machine learning & knowledge extraction

MDPI

Article

Factorizable Joint Shift in Multinomial Classification

Dirk Tasche

Independent Researcher, 8032 Zurich, Switzerland; dirk.tasche@gmx.net

Abstract: Factorizable joint shift (FJS) was recently proposed as a type of dataset shift for which the complete characteristics can be estimated from feature data observations on the test dataset by a method called Joint Importance Aligning. For the multinomial (multiclass) classification setting, we derive a representation of factorizable joint shift in terms of the source (training) distribution, the target (test) prior class probabilities and the target marginal distribution of the features. On the basis of this result, we propose alternatives to joint importance aligning and, at the same time, point out that factorizable joint shift is not fully identifiable if no class label information on the test dataset is available and no additional assumptions are made. Other results of the paper include correction formulae for the posterior class probabilities both under general dataset shift and factorizable joint shift. In addition, we investigate the consequences of assuming factorizable joint shift for the bias caused by sample selection.

Keywords: dataset shift; factorizable joint shift; multinomial classification; covariate shift; prior probability shift; sample selection bias

check for updates

Citation: Tasche, D. Factorizable Joint Shift in Multinomial Classification. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 779–802. https://doi.org/10.3390/make4030038

Academic Editor: Gianluigi Greco

Received: 6 August 2022
Accepted: 7 September 2022
Published: 10 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In machine learning terminology, dataset shift refers to the phenomenon that the joint distribution of features and labels on the training dataset used for learning a model may differ from the related joint distribution on the test dataset to which the model is going to be applied; see Storkey [1] or Moreno-Torres et al. [2] for surveys and background information on dataset shift. Dataset shift can be the consequence of very different causes. For that reason, a catch-all treatment of general dataset shift is difficult if not impossible. As a workaround a number of specific types of dataset shift have been defined in order to introduce additional assumptions that allow for different tailor-made approaches to deal with the problem. The most familiar subtypes of dataset shift are prior probability shift and covariate shift, but more types are introduced on a continuing basis as there is a practice-driven need to do so.

Typically, under dataset shift, the test dataset observations of features are available, but the class labels cannot be observed. In this situation, it is impossible to know ex ante if covariate shift or prior probability shift (or something in between) has occurred. However, estimates of models under assumptions of covariate shift and prior probability shift, respectively, tend to differ conspicuously. As a consequence, additional assumptions need to be made in order to be able to choose between modelling options related to covariate shift and prior probability shift. Such additional assumptions may be phrased in terms of causality (Storkey [1]): if the features can be considered "causing" the class labels, then models designed to deal with covariate shift are appropriate. Otherwise, if the class "causes" features, models targeting prior probability shift should be preferred.

He et al. [3] recently proposed "factorizable joint shift" (FJS) which generalises both prior probability shift and covariate shift. They went on with presenting the "joint importance aligning" method for estimating the characteristics of this type of shift. At first glance, He et al. hence seemed to provide a way to avoid choosing ex ante between covariate shift and prior probability shift models. Instead, "joint importance aligning" (plus some regularisation) appeared to be a method that functioned as a covariate shift model, prior

probability shift model, or combined covariate and label shift model, as required by the characteristics of the test dataset.

By a detailed analysis of factorizable joint shift in multinomial classification settings, in this paper we point out that general factorizable joint shift is not fully identifiable if no class label information on the test dataset is available and no additional assumptions are made. This is in contrast to the situations with covariate shift or prior probability shift. Therefore, circumspection is recommended with regard to potential deployment of "joint importance aligning" as proposed by He et al. [3].

He et al. characterised factorizable joint shift by claiming that "the biases coming from the data and the label are statistically independent". This description might not fully hit the mark. As we demonstrate in this paper, factorizable joint shift has little to do with statistical independence but should rather 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. We also argue that, in probabilistic terms, factorizable joint shift perhaps is better described as "scaled density ratios" shift.

The plan of this paper and its main research contributions are as follows:

- Section 2 "Setting the scene" presents the assumptions, concepts and notation for the multinomial (or multiclass) classification setting of this paper.
- Section 3 "General dataset shift in multinomial classification" introduces a normal form for the joint density of features and class labels (Theorem 1) and derives in Corollary 2 a generalisation of the correction formula for class posterior probabilities of Saerens et al. [4] and Elkan [5].
- Section 4 "Factorizable joint shift" defines this kind of dataset shift in a mathematically rigorous manner and presents a full representation in terms of the source (training) distribution, the target (test) prior class probabilities and the target marginal distribution of the features (Theorem 2). In addition, a specific version of the posterior correction formula is given (Corollary 4), and the description of factorizable joint shift as "scaled density ratios" shift is motivated. Moreover, alternatives to the "joint importance aligning" of He et al. [3] are proposed (Section 4.1).
- Section 5 "Common types of dataset shift" examines in a mathematically rigorous manner for a number of types of dataset shift mentioned in the literature if they are implied by or imply factorizable joint shift. The types of dataset shift treated in this section are prior probability shift, covariate shift, covariate shift with posterior drift, domain invariance and generalised label shift. In addition, the posterior correction formulae specific for these types of dataset shift are presented.
- Section 6 "Sample selection bias" revisits the topic of dataset shift caused by sample selection bias and looks at the question of how the class-wise selection probabilities look like if the induced dataset shift is factorizable joint shift (Theorem 3).
- Section 7 "Conclusions" provides a short discussion of the important findings of the paper and points to some open research questions.

2. Setting the Scene

In this paper, we use the following population-level description of the multinomial classification problem under dataset shift in terms of measure theory. See standard textbooks on probability theory like Billingsley [6] or Klenke [7] for formal definitions and background of the notions introduced in Assumption 1. See Tasche [8] for a detailed reconciliation of the setting of this paper with the concepts and notation used in the mainstream machine learning literature.

Assumption 1. *$(\Omega, \mathcal{F})$ is a measurable space. The* source distribution *P and the* target distribution *Q are probability measures on $(\Omega, \mathcal{F})$. For some positive integer $d \geq 2$, events $A_1, \ldots, A_d \in \mathcal{F}$ and a sub-σ-algebra $\mathcal{H} \subset \mathcal{F}$ are given. The events A_i, $i = 1, \ldots, d$, and $\mathcal{H}$ have the following properties:*

(i) $\bigcup_{i=1}^{d} A_i = \Omega$.

(ii) $A_i \cap A_j = \emptyset, i, j = 1, \dots d, i \neq j.$
(iii) $0 < P[A_i], i = 1, \dots, d.$
(iv) $0 < Q[A_i], i = 1, \dots, d.$
(v) $A_i \notin \mathcal{H}, i = 1, \dots, d.$

In the literature, P is also called "source domain" or "training distribution" while Q is also referred to as "target domain" or "test distribution'.

The elements ω of Ω are objects (or instances) with class (label) and covariate (or feature) attributes. $\omega \in A_i$ means that ω belongs to class i (or the positive class in the binary case if $i = 1$).

The σ-algebra $\mathcal{F}$ of events $F \in \mathcal{F}$ is a collection of subsets F of Ω with the property that they can be assigned probabilities $P[F]$ and $Q[F]$ in a logically consistent way. In the literature, thanks to their role of reflecting the available information, σ-algebras are sometimes also called "information set" (Holzmann and Eulert [9]). In the following, we use both terms exchangeably.

The sub-σ-algebra $\mathcal{H} \subset \mathcal{F}$ generated by the covariates (features) contains the events which are observable at the time when the class of an object ω has to be predicted. Since $A_i \notin \mathcal{H}, i = 1, \dots, d$, then the class of the object may not yet be known. In this paper, we assume that under the source distribution P, the class events A_i can be observed such that the prior class probabilities can be estimated. In contrast, under the target distribution Q, the events A_i cannot be directly observed and can only be predicted on the basis of the events $H \in \mathcal{H}$, which are assumed to reflect the features of the object.

For technical reasons, it is convenient to define the joint information set $\overline{\mathcal{H}}$ of features and class labels:

Definition 1. *We denote by* $\mathcal{A} = \sigma(\{A_1, \dots, A_d\})$ *the minimal sub-*σ*-algebra of* $\mathcal{F}$ *containing all* A_i, $i = 1, \dots, d$ *and by* $\overline{\mathcal{H}}$ *the minimal sub-*σ*-algebra of* $\mathcal{F}$ *containing both* $\mathcal{H}$ *and* $\mathcal{A}$*, i.e.,* $\overline{\mathcal{H}} = \sigma(\mathcal{H} \cup \mathcal{A})$.

Note that the σ-algebra $\mathcal{A}$ can be represented as

$$\mathcal{A} \;=\; \{\bigcup_{i=1}^{d} (A_i \cap F_i) : F_1, \dots, F_d \in \{\emptyset, \Omega\}\}, \tag{1a}$$

while the σ-algebra $\overline{\mathcal{H}}$ can be written as

$$\overline{\mathcal{H}} \;=\; \{\bigcup_{i=1}^{d} (A_i \cap H_i) : H_1, \dots, H_d \in \mathcal{H}\}. \tag{1b}$$

A standard assumption in machine learning is that source and target distribution are the same, i.e., $P = Q$. The situation where $P[F] \neq Q[F]$ holds for at least one $F \in \overline{\mathcal{H}}$ is called *dataset shift* (Moreno-Torres et al. [2], Definition 1).

Under dataset shift as defined this way, typically, classifiers or posterior class probabilities learnt under the source distribution stop working properly under the target distribution. Finding algorithms to deal with this problem is one of the tasks in the field of *domain adaptation*.

In this paper, we are mostly interested in exploring how posterior class probabilities change between a source and a target distribution as described in Assumption 1. In particular, we provide generalisations of the posterior correction formula (2.4) of Saerens et al. [4] (see also Theorem 2 of Elkan [5]). For this purpose, the notions of conditional expectation and conditional probability are crucial.

In the following, E_P denotes conditional or unconditional expectation with respect to the probability measure P. For a given probability space $(\Omega, \mathcal{F}, P)$, we refer to Section 8.2 of Klenke [7] for the formal definitions and properties of

- The expectation $E_P[X \mid \mathcal{H}]$ of a real-valued random variable X conditional on a sub-σ-algebra $\mathcal{H}$;
- The probability $P[F \mid \mathcal{H}]$ of an event $F \in \mathcal{F}$ conditional on $\mathcal{H}$.

In the machine learning literature, often the term *posterior class probability* rather than conditional probability is used to refer to the conditional probabilities $P[A_i \mid \mathcal{H}]$ and $Q[A_i \mid \mathcal{H}]$, $i = 1, \ldots, d$, in the context of Assumption 1. In contrast, the term *prior probability* is used for the probabilities $P[A_i]$ and $Q[A_i]$, which in our measure-theoretic setting should rather be called unconditional probabilities of A_i.

An assumption of absolute continuity is also crucial for an investigation of how the posterior class probabilities are impacted by a change from the source distribution to the target distribution. Formally, this assumption reads as follows:

Assumption 2. *Assumption 1 holds, and Q is absolutely continuous with respect to P on $\overline{\mathcal{H}}$, i.e.,*

$$Q|\overline{\mathcal{H}} \ll P|\overline{\mathcal{H}},$$

where $M|\mathcal{H}$ stands for the measure M with domain restricted to $\mathcal{H}$.

The statement "Q is absolutely continuous with respect to P on $\overline{\mathcal{H}}$" means that for all events $N \in \overline{\mathcal{H}}$, $P[N] = 0$ implies $Q[N] = 0$. Hence, "impossible" events under P are also impossible under Q. Measure-theoretic impossibility is somewhat unintuitive because for continuous distributions each single outcome has probability 0 and therefore is impossible. Nonetheless, sampled values from such distributions are single outcomes and occur despite having probability 0.

However, the statement "for all events $N \in \overline{\mathcal{H}}$, $P[N] = 0$ implies $Q[N] = 0$" is equivalent to saying: for all events $N \in \overline{\mathcal{H}}$, $Q[N] > 0$ implies $P[N] > 0$. This means that "possible" events under Q are also possible events under P, even if with very tiny probabilities of occurrence. This phrasing of absolute continuity is more intuitive and is preferred by some authors, for instance by He et al. [3] who in Section 2 make the assumption $\mathcal{D}_T(x, y) > 0 \Rightarrow \mathcal{D}_S(x, y) > 0$, which they seem to understand in the sense of Assumption 2.

As mentioned before, if the target distribution Q is absolutely continuous with respect to P, there may be events whose probabilities under Q are much greater than their probabilities under P. From a practical point of view, such events may even appear to be "impossible" under P. Notions such as "sufficient support" and "support sufficiency divergence" (Johansson et al. [10]) suggest that such is the view of the machine learning community. Hence, Assumption 2 is not necessarily in contrast to the working assumption of partially or fully nonoverlapping source and target domains made by many researchers in unsupervised domain adaptation.

For analyses of the case of domains where the source does not completely cover the target (such that Assumption 2 may be violated), see Johannsson et al. [10]. However, the statement of Johannsson et al., Section 5, "If this overlap is increased without losing information, such as through collection of additional samples, this is usually preferable." suggests that an assumption of nonoverlapping support is not the same as an assumption on a lack of absolute continuity. For according to the statement by Johannsson et al., events outside of the source support do not appear to be impossible because in that case the "collection of additional samples" could not increase the support overlap between source and target.

Assumption 2 is stronger than the common assumption of absolute continuity on $\mathcal{H}$ (see for instance, Scott [11]), but in terms of interpretation there is no big difference: all events possible under the target distribution (including in label space) are also possible under the source distribution.

An important consequence of Assumption 2 is that we can use the source distribution P as a reference measure for the target distribution Q. This is more natural than introducing another measure without real-world meaning as a reference for both P and

Q. In addition, renouncing another measure as a reference has the advantageous effect of simplifying notation.

Recall the following common conventions intended to make the measure-theoretic notation more incisive:

Notation 1. *An important consequence of deploying a measure-theoretic framework as in this paper is that real-valued random variables X on a fixed probability space $(\Omega, \mathcal{F}, P)$ are uniquely defined only up to events of probability 0 and may be undefined or ill-defined on such events or when being multiplied with the factor 0. To be more specific:*

- *If X' is another random variable such that $P[X \neq X'] = 0$, then $E_P[X']$ exists if and only if $E_P[X]$ exists. In this case, $E_P[X] = E_P[X']$ follows.*
- *If X is undefined or ill-defined on an event $N \in \mathcal{F}$ with $P[N] = 0$, then by definition $E_p[X]$ exists if and only if $E_P[X']$ exists for*

$$X' = \begin{cases} X, & \text{on } \Omega \setminus N, \\ 0, & \text{on } N. \end{cases}$$

  In this case, $E_P[X]$ is defined as $E_P[X']$.
- *If X is undefined or ill-defined on an event $F \in \mathcal{F}$ but is multiplied with another random variable Z which takes the value 0 on F, then, by definition, $E_p[X\,Z]$ exists if and only if $E_P[X']$ exists for*

$$X' = \begin{cases} X\,Z, & \text{on } \Omega \setminus F, \\ 0, & \text{on } F. \end{cases}$$

  In this case, $E_P[X\,Z]$ is defined as $E_P[X']$.

The conventions listed in Notation 1 are convenient and used frequently in the following text. Note, however, that they are only valid in the context of a fixed probability measure P. For instance, under Assumption 2, if the event N where the random variable X has probability 0 under the source distribution P of being undefined, i.e., $P[N] = 0$, then $Q[N] = 0$ follows as well, such that $E_Q[X]$ should be well-defined. Nonetheless, $Q[N] = 0$ does not necessarily imply $P[N] = 0$ such that $E_Q[X]$ might be well-defined despite $E_P[X]$ being ill-defined.

In the same vein, under Assumption 2, for the posterior class probabilities $P[A_i \,|\, \mathcal{H}]$, $i = 1, \ldots, d$, the expectations $E_Q\big[P[A_i \,|\, \mathcal{H}]\big]$ are well-defined. However, for the posterior class probabilities $Q[A_i \,|\, \mathcal{H}]$, $i = 1, \ldots, d$, the expectations $E_P\big[Q[A_i \,|\, \mathcal{H}]\big]$ are potentially ill-defined because there could be versions of $Q[A_i \,|\, \mathcal{H}]$ which are indistinguishable under Q but different with positive probability under P. In the following, we are careful to avoid such issues whenever the discussion involves more than one probability measure.

3. General Dataset Shift in Multinomial Classification

Under Assumption 2, by the Radon–Nikodym theorem, there is an $\overline{\mathcal{H}}$-measurable density $\overline{h} = \frac{dQ}{dP}\Big|\overline{\mathcal{H}}$ of the target distribution Q with respect to the target distribution P on the joint information set $\overline{\mathcal{H}}$ defined by (1b). This density links Q to P by Equation (2):

$$Q[F] \;=\; E_P[\overline{h}\,\mathbf{1}_F], \qquad \text{for all } F \in \overline{\mathcal{H}}. \tag{2}$$

In (2) and in the remainder of the paper, $\mathbf{1}_F$ denotes the indicator function of F, defined by $\mathbf{1}_F(\omega) = 1$ if $\omega \in F$ and $\mathbf{1}_F(\omega) = 0$ if $\omega \notin F$.

Unfortunately, in practice $\overline{h}$ is more or less unobservable. Therefore, it is desirable to decompose it into smaller parts which may be observable or can perhaps be determined through reasonable assumptions. The key step to such a decomposition is made with the following combination of definitions and lemma.

Definition 2. *Under Assumption 1, define the following class-conditional distributions, by letting for $F \in \mathcal{F}$ and $i = 1, \ldots, d$*

$$P_i[F] = P[F \mid A_i] = \frac{P[A_i \cap F]}{P[A_i]} \quad \text{and} \quad Q_i[F] = Q[F \mid A_i] = \frac{Q[A_i \cap F]}{Q[A_i]}. \tag{3}$$

In the literature, when restricted to the feature information set $\mathcal{H}$, the P_i and Q_i sometimes are called *class-conditional feature distributions*.

Lemma 1. *Under Assumption 2, for $i = 1, \ldots, d$, the class-conditional feature distribution Q_i is absolutely continuous with respect to P_i on $\mathcal{H}$.*

Denote by $h_i = \frac{dQ_i}{dP_i}\big|\mathcal{H}$ a Radon–Nikodym derivative (or density) of Q_i with respect to P_i. If there is another $\mathcal{H}$-measurable function $h_i^* \geq 0$ with the density property, i.e., $Q_i[H] = E_{P_i}[h_i^* \, \mathbf{1}_H]$ for all $H \in \mathcal{H}$, then it follows that

$$P\big[h_i \neq h_i^*, \; P[A_i \mid \mathcal{H}] > 0\big] \;=\; 0 \;=\; P[\{h_i \neq h_i^*\} \cap A_i]. \tag{4}$$

Proof. Fix i and choose any $N \in \overline{\mathcal{H}}$ with $P_i[N] = 0$. Then, it follows that $N \cap A_i \in \overline{\mathcal{H}}$ and $P[N \cap A_i] = 0$. By Assumption 2, $Q[N \cap A_i] = 0$ follows, which implies

$$Q_i[N] \;=\; \frac{Q[N \cap A_i]}{Q[A_i]} \;=\; 0.$$

Hence, we have $Q_i|\overline{\mathcal{H}} \ll P_i|\overline{\mathcal{H}}$ from which $Q_i|\mathcal{H} \ll P_i|\mathcal{H}$ follows. The uniqueness of Radon–Nikodym derivatives implies

$$0 \;=\; P_i[h_i \neq h_i^*] \;=\; \frac{P[\{h_i \neq h_i^*\} \cap A_i]}{P[A_i]},$$

and hence the right-hand side of (4). However, by the definition of conditional probability it also follows that

$$0 \;=\; P[\{h_i \neq h_i^*\} \cap A_i] \;=\; E_P\big[\mathbf{1}_{\{h_i \neq h_i^*\}} \, P[A_i \mid \mathcal{H}]\big].$$

This implies the left-hand side of (4). □

With Lemma 1 as preparation, we are in a position to state the following key representation result and some corollaries for the joint density $\overline{h}$ of features and class labels. In the remainder of this paper, we make use of (5) as a normal form for $\overline{h}$.

Theorem 1. *Under Assumption 2, the density $\overline{h}$ of Q with respect to P on $\overline{\mathcal{H}}$ can be represented as*

$$\overline{h} \;=\; \sum_{i=1}^{d} h_i \, \frac{Q[A_i]}{P[A_i]} \, \mathbf{1}_{A_i} \tag{5}$$

where the h_i are any densities of Q_i with respect to P_i on $\mathcal{H}$ as introduced in Lemma 1, for $i = 1, \ldots, d$.

Proof. Let $F \in \overline{\mathcal{H}}$. By (1b), then it holds that

$$F \;=\; \bigcup_{i=1}^{d} (A_i \cap H_i) \quad \text{for some } H_1, \ldots, H_d \in \mathcal{H}.$$

This implies

$$\begin{aligned} Q[F] &= \sum_{i=1}^{d} Q[A_i]\, Q_i[H_i] \\ &= \sum_{i=1}^{d} Q[A_i]\, E_{P_i}[h_i\, \mathbf{1}_{H_i}] \\ &= \sum_{i=1}^{d} \frac{Q[A_i]}{P[A_i]} E_P[h_i\, \mathbf{1}_{H_i \cap A_i}] \\ &= E_P\left[\left(\sum_{i=1}^{d} h_i\, \frac{Q[A_i]}{P[A_i]}\, \mathbf{1}_{A_i}\right) \mathbf{1}_F\right]. \end{aligned}$$

Equation (5) follows from this by the definition of Radon–Nikodym derivatives. □

Corollary 1. *Under Assumption 2, the density h of Q with respect to P on $\mathcal{H}$ can be written as*

$$h \;=\; \sum_{i=1}^{d} h_i\, \frac{Q[A_i]}{P[A_i]}\, P[A_i \,|\, \mathcal{H}].$$

Proof. The corollary follows from Theorem 1 because $h = E_P[\overline{h} \,|\, \mathcal{H}]$. □

Corollary 2. *Under Assumption 2, for $i = 1, \ldots, d$, the conditional probability (posterior class probability) $Q[A_i \,|\, \mathcal{H}]$ can be represented as*

$$Q[A_i \,|\, \mathcal{H}] \;=\; \frac{h_i\, \frac{Q[A_i]}{P[A_i]}\, P[A_i \,|\, \mathcal{H}]}{\sum_{j=1}^{d} h_j\, \frac{Q[A_j]}{P[A_j]}\, P[A_j \,|\, \mathcal{H}]}, \tag{6}$$

on the set $\{h > 0\}$, where h denotes the denominator of the right-hand side of (6) (and the density of Q with respect to P on $\mathcal{H}$, as introduced in Corollary 1).

Equation (6) generalises Equation (2.4) of Saerens et al. [4] and Theorem 2 of Elkan [5] from prior probability shift to general dataset shift. Saerens et al. commented on their Equation (2.4) as follows: "This well-known formula can be used to compute the corrected a posteriori probabilities, ...". Hence, in this paper we call (6) the *posterior correction formula*.

Recall that under Assumption 2, it holds that $Q[h > 0] = 1$ while $P[h > 0] < 1$ is possible. Hence, $Q[A_i \,|\, \mathcal{H}]$ is fully specified by (6) under Q but possibly only incompletely specified under P.

Proof of Corollary 2. Apply the generalised Bayes formula (see Lemma A1 in Appendix A) with $\mathcal{F} = \overline{\mathcal{H}}$, $f = \overline{h}$, $\mathcal{G} = \mathcal{H}$ and $X = \mathbf{1}_{A_i}$. □

A direct application of the posterior correction formula (6) is not possible because the target prior probabilities $Q[A_i]$ and the target class conditional feature densities h_i typically are unknown. However, in some cases the target priors might be known from external sources such as central banks, IMF or national offices of statistics. Under more specific assumptions on the type of dataset shift, it may be possible to estimate the target priors from the target dataset. See González et al. [12] for a survey of estimation methods under the assumption of prior probability shift.

Under prior probability shift, $h_i = 1$ is assumed for all i (see Section 5.1 below). This means there is no change of the conditional feature distributions. This assumption might be too strong in some situations. It might be more promising to assume similar changes for all classes (i.e., $h_i \approx h_j$ for $i \neq j$), for instance, by assuming factorizable joint shift (see

Section 4 below), or by trying to find transformations (or representations) of the features that make the resulting feature densities similar (see Sections 5.4 and 5.5 below).

For the sake of completeness, we also mention the following alternative representation (7b) of $\overline{h} = \frac{dQ}{dP}|\overline{\mathcal{H}}$. Compared to (7b), (5) provides more structural information, in particular when taking into account Corollary 2 above and, therefore, is potentially more useful.

Corollary 3. *Under Assumption 2, let h be a density of Q with respect to P on $\mathcal{H}$. Then, the target posterior class probabilities $Q[A_i \,|\, \mathcal{H}]$ vanish on the event $\{h > 0\}$ if the source posterior class probabilities $P[A_i \,|\, \mathcal{H}]$ vanish on $\{h > 0\}$, i.e., it holds on $\{h > 0\}$ that*

$$P[A_i \,|\, \mathcal{H}] = 0 \quad \Rightarrow \quad Q[A_i \,|\, \mathcal{H}] = 0. \tag{7a}$$

Moreover, the density $\overline{h}$ of Q with respect to P on $\overline{\mathcal{H}}$ can be represented as

$$\overline{h} \;=\; h \sum_{i=1}^{d} \frac{Q[A_i \,|\, \mathcal{H}]}{P[A_i \,|\, \mathcal{H}]} \mathbf{1}_{A_i}. \tag{7b}$$

Proof. Equation (7a) follows immediately from Corollary 2. Taking into account Notation 1 for the meaning of (7b) on the event $\{P[A_i \,|\, \mathcal{H}] = 0\}$, the equation follows from (1b) and the definition of the posterior class probabilities. □

The following result may be considered an inversion of the previous results and in particular Corollary 2 on the relationship between source and target distributions. It is of interest mostly for dealing with sample selection bias (see Section 6 below).

Proposition 1. *In the setting of Theorem 1, assume additionally that $P[\overline{h} = 0] = 0$ holds. Then, the following statements hold true:*

(i) *P is absolutely continuous with respect to Q on $\overline{\mathcal{H}}$, with $\frac{dP}{dQ}|\overline{\mathcal{H}} = 1/\overline{h}$.*

(ii) *For $i = 1, \ldots, d$, the source class-conditional feature distribution P_i is absolutely continuous with respect to Q_i on $\mathcal{H}$, with $Q_i[h_i = 0] = 0 = P[h_i = 0]$ and*

$$\frac{dP_i}{dQ_i}\Big|\mathcal{H} \;=\; \frac{1}{h_i}.$$

(iii) *The density $\frac{dP}{dQ}|\overline{\mathcal{H}}$ can also be represented as*

$$\frac{dP}{dQ}\Big|\overline{\mathcal{H}} \;=\; \sum_{i=1}^{d} \frac{1}{h_i} \frac{P[A_i]}{Q[A_i]} \mathbf{1}_{A_i}.$$

(iv) *The density $\frac{dP}{dQ}|\mathcal{H}$ can be represented as*

$$\frac{dP}{dQ}\Big|\mathcal{H} \;=\; \sum_{i=1}^{d} \frac{1}{h_i} \frac{P[A_i]}{Q[A_i]} Q[A_i \,|\, \mathcal{H}].$$

(v) *For $i = 1, \ldots, d$, it holds that*

$$P[A_i \,|\, \mathcal{H}] \;=\; \frac{\frac{1}{h_i} Q[A_i \,|\, \mathcal{H}] \frac{P[A_i]}{Q[A_i]}}{\sum_{j=1}^{d} \frac{1}{h_j} Q[A_j \,|\, \mathcal{H}] \frac{P[A_j]}{Q[A_j]}}.$$

Proof. (i) is a well-known property of equivalent probability measures (see Problem 32.6 of Billingsley [6]).

By (i), P is absolutely continuous with respect to Q on $\overline{\mathcal{H}}$. This implies that P_i is absolutely continuous with respect to Q_i on $\mathcal{H}$ and, again by Problem 32.6 of [6], the rest of (ii) follows as well.

Properties (iii), (iv) and (v) follow from (i) and (ii), by making use of Theorem 1 and Corollaries 1 and 2 with swapped roles of P and Q. □

4. Factorizable Joint Shift

The following definition translates Definition 2.2 of He et al. [3] into the setting of this paper.

Definition 3. *Under Assumption 2, we say that the target distribution Q is related to the source distribution P by* factorizable joint shift (FJS)*, if there are a non-negative $\mathcal{H}$-measurable function g and a non-negative $\mathcal{A}$-measurable function b such that the density $\overline{h}$ of Q with respect to P on $\overline{\mathcal{H}}$ can be represented as*

$$\overline{h} \;=\; g\, b. \tag{8a}$$

Observe that the functions g and b of Definition 3 are not uniquely determined because for any $c > 0$ the functions $g_c = c\, g$ and $b_c = b/c$ are also $\mathcal{H}$-measurable and $\mathcal{A}$-measurable, respectively, and satisfy

$$\overline{h} \;=\; g_c\, b_c. \tag{8b}$$

In the remainder of this section, we show that the functions g and b depend on the source distribution P as well as the marginal distributions of Q on $\mathcal{H}$ and $\mathcal{A}$, respectively, but not on the joint distribution $Q|\overline{\mathcal{H}}$. For the case $d = 2$, in Section 4.2 below we obtain the stronger result that

- g and b are uniquely determined (up to the ambiguity expressed by (8b)) by the marginal distributions of Q on $\mathcal{H}$ and $\mathcal{A}$ and the source distribution P;
- With fixed source distribution P, for each pair of marginal distributions of Q on $\mathcal{H}$ and $\mathcal{A}$, there exists (up to a constant factor) a factorization (8a).

Theorem 2. *Under Assumption 2, let the source distribution P and the target distribution Q be related by joint factorizable shift in the sense of Definition 3. Denote by h the density of Q with respect to P on $\mathcal{H}$ and let $q_i = Q[A_i]$ and $p_i = P[A_i]$, $i = 1, \ldots, d$.*

Then, up to a constant factor c as in (8b), it follows that

$$b = \sum_{i=1}^{d-1} \varrho_i \, \frac{q_i}{p_i}\, \mathbf{1}_{A_i} + \frac{q_d}{p_d}\, \mathbf{1}_{A_d} \quad \textit{and} \tag{9a}$$

$$g = \frac{h}{\sum_{i=1}^{d-1} \varrho_i \, \frac{q_i}{p_i}\, P[A_i \,|\, \mathcal{H}] + \frac{q_d}{p_d}\, P[A_d \,|\, \mathcal{H}]}, \tag{9b}$$

where the constants $\varrho_1, \ldots, \varrho_{d-1}$ are positive and finite and satisfy the following equation system:

$$p_j \;=\; \varrho_j \, E_P\left[\frac{h\, P[A_j \,|\, \mathcal{H}]}{\sum_{i=1}^{d-1} \varrho_i \, \frac{q_i}{p_i}\, P[A_i \,|\, \mathcal{H}] + \frac{q_d}{p_d}\, P[A_d \,|\, \mathcal{H}]} \right], \quad j = 1, \ldots, d-1. \tag{9c}$$

Conversely, let an $\mathcal{H}$-measurable function $h \geq 0$ with $E_P[h] = 1$ and $(q_i)_{i=1,\ldots,d} \in (0,1)^d$ with $\sum_{i=1}^{d} q_i = 1$ be given. If $\varrho_1 > 0, \ldots, \varrho_{d-1} > 0$ are solutions of the equation system (9c) and b and g are defined by (9a) and (9b), respectively, then $g\, b$ is a density of a probability measure Q with respect to P on $\overline{\mathcal{H}}$, such that h is the marginal density of Q with respect to P on $\mathcal{H}$ and $Q[A_i] = q_i$ holds for $i = 1, \ldots, d$.

Proof. First, we show that (9a)–(9c) are necessary if Q and P are related by factorizable joint shift as in (8a).

Since b is $\mathcal{A}$-measurable by assumption, there are constants $\beta_1, \dots, \beta_d \in \mathbb{R}$ such that

$$b \;=\; \sum_{i=1}^{d} \beta_i \, \mathbf{1}_{A_i}. \tag{10a}$$

For fixed $k \in \{1, \dots, d\}$, this implies

$$q_k = E_P[g\, b\, \mathbf{1}_{A_k}] = \beta_k \, E_P[g\, \mathbf{1}_{A_k}] = \beta_k \, p_k \, E_{P_k}[g].$$

By Assumption 2, we have $q_k > 0$ and $p_k > 0$. Hence, it follows $E_{P_k}[g] > 0$ and

$$\beta_k \;=\; \frac{q_k}{p_k \, E_{P_k}[g]} \;>\; 0. \tag{10b}$$

As $g\, b$ is by assumption an $\overline{\mathcal{H}}$-density of Q with respect to P, it follows that

$$h \;=\; E_p[g\, b \,|\, \mathcal{H}] \;=\; g \sum_{i=1}^{d} \frac{q_i}{p_i \, E_{P_i}[g]} \, P[A_i \,|\, \mathcal{H}].$$

The relations $1 = \sum_{i=1}^{d} P[A_i \,|\, \mathcal{H}]$ and (10b) imply

$$\sum_{i=1}^{d} \frac{q_i}{p_i \, E_{P_i}[g]} \, P[A_i \,|\, \mathcal{H}] \;>\; 0.$$

Therefore, we obtain

$$g \;=\; \frac{h}{\sum_{i=1}^{d} \frac{q_i}{p_i \, E_{P_i}[g]} \, P[A_i \,|\, \mathcal{H}]}. \tag{11}$$

For $k \in \{1, \dots, d-1\}$, (11) implies

$$\begin{aligned} E_{P_k}[g] &= \frac{E_P\big[g\, P[A_k \,|\, \mathcal{H}]\big]}{p_k} \\ &= \frac{1}{p_k} E_P\left[\frac{h\, P[A_k \,|\, \mathcal{H}]}{\sum_{i=1}^{d} \frac{q_i}{p_i \, E_{P_i}[g]} \, P[A_i \,|\, \mathcal{H}]}\right], \end{aligned}$$

and, equivalently,

$$p_k = \frac{E_{P_d}[g]}{E_{P_k}[g]} E_P\left[\frac{h\, P[A_k \,|\, \mathcal{H}]}{\sum_{i=1}^{d-1} \frac{E_{P_d}[g]}{E_{P_i}[g]} \, \frac{q_i}{p_i} \, P[A_i \,|\, \mathcal{H}] + \frac{q_d}{p_d} \, P[A_d \,|\, \mathcal{H}]}\right].$$

With $\varrho_k = \frac{E_{P_d}[g]}{E_{P_k}[g]} > 0$, this implies (9c). Equations (9a) and (9b) follow from multiplying (10a) with $E_{P_d}[g]$ and (11) with $1/E_{P_d}[g]$, respectively.

The converse statement follows from the following observations:

- With b and g as in (9a) and (9b), $E_P[g\, b] = 1$ holds such that $g\, b$ is an $\overline{\mathcal{H}}$-measurable density with respect to P.
- Furthermore, $E_P[g\, b \,|\, \mathcal{H}] = h$ holds such that h is the marginal density of $g\, b$ on $\mathcal{H}$ with respect to P.
- For $j \in \{1, \dots, d-1\}$, (9c) is actually equivalent to

$$Q[A_j] \;=\; E_P[g\, b\, \mathbf{1}_{A_j}] \;=\; q_j.$$

Finally, $q_d = Q[A_d]$ is implied by $\sum_{i=1}^{d} q_i = 1$. □

Thanks to Theorem 2, the following version of the posterior correction formula (6) can be given for factorizable joint shift.

Corollary 4. *Under Assumption 2, let the source distribution P and the target distribution Q be related by joint factorizable shift in the sense of Definition 3. Denote by h the density of Q with respect to P on $\mathcal{H}$. Then, the target posterior probabilities $Q[A_j \,|\, \mathcal{H}]$, $j = 1, \ldots, d$, can be represented as functions of the source posterior probabilities $P[A_j \,|\, \mathcal{H}]$, $j = 1, \ldots, d$, in the following way on the event $\{h > 0\}$:*

$$\begin{aligned} Q[A_j \,|\, \mathcal{H}] &= \frac{\varrho_j \, \frac{Q[A_j]}{P[A_j]} P[A_j \,|\, \mathcal{H}]}{\sum_{i=1}^{d-1} \varrho_i \, \frac{Q[A_i]}{P[A_i]} P[A_i \,|\, \mathcal{H}] + \frac{Q[A_d]}{P[A_d]} P[A_d \,|\, \mathcal{H}]}, \quad j = 1, \ldots, d-1, \\ Q[A_d \,|\, \mathcal{H}] &= \frac{\frac{Q[A_d]}{P[A_d]} P[A_d \,|\, \mathcal{H}]}{\sum_{i=1}^{d-1} \varrho_i \, \frac{Q[A_i]}{P[A_i]} P[A_i \,|\, \mathcal{H}] + \frac{Q[A_d]}{P[A_d]} P[A_d \,|\, \mathcal{H}]}, \end{aligned} \tag{12}$$

where the positive constants $\varrho_1, \ldots, \varrho_{d-1}$ satisfy the equation system (9c).

Proof. Apply the generalised Bayes formula (Lemma A1 in Appendix A) for $\mathcal{G} = \mathcal{H}$, $X = \mathbf{1}_{A_j}$ and $f = g\, b$, with g and b specified by (9b) and (9a), respectively. □

Remark 1. *Assuming $P[A_d \,|\, \mathcal{H}] > 0$, (12) implies*

$$\frac{Q[A_j \,|\, \mathcal{H}]}{Q[A_d \,|\, \mathcal{H}]} \frac{Q[A_d]}{Q[A_j]} = \varrho_j \frac{P[A_j \,|\, \mathcal{H}]}{P[A_d \,|\, \mathcal{H}]} \frac{P[A_d]}{P[A_j]}, \quad j = 1, \ldots, d-1. \tag{13}$$

Recall that $P[A_k \,|\, \mathcal{H}]/P[A_k]$ is the density with respect to P of the class-conditional feature distribution P_k, as defined by (3), on the feature information set $\mathcal{H}$. Similarly, $Q[A_k \,|\, \mathcal{H}]/Q[A_k]$ is the density with respect to Q of the class-conditional feature distribution Q_k on $\mathcal{H}$. Therefore, (13) states that under factorizable joint shift, the ratios of the class-conditional feature densities are invariant up to a constant factor.

Remark 1 suggests joint factorizable shift could also be called *scaled density ratios* shift. This term would emphasise a probabilistic interpretation of this kind of dataset shift, in contrast to "factorizable joint shift" with its focus on the technical aspect of separation of input and output variables.

4.1. Alternatives to Joint Importance Aligning

He et al. [3] proposed in Section 3 the "joint importance aligning" method for estimating a factorized version of the ratio of source and target domain densities which they called "joint importance weight". He et al. presented a "supervised" and an "unsupervised" version of their method. The "unsupervised" version was intended for the case where no class labels were observed in the target domain, i.e., the case considered primarily in this paper.

Regarding the performance of the "unsupervised" version of their proposal, He et al. indicated that the proposed method tended to present simple covariate shift (see Section 5.2 below) as a solution. This does not come as a surprise because He et al. [3] stated "... in unsupervised objective, we define $\tilde{V}(x) \triangleq \mathbb{E}_{y \sim \mathcal{D}_S(y|x)} V(y) \ldots$", which suggests that the authors implicitly assumed $\mathcal{D}_S(y|x) = \mathcal{D}_T(y|x)$, i.e., covariate shift. Without providing an explanation, He et al. proposed a discretisation of the data (covariate) space in order to prevent the algorithm from converging to covariate shift as solution.

Given these qualms about "joint importance aligning', it might be useful to point out alternative approaches to finding the factorization (8a), based on Theorem 2. The theorem suggests two obvious ways to learn the characteristics of factorizable joint shift:

(a) If the target prior class probabilities $Q[A_i]$ are known (for instance from external sources), solve (9c) for the constants ϱ_i.
(b) If the target prior class probabilities $Q[A_i]$ are unknown, fix values for the constants ϱ_i and solve (9c) for the $Q[A_i]$. Letting $\varrho_i = 1$ for all i is a natural choice that converts (9c) into the system of maximum likelihood equations for the $Q[A_i]$ under the prior probability shift assumption.

See Section 4.2.4 of Tasche [13] for an example of approach (a) from the area of credit risk. Regarding the interpretation of (9c) in approach (b) as maximum likelihood equations, see Du Plessis and Sugiyama [14] or Tasche [15]. This interpretation, in particular, implies that an EM (expectation maximisation) algorithm can be deployed for solving the equation system (Saerens et al. [4]).

4.2. The Binary Case

Theorem 2 does not provide sufficient or necessary conditions for the existence or uniqueness of solutions to equation system (9c) if a density h and a candidate class distribution $(q_i)_{i=1,\dots,d}$ are given. In the special case $d = 2$, such an existence and uniqueness statement can be made as the following proposition shows. The following proposition is a generalisation of Section 4.2.4 of Tasche [13].

Proposition 2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{H} \subset \mathcal{F}$ a sub-σ-algebra of $\mathcal{F}$ and $A \in \mathcal{F} \setminus \mathcal{H}$ with $0 < p = P[A] < 1$. Assume that $P\big[P[A \mid \mathcal{H}] \in \{0, 1\}\big] = 0$.*

Then, there exists a solution $\varrho = \varrho_1 > 0$ to (9c) with $A_1 = A$, $A_2 = \Omega \setminus A$, $p_1 = p = 1 - p_2$ and $q_1 = q = 1 - q_2$, if an $\mathcal{H}$-measurable function $h : \Omega \to [0, \infty)$ with $E_P[h] = 1$ and a number $0 < q < 1$ are given.

Assume additionally that $\mathcal{H}$ and A are not independent under P. Then, the solution ϱ to (9c) is unique. Denote by $\phi : (0, 1) \to (0, \infty)$ the function that maps, for a fixed density h, the number $0 < q < 1$ to ϱ, i.e., $\phi(q) = \varrho$. Then, ϕ has the following properties:

(i) *ϕ is strictly increasing and continuous on $(0, 1)$.*
(ii) $\lim_{q \to 0} \phi(q) = \frac{P[A]}{(1-P[A])\, E_P\left[h \frac{P[A \mid \mathcal{H}]}{1-P[A \mid \mathcal{H}]}\right]}$.
(iii) $\lim_{q \to 1} \phi(q) = \frac{P[A]}{1-P[A]} E_P\left[h \frac{1-P[A \mid \mathcal{H}]}{P[A \mid \mathcal{H}]}\right]$.

See Appendix B for a proof of Proposition 2. The uniqueness statement of Proposition 2 is interesting because it implies an answer to the question of whether proper *concept shift* (dataset shift where the marginal distributions of the features and labels, respectively, remain unchanged) can be modelled as factorizable joint shift. The answer—at least for the binary case—is no, because "no shift" then provides the only solution to Equation (9c).

5. Common Types of Dataset Shift

In this section, we revisit some popular special cases of dataset shift. In each case, we discuss the question if factorizable joint shift is implied or if the special type of shift is implied by factorizable joint shift. In addition, we provide in each case an adapted version of the posterior correction formula (6).

5.1. Prior Probability Shift

Moreno-Torres et al. [2] defined *prior probability shift* as invariance of the class-conditional feature distributions between source and target, i.e.,

$$Q_i[H] \;=\; P_i[H], \qquad H \in \mathcal{H}, i = 1, \dots, d, \tag{14a}$$

with Q_i and P_i defined as in (3) above, and $Q[A_i] \neq P[A_i]$ for at least one i. This type of dataset shift is also known as "target shift" [16], "global drift" [17], "label shift" [18] and under other names. In terms of the notation used in Theorem 1, (14a) is equivalent to

having the densities of the Q_i with respect to the P_i on the feature information set $\mathcal{H}$ equal to 1, i.e.,

$$h_i \;=\; 1, \qquad i = 1, \ldots, d. \tag{14b}$$

By Theorem 1, (14b) implies for the density $\overline{h}$ of Q with respect to P on $\overline{\mathcal{H}}$ that

$$\overline{h} \;=\; \sum_{i=1}^{d} \frac{Q[A_i]}{P[A_i]}\, \mathbf{1}_{A_i} \;=\; \frac{\sum_{i=1}^{d} Q[A_i]\, \mathbf{1}_{A_i}}{\sum_{i=1}^{d} P[A_i]\, \mathbf{1}_{A_i}}, \tag{15}$$

which obviously is an $\mathcal{A}$-measurable function. Definition 3 of factorizable joint shift, therefore, is satisfied—as stated by He et al. [3] in Table 1.

The posterior correction formula (6) in this case takes the well-known form

$$Q[A_i \,|\, \mathcal{H}] \;=\; \frac{\frac{Q[A_i]}{P[A_i]}\, P[A_i \,|\, \mathcal{H}]}{\sum_{j=1}^{d} \frac{Q[A_j]}{P[A_j]}\, P[A_j \,|\, \mathcal{H}]}, \tag{16}$$

as noted before, e.g., by Saerens et al. [4] and Elkan [5].

5.2. Covariate Shift

Moreno-Torres et al. [2] defined *covariate shift* as invariance of the posterior class probabilities between source and target, i.e.,

$$Q[A_i \,|\, \mathcal{H}] \;=\; P[A_i \,|\, \mathcal{H}], \qquad i = 1, \ldots, d, \tag{17}$$

and $Q[H] \neq P[H]$ for at least one $H \in \mathcal{H}$.

Proposition 3. *Under Assumption 2, denote by $\overline{h}$ and h, as in Section 3, the densities of Q with respect to P on $\overline{\mathcal{H}}$ and $\mathcal{H}$, respectively. Then, (17) holds true if and only if h is also a density of Q with respect to P on $\overline{\mathcal{H}}$, i.e., $P[\overline{h} = h] = 1$.*

Proof. The "if" part of the assertion is Lemma 1 of Tasche [8]. Taking into account Notation 1, the "only if" is implied by Corollary 3. □

Proposition 3 implies that covariate shift is a special case of factorizable joint shift in the sense of Definition 3, with $b = 1$ and $g = h$, as noted in Table 1 of He et al. [3].

Then, observe that the fact that b is constant implies by (9a) that

$$\varrho_i \;=\; \frac{Q[A_d]}{P[A_d]}\, \frac{P[A_i]}{Q[A_i]}, \qquad \text{for all } i = 1, \ldots, d-1. \tag{18}$$

It can readily be checked that under the assumption of covariate shift the ϱ_i defined by (18) indeed solve equation system (9c).

5.3. Covariate Shift with Posterior Drift

Scott [11] defined *covariate shift with posterior drift* (CSPD) for the binary special case ($d = 2$) of Assumption 1 as the following variant of (17):

there exists a strictly increasing function φ such that

$$Q[A_1 \,|\, \mathcal{H}] \;=\; \varphi\big(P[A_1 \,|\, \mathcal{H}]\big). \tag{19}$$

Equation (19) implies that $Q[A_1 \,|\, \mathcal{H}]$ and $P[A_1 \,|\, \mathcal{H}]$ are strongly comonotonic. As shown in Tasche [19], the converse implication also holds true.

Note that from (19), it also follows that

$$Q[A_2 \,|\, \mathcal{H}] \;=\; 1 - \varphi\big(1 - P[A_2 \,|\, \mathcal{H}]\big).$$

Hence, the increasing link between the posterior positive class probabilities defining CSPD does not only apply to class A_1 but automatically also to the negative class A_2.

CSPD is implied by factorizable joint shift. This follows from (12) because of

$$Q[A_2 \,|\, \mathcal{H}] \;=\; \varphi^*\big(P[A_2 \,|\, \mathcal{H}]\big),$$

with $\varphi^*(x) = \dfrac{\frac{Q[A_2]}{P[A_2]}\, x}{\varrho_1 \frac{Q[A_1]}{P[A_1]}\,(1-x) + \frac{Q[A_2]}{P[A_2]}\, x}$ which is strictly increasing in x.

Under CSPD, the class-conditional densities $h_i = \frac{dQ_i}{dP_i}|\mathcal{H}$, $i = 1,2$, introduced in Lemma 1 can be shown to be

$$\begin{aligned} h_1 &= \frac{Q[A_1]}{P[A_1]}\, h\, \frac{\varphi\big(P[A_1 \,|\, \mathcal{H}]\big)}{P[A_1 \,|\, \mathcal{H}])} \quad \text{and} \\ h_2 &= \frac{1-Q[A_1]}{1-P[A_1]}\, h\, \frac{1-\varphi\big(P[A_1 \,|\, \mathcal{H}]\big)}{1-P[A_1 \,|\, \mathcal{H}])}, \end{aligned} \tag{20}$$

where h is the density of Q with respect to P on the feature information set $\mathcal{H}$. Alas, when used in connection with Theorem 1, (20) does not provide a very useful representation of $\overline{h}$.

5.4. Domain Invariance

Translated into the concepts and notation of this paper, *domain invariance* (see Table 1 of He et al. [3]) is defined as follows:

- There is an $\mathcal{H}$-measurable mapping (transformation) T into some measurable space with the property that

$$Q[M] \;=\; P[M] \qquad \text{for all } M \in \sigma\big(\mathcal{A} \cup \mathcal{G}\big), \tag{21a}$$

  where $\mathcal{G} = \sigma(T)$ denotes the smallest sub-σ-algebra of $\mathcal{H}$ such that T is still $\mathcal{G}$-measurable.
- For all $i = 1, \ldots d$ it holds that:

$$Q[A_i \,|\, \mathcal{H}] \;=\; P[A_i \,|\, \mathcal{G}] \quad \text{and} \quad Q[A_i \,|\, \mathcal{H}] \;=\; Q[A_i \,|\, \mathcal{G}]. \tag{21b}$$

Property (21b) means that T is *sufficient* for $\mathcal{H}$ under both P and Q in the sense of Section 32.3 of Devroye et al. [20].

As mentioned in He et al. [3], (21a) implies covariate shift with respect to $\mathcal{G}$, i.e.,

$$Q[A_i \,|\, \mathcal{G}] \;=\; P[A_i \,|\, \mathcal{G}], \quad i = 1, \ldots, d. \tag{21c}$$

From (21b) then follows covariate shift with respect to $\mathcal{H}$.

Actually, this reasoning shows that in the definition of domain invariance according to He et al. [3], (21a) could be replaced by the weaker assumption (21c), without losing the consequence that covariate shift holds on the whole information set $\mathcal{H}$.

5.5. Generalised Label Shift

Tachet des Combes et al. [21] defined *generalised label shift (GLS)* as follows: there is an $\mathcal{H}$-measurable mapping (transformation) T into some measurable space with the property that

$$Q[G \,|\, A_i] \;=\; P[G \,|\, A_i], \qquad i = 1, \ldots, d, G \in \mathcal{G} = \sigma(T). \tag{22}$$

Since $\sigma(T) \subset \mathcal{H}$ holds, this is weaker than requiring (14a) as for prior probability shift. In this sense, GLS generalises prior probability shift.

He et al. [3] gave in Table 1 a narrower definition of GLS, by requiring in addition to (22) also (21b), and went on to prove that GLS implied factorizable joint shift. We provide an alternative proof of this result, providing mathematically rigorous meaning for the factorisation proposed by He et al.

Proposition 4. *Under Assumption 2, let there be an $\mathcal{H}$-measurable mapping T into some measurable space such that (22) and (21b) hold. Denote by h the density of the target distribution Q with respect to the source distribution P on $\mathcal{H}$. Then, Q and P are related by factorizable joint shift in the sense of Definition 3, with*

$$b = \sum_{i=1}^{d} \frac{Q[A_i]}{P[A_i]} \mathbf{1}_{A_i} \quad \text{and} \qquad g = \frac{h}{\sum_{i=1}^{d} \frac{Q[A_i]}{P[A_i]} P[A_i \mid \mathcal{H}]}. \tag{23}$$

See Appendix B for a proof of Proposition 4. Observe that Proposition 4 and Corollary 4 together imply that the same class posterior correction formula (16) applies for generalised label shift and prior probability shift.

The factorisation presented in (23) of Proposition 4 corresponds to the factorisation of generalised label shift proposed by He et al. [3] in Table 1 in the following way:

- Function b matches $\mathcal{D}_T(Y)$ of He et al. Because in this paper the reference measure is the source distribution P, $\mathcal{D}_S(Y)$ of He et al. corresponds to constant 1 in (23).
- Function h matches $\mathcal{D}_T(X)$ and function γ (the denominator of g) matches $\mathcal{D}_T(Z)$ of He et al. Due to our reference measure being P, $\mathcal{D}_S(X)$ and $\mathcal{D}_S(Z)$ of He et al. are both matched by constant 1. The term $\mathcal{D}_T(X = x \mid Z = g(x))$ appears in (23) as the density ratio $g = h/\gamma$, hence with a well-defined mathematical meaning.

Remark 2. *Proposition 4 combined with Remark 1 shows that "generalized label shift" in the sense of He et al. [3] is the same type of dataset shift that was discussed as "invariant density ratio"-type dataset shift in Tasche [22].*

6. Sample Selection Bias

Sample selection bias is an important cause of dataset shift. In this subsection, we revisit parts of Hein [23] in order to illustrate some of the concepts and results presented before. We basically work under Assumption 1 but without the interpretation of P as source and Q as target distribution. Instead, P is interpreted as the distribution of a population from which a potentially biased random sample is taken, resulting in the distribution Q. When studying sample selection bias in this setting, the goal is to infer properties of P from properties of the sample distribution Q.

The following assumption describes the setting of this section. The idea is that under the population distribution, each object has a positive chance to be selected. This chance may depend upon the features (covariates) and the class of the object.

Assumption 3 (Sample selection)**.** *$(\Omega, \mathcal{F})$ is a measurable space. The* population distribution *P is a probability measure on $(\Omega, \mathcal{F})$. For some positive integer $d \geq 2$, events $A_1, \ldots, A_d \in \mathcal{F}$ and a sub-σ-algebra $\mathcal{H} \subset \mathcal{F}$ are given. The events A_i, $i = 1, \ldots, d$ and $\mathcal{H}$ have the following properties:*

(i) $\bigcup_{i=1}^{d} A_i = \Omega$.
(ii) $A_i \cap A_j = \emptyset$, $i, j = 1, \ldots d$, $i \neq j$.
(iii) $0 < P[A_i]$, $i = 1, \ldots, d$.
(iv) $A_i \notin \mathcal{H}$, $i = 1, \ldots, d$.

The selection probability *is an $\overline{\mathcal{H}}$-measurable random variable $0 < \varphi \leq 1$ where the sub-σ-algebra $\overline{\mathcal{H}}$ is defined as in (1b).*

The probability space $(\Omega, \mathcal{F}, P)$ also supports a random variable U which is uniformly distributed on $[0, 1]$ such that U and $\overline{\mathcal{H}}$ are independent.

Definition 4 (Sample distribution)**.** *Under Assumption 3, define the* event of being selected *by $S = \{U \leq \varphi\}$. The probability measure Q on $(\Omega, \mathcal{F})$, defined by*

$$Q[F] \;=\; P[F \,|\, S] \;=\; \frac{P[F \cap S]}{P[S]}, \qquad \textit{for } F \in \mathcal{F},$$

is called sample distribution.

Note that the measure Q is well-defined because from the independence of U and $\overline{\mathcal{H}}$, it follows that

$$P[S] \;=\; E_P\left[\int_0^1 \mathbf{1}_{[0,\varphi]}(u)\, du\right] \;=\; E_P[\varphi] > 0.$$

Another consequence of the independence of U and $\overline{\mathcal{H}}$ is

$$P[S \,|\, \overline{\mathcal{H}}] \;=\; P[U \leq \varphi \,|\, \overline{\mathcal{H}}] \;=\; \varphi \;>\; 0. \tag{24}$$

Proposition 5. *P and Q as described in Assumption 3 and Definition 4 satisfy Assumptions 1 and 2 with P as source distribution and Q as target distribution. Moreover, P is absolutely continuous with respect to Q on $\overline{\mathcal{H}}$.*

Proof. It remains to show that

- Q is absolutely continuous with respect to P on $\overline{\mathcal{H}}$, with density $\overline{h} = \frac{P[S \,|\, \overline{\mathcal{H}}]}{P[S]}$;
- P is absolutely continuous with respect to Q on $\overline{\mathcal{H}}$;
- $0 < Q[A_i]$ for $i = 1, \ldots, d$.

By definition of Q as P conditional on S, the sample distribution Q is absolutely continuous with respect to P on $\mathcal{F}$ and hence also on $\overline{\mathcal{H}} \subset \mathcal{F}$. For the density $\overline{h}$, we obtain

$$\overline{h} \;=\; E_P\left[\frac{\mathbf{1}_S}{P[S]} \,\middle|\, \overline{\mathcal{H}}\right] \;=\; \frac{P[S \,|\, \overline{\mathcal{H}}]}{P[S]} \;>\; 0. \tag{25}$$

The fact that $\overline{h}$ is positive implies that P is absolutely continuous with respect to Q on $\overline{\mathcal{H}}$. Since $A_i \in \overline{\mathcal{H}}$ for $i = 1, \ldots, d$, the absolute continuity of P with respect to Q implies $Q[A_i] > 0, i = 1, \ldots, d$. □

6.1. Properties of the Sample Selection Model

Equation (25) implies for the density h of Q with respect to P on $\mathcal{H}$

$$h \;=\; E_P[\overline{h} \,|\, \mathcal{H}] \;=\; \frac{P[S \,|\, \mathcal{H}]}{P[S]} \;>\; 0. \tag{26}$$

From representation (1b) of $\overline{\mathcal{H}}$, the following alternative description for $P[S \,|\, \overline{\mathcal{H}}]$ follows:

$$\begin{aligned} P[S \,|\, \overline{\mathcal{H}}] &= \sum_{i=1}^{d} \frac{P[A_i \cap S \,|\, \mathcal{H}]}{P[A_i \,|\, \mathcal{H}]}\, \mathbf{1}_{A_i} \\ &= \sum_{i=1}^{d} P_i[S \,|\, \mathcal{H}]\, \mathbf{1}_{A_i}, \end{aligned} \tag{27}$$

where the P_i denote the class-conditional feature distributions under P, see Definition 2. $P_i[S \,|\, \mathcal{H}]$ is accordingly the feature-conditional probability of being selected on the subpopulation of objects with class A_i.

For $i = 1, \ldots, d$ and $H \in \mathcal{H}$, a short calculation shows:

$$\begin{aligned} Q_i[H] &= \frac{P[A_i \cap H \cap S]}{P[A_i \cap S]} \\ &= \frac{E_P\left[\mathbf{1}_H\, P[A_i \mid \mathcal{H}]\right]}{P[A_i \cap S]} \\ &= \frac{E_P\left[\mathbf{1}_H\, P[A_i \mid \mathcal{H}]\, P_i[S \mid \mathcal{H}]\right]}{P[A_i]\, P_i[S]} \\ &= \frac{E_P\left[\mathbf{1}_{H \cap A_i}\, P_i[S \mid \mathcal{H}]\right]}{P[A_i]\, P_i[S]} \\ &= E_{P_i}\left[\mathbf{1}_H\, \frac{P_i[S \mid \mathcal{H}]}{P_i[S]}\right]. \end{aligned}$$

This implies

$$h_i \;=\; \frac{dQ_i}{dP_i}\Big|\mathcal{H} \;=\; \frac{P_i[S \mid \mathcal{H}]}{P_i[S]}, \quad i = 1, \ldots, d. \tag{28}$$

Equation (28) and Theorem 1 together imply the following alternative representation of $\overline{h}$:

$$\overline{h} \;=\; \sum_{i=1}^{d} \frac{P_i[S \mid \mathcal{H}]}{P_i[S]}\, \frac{Q[A_i]}{P[A_i]}\, \mathbf{1}_{A_i}. \tag{29}$$

By the generalised Bayes formula (Lemma A1 in Appendix A), (25) implies the following representation of the posterior class probabilities $Q[A_i \mid \mathcal{H}]$, $i = 1, \ldots, d$, under Q:

$$\begin{aligned} Q[A_i \mid \mathcal{H}] &= \frac{E_P[\mathbf{1}_{A_i}\, \overline{h} \mid \mathcal{H}]}{E_P[\overline{h} \mid \mathcal{H}]} \\ &= \frac{E_P\left[\mathbf{1}_{A_i}\, P[S \mid \overline{\mathcal{H}}] \,\middle|\, \mathcal{H}\right]}{P[S \mid \mathcal{H}]} \\ &= \frac{P[S \cap A_i \mid \mathcal{H}]}{P[S \mid \mathcal{H}]}. \end{aligned} \tag{30}$$

Zadrozny [24] and Hein [23] observed that if the event S of being selected and the class labels as expressed by the σ-algebra $\mathcal{A}$ were independent conditional on $\mathcal{H}$, the information set reflecting the features, then the population distribution P and the sample distribution Q were related by covariate shift. A consequence of (30) is that the converse of this observation actually also holds true, as stated in the following proposition.

Proposition 6. *In the sample selection model, as specified by Assumption 3 and Definition 4, the population distribution P and the sample distribution Q are related by covariate shift if and only if*

$$P[S \cap A_i \mid \mathcal{H}] \;=\; P[S \mid \mathcal{H}]\, P[A_i \mid \mathcal{H}], \qquad i = 1, \ldots, d,$$

i.e., if the event of being selected and the class labels are independent conditional on the features under the population distribution P.

Proof. Proposition 6 is obvious from (30) and the definition of covariate shift (17). □

In the case of general dataset shift caused by sample selection, Equation (30) does not provide information about how to compute the population posterior class probabilities $P[A_i \mid \mathcal{H}]$ from the sample posterior class probabilities $Q[A_i \mid \mathcal{H}]$. Translated into the setting of this paper, Hein [23] presented in Equation (3.2) the following two ways to do so:

- Define Q^* as the distribution of the not-selected sample, i.e.,

$$Q^*[F] \;=\; P[F \mid (\Omega \setminus S)] \;=\; \frac{P[F] - P[F \cap S]}{1 - P[S]}, \qquad F \in \mathcal{F}.$$

Then, it holds that

$$\begin{aligned} P[A_i \mid \mathcal{H}] &= P[A_i \cap S \mid \mathcal{H}] + P[A_i \cap (\Omega \setminus S) \mid \mathcal{H}] \\ &= Q[A_i \mid \mathcal{H}]\, P[S \mid \mathcal{H}] + Q^*[A_i \mid \mathcal{H}]\, (1 - P[S \mid \mathcal{H}]), \end{aligned} \tag{31a}$$

for $i = 1, \ldots, d$.

- Equation (30) can be written equivalently as

$$Q[A_i \mid \mathcal{H}] \;=\; \frac{P_i[S \mid \mathcal{H}]\, P[A_i \mid \mathcal{H}]}{P[S \mid \mathcal{H}]}.$$

Hence, on the event $\{P_i[S \mid \mathcal{H}] > 0\}$, the following representation of $P[A_i \mid \mathcal{H}]$, $i = 1, \ldots, d$, is obtained:

$$P[A_i \mid \mathcal{H}] \;=\; \frac{P[S \mid \mathcal{H}]}{P_i[S \mid \mathcal{H}]}\, Q[A_i \mid \mathcal{H}]. \tag{31b}$$

Both (31a) and (31b) are of limited practical usefulness, however, as on the one hand, (31a) requires knowledge of the class labels in the not-selected sample, which usually are not available. On the other hand, for (31b) to be applicable, class-wise probabilities of selection $P_i[S \mid \mathcal{H}]$ must be estimated, which again requires knowledge of the class labels in the not-selected sample.

6.2. Sample Selection Bias and Factorizable Joint Shift

Proposition 6 provides an example of a condition for the sample selection process that makes the resulting bias between population and sample representable as covariate shift and, consequently, according to Section 5.2, as a special case of factorizable joint shift. Are there other selection procedures that entail factorizable joint shift?

We investigate this question by assuming that the population distribution P and the sample distribution Q are related by factorizable joint shift and then identifying the consequences this assumption implies for the class-wise feature-conditional selection probabilities $P_i[S \mid \mathcal{H}]$, $i = 1, \ldots, d$.

Theorem 3. *Under Assumption 3 and Definition 4, let P and Q be related by factorizable joint shift in the sense of Definition 3, i.e., there are an $\mathcal{H}$-measurable function $g \geq 0$ and an $\mathcal{A}$-measurable function $b \geq 0$ such that the density $\overline{h}$ of Q with respect to P on $\overline{\mathcal{H}}$ can be represented as $\overline{h} = g\, b$. Then, the following statements hold true:*

(i) Q and P are related by factorizable joint shift with an $\mathcal{H}$-measurable function $g^* > 0$ and an $\mathcal{A}$-measurable function $b^* > 0$ that can be represented up to a constant factor in the sense of (8b) as

$$\begin{aligned} b^* &= \sum_{i=1}^{d} \alpha_i \frac{P[A_i]}{Q[A_i]}\, \mathbf{1}_{A_i} + \frac{P[A_d]}{Q[A_d]}\, \mathbf{1}_{A_d} \quad \textit{and} \\ g^* &= \frac{P[S]}{P[S \mid \mathcal{H}] \left(\sum_{i=1}^{d} \alpha_i \frac{P[A_i]}{Q[A_i]}\, Q[A_i \mid \mathcal{H}] + \frac{P[A_d]}{Q[A_d]}\, Q[A_d \mid \mathcal{H}] \right)}, \end{aligned} \tag{32a}$$

where the constants $0 < \alpha_1, \ldots, \alpha_{d-1} < \infty$ satisfy the following equation system, with $i = 1, \ldots, d-1$:

$$Q[A_i] \;=\; P[S]\, \alpha_i\, E_Q \left[\frac{Q[A_i \mid \mathcal{H}]}{P[S \mid \mathcal{H}] \left(\sum_{j=1}^{d} \alpha_j \frac{P[A_j]}{Q[A_j]}\, Q[A_j \mid \mathcal{H}] + \frac{P[A_d]}{Q[A_d]}\, Q[A_d \mid \mathcal{H}] \right)} \right]. \tag{32b}$$

(ii) The population posterior probabilities $P[A_i \mid \mathcal{H}]$, $i = 1, \ldots, d$, can be represented as functions of the sample posterior probabilities $Q[A_i \mid \mathcal{H}]$, $i = 1, \ldots, d$, in the following way:

$$
\begin{aligned}
P[A_i \mid \mathcal{H}] &= \frac{\alpha_i \, \frac{P[A_i]}{Q[A_i]} \, Q[A_i \mid \mathcal{H}]}{\sum_{j=1}^{d} \alpha_j \, \frac{P[A_j]}{Q[A_j]} \, Q[A_j \mid \mathcal{H}] + \frac{P[A_d]}{Q[A_d]} \, Q[A_d \mid \mathcal{H}]}, \quad i = 1, \ldots, d-1, \\
P[A_d \mid \mathcal{H}] &= \frac{\frac{P[A_d]}{Q[A_d]} \, Q[A_d \mid \mathcal{H}]}{\sum_{j=1}^{d} \alpha_j \, \frac{P[A_j]}{Q[A_j]} \, Q[A_j \mid \mathcal{H}] + \frac{P[A_d]}{Q[A_d]} \, Q[A_d \mid \mathcal{H}]},
\end{aligned}
\tag{33}
$$

where the constants $0 < \alpha_1, \ldots, \alpha_{d-1} < \infty$ *satisfy equation system* (32b).

(iii) The class-wise feature-conditional selection probabilities $P_i[S \mid \mathcal{H}]$, $i = 1, \ldots, d$, *can be represented as*

$$
P_i[S \mid \mathcal{H}] \;=\; \frac{Q[A_i]}{\alpha_i \, P[A_i]} \, P[S \mid \mathcal{H}] \left(\sum_{j=1}^{d} \alpha_j \, \frac{P[A_j]}{Q[A_j]} \, Q[A_j \mid \mathcal{H}] + \frac{P[A_d]}{Q[A_d]} \, Q[A_d \mid \mathcal{H}] \right), \tag{34}
$$

where the constants $0 < \alpha_1, \ldots, \alpha_{d-1} < \infty$ *satisfy equation system* (32b) *and* $\alpha_d = 1$.

Proof. Functions g and b must be positive since $\overline{h}$ is positive according to Proposition 5. Hence, Q and P are related by factorizable joint shift with decomposition $b^* = 1/b$ and $g^* = 1/g$. Apply Theorem 2 with swapped roles of P and Q to obtain representation (32a) and equation system (32b). Statement (ii) follows immediately from Corollary 4.

Regarding (iii), use (28) and Proposition 1 (iv) together with (32a) to obtain

$$
\frac{P_i[S]}{P_i[S \mid \mathcal{H}]} \;=\; \alpha_i \, g^*, \quad i = 1, \ldots, d.
$$

This is equivalent to (34). □

As mentioned in Section 4.1 as a potential application of Theorem 2, assuming that the posterior probabilities $Q[A_i \mid \mathcal{H}]$ under the sample distribution can be estimated, Theorem 3 offers two obvious ways to learn the characteristics of factorizable joint shift:

(a) If the population prior class probabilities $P[A_i]$ are known (for instance from external sources) solve (32b) for the constants α_i.

(b) If the population prior class probabilities $P[A_i]$ are unknown, fix values for the constants α_i and solve (32b) for the $P[A_i]$. Letting $\alpha_i = 1$ for all i is a natural choice that converts (32b) into the system of maximum likelihood equations for the $P[A_i]$ under the prior probability shift assumption.

In case (a), (34) may serve as an admissibility check for the solutions found. If the class-wise selection probabilities $P_i[S \mid \mathcal{H}]$ obtained from (34) can take values greater than 100%, the corresponding set of values $(\alpha_1, \ldots, \alpha_{d-1})$ is not an admissible solution of (32b). If all solutions $(\alpha_1, \ldots, \alpha_{d-1})$ of (32b) turn out to be inadmissible, it must be concluded that the assumption of factorizable joint shift for the sample selection process is wrong.

In case (b), from (34) follows for all $i, j = 1, \ldots, d$

$$
P_i[S \mid \mathcal{H}] \, \frac{P[A_i]}{Q[A_i]} \;=\; P_j[S \mid \mathcal{H}] \, \frac{P[A_j]}{Q[A_j]},
$$

which implies

$$
P_i[S \mid \mathcal{H}] \;\leq\; \frac{Q[A_i]}{P[A_i]} \, \min\left(\frac{P[A_1]}{Q[A_1]}, \ldots, \frac{P[A_d]}{Q[A_d]} \right), \quad \text{for all } i = 1, \ldots, d. \tag{35}
$$

Inequality (35) provides a simple necessary criterion for the presence of factorizable joint shift with constants α_i all equal to 1.

A further, less obvious special case of Theorem 3 is encountered if is assumed that

$$\alpha_i \frac{P[A_i]}{Q[A_i]} = \frac{P[A_d]}{Q[A_d]}, \qquad \text{for all } i = 1, \ldots, d-1. \tag{36}$$

Then, (34) implies $P_i[S \mid \mathcal{H}] = P[S \mid \mathcal{H}]$ for all $i = 1, \ldots, d$. By (31b), this means that population distribution and sample distribution are related by covariate shift, as already observed by Hein [23].

7. Conclusions

We revisited the notion of "factorizable joint shift" recently introduced by He et al. [3]. A main finding is that factorizable joint shift is actually not much more general than prior probability shift or covariate shift. However, in contrast to these two types of shifts, factorizable joint shift is not fully identifiable if no class label information on the test (target) dataset is available and no additional assumptions are made. These findings are based on a representation result (Theorem 2) and a comparison of the class posterior correction formula (12) for factorizable joint shift to the related correction formulae (16) and (17) for prior probability and covariate shifts, respectively. Formula (12) is structurally identical with formula (16) but includes additional constants which can be found by solving the nonlinear equation system (9c).

He et al. [3] did not present the full rationale for their joint importance aligning approach to estimating the characteristics of factorizable joint shift. Hence, solving equation system (9c) for the additional constants in the posterior correction formula or for the prior class probabilities under the target distribution can be considered attractive alternative approaches.

Some open research questions remain:

- Under what conditions can the existence and the uniqueness of solutions $(\varrho_1, \ldots, \varrho_{d-1})$ to equation system (9c) be guaranteed in the case of more than two classes?
- Is there any manageable—in the sense of having observable characteristics—type of dataset shift which is both more complex than factorizable joint shift and less complex than covariate shift with posterior drift?
- To which extent can Theorem 2 be adapted for a more general regression setting?

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author thanks three anonymous reviewers for suggestions that helped to improve an earlier version of this paper.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. The Generalized Bayes Formula

Lemma A1 is Theorem 10.8 of Klebaner [25], slightly extended to explicitly cover the case when the denominator in the formula for the density can be 0.

Lemma A1. *Let $(\Omega, \mathcal{F})$ be a measurable space and P and Q probability measures on $(\Omega, \mathcal{F})$. Assume that $f = \frac{dQ}{dP}$ is a density of Q with respect to P on $\mathcal{F}$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and X be a non-negative random variable on $(\Omega, \mathcal{F})$ or a random variable on $(\Omega, \mathcal{F})$ such that $f\,X$ is P-integrable. Then, the following two statements hold:*

(i) $\{f > 0\} \subset \{E_P[f \mid \mathcal{G}] > 0\}$, *in the sense of* $P\big[f > 0,\, E_P[f \mid \mathcal{G}] = 0\big] = 0$.

(ii) $E_Q[X \mid \mathcal{G}] = \dfrac{E_P[f\,X \mid \mathcal{G}]}{E_P[f \mid \mathcal{G}]} \mathbf{1}_{\{E_P[f \mid \mathcal{G}] > 0\}}$.

Proof. For (i): Observe that

$$E_P\big[f\,\mathbf{1}_{\{E_P[f \mid \mathcal{G}]=0\}}\big] = E_P\big[E_P[f \mid \mathcal{G}]\,\mathbf{1}_{\{E_P[f \mid \mathcal{G}]=0\}}\big] = 0.$$

This implies

$$0 \;=\; P\big[f\,\mathbf{1}_{\{E_P[f\,|\,\mathcal{G}]=0\}} > 0\big] \;=\; P\big[f > 0,\, E_P[f\,|\,\mathcal{G}] = 0\big].$$

For (ii): see Klebaner [25], proof of Theorem 10.8. □

Appendix B. Proofs

Appendix B.1. Proof of Proposition 2

For a more concise notation, define the non-negative, $\mathcal{H}$-measurable random variables R_1 and R_2 by

$$R_1 \;=\; \frac{P[A\,|\,\mathcal{H}]}{p} \quad \text{and} \quad R_2 \;=\; \frac{1 - P[A\,|\,\mathcal{H}]}{1-p.}$$

Then, (9c) can be written as

$$1 \;=\; \varrho\, E_P\left[\frac{h\, R_1}{\varrho\, q\, R_1 + (1-q)\, R_2}\right]. \tag{A1a}$$

Some algebra shows that (A1a) is equivalent to

$$1 \;=\; E_P\left[\frac{h\, R_2}{\varrho\, q\, R_1 + (1-q)\, R_2}\right], \tag{A1b}$$

and that it is also equivalent to

$$0 \;=\; E_P\left[h\frac{\varrho\, R_1 - R_2}{\varrho\, q\, R_1 + (1-q)\, R_2}\right]. \tag{A1c}$$

Define the function $g(\varrho) = E_P\left[\frac{h\, R_2}{\varrho\, q\, R_1 + (1-q)\, R_2}\right]$ for $\varrho \geq 0$. Then, it holds that

- $g(\varrho) \leq \frac{1}{1-q} < \infty$ for all $\varrho \geq 0$;
- $g(0) = \frac{1}{1-q} > 1$;
- By the dominated convergence theorem, g is continuous for $0 \leq \varrho < \infty$ with $\lim\limits_{\varrho\to\infty} g(\varrho) = 0$.

By the mean value theorem, these properties of g imply the existence of some $\varrho > 0$ with $g(\varrho) = 1$. By the equivalence of (A1a) and (A1b), the existence of a positive solution ϱ to (9c) follows.

Regarding the uniqueness of the solution to (9c), define for $q \in (0,1)$ and $\varrho \in (0,\infty)$ the function $f(q,\varrho) = E_P\left[h\frac{\varrho\, R_1 - R_2}{\varrho\, q\, R_1 + (1-q)\, R_2}\right]$. Then, f is continuously partially differentiable with

$$\begin{aligned}
\frac{\partial f}{\partial q}(q,\varrho) &= -E_P\left[h\frac{(\varrho\, R_1 - R_2)^2}{(\varrho\, q\, R_1 + (1-q)\, R_2)^2}\right], \quad \text{and} \\
\frac{\partial f}{\partial \varrho}(q,\varrho) &= E_P\left[h\frac{R_1\, R_2}{(\varrho\, q\, R_1 + (1-q)\, R_2)^2}\right].
\end{aligned}$$

The assumption $P\big[P[A\,|\,\mathcal{H}] \in \{0,1\}\big] = 0$ implies $\frac{\partial f}{\partial \varrho}(q,\varrho) > 0$ for all $0 < q < 1$ and $\varrho > 0$.

$P[\varrho\, R_1 - R_2 = 0] = 1$ would imply $P[A\,|\,\mathcal{H}] = p$ and as a further consequence A and $\mathcal{H}$ would be independent. By assumption, this is not the case, and hence, $P[\varrho\, R_1 - R_2 = 0] < 1$. This implies also $\frac{\partial f}{\partial q}(q,\varrho) > 0$ for all $0 < q < 1$ and $\varrho > 0$.

Consequently, by the implicit function theorem, there exists a continuously differentiable function $\phi : (0,1) \to (0,\infty)$, $q \mapsto \phi(q) = \varrho$ such that $f(q,\phi(q)) = 0$ for all $0 < q < 1$ and

$$\phi'(q) \;=\; \frac{E_P\left[h\frac{(\phi(q)\,R_1-R_2)^2}{(\phi(q)\,q\,R_1+(1-q)\,R_2)^2}\right]}{E_P\left[h\frac{R_1\,R_2}{(\phi(q)\,q\,R_1+(1-q)\,R_2)^2}\right]} \;>\; 0.$$

This proves claim (i) on ϕ and the existence of $\lim_{q\to 0}\phi(q) < \infty$ and $\lim_{q\to 1}\phi(q) > 0$. Making use again of the equivalence of (A1c) and (A1a) and invoking Lemma 4.1 of Tasche [15] now implies for all $0 < q < 1$

$$\frac{1}{E_P\left[h\frac{R_1}{R_2}\right]} \;<\; \phi(q) \;<\; E_P\left[h\frac{R_2}{R_1}\right]. \tag{A2}$$

These inequalities also hold true if $E_P\left[h\frac{R_1}{R_2}\right] = \infty$ or $E_P\left[h\frac{R_2}{R_1}\right] = \infty$.

Now, apply Fatou's lemma to obtain

$$\begin{aligned} 1 &= \liminf_{q\to 1} E_P\left[h\frac{R_2}{\phi(q)\,q\,R_1+(1-q)\,R_2}\right] \\ &\geq E_P\left[h\frac{R_2}{R_1}\frac{1}{\lim\limits_{q\to 1}\phi(q)}\right]. \end{aligned}$$

From this, $\lim_{q\to 1}\phi(q) \geq E_P\left[h\frac{R_2}{R_1}\right]$ follows, and by (A2), also claim (iii).

Another application of Fatou's lemma gives

$$\begin{aligned} 1 &= \liminf_{q\to 0}\left(\phi(q)\,E_P\left[h\frac{R_1}{\phi(q)\,q\,R_1+(1-q)\,R_2}\right]\right) \\ &\geq E_P\left[h\frac{R_1}{R_2}\lim_{q\to 1}\phi(q)\right]. \end{aligned}$$

Together with (A2), this proves claim (ii) and completes the proof.

Appendix B.2. Proof of Proposition 4

Observe that function g in (23) is well-defined because the denominator

$$\gamma = \sum_{i=1}^{d}\frac{Q[A_i]}{P[A_i]}P[A_i\,|\,\mathcal{G}] = \sum_{i=1}^{d}\frac{Q[A_i]}{P[A_i]}P[A_i\,|\,\mathcal{H}] \tag{A3}$$

on the right-hand side of the equation is always positive. On the one hand, by Corollary 2, we obtain for $i = 1,\ldots,d$ on the set $\{h > 0\}$

$$Q[A_i\,|\,\mathcal{H}] \;=\; h_i\,\frac{Q[A_i]}{P[A_i]}\,\frac{P[A_i\,|\,\mathcal{H}]}{h}, \tag{A4a}$$

where h_i denotes the density of the target class-conditional feature distribution Q_i with respect to the source class-conditional feature distribution P_i on $\mathcal{H}$.

On the other hand, by combining the prior probability shift property (22) on $\mathcal{G} = \sigma(T)$, the sufficiency property (21b) and (14b), Corollary 2 implies

$$Q[A_i\,|\,\mathcal{H}] \;=\; \frac{Q[A_i]}{P[A_i]}\,\frac{P[A_i\,|\,\mathcal{H}]}{\gamma}, \tag{A4b}$$

Hence, from (A4a) and (A4b), it follows for $i = 1,\ldots,d$

$$\frac{P[A_i\,|\,\mathcal{H}]}{\gamma} \;=\; h_i\,\frac{P[A_i\,|\,\mathcal{H}]}{h} \quad \text{on } \{h > 0\}. \tag{A4c}$$

Making use of (A4c), we obtain for any $F = \bigcup_{i=1}^{d}(A_i \cap H_i) \in \overline{\mathcal{H}}$

$$\begin{aligned}
Q[F] &= E_Q\big[Q[F \mid \mathcal{F}]\big] \\
&= \sum_{i=1}^{d} E_P\big[h\,\mathbf{1}_{H_i}\,Q[A_i \mid \mathcal{H}]\big] \\
&= \sum_{i=1}^{d} E_P\left[h\,\mathbf{1}_{\{h>0\}}\,\mathbf{1}_{H_i}\,h_i\,\frac{Q[A_i]}{P[A_i]}\,\frac{P[A_i \mid \mathcal{H}]}{h}\right] \\
&= \sum_{i=1}^{d} E_P\left[h\,\mathbf{1}_{H_i}\,\frac{Q[A_i]}{P[A_i]}\,\frac{P[A_i \mid \mathcal{H}]}{\gamma}\right] \\
&= E_P\left[\frac{h}{\gamma}\sum_{i=1}^{d}\frac{Q[A_i]}{P[A_i]}\,\mathbf{1}_{H_i}\,P[A_i \mid \mathcal{H}]\right] \\
&= E_P\left[\frac{h}{\gamma}\sum_{i=1}^{d}\frac{Q[A_i]}{P[A_i]}\,\mathbf{1}_{A_i\cap H_i}\right] \\
&= E_P\left[\mathbf{1}_F\,\frac{h}{\gamma}\sum_{i=1}^{d}\frac{Q[A_i]}{P[A_i]}\,\mathbf{1}_{A_i}\right].
\end{aligned}$$

This proves that $b\,g$ with b and g as defined by (23) is a density of Q with respect to P on $\overline{\mathcal{H}}$. As the $\mathcal{A}$-measurability of b and $\mathcal{H}$-measurability of g are obvious, the proof is complete.

References

1. Storkey, A. When Training and Test Sets Are Different: Characterizing Learning Transfer. In *Dataset Shift in Machine Learning*; Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N., Eds.; The MIT Press: Cambridge, MA, USA, 2009; Chapter 1, pp. 3–28.
2. Moreno-Torres, J.; Raeder, T.; Alaiz-Rodriguez, R.; Chawla, N.; Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **2012**, *45*, 521–530. [CrossRef]
3. He, H.; Yang, Y.; Wang, H. Domain Adaptation with Factorizable Joint Shift. *arXiv* **2021**, arXiv:2203.02902. [CrossRef]
4. Saerens, M.; Latinne, P.; Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Comput.* **2001**, *14*, 21–41. [CrossRef] [PubMed]
5. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, WA, USA, 4–10 August 2001; Nebel, B., Ed.; Morgan Kaufmann: San Francisco, CA, USA, 2001; pp. 973–978.
6. Billingsley, P. *Probability and Measure*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1986.
7. Klenke, A. *Probability Theory: A Comprehensive Course*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
8. Tasche, D. Class Prior Estimation under Covariate Shift: No Problem? *arXiv* **2022**, arXiv:2206.02449. [CrossRef]
9. Holzmann, H.; Eulert, M. The role of the information set for forecasting—With applications to risk management. *Ann. Appl. Stat.* **2014**, *8*, 595–621. [CrossRef]
10. Johansson, F.; Sontag, D.; Ranganath, R. Support and Invertibility in Domain-Invariant Representations. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Okinawa, Japan, 16–18 April 2019; Chaudhuri, K., Sugiyama, M., Eds.; Volume 89, pp. 527–536.
11. Scott, C. A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation. In Proceedings of the Machine Learning Research, 30th International Conference on Algorithmic Learning Theory, Chicago, IL, USA, 22–24 March 2019; Volume 98, pp. 1–24.
12. González, P.; Castaño, A.; Chawla, N.; Coz, J.D. A Review on Quantification Learning. *ACM Comput. Surv.* **2017**, *50*, 74:1–74:40. [CrossRef]
13. Tasche, D. The art of probability-of-default curve calibration. *J. Credit. Risk* **2013**, *9*, 63–103. [CrossRef]
14. Du Plessis, M.; Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Netw.* **2014**, *50*, 110–119. [CrossRef] [PubMed]
15. Tasche, D. The Law of Total Odds. *arXiv* **2013**, arXiv:1312.0365. [CrossRef]
16. Zhang, K.; Schölkopf, B.; Muandet, K.; Wang, Z. Domain Adaptation Under Target and Conditional Shift. In Proceedings of the 30th International Conference on International Conference on Machine Learning—Volume 28, ICML'13, Atlanta, GA, USA, 17–19 June 2013; pp. III-819–III-827.

17. Hofer, V.; Krempl, G. Drift mining in data: A framework for addressing drift in classification. *Comput. Stat. Data Anal.* **2013**, *57*, 377–391. [CrossRef]
18. Lipton, Z.; Wang, Y.X.; Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; Volume 80, pp. 3122–3130.
19. Tasche, D. Calibrating sufficiently. *Statistics* **2021**, *55*, 1356–1386. [CrossRef]
20. Devroye, L.; Györfi, L.; Lugosi, G. *A Probabilistic Theory of Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 1996.
21. Tachet des Combes, R.; Zhao, H.; Wang, Y.X.; Gordon, G. Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 19276–19289.
22. Tasche, D. Fisher Consistency for Prior Probability Shift. *J. Mach. Learn. Res.* **2017**, *18*, 1–32.
23. Hein, M. Binary Classification under Sample Selection Bias. In *Dataset Shift in Machine Learning*; Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N., Eds.; The MIT Press: Cambridge, MA, USA, 2009; Chapter 3, pp. 41–64.
24. Zadrozny, B. Learning and Evaluating Classifiers under Sample Selection Bias. In Proceedings of the Twenty-First International Conference on Machine Learning, ICML'04, Banff, AB, Canada, 4–8 July 2004; Association for Computing Machinery: New York, NY, USA, 2004.
25. Klebaner, F. *Introduction to Stochastic Calculus with Applications*, 2nd ed.; Imperial College Press: London, UK, 2005.