FAILING LOUDLY: AN EMPIRICAL STUDY OF METHODS FOR DETECTING DATASET SHIFT

Stephan Rabanser*, **Stephan Günnemann**
Technical University of Munich, Germany
{rabanser,guennemann}@in.tum.de

Zachary C. Lipton
Carnegie Mellon University, Pittsburgh, PA
zlipton@cmu.edu

ABSTRACT

We might hope that when faced with unexpected inputs, well-designed software systems would fire off warnings. Machine learning (ML) systems, however, which depend strongly on properties of their inputs (e.g. the i.i.d. assumption), tend to fail silently. This paper explores the problem of building ML systems that fail loudly, investigating methods for detecting dataset shift and identifying exemplars that most typify the shift. We focus on several datasets and various perturbations to both covariates and label distributions with varying magnitudes and fractions of data affected. Interestingly, we show that while classifier-based methods perform well in high-data settings, they perform poorly in low-data settings. Moreover, across the dataset shifts that we explore, a two-sample-testing-based approach, using pre-trained classifiers for dimensionality reduction performs best.

1 INTRODUCTION

Even subtle changes in the data distribution can destroy the performance of otherwise state-of-the-art classifiers, a phenomenon exemplified by adversarial examples (Szegedy et al., 2013; Zügner et al., 2018). When decisions are made under uncertainty, even shifts in the label distribution can significantly compromise accuracy (Zhang et al., 2013; Lipton et al., 2018). Unfortunately, in practice, ML pipelines rarely inspect incoming data for signs of distribution shift, and for detecting shift in high-dimensional real-world data, best practices have not been established yet[1]. The first indications that something has gone awry might come when customers complain.

This paper investigates methods for efficiently detecting distribution shift, a problem naturally cast as two-sample testing. We wish to test the equivalence of the *source* distribution (from which training data is sampled) and *target* distribution (from which real-world data is sampled). For simple univariate distributions, such hypothesis testing is a mature science. One might be tempted to use off-the-shelf methods for multivariate two-sample tests to handle high-dimensional data but these kernel-based approaches do not scale with the dataset size and their statistical power decays badly when the ambient dimension is high (Ramdas et al., 2015).

For ML practitioners, another intuitive approach might be to train a classifier to distinguish between examples from source and target distributions. We can then look to see if the classifier achieves significantly greater than 50% accuracy. Analyzing the simple case where one wishes to test the means of two Gaussians, Ramdas et al. (2016) recently made the intriguing discovery that the power of a classification-based strategy using Fisher's Linear Discriminant Analysis classifier achieves minimax rate-optimal performance. However, to date, we know of no rigorous empirical investigation characterizing classifier-based approaches to recognize dataset shift in the real high-dimensional data distributions with no known parametric form on which modern machine learning is routinely deployed. Providing this analysis is a key contribution of this paper. To avoid confusion, we will denote any source-vs-target classifier a *domain classifier* and refer to the original classifier, (trained on source data) to predict the classes as the *label classifier*.

*Work done while a visiting research scholar at Carnegie Mellon University.

[1]TensorFlow's data validation tools only compare summary stats of training vs incoming data—https://www.tensorflow.org/tfx/data_validation/get_started#checking_data_skew_and_drift

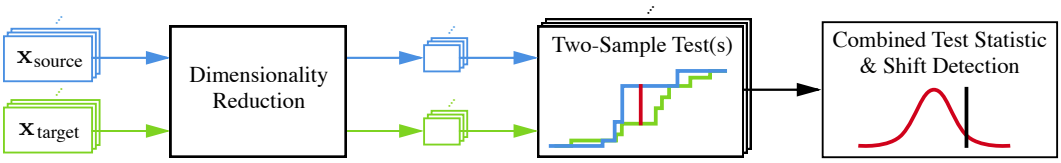


Figure 1: Our pipeline for detecting dataset shift. We consider various choices for how to represent the data and how to perform two-sample tests.

A key benefit of the classifier-based approach is that the *domain classifier* reduces dimensionality to a single dimension, learned precisely for the purpose of discriminating between source and target data. However, a major drawback is that deep neural networks, the precise classifiers that are effective on the high-dimensional data that interests us, require large amounts of training data. Adding to the problem, the domain-classifier approach requires partitioning our (scarce) target data using, e.g., half for training and leaving only the remainder for two-sample testing. Thus, as an alternative we also explore the *black box shift detection (BBSD)* approach due to Lipton et al. (2018), which addresses shift detection under the label shift assumption. They show that if one possesses an off-the-shelf label classifier $f(x)$ with an invertible confusion matrix (verifiable on training data), then detecting that the source distribution p is different from the target distribution q requires only detecting that $p(f(x)) \neq q(f(x))$. This insight enables efficient shift detection, using a pre-trained (label) classifier for dimensionality reduction.

Building on these ideas of combining (black-box) dimensionality reduction with subsequent two-sample testing, we explore a range of dimensionality reduction techniques and compare them under a wide variety of shifts (Figure 1 illustrates our general framework). Interestingly, we show (empirically) that BBSD works surprisingly well under a broad set of shifts, outperforming other methods, even when its assumptions are not met.

Related work Given just one example from the test data our problem simplifies to anomaly detection, surveyed thoroughly by Chandola et al. (2009); Markou & Singh (2003). Popular approaches to anomaly detection include density estimation (Breunig et al., 2000), margin-based approaches such as one-class SVMs (Schölkopf et al., 2000), and the tree-based isolation forest method due to (Liu et al., 2008). Recently, GANs have been explored for this task (Schlegl et al., 2017). Given simple streams of data arriving in a time-dependent fashion where the signal is piece-wise stationary, with stationary periods separated by abrupt changes, the problem of detecting a dataset shift becomes the classic time series problem of change point detection, with existing methods summarized succinctly in an excellent survey by Truong et al. (2018). An extensive literature addresses dataset shift in machine learning, typically in the larger context of domain adaptation, often through importance-weighted risk minimization. Owing to the impossibility of correcting for shift absent assumptions (Ben-David et al., 2010), these papers often assume either covariate shift $q(x, y) = q(x)p(y|x)$ (Shimodaira, 2000; Sugiyama et al., 2008; Gretton et al., 2009) or label shift $q(x, y) = q(y)p(x|y)$ (Saerens et al., 2002; Chan & Ng, 2005; Storkey, 2009; Zhang et al., 2013; Lipton et al., 2018). Schölkopf et al. (2012) provides a unifying view of these shifts, showing how assumed invariances in conditional probabilities correspond to causal assumptions about the inputs and outputs.

2 Shift Detection Techniques

Given labeled data $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N) \sim p$ and unlabeled data $\boldsymbol{x}'_1, ..., \boldsymbol{x}'_M \sim q$, our task is to determine whether $p(\boldsymbol{x})$ equals $q(\boldsymbol{x}')$. Formally, $H_0 : p(\boldsymbol{x}) = q(\boldsymbol{x}')$ vs $H_1 : p(\boldsymbol{x}) \neq q(\boldsymbol{x}')$. Chiefly, we explore the following design considerations: (i) what **representation** to run the test on; (ii) which **two-sample test** to run; (iii) when the representation is multidimensional; whether to run **multivariate or multiple univariate two-sample tests**; and (iv) **how to combine** their results. Additionally, we share some preliminary work on qualitatively characterizing the shift, e.g. by presenting exemplars, or identifying salient features.

2.1 Dimensionality Reduction

Building on recent results in Lipton et al. (2018); Ramdas et al. (2015) suggesting the benefits of low-dimensional two-sample testing, we consider the following representations: (i) **No Reduction**

(***NoRed***): To justify any dimensionality reduction techniques, we include tests on the original raw features; (ii) **SRP**: sparse random projections (Achlioptas, 2003; Li et al., 2006); (iii) **PCA**: principal components analysis; (vi) **TAE**: We extract representations by running an autoencoder trained on source data; (v) **UAE**: the same approach but with an untrained autoencoder; (vi) **BBSD**: Here, we adopt the approach of Lipton et al. (2018), using the outputs of a *label classifier* trained on source data. Two variations of this approach are to use the hard-thresholded predictions (BBSDh) of the label classifier enabling a chi-squared test of independence, or to use the softmax outputs (BBSDs), requiring a subsequent multivariate test; and (vii) **Classifier (*Classif*)**: We partition both the source data and target data into two halves, training a *domain classifier* to distinguish source from target (trained with balanced classes). We then apply this model to the remaining data, performing a subsequent binomial test on the hard-thresholded predictions.

2.2 TWO-SAMPLE TESTING

The dimensionality reduction techniques each yield a representation, either uni- or multidimensional, either continuous or discrete. Among categorical output, we may have binary outputs (as from the domain classifier) or multiple categories (the results from the hard label classifier). The next step is to choose a suitable two sample test. In all experiments, we adopt a high-significance level of $\alpha = 0.05$ for hypothesis rejection. For representation methods that yield multidimensional outputs, we have two choices: to perform a multivariate two-sample test, such as kernel two-sample tests due to Gretton et al. (2012) or to perform uni-variate tests separately on each component. In the latter case, we must subsequently combine the p-values from each test, encountering the problem of multiple hypothesis testing. Unable to make strong assumptions about the dependence among the tests, we must rely on a conservative aggregation method, such as the Bonferroni correction (Bland & Altman, 1995). While a number of less conservative aggregations methods have been proposed (Simes, 1986; Zaykin et al., 2002; Loughin, 2004; Heard & Rubin-Delanchy, 2018; Vovk & Wang, 2018), we found that even using the Bonferroni method, dimensionality reduction plus aggregation, generally outperformed kernel two-sample testing on those same representations.

When performing univariate tests on continuous variables, we adopt the *Kolmogorov-Smirnov (KS) test*, a non-parametric test whose statistic is calculated by taking the supremum over all values z of the differences of the cumulative density functions (CDFs) as follows: $D = \sup_z |F_s(z) - F_t(z)|$ where F_s and F_t are the empirical CDFs of the source and target data, respectively. When aggregating K univariate tests together, we apply the Bonferroni correction rejecting the null hypothesis if the minimum p-value among all tests is less than α/K.

For all methods yielding multidimensional representations (NoRed, PCA, SRP, UAE, TAE, and BBSDs), we tried both the kernel two-sample tests and Bonferroni-corrected univariate KS tests, finding to our surprise, that the aggregated KS tests provided superior shift detection in most cases (see experiments). For BBSDh (using the hard-thresholded label classifier predictions) we employ a chi-squared test on the class frequencies. For the domain-classifier, we simply compare its accuracy on held-out data to random chance via a binomial test.

3 EXPERIMENTS

We briefly summarize our experimental setup. Our experiments address the MNIST and CIFAR-10 datasets. For autoencoder (UAE & TAE) experiments, we employ a convolutional architecture with 3 conv and 1 fully-connected layers. For the *label classifier* and *domain classifier* we use a ResNet-18 (He et al., 2016). We train all networks (TAE, BBSDs, BBSDh, Classif) with stochastic gradient descent (SGD) with momentum and a batch size of 128, (decaying the learning rate with $1/\sqrt{t}$) over 200 epochs with early stopping. For PCA, SRP, UAE, and TAE, we reduce dimensionality to $K = 32$ latent dimensions, which explains roughly 80% of PCA variance in both datasets used. The label classifier BBSDs reduces dimensionality to the number of classes C. In these experiments, we simulate a number of varieties of shift, affecting both the covariates and the label proportions. For all shifts, we evaluate the various methods' abilities to detect shift (including no-shift cases to check against false positives). We evaluate the models with various amounts of samples from the target dataset $s \in \{10, 20, 50, 100, 200, 500, 1000, 10000\}$. Because of the unfavorable dependence of kernel methods on the dataset size we run these methods only up until 1000 target samples.

Table 1: Dimensionality reduction methods (a) and shift-type (b) comparison. Underlined entries indicate accuracy values larger than 0.5.

(a) Detection accuracy of different dimensionality reduction techniques across all simulated shifts on MNIST and CIFAR-10. **Green bold** entries indicate the best DR method at a given sample size, *red italic* the worst. BBSDs performs best for univariate testing, while both UAE and TAE perform best for multivariate testing.

Test	DR	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univ. tests	NoRed	0.18	0.25	0.36	0.39	0.45	0.49	0.57	0.70
	PCA	0.13	0.22	0.25	0.30	0.35	0.41	*0.46*	0.58
	SRP	0.18	0.21	0.27	0.32	0.37	0.46	0.53	0.61
	UAE	0.20	0.25	0.33	0.42	0.48	0.54	0.65	**0.74**
	TAE	0.20	0.26	0.37	0.45	0.44	0.53	0.58	0.67
	BBSDs	**0.29**	**0.38**	**0.43**	**0.49**	**0.55**	**0.61**	**0.66**	0.72
χ^2	BBSDh	0.14	0.18	0.23	0.26	0.32	0.41	0.47	*0.48*
Bin	*Classif*	*0.04*	*0.09*	*0.10*	*0.10*	*0.28*	*0.38*	0.47	0.66
Multiv. tests	*NoRed*	*0.00*	*0.00*	*0.00*	*0.04*	0.15	0.15	*0.18*	–
	PCA	0.01	0.05	0.09	0.11	0.15	0.22	0.28	–
	SRP	*0.00*	*0.00*	0.03	0.08	*0.13*	*0.14*	0.19	–
	UAE	**0.19**	**0.26**	**0.36**	0.36	0.42	0.49	**0.59**	–
	TAE	**0.19**	0.22	**0.36**	**0.44**	**0.46**	**0.50**	0.58	–
	BBSDs	0.16	0.19	0.23	0.31	0.30	0.43	0.48	–

(b) Detection accuracy of different shifts on MNIST and CIFAR-10 using the best-performing DR technique (univariate: BBSDs, multivariate: average of UAE and TAE).

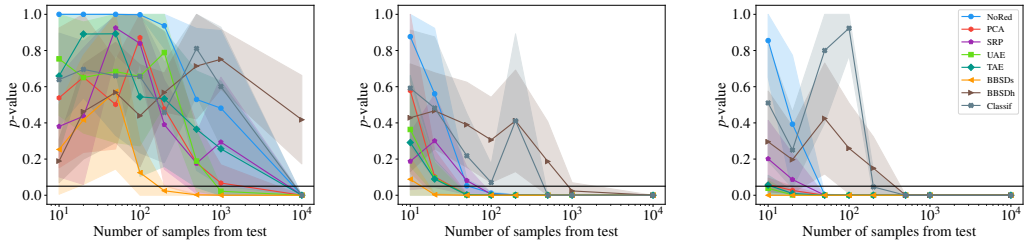
Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univariate tests	s_gn	0.00	0.03	0.07	0.10	0.10	0.10	0.10	0.10
	m_gn	0.00	0.03	0.10	0.10	0.14	0.24	0.24	0.38
	l_gn	0.45	0.52	0.59	0.72	0.83	0.86	0.97	1.00
	s_img	0.14	0.21	0.31	0.45	0.59	0.59	0.69	0.97
	m_img	0.34	0.55	0.66	0.79	0.83	0.90	0.93	1.00
	l_img	0.48	0.66	0.72	0.83	0.83	0.93	1.00	1.00
	adv	0.07	0.10	0.10	0.10	0.17	0.21	0.34	0.45
	ko	0.00	0.00	0.00	0.07	0.10	0.34	0.48	0.72
	m_img+ko	0.45	0.66	0.79	1.00	1.00	1.00	1.00	1.00
	oz+m_img	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Multivariate kernel tests	s_gn	0.02	0.02	0.05	0.05	0.08	0.10	0.12	–
	m_gn	0.03	0.08	0.17	0.19	0.29	0.32	0.39	–
	l_gn	0.51	0.53	0.71	0.75	0.81	0.88	0.97	–
	s_img	0.12	0.15	0.20	0.29	0.39	0.41	0.51	–
	m_img	0.29	0.31	0.36	0.39	0.44	0.54	0.66	–
	l_img	0.29	0.32	0.36	0.46	0.59	0.75	0.81	–
	adv	0.03	0.08	0.20	0.22	0.22	0.25	0.32	–
	ko	0.05	0.10	0.17	0.19	0.20	0.24	0.34	–
	m_img+ko	0.17	0.22	0.39	0.51	0.54	0.58	0.78	–
	oz+m_img	0.36	0.75	0.86	1.00	1.00	1.00	1.00	–

For each shift type (as appropriate) we explored three levels of shift intensity (indicated by s_, m_, l_ prefixes) and various percentages of affected data $\delta \in \{0.1, 0.5, 1.0\}$. We explore the following types of shift: **Adversarial examples (*adv*)**: as introduced by Szegedy et al. (2013) and created via the FGSM method (Goodfellow et al., 2014); **Knock-out shift (*ko*)**: a form of label shift introduced by Lipton et al. (2018), where a fraction δ of data points from class $c = 0$ are removed, creating class imbalance; **Gaussian shift (*gn*)**: covariates corrupted by noise with standard deviation $\sigma \in \{1, 10, 100\}$; **Image shift (*img*)**: more natural shifts to images using random amounts of rotations $\{10, 40, 90\}$, (x, y)-axis-translation percentages $\{0.05, 0.2, 0.4\}$, as well as zoom-in percentages $\{0.1, 0.2, 0.4\}$. We also explored combinations of image shift with label shift (e.g. m_img+ko). **Original splits**: As a sanity check, we also evaluated our detectors on the original source/target splits provided by in MNIST, CIFAR-10, Fashion MNIST, and SVHN datasets typically regarded being i.i.d.; and **Domain adaptation datasets**: We tested our detection method on the domain adaptation task from MNIST (source) to USPS (target) (Long et al., 2013) as well as on the COIL-10 dataset (an alternation of the COIL-100 dataset which only includes the first 10 classes) (Nene et al., 1996) where images between $0°$ and $175°$ are sampled by the source and images between $180°$ and $355°$ are sampled by the target distribution.

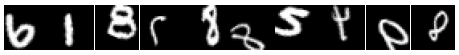
4 Discussion

Aggregating results across the broad spectrum of explored shifts (Table 1a), we see that univariate tests on BBSDs representations performs best. The domain-classifier approach struggles the most to detect shift in the low-sample regime, but performs better with more sample data. One benefit of the classifier-based approach is that it gives us an intuitive way to quantify the amount of shift (the accuracy of the classifier), and also yields a mechanism for identifying exemplars most likely to occur in either the source or target distributions. In Appendix A, we break out all results by shift type and provide top different/similar samples. Detection results by shift type are summarized in Table 1b. We were surprised to find that across our dimensional-reduction methods, aggregated univariate tests performed separately on each component of the latent vectors outperformed multivariate tests. Overall, we observed that large shifts can on average already be detected with better than chance accuracy at only 20 samples in the multiple univariate testing setting, while medium and small shifts required orders of magnitude more samples in order to achieve similar accuracy.

Aside from our synthetically generated shifts applied to MNIST and CIFAR-10 (example shown in Figure 2), we also ran our detection scheme on the COIL-10 dataset. The key results are reported



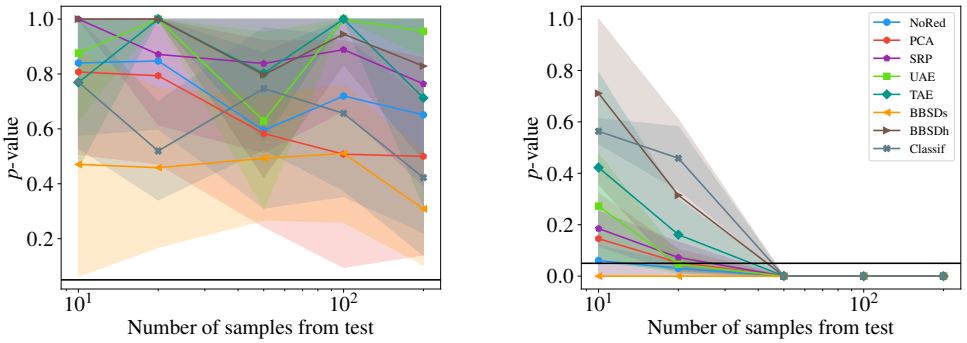
(a) 10% perturbed data in test set. (b) 50% perturbed data in test set. (c) 100% perturbed data in test set.



(d) Top different samples.

(e) Top similar samples.

Figure 2: Shift detection results for medium image shift on MNIST.



(a) Randomly shuffled dataset containing images of all angles in both training and test set. (b) Angle-partitioned dataset as described in Sec. 3.

(c) Top different samples.

(d) Top similar samples.

Figure 3: Shift detection results on COIL-10.

in Figure 3. We can easily see in subfigures (a) and (b) that our testing procedure does not return significant p-values in case the source and target distributions align, while the different dimensionality reduction techniques show varying speeds at detecting a shift when the two distributions do not match. In alignment with our artificially generated shifts, BBSDs is again in the lead, consistently being able to already detect a shift at only 10 samples from the target distribution. Furthermore, the samples provided by the difference classifier (see subfigures (c) and (d)) clearly show images from the first half of the rotation spectrum ($0° - 175°$) being mostly similar to the source distribution and images from the second half of the rotation spectrum ($180° - 355°$) being mostly different.

One surprising finding discovered early in this study was that the original MNIST train/test split appears not to be i.i.d., detected by nearly all methods. As a sanity check we reran the test on a random split, not detecting a shift. Closer inspection revealed significant differences in the means of 6's. Corroborating this finding, the difference classifier singled out test set 6's most confidently.

We see several promising paths for future work, the most crucial ones being: (i) We must extend our work to account intelligently for repeated-two sample tests over time as data streams in; and exploiting the high degree of correlation between adjacent time steps; (ii) Our detection scheme needs to be tested in more machine learning domains; and (ii) In reality all datasets shift, so the bigger challenge is to quantify/characterize the shift, and to decide when practitioners should be alarmed and what actions they should take.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66, 2003.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *BMJ*, 1995.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, 2000.

Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In *International Joint Conference on Artificial intelligence (IJCAI)*, 2005.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 2009.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2014.

Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Journal of Machine Learning Research (JMLR)*, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer vision and pattern recognition (CVPR)*, 2016.

Nicholas A Heard and Patrick Rubin-Delanchy. Choosing between methods of combining-values. *Biometrika*, 2018.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *International Conference on Data Mining (ICDM)*, 2008.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *International conference on computer vision (ICCV)*, 2013.

Thomas M Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis*, 2004.

Markos Markou and Sameer Singh. Novelty detection: a reviewpart 1: statistical approaches. *Signal processing*, 2003.

Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia object image library (coil-100. 1996.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2015.

Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*, 2016.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 2002.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, 2017.

Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems (NIPS)*, 2000.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *International Conference on Machine Learning (ICML)*, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 2000.

R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 1986.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 2009.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems (NIPS)*, 2008.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. A review of change point detection methods. *arXiv preprint arXiv:1801.00718*, 2018.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *arXiv preprint arXiv:1212.4966*, 2018.

Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 2002.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*, 2013.

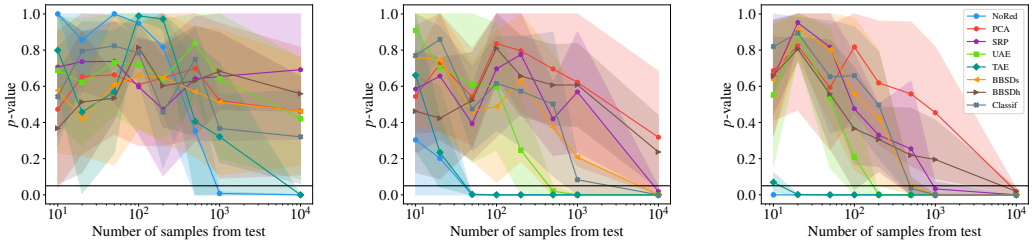
Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018.

A Detailed Shift Detection Results

Our complete shift detection results in which we evaluate different kinds of target shifts on MNIST and CIFAR-10 using the proposed methods are documented below. In addition to our artificially generated shifts, we also evaluated our testing procedure on the original splits provided by MNIST, Fashion MNIST, CIFAR-10, and SVHN.

A.1 Artificially Generated Shifts

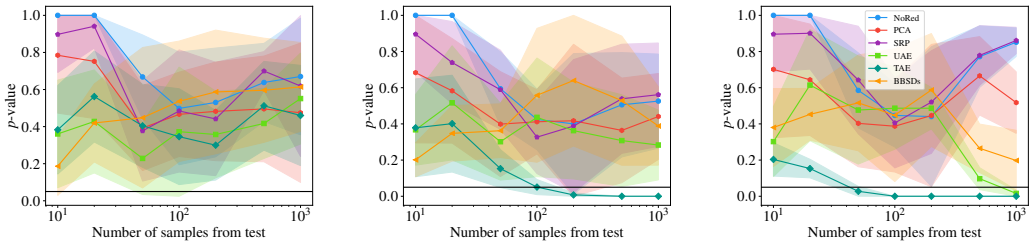
A.1.1 MNIST



(a) 10% adversarial samples. (b) 50% adversarial samples. (c) 100% adversarial samples.

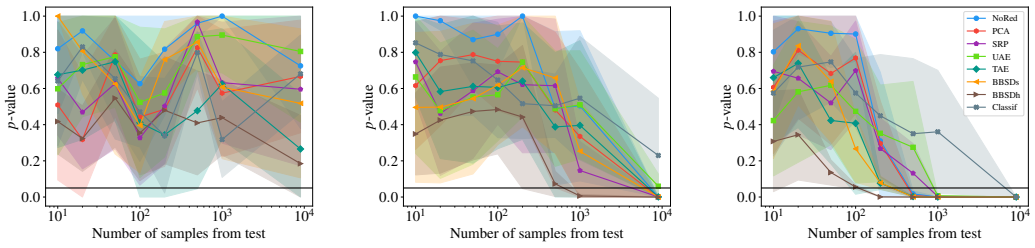
(d) Top different samples. (e) Top similar samples.

Figure 4: MNIST adversarial shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% adversarial samples. (b) 50% adversarial samples. (c) 100% adversarial samples.

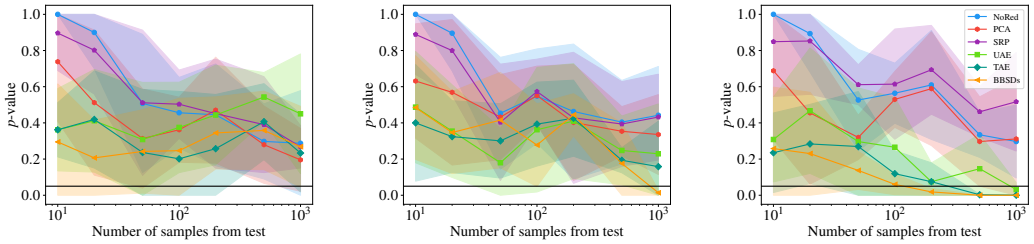
Figure 5: MNIST adversarial shift, multivariate two-sample tests.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

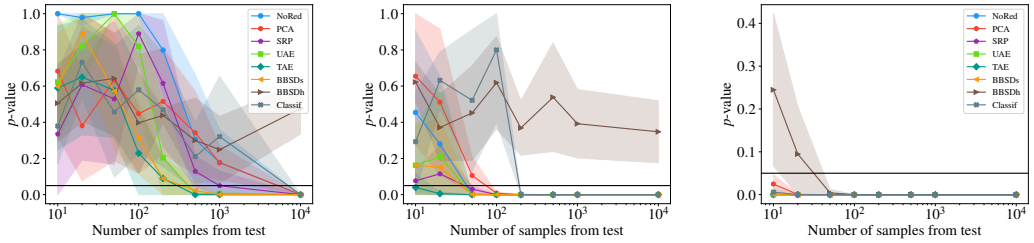
(d) Top different samples. (e) Top similar samples.

Figure 6: MNIST knock-out shift, univariate two-sample tests + Bonferroni aggregration.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

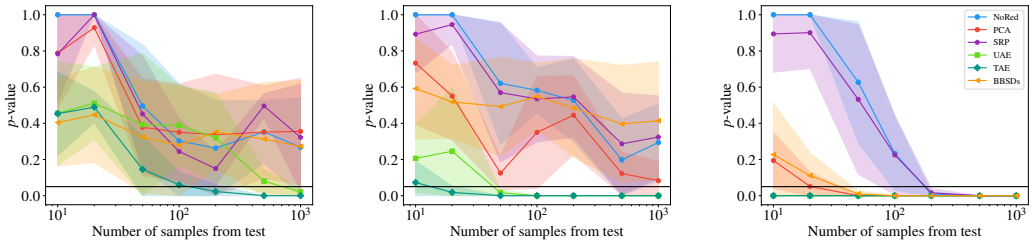
Figure 7: MNIST knock-out shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

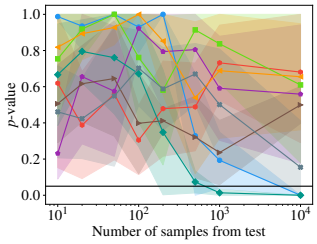
(d) Top different samples. (e) Top similar samples.

Figure 8: MNIST large Gaussian noise shift, univariate two-sample tests + Bonferroni aggregration.

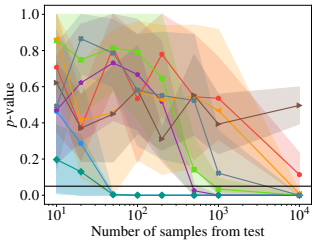


(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

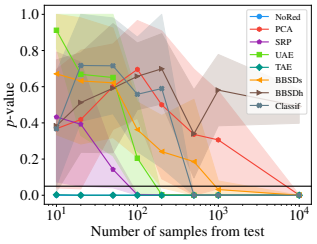
Figure 9: MNIST large Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples.



(b) 50% perturbed samples.

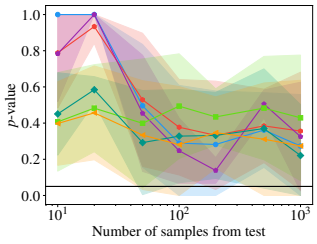


(c) 100% perturbed samples.

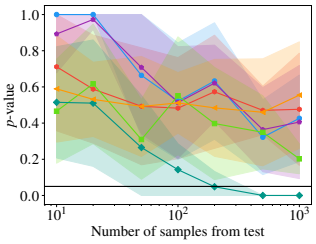
(d) Top different samples.

(e) Top similar samples.

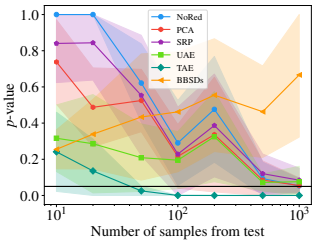
Figure 10: MNIST medium Gaussian noise shift, univariate two-sample tests + Bonferroni aggregation.



(a) 10% perturbed samples.

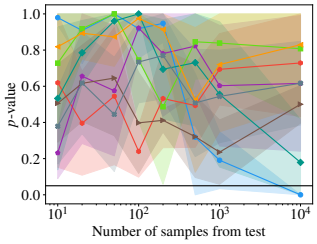


(b) 50% perturbed samples.

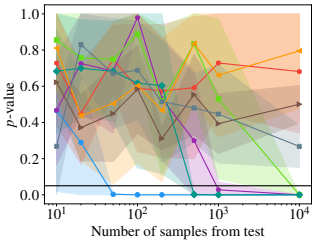


(c) 100% perturbed samples.

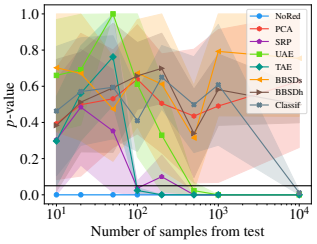
Figure 11: MNIST medium Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples.



(b) 50% perturbed samples.

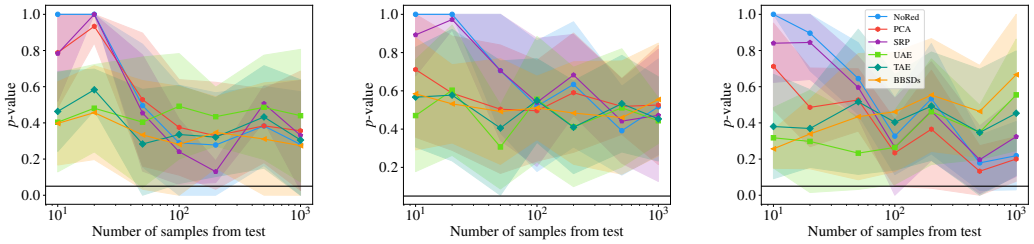


(c) 100% perturbed samples.

(d) Top different samples.

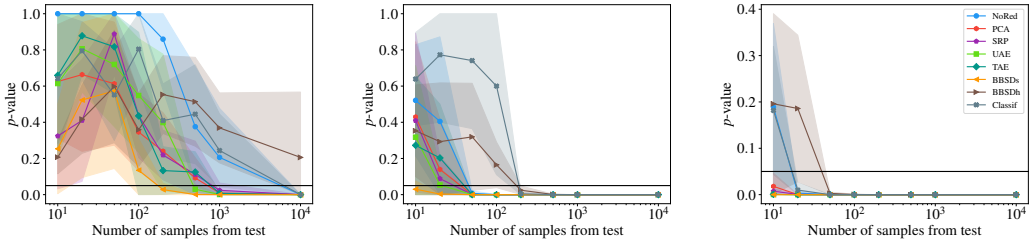
(e) Top similar samples.

Figure 12: MNIST small Gaussian noise shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

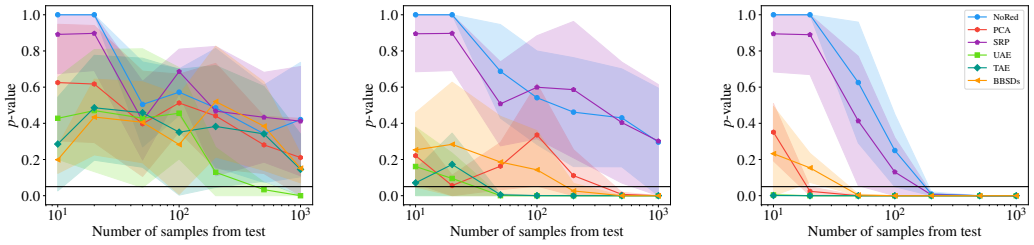
Figure 13: MNIST small Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

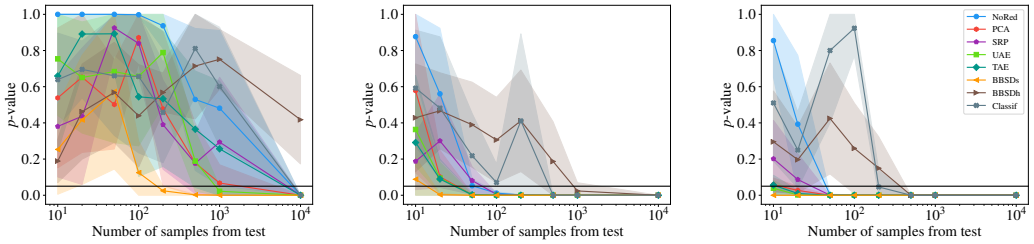
(d) Top different samples. (e) Top similar samples.

Figure 14: MNIST large image shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

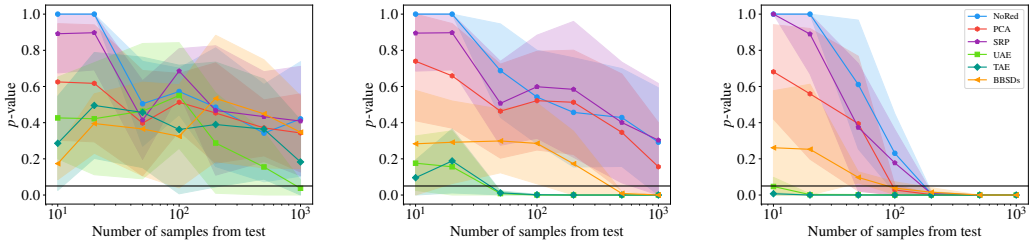
Figure 15: MNIST large image shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

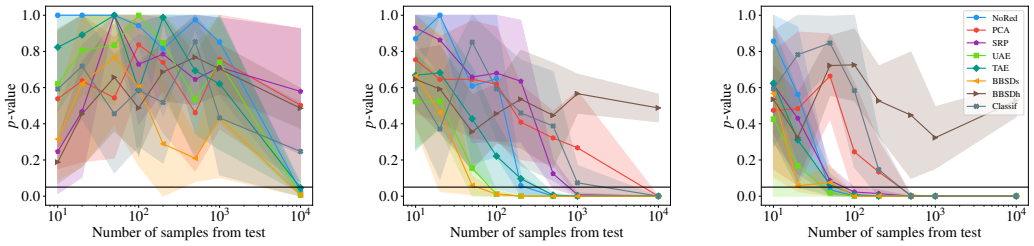
(d) Top different samples. (e) Top similar samples.

Figure 16: MNIST medium image shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

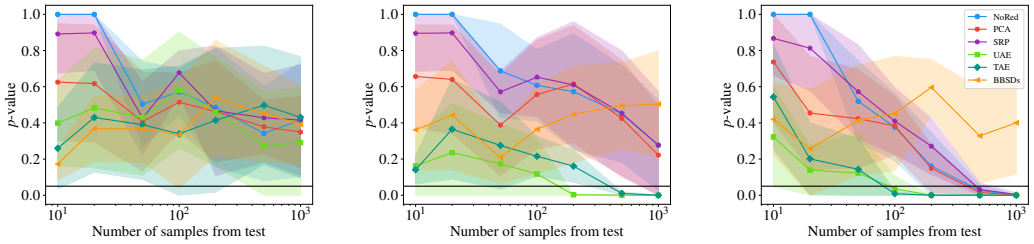
Figure 17: MNIST medium image shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

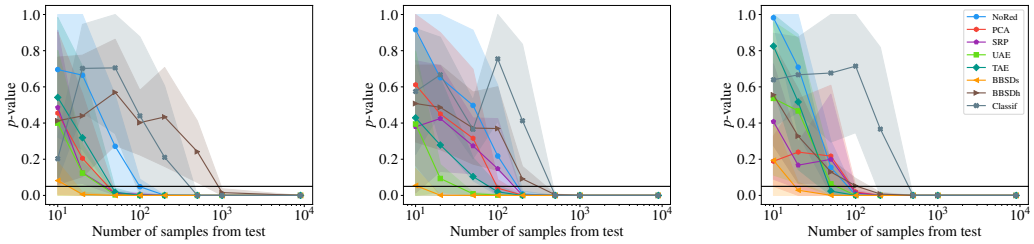
(d) Top different samples. (e) Top similar samples.

Figure 18: MNIST small image shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

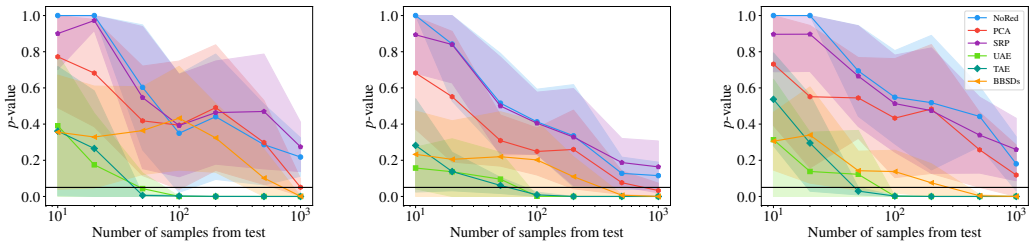
Figure 19: MNIST small image shift, multivariate two-sample tests.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

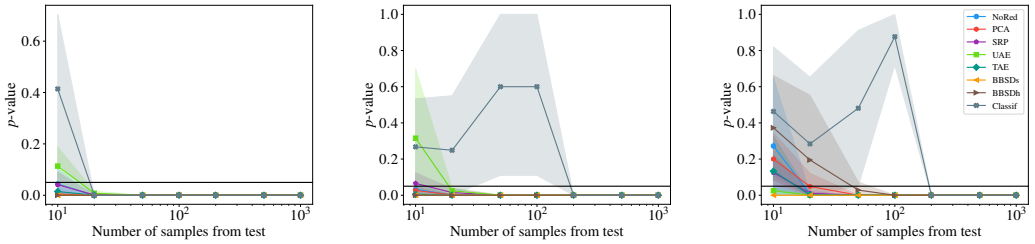
(d) Top different samples. (e) Top similar samples.

Figure 20: MNIST medium image shift (50%, fixed) plus knock-out shift (variable), univariate two-sample tests + Bonferroni aggregration.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

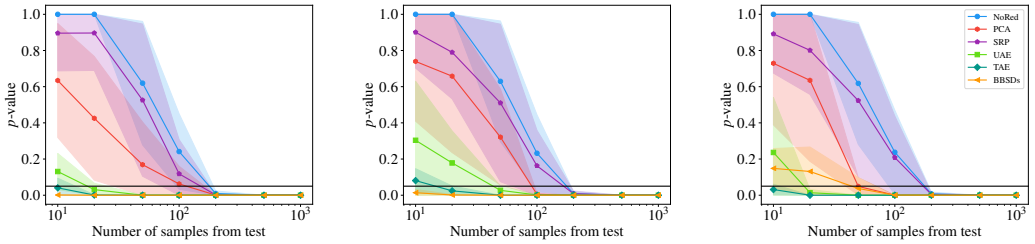
Figure 21: MNIST medium image shift (50%, fixed) plus knock-out shift (variable), multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

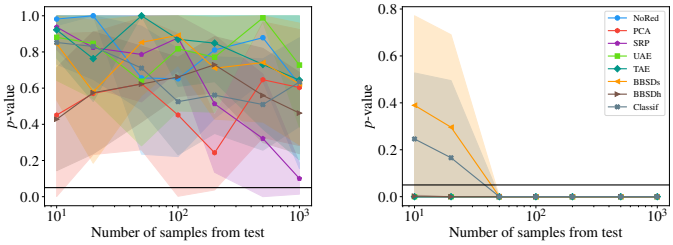
(d) Top different samples. (e) Top similar samples.

Figure 22: MNIST only-zero shift (fixed) plus medium image shift (variable), univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

Figure 23: MNIST only-zero shift (fixed) plus medium image shift (variable), multivariate two-sample tests.

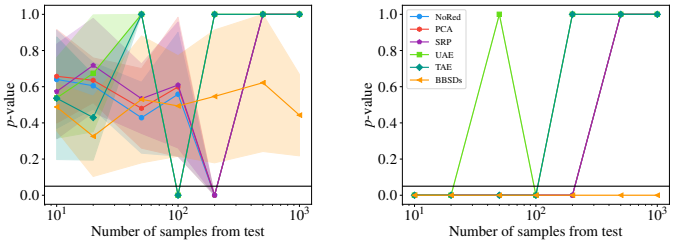


(a) Randomly shuffled dataset with same split proportions as original dataset.

(b) Original split.

(c) Top different samples. (d) Top similar samples.

Figure 24: MNIST to USPS domain adaptation, univariate two-sample tests + Bonferroni aggregration.

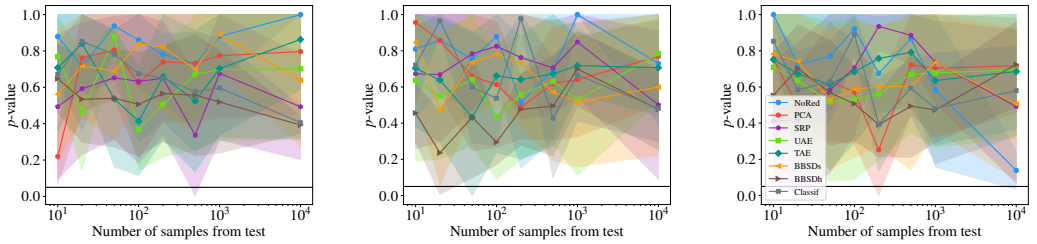


(a) Randomly shuffled dataset with same split proportions as original dataset.

(b) Original split.

Figure 25: MNIST to USPS domain adaptation, multivariate two-sample tests.

A.1.2 CIFAR-10



(a) 10% adversarial samples.

(b) 50% adversarial samples.

(c) 100% adversarial samples.

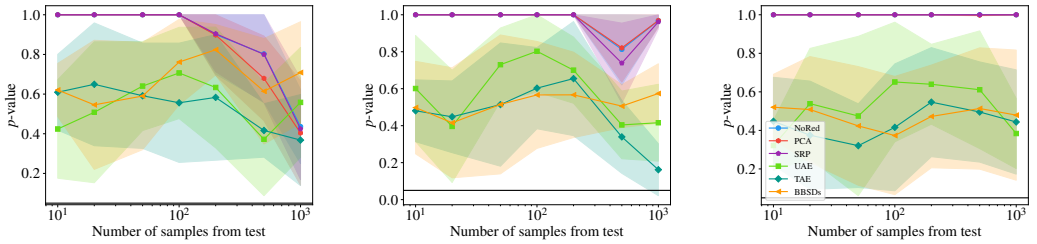
No samples available as *Classif* did not detect a shift.

(d) Top different samples.

No samples available as *Classif* did not detect a shift.

(e) Top similar samples.

Figure 26: CIFAR-10 adversarial shift, univariate two-sample tests + Bonferroni aggregration.

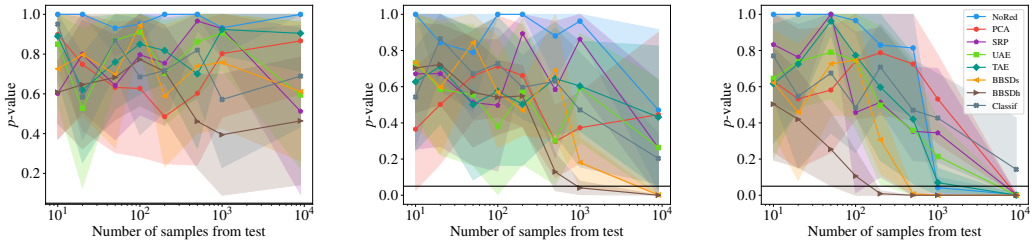


(a) 10% adversarial samples.

(b) 50% adversarial samples.

(c) 100% adversarial samples.

Figure 27: CIFAR-10 adversarial shift, multivariate two-sample tests.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

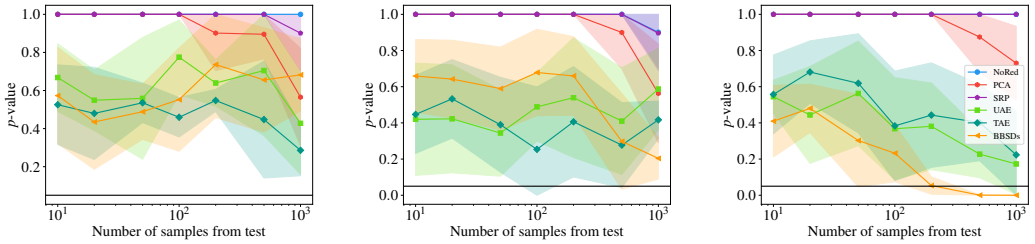
No samples available as *Classif* did not detect a shift.

(d) Top different samples.

No samples available as *Classif* did not detect a shift.

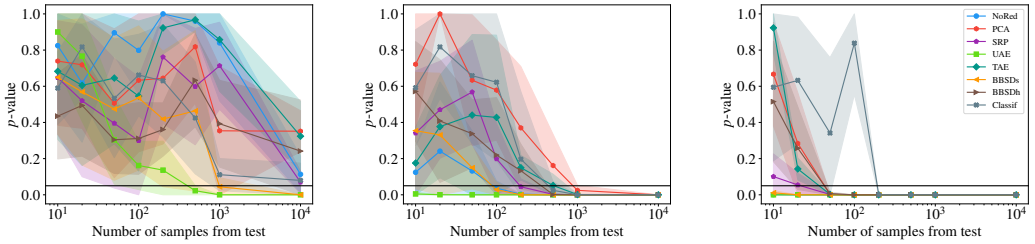
(e) Top similar samples.

Figure 28: CIFAR-10 knock-out shift, univariate two-sample tests + Bonferroni aggregration.



(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

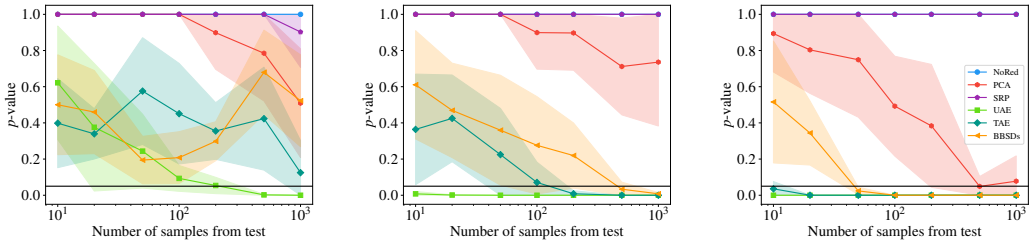
Figure 29: CIFAR-10 knock-out shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

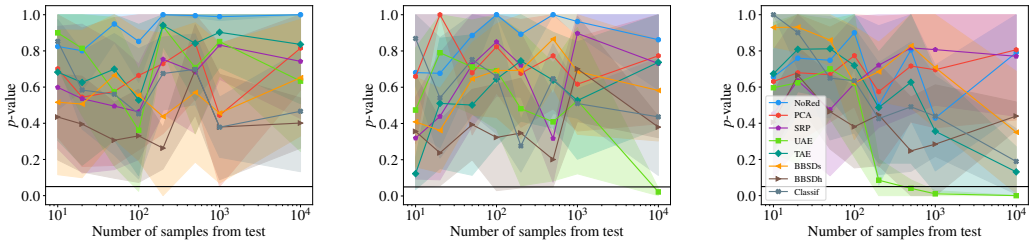
(d) Top different samples. (e) Top similar samples.

Figure 30: CIFAR-10 large Gaussian noise shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

Figure 31: CIFAR-10 large Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

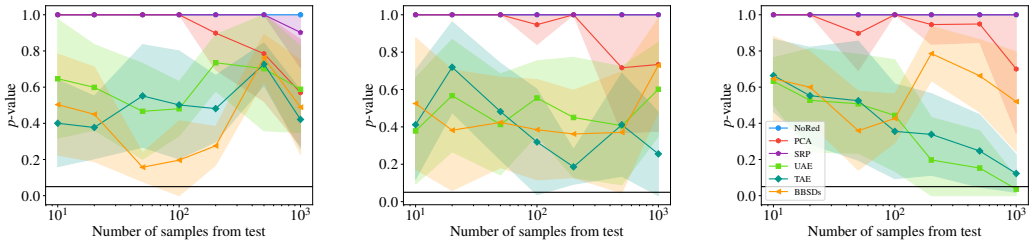
No samples available as *Classif* did not detect a shift.

(d) Top different samples.

No samples available as *Classif* did not detect a shift.

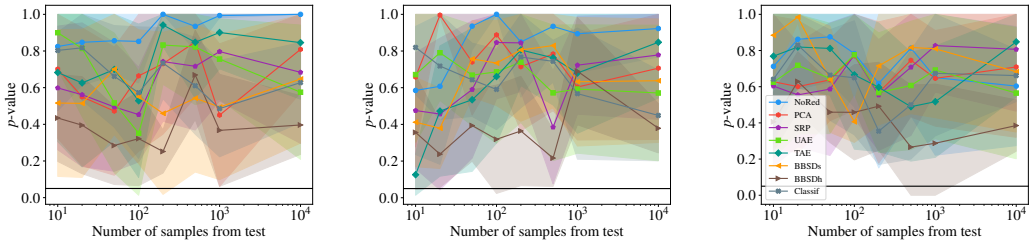
(e) Top similar samples.

Figure 32: CIFAR-10 medium Gaussian noise shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

Figure 33: CIFAR-10 medium Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

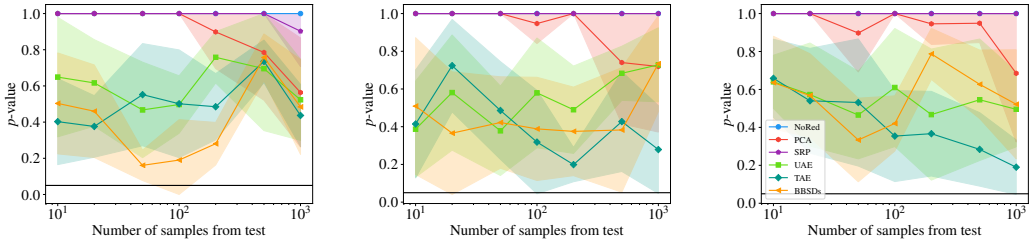
No samples available as *Classif* did not detect a shift.

(d) Top different samples.

No samples available as *Classif* did not detect a shift.

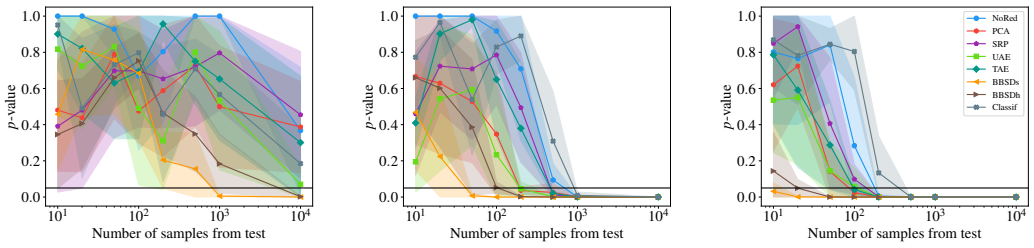
(e) Top similar samples.

Figure 34: CIFAR-10 small Gaussian noise shift, univariate two-sample tests + Bonferroni aggregation.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

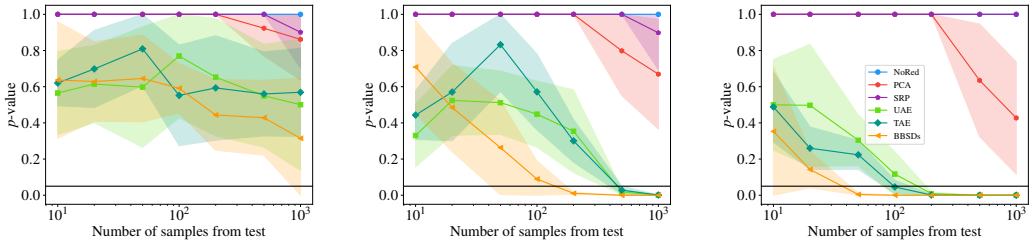
Figure 35: CIFAR-10 small Gaussian noise shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

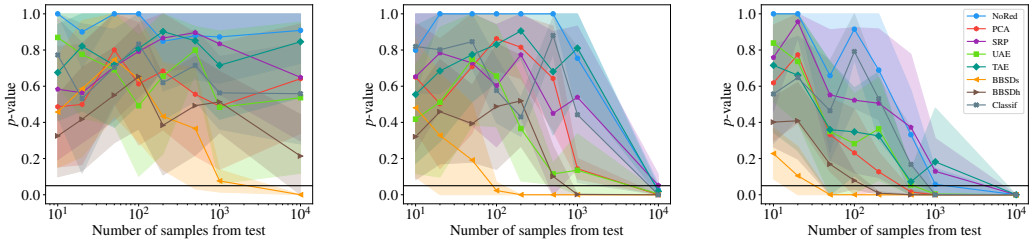
(d) Top different samples. (e) Top similar samples.

Figure 36: CIFAR-10 large image shift, univariate two-sample tests + Bonferroni aggregation.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

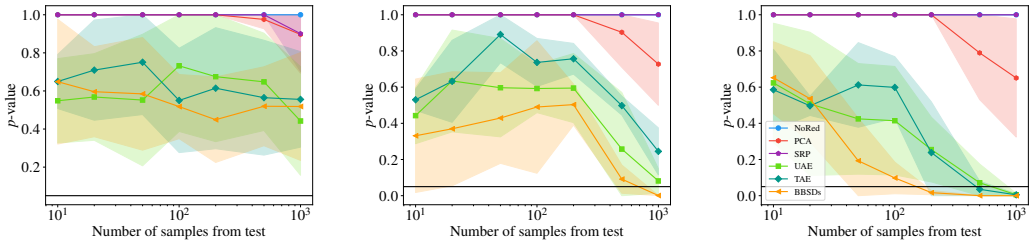
Figure 37: CIFAR-10 large image shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

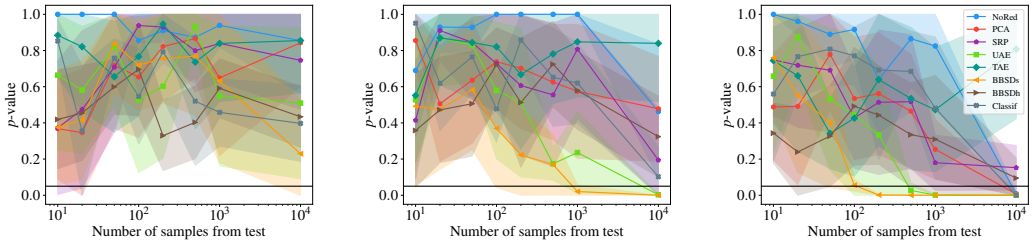
(d) Top different samples. (e) Top similar samples.

Figure 38: CIFAR-10 medium image shift, univariate two-sample tests + Bonferroni aggregration.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

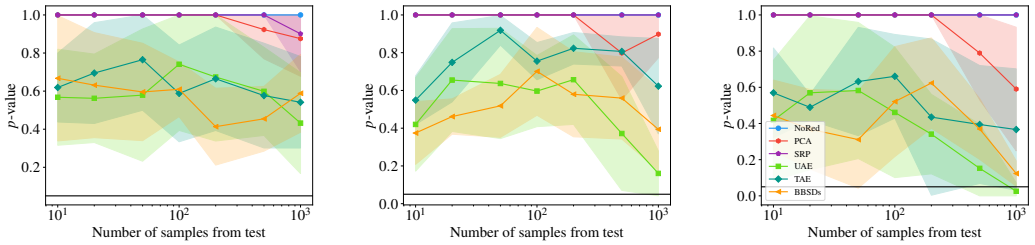
Figure 39: CIFAR-10 medium image shift, multivariate two-sample tests.



(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

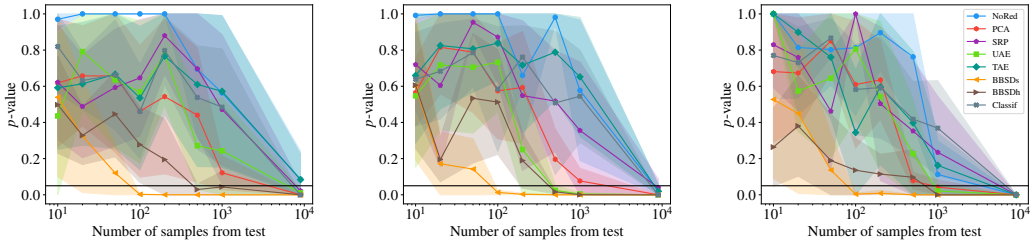
(d) Top different samples. (e) Top similar samples.

Figure 40: CIFAR-10 small image shift, univariate two-sample tests + Bonferroni aggregration.

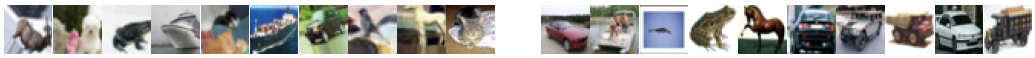


(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

Figure 41: CIFAR-10 small image shift, multivariate two-sample tests.

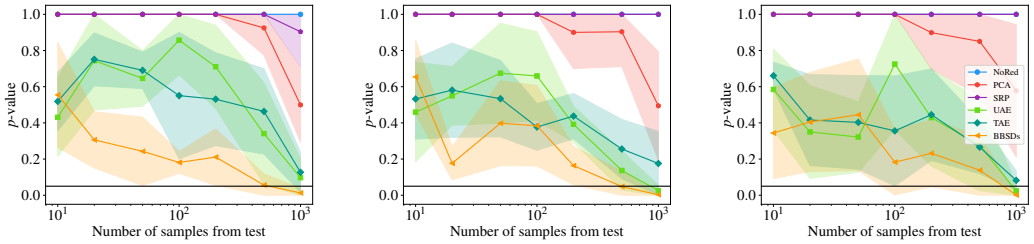


(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.



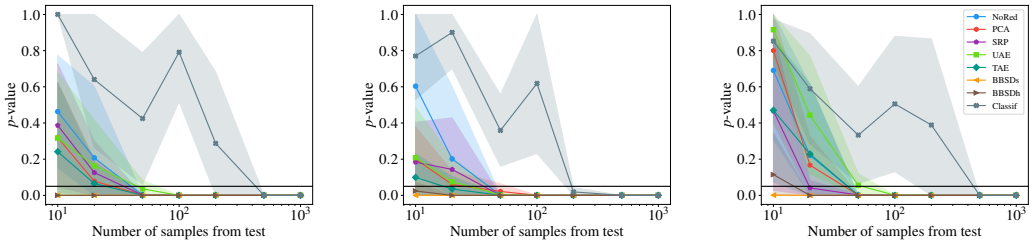
(d) Top different samples. (e) Top similar samples.

Figure 42: CIFAR-10 medium image shift (50%, fixed) plus knock-out shift (variable), univariate two-sample tests + Bonferroni aggregration.

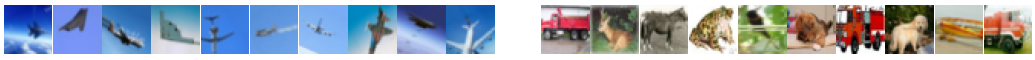


(a) Knock out 10% of class 0. (b) Knock out 50% of class 0. (c) Knock out 100% of class 0.

Figure 43: CIFAR-10 medium image shift (50%, fixed) plus knock-out shift (variable), multivariate two-sample tests.

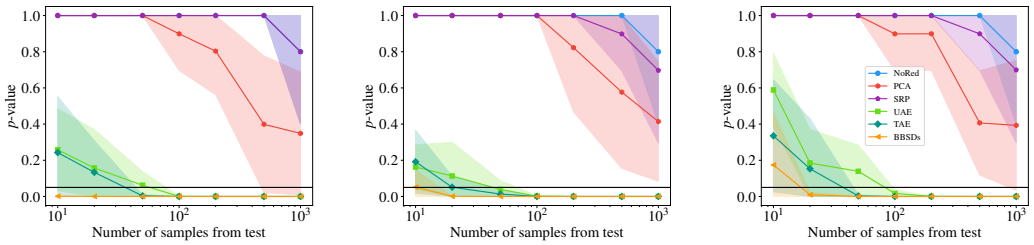


(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.



(d) Top different samples. (e) Top similar samples.

Figure 44: CIFAR-10 only-zero shift (fixed) plus medium image shift (variable), univariate two-sample tests + Bonferroni aggregation.

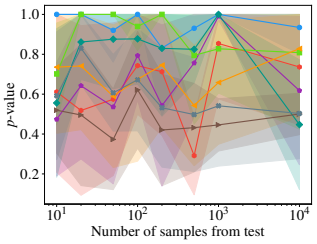


(a) 10% perturbed samples. (b) 50% perturbed samples. (c) 100% perturbed samples.

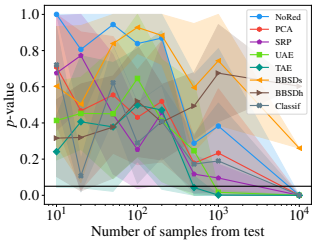
Figure 45: CIFAR-10 only-zero shift (fixed) plus medium image shift (variable), multivariate two-sample tests.

A.2 ORIGINAL SPLITS

A.2.1 MNIST



(a) Randomly shuffled dataset with same split proportions as original dataset.

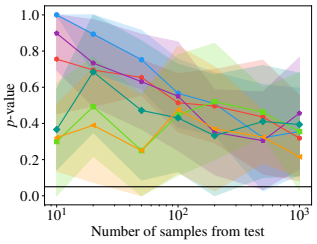


(b) Original split.

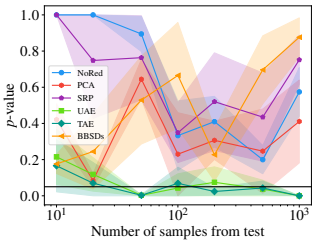
(c) Top different samples.

(d) Top similar samples.

Figure 46: MNIST randomized and original split, univariate two-sample tests + Bonferroni aggregation.



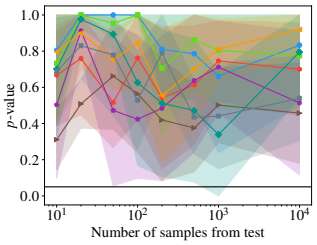
(a) Randomly shuffled dataset with same split proportions as original dataset.



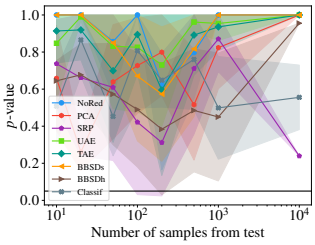
(b) Original split.

Figure 47: MNIST randomized and original split, multivariate two-sample tests.

A.2.2 FASHION MNIST



(a) Randomly shuffled dataset with same split proportions as original dataset.



(b) Original split.

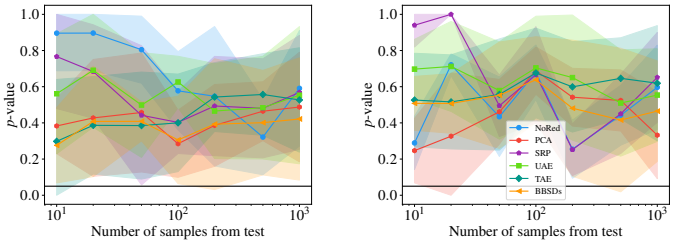
No samples available as *Classif* did not detect a shift.

(c) Top different samples.

No samples available as *Classif* did not detect a shift.

(d) Top similar samples.

Figure 48: Fashion MNIST randomized and original split, univariate two-sample tests + Bonferroni aggregration.

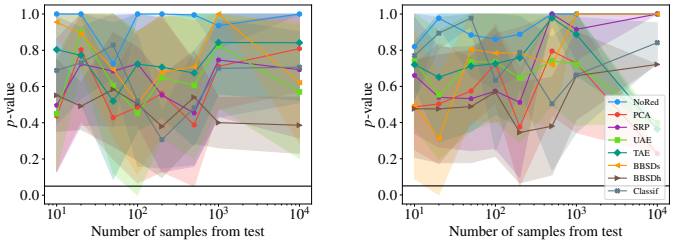


(a) Randomly shuffled dataset with same split proportions as original dataset.

(b) Original split.

Figure 49: Fashion MNIST randomized and original split, multivariate two-sample tests.

A.2.3 CIFAR-10



(a) Randomly shuffled dataset with same split proportions as original dataset.

(b) Original split.

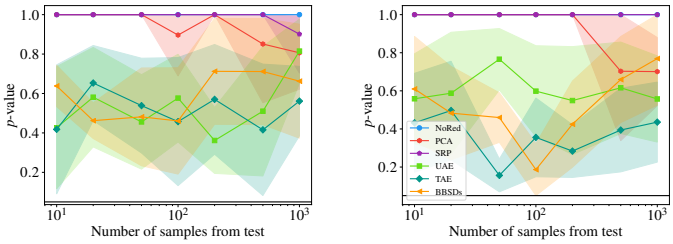
No samples available as *Classif* did not detect a shift.

(c) Top different samples.

No samples available as *Classif* did not detect a shift.

(d) Top similar samples.

Figure 50: CIFAR-10 randomized and original split, univariate two-sample tests + Bonferroni aggregration.

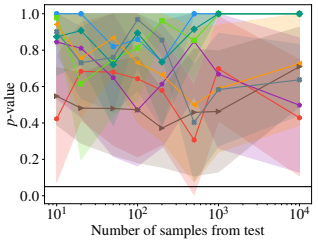


(a) Randomly shuffled dataset with same split proportions as original dataset.

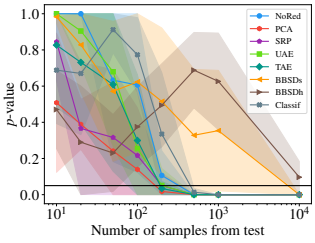
(b) Original split.

Figure 51: CIFAR-10 randomized and original split, multivariate two-sample tests.

A.2.4 SVHN



(a) Randomly shuffled dataset with same split proportions as original dataset.

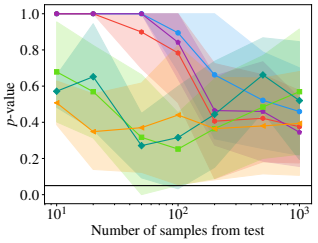


(b) Original split.

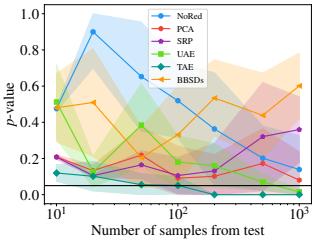
(c) Top different samples.

(d) Top similar samples.

Figure 52: SVHN randomized and original split, univariate two-sample tests + Bonferroni aggregation.



(a) Randomly shuffled dataset with same split proportions as original dataset.



(b) Original split.

Figure 53: SVHN randomized and original split, multivariate two-sample tests.