Neural Computing and Applications (2023) 35:18187–18199
https://doi.org/10.1007/s00521-021-06615-1

S.I.: LATINX IN AI RESEARCH

Effective sample size, dimensionality, and generalization in covariate shift adaptation

Felipe Maia Polo[1] · Renato Vicente[2]

Received: 3 February 2021 / Accepted: 4 October 2021 / Published online: 8 January 2022

Abstract
In supervised learning, training and test datasets are often sampled from distinct distributions. Domain adaptation techniques are thus required. Covariate shift adaptation yields good generalization performance when domains differ only by the marginal distribution of features. Covariate shift adaptation is usually implemented using importance weighting, which may fail, according to common wisdom, due to small effective sample sizes (ESS). Previous research argues this scenario is more common in high-dimensional settings. However, how effective sample size, dimensionality, and model performance/generalization are formally related in supervised learning, considering the context of covariate shift adaptation, is still somewhat obscure in the literature. Thus, a main challenge is presenting a unified theory connecting those points. Hence, in this paper, we focus on building a unified view connecting the ESS, data dimensionality, and generalization in the context of covariate shift adaptation. Moreover, we also demonstrate how dimensionality reduction or feature selection can increase the ESS and argue that our results support dimensionality reduction before covariate shift adaptation as a good practice.

Keywords Covariate shift adaptation · Effective sample size · High-dimensional data · Dimensionality reduction

1 Introduction

A fundamental assumption in supervised statistical learning is that the data used to train our models and the data we want to make predictions for are sampled from the same distribution. Usually, real-world machine leaning applications, explicitly or implicitly, rely on this assumption[1]. However, that assumption is violated when there is covariate shift [7, 8]. In this scenario, we have a training/source joint distribution $Q_{\mathbf{x},\mathrm{y}}$ which differs from the test/target distribution $P_{\mathbf{x},\mathrm{y}}$. Features are sampled from different marginals $Q_{\mathbf{x}} \neq P_{\mathbf{x}}$ while labels are sampled according to the same conditional distribution $Q_{\mathrm{y}|\mathbf{x}} = P_{\mathrm{y}|\mathbf{x}}$. In the training phase, labeled pairs $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^{n}$ are identically and independently sampled from $Q_{\mathbf{x},\mathrm{y}}$, while unlabeled vectors $\{\mathbf{x}'_i\}_{i=1}^{m}$ are identically and independently sampled from $P_{\mathbf{x}}$. If the marginal distributions of features have density functions $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$, such that $\mathrm{support}(p_{\mathbf{x}}) \subseteq \mathrm{support}(q_{\mathbf{x}})$, the most common approach to adapt a model for the target distribution is to employ an empirical error weighted by $w(\boldsymbol{x}) = p_{\mathbf{x}}(\boldsymbol{x})/q_{\mathbf{x}}(\boldsymbol{x})$ [7–11].

The weighting scheme may fail when the effective sample sizes (ESS) are small. According to common wisdom, a small ESS hurts model's performance in the target distribution. As previous research argues, e.g., [12], that kind of scenario is common when working with high-dimensional data. However, to the best of our knowledge, there is no unified and rigorous view on how the three key concepts (i) effective sample size (ESS), (ii) data dimensionality, and (iii) generalization of supervised models

✉ Felipe Maia Polo
felipemaiapolo@gmail.com

Renato Vicente
rvicente@usp.br

[1] Department of Statistics, Institute of Mathematics and Statistics of the University of São Paulo Advanced Institute for Artificial Intelligence (AI2), São Paulo, Brazil

[2] Department of Applied Mathematics, Institute of Mathematics and Statistics of the University of São Paulo Experian DataLab LatAm Advanced Institute for Artificial Intelligence (AI2), São Paulo, Brazil

[1] See, for example, [1–6].

under covariate shift are connected to each other. In this paper, we present a unified theory connecting the three concepts. Moreover, we also explore how dimensionality reduction or feature selection can increase the effective sample size.

This paper is organized as follows. In Sect. 2, we discuss previous results and explain our contribution to the debate. In Sect. 3, we briefly review importance weighting and introduce a new connection between effective sample size and generalization in the context of covariate shift adaptation. In Sect. 4, we introduce dimensionality to the problem showing how it connects to the other two concepts and then illustrate these connections with a toy experiment. Finally, in Sect. 5, we show how dimensionality reduction and feature selection can lead to a larger effective sample size. We conclude our discussion with real-data experiments that supports feature selection before covariate shift adaptation as a good practice.

2 Related work

There is a rich literature on the problem of covariate shift adaptation[2] or related subjects. The main interest has been to develop methods to estimate the density ratio w [9, 11, 13–15]. Some of the proposed methods aim to reliably estimate w in high-dimensional and unstable settings [14, 15], when the more traditional approaches may fail. However, according to the common wisdom of the area, even if we could perfectly estimate w, we would still have to deal with poor performance due to small effective sample sizes (ESS), especially in high-dimensional settings. Understanding the role of small ESS and possible ways to attenuate it may, therefore, be productive. The covariate shift adaptation literature has already tried to articulate the relationships between ESS and generalization in high-dimensional settings, also proposing dimensionality reduction as a cure. In spite of that, we believe these previous attempts fail in connecting these concepts in a unified manner and as explicitly as we propose to do in this paper.

In recent years, [16] proposed a regularization method that controls the ESS and offers sharper generalization bounds while correcting for covariate shift. However, the authors do not explore how the number of features plays an essential role. Another work that explores the concept of ESS in the context of covariate shift adaptation is [17]. In that work, the authors present the relationship between ESS and generalization bounds in a transductive learning scenario. Besides transductive learning not being as common as inductive learning in practice, the authors also do not explore how dimensionality plays an essential role in the problem.

The idea of features dimensionality being related to ESS is explored in [12], without formalizing the connection to generalization. The authors also motivate how dimensionality reduction can make ESS bigger; however, the central hypothesis adopted in this case is that dimensionality reduction does not depend on the data, which, in most cases, is not valid. In a more recent paper, [18] proposes a dimensionality reduction method to make covariate shift adaptation feasible, especially when estimating weights. The authors show how the number of features is indirectly related to transductive generalization bounds and effective sample size when the correction is made by Kernel Mean Matching [9]. In addition to the results being restricted to a particular case, the authors implicitly assume that the mapping that defines dimensionality reduction is given beforehand and does not depend on the training data, what is not realistic.

In this paper, we complement previous works by formally articulating the relationship among ESS, generalization of predictive models in the inductive scenario, and dimensionality as explicitly as possible. We present a unified theory connecting the three concepts, which was not observed by us in the literature. We also show that dimensionality reduction, even considering that the mapping may depend on the data, mitigates low ESS by making the source and target domains less divergent.

3 Effective sample size (ESS) and generalization in covariate shift adaptation

3.1 Importance weighting

To keep our discussion as self-contained as possible, we first use this subsection to quickly summarize key points behind importance weighting.

Given a hypothesis class $\mathcal{H}$ and a loss function L, our goal is finding a hypothesis $h^* \in \mathcal{H}$ that minimizes the risk R assessed in the target distribution $P_{\mathbf{x},\mathrm{y}}$ using data from source distribution $Q_{\mathbf{x},\mathrm{y}}$. From now on we assume: (i) $Q_{\mathrm{y}|\mathbf{x}} = P_{\mathrm{y}|\mathbf{x}}$ and $Q_{\mathbf{x}} \neq P_{\mathbf{x}}$; (ii) distributions $P_{\mathbf{x}}$ and $Q_{\mathbf{x}}$ have probability density functions (p.d.f.s) $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$ such that $\mathrm{support}(p_{\mathbf{x}}) \subseteq \mathrm{support}(q_{\mathbf{x}})$. Then, the risk can be written in terms of the source distribution:

$$R(h) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}}}\mathbb{E}_{\mathrm{y}|\mathbf{x}}[L(h(\mathbf{x}),\mathrm{y})] \tag{1}$$

$$= \int \frac{p_{\mathbf{x}}(\boldsymbol{x})}{q_{\mathbf{x}}(\boldsymbol{x})} q_{\mathbf{x}}(\boldsymbol{x})\mathbb{E}_{\mathrm{y}|\mathbf{x}}[L(h(\boldsymbol{x}),\mathrm{y})]\mathrm{d}\boldsymbol{x} \tag{2}$$

[2] See [8] for a general view.

$$= \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\mathbb{E}_{\mathrm{y}|\mathbf{x}}[w(\mathbf{x})\cdot L(h(\mathbf{x}),\mathrm{y})] \quad (3)$$

We would like to find a hypothesis $h^{\mathrm{ERM}}_{\hat{w}} \in \mathcal{H}$ that minimizes a weighted version of the empirical risk while also obtaining a low value for R. Assume we have an estimate $\hat{w}$ for the "true" weighting function $w = p_{\mathbf{x}}/q_{\mathbf{x}}$ and that we have pairs $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$ that are identically and independently (i.i.d.) sampled from $Q_{\mathbf{x},\mathrm{y}}$. The weighted empirical risk is thus given by

$$\widehat{R}_{\hat{w}}(h) = \frac{1}{n}\sum_{i=1}^{n}\hat{w}(\mathbf{x}_i)\cdot L(h(\mathbf{x}_i),\mathrm{y}_i) \quad (4)$$

In practice, we might also want to add a regularization term $\Omega(h)$ to penalize for the complexity of the hypothesis h.

3.2 Relationship of effective sample size (ESS) and generalization in covariate shift adaptation

To introduce the concept of effective sample size in the context of covariate shift adaptation, we first describe how this heuristic is employed within the importance sampling literature [19–21], where it originally comes from. We assume the "true" importance function (density ratio) is known up to a constant. This assumption enables us to achieve some theoretical results and is also adopted in previous works [12, 22, 23]. The strategy we use to show the relevance of the effective sample size in covariate shift adaptation is to find an asymptotic approximation for that quantity and then connect it to a known generalization bound.

The ESS formulation we use is slightly different from the most usual one [19–21] in the sense we are concerned with percentage of effective samples and not with the number of effective samples[3]. Given the two definitions are not very different, the intuitions and some results regarding ESS are easily adaptable. We present our definition in the following.

Consider two probability distributions $P_{\mathbf{z}}$ and $Q_{\mathbf{z}}$ over $\mathcal{Z} \subseteq \mathbb{R}^d$ with probability density functions $p_{\mathbf{z}}$ and $q_{\mathbf{z}}$ such that $\mathrm{support}(p_{\mathbf{z}}) \subseteq \mathrm{support}(q_{\mathbf{z}})$. From now on, we call $P_{\mathbf{z}}$ the *target* distribution and $Q_{\mathbf{z}}$ the *source* distribution. We thus sample from $Q_{\mathbf{z}}$ in order to estimate the integral $\int_{\mathcal{Z}} g(z)p_{\mathbf{z}}(z)dz = \int_{\mathcal{Z}} \frac{p_{\mathbf{z}}(z)}{q_{\mathbf{z}}(z)} g(z)q_{\mathbf{z}}(z)dz$, with $g : \mathcal{Z} \to \mathbb{R}$ integrable. A key quantity in this problem is the importance function, which is given by $w \propto p_{\mathbf{z}}/q_{\mathbf{z}}$.

Suppose we have an independent and identically distributed (i.i.d.) sample $\{\mathbf{z}_i\}_{i=1}^n$ from the source distribution $Q_{\mathbf{z}}$ and we want to use the (self-normalized[4]) importance sampling estimator $n^{-1}\sum_{i=1}^n \bar{\mathrm{w}}_i g(\mathbf{z}_i)$ in order to estimate the integral of interest. The weights are given by $\bar{\mathrm{w}}_i = \mathrm{w}_i/\sum_j \mathrm{w}_j$, where $\mathrm{w}_i = w(\mathbf{z}_i) \propto p_{\mathbf{z}}(\mathbf{z}_i)/q_{\mathbf{z}}(\mathbf{z}_i)$, $i \in [n] := \{1, ..., n\}$. Then, the effective sample size is defined as

$$\widehat{ESS}_n(P_{\mathbf{z}}, Q_{\mathbf{z}}) := \frac{1}{n\sum_{i=1}^n \bar{\mathrm{w}}_i^2} \quad (5)$$

$$= \frac{\left(\sum_{i=1}^n \mathrm{w}_i\right)^2}{n\sum_{i=1}^n \mathrm{w}_i^2} \quad (6)$$

Intuitively, the effective sample size is the percentage of effective samples. For example, if the effective sample size equals 1/2, then the importance sampling estimator *effectiveness* is the same of a Monte Carlo estimator with $n/2$ samples. That formulation can be used to approximate, via Delta Method, the ratio of Monte Carlo estimators' variance and the self-normalized importance sampling estimator' variance, using the derivation made by [24]. While that work motivates the use of the ESS, other approaches can be derived from [20] and [21]. The latter presents the relationship between effective sample size and the euclidean distance between the vector $(\bar{\mathrm{w}}_1, ..., \bar{\mathrm{w}}_n)$ and the "ideal" balanced vector $(1/n, ..., 1/n)$. Furthermore, effective sample size informs about the importance sampling estimator's convergence rate [25]. Said that, the results presented in this section, in the context of covariate shift adaptation, resembles the results presented by [25] in a different context.

To move forward, we introduce the concept of Rényi Divergence, which plays a central role in our analysis:

Definition 1 (Rényi Divergence [26]) Consider two probability distributions $P_{\mathbf{x}}$ and $Q_{\mathbf{x}}$ over $\mathcal{X} \subseteq \mathbb{R}^d$, with probability density functions $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$ such that $\mathrm{support}(p_{\mathbf{x}}) \subseteq \mathrm{support}(q_{\mathbf{x}})$. The Rényi Divergence of order $\alpha > 1$ of $P_{\mathbf{x}}$ from $Q_{\mathbf{x}}$ is given by:

$$D_\alpha(P_{\mathbf{x}}||Q_{\mathbf{x}}) := \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\left(\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right)^{\alpha}\right] \quad (7)$$

Consequently, the Rényi Divergence of order 2 of $P_{\mathbf{x}}$ from $Q_{\mathbf{x}}$ is given by $D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) = \log \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}}}[\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}]$.

Despite all previous work, the question of how we should transpose the effective sample size concept to the covariate shift adaptation framework remains. In the following, we make explicit the close relationship between the ESS and generalization bounds under covariate shift

[3] In the literature, it is common to present the ESS as $n \cdot \widehat{ESS}_n$, while we are concerned only with $\widehat{ESS}_n$ (Equation 5).

[4] We show the case of the self-normalized estimator because it returns the most usual definition for the ESS, which is also used in the context of covariate shift [16]. In spite of that, we show that this definition for the ESS is still useful for the non-normalized case while performing covariate shift adaptation.

adaptation. As we start talking about covariate shift adaptation, we substitute $\mathbf{z}$ by a vector of features $\mathbf{x}$, the set $\mathcal{Z}$ by $\mathcal{X}$ or $\mathcal{X} \times \mathcal{Y}$ and the function g by the loss function L. Before we move on, we must establish that the effective sample size $\widehat{ESS}_n(P_\mathbf{x}, Q_\mathbf{x})$ converges almost surely to the quantity $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$, which plays a central role in our analysis. $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$ can be considered a population version for the effective sample size. From now on, we may call it by population effective sample size or only effective sample size, when it is not ambiguous.

Theorem 1 *Consider two probability distributions $P_\mathbf{x}$ and $Q_\mathbf{x}$ over $\mathcal{X} \subseteq \mathbb{R}^d$, with probability density functions $p_\mathbf{x}$ and $q_\mathbf{x}$ such that* $\text{support}(p_\mathbf{x}) \subseteq \text{support}(q_\mathbf{x})$. *Suppose we have a random sample $\{\mathbf{x}_i\}_{i=1}^n$, identically and independently sampled from the distribution $Q_\mathbf{x}$, and we define $\mathrm{w}_i = w(\mathbf{x}_i) \propto p_\mathbf{x}(\mathbf{x}_i)/q_\mathbf{x}(\mathbf{x}_i)$. Assume that $0 < \mathbb{E}_{\mathbf{x}\sim Q_\mathbf{x}}\left[w(\mathbf{x})^2\right] < \infty$. Then,*

$$\widehat{ESS}_n(P_x, Q_x) \xrightarrow[n\to\infty]{a.s.} ESS^*(P_x, Q_x) \tag{8}$$

where

$$ESS^*(P_\mathbf{x}, Q_\mathbf{x}) := exp[-D_2(P_\mathbf{x}||Q_\mathbf{x})] \tag{9}$$

The quantity $D_2(P_\mathbf{x}||Q_\mathbf{x})$ is the Rényi Divergence of order 2 of $P_\mathbf{x}$ from $Q_\mathbf{x}$ [26].

The proof can be found in the "Appendix". The last theorem can be seen as a variation of some of the results presented by Agapiou et al. [25]. While the authors focus on related but different divergences, we choose to present this result in terms of the Rényi Divergence because, in that way, we can connect it to other results in the literature. Furthermore, it is essential to state that similar results hold for other effective sample size definitions as, for example, the one used by [12] divided by n, to give the percentage of effective samples considering the non normalized weights for covariate shift adaptation.

It is fascinating how Rényi Divergence naturally emerges when working with the effective sample size. It is a crucial point to understand that, when calculating the effective sample size, we are approximating a quantity inversely proportional to the exponential of Rényi Divergence of order 2 of $P_\mathbf{x}$ from $Q_\mathbf{x}$.

Now we focus on the understanding of how effective sample size relates to generalization of adapted supervised models. For Theorem 2, consider some conditions. Let $\mathcal{X}$ denote the input space, $\mathcal{Y}$ the label set, and let $L : \mathcal{Y}^2 \to [0, 1]$ be a bounded loss function. Denote the *target* distribution of features by $P_\mathbf{x}$ and the *source* distribution of features by $Q_\mathbf{x}$, such that $P_\mathbf{x}$ is dominated by $Q_\mathbf{x}$. Consider $\mathcal{H}$ to be the hypothesis class used by the learning algorithm and $f : \mathcal{X} \to \mathcal{Y}$ to be the labeling function we want to learn about. We denote by $\text{Pdim}(U)$ the pseudo-dimension[5] of a real-valued function class U [27]. Pdim is used here to quantify the complexity of a hypothesis class through the loss function. Finally, R is the risk assessed in the target distribution $P_\mathbf{x}$ and $\hat{R}_w$ is the weighted empirical error calculated using the true weighting function (density ratio) and samples $\{\mathbf{x}_i\}_{i=1}^n$, identically and independently sampled from the source distribution $Q_\mathbf{x}$.

Theorem 2 (*Adapted from* [22]) *Define the function $L_h(\boldsymbol{x}) := L[h(\boldsymbol{x}), f(\boldsymbol{x})]$ and let $\mathcal{H}$ be a hypothesis set such that* $\text{Pdim}(\{L_h : h \in \mathcal{H}\}) = p < \infty$. *Assume that $ESS^*(P_\mathbf{x}, Q_\mathbf{x}) = exp[-D_2(P_\mathbf{x}||Q_\mathbf{x})]$, $D_2(P_\mathbf{x}||Q_\mathbf{x}) < \infty$, and the target/source density ratio $w > 0$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have that:*

$$\sup_{h\in\mathcal{H}}[R(h) - \hat{R}_w(h)] \leq \tag{10}$$

$$\leq \frac{2^{\frac{5}{4}}}{\sqrt{ESS^*(P_\mathbf{x}, Q_\mathbf{x})}} \cdot \left[\frac{p \cdot log\frac{2\cdot e\cdot n}{p} + log\frac{4}{\delta}}{n}\right]^{\frac{3}{8}} \tag{11}$$

See [22] for the proof and replace D_2 by ESS^* to get this version of the theorem.

It is clear from Theorem 2 that $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$ plays a fundamental role when learning f from data. A larger $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$ leads to a tighter generalization bound. Consequently, if $\widehat{ESS}_n(P_\mathbf{x}, Q_\mathbf{x})$ is a good approximation for $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$, the rationale behind using effective sample size as a heuristic for diagnosis of covariate shift adaptation becomes clearer. To conclude, we should mention that [23] shows a similar result to Theorem 2 with less assumptions, namely, assuming the existence of a labeling function f and that $w > 0$. However, we chose the form provided by [22], as it gives us a more straightforward expression without losing the property that is key to our approach, to say, that a larger $ESS^*(P_\mathbf{x}, Q_\mathbf{x})$ leads to a sharper generalization bound.

4 The role of dimensionality

In Sect. 3, we showed the effective sample size's role in the context of covariate shift adaptation exploring its asymptotic relationship with generalization bounds. However, we still need to understand the role that dimensionality plays during covariate shift adaptation. In Theorem 3, we demonstrate that the Rényi Divergence of source and target distributions does not decrease with the number of features, and, consequently, the population effective

[5] A pseudo-dimension is an extension of VC Dimension for real-valued classes of functions.

sample size does not increase with the number of features, which explains potential adaptation problems for high-dimensional data.

Theorem 3 *Given two joint probability distributions* $P_{\mathbf{x}_1,\mathbf{x}_2}$ *(target) and* $Q_{\mathbf{x}_1,\mathbf{x}_2}$ *(source) over* $\mathcal{X} \subseteq \mathbb{R}^d$, $D_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2}) < \infty$, *with joint probability density functions* $p_{\mathbf{x}_1,\mathbf{x}_2}$ *and* $q_{\mathbf{x}_1,\mathbf{x}_2}$, *such that* $\text{support}(p_{\mathbf{x}_1,\mathbf{x}_2}) \subseteq \text{support}(q_{\mathbf{x}_1,\mathbf{x}_2})$, *we have that*

$$D_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2}) \geq D_2(P_{\mathbf{x}_1}||Q_{\mathbf{x}_1}) \tag{12}$$

And, consequently,

$$ESS^*(P_{\mathbf{x}_1}, Q_{\mathbf{x}_1}) \geq ESS^*(P_{\mathbf{x}_1,\mathbf{x}_2}, Q_{\mathbf{x}_1,\mathbf{x}_2}) \tag{13}$$

The proof can be found in the "Appendix". This theorem can be seen as a particular case of the Data Processing Inequality [26].

Combining the results of Theorems 2 and 3, we conclude that performing covariate shift adaptation with many features may not be feasible, as we would potentially have loose generalization bounds.

Note that Theorem 3 does not necessarily say that by reducing dimensions or selecting the most relevant features, we will have a bigger effective sample size. Reducing dimensions or selecting features is a random process that depends on data, and we have ignored this fact so far. In Sect. 5, we consider the randomness of the dimensionality reduction or feature selection step to prove that we can increase the effective sample size by following these procedures before conducting covariate shift adaptation.

4.1 A toy experiment

In this section, we present a toy experiment in order to illustrate the relationship between effective sample size, Rényi divergence, dimensionality, and performance of supervised methods.

Assume there are two joint distributions of features and labels P_λ and Q with densities p_λ and q, being the case that Q describes the source/training population and that P_λ describes the target/test population. Moreover, we assume that we are facing the classical covariate shift problem, that is, $p_\lambda(y|\boldsymbol{x}) = q(y|\boldsymbol{x}) = p(y|\boldsymbol{x})$ but $p_\lambda(\boldsymbol{x}) \neq q(\boldsymbol{x})$, plus the fact that we cannot sample the labels from the test population. Finally, consider $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \boldsymbol{I}_d)$ and $p_\lambda(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\lambda \cdot \mathbf{1}, \boldsymbol{I}_d)$, for $\lambda \neq 0$, with d indicating the number of dimensions. Suppose $p(y|\boldsymbol{x}) = \mathcal{N}(y|100 \cdot x_1, 1)$, that is, y depends on $\mathbf{x}$ only through its first coordinate x_1.

Firstly, we calculate $D_2(P_\lambda||Q)$ and $\text{ESS}^*(P_\lambda, Q)$ as functions of d and then simulate how the predictive power of a decision tree regressor deteriorates as d increases and $\text{ESS}^*(P_\lambda, Q)$ decreases. We train the trees by minimizing the empirical error weighted by the true weighting function w in the training set, also imposing a minimum of 10 samples per leaf as a regularization strategy. We choose to work with decision trees since they are fast to train and robust against irrelevant features. Thus, it is reasonable to expect that a great part of performance deterioration is not due to noisy features but because of small ESSs.

The first step to calculate $\text{ESS}^*(P_\lambda, Q)$ and $D_2(P_\lambda||Q)$ is to calculate $\exp[D_2(P_\lambda||Q)]$:

$$\exp[D_2(P_\lambda||Q)] = \mathbb{E}_{\mathbf{x}\sim P_\lambda}\left[\frac{p_\lambda(\mathbf{x})}{q(\mathbf{x})}\right] \tag{14}$$

$$= \mathbb{E}_{\mathbf{x}\sim P_\lambda}\left\{\frac{\exp[-\frac{1}{2}(\mathbf{x}-\lambda\mathbf{1})^\top(\mathbf{x}-\lambda\mathbf{1})]}{\exp[-\frac{1}{2}\mathbf{x}^\top\mathbf{x}]}\right\} \tag{15}$$

$$= \exp\left(-\frac{d\lambda^2}{2}\right) \cdot \mathbb{E}_{\mathbf{x}\sim P_\lambda}\left[\exp\left(\lambda\sum_{j=1}^{d}\mathrm{x}_j\right)\right] \tag{16}$$

$$= \exp(d\lambda^2) \tag{17}$$

The last equality is true since $\exp(\lambda\sum_{j=1}^{d}\mathrm{x}_j) \sim \text{LogNormal}(d\lambda^2, d\lambda^2)$. Then, $D_2(P_\lambda||Q) = d\lambda^2$ and $\text{ESS}^*(P_\lambda, Q) = \exp(-d\lambda^2)$.

Figure 1 depicts the behavior of Rényi Divergence and $\text{ESS}^*(P_\lambda, Q)$ as functions of d. We also vary the value for λ. Given that $D_2(P_\lambda||Q)$ only depends on $|\lambda|$ and not on $\text{sign}(\lambda)$, we consider the case where $\lambda > 0$. When $|\lambda|$ is bigger, the divergence between the source and target distributions also increases. Finally, to check how large d affects performance of a regressor, we, for each d, (i) sample 50 training and test sets of size 10^6, (ii) train the trees on the training set minimizing the weighted empirical error and (iii) assess the regressors on the test sets. The third plot of Fig. 1 represents the average root-mean-square test error ($\pm$ standard deviation). Clearly the regressor deteriorates as the divergence between domains grows and the ESS decreases.

5 The use of dimensionality reduction/ feature selection to make effective sample size bigger

In this section, we present dimensionality reduction and feature selection as ways to obtain a bigger effective sample size. The two main results of this section are given by Theorems 4 and 5. We show that linear dimensionality reduction and feature selection, under some conditions, decrease Rényi divergence between the target and source probability distributions, leading to a bigger effective sample size. This result accounts for the dimensionality

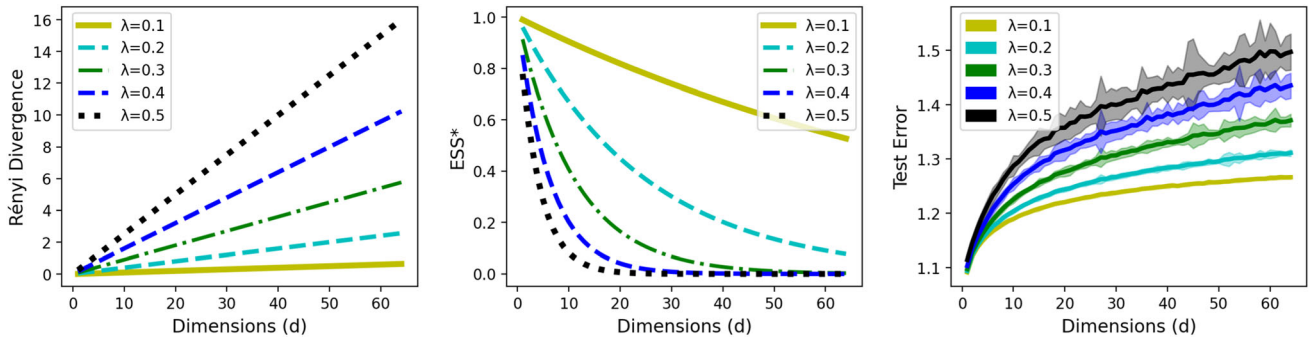

Fig. 1 (i) We plot the Rényi Divergence of the target distribution P_λ from the source distribution Q as a function of the number of features. Both distributions are normal with the same covariance matrix but located $\sqrt{d\lambda^2}$ units apart from each other, i.e., the divergence also depends on $|\lambda|$; (ii) We plot the $\text{ESS}^*(P_\lambda, Q)$ as a function of d and also varying λ. As expected, $\text{ESS}^*(P_\lambda, Q)$ exponentially decays in d as long as the divergence is linearly related with d; (iii) in 50 simulations for each pair (λ, d), we observe how decision trees' performances deteriorate as the divergence between domains grows and the ESS decreases

reduction or feature selection's randomness; that is, the transformation can depend on data in some specific ways.

To arrive at our main results, we first show the intermediate result given by Lemma 1. In the following result, $\boldsymbol{A}$ represents a constant dimensionality reduction matrix and the vector $\boldsymbol{b}$ represents a translation in data before dimensionality reduction, which is common when performing principal components analysis (PCA) [28], for example. When there is no need for considering a translation, we just can adopt $\boldsymbol{b} = \mathbf{0}$. Also, $\boldsymbol{A}$ can represent a feature selector, as we explain in the coming paragraphs.

Lemma 1 *Consider* (*i*) *two absolutely continuous random vectors* $\mathbf{x} \sim Q_\mathbf{x}$ *and* $\mathbf{x}' \sim P_\mathbf{x}$ *of size* $d \geq 2$, $D_2(P_\mathbf{x}||Q_\mathbf{x}) < \infty$, (*ii*) *a nonrandom constant vector* $\boldsymbol{b} \in \mathbb{R}^d$, *and* (*iii*) *a nonrandom constant matrix* $\boldsymbol{A} \in \mathbb{R}^{d' \times d}$ *with rank* d' (*and* $d' \leq d$). *Suppose* $Q_\mathbf{x}$ *and* $P_\mathbf{x}$ *measure events in* $\mathcal{X} \subseteq \mathbb{R}^d$, $d \geq 2$, *and have probability density functions* $q_\mathbf{x}$ *and* $p_\mathbf{x}$, *such that* $\text{support}(p_\mathbf{x}) \subseteq \text{support}(q_\mathbf{x})$. *Also, assume* $\boldsymbol{A}(\mathbf{x} - \boldsymbol{b}) \sim Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}$ *and* $\boldsymbol{A}(\mathbf{x}' - \boldsymbol{b}) \sim P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}$. *Then,*

$$D_2(P_\mathbf{x}||Q_\mathbf{x}) \geq D_2(P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}||Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}) \tag{18}$$

And, consequently,

$$ESS^*(P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}, Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}) \geq ESS^*(P_\mathbf{x}, Q_\mathbf{x}) \tag{19}$$

The proof can be found in the "Appendix". Like Theorem 3, this result can be seen as a particular case of the Data Processing Inequality [26].

Although Lemma 1 gives us a way out in cases which the dimensionality reduction is not random, this case is not realistic. We know that, in practice, $\boldsymbol{A}$ and $\boldsymbol{b}$ are obtained using data.

In the next results, linear dimensionality reduction and feature selection are represented by the random matrix $\mathbf{A}$. If we assume in advance that $\mathbf{A}$ is absolutely continuous, then it represents an ordinary dimensionality reduction matrix. On the other hand, if $\mathbf{A}$ is composed of zeros except for a single entry in each of its columns, which is given by one, then it represents a feature selector. Also, we can consider a random data translator $\mathbf{b}$ instead of the deterministic $\boldsymbol{b}$.

Theorem 4 (*Linear dimensionality reduction*) *Firstly, consider the training random samples of absolutely continuous vectors* $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} Q_\mathbf{x}$ *and an absolutely continuous random vector from target domain* $\mathbf{x}' \sim P_\mathbf{x}$. *Assume* $Q_\mathbf{x}$ *and* $P_\mathbf{x}$ *measure events in* $\mathcal{X} \subseteq \mathbb{R}^d$, $d \geq 2$, *and have probability density functions* $q_\mathbf{x}$ *and* $p_\mathbf{x}$, *such that* $\text{support}(p_\mathbf{x}) \subseteq \text{support}(q_\mathbf{x})$. *Also, assume that* $D_2(P_\mathbf{x}||Q_\mathbf{x}) < \infty$. *Secondly, consider an absolutely continuous random vector* $\mathbf{b} \in \mathbb{R}^d$ *and an absolutely continuous random matrix* $\mathbf{A} \in \mathbb{R}^{d' \times d}$, $rank(\boldsymbol{A}) = d'$, *jointly distributed according to the p.d.f.* $p_{\mathbf{b},\mathbf{A}}$, *such that* $(\mathbf{b}, \mathbf{A}), \mathbf{x}_i$, *and* $\mathbf{x}'$ *are pairwise independent, for every* $i \in [n]$. *Assume that* $d' \leq d$. *Suppose* $\mathbf{A}(\mathbf{x}_i - \mathbf{b}) \sim Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}$ *and* $\mathbf{A}(\mathbf{x}' - \mathbf{b}) \sim P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}$, *for every* $i \in [n]$, *then*

$$D_2(P_\mathbf{x}||Q_\mathbf{x}) \geq D_2(P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}||Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}) \tag{20}$$

And, consequently,

$$ESS^*(P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}, Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}) \geq ESS^*(P_\mathbf{x}, Q_\mathbf{x}) \tag{21}$$

The proof can be found in the "Appendix".

Theorem 4 tells us that a dimensionality reduction procedure before performing covariate shift adaptation increases the population effective sample size. It is important to state that Theorem 4 also holds when disconsidering $\mathbf{b}$ and the proof's adaptation is straightforward. In that case, we would have that $D_2(P_\mathbf{x}||Q_\mathbf{x}) \geq D_2(P_{\mathrm{A}\mathbf{x}}||Q_{\mathrm{A}\mathbf{x}})$ and $ESS^*(P_{\mathrm{A}\mathbf{x}}, Q_{\mathrm{A}\mathbf{x}}) \geq ESS^*(P_\mathbf{x}, Q_\mathbf{x})$.

Next, in Theorem 5, we state a result regarding feature selection.

Theorem 5 (*Feature selection*)*Firstly, consider the training random samples of absolutely continuous vectors* $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} Q_{\mathbf{x}}$ *and an absolutely continuous random vector from target domain* $\mathbf{x}' \sim P_{\mathbf{x}}$. *Assume* $Q_{\mathbf{x}}$ *and* $P_{\mathbf{x}}$ *measure events in* $\mathcal{X} \subseteq \mathbb{R}^d$, $d \geq 2$, *and have probability density functions* $q_{\mathbf{x}}$ *and* $p_{\mathbf{x}}$, *such that* $\text{support}(p_{\mathbf{x}}) \subseteq \text{support}(q_{\mathbf{x}})$. *Also, assume that* $D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) < \infty$. *Secondly, consider a discrete random matrix* $\mathbf{A} \in \mathbb{R}^{d' \times d}$, *that represents a feature selector with* $rank(\mathbf{A}) = d'$, *distributed according to the probability mass function* (*p.m.f.*) p_{A}, *such that* $\mathbf{A}, \mathbf{x}_i$, *and* $\mathbf{x}'$ *are pairwise independent, for every* $i \in [n]$. *Assume that* $d' \leq d$. *Suppose* $\mathbf{A}\mathbf{x}_i \sim Q_{\mathbf{A}\mathbf{x}}$ *and* $\mathbf{A}\mathbf{x}' \sim P_{\mathbf{A}\mathbf{x}}$, *for every* $i \in [n]$, *then*

$$D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \geq D_2(P_{\mathbf{A}\mathbf{x}}||Q_{\mathbf{A}\mathbf{x}}) \tag{22}$$

And, consequently,

$$ESS^*(P_{\mathbf{A}\mathbf{x}}, Q_{\mathbf{A}\mathbf{x}}) \geq ESS^*(P_{\mathbf{x}}, Q_{\mathbf{x}}) \tag{23}$$

The proof can be found in the "Appendix".

Theorems 4 and 5 hold when the data used to obtain $\mathbf{A}$ and $\mathbf{b}$ do not depend on training data that will be used to train the supervised models or data points that represent the target domain we want to make generalizations for. That does not mean we cannot use some portion of the dataset to obtain $\mathbf{A}$ and $\mathbf{b}$, but it only means the results are not valid for those specific used data points, being from source or target domains.

Before closing this section, it is worth mentioning three points. Firstly, at the same time dimensionality reduction/feature selection solve the problem of low effective sample sizes, it might impose other problems. For example, when performing principal components analysis (PCA) [28] for dimensionality reduction, it is not guaranteed the method will not discard useful information for the supervised task. Also, it is not even possible to ensure the covariate shift main assumption, that the conditional distribution of the labels are the same in source and target domains, still holds. In this direction, [18] offers a clever solution to overcome these specific problems, applying sufficient dimension reduction (SDR), which is a supervised method, to reduce dimensions. Secondly, given that $\mathbf{A}$ and $\mathbf{b}$ are random quantities[6], $\{\mathbf{A}(\mathbf{x}_i - \mathbf{b})\}_{i=1}^n$ or $\{\mathbf{A}\mathbf{x}_i\}_{i=1}^n$ may not form independent samples, even when $\mathbf{x}_i \perp\!\!\!\perp (\mathbf{A}, \mathbf{b}), \forall i \in [n]$, and $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} Q_{\mathbf{x}}$. If samples are not independent, then the results presented in Sect. 3 might not hold. Finally, it is true that the results presented in this section can be extended to include more general dimensionality reduction transformations, i.e., nonlinear transformations, and the validity of other transformations might be proven using the Data Processing Inequality [26]. Unfortunately, exploring the two last points is beyond the scope of the present paper and might be treated in future work.

[6] This is not true when $\mathbf{A}$ and $\mathbf{b}$ are fixed.

6 Numerical experiments with real data

In this section, we present regression and classification experiments in which we perform feature selection before covariate shift adaptation. When designing the experiments, we choose to work with the least possible number of assumptions, searching for evidence that the theoretical results presented so far can be extended to more general cases, which will be treated in future work. Namely, we did not assume (i) the true importance function is always known, (ii) that training data are independent of the feature selector, and (iii) that training data are formed with independent data points after the feature selection procedure.

For the following experiments, 10 regression datasets with no missing values were selected[7]. Each experiment consisted of (i) introducing covariate shift[8], (ii) estimating the weights, (iii) correcting the shift by the importance weighting method, and finally (iv) assessing the performance of the predictors and the effective sample size. We also studied the classification case by binarizing the target variables using their medians as a threshold. We use the same datasets for both regression and classification experiments to make comparisons easier. For each one of the 10 datasets, we repeated the following preprocessing steps: (i) we kept up to 8000 data points per dataset[9], (ii) generated new features using independent standard Gaussian noise and (iii) standardized each column in every dataset. By augmenting the dataset to 32 features using noise, we can explore a scenario in which performance deterioration is mainly due to small effective sample sizes. We give more details on this point in the next paragraph.

The following procedure is used to create divergent training and test sets after the preprocessing steps. For each of the datasets, we sampled a sequence of vectors uniformly from $[-1, 1]^d$. We projected the data points onto the subspace generated by each vector, resulting in only one feature $\boldsymbol{x}_i^{(j)}$ per sample i for each subspace/simulation j. For

[7] From www.dcc.fc.up.pt/∼ltorgo/Regression/DataSets.html and https://archive.ics.uci.edu/ml/datasets.php.

[8] Similarly to previous research, e.g., [9, 12, 16, 18].

[9] The datasets "Abalone," "Delta Ailerons," and "Wine Quality" had 4177, 7129, and 6497 data points, respectively. All the others were undersampled to have 8000 data points.

each $\boldsymbol{x}_i^{(j)}$, we calculated the score $s_{ij} = \Phi\left([x_i^{(j)} - \text{median}(\boldsymbol{x}^{(j)})]/\sigma_j\right)$, which is the probability that the data point i from simulation j is in the training set. According to that score, we randomly allocated each data point in either the training or test set in simulation j. The constant σ_j was adjusted until the empirical effective sample size, as defined in Sect. 3, is less than 0.01. Following this procedure, the training and test sets are approximately of the same sizes in each simulation j. Then, we fit two decision trees for each of the training/test sets: one in the training set and one in a subset of the test set. Then, we tested both decision trees in the unused portion of the test set and compared their performance according to the mean squared error for regression and classification error (1 - accuracy) for classification. We selected the 75 simulations[10] in which decision trees trained in the test sets did best, relatively to the training set trees. We chose decisiontrees because they are fast to train and robust against irrelevant features. Thus, the noisy features added in the datasets are not likely to directly affect predictive power but only by making the effective sample size smaller. It is important to state that, during the whole experimenting phase, decisions trees were twofold cross-validated in order to choose the minimum number of samples per leaf[11].

For the feature selection step, we were inspired by [18] and the idea of Sufficient Dimension Reduction [29], which is a supervised approach to dimensionality reduction and feature selection, contrasting to Principal Component Analysis, for example. Supervised approaches to dimensionality reduction and feature selection are preferable since we are able to keep important information for a supervised task performed afterwards. Using training data, we apply a combination of the methods described by [30, 31] and the *Forward Selection* algorithm [32]. The approach uses gaussian mixture models (GMMs) to estimate, using the whole training set, the mutual information between a subset of features and the target variable. In this case, the number of GMMs' components are chosen evenly splitting the training data and performing a simple holdout set hyperparameter tuning phase[12]. After training the GMMs, the procedure follows these steps: we start by choosing the feature that has the largest estimated mutual information with the target variable, and, at each subsequent step, we select the feature that marginally maximizes the estimated mutual information of target variable and selected features. We repeat the process until we reach a stop criteria. Our stopping criteria is that we should stop selecting features when the marginal improvement in the empirical mutual information is less than 1% relative to the last level or when we select the first 15 features. An implementation of the feature selection method is available in the Python package *InfoSelect*[13].

To estimate the weighting function for covariate shift adaptation, we use the probabilistic classification approach [8, 33] with a logistic regression model combined with a quadratic polynomial expansion of the original features. We choose to work with this approach since it is simple and fast to implement, besides being promising for high-dimensional settings. Others approaches are possible though [8]. In order to prepare the data for training the logistic regression model, we first append the whole training set and randomly select rows (80%) from the test set, and create the artificial labels for both groups. Then, we randomly/evenly split that dataset in order to choose the best value for the $l1$ regularization hyperparameter of the logistic regression, using the simple holdout validation approach[14]. After getting the optimal values for the hyperparameter, we train a final model using the whole appended dataset.

In the experiments, we work with four training scenarios. In the first one, we use the whole set of features and no weighting method. In the second one, we use the entire set of features and importance weighting combined with the "true" weights $(1 - s_{ij})/s_{ij}$. In the third, we use the whole set of features and estimated weights using the probabilistic classification approach. In the fourth scenario, we use only selected features and estimated weights using the probabilistic classification approach. Comparing the four scenarios enables us to see how importance weighting may fail in high-dimensional settings due to low ESS, even when we know the "true" weighting function.

Table 1 shows, for each one of the employed datasets, (i) the original number of features, (ii) the augmented number of features, (iii) the average number (± standard deviation) of selected features for the regression and (iv) classification experiments.

In Fig. 2, one can see the distribution of effective sample sizes in all the weighted approaches, calculated in the entire set of experiments. It is possible to notice how small the ESSs can be by adopting the pure weighting strategy. The feature selection step allows bigger ESSs.

In Table 2, we see the average test errors (± standard deviation). To compute the errors, we use the test set portion (20%) not used to train the importance function.

[10] From the total of 7200 simulations.

[11] More details on hyperparameter tuning can be found in the "Appendix".

[12] More details on hyperparameter tuning can be found in the "Appendix".

[13] See https://github.com/felipemaiapolo/infoselect or https://pypi.org/project/infoselect/.

[14] More details on hyperparameter tuning can be found in the "Appendix"

Table 1 Average Numbers of features (± standard deviation) - in this table we compare the numbers of original, augmented and selected features for regression (reg.) and classification (class.) tasks

Dataset	Original	Augmented	Selected (Reg.)	Selected (Class.)
Abalone	7	32	4.19 ± 1.26	9.87 ± 5.64
Ailerons	40	40	5.16 ± 0.54	3.79 ± 0.64
Bank32nh	32	32	10.00 ± 1.82	13.91 ± 0.61
Cal housing	8	32	5.29 ± 1.29	7.45 ± 4.92
CPU act	21	32	9.88 ± 1.20	2.56 ± 0.72
Delta ailerons	5	32	3.16 ± 0.49	3.75 ± 0.63
Elevators	18	32	7.97 ± 1.11	13.08 ± 2.16
Fried delve	10	32	4.45 ± 0.50	5.00 ± 0.00
Puma32H	32	32	1.88 ± 0.32	14.00 ± 0.00
Wine quality	11	32	9.60 ± 1.02	14.00 ± 0.00

It is possible to note that, on average, we select small subsets of features, even smaller than the original sets

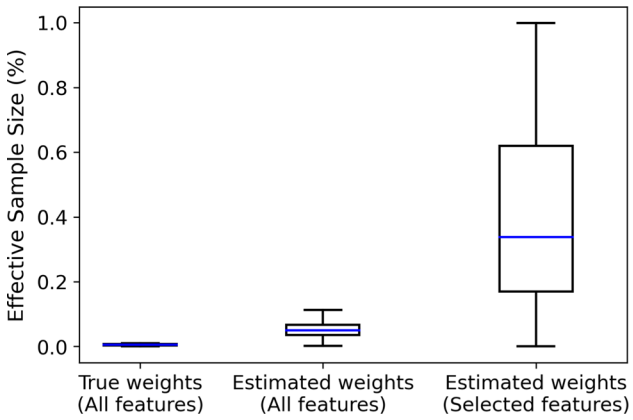

Fig. 2 Effective Sample Size distributions across all experiments. Notice higher ESSs can be achieved by a prior feature selection stage

The errors reported are the mean squared error and classification error relative to the first scenario. From Table 2, it is noticeable that our feature selection approach and posterior weighting systematically outperforms all the other benchmarks, especially the pure weighting method when the whole set of features is used. Even the benchmarks that used true weights are often beaten by large margins. That suggests that poor models' performances are mainly due to small effective sample sizes instead of difficulties estimating the weighting function.

Through our experiments, we were able to verify that the feature selection stage tends to increase the effective sample size, consequently allowing better performance of supervised methods.

7 Conclusion

In this paper, we have made two main contributions. The first is that we explicitly and formally connected three key concepts in the context of covariate shift adaptation: (i) effective sample size, (ii) dimensionality of data, and (iii) generalization of a supervised model. Since, to the best of our knowledge, there is no unified and rigorous view on how the three key concepts connect to each other, we consider this to be the first contribution of the paper. The second contribution of the paper is that we show dimensionality reduction or feature selection, even considering data dependent mappings, corrects small effective sample sizes by making the source and target distributions less divergent. This suggests that it is a good practice to perform dimensionality reduction or feature selection before covariate adaptation. We also present numerical experiments using real and artificial data to complement our theoretical results.

Regarding possible future research paths and improvements, we point to Sects. 3 and 5. Concerning Sect. 3, perhaps the three most relevant points to be considered for future research relate to the following assumptions: the first one is assuming the importance function is known up to a constant, the second is assuming the sample ESS is close to its population version, and the third is assuming independent samples. While the first hardly applies in practice, the second may hold in many situations, and the third could be relaxed to include dependent samples, thus solving one of the problems discussed in Sect. 5. Considering Sect. 5, we think there is one main point to be explored in future work, which is extending the theorems to include more general transformations, i.e., nonlinear or training data dependent transformations. Said that, future work and improvements could focus on relaxing assumptions.

8 Computing infrastructure

All the experiments were carried out using a Google Cloud Platform's (GCP) Virtual Machine with 96 vCPUs and 86.4 GB of memory. All the experiments took around 4 h to run.

Table 2 Average Test Errors (± std. deviation) - here we compared the predictive performance of decision trees in the test set of 75 different simulations for each dataset. We have four basic scenarios: (i) whole set of features and no weighting method; (ii) whole set of features and use of "true" weights; (iii) whole set of features and estimated weights; (iv) selected features and estimated weights

		All features			Selected features
	Dataset	Unweighted	True weights	Estimated weights	Estimated weights
Regression	Abalone	1.00	1.42 ± 0.24	1.25 ± 0.19	$\mathbf{0.92 \pm 0.07}$
	Ailerons	1.00	1.01 ± 0.13	0.99 ± 0.11	$\mathbf{0.87 \pm 0.11}$
	Bank32nh	1.00	1.29 ± 0.14	1.20 ± 0.11	$\mathbf{0.98 \pm 0.06}$
	Cal housing	1.00	1.50 ± 0.24	1.35 ± 0.20	$\mathbf{0.84 \pm 0.09}$
	CPU act	1.00	0.52 ± 0.55	0.55 ± 0.59	$\mathbf{0.15 \pm 0.21}$
	Delta ailerons	1.00	1.39 ± 0.18	1.25 ± 0.12	$\mathbf{0.92 \pm 0.06}$
	Elevators	1.00	1.10 ± 0.15	1.05 ± 0.13	$\mathbf{0.85 \pm 0.15}$
	Fried delve	1.00	1.60 ± 0.22	1.40 ± 0.15	$\mathbf{0.90 \pm 0.11}$
	Puma32H	**1.00**	2.24 ± 1.18	1.45 ± 0.22	1.77 ± 2.42
	Wine quality	1.00	1.31 ± 0.12	1.24 ± 0.11	$\mathbf{0.97 \pm 0.04}$
Classification	Abalone	**1.00**	1.29 ± 0.19	1.22 ± 0.16	1.05 ± 0.15
	Ailerons	1.00	1.03 ± 0.27	1.01 ± 0.20	$\mathbf{0.86 \pm 0.13}$
	Bank32nh	**1.00**	1.25 ± 0.13	1.20 ± 0.13	$\mathbf{1.00 \pm 0.09}$
	Cal housing	1.00	1.43 ± 0.23	1.36 ± 0.19	$\mathbf{0.87 \pm 0.14}$
	CPU act	1.00	1.09 ± 0.16	1.06 ± 0.16	$\mathbf{0.99 \pm 0.15}$
	Delta ailerons	1.00	1.38 ± 0.40	1.25 ± 0.31	$\mathbf{0.84 \pm 0.12}$
	Elevators	1.00	1.07 ± 0.15	1.04 ± 0.14	$\mathbf{0.89 \pm 0.13}$
	Fried delve	1.00	1.34 ± 0.22	1.22 ± 0.18	$\mathbf{0.85 \pm 0.09}$
	Puma32H	**1.00**	1.73 ± 0.59	1.22 ± 0.18	1.10 ± 0.42
	Wine quality	**1.00**	1.20 ± 0.13	1.13 ± 0.10	1.07 ± 0.10

The numbers reported are the mean squared error and classification error averages and their std. deviations. All the results were normalized w.r.t. the first scenario

Appendix

Proofs and derivations

Proof of Theorem 1

Proof Assume the hypothesis stated are valid. Being $c \neq 0$ a real constant, see we can re-wright the ESS as follows:

$$\widehat{ESS}_n(P_{\mathbf{x}}, Q_{\mathbf{x}}) = \frac{(\sum_{i=1}^n \mathrm{w}_i)^2}{n\sum_{i=1}^n \mathrm{w}_i^2} = \frac{\left[\sum_{i=1}^n c \cdot \frac{p_{\mathbf{x}}(\mathbf{x}_i)}{q_{\mathbf{x}}(\mathbf{x}_i)}\right]^2}{n\sum_{i=1}^n \left[c \cdot \frac{p_{\mathbf{x}}(\mathbf{x}_i)}{q_{\mathbf{x}}(\mathbf{x}_i)}\right]^2} = \frac{\left[\frac{1}{n}\sum_{i=1}^n \frac{p_{\mathbf{x}}(\mathbf{x}_i)}{q_{\mathbf{x}}(\mathbf{x}_i)}\right]^2}{\frac{1}{n}\sum_{i=1}^n \left[\frac{p_{\mathbf{x}}(\mathbf{x}_i)}{q_{\mathbf{x}}(\mathbf{x}_i)}\right]^2}$$

Then, by the Strong Law of Large Numbers and almost-sure convergence properties [34], we verify that $\widehat{ESS}_n(P_{\mathbf{x}}, Q_{\mathbf{x}}) \xrightarrow{a.s.} \frac{\mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right]^2}{\mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\left(\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right)^2\right]}$ when $n \to \infty$. To complete the proof, we state the following

$$\frac{\mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right]^2}{\mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\left(\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right)^2\right]} = \frac{1}{\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}}}\left[\frac{p_{\mathbf{x}}(\mathbf{x})}{q_{\mathbf{x}}(\mathbf{x})}\right]} = \frac{1}{exp[D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})]} = \mathrm{ESS}^*(P_{\mathbf{x}}, Q_{\mathbf{x}})$$

□

Proof of Theorem 3

Proof Assume the hypothesis are valid and let $d_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2}) = \exp[D_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2})]$. See that:

$$
\begin{aligned}
d_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2}) &= \mathbb{E}_{P_{\mathbf{x}_1,\mathbf{x}_2}}\left[\frac{p_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{x}_1,\mathbf{x}_2)}{q_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{x}_1,\mathbf{x}_2)}\right] \\
&= \mathbb{E}_{P_{\mathbf{x}_1}}\left[\frac{p_{\mathbf{x}_1}(\mathbf{x}_1)}{q_{\mathbf{x}_1}(\mathbf{x}_1)} \cdot \mathbb{E}_{P_{\mathbf{x}_2|\mathbf{x}_1}}\left[\frac{p_{\mathbf{x}_2|\mathbf{x}_1}(\mathbf{x}_2|\mathbf{x}_1)}{q_{\mathbf{x}_2|\mathbf{x}_1}(\mathbf{x}_2|\mathbf{x}_1)}\right]\right] \\
&= \mathbb{E}_{P_{\mathbf{x}_1}}\left[\frac{p_{\mathbf{x}_1}(\mathbf{x}_1)}{q_{\mathbf{x}_1}(\mathbf{x}_1)} \cdot d_2(P_{\mathbf{x}_2|\mathbf{x}_1}||Q_{\mathbf{x}_2|\mathbf{x}_1})\right] \\
&\geq \mathbb{E}_{P_{\mathbf{x}_1}}\left[\frac{p_{\mathbf{x}_1}(\mathbf{x}_1)}{q_{\mathbf{x}_1}(\mathbf{x}_1)}\right] = d_2(P_{\mathbf{x}_1}||Q_{\mathbf{x}_1})
\end{aligned}
$$

The inequality is obtained by the fact that the exponential of the Rényi Divergence must be greater or equal to one. To complete the proof, see that $ESS^*(P_{\mathbf{x}_1,\mathbf{x}_2}, Q_{\mathbf{x}_1,\mathbf{x}_2}) = d_2(P_{\mathbf{x}_1,\mathbf{x}_2}||Q_{\mathbf{x}_1,\mathbf{x}_2})^{-1}$. Therefore,

$$
ESS^*(P_{\mathbf{x}_1}, Q_{\mathbf{x}_1}) \geq ESS^*(P_{\mathbf{x}_1,\mathbf{x}_2}, Q_{\mathbf{x}_1,\mathbf{x}_2})
$$

□

Proof of Lemma 1

Proof If $d = d'$, the result is direct, considering the arguments used by Qiao and Minematsu [35] to prove their Theorem 1, because $\boldsymbol{A}$ represents an invertible linear (and differentiable) transformation. Otherwise, consider a full rank matrix $\boldsymbol{C} = \begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix} \in \mathbb{R}^{d\times d}$, where $\boldsymbol{B} \in \mathbb{R}^{d-d',d}$. Given that $\boldsymbol{C}$ is full rank, it represents an invertible linear (and differentiable) transformation. If $\boldsymbol{C}(\mathbf{x}-\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})\\ \boldsymbol{B}(\mathbf{x}-\boldsymbol{b})\end{bmatrix} \sim Q_{\boldsymbol{C}(\mathbf{x}-\boldsymbol{b})}$ and $\boldsymbol{C}(\mathbf{x}'-\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{A}(\mathbf{x}'-\boldsymbol{b})\\ \boldsymbol{B}(\mathbf{x}'-\boldsymbol{b})\end{bmatrix} \sim P_{\boldsymbol{C}(\mathbf{x}'-\boldsymbol{b})}$, then by the arguments used by Qiao and Minematsu [35] to prove[15] their Theorem 1 we have that $D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) = D_2(P_{\boldsymbol{C}(\mathbf{x}-\boldsymbol{b})}||Q_{\boldsymbol{C}(\mathbf{x}-\boldsymbol{b})})$. Discarding $\boldsymbol{B}(\mathbf{x}-\boldsymbol{b})$ and $\boldsymbol{B}(\mathbf{x}'-\boldsymbol{b})$ from random vectors $\boldsymbol{C}(\mathbf{x}-\boldsymbol{b})$ and $\boldsymbol{C}(\mathbf{x}'-\boldsymbol{b})$, by Theorem 3, we have that

$$
D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \geq D_2(P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}||Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})})
$$

Therefore,

$$
\begin{aligned}
&ESS^*(P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}, Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}) = \exp\left[-D_2(P_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})}||Q_{\boldsymbol{A}(\mathbf{x}-\boldsymbol{b})})\right] \\
&\geq \exp[-D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})] = ESS^*(P_{\mathbf{x}}, Q_{\mathbf{x}})
\end{aligned}
$$

□

Proof of Theorem 4

Proof Firstly, we define $\mathbf{v} := \mathbf{A}(\mathbf{x}_i - \mathbf{b}) \sim Q_{\mathbf{v}} \equiv Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}$ and $\mathbf{u} := \mathbf{A}(\mathbf{x}' - \mathbf{b}) \sim P_{\mathbf{u}} \equiv P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}$, for an arbitrary $i \in [n]$. Let $q_{\mathbf{v}}$ and $p_{\mathbf{u}}$ be probability density functions associated with distributions $Q_{\mathbf{v}}$ and $P_{\mathbf{u}}$. From Lemma 1, we know that $D_2\left(P_{\mathbf{u}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}), \forall \boldsymbol{b} \in \mathbb{R}^d, \ \forall \boldsymbol{A} \in \mathbb{R}^{d'\times d}$ such that $\mathrm{rank}(\boldsymbol{A}) = d'$. That statement implies the following:

$$
\begin{aligned}
&D_2\left(P_{\mathbf{u}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \exp D_2\left(P_{\mathbf{u}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{b}=\boldsymbol{b},\mathbf{A}=\boldsymbol{A}}\right) \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \mathbb{E}_{p_{\mathbf{b},\mathbf{A}}}\left[\exp D_2\left(P_{\mathbf{u}|\mathbf{b},\mathbf{A}}||Q_{\mathbf{v}|\mathbf{b},\mathbf{A}}\right)\right] \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \int p_{\mathbf{b},\mathbf{A}}(\boldsymbol{b},\boldsymbol{A}) \int p_{\mathbf{u}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A}) \frac{p_{\mathbf{u}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A})}{q_{\mathbf{v}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A})} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{b}\mathrm{d}\boldsymbol{A} \\
&\leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \int p_{\mathbf{u}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A}) p_{\mathbf{b},\mathbf{A}}(\boldsymbol{b},\boldsymbol{A}) \frac{p_{\mathbf{u}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A})}{q_{\mathbf{v}|\mathbf{b},\mathbf{A}}(\boldsymbol{u}|\boldsymbol{b},\boldsymbol{A})} \frac{p_{\mathbf{b},\mathbf{A}}(\boldsymbol{b},\boldsymbol{A})}{p_{\mathbf{b},\mathbf{A}}(\boldsymbol{b},\boldsymbol{A})} \\
&\mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{b}\mathrm{d}\boldsymbol{A} \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow D_2\left(P_{\mathbf{u},\mathbf{b},\mathbf{A}}||Q_{\mathbf{v},\mathbf{b},\mathbf{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow D_2\left(P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}||Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}\right) = D_2(P_{\mathbf{u}}||Q_{\mathbf{v}}) \\
&\leq D_2\left(P_{\mathbf{u},\mathbf{b},\mathbf{A}}||Q_{\mathbf{v},\mathbf{b},\mathbf{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})
\end{aligned}
$$

The last step is due to Theorem 3 (extending to random matrices). To complete the proof, we state the following:

$$
\begin{aligned}
&ESS^*(P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}, Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})}) = \exp\left[-D_2(P_{\mathrm{A}(\mathbf{x}-\mathbf{b})}||Q_{\mathrm{A}(\mathbf{x}-\mathbf{b})})\right] \\
&\geq \exp[-D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})] = ESS^*(P_{\mathbf{x}}, Q_{\mathbf{x}})
\end{aligned}
$$

□

Proof of Theorem 5

Proof Firstly, we define $\mathbf{v} := \mathbf{A}\mathbf{x}_i \sim Q_{\mathbf{v}} \equiv Q_{\mathrm{A}\mathbf{x}}$ and $\mathbf{u} := \mathbf{A}\mathbf{x}' \sim P_{\mathbf{u}} \equiv P_{\mathrm{A}\mathbf{x}}$, for an arbitrary $i \in [n]$. Let $q_{\mathbf{v}}$ and $p_{\mathbf{u}}$ be probability density functions associated with distributions $Q_{\mathbf{v}}$ and $P_{\mathbf{u}}$. From Lemma 1, we know that $D_2\left(P_{\mathbf{u}|\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{A}=\boldsymbol{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})$, $\forall \boldsymbol{A} \in \mathbb{R}^{d'\times d}$ such that $\mathrm{rank}(\boldsymbol{A}) = d'$. That statement implies the following:

$$
\begin{aligned}
&D_2\left(P_{\mathbf{u}|\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{A}=\boldsymbol{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \exp D_2\left(P_{\mathbf{u}|\mathbf{A}=\boldsymbol{A}}||Q_{\mathbf{v}|\mathbf{A}=\boldsymbol{A}}\right) \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \mathbb{E}_{p_{\mathbf{A}}}\left[\exp D_2\left(P_{\mathbf{u}|\mathbf{A}}||Q_{\mathbf{v}|\mathbf{A}}\right)\right] \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \sum_{\boldsymbol{A}} p_{\mathbf{A}}(\boldsymbol{A}) \int p_{\mathbf{u}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A}) \frac{p_{\mathbf{u}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A})}{q_{\mathbf{v}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A})} \mathrm{d}\boldsymbol{u} \leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow \sum_{\boldsymbol{A}} \int p_{\mathbf{u}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A}) p_{\mathbf{A}}(\boldsymbol{A}) \frac{p_{\mathbf{u}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A})}{q_{\mathbf{v}|\mathbf{A}}(\boldsymbol{u}|\boldsymbol{A})} \frac{p_{\mathbf{A}}(\boldsymbol{A})}{p_{\mathbf{A}}(\boldsymbol{A})} \mathrm{d}\boldsymbol{u} \\
&\leq \exp D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \Rightarrow D_2\left(P_{\mathbf{u},\mathbf{A}}||Q_{\mathbf{v},\mathbf{A}}\right) \leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) \\
&\Rightarrow D_2(P_{\mathrm{A}\mathbf{x}}||Q_{\mathrm{A}\mathbf{x}}) = D_2(P_{\mathbf{u}}||Q_{\mathbf{v}}) \leq D_2\left(P_{\mathbf{u},\mathbf{A}}||Q_{\mathbf{v},\mathbf{A}}\right) \\
&\leq D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})
\end{aligned}
$$

Given the matrix $\boldsymbol{A}$ represents a feature selector, it can only assume a finite number of values. Thus, the sum is given over a finite number of values of $\boldsymbol{A}$. The last step is due to

[15] Even though D_2 is not an f-divergence, the thoughts presented by Qiao and Minematsu [35] in their proof can readily be applied in this case. Furthermore, we can write $D_2(P_{\mathbf{x}}||Q_{\mathbf{x}}) = \log(\chi^2(P_{\mathbf{x}}||Q_{\mathbf{x}}) + 1)$, where χ^2 is a f-divergence [36]. This is an another reason on why this is valid.

the Theorem 3 (extending to random matrices). To complete the proof, we state the following:

$$ESS^*(P_{\mathbf{Ax}}, Q_{\mathbf{Ax}}) = \exp[-D_2(P_{\mathbf{Ax}}||Q_{\mathbf{Ax}})] \geq \exp[-D_2(P_{\mathbf{x}}||Q_{\mathbf{x}})] = ESS^*(P_{\mathbf{x}}, Q_{\mathbf{x}})$$

□

Experiments

In the experiments section, we tune three hyperparameters: (i) $l1$ regularization parameter used to train the logistic regression model when estimating w, (ii) the minimum number of samples per leaf in each regression/classification tree, and (iii) number of GMM components. We use the Scikit-Learn [37] implementations to train the logistic regressions, regression/classification trees and GMMs. Firstly, we choose the $l1$ logistic regression regularization parameter C from values in $[10^{-4}, 5]$, in order to minimize the log loss in a holdout dataset. Secondly, we choose the minimum number of samples per leaf in each regression/classification tree from values in the list (5, 15, 25, 40, 50), in order to minimize the mean squared error or classification error within a twofold cross-validation procedure. Finally, we maximize the log-likelihood in a holdout dataset to choose the number of GMM components, varying the possible number of components within the list (1, 3, 5, 7, 9, 11, 13, 15).

Acknowledgements We would like to thank Evanildo Lacerda Júnior, Felipe Leno da Silva, Raphael Mendes de Oliveira Cobe, Sami Yamouni, and Juan Pablo Ibieta Jimenez for valuable feedback and support.

Funding We gratefully acknowledge financial support from Conselho Nacional de Desenvolvimento Científico and Tecnológico (CNPq) (process 132857/2019-7) and from the Fundunesp/Advanced Institute for Artificial Intelligence (AI2) (process 3061/2019-CCP), Brazil. Felipe Maia Polo was supported by the two institutions during his master's degree while writing this work.

Availability of data and material All the datasets used are open datasets and are downloaded while running the code.

Code availability The code and material used can be found on GitHub (https://github.com/felipemaiapolo/ess_dimensionality_covariate_shift). Also, we have made our feature selection implementation available as a Python package called *InfoSelect* (https://github.com/felipemaiapolo/infoselect).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Wu CL, Chau K-W (2013) Prediction of rainfall time series using modular soft computingmethods. Eng Appl Artif Intell 26(3):997–1007
2. Ashkan B, Amin N, Amin T-G (2020) Deep learning-based appearance features extraction for automated carp species identification. Aquacult Eng 89:102053
3. Cheng X, Feng Z-K, Niu W-J (2020) Forecasting monthly runoff time series by single-layer feedforward artificial neural network and grey wolf optimizer. IEEE Access 8:157346–157355
4. Ghalandari M, Ziamolki A, Mosavi A, Shamshirband S, Chau K-W, Bornassi S (2019) Aeromechanical optimization of first row compressor test stand blades using a hybrid machine learning model of genetic algorithm, artificial neural networks and design of experiments. Eng Appl Comput Fluid Mech 13(1):892–904
5. Fan Y, Kangkang X, Hui W, Zheng Y, Tao B (2020) Spatiotemporal modeling for nonlinear distributed thermal processes based on kl decomposition, mlp and lstm network. IEEE Access 8:25111–25121
6. Taormina R, Chau K-W (2015) Ann-based interval forecasting of streamflow discharges using the lube method and mofips. Eng Appl Artif Intell 45:429–440
7. Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plann Inference 90(2):227–244
8. Masashi S, Motoaki K (2012) Machine learning in non-stationary environments: introduction to covariate shift adaptation. MIT Press, Cambridge
9. Jiayuan H, Arthur G, Karsten B, Bernhard S, Alex JS (2007) Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems, pp 601–608
10. Masashi S, Matthias K, Klaus-Robert M (2007) Covariate shift adaptation by importance weighted cross validation. J Mach Learn Res 8(May):985–1005
11. Takafumi K, Shohei H, Masashi S (2009) A least-squares approach to direct importance estimation. J Mach Learn Res 10(Jul):1391–1445
12. Fulton W, Cynthia R (2017) Extreme dimension reduction for handling covariate shift. arXiv:1711.10938
13. Sugiyama M, Suzuki T, Nakajima S, Kashima H, von Bünau P, Kawanabe M (2008) Direct importance estimation for covariate shift adaptation. Ann Inst Stat Math 60(4):699–746
14. Rafael I, Ann L, Chad S (2014) High-dimensional density ratio estimation with extensions to approximate likelihood computation. In: Artificial intelligence and statistics, pp 420–429
15. Song L, Akiko T, Taiji S, Kenji F (2017) Trimmed density ratio estimation. In: Advances in neural information processing systems, pp 4518–4528
16. Sashank JR, Barnabas P, Alex S (2015) Doubly robust covariate shift correction. In: Twenty-Ninth AAAI conference on artificial intelligence
17. Arthur G, Alex S, Jiayuan H, Marcel S, Karsten B, Bernhard S (2009) Covariate shift by kernel mean matching. Dataset Shift Mach Learn 3(4):5
18. Petar S, Mingming G, Jaime GC, Kun Z (2019) Low-dimensional density ratio estimation for covariate shift correction. Proc Mach Learn Res 89:3449
19. Christian PR, George C, George C (2010) Introducing Monte Carlo methods with r, volume 18. Springer
20. Art BO (2013) Monte Carlo theory, methods and examples.
21. Martino L, Elvira V, Louzada F (2017) Effective sample size for importance sampling based on discrepancy measures. Signal Process 131:386–401

22. Cortes C, Mansour Y, Mohri M (2010) Learning bounds for importance weighting. In: Advances in neural information processing systems, pp 442–450
23. Cortes C, Greenberg S, Mohri M (2019) Relative deviation learning bounds and generalization with unbounded loss functions. Ann Math Artif Intell 85(1):45–70
24. Víctor E, Luca M, Christian PR (2018) Rethinking the effective sample size. arXiv:1809.04129
25. Agapiou S, Omiros P, Sanz-Alonso D, Stuart AM et al (2017) Importance sampling: intrinsic dimension and computational cost. Stat Sci 32(3):405–431
26. Van Erven T, Harremos P (2014) Rényi divergence and Kullback-Leibler divergence. IEEE Trans Inf Theory 60(7):3797–3820
27. Mathukumalli V (2002) A theory of learning and generalization. Springer, Berlin
28. Trevor H, Robert T, Jerome F (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
29. Taiji S, Masashi S (2010) Sufficient dimension reduction via squared-loss mutual information estimation. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 804–811
30. Tian L, Deniz E, Umut O, Yonghong H (2006) Estimating mutual information using gaussian mixture model for feature ranking and selection. In: The 2006 IEEE international joint conference on neural network proceedings, pp 5034–5039. IEEE
31. Emil E, Amaury L, Juha K (2014) Variable selection for regression problems using gaussian mixture models to estimate mutual information. In: 2014 international joint conference on neural networks (IJCNN), pp 1606–1613. IEEE
32. Isabelle G, André E (2003)An introduction to variable and feature selection. J Mach Learn Res 3 (Mar):1157–1182
33. Masashi S, Taiji S, Takafumi K (2012) Density ratio estimation in machine learning. Cambridge University Press, Cambridge
34. George GR (1997) A course in mathematical statistics. Elsevier, Amsterdam
35. Qiao Yu, Minematsu N (2010) A study on invariance of f-divergence and its application to speech recognition. IEEE Trans Signal Process 58(7):3884–3890
36. Sason I, Verdú S (2016) f-divergence inequalities. IEEE Trans InfTheory 62(11):5973–6006
37. Fabian P, Gaël V, Alexandre G, Vincent M, Bertrand T, Olivier G, Mathieu B, Peter P, Ron W, Vincent D et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830