Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum

Xiaoyu Hu & Jing Lei

To cite this article: Xiaoyu Hu & Jing Lei (2024) A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum, Journal of the American Statistical Association, 119:546, 1136-1154, DOI: 10.1080/01621459.2023.2177165

To link to this article: https://doi.org/10.1080/01621459.2023.2177165

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 546, 1136–1154: Theory and Methods
https://doi.org/10.1080/01621459.2023.2177165

A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum

Xiaoyu Hu[a] and Jing Lei[b]

[a]School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing, China; [b]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

We consider the problem of testing the equality of conditional distributions of a response variable given a vector of covariates between two populations. Such a hypothesis testing problem can be motivated from various machine learning and statistical inference scenarios, including transfer learning and causal predictive inference. We develop a nonparametric test procedure inspired from the conformal prediction framework. The construction of our test statistic combines recent developments in conformal prediction with a novel choice of conformity score, resulting in a weighted rank-sum test statistic that is valid and powerful under general settings. To our knowledge, this is the first successful attempt of using conformal prediction for testing statistical hypotheses beyond exchangeability. Our method is suitable for modern machine learning scenarios where the data has high dimensionality and large sample sizes, and can be effectively combined with existing classification algorithms to find good conformity score functions. The performance of the proposed method is demonstrated in various numerical examples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2021
Accepted January 2023

KEYWORDS
Conformal prediction; Covariate shift; Distribution-free; Model misspecification; Rank-sum test

1. Introduction

Suppose we have two independent random samples

$$\{(X_{1i}, Y_{1i})\}_{i=1}^{n_1} \overset{\text{iid}}{\sim} P_1 \quad \text{and} \quad \{(X_{2j}, Y_{2j})\}_{j=1}^{n_2} \overset{\text{iid}}{\sim} P_2\,,$$

where P_1, P_2 are distributions on a product space $\mathcal{X} \times \mathcal{Y}$. Here we consider a regression or classification setting where X is the free variable (covariate), and Y is the response variable. We allow the spaces $\mathcal{X}$, $\mathcal{Y}$ to be general, and do not assume specific forms such as smoothness or parametric forms of P_1, P_2. For example, $\mathcal{X}$ and $\mathcal{Y}$ can be multi-dimensional Euclidean spaces, manifolds, or discrete sets.

For $j = 1, 2$, let $P_j(\cdot|x)$ be the conditional distribution of Y given $X = x$ under P_j, and $P_{j,X}(\cdot)$ be the corresponding marginal distribution of X. We are interested in testing whether these two conditional distributions are the same.

$$H_0 : P_1(\cdot|x) = P_2(\cdot|x) \text{ for all } x \in \mathcal{X}, \text{ versus } H_1 : \text{otherwise}. \tag{1}$$

We illustrate P_1 and P_2 in the two motivating examples below.

Example 1: Covariate shift in transfer learning. In traditional machine learning, a central task is to build a predictor $\widehat{f} : \mathcal{X} \mapsto \mathcal{Y}$ from a training sample $\{(X_{1i}, Y_{1i})\}_{i=1}^{n_1} \overset{\text{iid}}{\sim} P_1$, and use it to predict the unseen response Y_2 given a future observation of the covariate X_2, where $(X_2, Y_2) \sim P_2$. A common assumption is that the training data distribution P_1 and testing data distribution P_2 are the same. In many modern applications, it is often the case that the testing data may come from a different distribution, and classical methods developed for iid data need to be modified to account for such distribution difference. Subfields known as domain adaptation and transfer learning have emerged to deal with scenarios in which the training data and testing data may come from different but related distributions (Pan and Yang 2009; Csurka 2017; Kouw and Loog 2018). To avoid arbitrary changes in the distribution, one situation of particular practical and theoretical interest assumes that the conditional distribution of the response given the covariate remains the same between the training and testing data while the covariates may follow different marginal distributions. This is the *covariate shift* assumption. It is mathematically equivalent to the null hypothesis in (1), and enables us to obtain improved predictive performance by weighting the training data according to the marginal density ratio. This approach has been widely studied in the machine learning literature. For examples, see Shimodaira (2000), Sugiyama, Krauledat, and Müller (2007), Sugiyama et al. (2008), Bickel, Brückner, and Scheffer (2009), Gretton et al. (2009), and Sugiyama and Kawanabe (2012). The two-sample conditional distribution testing problem is a formal way to verify the covariate shift assumption, assuming an independent sample $\{(X_{2j}, Y_{2j}) : 1 \le j \le n_2\}$ from P_2 is also available. If we do not reject the null hypothesis, the methods and theory based on the covariate shift assumption may be plausible. Otherwise, those methods should be used with caution.

CONTACT Jing Lei *jinglei@andrew.cmu.edu* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Example 2: Causal predictive inference. In the literature of causal inference, a causal predictive model is one that will work equally well under different experimental or observational environments (Peters, Bühlmann, and Meinshausen 2016; Bühlmann 2020; Li et al. 2021). Formally, a covariate vector X is called causal for the prediction of Y if $Y = h(X, \epsilon)$, for a fixed function h and an independent random variable ϵ with a fixed distribution. This definition highlights the idea that if X contains all the causal factors of Y, then the way X affects Y does not depend on any other variables (potential confounders). Therefore, the conditional distribution of Y given X will remain the same under different experimental conditions, regardless whether these conditions are fully controlled or simply observed. In this context, the two distributions P_1 and P_2 represent the joint distribution of (X, Y) under two different experimental conditions. If we reject H_0 in (1), then X is unlikely to be causal for the prediction of Y.

Most existing methods for testing conditional distributions follow one of two directions. The first is to test the equality of conditional moments, including semiparametric methods (Hardle and Marron 1990), nonparametric methods (Hall and Hart 1990; Kulasekera 1995; Kulasekera and Wang 1997; Neumeyer and Dette 2003), and second order moments (Pardo-Fernández, Jiménez-Gamero, and El Ghouch 2015). However, in many applications such as risk management and insurance, it is not enough to just consider mean and variance terms, and it is necessary to consider the whole conditional distribution of the response given the covariates. Another direction is to extend methods for unconditional distribution tests to conditional distribution tests. Andrews (1997) extended the Kolmogorov-Smirnov test to the conditional distribution case. Zheng (2000) proposed a test statistic based on the first-order linear expansion of Kullback-Leibler divergence. Fan, Li, and Min (2006) proposed a bootstrap test. Bai (2003) and Corradi and Swanson (2006) studied the problem of testing conditional distributions in a dynamic model. Most of the aforementioned methods rely on some assumptions that are hard to verify in a data-driven manner, such as smoothness of density and regression functions, and/or correctly specified parametric models. In addition, existing nonparametric methods usually involve nonparametric density estimation as an intermediate step, making them less feasible when the dimensionality is moderately high.

Our new method for two-sample conditional distribution test has three remarkable features. First, our test statistic is inspired from conformal prediction (Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013; Lei et al. 2018), a framework of converting point estimators to prediction sets by exploiting the symmetry of data. The Type I error control is guaranteed by a weighted exchangeability that is tailored to the null hypothesis, assuming an accurate estimate of the marginal density ratio of the covariates is available. Second, our method does not require estimating the density functions. Instead, it uses a classification algorithm to estimate the density ratio, and can incorporate almost any existing classification algorithms ranging from classical parametric estimators to modern black-box neural nets. In practice, the validity of the data-driven p-value depends on the accuracy of the classifier, which can be empirically validated. This makes our method particularly useful in modern machine learning scenarios with high dimensionality and large sample sizes. 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. To our knowledge, this is the first successful extension of conformal prediction to statistical hypothesis testing beyond exchangeability with provable size and power guarantees for a data-driven procedure. These theoretical results are supported by our simulation and real data examples.

Related work in conformal prediction. We provide a general review of conformal prediction in Section 2.2. Roughly speaking, conformal prediction uses a conformity score to determine a sample point's agreement, or conformity, with the current dataset and fitting procedure, and the resulting prediction set is the subset of the sample space with high conformity scores. Conformal prediction in regression has been studied by Lei and Wasserman (2014) in a nonparametric setting and Lei et al. (2018) in the high dimensional setting, where the conformity scores are chosen to be either the conditional density of Y given X, or the absolute fitted residual. However, these conformity scores do not guarantee power against general alternatives. A main methodological contribution of this article is to show that using the conditional likelihood ratio as the conformity score can provide universal power guarantee against any alternatives. Such a choice of conformity score is partially inspired by a recent work of conformal prediction in classification by Guan and Tibshirani (2019). In the context of transfer learning, under the covariate shift assumption the data points are often exchangeable within the training set or the testing set alone, but not exchangeable when the training and testing datasets are merged together. This nonexchangeability issue can be treated using the weighted conformal prediction method developed in Tibshirani et al. (2019), who construct a valid p-value for a single observation in the test sample assuming the marginal density ratio is known. Our method combines these ideas, with further theoretical development to allow the marginal density ratios and conformity scores to carry estimation errors. Moreover, these existing methods only consider prediction, and our method transforms conformal prediction from a prediction tool to a hypothesis testing tool.

There is a line of work on applying conformal inference in other contexts, such as testing the global null for streaming data (Vovk, Nouretdinov, and Gammerman 2003; Fedorova et al. 2012; Vovk 2019, 2020; Vovk et al. 2021) and outlier detection (Bates et al. 2021). Although both the work of Bates et al. (2021) and ours involve the conformal p-values and sample splitting, they are different in various aspects. Bates et al. (2021) studied the nonparametric outlier detection problem, which is different from our goal of two-sample conditional distribution testing. The score functions to construct the conformal p-values are different. Specifically, they used the one-class classification score, while we used the conditional density ratio. Moreover, the ways to exploit the conformal p-values are distinct. They constructed the p-values for the multiple-testing procedure, while we combined the p-values to eventually perform a one sided mean test. None of the three features of our method (testing conditional distributions, conformity score function with universal power,

and asymptotic error bounds for data-driven procedures) is considered in these papers.

2. Background

2.1. Problem Formulation

If an $x \in \mathcal{X}$ is not in the support of $P_{1,X}$ or $P_{2,X}$, then the point x should not matter in testing the conditional distribution, since the conditional distribution given this value of x can be arbitrarily defined. Therefore, to facilitate discussion we assume that $P_{1,X}$ and $P_{2,X}$ are equivalent to each other in the following sense,

$$P_{1,X} \ll P_{2,X}, \quad \text{and} \quad P_{2,X} \ll P_{1,X},$$

where "$\ll$" stands for absolute continuity. We further assume, without loss of generality, that $P_{1,X}, P_{2,X}$ have density functions $f_{1,X}, f_{2,X}$, respectively, under a common base measure. Now we can formally state the null and alternative hypotheses as follows.

$$H_0 : P_{1,X}\{P_1(\cdot|X) = P_2(\cdot|X)\} = 1, \quad \text{versus} \quad H_1 : P_{1,X}\{P_1(\cdot|X) = P_2(\cdot|X)\} < 1. \tag{2}$$

Due to the assumed equivalence between $P_{1,X}$ and $P_{2,X}$, the hypotheses in (2) can be equivalently stated by replacing $P_{1,X}$ with $P_{2,X}$. For a similar consideration of avoiding triviality and the ease of discussion, we also assume that $P_1(\cdot|x)$ and $P_2(\cdot|x)$ are equivalent, with density functions $f_1(y|x)$ and $f_2(y|x)$ under a common base measure.

2.2. Conformal Prediction

Here we briefly introduce conformal prediction in a regression setting. For conformal prediction in other contexts, such as unsupervised learning, see (Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013; Lei, Rinaldo, and Wasserman 2015).

Given iid data $\{(X_i, Y_i)\}_{i=1}^m$, conformal prediction converts a point estimator of the regression function $\widehat{\theta} : \mathcal{X} \mapsto \mathcal{Y}$ to a prediction set $\widehat{C} \in \mathcal{X} \times \mathcal{Y}$, with guaranteed finite-sample expected prediction coverage:

$$P\left\{Y_{m+1} \in \widehat{C}(X_{m+1})\right\} \geq 1 - \alpha, \tag{3}$$

where $\widehat{C}(x) = \{y \in \mathcal{Y} : (x, y) \in \widehat{C}\}$, and the probability is taken over the $(m+1)$-tuple of iid data $\{(X_i, Y_i) : 1 \leq i \leq m+1\}$.

Let $\mathcal{D}$ denote the sample $\{(X_i, Y_i) : 1 \leq i \leq m\}$, and $\mathcal{D}_{-i}$ the sample obtained by removing (X_i, Y_i) from $\mathcal{D}$. A conformal prediction set $\widehat{C}$ is constructed using a conformity score function $v : (\mathcal{X} \times \mathcal{Y})^{m+1} \mapsto \mathbb{R}$ that is symmetric in its first m inputs. For a given dataset $\mathcal{D}$, a new X_{m+1} for which a prediction of the corresponding Y_{m+1} is wanted, and a $y \in \mathcal{Y}$, let $\mathcal{D}(y)$ be the augmented dataset with the $(m+1)$th data point being (X_{m+1}, y), and $\mathcal{D}_{-i}(y)$ the corresponding m-tuple dataset with the ith sample removed, where the $(m+1)$th sample is (X_{m+1}, y). Let

$$V_i(y) = v(\mathcal{D}_{-i}(y), (X_i, Y_i)), \quad i = 1, \ldots, m,$$
$$V_{m+1}(y) = v(\mathcal{D}, (X_{m+1}, y)) \tag{4}$$

be conformity scores for each sample point in the augmented data $\mathcal{D}(y)$. The conformal prediction set using the conformity score function $v(\cdot)$ is

$$\widehat{C}(X_{m+1}) = \left\{y \in \mathcal{Y} : \sum_{i=1}^{m+1} \mathbb{1}\left[V_i(y) \leq V_{m+1}(y)\right] \geq \lfloor (m+1)\alpha \rfloor\right\}. \tag{5}$$

The finite sample coverage (3) can be easily derived from the iid assumption, the symmetry of $v(\cdot)$, and the construction of $\widehat{C}$ in (5). To see this, if we replace y by Y_{m+1}, then the iid assumption and symmetry of $v(\cdot)$ implies exchangeability of $(V_i(Y_{m+1}) : 1 \leq i \leq m+1)$. Thus, the rank of V_{m+1} being lower than $\lfloor (m+1)\alpha \rfloor$ has probability no more than α.

Although the finite sample coverage guarantee only requires $v(\cdot)$ to satisfy a symmetry condition, its choice will have a crucial impact on the quality of the resulting prediction set. A good choice of $v(\cdot)$ needs to reflect the structure of the underlying distribution of (X, Y) and be able to tell whether a sample point is likely drawn from this distribution. Such a function $v(\cdot)$ is often constructed from a point estimate $\widehat{\theta}$ of the regression function. For example, in nonparametric regression, one can choose $v(\mathcal{D}, (x, y)) = \widehat{f}(y|x)$, where $\widehat{f}(\cdot|\cdot)$ is an estimated conditional density function of y given x using the sample $\mathcal{D} \cup \{(x, y)\}$ (Lei and Wasserman 2014). In high dimensional regression, one can use $v(\mathcal{D}, (x, y)) = -|y - \widehat{\theta}(x)|$ where $\widehat{\theta}$ is an estimated regression function using $\mathcal{D} \cup \{(x, y)\}$. More recently, some conformal prediction methods adaptive to heteroscedasticity based on quantile regression have been proposed (Romano, Patterson, and Candes 2019; Kivaranovic, Johnson, and Leeb 2020; Sesia and Romano 2021; Chernozhukov, Wüthrich, and Zhu 2021). In this work, we develop a new conformity score based on conditional density ratios, which is particularly suited for the two-sample conditional testing problem.

The definition of $\widehat{C}$ in (5) is only conceptual and not practical if $\mathcal{Y}$ is infinite, as it requires to evaluate $V_i(y)$ for all y and all $1 \leq i \leq m+1$. For practical implementation of conformal prediction, we refer to Lei et al. (2018) and Barber et al. (2019). However, in our hypothesis testing problem, we do not need to actually construct a prediction set. Instead, we only need to compute the corresponding p-values for a subset of sample points, and evaluate their deviance from the null distribution. The details are given in the next section as we develop our testing procedure.

3. A Conformal Test of Two-Sample Conditional Distributions

3.1. The Conformal p-value

Now we put the conformal prediction method described in Section 2 under the context of our two-sample testing problem. Consider a subset of the data $\mathcal{D}_{(1)} = \{(X_{1i}, Y_{1i}) : 1 \leq i \leq n_{11}\}$ iid from P_1, and just a single pair $(X_{21}, Y_{21}) \sim P_2$. Here $n_{11} < n_1$ is a subsample size, whose value will be specified later. With the correspondence $m = n_{11}$, $(X_i, Y_i) = (X_{1i}, Y_{1i})$ for $1 \leq i \leq n_{11}$ and $(X_{n_{11}+1}, Y_{n_{11}+1}) = (X_{21}, Y_{21})$, the conformal prediction

procedure described in the previous section implies that, under the simplified scenario that $P_{1,X} = P_{2,X}$, the ranking statistic

$$\widetilde{U} = \frac{1}{n_{11}+1} \sum_{i=1}^{n_{11}+1} \mathbb{1}(V_i(Y_{21}) \le V_{n_{11}+1}(Y_{21})) \tag{6}$$

has an approximately $U(0,1)$ distribution under H_0.

With a random tie-break we can make $\widetilde{U}$ have the exact $U(0,1)$ distribution. Let $R_- = 1 + \sum_{i=1}^{n_{11}+1} \mathbb{1}(V_i(Y_{n_{11}+1}) < V_{n_{11}+1}(Y_{n_{11}+1}))$ and $R_+ = \sum_{i=1}^{n_{11}+1} \mathbb{1}(V_i(Y_{n_{11}+1}) \le V_{n_{11}+1}(Y_{n_{11}+1}))$. Let R be uniformly and independently sampled from the integers in $[R_-, R_+]$. Now we can construct a uniform random variable

$$U = \frac{R-1+\zeta}{n_{11}+1} \tag{7}$$

where $\zeta \sim U(0,1)$ is independent of everything else. By exchangeability R has a uniform distribution on $\{1,\dots,n_{11}+1\}$, and hence U has a uniform distribution on $[0,1]$. This U can be viewed as a continuous version of $\widetilde{U}$ in (6), and can serve as a p-value for our testing problem. Thus, we call the statistic U a *conformal p-value*.

In order to develop this idea into a useful test procedure, we need to resolve the following three issues.

1. How to choose a conformity score function v?
2. How to allow for $P_{1,X} \neq P_{2,X}$?
3. How to make use of multiple data points from P_2 to have increased power under H_1?

These three issues are addressed in the next three sections, respectively.

3.2. A Choice of v that Separates H_0 and H_1

For now we still make the assumption $P_{1,X} = P_{2,X}$. A good choice of $v(\cdot)$ will be such that the conformal p-value U constructed in (7) has a nonuniform distribution under the alternative hypothesis H_1. Common existing choices such as conditional density and absolute residual do not satisfy this. Our choice of $v(\cdot)$ is the conditional Radon-Nikodym derivative between P_1 and P_2:

$$v(x,y) = \frac{dP_1(y|x)}{dP_2(y|x)} = \frac{f_1(y|x)}{f_2(y|x)}.$$

This v function is different from the conformity score functions introduced in (4), as it only involves a single data pair (x,y). This is not a real problem as we can let v be independent of the first m arguments and treat the input (x,y) as the last argument.

Remark 1. The function v involves the unknown density ratio $f_1(y|x)/f_2(y|x)$. Our method will need to use an empirical version of v:

$$\widehat{v}(x,y) = \frac{\widehat{f_1(\cdot|\cdot)}}{f_2(\cdot|\cdot)}(x,y),$$

where $\widehat{\frac{f_1(\cdot|\cdot)}{f_2(\cdot|\cdot)}}$ is an estimate of the conditional density ratio, independent of $\{(X_{1i}, Y_{1i}) : 1 \le i \le n_{11}\}$ and (X_{21}, Y_{21}). A remarkable advantage of our choice of v and $\widehat{v}$ is that the density ratio $f_1(\cdot|\cdot)/f_2(\cdot|\cdot)$ can be conveniently estimated using classification algorithms, which is both theoretically and practically much easier than estimating the density functions themselves. There is a rich literature on classification and density ratio estimation, with many powerful algorithms even in high dimensional settings. Further discussion of estimating the conditional density ratio is provided in Section 3.5 when we summarize our algorithm.

The ability of $v(x,y)$ to separate H_0 and H_1 is established by the following lemma.

Lemma 1 (Separation of H_0 and H_1 by v under equal X-marginal). If $P_{1,X} = P_{2,X}$ and $v(x,y) = \frac{f_1(y|x)}{f_2(y|x)}$, then under H_1, $\mathbb{E}U = \frac{1}{2} - \frac{1}{4}\mathbb{E}|v(X_2,Y_2) - v(X_2',Y_2')| < \frac{1}{2}$ for all values of $n_{11} \ge 1$, where (X_2,Y_2), (X_2',Y_2') are iid realizations from P_2.

Lemma 1 can be viewed as a special case of Lemma 2(c), for which a complete proof is provided in Appendix D. Under H_0, $v(x,y) \equiv 1$. Under H_1, the conditional likelihood ratio $v(X_2,Y_2)$ cannot be degenerate, so $\mathbb{E}|v(X_2,Y_2) - v(X_2',Y_2')|$ is strictly positive and hence we have $\mathbb{E}U < 1/2$ under H_1. This also suggests a simple one sided rejection rule for our test.

Remark 2. The choice of v is motivated from an information theoretical perspective. It can be directly verified that $\mathbb{E}_{P_1} v(X,Y) - \mathbb{E}_{P_2} v(X,Y) = \mathbb{E}_{P_1}\left[D_{\chi^2}(f_1(\cdot|X), f_2(\cdot|X))\right]$, where $D_{\chi^2}(f_1,f_2) = \int f_1^2/f_2 - 1$ is the Neyman's χ^2 divergence between two densities f_1, f_2. As a result, $\mathbb{E}_{P_1} v(X,Y) - \mathbb{E}_{P_2} v(X,Y) \ge 0$ with equality holds if and only if H_0 is true. This suggests that under the alternative, $v(X,Y)$ tends to take larger values under P_1 than under P_2. A more involved argument is needed in order to carry over this intuition rigorously to analyze the rank of $V_{n_{11}+1}(Y_{21})$ in (6) and the continuous version in (7). It is made clear in (11), Section 4.2, that the conformal p-value based on $v(x,y) = f_1(y|x)/f_2(y|x)$ is closely related to the expected conditional total variation distance between P_1 and P_2.

3.3. 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. Recall the notation $(X_i,Y_i)_{i=1}^{n_{11}+1}$, with $(X_i,Y_i) \sim P_1$ for $1 \le i \le n_{11}$ and $(X_{n_{11}+1}, Y_{n_{11}+1}) \sim P_2$. Now the $(n_{11}+1)$-tuple used to construct U in (7) are no longer exchangeable under the null hypothesis and the distribution of U will in general not be uniform. Here we use the "weighted conformal prediction" idea developed in Tibshirani et al. (2019) to obtain a modified version of U with valid uniform sampling distribution under H_0. The key idea is to condition on a randomly permuted data sequence.

In the subsequent discussion, we will focus on the conformity score functions v that only depends on the last argument, and write $V_i = v(X_i,Y_i)$.

We begin by imagining that the data $\mathbf{Z} = (X_i,Y_i)_{i=1}^{n_{11}+1}$ are stored in two parts: a randomly permuted sequence $\widetilde{\mathbf{Z}} = (\widetilde{X}_i,\widetilde{Y}_i)_{i=1}^{n_{11}+1}$, and the permutation $\sigma : [n_{11}+1] \mapsto [n_{11}+1]$ with the correspondence $(X_i,Y_i) \leftrightarrow (\widetilde{X}_{\sigma(i)}, \widetilde{Y}_{\sigma(i)})$. By construction, the vector $(V_i : 1 \le i \le n_{11}+1)$ is a deterministic function

of $(\widetilde{\mathbf{Z}}, \sigma)$. Given $\widetilde{\mathbf{Z}}$, $V_{n_{11}+1}$ may take $n_{11}+1$ possible values, and $V_{n_{11}+1} = v(\widetilde{X}_i, \widetilde{Y}_i)$ if $\sigma(n_{11}+1) = i$.

Now we are ready to derive the conditional distribution of $V_{n_{11}+1}$ given $\widetilde{\mathbf{Z}}$, 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

$$(V_{n_{11}+1}|\widetilde{\mathbf{Z}}) \sim \sum_{i=1}^{n_{11}+1} p_i(\mathbf{Z})\delta_{V_i}$$

with

$$p_i(\mathbf{Z}) = \frac{\frac{f_{2,X}(X_i)}{f_{1,X}(X_i)}}{\sum_{l=1}^{n_{11}+1}\frac{f_{2,X}(X_l)}{f_{1,X}(X_l)}}, \quad i = 1, \ldots, n_{11}+1,$$

where $f_{k,X}(\cdot)$ denotes the marginal density function of X under P_k $(k = 1, 2)$, and δ_v denotes the point mass at v.

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

$$U = \sum_{i=1}^{n_{11}+1} p_i(\mathbf{Z})\mathbb{1}(V_i < V_{n_{11}+1}) + \zeta \sum_{i=1}^{n_{11}+1} p_i(\mathbf{Z})\mathbb{1}(V_i = V_{n_{11}+1}) \tag{8}$$

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

(c) Under H_1, if $v(x,y) = f_1(y|x)/f_2(y|x)$, there exist $\delta > 0$ and $m_0 > 0$, depending only on P_1 and P_2, such that $\mathbb{E}U \le 1/2 - \delta$ when $n_{11} \ge m_0$.

The definitions of U in (8) and (7) are compatible, as the construction with random tie-breaking in (7) can be viewed as a special case of (8) with $p_i(\mathbf{Z}) = (n_{11}+1)^{-1}$.

Part (a) of Lemma 2 is due to Tibshirani et al. (2019), who first considered weighted conformal prediction under the covariate shift assumption. Part (b) is a simple consequence of part (a), which can be viewed as a discrete version of the CDF transform. The most non-obvious part of the proof is that of part (c), which exploits the form of $v(x,y) = f_1(y|x)/f_2(y|x)$. The detailed proofs are given in Appendix D.

The conformal p-value U will exhibit a constant difference from the null distribution once the ranking sample size n_{11} exceeds a finite threshold. Such a separation between the null and alternative hypotheses can be amplified using multiple such weighted conformal p-values, as we discuss in the next section.

3.4. Incorporating Multiple Testing Sample Points for Better Power

So far we have only focused on obtaining a single conformal p-value from a single sample point in P_2. Such a single p-value often has limited power in distinguishing H_1 from H_0. To have a consistent test that rejects H_0 under the alternative hypothesis with probability tending to 1 as the sample size increases to ∞, we must consider multiple testing sample points: $\{(X_{2j}, Y_{2j}) : 1 \le j \le n_{21}\}$, where n_{21} is a subsample size whose relationship with the original sample size n_2 will be discussed later.

Now assume that we have obtained estimates $\widehat{g}$ and $\widehat{v}$ for the marginal and conditional density ratios $g(x) \equiv f_{2,X}(x)/f_{1,X}(x)$ and $v(x,y) = f_1(y|x)/f_2(y|x)$, respectively. Given $(X_{1i}, Y_{1i})_{i=1}^{n_{11}}$ from P_1 and $(X_{2j}, Y_{2j})_{j=1}^{n_{21}}$ from P_2, we can repeat the procedure used to obtain U in (7) for each sample point (X_{2j}, Y_{2j}) for $1 \le j \le n_{21}$, resulting in $(\widehat{U}_j : 1 \le j \le n_{21})$. If the function estimates $\widehat{g}$ and $\widehat{v}$ are accurate enough, then approximately each $\mathbb{E}\widehat{U}_j = 1/2$ under H_0 and $\mathbb{E}\widehat{U}_j < 1/2$ under H_1. However, these $\widehat{U}_j$'s are dependent as they use the same set of ranking sample $(X_{1i}, Y_{1i})_{i=1}^{n_{11}}$ from the first population. To obtain a valid p-value for one-sided mean test over the $\widehat{U}_j$'s, we must take their dependence into account. To this end, we redefine $\widehat{U}_j$ as

$$\widehat{U}_j = \frac{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}}\widehat{g}(X_{1i})\widehat{D}_{ij}}{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}}\widehat{g}(X_{1i})}, \quad j = 1, \ldots, n_{21},$$

where $\widehat{D}_{ij} = \{\mathbb{1}(\widehat{V}_{1i} < \widehat{V}_{2j}) + \zeta_j\mathbb{1}(\widehat{V}_{1i} = \widehat{V}_{2j})\}$, $\widehat{V}_{1i} = \widehat{v}(X_{1i}, Y_{1i})$ and $\widehat{V}_{2j} = \widehat{v}(X_{2j}, Y_{2j})$. Such a $\widehat{U}_j$ can be viewed as an approximate plug-in version of the conformal p-value in (8), with estimated versions of g and v, and omitting the vanishing terms $\widehat{g}(X_{2j})/n_{11}$.

The key observation is that despite the dependence due to a common ranking sample, the average of these p-values $n_{21}^{-1}\sum_{j=1}^{n_{21}}\widehat{U}_j$ is a two-sample U-statistic conditioning on the estimated density ratios $\widehat{g}, \widehat{v}$, whose asymptotic distribution can be readily estimated using plug-in estimators.

Formally, we use statistic

$$\widehat{T} = \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_{j=1}^{n_{21}}\widehat{U}_j}{\widehat{\sigma}/\sqrt{n_{11}}} \tag{9}$$

where $\widehat{\sigma}$ is the estimated asymptotic standard deviation of $\sqrt{n_{11}}n_{21}^{-1}\sum_{j=1}^{n_{21}}\widehat{U}_j$.

Let $\widehat{F}_n$ be the empirical CDF of $\{\widehat{V}_{2j} : 1 \le j \le n_{21}\}$, and $\widehat{F}_{n,1/2} = (\widehat{F}_n + \widehat{F}_{n,-})/2$ where F_- is the left limit of a function F. The asymptotic variance used in $\widehat{T}$ can be estimated as follows.

$$\widehat{\sigma}^2 = \widehat{\sigma}_1^2 + \frac{n_{11}}{12n_{21}} + \frac{1}{4}\widehat{\sigma}_2^2 - \widehat{\rho}_{12}, \tag{10}$$

where $\widehat{\sigma}_1^2$ is the empirical variance of $\{\widehat{g}(X_{1i})[1 - \widehat{F}_{n,1/2}(\widehat{V}_{1i})] : 1 \le i \le n_{11}\}$, $\widehat{\sigma}_2^2$ is the empirical variance of $\{\widehat{g}(X_{1i}) : 1 \le i \le n_{11}\}$, and $\widehat{\rho}_{12}$ is the empirical covariance between $\{\widehat{g}(X_{1i})[1 - \widehat{F}_{n,1/2}(\widehat{V}_{1i})] : 1 \le i \le n_{11}\}$ and $\{\widehat{g}(X_{1i}) : 1 \le i \le n_{11}\}$. The derivation of the asymptotic variance is provided in Theorem 1 and Lemma 7.

Remark 3. The asymptotic variance of $\widehat{T}$ can be alternatively estimated using $\widehat{g}(X_{2j})$ and $[1 - \widehat{F}_{n,1/2}(\widehat{V}_{2j})]$ using an importance sampling technique. This provides a larger sample size for asymptotic variance estimation. Practically, we found using the harmonic mean of the original estimate and the importance sampling estimate to have good performance in simulations. The details of the importance sampling estimate are given in Appendix A.

3.5. The Conformal Conditional Distribution Test Algorithm

Given the ideas and methods presented in the previous sections, we can now describe the full testing procedure in Algorithm 1. The algorithm assumes availability of two classification subroutines: (i) a marginal classification algorithm $\mathcal{A}_1$ that takes input

two labeled samples $\{X_{1i} : n_{11}+1 \le i \le n_1\}$, $\{X_{2j} : n_{21}+1 \le j \le n_2\}$ with sample sizes $n_{12} = n_1 - n_{11}$, $n_{22} = n_2 - n_{21}$, respectively, and outputs an estimate of the marginal density ratio $g = f_{2,X}/f_{1,X}$; and (ii) a joint classification algorithm $\mathcal{A}_2$ that takes input two labeled samples $\{(X_{1i}, Y_{1i}) : n_{11}+1 \le i \le n_1\}$ and $\{(X_{2j}, Y_{2j}) : n_{21}+1 \le j \le n_2\}$, and outputs an estimate of the conditional density ratio $v = f_1(y|x)/f_2(y|x)$. Our numerical experiments use a equal split ratio: $(n_{11}, n_{21}) = (n_1/2, n_2/2)$, which yields reasonable performance in all the scenarios considered.

Given $\mathcal{A}_1$, $\mathcal{A}_2$, the testing procedure first splits the sample, applying the density ratio estimation subroutines $\mathcal{A}_1, \mathcal{A}_2$ on one part to obtain approximate versions of the density ratios. Then the other part is used to obtain the final test statistic $\widehat{T}$.

Algorithm 1 Two-sample test of conditional distribution

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

For $k = 1, 2$, randomly split $\{1, \ldots, n_k\}$ into subsets $\mathcal{I}_{k1} = \{1, \ldots, n_{k1}\}$, $\mathcal{I}_{k2} = \{n_{k1}+1, \ldots, n_k\}$

$\widehat{g}(\cdot) = \mathcal{A}_1[\{X_{1i}, i \in \mathcal{I}_{12}, X_{2j}, j \in \mathcal{I}_{22}\}]$

$\widehat{v}(\cdot,\cdot) = \mathcal{A}_2[\{(X_{1i}, Y_{1i}), i \in \mathcal{I}_{12}, (X_{2j}, Y_{2j}), j \in \mathcal{I}_{22}\}]$

for $j \in \mathcal{I}_{21}$ **do**

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

$\widehat{D}_{ij} = \mathbb{1}(\widehat{V}_{1i} < \widehat{V}_{2j}) + \zeta_j \mathbb{1}(\widehat{V}_{1i} = \widehat{V}_{2j})$, where $\widehat{V}_{ki} = \widehat{v}(X_{ki}, Y_{ki})$

$\widehat{U}_j = \sum_{i=1}^{n_{11}} \widehat{g}(X_{1i}) \widehat{D}_{ij} / \sum_{i=1}^{n_{11}} \widehat{g}(X_{1i})$

end for

$\widehat{T} = (1/2 - n_{21}^{-1} \sum_{j=1}^{n_{21}} \widehat{U}_j)/(\widehat{\sigma}/\sqrt{n_{11}})$, where $\widehat{\sigma}^2$ is given in (10)

Reject H_0 if $\widehat{T} \ge \Phi^{-1}(1-\alpha)$ (Φ is the CDF of $N(0,1)$, α is the nominal Type I error level)

Simplification when the marginals are equal. Sometimes it is plausible to assume that the marginal distributions $f_{1,X}$, $f_{2,X}$ are equal. This can happen, for example, when the sampling schemes and environments of X are known to be the same, or they come from the same experimental design. In this scenario, the algorithm becomes much simpler, as we know that $f_{1,X}/f_{2,X} \equiv 1$. As a result, the algorithm does not need to use the marginal density ratio subroutine $\mathcal{A}_1$ and $g \equiv 1$.

Choice of classification algorithms. As mentioned in Remark 1, our method does not require estimating the densities $f_{k,X}(\cdot)$ or conditional densities $f_k(\cdot|\cdot)$ for $k = 1, 2$. Instead, it only requires estimating the marginal density ratio $f_{1,X}(x)/f_{2,X}(x)$, and the conditional density ratio $f_1(y|x)/f_2(y|x)$ only needs to be estimated up to a monotone transform, since only the ranking information is needed in the test statistic. Estimating density ratios is often easier than estimating the density functions themselves, and has been well studied in the statistics and machine learning literature, including moment matching approach (Gretton et al. 2009), the ratio matching approach (Sugiyama et al. 2008; Kanamori, Hido, and Sugiyama 2009; Tsuboi et al. 2009), and probabilistic classification approach (Qin 1998; Cheng and Chu 2004; Bickel, Brückner, and Scheffer 2007).

Our algorithm can be implemented with any available density ratio estimators. Here we provide some further detail about the probabilistic classification estimator due to its simplicity. In the case of $f_{1,X} = f_{2,X}$, we only need to consider a single classification problem over the joint distribution (X, Y), where class "1" represents the subsample $\{(X_{1i}, Y_{1i}) : i \in \mathcal{I}_{12}\}$, and class "2" represents the subsample $\{(X_{2j}, Y_{2j}) : j \in \mathcal{I}_{22}\}$. Let $\eta(x,y)$ be the true conditional probability $P(1|x,y)$, then $\eta(x,y)/(1-\eta(x,y)) \propto f_1(x,y)/f_2(x,y)$, which also equals $f_1(y|x)/f_2(y|x)$ since $f_{1,X} = f_{2,X}$. When $f_{1,X} \ne f_{2,X}$, we can consider an additional classification problem using only X. Let $\eta(x) = P(1|x)$, then $f_{2,X}(x)/f_{1,X}(x) = n_{12}(1-\eta(x))/(n_{22}\eta(x))$. With probabilistic classifiers providing $\widehat{\eta}(x,y)$ and $\widehat{\eta}(x)$, the corresponding joint and marginal density ratios can be estimated by plugging in $\widehat{\eta}(x,y)$ and $\widehat{\eta}(x)$. The conditional density ratio can be obtained by taking a further ratio between the joint and marginal density ratios. Many commonly used classification methods offer a probability output, including the classical linear and quadratic discriminant analysis, logistic regression, popular machine learning algorithms such as random forest and support vector machines (Sollich 2000), and modern deep neural nets.

4. Asymptotic Properties

In this section, we investigate the theoretical properties of the testing procedure described in Algorithm 1 under (i) a standard fixed population asymptotic framework, and (ii) a local alternative perspective. Since the test statistic is constructed from the ranking subsamples, we assume that the two ranking subsample sizes are proportional: n_{11}/n_{21} stays bounded and bounded away from 0 as $n_{11} \to \infty$. The asymptotic behavior of our test will depend on the estimated functions $\widehat{g}$ and $\widehat{v}$, which depends on the fitting sample sizes (n_{12}, n_{22}). It is natural to have (n_{12}, n_{22}) increasing with the ranking sample size (n_{11}, n_{21}). We quantify the required accuracy of the density ratio estimates $\widehat{g}$, $\widehat{v}$ in Assumption 2. Designing a high quality density ratio estimator is a rich and context-dependent topic, and is beyond the scope of this article.

4.1. Fixed Population Asymptotics

We first consider the classical setting where the two distributions P_1, P_2 are fixed, and study the limiting behavior of the test statistic as the ranking sample size n_{11} grows to infinity and n_{11}/n_{21} stays bounded and bounded away from 0. A local alternative analysis with varying signal strength is presented in Section 4.2.

Recall that we use the notation $g(x) = f_{2,X}(x)/f_{1,X}(x)$ and $v(x,y) = f_1(y|x)/f_2(y|x)$. For $k = 1, 2$, let $G_{ki} = g(X_{ki})$ and $V_{ki} = v(X_{ki}, Y_{ki})$. Let $D_{ij} = \mathbb{1}(V_{1i} < V_{2j}) + \zeta_j \mathbb{1}(V_{1i} = V_{2j})$ where ζ_j's are auxiliary $U(0,1)$ random variables independent of everything else. Define $\widehat{G}_{ki}$, $\widehat{V}_{ki}$ and $\widehat{D}_{ij}$ similarly as G_{ki}, V_{ki} and D_{ij} using the estimated functions $\widehat{g}$, $\widehat{v}$. For a random variable Z and constant $q > 0$, $\|Z\|_q$ denotes the ℓ_q norm of Z: $\|Z\|_q^q = \mathbb{E}(|Z|^q)$. Much of our analysis will involve the estimation errors in $\widehat{g}, \widehat{v}$ reflected through the random variables $\widehat{G}_{11} - G_{11}$ and $\widehat{D}_{11} - D_{11}$. We use $\mathbb{E}_*(\cdot)$, $\mathrm{var}_*(\cdot)$, and $\|\cdot\|_{q,*}$ to denote the conditional expectation, variance, and ℓ_q norm given the density ratio estimates $\widehat{v}, \widehat{g}$ (or, equivalently, given the fitting

subsample). For example, $\mathbb{E}_*(\widehat{\nu}(X_1, Y_1)) = \mathbb{E}(\widehat{\nu}(X_1, Y_1)|\widehat{\nu})$, and $\|\widehat{G}_{11} - G_{11}\|_{2,*} = [\mathbb{E}(\widehat{g}(X_{11}) - g(X_{11}))^2|\widehat{g}]^{1/2}$.

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 $\|G_{11}\|_2 < \infty$.

Our next assumption is on the asymptotic accuracy of the density ratio estimators.

Assumption 2. (a) $\|\widehat{G}_{11} - G_{11}\|_{2,*} = o_P(1)$.
(b) $\left|\mathbb{E}_*(\widehat{G}_{11} - G_{11})\widehat{D}_{11} - \mathbb{E}_*(\widehat{G}_{11} - G_{11})\mathbb{E}_*(G_{11}\widehat{D}_{11})\right| = o_P(1/\sqrt{n_{11}})$.

To help understand the notation, consider Assumption 2(a) for example. By construction and the definition of $\|\cdot\|_{2,*}$, we have $\|\widehat{G}_{11} - G_{11}\|_{2,*} = [\mathbb{E}(\widehat{g}(X_{11}) - g(X_{11}))^2|\widehat{g}]^{1/2}$, which is a random quantity, whose randomness comes from $\widehat{g}$, which is itself a function of the fitting subsamples. Therefore, Assumption 2(a) requires that with high probability over the randomness of the fitting subsamples, the estimate $\widehat{g}$ is close to g when their distance is measured using the ℓ_2 norm of $\widehat{g} - g$ under $f_{1,X}$. Notation in part (b) of Assumption 2 is interpreted accordingly.

Assumption 2(a) requires consistent estimation of the marginal density ratio $f_{1,X}(x)/f_{2,X}(x)$, which is mild. Assumption 2(b) deserves further discussion, which is deferred after the presentation of the main theorem on the asymptotic behavior of the test statistic output by Algorithm 1.

Theorem 1. Suppose that Assumptions 1 and 2 hold. The test statistic $\widehat{T}$ output by Algorithm 1 converges in distribution to the standard normal as $n_{11} \to \infty$ under H_0.

Let $\Delta = -\mathbb{E}_*(\widehat{G}_{11} - G_{11})\widehat{D}_{11} + \mathbb{E}_*(\widehat{G}_{11} - G_{11})\mathbb{E}_*(G_{11}\widehat{D}_{11}) + \mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})$. If there exists a constant $c > 0$ such that

$$\mathbb{P}\left[\Delta < (1/4)\mathbb{E}|\nu(X_2, Y_2) - \nu(X_2', Y_2')| - c\right] \to 1\,,$$

where (X_2, Y_2) and (X_2', Y_2') are iid realizations from P_2, then under H_1, $\widehat{T} \to \infty$ in probability.

Now we discuss Assumption 2(b), which is needed to ensure that the approximation error in the estimated weights does not break the central limit theorem. Consider the following two scenarios.

1. $\sqrt{n_{11}}\|\widehat{G}_{11} - G_{11}\|_{2,*} = o_P(1)$. In this case Assumption 2(b) follows immediately from boundedness of $\widehat{D}$ and Cauchy-Schwartz, regardless of $\widehat{\nu}$, the estimated conditional density ratio. This reflects the typical validity guarantee for conformal methods: when the weights are accurate enough, the Type I error is always controlled for all conformity score functions. However, in order to achieve such a convergence rate of $\|\widehat{G}_{11} - G_{11}\|_{2,*}$, we typically will need the fitting subsample size n_{12} to be much larger than the ranking subsample size n_{11}. In the special case where we have side information $f_{1,X} = f_{2,X}$, then there is no need to estimate $\widehat{g}$, and $\widehat{g} \equiv 1$ and Assumption 2 holds trivially.
2. $\sqrt{n_{11}}\|\widehat{G}_{11} - G_{11}\|_{2,*}$ does not converge to 0 in probability. This is more likely to be the case when $n_{12} \asymp n_{11}$, and is of major practical interest. In this case, the convergence of $\widehat{G}_{11} - G_{11}$ alone is not enough to control the approximation error in the weights. In order for Assumption 2(b) to hold we will need the random variable $\widehat{D}_{11}$, which involves the estimated conditional density ratio $\widehat{\nu}$, to behave reasonably. Specifically, after some simple algebra, the left hand side of Assumption 2(b) equals

$$|\mathrm{cov}_*(\widehat{G}_{11} - G_{11}, \widehat{D}_{11}) - \mathbb{E}_*(\widehat{G}_{11} - G_{11})\mathrm{cov}_*(G_{11}, \widehat{D}_{11})|\,,$$

which is upper bounded by (ignoring constant factors) $\rho\|\widehat{G}_{11} - G_{11}\|_{2,*}$, with

$$\rho = \max\left(|\mathrm{corr}_*(\widehat{G}_{11} - G_{11}, \widehat{D}_{11})|,\ |\mathrm{corr}_*(G_{11}, \widehat{D}_{11})|\right).$$

There is good reason to expect $\mathrm{corr}_*(\widehat{G}_{11} - G_{11}, \widehat{D}_{11})$ and $\mathrm{corr}_*(G_{11}, \widehat{D}_{11})$ to be close to 0, because for a good estimate $\widehat{\nu}$, the randomness in $\widehat{\nu}(X_{11}, Y_{11})$ will be mostly from the conditional randomness of Y_{11} given X_{11}, which is independent of X_{11} itself. But both $\widehat{G}_{11}$ and G_{11} are functions of X_{11}, so we should expect $\widehat{\nu}(X_{11}, Y_{11})$ (and hence $\widehat{D}_{11}$) to be nearly independent of $\widehat{G}_{11}$ and G_{11}. Consider a simple Gaussian mean shift example, where $X_{11} \sim N(-\delta/2, 1)$, $(Y_{11}|X_{11}) \sim N(X_{11} - \mu/2, 1)$, $X_{21} \sim N(\delta/2, 1)$, $(Y_{21}|X_{21}) \sim N(X_{21} + \mu/2, 1)$. Suppose all functions are estimated using the parametric maximum likelihood, then $\widehat{D}_{11} = \mathbb{1}[\widehat{\mu}(Y_{11} - X_{11}) > \widehat{\mu}(Y_{21} - X_{21})]$, which is independent of X_{11}, because, by construction and joint Gaussianity, $Y_{11} - X_{11}$ is independent of X_{11}. This example can be extended to more complex versions, such as heteroscedastic responses. We expect that a weak dependence between $\widehat{D}_{11}$ and X_{11} also holds true for other good conditional density ratio estimators. In our numerical experiments, we observe that such weak dependence assumption is indeed plausible in many settings. See Appendix B.2 for detailed empirical evidences. In practice, if we are confident about the density ratio estimate, such as when using a correctly specified parametric model, then an equal-sized sample split for fitting and ranking is recommended. Otherwise, one could use a larger sample size for fitting and a smaller sample size for ranking. We report empirical results for different fitting-ranking split ratios in Appendix B.1.

The asymptotic power guarantee only requires $\widehat{D}_{11}$ and $\widehat{G}_{11}$ to be within a constant distance from their corresponding population versions. This is because when there is a constant separation between the data distribution and the null model, a small constant distortion in the test statistic will not remove all the signal. Below we provide a more delicate analysis under a local alternative framework.

4.2. A Local Alternative Analysis

Now we consider a local alternative scenario such that the two joint distributions (P_1, P_2) may change with the sample size (n_{11}, n_{21}). This provides a more refined view of different sources of the approximation error and the accumulation of signal strength when the sample size increases.

Following Theorem 1 and Lemma 2(c), we use the following quantity to quantify the deviation from the null,

$$\delta = \delta(P_1, P_2) = \mathbb{E}|\nu(X_2, Y_2) - \nu(X_2', Y_2')|\,,$$

where $(X_2, Y_2), (X_2', Y_2') \overset{\text{iid}}{\sim} P_2$. By construction $\mathbb{E}\nu(X_2, Y_2) = 1$ and $\delta = 0$ if and only if $\mathbb{P}(\nu(X_2, Y_2) = 1) = 1$, which is equivalent to H_0. A larger value of δ indicates the conditional density ratio $f_1(Y|X)/f_2(Y|X)$ is likely to be far away from 1. In fact, using the triangle inequality and Jensen's inequality we can show that

$$\mathbb{E}_{X\sim f_{2,X}} D_{\text{tv}}(f_1(\cdot|X), f_2(\cdot|X)) = \frac{1}{2}\mathbb{E}|\nu(X_2, Y_2) - 1| \in [\delta/4, \delta/2], \tag{11}$$

where D_{tv} is the total variation distance between two distributions. A detailed proof of this claim is given in Appendix D.

Under a slightly stronger technical condition than those in Theorem 1, we have the following local alternative result.

Proposition 1. In addition to Assumptions 1 and 2, suppose that $\left|\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})\right| = o_P(n_{11}^{-1/2})$, $\widehat{\nu}(X_{21}, Y_{21}) - \nu(X_{21}, Y_{21}) = o_P(1)$, and $\nu(X_{21}, Y_{21})$ has a continuous distribution with bounded density. Then, we have

$$\widehat{T} = \frac{\sqrt{n_{11}}\delta}{4\sigma}(1 + o_P(1)) + Z + o_P(1),$$

where σ is the population version of $\widehat{\sigma}$ in (10) using the true density ratios g and ν, and $Z \rightsquigarrow N(0, 1)$ as $n_{11} \to \infty$.

The most nontrivial additional assumption here is $\left|\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})\right| = o_P(n_{11}^{-1/2})$, whereas only a constant error bound is required in Theorem 1 for the test to be powerful against a constant alternative. This is because the local alternative is very close to the null, while in the fixed population analysis considered in Theorem 1, the difference between the null and alternative is much larger. This more stringent assumption can still be realistic for the same reason explained in the discussion after Theorem 1. Numerical evidences are provided in Appendix B.2. The continuity of $\nu(X_2, Y_2)$ and bounded density allow us to provide more refined control on the difference between indicators $\widehat{D}_{11} - D_{11}$.

Proposition 1 suggests the following local asymptotic behavior of our test statistic.

1. If $\delta\sqrt{n_{11}} \to \infty$, then $\widehat{T} \to \infty$ in probability.
2. If $\delta\sqrt{n_{11}} \to a \in [0, \infty)$, then $\widehat{T} \rightsquigarrow N(a/(4\sigma), 1)$.

As a result, in the asymptotic regime considered here, the power is mostly determined by $\delta\sqrt{n_{11}}$.

5. Simulation Study

In this section, we illustrate the performance of our method in several simulation settings. For brevity, we focus on the more challenging and interesting case where the X-marginals are different. Denote $x_i = (x_i(1), \ldots, x_i(p))^{\mathrm{T}}$, and I_p is a $p \times p$ identity matrix. We first consider three prototypical regression models with $p = 5$ that are similar to those in Lei et al. (2018) and Zheng (2000). A higher dimensional case is presented in Section 5.2.

Model A (Gaussian, linear). Let $y_{\ell i} = \alpha_\ell + \beta^{\mathrm{T}} x_{\ell i} + \epsilon_{\ell i}$, $i = 1, \ldots, n_\ell, \ell = 1, 2$, where $x_{1i} \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $x_{2i} \overset{\text{iid}}{\sim} N(\mu, I_p)$ where $\mu = (1, 1, -1, -1, 0)^{\mathrm{T}}$, and $\epsilon_{1i}, \epsilon_{2i} \overset{\text{iid}}{\sim} N(0, 1)$, independent of the features. Set $\alpha_1 = \alpha_2 = 0$ under the null and $\alpha_1 = 0, \alpha_2 = 0.5$ under the alternative.

Model B (Gaussian mixture, nonlinear, heavy-tailed). Let $y_{\ell i} = \alpha_\ell + \beta_1 x_{\ell i}(1) + \beta_2 x_{\ell i}(2) + \beta_3 x_{\ell i}^2(3) + \beta_4 x_{\ell i}^2(4) + \beta_5 x_{\ell i}^3(5) + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2$, where $x_{1i} \overset{\text{iid}}{\sim} 0.5N(\mathbf{0}, I_p) + 0.5N(\mu, I_p)$, $x_{2i} \overset{\text{iid}}{\sim} 0.5N(\mathbf{0}, I_p) + 0.5N(\mathbf{0}, 1.5I_p)$ where $\mu = (0.5, 0.5, -0.5, -0.5, 0)^{\mathrm{T}}$, and $\epsilon_{1i}, \epsilon_{2i} \overset{\text{iid}}{\sim} t(5)$, the student's t-distribution with 5 degrees of freedom, independent of the features. Set $\alpha_1 = \alpha_2 = 0$ under the null and $\alpha_1 = 0, \alpha_2 = 0.5$ under the alternative.

Model C (Gaussian mixture, additive spline, heteroscedastic). Let $y_{\ell i} = \theta(x_{\ell i}) + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2$, where $\theta(x) = \mathbb{E}(y|x)$ is an additive function of B-splines of covariates, $x_{1i} \overset{\text{iid}}{\sim} 0.5N(\mathbf{0}, I_p) + 0.5N(\mu, I_p)$, $x_{2i} \overset{\text{iid}}{\sim} 0.5N(\mathbf{0}, I_p) + 0.5N(\mathbf{0}, 1.5I_p)$ where $\mu = (0.5, 0.5, -0.5, -0.5, 0)^{\mathrm{T}}$. Set $\epsilon_{\ell i} \sim N(0, 4/(1 + x_{\ell i}^2(1)))$, $\ell = 1, 2$, under the null and $\epsilon_{1i} \sim N(0, 4/(1 + x_{1i}^2(1)))$, $\epsilon_{2i} \sim N(0, 2/(1 + x_{2i}^2(1)))$ under the alternative. Here the noises are not independent of the covariates.

In order to make density ratio estimation stable, we remove sample points whose marginal density ratio $f_{1,X}(x)/f_{2,X}(x)$ or the joint density ratio $f_1(x, y)/f_2(x, y)$ are outside of the interval $[1/100, 100]$.

5.1. The Low-Dimensional Case

We first consider low-dimensional cases with $p = 5$, setting the entries of β to ± 1 with random signs in Models A and B. In Model C, we multiply the coefficients to the B-spline transformation of predictors: $\theta(x) = \sum_{j=1}^{5}\sum_{l=1}^{4} \beta_{jl} b_l(x(j))$, where b_l's are B-spline functions and $\beta_{jl} = \pm 1$ with randomly chosen signs. We consider sample sizes $n_1 = n_2 \in \{200, 500, 1000, 2000\}$ and randomly split each sample into two equal-sized subsets for the fitting and ranking steps. Results for other split ratios are deferred to Appendix B.1. In addition, we use four different probabilistic classification methods including Linear Logistic (LL), Quadratic Logistic (QL), Neural Network (NN) and Kernel Logistic Regression (KLR). Let L be the class label such that $L_{1i} = 1$ for $i \in \mathcal{I}_{12}$ and $L_{2i} = 0$ for $i \in \mathcal{I}_{22}$. The KLR method (Zhu and Hastie 2005) learns a kernel logistic regression classifier by minimizing

$$-\sum_{k\in\{1,2\}}\sum_{i\in\mathcal{I}_{k2}} \left[L_{ki}\theta(x_{ki}; \boldsymbol{\beta}) - \log\{1 + \exp(\theta(x_{ki}; \boldsymbol{\beta}))\}\right] + \frac{\lambda}{2}\|\theta\|_{\mathcal{H}_K}^2,$$

where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Spaces (RKHS) generated by the kernel $K(x, y) = \exp(-\|x-y\|^2/\sigma^2)$, $\theta(x; \boldsymbol{\beta}) = \beta_0 + \sum_{k\in\{1,2\}}\sum_{i\in\mathcal{I}_{k2}} \beta_{ki} K(x_{ki}, x)$ and $\eta(x; \boldsymbol{\beta}) = P(L = 1|x) = 1/[1 + \exp\{-\theta(x; \boldsymbol{\beta})\}]$. Then we obtain the marginal ratio estimator $\widehat{g}(x) = n_{12}\{1 - \eta(x; \widehat{\boldsymbol{\beta}})\}/\left\{n_{22}\eta(x; \widehat{\boldsymbol{\beta}})\right\}$. The joint classifier is obtained similarly using (x_i, y_i). For the KLR method, the tuning parameter $\sigma^2 = 200$ is used in all settings. The more important tuning parameter λ is chosen by minimizing the out-of-sample cross entropy loss. For large sample sizes such data-driven tuning is time-consuming to run for all repetitions. To reduce the overall running time for large sample sizes ($n_2 = 1000, 2000$), we use data-driven out-of-sample cross entropy loss to select λ in first 20 repetitions, and then use the median of these 20 λ values for the rest of the simulation. Specifically, when $n_2 = 1000, 2000$, we use $\lambda = 0.05$ for both joint and marginal ratio estimates in Model A, $\lambda = 0.0002$ for joint estimates and

$\lambda = 0.015$ for marginal estimates in Model B, $\lambda = 0.015$ for joint estimates and $\lambda = 0.02$ for marginal estimates in Model C. For the neural network method, we use the sigmoid activation function and the stochastic gradient descent algorithm, and two hidden layers are used in all considered models. Moreover, we compare with the parametric Likelihood Ratio Test (LRT) for hypotheses $H_0 : \alpha_1 = \alpha_2$ versus $H_1 : \alpha_1 \neq \alpha_2$ under the model $(Y_\ell|X) \sim N(\alpha_\ell + \beta^{\mathrm{T}}X, \sigma^2), \ell = 1, 2$. Such a model is correctly specified under model A, where we expect LRT to have the best performance. Under models B, C, this is misspecified. The code for the simulation is available to ensure reproducibility.

Among the four classification methods and three model settings considered, LL and QL rely on parametric models, which is correctly specified under model A, but incorrectly specified under models B and C, where the marginal distributions of X are Gaussian mixtures, and the corresponding density ratios cannot be expressed as a logit transform of second order polynomials of X. Moreover, the noise distribution is non-Gaussian under model B, and Gaussian but heteroscedastic in model C. The KLR and NN methods do not rely on any parametric model specification. We summarize the model specification scenarios in Table 1.

All the simulation results in the following are computed over 500 repetitions with nominal Type I error level $\alpha = 0.05$. The results are summarized in Table 2. In Model A, the LL estimator is the correctly specified parametric method, and has very good performance even with small sample sizes. QL has relatively inaccurate empirical size for very small sample size $n_2 = 200$, because this model involves more parameters including all the linear and quadratic terms. The NN and KLR estimators are fully nonparametric, but still yield satisfying control of the Type I error even under moderate sample sizes. When the alternative hypothesis is true, the power increases as the sample size increases. The NN and KLR methods yield comparable or larger power against the LL approach even when the sample size is small, thanks to accurately estimated density ratios v and g. The QL estimator requires a larger sample size to obtain reasonable power. The performance using the estimated $\widehat{v}, \widehat{g}$ is indeed close to using the true functions v, g, which is presented in Table 3.

Models B and C represent more challenging scenarios where the marginal and joint ratios are nonparametric. The parametric methods such as LL and QL fail to yield correct Type I error control or have limited ability to capture the difference under the alternative in at least one model setting, which demonstrates the limitation of such parametric approaches. The nonparametric methods NN and KLR both provide robust Type I and II error control, with empirical Type I error rate close to the nominal level. Moreover, the simulation results demonstrate that the data-driven tuning by minimizing out-of-sample cross entropy can have good practical performance.

Note that in Model A, $\alpha_1 = \alpha_2$ under the null and $\alpha_1 \neq \alpha_2$ under the alternative hypothesis, while in Models B and C, the parametric assumption for the LRT is violated. As shown in Table 2, the LRT method performs the best in Model A, which makes sense since it maximizes the usage of known information about the data and does not require sample splitting. In Model B, the LRT is not correctly specified because the true conditional distribution of Y given X is t-distribution and the conditional mean is not a linear function. Although it shows some ability to capture the distributional difference, it also has inflated Type I error due to model misspecification. Further, the LL, KLR, and NN have comparable power with the LRT, demonstrating the advantage of the proposed approach by efficiently aggregating multiple conformal p-values, which reduces the impact of the sample splitting. Moreover, in Model C, the LRT method fails to control the Type I error, and is significantly less powerful than the QL, NN and KLR methods, demonstrating the benefit brought by the flexibility of our method in choosing the estimation algorithms in nonparametric models.

Table 1. The model misspecification for all methods: "✓" means the estimator uses a correctly specified parametric model to estimate the (conditional) density ratio; "✗" means that the estimator uses a misspecified parametric model to estimate the (conditional) density ratio; "–" means that the estimator is nonparametric and does not rely on any parametric model specification.

Task	$\widehat{g}$					$\widehat{v}$				
Method	LRT	LL	QL	KLR	NN	LRT	LL	QL	KLR	NN
Model A	✓	✓	✓	–	–	✓	✓	✓	–	–
Model B	✗	✗	✗	–	–	✗	✗	✗	–	–
Model C	✗	✗	✗	–	–	✗	✗	✗	–	–

5.2. The High-Dimensional Case

The flexibility of choosing probabilistic classification algorithms makes our method applicable in high-dimensional problems.

Table 2. Percentage of rejections over 500 repetitions using methods LL, QL, NN, and KLR under the split ratio 0.5 and the Likelihood Ratio Test (LRT) with $\alpha = 0.05$.

		Null					Alternative				
		LL	QL	NN	KLR	LRT	LL	QL	NN	KLR	LRT
Model A	$n_2 = 200$	0.038	0.022	0.066	0.038	0.054	0.354	0.082	0.416	0.406	0.992
	$n_2 = 500$	0.044	0.020	0.058	0.066	0.054	0.654	0.392	0.644	0.698	1.000
	$n_2 = 1000$	0.042	0.038	0.058	0.056	0.032	0.898	0.770	0.866	0.912	1.000
	$n_2 = 2000$	0.048	0.056	0.068	0.068	0.038	0.980	0.974	0.990	0.990	1.000
Model B	$n_2 = 200$	0.044	0.058	0.052	0.062	0.054	0.200	0.116	0.180	0.224	0.236
	$n_2 = 500$	0.058	0.046	0.076	0.062	0.040	0.456	0.268	0.470	0.506	0.478
	$n_2 = 1000$	0.056	0.074	0.024	0.024	0.044	0.726	0.586	0.722	0.802	0.764
	$n_2 = 2000$	0.066	0.062	0.038	0.026	0.080	0.946	0.936	1.000	0.994	0.970
Model C	$n_2 = 200$	0.080	0.066	0.056	0.064	0.090	0.074	0.204	0.062	0.300	0.088
	$n_2 = 500$	0.070	0.036	0.066	0.040	0.096	0.080	0.624	0.504	0.796	0.116
	$n_2 = 1000$	0.098	0.046	0.030	0.064	0.142	0.088	0.964	0.900	0.988	0.144
	$n_2 = 2000$	0.110	0.050	0.080	0.064	0.182	0.094	0.998	1.000	1.000	0.176

Table 3. Percentage of rejections using methods LL, QL, NN, and KLR with $\alpha = 0.05$ under different sample splitting ratios for Model A.

		Null					Alternative				
		LL	QL	NN	KLR	Oracle	LL	QL	NN	KLR	Oracle
$n_2 = 200$	r=0.3	0.024	0.024	0.034	0.046	0.042	0.282	0.092	0.338	0.336	0.300
	r=0.5	0.038	0.022	0.066	0.038	0.052	0.354	0.082	0.416	0.406	0.444
	r=0.8	0.034	0.010	0.136	0.042	0.042	0.350	0.010	0.398	0.408	0.578
$n_2 = 500$	r=0.3	0.050	0.026	0.034	0.052	0.038	0.550	0.382	0.496	0.534	0.580
	r=0.5	0.044	0.020	0.058	0.066	0.038	0.654	0.392	0.644	0.698	0.720
	r=0.8	0.032	0.024	0.086	0.042	0.066	0.770	0.216	0.822	0.752	0.864
$n_2 = 1000$	r=0.3	0.048	0.044	0.068	0.044	0.024	0.744	0.672	0.730	0.770	0.754
	r=0.5	0.042	0.038	0.058	0.056	0.050	0.898	0.770	0.866	0.912	0.902
	r=0.8	0.048	0.030	0.088	0.070	0.054	0.962	0.700	0.932	0.974	0.980
$n_2 = 2000$	r=0.3	0.042	0.052	0.050	0.066	0.040	0.938	0.902	0.920	0.932	0.934
	r=0.5	0.048	0.056	0.068	0.068	0.070	0.980	0.974	0.990	0.990	0.986
	r=0.8	0.064	0.044	0.088	0.066	0.064	1.000	0.968	0.994	0.998	1.000

Here we illustrate its performance in a high-dimensional scenario, which is similar to Model A in the low dimensional case but with ambient dimensionality $p = 500$ and signal dimensionality $s = 5$. The additional coordinates of X are generated by iid standard normal, and the corresponding coefficients of $\boldsymbol{\beta}$ are filled with zeros. Here we focus on a sparse linear classifier and investigate the effect of tuning and regularization. Letting L be the class label with $L = 1$ for training data and $L = 0$ for testing data, we learn a sparse logistic regression model by minimizing

$$-\frac{1}{n_{12}+n_{22}}\sum_{k\in\{1,2\}}\sum_{i\in\mathcal{I}_{k2}}\big[L_{ki}\log\eta(x_{ki};\boldsymbol{\beta}) + (1-L_{ki})\log\{1-\eta(x_{ki};\boldsymbol{\beta})\}\big]+\lambda\|\boldsymbol{\beta}\|_1,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\eta(x;\boldsymbol{\beta}) = P(L = 1|x) = 1/[1 + \exp\{-\beta_0 - \sum_{j=1}^{p}\beta_j x(j)\}]$. Then we obtain the marginal ratio estimator with $\widehat{g}(x) = n_{12}\{1-\eta(x;\widehat{\boldsymbol{\beta}})\}/\left\{n_{22}\eta(x;\widehat{\boldsymbol{\beta}})\right\}$. In a similar manner, we can the estimate the joint density ratio and hence the conditional density ratio.

We consider sample sizes $n_1 = n_2 = 1000$ and 2000. The empirical rejection frequency and estimation errors of the conformal weights, $\mathrm{Err}_{\widehat{g}} = n_{11}^{-1}\sum_i|\widehat{G}_{1i}/\sum_i\widehat{G}_{1i} - G_{1i}/\sum_i G_{1i}|$, are shown in Figure 1. Since the estimation errors are not observable in practice, we plot the out-of-sample marginal cross entropy error (MCEntropy, defined as $-L\log\widehat{p} - (1-L)\log(1-\widehat{p})$) in the classification problem involved in estimating $\widehat{g}$ (the solid lines with star-shaped marks). Moreover, under the alternative, we also report the empirical out-of-sample estimation error of the conditional density ratio v, defined by $Err_{\widehat{v}} = (n_{11}+n_{21})^{-1}\left\{\sum_i(\widehat{V}_{1i}-V_{1i})^2 + \sum_j(\widehat{V}_{2j}-V_{2j})^2\right\}$.

Again, when the true marginal density ratios are used, the empirical sizes are close to the nominal level $\alpha = 0.05$ as expected. When the marginal density ratios 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 out-of-sample marginal cross entropy error suggests that in practice one can choose the tuning parameter value near the elbow of the error plot. Under the alternative hypotheses, the power is maximized at tuning parameter values corresponding to the smallest estimation error in $\widehat{v}$, which can also be chosen using its out-of-sample cross entropy error plot (not shown in the plots). Practically one can also use separate tuning parameters for the marginal classification and joint classification.

6. A Synthetic Data Example

We consider the airfoil dataset from the UCI Machine Learning Repository (Dua and Graff 2019), which has $n = 1503$ observations of a response Y (scaled sound pressure level of NASA airfoils), and a covariate X with $p = 5$ dimensions (log frequency, angle of attack, chord length, free-stream velocity, and suction side log displacement thickness). This dataset has been used by Tibshirani et al. (2019), who first studied weighted conformalization.

The original data does not have a two-sample separation. We consider five experiments based on different ways to generate the two populations.

(i) Random partition. We randomly partition the dataset with $n_1 = 751$ and $n_2 = 752$.

(ii) Random partition and exponential tilting. We first randomly partition the data into twosets $\mathcal{D}_1, \widetilde{\mathcal{D}}_2$. Then following Tibshirani et al. (2019), we construct $\mathcal{D}_2$ by sampling 25% of the points from $\widetilde{\mathcal{D}}_2$ with replacement, with probabilities proportional to

$$w(x) = \exp(x^{\mathrm{T}}\alpha), \quad \text{where } \alpha = (-1, 0, 0, 0, 1).$$

The final sample sizes are $n_1 = 301$ and $n_2 = 301$.

(iii) Chord-based partition. We split the dataset into two subsets where the values of the "chord" variable in $\mathcal{D}_1$ are smaller than the 50% quantile and exclude the "chord" variable in subsequent analyses, resulting in $n_1 = 778$ and $n_2 = 725$. To avoid singularity between the two populations, we randomly select $0.05n_1$ samples in $\mathcal{D}_1$ and the same amount of samples in $\mathcal{D}_2$ to flip their groups.

(iv) Velocity-based partition. We partition the dataset into two subsets where the values of the "velocity" variable in $\mathcal{D}_1$ are smaller than the 50% quantile and remove the covariate "velocity" from subsequent analyses, with $n_1 = 761$ and $n_2 = 742$. We randomly select $0.05n_1$ samples in $\mathcal{D}_1$ and the same amount of samples in $\mathcal{D}_2$ to flip their groups.

(v) Response-based partition. We split the data according to the value of the response variable, where the first group contains the sample points with smaller response values, while $\mathcal{D}_2$ contains the rest, with $n_1 = 752$ and $n_2 = 751$. A similar label flipping is applied to avoid singularity.

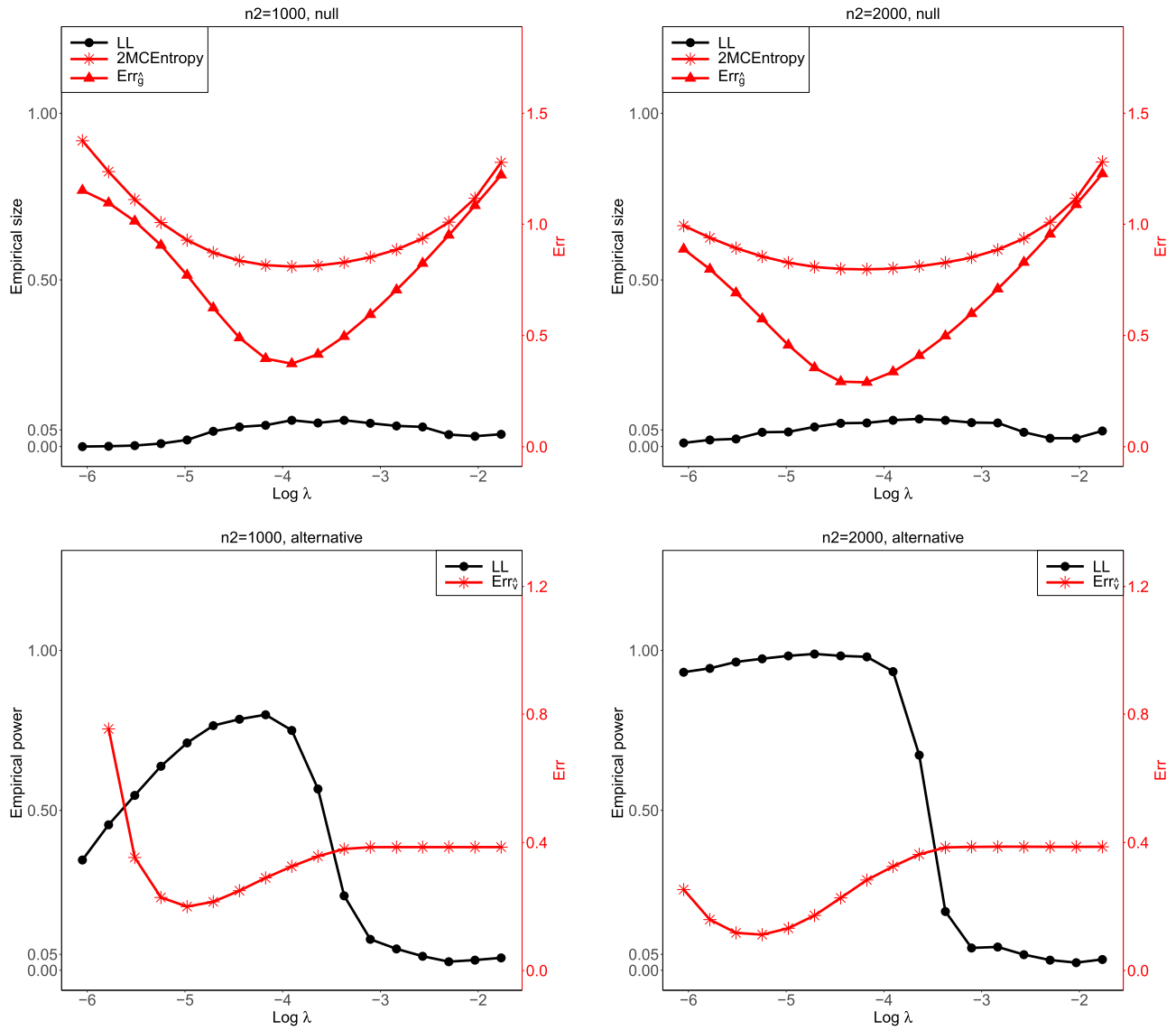

Figure 1. Performance of empirical rejection frequency in Model A with $p = 500, s = 5$ under the null (top) and alternative (bottom) over 500 repetitions, across a variety of regularization parameters using sparse LL with $\alpha = 0.05$ and split ratio 0.5.

Table 4. Percentage of rejections (PR) and average classification error of estimating g in Airfoil dataset for cases (i–ii) using methods LL and NN over 500 repetitions, with $\alpha = 0.05$ and split ratio 0.5.

	PR_{LL}	PR_{NN}	MCE_{LL}	MCE_{NN}
case (i)	0.048	0.056	0.500	0.500
case (ii)	0.052	0.068	0.208	0.208

Table 5. Median p-values (Pval) and average classification error of estimating g in Airfoil dataset for cases (iii–v) using methods LL and NN over 500 random splits, with $\alpha = 0.05$ and split ratio 0.5.

	$Pval_{LL}$	$Pval_{NN}$	MCE_{LL}	MCE_{NN}
case (iii)	0.005	0.472	0.160	0.053
case (iv)	0.000	0.412	0.444	0.066
case (v)	0.000	0.002	0.233	0.136

As in the simulation study, we split each group into two equal-sized subsets, and conduct the two-sample conditional distribution test at significance level $\alpha = 0.05$. Linear LL and NN are used to estimate the density ratios. The neural network uses two hidden layers with ten nodes in all cases. Each experiment is repeated for 500 trials.

For experiments (i) and (ii), which are clearly under the null hypothesis, we can regenerate the data by repeating the random generation of the two subsamples. With these repeatedly generated datasets, we can compute the empirical frequency of rejections. As shown in Table 4, the Type I errors are close to the nominal level.

For experiments (iii)–(v) (Table 5), we can only have a single deterministic generation of the training and testing data, except a small fraction of group flipping, and only a single p-value can be computed. We use these experiments to illustrate the effect of multiple realizations of auxiliary randomization. Recall that our method uses auxiliary randomization to split the datasets into fitting and ranking subsamples. Such auxiliary randomization may lead to different results on the same dataset if the inference is carried out independently by different researchers. To mitigate this effect, one can obtain multiple p-values using multiple realizations of auxiliary randomization. Although each

single p-value has asymptotically valid null distribution, their dependence requires a careful aggregation of these p-values. Here we use a median p-value approach in DiCiccio, DiCiccio, and Romano (2020). Formally, suppose we repeat the auxiliary randomization B times, obtaining p-values $\widehat{p}_1, \ldots, \widehat{p}_B$. Then $\widehat{p} = 1 \wedge \left[2 \times \text{Median}(\widehat{p}_1, \ldots, \widehat{p}_B)\right]$ is a valid p-value.

In all experiments (i–v), we use out-of-sample Marginal Classification Error (MCE) as a proxy of the accuracy of marginal density ratio estimation. In experiment (iii), both LL and NN methods give large p-values, and the NN method also has small marginal classification errors. Thus, there is no strong evidence against the covariate shift assumption. In experiment (iv), the LL method gives small p-values while the NN method suggests otherwise. But the marginal classification errors indicate that the NN method is likely to be more accurate in estimating the marginal density ratios, and hence, provides more trustworthy p-values. In experiment (v), both methods agree to reject the null hypothesis, which is the correct decision by the construction of training and testing samples. The neural net method also gives marginal classification errors comparable to those in the null cases, further confirming the validity of p-value.

7. Discussion

In applications it is often the case that the training data $(X_{1i}, Y_{1i})_{i=1}^{n_1}$ has a large sample size, whereas the testing data $(X_{2j}, Y_{2j})_{j=1}^{n_2}$ has a limited sample size. As our theory and experiments have demonstrated, 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. In this case we can use these unlabeled testing sample points to estimate the marginal density ratio, and save the scarce labeled testing sample to estimate the joint density ratio.

The use of sample splitting and auxiliary randomization for valid and efficient statistical inference has been studied by many authors in the high dimensional regression literature (Wasserman and Roeder 2009; Meinshausen, Meier, and Bühlmann 2009; Rinaldo, Wasserman, and G'Sell 2019), and more recently in the conformal inference literature (Kuchibhotla and Ramdas 2019; Kim, Xu, and Barber 2020). The theory in Kim, Xu, and Barber (2020) also required an inflated noncoverage by a factor of two after aggregating multiple subsamples. It is unclear whether such a loss of coverage is unavoidable. It would be interesting and important to better understand the dependence between the p-values from multiple splits, and improve the current conservative inflation method when combining multiple p-values.

Conformal methods are initially developed without sample splitting, but in a leave-one-out manner. The validity of such conformal p-values comes from the symmetry among the data points. A straightforward leave-one-out version of our method would be to leave out each pair of sample points (X_{1i}, Y_{1i}), (X_{2j}, Y_{2j}) for $i \in [n_1]$ and $j \in [n_2]$, and fit $\widehat{g}, \widehat{v}$ using all remaining data. Such an implementation will require $n_1 n_2$ refitting of $\widehat{g}$, $\widehat{v}$, which is computationally prohibitive. Some efficient implementation of leave-one-out update or warm start techniques would be necessary. Also, the complex dependencies among the resulting conformity scores bring further theoretical challenges in understanding the aggregated conformal p-values. Given the potential improved sample efficiency, these questions may be investigated in future works.

Appendix A: Asymptotic Variance Using Importance Sampling

We illustrate the importance sampling idea for estimating the asymptotic variance using the samples from P_2. For simplicity we focus on the ideal statistic T which uses the true density ratio functions. The same idea can be directly carried over to the actual statistic that uses estimated density ratios.

Recall that the asymptotic variance of $\sqrt{n_{11}} n_{21}^{-1} \sum_j U_j$ is given by $\sigma^2 = \sigma_1^2 + n_{11}/(12 n_{21}) + \sigma_2^2/4 - \rho_{12}$ where $\sigma_1^2 = \text{var}\left[G_{11}\{1 - F_{1/2}(V_{11})\}\right]$, $\sigma_2^2 = \text{var}(G_{11})$ and $\rho_{12} = \text{cov}[G_{11}\{1 - F_{1/2}(V_{11})\}, G_{11}]$.

Note that under H_0,

$$\begin{aligned}\sigma_1^2 &= \int \frac{f_2^2(z)}{f_1^2(z)}\{1 - F_{1/2}(v(z))\}^2 f_1(z)dz - \frac{1}{4} \\ &= \int \frac{f_2(z)}{f_1(z)}\{1 - F_{1/2}(v(z))\}^2 f_2(z)dz - \frac{1}{4} \\ &= \mathbb{E} G_{21}\{1 - F_{1/2}(V_{21})\}^2 - \frac{1}{4}.\end{aligned}$$

Analogously, we have

$$\sigma_2^2 = \mathbb{E} G_{21} - 1,$$

$$\rho_{12} = \mathbb{E} G_{21}\{1 - F_{1/2}(V_{21})\} - 1/2.$$

Essentially, the change of base measure allows us to represent the expectations under P_1 to corresponding expectations under P_2. So we can use the sample $(X_{2j}, Y_{2j})_{j=1}^{n_{21}}$ to estimate the asymptotic variance. The consistency of the importance sample estimate follows the same strategy as the proof for the original estimate and is omitted.

Appendix B: More Simulation Results

B.1. Additional Simulation Results Under Different Splitting Ratios

To investigate the effect of the splitting ratio on the performance of the test, we consider $r = 0.3, 0.5$, and 0.8, respectively, where $n_{11} = \lceil n_1 * r \rceil$ and $n_{21} = \lceil n_2 * r \rceil$. We provide the results for Models A, B, and C in Tables 3–7. Additionally, the results of the proposed approach with the true functions g, v are also included in these tables, denoted by "Oracle". We should note that the "Oracle" requires no estimation, thus, it achieves higher power as r increases while still controlling the Type I error. By contrast, when r increases, the sample size of the fitting data decreases, and the estimation algorithms may produce estimates with limited accuracy for the testing procedure. For example, as shown in Table 3, the NN fails to control the Type I error in model A under $n_2 = 200$ and $r = 0.8$. Moreover, a larger splitting ratio r may lead to lower power due to the inaccurate estimates. From Tables 6 and 7, when $n_2 = 2000$, the NN achieves lower power under $r = 0.8$ than $r = 0.5$. In contrast, a smaller r will lead to more accurate estimates, but the test may suffer from power loss because less data are used for constructing the test statistic. More specifically, when $r = 0.3$, though the Type I error is well controlled for NN and KLR in all considered models, the power is lower than $r = 0.5$ especially for the small sample size. Overall, we suggest using $r = 0.5$, and a smaller r is allowed if one is not very confident in the resulting estimators in practice.

Table 6. Percentage of rejections using methods LL, QL, NN, and KLR with $\alpha = 0.05$ under different sample splitting ratios for Model B.

		Null					Alternative				
		LL	QL	NN	KLR	Oracle	LL	QL	NN	KLR	Oracle
$n_2 = 200$	r=0.3	0.064	0.048	0.056	0.044	0.060	0.176	0.128	0.192	0.178	0.680
	r=0.5	0.044	0.058	0.052	0.062	0.066	0.200	0.116	0.180	0.224	0.858
	r=0.8	0.066	0.032	0.032	0.112	0.052	0.194	0.068	0.086	0.302	0.970
$n_2 = 500$	r=0.3	0.070	0.056	0.050	0.042	0.058	0.338	0.246	0.356	0.352	0.942
	r=0.5	0.058	0.046	0.076	0.062	0.044	0.456	0.268	0.470	0.506	0.994
	r=0.8	0.058	0.070	0.038	0.152	0.064	0.518	0.206	0.458	0.560	0.998
$n_2 = 1000$	r=0.3	0.052	0.054	0.030	0.030	0.056	0.584	0.514	0.714	0.728	1.000
	r=0.5	0.056	0.074	0.024	0.024	0.054	0.726	0.586	0.722	0.802	1.000
	r=0.8	0.062	0.066	0.180	0.024	0.052	0.774	0.390	0.774	0.608	1.000
$n_2 = 2000$	r=0.3	0.064	0.050	0.054	0.054	0.062	0.822	0.858	0.998	0.986	1.000
	r=0.5	0.066	0.062	0.038	0.026	0.060	0.946	0.936	1.000	0.994	1.000
	r=0.8	0.056	0.078	0.054	0.022	0.064	0.952	0.810	0.866	0.972	1.000

Table 7. Percentage of rejections using methods LL, QL, NN, and KLR with $\alpha = 0.05$ under different sample splitting ratios for Model C.

		Null					Alternative				
		LL	QL	NN	KLR	Oracle	LL	QL	NN	KLR	Oracle
$n_2 = 200$	r=0.3	0.074	0.064	0.048	0.062	0.050	0.062	0.190	0.072	0.202	0.530
	r=0.5	0.080	0.066	0.056	0.064	0.046	0.074	0.204	0.062	0.300	0.760
	r=0.8	0.062	0.022	0.040	0.098	0.060	0.062	0.118	0.040	0.230	0.908
$n_2 = 500$	r=0.3	0.080	0.042	0.038	0.054	0.038	0.062	0.580	0.468	0.684	0.900
	r=0.5	0.070	0.036	0.066	0.040	0.036	0.080	0.624	0.504	0.796	0.984
	r=0.8	0.074	0.034	0.056	0.074	0.044	0.060	0.366	0.132	0.702	0.994
$n_2 = 1000$	r=0.3	0.092	0.044	0.032	0.056	0.034	0.070	0.906	0.866	0.950	0.994
	r=0.5	0.098	0.046	0.030	0.064	0.058	0.088	0.964	0.900	0.988	1.000
	r=0.8	0.116	0.052	0.102	0.066	0.052	0.090	0.800	0.626	0.972	1.000
$n_2 = 2000$	r=0.3	0.100	0.046	0.068	0.046	0.062	0.090	0.994	1.000	0.998	1.000
	r=0.5	0.110	0.050	0.080	0.064	0.044	0.094	0.998	1.000	1.000	1.000
	r=0.8	0.138	0.048	0.044	0.050	0.030	0.092	0.990	0.974	1.000	1.000

B.2. Error Quantities

Let $\rho_1 = |\text{corr}_*(\widehat{G}_{11} - G_{11}, \widehat{D}_{11})|$, $\rho_2 = |\text{corr}_*(G_{11}, \widehat{D}_{11})||\mathbb{E}_*(\widehat{G}_{11} - G_{11})|/\|\widehat{G}_{11} - G_{11}\|_{1,*}$ and $\rho_3 = |\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})|$. As discussed in Section 4, to control the Type I error asymptotically, the ρ_1 and ρ_2 should be sufficiently small. To guarantee the consistency of the test, we require the ρ_3 to be small enough so it does not destroy the signal in the data. Since it is hard to verify if these quantities converges theoretically because of the complex interactions of the variables in the simulation studies, we provide the empirical estimation of the quantities for the LL and KLR, see Figure 2. Note that $\rho_{1,LL}$ denotes the empirical estimate of ρ_1 with the estimators obtained by the LL. Other quantities are defined similarly.

In Model A, the LL leads to smaller estimation error compared to the KLR method, which makes sense since the LL is the correctly specified parametric method. The error quantities for both LL and KLR decreases to nearly zero as the sample size increases, which provides empirical support for the assumptions. In Model B, though the LL and KLR are not correct models, they perform decently. Under the alternative, the $\widehat{\rho}_{3,LL}$ converges slower than $\widehat{\rho}_{3,KLR}$, which agrees with the power performance that the power of the LL is a bit lower than the KLR for the large sample size. In Model C, the $\widehat{\rho}_{1,LL}$ is larger than $\widehat{\rho}_{1,KLR}$ for large sample sizes, which explains why the empirical Type I error of the LL is large. Under the alternative of Model C, $\widehat{\rho}_{3,LL}$ is not convergent, and the error seems too large that it destroys the signal. It explains the phenomenon that the LL barely shows any power in this case.

Appendix C: Auxiliary Lemmas

In the rest of the Appendix we provide proofs of main results and related auxiliary lemmas. The notation will mostly follow the main text, although some proofs involve their own notation. Throughout the proofs, we use C to denote a constant whose value only depends on (P_1, P_2) but not the sample sizes (n_1, n_2). The value of C may also change from line to line.

Our first auxiliary lemma provides an analogous CDF transformation for discrete random variable.

Lemma 3. Let V be a discrete random variable with finite support. Let $F(\cdot)$ be the CDF of V, with $F_-(\cdot)$ being its left limit. Let $p(\cdot)$ be the corresponding probability mass function. Let ζ be an independent $U(0,1)$ random variable. Then $U = F_-(V) + \zeta p(V) \sim U(0,1)$.

Proof of Lemma 3. Let $v_1 < \cdots < v_s$ be the support of V. Let $q_0 = 0$, $q_j = p(v_1) + \cdots + p(v_j)$ for $j = 1, \ldots, s$. Then $q_s = 1$. For $1 \le j \le s$, it is direct to verify that the density of U on (q_{j-1}, q_j) is a constant. Moreover, the length of this interval is $q_j - q_{j-1} = p(v_j)$, which is the same as the probability of U falling in (q_{j-1}, q_j). So the density of U on this interval is 1. Since this holds for each j, we conclude that $U \sim U(0,1)$. □

The next lemma is useful in establishing separation between H_0 and H_1 when the correct V function is used. Lemma 4 provides a "change-of-variable" trick to simplify an integral involved in calculating $\mathbb{E}U$, the expected value of the conformal p-value given in (8).

Lemma 4. Let $(X_j, Y_j) \sim P_j$ be independent for $j = 1, 2$, and (X_2', Y_2') be another independent draw from P_2. Let $v(x,y) = \frac{f_1(y|x)}{f_2(y|x)}$ and $g(x) = f_{2,X}(x)/f_{1,X}(x)$. Define $V_1 = v(X_1, Y_1)$, $V_2 = v(X_2, Y_2)$, and $V_2' = v(X_2', Y_2')$. Let $\widehat{v} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be an arbitrary nonrandom function and define $\widehat{V}_1$, $\widehat{V}_2$, $\widehat{V}_2'$ similarly as V_1, V_2, V_2' using $\widehat{v}$. We have

$$\mathbb{E}g(X_1)\mathbb{1}(\widehat{V}_1 < \widehat{V}_2) = \mathbb{E}V_2'\mathbb{1}(\widehat{V}_2' < \widehat{V}_2)$$

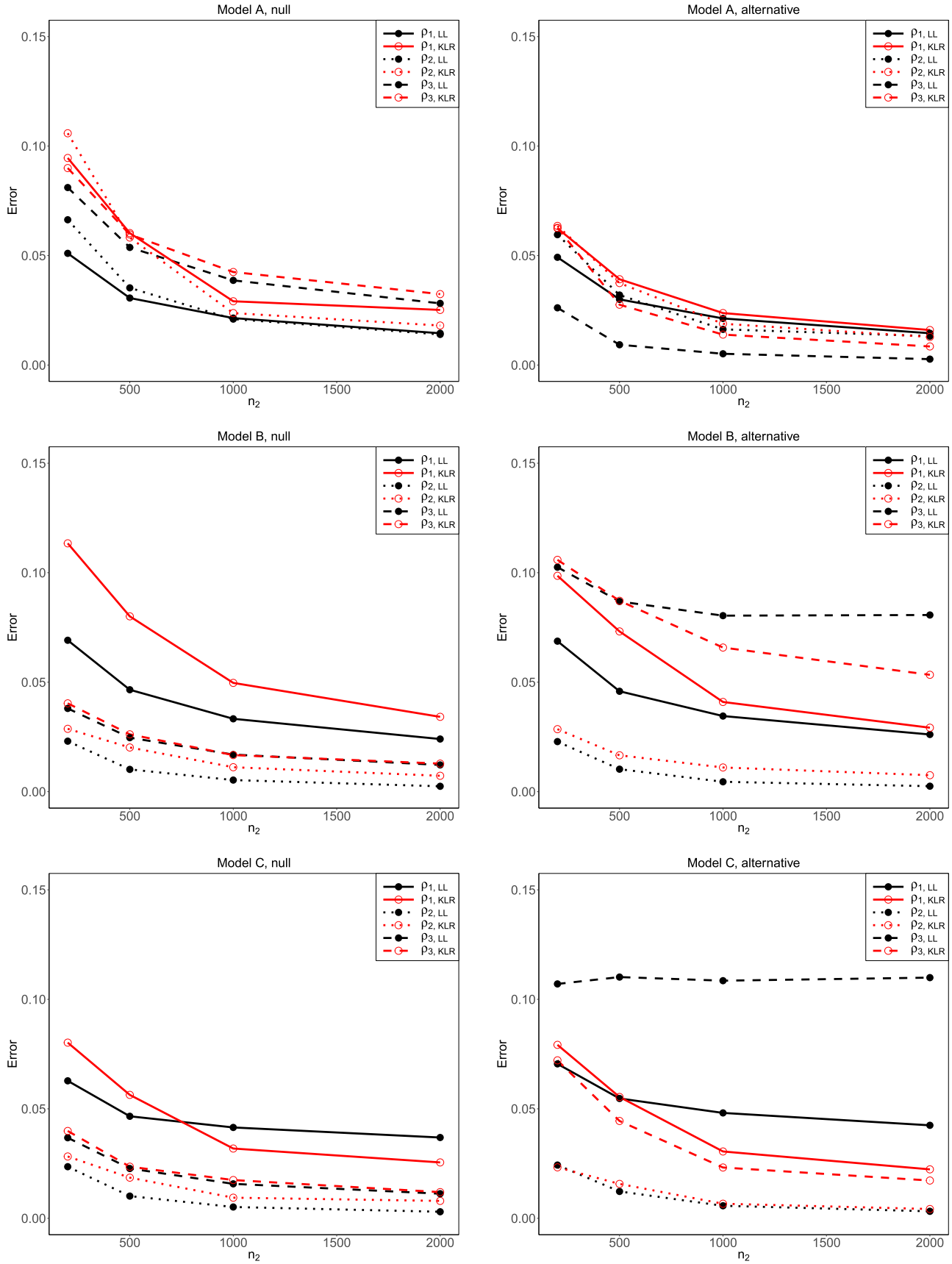

Figure 2. Error quantities under LL and KLR for all models with $p = 500, s = 5$ over 500 repetitions with $\alpha = 0.05$ and the split ratio $r = 0.5$.

and

$$\mathbb{E}g(X_1)\mathbb{1}(\widehat{V}_1 \le \widehat{V}_2) = \mathbb{E}V_2'\mathbb{1}(\widehat{V}_2' \le \widehat{V}_2)\,.$$

Proof. By definition,

$$\begin{aligned}
&\mathbb{E}g(X_1)\mathbb{1}(\widehat{V}_1 < \widehat{V}_2)\\
&= \mathbb{E}\left\{\frac{f_2(X_1,Y_1)}{f_1(X_1,Y_1)}\frac{f_1(Y_1|X_1)}{f_2(Y_1|X_1)}\mathbb{1}(\widehat{V}_1 < \widehat{V}_2)\right\}\\
&= \int f_2(x_2,y_2)\int_{\widehat{v}(x_1,y_1)<\widehat{v}(x_2,y_2)} f_2(x_1,y_1)v(x_1,y_1)dx_1dy_1dx_2dy_2\\
&= \mathbb{E}\{V_2'\mathbb{1}(\widehat{V}_2' < \widehat{V}_2)\}\,.
\end{aligned}$$

The proof of the second equation is identical. □

Lemma 5. Let $\widehat{\sigma}^2 = \widehat{\sigma}_1^2 + n_{11}/(12n_{21}) + \widehat{\sigma}_2^2/4 - \widehat{\rho}_{12}$ be defined as in (10), and $\widetilde{\sigma}^2 = \widetilde{\sigma}_1^2 + n_{11}/(12n_{21}) + \sigma_2^2/4 - \widetilde{\rho}_{12}$ where $\widetilde{\sigma}_1^2 = \mathrm{var}_*[G_{11}\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\}]$, $\sigma_2^2 = \mathrm{var}(G_{11})$ and $\widetilde{\rho}_{12} = \mathrm{cov}_*(G_{11}\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\}, G_{11})$. Then we have $\widetilde{\sigma}/\widehat{\sigma} = 1 + o_p(1)$ under Assumptions 1, 2.

Proof. Recall that $\widehat{F}$ is the CDF of $\widehat{v}(X_2,Y_2)$ by treating $\widehat{v}$ as fixed. Let $\widehat{F}_n$ be the empirical CDF of $\widehat{v}(X_2,Y_2)$ from the ranking sample. Because the ranking sample and $\widehat{v}$ are independent, we have

$$\|\widehat{F}_n - \widehat{F}\|_\infty = O_P(n_{21}^{-1/2})\,.$$

Consider the estimate

$$\begin{aligned}
\widehat{\sigma}_1^2 = &\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}^2\left\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\right\}^2\\
&-\left[\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\}\right]^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\widehat{\sigma}_1^2 - \widetilde{\sigma}_1^2\\
&= \frac{1}{n_{11}}\sum_i \widehat{G}_{1i}^2\left\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\right\}^2 - \mathbb{E}_*\left[G_{11}^2\left\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\right\}^2\right]\\
&\quad + \mathbb{E}_*^2[G_{11}\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\}] - \left[\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\}\right]^2\\
&= I + II.
\end{aligned}$$

To control the term I, we have

$$\begin{aligned}
I = &\frac{1}{n_{11}}\sum_i (\widehat{G}_{1i}^2 - G_{1i}^2)\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\}^2\\
&+\frac{1}{n_{11}}\sum_i G_{1i}^2\left[\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\}^2 - \{1-\widehat{F}_{1/2}(\widehat{V}_{1i})\}^2\right]\\
&+\frac{1}{n_{11}}\sum_i G_{1i}^2\{1-\widehat{F}_{1/2}(\widehat{V}_{1i})\}^2 - \mathbb{E}_*\left[G_{11}^2\left\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\right\}^2\right]\\
= &I_1 + I_2 + I_3.
\end{aligned}$$

Due to the fact that $\{1-\widehat{F}_{n,1/2}(\widehat{V}_{1i})\}^2 \le 1$ and by Assumption 2(a),

$$|I_1| \le \frac{1}{n_{11}}\sum_i |G_{1i}^2 - \widehat{G}_{1i}^2| = o_P(1)\,.$$

To control the term I_2, we have

$$|I_2| \le \frac{2\|\widehat{F}_n - \widehat{F}\|_\infty}{n_{11}}\sum_i G_{1i}^2 = O_P(n_{21}^{-1/2}).$$

Controlling the term I_3 follows from the Weak Law of Large Numbers (WLLN), which gives $|I_3| = o_P(1)$. Putting these pieces together, we obtain $I = o_P(1)$. In a similar way, we conclude that $II = o_P(1)$.

Analogously, we could establish that $\sigma_2^2 - \widehat{\sigma}_2^2 = o_P(1)$ and $\widetilde{\rho}_{12} - \widehat{\rho}_{12} = o_P(1)$. We complete the proof by using the continuous mapping theorem. □

Lemma 6. Let $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$ be defined as in (10) and Lemma 5. Let $\sigma^2 = \sigma_1^2 + n_{11}/(12n_{21}) + \sigma_2^2/4 - \rho_{12}$ be the ideal version of $\widetilde{\sigma}^2$ using the true conditional density ratio v. If $|\mathbb{E}_*G_{11}(\widehat{D}_{11} - D_{11})| = o_P(n_{11}^{-1/2})$, then we have $\sigma/\widehat{\sigma} = 1 + o_p(1)$.

Proof. According to Lemma 5 and Slutsky's theorem, it suffices to prove that $\sigma/\widetilde{\sigma} = 1 + o_P(1)$.

Note that

$$\begin{aligned}
\sigma_1^2 - \widetilde{\sigma}_1^2 =&\mathbb{E}_*G_{11}^2\left[\{1-F_{1/2}(V_{11})\}^2 - \{1-\widehat{F}_{1/2}(\widehat{V}_{11})\}^2\right]\\
&+\mathbb{E}_*^2G_{11}\{1-\widehat{F}_{1/2}(\widehat{V}_{11})\} - \mathbb{E}^2G_{11}\{1-F_{1/2}(V_{11})\}\\
=&I + II.
\end{aligned}$$

To bound the terms I and II, we have

$$|I| \le 2\mathbb{E}_*G_{11}^2|F_{1/2}(V_{11}) - \widehat{F}_{1/2}(\widehat{V}_{11})|\,,$$

$$|II| \le 2\mathbb{E}G_{11}\mathbb{E}_*G_{11}|F_{1/2}(V_{11}) - \widehat{F}_{1/2}(\widehat{V}_{11})|\,.$$

An application of Hölder's inequality implies that $|I| = o_P(1)$ and $II = o_P(1)$. Analogously, we obtain $\widetilde{\rho}_{12} - \rho_{12} = o_P(1)$, which completes the proof. □

Lemma 7. Let $U_j' = \left(n_{11}^{-1}\sum_{i=1}^{n_{11}} G_{1i}\widehat{D}_{ij}\right)/\left(n_{11}^{-1}\sum_{i=1}^{n_{11}} G_{1i}\right)$, and

$$T' = \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_{j=1}^{n_{21}} U_j'}{\widehat{\sigma}/\sqrt{n_{11}}}\,.$$

Then, under Assumptions 1, 2,

$$\widehat{T} - T' = \frac{\mathbb{E}_*(\widehat{G}_{11} - G_{11})\widehat{D}_{11} - \mathbb{E}_*(\widehat{G}_{11} - G_{11})\mathbb{E}_*(G_{11}\widehat{D}_{11})}{\widehat{\sigma}/\sqrt{n_{11}}} + o_P(1)\,,$$

where the expectation is taken over the ranking subsample while $\widehat{g}$ and $\widehat{v}$ are treated as fixed. Moreover, we have $T' \rightsquigarrow N(0,1)$ as $n_{11} \to \infty$ under H_0.

Proof of Lemma 7. Note that

$$\widehat{U}_j - U_j' = \frac{\frac{1}{n_{11}}\sum_i(\widehat{G}_{1i} - G_{1i})\widehat{D}_{ij}}{\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}} + \frac{\frac{1}{n_{11}}\sum_i G_{1i}\widehat{D}_{ij}}{\frac{1}{n_{11}}\sum_i G_{1i}}\left(\frac{\frac{1}{n_{11}}\sum_i G_{1i}}{\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}} - 1\right).$$

By law of large numbers, we have $n_{11}^{-1}\sum_i G_{1i} = 1 + o_P(1)$. Under Assumption 2(a), we obtain $n_{11}^{-1}\sum_i \widehat{G}_{1i} = 1 + o_P(1)$ and $n_{11}^{-1}\sum_i(G_{1i} - \widehat{G}_{1i}) = \mathbb{E}_*(G_{11} - \widehat{G}_{11}) + o_P(n_{11}^{-1/2})$. Since $|\widehat{D}_{ij}| \le 1$ and $|n_{21}^{-1}\sum_j \widehat{D}_{ij}| \le 1$, we have $(n_{11}n_{21})^{-1}\sum_{i,j}(\widehat{G}_{1i} - G_{1i})\widehat{D}_{ij} = \mathbb{E}_*(\widehat{G}_{11} - G_{11})\widehat{D}_{11} + o_P(n_{11}^{-1/2})$ and $n_{11}^{-1}\sum_i G_{1i}\widehat{D}_{ij} = \mathbb{E}_*G_{11}\widehat{D}_{11} + O_P(n_{11}^{-1/2})$.

Thus, by continuous mapping theorem,

$$\begin{aligned}
\frac{1}{n_{21}}\sum_j(\widehat{U}_j - U_j') = &\frac{\frac{1}{n_{11}n_{21}}\sum_{i,j}(\widehat{G}_{1i} - G_{1i})\widehat{D}_{ij}}{\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}}\\
&+\frac{\frac{1}{n_{11}n_{21}}\sum_{i,j} G_{1i}\widehat{D}_{ij}}{\frac{1}{n_{11}}\sum_i G_{1i}}\left(\frac{\frac{1}{n_{11}}\sum_i G_{1i}}{\frac{1}{n_{11}}\sum_i \widehat{G}_{1i}} - 1\right)\\
= &\left[\mathbb{E}_*(\widehat{G}_{11} - G_{11})\widehat{D}_{11} + o_P(n_{11}^{-1/2})\right](1 + o_P(1))
\end{aligned}$$

$$+\left[\mathbb{E}_* G_{11}\widehat{D}_{11} + O_P(n_{11}^{-1/2})\right]$$
$$\left[\mathbb{E}_*(G_{11}-\widehat{G}_{11}) + o_P(n_{11}^{-1/2})\right](1+o_P(1)).$$

Then we have

$$\widehat{T}-T' = \frac{\mathbb{E}_*(\widehat{G}_{11}-G_{11})\widehat{D}_{11} - \mathbb{E}_*(\widehat{G}_{11}-G_{11})\mathbb{E}_*(G_{11}\widehat{D}_{11})}{\widehat{\sigma}/\sqrt{n_{11}}} + o_P(1)\,.$$

Next, we prove $T' \rightsquigarrow N(0,1)$ as $n_{11}\to\infty$ under H_0. Define

$$T'' = \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_{j=1}^{n_{21}} U'_j}{\widetilde{\sigma}/\sqrt{n_{11}}}\,,$$

where $\widetilde{\sigma}^2$ is defined as in Lemma 5.

Recall that $\widehat{F}$ is the conditional CDF of $\widehat{V}_{21}$ given $\widehat{\nu}$, $\widehat{F}_-$ its left limit, and $\widehat{F}_\zeta = (1-\zeta)\widehat{F}_- + \zeta\widehat{F}$ for $\zeta\in[0,1]$.

Using the marginal projection of a two sample U-statistic, we have

$$\frac{1}{n_{11}n_{21}}\sum_{i,j} G_{1i}\widehat{D}_{ij} = \frac{1}{n_{11}}\sum_{i=1}^{n_{11}}\widehat{H}_{1i} + \frac{1}{n_{21}}\sum_{j=1}^{n_{21}}\widehat{H}_{2j} + \frac{1}{2} + R \qquad (12)$$

where

$$\widehat{H}_{1i} = \mathbb{E}_*\left[G_{1i}\widehat{D}_{ij}|(X_{1i},Y_{1i})\right] - 1/2 = g(X_{1i})\{1-\widehat{F}_{1/2}(\widehat{V}_{1i})\} - 1/2\,,$$
$$\widehat{H}_{2j} = \mathbb{E}_*\left[G_{1i}\widehat{D}_{ij}|(X_{2j},Y_{2j}),\zeta_j\right] - 1/2 = \widehat{F}_{\zeta_j}(\widehat{V}_{2j}) - 1/2\,,$$

and R is a remainder term to make (12) hold, which satisfies

$$\mathbb{E}_* R^2 = \frac{1}{n_{11}^2 n_{21}^2}\sum_{i,j,i',j'}\mathbb{E}_*\widetilde{H}_{ij}\widetilde{H}_{i'j'}\,,$$

where $\widetilde{H}_{ij} = G_{1i}\widehat{D}_{ij} - \widehat{H}_{1i} - \widehat{H}_{2j} - 1/2$. By construction, $\mathbb{E}_*\widetilde{H}_{ij}\widetilde{H}_{i'j'}$ is nonzero only if $(i,j) = (i',j')$, and $\mathbb{E}_* R^2$ reduces to

$$\mathbb{E}_* R^2 = \frac{1}{n_{11}^2 n_{21}^2}\sum_{i,j}\mathbb{E}_*\widetilde{H}_{ij}^2 = O((n_{11}n_{21})^{-1})\,.$$

Now we can write the main part of the numerator of T'' as

$$\frac{1}{n_{21}}\sum_{j=1}^{n_{21}} U'_j = \frac{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}}\widehat{H}_{1i} + \frac{1}{n_{21}}\sum_{j=1}^{n_{21}}\widehat{H}_{2j} + \frac{1}{2} + R}{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}} G_{1i}}\,.$$

Now apply the Lindeberg-Feller CLT to the triangular array $\{(\widehat{H}_{1i}, G_{1i}), 1\le i\le n_{11}, \widehat{H}_{2j} : 1\le j\le n_{21}\}$ indexed by n_{11}, combining with the delta method, we have for any $\widehat{\nu}$,

$$T''|\widehat{\nu} \rightsquigarrow N(0,1)\,.$$

Then, for any bounded continuous function f and $Z\sim N(0,1)$, we have $\mathbb{E}_* f(T'') \xrightarrow{a.s.} \mathbb{E}f(Z)$. According to the bounded convergence theorem, $\mathbb{E}f(T'')\to\mathbb{E}f(Z)$. Thus, we conclude that $T''\rightsquigarrow N(0,1)$ as $n_{11}\to\infty$. Combining with Lemma 5 and Slutsky's theorem yields $T'\rightsquigarrow N(0,1)$ as $n_{11}\to\infty$ under H_0. □

Appendix D: Proofs of Main Results

Proof of Lemma 2. For part (a), the proof is a standard application of the Bayes theorem, and is implicitly given in (Tibshirani et al. 2019, eq. (6) and sec. 3.2). Here we provide the calculation for the readers' convenience.

Recall that given $\widetilde{\mathbf{Z}}$, we have $V_{n_{11}+1} = \nu(\widetilde{X}_i, \widetilde{Y}_i)$ if $\sigma(n_{11}+1) = i$. Thus,

$$(V_{n_{11}+1}|\widetilde{\mathbf{Z}}) \sim \sum_\sigma P(\sigma|\widetilde{\mathbf{Z}})\delta_{\nu(\widetilde{X}_{\sigma(n_{11}+1)},\widetilde{Y}_{\sigma(n_{11}+1)})}$$
$$= \sum_{i=1}^{n_{11}+1} P\left[\sigma(n_{11}+1)=i|\widetilde{\mathbf{Z}}\right]\delta_{\nu(\widetilde{X}_i,\widetilde{Y}_i)}\,. \qquad (13)$$

Using Bayes rule,

$$P\left[\sigma(n_{11}+1)=i|\widetilde{\mathbf{Z}}\right]$$
$$= \frac{\sum_{\sigma(n_{11}+1)=i}\Pi_{l=1}^{n_{11}+1} f_{1,X}(\widetilde{X}_l)f(\widetilde{Y}_l|\widetilde{X}_l)\frac{f_{2,X}(\widetilde{X}_i)}{f_{1,X}(\widetilde{X}_i)}}{\sum_{j=1}^{n_{11}+1}\sum_{\sigma(n_{11}+1)=j}\Pi_{l=1}^{n_{11}+1} f_{1,X}(\widetilde{X}_l)f(\widetilde{Y}_l|\widetilde{X}_l)\frac{f_{2,X}(\widetilde{X}_j)}{f_{1,X}(\widetilde{X}_j)}}$$
$$= \frac{\frac{f_{2,X}(\widetilde{X}_i)}{f_{1,X}(\widetilde{X}_i)}}{\sum_{j=1}^{n_{11}+1}\frac{f_{2,X}(\widetilde{X}_j)}{f_{1,X}(\widetilde{X}_j)}} = p_i(\widetilde{\mathbf{Z}})\,.$$

As a result, (13) becomes

$$(V_{n_{11}+1}|\widetilde{\mathbf{Z}}) \sim \sum_{i=1}^{n_{11}+1} p_i(\widetilde{\mathbf{Z}})\delta_{\nu(\widetilde{X}_i,\widetilde{Y}_i)} = \sum_{i=1}^{n_{11}+1} p_i(\mathbf{Z})\delta_{\nu(X_i,Y_i)}\,.$$

Part (b) follows directly by combining part (a) with Lemma 3.

For part (c), we use notation $V_{1l} = \nu(X_l,Y_l) = f_1(Y_l|X_l)/f_2(Y_l|X_l)$, $l=1,\dots,n_{11}$, and $V_2 = \nu(X_{n_{11}+1},Y_{n_{11}+1}) = f_1(Y_{n_{11}+1}|X_{n_{11}+1})/f_2(Y_{n_{11}+1}|X_{n_{11}+1})$. Recall that $g(x) = f_{2,X}(x)/f_{1,X}(x)$.

For a given X_2, we consider

$$2\mathbb{E}_\zeta U = \frac{\sum_{l=1}^{n_{11}} g(X_l)[\mathbb{1}(V_{1l}<V_2) + \mathbb{1}(V_{1l}\le V_2)] + g(X_{n_{11}+1})}{\sum_{l=1}^{n_{11}} g(X_l) + g(X_{n_{11}+1})}\,,$$

where $\mathbb{E}_\zeta$ denotes expectation taken only over ζ, keeping everything else as given. Because $\mathbb{E}g(X_l) = 1$ for $1\le l\le n_{11}$, it can be directly verified from the strong law of large numbers on the iid random variables $X_1,\dots,X_{n_{11}}$ that

$$2\mathbb{E}_\zeta U \to \mathbb{E}\left[g(X_1)\mathbb{1}(V_{11}<V_2)\middle|X_{n_{11}+1},Y_{n_{11}+1}\right]$$
$$+\mathbb{E}\left[g(X_1)\mathbb{1}(V_{11}\le V_2)\middle|X_{n_{11}+1},Y_{n_{11}+1}\right]$$

a.e. over $(X_1,\dots,X_{n_{11}},X_{n_{11}+1})$ as $n_{11}\to\infty$. By construction we have $U\in[0,1]$. By the dominated convergence theorem, we have

$$2\mathbb{E}(U) \to \mathbb{E}g(X_{11})\mathbb{1}(V_{11}<V_2) + \mathbb{E}g(X_{11})\mathbb{1}(V_{11}\le V_2)\,. \qquad (14)$$

Lemma 4 implies that the right hand side of (14) is (letting V'_2 be an iid copy of V_2)

$$\mathbb{E}V'_2\mathbb{1}(V'_2<V_2) + \mathbb{E}V'_2\mathbb{1}(V'_2\le V_2)$$
$$= \mathbb{E}V_2\mathbb{1}(V_2<V'_2) + 1 - \mathbb{E}V'_2\mathbb{1}(V'_2>V_2)$$
$$= 1 - \left[\mathbb{E}V'_2\mathbb{1}(V'_2>V_2) - \mathbb{E}V_2\mathbb{1}(V'_2>V_2)\right]$$
$$= 1 - \mathbb{E}(V'_2-V_2)\mathbb{1}(V'_2>V_2)$$
$$= 1 - \frac{1}{2}\mathbb{E}|V'_2-V_2|\,,$$

which completes the proof. □

Proof of Theorem 1. Define

$$T' = \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_{j=1}^{n_{21}} U'_j}{\widehat{\sigma}/\sqrt{n_{11}}},$$

where $U'_j = \left(n_{11}^{-1}\sum_{i=1}^{n_{11}} G_{1i}\widehat{D}_{ij}\right) / \left(n_{11}^{-1}\sum_{i=1}^{n_{11}} G_{1i}\right)$ and $\widehat{D}_{ij} = \mathbb{1}(\widehat{V}_{1i} < \widehat{V}_{2j}) + \zeta_j \mathbb{1}(\widehat{V}_{1i} = \widehat{V}_{2j})$. Then Assumption 2(b) and Lemma 7 imply that $\widehat{T} = T' + \widehat{T} - T' \rightsquigarrow N(0,1)$ under H_0.

Next we prove the asymptotic power under the alternative. Note that, letting $Z_{k1} = (X_{k1}, Y_{k1})$ for $k = 1, 2$,

$$\begin{aligned}
&\mathrm{var}_*\left\{\frac{1}{n_{11}n_{21}}\sum_{i,j} G_{1i}(D_{ij} - \widehat{D}_{ij})\right\} \\
&= \left(\frac{1}{n_{11}}\mathrm{var}_*[\mathbb{E}_*\{G_{11}(D_{11} - \widehat{D}_{11})|Z_{11}\}]\right. \\
&\quad \left. + \frac{1}{n_{21}}\mathrm{var}_*[\mathbb{E}_*\{G_{11}(D_{11} - \widehat{D}_{11})|Z_{21}, \zeta_1\}]\right)(1 + o_P(1)) \\
&\le \left[\frac{C}{n_{11}}\mathrm{var}_*\{G_{11}(D_{11} - \widehat{D}_{11})\}\right](1 + o_P(1)), \qquad (15)
\end{aligned}$$

due to Assumption 1 and the boundedness of $\widehat{F}_\zeta$ and F_ζ for any $\zeta \in [0,1]$. Thus, we have

$$\begin{aligned}
\frac{1}{n_{21}}\sum_j (U_j - U'_j) &= \frac{\frac{1}{n_{11}n_{21}}\sum_{i,j} G_{1i}(D_{ij} - \widehat{D}_{ij})}{\frac{1}{n_{11}}\sum_i G_{1i}} \\
&= \{\mathbb{E}_* G_{11}(D_{11} - \widehat{D}_{11}) + O_P(1/\sqrt{n_{11}})\}(1 + o_P(1)).
\end{aligned}$$

Then,

$$\begin{aligned}
T' &= \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_j U'_j}{\widehat{\sigma}/\sqrt{n_{11}}} \\
&= \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_j U_j}{\widehat{\sigma}/\sqrt{n_{11}}} + \frac{\frac{1}{n_{21}}\sum_j (U_j - U'_j)}{\widehat{\sigma}/\sqrt{n_{11}}} \\
&= \frac{\frac{1}{2} - \frac{1}{n_{21}}\sum_j U_j}{\widehat{\sigma}/\sqrt{n_{11}}} - \frac{\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})}{\widehat{\sigma}/\sqrt{n_{11}}}(1 + o_P(1)) + O_P(1) \\
&= \frac{\frac{1}{2} - \mathbb{E}G_{11}D_{11}}{\widehat{\sigma}/\sqrt{n_{11}}} + \frac{\mathbb{E}G_{11}D_{11} - \frac{1}{n_{21}}\sum_j U_j}{\widehat{\sigma}/\sqrt{n_{11}}} \\
&\quad - \frac{\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})}{\widehat{\sigma}/\sqrt{n_{11}}}(1 + o_P(1)) + O_P(1) \\
&= \frac{\frac{1}{4}\delta}{\widehat{\sigma}/\sqrt{n_{11}}} + \frac{\mathbb{E}G_{11}D_{11} - \frac{1}{n_{21}}\sum_j U_j}{\sigma/\sqrt{n_{11}}}\frac{\sigma}{\widehat{\sigma}} \\
&\quad - \frac{\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})}{\widehat{\sigma}/\sqrt{n_{11}}}(1 + o_P(1)) + O_P(1), \qquad (16)
\end{aligned}$$

where $\delta = \mathbb{E}|v(X_2, Y_2) - v(X'_2, Y'_2)|$, and the last equality holds due to the fact that $\delta/4 = 1/2 - \mathbb{E}G_{11}D_{11}$ as explained in the proof of Lemma 2(c). Note that

$$\frac{\mathbb{E}G_{11}D_{11} - \frac{1}{n_{21}}\sum_j U_j}{\sigma/\sqrt{n_{11}}} \rightsquigarrow N(0,1)$$

and $\sigma/\widehat{\sigma} = O_P(1)$. If there exists a constant $c > 0$ such that

$$\mathbb{P}\left[\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11}) < (1/4)\mathbb{E}|v(X_2, Y_2) - v(X'_2, Y'_2)| - c\right] \to 1,$$

where (X_2, Y_2) and (X'_2, Y'_2) are iid realizations from P_2, then we have $T' \to \infty$ in probability. Together with $\widehat{T} - T' = o_P(1)$, we establish the desired result. □

Proof of Equation (11). First realize that

$$\begin{aligned}
&\mathbb{E}_{P_2} D_{\mathrm{tv}}(f_1(\cdot|X), f_2(\cdot|X)) \\
&= \frac{1}{2}\int\int |f_1(y|x) - f_2(y|x)|\, dy f_2(x) dx \\
&= \frac{1}{2}\int\int \left|\frac{f_1(y|x)}{f_2(y|x)} - 1\right| f_2(y|x) f_2(x) dy dx = \frac{1}{2}\mathbb{E}_{P_2}|v(X,Y) - 1|.
\end{aligned}$$

The lower bound follows from a conditional Jensen's inequality:

$$\mathbb{E}|v(X,Y) - 1| = \mathbb{E}\left|v(X,Y) - \mathbb{E}v(X',Y')\right| \le \mathbb{E}|v(X,Y) - v(X',Y')|,$$

where (X,Y), (X',Y') are iid copies from P_2.

The upper bound follows from triangle inequality:

$$\begin{aligned}
\mathbb{E}|v(X,Y) - v(X',Y')| &= \mathbb{E}\left|v(X,Y) - 1 - [v(X',Y') - 1)]\right| \\
&\le \mathbb{E}|v(X,Y) - 1| + \mathbb{E}|v(X',Y') - 1| = 2\mathbb{E}|v(X,Y) - 1|. \quad \square
\end{aligned}$$

Proof of Proposition 1. From (15), we have

$$\begin{aligned}
&\mathrm{var}_*\left\{\frac{1}{n_{11}n_{21}}\sum_{i,j} G_{1i}(D_{ij} - \widehat{D}_{ij})\right\} \\
&\le \left[\frac{C}{n_{11}}\mathrm{var}_*\{G_{11}(D_{11} - \widehat{D}_{11})\}\right](1 + o_P(1)).
\end{aligned}$$

Since $|D_{11} - \widehat{D}_{11}| \le 1$,

$$\mathrm{var}_*\{G_{11}(D_{11} - \widehat{D}_{11})\} \le \mathbb{E}_* G_{11}^2 (D_{11} - \widehat{D}_{11})^2 \le \mathbb{E}_* G_{11}^2 |D_{11} - \widehat{D}_{11}|.$$

Now we study $\widehat{D}_{11} - D_{11}$. Let $\xi = V_{21} - \widehat{V}_{21} - (V_{11} - \widehat{V}_{11})$. Then, for any $\epsilon > 0$

$$\begin{aligned}
&\left|\mathbb{1}(\widehat{V}_{11} < \widehat{V}_{21}) - \mathbb{1}(V_{11} < V_{21})\right| \\
&= \mathbb{1}(V_{11} + \xi < V_{21} \le V_{11},\ \xi < 0) \\
&\quad + \mathbb{1}(V_{11} < V_{21} \le V_{11} + \xi,\ \xi > 0) \\
&\le \mathbb{1}(V_{11} - |\xi| \le V_{21} \le V_{11} + |\xi|) \\
&\le \mathbb{1}(V_{11} - \epsilon \le V_{21} \le V_{11} + \epsilon) + \mathbb{1}(|\xi| \ge \epsilon).
\end{aligned}$$

The same upper bound holds for $\mathbb{1}(\widehat{V}_{11} \le \widehat{V}_{21}) - \mathbb{1}(V_{11} \le V_{21})$ using the same argument, and we conclude that

$$\begin{aligned}
|\widehat{D}_{11} - D_{11}| &= \left|(1 - \zeta_1)\left[\mathbb{1}(\widehat{V}_{11} < \widehat{V}_{21}) - \mathbb{1}(V_{11} < V_{21})\right]\right. \\
&\quad \left. + \zeta_1\left[\mathbb{1}(\widehat{V}_{11} \le \widehat{V}_{21}) - \mathbb{1}(V_{11} \le V_{21})\right]\right| \\
&\le (1 - \zeta_1)\left|\mathbb{1}(\widehat{V}_{11} < \widehat{V}_{21}) - \mathbb{1}(V_{11} < V_{21})\right| \\
&\quad + \zeta_1\left|\mathbb{1}(\widehat{V}_{11} \le \widehat{V}_{21}) - \mathbb{1}(V_{11} \le V_{21})\right| \\
&\le \mathbb{1}(V_{11} - \epsilon \le V_{21} \le V_{11} + \epsilon) + \mathbb{1}(|\xi| \ge \epsilon).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_* G_{11}^2 |\widehat{D}_{11} - D_{11}| \\
&\le \mathbb{E}_* G_{11}^2 \mathbb{1}(V_{11} - \epsilon \le V_2 \le V_{11} + \epsilon) + \mathbb{E}_* G_{11}^2 \mathbb{1}(|\xi| \ge \epsilon) \\
&\le \mathbb{E}_*\left\{G_{11}^2 \mathbb{E}_*[\mathbb{1}(V_{11} - \epsilon \le V_2 \le V_{11} + \epsilon)|X_{11}, Y_{11}]\right\} \\
&\quad + \mathbb{E}_* G_{11}^2 \mathbb{1}(|\xi| \ge \epsilon) \\
&\le 2C\epsilon\mathbb{E}G_{11}^2 + \mathbb{E}_* G_{11}^2 \mathbb{1}(|\xi| \ge \epsilon),
\end{aligned}$$

where the first term in the last inequality follows from the assumption of bounded density of V_{21} with the constant C being a finite upper bound of the density. Because $\xi = o_P(1)$ the support of $\mathbb{1}(|\xi| \ge \epsilon)$ has vanishing probability measure. The integrability of G_{11}^2 implies it is uniformly integrable. So we have $\mathbb{E}_* G_{11}^2 \mathbb{1}(|\xi| \ge \epsilon) = o_P(1)$ for arbitrary $\epsilon > 0$.

Therefore, we conclude that $\mathbb{E}_* G_{11}^2 |\widehat{D}_{11} - D_{11}| = o_P(1)$. Combining this variance bound with the mean bound $|\mathbb{E}_* G_{11}(\widehat{D}_{11} - D_{11})| = o_P(1/\sqrt{n_{11}})$ assumed in the proposition, we have

$$\frac{1}{n_{21}} \sum_j (U_j - U_j') = \frac{\frac{1}{n_{11} n_{21}} \sum_{i,j} G_{1i}(D_{ij} - \widehat{D}_{ij})}{\frac{1}{n_{11}} \sum_i G_{1i}} = o_P(1/\sqrt{n_{11}}).$$

As in the proof of Theorem 1, Assumption 2(b) and Lemma 7 imply that $\widehat{T} - T' = o_P(1)$. According to (16) and Lemma 6, we have

$$\begin{aligned} \widehat{T} &= \widehat{T} - T' + T' \\ &= \frac{\frac{1}{4}\delta}{\widehat{\sigma}/\sqrt{n_{11}}} + \frac{\mathbb{E} G_{11} D_{11} - \frac{1}{n_{21}} \sum_j U_j}{\sigma/\sqrt{n_{11}}} \frac{\sigma}{\widehat{\sigma}} \\ &\quad + \frac{\frac{1}{n_{21}} \sum_j (U_j - U_j')}{\widehat{\sigma}/\sqrt{n_{11}}} + o_P(1) \\ &= \frac{\sqrt{n_{11}}\delta}{4\sigma}(1 + o_P(1)) + Z + o_P(1), \end{aligned}$$

where $Z \rightsquigarrow N(0,1)$ as $n_{11} \to \infty$. □

Supplementary Materials

The supplementary materials contain data and code to conduct the experiments in the article.

Funding

Xiaoyu Hu's research is partially supported by a scholarship from the China Scholarship Council. Jing Lei's research is partially supported by NSF grant DMS-2015492.

References

Andrews, D. W. (1997), "A Conditional Kolmogorov Test," *Econometrica: Journal of the Econometric Society*, 65, 1097–1128. [1137]

Bai, J. (2003), "Testing Parametric Conditional Distributions of Dynamic Models," *Review of Economics and Statistics*, 85, 531–549. [1137]

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019), "Predictive Inference with the jackknife+," arXiv preprint arXiv:1905.02928. [1138]

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2021), "Testing for Outliers with Conformal *p*-values," arXiv preprint arXiv:2104.08279. [1137]

Bickel, S., Brückner, M., and Scheffer, T. (2007), "Discriminative learning for differing training and test distributions," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 81–88. [1141]

——— (2009), "Discriminative Learning Under Covariate Shift," *Journal of Machine Learning Research*, 10, 2137–2155. [1136]

Bühlmann, P. (2020), "Invariance, Causality and Robustness," *Statistical Science*, 35, 404–426. [1137]

Cheng, K. F., and Chu, C.-K. (2004), "Semiparametric Density Estimation Under a Two-Sample Density Ratio Model," *Bernoulli*, 10, 583–604. [1141]

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021), "Distributional Conformal Prediction," *Proceedings of the National Academy of Sciences*, 118, e2107794118. [1138]

Corradi, V., and Swanson, N. R. (2006), "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification," *Journal of Econometrics*, 133, 779–806. [1137]

Csurka, G. (2017), *Domain Adaptation in Computer Vision Applications* (Vol. 2), Cham: Springer. [1136]

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020), "Exact Tests via Multiple Data Splitting," *Statistics & Probability Letters*, 166, 108865. [1147]

Dua, D., and Graff, C. (2019), "UCI Machine Learning Repository." [1145]

Fan, Y., Li, Q., and Min, I. (2006), "A Nonparametric Bootstrap Test of Conditional Distributions," *Econometric Theory*, 22, 587–613. [1137]

Fedorova, V., Gammerman, A., Nouretdinov, I., and Vovk, V. (2012), "Plug-in Martingales for Testing Exchangeability On-line," arXiv preprint arXiv:1204.3251. [1137]

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009), "Covariate Shift by Kernel Mean Matching," *Dataset Shift in Machine Learning*, eds. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, pp. 131–160, Cambridge, MA: MIT Press. [1136,1141]

Guan, L., and Tibshirani, R. (2019), "Prediction and Outlier Detection in Classification Problems," arXiv preprint arXiv:1905.04396. [1137]

Hall, P., and Hart, J. D. (1990), "Bootstrap Test for Difference between Means in Nonparametric Regression," *Journal of the American Statistical Association*, 85, 1039–1049. [1137]

Hardle, W., and Marron, J. S. (1990), "Semiparametric Comparison of Regression Curves," *The Annals of Statistics*, 18, 63–89. [1137]

Kanamori, T., Hido, S., and Sugiyama, M. (2009), "A Least-Squares Approach to Direct Importance Estimation," *Journal of Machine Learning Research*, 10, 1391–1445. [1141]

Kim, B., Xu, C., and Barber, R. F. (2020), "Predictive Inference is Free with the Jackknife+-after-Bootstrap," arXiv preprint arXiv:2002.09025. [1147]

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020), "Adaptive, Distribution-Free Prediction Intervals for Deep Networks," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 4346–4356. [1138]

Kouw, W. M., and Loog, M. (2018), "An Introduction to Domain Adaptation and Transfer Learning," arXiv preprint arXiv:1812.11806. [1136]

Kuchibhotla, A. K., and Ramdas, A. K. (2019), "Nested Conformal Prediction and the Generalized Jackknife+," arXiv preprint arXiv:1910.10562. [1147]

Kulasekera, K. (1995), "Comparison of Regression Curves Using Quasi-Residuals," *Journal of the American Statistical Association*, 90, 1085–1093. [1137]

Kulasekera, K., and Wang, J. (1997), "Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves," *Journal of the American Statistical Association*, 92, 500–511. [1137]

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111. [1137,1138,1143]

Lei, J., Rinaldo, A., and Wasserman, L. (2015), "A Conformal Prediction Approach to Explore Functional Data," *Annals of Mathematics and Artificial Intelligence*, 74, 29–43. [1138]

Lei, J., Robins, J., and Wasserman, L. (2013), "Distribution-Free Prediction Sets," *Journal of the American Statistical Association*, 108, 278–287. [1137,1138]

Lei, J., and Wasserman, L. (2014), "Distribution-Free Prediction Bands for Non-parametric Regression," *Journal of the Royal Statistical Society*, Series B, 76, 71–96. [1137,1138]

Li, S., Sesia, M., Romano, Y., Candès, E., and Sabatti, C. (2021), "Searching for Robust Associations with a Multi-Environment Knockoff Filter," *Biometrika*, 109, 611–629. [1137]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [1147]

Neumeyer, N., and Dette, H. (2003), "Nonparametric Comparison of Regression Curves: An Empirical Process Approach," *The Annals of Statistics*, 31, 880–920. [1137]

Pan, S. J., and Yang, Q. (2009), "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345–1359. [1136]

Pardo-Fernández, J. C., Jiménez-Gamero, M. D., and El Ghouch, A. (2015), "Tests for the Equality of Conditional Variance Functions in Nonparametric Regression," *Electronic Journal of Statistics*, 9, 1826–1851. [1137]

Peters, J., Bühlmann, P., and Meinshausen, N. (2016), "Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 78, 947–1012. [1137]

Qin, J. (1998), "Inferences for Case-Control and Semiparametric Two-Sample Density Ratio Models," *Biometrika*, 85, 619–630. [1141]

Rinaldo, A., Wasserman, L., G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438–3469. [1147]

Romano, Y., Patterson, E., and Candes, E. (2019), "Conformalized Quantile Regression," in *Advances in Neural Information Processing Systems* (Vol. 32). [1138]

Sesia, M., and Romano, Y. (2021), "Conformal Prediction using Conditional Histograms," in *Advances in Neural Information Processing Systems* (Vol. 34). [1138]

Shimodaira, H. (2000), "Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function," *Journal of Statistical Planning and Inference*, 90, 227–244. [1136]

Sollich, P. (2000), "Probabilistic Methods for Support Vector Machines," in *Advances in Neural Information Processing Systems*, pp. 349–355. [1141]

Sugiyama, M., and Kawanabe, M. (2012), *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*, Cambridge, MA: MIT Press. [1136]

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007), "Covariate Shift Adaptation by Importance Weighted Cross Validation," *Journal of Machine Learning Research*, 8, 985–1005. [1136]

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008), "Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation," in *Advances in neural information processing systems*, pp. 1433–1440. [1136,1141]

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019), "Conformal Prediction Under Covariate Shift," in *Advances in Neural Information Processing Systems*, pp. 2526–2536. [1137,1139,1140,1145,1151]

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009), "Direct Density Ratio Estimation for Large-Scale Covariate Shift Adaptation," *Journal of Information Processing*, 17, 138–155. [1141]

Vovk, V. (2019), "Testing Randomness," arXiv preprint arXiv:1906.09256. [1137]

——— (2020), "Testing for Concept Shift Online," arXiv preprint arXiv:2012.14246. [1137]

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, New York: Springer. [1137,1138]

Vovk, V., Nouretdinov, I., and Gammerman, A. (2003), "Testing Exchangeability On-line," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 768–775. [1137]

Vovk, V., Petej, I., Nouretdinov, I., Ahlberg, E., Carlsson, L., and Gammerman, A. (2021), "Retrain or Not Retrain: Conformal Test Martingales for Change-Point Detection," in *Conformal and Probabilistic Prediction and Applications*, pp. 191–210, PMLR. [1137]

Wasserman, L., and Roeder, K. (2009), "High Dimensional Variable Selection," *Annals of Statistics*, 37, 2178–2201. [1147]

Zheng, J. X. (2000), "A Cconsistent Test of Conditional Parametric Distributions," *Econometric Theory*, 16, 667–691. [1137,1143]

Zhu, J., and Hastie, T. (2005), "Kernel Logistic Regression and the Import Vector Machine," *Journal of Computational and Graphical Statistics*, 14, 185–205. [1143]