A Novel Non-parametric Two-Sample Test on Imprecise Observations

Feng Liu, Guangquan Zhang and Jie Lu

Decision System&e-Service Intelligence (DeSI) Lab
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology
University of Technology Sydney, Sydney, Australia
{*Feng.Liu, Guangquan.Zhang, Jie.Lu*}*@uts.edu.au*

***Abstract*—In kernel non-parametric two-sample test, we aim to determine whether two sets of *precise* observations (i.e., samples) are from the same distribution based on a selected kernel. However, in real world, precise observations may be unavailable. For example, readings on an analogue measurement equipment are not precise numbers but intervals since there is only a finite number of decimals available. Hence, we consider a new and more realistic problem setting—*two-sample test on imprecise observations*. We show that the test power of existing kernel two-sample tests will drop significantly if they do not take care of the vagueness of the imprecise observations, and to this end, we propose a *fuzzy-based maximum mean discrepancy* (F-MMD), a powerful two-sample test on imprecise observations. F-MMD is based on a novel fuzzy-based kernel function that can measure the discrepancy between two imprecise observations. This novel kernel function takes care of the vagueness of the imprecise observations and its parameters are optimized to maximize the approximate test power of F-MMD. Experiments demonstrate that F-MMD significantly outperforms competitive two-sample test methods when facing imprecise observations.**

***Index Terms*—Hypothesis test, kernel, imprecise observations**

I. Introduction

Two-sample test methods aim to determine whether two sets of samples are drawn from the same distribution [1]–[8]. Traditional parametric two-sample test methods such as t-tests are mainstays of statistical applications, but require strong parametric assumptions about the distributions being studied and/or are only effective on data in extremely low-dimensional spaces [9]. By contrast, non-parametric test methods [10]–[13] make only mild assumptions about distributions, and thus are far more broadly applicable. For this reason, non-parametric test methods exhibit the greatest potential in real world [3].

Kernel-based two-sample test methods [2]–[5], as well-known non-parametric two-sample tests, have been well studied in the last decade. Researchers constructed representative features of two distributions using kernel mean embeddings (ME) [2], [3] and smooth characteristic functions (SCF) [3], [4]. The discrepancy between two distributions can be ascertained by comparing these features [3], [14], [15]. *Maximum mean discrepancy* (MMD) is the most common statistic used to compare two distributions. Its contribution in such fields as domain adaptation [16]–[18], concept drift [19], [20] and generative adversarial networks [21], [22] is significant.

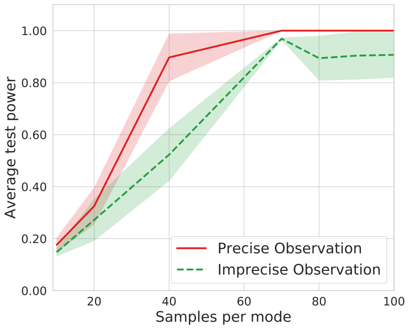


Fig. 1. Average test power of the MMD test when facing precise observations (red line) and imprecise observations (green dash line). Clearly, the test power of MMD will drop significantly if we do not take care of vagueness of the imprecise observations.

However, existing kernel two-sample test methods all share the assumption that observations from distributions are *precise*, which is not realistic in the real world. For example, readings on an analogue measurement equipment are not precise numbers but intervals since there is only a finite number of decimals available [23], [24]. This motivates us to consider a more realistic problem, named as *two-sample test on imprecise observations*. In this problem, we can only obtain imprecise observations (the precise observations are unavailable) and have to determine if two sets of imprecise observations are from the same distribution.

To validate the effects brought from such imprecise observations, in Figure 1, we show the test power of MMD when facing precise and imprecise observations. Following [5], MMD used in this figure has been optimized (i.e., MMD with the best kernel) and has the highest approximate test power. Clearly, the test power of MMD drops significantly when we can only obtain the imprecise observations. Namely, MMD, as a most common two-sample test method, cannot handle this new problem well even when we feed the best kernel to it.

To address this new problem, we propose a *fuzzy-based maximum mean discrepancy* (F-MMD), a powerful two-sample test on imprecise observations. F-MMD is based on a novel fuzzy-based kernel function that can measure the discrepancy between two imprecise observations (modeled by the fuzzy vector). This novel kernel function takes care of the

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

vagueness of the imprecise observations and its parameters are optimized to maximize the approximate test power of F-MMD.

The main contributions of this paper is listed as follows.

- This paper presents a new problem setting in the field of two-sample test and empirically shows existing kernel two-sample test methods cannot handle this problem well;
- We propose a novel two-sample test method, called *fuzzy-based maximum mean discrepancy* (F-MMD), which can take care of the vagueness of the imprecise observations and have a higher test power than existing methods.

We test our method on the Blob dataset that is commonly used in the field of two-sample test [2], [3], [5]. We find that our method has higher average test power compared to competitive baselines.

II. Preliminary

In this section, we introduce five concepts related to kernel two-sample test, which includes two-sample test problem, definition of *maximum mean discrepancy* (MMD), Asymptotics of MMD, test power of MMD and how to maximize test power of MMD.

A. Two-sample test

Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $\mathcal{P}$, $\mathcal{Q}$ be Borel probability measures on $\mathcal{X}$. Given independent identically distributed (*i.i.d.*) samples $S_X = \{x_i\}_{i=1}^n \sim \mathcal{P}^n$ and $S_Y = \{y_j\}_{j=1}^m \sim \mathcal{Q}^m$, in two-sample test problem, we aim to determine if S_X and S_Y come from the same distribution, i.e., if $\mathcal{P} = \mathcal{Q}$.

In two-sample test problem, we consider two hypothesis, where the null hypothesis $H_0 : \mathcal{P} = \mathcal{Q}$ is tested against the alternative hypothesis $H_1 : \mathcal{P} \neq \mathcal{Q}$. Generally, a two-sample test method is performed in four steps: 1) confirm a significance level $\alpha \in [0, 1]$; 2) calculate a test statistic $\hat{t}(S_X, S_Y)$; 3) calculate the p-value $\hat{p} = \Pr_{H_0}(T > \hat{t}(S_X, S_Y))$, the probability of the two-sample test method returning a T as large as $\hat{t}(S_X, S_Y)$ when H_0 is true; 4) reject H_0 if $\hat{p} < \alpha$.

B. Maximum mean discrepancy

MMD is an estimator of a distance between distributions, which is called integral probability metrics:

Definition 1 (Integral Probability Metrics [25]). *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. The integral probability metric based on $\mathcal{F}$ is*

$$\mathbb{D}(\mathcal{P}, \mathcal{Q}; \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim \mathcal{P}}[f(X)] - \mathbb{E}_{Y \sim \mathcal{Q}}[f(Y)]|, \tag{1}$$

where $X \sim \mathcal{P}$ and $Y \sim \mathcal{Q}$ are random variables on $\mathcal{X}$.

This class of metrics has been well studied in probability theory [26], [27]. A simple (but biased) estimator of Eq. (1) based on samples S_X and S_Y is

$$\widehat{\mathbb{D}}(S_X, S_Y; \mathcal{F}) = \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_i f(x_i) - \frac{1}{m} \sum_j f(y_j) \Big|.$$

However, the sup can not be well estimated, i.e., we cannot directly use this biased estimation to calculate the distance between $\mathcal{P}$ and $\mathcal{Q}$ through samples (i.e., $S_\mathcal{P}$ and $S_\mathcal{Q}$) drawn from them. To overcome this issue, taking the advantage of *reproducing kernel Hilbert space* (RKHS), *maxiximum mean discrepancy* (MMD) is proposed by taking $\mathcal{F}$ to be a unit ball of a RKHS, which allows for efficient estimation of the IPM.

Definition 2 (MMD [2]). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel of a RKHS $\mathcal{H}_k$ (i.e., $|k(\cdot, \cdot)| < +\infty$). Let $X, X' \sim \mathcal{P}$ and $Y, Y' \sim \mathcal{Q}$ be random variables on $\mathcal{X}$, MMD is defined as follows.*

$$\mathrm{MMD}(\mathcal{P}, \mathcal{Q}; \mathcal{H}_k) := \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]|$$
$$= \|\mu_\mathcal{P} - \mu_\mathcal{Q}\|_{\mathcal{H}_k} = \sqrt{\mathbb{E}\left[k(X, X') + k(Y, Y') - 2k(X, Y)\right]}.$$

where $\mu_\mathcal{P} := \mathbb{E}[k(\cdot, X)]$ and $\mu_\mathcal{Q} := \mathbb{E}[k(\cdot, Y)]$ are meaning embeddings of $\mathcal{P}$ and $\mathcal{Q}$, respectively. If k is a characteristic kernel *and $\mu_\mathcal{P} = \mu_\mathcal{Q}$, we have $\mathrm{MMD}(\mathcal{P}, \mathcal{Q}; \mathcal{H}_k) = 0$ if and only if $\mathcal{P} = \mathcal{Q}$.*

We can estimate MMD using the U-statistic estimator, which is unbiased for MMD^2 and has nearly minimal variance among unbiased estimators [2]:

$$\widehat{\mathrm{MMD}}_u^2(S_X, S_Y; k) = \frac{1}{n(n-1)} \sum_{i \neq j} H_{ij} \tag{2}$$
$$H_{ij} = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j).$$

Based on H_{ij}, the biased estimator of the squared MMD is $\widehat{\mathrm{MMD}}_b^2 := \frac{1}{n^2} \sum_{ij} H_{ij}$. Compared to $\widehat{\mathrm{MMD}}_u$, $\widehat{\mathrm{MMD}}_b$ allows $m \neq n$ but is not convenient for analysis and to efficiently implement permutations.

C. Asymptotics of the MMD

Based on the unbiased estimator of MMD, we can analyze the asymptotics of the MMD under H_0 and H_1 as follows.

Proposition 1 (Asymptotics of $\widehat{\mathrm{MMD}}_u^2$). *Under the null hypothesis, $H_0 : \mathcal{P} = \mathcal{Q}$, $\widehat{\mathrm{MMD}}_u^2$ is $\mathcal{O}_P(1/n^2)$, with*

$$n\widehat{\mathrm{MMD}}_u^2 \xrightarrow{d} \sum_i \lambda_i (Z_i^2 - 2);$$

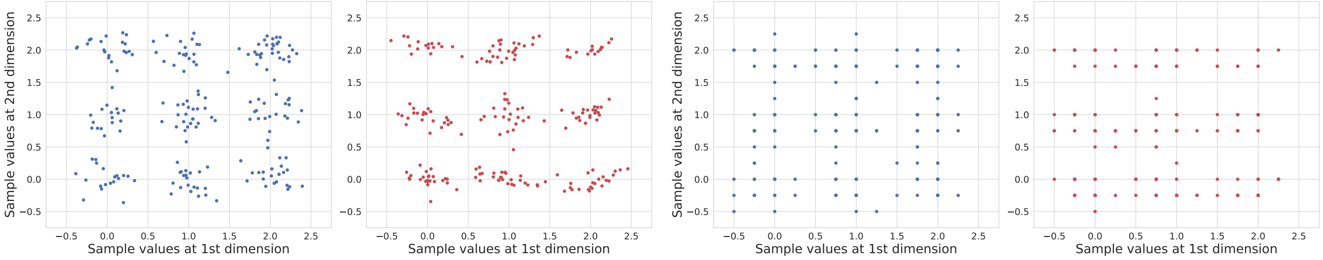
here λ_i are the eigenvalues of the covariance operator under $\mathcal{P}$ of the centered kernel, and the Z_i are i.i.d. Gaussian with mean 0 and variance 2 [2, Theorem 12].

Under the alternative, $H_1 : \mathcal{P} \neq \mathcal{Q}$, $\widehat{\mathrm{MMD}}_u^2$ is $\mathcal{O}_P(1/n)$, and in particular [28, Section 5.5.1]

$$\sqrt{n}(\widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2),$$

where $\sigma_{H_1}^2 = 4(\mathbb{E}_z[(\mathbb{E}_{z'} h(z, z'))^2] - [(\mathbb{E}_{z,z'} h(z, z'))^2]) = 4\left(\mathbb{E}[H_{12}H_{13}] - \mathbb{E}[H_{12}]^2\right)$, $h(z, z') = k(x, x') + k(y, y') - k(x, y') - k(x', y)$ and $z := (x, y)$.

Based on this Proposition (especially on asymptotics under H_1), we can analyze the test power of MMD (next subsection).



(a) Precise observations from $\mathcal{P}$ (left) and $\mathcal{Q}$ (right).

(b) Imprecise observations from $\mathcal{P}$ (left) and $\mathcal{Q}$ (right).

Fig. 2. Blob dataset with precise observations (a) and imprecise observations (b).

D. Test power of MMD

The main criterion of efficacy of a two-sample test method is its *power*: the probability that, for a particular $\mathcal{P} \neq \mathcal{Q}$ and n, we correctly reject H_0. Using Proposition 1, we have that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}_u^2 > r\right) \to \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{r}{\sqrt{n}\,\sigma_{H_1}}\right),$$

where Φ is the standard normal CDF. Via Proposition 1, We know that this r will not increase as increasing n, and $\mathrm{MMD}(\mathcal{P}, \mathcal{Q})$, σ_{H_1} are also constants. Thus, for reasonably large n, the test power of MMD is dominated by the first term (inside Φ), and the kernel yielding the most powerful test will approximately maximize [5]

$$J(\mathcal{P}, \mathcal{Q}; k) := \mathrm{MMD}^2(\mathcal{P}, \mathcal{Q}; k)/\sigma_{H_1}(\mathcal{P}, \mathcal{Q}; k). \tag{3}$$

E. Maximizing Test Power

Although the higher value of criterion $J(\mathcal{P}, \mathcal{Q}; k)$ means higher test power of MMD, we cannot directly maximize $J(\mathcal{P}, \mathcal{Q}; k)$ since $\mathrm{MMD}^2(\mathcal{P}, \mathcal{Q}; k)$ and $\sigma_{H_1}(\mathcal{P}, \mathcal{Q}; k)$ depends on the particular $\mathcal{P}$ and $\mathcal{Q}$ that are unknown. However, We can estimate it with

$$\hat{J}_\lambda(S_X, S_Y; k) := \frac{\widehat{\mathrm{MMD}}_u^2(S_X, S_Y; k)}{\hat{\sigma}_{H_1,\lambda}(S_X, S_Y; k)}, \tag{4}$$

where $\hat{\sigma}^2_{H_1,\lambda}$ is a regularized estimator of $\sigma^2_{H_1}$ given by[1]

$$\frac{4}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} H_{ij}\right)^2 - \frac{4}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}\right)^2 + \lambda, \tag{5}$$

where S_X and S_Y are observations from $\mathcal{P}$ and $\mathcal{Q}$. Researchers can construct a test by choosing k to maximize $\hat{J}_\lambda(S_X, S_Y; k)$, then using the chosen k to calculate $\hat{J}_\lambda(S_X, S_Y; k)$.

III. Two-Sample Test via Imprecise Observations

Although existing kernel two-sample test methods have been well developed (e.g., MMD mentioned above), they all share the assumption that observations from distributions are *precise*, which is not realistic in the real world. For example, readings on an analogue measurement equipment are not precise numbers but intervals since there is only a finite number of decimals available. This motivates us to consider a more realistic problem, named as two-sample test on imprecise observations.

Problem 1 (Two-sample test on imprecise observations)**.** *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $\mathcal{P}$, $\mathcal{Q}$ be Borel probability measures on $\mathcal{X}$, and $S_X = \{x_i\}_{i=1}^n \sim \mathcal{P}^n$ and $S_Y = \{y_j\}_{j=1}^m \sim \mathcal{Q}^m$ be precise independent identically distributed* (i.i.d.) *observations from $\mathcal{P}$ and $\mathcal{Q}$, and $T : \mathcal{X} \to \mathcal{X}$ be the imprecise-observation function. We aim to determine if S_X and S_Y come from the same distribution through the imprecise observations $\tilde{S}_X = \{\tilde{x}_i := T(x_i)\}_{i=1}^n$ and $\tilde{S}_Y = \{\tilde{y}_j : T(y_j)\}_{i=1}^m$.*

In Problem 1, S_X and S_Y are unknown, and T is not a bijective function in general. Namely, we cannot find a T^{-1} to recover S_X and S_Y from $\tilde{S}_X$ and $\tilde{S}_Y$. In this paper, we consider the following T.

$$\tilde{x} := T_v(x) = \lfloor x * v \rfloor / v, \tag{6}$$

where the higher $v \in \mathbb{Z}^+$ means that the imprecise observations are closer to the precise observations. If $v = 2$, the decimal places of $T_v(x)$ can only be 0 or 5 (e.g., $0.5, 1.0, 1.5, 2.0$). Clearly, we cannot obtain T^{-1} to recover x.

In Figure 2, we show the difference between precise observations and imprecise observations in the Blob example. Moreover, we show that the test power of MMD drops significantly when we can only obtain the imprecise observations (see Figure 1), which motivates us to propose a novel two-sample test method to address this new problem.

IV. Fuzzy-based MMD Test

To handle such imprecise observations in two-sample test problem, in this paper, we first model $T_v(x)$ using fuzzy geometry and then propose a fuzzy-based kernel function. Finally, we can optimize parameters of the fuzzy-based kernel function by maximizing the test power. Based on the optimized fuzzy-based kernel, we can perform the two-sample test on the imprecise observations.

A. n-D Fuzzy Geometry

The geometric properties of fuzzy sets have been studied from various aspects such as fuzzy point, fuzzy line and fuzzy circle [30]–[34]. The *n-D fuzzy geometry* (n-D FG) theory provides an effective way to analyze and compute fuzzy

[1] This estimator, as a V-statistic, is biased even when $\lambda = 0$. Although [5], [29] give a quadratic-time estimator unbiased for $\sigma^2_{H_1}$, it is much more complicated to implement and analyze, likely has higher variance, and is often negative.

information in a geometric form [35]. This subsection presents the definition of fuzzy vector, which is the key to the n-D FG. Without loss of generality, this paper uses capital or small letters with a bar to represent the fuzzy subsets or fuzzy points of $\mathbb{R}^n$. The membership function of a fuzzy set $\bar{A}$ is expressed by $\mu(x|\bar{A})$, $x \in R^n$, with $\mu(x|\bar{A}) \subseteq$ [0,1]. The definition of fuzzy vector at the n-D real valued vector A is presented as follows.

Definition 1 (Fuzzy vector [36])**.** *A fuzzy set $\bar{A}(a_1, a_2, ..., a_n)$ of $\mathbb{R}^n$ is called a fuzzy vector at $A = (a_1, a_2, ..., a_n) \in \mathbb{R}^n$ if its membership function μ satisfies the following properties.*

1. $\mu((x_1, x_2, ..., x_n)|\bar{A}(a_1, a_2, ..., a_n)) = 1$ *is upper semi-continuous in $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$.*

2. $\mu((x_1, x_2, ..., x_n)|\bar{A}(a_1, a_2, ..., a_n)) = 1$ *if and only if* $(x_1, x_2, ..., x_n) = (a_1, a_2, ..., a_n)$.

3. $\bar{A}(\alpha) = \{x|\mu(x|\bar{A}(a_1, a_2, ..., a_n)) = \alpha,\ x \in \mathbb{R}^n\}$ *is a compact convex subset of $\mathbb{R}^n$ for all α in* $[0, 1]$.

The fuzzy vector is the elementary concept to study properties of the n-D FG and the set of all n-D fuzzy vectors is denoted by $F(\mathbb{R}^n)$. The third property of n-D fuzzy vectors means that $F(\mathbb{R}^n)$ can be connected with $\mathbb{R}^n$ using the membership α. In this paper, we use the triangular membership function to construct the membership function of each n-D fuzzy vector in $F(\mathbb{R}^n)$ (see details in the next subsection).

B. Distance between Fuzzy Vectors

In [37], a new metric between two n-D fuzzy vectors was proposed to construct the similarity between two fuzzy vectors. This subsection briefly presents the metric, which is a key component of the proposed fuzzy-based kernel function. First, the detailed expression of a fuzzy vector $\bar{A}_i(a_{i1}, a_{i2}, ..., a_{in})$ (with the triangular membership function) is expressed in the following formula: for each $\bar{a}_{ij} \in F(\mathbb{R})$, its membership function is

$$\mu_{ij}(x|\bar{a}_{ij}) = \begin{cases} 0, & \forall x < a_{ij} - \rho_i \\ 1 - \frac{|x - a_{ij}|}{\rho_i}, & \forall |x - a_{ij}| \leq \rho_i \ , x \in \mathbb{R}, \\ 0, & \forall x > a_{ij} + \rho_i \end{cases} \tag{7}$$

Based on the $\mu_{ij}(x|\bar{a}_{ij})$, $\mu_i(x|\bar{A}_i)$ is expressed by the following term.

$$\mu_i(x|\bar{A}_i) = \begin{cases} 0, & \exists x_j, x_j < a_{ij} - \rho_i \\ 1 - \frac{\|x - a_{ij}\|_1}{n\rho_i}, & \forall x_j, |x_j - a_{ij}| \leq \rho_i \ , \\ 0, & \exists x_j, x_j > a_{ij} + \rho_i \end{cases} \tag{8}$$

where $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ and $\rho_i > 0$. Then, we define a metric to measure the distance between two fuzzy vectors.

Definition 2. *Given two fuzzy vectors $\bar{A}_i \in F(\mathbb{R}^n)$ and $\bar{A}_j \in F(\mathbb{R}^n)$, the metric between $\bar{A}_i$ and $\bar{A}_j$ is defined by the map $\mathcal{D} : F(\mathbb{R}^n) \times F(\mathbb{R}^n) \to [0, +\infty)$:*

$$\mathcal{D}(\bar{A}_i, \bar{A}_j) = \frac{1}{n}\int_0^1 sup\{\mathcal{D}_\lambda(u, v) : \mathcal{D}_\lambda(u, v) \in \Omega(\lambda)\}d\lambda,$$

where

$$\Omega(\lambda) = \{d(u, \bar{A}_j(\lambda))\} \cup \{d(v, \bar{A}_i(\lambda))\},$$

$u \in \bar{A}_i(\lambda), v \in \bar{A}_j(\lambda)$ *and the first part of $\Omega(\lambda)$ collects L_1 distances between each u and $\bar{A}_j(\lambda)$ ($d(u, \bar{A}_j(\lambda)) = \min\{d(u, v), v \in \bar{A}_j(\lambda)\}$ means the minimum L_1 distances between u and all elements in $\bar{A}_j(\lambda)$), and the second part of $\Omega(\lambda)$ collects L_1 distances between v and $\bar{A}_i(\lambda)$ ($d(v, \bar{A}_i(\lambda)) = min\{d(v, u), u \in \bar{A}_i(\lambda)\}$ means the minimum L_1 distances between v and all elements in $\bar{A}_i(\lambda)$), and $d(u, v)$ represents the L_1 distance (ℓ_1−norm) between two n-dimension vector (u and v).*

Then, the following equation is obtained to calculate $\mathcal{D}$ [37].

$$\mathcal{D}(\bar{A}_i, \bar{A}_j) = \frac{1}{n}d(A_i, A_j) + \frac{1}{4}|\rho_i - \rho_j|. \tag{9}$$

Moreover, $(F(\mathbb{R}^n), \mathcal{D})$ is a metric space [37].

C. Fuzzy-based Kernel Function

Since we can only access the imprecise observations $\tilde{S}_X$ and $\tilde{S}_Y$, we regard each observation in $\tilde{S}_X$ as a fuzzy vector. Thus, for each observation (e.g., $\tilde{x}_1 = T(x_1)$), it consists of d fuzzy numbers:

$$\bar{\tilde{x}}_i = [(\tilde{x}_{i1} - \rho_i, \tilde{x}_{i1}, \tilde{x}_{i1} + \rho_i), \ldots, (\tilde{x}_{id} - \rho_i, \tilde{x}_{id}, \tilde{x}_{id} + \rho_i)]$$

and $(\tilde{x}_{ij} - \rho_i, \tilde{x}_{ij}, \tilde{x}_{ij} + \rho_i)$ is a fuzzy number with a triangular membership function, where $j = 1, \ldots d$. Then, based on the metric $\mathcal{D}$, we define the fuzzy-based kernel function as

$$\begin{aligned} k_{\theta'}(\tilde{x}_i, \tilde{x}_j) &= \exp\left(-\frac{\mathcal{D}(\bar{\tilde{x}}_i, \bar{\tilde{x}}_j)}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\frac{1}{d}\|\tilde{x}_i - \tilde{x}_j\|_1 + \frac{1}{4}|\rho_i^x - \rho_j^x|}{2\sigma^2}\right), \end{aligned} \tag{10}$$

where $\theta' = \{\rho_i^x\}_{i=1}^n \cup \{\rho_j^y\}_{j=1}^m \cup \sigma$ is the parameter set of the function $k_{\theta'}(\cdot, \cdot)$ associated to $\tilde{S}_X$ and $\tilde{S}_Y$. However, since the number of parameters of $k_{\theta'}$ will grow when increasing the number of samples, it is not practical to optimize θ'. To address this issue, we assume that there is a real-valued function f that can map each $\tilde{x}_i$ to its ρ_i^x. Based on this assumption, the fuzzy-based kernel function k_θ can be expressed in the following.

$$\begin{aligned} k_\theta(\tilde{x}_i, \tilde{x}_j) &= \exp\left(-\frac{\frac{1}{d}\|\tilde{x}_i - \tilde{x}_j\|_1 + \frac{1}{4}|\rho_i^x - \rho_j^x|}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\frac{1}{d}\|\tilde{x}_i - \tilde{x}_j\|_1 + \frac{1}{4}|f(\tilde{x}_i) - f(\tilde{x}_j)|}{2\sigma^2}\right), \end{aligned} \tag{11}$$

where θ is the set containing σ and parameters of f.

Algorithm 1 Learning fuzzy-based kernel function

```
1: Input $\tilde{S}_X$, $\tilde{S}_Y$, learning rate $\eta$, $\theta_0$, $\sigma_0$, epoch $T_k$ and $T_{max}$;
2: Initial $\theta = \theta_0$, $\sigma = \sigma_0$ and $\lambda = 10^{-8}$;
3: Split $\tilde{S}_X = \tilde{S}_X^{tr} \cup \tilde{S}_X^{te}$ and $\tilde{S}_Y = \tilde{S}_Y^{tr} \cup \tilde{S}_Y^{te}$;
for $T = 1, 2, \ldots, T_{max}$ do
    4: Obtain $k_\theta(\cdot, \cdot)$ using $\theta$ in (11);    // update kernel
    5: Compute $M(\theta, \sigma) = \widehat{\mathrm{MMD}}_u^2(\tilde{S}_X^{tr}, \tilde{S}_Y^{tr}; k_\theta)$ via (2);
    6: Compute $V_\lambda(\theta, \sigma) = \hat{\sigma}^2_{H_1,\lambda}(\tilde{S}_X^{tr}, \tilde{S}_Y^{tr}; k_\theta)$ via (5);
    7: Compute $J_\lambda(\theta, \sigma) = M(\theta, \sigma)/\sqrt{V_\lambda(\theta, \sigma)}$;
    8: Update $\theta = \theta + \eta \nabla_{\mathrm{Adam}} J(\theta, \sigma)$;  // maximize $J(\theta, \sigma)$
    9: Update $\sigma = \sigma + \eta \nabla_{\mathrm{Adam}} J(\theta, \sigma)$;  // maximize $J(\theta, \sigma)$
end
9: Output $k_\omega$, $S_P^{te}$ and $S_Q^{te}$
```

D. Fuzzy-based MMD

To find the best parameter set θ for the two-sample test problem, we adopt the strategy of maximizing test power (as introduced in Section II-E). Namely, we want to maximize the following function with respective to θ.

$$\hat{J}_\lambda(\tilde{S}_X, \tilde{S}_Y; k_\theta) := \frac{\widehat{\mathrm{MMD}}_u^2(\tilde{S}_X, \tilde{S}_Y; k_\theta)}{\hat{\sigma}_{H_1,\lambda}(\tilde{S}_X, \tilde{S}_Y; k_\theta)}, \tag{12}$$

where all notations are introduced in Section II. Algorithm 1 shows detailed procedures to learn the fuzzy-based kernel function k_θ. Since $\bar{\tilde{x}}_i$ can be regarded as a fuzzy representation of $\tilde{x}_i$, we also call this optimization procedure as finding the best *fuzzy representations* for $\tilde{S}_X$ and $\tilde{S}_Y$.

E. Fuzzy-based MMD for Two-sample Test

To use the learned fuzzy-based kernel k_θ^* to address two-sample test problem, we proceed in three steps.

First, compute $\hat{t} = \widehat{\mathrm{MMD}}_u^2(\tilde{S}_X^{te}, \tilde{S}_Y^{te}; k_\theta^*)$ via Eq. (2).

Second, we use permutation test to compute the test threshold r_α. Let $\tilde{S}^{te} = \tilde{S}_X^{te} \cup \tilde{S}_Y^{te}$. At the i^{th} permutation, we randomly take $|\tilde{S}_X^{te}|$ samples out from $\tilde{S}^{te}$ to form $\tilde{S}_{i,X'}^{te}$ and regard remaining samples in $\tilde{S}^{te}$ as $\tilde{S}_{i,Y'}^{te}$. Then, compute the test statistic

$$\hat{t}_i^{\text{null}} = \widehat{\mathrm{MMD}}_u^2(\tilde{S}_{i,X'}^{te}, \tilde{S}_{i,Y'}^{te}; k_\theta^*)$$

via Eq. (2). So, after N_{perm} permutations, we can obtain $\{\hat{t}_i^{\text{null}}\}_{i=1}^{N_{\text{perm}}}$. The test threshold r_α will be the $(1-\alpha)-$quantile of $\{\hat{t}_i^{\text{null}}\}_{i=1}^{N_{\text{perm}}}$.

Third, reject H_0 if $\hat{t} > r_\alpha$, and accept it otherwise.

V. Experiment Evaluation

We compare our test method against four state-of-the-art two-sample tests in this section.

A. Baselines

We compare the following tests on several datasets:

- F-MMD: Fuzzy-based MMD with maximum test power.
- MMD-O: MMD with a Gaussian kernel whose lengthscale is optimized to maximize the test power.
- Mean embedding (ME): a state-of-the-art test [3], [4] based on differences in Gaussian kernel mean embeddings at a set of optimized points.
- Smooth characteristic functions (SCF): a state-of-the-art test [3], [4] based on differences in Gaussian mean embeddings at a set of optimized frequencies.

B. Experiments Setup

We take a single sample set as $\tilde{S}_X^{tr}$ and $\tilde{S}_X^{te}$ and learn a kernel or test locations once for each method on $\tilde{S}_X^{tr}$. We then evaluate its test power on 100 new sample sets $\tilde{S}_X^{te}$, $\tilde{S}_Y^{te}$ from the same distribution. We repeat each experiment 10 times, and report the mean test power.

We implement all methods on Python 3.7 (Pytorch 1.1) with an NIVIDIA Titan V GPU. We run ME and SCF using the official code [3] (we select the number of test locations of ME and SCF as 15 to ensure they have enough test power), and implement MMD-D and MMD-O by ourselves. We use permutation test to compute p-values of F-MMD, MMD-O. We set $\alpha = 0.05$ for all experiments.

To ensure that f can learn the best mapping between the imprecise observation and its membership function's parameter (i.e., ρ_i), we let f be a four-layer neural network with softplus activation function. The number of neurons in the two hidden layers is set to 20. The Adam optimizer [38] is used to optimize the θ (for F-MMD) and the bandwidth of Gaussian kernel function (for MMD-O) to maximum the approximate test power, and the learning rate of the Adam optimizer is set to 0.0005 for F-MMD and MMD-O.

C. Evaluation on Blob Dataset

Blob-D is the dataset shown in Figure 2; *Blob-S* has both P also equal to the distribution shown in Figure 2a, so that the null hypothesis holds. Results are shown in Tabel II. *F-MMD* has the highest average test power among all considered test methods, as well as generally consistent power (low variance). All methods have appropriate Type I error rates.

VI. Conclusion

This paper considers a new problem called *two-sample test using imprecise observations*. Although existing two-sample test methods can directly handle such imprecise observations, the test power of them will drop significantly. To address this new problem, we first propose a fuzzy-based kernel function to reveal the relation between two imprecise observations. Parameters of the fuzzy-based kernel function are optimized by maximizing the approximate test power. Eventually, based on the optimized fuzzy-based kernel function, we propose the fuzzy-based maximum mean discrepancy to handle the imprecise observations in two-sample test problem. Empirical results demonstrate that the test power of our test method is higher than that of existing two-sample test methods when facing imprecise observations.

Acknowledgment

The work presented in this paper was supported by the Australian Research Council (ARC) under FL190100149.

TABLE I

Specifications of P and Q of the Blob dataset. $\mu_1^b = [0,0], \mu_2^b = [0,1], \mu_3^b = [0,2], \ldots, \mu_8^b = [2,1], \mu_9^b = [2,2]$ (same with Figure 2a). $\Delta_i^b = -0.02 - 0.002 \times (i-1)$ if $i < 5$ and $\Delta_i^b = 0.02 + 0.002 \times (i-6)$ if $i > 5$. if $i = 5$, $\Delta_i^b = 0$ (same with Figure 2a).

Datasets	P	Q
Blob-S	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$
Blob-D	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}\left(\mu_i^b, \begin{bmatrix} 0.03 & \Delta_i^b \\ \Delta_i^b & 0.03 \end{bmatrix}\right)$

TABLE II

Results on *Blob* ($\alpha = 0.05$): average test power±standard errors for increasing numbers of samples per mode (N). In *Blob*, each set of samples has 9 modes (see Table I).

N	F-MMD	MMD-O	ME	SCF
10	0.164±0.011	0.148±0.015	0.092±0.012	0.114±0.007
20	0.268±0.019	0.272±0.08	0.135±0.021	0.134±0.008
40	0.716±0.009	0.523±0.10	0.177±0.044	0.253±0.04
60	0.950±0.002	0.949±0.003	0.149±0.065	0.349±0.041
80	0.979±0.001	0.929±0.085	0.584±0.117	0.537±0.471
100	1.000±0.000	0.907±0.087	0.637±0.098	0.723±0.053
Avg.	0.679	0.621	0.296	0.352

References

[1] M. G. Akbari and G. Hesamian, "Testing statistical hypotheses for intuitionistic fuzzy data," *Soft Computing*, vol. 23, no. 20, pp. 10 385–10 392, 2019.

[2] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[3] W. Jitkrittum, Z. Szabo, K. Chwialkowski, and A. Gretton, "Interpretable distribution features with maximum testing power," in *NeurIPS*, 2016.

[4] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, "Fast two-sample testing with analytic representations of probability measures," in *NeurIPS*, 2015.

[5] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton, "Generative models and model criticism via optimized maximum mean discrepancy," in *ICLR*, 2017.

[6] G. Hesamian and M. Shams, "Parametric testing statistical hypotheses for fuzzy random variables," *Soft Comput.*, vol. 20, no. 4, pp. 1537–1548, 2016.

[7] G. Hesamian and J. Chachi, "Two-sample kolmogorov–smirnov fuzzy test for fuzzy random variables," *Statistical Papers*, vol. 56, no. 1, pp. 61–82, 2015.

[8] A. Liu, J. Lu, F. Liu, and G. Zhang, "Accumulating regional density dissimilarity for concept drift detection in data streams," *Pattern Recognition*, vol. 76, pp. 256–272, 2018.

[9] A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, and M. Pontil, "Optimal kernel choice for large-scale two-sample tests," in *NeurIPS*, 2012.

[10] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[11] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[12] W. Pan, Y. Tian, X. Wang, and H. Zhang, "Ball divergence: Nonparametric two sample test," *Annals of statistics*, vol. 46, no. 3, pp. 1109–1137, 2018.

[13] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu *et al.*, "Equivalence of distance-based and rkhs-based statistics in hypothesis testing," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.

[14] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for rkhs embeddings of probability distributions," in *NeurIPS*, 2009.

[15] K. Muandet, B. K. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf, "Kernel mean shrinkage estimators," *Journal of Machine Learning Research*, vol. 17, pp. 48:1–48:41, 2016.

[16] F. Liu, G. Zhang, and J. Lu, "Heterogeneous domain adaptation: An unsupervised approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. Early Access, pp. 1–15, 2020.

[17] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and I. Systems, "Domain adaptation with conditional transferable components," in *ICML*, 2016.

[18] F. Liu, G. Zhang, and J. Lu, "A novel fuzzy neural network for unsupervised domain adaptation in heterogeneous scenarios," in *FUZZ-IEEE*, 2019.

[19] A. Liu, G. Zhang, and J. Lu, "Concept drift detection via equal intensity k-means space partitioning," *IEEE Transactions on Cybernetics*, vol. Early Access, pp. 1–14, 2020.

[20] ——, "Region drift disagreement-based diverse instances weighting ensemble for concept drift adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. Early Access, pp. 1–15, 2020.

[21] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *ICLR*, 2018.

[22] M. Arbel, D. J. Sutherland, M. Binkowski, and A. Gretton, "On gradient regularizers for MMD GANs," in *NeurIPS*, 2018.

[23] P. Filzmoser and R. Viertl, "Testing hypotheses with fuzzy data: the fuzzy p-value," *Metrika*, vol. 59, no. 1, pp. 21–29, 2004.

[24] J. Chachi, "On distribution characteristics of a fuzzy random variable," *Austrian Journal of Statistics*, vol. 47, no. 2, pp. 53–67, 2018.

[25] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.

[26] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "On integral probability metrics, ϕ-divergences and binary classification," 2009, arXiv:0901.2698.

[27] S. T. Rachev, "The monge–kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[28] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

[29] D. J. Sutherland, "Unbiased estimators for the variance of MMD estimators," *arXiv:1906.02104*, 2019.

[30] D. Ghosh and D. Chakraborty, "Analytical fuzzy plane geometry I," *Fuzzy Sets and Systems*, vol. 209, pp. 66–83, 2012.

[31] D. Chakraborty and D. Ghosh, "Analytical fuzzy plane geometry II," *Fuzzy Sets and Systems*, vol. 243, pp. 84–109, 2014.

[32] D. Ghosh and D. Chakraborty, "Analytical fuzzy plane geometry III," *Fuzzy Sets and Systems*, vol. 283, pp. 83–107, 2016.

[33] J. Buckley and E. Eslami, "Fuzzy plane geometry I : Points and lines," *Fuzzy Sets and Systems*, vol. 86, pp. 179–187, 1997.

[34] ——, "Fuzzy plane geometry II : Circles and polygons," *Fuzzy Sets and Systems*, vol. 87, pp. 79–85, 1997.

[35] Y. Li, Q. Huang, W. Xie, and X. Li, "A novel visual codebook model based on fuzzy geometry for large-scale image classification," *Pattern Recognition*, vol. 48, no. 10, pp. 3125–3134, 2015.

[36] R. Goetschel and W. Voxman, "Topological properties of fuzzy numbers," *Fuzzy Sets and Systems*, vol. 10, pp. 87–99, 1983.

[37] F. Liu, J. Lu, and G. Zhang, "Unsupervised heterogeneous domain adaptation via shared fuzzy relations," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3555–3568, 2018.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.