A Distribution-Free Test of Covariate Shift Using Conformal Prediction

Xiaoyu Hu and Jing Lei

Kiho Park
parkkiho@uchicago.edu

1 Introduction

Covariate shift is a common and important assumption in transfer learning and domain adaptation to treat the distributional difference between the training and testing data. The purpose of this paper is to provide a statistical test for the hypothesis of covariate shift. We are interested in testing the null hypothesis that the two conditional distributions are the same.

Our new method for testing covariate shift has three remarkable features. First, out test statistic is constructed using the conformal prediction technique. Second, our method does not require estimating the density functions. Third, the asymptotic null distribution of our test statistic and its universal power guarantee are rigorously established under certain moment conditions on the density ratios and the accuracy of classification algorithms.

2 Problem formulation and preliminaries

2.1 The two-sample conditional distribution testing problem

Consider two independent random samples $\{(X_{1i}, Y_{1i})\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} P_1$ and $\{(X_{2i}, Y_{2i})\}_{i=1}^{n_2} \overset{i.i.d.}{\sim} P_2$ where P_1, P_2 are distributions on $\mathcal{X} \times \mathcal{Y}$. We assume that we have a sample from P_2 to make testing covariate shift possible. For $j = 1, 2$, let $P_j(\cdot \mid x)$ be the conditional distribution of Y given $X = x$ under P_j, and let $P_{j,X}(\cdot)$ be the corresponding marginal distribution of X. We are interested in testing $P_1(\cdot \mid x) = P_2(\cdot \mid x)$ for all $x \in \mathcal{X}$. We assume that $P_{1,X}$ and $P_{2,X}$ are equivalent: $P_{1,X} \ll P_{2,X}$ and $P_{2,X} \ll P_{1,X}$ where "$\ll$" stands for absolute continuity. We further assume $P_{1,X}, P_{2,X}$ have density functions $f_{1,X}, f_{2,X}$ respectively, under a common base measure. We also assume that $P_1(\cdot \mid x)$ and $P_2(\cdot \mid x)$ are equivalent, with density functions $f_1(y \mid x)$ and $f_2(y \mid x)$. We consider the two-sample conditional distribution testing problem:

$$H_0 : P_{1,X}\{P_1(\cdot \mid X) = P_2(\cdot \mid X)\} = 1, \text{ versus } H_1 : P_{1,X}\{P_1(\cdot \mid X) = P_2(\cdot \mid X)\} < 1$$

2.2 Background on conformal prediction

Given iid data $\{(X_i, Y_i)\}_{i=1}^m$, conformal prediction converts a point estimator of the regression function $\widehat{f} : \mathcal{X} \to \mathcal{Y}$ to a prediction set $\widehat{C} \in \mathcal{X} \times \mathcal{Y}$ with guaranteed finite-sample expected prediction coverage: $\mathbb{P}\left\{Y_{m+1} \in \widehat{C}(X_{m+1})\right\} \geq 1 - \alpha$, where $\widehat{C}(x) = \left\{y \in \mathcal{Y} : (x, y) \in \widehat{C}\right\}$, and the probability is taken over the $(m+1)$-tuple of iid data.

Let $\mathcal{D}_m$ denote the sample $\{(X_i, Y_i) : 1 \leq i \leq m\}$, and $\mathcal{D}_{m,-i}$ the sample obtained by removing (X_i, Y_i) from $\mathcal{D}_m$. Let $\mathcal{D}_{m+1}(y)$ be the augmented data set with the $(m+1)$th data point (X_{m+1}, y). Using a conformity score function $V : (\mathcal{X} \times \mathcal{Y})^{m+1} \to \mathbb{R}$ that is symmetric in its first m inputs, we let $V_i(y) = V(\mathcal{D}_{m+1,-i}(y), (X_i, Y_i))$ for $i = 1, \cdots, m+1$ be the conformity score for the ith sample point in the augmented

data with $Y_{m+1} = y$. Then, the conformal prediction set

$$\widehat{C}(X_{m+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^{m+1} \mathbb{1}[V_i(y) \leq V_{n+1}(y)] \geq \lfloor (m+1)\alpha \rfloor \right\}$$

satisfies the finite sample coverage.

A good choice of V needs to reflect the structure of the underlying distribution of (X, Y). The conformity score function V consists of a fitting part and a conformity assessment part. If both parts accurately reflect the underlying joint distribution, the resulting prediction set will not only have coverage guarantee, but will also be nearly optimal.

3 A conformal Test of covariate shift

3.1 The conformal p-value

The construction of conformal prediction set can be viewed as inverting a hypothesis test $(\mathcal{D}_m, (X_{m+1}, Y_{m+1})) \overset{i.i.d.}{\sim} P$ with observed value $Y_{m+1} = y$, using the test statistic

$$U(y) = \frac{1}{m+1} \sum_{i=1}^{m} \mathbb{1}(V_i(y) \leq V_{m+1}(y))$$

which behaves like an $U(0,1)$ random variable under the null hypothesis that $(\mathcal{D}_m, (X_{m+1}.Y_{m+1}))$ is a realized iid sample from P. This provides a way to generate a valid p-value.

We first treat the simpler case where $P_{1,X} = P_{2,X}$. If the covariate shift hypothesis H_0 is true, $(X_1, Y_1), \cdots, (X_m, Y_m), (X_{m+1}, Y_{m+1})$ are iid from P_1. Let $R_- = 1 + \sum_{i=1}^{m+1} \mathbb{1}(V_i(Y_{m+1}) < V_{m+1}(Y_{m+1}))$ and $R_+ = \sum_{i=1}^{m+1} \mathbb{1}(V_i(Y_{m+1}) \leq V_{m+1}(Y_{m+1}))$. Let R be uniformly and independently sample from the integers in $[R_-, R_+]$. Now we can construct a uniform random variable

$$U = \frac{R - 1 + \zeta}{m+1}$$

where $\zeta \sim U(0,1)$ is independent of everything else. Since R has a uniform distribution on $\{1, \cdots, m+1\}$ by exchangeability under the null hypothesis, U has a uniform distribution on $[0,1]$. Thus, it can be used as an exact p-value to test the covariate shift hypothesis. We call such a statistics U a conformal p-value.

3.2 A choice of V that separates H_0 and H_1

Assume $P_{1,X} = P_{2,X}$. We want to choose V such that the conformal p-value U has a non-uniform distribution under the alternative hypothesis H_1. Our choice of V is the conditional Radon-Nikodym derivative between P_1 and P_2:

$$V(x, y) = \frac{dP_1(y \mid x)}{dP_2(y \mid x)} = \frac{f_1(y \mid x)}{f_2(y \mid x)}$$

Lemma 1 (Separation of H_0 and H_1 by V under equal X-marginal)**.** *If $P_{1,X} = P_{2,X}$, then there exist $\delta > 0$ and m_0, depending only on (P_1, P_2), such that the statistic U with $V(\mathcal{D}_{m+1,-i}, (x, y)) = \frac{f_1(y|x)}{f_2(y|x)}$ satisfies $\mathbb{E}(U) \leq 1/2 - \delta$ under H_1 when $m \geq m_0$.*

3.3 Incorporating multiple testing sample points for better power

A single p-value often has limited power in distinguishing H_1 from H_0. To have a test with asymptotically full power, we consider multiple testing sample points $\{(X_{2k}, Y_{2k})\}_{k=1}^K$. Given $\{(X_{1i}, Y_{1i})\}_{i=1}^m$ from P_1, we can repeat the procedure for each testing sample point, resulting in $U_1, \cdots, U_K$. However, U_k are dependent as their constructions all involve the same set of training data $\{(X_{1i}, Y_{1i})\}_{i=1}^m$. To make a one-sided mean test over the U_k's work, we construct U_k for each k by using (X_{2k}, Y_{2k}) and $(X_{1i_{k,l}}, Y_{1i_{k,l}})_{l=1}^m$, where each $(X_{1i_{k,l}}, Y_{1i_{k,l}})_{l=1}^m$ is a disjoint batch of training data. Then, U_k are iid and the standard normal approximation theory would be valid in testing the mean of U_k.

3.4 Allowing for $P_{1,X} \neq P_{2,X}$ using weighted conformalization

Now we drop the assumption of equal marginal distribution of X under P_1 and P_2. The $(m+1)$-tuple used to construct U are no longer exchangeable under the null hypothesis, so we use the "weighted conformal prediction" developed in Tibshirani et al. (2019).

We focus on conformity score functions V that only depends on the last argument, and write $V_i = V(X_i, Y_i)$. We assume the data $\mathbf{Z} = (X_i, Y_i)_{i=1}^{m+1}$ is stored in two parts: a randomly permuted sequence $\tilde{\mathbf{Z}}$ and the permutation $\sigma : [m+1] \to [m+1]$. Given $\tilde{\mathbf{Z}}$, $V_{m+1} = V(\tilde{X}_i, \tilde{Y}_i)$ if $\sigma(m+1) = i$.

In the following lemma, we derive the conditional distribution of V_{m+1} given $\tilde{\mathbf{Z}}$, construct the uniformly distributed weighted conformal p-value, and establish its ability to separate H_0 and H_1.

Lemma 2. *(a) Under H_0, for any choice of $V(x, y)$, we have*

$$V_{m+1} \mid \tilde{\mathbf{Z}} \sim \sum_{i=1}^{m+1} p_i(\mathbf{Z}) \delta_{V_i}$$

with

$$p_i(\mathbf{Z}) = \frac{f_{2,X}(X_i)/f_{1,X}(X_i)}{\sum_{l=1}^{m+1} f_{2,X}(X_l)/f_{1,X}(X_l)}, \quad i = 1, \cdots, m+1$$

(b) For any choice of $V(x, y)$, the randomized statistic

$$U = \sum_{i=1}^{m+1} p_i(\mathbf{Z}) \mathbb{1}(V_i < V_{m+1}) + \zeta \sum_{i=1}^{m+1} p_i(\mathbf{Z}) \mathbb{1}(V_i = V_{m+1})$$

has a uniform distribution under H_0, where ζ is an $U(0,1)$ random variable independent of everything else.

(c) Under H_1, if $V(x, y) = \frac{f_1(y|x)}{f_2(y|x)}$, there exist $\delta > 0$ and m_0, depending only P_1, P_2, such that $\mathbb{E}U \leq 1/2 - \delta$ when $m \geq m_0$.

3.5 The conformal covariate shift test algorithm

The following algorithm presents whole steps of the methods. We assume $\mathcal{A}_1$ is an algorithm that outputs an estimate of the function $V(x, y) = f_1(y \mid x)/f_2(y \mid x)$. $\mathcal{A}_2$ is an algorithm that outputs an estimate of the marginal density ratio $g(x) = f_{2,X}(x)/f_{1,X}(x)$. Theoretical results of the asymptotic behavior of the test statistics T under H_0 and H_1 are presented in Section 4. Empirical performance in simulation and a real data example will be presented in Section 5 and Section 6 respectively.

Algorithm 1 Conformal test of covaraite shift

Require: Training data $(X_{1i}, Y_{1i})_{i=1}^{n_1}$; testing data $(X_{2j}, Y_{2j})_{j=1}^{n_2}$; number of minibatches K, minibatch size m; density ratio estimation subroutines $\mathcal{A}_1$, $\mathcal{A}_2$

Randomly split $\{1, \cdots, n_1\}$ into subsets $\mathcal{I}_{11}, \mathcal{I}_{12}$ such that $|\mathcal{I}_{12}| = mK$

Randomly split $\{1, \cdots, n_2\}$ into subsets $\mathcal{I}_{21}, \mathcal{I}_{22}$ such that $|\mathcal{I}_{22}| = K$

Randomly split $\mathcal{I}_{12}$ into K equal-sized batches $\mathcal{J}_k$, $k = 1, \cdots, K$

$\widehat{V}(\cdot, \cdot) = \mathcal{A}_1[\{(X_{1i}, Y_{1i}), i \in \mathcal{I}_{11}, (X_{2j}, Y_{2j}), j \in \mathcal{I}_{21}\}]$

$\widehat{g}(\cdot) = \mathcal{A}_2[\{X_{1i}, i \in \mathcal{I}_{11}, X_{2j}, j \in \mathcal{I}_{21}\}]$

for $k \in \mathcal{I}_{22}$ and $\mathcal{J}_k = \{i_{k,1}, \cdots, i_{k,m}\}$ **do**

  for $l \in \{1, \cdots, m\}$ **do**

    $\widehat{V}_{k,l} = \widehat{V}(X_{1i_{k,l}}, Y_{1i_{k,l}})$ and $\widehat{p}_{k,l} = \widehat{g}(X_{1i_{k,l}}) / \{\sum_{l=1}^m \widehat{g}(X_{1i_{k,l}}) + \widehat{g}(X_{2k})\}$

  end for

  $\widehat{V}_{k,m+1} = \widehat{V}(X_{2k}, Y_{2k})$ and $\widehat{p}_{k,m+1} = \widehat{g}(X_{2k}) / \{\sum_{l=1}^m \widehat{g}(X_{1i_{k,l}}) + \widehat{g}(X_{2k})\}$

  Generate $\zeta_k \sim U(0,1)$, independent of everything else

  $\widehat{U}_k = \sum_{l=1}^m \widehat{p}_{k,l} \mathbb{1}\left(\widehat{V}_{k,l} < \widehat{V}_{k,m+1}\right) + \zeta_k \left\{\widehat{p}_{k,m+1} + \sum_{l=1}^m \widehat{p}_{k,l} \mathbb{1}\left(\widehat{V}_{k,l} = \widehat{V}_{k,m+1}\right)\right\}$

end for

$T = \sqrt{12K}\left(1/2 - \sum_{k=1}^K \widehat{U}_k / K\right)$

Reject H_0 if $T \geq \Phi^{-1}(1-\alpha)$

Remark. *1) If we assume that the marginal distributions $P_{1,X}$, $P_{2,X}$ are equal, we can just assign all $\widehat{p}_{k,l}$ the theoretical value $1/(m+1)$ without using $\mathcal{A}_2$.*

2) The test will be more powerful when m is large, and K needs to be large enough for the normal approximation. In practice, we recommend setting K to be a small fraction of n_2 such as $n_2/\log(n_2)$ while not exceeding $n_1/(2m)$.

3) The problem of estimating density ratios is often easier than estimating the density functions themselves. There are some methods such as moment matching approach, the ratio matching approach, and probabilistic classification approach.

4 Asymptotic properties

Three quantities play different roles in determing different sample sizes: the fitting subsample size $n \equiv \min(n_1 - mK, n_2 - K)$; the ranking batch size m; and the number of minibatches K. We denote the subsample used to fit $\widehat{V}, \widehat{g}$ as the fitting subsample:

$$\mathcal{D}_{\text{fit}} \equiv (\{(X_{1i}, Y_{1i}) : i \in \mathcal{I}_{11}\}, \{(X_{2j}, Y_{2j}) : j \in \mathcal{I}_{21}\})$$

For $q \geq 1$, $j = 1, 2$, denote the ℓ_q norm of a function γ under distribution P_j by $\|\gamma\|_{q,P_j}^q \equiv \mathbb{E}_{(X,Y)\sim P_j} |\gamma(X,Y)|^q$. Our first assumption puts some moment conditions on the marginal density ratio $g(\cdot)$.

Assumption 1. *The marginal likelihood ratio $g(x) = f_{2,X}(x)/f_{1,X}(x)$ satisfies $\|g\|_{2,P_j} < \infty$ and $\|g^{-1}\|_{2,P_j} < \infty$ for $j = 1, 2$.*

Our next assumption quantifies the required accuracy of the density ratio estimators.

Assumption 2. *When* $n = \min\{n_1 - mK, n_2 - K\} \to \infty$ *and* $K \to \infty$*, we have*

$$(a) \quad \left\| \widehat{V} - V \right\|_{1,P_2} = o_P(1)$$

$$(b) \quad r_n \equiv \max\left\{ \|\widehat{g} - g\|_{2,P_1}, \|\widehat{g} - g\|_{2,P_2} \right\} = o_P(K^{-1/2})$$

Theorem 1 (Asymptotic null distribution)**.** *Suppose that Assumptions 1 and 2(b) hold. For any estimate* $\widehat{V}$ *and value of* m*, the statistic* T *output by Algorithm 1 converges in distribution to the standard normal as* $n, K \to \infty$ *under* H_0*.*

In the case $P_{1,X} = P_{2,X}$, we have $\widehat{g} \equiv 1$ and both assumptions hold. When $P_{1,X} \neq P_{2,X}$, one can use the out-of-sample classification error as a proxy to the accuracy of $\widehat{g}$.

Theorem 2 (Asymptotic power guarantee)**.** *Suppose that Assumptions 1, 2 hold. There exists constants* $C > 0$*,* $\delta > 0$ *and* m_0 *depending only on* P_1, P_2*, such that under* H_1*, the empirical conformal p-values* $\widehat{U}_k$ *in Algorithm 1 satisfies* $\mathbb{E}\left(\widehat{U}_k \mid \mathcal{D}_{\text{fit}}\right) \leq 1/2 - \delta + Cr_n$ *when* $m \geq m_0$*. As a consequence, when* $m \geq m_0$*, the test statistic* T *output by Algorithm 1 satisfies* $T \xrightarrow{p} +\infty$ *as* $n, K \to \infty$*.*

As a consequence, for any nominal level $\alpha \in (0, 1)$, the probability of rejecting the null hypothesis H_0 tends to 1 under the alternative hypothesis H_1.

5 Simulation study

We illustrate the performance of our method in several simulation settings: (A) Gaussian, linear; (B) non-linear, heavy-tailed; (C) linear, heteroskedastic, correlated features; (D) linear, quadratic shift.

5.1 The low dimensional case

We first consider low-dimensional cases with $p = 5$. We use four different probabilistic classification methods including linear logistic (LL), quadratic logistic (QL), neural network (NN), and random forest (RF) to estimate V and g. All the simulation results are computed over 1000 repetitions with nominal type I error level $\alpha = 0.05$. We calculate probability weight estimation error $Err_{\widehat{p}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{m+1}|\widehat{p}_{k,l} - p_{k,l}|$ to evaluate the accuracy of estimated probability weights.

As we can see in the results, the empirical sizes are always close to the nominal level α as long as the true g is used or a good estimate is plugged in. The LL estimator have very good performance even with small sample sizes. The NN and QL estimators requires larger sample sizes to achieve comparable estimation accuracy. In contrast, the RF estimator fails to control the type I error in all models because of higher error in estimating $p_{k,l}$.

When the alternative hypothesis is true, the power increases as the sample size increases. In Model A, the LL and NN methods yield comparable power even when the sample size is small. In model B, the NN method is flexible to capture the complicated relationships and yields superior power. Although the LL and QL estimators are misspecified, they still have non-trivial power.

For Model C, the observations are very similar to those in Models A and B. In Model D, the QL is the correct model and works well. However, LL method has poor power.

5.2 The high dimensional case

We illustrate the performance in a high-dimensional scenario: Model A with ambient dimensionality $p = 500$ and signal dimensionality $s = 5$. We focus on a sparse linear classifier and investigate the effect of tuning and regularization. The figures show the empirical rejection frequency, estimation errors of the probability weights $p_{k,l}$, out-of-sample marginal classification error (MCE) involved in estimating $\widehat{g}$, and empirical out-of-sample estimation error of V defined by $Err_{\widehat{V}} = ((m+1)K)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{m+1} \left(\widehat{V}_{k,l} - V_{k,l}\right)^2$.

When the true marginal density ratios are used, the empirical sizes are close to the nominal level $\alpha = 0.05$ as expected. When the probability weights are estimated, the type I error is well controlled for a wide range of tuning parameter values, indicating good robustness of validity. The plot of MCE suggets that in practice one can choose the tuning parameter λ value near the elbow of the error plot. Under the alternative hypotheses, the power is maximized at λ corresponding to the smallest estimation error in $\widehat{V}$.

6 Real data

We consider the airfoil data set from UCI Machine learning Repository, which has been used by Tibshirani et al. (2019). We consider five experiments based on different ways to generate the training and testing datasets: (i) Random partition, (ii) Random partition and exponential tilting, (iii) Chord-based partition, (iv) Velocity-based partition, and (v) Response-based partition. LL and NN are used to estimate the density ratios. Each experiment is repeated for 1000 trials.

For experiments (i) and (ii), the type I error is close to the nominal level. For experiments (iii-v), we use a median p-value approach with auxiliary randomization. For experiments (iii), both LL and NN methods give large p-values, and the NN method also has small MCE. Thus, there is no strong evidence against the covariate shift assumption. In experiment (iv), the MCE indicate that the NN method is likely more accurate, and hence providing more trustworthy large p-value. In experiment (v), both methods agree to reject the null hypothesis, which is the correct decision by the construction of training and testing samples.

7 Discussion

A valid type I error control of the proposed test only depends on the accuracy of marginal density ratio estimation. Our method would be particularly useful in the semi-supervised scenario, where unlabeled testing sample points X_{2j} are easy to obtain.

The covariate shift hypothesis can be viewed as a special case of conditional independence given a binary random variable. For a triplet of random variables (X, Y, A), the conditional distributions $(Y \mid X, A = 0)$ and $(Y \mid X, A = 1)$ are the same if and only if Y and A are conditionally independent given X. One can consider the alternative hypothesis, such as that the treatment is expected to increase the mean response. One can extend the method to new nonparametric conditional independence tests.

References

Tibshirani, R. J., Barber, R. F., Candès, E., & Ramdas, A. (2019). *Conformal prediction under covariate shift.* In Advances in Neural Information Processing Systems, 2019.