ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251147

A Generalized Approach to Label Shift: The Conditional Probability Shift Model

Paweł Teisseyre [a,*] **and Jan Mielniczuk** [b]

[a]Warsaw University of Technology, Warsaw, Poland
[b]Polish Academy of Sciences, Warsaw, Poland

Abstract. In many practical applications of machine learning, a discrepancy often arises between a source distribution from which labeled training examples are drawn and a target distribution for which only unlabeled data is observed. Traditionally, two main scenarios have been considered to address this issue: covariate shift (CS), where only the marginal distribution of features changes, and label shift (LS), which involves a change in the class variable's prior distribution. However, these frameworks do not encompass all forms of distributional shift. This paper introduces a new setting, Conditional Probability Shift (CPS), which captures the case when the conditional distribution of the class variable given some specific features changes while the distribution of remaining features given the specific features and the class is preserved. For this scenario we present the Conditional Probability Shift Model (CPSM) based on modeling the class variable's conditional probabilities using multinomial regression. Since the class variable is not observed for the target data, the parameters of the multinomial model for its distribution are estimated using the Expectation-Maximization algorithm. The proposed method is generic and can be combined with any probabilistic classifier. The effectiveness of CPSM is demonstrated through experiments on synthetic datasets and a case study using the MIMIC medical database, revealing its superior balanced classification accuracy on the target data compared to existing methods, particularly in situations of conditional distribution shift and no prior distribution shift, which are not detected by LS-based methods.

1 Introduction

1.1 Motivation

In many real world applications aiming at classification of objects, the source distribution P, from which we sample labeled training examples, with a label denoting a class variable, differs from the target distribution Q, where only unlabeled data is observed [37, 3, 14]. This situation occurs, for example, in medical applications, where, models trained on patients from one hospital do not necessarily adequately generalize to new data from other hospitals [38, 28]. The shift in the distribution of diseases and/or patient parameters may be related, among other factors, to socio-economic differences in the studied populations or outbreak of an epidemic [25]. Also, in banking, distribution of credit repayments and defaults may change in time [30].

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

In general, it is not possible to make inferences for the target data based on information obtained solely from the source dataset [8], and further assumptions are needed to enable training the classifier. There are two natural scenarios, which can be considered: covariate shift (CS) and label shift (LS). Under the covariate shift (CS) scheme, we assume that only the marginal distribution of features changes, and the posterior distribution remains unchanged. Label shift (LS) stipulates that the prior distribution of the class variable changes, while the conditional distribution of features given labels remains the same for the source and target datasets. LS aligns with the anti-causal setting in which the class variable determines the features, e.g. in medical applications, where symptoms are caused by disease [27, 16]. The LS assumption stipulates that the distribution of symptoms in the presence of the disease is the same in the source and target data.

Obviously, the above two scenarios do not cover all possibilities of distribution shifts. Recently a novel SJS (Sparse Joint Shift) assumption [6, 31, 32] has been introduced, which is a generalization of label shift and sparse covariate shift. Namely, it is assumed that distribution of class variable and some features change, but the remaining features' distribution conditional on the shifted features and class variable is fixed.

SJS encompasses, as a special case the shift in the marginal distribution of the chosen features only. This situation, however, is of minor interest for the classification problem because, as we show in the following, the class posterior probabilities for the source and the target datasets match in this case. Much more interesting and challenging is the shift of the conditional distributions of the class variable given some features. Motivated by this, in this paper we introduce a new setting called CPS (conditional probability shift), which covers the most interesting case within the SJS framework.

In particular, we can observe a shift in the conditional distribution of the class variable given some features, while its marginal (prior) distribution is preserved, which means that there is no label shift. To stress this important point, we discuss the following example (see Figure 1) in which the class variable is the occurrence of a disease, and additionally we consider the age variable with two categories: '$\leq$40 years' and '$>$40 years'. For the source data, the probability of succumbing to the disease in the group of older people (equal 0.4) is twice as high as in the group of younger people (equal 0.2). In the target data, on the other hand, the probability increases in the older group to 0.5, while in the younger group it equals 0.1. Therefore, we observe a change in the conditional distribution of the disease, while its prior distribution, in the case where both ages of the group are equally likely, remains the same for the source and target data and is

equal 0.3. The above situation may occur when the target data refers to a population in which disease prevention for the elderly is at a low level compared to the source population.

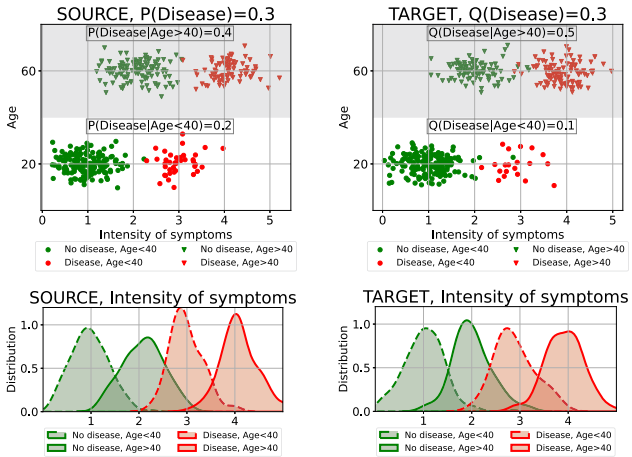

Figure 1. Example of a shift in the conditional distribution of a class variable denoting the occurrence of a disease given the age (1st row) : $P(Disease|Age) \neq Q(Disease|Age)$. The prior distribution of the disease is the same for the target and source distributions $P(Disease) = Q(Disease)$ (no label shift). The empirical conditional distribution of the feature (intensity of symptoms) given disease and age are approximately the same for source and target data up to estimation error (2nd row).

1.2 Contribution

In this work, we propose a new model, called CPSM (Conditional Probability Shift Model) that relies on the CPS assumption. The main idea of the method is to model the conditional probability of a class variable given some of the features using multinomial models for the source and target data. Estimating the model parameters for the target data is challenging because we do not observe the class variable. Therefore, we use the Expectation-Maximisation algorithm, which allows us to iteratively estimate the parameters of the multinomial model as well as the posterior probability of the class on the target data.

Experiments conducted on artificial data as well as a case study on the medical MIMIC database show that the method provides higher classification accuracy than competing methods dedicated to LS and SJS schemes. In particular, there is a significant advantage over methods dedicated to the LS scheme (such as BBSC or MLLS described below) in the case when the conditional distribution is shifted but the prior probability of the target variable does not change.

1.3 Related work

The shift of distributions between source and target data is an important problem in machine learning and therefore in recent years much attention has been paid to detecting such a situation [24, 15] and adapting models based on the source data [7, 29, 16].

The CS scheme involving a shift in feature probability distribution is less conceptually demanding, since the feature vector is observed for both the source and target data and posterior probabilities do not change. The dominant approach is based on weighted risk minimization [29], where the weights are usually determined using matching techniques, such as Kernel Mean Matching [12].

Label Shift (LS) and a more general sparse joint shift (SJS) are the two most closely related frameworks to the scenario studied here. Under LS assumption, there are two dominant approaches: BBSC (Black-Box Shift Correction) [16] and MLLS (Maximum Likelihood Label Shift) [26]. BBSC method introduces importance weights that are estimated using the confusion matrix. The weights are then incorporated into the risk function, which is optimized on the source data to train a final classifier. Regularized Likelihood Label Shift (RLLS) [2] is a variant of BBSC which introduces novel regularization of weighted likelihood to compensate for the high approximation error of the importance weights. The approach based on weighted risk function is also used in other methods, such as the method called ExTRA [18], where the weights are estimated using exponential tilt model and distribution matching technique. The ExTRA method allows for covariate and label shift to occur at the same time. The above methods require retraining; the first model is trained on the original source dataset, and the second model is trained on the dataset with assigned sample weights. This can be computationally expensive, especially for large-scale data. On the other hand, the MLLS method [26] avoids this drawback by calibrating class posterior probabilities. The method exploits the fact that the class posterior probability for the target data can be written as a transformation of the class posterior probability for the source data, with the scaling factor depending on the unknown target class prior. Since the class variable is not observable for the target data, the EM procedure is used, which allows iteratively estimating the posterior and prior probabilities for the target data. The method proposed in this paper is based on a similar idea, but the main difference is that instead of estimating the class prior, we estimate the parameters of a multinomial model associated with the conditional probabilities. When combined with deep neural networks, the MLLS method can be further improved by using a calibration heuristic called Bias-Corrected Temperature Scaling (BCTS) [1]. In [9], the BBSC and MLLS are analyzed theoretically, in particular consistency conditions for MLLS, including calibration of the classifier and a confusion matrix invertibility, are provided. LS problem is recently modified to allow for partial observability of the source data, e.g. in [20] it is assumed that it is Positive-Unlabeled (PU).

The SJS scheme, which generalizes LS, was proposed very recently and therefore not as many algorithms have been developed yet as for LS. SEES [6] is, to the best of our knowledge, the sole method developed under SJS assumption. Discussions on the theoretical properties of the SJS scheme can be found in [32]. SEES determines sample weights by minimizing distance between the induced feature density and the target feature density.

Finally, let's mention that there are other interesting issues related to the distribution shift. For example, domain adaptation is a research area concerned with a problem for which data availability is similar to that discussed here. Namely, for Unsupervised Domain Adaptation (UDA) the source data consists of a set of labeled data and unlabeled data, the latter possibly coming from different distribution than the labeled data. The task then consists of building a classifier based *jointly* on those two types of source data, which is meant to classify new observations following a target distribution, usually equal to distribution of unlabeled source data. For recent advances of UDA see e.g. [17]. Another related problem is out-of-distribution (OOD) detection, where the target unlabeled observations follow a mixture of the source distribution and an unknown distribution of outliers. The goal is to decide whether the given target observation is drawn from the source distribution or from the distribution of outliers, based on a classifier trained on the source data [23, 36].

2 Conditional Probability Shift Model

We consider a K-class classification problem, where each instance is described by a feature vector $(x, z) \in \mathcal{X} \times \mathcal{Z}$ and class variable (label) $y \in \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$ is the output space. Feature vector z is considered separately from the remaining feature vector x since it will determine the shift of the conditional distribution of labels as a conditioning variable. Furthermore, let P and Q denote respectively the source and the target probability distributions on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ and let $p(x, z, y)$ and $q(x, z, y)$ denote the corresponding probability densities or probability mass functions. It is assumed that the source and the target distributions differ. Importantly, for the source dataset we observe x, z, y, whereas for the target dataset, we only observe feature vector (x, z), whereas class y is not observed. The goal is to predict y in the target dataset.

There are various ways to approach distribution shift between a source data distribution P and a target data distribution Q [37]. Without any assumptions concerning the relation between P and Q, the task is clearly infeasible. The existing works focus mostly on covariate shift and label shift. The covariate shift (CS) [12, 4] means that the difference in distributions of (x, z, y) is due to a difference in distribution of covariates: $p(x, z) \neq q(x, z)$, whereas the conditional posterior distribution is preserved, i.e. $p(y|x, z) = q(y|x, z)$. On the other hand, for label shift (LS) [16, 26, 1, 2] it is assumed that the difference in distributions results from the difference in prior distributions: $p(y) \neq q(y)$, but the conditional distribution of the feature vector given class variable is preserved, i.e. $p(x, z|y) = q(x, z|y)$. Recently, in [6, 32], more general assumption has been introduced, called Sparse Joint Shift (SJS). Under SJS, we have that $p(x, z, y) \neq q(x, z, y)$ is due to the fact that the joint distribution of label y and part of the features z is shifted $p(y, z) \neq q(y, z)$, but the conditional distribution of x given y and z is preserved, i.e., $p(x|y, z) = q(x|y, z)$. The considered extension is important from practical point of view: consider e.g. y being a certain disease, x its symptoms and z an age category. Assuming that the data distribution changes in time, condition $p(y, z) \neq q(y, z)$ means that change of occurrence of disease differs in various age categories, but the assumption $p(x|y, z) = q(x|y, z)$ signifies that characteristics of symptoms is unchanged given illness and age. Figure 1 (bottom row) illustrates this situation. The distribution of symptom intensity (variable x) conditional on age (z) and disease occurence (y) remains approximately the same for the source and target data; slight differences result only from the fact that these are empirical distributions, estimated from the data. We stress that even in anti-causal setting it may happen that $p(x|y, z) = q(x|y, z)$ but $p(x|y) \neq q(x|y)$ due to $p(z) \neq q(z)$.

Table 1. Types of shift in joint distributions between source and target domains. CPS stands for the Conditional Probability Shift scheme considered in this paper. MS denotes marginal shift of z.

Distribution shift	**Source domain** $p(x, z, y)$	**Target domain** $q(x, z, y)$
Covariate shift	$\mathbf{p(x, z)}p(y|x, z)$	$\mathbf{q(x, z)}p(y|x, z)$
Label shift	$\mathbf{p(y)}p(x, z|y)$	$\mathbf{q(y)}p(x, z|y)$
Sparse Joint Shift:	$\mathbf{p(z, y)}p(x|z, y)$	$\mathbf{q(z, y)}p(x|z, y)$
-case 1 CPS	$\mathbf{p(y|z)}p(z)p(x|z, y)$	$\mathbf{q(y|z)}p(z)p(x|z, y)$
-case 2 CPS + MS	$\mathbf{p(y|z)p(z)}p(x|z, y)$	$\mathbf{q(y|z)q(z)}p(x|z, y)$
-case 3 MS	$p(y|z)\mathbf{p(z)}p(x|z, y)$	$p(y|z)\mathbf{q(z)}p(x|z, y)$

Observe that under SJS assumption, in view of $p(y, z) = p(y|z)p(z)$, the shift of (y, z) distribution can be associated with two possible shifts. First, it can happen that the conditional distributions do not change, i.e. $p(y|z) = q(y|z)$, and the shift $p(z, y)$ results only from the fact that $p(z) \neq q(z)$. However, provided $p(x|y, z) = q(x|y, z)$, such situation does not cause any new problems for classification, because then $q(y|x, z) = p(y|x, z)$. This follows e.g. from Theorem 1 below. The second case when $p(y|z) \neq q(y|z)$ is much more interesting. So, in this paper we introduce a novel shift assumption, which extracts the most important case from SJS assumption, corresponding to the change in conditional distributions.

Assumption (CPS). *Under Conditional Probability Shift (CPS) assumption, the difference in distributions $p(x, z, y) \neq q(x, z, y)$ is solely due to the fact that the conditional probability $q(y|z)$ of label y given features z is shifted : $p(y|z) \neq q(y|z)$, but $p(z) = q(z)$ and the conditional distribution of the remaining features x given y and z is preserved, i.e.*

$$p(x|y, z) = q(x|y, z). \tag{1}$$

Table 1 summarizes the different types of shift in probability distributions. Bold lettering corresponds to quantities which change for P and Q. We stress that (1) is assumed in both CPS and SJS introduced previously. The difference is that in SJS the shift is associated with the distribution (y, z) while in CPS specifically with the conditional distribution y given z. The CPS assumption excludes the case of marginal interest from the SJS assumption (case 3 in Table 1) and covers more interesting situations (cases 1 and 2 in Table 1). In this paper, in the CPS setting, inference is based on the modeling of conditional distributions $p(y|z)$ and $q(y|z)$. We will show in the following that (1) is a valuable side information that allows classification to be performed despite the occurrence of distributional shift.

Let us also note that the term *sparse*, used in SJS, referred to the fact that the z-dimension is significantly smaller than the x-dimension. Unlike in SJS, in our approach, this assumption is not needed and the z-dimension with respect to the x-dimension can be arbitrary, although in practice it is usually low. Obviously, when z is the null vector, the assumption of SJS coincides with LS.

The following theorem will be crucial to our method. It shows the relationship between the posterior probabilities for the source and target domains. It is stated in [32] using different formalism and assumptions, here we state and prove it in probabilistic setting. It generalizes the result for the LS scheme, provided in [26].

Theorem 1. *Assume that (1) holds. Then we have for $k = 1, \ldots, K$*

$$q(y = k|x, z) = \frac{p(y = k|x, z)\frac{q(y=k|z)}{p(y=k|z)}}{\sum_{l=1}^{K} p(y = l|x, z)\frac{q(y=l|z)}{p(y=l|z)}}.$$

Proof. Let us denote $r(x, z) := \frac{p(x|z)}{q(x|z)}$. Using Bayes' theorem and assumption (1) we get

$$\begin{aligned} q(y = k|x, z) =& \frac{q(x|z, y = k)q(y = k|z)}{q(x|z)} = \frac{p(x|z, y = k)q(y = k|z)}{q(x|z)} \\ =& \frac{p(y = k|x, z)p(x|z)q(y = k|z)}{p(y = k|z)q(x|z)} \\ =& p(y = k|x, z) \cdot r(x, z) \cdot \frac{q(y = k|z)}{p(y = k|z)}. \end{aligned}$$

Then using the fact that $\sum_{k=1}^{K} q(y = k|x, z) = 1$, summation over k of the above expression yields

$$r(x, z) = \left[\sum_{l=1}^{K} p(y = l|x, z)\frac{q(y = l|z)}{p(y = l|z)}\right]^{-1},$$

which ends the proof. □

The theorem justifies why considering the CPS scheme is important within the general sparse joint-shift situation. It indicates that provided (1) holds, in order to recover the class posterior $q(y|x,z)$, we only need to estimate quantities $p(y|x,z)$, $p(y|z)$ and $q(y|z)$, but neither $p(z)$ nor $q(z)$. The class posterior $q(y|x,z)$ is then used to classify observations on the target set based on the standard Bayes rule:

$$\hat{y} = \arg\max_k q(y=k|x,z) \tag{2}$$

or its variants, taking into account imbalance of the classes.

3 Modeling shifted conditional distribution of labels: EM approach

We propose to model the conditional probabilities $p(y=k|z)$ and $q(y=k|z)$ using a multinomial regression models. Estimation of the first probability $p(y=k|z)$ is straightforward using multinomial regression because we observe both z and y for the source data. However, it is not evident how to estimate $q(y=k|z)$, as class indicator y is not observed for target data. We consider a family of soft-max functions indexed by θ:

$$q_\theta(y=k|z) = \frac{\exp(\theta_{k,0} + z^T\theta_k)}{1+\sum_{l=1}^{K-1}\exp(\theta_{l,0}+z^T\theta_l)}, \quad 1 \le k < K, \tag{3}$$

where $\theta = (\theta_{1,0},\ldots,\theta_{K-1,0},\theta_1,\ldots,\theta_{K-1})^T$ is a vector of parameters. Additionally, we have $q_\theta(y=K|z) = (1+\sum_{l=1}^{K-1}\exp(\theta_{l0}+z^T\theta_l))^{-1}$. Assume that the true probability $q(y=k|z) = q_{\theta^*}(y=k|z)$, for some unknown ground-truth parameter θ^*. In the following, we propose an estimation method of θ^*, which in turn will allow us to estimate $q(y=k|x,z)$ using Theorem 1. Define indicator of the k-th class as $y_k = I(y=k)$. Using $q(x,z,y) = q(z)q(y|z)q(x|y,z)$ the log-likelihood function for a single observation in target population can be written as

$$l(\theta;x,y,z) = \log\prod_{k=1}^{K} q_\theta(x,z,y=k)^{y_k} \tag{4}$$

$$= \sum_{k=1}^{K} y_k \log[q(z)] + \sum_{k=1}^{K} y_k \log[q_\theta(y=k|z)]$$

$$+ \sum_{k=1}^{K} y_k \log[q(x|z,y=k)] = \log[q(z)] + \sum_{k=1}^{K-1} y_k[\theta_{k,0}+z^T\theta_k]$$

$$- \log\left[1+\sum_{l=1}^{K-1}\exp(\theta_{l,0}+z^T\theta_l)\right] + \sum_{k=1}^{K} y_k \log q(x|z,y=k).$$

Note that the first and last term in the expression above do not depend on θ due to CPS assumption 1. Since y_k is not observed for the target data, we can treat it as a latent variable and maximize the above function using the Expectation-Maximisation (EM) algorithm (see e.g. [11], Chapter 8). Let us denote by $\mathcal{L}(\theta,\mathcal{D}_t) = \sum_{(x,z,y)\in\mathcal{D}_t} l(\theta;x,y,z)$ the (unobservable) log-likelihood function for all observations in the target dataset $\mathcal{D}_t$.

1. **Expectation step (E step):** determine the expected value of the log likelihood function, with respect to the current conditional distribution of y given x and z corresponding to the current estimate of the parameter $\widehat{\theta}^{(t)}$:

$$Q(\theta|\widehat{\theta}^{(t)}) = \mathbb{E}_{y\sim q_{\widehat{\theta}^{(t)}}(y|x,z)}\mathcal{L}(\theta;\mathcal{D}_t)$$

$$= \sum_{(x,z)\in\mathcal{D}_t}\sum_{k=1}^{K} q_{\widehat{\theta}^{(t)}}(y=k|x,z)\log\left[q_\theta(x,z,y=k)\right],$$

where $q_{\widehat{\theta}^{(t)}}$ is given by the formula

$$q_{\widehat{\theta}^{(t)}}(y=k|x,z) := \frac{p(y=k|x,z)\frac{q_{\widehat{\theta}^{(t)}}(y=k|z)}{p(y=k|z)}}{\sum_{l=1}^{K} p(y=l|x,z)\frac{q_{\widehat{\theta}^{(t)}}(y=l|z)}{p(y=l|z)}}$$

which uses Theorem 1. Here, we assume that $p(y=k|x,z)$ and $p(y=k|z)$ are known. Obviously, in practice, they have to be estimated. In the proposed method, we first estimate $p(y=k|x,z)$ and $p(y=k|z)$ using source data, and then apply the EM procedure using target data.

2. **Maximization step (M-step):** find the parameters that maximize

$$\widehat{\theta}^{(t+1)} = \arg\max_\theta Q(\theta|\widehat{\theta}^{(t)})$$

$$= \arg\max_\theta \sum_{(x,z)\in\mathcal{D}_t}\Big[\sum_{k=1}^{K-1} q_{\widehat{\theta}^{(t)}}(y=k|x,z)[\theta_{k,0}+z^T\theta_k]$$

$$- \log[1+\sum_{l=1}^{K-1}\exp(\theta_{l,0}+z^T\theta_l)]\Big],$$

where the second equality follows from the reasoning in (4). Thus, in step M, instead of maximizing the log-likelihood which is unobservable, we optimize its expected value with respect to y, given the current value of the estimated parameter. The above optimization problem is concave and thus can be solved using standard gradient methods such as SGD or ADAM. The following theorem states the main property of the EM algorithm, namely that at each step of the EM procedure, the value of the marginal log-likelihood function for the observed data (x,z) does not decrease. In order to facilitate reading, in the supplement [33], we give the proof for our setting when y is a latent variable.

Theorem 2. *Let $l_{obs}(\theta;x,z) = \log q_\theta(x,z) = \log\sum_k q(z)q_\theta(y=k|z)q(x|y=k,z)$ and $\mathcal{L}_{obs}(\theta,D_t) = \sum_{(x,z)\in\mathcal{D}_t} l_{obs}(\theta;x,z)$. The following inequality holds*

$$\mathcal{L}_{obs}(\widehat{\theta}^{(t+1)},\mathcal{D}_t) \ge \mathcal{L}_{obs}(\widehat{\theta}^{(t)},\mathcal{D}_t)$$

If optimization in M-step is performed over a a bounded set, it follows from Theorem 2 in [35] that all limit points θ^* of the sequence $(\widehat{\theta}^{(t)})$ are stationary points i.e. $\nabla\mathcal{L}_{\text{obs}}(\theta^*,\mathcal{D}_t) = 0$.

As discussed above, application of the EM procedure is preceded by the estimation of $p(y=1|z)$ and $p(y=1|x,z)$ based on the source data. We assume that $p(y=1|z)$ is modeled using multinomial logistic regression (MLR), which is natural since z is usually a low-dimensional vector containing only a few variables. The question arises what model we should use to estimate $p(y=1|x,z)$? Of course, allowing for incorrect specification, we can use any classification model which yields estimates of posterior probabilities, but the following Theorem shows that, with the additional assumption that the distribution of x given features z and label y is Gaussian, the adequate choice is also MLR. In view of the theorem below, we use MLR as a basic model in our experiments to estimate $p(y=1|x,z)$, although other models such as deep neural networks are also applied. The proof is in the supplement [33].

Theorem 3. *Consider class variable $y \in \{1,\ldots,K\}$, $z \in R^d$ and assume that $p(y=k|z) = \exp(\omega^*_{k0}+z^T\omega^*_k)/[1+\sum_{l=1}^{K-1}\exp(\omega^*_{l0}+z^T\omega^*_l)]$, for $1 \le k \le K-1$, where ω^*_{k0},ω^*_k are unknown ground-truth parameters. Moreover, assume that conditional distribution of x given y,z is $\mathcal{N}(Mz+a_ky_k,I)$, where*

$M \in R^{p\times d}, a_k \in R^p$ and $y_k = I(y = k)$. Then the posterior probability of y, given x, z is also described by softmax function

$$p(y=k|x,z) = \frac{\exp[x^T(a_k - a_K) + z^T(M^T(a_K - a_k) + \omega_k^*) + c]}{1 + \sum_{l=1}^{K-1} \exp[x^T(a_l - a_K) + z^T(M^T(a_K - a_l) + \omega_k^*) + c]}$$

where constant $c = 0.5(a_K^T a_K - a_k^T a_k) + \omega_{k0}^*$.

4 Experiments

In the experiments, we aim to address the following research questions. 1) What is the accuracy of the proposed CPSM method compared to existing methods dedicated to the LS and SJS scenarios? 2) How accurately do we approximate the posterior probability for the target data using the methods considered? 3) How does the magnitude of the shift in the conditional distributions affect the results? 4) What is the performance of the method in a situation where we observe a shift in the conditional distributions $q(y|z) \neq p(y|z)$ but there is no change in the marginal distributions, i.e. $q(y) = p(y)$? 5) How does the number of instances in the data and the class prior probabilities affect the performance of the method?

4.1 Methods and evaluation

We compare the method CPSM proposed here [1] with the following baselines: ExTRA (Exponential Tilting Model) [18], SEES (Shift Estimation and Explanation under SJS) [6], MLLS (Maximum Likelihood Label Shift) [26], BBSC (Black-Box Shift Correction) [16] and also the NAIVE method in which a classification model is trained on the source data and applied to the target data without any corrections. Recall, that MLLS and BBSC are methods developed assuming the LS scenario.

In the case of competing methods, we use implementations made publicly available by the authors. To make the comparison reliable, in all iterative methods (CPSM, MLLS, ExTRA), we set the number of iterations equal to 500. As base classifiers, we use logistic regression (LR) model (for artificial data and MIMIC data) and also a neural network (NN), for MIMIC data. In the case of LR, we use the implementation from the scikit-learn library [22], with default values of the hyper-parameters. The NN consists of 3 layers, each of which has the number of nodes equal to the number of input features. For optimization, we apply the ADAM algorithm, the learning constant is equal to 0.0001, and the batch size is 10. We use the implementation from the PyTorch library [21]. In the ExTRA method, we used $T(x, z) = z$ as the sufficient statistic. We emphasize that all described methods are generic and can be combined with any probabilistic classifier.

In the experiments, we focus on the case of binary classification, for which the classification rule has the form: $q(y = 1|x, z) > t$, where t is a threshold. The choice of $t = 0.5$ corresponds to Bayes' rule (2) and maximization of classification accuracy. Since in the considered data, classes are usually strongly imbalanced, instead of the accuracy score on the target data, we consider the Balanced Accuracy, the maximization of which results in the rule $q(y = 1|x, z) > q(y = 1)$ (see [19], Section 4), and $q(y = 1)$ is estimated by the respective algorithm as an average of $\hat{q}(y = 1|x, z)$ over all samples. In addition we consider the Approximation Error defined as $|\hat{q}_{\text{method}}(y = 1|x, z) - \hat{q}_{\text{oracle}}(y = 1|x, z)|$, where $\hat{q}_{\text{method}}(y = 1|x, z)$ is estimator pertaining to the considered method, whereas $\hat{q}_{\text{oracle}}(y = 1|x, z)$ is estimator which results from the model trained on target data and assuming knowledge of y. The Approximation Errors are averaged over all instances (x, z) in target data. We also analyzed (see Appendix0 the Expected Calibration Error (ECE) [10], which allows us to check how well the estimated probabilities are calibrated.

[1] The source code: https://github.com/teisseyrep/CPSM

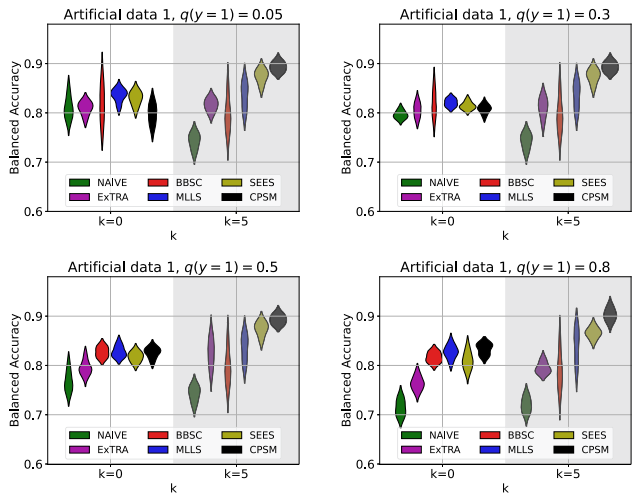

Figure 2. Balanced Accuracy for artificial data 1, for $p(y = 1) = 0.05$ and varying class prior for target data $q(y = 1) = 0.05, 0.3, 0.5, 0.8$.

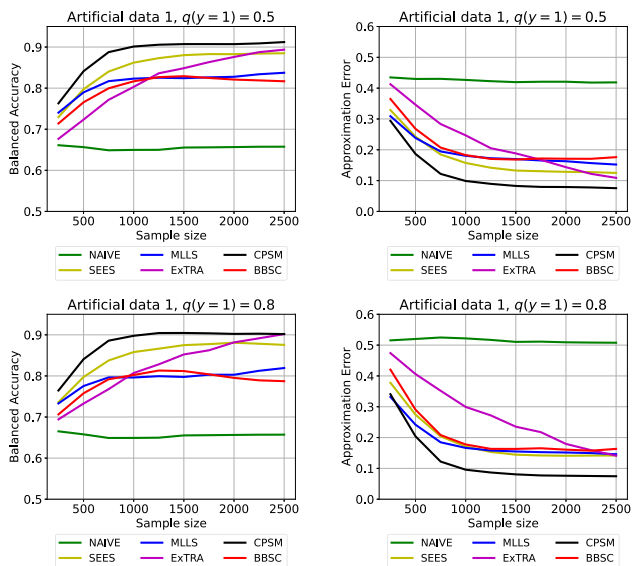

Figure 3. Balanced Accuracy and Approximation Error wrt sample size for artificial data 1, $p(y = 1) = 0.05$ and $k = 5$. The class prior for the target data is $q(y = 1) = 0.5$ (top panels) and $q(y = 1) = 0.8$ (bottom panels).

4.2 Experiments on synthetic datasets

For datasets generated synthetically, we can easily control the values of various parameters such as $p(y)$, $q(y)$ and the relationship between y and z, as well as between x and (y, z). In synthetic dataset 1, we consider a binary class variable $y \in \{0, 1\}$, and binary variables $z_j \in \{0, 1\}$ being independent Bernoulli random variables such that

$p(z_j = 1) = q(z_j = 1) = 0.5$ and let $z = (z_1, z_2, z_3, z_4, z_5)^T$. For synthetic dataset 2, $z_j \sim N(0, 1)$, for $1 \leq j \leq 5$ and are independent. Moreover, in both cases, we assume that

$$p(y = 1|z) = 0.05, \qquad q(y = 1|z) = \sigma(\theta_0 + kz^T\mathbb{1})$$

where $\mathbb{1} = (1, \ldots, 1)^T$ and $\sigma : R \to (0, 1)$ is logistic sigmoid function. For fixed value k, we choose θ_0 to control $q(y = 1)$. In total, we have 2 parameters whose values we will change in the experiment: $q(y = 1)$ and k. Finally, given y and z we generate x from 10-dimensional Gaussian distribution $\mathcal{N}(\mu, I)$, where $\mu = (y, z_1, z_2, z_3, z_4, z_5, 0, \ldots, 0)^T$. This corresponds to the situation described in Theorem 3. Note that, given z and y, vector x is generated in the same way for the source and target datasets, which is consistent with the assumption (1) (a preliminary analysis of robustness of the CPSM method to violations of (1) is provided in the supplement [33]). Importantly, we can distinguish 4 cases of interest:

1. **No shift**: for $p(y = 1) = q(y = 1)$ and $k = 0$ there is no shift in probability distributions.
2. **LS**: for $p(y = 1) \neq q(y = 1)$ and $k = 0$, label shift (LS) occurs, which follows from $p(x, z|y) = p(x|z, y)p(z|y) = q(x|z, y)p(z) = q(x|z, y)q(z) = q(x|z, y)$, as y and z are independent wrt P and Q and $p(z) = q(z)$.
3. **CPS**: for $p(y = 1) = q(y = 1)$ and $k \neq 0$ there is no label shift (LS), but there is shift in conditional distributions $p(y|z) \neq q(y|z)$.
4. **CPS + LS**: for $p(y = 1) \neq q(y = 1)$ and $k \neq 0$ there is label shift (LS), and there is a shift in conditional distributions $p(y|z) = p(y) \neq q(y|z)$ but $p(x|y, z) = q(x|y, z)$.

We show results for dataset 1, and results for dataset 2 are in the supplement [33]. In case 1, when there is no distributional shift, as expected, all methods, including the naive method, perform comparably (Figure 2, top left panel for $k = 0$). In case 2, when the label shift occurs, we observe an increase in the accuracy of the investigated methods in relation to the naive method (see Figure 2, all panels except top left and $k = 0$). This effect becomes more pronounced when $q(y = 1)$ increases, which is associated with an increase in the difference between $p(y)$ and $q(y)$. In cases 3 and 4 ($k \neq 0$), when the distribution of y given z changes between the source and target datasets, we see a clear advantage of the proposed method. Moreover, the proposed method and SEES work better than the BBSC and MLLS methods, because the former are based on the more general SJS assumption, so they are adapted to detect and account for the change in the conditional distribution of y given z. ExTRA is much more variable than CPSM and SEES for $q(y = 1) = 0.3$ and $q(y = 1) = 0.5$. We also note that CPSM performs similarly for the first panel and $k = 5$ (case 3) and the other panels for $k = 5$ (case 4) which shows that it is robust to the change in a marginal distribution of y.

Figure 3 shows how the varying number of instances influences the results for $k = 5$ and $p(y = 1) = 0.05$ (we assume that the target data have the same number of observations as the source data). This corresponds to case 4 listed above. As expected, Balanced Accuracy increases with the sample size, while the Approximation Error decreases. Importantly, CPSM outperforms the other methods for all sample sizes. Method ExTRA requires more data to reach the error level of the other methods. For the naive method, the approximation error remains high, regardless of the sample size due to misspecification effects.

4.3 Case study: MIMIC database

We perform experiments on large clinical database MIMIC [13], containing information on 33166 patients of various intensive care units whose diagnosis is recorded using the coding scheme ICD-9. In the analyses we focused on adult patients only (>16 years). As class variables we consider indicators of 4 families of diseases, which were already used in previous studies: COPD (Chronic Obstructive Pulmonary Disease; ICD-9 codes 490–496), DIABETES (Diabetes mellitus, ICD-9 codes: 249–250), KIDNEY (Kidney disease; ICD-9 codes: 580–589) and FLUID (Fluid electrolyte disease, ICD-9 codes: 276) [5]. The percentage of patients with individual diseases was 7%, 14%, 38% and 37% for COPD, DIABETES, KIDNEY and FLUID, respectively. The considered dataset contains 308 numerical features, corresponding to some laboratory readings, medical scores and administrative information such as age, gender, and marital status. Preprocessing (normalization of variables, removal of missing values) was performed following the steps used by other authors using this dataset; details can be found in [5, 39, 34].

For the MIMIC dataset we introduce conditional probability shift as follows. The conditioning variable z is either age or gender. The patient's age is discretized into two categories '<60 years' ($z = 0$) and '>60 years' ($z = 1$). In case of the gender variable, the value $z = 1$ refers to males, and the value $z = 0$ refers to females. Age and gender are natural choices for a conditioning variable because most of the other variables are clinical symptoms and diagnostic test results. We artificially sample source data in such a way that $p(y = 1|z = 0) = p(y = 1|z = 1) = p(y = 1) = a$, where a is a varying parameter. Moreover, for the target data, in order to allow for shifts, we sample observations in such a way that $q(y = 1|z = 0) = a$, and $q(y = 1|z = 1) = a + k$, where k is an additional parameter that is controlled. By choosing larger k, we increase the dependence between y and z in the target data. The most interesting scenario is when the prevalence of the disease in the source data is low. In this way, the source data can resemble data collected before the increase in disease incidence, e.g. before the outbreak of the pandemic. Therefore, in the experiments, we analyze the results for $a = 0.05, 0.1, 0.2$. Moreover, we show the results for $k = 0.3, 0.5, 0.7$. Note that $k > 0$ corresponds to a shift of conditional distribution of y given z.

In the main part of the paper we show and discuss in detail the results for one chosen target variable (DIABETES) and the situation when the basic model is logistic regression and the conditioning variable is age. The full results for all 4 diseases and other classifiers and conditioning variables can be found in the appendix. Tables 2 and 3 contain the values of Balanced Accuracy and Approximation Error for the considered method, when logistic regression (LOGIT) is used as a base classifier. The proposed CPSM emerges as the winner in terms of Balanced Accuracy and Approximation Error for most settings of parameters a and k, while SEES usually is the second best. The advantage of CPSM over SEES is observed for $a = 0.05$ and $a = 0.1$, while for $a = 2$, SEES usually performs better. As expected, in the case of CPSM and SEES, the classification accuracy increases with increasing k, which is related to the fact that the discrepancy between the conditional distributions $p(y = 1|z)$ and $q(y = 1|z)$ increases in such a case. We also observe a similar effect for other methods, such as BBSC and MLLS. The conclusions are very similar when the conditioning variable is gender (Tables 5 and 6 in the supplement [33]).

Tables 3-4 and 7-8 in the supplement [33] show analogous results when the neural network (NN) is used as the base classifier. The per-

Table 2. Balanced Accuracy score for MIMIC case study, for the **conditioning variable: age**. Predicted disease: DIABETES. The method with the highest Balanced Accuracy is in red and the second best in blue. **Logistic regression** is used as a base classifier.

a	k	NAIVE	ExTRA [18]	BBSC [16]	MLLS [26]	SEES [6]	CPSM
0.05	0.3	0.548 ± 0.012	0.592 ± 0.065	0.625 ± 0.019	0.616 ± 0.017	0.579 ± 0.021	0.637 ± 0.007
0.05	0.5	0.549 ± 0.009	0.627 ± 0.049	0.656 ± 0.015	0.641 ± 0.014	0.668 ± 0.034	0.73 ± 0.025
0.05	0.7	0.548 ± 0.01	0.609 ± 0.062	0.663 ± 0.02	0.661 ± 0.016	0.716 ± 0.037	0.828 ± 0.03
0.1	0.3	0.577 ± 0.01	0.614 ± 0.035	0.639 ± 0.015	0.638 ± 0.011	0.644 ± 0.01	0.669 ± 0.012
0.1	0.5	0.579 ± 0.011	0.631 ± 0.038	0.686 ± 0.015	0.665 ± 0.015	0.772 ± 0.014	0.776 ± 0.013
0.1	0.7	0.577 ± 0.007	0.647 ± 0.081	0.697 ± 0.019	0.688 ± 0.028	0.852 ± 0.01	0.861 ± 0.004
0.2	0.3	0.629 ± 0.01	0.679 ± 0.081	0.663 ± 0.008	0.648 ± 0.01	0.698 ± 0.017	0.708 ± 0.007
0.2	0.5	0.63 ± 0.011	0.684 ± 0.045	0.683 ± 0.012	0.691 ± 0.019	0.811 ± 0.013	0.797 ± 0.013
0.2	0.7	0.631 ± 0.009	0.714 ± 0.065	0.727 ± 0.028	0.707 ± 0.012	0.874 ± 0.006	0.87 ± 0.004

Table 3. Approximation Error for MIMIC case study, for the **conditioning variable: age**. Predicted disease: DIABETES. The method with the smallest approximation error is in red and the second best in blue. **Logistic regression** is used as a base classifier.

a	k	NAIVE	ExTRA [18]	BBSC [16]	MLLS [26]	SEES [6]	CPSM
0.05	0.3	0.219 ± 0.008	0.177 ± 0.061	0.126 ± 0.012	0.135 ± 0.009	0.16 ± 0.014	0.09 ± 0.014
0.05	0.5	0.29 ± 0.011	0.2 ± 0.038	0.192 ± 0.009	0.209 ± 0.009	0.158 ± 0.024	0.101 ± 0.022
0.05	0.7	0.343 ± 0.014	0.28 ± 0.064	0.25 ± 0.005	0.266 ± 0.013	0.178 ± 0.03	0.098 ± 0.031
0.1	0.3	0.174 ± 0.007	0.138 ± 0.03	0.117 ± 0.007	0.12 ± 0.003	0.082 ± 0.01	0.064 ± 0.005
0.1	0.5	0.245 ± 0.009	0.198 ± 0.038	0.179 ± 0.002	0.194 ± 0.007	0.069 ± 0.011	0.068 ± 0.008
0.1	0.7	0.295 ± 0.013	0.252 ± 0.065	0.24 ± 0.011	0.259 ± 0.011	0.072 ± 0.011	0.063 ± 0.018
0.2	0.3	0.124 ± 0.004	0.109 ± 0.061	0.107 ± 0.001	0.109 ± 0.004	0.038 ± 0.012	0.041 ± 0.007
0.2	0.5	0.201 ± 0.007	0.168 ± 0.05	0.181 ± 0.007	0.18 ± 0.005	0.047 ± 0.008	0.044 ± 0.004
0.2	0.7	0.26 ± 0.012	0.236 ± 0.037	0.235 ± 0.008	0.247 ± 0.01	0.036 ± 0.007	0.048 ± 0.011

formance when using NN as a base classifier compared to the logistic model depends on the class variable considered and the parameter setting (see Tables 1 and 3 in the supplement [33]). For example, in the case of COPD, for $a = 0.1$ and $k = 0.7$, the Balanced Accuracy is 0.842 for CPSM + LOGIT and 0.752 for CPSM + NN. On the other hand, in the case of COPD, for $a = 0.2$ and $k = 0.3$, the Balanced Accuracy is 0.65 for CPSM + LOGIT and 0.799 for CPSM + NN. In general, when using NN as a base classifier, SEES usually performs slightly better than CPSM in terms of Balanced Accuracy, but CPSM still has the smallest approximation errors in most cases.

We also analyzed the computation times for individual methods. Table 4 shows the times averaged over 5 runs for the MIMIC dataset. As expected, the proposed CPSM method is significantly slower than the MLLS method, which is also based on the EM scheme. This is due to the fact that in CPSM, the M-step requires the estimation of the multinomial model parameters, while in MLLS, the M-step is straightforward and only inlolves computing prior probability by averaging the posterior probabilities. Importantly, however, MLLS and BBSC are based on the more restrictive LS scheme. CPSM is clearly faster than its main competitor, the SEES method. The computation times depend, of course, on the choice of the base classifier, in the case of using NN they are significantly higher than in the case of the logistic model.

Table 4. Computation times [sec] for the considered methods in the case of MIMIC data and two base classifiers: logistic regression (LR) and neural network (NN). The results are averaged over 5 runs.

Clf	ExTRA	BBSC	MLLS	SEES	CPSM
LR	28.501	0.142	0.123	428.818	91.416
NN	182.979	235.506	158.56	645.358	248.322

5 Conclusions

In this work, we consider the practically important case of distribution shift between the source and target datasets, such that the conditional distribution of the class variable, given selected features, changes. The proposed CPSM method is based on modeling the conditional distributions using multinomial regression. We show that estimation of the multinomial regression parameters for the target data is feasible using the EM algorithm. This leads to an algorithm for estimating the posterior distribution for the target data.

Experiments, conducted on artificial and medical data, show that, in the considered cases, CPSM is superior with respect to the balanced classification accuracy and approximation errors when compared to methods based on the LS and SJS assumptions (BBSC, MLLS and SEES). In particular, CPSM performs consistently better in the case when the source and target conditional distributions differ, while the priors remain unchanged. The advantage of CPSM over competing methods is significant, regardless of the sample size and the prior probability of the class variable in the source data. Finally, CPSM is also computationally faster than the competing method SEES, which is also based on the SJS assumption. CPSM is slower than ExTRA, but it achieves consistently larger accuracy.

Several problems of interest for future research can be identified. It would be valuable to test whether the shift in the conditional distributions is statistically significant, e.g. by testing whether the parameters of the multinomial model for the source and target data differ. Moreover, as in the present work we have shown the advantage of (1) for inference, it is of interest to identify a set of conditioning variables z for which it holds, based on data. It is also desirable to investigate in depth whether the procedures considered are robust to small violations of assumption (1) (see Appendix for partial results).

References

[1] A. Alexandari, A. Kundaje, and A. Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, ICML' 20, pages 1–11, 2020.

[2] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. Regularized learning for domain adaptation under label shifts. In *Proceedings of the International Conference on Learning Representations*, ICLR' 19, pages 1–25, 2019.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

[4] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009.

[5] S. Bromuri, D. Zufferey, J. Hennebert, and M. Schumacher. Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *Journal of Biomedical Informatics*, 51:165–175, 2014.

[6] L. Chen, M. Zaharia, and J. Zou. Estimating and explaining model performance when both covariates and labels shift. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS' 22, pages 1–13, 2024.

[7] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT '08, page 38–53, 2008.

[8] S. B. David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. In *Proceedings of the 13-th International Conference on Artificial Intelligence and Statistics*, AISTATS'20, pages 129–136, 2010.

[9] S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton. A unified view of label shift estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS' 20, pages 1–11, 2020.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17, page 1321–1330, 2017.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2009.

[12] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19-th International Conference on Neural Information Processing Systems*, NIPS' 06, pages 1–8, 2006.

[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, A. C. L., and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9, 2016.

[14] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *ICML' 21*, pages 5637–5664, 2021.

[15] S. M. Kulinski, S. Bagchi, and D. I. Inouye. Feature shift detection: localizing which features have shifted via conditional distribution tests. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS' 20, 2020.

[16] Z. C. Lipton, Y. Wang, and A. J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, ICML' 18, pages 3128–3136, 2018.

[17] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, and J. Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives, 2022. URL https://arxiv.org/abs/2208.07422.

[18] S. Maity, M. Yurochkin, M. Banerjee, and Y. Sun. Understanding new tasks through the lens of training data via exponential tilting. In *Proceedings of the International Conference on Learning Representations*, ICRL' 23, pages 1–29, 2023.

[19] A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning*, ICML' 13, pages 603–611, 2013.

[20] S. Nakajima and M. Siguyama. Positive-unlabeled classification under class-prior shift: a prior-invariant approach based on density ratio estimation. *Machine Learning*, 112:889–919, 2023.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS' 19, 2019.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. ISSN 0165-1684.

[24] S. Rabanser, S. Günnemann, and Z. C. Lipton. Failing loudly: an empirical study of methods for detecting dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS' 19, 2019.

[25] T. Roland, C. Bock, T. Tschoellitsch, A. Maletzky, S. Hochreiter, J. Meier, and G. Klambauer. Domain shifts in machine learning based covid-19 diagnosis from blood tests. *Journal of Medical Systems*, 46 (5):1–12, 2022.

[26] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Comput.*, 14(1):21–41, 2002.

[27] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML' 12, page 459–466, 2012.

[28] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021.

[29] M. Sugiyama, M. Krauledat, and K. R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[30] H. Suryanto, A. Mahidadia, M. Bain, C. Guan, and A. F. A. I. . Guan. Credit risk modeling using transfer learning and domain adaptation. *Frontiers in Artificial Intelligence*, 5:1–16, 2011.

[31] D. Tasche. Invariance assumptions for class distribution estimation. In *Proceedings of the Workshop Learning to Quantify: Methods and Applications*, LQ' 23, 2023.

[32] D. Tasche. Sparse joint shift in multinomial classification. *Foundations of Data Science*, 6(3):251–277, 2024.

[33] P. Teisseyre and J. Mielniczuk. Supplement to the paper: A generalized approach to label shift: the conditional probability shift model, 2025. URL https://doi.org/10.5281/zenodo.16674708.

[34] P. Teisseyre, D. Zufferey, and M. Slomka. Cost-sensitive classifier chains: Selecting low-cost features in multi-label classification. *Pattern Recognition*, 86:290–319, 2019.

[35] J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

[36] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, page 1–28, 2024.

[37] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML' 04, page 114, 2004.

[38] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):1–17, 2018.

[39] D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Computers in Biology and Medicine*, 65:34–43, 2015.