Test for non-negligible adverse shifts

Abstract

Statistical tests for dataset shift are susceptible to false alarms: they are sensitive to minor differences when there is in fact adequate sample coverage and predictive performance. We propose instead a framework to detect adverse shifts based on outlier scores, D-SOS for short. D-SOS holds that the new (test) sample is not substantively worse than the reference (training) sample, and not that the two are equal. The key idea is to reduce observations to outlier scores and compare contamination rates at varying weighted thresholds. Users can define what *worse* means in terms of relevant notions of outlyingness, including proxies for predictive performance. Compared to tests of equal distribution, our approach is uniquely tailored to serve as a robust metric for model monitoring and data validation. We show how versatile and practical D-SOS is on a wide range of real and simulated data.

1 INTRODUCTION

Suppose we fit a predictive model on a training set and predict on a test set. Dataset shift, also known as data or population drift, occurs when training and test distributions are not alike [Kelly et al., 1999, Quionero-Candela et al., 2009]. This is essentially a sample mismatch problem. Some regions of the data space are either too sparse or absent during training and gain importance at test time. We want methods that alert users to the presence of unexpected inputs in the test set [Rabanser et al., 2019]. To do so, a measure of divergence between training and test set is required. Can we not use the many modern off-the-shelf multivariate tests of equal distributions for this?

One reason for moving beyond tests of equal distributions is that they are often too strict. They require high fidelity between training and test set everywhere in the input domain. However, not *all* changes in distribution are a cause for concern – some changes are benign. Practitioners distrust these tests because of false alarms. Polyzotis et al. [2019] comment:

> statistical tests for detecting changes in the data distribution [...] are too sensitive and also uninformative for the typical scale of data in machine learning pipelines, which led us to seek alternative methods to quantify changes between data distributions.

Even when the difference is small or negligible, tests of equal distributions reject the null hypothesis of no difference. Monitoring model performance and data quality is a critical part of deploying safe and mature models in production [Paleyes et al., 2020, Zhang et al., 2020]. An alarm should only be raised if a shift warrants intervention. Retraining models when distribution changes are benign is both costly and ineffective. To tackle these challenges, we propose D-SOS instead.

In comparing the test set to the training set, D-SOS pays more attention to the regions — typically, the outlying regions — where we are most vulnerable. To confront false alarms, it uses a robust test statistic, the weighted area under the receiver operating characteristic curve (WAUC). The weights in the WAUC discount the safe regions of the distribution. As best we know, this is the first time that the WAUC is being used as a test statistic in this context. The goal of D-SOS is to detect *non-negligible adverse shifts*. This is reminiscent of noninferiority tests [Wellek, 2010], widely used in healthcare to determine if a new treatment is in fact not inferior to an older one. Colloquially, D-SOS holds that the new sample is not substantively worse than the old sample, and not that the two are equal.

D-SOS moves beyond tests of equal distributions and lets users specify which notions of outlyingness to probe. The choice of the score function plays a central role in formaliz-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

ing what we mean by *worse*. These scores can come from out-of-distribution (outlier) detection, two-sample classification, uncertainty quantification, residual diagnostics, density estimation, dimension reduction, and more. While some of these scores are underused and underappreciated in two-sample statistical tests, they can be more telling than the density-based scores. In this paper, we make the following contributions:

1. We derive D−SOS, a novel multivariate two-sample test for no adverse shift, from tests of goodness-of-fit. D−SOS computes its p-values based on sample splitting, permutation and/or out-of-bag predictions. The out-of-bag variant is often the most convenient. This approach improves on sample splitting, which sacrifices calibration accuracy for inferential robustness.
2. We compare D−SOS performance to two modern tests of equal distribution in unsupervised settings. To do so, we transform p-values to s-values. These s-values give us the means to define a region of practical equivalence. We show via simulations when D−SOS matches or exceeds the performance of these tests. Like tests of goodness-of-fit, D−SOS benefits from explicitly taking the training set as the reference distribution while being robust to outliers.
3. Based on 62 real-world classification tasks from the OpenML-CC18 benchmark, we show that different notions of outlyingness are complementary for detecting partition-induced dataset shift. In these supervised settings, D−SOS easily extends to notions of outlyingness such as residuals and prediction uncertainty. These help to determine whether the shift has an adverse impact on predictive performance.

The main takeaway is that given a generic method to assign an outlier score to a data point, D−SOS turns these scores into a two-sample test for no adverse shift. We have created an R package dsos for our method. In addition, all code and data used in this paper are available.

2 STATISTICAL FRAMEWORK

The theoretical framework builds on Zhang [2002]. Take an i.i.d. training set $X^{tr} = \{x_i^{tr}\}_{i=1}^{n^{tr}}$ and a test set $X^{te} = \{x_i^{te}\}_{i=1}^{n^{te}}$. Each dataset X^o with origin $o \in \{tr(ain), te(st)\}$ lies in d-dimensional domain $\mathcal{X} \subseteq \mathbf{R}^d$ with sample sizes n^o, cumulative distribution function (CDF) F_X^o and probability density function (PDF) f_X^o. Let $\phi : \mathcal{X} \to \mathcal{S} \subseteq \mathbf{R}$ be a score function and define the threshold score $s \in \mathcal{S}$. The proportion above the threshold s in dataset o is in effect the contamination rate $C^o(s) = \Pr(\phi(x_i^o) \geq s)$. The contamination rate is the complementary CDF of the scores, $F^o(s) = 1 - C^o(s)$. As before, F^o and f^o now denote the CDF and PDF of the scores.

Consider the null hypothesis $\mathcal{H}_0 : F_X^{te} = F_X^{tr}$ of equal distribution against the alternative $\mathcal{H}_1 : F_X^{te} \neq F_X^{tr}$. In tandem, consider the null of equal contamination $\mathcal{H}_0^s : C^{te}(s) = C^{tr}(s)$ against the alternative $\mathcal{H}_1^s : C^{te}(s) \neq C^{tr}(s)$ at a given threshold score s. To evaluate goodness-of-fit, Zhang [2002] shows that testing $\mathcal{H}_0$ is equivalent to testing $\mathcal{H}_0^s$ for $\forall s \in \mathcal{S}$. If $z(s)$ is the relevant test statistic for equal contamination, a global test statistic Z for goodness-of-fit can be constructed from $z(s)$, its local counterpart. One such Z is

$$Z = \int z(s) \cdot w(s) \cdot d(s) \tag{1}$$

where $w(s)$ are threshold-dependent weights. For concision, we sometimes suppress the dependence on the threshold score s and denote weights, contamination statistics and rates as w, z and C^o respectively. D−SOS differs from the Z statistic in Equation (1) in three ways: the score function ϕ, the weights w and the contamination statistic z. We address each in turn.

D−SOS scores instances from least to most *abnormal* according to a specified notion of outlyingness. To be concrete, for density estimation, the negative log density is a natural score for outlyingness [Kandanaarachchi and Hyndman, 2021]. This property of the score function ϕ can be expressed as

$$\phi(x_i) \leq \phi(x_j) \Rightarrow \Pr(f_X^{tr}(x_i) \geq f_X^{tr}(x_j)) \geq 1 - \epsilon \tag{2}$$

for $x_i, x_j \in \mathcal{X}$ and ϵ, a (sufficiently small) approximation error. Accordingly, instances in high-density regions of the training set X^{tr} (nominal points or inliers) score low; those in low-density regions (outliers) score high. Here, the score function ϕ can be thought of as a density-preserving projection. More generally, higher scores, e.g. wider prediction intervals and larger residuals, indicate worse outcomes; the higher the score, the more *unusual* the observation. The structure in ϕ is the catalyst for adjusting the weights w and the statistic z.

D−SOS updates the weights w to be congruent with the score function ϕ. When projecting to the outlier subspace, high scores imply unusual points, whereas both tails are viewed as extremes in the Zhang framework. Univariate tests such as the Anderson-Darling and the Cramér-von Mises tests fit the framework in Equation (1): they make different choices for ϕ, z and w. These classical tests place more weight at the tails to reflect the severity of tail exceedances relative to deviations in the center of the distribution. D−SOS corrects for outliers being projected to the upper (right) tail of scores via the score function. The D−SOS weights w are specified as

$$w(s) = \left(F^{tr}(s)\right)^2 \tag{3}$$

The weights in Equation (3) shift most of the mass from low to high-threshold regions. As a result, D−SOS is highly tolerant of negligible shifts associated with low scores, and conversely, it is attentive to the shifts associated with high scores.

D−SOS constructs a test statistic based on the score ranks, not on the levels. The weighted area under the receiver operating characteristic curve (WAUC) is a robust statistic that is invariant to changes in levels so long as the underlying ranking is unaffected. The WAUC, denoted T, can be written as

$$T = \int F^{tr}(s) \cdot f^{te}(s) \cdot w(s) \cdot d(s) \tag{4}$$

See Equation (1) in Hand [2009] for details. T in Equation (4) is formally the D−SOS test statistic for adverse shift. We do not assume that the training set only consists of inliers. The training set may be contaminated, but hopefully not too badly compromised i.e. it is clean enough to be the reference distribution.

The null of no adverse shift $\mathcal{H}_0^{ds}$ is that most instances in the test set are not worse than those in the training set. The alternative $\mathcal{H}_1^{ds}$ is that the test set contains more outliers than expected, if the training set is the reference distribution. D−SOS specifies its null as $\mathcal{H}_0^{ds} : T_0 \leq T \mid \mathcal{H}_0$ against the alternative $\mathcal{H}_1^{ds} : T_0 > T \mid \mathcal{H}_0$. T_0 is the observed WAUC and $T \mid \mathcal{H}_0$, the WAUC under the null of exchangeable samples. When the test set is cleaner than the training set, tests of goodness-of-fit and of equal distributions both reject the null of no difference, whereas D−SOS by design does not reject its null of no adverse shift. These shifts do not trigger the alarm as they otherwise would had D−SOS been a two-tailed test.

Moreover, firstly tests of equal distributions ignore that the training set precedes the test set, and not vice versa. D−SOS, like tests of goodness-of-fit, improve on the former because they use the training set as the reference distribution. It takes advantage of and incorporates this information into the problem. Secondly, we note that D−SOS does not explicitly specify the equivalence margin, the minimum meaningful difference needed to sound the alarm [Wellek, 2010]. Choosing this margin or minimum effect size is often a difficult task, requiring substantive domain knowledge [Betensky, 2019]. The data-driven (adaptive) weights partially relieves us of this burden. They implicitly encode the principle that higher scores are worse without recourse to a fixed margin.

3 MOTIVATIONAL EXAMPLE

For illustration, we apply D−SOS to the canonical iris dataset [Anderson, 1935]. The task is to classify the species of Iris flowers based on $d = 4$ covariates (features) and $n = 50$ observations for each species. We show how D−SOS helps diagnose false alarms. We highlight that (1) changes in distribution do not necessarily hurt predictive performance, and (2) points in the densest regions of the distribution can be the most difficult – unsafe – to predict.

We consider four tests of no adverse shift. Each test uses a different score. For two-sample classification, this score is the probability of belonging to the test set. For density-based out-of-distribution (OOD) detection, the score comes from isolation forest; this score is inversely related to the local density. For residual diagnostics, it is the out-of-sample (out-of-bag) prediction error from random forests. Finally, for confidence-based OOD detection, it is the standard error of the mean prediction from random forests, a proxy for prediction (resampling) uncertainty similar to [Schulam and Saria, 2019]. Only the first notion of outlyingness – two-sample classification – pertains to modern tests of equal distributions; the others capture other meaningful notions of adverse shifts. For all these scores, higher is *worse*: higher scores indicate that the observation is diverging from the desired outcome or that it does not conform to the training set.

For the subsequent tests, we split iris into 2/3 training and 1/3 test set. The train-test pairs correspond to two partitioning strategies: (1) random sampling and (2) in-distribution (most dense) examples in the test set. How do these sample splits fare with respect to the aforementioned tests? Let s and p denote $s-$value and $p-$value. The results are reported on the $s = -\log_2(p)$ scale because it is intuitive and lends itself to comparison. We return to the advantages of using $s-$value for comparison later. An $s-$value of k can be interpreted as seeing k independent coin flips with the same outcome – all heads or all tails – if the null is that of a *fair* coin [Greenland, 2019]. This conveys how incompatible the data is with the null.

In Figure 1, the case with (1) random sampling exemplifies the type of false alarms we want to avoid. Two-sample classification, standing in for tests of equal distributions, is incompatible with the null of no adverse shift (a $s-$value of around 8). But this shift does not carry over to the other tests. Residual diagnostics, density-based and confidence-based OOD detection are all fairly compatible with the view that the test set is not worse. Had we been entirely reliant on two-sample classification, we may not have realized that this shift is essentially benign. Tests of equal distributions alone give a narrow perspective on dataset shift.

Returning to Figure 1, density-based OOD detection does not flag the (2) in-distribution test set as expected. We might be tempted to conclude that the in-distribution observations are *safe*, and yet, the tests based on residual diagnostics and confidence-based OOD detection are fairly incompatible with this view. Some of the densest points are concentrated in a region where the classifier does not discriminate well:

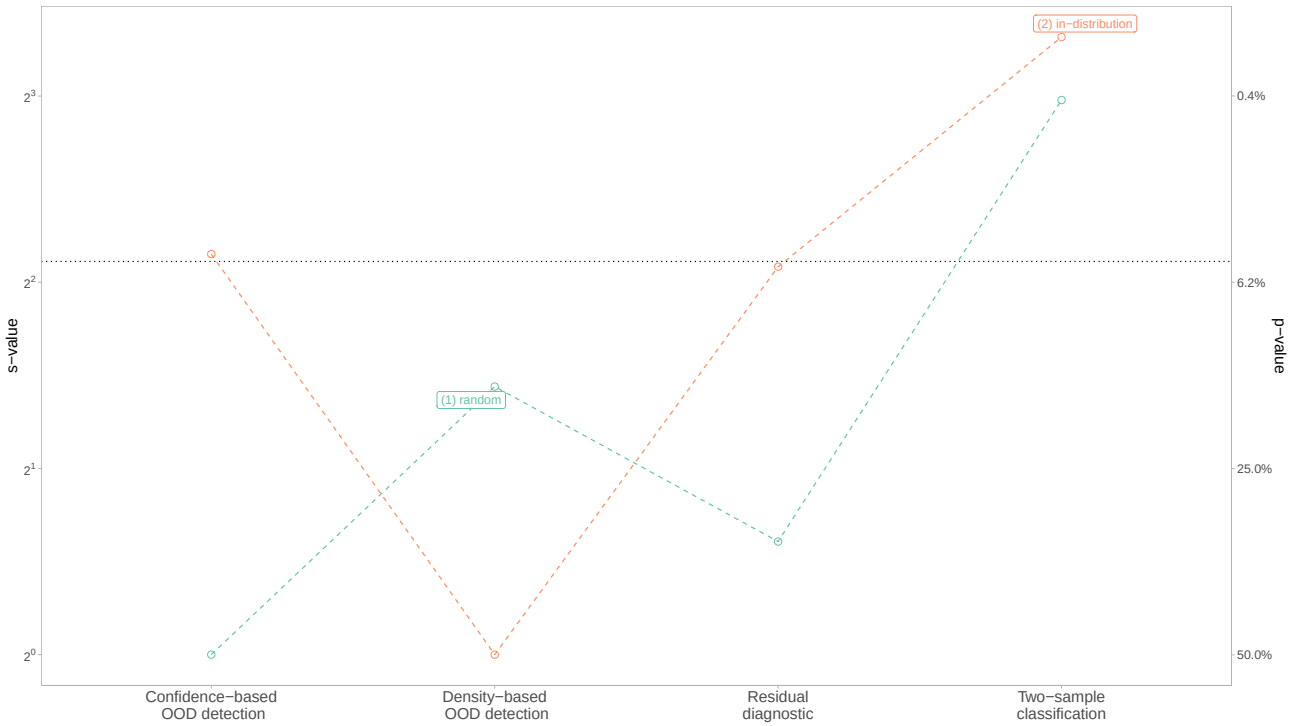

Figure 1: Tests of no adverse shift for `iris`. The tests cover 4 notions of outlyingness for two train-test splits. The x-axis indicates the test type and the colour, the sampling (splitting) strategy. The dotted black line is the common, if not commonly abused, $p-$value $= 0.05$ threshold. We clip $s-$values to a low and high of 1 and 10 respectively and display a secondary y-axis with the $p-$value as a cognitive bridge.

the species 'versicolor' and 'virginica' overlap. That is, the densest observations are not necessarily safe. Density-based OOD detection glosses over this: the trouble may well come from inliers that are difficult to predict. We get a more holistic perspective of dataset shift because of these complementary notions of outlyingness.

4 RELATED WORK

Outlier scores. Density ratios and class probabilities often serve as scores in comparing distributions [Menon and Ong, 2016] These scores, however, do not directly account for the predictive performance. Prediction intervals and residuals, on the other hand, do. Intuitively, they both reflect poor predictive performance in some regions of the feature space. Confidence-based OOD detection leans on this insight and tracks *uncertain* predictions e.g. Snoek et al. [2019], Berger et al. [2021]. This is in contrast to density-based OOD detection e.g. Morningstar et al. [2021]. The classical approach, which also accounts for the predictive performance, is based on residual diagnostics and underpin misspecification tests. Janková et al. [2020] is a recent example of this approach with a machine learning twist. Other methods such as trust scores can also flag unreliable predictions [Jiang et al., 2018]. Because all these scores represent distinct notions of outlyingness, the contrasts are often insightful. In this respect, `D-SOS` is in some sense a unifying framework. Bring your own outlier scores and `D-SOS` would morph into the equivalent two-sample test for no adverse shift.

Dimension reduction. In practice, the reconstruction errors from dimension reduction can separate inliers from outliers. Recently, Rabanser et al. [2019] uses dimension reduction, as a preprocessing step, to detect distributional shifts. Some methods gainfully rely on the supervised predictions for their low-rank representation [Cieslak and Chawla, 2009, Lipton et al., 2018]. This last approach, to add a cautionary note, entangles the supervised model, subject to its own sources of errors such as misspecification and overfitting, with dataset shift [Wen et al., 2014]. But it also points to the effectiveness of projecting to a lower and more informative subspace to circumvent issues related to high dimensions. Indeed, the classifier two-sample test uses *univariate* scores to detect changes in *multivariate* distributions [Cai et al., 2020]. Inspired by this approach, `D-SOS` uses outlier scores as a device for dimension reduction.

Statistical tests. The area under the receiver operating characteristic curve (AUC) has a long tradition of being used as a robust test statistic in two-sample comparison. Demidenko [2016] proposes an AUC-like test statistic as an alterna-

tive to the classical tests at scale because the latter "do not make sense with big data: everything becomes statistically significant" while the former attenuates the strong bias toward large sample size. As a generalization of the AUC, the weighted AUC (WAUC) puts non-uniform weights on decision thresholds [Li and Fine, 2010]. `D-SOS` seizes on this to give more weight to the outlying regions of the data. Weng and Poon [2008] also advocates for the WAUC as a performance metric for imbalanced datasets because it captures disparate misclassification costs. `D-SOS` is firmly rooted in the tradition of two-sample comparison. Inference is at the sample level. At the instance level, we may want to check if a given test point is an outlier to the training set. Bates et al. [2021] for example tests for individual outliers via conformal $p-$values. Podkopaev and Ramdas [2021], on the other hand, casts the problem of detecting harmful shifts in a sequential setting and controls for the false alarm rate due to continuous monitoring.

5 IMPLEMENTATION

In this section, we turn our attention to deriving valid $p-$values. Without loss of generality, let $\mathcal{D} = [X^{tr}, X^{te}]$ be the pooled training and test set. The score function $\phi = \Phi(\mathcal{D}, \lambda)$ is estimated from data $\mathcal{D}$ and hyperparameters $\lambda \in \Lambda$. This calibration procedure $\Phi : \mathcal{X} \times \Lambda \to \phi$ returns the requisite score function $\phi : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}$. The asymptotic null distribution of the WAUC, however, is invalid when the same data is used both for calibration and scoring. We circumvent this issue with permutations, sample splitting, and/or out-of-bag predictions.

Permutations use the empirical, rather than the asymptotic, null distribution. The maximum number of permutations is set to $R = 1000$. We refer to this variant as `DSOS-PT`. Unless stated otherwise, this is the default used in Section 6. For speed, `DSOS-PT` is implemented as a sequential Monte Carlo test, which terminates early when the resampling risk is unlikely to affect the final result materially [Gandy, 2009]. Even so, permutations can be computationally prohibitive.

A faster alternative, based on sample splitting, relies on the asymptotic null distribution but incurs a cost in calibration accuracy because it needs to set aside some data. This tradeoff is common in two-sample tests, which requires calibration. We randomly split each dataset in half: $X^{tr} = X_1^{tr} \cup X_2^{tr}$ and $X^{te} = X_1^{te} \cup X_2^{te}$. The first halves are used for calibration, and the second for scoring. We describe this split-sample procedure in Algorithm 1, and refer to it as `DSOS-SS` in Section 6. Given the weights in Equation 3, the WAUC under the null $T \mid \mathcal{H}_0$ is asymptotically normally distributed as

$$\begin{aligned} \sigma(n^{tr}, n^{te}) &= \sqrt{\frac{a_1 \cdot (\delta + 1)}{n^{te}}} \\ T \mid \mathcal{H}_0 &\sim \mathcal{N}(\frac{1}{12}, \sigma(n^{tr}, n^{te})) \end{aligned} \tag{5}$$

with mean $= \frac{1}{12}$ and standard deviation $\sigma(n^{tr}, n^{te})$, which depends on the sample sizes. The constants in Equation (5) work out to $\delta = \frac{n^{te}}{n^{tr}}$ and $a_1 \approx 0.008928571$. Equation 5 follows trivially from results in Li and Fine [2010] with a little bit of tedious algebra – see Equation (9) therein for the variance.

A third option, an obvious extension to sample splitting, is to use cross-validation instead. In $k-$fold cross-validation, the number of folds k mediates between calibration accuracy and inferential robustness. At the expense of refitting k times, this approach uses most of the data for calibration and can leverage the asymptotic null distribution, provided that the scores are out-of-sample predictions. This strategy is also called cross-fitting in semiparametric inference. It combines the best of both worlds, namely calibration accuracy and inferential speed. We refer to this variant as `DSOS-CV` in Section 6. We show in simulations that this approach either matches or exceeds the performance of `DSOS-PT` and `DSOS-SS`.

Algorithm 1: Split-sample test (DSOS-SS)

Data: Calibration set $\mathcal{D}_1 = [X_1^{tr}, X_1^{te}]$, Inference set $\mathcal{D}_2 = [X_2^{tr}, X_2^{te}]$, calibration procedure Φ and hyperparameters λ

1 **Label:** Assign labels $y = \mathbb{I}(x \in X_2^{te}), \forall x \in \mathcal{D}_2$;
2 **Calibrate:** Fit the score function $\phi = \Phi(\mathcal{D}_1, \lambda)$;
3 **Score:** Score observations $s = \phi(x), \forall x \in \mathcal{D}_2$;
4 **Test:** Compute the observed WAUC $T_0 = \texttt{WeightedAUC}(s, y)$ using Equation (4);
5 **Asymptotic null:** Under the null, T is asymptotically normally distributed as in Equation (5);

Output: $p - value = 1 - \Pr(T \leq T_0)$

6 EXPERIMENTS

We make the following pragmatic choices for ease of use; typically, the selected score functions perform well out-of-the-box with little to no costly hyperparameter tuning. For density-based OOD detection, we use isolation forest from the `isotree` package [Cortes, 2020]. We take the default hyperparamters from `isotree` as given, but increase the ensemble size to 500. To investigate other notions of outlyingness, we use random forests from the `ranger` package [Wright and Ziegler, 2017]. We take its default hyperparameters from Probst et al. [2019]. As in Hediger et al. [2022],

Table 1: Number of wins for test comparisons based on simulations. [1]Number of ties, not shown, when s-value difference is within ROPE.

Contender		Outcome[1]	
(1)	(2)	Win (1)	Win (2)
ctst	energy	12	43
ctst	DSOS-SS	0	34
ctst	DSOS-PT	0	60
ctst	DSOS-CV	0	64
DSOS-CV	energy	37	26
DSOS-SS	DSOS-CV	0	36
DSOS-SS	DSOS-PT	0	32
DSOS-CV	DSOS-PT	0	0

random forests allow us to use the out-of-sample variant of D-SOS (DSOS-CV) for free, so to speak, since out-of-bag predictions are viable surrogates for out-of-sample scores.

All experiments were run on a commodity desktop computer with a 12-core Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz processor and 64 GB RAM in R version 3.6.1 (2019-07-05). We stress that no hyperparameter tuning was performed - we set the hyperparameters to reasonable defaults as previously discussed. To avoid ambiguity, we explicitly say when the results pertain to DSOS-SS, DSOS-PT or DSOS-CV.

6.1 SIMULATED SHIFTS

We compare D-SOS to two modern tests of equal distribution. The first is the energy test, a type of kernel-based (distance-based) test [Székely et al., 2004]. The second, arguably the main inspiration for D-SOS, is a classifier two-sample test, ctst [Kim et al., 2021]. ctst tests whether a classifier can reliably identify training from test instances. If so, it is taken as evidence against the null. Both D-SOS and ctst use the same underlying classifier and hyperparameters and leverage classification probabilities as scores. As this score increases, so does the likelihood that an observation belongs to the test set, as opposed to the training set. Following Clémençon et al. [2009], this ctst variant uses sample splitting, as does DSOS-SS, but the AUC rather than the WAUC as a test statistic. That is, the weighting scheme is what differentiates the two. We show that overall, D-SOS performs favorably. We lose little by using a test for no adverse shift instead of an explicit test of equal distribution.

We simulate distribution shifts from a two-component multivariate Gaussian mixture model (GMM). The training and test set are drawn from:

$$X^{tr} \sim \phi^{tr} \cdot \mathcal{N}_d(\mu_{\mathbf{1}}^{\mathbf{tr}}, \mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{tr}}) + (1-\phi^{tr}) \cdot \mathcal{N}_d(\mu_{\mathbf{2}}^{\mathbf{tr}}, \mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{tr}})$$
$$X^{te} \sim \phi^{te} \cdot \mathcal{N}_d(\mu_{\mathbf{1}}^{\mathbf{te}}, \mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{te}}) + (1-\phi^{te}) \cdot \mathcal{N}_d(\mu_{\mathbf{2}}^{\mathbf{te}}, \mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{te}})$$

Omitting subscripts and superscripts for brevity, $\phi \in [0,1]$, μ and $\mathbf{\Sigma}$ are the component weight, mean vector and covariance matrix respectively. The baseline specifies the training and test sample size $n^{tr} = n^{te} \in \{400, 800, 1600\}$, the number of dimensions $d \in \{4, 8, 16\}$, the component weights $\phi^{tr} = \phi^{te} = 0.5$, the mean vectors $\mu_{\mathbf{1}}^{\mathbf{tr}} = \mu_{\mathbf{1}}^{\mathbf{te}} = \mathbf{1}_d$ and $\mu_{\mathbf{2}}^{\mathbf{tr}} = \mu_{\mathbf{2}}^{\mathbf{te}} = -\mathbf{1}_d$ and the covariance matrices $\mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{tr}} = \mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{te}} = \mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{tr}} = \mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{te}} = \mathbf{I}_d$. $\mathbf{I}_d$ is the $d \times d$ identity matrix and $\mathbf{1}_d$ is the d-dimensional all-ones vector. The baseline configurations enforce that training and test set are drawn from the same distribution, i.e. no shift. There is a total of 9 such configurations (3 dimensions times 3 sample sizes).

We generally shift the distribution so that the dimension of change $d^* = 1$ is fixed as the ambient dimension d increases. The power of multivariate tests based on kernels and distances decays with increasing dimension when differences only exist along a few intrinsic dimensions $d^* \ll d$. We vary one or more parameters, namely ϕ^{tr}, ϕ^{te}, $\mu_{\mathbf{2}}^{\mathbf{te}}$ and $\mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{te}}$, to simulate the desired shifts, all else constant. We change the following parameters to pre-set intensity levels:

1. Label (prior) shift – We flip the weights so that $\phi^{tr} \in \{0.49, 0.47, 0.45\}$ goes with $\phi^{te} = 1 - \phi^{tr}$. The majority component in training becomes the minority in the test sample.
2. Corrupted sample – We draw a fraction $\omega \in \{0.01, 0.02, 0.04\}$ of examples in the test set from the component that is absent in training such that $\phi^{tr} = 1$ and $\phi^{te} = 1 - \omega$.
3. Mean shift – We change the mean vector in the test set so that $\mu_{\mathbf{2}}^{\mathbf{te}} = [-\frac{\kappa}{10}, -\mathbf{1}_{d-1}]$, where $\kappa \in \{11, 12, 14\}$.
4. Noise shift – We change the covariance matrix in the test set so that $diag(\mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{te}}) = [\frac{\theta}{10}, \mathbf{1}_{d-1}]$, where $\theta \in \{11, 12, 14\}$ and $diag(\cdot)$ is the assignment operator for the diagonal elements of the d-by-d covariance matrix.
5. Dependency shift – We induce a positive relationship between the first two covariates. We change the covariance structure in the test set so that $\mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{te}}[1,2] = \mathbf{\Sigma}_{\mathbf{2}}^{\mathbf{te}}[2,1] = \gamma$, where $\gamma \in \{0.1, 0.2, 0.4\}$.

There is a total of 27 configurations for each shift type. For each type and shift intensity, we repeat the experiment 500 times.

To compare the D-SOS, ctst and energy tests, we employ the Bayesian signed rank test [Benavoli et al., 2014]. To do so, we specify a region of practical equivalence (ROPE) on the s-value scale and a suitable prior. We deem two statistical tests practically equivalent if the absolute difference in

s-value is ≤ 1. This difference corresponds to one more coin flip turning up as heads (or tails), keeping the streak alive. Anything greater arouses suspicion that one test is more powerful, and that the difference is not negligible. Specifying a ROPE on the s-value scale is less cumbersome than the p-value scale. p-values, let alone p-value differences, are notoriously difficult to interpret [Wasserstein et al., 2016]. The Bayesian signed rank test yields the posterior probability that one method is better, practically equivalent or worse than the other. The prior for these experiments adds one *pseudo* experiment, distinct from the 500 *real* simulations, where the difference is 0, i.e. no practical difference exists out of the gate.

The full tables for comparison are provided as supplementary material. Table 1 summarizes these findings across all settings: dimension, sample size, shift type and shift intensity. For simplicity, we say that one method wins if the posterior probability that it is better is ≥ 0.5; similarly, they draw if the posterior probability that they are practically equivalent is ≥ 0.5. This way of comparing methods resembles the Pitman closeness criterion. We make several observations:

1. `DSOS-PT` and `DSOS-CV` are practically equivalent across all settings. Notwithstanding the gains in inferential speed that one may hold over the other, the inferential accuracy (power) is – should be – the same. This is a good sanity check that the implementation is correct. Similarly, for the 9 (out of 144) configurations with no shift, all tests are practically equivalent.
2. `DSOS-PT` and `DSOS-CV` dominate `DSOS-SS`. `DSOS-SS` pays a hefty price in inferential accuracy due to sample splitting. Sample splitting, as evidenced in `DSOS-SS` and `ctst`, clearly trails behind the competition. Even so, `DSOS-SS` outperforms this `ctst` variant: the weights in `DSOS-SS` clearly pay dividends.
3. `DSOS-CV` (or `DSOS-PT`) lag behind `energy` in settings with (1) label shift and (2) corrupted samples. There are 26 such cases. As previously noted, that `DSOS` is robust to outliers makes it more forgiving than nonrobust tests when the bulk of the distributions is largely unaffected. This is not a bug; it is a feature. In all other configurations, `DSOS-CV` either wins or draws. In particular, there are 37 cases where `DSOS-CV` beats `energy`. This is because it conditions on the training set as the reference distribution to gain an edge.

6.2 PARTITION-INDUCED SHIFTS

Cross-validation mimics random sampling variation, and is helpful in assessing stability [Moreno-Torres et al., 2012, Lim and Yu, 2016]. It generates partition-induced shifts, rather than worst-case (adversarial) perturbations [Subbaswamy et al., 2021]. To investigate how different notions of outlyingness are correlated, we analyze via cross-validation 62 classification tasks from the OpenML-CC18 benchmark [Casalicchio et al., 2017][1]. As in Section 3, we look at tests of no adverse shift based on two-sample classification, residual diagnostics, and predictive uncertainty (no density-based OOD detection in this round). Because the core task *is* prediction, the `D-SOS` tests based on residual diagnostics and/or predictive uncertainty are more informative than, and can be at odds with, two-sample classification (density-based tests).

For each OpenML-CC18 dataset, we perform stratified 10-fold cross-validation, repeated twice. We end up with 20 train-test splits per task. In total, we run 3720 tests of no adverse shifts (62 datasets, 20 random splits, and 3 tests). Summary statistics for these datasets and granular test results are in the supplementary material. We expect the statistical tests to be correlated within but not across datasets. To formalize this setup, we use the following model:

$$\log s_i^j \sim \mathcal{N}(\mu_i^j, \Sigma_i^j) \tag{6}$$

where for each dataset $i = 1, 2, \cdots, 62$ and test $j = 1, 2, 3$, the $s-$value s_i^j is lognormally distributed: positive and skewed to the right. The $s-$value s_i^j consists of a dataset-specific (fixed) effect μ_i^j, subject to noise in Σ_i^j; Σ_i^j accounts for within-dataset covariance.

We fit the model in 6 using the `clubSandwich` package to obtain robust estimates of the fixed effects μ_i^j even with arbitrary covariance structure in Σ_i^j left unspecified [Pustejovsky and Tipton, 2018]. These fixed effects μ_i^j in 6 measure how sensitive a dataset is to these partition-induced shifts. The higher the value, the more susceptible a dataset is to adverse shifts caused by sampling variation. The exponentiated fixed effects are on the s-value scale and can be interpreted in the usual way as the strength of evidence against the null of no adverse shift.

At a glance, Figure 2, created with the `superheat` package [Barter and Yu, 2018], tells the story quite succintly. Across these 62 datasets, we see that tests of no adverse shift based on two-sample classification, residual diagnostics, and predictive uncertainty (confidence-based OOD detection) are indeed highly correlated. Residual diagnostics and predictive uncertainty, both of which mirror the performance of the underlying supervised algorithm, are the most correlated (Pearson correlation coefficient is 0.82). This suggests that they can be used interchangeably, and by virtue of focusing on predictive performance, are suitable metrics for model monitoring.

[1] We exclude datasets with more than 500 features because of long runtimes.

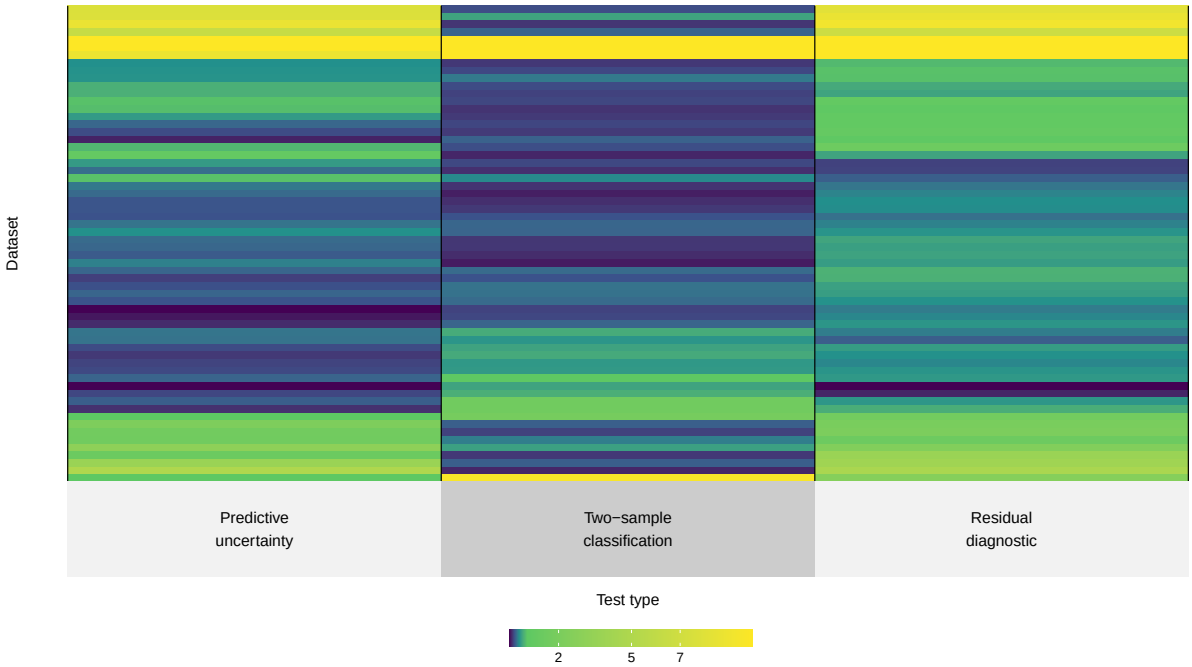

Figure 2: Heatmap of fixed effects for 62 datasets from the OpenML-CC18 benchmark, ordered via hierarchical clustering. Fixed effects are exponentiated and reported on the s-value scale for 3 types of `D−SOS` test. The brighter the color, the higher the s-value, the more prone the dataset to adverse shift caused by sampling variation.

Two-sample classification is not as strongly associated with residual diagnostics and predictive uncertainty (Pearson correlation coefficient is 0.5 and 0.47 respectively). This arcs back to the point made in Section 3; namely, tests of equal distribution alone are ill-equiped, by definition, to detect whether a shift is benign or harmful for predictive tasks. Notice the datasets clustered together at the very top of Figure 2 with low $s-$values for two-sample classification but high $s-$values for residual diagnostics and predictive uncertainty. For these datasets, the distribution seems fine but prediction suffers. To a lesser extent, also notice the reverse effect at the bottom of Figure 2. These are the false alarms we previously highlighted where we have relatively higher $s-$values for two-sample classification than for the other two notions of outlyingness. That is, for these datasets, prediction seems fine but the distribution is not.

7 CONCLUSION

`D−SOS` is a wrapper to turn outlier scores into a statistical test. This framework, derived from tests of goodness-of-fit, detects non-negligible adverse shifts. It can confront the data with more pertinent hypotheses than tests of equal distribution. This works well when mapping to the relevant outlier subspace discriminates between safe and unsafe regions of the data distribution. Our method accommodates different notions of outlyingness. It lets users define, via a score function, what is meant by *worse* off. Besides the outlier scores explored in this paper, we stress that other sensible choices for the score functions ϕ and the weights w abound. These can be adjusted to account for domain knowledge. Looking ahead, future research could investigate other weighting schemes. The functional form of the postulated weights could be worth tuning. Moreover, *composite* score function, which would combine several notions of outlyingness together, would enrich the types of hypotheses we can test. We imagine situations where we want to combine the strengths of confidence-based and density-based OOD detection into a single aggregate score for example.

References

Edgar Anderson. The irises of the gaspe peninsula. *Bull. Am. Iris Soc.*, 59:2–5, 1935.

Rebecca L Barter and Bin Yu. Superheat: An r package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of Computational and Graphical Statistics*, 27(4):910–922, 2018.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021.

Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pages 1026–1034. PMLR, 2014.

Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: A comparative study and analysis. *arXiv preprint arXiv:2107.02568*, 2021.

Rebecca A Betensky. The p-value requires context, not a threshold. *The American Statistician*, 73(sup1):115–117, 2019.

Haiyan Cai, Bryan Goggin, and Qingtang Jiang. Two-sample test based on classification probability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(1):5–13, 2020.

Giuseppe Casalicchio, Jakob Bossek, Michel Lang, Dominik Kirchhoff, Pascal Kerschke, Benjamin Hofner, Heidi Seibold, Joaquin Vanschoren, and Bernd Bischl. Openml: An r package to connect to the machine learning platform openml. *Computational Statistics*, 32(3):1–15, 2017. doi: 10.1007/s00180-017-0742-2. URL `http://doi.acm.org/10.1007/s00180-017-0742-2`.

David A Cieslak and Nitesh V Chawla. A framework for monitoring classifiers' performance: when and why failure occurs? *Knowledge and Information Systems*, 18(1): 83–108, 2009.

Stéphan Clémençon, Marine Depecker, and Nicolas Vayatis. Auc optimization and the two-sample problem. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 360–368, 2009.

David Cortes. *isotree: Isolation-Based Outlier Detection*, 2020. URL `https://CRAN.R-project.org/package=isotree`. R package version 0.1.18.

Eugene Demidenko. The p-value you can't buy. *The American Statistician*, 70(1):33–38, 2016.

Axel Gandy. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009.

Sander Greenland. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73 (sup1):106–114, 2019.

David J Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, 2009.

Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, page 107435, 2022.

Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):773–795, 2020.

Heinrich Jiang, Been Kim, Melody Y Guan, and Maya R Gupta. To trust or not to trust a classifier. In *NeurIPS*, pages 5546–5557, 2018.

Sevvandi Kandanaarachchi and Rob J Hyndman. Leave-one-out kernel density estimates for outlier detection. *Journal of Computational and Graphical Statistics*, (just-accepted):1–28, 2021.

Mark G Kelly, David J Hand, and Niall M Adams. The impact of changing populations on classifier performance. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–371, 1999.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

Jialiang Li and Jason P Fine. Weighted area under the receiver operating characteristic curve and its application to gene selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(4):673–692, 2010.

Chinghway Lim and Bin Yu. Estimation stability with cross-validation (escv). *Journal of Computational and Graphical Statistics*, 25(2):464–492, 2016.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.

Jose García Moreno-Torres, José A Sáez, and Francisco Herrera. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021.

Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*, 2020.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. *arXiv preprint arXiv:2110.06177*, 2021.

Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. Data validation for machine learning. *Proceedings of Machine Learning and Systems*, 1:334–347, 2019.

Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019.

James E Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: an empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, pages 1394–1406, 2019.

Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1022–1031. PMLR, 2019.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021.

Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

Ronald L Wasserstein, Nicole A Lazar, et al. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, 2010.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *ICML*, pages 631–639, 2014.

Cheng G Weng and Josiah Poon. A new evaluation measure for imbalanced datasets. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 27–32, 2008.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1): 1–17, 2017. doi: 10.18637/jss.v077.i01.

Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.

Jin Zhang. Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):281–294, 2002.