Received: 31 May 2023 | Revised: 16 August 2023 | Accepted: 24 August 2023

DOI: 10.1002/sam.11643

WILEY

RESEARCH ARTICLE

Stratified learning: A general-purpose statistical method for improved learning under covariate shift

Maximilian Autenrieth[1] | **David A. van Dyk**[1] | **Roberto Trotta**[2,3,4] | **David C. Stenning**[5]

[1]Department of Mathematics, Imperial College London, London, UK

[2]Department of Physics, SISSA, Trieste, Italy

[3]Department of Physics, Imperial College London, London, UK

[4]Centro Nazionale "High Performance Computer Big Data and Quantum Computing", Italy

[5]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia Canada

Correspondence
Maximilian Autenrieth, Department of Mathematics, Imperial College London, London, UK.
Email: m.autenrieth19@imperial.ac.uk

Funding information
UK Engineering and Physical Sciences Research Council, Grant/Award Number: EP/W015080/1; STFC, Grant/Award Numbers: ST/T000791/1, ST/P000762/1; Next Generation EU, Grant/Award Number: (DM1555del11.10.22); Fondazione ICSC, Spoke 3, Grant/Award Number: CN00000013; MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19).."; Natural Sciences and Engineering Research Council of Canada (NSERC), Grant/Award Number: RGPIN-2021-03985; Marie-Skodowska-Curie RISE, Grant/Award Number: H2020-MSCA-RISE-2019-873089

Abstract

We propose a simple, statistically principled, and theoretically justified method to improve supervised learning when the training set is not representative, a situation known as covariate shift. We build upon a well-established methodology in causal inference and show that the effects of covariate shift can be reduced or eliminated by conditioning on propensity scores. In practice, this is achieved by fitting learners within strata constructed by partitioning the data based on the estimated propensity scores, leading to approximately balanced covariates and much-improved target prediction. We refer to the overall method as Stratified Learning, or *StratLearn*. We demonstrate the effectiveness of this general-purpose method on two contemporary research questions in cosmology, outperforming state-of-the-art importance weighting methods. We obtain the best-reported AUC (0.958) on the updated "Supernovae photometric classification challenge," and we improve upon existing conditional density estimation of galaxy redshift from Sloan Digital Sky Survey (SDSS) data.

KEYWORDS

astrostatistics, bias reduction, domain adaptation, machine learning, propensity scores

1 | INTRODUCTION

In supervised learning, a model is fit to a dataset consisting of covariates and outcomes and used to predict the outcome variables in a second dataset where only the covariates are observed. Domain shift refers to the situation where the initial data used to fit the model are systematically different from the second data set, whose outcome variables are predicted with the model. In this case, which occurs commonly in many situations, the learning model is unlikely to generalize well, leading to unreliable pre-

David A. van Dyk, Roberto Trotta, and David C. Stenning contributed equally to this study.

dictions. This work proposes a new, general method to address this problem. Following standard nomenclature, we refer to the partially observed outcome variables as labels, to the labeled data used to fit the learning model as the training or source data, and to the unlabeled data as the target data.

Domain adaptation methods aim to obtain accurate target predictions under domain shift [38], arising in applications such as in medical imaging, where mechanical configurations may vary between medical centers [22]; natural language processing, where annotated training data are often highly specialized and thus different from the target data [21]; robotics and computer vision, where simulated and observed data are often combined to improve classification performance on the unlabeled target data [12]; in cyber security; epidemiology; astronomy; among others. A variety of methods designed to tackle domain shift have been proposed. Following Ref. [27], these methods can be organized into three categories: covariate-based methods [e.g., 37]; inference-based methods [e.g., 31]; and sample-based approaches, with a focus on importance weighting [e.g., 11, 51, 55], mainly in the covariate shift framework, which is our focus (Section S1 of the Supplement includes additional background and literature review.)

1.1 | Covariate shift

In this paper, we present a new method to address covariate shift, a common case of domain shift, characterized by the fact that the conditional distribution of the labels given the predictive covariates is the same for source and target data, but the distribution of source and target covariates differ [36]. In the world of machine learning, with its high-dimensional covariate spaces, the problem is both widespread and often difficult to diagnose [27, 36].

Covariate shift usually occurs when the training sample is not selected at random, but is biased in terms of certain covariates. For instance, brighter astronomical objects are more likely to be better observed and therefore selected in the training set [29, 43]. Selection bias has been widely studied in the statistical literature [14, 30], for example, when estimating treatment effects via observational studies, where the treatment assignment is often not random but is biased with regards to certain covariates. The transfer of causal inference techniques to domain adaptation has received more attention in recent years—both fields share the goal of obtaining accurate estimators under distribution shift [35]. We build on this work in the present paper, via the transfer of the propensity score framework from causal inference to the covariate shift setting.

1.2 | Propensity scores

The introduction of propensity scores [47] was groundbreaking in causal inference for obtaining unbiased treatment effect estimates from confounded, observational data. Previous study [47] defines the propensity score as the probability of treatment assignment given the observed covariates. They show that, under certain assumptions, conditional on the propensity scores, the treatment and control group have balanced covariates, which allows unbiased treatment effect estimation. Four main methods are used to condition on the propensity scores: inverse probability of treatment weighting (IPTW); using the propensity score for covariate adjustment; matching; and stratification on the propensity scores [47, 48]. Extensive work has been done on best practice and generalization of propensity scores in causal inference, too much to list here, and we refer to [18] for an overview. Propensity scores have also found wide application to related areas, such as classification with imbalanced classes [46], or fairness-aware machine learning [7], among others. In the covariate shift framework, estimated propensity scores are used only implicitly for importance weighting [e.g., 23, 56] analogously to IPTW, and for matching to obtain validation data [8]. There has been no effort, however, to transfer the general methodology.

In the causal inference literature, there is an ongoing debate on the relative merits of using weighted estimators versus stratification or matching [e.g., 4, 34, 49]. While, under correct specification of the propensity score model, weighting leads to consistent estimation of treatment effects, this may not hold for stratification due to potential residual confounding within strata [34]. However, the bias introduced by stratified estimators is traded with reduced variance compared with weighted estimators [34]. In addition, estimates of the propensity score are less variable than estimates of their reciprocal, or of a density ratio (to form weights). A small change in an estimated propensity score that is near zero can lead to a large difference in the computed inverse-propensity weight, causing massive variance in the estimates based on the weights [4, 34].

1.3 | Contribution

We propose stratified learning (*StratLearn*), a simple and statistically principled framework to improve learning under covariate shift, based on propensity score stratification. On the theory side, we show that conditioning on estimated propensity scores eliminates the effects of covariate shift. In practice, we partition (stratify) the (source and target) data into subgroups based on the estimated propensity scores, improving covariate

balance within strata. We show that supervised learning models can then be optimized within source strata without further adjustment for covariate shift, leading to reduced bias in the predictions for each stratum. *StratLearn* is general-purpose, meaning it is in principle applicable to any supervised regression or classification task under any model. We provide theoretical evidence for the effectiveness of *StratLearn* and demonstrate it in a range of low- and high-dimensional applications. We show that the principled transfer of the propensity score methodology from causal inference to the covariate shift framework allows the statistical learning community to employ hard-won practical advice from causal inference, for example, balance diagnostics and propensity score model assessment/selection. [e.g., 2, 17, 41, 48]. We stratify to condition on propensity scores instead of using importance weighting to avoid the massive variance sometimes associated with the latter. *StratLearn* (stratification) does not use individually estimated propensity score values except to form strata, leading to a more robust method [49], as demonstrated in our numerical studies.

This article is organized as follows. In Section 2, we formally introduce the target risk minimization task and summarize the related literature on covariate shift, particularly on importance weighting methods, concluding with an overview of propensity scores in causal inference. We develop our new methodology in Section 3. In Section 4, we evaluate our method on numerical examples. In Section 4.2, we apply *StratLearn* to an astronomical problem that has attracted broad interest in recent years, namely, Type Ia supernovae (SNIa) classification [6, 25]. Improving upon [43], *StratLearn* obtains the best-reported AUC[1] (0.958) on the updated "Supernovae photometric classification challenge" (SPCC) [24]. In Section 4.3, we improve upon non-parametric full-conditional density estimation of galaxy photometric redshift (so-called photo-z estimation) [20], a key quantity in cosmology. We conclude by summarizing and discussing our work in Section 5.

Supplemental Materials (numbers appearing with a prepended 'S') provide further information: a bibliographic note, Section S1, gives additional background; Section S2 provides further details on our proposed methodology; Section S3 delves into related methods, while data and software used in this paper are given in Section S4. Section S5 illustrates *StratLearn* on simulated data from a well-known univariate toy regression example [51], which might serve as a helpful demonstration of the *StratLearn* framework, particularly for readers less familiar with the topic. The numerical results in Section 4 are further expanded upon in Sections S6 (additional numerical evidence using data from the UCI repository [13]), S7 (SNIa classification), S8 (a variation of the SPCC data [25]), and S9 (photo-z regression).

StratLearn is computationally efficient, easy to implement, and readily adaptable to various applications. Our investigations show that *StratLearn* is competitive with state-of-the-art importance weighting methods in lower dimensions and greatly beneficial for higher-dimensional applications.

2 | PRELIMINARIES

2.1 | Notation

Adapting a model trained on unrepresentative source data to accurately predict the labels of the target data requires information about how the distributions of the source and target data differ. Let $\mathcal{X} \subset \mathbb{R}^F$, $F > 0$, be the covariate space that is observed for all data, and let $\mathcal{Y}$ be the label space, observed only for the source data. The labels typically consist of $K > 1$ classes or, in multivariate regression with K dependent variables, a subset of $\mathbb{R}^K$. Different domains are defined as different joint distributions $p(x, y)$ over the same joint covariate-label space $\mathcal{X} \times \mathcal{Y}$ [27]. Let $D_S = \left\{ \left(x_S^{(i)}, y_S^{(i)} \right) \right\}_{i=1}^{n_s}$ denote the source data, a sample of size n_s from the joint distribution p_S, and let $D_T = \left\{ x_T^{(i)} \right\}_{i=1}^{n_t}$ denote the target data, an unlabelled sample of size n_t from the distribution p_T. To avoid the trivial case, we assume that $p_S(x, y) \neq p_T(x, y)$, which means the joint distribution of the covariates x and the outcome labels y differs for the source and target data. For ease of notation, we implicitly condition on an indicator variable S, with $p_S(x, y) := p(x, y|s = 1)$ representing source data (analogously $p_T(x, y) := p(x, y|s = 0)$ for target data).

Covariate shift refers to the case where the conditional distribution of the labels given the covariates is the same for all data, but the marginal distribution of the covariates differs between the source and target data. In this notation, covariate shift corresponds to $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$. Thus, if an object in the source data has the same set of covariates as an object in the target data, then the conditional distribution of the outcomes of both objects is the same (i.e., $p_S(y|x) = p_T(y|x)$). However, the distribution of covariates for objects in the source data is systematically different to the distribution of covariates for objects in the target data (i.e., $p_S(x) \neq p_T(x)$).

[1] The AUC is the area under the Receiver Operator Characteristic (ROC) curve, obtained by plotting classifier efficiency against the false positive rate for different classification thresholds (between [0, 1]).

2.2 | Target risk minimization

In a supervised learning task, let $f : \mathcal{X} \to \mathbb{R}^K$ be the training function, and $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, \infty)$ be the loss function comparing the output of f with the true label, y_T in $\mathcal{Y}$. (This describes a general multivariate regression case; in a probabilistic classification task with K classes, we usually have $f : \mathcal{X} \to [0,1]^K$). The risk function associated with our supervised learning task is $\mathcal{R}(f) := \mathbb{E}[\ell(f(x), y)]$. We cannot generally compute $\mathcal{R}(f)$, since the exact joint distribution $p(x, y)$ is unknown. However, an approximation of the risk can be obtained by computing the empirical risk by averaging the loss on the training sample, where we have access to both the covariates x_S and the labels y_S.

In the covariate shift setting, the objective is to minimize the target risk $\mathcal{R}_T(f) := \mathbb{E}_{p_T(x,y)}[\ell(f(x), y)]$, via the labeled source data D_S and unlabeled target data D_T, assuming that $p_S(x, y) \neq p_T(x, y)$, with $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$. More precisely, our task is to train a model function f that minimizes $\mathcal{R}_T(f)$, being able to compute only the source loss $\ell(f(x_S), y_S)$, but not the target loss $\ell(f(x_T), y_T)$, since y_T is unavailable in practice. Section 2.3 reviews importance weighting methods to minimize the target risk under covariate shift.

2.3 | Related literature—Importance weighting

In an influential work, the author of Ref. [51] proposes a weighted maximum likelihood estimation (MWLE) and shows that this MWLE converges in probability to the minimizer of the target risk under covariate shift. Following Ref. [51], assuming that the support of $p_T(x)$ is contained in $p_S(x)$, the expected loss (risk) w.r.t. $p_T(x, y)$ equals that w.r.t. $p_S(x, y)$ with weights $w(x) := p_T(x)/p_S(x)$ for the loss incurred by each x:

$$\mathbb{E}_{p_T(x,y)}[\ell(f(x), y)] = \mathbb{E}_{p_S(x,y)}[w(x)\ell(f(x), y)]. \quad (1)$$

In short, the target risk can be minimized by weighting the source domain loss by a ratio of the densities of target and source domain covariates. The importance weights $w(x)$ are paramount in the covariate shift literature and several approaches optimize the estimation of the weights. One approach estimates the densities $p_T(x)$ and $p_S(x)$ separately [51], for example, through kernel density estimators [54]. Others estimate the density ratio directly, for example, via Kernel-Mean-Matching [16], Kullback–Leibler importance estimation (KLIEP) [55], and variations of unconstrained least-squares importance fitting (uLSIF) [23]. Given $w(x)$, the authors of Ref. [53] propose importance-weighted cross validation (IWCV) and show that in theory this can deliver an almost unbiased estimate of the target risk. Previous study [56] links covariate shift with selection bias and shows that the target risk can be minimized by importance sampling of source domain data, employing the inverse probability of source set assignment for importance weights. This allows any probabilistic classifier to be used to obtain the weights, for example, logistic regression [5].

Although importance weighting in theory enables minimization of the target risk, there are challenges. Based on a measure of domain dissimilarity (e.g., Rényi divergence), the authors of Ref. [11] show that weighting leads to high generalization upper error bounds, making predictions unreliable, especially with large importance weights. In addition, the authors of Ref. [42] point out that while weighting can reduce bias, it can also greatly increase variance. Unfortunately, with increasing covariate space dimension, the variance of the importance-weighted empirical risk estimates may increase sharply [19, 52]. This can be partly tackled by dimensionality reduction methods [52]; see Ref. [27] for a detailed discussion. We address these variance concerns via propensity scores.

2.4 | Related literature—Propensity scores in causal inference

Before conducting the transfer of propensity scores to the covariate shift framework in Section 3, we provide an overview of the propensity score methodology in its original causal inference framework [47]. Propensity scores are a pivotal methodology to account for selection bias in observational studies to perform treatment effect estimation. In observational studies, the treatment assignment can typically be observed, but treatment is not randomly allocated. Confounding covariates, associated with both the treatment assignment and the outcome variable, can systematically bias average treatment effect estimates. Propensity score methods aim to generate balance between the covariates of the treatment and the control groups, eliminating or mitigating bias in treatment effect estimates.

More precisely, given a set of observed covariates X and a binary indicator Z for treatment assignment (treatment vs. control), the authors of Ref. [47] introduce the propensity score as

$$e(X) := P(Z = 1|X) \quad (2)$$

and define treatment allocation Z as strongly ignorable, if

$$\text{(i)} \quad (Y_1, Y_0) \perp\!\!\!\perp Z|X \quad \text{and} \quad \text{(ii)} \quad 0 < e(X) < 1. \quad (3)$$

Condition (i) means that treatment assignment Z is conditionally independent of the potential outcome (Y_1, Y_0), given the observed covariates. The potential outcomes are the possible outcomes for an object, depending on its treatment status, and at most one is observed (e.g., for a treated object the observed outcome is $Y = Y_1$). In practice, condition (i) means that no confounders (covariates that are associated with the treatment and outcome) are unmeasured. The authors of Ref. [47] show that if (3) holds, the propensity score is a balancing score. That is, given the propensity score, the distribution of the covariates in treatment and control are the same, that is, $p(X|e(X), Z = 1) = p(X|e(x), Z = 0)$. Thus, conditional on the propensity score, unbiased average treatment effect estimates can be obtained, i.e., $\mathbb{E}[Y_1|e(x), Z = 1] - \mathbb{E}[Y_0|e(x), Z = 0] = \mathbb{E}[Y_1 - Y_0|e(x)]$. In practice, conditioning on the estimated (rather than true) propensity score can achieve better empirical balance as this corrects for statistical fluctuations in the sample as well [15, 47].

Below, we show how the balancing property of propensity scores can be employed to transfer the propensity score methodology to the covariate shift framework, for target risk minimization in supervised learning tasks.

3 | A NEW METHOD: STRATLEARN

In this section, we introduce our novel *StratLearn* methodology for principled learning under covariate shift via the transfer of the propensity score framework from causal inference to the covariate shift setting. In Section 3.1, we provide theoretical justification that conditioning on propensity scores eliminates covariate shift, and we devise how propensity scores can be employed in practice to mitigate the effects of covariate shift via propensity score stratification. The flowchart in Figure 1 outlines the *StratLearn* framework, with technical details described in Section 3.2. In Section 3.3, we describe covariate balance diagnostics which can readily be transferred from the causal inference framework to assess the propensity score model, and we introduce outcome balance diagnostics, employing additional structure in the covariate shift setting.

3.1 | StratLearn—Methodology

In the covariate shift framework, we define the propensity score to be the probability that object i is in the source data, given its observed covariates, that is,

$$e(x_i) := P(s_i = 1|x_i), \quad \text{with} \quad 0 < e(x_i) < 1. \tag{4}$$

Proposition 1 (Learning conditional on the propensity score)**.** *If $p_S(x, y)$ and $p_T(x, y)$ satisfy the covariate shift definition and $0 < e(x) < 1$, then it holds that*

$$p_T(x, y|e(x)) = p_S(x, y|e(x)). \tag{5}$$

That is, conditional on $e(x)$, the joint source and target distributions are the same, eliminating covariate shift. It follows, for any loss function $\ell = \ell(f(x), y)$,

$$\mathbb{E}_{p_T(x,y|e(x))}[\ell(f(x), y)] = \mathbb{E}_{p_S(x,y|e(x))}[\ell(f(x), y)]. \tag{6}$$

Proposition 1 is verified in Section S2. Note that its condition, $0 < e(x) < 1$, is no stronger than the conditions required for (1). The support of $p_T(x)$ being contained in $p_S(x)$ implies $0 < e(x)$, and $e(x) = 1$ implies $p_T(x) = 0$, in which case the importance weight $w(x) = 0$, which is equivalent to discarding the sample.

With Proposition 1, we extend the basic causal inference theory to use propensity scores in the covariate shift framework. Conditioning on estimated propensity scores enables statistically principled minimization of the target risk based on source data. According to Proposition 1, if we were to condition on any single value of the propensity score, the distribution of x and y in the source and target domains would be identical and we could minimize their target risk using the source data alone. Because sample sizes with identical propensity scores are too small in practice for model fitting, we employ an approximation.

StratLearn takes advantage of Proposition 1 via propensity score stratification; source data D_S and target data D_T are divided into k non-overlapping subgroups (strata) based on quantiles of the estimated propensity scores. More precisely, letting q_j be the jth k-quantile of $\{e(x_i) : x_i \in (x_S \cup x_T)\}$, for $j \in 1, \dots, k$, we divide D_S and D_T into

$$\begin{aligned} D_{Sj}^{(k)} &= \left\{(x, y) \in D_S : q_{k-j} < e(x) \leq q_{k-j+1}\right\} \text{ and} \\ D_{Tj}^{(k)} &= \left\{x \in D_T : q_{k-j} < e(x) \leq q_{k-j+1}\right\}, \end{aligned} \tag{7}$$

where $q_0 = 0$ and $q_k = 1$. By Proposition 1, within strata,

$$p_{T_j}(y, x) \approx p_{S_j}(y, x), \quad \text{for } j \in 1, \dots, k, \tag{8}$$

where S_j indicates conditioning on assignment to the jth source stratum (analogously for target T_j). It follows that for $j \in 1, \dots, k$,

$$\mathbb{E}_{p_{T_j}(x,y)}[\ell(f(x), y)] \approx \mathbb{E}_{p_{S_j}(x,y)}[\ell(f(x), y)]. \tag{9}$$

Thus, we can minimize the target risk within strata by minimizing the source risk within strata. In this way, we

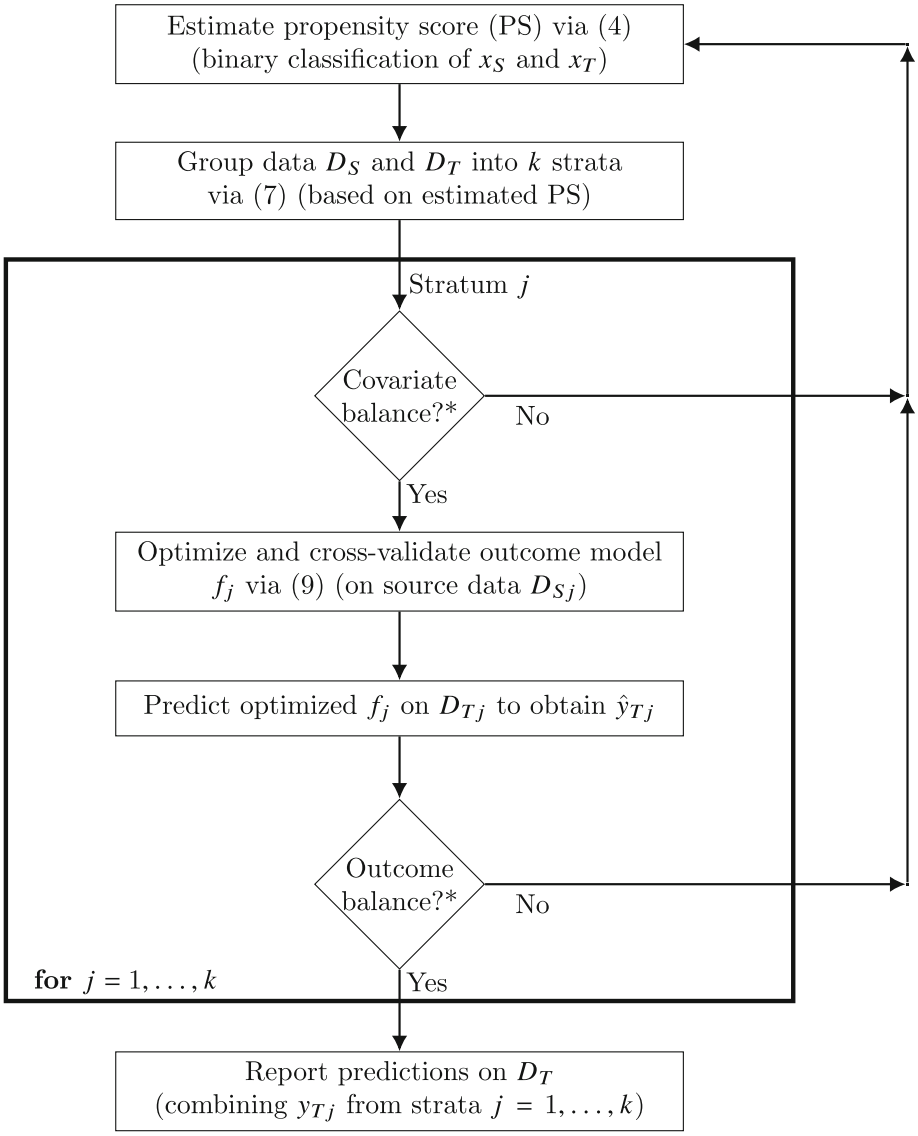

FIGURE 1 *StratLearn* flow chart (*Covariate balance and outcome balance is assessed as described in Section 3.3, with a numerical example given in Section 4.2.).

reduce the covariate shift problem to non-overlapping subgroups where the source and target domain are approximately the same, which in principle allows us to fit any supervised learner to D_{S_j} to predict the target objects in D_{T_j}. Figure 1 presents a flow chart illustrating the steps of our proposed *StratLearn* methodology.

3.2 | StratLearn—Technical details

In general, any probabilistic classifier could be used to estimate propensity scores (e.g., [41]). Logistic regression is commonly used in causal inference, and we adopt it for the applications in this paper. In practice, the covariate shift assumption, $p_S(y|x) = p_T(y|x)$, requires there be no unobserved confounding covariates. To meet this requirement, we include all potential confounders as main effects. An estimate of the propensity scores in (4) is then obtained by probabilistic classification of source assignment, given source data x_S and target data x_T.

Using the estimated propensity scores, the source and target data are grouped into strata, following (7). We use $k = 5$ strata based on empirical evidence provided by Ref. [9], showing that sub-grouping into five strata is enough to remove at least 90% of the bias for many continuous distributions [47]. Given the stratified data, we fit a model f_j to source data D_{S_j} and predict the respective target samples in D_{T_j}, for $j \in 1, \ldots, k$. Model hyperparameters for f_j can be selected through empirical risk minimization on source data D_{S_j}, for instance through cross validation on D_{S_j}. The model functions f_j are trained independently and can be computed in parallel to reduce computational time. If the source distribution does not cover the target distribution well enough, some of the strata may contain too little source data to reliably train the model. In this case, we add source data from one or more adjacent strata to

avoid highly variable predictions. There is a bias-variance trade-off here in that this reduction in variance requires a relaxation of the approximation in (8), which inevitably increases bias somewhat. Although a general and precise criterion for combining the strata is elusive (more complex models require more data and data sets of the same size may be more or less informative for the same model), we illustrate the combination of source strata in Section 4.2 (and Section S6), where one or more source strata have insufficient data.

3.3 | StratLearn—Balance diagnostics

A key advantage of propensity scores derived in causal inference is their covariate balancing score property [47], that is, $p_S(x|e(x)) = p_T(x|e(x))$. In causal inference, this property is used to verify the propensity score model and/or the choice of covariates, x, for example, by checking that x has the same within-strata distribution in the treatment and control groups. Employing the balancing property in the derivation of Proposition 1 allows us to take advantage of such diagnostic tools in our framework. We refer to the large literature on this [e.g., 1, 2, 17, 48] and provide an example of such a balance check in Section 4.2 (and in Sections S8 and S9).

In Remark 1, we detail how additional structure in the covariate shift setting can be exploited to justify a corollary model diagnostic.

> *Remark* 1. *In the propensity score framework of causal inference* [47]*, we have potential outcomes* Y_0 *and* Y_1*. In the covariate shift framework, the potential outcomes are identical* ($Y_0 \equiv Y_1$)*. That is, there is no "treatment effect" from being assigned to the source or target set, though only the source data are observed* ($Y_1 \equiv Y$)*. Now, given the propensity score* $e(x)$*, with* $0 < e(x) < 1$*, and the covariate shift condition* $p(y|x, s = 1) = p(y|x, s = 0)$*, source data assignment is 'strongly ignorable' (using the terminology of Ref*. [47]*). It follows through Theorem 4 in Ref*. [47] *that, conditional on the propensity score, source and target outcome are the same in expectation.*

In cases where labels are observed for (part of) the target group, we can use Remark 1 as a model diagnostic. Although in practice the labels are mostly unobserved in the target group, they are available in our real-world scientific/experimental settings described in Sections 4.2 and 4.3. In Section 4.2 (as well as Sections S8, S6, and S9), we use Remark 1 to demonstrate a reduction of within-strata covariate shift (i.e., by conditioning on the propensity score).

We further demonstrate the possibility of similarly using predicted labels instead of actual labels as a model diagnostic. While the actual target labels y_T are usually not available in real-world data applications, the distribution of the model predicted outcome labels ($\widetilde{y} = f(x)$) can be evaluated for source $f(x_S)$ and target $f(x_T)$. With f being a measurable function of the covariates x, and by employing the balancing property of propensity scores, it holds $p_S(f(x)|e(x)) = p_T(f(x)|e(x))$. Consequently, a discrepancy between the distributions of predicted source outcome $f(x_S)$ and predicted target outcome $f(x_T)$ is an indication of residual (covariate) shift in the source and target distribution.[2] An advantage of assessing the balance in the predicted outcome $f(x)$ (in addition to covariate balance) is that $f(x)$ is designed to approximate the outcome y. Thus, a discrepancy of $f(x_S)$ and $f(x_T)$ indicates an imbalance in strongly predictive covariates, a straightforward sign for remaining (likely), concerning confounding. We demonstrate the application of balance diagnostics via predicted labels in Section 4.2.

4 | NUMERICAL DEMONSTRATIONS

This section provides numerical evaluation of *StratLearn* with a comparison with state-of-the-art importance weighting methods on two topical scientific questions in cosmology. We first introduce the comparison methods in Section 4.1. In Section 4.2, we demonstrate the benefit of *StratLearn* on a classification task with a large number of covariates. We further illustrate how balance diagnostics can be employed in practice to detect potentially remaining confounding. In Section 4.3, we demonstrate how the *StratLearn* framework can improve conditional density estimation under covariate shift, investigating the effect of increasing (noisy) covariate dimensions on the performance of *StratLearn* and comparison methods.

4.1 | Comparison methods

We compare *StratLearn* to a range of well-established importance weighting methods.

- KLIEP – Kullback–Leibler importance estimation procedure [55].

[2] In practice, one has to ensure that f is not overfitted on the available training (source) data sample, a standard check in supervised learning, which can readily be assessed via standard validation tools, such as cross-validation or bootstrapping of source data.

- uLSIF – Unconstrained least-squares importance fitting [23].[3]
- NN – Several versions of the nearest-neighbor importance weight estimator [28, 29, 33], varying the number of neighbors.
- IPS – Importance weight estimation through probabilistic classification of source set assignment [23].

In Section 4.3, we incorporate the estimated weights as in the corresponding benchmark publication. Following Ref. [20], the estimated weights are used for loss weighting as in (Section 2.1). In Section 4.2, importance weighting has not previously been applied. We implement IWCV, importance sampling, and a combination of both, to demonstrate the advantage of *StratLearn* with respect to either; see Section S3.

4.2 | Classification—SNIa identification

4.2.1 | Objective

Type Ia supernovae (SNIa) are invaluable for the study of the accelerated expansion history of the universe [e.g., 40, 45]. SNIa are exploding stars that can be seen at cosmological distances (billions of light years away), occurring in a particular physical scenario which causes their intrinsic luminosity to correlate with observable properties of their light curve (apparent brightness at Earth as a function of time). This "standardizable candle" property of SNIa makes it possible to measure their distance, which in turn depends on parameters describing the physical contents of the universe.

To take advantage of this, reliable identification of SNIa based on photometric light curve (LC) data is a major challenge in modern observational cosmology. Photometric LC data are easily collectable, consisting of measurements of an astronomical object's apparent brightness (i.e., flux), filtered through different passbands (wavelength ranges), at a sequence of time points (as illustrated in Figure 2). Only a small subset of the objects are labeled via expensive and time-consuming spectroscopical observations, which enable SNIa identification thanks to its characteristic spectral lines. The labeled source data, D_S, are therefore not representative of the photometric target data, D_T, as the selection of spectroscopic source samples is biased towards brighter and bluer objects. The automatic classification of unlabeled objects, based on biased spectroscopically confirmed source data, is the subject of much research, including public classification challenges [25, 26].

[3] KLIEP and uLSIF were implemented with the original author's public domain MATLAB code (link).

Leading SNIa classification approaches are based on data augmentation; they sample synthetic objects from Gaussian process (GP) fits of the LCs to overcome covariate shift [6, 43]. The method of Ref. [43] can be viewed as a prototype of *StratLearn*, as it augments the source data separately in strata based on the estimated propensity scores. However, to optimize data augmentation within strata, [43] requires a subsample of labeled target data that are unavailable in practice. While effective in this particular case, GP data augmentation is not an option in most covariate shift tasks. We show that *StratLearn* makes augmentation unnecessary. We use target prediction AUC to compare performance to published results.

4.2.2 | Data and preprocessing

We use data from the updated "Supernova photometric classification challenge" (SPCC) [24], containing a total of 21,318 simulated SNIa and of other types (Ib, Ic, and II). For each supernova (SN), LCs are given in four color bands, $\{g, r, i, z\}$. The data include a source set D_S of 1102 spectroscopically confirmed SNe with known types and 20,216 SNe with unknown types (target set D_T). Notably, 51% of the source objects are SNIa, while only 23% of the target date are SNIa, a consequence of the strong covariate shift in the data.

We follow the approach in Ref. [43], which was applied to an earlier release of the SPCC data [25, discussed in Section S8], to extract a set of covariates from the LC data that can be used for classification. First, a GP with a squared exponential kernel is used to model the LCs. Then, a diffusion map [10] (as used in Ref. [44]) is applied, resulting in a vector of 100 similarity measures between the LCs that we use as predictor variables. Combining these with redshift (a measure of cosmological distance, defined in Section 4.3) and a measure of overall brightness, we obtain 102 predictive covariates.

4.2.3 | Results

To evaluate the impact of covariate shift on classification, we first consider a "biased fit" by training a random forest classifier (as in Ref. [43]) on the source covariates ignoring covariate shift, resulting in an AUC of 0.902 on the target data (black ROC curve in Figure 3). Next, we obtain a "gold standard" benchmark by randomly selecting 1102 objects from target data as a representative source set. The same classification procedure with the unbiased "gold standard" training data (unavailable in practice) yields an AUC of 0.972 on the remaining 19,114 target objects.

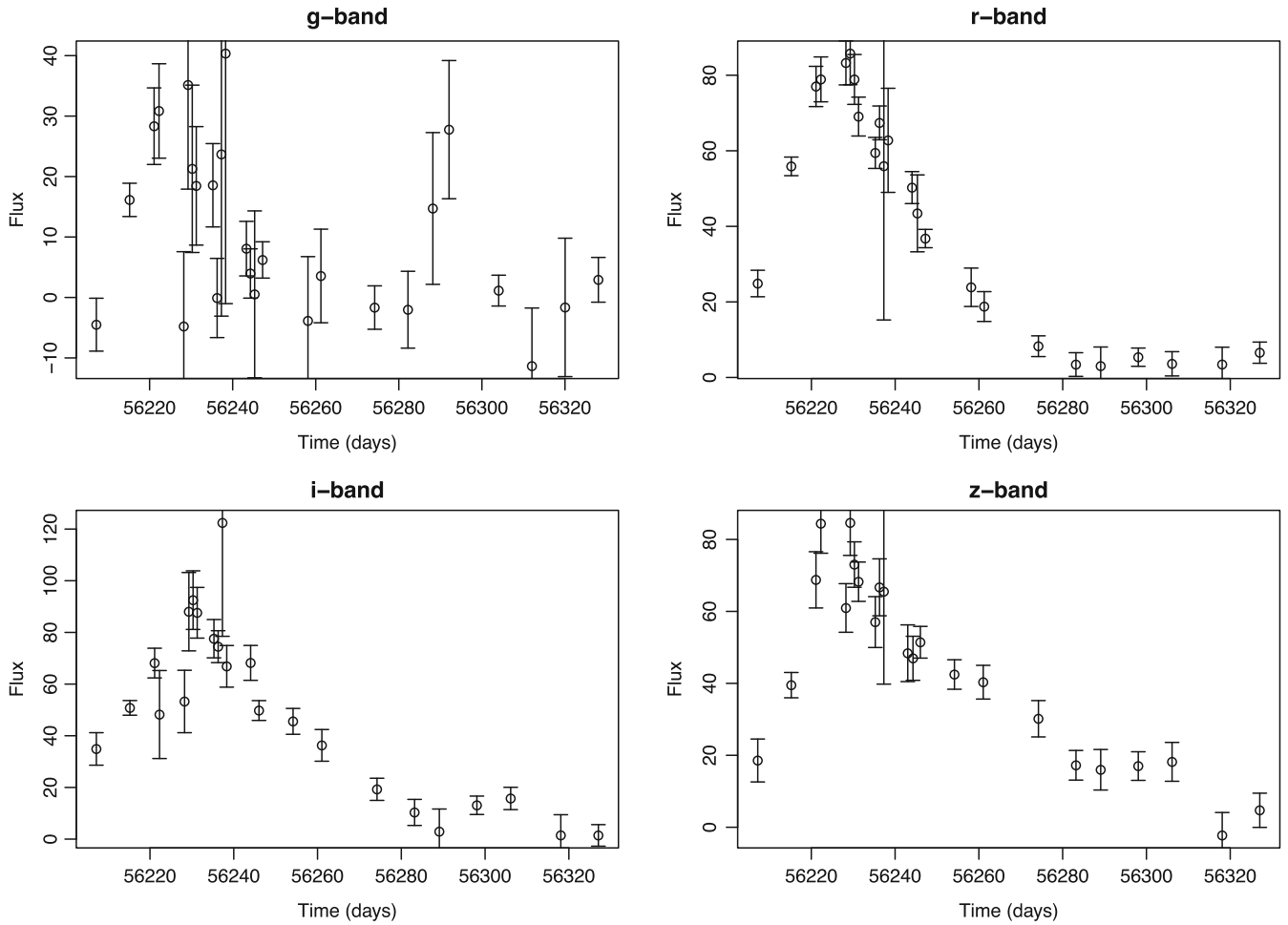

FIGURE 2 Example of photometric LC data, including 1σ error bars, for a typical SNIa (specifically, SN2475 from the updated [24] simulated SPCC data).

Given the biased source data, *StratLearn* is implemented as described in Section 3, including all 102 covariates in the logistic propensity score estimation model. After stratification, a random forest classifier is trained and optimized on source strata D_{S_1} and D_{S_2} separately to predict samples in target strata D_{T_1} and D_{T_2}. We use repeated 10-fold cross validation with a large hyperparameter grid to minimize the empirical risk of (9) within each strata, employing log loss[4] as our loss function; details appear in Section S7. Source strata D_{S_j} for $j \in \{3,4,5\}$ have a small sample size (13,7,4), respectively. Thus, source strata D_{S_j} for $j \in \{3,4,5\}$ are merged with D_{S_2} to train the random forest to predict D_{T_j} for $j \in \{3,4,5\}$. With *StratLearn*, we obtain an AUC of 0.958 on the target data (blue ROC curve in Figure 3), very near the optimal "gold standard" benchmark.

[4]The log-loss (also referred to as cross-entropy loss) compares the output of a classification $f(x) \in [0,1]$ with the true output y for an observation (x,y) via $\ell_{\text{logloss}}(f(x),y) := -(y\log(f(x)) + (1-y)\log(1-f(x)))$.

Figure 3 compares *StratLearn* to importance sampling methods designed to adjust for covariate shift. For importance sampling, the bootstrapped samples in the random forest fit were resampled with probabilities proportional to the estimated importance weights (see Section S3). NN and IPS led to the best importance weighted classifier (AUC = 0.923, 0.921)—an improvement over the biased fit, but substantially lower than *StratLearn*. AUC standard errors (see Figure 3) are small relative to the large performance improvement of *StratLearn*. KLIEP failed to fit importance weights and is thus not included in the results. We also implemented IWCV using the same hyperparameter grid as for *StratLearn*, and a combination of IWCV and importance sampling, which both led to lower AUC than the ones reported in Figure 3 (see Section S7).

Previous state-of-the-art methods report an AUC of 0.855 [32] using boosted decision trees, 0.939 [39] using a framework of an autoencoder and a convolutional neural network, and 0.94 [43] using LC augmentation and target data leakage, all lower than *StratLearn*.

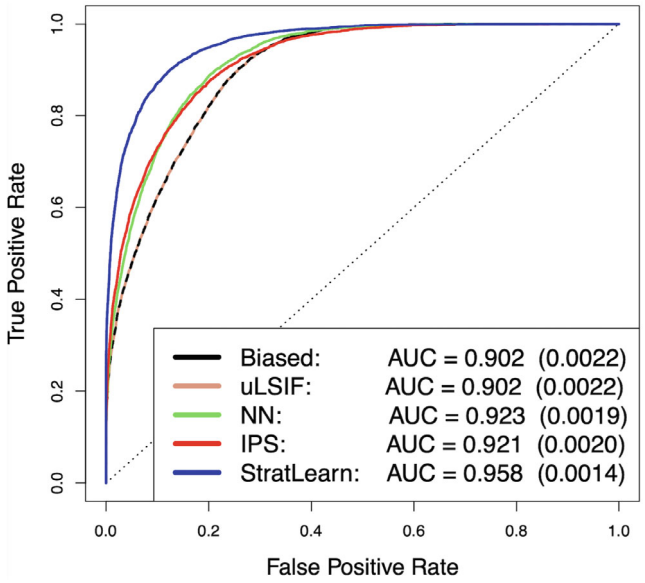

FIGURE 3 Comparison of ROC curves for SNIa classification using the updated SPCC data. Here, Biased and uLSIF are identical. Bootstrap AUC standard errors (from 400 bootstrap samples) are given in parentheses.

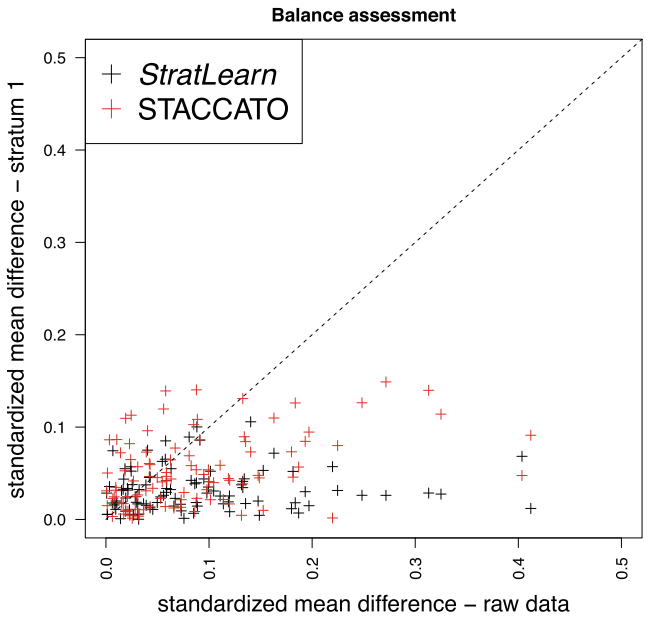

FIGURE 4 Absolute standardized mean differences between source and target data of stratum 1 plotted against "raw" data absolute standardized mean differences for *StratLearn* and STACCATO.

4.2.4 | Balance assessment on updated SPCC data

To illustrate the balancing property of propensity scores (see Section 3.3) and its effect on predictive target performance, we assess the covariate balance in the updated SPCC data within strata conditional on the estimated propensity scores, by means of two commonly used balance measures: absolute standardized mean differences (SMD) and the Kolmogorov–Smirnov test statistics (KS-stats) [1, 3].

Figure 4 provides a detailed covariate balance comparison, by plotting the "raw" SMD against the *StratLearn* SMD in stratum 1 (black) for each covariate. We remove two outliers (redshift and brightness) with very large "raw" SMD (1.1 and 1.7), because including them in the Figure makes it more difficult to illustrate the balance of the bulk of the covariates; both are well balanced in stratum 1 using *StratLearn* (SMD equals 0.12 and 0.17). Points below the diagonal line are better balanced in the stratum than those in the "raw" non-stratified data. This is the case for the vast majority (71%) of black points in Figure 4, illustrating the balance improvement achieved with *StratLearn*.

Figure 4 also plots (red) the SMD achieved by STACCATO [43], which uses two covariates (redshift and brightness, as opposed to the 102 used by *StratLearn*) in the logistic regression to estimate propensity scores. While STACCATO improves the balance of the majority (69%) of the covariates, most (66%) black (*StratLearn*) SMD have smaller vertical values, indicating better balance than STACCATO (red).

On average across the 102 covariates, *StratLearn* improves covariate balance compared to the "raw" non-stratified data measured by SMD by ∼70% in stratum 1 and ∼10% in stratum 2 (KS-stats: ∼70% in stratum 1 and ∼30% in stratum 2).[5] It further improves upon STACCATO by ∼36% in stratum 1 and ∼46% in stratum 2 using SMD (KS-stats: ∼24% in stratum 1 and ∼36% in stratum 2). The remaining strata contain too few source data to assess covariate balance. Details are provided in Table S5.

The improved covariate balance (reduced covariate shift) directly translates into improved predictive performance. STACCATO (including data augmentation and target data leakage) yields a target AUC of 0.94, whereas with *StratLearn*, we obtain a target AUC of 0.958 (without data augmentation and no target data leakage)—a substantial improvement resulting from the improved covariate balance by accounting for potentially confounding covariates. In general, we note that balance is particularly important for covariates that are strongly predictive for the outcome. Domain-specific expertise might be necessary to identify such covariates in the individual cases in practice. In Section S6, we demonstrate how covariate balance can be improved by adjusting the propensity score model.

[5] Percentages are calculated by taking the ratio of the average SMD (average KS-stats) of all 102 covariates.

TABLE 1 Strata composition on the updated SPCC data (Section 4.2), applying STACCATO (left) and *StratLearn* (right).

		STACCATO			StratLearn		
Stratum	**Set**	**Number of SNe**	**Number of SNIa**	**Prop. of SNIa**	**Number of SNe**	**Number of SNIa**	**Prop. of SNIa**
1	Source	924	414	**0.45**	958	518	**0.54**
	Target	3340	1125	**0.34**	3306	1790	**0.54**
2	Source	153	125	**0.82**	120	28	**0.23**
	Target	4111	973	**0.24**	4144	927	**0.22**
3 to 5	Source	25	19	0.76	24	12	0.5
	Target	12,765	2431	0.19	12,766	1812	0.14

Note: The number of SNe, as well as the number and proportion of SNIa are presented in source and target stratum 1 and 2. For conciseness, we present the combined strata 3 to 5, containing too little source data for meaningful comparison of the SNIa proportions.

TABLE 2 Outcome balance diagnostics via predicted labels on the updated SPCC data (Section 4.2), applying STACCATO (left) and *StratLearn* (right).

		STACCATO (predicted)			*StratLearn* (predicted)		
Stratum	**Set**	**Number of SNIa**	**Prop. of SNIa**	***p* value**	**Number of SNIa**	**Prop. of SNIa**	***p* value**
1	Source	414	**0.45**	8.4e-11	518	**0.54**	0.284
	Target	1106	**0.33**		1853	**0.56**	
2	Source	125	**0.82**	2.8e-13	28	**0.23**	0.749
	Target	2166	**0.53**		1040	**0.25**	

Note: The number and proportion of predicted SNIa are presented in source and target stratum 1 and 2. *p* values are computed via Fisher's exact test of independence between predicted SNIa target and source proportions within strata.

Table 1 presents the composition of the five *StratLearn* strata. Recall that according to Remark 1, conditional on the propensity score the marginal distributions of source and target outcome are the same in expectation. Table 1 shows that the proportion of SNIa in the source and target data (which in this case can be computed from knowledge of the true target labels in the simulation) align well for *StratLearn* in the first two strata, indicating the expected reduction in covariate shift. The source sample sizes in strata 3–5 are quite small, rendering meaningful comparison of the SNIa proportions impossible. In strata 1 and 2, however, *StratLearn* achieves much better balance than either STACCATO or the raw (unstratified) data (51% SNIa in source, 23% SNIa in target).

In Table 2, we demonstrate how predicted outcomes can be employed for balance diagnostics by assessing the predicted proportions of SNIa within strata obtained by STACCATO and by *StratLearn*. We compute the predicted outcomes by classifying objects to be SNIa if the (random forest) predicted SNIa probabilities are above 0.5. While STACCATO leads to a strong discrepancy between predicted SNIa proportions in the first two strata (indicating remaining confounding), *StratLearn* leads to well-matched predicted SNIa proportions. We further quantify the discrepancy by performing a two-sided Fisher's exact test of independence, with the null hypothesis that there is no association of source/target set assignment and predicted SNIa proportion. Comparing different propensity score models, a higher *p*-value is an indicator for better balance in the predicted outcomes and should thus be desirable. *StratLearn* leads to much higher *p*-values than STACCATO (failing to reject the null hypothesis for reasonable significance levels), which implies much weaker relation between source/target assignment and predicted outcomes.

In this particular example, with *StratLearn*, we fail to reject the null hypothesis for most significance levels. This may not always be the case (e.g., Section S8). We recall that the strategy of conditioning on propensity scores via stratification leads to subgroups with similar (not identical) propensity scores and thus to similar (not identical) joint distributions within strata (this is the approximation in (8)). This in turn might lead to differences in the distributions of the covariates and the (predicted) outcomes,

even if we could condition on the true propensity scores. We thus employ the p-values of (predicted) outcomes as an additional tool to asses, and primarily to compare, propensity score models to detect and reduce confounding of highly predictive and thus most relevant covariates.

4.3 | Conditional density estimates—Photo-z Regression

4.3.1 | Objective

The wavelength of light from extragalactic objects is stretched because of the expansion of the universe—a phenomenon called 'redshift'. This fractional shift towards the red end of the spectrum is denoted by z. A precise measurement of redshift allows cosmologists to estimate distances to astronomical sources, and its accurate quantification is essential for cosmological inference (e.g., redshift is a key component of the Big Bang theory). Because of instrumental limitations, redshift can be precisely measured only for a small fraction of the $\sim 10^7$ galaxies observed to date (set to grow to $\sim 10^9$ within a decade). These source data are subject to covariate shift relative to the set of galaxies with unknown redshift (target). The authors of Ref. [20] employed importance weighting to adjust for covariate shift in x, a set of observed photometric magnitudes (a logarithmic measure of passband-filtered brightness), when estimating z. They obtain a non-parametric estimate of the full conditional density, $f(z|x)$, to quantify predictive uncertainty of redshift estimates. Proper quantification of predictive uncertainties is crucial to avoid systematic errors in the scientific downstream analysis [20, 50]. Using the same setup and conditional density estimation models (hist-NN, ker-NN, Series, and Comb, detailed in Ref. [20]),[6] we show that *StratLearn* leads to better overall predictive performance than importance weighting.

Assuming that source and target data follow the same distribution, under the L^2-loss, conditional density estimators typically aim to minimize the *generalized* risk (generalized in that the underlying loss can be negative):

$$\widehat{R}(\widehat{f}) = \frac{1}{n_S}\sum_{k=1}^{n_S}\int \widehat{f}^2\left(z|x_S^{(k)}\right)dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\widehat{f}\left(z_S^{(k)}|x_S^{(k)}\right), \quad (10)$$

The authors of Ref. [20] propose to adjust for covariate shift by adapting (10), via optimizing weighted versions of the conditional density estimators [20, sections 5.1–5.3] with respect to an importance weighted *generalized* risk:

$$\widehat{R}_S(\widehat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \widehat{f}^2\left(z|x_T^{(k)}\right)dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\widehat{f}\left(z_S^{(k)}|x_S^{(k)}\right)\widehat{w}\left(x_S^{(k)}\right), \quad (11)$$

where the weights, $\widehat{w}(x_S) = p_T(x)/p_S(x)$, are estimated using the methods described in Section 4.1. As their best performing model for $f(z|x)$, the authors of Ref. [20] propose an average of importance weighted ker-NN and Series,

$$\widehat{f}^{\alpha}(z|x) = \sum_{k=1}^{p}\alpha_k\widehat{f}_k(z|x), \text{with constraints}$$
$$\text{(i)}: \ \alpha_i \geq 0, \text{and (ii)}: \ \sum_{k=1}^{p}\alpha_k = 1, \quad (12)$$

referred to as "Comb" (i.e., combination), where $p = 2$ and α_i is optimized to minimize (11).

With *StratLearn*, we optimize the unweighted conditional density estimators (hist-NN, ker-NN, Series) by minimizing (10) in each source stratum separately (accounting for covariate shift following Proposition 1). We also propose a *StratLearn* version of Comb by optimizing (12) on each source stratum separately (via the generalized risk in (11) with $w(x) \equiv 1$), including ker-NN and Series (each optimized via *StratLearn* beforehand). *StratLearn* and the other methods are compared with a 'Biased' (unweighted) method that simply optimizes (10). We abbreviate the combination of each method (*StratLearn*, Biased, and each of the weighting methods in Section 4.1) with the models for $f(z|x)$ (hist-NN, ker-NN, Series and Comb) as $\text{Method}_{\text{Model}}$.

4.3.2 | Data

We use the same data as Ref. [20], consisting of 467,710 galaxies from Ref. [50], each with spectroscopic redshift z (measured with negligible error), and five photometric covariates x. As in Ref. [20], we use the r-band magnitude and the four colors (differences of magnitude values in adjacent photometric bands) as our covariates. We denote this spectroscopic source sample by D_S. To simulate realistic covariate shift, we follow Ref. [20]: starting from D_S, we use rejection sampling to simulate a photometric, unrepresentative target sample D_T, with the prescription $p(s = 0|x) = f_{B(13,4)}\left(x_{(r)}\right) / \max_{x_{(r)}} f_{B(13,4)}\left(x_{(r)}\right)$, where $x_{(r)}$ is the r-band magnitude and $f_{B(13,4)}$ is a beta density with parameters (13,4). Additionally, we investigated adding $k \in \{10, 50\}$ i.i.d. standard normal covariates as potential predictors to the five photometric covariates. This

[6] For the computation of the conditional density estimators we used code by [20] (link).

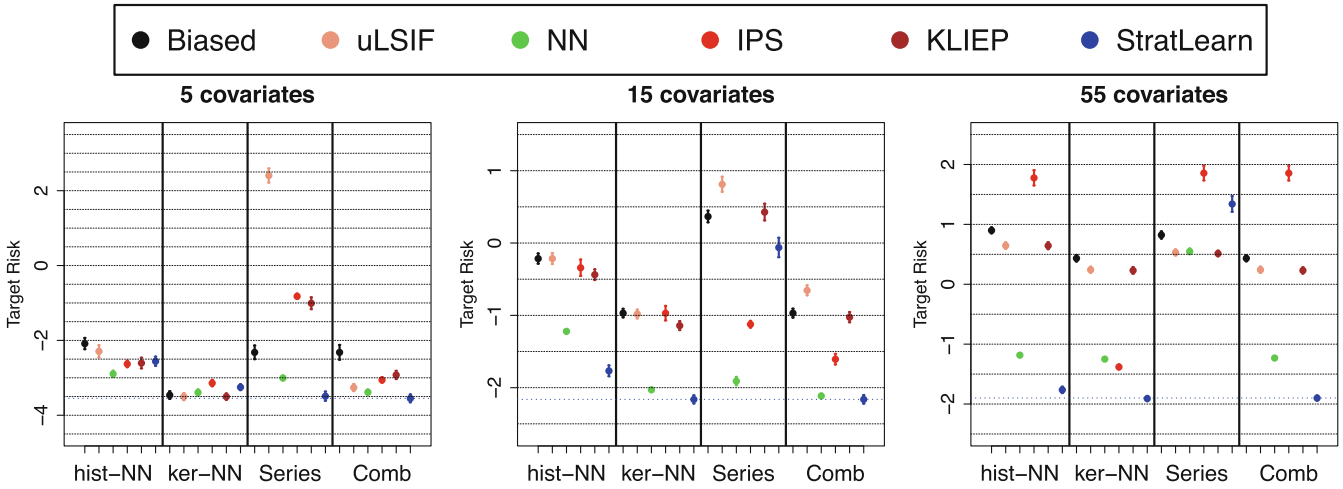

FIGURE 5 Target risk ($\widehat{R}_T$) of the four photo-z estimation models under each method (different colors), using different sets of predictors. Bars give the mean ± 2 bootstrap standard errors (from 400 bootstrap samples).

simulates the realistic case where additional potentially confounding covariates are present. For comparability, we follow Ref. [20] and use $|D_S^{\text{train}}| = 2800$ galaxies randomly sampled from D_S as training data, plus a validation set of $|D_S^{\text{val}}| = 1200$ galaxies. We assess the performance of each $\text{Method}_{\text{Model}}$ pair using a random subset of D_T, that is, $|D_T^{\text{test}}| = 6000$.

4.3.3 | Results

For evaluation of $\widehat{f}$ under each $\text{Method}_{\text{Model}}$ pair, we use the (in our simulation) known target redshifts, z_T, to compute the target risk, $\widehat{R}_T(\widehat{f})$, via a non-weighted version of (11) with $x_S^{(k)}$ and $y_S^{(k)}$ replaced by $x_T^{(k)}$ and $y_T^{(k)}$, given by

$$\widehat{R}_T(\widehat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \widehat{f}^2\left(z|x_T^{(k)}\right)dz - 2\frac{1}{n_T}\sum_{k=1}^{n_T}\widehat{f}\left(z_T^{(k)}|x_T^{(k)}\right). \tag{13}$$

Figure 5 compares the resulting target risk $\widehat{R}_T$ across models and covariate sets, showing that $\textit{StratLearn}_{Comb}$ gives the best performance in all three covariate setups.

For small covariate space dimension (Figure 5, left panel), $\textit{StratLearn}_{Comb}$ improves upon $\textit{StratLearn}_{ker-NN}$ and $\textit{StratLearn}_{Series}$, optimizing the source risk in each stratum separately and combining their predictions. In the presence of potential additional confounding covariates (Figure 5, middle and right panels), the performance of the Series estimator degrades strongly under most methods. In these cases, $\textit{StratLearn}_{Comb}$ exploits the higher performance of $\textit{StratLearn}_{ker-NN}$. In contrast, for the non-adjusted (Biased) and importance-weighted methods (e.g. IPS), the combination of approaches (Comb) does not necessarily lead to improved performance (e.g., IPS_{Comb} exhibits a higher target risk than $\text{IPS}_{\text{ker-NN}}$ on its own (Figure 5, right panel)), indicating that the optimization in (12) fails due to remaining covariate shift in the data. More precisely, the weighted empirical *source* risk minimization ((11), as a form of (1)) does not lead to *target* risk minimization in these situations. In general, the improvement of *StratLearn* relative to weighting methods increases with the dimensionality of the covariate space, leading to a more robust regime.

5 | DISCUSSION

We provide a simple, though statistically principled and theoretically justified method for learning under covariate shift conditions. We show that *StratLearn* outperforms a range of state-of-the-art importance weighting methods on two contemporary research questions in cosmology (and on toy covariate shift examples, Sections S5 and S6), especially in the presence of a high-dimensional covariate space. The assumption of covariate shift is rather strong, requiring that there are no unmeasured confounding covariates—something that cannot be guaranteed in general. In Section S6, however, we demonstrate a certain robustness of our method against violation of this assumption. Further work is necessary to assess the performance of *StratLearn* more fully when this assumption is only approximately fulfilled. We emphasize that the covariate shift framework is best justified in the presence of a large number of covariates mitigating the risk of unmeasured confounders—in which case it is critical to adopt a method that, like *StratLearn*, can robustly handle many covariates. Our framework is entirely general

and versatile, as illustrated with examples of regression, conditional density estimation and classification. Notably, our numerical demonstrations illustrate the advantage of using only a subset of the source data when formulating predictions for individual objects in the target, where the subset is chosen for its similarity to the target data in question (through stratification). This is a markedly different strategy to the widespread practice of including all possible available observations when fitting learning models.

The novelty of our approach is grounded in the transfer of the well-established causal inference propensity score framework [47] to the domain adaptation/covariate shift setting, by demonstrating that a method established to obtain unbiased treatment effect estimates can be adapted to optimize the target risk of a supervised learner under covariate shift. In future work, this extension offers the opportunity to transfer hard-won practical advice from causal inference (e.g., balance diagnostics, estimation of propensity scores, and choice of included covariates [2, 41, 48]) to the covariate shift framework. We will also explore the possibility of taking advantage of Proposition 1 through a matching approach [18], which could prove more sensitive to the underlying propensity score distribution. We believe *StratLearn* may become a powerful alternative to importance weighting, with a myriad of possible extensions and applications.

ACKNOWLEDGMENTS

David van Dyk acknowledges partial support from the UK Engineering and Physical Sciences Research Council [EP/W015080/1]; Roberto Trotta's work was partially supported by STFC in the UK [ST/P000762/1, ST/T000791/1], and co-funded in Italy from Next Generation EU (DM 1555 del 11.10.22), in the context of the National Recovery and Resilience Plan, Investment PE1 (Project FAIR "Future Artificial Intelligence Research"), as well as partially supported by the Fondazione ICSC, Spoke 3 "Astrophysics and Cosmos Observations," Piano Nazionale di Ripresa e Resilienza Project ID CN00000013, "Italian Research Center on High-Performance Computing, Big Data and Quantum Computing," funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19)." David Stenning acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2021-03985]. Finally, van Dyk, Stenning, and Autenrieth acknowledge support from the Marie-Skodowska-Curie RISE [H2020-MSCA-RISE-2019-873089] Grant provided by the European Commission. *The authors report there are no competing interests to declare.*

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Maximilian Autenrieth https://orcid.org/0009-0006-2068-5950
David A. van Dyk https://orcid.org/0000-0002-0816-331X
Roberto Trotta https://orcid.org/0000-0002-3415-0707
David C. Stenning https://orcid.org/0000-0002-9761-4353

REFERENCES

1. P. C. Austin, *An introduction to propensity score methods for reducing the effects of confounding in observational studies*, Multivar. Behav. Res. 46 (2011), no. 3, 399–424.
2. P. C. Austin, P. Grootendorst, and G. M. Anderson, *A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study*, Stat. Med. 26 (2007), no. 4, 734–753.
3. P. C. Austin and E. A. Stuart, *Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies*, Stat. Med. 34 (2015), no. 28, 3661–3679.
4. P. C. Austin and E. A. Stuart, *The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes*, Stat. Methods Med. Res. 26 (2017), no. 4, 1654–1670.
5. S. Bickel and T. Scheffer, "*Dirichlet-enhanced spam filtering based on biased samples*," *Advances in neural information processing systems*, MIT Press, Cambridge, USA, 2007, pp. 161–168.
6. K. Boone, *Avocado: Photometric classification of astronomical transients with gaussian process augmentation*, Astron. J. 158 (2019), no. 6, 257.
7. T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "*Controlling attribute effect in linear regression*," *2013 IEEE 13th international conference on data mining*, IEEE, New York, USA, 2013, pp. 71–80.
8. A. Chan, A. Alaa, Z. Qian, and M. Van Der Schaar, "*Unlabelled data improves bayesian uncertainty calibration under covariate shift*," *International conference on machine learning*, PMLR, Cambridge, USA, 2020, pp. 1392–1402.
9. W. G. Cochran, *The effectiveness of adjustment by subclassification in removing bias in observational studies*, Biometrics 24 (1968), no. 2, 295–313.
10. R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal. 21 (2006), no. 1, 5–30.
11. C. Cortes, Y. Mansour, and M. Mohri, "*Learning bounds for importance weighting*," *Advances in neural information processing system*, Curran Associates, Inc., New York, USA, 2010, pp. 442–450.
12. G. Csurka, "*A comprehensive survey on domain adaptation for visual applications*," *Advances in Computer Vision and Pattern Recognition*, Springer, Cham, 2017, pp. 1–35.
13. Dua, D. and C. Graff. 2017. *UCI machine learning repository,* UC Irvine, Irvine, CA.

14. J. J. Heckman, *Sample selection bias as a specification error*, Econom. J. Econom. Soc. 47 (1979), 153–161.
15. K. Hirano, G. W. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica 71 (2003), no. 4, 1161–1189.
16. J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, "*Correcting sample selection bias by unlabeled data*," *Advances in neural information processing systems*, MIT Press, Cambridge, USA, 2007, pp. 601–608.
17. K. Imai and D. A. van Dyk, *Causal inference with general treatment regimes: Generalizing the propensity score*, J. Am. Stat. Assoc. 99 (2004), no. 467, 854–866.
18. G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, Cambridge, UK, 2015.
19. R. Izbicki, A. Lee, and C. Schafer, "*High-dimensional density ratio estimation with extensions to approximate likelihood computation*," *Artificial intelligence and statistics*, PMLR, Cambridge, USA, 2014, pp. 420–429.
20. R. Izbicki, A. B. Lee, P. E. Freeman, et al., *Photo-z estimation: An example of nonparametric conditional density estimation under selection bias*, Ann. Appl. Stat. 11 (2017), no. 2, 698–724.
21. J. Jiang and C. Zhai, *Instance weighting for domain adaptation in nlp*, ACL, Prague, Czech Republic, 2007.
22. K. Kamnitsas, C. Baumgartner, C. Ledig, et al., "*Unsupervised domain adaptation in brain lesion segmentation with adversarial networks*," *International conference on information processing in medical imaging*, Springer, Berlin, Germany, 2017, pp. 597–609.
23. T. Kanamori, S. Hido, and M. Sugiyama, *A least-squares approach to direct importance estimation*, J. Mach. Learn. Res. 10 (2009), no. Jul, 1391–1445.
24. R. Kessler, B. Bassett, P. Belov, V. Bhatnagar, H. Campbell, A. Conley, J. A. Frieman, A. Glazov, S. González-Gaitán, R. Hlozek, S. Jha, S. Kuhlmann, M. Kunz, H. Lampeitl, A. Mahabal, J. Newling, R. C. Nichol, D. Parkinson, N. S. Philip, D. Poznanski, J. W. Richards, S. A. Rodney, M. Sako, D. P. Schneider, M. Smith, M. Stritzinger, and M. Varughese, *Results from the supernova photometric classification challenge*, Publ. Astron. Soc. Pac. 122 (2010), no. 898, 1415–1431.
25. R. Kessler, A. Conley, S. Jha, and S. Kuhlmann. Supernova photometric classification challenge. 2010 arXiv preprint arXiv:1001.5210.
26. R. Kessler, G. Narayan, A. Avelino, et al., *Models and simulations for the photometric lst astronomical time series classification challenge (plastic)*, Publ. Astron. Soc. Pac. 131 (2019), no. 1003, 094501.
27. W. M. Kouw and M. Loog, *A review of domain adaptation without target labels*, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2019), 766–785.
28. J. Kremer, F. Gieseke, K. S. Pedersen, and C. Igel, *Nearest neighbor density ratio estimation for large-scale applications in astronomy*, Astronom. Comput. 12 (2015), 67–72.
29. M. Lima, C. E. Cunha, H. Oyaizu, J. Frieman, H. Lin, and E. S. Sheldon, *Estimating the redshift distribution of photometric galaxy samples*, Mon. Not. R. Astron. Soc. 390 (2008), no. 1, 118–130.
30. R. J. Little and D. B. Rubin, *Statistical analysis with missing data, Volume 793*, John Wiley & Sons, Hoboken, New Jersey, 2019.
31. A. Liu and B. D. Ziebart, "*Robust classification under sample selection bias*," *Advances in Neural Information Processing Systems*, Curran Associates, Inc., New York, USA, 2014, 37–45.
32. M. Lochner, J. D. McEwen, H. V. Peiris, O. Lahav, and M. K. Winter, *Photometric supernova classification with machine learning*, Astrophys. J. Suppl. Ser. 225 (2016), no. 2, 31.
33. M. Loog, "*Nearest neighbor-based importance weighting*," *2012 IEEE international workshop on machine learning for signal processing*, IEEE, New York, USA, 2012, pp. 1–6.
34. J. K. Lunceford and M. Davidian, *Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study*, Stat. Med. 23 (2004), no. 19, 2937–2960.
35. S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "*Domain adaptation by using causal inference to predict invariant conditional distributions*," *Advances in neural information processing systems*, Curran Associates, Inc., New York, USA, 2018, pp. 10846–10856.
36. J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera, *A unifying view on dataset shift in classification*, Pattern Recogn. 45 (2012), no. 1, 521–530.
37. S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, *Domain adaptation via transfer component analysis*, IEEE Trans. Neural Netw. 22 (2010), no. 2, 199–210.
38. S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Trans. Knowl. Data Eng. 22 (2009), no. 10, 1345–1359.
39. J. Pasquet, J. Pasquet, M. Chaumont, and D. Fouchez, *Pelican: Deep architecture for the light curve analysis*, Astronom. Astrophys. 627 (2019), A21.
40. S. Perlmutter, G. Aldering, G. Goldhaber, et al., *Measurements of ω and λ from 42 high-redshift supernovae*, Astrophys. J. 517 (1999), no. 2, 565.
41. R. Pirracchio, M. L. Petersen, and M. van der Laan, *Improving propensity score estimators' robustness to model misspecification using super learner*, Am. J. Epidemiol. 181 (2014), no. 2, 108–119.
42. S. Reddi, B. Poczos, and A. Smola, "*Doubly robust covariate shift correction*," Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 29*, MIT Press, Cambridge, USA, 2015.
43. E. A. Revsbech, R. Trotta, and D. A. van Dyk, *Staccato: A novel solution to supernova photometric classification with biased training sets*, Mon. Not. R. Astron. Soc. 473 (2018), no. 3, 3969–3986.
44. J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer, and D. Poznanski, *Semi-supervised learning for photometric supernova classification*, Mon. Not. R. Astron. Soc. 419 (2012), no. 2, 1121–1135.
45. A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry, *Observational evidence from supernovae for an accelerating universe and a cosmological constant*, Astron. J. 116 (1998), no. 3, 1009–1038.
46. W. A. Rivera, A. Goel, and J. P. Kincaid, "*Oups: A combined approach using smote and propensity score matching*," *2014 13th international conference on machine learning and applications*, IEEE, New York, USA, 2014, pp. 424–427.
47. P. R. Rosenbaum and D. B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika 70 (1983), no. 1, 41–55.

48. P. R. Rosenbaum and D. B. Rubin, *Reducing bias in observational studies using subclassification on the propensity score*, J. Am. Stat. Assoc. 79 (1984), no. 387, 516–524.
49. D. B. Rubin, *On principles for modeling propensity scores in medical research*, Pharmacoepidemiol. Drug Saf. 13 (2004), no. 12, 855–857.
50. E. S. Sheldon, C. E. Cunha, R. Mandelbaum, J. Brinkmann, and B. A. Weaver, *Photometric redshift probability distributions for galaxies in the sdss dr8*, Astrophys. J. Suppl. Ser. 201 (2012), no. 2, 32.
51. H. Shimodaira, *Improving predictive inference under covariate shift by weighting the log-likelihood function*, J. Stat. Plan. Inference 90 (2000), no. 2, 227–244.
52. P. Stojanov, M. Gong, J. Carbonell, and K. Zhang, "*Low-dimensional density ratio estimation for covariate shift correction*," *The 22nd international conference on artificial intelligence and statistics*, PMLR, Cambridge, USA, 2019, pp. 3449–3458.
53. M. Sugiyama, M. Krauledat, and K. R. Müller, *Covariate shift adaptation by importance weighted cross validation*, J. Mach. Learn. Res. 8 (2007), no. May, 985–1005.
54. M. Sugiyama and K. R. Müller, *Input-dependent estimation of generalization error under covariate shift*, Stat. Risk Model. 23 (2005), no. 4, 249–279.
55. Sugiyama, M., S. Nakajima, H. Kashima, P.V. Buenau, and M. Kawanabe, *Direct importance estimation with model selection and its application to covariate shift adaptation*. In *Advances in neural information processing systems*, Curran Associates, Inc., New York, USA, 2008, Pp. 1433–1440.
56. B. Zadrozny, "*Learning and evaluating classifiers under sample selection bias*," *Proceedings of the twenty-first international conference on Machine learning*, ACM, New York, USA, 2004, pp. 114.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: M. Autenrieth, D. A. van Dyk, R. Trotta, and D. C. Stenning, *Stratified learning: A general-purpose statistical method for improved learning under covariate shift*, Stat. Anal. Data Min.: ASA Data Sci. J. (2023), 1–16. https://doi.org/10.1002/sam.11643