$ be a true class label and $S \in \{0, 1\}$ an indicator of an example being 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 iid sample drawn from it. Observed data consists of $(X_i, S_i), i = 1, \ldots, n$ (so called, the single sample scenario). 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 $S = 0$, label Y can be either 1 or 0. Our aim is to learn binary posterior distribution of Y given $X = x$ i.e. $y(x) = P(Y = 1|X = x)$ and we only observe samples from distribution of (X, S), where $S = Y$ with a certain probability. To this end we define a binary *posterior distribution function of* S *given* x as $s(x) = P(S = 1|x)$ and *a propensity score function* $e(x) = P(S = 1|Y = 1, x)$. We note that

$$s(x) = e(x)y(x) \tag{1}$$

as $P(S = 1|Y = 0, x) = 0$. In the following we assume that $e(x)$ may depend on x that is we do not impose restrictive and hard to verify Selected Completely at Random (SCAR) assumption. We note that SCAR assumption implies that the distribution of X for labelled data coincides with its distribution in the positive class, but this is not true in general. This makes inference much harder task in the general setting as the distribution of the labelled data is biased. We note that $e(x)$ plays a role of a nuisance functional parameter and our primary objective is to estimate $y(x)$.

We stress that in parallel to the single sample scenario, the case-control (c-c) scenario is frequently considered for PU data. In this scenario in addition to the labelled data from the positive class we have at our disposal unlabelled sample drawn from *the marginal distribution* of X. The form of the second sample makes the inference problem different (and in general easier) than that studied here. Moreover, solutions obtained for c-c case are not transferable to the single sample case. For approaches developed for c-c case see e.g. [21] and [18].

3 Joint estimation of posterior probability and propensity score function

As only $s(x)$ is observable and $e(x)$ is an unknown function which is not constant, identification of posterior $y(x)$ in view of (1) is clearly impossible in general. However, we will show that if certain parametric assumptions are imposed on $y(x)$ and $e(x)$ then both functions are identifiable up to an interchange of $y(x)$ and $e(x)$. Namely, let $\sigma(s) = 1/(1 + e^{-s})$ be a logistic function and assume that both $y(x)$ and $e(x)$ are governed by the logistic model:

$$y(x) = \sigma(\beta_0^* + \beta^{*T} x) \qquad e(x) = \sigma(\gamma_0^* + \gamma^{*T} x). \tag{2}$$

We will call PU model for which (2) is satisfied a double logistic model. Note that no assumptions on the distribution of the vector of features X is imposed. As logistic model is quite versatile, it is not unrealistic to assume, that in many situations both $y(x)$ and $e(x)$ may follow it, at least approximately. We also note that Two Models method proposed below can be combined with other classifiers (e.g. neural networks), but this is left for a future research.

For any $b_0 \in R$ and $b \in R^p$ with some abuse of notation we let $\tilde{b} = (b_0, b^T)^T$. We have the following result which plays an important role in proving the consistency of joint maximum likelihood estimation.

Theorem 1 *Consider $s(x)$ defined as in (1) and assume that $y(x)$ and $e(x)$ satisfy (2). Then parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ are uniquely defined up to an interchange of $y(x)$ and $e(x)$ i.e. if for some $\tilde{\beta}$ and $\tilde{\gamma}$ we have $s(x) = \sigma(\beta_0 + \beta^T x)\sigma(\gamma_0 + \gamma^T x)$ for all $x \in R^p$, then $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\beta}^*, \tilde{\gamma}^*)$ or $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\gamma}^*, \tilde{\beta}^*)$.*

The proof of Theorem 1 is contained in the Appendix 1. Moreover, we show in Theorem 5 in the Appendix 1 that this result actually holds for a general function response p replacing σ in definitions of $y(x)$ and $e(x)$ under certain mild assumptions imposed on the logarithmic derivative $p'(s)/p(s)$.

Assume now that the logistic model is fitted both to $y(x)$ and $e(x)$ and consider the risk function corresponding to logistic loss for (X, S)

$$Q(\tilde{\beta}, \tilde{\gamma}) = E_{X,S}[S \log s_{\tilde{\beta},\tilde{\gamma}}(X) + (1 - S) \log(1 - s_{\tilde{\beta},\tilde{\gamma}}(X))], \tag{3}$$

where $s_{\tilde{\beta},\tilde{\gamma}}(x) = \sigma(\beta_0 + \beta^T x)\sigma(\gamma_0 + \gamma^T x)$. We have the following result in which $|\tilde{b}|_1 = \sum_{i=0}^{p} |b_i|$ for $\tilde{b} = (b_0, b_1, \ldots, b_p)^T$ denotes the l_1 norm.

Lemma 1 *Let assumptions of Theorem 1 hold and* $|\tilde{\beta}^*|_1 > |\tilde{\gamma}^*|_1$. *Then*

$$(\tilde{\beta}^{*T}, \tilde{\gamma}^{*T})^T = \arg \max_{(\tilde{\beta},\tilde{\gamma}):|\tilde{\beta}|_1>|\tilde{\gamma}|_1} Q(\tilde{\beta}, \tilde{\gamma})$$

and $(\tilde{\beta}^{*T}, \tilde{\gamma}^{*T})^T$ *is the unique maximiser of* $Q(\tilde{\beta}, \tilde{\gamma})$.

The assumption $|\tilde{\beta}^*|_1 > |\tilde{\gamma}^*|_1$ is imposed due to the possibility of the fact that Q is a symmetric function: $Q(\tilde{\beta}, \tilde{\gamma}) = Q(\tilde{\gamma}, \tilde{\beta})$. We note that the l_1 norm in this condition is not essential and may be replaced by any norm. The proof of the Lemma 1 is relegated to the Appendix 1.

Define an empirical counterpart of $Q(\tilde{\beta}, \tilde{\gamma})$ given in (3) as

$$Q_n(\tilde{\beta}, \tilde{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}[S_i \log s_{\tilde{\beta},\tilde{\gamma}}(X_i) + (1 - S_i)\log(1 - s_{\tilde{\beta},\tilde{\gamma}}(X_i))].$$

In the view of Lemma 1 it is intuitive to expect that maximisers of $Q_n(\tilde{\beta}, \tilde{\gamma})$ will approximate true parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ of the generating mechanism. Indeed, we have the following result.

Theorem 2 *(Strong consistency of joint ML estimation) Let assumptions of Lemma 1 hold and* $K((\tilde{\beta}^*, \tilde{\gamma}^*), r)$ *be a closed ball with the centre* $(\tilde{\beta}^*, \tilde{\gamma}^*)$ *and a radius* $r > 0$. *Suppose that for each* x *and* $(\tilde{\beta}, \tilde{\gamma}) \in K((\tilde{\beta}^*, \tilde{\gamma}^*), r)$ *functions* $\log s_{\tilde{\beta},\tilde{\gamma}}(x)$ *and* $\log(1 - s_{\tilde{\beta},\tilde{\gamma}}(x))$ *are bounded from below by a function* $\eta(x)$ *such that* $E|\eta(X)| < \infty$. *Then with probability one, for sufficiently large* n *there exists a sequence* $(\hat{\beta}_n, \hat{\gamma}_n)$ *of local maximisers of* $Q_n(\tilde{\beta}, \tilde{\gamma})$ *such that* $(\hat{\beta}_n, \hat{\gamma}_n) \to (\tilde{\beta}^*, \tilde{\gamma}^*)$.

Note that the assumption $E|\eta(X)| < \infty$ imposed in Theorem 2 does not force $s(x)$ to be bounded way from 0 and 1 which is frequently assumed while dealing with consistency issues of estimates in the logistic model.

Finding the global maximiser of $Q_n(\tilde{\beta}, \tilde{\gamma})$ defined above is a complicated task as the optimised function is not concave in either $\tilde{\beta}$ or $\tilde{\gamma}$; see e.g. [23], where it is shown that Q_n is not concave even under SCAR when $e(x)$ is assumed constant. Thus we also introduce here a second approach which consists in iterative alternate solving for maxima of *concave* empirical likelihoods of $y(x)$ and $e(x)$.

Let $y(x, \tilde{\beta}) = \sigma(\beta_0 + \beta^T x)$ and $e(x, \tilde{\gamma}) = \sigma(\gamma_0 + \gamma^T x)$. We will thus look for solutions of empirical counterparts of two optimisation problems. Optimisation problem for $y(x, \tilde{\beta})$ is to maximise wrt $\tilde{\beta}$

$$\begin{aligned} &E_X W(X, \tilde{\beta}) = E_X[y(X, \tilde{\beta}^*)\log y(X, \tilde{\beta}) + \\ &(1 - y(X, \tilde{\beta}^*))\log(1 - y(X, \tilde{\beta}))], \end{aligned} \tag{4}$$

where $W(X, \tilde{\beta})$ is the bracketed expression above. Note that (4) is the expected value of the loglikelihood of (Y, X) in the double logistic model. Let $K(s, x, \tilde{\gamma}) = s \log e(x, \tilde{\gamma}) + (1 - s)\log(1 - e(x, \tilde{\gamma}))$. and note $E_{S|Y=1,X} S = e(X, \tilde{\gamma}^*)$. Using this equality, we note that optimisation problem for $e(x, \tilde{\gamma})$ can be approached via maximising wrt to $\tilde{\gamma}$

$$\begin{aligned} &E_{X|Y=1}[e(X, \tilde{\gamma}^*)\log e(X, \tilde{\gamma}) + \\ &(1 - e(X, \tilde{\gamma}^*))\log(1 - e(X, \tilde{\gamma}))] = \\ &E_{X|Y=1} E_{S|Y=1,X} K(S, X, \tilde{\gamma}) = \\ &E_{S,X|Y=1} K(S, X, \tilde{\gamma}), \end{aligned} \tag{5}$$

Notice that in the case of (4) for any $X = x$ maximiser of $W(x, \tilde{\beta})$ is $\tilde{\beta}^*$ (this can be seen reasoning analogously as in the case of Lemma 1), whereas in the case of (5) maximiser of $E_{S,X|Y=1} K(S, X, \tilde{\gamma})$ is $\tilde{\gamma}^*$. Thus $E_X W(X, \tilde{\beta})$ and $E_{S,X|Y=1} K(S, X, \tilde{\gamma})$ are Fisher consistent in the sense that maximisation over $\tilde{\beta}$ and $\tilde{\gamma}$ yields true parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$, see [25] for discussion of Fisher consistency. This is an important property as Fisher consistency implies strong consistency of empirical maximisers under mild assumptions. The obvious problem is that neither (4) nor (5) have direct empirical counterparts due to dependence on $y(x, \tilde{\beta}^*)$ in the first case and in the second case due to averaging over the unknown conditional distribution of (S, X) given $Y = 1$. In case of (4) we will solve this problem by introducing weights depending on propensity score such that the weighted risk based on these weights will equal $E_X W(X, \tilde{\beta})$. Then using the current estimator of $e(x, \tilde{\gamma}^*)$ we will define an approximation to its empirical counterpart which will be maximised. The obtained estimator of posterior probability will be used to approximate the stratum $\{Y = 1\}$ and the expected value with respect to $S, X|Y = 1$ and thus making evaluation of empirical counterpart of (5) feasible. Namely, for the first problem we want to find weights $w_1(s, x)$ and $w_0(s, x)$ such that

$$\begin{aligned} &W(x, \tilde{\beta}) = E_{S|X=x}[w_1(S, x)\log y(x, \tilde{\beta}) + \\ &w_0(S, x)\log(1 - y(x, \tilde{\beta}))]. \end{aligned} \tag{6}$$

Then maximising an empirical counterpart of (6) yields consistent estimator of $\tilde{\beta}^*$. In the case of the second optimisation we have to approximate expectation $E_{S,X|Y=1}$. For the first problem we have

Lemma 2 *Let* $w_1(S, x) = I\{S = 1\} + I\{S = 0\}P(Y = 1|S = 0, x)$ *and* $w_0(S, x) = I\{S = 0\}P(Y = 0|S = 0, x)$. *Then (6) holds.*

*P*roof. Observe that $W(x, \tilde{\beta})$ equals

$$\begin{aligned} &[P(S = 1|x) + P(S = 0|x)P(Y = 1|S = 0, x)]\log y(x, \tilde{\beta})] \\ &+[P(S = 0|x)P(Y = 0|S = 0, x)]\log(1 - y(x, \tilde{\beta}))] \end{aligned}$$

and the bracketed terms are equal $P(Y = 1|x)$ and $P(Y = 0|x)$, respectively.

The weights $w_i(s, x)$ were introduced in [8]. The lemma above states that they yield unbiased estimator of $W(x, \tilde{\beta})$. Note that the factor $P(Y = 1|S = 0, x)$ appearing in $w_1(S, x)$ equals

$$OR(x) = \frac{1 - e(x)}{e(x)} / \frac{1 - s(x)}{s(x)}$$

and thus is the odds ratio equal to the ratio of the odds of being unlabelled among positive observations and the odds of being unlabelled in the general population. In the view of this and (6) the empirical counterpart of $E_X W(X, \tilde{\beta})$ is defined as

$$\begin{aligned} &W_n(\tilde{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_1(S_i, X_i)\log y(X_i, \tilde{\beta}) + \\ &\hat{w}_0(S_i, X_i)\log(1 - y(X_i, \tilde{\beta})), \end{aligned} \tag{7}$$

where $\hat{w}_1(S_i, X_i) = I\{S_i = 1\} + I\{S_i = 0\}\widehat{OR}(X_i)$, $\hat{w}_0(S_i, X_i) = I\{S_i = 0\}(1 - \widehat{OR}(X_i))$ and $\widehat{OR}(x) = \frac{1-\hat{e}(x)}{\hat{e}(x)} / \frac{1-\hat{s}(x)}{\hat{s}(x)}$ is an estimator of $OR(x)$, which is discussed in Subsection 4.3. We will prove below that if $OR(x)$ is consistently estimated, then any maximiser of $W_n(\cdot)$ is consistent estimator of $\tilde{\beta}^*$. Indeed, notice that the key assumption (8) below is satisfied if $\sup_x |\widehat{OR}(x) - OR(x)| \to_P 0$.

Theorem 3 *Let* $\tilde{\beta}^*$ *be the unique maximiser of* $E_X W(X, \tilde{\beta})$ *and for each* $\tilde{\beta}$

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{OR}(X_i) I(S_i = 0)(\beta^T X_i + \beta_0) \to_P$$

$$E\left[OR(X)I(S=0)(\beta^T X+\beta_0)\right]. \quad (8)$$

Then every $\hat{\beta}_n = \arg\max_{\tilde{\beta}} W_n(\tilde{\beta})$ *tends to* β^* *in probability.*

The proof of Theorem 3 can be found in the Appendix 1.

We consider now the second problem that is consistent estimation of $\tilde{\gamma}^*$. Let $n_1 = \#\{1 \le i \le n : Y_i = 1\}$ be a number of positive observations in a data set. Let $\hat{Y}_i$'s be some predictors of unknown Y_i's and $\hat{n}_1 = \#\{1 \le i \le n : \hat{Y}_i = 1\}$. Besides, we consider a function $\hat{R}_n(\tilde{\gamma}) = \frac{1}{\hat{n}_1} \sum_{1\le i\le n:\hat{Y}_i=1} K(S_i, X_i, \tilde{\gamma})$, which we use to approximate

$$R(\tilde{\gamma}) := E_{S,X|Y=1} K(S, X, \tilde{\gamma})$$

in (5). Finally, changing $\hat{Y}_i$ to Y_i in the definition of $\hat{R}_n(\tilde{\gamma})$ we define $R_n(\tilde{\gamma}) = \frac{1}{n_1} \sum_{1\le i\le n:Y_i=1} K(S_i, X_i, \tilde{\gamma})$. Next, consistent estimation of $\tilde{\gamma}^*$ is considered.

Theorem 4 *Let* $\tilde{\gamma}^*$ *be the unique maximiser of* $R(\tilde{\gamma})$ *and for each* $\tilde{\gamma}$ *we have* $\hat{R}_n(\tilde{\gamma}) - R_n(\tilde{\gamma}) \to_P 0$. *Then every maximiser of* $\hat{R}_n(\tilde{\gamma})$ *tends to* $\tilde{\gamma}^*$ *in probability.*

The proof of Theorem 4 can be found in the Appendix 1.

4 Algorithms

4.1 NAIVE method

We first describe the NAIVE method which is the simplest approach in PU learning. In this method, estimator of $s(x) = P(S = 1|X = x)$ is substituted for estimator of $y(x) = P(Y = 1|X = x)$. To this end misspecified empirical loglikelihood

$$\sum_{i=1}^{n} S_i \log(y(X_i, \tilde{\alpha})) + (1 - S_i) \log(1 - y(X_i, \tilde{\alpha}))$$

is optimised with respect to $\tilde{\alpha}$. The estimator of $s(x)$ is defined as $\hat{s}_{\text{naive}}(x) = y(x, \hat{\alpha})$, where $\hat{\alpha}$ is the maximizer of the above function. Obviously, this method underestimates $y(x)$ as the positive unlabelled observations are treated as the negative ones and the bias increases with decreasing label frequency $c = P(S = 1|Y = 1)$. The NAIVE method serves as a baseline in our experiments. Moreover, using the naive method, one can consider a very simple estimator of the propensity score function $e(x)$ which will serve as initial estimator in the methods described in next subsections. It is based on inequality $s(x) \le e(x) \le 1$ and is defined as an average of two endpoints of the interval $[\hat{s}_{\text{naive}}(x), 1]$, i.e. $\hat{e}_{\text{naive}}(x) = 0.5(\hat{s}_{\text{naive}}(x) + 1)$.

4.2 JOINT method

We optimize function $Q_n(\tilde{\beta}, \tilde{\gamma})$ defined in Section 3 with respect to $\tilde{\beta}$ and $\tilde{\gamma}$, alternately. We repeat the following two steps for $k = 1, 2, \ldots,$ until convergence:

1. Solve $\hat{\beta}_n^{(k)} = \arg\max_{\tilde{\beta}} Q_n(\tilde{\beta}, \hat{\gamma}_n^{(k-1)})$.
2. Solve $\hat{\gamma}_n^{(k)} = \arg\max_{\tilde{\gamma}} Q_n(\hat{\beta}_n^{(k)}, \tilde{\gamma})$.

In the first iteration we need some initial estimator $\hat{\gamma}_n^{(0)}$ or equivalently initial estimator of $e(X_i, \hat{\gamma}_n^{(0)})$, because $Q_n(\tilde{\beta}, \hat{\gamma}_n^{(0)})$ involves an unknown term $e(X_i, \hat{\gamma}_n^{(0)})$. For this we use $\hat{e}_{\text{naive}}(x)$ defined in Subsection 4.1. As $Q_n(\tilde{\beta}, \tilde{\gamma})$ is not concave in either $\tilde{\beta}$ or $\tilde{\gamma}$, Minorisation-Maximisation (MM) algorithm (see e.g. [12], Section 5.8) is used to find the maximisers in 1 and 2. The analogous idea was used in [23, 34], who assumed SCAR and optimized jointly with respect to $\tilde{\beta}$ and label frequency c.

4.3 TWO MODELS method (TM)

The proposed method involves fitting two models in each iteration. The first model aims to estimate $y(x)$, whereas the second model corresponds to $e(x)$. In the case of the second model, our goal is to first approximate the stratum $\mathcal{P} := \{i : Y_i = 1\}$. We define its estimator as $\hat{\mathcal{P}} = \{i : S_i = 1 \text{ or } \hat{y}(X_i) > t\}$ where $\hat{y}(X_i)$ is an estimator of $P(Y = 1|X_i)$ obtained from the first model and t is a threshold. For the threshold we use a quantile of order α of the set $\{\hat{y}(X_i)$ for i such that $S_i = 1\}$ (a data-adaptive choice of α is discussed below). Next, we estimate $e(X_i)$ by fitting the logistic model using observations (X_i, S_i) for $i \in \hat{\mathcal{P}}$. More specifically, we repeat the following steps until convergence:

1. **Model 1.** Solve $\hat{\beta}_n = \arg\max_{\tilde{\beta}} W_n(\tilde{\beta})$, where $W_n(\tilde{\beta})$ is defined in (7).
2. Calculate $\hat{y}(X_i) = y(X_i, \hat{\beta}_n)$.
3. **Model 2.** Solve $\hat{\gamma}_n = \arg\max_{\gamma} \hat{R}_n(\tilde{\gamma})$, where

$$\hat{R}_n(\tilde{\gamma}) = \sum_{i=1}^{n} I(i \in \hat{\mathcal{P}}) K(S_i, X_i, \tilde{\gamma}),$$

   where K is defined below (4).
4. Calculate $\hat{e}(X_i) = e(X_i, \hat{\gamma})$.
5. Update $\hat{s}(X_i) = \hat{e}(X_i)\hat{y}(X_i)$ and $\widehat{OR}(X_i) = \frac{1-\hat{e}(X_i)}{\hat{e}(X_i)} / \frac{(1-\hat{s}(X_i)}{\hat{s}(X_i)}$.

Note that in step 1, function $W_n(\tilde{\beta})$ depends on $\widehat{OR}(X_i)$, thus in the first iteration some initial estimators of $s(x)$ and $e(x)$ are required. Similarly to the JOINT method, we use the naive estimators $\hat{s}_{\text{naive}}(x)$ and $\hat{e}_{\text{naive}}(x)$ described in Subsection 4.1. The significant advantage of TM is concavity, which allows to avoid problems with local minima. It is especially important when working with 'larger' data sets, say $p \ge 50$ and $n \ge 1000$.

An important issue is the order of the quantile α of $\{\hat{y}(X_i)$ for i such that $S_i = 1\}$ used in $\hat{\mathcal{P}}$. Small value of α allows to detect significant portion of the positive observations among unlabelled ones. On the other hand, a larger value of α reduces the risk of including negative examples from the set of unlabelled ones. The choice of the optimal value of α is a challenging task as it should depend on two factors: the difficulty of the classification problem, i.e. on how much the distributions $X|Y = 1$ and $X|Y = 0$ overlap, as well as on label frequency $c = P(S = 1|Y = 1)$. It follows from our experiments that when the distributions $X|Y = 1$ and $X|Y = 0$ are practically disjoint it is better to take a smaller α (for example $\alpha = 0.1$), especially for small c. When the distributions of $X|Y = 1$ and $X|Y = 0$ significantly overlap, large α is preferable (i.e. selection should become more conservative), especially when c approaches 1. In the latter case, small α results in a large number of 'false positive' observations, i.e. the set $\hat{\mathcal{P}}$ contains a significant number of negative examples, which deteriorates the performance of the method. The above insights suggest that α should increase in a certain manner when label frequency increases. In our method we use $\alpha = \hat{P}(S = 1)$, which is motivated by a simple inequality $P(S = 1) \le P(S = 1|Y = 1)$. Although this choice of α gives very good results, we believe that the problem is worth further analysis as it is crucial for estimation of $e(x)$.

In addition to TM method, we also consider its simplified version (called TM SIMPLE) in which we do not estimate $e(x)$ in iterative manner. Instead, we estimate the propensity score by $\hat{e}_{\text{naive}}(\cdot)$ and then we solve $\arg\max_{\tilde{\beta}} W_n(\tilde{\beta})$. Comparison between TM and TM SIMPLE allows to explore the effectiveness of employing the

Model 2 in TM. In addition, TM SIMPLE is much faster than TM as it does not require running many iterations. For example, for the largest considered dataset Adult, the average computation time is 0.5 sec for TM simple and 11.5 sec for TM (PC Intel Core i7-10850H CPU 2.70GHz, 32.0 GB RAM).

5 The related methods

In this section we describe two existing methods, which are most related to our proposals: EM method proposed in [5] and LBE method proposed in [11]. We keep the names of the methods according to the way they where christened by the authors, although we note (see below) that LBE is a classical EM algorithm applied in PU setting and thus the name 'EM method' would be actually more appropriate for LBE.

LBE method relies on parametric assumptions (2) and is based on considering an averaged conditional likelihood for the sample $(S_i, Y_i), i = 1, \ldots, n$ given $X_1, \ldots, X_n$, namely

$$E_{Y_1,\ldots,Y_n|S_1,\ldots,S_n,X_1,\ldots,X_n} \log \mathcal{L}(\beta,\gamma) = \sum_{i=1}^{n} E_{\tilde{P}(Y_i)}\{\log P(Y_i|X_i,\beta)P(S_i|Y_i,X_i,\gamma)\}, \quad (9)$$

where $\tilde{P}(Y_i) = P(Y_i|X_i, S_i)$ and

$$\mathcal{L}(\beta,\gamma) = P(S_1, Y_1, \ldots, S_n, Y_n|X_1, \ldots, X_n, \beta, \gamma) = \prod_{i=1}^{n} P(Y_i|X_i,\beta)P(S_i|Y_i,X_i,\gamma),$$

where the last equality is due to independence of observations. In the expectation step (E-step) binary distribution $\tilde{P}(Y_i)$ is estimated based on current estimates of β and γ using (2), Bayes formula and a normalisation trick which is applied to calculate estimate of $P(S_i|X_i)$. In the maximisation step (M-step) current estimate of $\tilde{P}(Y_i)$ is employed to calculate (9) which is then maximised using Adam ([20]) algorithm yielding the values of β and γ for the next E-step.

The main difference between our proposal TM and LBE method is difference in criterion functions to be optimised. In [11] it is (9), whereas our TM method is based on alternate maximisation of concave log-likelihoods pertaining to posterior probabilities $y(X_i)$ and propensity scores $e(X_i)$, respectively. What is more, the theoretical analysis in [11] does not address the identification issue studied in Theorem 1 and thus leaves the question of consistent estimation of the true vector of parameters $(\beta^{*T}, \gamma^{*T})$ unanswered.

In EM algorithm proposed in [5], the maximization step in EM approach is similar to ours but with one crucial difference. Namely, in their approach estimation of $e(x)$ is based on maximisation of estimated value of $E_{S,X,Y} YK(S, X, \gamma) = E_{S,X} P(Y = 1|S, X)K(S, X, \gamma)$, whereas we consider maximisation of estimated $E_{S,X|Y=1}K(S, X, \gamma)$. Both expressions are Fisher consistent. Their approach leads to consideration of weights being estimates of $P(Y = 1|S, X)$ instead of $I(i \in \hat{\mathcal{P}})$ in step 3 of TM algorithm. Criterion function $W_n(\tilde{\beta})$ used in step 1 is the same for both methods. We show in numerical experiments below that by using the proposed Fisher consistent expression for $e(x)$ and suitable approximation of the stratum $\mathcal{P}$ we obtain the method which is in most cases superior to EM algorithm.

6 Numerical experiments

In the experiments we compare the following methods: NAIVE, JOINT, TM [1], TM SIMPLE (described in Section 4) and two most related methods EM [5] and LBE [11] (described in Section 5). We also consider ORACLE method which assumes the full knowledge of a class variable Y and calculates maximum likelihood estimator for the logistic fit. The ORACLE method serves as a reference method and obviously it cannot be used in practise for PU data. To make a comparison fair, in TM, JOINT, EM and LBE we use the same stopping criterion, namely the convergence is reached when the relative change in the consecutive values of the objective function used in all methods to estimate $\tilde{\beta}$, is less than 10^{-6} or the number of iterations exceeds 1000. Importantly, all methods are based on logistic model and thus the comparison between them is reliable.

In each experiment, the data sets are randomly split into training set (70%) and testing set (30%). We repeat the experiments 100 times and average the results. Two evaluation measures are considered: accuracy calculated on the testing set and approximation error defined as

$$AE = n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} |y(x_i, \hat{\beta}) - y(x_i, \hat{\beta}_O)|,$$

where $\hat{\beta}$ is the solution of the considered method, $\hat{\beta}_O$ is the solution of the ORACLE method and n_{test} is the size of the testing set. In the case of artificial data sets, we replace $y(x_i, \hat{\beta}_O)$ by the true posterior probability $P(Y = 1|X_i)$. The AE shows how close are the predictions of the considered methods to the predictions of ORACLE method (or to the true posterior probabilities for artificial datasets). Small value of the AE indicates that the performance of the considered PU method is similar to that of the oracle. In the experiments, we observe that for some datasets the differences between the methods are more pronounced for AE than for standard accuracy.

There are several important questions which are addressed in the experiments. First, how much do we lose compared to the oracle method and how much the proposed methods improve the prediction accuracy of the naive approach? Secondly, what is the impact of various labelling schemes and how the performance deteriorates with decreasing label frequency? Thirdly, how robust are the considered methods against deviations from the logistic model?

6.1 Data sets

We consider two artificial data sets: Artif1 and Artif2. They are obtained as follows. We first generate feature vector $X \sim N(0, I)$, where I is $p \times p$ identity matrix. Then we pick the true class variable from Bernoulli distribution with success probability $P(Y = 1|X) = F(X^T\beta)$, where $\beta = p^{-1/2}(1, \ldots, 1)^T$ and F denotes the probability distribution function (PDF). We consider two forms of F: in the case of Artif1 we use PDF of the standard logistic distribution ($F(s) = \sigma(s)$) and in the case of Artif2 we use PDF of the standard Cauchy distribution (location 0 and scale 1). Since all considered methods are based on logistic model, Artif1 corresponds to the correct specification of the fitted model, whereas Artif2 corresponds to the misspecified model. Note that as Cauchy distribution significantly differs from logistic distribution in the tails, Artif2 is strongly misspecified. The advantage of using artificial datasets is that the true posterior probability is known (as opposed to the real datasets) and therefore it is possible to assess how accurately the

[1] Source code of the proposed method is available at GitHub: https://github.com/teisseyrep/putm

considered methods estimate the posterior probability. In addition to artificial datasets, we consider 8 benchmark datasets from UCI Machine Learning Repository. A short summary of each dataset can be found in Table 1 in Appendix 2. They were chosen to account for variable characteristics of data (number of observations, number of features and difficulty of the classification problem). The penultimate column of the Table 1 contains the values of R^2 (proportion of explained deviance) calculated for the ORACLE method; the R^2 describes the goodness of fit of the model (the larger the better) and can be treated as a measure of the difficulty of classification problem. For the considered datasets R^2 ranges from 0.23 to 0.93.

6.2 Labelling scenarios

We consider three methods of generating observed target variable S based on the true target variable Y. The first scenario corresponds to SCAR assumption whereas for the two remaining schemes SCAR is violated. The second scenario is similar to the one considered in [11] where the logistic sigmoid function is also used to describe the propensity score. The last scenario was considered in [2] and is similar to the one considered in [5].

1. **Scenario 1.** We consider constant propensity score function $e(x) = P(S = 1|Y = 1, X = x) = c$, where c is label frequency which varies in simulations (SCAR scenario).
2. **Scenario 2.** Logistic propensity score function $e(x) = \sigma(x^T\gamma)$ is considered, where $\gamma = p^{-1/2}(g, \ldots, g)^T$ and g is a parameter which varies in simulations. Obviously, for $g = 0$ the scenario reduces to scenario 1 with $c = 0.5$.
3. **Scenario 3.** Propensity score function is defined as $e(x) = \prod_{j=1}^{k}[sc(x(j), p^-, p^+)]^{1/k}$, where $x(j)$ is j-th coordinate of x and $sc(x(j), p^-, p^+) := p^- + \frac{x(j)-\min x(j)}{\max x(j)-\min x(j)}(p^+ - p^-)$ and k is a chosen integer smaller than p.

6.3 Results

We first analyse labelling scenario 1. Figures 1, 3 show the impact of label frequency c on the accuracy and the approximation error for labelling scheme 1; see also Fig. 1 in Appendix 2 which shows approximation errors for scenario 1 for benchmark datasets. As expected, the performance deteriorates with decreasing c and all the curves approach the oracle curve for $c \approx 1$. The proposed TM method is the clear winner for both artificial data sets and is among the best performing methods for most benchmark datasets. The advantage of TM becomes significant for small c, e.g. for Breast Cancer and spambase data sets its approximation errors for $c = 0.25$ are around 50% smaller than approximation errors for EM and LBE. Importantly, TM usually works better than two most related methods: EM and LBE as well as TM SIMPLE, which indicates that the proposed method of the propensity score estimation is crucial for the improved performance. Generally, the ranking of the method depends on particular dataset and value of c. For example, the LBE method is a clear winner for heart-c and wdbc for larger c, but it works significantly worse than other methods for Breast Cancer and spambase when c is small. On the other hand, the TM is a clear winner for these two datasets (Breast Cancer and spambase) and small c, but it works much worse for adult dataset and larger c. As expected, we observe the highest approximation errors for the NAIVE method which is due to the fact that it tends to underestimate $y(x)$ and the bias increases with decreasing c. Interestingly, we obtain similar results for Artif1 and Artif2 datasets, which shows that the methods are robust against model misspecification, even when the misspecification is strong. The JOINT method works worse than the TM which is probably associated with the optimization issues in the case for the JOINT method, which uses MM algorithm to search of the maximizer. There is certainly room for improvement in the JOINT method. Tables 2 and 3 in Appendix 2 show the performance measures averaged over different $c = 0.05, 0.1, \ldots, 0.95$. We used the simple t-test to verify whether the difference between the winner method (in bold) and the second best is significant. The last column contains the p-value of the t-test. Importantly, in most cases, the p-value is smaller than $0.05/15 = 0.003$. Note that 0.003 is a level corresponding to Bonferroni correction as 0.05 is a standard significance level and 15 is the number of pairs among 6 considered methods. We remark that the proposed TM method has the lowest rank averaged over data sets; it is a winner for 6/10 datasets (with respect to accuracy) and for 8/10 datasets (with respect to AE).

The results for different values of parameter g for scenario 2 are presented in Figure 2. For artificial datasets, we observe stable results for all methods (except the naive method), i.e. the accuracy and AE do not vary significantly for different values of parameter g. In addition, we can see the superior performance of the TM method. The TM exhibits the most stable performance among the methods. For naive method, the accuracy increases with g and the AE decreases with g. Tables 4 and 5 in Appendix 2 show the overall results for scenario 2 averaged over different $g = 0.1, 0.2, \ldots, 1$. Here, the LBE methods achieves on average the highest accuracy (it has the smallest averaged ranks), whereas the TM achieves on average the smallest approximation error.

Tables 6 and 7 in the Appendix 2 show the results for scenario 3, for $p^- = 0.2$, $p^+ = 0.6$ and $k = 5$. The conclusions remain very similar to those of the first and the second scheme. Namely, the TM method achieves the smallest approximation error (whereas the LBE is the second best); it is a winner for 8/10 datasets (wrt to AE) and for 5/10 datasetes (wrt accuracy).

In summary, we conclude that the averaged ranks over data sets are the smallest for TM and LBE, with EM and TM SIMPLE being the third and the fourth method, respectively. This is a positive message, it means that the TM method (as well as LBE and EM) allows to estimate $y(x)$ accurately in a general case when the SCAR assumption is not necessarily satisfied. Surprisingly, the TM SIMPLE method works quite well although it is based on a crude estimator of the propensity score function. For example it outperforms other methods for diabetes dataset and for scenario 2, for which the propensity score is not constant. As the TM SIMPLE is much faster than the remaining competitors (TM, EM, LBE and JOINT), it can be recommended in applications where reducing computational time is crucial.

7 Conclusions

We have considered estimation problem for PU data when SCAR assumption is not necessarily satisfied and have shown that when posterior probability $y(x)$ and propensity function $e(x)$ are both governed by the double logistic model, their corresponding parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ are identifiable. This motivates JOINT method of estimation of $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ by alternately maximising loglikelihood of the product logistic model. We have also proposed the second method, called TM, which relies on iterative maximisation of two estimated Fisher consistent expressions for the unknown parameters. For both approaches under certain assumptions we have proved consistency of the underlying estimators. Analysis of their behaviour indicates that considering non-constant propensity function is crucial for estimation of the

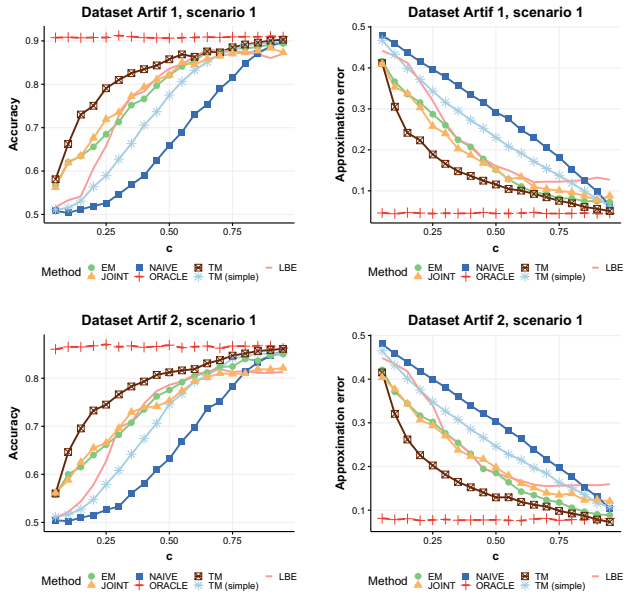

Figure 1. Accuracy and approximation errors for datasets Artif1 and Artif2 for scenario 1 and different values of c.

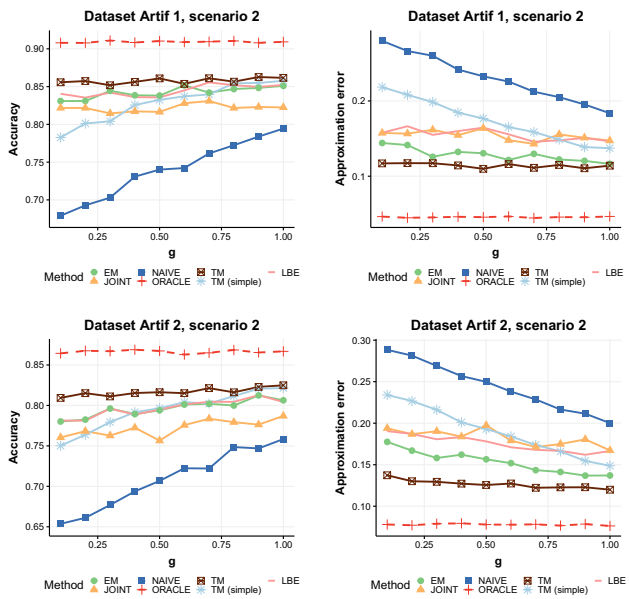

Figure 2. Accuracy and approximation errors for datasets Artif1 and Artif2 for scenario 2 and different values of g.

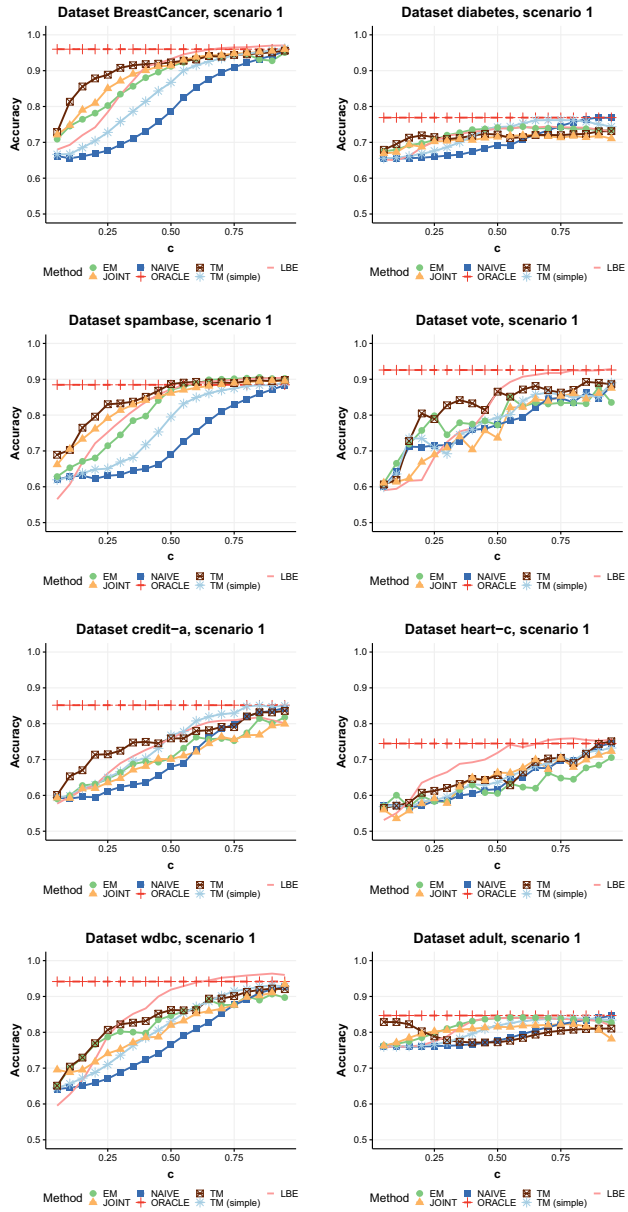

Figure 3. Accuracy for benchmark datasets for scenario 1 and different values of c.

aposteriori probability as well as for performance of the corresponding classifiers. In particular, the results show promising behaviour of TM method; it outperforms EM algorithm for most datasets and works on par with LBE method in the case of non-constant propensity score function. For constant propensity score, we observe the superior performance of the TM compared to the LBE.

There are still interesting issues left for future research. First, in addition to MM algorithm, other non-convex optimization procedures can be used in the JOINT method.This is important as its underperformance is likely caused by optimisation issues. As for TM method, although the proposed method of estimation of the stratum $\mathcal{P}$ described in Section 4.3 works effectively, we believe that this crucial problem is worth further studying and there is still room for improvement. For the high-dimensional X consideration of the regularised versions of the introduced methods is of interest. Moreover, note that the presented developments open the way to test SCAR assumption, which under the considered model is equivalent to $\tilde{\gamma} = (\tilde{\gamma}, 0^T)^T$. Finally, the proposed methods can be possibly adapted to multi-label PU data, where multiple target variables are considered simultaneously and in many aplications, such as recommender systems, selection bias is frequent.

Acknowledgements

The comments of the reviewers, which helped to improve the final version of paper, are acknowledged. The research of W.R. was supported by the Polish National Science Center grant: NCN UMO-2018/31/B/ST1/00253.

References

[1] A. L. Bahorik, C. E. Newhill, C. C. Queen, and S. M. Eack, 'Under-reporting of drug use among individuals with schizophrenia: prevalence and predictors', *Psychological Medicine*, **44**(1), 61–69, (2014).

[2] J. Bekker, *Learning from Positive and Unlabeled Data*, Ph.D. dissertation, Informatics Section, Faculty of Engineering Science, Science, Engineering and Technology Group, December 2018. Davis, J. (supervisor).

[3] J. Bekker and J. Davis, 'Estimating the class prior in positive and unlabeled data through decision tree induction', in *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, (February 2018).

[4] J. Bekker and J. Davis, 'Learning from positive and unlabeled data: a survey', *Machine Learning*, **109**, 719–760, (2020).

[5] J. Bekker, P. Robberechts, and J. Davis, 'Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data', in *Proceedings of the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML'19, pp. 71–85, (2019).

[6] L. Cerulo, C. Elkan, and M. Ceccarelli, 'Learning gene regulatory networks from only positive and unlabeled data', *BMC Bioinformatics*, **11**, 228, (2010).

[7] M. du Plessis and M. Sugiyama, 'Class prior estimation for positive and unlabeled data', *IEICE Transactions on Information and Systems*, **E-97D**(5), 1358–1372, (2014).

[8] C. Elkan and K. Noto, '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, (2008).

[9] G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu, 'Text classification without negative examples revisit', *IEEE Transactions on Knowledge and Data Engineering*, **18**(1), 6–20, (2006).

[10] W. Gerych, T. Hartvigsen, L. Buquicchio, E. Agu, and E. Rundensteiner, 'Recovering the propensity score from biased positive unlabeled data', in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'22, pp. 6694–6702, (2022).

[11] C. Gong, Q. Wang, T. Liu, B. Han, J. You, J. Yang, and D. Tao, 'Instance-dependent positive and unlabeled learning with labeling bias estimation', *IEEE Trans Pattern Anal Mach Intell*, 1–16, (2021).

[12] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC, London, 2015.

[13] F. He, T. Liu, J. Webb, and D. Tao, 'Instance-dependent pu learning by bayesian optimal relabeling'. manuscript, 2018.

[14] J. Huang, J. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, 'Correcting sample selection bias by unlabeled data', in *Proceedings of the 2006 Conference: Advances in Neural Information Processing Systems*, NIPS'07, pp. 71–85, (2007).

[15] H. Jain, J. Prahbu, and M. Varma, 'Extreme multi-label loss function for recommendation', KDD, pp. 935–944, (2018).

[16] S. Jain, M. White, and P. Radivojac, 'Estimating the class prior and posterior from noisy positives and unlabeled data', in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, p. 2693–2701, (2016).

[17] K. Jaskie, C. Elkan, and A. Spanias, 'A modified logistc regression for positive and unlabeld learning', in *53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 2007–2011, (2020).

[18] M. Kato, T. Teshima, and J. Honda, 'Learning from positive and unlabeled data with a selection bias', in *Proceedings of International Conference on Learning Representations*, pp. 1–12, (2019).

[19] U. Kesmodel, 'Information bias in epidemiological studies with a special focus on obstetrics and gynecology', *Journal of Obsteyrics and Ginecology*, **97**, 417–423, (2018).

[20] D. Kingma and J. Ba, 'Method of stochastic optimisation'. CORR, vol.abs/1412.6980, 2015.

[21] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, '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, (2017).

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

[23] M. Łazecka, J. Mielniczuk, and P. Teisseyre, 'Estimating the class prior for positive and unlabelled data via logistic regression', *Advances in Data Analysis and Classification*, **15**, 1039–1068, (2021).

[24] M.F. Leitzmann and S Rohrmann, 'Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates', *Clinical Epidemiology*, **4**, 1–11, (2012).

[25] K. Li and N. Duan, 'Regression analysis under link violation', *The Annals of Statistics*, **17**(3), 1009–1052, (1989).

[26] X. Li and B. Liu, 'Learning to classify texts using positive and unlabeled data', in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, p. 587–592, (2003).

[27] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, 'Building text classifiers using positive and unlabeled examples', in *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pp. 179–186, (2003).

[28] B. Na, H. Kim, K. Song, Kim Y.-Y., and I.-C. Moon, 'Deep generative positive-unlabeled learning under selection bias', in *Proceedings of Conference on Information and Knowledge Management*, CIKM'20, pp. 1155–1164, (2020).

[29] J. L. Pearce and M. S. Boyce, 'Modelling distribution and abundance with presence-only data', *Journal of Applied Ecology*, **43**(3), 405–412, (2006).

[30] M. C. Plessis, G. Niu, and M. Sugiyama, 'Class-prior estimation for learning from positive and unlabeled data', *Machine Learning*, **106**(4), 463–492, (2017).

[31] H. Ramaswamy, C. Scott, and A. Tewari, 'Mixture proportion estimation via kernel embeddings of distributions', in *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 2052–2060, (2016).

[32] E. Sansone, F. De Natale, and Z.-H. Zhou, 'Efficient training for positive unlabeled learning', *IEEE Transactions on Pattern Recognition and Machine Inteligence*, **78**, (2018).

[33] K. Sechidis, M. Sperrin, E. S. Petherick, M. Luján, and G. Brown, 'Dealing with under-reported variables: An information theoretic solution', *International Journal of Approximate Reasoning*, **85**, 159 – 177, (2017).

[34] P. Teisseyre, J. Mielniczuk, and M. Łazecka, 'Different strategies of fitting logistic regression for positive and unlabelled data', in *Proceedings of the International Conference on Computational Science*, ICCS'20, (2020).

[35] D. Vistisen, D. R. Witte, A. G. Tabák, Ch. Herder, E. J. Brunner, M. Kivimäki, and K. Færch, 'Patterns of obesity development before the diagnosis of type 2 diabetes: The whitehall ii cohort study', *PLOS Medicine*, **11**(2), 1–14, (2014).

[36] N. M. Walley et al., 'Characteristics of undiagnosed diseases network applicants: implications for referring providers', *BMC Health Services Research*, **18**(1), 1–8, (2018).

[37] G. Ward, T. Hastie, S. Barry, J. Elith, and J. Leathwick, 'Presence-only data and the EM algorithm', *Biometrics*, **65**, 554–563, (2009).

[38] P. Yang, X. Li, H. N. Chua, C.K. Kwoh, and S.K. Ng, 'Ensemble positive unlabeled learning for disease gene identification', *PLOS ONE*, **9**(5), 1–11, (2014).

[39] C. Zhang, T. Ren, T Liu, J. Yang, and C. Gong, 'Positive and unlabeled learning via label disambiguation', in *Proceedings of International Joint Conference on Artificial Intelligence*, ICAI'19, pp. 420–4256, (2019).