Invariance assumptions for class distribution estimation

Dirk Tasche[0000−0002−2750−2970]

Independent Scholar, dirk.tasche@gmx.net
First version: May 25, 2023
This version: August 27, 2023

Abstract. We study the problem of class distribution estimation under dataset shift. On the training dataset, both features and class labels are observed while on the test dataset only the features can be observed. The task then is the estimation of the distribution of the class labels, i.e. the estimation of the class prior probabilities, in the test dataset. Assumptions of invariance between the training joint distribution of features and labels and the test distribution can considerably facilitate this task. We discuss the assumptions of covariate shift, factorizable joint shift, and sparse joint shift and their implications for class distribution estimation.

Keywords: Class prior estimation · quantification · prevalence estimation · dataset shift · distribution shift · covariate shift · factorizable joint shift · sparse joint shift.

1 Introduction

We consider class distribution estimation against the backdrop of dataset shift (also called distribution shift) between training and test dataset. On the training dataset, both features and class labels are observed while on the test dataset only the features can be observed. In this context, important tasks of interest are the prediction of the labels (classification) and the estimation of the label distribution (class distribution estimation) in the test dataset. In the literature, class distribution estimation is also referred to as class prior estimation, class prevalence estimation, quantification, and with a number of other terms.

Referring to Forman (2005), Esuli et al. (2023, Preface) made the following case for class distribution estimation as a research topic of its own: "In a number of applications involving classification, the final goal is not determining which class (or classes) individual unlabelled instances belong to, but estimating the prevalence (or 'relative frequency', or 'prior probability') of each class in the unlabelled data."

Class distribution estimation for the target (test) dataset when its distribution is allowed to differ from the distribution of the training (source) dataset, in general, is an ill-posed problem, because joint target (test) distributions of features and labels whose marginal feature distributions perfectly match the observed target feature distribution cannot be distinguished. Constraints are

needed on the range of joint target distributions taken into account for the estimation exercise in order to make the problem well-posed. The consideration of causality is a popular approach for specifying such constraints. Typically, this approach leads to making a decision either for prior probability shift (label shift) or for covariate shift as the model for the joint target distribution (Fawcett and Flach, 2005).

Other approaches to the problem include

- Assumptions on the evolution of parts of the joint distribution of labels and features between training and test times (e.g. Zhang et al., 2013; Krempl et al., 2019).
- Implicit assumptions, for instance by the choice of the distance function for measuring the difference of the source and the target feature distributions (e.g. Hofer, 2015; Kirchmeyer et al., 2021).

In this paper, we revisit three approaches to class distribution estimation and, more generally, to modelling dataset shift under invariance assumptions between the joint source and target distributions: Covariate shift (Shimodaira, 2000), factorizable joint shift (FJS, He et al., 2021), and sparse joint shift (SJS, Chen et al., 2022).

The contribution of this paper to the literature is twofold. On the one hand, two new approaches to class distribution estimation under covariate shift are presented. These approaches may prove useful for cross-checking estimates obtained by application of the popular 'probabilistic classify and count' approach. On the other hand, some results on FJS and SJS which were presented in Tasche (2022b) and Tasche (2023) in uncommon notation are revisited in a notation more familiar to the machine learning community.

Class distribution estimation under prior probability shift has been receiving a lot of attention by the research community for at least the last sixty years, beginning with Gart and Buck (1966) if not earlier. For this reason, in this paper we do not dive into any detail of prior probability shift. Regarding this topic, we refer to the recent overviews by González et al. (2017) and Esuli et al. (2023) of the literature on class distribution estimation under prior probability shift and the references therein.

This paper is organised as follows:

- Section 2 'Notation and general assumptions' sets the scene in technical terms for the remainder of the paper.
- Section 3 'Types of dataset shift with invariance assumptions' provides the formal definitions of the four most important types of distribution shift considered in more or less detail in the following: Prior probability shift, covariate shift, factorizable joint shift (FJS), and sparse joint shift (SJS).
- Section 4 'Covariate shift' looks at class distribution estimation under covariate shift, based on previous work by Card and Smith (2018) and Tasche (2022a). Eq. (9b) and Proposition 1 are new results.
- Section 5 'Factorizable joint shift (FJS)' revisits the notion of distribution shift proposed by He et al. (2021). FJS is found to be unsuitable for class

distribution estimation due to lack of identifiability unless additional constraints are applied.
- Section 6 'Sparse joint shift (SJS)' summarises findings of Chen et al. (2022) and Tasche (2023). Proposition 3 on the 'conditional confusion matrix approach' presents a new interpretation of a result of Tasche (2023). SJS is shown to be a generalisation of prior probability shift and found to be a suitable assumption for designing class distribution estimators.
- The paper concludes with a brief assessment of the findings in Section 7.

2 Notation and general assumptions

We adopt notation and assumptions similar to the setting used in Scott (2019):

There are a feature space $\mathcal{X}$ (not necessarily with $\mathcal{X} \subset \mathbb{R}^d$ for any fixed d) and a label space $\mathcal{Y} = \{1, \ldots, \ell\}$ for some integer $\ell \geq 2$. This is the common machine learning setting for multinomial classification and class distribution estimation.

As in Scott (2019, Section 1.2), "... there are two distributions, P and Q, referred to as the *source and target distributions*. We consider the semi-supervised setting where the learner observes $(X_1, Y_1), \ldots, (X_m, Y_m) \sim P$ and $X_{m+1}, \ldots, X_{m+n} \sim Q_X$ ...".

P, Q are probability distributions on $\mathcal{X} \times \mathcal{Y}$. P is also called *training distribution*, Q *test distribution*. X is a generic random variable which shows the features of an object (or instance), Y is a generic random variable showing the class label of an object. Q_X stands for the marginal distribution of the features under the target distribution.

We suppose for the purpose of this paper that the sample sizes m of the training sample and n of the test sample are sufficiently large if not infinite such that P and Q_X can be perfectly inferred and assumed to be known.

Class distribution estimation then may be phrased as the problem of how to find the marginal distribution Q_Y of the labels (i.e. the class distribution) under the target distribution, i.e. the prior probabilities $Q[Y = 1]$, ..., $Q[Y = \ell]$.

Densities. In the following, we assume that the joint target distribution Q of features and labels (X, Y) is absolutely continuous (see Klenke, 2013, Definition 7.30) with respect to the joint source distribution P of (X, Y). We also suppose that $p = p(x, y)$ is a joint density of (X, Y) under P and $q = q(x, y)$ is a joint density of (X, Y) under Q, with respect to some third measure. Absolute continuity of Q with respect to P is implied in particular if the support of Q is a subset of the support of P, i.e. if it holds that

$$q(x, y) > 0 \quad \Rightarrow \quad p(x, y) > 0. \tag{1}$$

For the sake of simplying the notation, for the remainder of the paper we assume that (1) is true.

Under the assumption that (1) holds, define the general *importance weight* function $w(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ by

$$w(x, y) = \begin{cases} \frac{q(x,y)}{p(x,y)}, & \text{for } p(x, y) > 0, \\ 0, & \text{for } p(x, y) = 0. \end{cases} \tag{2a}$$

Function w reflects the change caused by transitioning from source P to target Q. It can also be interpreted as the density of Q with respect to P on $\mathcal{X} \times \mathcal{Y}$.

Besides the full densities p and q also the marginal densities p_X, q_X of the feature variable X are of interest:

$$p_X(x) = \sum_{y=1}^{\ell} p(x, y), \quad q_X(x) = \sum_{y=1}^{\ell} q(x, y).$$

The feature densities p_X, q_X give rise to the *feature importance weight* function $w_X(x)$ for $x \in \mathcal{X}$ which is defined by

$$w_X(x) = \begin{cases} \frac{q_X(x)}{p_X(x)}, & \text{for } p_X(x) > 0, \\ 0, & \text{for } p_X(x) = 0. \end{cases} \tag{2b}$$

Posterior probabilities. We denote the *posterior probability* (conditional probability) of class $y \in \mathcal{Y}$ given the feature variable x under the source distribution P by $P[Y = y \,|\, X = x]$. This is a single number. $P[Y = y \,|\, X]$ stands for the random variable created by sampling x from the feature distribution P_X and evaluating $P[Y = y \,|\, X = x]$ at x.
$Q[Y = y \,|\, X = x]$ and $Q[Y = y \,|\, X]$ respectively denote the corresponding posterior probabilities under the target distribution Q.

Recall also the definition of the *class-conditional feature distributions* $P_{Y=y}$ and $Q_{Y=y}$ under the source distribution P and target distribution Q respectively by

$$\begin{aligned} P_{Y=y}[X \in M] &= P[X \in M \,|\, Y = y] = \frac{P[X \in M, Y = y]}{P[Y = y]}, \\ Q_{Y=y}[X \in M] &= Q[X \in M \,|\, Y = y] = \frac{Q[X \in M, Y = y]}{P[Y = y]}, \end{aligned} \tag{3}$$

for $M \subset \mathcal{X}$.

Further notation. In the following, we denote by $\mathbf{C} = (C_1, \ldots, C_\ell)$ hard *multinomial classifiers* in the sense that

$$\begin{aligned} &C_i \subset \mathcal{X} \text{ for all } i = 1, \ldots, \ell, \\ &C_1, \ldots, C_\ell \text{ is a disjoint decomposition of } \mathcal{X}, \text{ and} \\ &Y = y \text{ is predicted when } X \in C_y \text{ is observed.} \end{aligned} \tag{4}$$

The *indicator function* $\mathbf{1}_S$ of a set S is defined as $\mathbf{1}_S(s) = 1$ for $s \in S$ and $\mathbf{1}_S(s) = 0$ for $s \notin S$.

3 Types of dataset shift with invariance assumptions

This section formally introduces the types of dataset shift to be discussed in the remainder of the paper.

The dataset shift type denoted here by prior probability shift is also called label shift, target shift, global drift, or named in other ways in the literature. Under this type of shift, the class-conditional feature distributions are invariant between source and target distribution. Its definition is given here mainly as a point of reference.

Definition 1 (Prior Probability Shift). *For each $y \in \mathcal{Y}$, the class-conditional feature distributions $P_{Y=y}[X \in M]$ and $Q_{Y=y}[X \in M]$ for measurable $M \subset \mathcal{X}$ as defined by (3) are equal, i.e. it holds that*

$$P_{Y=y}[X \in M] = Q_{Y=y}[X \in M], \quad \text{for } y \in \mathcal{Y},\ M \subset \mathcal{X}.$$

The notion of covariate shift was introduced by Shimodaira (2000). It is based on the possibly most popular invariance assumption for the relationship between source distribution and target distribution: The posterior class probabilities (sometimes called the 'concept') remain unchanged. We quote mutandis mutatis the definition of covariate shift from Kpotufe and Martinet (2021).

Definition 2 (Covariate Shift). *For each $y \in \mathcal{Y}$, there exists a measurable function $\eta_y : \mathcal{X} \to [0,1]$, called* posterior class probability*, such that*

$$P[Y = y \,|\, X = x] = \eta_y(x) = Q[Y = y \,|\, X = x], \tag{5}$$

almost surely for all x under P_X and under Q_X.

Class distribution estimation in the presence of covariate shift is discussed below in Section 4.

Against the backdrop that, under the assumptions of this paper, it is impossible to distinguish prior probability shift and covariate shift solely on the basis of data, the following notion of factorizable joint shift (FJS) as proposed by He et al. (2021) is very appealing at first glance. For it includes both prior probability shift and covariate shift as special cases and, thus, may be interpreted as interpolating between these two poles of dataset shift.

Definition 3 (Factorizable joint shift (FJS)). *There exist non-negative functions u on $\mathcal{X}$ and v on $\mathcal{Y}$ such that for the importance weight function w as defined in (2a), it holds that*

$$w(x, y) = u(x)\, v(y), \tag{6a}$$

almost surely for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ under P.

Observe that the functions u and v of Definition 3 are not uniquely determined because for any $c > 0$ the functions $u_c = c\, u$ and $v_c = v/c$ also satisfy (6a):

$$w(x, y) = u_c(x)\, v_c(y). \tag{6b}$$

No invariance property between the source and target distributions is obvious from Definition 3. Such a property, nonetheless, is implied by Theorem 1 below in Section 5 which is devoted to a discussion of FJS.

Chen et al. (2022) proposed "a new distribution shift model, Sparse Joint Shift (SJS), which considers the joint shift of both labels and a few features. This unifies and generalizes existing shift models including label shift and sparse covariate shift[1], where only marginal feature or label distribution shifts are considered."

Definition 4 (Sparse Joint Shift (SJS)). *Let $T : \mathcal{X} \to \mathcal{T}$ be a measurable transformation of the feature values x. The source distribution P and the target distribution Q are related through T-SJS if it holds for all $y \in \mathcal{Y}$ and $M \subset \mathcal{X}$ that*

$$P_{Y=y}[X \in M \,|\, T(X) = t] = Q_{Y=y}[X \in M \,|\, T(X) = t] \tag{7}$$

for all $t \in \mathcal{T}$ almost surely under $P_{T(X)}$ and $Q_{T(X)}$.

Under SJS, the doubly conditioned (by class and by a transformation of the features) feature distributions are invariant between source distribution and target distribution. Note that $T(X)$ in general creates a 'sparse' or 'thinned out' version of the features. Chen et al. (2022, Section 3.1) called this type of shift 'sparse' because "the sparsity is necessary for the shift to be identifiable".
Choosing T in Definition 4 as $T(x) = c$ for all $x \in \mathcal{X}$, where c is some fixed value, shows that prior probability shift in the sense of Definition 1 is a special case of SJS. In certain limited circumstances, covariate shift implies SJS and vice versa, as is discussed below in Section 6. In general, however, covariate shift is not a special case of SJS.
If P and Q are related through an 'exponential tilt model' as defined in Section 3 of Maity et al. (2023) then P and Q are also related through SJS.

4 Covariate shift

This section gives a brief overview of class distribution estimation under covariate shift. The topic appears to not have received much attention in the literature, with the exceptions of Card and Smith (2018) and Tasche (2022a).

Class prior estimators. If $\mathbf{C} = (C_1, \ldots, C_\ell)$ is a multinomial classifier as defined by (4), *classify & count* (Forman, 2005) might be the most obvious class prior estimator $\widetilde{Q}_n[Y = y]$, $y = 1, \ldots, \ell$, under any type of dataset shift:

$$\widetilde{Q}_n[Y = y] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{C_y}(x_i),$$

where $x_1, \ldots, x_n$ is a test sample of feature values, assumed to have been generated with the target feature distribution Q_X. If $x_1, \ldots, x_n$ is an i.i.d. sample from Q_X, it follows that $\widetilde{Q}_n[Y = y] \to Q[X \in C_y]$ for $n \to \infty$. However, given that Q_X may be any distribution on $\mathcal{X}$, under covariate shift there is no reason

[1] See Definition 6 below for a definition of sparse covariate shift.

why $Q[X \in C_y]$ should equal $Q[Y = y]$ unless $\mathbf{C}$ is a perfect classifier under the target distribution Q – which is an unrealistic assumption.

As noted by Card and Smith (2018), valid estimates $\widehat{Q}_n[Y = y]$ of the target prior probabilities $Q[Y = y]$, $y = 1, \ldots, \ell$, under covariate shift can be obtained by taking recourse to the law of total probability. The law of total probability implies

$$Q[Y = y] = E_Q\big[P[Y = y \,|\, X]\big] = \int_{\mathcal{X}} P[Y = y \,|\, X = x]\, Q_X(dx). \tag{8a}$$

This gives the estimator

$$\widehat{Q}_n[Y = y] = \frac{1}{n} \sum_{i=1}^{n} \widehat{P}[Y = y \,|\, X = x_i], \tag{8b}$$

where $x_1, \ldots, x_n$ is a test sample of feature values, as described above, and $\widehat{P}[Y = y \,|\, X = x]$ denotes an estimate of the posterior probability $P[Y = y \,|\, X = x]$ under the source distribution P, evaluated at the feature value x. Estimator (8b) was called *probabilistic classify and count (PCC)* by Card and Smith (2018) and *probability estimation & average (P&A)* by Bella et al. (2010).

With the feature importance weight function w_X defined by (2b), under covariate shift it holds true that

$$Q[Y = y] = E_P[w_X(X)\, \mathbf{1}_{\{y\}}(Y)], \quad y \in \mathcal{Y}. \tag{9a}$$

Hence, once the importance weight function w_X has been estimated from a sample of features generated under P and another sample of features generated under Q, the class prior probabilities $Q[Y = y]$ can be estimated by means of the estimator

$$\bar{Q}_m[Y = y] = \frac{1}{m} \sum_{i=1}^{m} w_X(x_i)\, \mathbf{1}_{\{y\}}(y_i), \tag{9b}$$

where $(x_1, y_1), \ldots, (x_m, y_m)$ is an i.i.d. sample of (X, Y) under the source distribution P. A variety of methods is available for estimating w_X, see e.g. Sugiyama et al. (2012) or Bickel et al. (2009). Card and Smith (2018) might have deployed estimator (9b), calling it *reweighting* estimator. They did not, however, provide an explicit formula for it. A potential application of (9b) would be to make use of it for cross-checking primary estimates of the target prior probabilities resulting from an application of (8b).

Dimension reduction. X may be a high dimensional random vector such that precisely estimating $x \mapsto P[Y = y \,|\, X = x]$ is difficult, and also the computation of the high-dimensional integral on the right-hand side of (8a) is a hard task. Hence, is it possible to reduce the dimension of X by applying a transformation T such that $T(X)$ has a lower dimension than X but some version of (8a), e.g. like (10a), still holds true:

$$Q[Y = y] \stackrel{?}{=} E_Q\big[P[Y = y \,|\, T(X)]\big] = \int_{\mathcal{T}} P[Y = y \,|\, T(X) = t]\, Q_{T(X)}(dt), \tag{10a}$$

supposing that the transformation T takes its values in $\mathcal{T}$.

Tasche (2022a, Theorem 1) showed that

$$P[Y = y \,|\, T(X)] = Q[Y = y \,|\, T(X)] \tag{10b}$$

is true under covariate shift with the same transformation $T(X)$ for all target distributions Q which are absolutely continuous with respect to the fixed source distribution P if and only if

$$P[Y = y \,|\, T(X) = T(x)] = P[Y = y \,|\, X = x], \tag{11}$$

almost surely for all x under P_X. (11) means that $T(X)$ is *sufficient* for X with respect to $Y = y$ (see Tasche, 2021, Section 3). In general, requesting sufficiency for $T(X)$ excludes simple approaches to dimension reduction for X. Hence, most of the time there is no guarantee that (10b) and consequently also (10a) are applicable.

Although (10b) is not true in general without an assumption of sufficiency, thanks to the generalised Bayes' theorem (Klebaner, 2005, Theorem 10.8) covariate shift can still be shown to imply the following variation of (5) for a fixed target distribution Q:

Proposition 1. *Suppose that Q is absolutely continuous with respect to P and Q and P are related through covariate shift in the sense of Definition 2. Then it follows for any measurable transformation $T : \mathcal{X} \to \mathcal{T}$ and all $y \in \mathcal{Y}$ that*

$$Q[Y = y \,|\, T(X) = t] = \frac{E_P\big[w_X(X)\,\mathbf{1}_{\{y\}}(Y) \,|\, T(X) = t\big]}{E_P[w_X(X) \,|\, T(X) = t]},$$

for all $t \in \mathcal{T}$ almost surely under $P_{T(X)}$, where w_X is defined as in (2b).

As a consequence of Proposition 1, (10b) holds true for fixed Q if and only if

$$E_P\big[w_X(X)\,\mathbf{1}_{\{y\}}(Y) \,|\, T(X)\big] = E_P[w_X(X) \,|\, T(X)]\, P[Y = y \,|\, T(X)], \tag{12}$$

i.e. if $w_X(X)$ and $\{Y = y\}$ are independent conditional on $T(X)$ under P. Such conditional independence, in particular, follows if $T(X)$ is sufficient for X with respect to $\{Y = y\}$. Accordingly, in principle it is possible to check by means of verification of (12) whether or not (10a) can be applied. This involves the estimation of w_X which, at first glance, might not be much easier or even harder than estimating $P[Y = y \,|\, X]$.

See, however, Stojanov et al. (2019, Section 3) for a method to identify a transformation T such that $T(X)$ is approximately sufficient for X with respect to all $\{Y = y\}$, $y \in \mathcal{Y}$. By (12), then (10b) holds for the target distribution Q in question such that (10a) is applicable.

5 Factorizable joint shift (FJS)

He et al. (2021) characterised FJS by claiming that "the biases coming from the data and the label are statistically independent", without specifying any

detail of the claim in technical terms. Tasche (2022b) suggested that FJS might be interpreted as a structural property similar to the 'separation of variables' which plays an important role for finding closed-form solutions to differential equations.

As noted by He et al. (2021), covariate shift is a special case of FJS because of

$$w(x, y) = w_X(x) \tag{13a}$$

for w_X defined by (2b), and prior probability shift is a special case of FJS because of

$$w(x, y) = \frac{Q[Y = y]}{P[Y = y]}. \tag{13b}$$

Characterising FJS. He et al. (2021) also noted that FJS is not fully identifiable in the unsupervised setting of this paper, i.e. if no labels are observed in the target dataset. In the remainder of this section, we summarise the analysis of FJS performed by Tasche (2022b) and clarify the additional assumptions needed to achieve identifiability for FJS.

The following theorem implies, among other things, an invariance property between source distribution P and target distribution Q thanks to FJS (see Eq. (15) below).

Theorem 1. *Suppose that the source distribution P and the target distribution Q are related by FJS in the sense of Definition 3. Denote by w_X the feature importance weight function defined by (2b) and let $q_i = Q[Y = i]$ and $p_i = P[Y = i]$, $i = 1, \ldots, \ell$.*
Then, up to a constant factor c as in (6b), it follows that

$$v(y) = \sum_{i=1}^{\ell-1} \varrho_i \, \frac{q_i}{p_i} \, \mathbf{1}_{\{i\}}(y) + \frac{q_\ell}{p_\ell} \, \mathbf{1}_{\{\ell\}}(y) \quad \textit{and} \tag{14a}$$

$$u(x) = \frac{w_X(x)}{\sum_{i=1}^{\ell-1} \varrho_i \, \frac{q_i}{p_i} \, P[Y = i \,|\, X = x] + \frac{q_\ell}{p_\ell} \, P[Y = \ell \,|\, X = x]}, \tag{14b}$$

where the constants $\varrho_1, \ldots, \varrho_{\ell-1}$ are positive and finite and satisfy the following equation system (with $j = 1, \ldots, \ell - 1$):

$$p_j = \varrho_j \, E_P \left[\frac{w_X(X) \, P[Y = j \,|\, X]}{\sum_{i=1}^{\ell-1} \varrho_i \, \frac{q_i}{p_i} \, P[Y = i \,|\, X] + \frac{q_\ell}{p_\ell} \, P[Y = \ell \,|\, X]} \right]. \tag{14c}$$

Conversely, suppose that for the source distribution P a function $w_X : \mathcal{X} \to [0, \infty)$ with $E_P\big[w_X(X)\big] = 1$ and $(q_i)_{i=1,\ldots,\ell} \in (0, 1)^\ell$ with $\sum_{i=1}^{\ell} q_i = 1$ are given. Assume also that $\varrho_1 > 0$, …, $\varrho_{\ell-1} > 0$ are solutions of the equation system (14c) and u and v are defined by (14b) and (14a), respectively. Then $w(x, y) = u(x)\, v(y)$ has the property that $w(x, y)\, p(x, y)$ is the density of a probability measure Q on $\mathcal{X} \times \mathcal{Y}$ such that $w_X(x)\, p_X(x)$ is the marginal density of the feature variable X under Q and $Q[Y = i] = q_i$ holds for $i = 1, \ldots, \ell$.

See Tasche (2022b, Theorem 2) for a proof of Theorem 1. The theorem characterises FJS through equations (14b), (14a) and (14c) but does not provide any information regarding the existence or uniqueness of solutions to (14c). A result on existence and uniqueness of the solutions to (14c) was proven for the binary case $\ell = 2$ by Tasche (2022b, Proposition 2).

It can be shown (Tasche, 2022b, Corollary 4) that Theorem 1 implies the following version of the correction formula for class posterior probabilities of Saerens et al. (2001, Eq. (2.4)) and Elkan (2001, Theorem 2) under FJS.

Corollary 1. *Suppose that the source distribution P and the target distribution Q are related through FJS in the sense of Definition 3. Then the target posterior probabilities $Q[Y = j \,|\, X = x]$, $j = 1, \ldots, \ell$, can be represented almost surely for all x under Q_X as functions of the source posterior probabilities $P[Y = j \,|\, X = x]$, $j = 1, \ldots, \ell$, in the following way:*

$$Q[Y = j \,|\, X = x] = \frac{\varrho_j \, \frac{Q[Y=j]}{P[Y=j]} P[Y = j \,|\, X = x]}{\sum_{i=1}^{\ell-1} \varrho_i \, \frac{Q[Y=i]}{P[Y=i]} P[Y = i \,|\, X = x] + \frac{Q[Y=\ell]}{P[Y=\ell]} P[Y = \ell \,|\, X = x]}, \quad j = 1, \ldots, \ell - 1,$$

$$Q[Y = \ell \,|\, X = x] = \frac{\frac{Q[Y=\ell]}{P[Y=\ell]} P[Y = \ell \,|\, X = x]}{\sum_{i=1}^{\ell-1} \varrho_i \, \frac{Q[Y=i]}{P[Y=i]} P[Y = i \,|\, X = x] + \frac{Q[Y=\ell]}{P[Y=\ell]} P[Y = \ell \,|\, X = x]},$$

where the positive constants $\varrho_1, \ldots, \varrho_{\ell-1}$ satisfy the equation system (14c).

Corollary 1 in turn implies that under FJS the following invariance property holds true:

$$\frac{Q[Y = j \,|\, X]}{Q[Y = \ell \,|\, X]} \frac{Q[Y = \ell]}{Q[Y = j]} = \varrho_j \, \frac{P[Y = j \,|\, X]}{P[Y = \ell \,|\, X]} \frac{P[Y = \ell]}{P[Y = j]}, \quad j = 1, \ldots, \ell - 1, \tag{15}$$

where the constants ϱ_j satisfy the equation system (14c). Eq. (15) may be interpreted as stating that under factorizable joint shift the ratios of the class-conditional feature densities are invariant between source and target distributions up to a constant factor (see Tasche, 2022b, Remark 1).

Class distribution estimation under FJS. Theorem 1 suggests two obvious ways to learn the characteristics of factorizable joint shift:

a) If the target prior class probabilities $Q[Y = i] = q_i$ are known (for instance from external sources) solve (14c) for the constants ϱ_i.
b) If the target prior class probabilities $Q[Y = i] = q_i$ are unknown (as would be the case for the problem of class distribution estimation), fix values for the constants ϱ_i and solve (14c) for the q_i. Letting $\varrho_i = 1$ for all i is a natural choice that converts (14c) into the system of maximum likelihood equations for the q_i under the prior probability shift assumption.

See Section 4.2.4 of Tasche (2013) for an example of approach a) from the area of credit risk. Whenever for a given marginal target feature distribution Q_X there

is more than one set of potential target class prior probabilities q_y, $y = 1, \ldots, \ell$, such that (14c) can be solved for the ϱ_i, then a case of unidentifiability of the joint target distribution Q under FJS is incurred. This always holds for the binary case $\ell = 2$ because for any given combination of joint source distribution P, target feature distribution Q_X and target prior probability $q_1 = Q[Y = 1]$, a constant ϱ_1 can be found such that P and Q are related through FJS (Tasche, 2022b, Proposition 2).

Regarding the interpretation of (14c) in approach b) as maximum likelihood equations, see Du Plessis and Sugiyama (2014). This interpretation, in particular, implies that an EM (expectation maximisation) algorithm can be deployed for solving the equation system (Saerens et al., 2001) in the case $1 = \varrho_1 = \ldots = \varrho_{\ell-1}$.

6 Sparse joint shift (SJS)

Definition 4 of SJS slightly generalises Definition 1 of Chen et al. (2022) as can be seen by choosing T as extractor of a subset of the components of the feature vector. The equivalence of this special case of Definition 4 and the definition of Chen et al. (2022) then follows from Proposition 3.8 of Tasche (2023).

Observe that by the generalised Bayes' theorem (Klebaner, 2005, Theorem 10.8), (7) can equivalently be stated as

$$\frac{P[X \in M, Y = y \,|\, T(X) = t]}{P[Y = y \,|\, T(X) = t]} = \frac{Q[X \in M, Y = y \,|\, T(X) = t]}{Q[Y = y \,|\, T(X) = t]}. \tag{16}$$

The following properties of SJS were first noted by Tasche (2023).

Proposition 2 (Properties of SJS). *Suppose that the source distribution P and the target distribution Q are related through T-SJS in the sense of Definition 4. Then the following two statements hold true:*

*(i) If $T' : \mathcal{X} \to \mathcal{T}'$ and $S : \mathcal{T}' \to \mathcal{T}$ are measurable transformations such that for all $x \in \mathcal{X}$ it holds that $T(x) = (S \circ T')(x) = S\big(T'(x)\big)$, then P and Q are also related through T'-SJS.**

(ii) For all $i \in \mathcal{Y}$, it holds that

$$Q[Y = i \,|\, X = x] = \frac{\frac{Q[Y=i \,|\, T(X)=T(x)]}{P[Y=i \,|\, T(X)=T(x)]}\, P[Y = i \,|\, X = x]}{\sum_{j=1}^{\ell} \frac{Q[Y=j \,|\, T(X)=T(x)]}{P[Y=j \,|\, T(X)=T(x)]}\, P[Y = j \,|\, X = x]},$$

for all $x \in \mathcal{X}$ almost surely under Q_X.

See Tasche (2023, Corollary 4.3) for a proof of Proposition 2 (i) and Tasche (2023, Proposition 4.5) for a proof of Proposition 2 (ii). By Proposition 2 (i), prior probability shift implies T-SJS for any transformation $T : \mathcal{X} \to \mathcal{T}$. Proposition 2 (ii) is another generalisation of the posterior correction formula of Saerens et al. (2001, Eq. (2.4)) and Elkan (2001, Theorem 2), this time under the assumption of SJS.

The next result rephrases the identifiability result of (Chen et al., 2022, Theorem 1) in terms of conditional expectations instead of joint densities.

Theorem 2 (Identifiability under SJS). *Suppose that there are distributions P, Q and Q' on $\mathcal{X} \times \mathcal{Y}$ as well as transformations $T : \mathcal{X} \to \mathcal{T}$ and $T' : \mathcal{X} \to \mathcal{T}'$ such that P and Q are related through T-SJS and P and Q' are related through T'-SJS. For given measurable functions $f_i : \mathcal{X} \to [0, \infty)$, $i = 1, \ldots, \ell$, define the random matrix $R(X) = \big(R_{ij}(X)\big)_{i,j\in\{1,\ldots,\ell\}}$ by*

$$R_{ij}(X) = \frac{E_P\big[f_i(X)\,\mathbf{1}_{\{j\}}(Y)\,|\,(T(X), T'(X))\big]}{P\big[Y = j\,|\,(T(X), T'(X))\big]}.$$

If $Q_X = Q'_X$ and $P\big[\operatorname{rank}\big(R(X)\big) = \ell\big] = 1$ is true, then it follows that $Q[Y = y,\, X \in M] = Q'[Y = y,\, X \in M]$ for all $y \in \mathcal{Y}$ and measurable $M \subset \mathcal{X}$.

See Tasche (2023, Theorem 4.7) for a proof of Theorem 2. The rank condition of Theorem 2 is likely to be satisfied for instance if $f_i(X) = \mathbf{1}_{C_i}(X)$ for some reasonably accurate classifier $\mathbf{C} = (C_1, \ldots, C_\ell)$ as in (4). Hence identifiability of SJS ought to be given most of the time.

SJS and covariate shift. As seen above, prior probability shift is not only a special case of SJS but also implies T-SJS for any transformation T of the features. In contrast, examples by Chen et al. (2022) and Tasche (2023) show that covariate shift and SJS are unrelated properties in the sense that they do not imply one another but do not exclude each other either.

For a full understanding of the relationship of covariate shift and SJS, we introduce two further types of dataset shift. The first of these was proposed by Tasche (2023, Definition 4.11).

Definition 5 (Conditional distribution invariance (CDI)). *Let $T : \mathcal{X} \to \mathcal{T}$ be a measurable transformation of the feature variable X. The source distribution P and the target distribution Q are related through T-*CDI *if it holds for all $M \subset \mathcal{X}$ that*

$$P[X \in M\,|\,T(X) = t] = Q[X \in M\,|\,T(X) = t] \tag{17}$$

for all $t \in \mathcal{T}$ almost surely under $P_{T(X)}$ and $Q_{T(X)}$.

The property CDI is interesting because in principle its presence can be evidenced by comparing statistics estimated from the feature observations in the training and test datasets. No label observations are needed. Moreover, in the presence of CDI, there is basically no difference between covariate shift and SJS, as we will see below.

The following additional type of dataset shift was introduced by Chen et al. (2022, Definition 3).

Definition 6 (Sparse Covariate Shift (SCS)). *Let $T : \mathcal{X} \to \mathcal{T}$ be a measurable transformation of the feature variable X. The source distribution P and the target distribution Q are related through T-*SCS *if it holds for all $y \in \mathcal{Y}$ and $M \subset \mathcal{X}$ that*

$$P[X \in M, Y = y\,|\,T(X) = t] = Q[X \in M, Y = y\,|\,T(X) = t] \tag{18}$$

for all $t \in \mathcal{T}$ almost surely under $P_{T(X)}$ and $Q_{T(X)}$.

The following theorem describes the interplay of SJS and covariate shift in the presence of CDI.

Theorem 3. *Let $T : \mathcal{X} \to \mathcal{T}$ be a measurable transformation of the feature variable X. Suppose that a source distribution P and a target distribution Q on $\mathcal{X} \times \mathcal{Y}$ are given. Then the following three statements hold true:*

(i) If P and Q are related through both T-CDI in the sense of Definition 5 and covariate shift in the sense of Definition 2, then P and Q are also related through T-SCS in the sense of Definition 6.

(ii) If P and Q are related through T-SCS, they are also related through both T-SJS and T-CDI.

(iii) For given measurable functions $f_i : \mathcal{X} \to [0, \infty)$, $i = 1, \ldots, \ell$, define the random matrix $R(X) = \big(R_{ij}(X)\big)_{i,j \in \{1,\ldots,\ell\}}$ by

$$R_{ij}(X) = \frac{E_p\big[f_i(X)\, \mathbf{1}_{\{j\}}(Y) \,|\, T(X)\big]}{P\big[Y = j \,|\, T(X)\big]}.$$

Suppose that $P\big[\operatorname{rank}\big(R(X)\big) = \ell\big] = 1$ holds true. Then, if P and Q are related through both T-SJS and T-CDI, they are also related through covariate shift.

For the derivation of Theorem 3, see Theorem 4.16 and Remark 4.18 of Tasche (2023). Somewhat oversimplifying, we might summarise Theorem 3 with the following 'equation': $SCS = \textit{covariate shift} \cap CDI = SJS \cap CDI$.

Class distribution estimation under SJS. Chen et al. (2022) proposed two methods for estimating SJS: SEES-c for the case of continuous features and SEES-d for the case of discrete features (SEES = "shift estimation and explanation under SJS"). In this paper, we briefly describe only an important special case of SEES-d (Tasche, 2023, Eq. (C.6)) because the results presented by Chen et al. (2022) appear to suggest that SEES-d is more efficient than SEES-c. By sufficiently fine discretisation of the feature space, SEES-d can also be applied to continuous or mixed continuous and discrete feature settings.

Proposition 3 (Conditional confusion matrix approach). *Let $T : \mathcal{X} \to \mathcal{T}$ be a measurable and discrete transformation of the feature variable X, i.e. with range $\mathcal{T} = \{t_1, \ldots, t_N\}$. Suppose that the source distribution P and a target distribution Q are related through T-SJS in the sense of Definition 4 and that $\mathbf{C} = (C_1, \ldots, C_\ell)$ is a classifier as in (4). Then for each $t \in \mathcal{T}$, the target posterior probabilities $q_{y,t} = Q[Y = y \,|\, T(X) = t]$, $y \in \mathcal{Y}$, satisfy the linear equation system (with $j = 1, \ldots, \ell$)*

$$\sum_{y=1}^{\ell} q_{y,t}\, P[X \in C_j \,|\, Y = y, T(X) = t] = Q[X \in C_j \,|\, T(X) = t]. \tag{19a}$$

Once the $q_{y,t}$, $y \in \mathcal{Y}, t \in \mathcal{T}$, have been determined, by the law of total probability the target class prior probabilities $Q[Y = y]$ can be calculated via

$$Q[Y = y] = \sum_{i=1}^{N} q_{y,t_i} \, Q[T(X) = t_i]. \tag{19b}$$

Therefore, Proposition 3 provides a solution to the class distribution estimation problem under an assumption of SJS, thereby generalising the confusion matrix approach as described by Saerens et al. (2001, Section 2.3.1). In particular, Proposition 3 could be deployed to check assumptions of prior probability shift. By Proposition 2 (i), prior probability shift implies T-SJS for any transformation T. Hence, in principle, results under prior probability shift by any suitable method of class distribution estimation must coincide with the results obtained by combining (19a) and (19b), for any choice of T taking discrete values.

In practice, develop the classifier on the full training dataset. Then stratify both training dataset and test dataset by T applied to the feature (or covariate) variable X. After that, treat each of the resulting sub-samples with the confusion matrix approach as in Saerens et al. (2001, Section 2.3.1) to estimate for each $t \in \mathcal{T}$ the posterior probabilities $Q[Y = y \,|\, T(X) = t] = q_{y,t}$, $y \in \mathcal{Y}$. Combine the posterior probabilities by means of (19b) to obtain estimates of the target prior probabilities $Q[Y = y]$, $y \in \mathcal{Y}$.

Examples for possible choices of the transformation T of Proposition 3 might be found in medical applications: It is plausible that the sensitivity and specificity of a test for an infection change between training and test datasets but that they are preserved within the strata when there is stratification by age group and gender. This would mean that the dataset shift can be described by T-sparse joint shift with T being the transformation that provides the age group and the gender of an instance (patient).

7 Conclusions

This paper provides analyses of invariance assumptions for distribution (dataset) shift, with focus on their suitability for designing class distribution estimators. Covariate shift, factorizable joint shift, and sparse joint shift are studied in some detail. Both the 'covariate' and the 'sparse joint' types of shift are found fit for designing class distribution estimators. In contrast, factorizable joint shift is found unsuitable due to lack of identifiability unless additional constraints are applied.

Sparse joint shift (SJS) is particularly appealing for the fact that it generalises prior probability shift (label shift) and, therefore, has the potential to provide meaningful estimates even in contexts where an assumption of prior probability shift is found untenable. An open research problem is how to identify feature transformations that entail SJS if they cannot be identified by theoretical considerations. Chen et al. (2022, Section 4.1) suggested two brute-force approaches but these approaches have issues which might make their application questionable (Tasche, 2023, Section 5).

Acknowledgement The author would like to thank three anonymous reviewers for their useful comments and suggestions.

References

A. Bella, C. Ferri, J. Hernandez-Orallo, and M.J. Ramírez-Quintana. Quantification via probability estimators. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 737–742. IEEE, 2010.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative Learning Under Covariate Shift. *The Journal of Machine Learning Research*, 10:2137–2155, 2009.

D. Card and N.A. Smith. The Importance of Calibration for Estimating Proportions from Annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1636–1646, 2018. https://doi.org/10.18653/v1/N18-1148.

L. Chen, M. Zaharia, and J.Y. Zou. Estimating and Explaining Model Performance When Both Covariates and Labels Shift. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems, NeurIPS 2022*, volume 35, pages 11467–11479. Curran Associates, Inc., 2022.

M.C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

C. Elkan. The foundations of cost-sensitive learning. In B. Nebel, editor, *Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001*, pages 973–978. Morgan Kaufmann, 2001.

A. Esuli, A. Fabris, A. Moreo, and F. Sebastiani. *Learning to Quantify*. Springer Cham, 2023. https://doi.org/https://doi.org/10.1007/978-3-031-20467-8.

T. Fawcett and P.A. Flach. A response to Webb and Ting's On the Application of ROC Analysis to Predict Classification Performance under Varying Class Distributions. *Machine Learning*, 58(1):33–38, 2005.

G. Forman. Counting Positives Accurately Despite Inaccurate Classification. In *European Conference on Machine Learning (ECML 2005)*, pages 564–575. Springer, 2005.

J.J. Gart and A.A. Buck. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602, 1966.

P. González, A. Castaño, N.V. Chawla, and J.J. Del Coz. A Review on Quantification Learning. *ACM Comput. Surv.*, 50(5):74:1–74:40, 2017.

H. He, Y. Yang, and H. Wang. Domain Adaptation with Factorizable Joint Shift. Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning, 2021.

V. Hofer. Adapting a classification rule to local and global shift when only unlabelled data are available. *European Journal of Operational Research*, 243(1):177–189, 2015.

M. Kirchmeyer, A. Rakotomamonjy, E. de Bezenac, and P. Gallinari. Mapping conditional distributions for domain adaptation under generalized target shift, 2021. URL https://arxiv.org/abs/2110.15057. Presented at ICLR 2022.

F.C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, second edition, 2005.

A. Klenke. *Probability Theory: A Comprehensive Course*. Springer Science & Business Media, 2013.

S. Kpotufe and G. Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021. https://doi.org/10.1214/21-AOS2084.

G. Krempl, D. Lang, and V. Hofer. Temporal density extrapolation using a dynamic basis approach. *Data mining and knowledge discovery*, 33:1323–1356, 2019.

S. Maity, M. Yurochkin, M. Banerjee, and Y. Sun. Understanding new tasks through the lens of training data via exponential tilting. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://openreview.net/forum?id=DBMttEEoLbw.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, 2001.

C. Scott. A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation. In *Proceedings of Machine Learning Research, 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 1–24, 2019.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

P. Stojanov, M. Gong, J. Carbonell, and K. Zhang. Low-Dimensional Density Ratio Estimation for Covariate Shift Correction. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3449–3458. PMLR, 2019.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

D. Tasche. The art of probability-of-default curve calibration. *Journal of Credit Risk*, 9(4):63–103, 2013. https://doi.org/10.21314/JCR.2013.169.

D. Tasche. Calibrating sufficiently. *Statistics*, 55(6):1356–1386, 2021. https://doi.org/10.1080/02331888.2021.2016767.

D. Tasche. Class Prior Estimation under Covariate Shift: No Problem? Working paper, presented at ECML/PKDD 2022 workshop Learning to Quantify: Methods and Applications (LQ 2022), 2022a.

D. Tasche. Factorizable Joint Shift in Multinomial Classification. *Machine Learning and Knowledge Extraction*, 4(3):779–802, 2022b. https://doi.org/10.3390/make4030038.

D. Tasche. Sparse joint shift in multinomial classification. *arXiv preprint arXiv:2303.16971*, 2023.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain Adaptation Under Target and Conditional Shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning – Volume 28*, ICML'13, pages III–819–III–827. JMLR.org, 2013.