Mixture Proportion Estimation and Weakly-supervised Kernel Test for Conditional Independence

Yushi Hirose
Institute of Science Tokyo

Akito Narahara
Institute of Science Tokyo

Takafumi Kanamori
Institute of Science Tokyo, RIKEN AIP

Abstract

Mixture proportion estimation (MPE) aims to estimate class priors from unlabeled data. This task is a critical component in weakly supervised learning, such as PU learning, learning with label noise, and domain adaptation. Existing MPE methods rely on the *irreducibility* assumption or its variant for identifiability. In this paper, we propose novel assumptions based on conditional independence (CI) given the class label, which ensure identifiability even when irreducibility does not hold. We develop method of moments estimators under these assumptions and analyze their asymptotic properties. Furthermore, we present weakly-supervised kernel tests to validate the CI assumptions, which are of independent interest in applications such as causal discovery and fairness evaluation. Empirically, we demonstrate the improved performance of our estimators compared with existing methods and that our tests successfully control both type I and type II errors.

1 INTRODUCTION

Mixture Proportion Estimation (MPE) is the problem of estimating the mixture proportions of underlying class distributions in unlabeled data. This work addresses a generalized MPE setting (Unlabeled-Unlabeled or UU setting) where we are given samples from two distinct mixture distributions, $U = \theta P + (1-\theta)N$ and $U' = \theta' P + (1-\theta')N$. Here, P and N represent the positive and negative class probability distributions, respectively, and the objective is to estimate the unknown class priors (θ, θ'). This formulation is a strict generalization of the standard MPE setting (Positive-Unlabeled or PU setting), which assumes access to P, corresponding to the case where $\theta = 1$.

MPE is a critical component in various machine learning tasks. For instance, weakly-supervised learning such as positive-unlabeled learning (Elkan and Noto, 2008; du Plessis et al., 2014; Kiryo et al., 2017), unlabeled-unlabeled learning (Lu et al., 2019, 2021), and learning from pairwise similarity (Bao et al., 2018) require known mixture proportions to train classifiers. Other applications of MPE include learning with label noise (Natarajan et al., 2013; Scott et al., 2013; Liu and Tao, 2016), anomaly detection (Sanderson and Scott, 2014), and domain adaptation under open set label shift (Garg et al., 2022).

Without any assumptions on the distributions P and N, (θ, θ') are not identifiable. To address this, Blanchard et al. (2010) introduced the irreducibility assumption, which posits that N is irreducible with respect to P. Intuitively, this means that the negative distribution cannot be expressed as a mixture containing the positive distribution. In the UU setting, the converse: irreducibility of P with respect to N is also required for identifiability (Scott, 2015). To date, most of the existing MPE algorithms (Scott, 2015; Scott et al., 2013; Jain et al., 2016; Ramaswamy et al., 2016; Ivanov, 2020; Bekker and Davis, 2020) have been developed under the irreducibility assumption or stricter conditions, such as the anchor set assumption (Ramaswamy et al., 2016).

However, the irreducibility assumption can be violated in practical applications as discussed by Zhu et al. (2023). To the best of our knowledge, Yao et al. (2022) and Zhu et al. (2023) are the only existing works that have attempted MPE beyond irreducibility, while they still have limitations. The regrouping method of Yao et al. (2022) can mitigate estimation bias, but is not statistically consistent. Zhu et al. (2023) derived a more general condition than irreducibility that requires the essential supremum of $P(Y|X = x)$, which is usually unavailable in practice.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

This work investigates alternative assumptions and presents new identifiability results for MPE. Our assumptions are based on the underlying data structure: conditional independence given class label. The CI assumption is widely adopted across various machine learning domains. For instance, in text classification and spam filtering (Jurafsky and Martin, 2024), features are often assumed to be independent given the label. Similarly, multi-view learning paradigms, including co-training (Blum and Mitchell, 1998) and unsupervised-learning (Song et al., 2014; Steinhardt and Liang, 2016; Anandkumar et al., 2014), frequently leverage conditional independence between feature sets. The applications include web page categorization (Blum and Mitchell, 1998), text-image categorization (Giesen et al., 2021) and biological data such as flow cytometry (Song et al., 2014). We consider two structural assumptions, conditional independence (CI) and multivariate conditional independence (MCI), on which we develop the method of moments estimators.

We also establish kernel test methods to verify the CI assumptions from only observed data from U and U' (weakly-supervised setting). Our tests are developed from non-parametric kernel tests such as HSIC (Gretton et al., 2007) and Kernel-based Conditional Independence test (Zhang et al., 2011). We show the test consistency under mild conditions. In contrast, existing MPE assumptions, such as irreducibility, are generally difficult to verify, and no such studies currently exist. Our tests have potential applications not only in MPE, but also in fairness evaluation (Mehrabi et al., 2021) and causal discovery (Gordon et al., 2023).

Our contributions are summarized as follows.

- We propose method of moments estimators for MPE, under class-specific CI (independence of two features given the class label) and MCI (independence of two features given the class label and additional features) assumptions. We show the asymptotic normality of estimators.
- We establish kernel tests for CI and MCI using unlabeled data from U and U', which is not possible with existing kernel tests.
- We investigate the testing methods under two settings: one where the true mixture proportions are known, and another where they are unknown. We derive the asymptotic distributions of the proposed test statistics under H_0 and propose gamma approximation methods.
- In the test setting of unknown mixture proportions, we propose a post-hoc testing method, where we plug estimated mixture proportions into the test statistics and estimate the modified mean and variance for gamma approximation.

2 PROBLEM SETTING

Let X and Y be feature and binary label random variables that take values in $\mathcal{X}$ and $\{-1, 1\}$ respectively. In this paper, we assume two unlabeled datasets are given:

$$\{x^{(i)}\}_{i=1}^{n} \sim U = \theta P + (1-\theta)N$$
$$\{x'^{(i)}\}_{i=1}^{n'} \sim U' = \theta' P + (1-\theta')N$$

where $P := P(X|Y=1)$ and $N := P(X|Y=-1)$ are probability distributions for positive and negative data. U and U' are unlabeled data distributions with different mixture proportions $\theta, \theta' \in [0,1]$ ($\theta > \theta'$, without loss of generality). Denote $M = n + n'$ and assume $\lim_{M\to\infty} M/n \to \nu$ and $\lim_{M\to\infty} M/n' \to \nu'$. The objective is to estimate θ, θ' from unlabeled datasets $\{x^{(i)}\}_{i=1}^{n}, \{x'^{(i)}\}_{i=1}^{n'}$. Under these settings, the labeled distributions P and N can be expressed as mixtures of the unlabeled distributions, U and U':

$$P = \alpha_+ U + (1-\alpha_+)U' \tag{1}$$
$$N = \alpha_- U + (1-\alpha_-)U' \tag{2}$$

where $\alpha_+ := \frac{1-\theta'}{\theta-\theta'}$ and $\alpha_- := \frac{-\theta'}{\theta-\theta'}$.

These equations demonstrate that the labeled distributions can be recovered from the unlabeled distributions if (θ, θ') are known. This property is utilized in weakly-supervised learning (Lu et al., 2019; du Plessis et al., 2014) and learning with label noise (Natarajan et al., 2013; Scott, 2015). We also leverage this property for our proposed MPE and CI tests. Specifically in our MPE, we estimate α_+ and α_-, which is equivalent to estimating θ, θ', and is more convenient for theoretical analysis.

3 MIXTURE PROPORTION ESTIMATION WITH CONDITIONAL INDEPENDENCE

In this section, we present a novel approach to mixture proportion estimation by leveraging conditional independence (CI) between features.

We assume a two-dimensional feature $X = (X_1, X_2)$ that takes values in $\mathcal{X}_1 \times \mathcal{X}_2$. We denote the positive and negative distributions as $P_{12} := P(X_1, X_2|Y=1)$ and $N_{12} := P(X_1, X_2|Y=-1)$, and unlabeled distributions $U_{12} := \theta P_{12} + (1-\theta)N_{12}$ and $U'_{12} := \theta' P_{12} + (1-\theta')N_{12}$. Similarly, $P_\tau, N_\tau, U_\tau, U'_\tau, \forall\tau \in \{1,2\}$ represent the corresponding marginal distributions. We define mixture distributions $F_\tau^\alpha := \alpha U_\tau + (1-\alpha)U'_\tau$ and

their empirical versions $\hat{F}_\tau^\alpha := \alpha\hat{U}_\tau + (1-\alpha)\hat{U}'_\tau, \forall\tau \in \{1,2,12\}$. Note that $F_\tau^{\alpha_+} = P_\tau$ and $F_\tau^{\alpha_-} = N_\tau$ from (1) and (2).

In the following procedure, we focus on estimating $\alpha^* \in \{\alpha_+, \alpha_-\}$ and define $\bar{\alpha}^* \in \{\alpha_+, \alpha_-\} \setminus \{\alpha^*\}$. Note that a similar assumption and procedure are required to estimate $\bar{\alpha}^*$ and subsequently recover (θ, θ'). We assume the following class-specific CI assumption:

Assumption 1. $F_{12}^{\alpha^*} = F_1^{\alpha^*} F_2^{\alpha^*}$

If $\alpha^* = \alpha_+$ (resp. α_-), the assumption corresponds to $P_{12} = P_1 P_2$ (resp. $N_{12} = N_1 N_2$). If we have multi-dimensional features (e.g., $\mathcal{X} = \mathbb{R}^d, d > 2$), we select two features that satisfy Assumption 1 depending on whether the target is α_+ or α_-.

Assumption 1 enables the identification of the mixture proportion α^*. This is because F_{12}^α is a mixture distribution of U and U', and it generally does not exhibit conditional independence unless $\alpha = \alpha^*$ under Assumption 1 as shown in Lemma 1. To derive a moment condition and the MPE estimator, we define two vector-valued functions, $g_1 : \mathcal{X}_1 \to \mathbb{R}^d$ and $g_2 : \mathcal{X}_2 \to \mathbb{R}^d$, along with their dot product $g_{12}(X) := g_1(X_1) \cdot g_2(X_2)$. For notational simplicity, we often omit function arguments when it is clear, e.g., $g(x)$ as g. Under Assumption 1, we can establish the following moment condition.

Lemma 1. *Define a moment function*

$$m_{CI}(\alpha) := E_{F_{12}^\alpha}[g_{12}] - E_{F_1^\alpha F_2^\alpha}[g_{12}]$$

Under Assumption 1, $m_{CI}(\alpha^*) = 0$. Moreover, the quadratic equation $m_{CI}(\alpha)$ has real solutions if $(E_{P_1}[g_1] - E_{N_1}[g_1]) \cdot (E_{P_2}[g_2] - E_{N_2}[g_2]) \neq 0$.

Note that for an arbitrary $\alpha \in \mathbb{R}$, the distribution F_τ^α is not necessarily a valid probability distribution, but rather a signed measure. Nevertheless, we define its expectation as an integral w.r.t. F_τ^α. Lemma 1 implies that α^* can be estimated by finding the roots of the equation $m_{CI}(\alpha) = 0$. Therefore, we define our empirical estimator as:

$$\hat{\alpha}_{CI} := \operatorname*{argmin}_{\alpha \in I_{\alpha^*}} \hat{m}_{CI}^2(\alpha)$$

where $\hat{m}_{CI}(\alpha) := E_{\hat{F}_{12}^\alpha}[g_{12}] - E_{\hat{F}_1^\alpha \hat{F}_2^\alpha}[g_{12}]$ and I_{α^*} is a bounded, closed set containing α^*.

The search space I_{α^*} should be chosen to ensure that α^* is the unique solution within this interval. If solving the equation yields two solutions within I_{α^*}, a disambiguation step may be necessary. This can be achieved by performing a second estimation with a different feature map g'_{12} and selecting the solution that yields a small $\hat{m}_{CI}^2(\alpha)$ for both g_{12} and g'_{12}.

$\hat{\alpha}_{CI}$ is a variant of method of moments estimator and we can derive its asymptotic normality as follows.

Theorem 1 (Asymptotic normality of $\hat{\alpha}_{CI}$)**.** *Assume g_1 and g_2 are bounded and continuous and α^* is the unique solution of $m_{CI}(\alpha) = 0$ in I_{α^*}. Then,*

$$\sqrt{M}(\hat{\alpha}_{CI} - \alpha^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\nu\alpha^{*2} V_{U_{12}}[\tilde{g}_{12}] + \nu'(1-\alpha^*)^2 V_{U'_{12}}[\tilde{g}_{12}]}{(\theta-\theta')^2 (E_{F_{12}^{\bar{\alpha}^*}}[\tilde{g}_{12}])^2}\right)$$

as $M \to \infty$, where $\tilde{g}_{12} := (g_1 - E_{F_1^{\alpha^*}}[g_1]) \cdot (g_2 - E_{F_2^{\alpha^*}}[g_2])$

This theorem indicates the conditions under which the asymptotic variance of the estimator is minimized. A small variance is achieved by maximizing the denominator and minimizing the numerator. The denominator is maximized when both $|\theta - \theta'|$ and $|E_{F_{12}^{\bar{\alpha}^*}}[\tilde{g}_{12}]|$ are large. $|E_{F_{12}^{\bar{\alpha}^*}}[\tilde{g}_{12}]|$ becomes large when the functions g_1 and g_2 fluctuate significantly on $F_{12}^{\bar{\alpha}^*}$ around their averages on $F_{12}^{\alpha^*}$. This implies the functions g_1 and g_2 should effectively capture the distributional difference between P and N. To minimize the numerator, the variance of $\tilde{g}_{12}$ should be small. In our experiments, we adopt identity functions for g_1 and g_2 as a practical choice, which is shown to perform well in Section 6.1.

4 MIXTURE PROPORTION ESTIMATION WITH MULTIVARIATE CONDITIONAL INDEPENDENCE

In this section, we consider a more general assumption, multivariate conditional independence (MCI), for MPE. Let us assume a multivariate feature $X = (X_1, X_2, X_S)$ that takes values in $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_S$. Similarly to the previous section, we denote the joint and marginal distributions of positive, negative, and unlabeled data as $P_\tau, N_\tau, U_\tau, U'_\tau$ with the appropriate subscripts τ. We define the mixture distribution $F_\tau^\alpha := \alpha U_\tau + (1-\alpha)U'_\tau$ and denote the conditional distribution by $F_{\tau|S}^\alpha := F_{\tau S}^\alpha / F_S^\alpha$ with a slight abuse of notation.

Suppose we aim to estimate $\alpha^* \in \{\alpha_+, \alpha_-\}$. We define the following class-specific MCI assumption:

Assumption 2. $F_{12S}^{\alpha^*} = F_{1|S}^{\alpha^*} F_{2|S}^{\alpha^*} F_S^{\alpha^*}$

As in CI MPE, we do not need to use the same feature triplet (X_1, X_2, X_S) to estimate both α_+ and α_-, if there are additional feature variables available. We can select a triplet that satisfies Assumption 2 depending on whether the target is α_+ or α_-. For MCI MPE, we use scalar-valued functions $g_1 : \mathcal{X}_1 \to \mathbb{R}$ and $g_2 : \mathcal{X}_2 \to \mathbb{R}$. By defining the conditional means $\mu_1^\alpha(x_S) := E_{F_{1S}^\alpha}[g_1(X_1)|X_S = x_S]$ and $\mu_2^\alpha(x_S) := E_{F_{2S}^\alpha}[g_2(X_2)|X_S = x_S]$, we can establish the follwoing lemma under Assumption 2.

Lemma 2. *Define*

$$m_{MCI}(\alpha) := E_{F^{\alpha}_{12S}}\left[(g_1(X_1) - \mu^{\alpha}_1(X_S))(g_2(X_2) - \mu^{\alpha}_2(X_S))\right]$$

Under Assumption 2, $m_{MCI}(\alpha^*) = 0$.

Based on this moment condition, we define the empirical estimator of α^* as

$$\hat{\alpha}_{MCI} := \underset{\alpha \in I_{\alpha^*}}{\operatorname{argmin}}\, \hat{m}^2_{MCI}(\alpha)$$

where $\hat{m}_{MCI}(\alpha) :=$

$$E_{\hat{F}^{\alpha}_{12S}}\left[(g_1(X_1) - \mu^{\alpha}_1(X_S))(g_2(X_2) - \mu^{\alpha}_2(X_S))\right]$$

and I_{α^*} is a bounded and closed set that contains α^*. Since we do not know the true conditional means $\mu^{\alpha}_1(x_S)$ and $\mu^{\alpha}_2(x_S)$, we estimate them using kernel ridge regression (KRR) in a weakly-supervised manner from the two sets of unlabeled data. The empirical mean squared error for $\mu^{\alpha}_{\tau}, \forall \tau \in \{1, 2\}$ is written as

$$MSE_{\mu^{\alpha}_{\tau}} = \frac{\alpha}{n}||\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})_{:n} - K_{:n,:}\mathbf{w}||^2 + \frac{1-\alpha}{n'}||\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})_{n:} - K_{n:,:}\mathbf{w}||^2 + \lambda \mathbf{w}^T K \mathbf{w}$$

where $\mathbf{w} \in \mathbb{R}^M$ is the weight vector, λ is a regularization parameter, and $\mathbf{v}_{x_{\tau}} := (x^{(1)}_{\tau}, ..., x^{(n)}_{\tau}, x'^{(1)}_{\tau}, ..., x'^{(n')}_{\tau})^T$ is the feature vector of x_{τ} for the unlabeled data. $\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})$ is a vector representing the element-wise application of g_{τ} to $\mathbf{v}_{x_{\tau}}$. $K \in \mathbb{R}^{M \times M}$ is the Gram matrix of x_S with a kernel $k(\cdot, \cdot)$, where $K_{ij} = k(\mathbf{v}_{x_S,i}, \mathbf{v}_{x_S,j})$. The terms $\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})_{:n} \in \mathbb{R}^n$, $\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})_{n:} \in \mathbb{R}^{n'}$, $K_{:n,:} \in \mathbb{R}^{n \times M}$ and $K_{n:,:} \in \mathbb{R}^{n' \times M}$ are subvectors of $\mathbf{g}_{\tau}(\mathbf{v}_{x_{\tau}})$ and submatrices of K.

Denoting the estimated parameter as $\mathbf{w}_{\alpha} = \operatorname{argmin}_{\mathbf{w}} MSE_{\mu^{\alpha}_{\tau}}$, the estimated conditional mean $\hat{\mu}^{\alpha}_{\tau}(\mathbf{v}_{x_S}) = K\mathbf{w}_{\alpha}$ is used for MCI MPE. Note that $MSE_{\mu^{\alpha}_{\tau}}$ can be non-convex when $\alpha \notin [0, 1]$. In such cases, we cannot obtain an explicit optimal solution as in standard KRR. In practice, we instead solve the first order condition to derive $\mathbf{w}_{\alpha}$. Given these procedures, $\operatorname{argmin}_{\alpha \in I_{\alpha^*}} \hat{m}^2_{MCI}(\alpha)$ can become a bilevel optimization. For the efficient computation, we can use iterative search methods, such as the golden-section search (Kiefer, 1953) to find the optimal α.

We now present the asymptotic normality of $\hat{\alpha}_{MCI}$, similarly to CI MPE.

Theorem 2 (Asymptotic normality of $\hat{\alpha}_{MCI}$)**.** *Assume g_1 and g_2 are bounded and continuous, and that $\mu^{\alpha}_{\tau}(x_S)$ is bounded and differentiable w.r.t. $\alpha \in I_{\alpha^*}$ and $x_S \in \mathcal{X}_S$. Assume $\frac{\partial}{\partial \alpha}\mu^{\alpha}_{\tau}(x_S)$ is bounded. Suppose α^* is the unique minimizer of $m^2_{MCI}(\alpha)$ in I_{α^*}, then*

$$\sqrt{M}(\hat{\alpha}_{MCI} - \alpha^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\nu {\alpha^*}^2 V_{U_{12S}}[\tilde{g}_{12S}] + \nu'(1-\alpha^*)^2 V_{U'_{12S}}[\tilde{g}_{12S}]}{(\theta - \theta')^2 (E_{F^{\alpha^*}_{12S}}[\tilde{g}_{12S}])^2}\right)$$

where $\tilde{g}_{12S}(x_1, x_2, x_S) := (g_1(x_1) - \mu^{\alpha^*}_1(x_S))(g_2(x_2) - \mu^{\alpha^*}_2(x_S))$.

This theorem suggests the conditions for small asymptotic variance, as discussed in Theorem 1. We use identity functions for g_1 and g_2 as a practical choice in our experiments.

5 CI AND MCI TEST UNDER WEAKLY-SUPERVISED SETTING

In this section, we introduce statistical testing methods to test the CI and MCI assumptions using only unlabeled data. These tests allow us to verify the applicability of our proposed MPE method. Beyond this primary objective, these tests have broader applications, including causal discovery and fairness evaluation. In causal discovery, conditional independence tests are required to infer the causal graph of the underlying data (Peters et al., 2017). In fairness evaluation, it is crucial to determine whether a classifier's output or representation $f(X)$ is independent of a protected variable Z (e.g., race or sex) given Y.

Our proposed tests can also be framed as a CI test with a single unobserved confounder in the context of recent work (Gordon et al., 2023; Mazaheri et al., 2023; Liu et al., 2024). However, existing methods have limitations; for instance, the methods of Gordon et al. (2023) and Mazaheri et al. (2023) are restricted to discrete variables. While the test by Liu et al. (2024) could be adopted by introducing an index variable to denote a sample's origin (U or U'), its application relies on a specific condition of integral equation.

In the following subsections, we first present a testing method that assumes the true mixture proportions are known. Since the proportions are unknown in advance, this setting does not directly verify Assumptions 1 and 2 for MPE. Nevertheless, the method remains valuable for the other applications mentioned above. Subsequently, we introduce a testing method without the true mixture proportions.

5.1 Weakly-supervised Kernel CI(WsKCI) Test with True Mixture Proportions

To verify the CI assumption, we test $H_0 : X_1 \perp\!\!\!\perp X_2 \mid Y = y$, against $H_1 : X_1 \not\perp\!\!\!\perp X_2 \mid Y = y$ using unlabeled data. In the first setting, we assume to know the true

mixture proportion $\alpha^* = \alpha_+$ for $y = 1$ or $\alpha^* = \alpha_-$ for $y = -1$. Let k_1 and k_2 be positive-definite and characteristic kernels (Gretton, 2015) on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively and let φ_1, φ_2 and $\mathcal{H}_1, \mathcal{H}_2$ be corresponding feature mappings and RKHSs. Our test is based on Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2007) for $F_{12}^{\alpha^*}$:

$$\left\| E_{F_{12}^{\alpha^*}}[\varphi_1(X_1) \otimes \varphi_2(X_2)] - E_{F_1^{\alpha^*} F_2^{\alpha^*}}[\varphi_1(X_1) \otimes \varphi_2(X_2)] \right\|_{\mathcal{H}}^2,$$

which is the squared Hilbert-Schmidt norm of cross-covariance operator and equals zero under H_0. Here, $\otimes$ denotes the tensor product of kernels (Schrab, 2025). By replacing the population distributions in the above statistic with their empirical counterparts, we define the following test statistic

$$T_{CI} := \left\| E_{\hat{F}_{12}^{\alpha^*}}[\varphi_1(X_1) \otimes \varphi_2(X_2)] - E_{\hat{F}_1^{\alpha^*} \hat{F}_2^{\alpha^*}}[\varphi_1(X_1) \otimes \varphi_2(X_2)] \right\|_{\mathcal{H}}^2$$

To implement the test, we require the null distribution of T_{CI}. We derive the following theorem using an approach similar to that of HSIC (Gretton et al., 2007).

Theorem 3 (Asymptotic distribution of T_{CI})**.** *Assume k_1 and k_2 are translation invariant c_0-kernels as defined in (Gretton, 2015). Then,*

(i) Under H_0, we have

$$MT_{CI} \xrightarrow{d} \sum_{i,j=1}^{\infty} \lambda_{1,i}\lambda_{2,j}\xi_{i,j}^2.$$

where $\lambda_{1,i}$ and $\lambda_{2,j}$ are the eigenvalues of the integral operators associated with k_1 and k_2, and $\xi_{i,j}$s follow a multivariate normal distribution with mean $\mathbf{0}$ and covariances defined in Appendix B.1.

(ii) Under H_1, we have $MT_{CI} \xrightarrow{p} \infty$.

The proof is provided in Appendix B.1. The null distribution is obtained by considering the Mercer's expansions of k_1 and k_2, and applying the Central Limit Theorem to them; the result under H_1 is given by Gretton (2015).

Empirically, we approximate the null distribution with a gamma distribution, an approach also used for the HSIC test (Gretton et al., 2007). The parameters for this gamma approximation are determined by estimating the mean and variance of MT_{CI}. The following theorem provides the asymptotic expressions for these moments.

Theorem 4 (Asymptotic mean and variance of MT_{CI})**.** *Define the centralized kernel $\tilde{k}_{12}$ associated with the feature map $\tilde{\varphi}_{12}(x) := (\varphi_1(x_1) - E_{F_1^{\alpha^*}}[\varphi_1(x_1)]) \otimes (\varphi_2(x_2) - E_{F_2^{\alpha^*}}[\varphi_2(x_2)])$. Under H_0 and as $M \to \infty$, we have*

$$\begin{aligned} E[MT_{CI}] \to & \nu\alpha^{*2} E_{x_{i_1},x_{i_2}}[\tilde{k}_{12}(x_{i_1}, x_{i_1}) - \tilde{k}_{12}(x_{i_1}, x_{i_2})] \\ & + \nu'(1-\alpha^*)^2 E_{x'_{q_1},x'_{q_2}}[\tilde{k}_{12}(x'_{q_1}, x'_{q_1}) - \tilde{k}_{12}(x'_{q_1}, x'_{q_2})] \\ V[MT_{CI}] \to & 2\nu^2\sigma_{CI,2,0}^2 + 2\nu'^2\sigma_{CI,0,2}^2 + 4\nu\nu'\sigma_{CI,1,1}^2, \end{aligned}$$

where x_{i_1},x_{i_2} and x'_{q_1},x'_{q_2} are i.i.d samples from U_{12} and U'_{12}, and

$$\begin{aligned} \sigma_{CI,2,0}^2 :=& E_{x_{i_1},x_{i_2}}\left[\left(E_{x_{q_1},x_{q_2}}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle]\right)^2\right], \\ \sigma_{CI,0,2}^2 :=& E_{x_{q_1},x_{q_2}}\left[\left(E_{x_{i_1},x_{i_2}}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle]\right)^2\right], \\ \sigma_{CI,1,1}^2 :=& E_{x_{i_1},x_{q_2}}\left[\left(E_{x_{i_2},x_{q_1}}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle]\right)^2\right]. \end{aligned}$$

Here, we define $\check{\varphi}_{i_1,q_1} := \alpha^*\tilde{\varphi}_{12}\left(x^{(i_1)}\right) + (1-\alpha^*)\tilde{\varphi}_{12}\left(x'^{(q_1)}\right)$.

The proof is based on the theory of two-sample V-statistics, as T_{CI} belongs to this class. By replacing the population distributions in the asymptotic expressions with their empirical counterparts, we can estimate the mean and variance from unlabeled data. The p-value of MT_{CI} is then computed using the approximated null distribution and compared against a predefined significance level.

5.2 Weakly-supervised Kernel MCI(WsKMCI) Test with True Mixture Proportions

To test the MCI assumption with unlabeled data, we define the null hypothesis $H_0 : X_1 \perp\!\!\!\perp X_2 \mid X_S, Y = y$ against the alternative $H_1 : X_1 \not\perp\!\!\!\perp X_2 \mid X_S, Y = y$. We assume the true mixture proportion $\alpha^* = \alpha_+$ for $y = 1$ or $\alpha^* = \alpha_-$ for $y = -1$ is given. Let k_S be a positive-definite and characteristic kernel on $\mathcal{X}_S$ with a feature mapping φ_S and RKHS $\mathcal{H}_S$. Our proposed test is based on the Kernel-based Conditional Independence (KCI) criterion (Zhang et al., 2011; Pogodin et al., 2025) applied to $F_{12S}^{\alpha^*}$:

$$\left\| E_{F_{12S}^{\alpha^*}}[(\varphi_1(X_1) - \mu_{X_1|X_S}(X_S)) \otimes \varphi_S(X_S) \otimes (\varphi_2(X_2) - \mu_{X_2|X_S}(X_S))] \right\|_{\mathcal{H}}^2$$

which is zero under H_0. Here, we define $\mu_{X_\tau|X_S}(X_S) := E_{F_{\tau S}^{\alpha^*}}[\varphi_\tau(X_\tau)|X_S], \forall\tau \in \{1, 2\}$. Analogously to the previous subsection, we define the empirical test statistic:

$$T_{MCI} := \left\| E_{\hat{F}_{12S}^{\alpha^*}}[(\varphi_1(X_1) - \mu_{X_1|X_S}(X_S)) \otimes \varphi_S(X_S) \otimes (\varphi_2(X_2) - \mu_{X_2|X_S}(X_S))] \right\|_{\mathcal{H}}^2.$$

The testing procedure is identical to that of T_{CI}. We approximate the null distribution with gamma distribution by estimating its asymptotic mean and variance as in Zhang et al. (2011). The following theorem establishes the asymptotic null distribution and the consistency of the test under the assumptions used in Fukumizu et al. (2007).

Theorem 5 (Asymptotic distribution of MT_{MCI}). *Denote $\ddot{X} := (X_1, X_S)$ and $k_{\ddot{\mathcal{X}}} := k_1 k_S$. Assume $\mathcal{H}_1 \subset L^2(F_1^{\alpha^*}), \mathcal{H}_2 \subset L^2(F_2^{\alpha^*})$, and $\mathcal{H}_S \subset L^2(F_S^{\alpha^*})$ where $L^2(P)$ is the space of the square integrable functions with probability P. Further assume that $k_{\ddot{\mathcal{X}}} k_2$ is a characteristic kernel on $(\mathcal{X}_1 \times \mathcal{X}_S) \times \mathcal{X}_2$, and that $\mathcal{H}_S + \mathbb{R}$ (the direct sum of the two RKHSs) is dense in $L^2\left(F_S^{\alpha^*}\right)$. Then,*

(i) Under H_0, we have

$$MT_{MCI} \xrightarrow{d} \sum_{i,j,q=1}^{\infty} \lambda_{1,i}\lambda_{2,j}\lambda_{S,q}\xi_{ijq}^2.$$

where $\lambda_{1,i}$, $\lambda_{2,j}$ and $\lambda_{S,q}$ are the eigenvalues of the integral operators associated with k_1, k_2 and k_S, and ξ_{ijq}s follow a multivariate normal distribution with mean $\mathbf{0}$ and covariances defined in Appendix B.2.

(ii) Under H_1, we have $MT_{MCI} \xrightarrow{p} \infty$.

In addition, the asymptotic mean and variance of MT_{MCI} are given in the following theorem. These theorems are proved similarly to Theorems 3 and 4.

Theorem 6 (Asymptotic mean and variance of T_{MCI}). *Define the centralized kernel $\tilde{k}_{12S}$ associated with the feature map $\tilde{\varphi}_{12S}(x) := (\varphi_1(x_1) - \mu_{X_1|X_S}(x_S)) \otimes \varphi_S(x_S) \otimes (\varphi_2(x_2) - \mu_{X_2|X_S}(x_S))$. Under H_0, as $M \to \infty$,*

$$\begin{aligned} E[MT_{MCI}] &\to \nu\alpha^{*2} E_{x_i,x_j}[\tilde{k}_{12S}(x_i,x_i) - \tilde{k}_{12S}(x_i,x_j)] \\ &\quad + \nu'(1-\alpha^*)^2 E_{x'_q,x'_r}[\tilde{k}_{12S}(x'_q,x'_q) - \tilde{k}_{12S}(x'_q,x'_r)] \\ V[MT_{MCI}] &\to 2\nu^2\sigma^2_{MCI,2,0} + 2\nu'^2\sigma^2_{MCI,0,2} \\ &\quad + 4\nu\nu'\sigma^2_{MCI,1,1} \end{aligned}$$

where x_{i_1},x_{i_2} and x'_{q_1},x'_{q_2} are i.i.d samples from U_{12S} and U'_{12S}, and

$$\begin{aligned} \sigma^2_{MCI,2,0} &:= E_{x_{i_1},x_{i_2}}\left[\left(E_{x_{q_1},x_{q_2}}\left[\langle\check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle\right]\right)^2\right], \\ \sigma^2_{MCI,0,2} &:= E_{x_{q_1},x_{q_2}}\left[\left(E_{x_{i_1},x_{i_2}}\left[\langle\check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle\right]\right)^2\right], \\ \sigma^2_{MCI,1,1} &:= E_{x_{i_1},x_{q_2}}\left[\left(E_{x_{i_2},x_{q_1}}\left[\langle\check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2}\rangle\right]\right)^2\right]. \end{aligned}$$

Here, we redefine $\check{\varphi}_{i_1,q_1} := \alpha^*\tilde{\varphi}_{12S}\left(x^{(i_1)}\right) + (1-\alpha^*)\tilde{\varphi}_{12S}\left(x'^{(q_1)}\right)$, analogously to the CI setting.

As with T_{CI}, the mean and variance are estimated using their empirical counterparts. For the MCI test, however, we must also estimate the conditional kernel mean $\mu_{X_\tau|X_S}$. Following the procedure of Zhang et al. (2011), we use the empirical kernel maps of k_1 and k_2 and then apply kernel ridge regression to estimate these conditional means. This estimation is performed in the weakly-supervised manner described in Section 4. Due to space constraints, full details are provided in Appendix D.1.

5.3 CI and MCI Test without True Mixture Proportions

The testing methods proposed in the previous subsections require known mixture proportions. Consequently, they cannot be used to verify Assumptions 1 and 2 to assess the MPE applicability, as these proportions are unknown in advance. Therefore, we propose a "plug-in" approach for the Cl and MCl tests without true mixture proportions. The validity of the plug-in test statistic is established in Lemma 3 and Theorem 7. In this subsection, we use the following definitions for each test.

$$\begin{aligned} T_{CI,\alpha} :=& \Big\| E_{\hat{F}^{\alpha}_{12}}\left[\varphi_1(X_1)\otimes\varphi_2(X_2)\right] - \\ & E_{\hat{F}^{\alpha}_1\hat{F}^{\alpha}_2}\left[\varphi_1(X_1)\otimes\varphi_2(X_2)\right]\Big\|^2_{\mathcal{H}} \\ T_{MCI,\alpha} :=& \Big\| E_{\hat{F}^{\alpha}_{12S}}[(\varphi_1(X_1) - \mu_{X_1|X_S}(X_S))\otimes\varphi_S(X_S)\otimes \\ & (\varphi_2(X_2) - \mu_{X_2|X_S}(X_S))]\Big\|^2_{\mathcal{H}} \end{aligned}$$

where $\mu_{X_\tau|X_S}(X_S) := E_{F^{\alpha^*}_{\tau S}}[\varphi_\tau(X_\tau)|X_S], \forall\tau \in \{1,2\}$. Note that $T_{CI,\alpha^*} = T_{CI}$ and $T_{MCI,\alpha^*} = T_{MCI}$.

Our proposed approach is as follows: we first estimate the mixture proportion by $\hat{\alpha}_{CI}$ (resp. $\hat{\alpha}_{MCI}$) and then use the plug-in test statistic $T_{CI,\hat{\alpha}_{CI}}$ (resp. $T_{MCI,\hat{\alpha}_{MCI}}$) [1] instead of the original statistic T_{CI,α^*} (resp. T_{MCI,α^*}). We use gamma approximation to derive the null distribution and conduct the statistical test, following a similar procedure to the ones from the previous subsections.

Since the mean and variance of the plug-in statistic deviate from those of the original statistic, we derive them based on the following lemma.

Lemma 3. *Denote $T_\alpha := T_{CI,\alpha}$ and $\hat{\alpha} := \hat{\alpha}_{CI}$ for the CI test ($T_\alpha := T_{MCI,\alpha}$ and $\hat{\alpha} := \hat{\alpha}_{MCI}$ for the MCI test). Then, the following convergence holds for both tests. Under H_0, as $M \to \infty$,*

$$MT_{\hat{\alpha}} - M\{T_{\alpha^*} + (\hat{\alpha}-\alpha^*)T'_{\alpha^*} + \tfrac{1}{2}(\hat{\alpha}-\alpha^*)^2 T''_{\alpha^*}\} \xrightarrow{p} 0$$

where $T'_{\alpha^*} := \frac{d}{d\alpha}T_\alpha|_{\alpha=\alpha^*}$ and $T''_{\alpha^*} := \frac{d^2}{d\alpha^2}T_\alpha|_{\alpha=\alpha^*}$.

[1] In practice, the true conditional kernel means required for the $T_{MCI,\hat{\alpha}_{MCI}}$ is estimated with $\hat{F}^{\hat{\alpha}_{MCI}}_{12S}$

This lemma is derived by the Taylor expansion of $T_{\hat{\alpha}}$ around α^*. Considering the probabilistic limit, we approximate the mean and variance of $MT_{\hat{\alpha}}$ for each test as follows. The asymptotic equalities hold under mild conditions such as uniform integrability.

$$E[MT_{\hat{\alpha}}] \simeq ME[T_{\alpha^*} + (\hat{\alpha} - \alpha^*)T'_{\alpha^*} + \frac{1}{2}(\hat{\alpha} - \alpha^*)^2 T''_{\alpha^*}]$$
$$V[MT_{\hat{\alpha}}] \simeq M^2 V[T_{\alpha^*} + (\hat{\alpha} - \alpha^*)T'_{\alpha^*} + \frac{1}{2}(\hat{\alpha} - \alpha^*)^2 T''_{\alpha^*}].$$

Each term on the r.h.s. can be estimated with the theory of U-statistics. However, as the derivation is rather complicated, we defer the full details to Appendix C.2.

To ensure the test correctly rejects H_0 under H_1 (the test consistency), the following additional assumption is required, since $\hat{\alpha} \xrightarrow{p} \alpha^*$ is not guaranteed under H_1.

Assumption 3. *(i) For the CI test, $\hat{\alpha}_{CI} \xrightarrow{p} \alpha_1$, such that $F_{12}^{\alpha_1}$ is a probability distribution and $F_{12}^{\alpha_1} \neq F_1^{\alpha_1} F_2^{\alpha_1}$.*

(ii) For the MCI test, $\hat{\alpha}_{MCI} \xrightarrow{p} \alpha_1$, such that $F_{12S}^{\alpha_1}$ is a probability distribution and $F_{12S}^{\alpha_1} \neq F_{1|S}^{\alpha_1} F_{2|S}^{\alpha_1} F_S^{\alpha_1}$.

Assumption 3 ensures that the limiting distributions $F_{12}^{\alpha_1}$ and $F_{12S}^{\alpha_1}$ do not exhibit spurious CI and MCI conditions, which could lead to accepting H_0 under H_1. This assumption is considered mild since $F_\tau^{\alpha_1}$ is a mixture distribution of P_τ and N_τ, and the features in X generally exhibit dependence. Under this assumption, we establish test consistency in the following theorem. The proof is provided in Appendix C.1.

Theorem 7. *Let Assumption 3 hold. Then,*

(i) For the CI test, under the assumptions in Theorem 3 and H_1, $MT_{CI,\hat{\alpha}} \xrightarrow{p} \infty$.

(ii) For the MCI test, under the assumptions in Theorem 5 and H_1, $MT_{MCI,\hat{\alpha}} \xrightarrow{p} \infty$.

6 EXPERIMENTS

6.1 Mixture Proportion Estimation

We implement our MPE methods with synthetic data and benchmark datasets taken from the UCI machine learning repository. The estimated $\alpha_\pm$ by our methods are converted to $(\hat{\theta}, \hat{\theta}') = (\frac{1-\hat{\alpha}_-}{\hat{\alpha}_+ - \hat{\alpha}_-}, \frac{-\hat{\alpha}_-}{\hat{\alpha}_+ - \hat{\alpha}_-})$ for comparison with existing MPE methods.

CI MPE with synthetic data To evaluate our method, we utilize Gaussian data and three datasets from the UCI machine learning repository. We compare our CI MPE method with three existing MPE algorithms: DEDPUL (Ivanov, 2020), EN (Elkan and Noto, 2008) and KM2 (Ramaswamy et al., 2016) that are developed under the irreducibility assumption or its variant. Since these baseline methods are designed for positive and unlabeled data, we set $\theta = 1$ and focused on estimating only θ'.

Gaussian data for each class is generated from a two-dimensional Gaussian distribution: $X_1 \sim \mathcal{N}(Y, 1), X_2 \sim \mathcal{N}(Y, 1)$. For UCI datasets, we choose Shuttle, Wine and Dry Bean datasets. The specific classes assigned as positive and negative for each dataset are detailed in Appendix D.2.1. The primary goal of this experiment is to validate our method when the CI assumption (Assumption 1) holds, while the irreducibility assumption is violated. To create this scenario, we modified the datasets by the following procedure. At first, 20% of the original positive data is transferred into negative data to break irreducibility. Then we split the features into two sets of equal dimension and sample each set independently, given the class Y, with replacement. This manually creates CI datasets. For each dataset, we conducted 3×10 experiments. We select θ' from $\{0.2, 0.5, 0.7\}$. For each θ', we performed 10 trials by randomly drawing $n = n' = 2000$ samples for the positive and unlabeled sets. Foxr our CI MPE method, we use identity functions for g_1 and g_2, as this setup performed well in preliminary experiments.

The results in Table 1 are the averages over the 10 trials for each θ'. The table shows that our method consistently outperforms the other methods, which yield unstable estimations when the irreducibility is violated.

CI MPE with real-world data We also implement experiments on real-world datasets to demonstrate the existence of CI features and the applicability of our MPE method.

In the experiments, we first use the HSIC test (Gretton et al., 2007) on labeled data to detect feature pairs (X_1, X_2) that are conditionally independent given the negative class, i.e., $X_1 \perp\!\!\!\perp X_2 | Y = -1$. Then, we conduct MPE experiments with our CI MPE method on the detected feature pairs. We use two datasets from the UCI repository: the Breast Cancer Wisconsin and Dry Bean datasets. For each dataset, we choose positive and negative classes and implement the experiments, switching the classes. The detailed procedure is provided in Appendix D.2.2.

For the MPE task, we set $n = n' = 2000$ and used a Positive-Unlabeled (PU) setting with class priors $(\theta, \theta') = (1, 0.5)$. The number of detected CI pairs and the resulting Mean Absolute Error (MAE) of all pairs and trials in each dataset are shown in Table 2. The results confirm the presence of CI pairs in real-world data and show that our MPE method is

Table 1: Mean absolute error for $\hat{\theta}'$ in the CI MPE experiment. The lowest value in each column is highlighted in bold.

	Gaussian	Shuttle	Wine	Dry Bean
DEDPUL	0.043	0.117	0.077	0.074
KM2	0.027	0.152	0.129	0.029
EN	0.063	0.075	0.110	0.107
CI MPE	**0.013**	**0.053**	**0.031**	**0.025**

Table 2: Number of detected CI pairs and mean absolute error of $\hat{\theta}'$

Dataset	Negative Class	# Detected CI pairs	MAE of $\hat{\theta}'$
Breast Cancer	B	88	0.0284 ± 0.0248
	M	86	0.0498 ± 0.0612
DryBean	SEKER	2	0.0255 ± 0.0252
	HOROZ	4	0.0179 ± 0.0075

Table 3: Mean and standard deviation of the absolute errors for $(\hat{\theta}, \hat{\theta}')$ in the MCI MPE experiment. The first row corresponds to the PU setting, in which only the results for θ' are shown.

(θ, θ')	$n = n' = 100$	500	1000
(1,0.2)	(− , 0.044 ± 0.037)	(− , 0.019 ± 0.012)	(− , 0.015 ± 0.010)
(0.8,0.2)	(0.048 ± 0.036, 0.044 ± 0.030)	(0.016 ± 0.013, 0.020 ± 0.014)	(0.015 ± 0.011, 0.014 ± 0.010)
(0.5,0.2)	(0.077 ± 0.079, 0.056 ± 0.048)	(0.031 ± 0.024, 0.025 ± 0.019)	(0.025 ± 0.017, 0.020 ± 0.014)

applicable in these scenarios.

MCI MPE with synthetic data In this experiment, we use Gaussian data that satisfies the MCI assumption (Assumption 2). We generate three-dimensional data as follows: $X_S \sim \mathcal{N}(0.5, 1), X_1 \sim \mathcal{N}(Y, 1) + X_S, X_2 \sim \mathcal{N}(Y, 1) + X_S$. We use identity functions for g_1 and g_2 as in CI MPE, and employ Gaussian kernels for KRR. The Golden section search method (Kiefer, 1953) is used to optimize $\hat{m}^2_{MCI}(\alpha)$. We performed 100 trials for each pair $(\theta, \theta') \in \{(1, 0.2), (0.8, 0.2), (0.5, 0.2)\}$. Further details, including the regularization parameter for KRR and the search range I_{α^*}, are available in Appendix D.2.3. The averaged results, presented in Table 3, suggest that our method successfully estimates θ and θ'. As shown in Theorem 2, the estimation errors tend to decrease as $(\theta - \theta')^2$ increases.

We also implement MCI MPE with real-world data. However, due to space constraints, these results are deferred to Appendix D.2.4.

6.2 Weakly-supervised Kernel CI and MCI Tests

We evaluate the performance of the proposed kernel CI and MCI tests with the following class-conditional Gaussian data for each test:

CI: $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} Y \\ Y \end{pmatrix}, \Sigma_Y\right)$

MCI: $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} Y \\ Y \end{pmatrix}, \Sigma_Y\right) + \begin{pmatrix} X_S \\ X_S \end{pmatrix}$ and $X_S \sim \mathcal{N}(0.5, 1)$.

Here, the covariance matrix is defined as $\Sigma_1 = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$ for the positive class, and the identity matrix $\Sigma_{-1} = I$ for the negative class. In these experiments, we specifically test the CI and MCI assumptions for the positive distribution.

To generate data under both the null and alternative hypotheses, we varied the covariance σ_{12}. The null hypothesis (H_0) corresponds to $\sigma_{12} = 0$, while the alternative (H_1) corresponds to $\sigma_{12} \neq 0$. A larger value of σ_{12} represents a greater deviation from the null hypothesis. For all experiments, we set the true

Table 4: Rejection rates for the kernel CI test (top) and MCI test (bottom) with true mixture proportions. They should be close to the significance level 0.05 under H_0 ($\sigma_{12} = 0$).

σ_{12}	$n = n' = 500$	1000	2000
0	0.051	0.055	0.052
0.2	0.399	0.748	0.996
0.5	1	1	1

σ_{12}	$n = n' = 500$	1000	2000
0	0.062	0.042	0.035
0.2	0.307	0.605	0.910
0.5	0.988	0.999	1

Table 5: Rejection rates for the kernel CI test (top) and MCI test (bottom) without true mixture proportions. They should be close to the significance level 0.05 under H_0 ($\sigma_{12} = 0$).

σ_{12}	$n = n' = 500$	1000	2000
0	0.041	0.038	0.042
0.2	0.573	0.915	0.994
0.5	1	1	1

σ_{12}	$n = n' = 1000$	2000	3000
0	0.040	0.050	0.055
0.2	0.207	0.484	0.743
0.5	0.726	0.92	0.951

mixture proportions to $(\theta, \theta') = (0.8, 0.2)$ and use a Gaussian kernel for our tests. We assess the tests' ability to control the Type I and Type II error rates by performing 1000 repetitions for each experimental setting, varying the sample size. The significance level is set to 0.05. Further details such as hyperparameters are provided in Appendix D.2.5.

Tests with true mixture proportions In this setting, the methods from Sections 5.1 and 5.2 are evaluated. As shown in Table 4, both methods effectively control the Type I error around the target significance level of 0.05. While statistical power is limited for smaller sample sizes ($n = 500$) when $\sigma_{12} = 0.2$, it improves significantly for larger n.

Tests without true mixture proportions We also evaluate the tests proposed in Section 5.3, with results presented in Table 5. This setting is more challenging as the true mixture proportions are unknown. Consequently, the Type II error rates are higher compared to the previous experiment, particularly for the MCI test. As we detail in Appendix D.3.2, the low power in the MCI test stems from the null and alternative distributions not being well-separated in small samples. This limitation can be mitigated by increasing the sample size at the expense of computational cost. From the perspective of our downstream MCI MPE application, this lower power is not highly problematic. Even when the test fails to reject H_0 for $\sigma_{12} = 0.2$, the resulting relative bias of the estimator $\hat{\alpha}_{MCI}$ is small (approximately 10%), as detailed in Appendix D.3.1.

7 CONCLUSIONS AND DISCUSSIONS

This work introduces novel identifiability conditions for mixture proportion estimation (MPE): the class-specific Conditional Independence (CI) and Multivariate Conditional Independence (MCI) assumptions. Based on these conditions, we propose method of moments estimators and establish their theoretical properties.

Another contribution is the development of weakly-supervised kernel-based statistical tests (WsKCI and WsKMCI) that validate these CI and MCI assumptions using only unlabeled data. These tests have potential applications such as causal discovery and fairness assessment. We empirically demonstrate the effectiveness of the proposed MPE estimators and statistical tests.

Several directions for future research arise from this study. First, the performance of our MPE methods depends on the choice of functions g_1 and g_2. Investigating how to choose g_1 and g_2 for the smaller estimation variance is an important direction. Second, conducting our statistical tests on high-dimensional data is challenging, as the number of candidate CI and MCI pairs can become large, which causes a multiple testing problem. Developing a method to efficiently find pairs that satisfy the MPE assumptions in high-dimensional data is left for future work. A third direction is to apply our test statistic as a regularizer for fair and domain-invariant representation learning, similarly to Pogodin et al. (2023).

Acknowledgement

TK was partially supported by JSPS KAKENHI Grant Numbers 20H00576, 23H03460, and 24K14849.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832.

Bao, H., Niu, G., and Sugiyama, M. (2018). Classification from pairwise similarity and unlabeled data. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 452–461. PMLR.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 109(4):719–760.

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA. Association for Computing Machinery.

du Plessis, M. C., Niu, G., and Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 213–220, New York, NY, USA. Association for Computing Machinery.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Garg, S., Balakrishnan, S., and Lipton, Z. (2022). Domain adaptation under open set label shift. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22531–22546. Curran Associates, Inc.

Giesen, J., Kahlmeyer, P., Laue, S., Mitterreiter, M., Nussbaum, F., Staudt, C., and Zarrieß, S. (2021). Method of moments for topic models with mixed discrete and continuous features. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2418–2424. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Gordon, S. L., Mazaheri, B., Rabani, Y., and Schulman, L. (2023). Causal inference despite limited global confounding via mixture models. In van der Schaar, M., Zhang, C., and Janzing, D., editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 574–601. PMLR.

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Huang, B., Liu, Y., and Peng, L. (2023). Weighted bootstrap for two-sample u-statistics. *Journal of Statistical Planning and Inference*, 226:86–99.

Ivanov, D. (2020). DEDPUL: Difference-of-estimated-densities-based positive-unlabeled learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 782–790. IEEE.

Jain, S., White, M., Trosset, M. W., and Radivojac, P. (2016). Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third edition draft edition.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4:502–506.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liu, M., Sun, X., Qiao, Y., and Wang, Y. (2024). Causal discovery via conditional independence testing with proxy variables. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Liu, T. and Tao, D. (2016). Classification with noisy labels by importance reweighting. In *IEEE Transactions on pattern analysis and machine intelligence*, pages 447–461.

Lu, N., Lei, S., Niu, G., Sato, I., and Sugiyama, M. (2021). Binary classification from multiple unlabeled datasets via surrogate set classification. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7134–7144. PMLR.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. (2019). On the minimal supervision for training any binary classifier from only unlabeled data. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mazaheri, B., Gordon, S., Rabani, Y., and Schulman, L. J. (2023). Causal discovery under latent class confounding. *CoRR*, abs/2311.07454.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Peters, J., Janzing, D., and Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Pogodin, R., Deka, N., Li, Y., Sutherland, D. J., Veitch, V., and Gretton, A. (2023). Efficient conditionally invariant representation learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pogodin, R., Schrab, A., Li, Y., Sutherland, D. J., and Gretton, A. (2025). Practical kernel tests of conditional independence.

Ramaswamy, H., Scott, C., and Tewari, A. (2016). Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR.

Sanderson, T. and Scott, C. (2014). Class Proportion Estimation with Application to Multiclass Anomaly Rejection. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 850–858, Reykjavik, Iceland. PMLR.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Schrab, A. (2025). A practical introduction to kernel discrepancies: MMD, HSIC & KSD.

Scott, C. (2015). A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846. PMLR.

Scott, C., Blanchard, G., and Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 489–511, Princeton, NJ, USA. PMLR.

Serfling, R. J. (1981). *Approximation theorems of mathematical statistics*. Wiley series in probability and statistics. Wiley, Hoboken, NJ.

Song, L., Anandkumar, A., Dai, B., and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 640–648, Bejing, China. PMLR.

Steinhardt, J. and Liang, P. (2016). Unsupervised risk estimation using only conditional independence structure. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3664–3672, Red Hook, NY, USA. Curran Associates Inc.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Yan, Y. and Lam, H. (2024). Data-driven solutions and uncertainty quantification for multistage stochastic optimization. In *2024 Winter Simulation Conference (WSC)*, pages 3300–3311.

Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. (2022). Rethinking class-prior estimation for positive-unlabeled learning. In *International Conference on Learning Representations*.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

Zhu, Y., Fjeldsted, A., Holland, D., Landon, G., Lintereur, A., and Scott, C. (2023). Mixture proportion estimation beyond irreducibility. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42962–42982. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Contents

A MIXTURE PROPORTION ESTIMATION WITH CI AND MCI **13**

A.1 Proofs for the MPE with CI 13

A.2 Proofs for the MPE with MCI 15

B WEAKLY-SUPERVISED KERNEL CI AND MCI TEST WITH TRUE MIXTURE PROPORTIONS **16**

B.1 Proofs for the CI test 16

B.2 Proofs for the MCI test 21

C WEAKLY-SUPERVISED KERNEL CI AND MCI TEST WITHOUT TRUE MIXTURE PROPORTIONS **23**

C.1 Proofs for the tests without true mixture proportions 23

C.2 Mean and variance estimation for the tests without true mixture proportions 23

D EXPERIMENTS **26**

D.1 Practical computation of test statistic 26

D.2 Experimental details 27

D.3 Additional experiments 28

A MIXTURE PROPORTION ESTIMATION WITH CI AND MCI

A.1 Proofs for the MPE with CI

Proof of Lemma 1. $m_{CI}(\alpha) = 0$ is a quadratic equation such that

$$m_{CI}(\alpha) = a\alpha^2 + b\alpha + c$$

where

$$\begin{aligned}
a &= E_{U_1U_2'}[g_{12}] + E_{U_1'U_2}[g_{12}] - E_{U_1U_2}[g_{12}] - E_{U_1'U_2'}[g_{12}],\\
b &= E_{U_{12}}[g_{12}] - E_{U_{12}'}[g_{12}] + 2E_{U_1'U_2'}[g_{12}] - E_{U_1U_2'}[g_{12}] - E_{U_1'U_2}[g_{12}],\\
c &= E_{U_{12}'}[g_{12}] - E_{U_1'U_2'}[g_{12}].
\end{aligned}$$

The distributions for the coefficient a are simplified as follows:

$$U_1U_2' + U_1U_2' - U_1U_2 - U_1'U_2' = -(U_1 - U_1')(U_2 - U_2') = -(\theta - \theta')^2(P_1 - N_1)(P_2 - N_2)$$

Therefore,

$$a = -(\theta - \theta')^2(E_{P_1}[g_1] - E_{N_1}[g_1]) \cdot (E_{P_2}[g_2] - E_{N_2}[g_2]).$$

Considering α^* is one solution of $m_{CI}(\alpha) = 0$, if $(E_{P_1}[g_1] - E_{N_1}[g_1]) \cdot (E_{P_2}[g_2] - E_{N_2}[g_2]) \neq 0$, $a \neq 0$ and there exist real solutions for $m_{CI}(\alpha) = 0$. □

Proof of Theorem 1. The proof first requires establishing the consistency of the estimator $\hat{\alpha}_{CI}$.

Lemma 4 (Consistency of $\hat{\alpha}_{CI}$)**.** *Assume α^* is the unique solution of $m_{CI}(\alpha) = 0$ in I_{α^*}. Then, $\hat{\alpha}_{CI}$ is a consistent estimator of α^*.*

Proof of Lemma 4. For a fixed $\alpha, \hat{m}^2_{CI}(\alpha)$ can be viewed as a two-sample V-statistic (Yan and Lam, 2024; Vaart, 1998). Following a similar procedure to the proof of Theorem 3 in Yan and Lam (2024), we can establish the uniform convergence: $\sup_{\alpha \in I_{\alpha^*}} \left|\hat{m}^2_{CI}(\alpha) - m^2_{CI}(\alpha)\right| \xrightarrow{p} 0$. Then, we can prove the consistency, $\hat{\alpha}_{CI} \xrightarrow{p} \alpha^*$, similarly to Theorem 5.7 of Vaart (1998). □

For simplicity, we denote $m_{CI}(\alpha)$, $\hat{m}_{CI}(\alpha)$, and $\hat{\alpha}_{CI}$ as $m(\alpha)$, $\hat{m}(\alpha)$ and $\hat{\alpha}$ respectively in this proof. Since $\hat{\alpha}_{CI}$ minimizes $\hat{m}^2(\alpha)$, by the first-order condition, we have

$$\frac{\partial \hat{m}^2}{\partial \alpha}(\hat{\alpha}) = 2\hat{m}(\hat{\alpha})\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha}) = 0, \tag{3}$$

where we denote $\frac{\partial \hat{m}^2}{\partial \alpha}(\hat{\alpha}) := \frac{\partial \hat{m}^2}{\partial \alpha}(\alpha)|_{\alpha=\hat{\alpha}}$.

On the other hand, by the mean value theorem, we get

$$\hat{m}(\hat{\alpha}) - \hat{m}(\alpha^*) = \frac{\partial \hat{m}}{\partial \alpha}(\tilde{\alpha})(\hat{\alpha} - \alpha^*) \tag{4}$$

for some $\tilde{\alpha}$ between α^* and $\hat{\alpha}$. Multiplying both sides of (4) by $\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})$ and applying (3) yields

$$-\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})\hat{m}(\alpha^*) = \frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})\frac{\partial \hat{m}}{\partial \alpha}(\tilde{\alpha})(\hat{\alpha} - \alpha^*),$$

and thus

$$\hat{\alpha} - \alpha^* = -\frac{\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})\hat{m}(\alpha^*)}{\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})\frac{\partial \hat{m}}{\partial \alpha}(\tilde{\alpha})}.$$

Here, $\frac{\partial \hat{m}}{\partial \alpha}(\hat{\alpha})$ and $\frac{\partial \hat{m}}{\partial \alpha}(\tilde{\alpha})$ converge to $\frac{\partial m}{\partial \alpha}(\alpha^*)$ in probability because $\hat{\alpha}, \tilde{\alpha} \xrightarrow{p} \alpha^*$ holds by consistency (Lemma 4). Therefore, assuming $\sqrt{M}\hat{m}(\alpha^*) \xrightarrow{d} \mathcal{D}$ for some distribution $\mathcal{D}$, we have

$$\sqrt{M}(\hat{\alpha} - \alpha^*) \xrightarrow{d} -\frac{1}{d_0}\mathcal{D},$$

where $d_0 := \frac{\partial m}{\partial \alpha}(\alpha^*)$.

Next, we identify the limiting distribution $\mathcal{D}$. We can write $\hat{m}(\alpha^*)$ as

$$\hat{m}(\alpha^*) = E_{\hat{F}^{\alpha^*}_{12}}[(g_1 - E_{\hat{F}^{\alpha^*}_1}[g_1]) \cdot (g_2 - E_{\hat{F}^{\alpha^*}_2}[g_2])].$$

We introduce $\hat{m}_c(\alpha^*)$ with the true centralization, defined as

$$\hat{m}_c(\alpha^*) := E_{\hat{F}^{\alpha^*}_{12}}[(g_1 - E_{F^{\alpha^*}_1}[g_1]) \cdot (g_2 - E_{F^{\alpha^*}_2}[g_2])].$$

These two quantities satisfy $\sqrt{M}(\hat{m}(\alpha^*) - \hat{m}_c(\alpha^*)) = o_p(1)$ because

$$\sqrt{M}(\hat{m}(\alpha^*) - \hat{m}_c(\alpha^*)) = \sqrt{M}(E_{\hat{F}^{\alpha^*}_1}[g_1] - E_{F^{\alpha^*}_1}[g_1]) \cdot (E_{\hat{F}^{\alpha^*}_2}[g_2] - E_{F^{\alpha^*}_2}[g_2]) \xrightarrow{p} 0,$$

where the first term,$\sqrt{M}(E_{\hat{F}^{\alpha^*}_1}[g_1] - E_{F^{\alpha^*}_1}[g_1])$ converges in distribution to a normal distribution by the Central Limit Theorem (CLT), while the second term $E_{\hat{F}^{\alpha^*}_2}[g_2] - E_{F^{\alpha^*}_2}[g_2]$ converges in probability to $\mathbf{0}$ by the Law of

Large number (LLN). Therefore, $\sqrt{M}\hat{m}(\alpha^*)$ and $\sqrt{M}\hat{m}_c(\alpha^*)$ converge in distribution to the same distribution $\mathcal{D}$ (if they converge). Let $\tilde{g}_{12} := (g_1 - E_{F_1^{\alpha^*}}[g_1]) \cdot (g_2 - E_{F_2^{\alpha^*}}[g_2])$ be the centralized function. Then we have

$$\begin{aligned}
\sqrt{M}\hat{m}_c(\alpha^*) =& \sqrt{M}E_{\hat{F}_{12}^{\alpha^*}}[\tilde{g}_{12}] \\
=& \alpha^*\sqrt{M}E_{\hat{U}_{12}}[\tilde{g}_{12}] + (1-\alpha^*)\sqrt{M}E_{\hat{U'}_{12}}[\tilde{g}_{12}] \\
=& \alpha^*\sqrt{M}E_{\hat{U}_{12}}[\tilde{g}_{12}] + (1-\alpha^*)\sqrt{M}E_{\hat{U'}_{12}}[\tilde{g}_{12}] - (\alpha^*\sqrt{M}E_{U_{12}}[\tilde{g}_{12}] + (1-\alpha^*)\sqrt{M}E_{U'_{12}}[\tilde{g}_{12}]) \\
=& \alpha^*\sqrt{M/n}\sqrt{n}(E_{\hat{U}_{12}}[\tilde{g}_{12}] - E_{U_{12}}[\tilde{g}_{12}]) + (1-\alpha^*)\sqrt{M/n'}\sqrt{n'}(E_{\hat{U}'_{12}}[\tilde{g}_{12}] - E_{U'_{12}}[\tilde{g}_{12}]) \\
\xrightarrow{d}& \mathcal{N}(0, \nu\alpha^{*2}V_{U_{12}}[\tilde{g}_{12}] + \nu'(1-\alpha^*)^2 V_{U'_{12}}[\tilde{g}_{12}]),
\end{aligned}$$

where the third equality holds under Assumption 1 and we used the CLT in the last convergence.

Combining these results, the asymptotic distribution of $\hat{\alpha}$ is

$$\sqrt{M}(\hat{\alpha} - \alpha^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\nu\alpha^{*2}V_{U_{12}}[\tilde{g}_{12}] + \nu'(1-\alpha^*)^2 V_{U'_{12}}[\tilde{g}_{12}]}{d_0^2}\right).$$

Finally, we analyze the derivative term d_0. For $\alpha^* = \alpha_+$, we have

$$\begin{aligned}
d_0 = \frac{\partial m}{\partial \alpha}(\alpha^*) &= E_{U_{12}}[g_{12}] - E_{U'_{12}}[g_{12}] - E_{F_1^{\alpha^*}(U_2 - U'_2)}[g_{12}] - E_{(U_1 - U'_1)F_2^{\alpha^*}}[g_{12}] \\
&= (\theta - \theta')(E_{F_{12}^{\alpha^*}}[g_{12}] - E_{F_{12}^{\bar{\alpha}^*}}[g_{12}]) - (\theta - \theta')(E_{F_1^{\alpha^*}F_2^{\alpha^*}}[g_{12}] - E_{F_1^{\alpha^*}F_2^{\bar{\alpha}^*}}[g_{12}]) \\
&\quad - (\theta - \theta')(E_{F_1^{\alpha^*}F_2^{\alpha^*}}[g_{12}] - E_{F_1^{\bar{\alpha}^*}F_2^{\alpha^*}}[g_{12}]) \\
&= -(\theta - \theta')E_{F_{12}^{\bar{\alpha}^*}}[\tilde{g}_{12}].
\end{aligned}$$

For $\alpha^* = \alpha_-$, we have $d_0 = (\theta - \theta')E_{F_{12}^{\bar{\alpha}^*}}[\tilde{g}_{12}]$. Then the desired asymptotic distribution is derived. □

A.2 Proofs for the MPE with MCI

Proof of Theorem 2. For simplicity, we denote $m_{MCI}(\alpha)$, $\hat{m}_{MCI}(\alpha)$, and $\hat{\alpha}_{MCI}$ as $m(\alpha)$, $\hat{m}(\alpha)$ and $\hat{\alpha}$ respectively in this proof. For any fixed α, $\hat{m}^2(\alpha)$ can be viewed as a two-sample V-statistic. We can show the uniform convergence $\sup_{\alpha \in I_{\alpha^*}} \left|\hat{m}^2(\alpha) - m^2(\alpha)\right| \xrightarrow{p} 0$ and prove the consistency of $\hat{\alpha}_{MCI}$, $\hat{\alpha} \xrightarrow{p} \alpha^*$ by an argument analogous to the proof of Lemma 4.

Furthermore, following the same procedure as in the proof of Theorem 1, if we assume $\sqrt{M}\hat{m}(\alpha^*) \xrightarrow{d} \mathcal{D}$ for some distribution $\mathcal{D}$, it follows that

$$\sqrt{M}(\hat{\alpha} - \alpha^*) \xrightarrow{d} -\frac{1}{d_0}\mathcal{D}$$

where $d_0 := \frac{\partial m}{\partial \alpha}(\alpha^*)$. By the CLT, we also have

$$\begin{aligned}
\sqrt{M}\hat{m}(\alpha^*) &= \alpha^*\sqrt{M}E_{\hat{U}_{12S}}[\tilde{g}_{12S}] + (1-\alpha^*)\sqrt{M}E_{\hat{U'}_{12S}}[\tilde{g}_{12S}] \\
&= \alpha^*\sqrt{M}E_{\hat{U}_{12S}}[\tilde{g}_{12S}] + (1-\alpha^*)\sqrt{M}E_{\hat{U'}_{12S}}[\tilde{g}_{12S}] - (\alpha^*\sqrt{M}E_{U_{12S}}[\tilde{g}_{12S}] \\
&\quad + (1-\alpha^*)\sqrt{M}E_{U'_{12S}}[\tilde{g}_{12S}]) \\
&= \alpha^*\sqrt{M/n}\sqrt{n}(E_{\hat{U}_{12S}}[\tilde{g}_{12S}] - E_{U_{12S}}[\tilde{g}_{12S}]) + (1-\alpha^*)\sqrt{M/n'}\sqrt{n'}(E_{\hat{U}'_{12S}}[\tilde{g}_{12S}] - E_{U'_{12S}}[\tilde{g}_{12S}]) \\
&\xrightarrow{d} \mathcal{N}(0, \nu\alpha^{*2}V_{U_{12S}}[\tilde{g}_{12S}] + \nu'(1-\alpha^*)^2 V_{U'_{12S}}[\tilde{g}_{12S}]).
\end{aligned}$$

The remaining part is d_0. We have

$$d_0 = \frac{\partial m}{\partial \alpha}(\alpha^*) = E_{U_{12}}[\tilde{g}_{12S}] - E_{U'_{12S}}[\tilde{g}_{12S}] + E_{F_{12S}^{\alpha^*}}[\frac{\partial}{\partial \alpha}(g_1 - \mu_1^\alpha)(g_2 - \mu_2^\alpha)|_{\alpha=\alpha^*}],$$

where we interchange differentiation and integration, assuming $\frac{\partial}{\partial \alpha}\mu_1^\alpha$ and $\frac{\partial}{\partial \alpha}\mu_2^\alpha$ are bounded.

Let us evaluate the derivative term inside the expectation:

$$\begin{aligned}\frac{\partial}{\partial\alpha}(g_1-\mu_1^\alpha)(g_2-\mu_2^\alpha)|_{\alpha=\alpha^*} =& -g_1\frac{\partial}{\partial\alpha}\mu_2^{\alpha^*}-g_2\frac{\partial}{\partial\alpha}\mu_1^{\alpha^*}+\mu_1^{\alpha^*}\frac{\partial}{\partial\alpha}\mu_2^{\alpha^*}+\mu_2^{\alpha^*}\frac{\partial}{\partial\alpha}\mu_1^{\alpha^*}\\ =&(\mu_1^{\alpha^*}-g_1)\frac{\partial}{\partial\alpha}\mu_2^{\alpha^*}+(\mu_2^{\alpha^*}-g_2)\frac{\partial}{\partial\alpha}\mu_1^{\alpha^*}\end{aligned}$$

and then $E_{F_{12S}^{\alpha^*}}[\frac{\partial}{\partial\alpha}(g_1-\mu_1^\alpha)(g_2-\mu_2^\alpha)|_{\alpha=\alpha^*}]=0$.

Therefore, the expression for d_0 is simplified to:

$$d_0=E_{U_{12S}}[\tilde{g}_{12S}]-E_{U'_{12S}}[\tilde{g}_{12S}]=\begin{cases}-(\theta-\theta')E_{F_{12S}^{\alpha^*}}[\tilde{g}_{12S}] & \text{if } \alpha^*=\alpha_+,\\ (\theta-\theta')E_{F_{12S}^{\alpha^*}}[\tilde{g}_{12S}] & \text{if } \alpha^*=\alpha_-.\end{cases}$$

□

B WEAKLY-SUPERVISED KERNEL CI AND MCI TEST WITH TRUE MIXTURE PROPORTIONS

In the proofs of this section, we denote $F_\tau^{\alpha^*}$ as F_τ for simplicity.

B.1 Proofs for the CI test

Proof of Theorem 3. We define the centralized kernel $\tilde{k}_{12}$ associated with the feature map $\tilde{\varphi}_{12}(x):=(\varphi_1(x_1)-E_{F_1}[\varphi_1(x_1)])\otimes(\varphi_2(x_2)-E_{F_2}[\varphi_2(x_2)])$. Then,

$$\begin{aligned}\tilde{k}_{12}(x,x')=&\left(k_1(x_1,x_1')-E_{z_1\sim F_1}k_1(x_1,z_1)-E_{z_1\sim F_1}k_1(x_1',z_1)+E_{z_1,z_1'\sim F_1}k_1(z_1,z_1')\right)\\ &\left(k_2(x_2,x_2')-E_{z_2\sim F_2}k_2(x_2,z_2)-E_{z_2\sim F_2}k_2(x_2',z_2)+E_{z_2,z_2'\sim F_2}k_2(z_2,z_2')\right).\end{aligned}$$

Since k_1 and k_2 are positive-definite kernels, by Mercer's theorem (Scholkopf and Smola, 2001), they can be expanded as $k_1(x_1,x_1')=\Sigma_{r=1}^\infty\lambda_{1,r}\phi_{1,r}(x_1)\phi_{1,r}(x_1')$ and $k_2(x_2,x_2')=\Sigma_{r=1}^\infty\lambda_{2,r}\phi_{2,r}(x_2)\phi_{2,r}(x_2')$ where $\lambda_{1,r},\lambda_{2,r}$ and $\phi_{1,r},\phi_{2,r}$ are eigenvalues and eigenfunctions. Since these expansions are absolutely convergent, applying Fubini-Tonelli theorem, we can write $\tilde{k}_{12}(x,x')$ as follows:

$$\begin{aligned}\tilde{k}_{12}(x,x')=&\left(k_1(x_1,x_1')-E_{z_1\sim F_1}k_1(x_1,z_1)-E_{z_1\sim F_1}k_1(x_1',z_1)+E_{z_1,z_1'\sim F_1}k_1(z_1,z_1')\right)\\ &\left(k_2(x_2,x_2')-E_{z_2\sim F_2}k_2(x_2,z_2)-E_{z_2\sim F_2}k_2(x_2',z_2)+E_{z_2,z_2'\sim F_2}k_2(z_2,z_2')\right)\\ =&\left(\sum_{r=1}^\infty\lambda_{1,r}\left(\phi_{1,r}(x_1)\phi_{1,r}(x_1')-\phi_{1,r}(x_1)E_{F_1}\phi_{1,r}(z_1)-\phi_{1,r}(x_1')E_{F_1}\phi_{1,r}(z_1)+E_{F_1}^2\phi_{1,r}(z_1)\right)\right)\\ &\left(\sum_{r=1}^\infty\lambda_{2,r}\left(\phi_{2,r}(x_2)\phi_{2,r}(x_2')-\phi_{2,r}(x_2)E_{F_2}\phi_{2,r}(z_2)-\phi_{2,r}(x_2')E_{F_2}\phi_{2,r}(z_2)+E_{F_2}^2\phi_{2,r}(z_2)\right)\right)\\ =&\left(\sum_{r=1}^\infty\lambda_{1,r}\left(\phi_{1,r}(x_1)-E_{F_1}\phi_{1,r}(z_1)\right)\left(\phi_{1,r}(x_1')-E_{F_1}\phi_{1,r}(z_1)\right)\right)\\ &\left(\sum_{r=1}^\infty\lambda_{2,r}\left(\phi_{2,r}(x_2)-E_{F_2}\phi_{2,r}(z_2)\right)\left(\phi_{2,r}(x_2')-E_{F_2}\phi_{2,r}(z_2)\right)\right)\\ =&\left(\sum_{r=1}^\infty\lambda_{1,r}\tilde{\phi}_{1,r}(x_1)\tilde{\phi}_{1,r}(x_1')\right)\left(\sum_{r=1}^\infty\lambda_{2,r}\tilde{\phi}_{2,r}(x_2)\tilde{\phi}_{2,r}(x_2')\right)\\ =&\sum_{i,j=1}^\infty\lambda_{1,i}\lambda_{2,j}\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{1,i}(x_1')\tilde{\phi}_{2,j}(x_2)\tilde{\phi}_{2,j}(x_2'),\end{aligned}$$

where we define $\tilde{\phi}_{1,r}(x_1)=\phi_{1,r}(x_1)-E_{F_1}\phi_{1,r}(z_1)$ and $\tilde{\phi}_{2,r}(x_2')=\phi_{2,r}(x_2')-E_{F_2}\phi_{2,r}(z_2)$.

Then the test statistic T_{CI} is written as follows with $\tilde{\phi}_{1,r}$ and $\tilde{\phi}_{2,r}$:

$$\begin{aligned}
T_{CI} &= \left\| E_{\hat{F}_{12}}\left[\varphi_1 \otimes \varphi_2\right] - E_{\hat{F}_1\hat{F}_2}\left[\varphi_1 \otimes \varphi_2\right] \right\|_{\mathcal{H}}^2 \\
&= \left\| E_{\hat{F}_{12}}\left[(\varphi_1 - E_{F_1}\varphi_1) \otimes (\varphi_2 - E_{F_2}\varphi_2)\right] - E_{\hat{F}_2\hat{F}_2}\left[(\varphi_1 - E_{F_1}\varphi_1) \otimes (\varphi_2 - E_{F_2}\varphi_2)\right] \right\|_{\mathcal{H}}^2 \\
&= E_{\hat{F}_{12},\hat{F}_{12}}\tilde{k}_{12}(x,x') - 2E_{\hat{F}_{12},\hat{F}_1\hat{F}_2}\tilde{k}_{12}(x,x') + E_{\hat{F}_1\hat{F}_2,\hat{F}_1\hat{F}_2}\tilde{k}_{12}(x,x') \\
&= \sum_{i,j=1}^{\infty} \lambda_{1,i}\lambda_{2,j}(E_{\hat{F}_{12},\hat{F}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{1,i}(x_1')\tilde{\phi}_{2,j}(x_2)\tilde{\phi}_{2,j}(x_2')\right] - 2E_{\hat{F}_{12},\hat{F}_1\hat{F}_2}\left[\cdots\right] \\
&\quad + E_{\hat{F}_1\hat{F}_2,\hat{F}_1\hat{F}_2}\left[\cdots\right]) \\
&= \sum_{i,j=1}^{\infty} \lambda_{1,i}\lambda_{2,j}\left(E_{\hat{F}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right] - E_{\hat{F}_1\hat{F}_2}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right]\right)^2.
\end{aligned}$$

Now we consider the asymptotic distribution of MT_{CI} under H_0. MT_{CI} can be written as

$$\begin{aligned}
MT_{CI} = \sum_{i,j=1}^{\infty} \lambda_{1,i}\lambda_{2,j} &\left(\sqrt{M}\alpha^* E_{\hat{U}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right] + \sqrt{M}(1-\alpha^*)E_{\hat{U}_{12}'}\left[\cdots\right]\right. \\
&- \left(\sqrt{M}\alpha^* E_{\hat{U}_1}\left[\tilde{\phi}_{1,i}(x_1)\right] + \sqrt{M}(1-\alpha^*)E_{\hat{U}_1'}\left[\cdots\right]\right) \\
&\left.\left(\alpha^* E_{\hat{U}_2}\left[\tilde{\phi}_{2,j}(x_2)\right] + (1-\alpha^*)E_{\hat{U}_2'}\left[\cdots\right]\right)\right)^2.
\end{aligned}$$

We denote T_{CI}^L as the partial sum of T_{CI} up to L-th eigenvalues and then

$$\begin{aligned}
MT_{CI}^L = \sum_{i,j=1}^{L} \lambda_{1,i}\lambda_{2,j} &\left(\underbrace{\sqrt{\frac{M}{n}}\sqrt{n}\alpha^* E_{\hat{U}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right] + \sqrt{\frac{M}{n'}}\sqrt{n'}(1-\alpha^*)E_{\hat{U}_{12}'}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right]}_{(a)}\right. \\
&- \left(\underbrace{\sqrt{\frac{M}{n}}\sqrt{n}\alpha^* E_{\hat{U}_1}\left[\tilde{\phi}_{1,i}(x_1)\right] + \sqrt{\frac{M}{n'}}\sqrt{n'}(1-\alpha^*)E_{\hat{U}_1'}\left[\tilde{\phi}_{1,i}(x_1)\right]}_{(b)}\right) \\
&\left.\left(\underbrace{\alpha^* E_{\hat{U}_2}\left[\tilde{\phi}_{2,j}(x_2)\right] + (1-\alpha^*)E_{\hat{U}_2'}\left[\tilde{\phi}_{2,j}(x_2)\right]}_{(c)}\right)\right)^2.
\end{aligned}$$

Under H_0, (a) and (b) are written as:

$$\begin{aligned}
(a) &= \sqrt{\frac{M}{n}}\sqrt{n}\alpha^*\left(E_{\hat{U}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right] - E_{U_{12}}[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)]\right) \\
&\quad + \sqrt{\frac{M}{n'}}\sqrt{n'}(1-\alpha^*)\left(E_{\hat{U}_{12}'}[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)] - E_{U_{12}'}[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)]\right), \\
(b) &= \sqrt{\frac{M}{n}}\sqrt{n}\alpha^*\left(E_{\hat{U}_1}\left[\tilde{\phi}_{1,i}(x_1)\right] - E_{U_1}[\tilde{\phi}_{1,i}(x_1)]\right) + \sqrt{\frac{M}{n'}}\sqrt{n'}(1-\alpha^*)\left(E_{\hat{U}_1'}[\tilde{\phi}_{1,i}(x_1)] - E_{U_1'}[\tilde{\phi}_{1,i}(x_1)]\right).
\end{aligned}$$

As $M \to \infty$, (a) and (b) converge to normal distributions from the CLT, while $(c) \xrightarrow{p} 0$ from the LLN. Combining these results, it follows that

$$MT_{CI}^L \xrightarrow{d} \sum_{i,j=1}^{L} \lambda_{1,i}\lambda_{2,j}\xi_{i,j}^2,$$

where $(\xi_{1,1} \dots \xi_{L,L})$ follows a multivariate normal distribution with a mean $\mathbf{0}$ and following covariances: $\forall i, j, i', j' \in [L]$,

$$Cov[\xi_{i,j}, \xi_{i',j'}] = \nu {\alpha^*}^2 Cov_{U_{12}}[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2), \tilde{\phi}_{1,i'}(x_1)\tilde{\phi}_{2,j'}(x_2)] + \nu'(1-\alpha^*)^2 Cov_{U'_{12}}[\tilde{\phi}_{1,i}(x'_1)\tilde{\phi}_{2,j}(x'_2), \tilde{\phi}_{1,i'}(x'_1)\tilde{\phi}_{2,j'}(x'_2)].$$

We now derived the asymptotic distribution of MT^L_{CI}. To derive the asymptotic distribution of MT_{CI}, we follow a similar procedure to the section 5.5.2 of Serfling (1981) . Then, we can show that the expectation of difference between MT_{CI} and MT^L_{CI} vanishes:

$$E\left[M|T_{CI} - T^L_{CI}|\right] = E\left[M \sum_{i,j>L}^{\infty} \lambda_{1,i}\lambda_{2,j}\left(E_{\hat{F}_{12}}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right] - E_{\hat{F}_1\hat{F}_2}\left[\tilde{\phi}_{1,i}(x_1)\tilde{\phi}_{2,j}(x_2)\right]\right)^2\right] \to 0,$$

as $L \to \infty$.

Next, we denote the limiting variable of MT^L_{CI} as $W^L_{\lim} := \sum_{i,j=1}^{L} \lambda_{1,i}\lambda_{2,j}\xi^2_{i,j}$, and define $W_{\lim} := \lim_{L\to\infty} W^L_{\lim}$. Then, we can show

$$E\left[|W^L_{\lim} - W_{\lim}|\right] = E\left[\sum_{i,j>L}^{\infty} \lambda_{1,i}\lambda_{2,j}\xi^2_{i,j}\right] = \sum_{i,j>L}^{\infty} \lambda_{1,i}\lambda_{2,j}E\left[\xi^2_{i,j}\right] \to 0$$

as $L \to \infty$. These results allow us to prove the pointwise convergence of the characteristic functions. Specifically, for any t and $\epsilon > 0$, and for all sufficiently large M and L,

$$\begin{aligned}
&\left|E\left[e^{itMT_{CI}}\right] - E\left[e^{itW_{\lim}}\right]\right| \\
&= \left|\left(E\left[e^{itMT_{CI}}\right] - E\left[e^{itMT^L_{CI}}\right]\right) + \left(E\left[e^{itMT^L_{CI}}\right] - E\left[e^{itW^L_{\lim}}\right]\right) + \left(E\left[e^{itW^L_{\lim}}\right] - E\left[e^{itW_{\lim}}\right]\right)\right| \\
&\le \left|E\left[e^{itMT_{CI}}\right] - E\left[e^{itMT^L_{CI}}\right]\right| + \left|E\left[e^{itMT^L_{CI}}\right] - E\left[e^{itW^L_{\lim}}\right]\right| + \left|E\left[e^{itW^L_{\lim}}\right] - E\left[e^{itW_{\lim}}\right]\right] \\
&\le |t|E\left[M\left|T_{CI} - T^L_{CI}\right|\right] + \left|E\left[e^{itMT^L_{CI}}\right] - E\left[e^{itW^L_{\lim}}\right]\right| + |t|E\left[\left|W^L_{\lim} - W_{\lim}\right|\right] \le \epsilon,
\end{aligned}$$

where we used the inequality $\left|e^{iz} - 1\right| \le |z|$. Thus, the asymptotic distribution of MT_{CI} is:

$$MT_{CI} \xrightarrow{d} W_{\lim} = \sum_{i,j=1}^{\infty} \lambda_{1,i}\lambda_{2,j}\xi^2_{i,j},$$

where $\xi_{i,j}$'s follow the multivariate normal distribution defined above.

We next consider the asymptotic behavior of MT_{CI} under H_1. In this case, from Theorem 2 in Gretton (2015), the population version of T_{CI} equals some positive value c. Since T_{CI} is a two-sample V-statistic and a consistent estimator (Huang et al., 2023), $T_{CI} \xrightarrow{p} c$, as $M \to \infty$. Therefore, $MT_{CI} \xrightarrow{p} \infty$, as $M \to \infty$. □

Proof of Theorem 4. In this proof, we use the property of V-statistics. T_{CI} is a two-sample V-statistic (Vaart, 1998) since it can be written as follows.

$$T_{CI} = \frac{1}{n^6 n'^6} \sum_{i_1,\dots,i_6=1}^{n} \sum_{q_1,\dots,q_6=1}^{n'} h_{i_1,\dots,i_6,q_1,\dots,q_6}$$

where $h_{i_1,\dots,i_6,q_1,\dots,q_6}$ is a symmetric function such that

$$h_{i_1,\dots,i_6,q_1,\dots,q_6} := \frac{1}{6!6!} \sum_{(j_1,\dots,j_6)}^{(i_1,\dots,i_6)} \sum_{(r_1,\dots,r_6)}^{(q_1,\dots,q_6)} \langle \varphi_{j_1,\dots,j_3,r_1,\dots,r_3}, \varphi_{j_4,\dots,j_6,r_4,\dots,r_6} \rangle$$

and

$$\varphi_{j_1,...,j_3,r_1,...,r_3} := \alpha^*\tilde{\varphi}_{12}(x^{(j_1)}) + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(r_1)}) - (\alpha^*\tilde{\varphi}_1(x_1^{(j_2)}) + (1-\alpha^*)\tilde{\varphi}_1(x_1'^{(r_2)})) \otimes (\alpha^*\tilde{\varphi}_2(x_2^{(j_3)}) + (1-\alpha^*)\tilde{\varphi}_2(x_2'^{(r_3)})).$$

Here, the summation $\sum_{(j_1,...,j_6)}^{(i_1,...,i_6)}$ is taken over all ordered $(j_1, ..., j_6)$ drawn without replacement from $(i_1, ..., i_6)$. This is a degenerate V-statistic, which means

$$E_{i_2,...,i_6,q_1,...,q_6}[h_{i_1,...,i_6,q_1,...,q_6}] = E_{i_1,...,i_6,q_2,...,q_6}[h_{i_1,...,i_6,q_1,...,q_6}] = 0$$

where we take the expectations by samples of each index in the subscripts.

Next, we define a related statistic, $\check{T}_{CI} := E_{x,x'\sim\hat{F}_{12}}[\tilde{k}_{12}(x,x')]$. We prove the limit mean and variance of MT_{CI} and $M\check{T}_{CI}$ are the same, and derive the mean and variance of $M\check{T}_{CI}$.

$\check{T}_{CI}$ is a two-sample V-statistics and written as

$$\check{T}_{CI} = \frac{1}{n^2n'^2}\sum_{i_1,i_2=1}^{n}\sum_{q_1,q_2=1}^{n'}\check{h}_{i_1,i_2,q_1,q_2}$$

where $\check{h}_{i_1,i_2,q_1,q_2}$ is a symmetric function such that

$$\check{h}_{i_1,i_2,q_1,q_2} := \frac{1}{2!2!}\sum_{(j_1,j_2)}^{(i_1,i_2)}\sum_{(r_1,r_2)}^{(q_1,q_2)}\langle\check{\varphi}_{j_1,r_1},\check{\varphi}_{j_2,r_2}\rangle$$

and $\check{\varphi}_{j_1,r_1} := \alpha^*\tilde{\varphi}_{12}(x^{(j_1)}) + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(r_1)})$. $\check{T}_{CI}$ is also degenerate since $E_{i_2,q_1,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] = E_{i_1,i_2,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] = 0$.

Furthermore, we consider the difference $T_{CI} - \check{T}_{CI}$, which itself is a V-statistic:

$$\begin{aligned} T_{CI} - \check{T}_{CI} &= \left\langle 2E_{\hat{F}_{12}}[\tilde{\varphi}_{12}(x)] - E_{\hat{F}_1}[\tilde{\varphi}_1(x_1)] \otimes E_{\hat{F}_2}[\tilde{\varphi}_2(x_2)], -E_{\hat{F}_1}[\tilde{\varphi}_1(x_1)] \otimes E_{\hat{F}_2}[\tilde{\varphi}_2(x_2)]\right\rangle \\ &= \frac{1}{n^5n'^5}\sum_{i_1,...,i_5=1}^{n}\sum_{q_1,...,q_5=1}^{n'}\bar{h}_{i_1,...,i_5,q_1,...,q_5} \end{aligned}$$

where $\bar{h}_{i_1,...,i_5,q_1,...,q_5}$ is a symmetric function such that

$$\bar{h}_{i_1,...,i_5,q_1,...,q_5} := \frac{1}{5!5!}\sum_{(j_1,...,j_5)}^{(i_1,...,i_5)}\sum_{(r_1,...,r_5)}^{(q_1,...,q_5)}\langle\bar{\varphi}_{j_1,...,j_3,r_1,...,r_3},\bar{\varphi}_{j_4,j_5,r_4,r_5}\rangle$$

and we define

$$\bar{\varphi}_{j_1,...,j_3,r_1,...,r_3} := 2\alpha^*\tilde{\varphi}_{12}(x^{(j_1)}) + 2(1-\alpha^*)\tilde{\varphi}_{12}(x'^{(r_1)}) - (\alpha^*\tilde{\varphi}_1(x_1^{(j_2)}) + (1-\alpha^*)\tilde{\varphi}_1(x_1'^{(r_2)})) \otimes (\alpha^*\tilde{\varphi}_2(x_2^{(j_3)}) + (1-\alpha^*)\tilde{\varphi}_2(x_2'^{(r_3)}))$$

and

$$\bar{\varphi}_{j_4,j_5,r_4,r_5} := -(\alpha^*\tilde{\varphi}_1(x_1^{(j_4)}) + (1-\alpha^*)\tilde{\varphi}_1(x_1'^{(r_4)})) \otimes (\alpha^*\tilde{\varphi}_2(x_2^{(j_5)}) + (1-\alpha^*)\tilde{\varphi}_2(x_2'^{(r_5)})).$$

We next analyze the expectation $E[T_{CI} - \check{T}_{CI}]$. Since $E[\langle\bar{\varphi}_{j_1,...,j_3,r_1,...,r_3},\bar{\varphi}_{j_4,j_5,r_4,r_5}\rangle] = 0$ unless at least two pairs of indices in $j_1, ..., j_5, r_1, ..., r_5$ are equivalent,

$$E[T_{CI} - \check{T}_{CI}] = \frac{1}{n^5n'^5}O(M^8).$$

Thus, $ME[T_{CI} - \check{T}_{CI}] \to 0$ as $M \to \infty$. Since the expectations of MT_{CI} and $M\check{T}_{CI}$ are asymptotically equivalent, we now focus on analyzing $E[\check{T}_{CI}]$, which is easier to obtain.

$E[\check{T}_{CI}]$ is derived by a similar approach to the estimation of $E[HSIC_b]$ in Gretton et al. (2007). First, the V-statistic $\check{T}_{CI}$ is expanded as

$$\check{T}_{CI} = \frac{\alpha^{*2}}{n^2} \sum_{i_1,i_2=1}^{n} \tilde{k}_{12}(x^{(i_1)}, x^{(i_2)}) + \frac{(1-\alpha^*)^2}{n'^2} \sum_{q_1,q_2=1}^{n'} \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_2)}) + 2\frac{\alpha^*(1-\alpha^*)}{nn'} \sum_{i_1=1}^{n} \sum_{q_1=1}^{n'} \tilde{k}_{12}(x^{(i_1)}, x'^{(q_1)}).$$

Next, we define the corresponding U-statistic for $\check{T}_{CI}$ as $\check{T}_{CI,U}$:

$$\check{T}_{CI,U} = \frac{\alpha^{*2}}{(n)_2} \sum_{i_1 \neq i_2} \tilde{k}_{12}(x^{(i_1)}, x^{(i_2)}) + \frac{(1-\alpha^*)^2}{(n')_2} \sum_{q_1 \neq q_2} \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_2)}) + 2\frac{\alpha^*(1-\alpha^*)}{nn'} \sum_{i_1=1}^{n} \sum_{q_1=1}^{n'} \tilde{k}_{12}(x^{(i_1)}, x'^{(q_1)}),$$

where $(n)_m := \frac{n!}{(n-m)!}$. Note that the U-statistic, $\check{T}_{CI,U}$ is an unbiased estimator of its population mean, and $E[\check{T}_{CI,U}] = 0$.

The difference between the V-statistic and the U-statistic is given by:

$$\begin{aligned}\check{T}_{CI} - \check{T}_{CI,U} = &\frac{\alpha^{*2}}{n^2} \sum_{i_1=1}^{n} \tilde{k}_{12}(x^{(i_1)}, x^{(i_1)}) - \frac{\alpha^{*2}}{n(n)_2} \sum_{i_1 \neq i_2} \tilde{k}_{12}(x^{(i_1)}, x^{(i_2)}) \\ &+ \frac{(1-\alpha^*)^2}{n'^2} \sum_{q_1=1}^{n'} \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_1)}) - \frac{(1-\alpha^*)^2}{n'(n')_2} \sum_{q_1 \neq q_2} \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_2)}).\end{aligned}$$

Since $E[\check{T}_{CI,U}] = 0$, taking the expectation of the equation above yields the desired result:

$$\begin{aligned}E[\check{T}_{CI}] &= E[\check{T}_{CI} - \check{T}_{CI,U}] \\ &= \frac{\alpha^{*2}}{n} E_{x^{(i_1)},x^{(i_2)} \sim U_{12}}[\tilde{k}_{12}(x^{(i_1)}, x^{(i_1)}) - \tilde{k}_{12}(x^{(i_1)}, x^{(i_2)})] \\ &+ \frac{(1-\alpha^*)^2}{n'} E_{x'^{(q_1)},x'^{(q_2)} \sim U'_{12}}[\tilde{k}_{12}(x'^{(q_1)}, x'^{(q_1)}) - \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_2)})],\end{aligned}$$

from which the limit of $E[MT_{CI}]$ is obtained.

Next, we derive $V[T_{CI}]$. Using the expression of the V-statistic, we have:

$$V[T_{CI}] = \frac{1}{n^{12}n'^{12}} \sum_{i_1,\ldots,i_6,i'_1,\ldots,i'_6=1}^{n} \sum_{q_1,\ldots,q_6,q'_1,\ldots,q'_6=1}^{n'} Cov[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}, h_{i'_1,\ldots,i'_6,q'_1,\ldots,q'_6}].$$

Recall that T_{CI} is a degenerate V-statistic and $E_{i_2,\ldots,i_6,q_1,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}] = E_{i_1,\ldots,i_6,q_2,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}] = 0$. Thus, in order for $Cov[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}, h_{i'_1,\ldots,i'_6,q'_1,\ldots,q'_6}]$ to be nonzero, at least two pairs of indices must be identical between the sets $\{i_1, \ldots, i_6, q_1, \ldots, q_6\}$ and $\{i'_1, \ldots, i'_6, q'_1, \ldots, q'_6\}$. Using this combinatorial constraint, we can identify the leading order terms as $M \to \infty$.

We restrict our focus to combinations where exactly two indices overlap, as sharing more variables reduces the free choices from the sample, making those terms asymptotically negligible. Specifically, there are three cases for sharing exactly two variables: (1) two shared i indices and zero shared q indices, (2) zero shared i indices and two shared q indices, and (3) one shared i index and one shared q index.

In case (1), for the i indices, we choose 6 distinct indices for the first h function ($(n)_6$ ways), select 2 of these to share (C_2^6 ways), arrange them in the second h function (6×5 ways), and fill its remaining 4 slots with unselected indices ($(n-6)_4$ ways). This yields a total of $\mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{10}(n')_{12}$ combinations. By symmetry, case (2) yields $\mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{12}(n')_{10}$ combinations. For case (3), sharing one i index and one q index yields

$(n)_6 \times 6 \times 6 \times (n-6)_5 \times (n')_6 \times 6 \times 6 \times (n'-6)_5 = 6^2 \cdot 6^2 (n)_{11}(n')_{11}$ combinations. By considering only these leading order terms, we obtain:

$$\begin{aligned}
V[T_{CI}] &= \frac{1}{n^{12}n'^{12}}(\mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{10}(n')_{12} Cov[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}, h_{i_1,i_2,i'_3,\ldots,i'_6,q'_1,\ldots,q'_6}] \\
&+ \mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{12}(n')_{10} Cov[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}, h_{i'_1,\ldots,i'_6,q_1,q_2,q'_3,\ldots,q'_6}] \\
&+ 6^2 \cdot 6^2 (n)_{11}(n')_{11} Cov[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}, h_{i_1,i'_2,\ldots,i'_6,q_1,q'_2,\ldots,q'_6}] + O(M^{21})) \\
&= \frac{1}{n^{12}n'^{12}}(\mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{10}(n')_{12} E_{i_1,i_2}[(E_{i_3,\ldots,i_6,q_1,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] \\
&+ \mathrm{C}_2^6 \cdot 6 \cdot 5(n)_{12}(n')_{10} E_{q_1,q_2}[(E_{i_1,\ldots,i_6,q_3,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] \\
&+ 6^2 \cdot 6^2 (n)_{11}(n')_{11} E_{i_1,q_1}[(E_{i_2,\ldots,i_6,q_2,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] + O(M^{21})).
\end{aligned}$$

We simplify the each term in the above equation. Recall that $h_{i_1,\ldots,i_6,q_1,\ldots,q_6}$ is the sum of inner products $\langle \varphi_{j_1,\ldots,j_3,r_1,\ldots,r_3}, \varphi_{j_4,\ldots,j_6,r_4,\ldots,r_6} \rangle$. Since the feature map $\varphi_{j_1,\ldots,j_3,r_1,\ldots,r_3}$ are centered, the expectation vanishes if all sample indices are distinct. Taking this into account, we obtain

$$\begin{aligned}
E_{i_1,i_2}[(E_{i_3,\ldots,i_6,q_1,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] &= (\frac{1}{6!6!} \cdot 2 \cdot 4!6!)^2 E_{i_1,i_2}[(E_{q_1,q_2}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2] \\
E_{q_1,q_2}[(E_{i_1,\ldots,i_6,q_3,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] &= (\frac{1}{6!6!} \cdot 2 \cdot 4!6!)^2 E_{q_1,q_2}[(E_{i_1,i_2}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2] \\
E_{i_1,q_1}[(E_{i_2,\ldots,i_6,q_2,\ldots,q_6}[h_{i_1,\ldots,i_6,q_1,\ldots,q_6}])^2] &= (\frac{1}{6!6!} \cdot 2 \cdot 5!5!)^2 E_{i_1,q_2}[(E_{i_2,q_1}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2].
\end{aligned}$$

Substituting these expressions into the formula for $V[T_{CI}]$ and taking the limit as $M \to \infty$, we derive the desired asymptotic variance:

$$\begin{aligned}
V[MT_{CI}] = M^2 V[T_{CI}] \to & 2\nu^2 E_{i_1,i_2}[(E_{q_1,q_2}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2] + 2\nu'^2 E_{q_1,q_2}[(E_{i_1,i_2}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2] \\
& + 4\nu\nu' E_{i_1,q_2}[(E_{i_2,q_1}[\langle \check{\varphi}_{i_1,q_1}, \check{\varphi}_{i_2,q_2} \rangle])^2].
\end{aligned}$$

□

B.2 Proofs for the MCI test

Proof of Theorem 5. We begin by defining the centered kernels $\tilde{k}_{1S}(x_{1S}, x'_{1S}) = \langle \varphi_1(x_1) - \mu_{X_1|X_S}(x_S), \varphi_1(x'_1) - \mu_{X_1|X_S}(x'_S) \rangle$ and $\tilde{k}_{2S}(x_{2S}, x'_{2S}) = \langle \varphi_2(x_2) - \mu_{X_2|X_S}(x_S), \varphi_2(x'_2) - \mu_{X_2|X_S}(x'_S) \rangle$. By Mercer's theorem, these kernels can be expanded

$$\begin{aligned}
\tilde{k}_{1S}(x_{1S}, x'_{1S}) &= \sum_{r=1}^{\infty} \lambda_{1,r}(\phi_{1,r}(x_1) - E_F[\phi_{1,r}(x_1)|x_S])(\phi_{1,r}(x'_1) - E_F[\phi_{1,r}(x_1)|x'_S]) \\
&= \sum_{r=1}^{\infty} \lambda_{1,r} \tilde{\phi}_{1,r}(x_{1S}) \tilde{\phi}_{1,r}(x'_{1S}), \\
\tilde{k}_{2S}(x_{2S}, x'_{2S}) &= \sum_{r=1}^{\infty} \lambda_{2,r}(\phi_{2,r}(x_2) - E_F[\phi_{2,r}(x_2)|x_S])(\phi_{2,r}(x'_2) - E_F[\phi_{2,r}(x_2)|x'_S]) \\
&= \sum_{r=1}^{\infty} \lambda_{2,r} \tilde{\phi}_{2,r}(x_{2S}) \tilde{\phi}_{2,r}(x'_{2S}), \\
k_S(x_S, x'_S) &= \sum_{r=1}^{\infty} \lambda_{S,r} \phi_{S,r}(x_S) \phi_{S,r}(x'_S),
\end{aligned}$$

where $\lambda_{S,r}$ and $\phi_{S,r}$ are eigenvalues and eigenfunctions of the operator associated to k_S. We also define centered eigenfunctions $\tilde{\phi}_{1,r}(x_{1S}) := \phi_{1,r}(x_1) - E_F[\phi_{1,r}(x_1)|x_S]$ and $\tilde{\phi}_{2,r}(x_{2S}) := \phi_{2,r}(x_2) - E_F[\phi_{2,r}(x_2)|x_S]$ in this proof.

Then MT_{MCI} can be expressed as

$$\begin{aligned}
MT_{MCI} &= ME_{\hat{F}_{12S},\hat{F}_{12S}}[\tilde{k}_{1S}(x_{1S},x'_{1S})k_S(x_S,x'_S)\tilde{k}_{2S}(x_{2S},x'_{2S})] \\
&= M\sum_{i,j,q=1}^{\infty}\lambda_{1,i}\lambda_{2,j}\lambda_{S,q}E^2_{\hat{F}_{12S}}[\tilde{\phi}_{1,r}(x_{1S})\phi_{S,r}(x_S)\tilde{\phi}_{2,r}(x_{2S})] \\
&= M\sum_{i,j,q=1}^{\infty}\lambda_{1,i}\lambda_{2,j}\lambda_{S,q}(\alpha^*E_{\hat{U}_{12S}}[\tilde{\phi}_{1,r}(x_{1S})\phi_{S,r}(x_S)\tilde{\phi}_{2,r}(x_{2S})]-\alpha^*E_{U_{12S}}[\cdots] \\
&+(1-\alpha^*)E_{\hat{U}'_{12S}}[\tilde{\phi}_{1,r}(x_{1S})\phi_{S,r}(x_S)\tilde{\phi}_{2,r}(x_{2S})]-(1-\alpha^*)E_{U'_{12S}}[\cdots])^2 \\
&= \sum_{i,j,q=1}^{\infty}\lambda_{1,i}\lambda_{2,j}\lambda_{S,q}\big(\sqrt{\frac{M}{n}}\sqrt{n}\alpha^*(E_{\hat{U}_{12S}}[\tilde{\phi}_{1,r}(x_{1S})\phi_{S,r}(x_S)\tilde{\phi}_{2,r}(x_{2S})]-E_{U_{12S}}[\cdots]) \\
&+\sqrt{\frac{M}{n'}}\sqrt{n'}(1-\alpha^*)(E_{\hat{U}'_{12S}}[\tilde{\phi}_{1,r}(x_{1S})\phi_{S,r}(x_S)\tilde{\phi}_{2,r}(x_{2S})]-E_{U'_{12S}}[\cdots])\big)^2.
\end{aligned}$$

Following a similar procedure to the proof of Theorem 3, we can show the distributional convergence:

$$MT_{CI}\xrightarrow{d}\sum_{i,j,q=1}^{\infty}\lambda_{1,i}\lambda_{2,j}\lambda_{S,q}\xi^2_{ijq},$$

where $(\dots,\xi_{ijq},\dots)$ follows a multivariate normal distribution of mean $\mathbf{0}$ and covariances

$$\begin{aligned}
Cov[\xi_{ijq},\xi_{i'j'q'}] &= \nu{\alpha^*}^2Cov_{U_{12S}}[\tilde{\phi}_{1,i}(x_{1S})\phi_{S,j}(x_S)\tilde{\phi}_{2,q}(x_{2S}),\tilde{\phi}_{1,i'}(x_{1S})\phi_{S,j'}(x_S)\tilde{\phi}_{2,q'}(x_{2S})] \\
&+\nu'(1-\alpha^*)^2Cov_{U'_{12S}}[\tilde{\phi}_{1,i}(x_{1S})\phi_{S,j}(x_S)\tilde{\phi}_{2,q}(x_{2S}),\tilde{\phi}_{1,i'}(x_{1S})\phi_{S,j'}(x_S)\tilde{\phi}_{2,q'}(x_{2S})].
\end{aligned}$$

Finally, we consider the behavior of MT_{MCI} under H_1. In this case, according to Theorem 3 in Fukumizu et al. (2007), T_{MCI} converges to a positive value. Therefore, $MT_{MCI}\xrightarrow{p}\infty$, as $M\to\infty$. □

Proof of Theorem 6. The statistic T_{MCI} can be written as

$$T_{MCI}=\frac{{\alpha^*}^2}{n^2}\sum_{i_1,i_2=1}^{n}\tilde{k}_{12S}(x^{(i_1)},x^{(i_2)})+\frac{(1-\alpha^*)^2}{n'^2}\sum_{q_1,q_2=1}^{n'}\tilde{k}_{12S}(x'^{(q_1)},x'^{(q_2)})+2\frac{\alpha^*(1-\alpha^*)}{nn'}\sum_{i_1=1}^{n}\sum_{q_1=1}^{n'}\tilde{k}_{12S}(x^{(i_1)},x'^{(q_1)}),$$

where $\tilde{k}_{12S}$ is a kernel associated with a feature map $\tilde{\varphi}_{12S}(x):=(\varphi_1(x_1)-\mu_{x_1|x_S}(x_S))\otimes\varphi_S(x_S)\otimes(\varphi_2(x_2)-\mu_{x_2|x_S}(x_S))$.

This expression corresponds to a two-sample V-statistic, which can be rewritten as

$$T_{MCI}=\frac{1}{n^2n'^2}\sum_{i_1,i_2=1}^{n}\sum_{q_1,q_2=1}^{n'}f_{i_1,i_2,q_1,q_2},$$

where we define a symmetric function

$$f_{i_1,i_2,q_1,q_2}:=\frac{1}{4}\sum_{(j_1,j_2)}^{(i_1,i_2)}\sum_{(r_1,r_2)}^{(q_1,q_2)}\langle\alpha^*\tilde{\varphi}_{12S}(x^{(j_1)})+(1-\alpha^*)\tilde{\varphi}_{12S}(x'^{(r_1)}),\alpha^*\tilde{\varphi}_{12S}(x^{(j_2)})+(1-\alpha^*)\tilde{\varphi}_{12S}(x'^{(r_2)})\rangle.$$

Then, similarly to the proof of Theorem 4, we can derive the desired limits of $E[MT_{MCI}]$ and $V[MT_{MCI}]$. □

C WEAKLY-SUPERVISED KERNEL CI AND MCI TEST WITHOUT TRUE MIXTURE PROPORTIONS

C.1 Proofs for the tests without true mixture proportions

Proof of Lemma 3. By the Taylor expansion of $T_{\hat{\alpha}}$ around α^*, we derive

$$MT_{\hat{\alpha}} = MT_{\alpha^*} + M(\hat{\alpha} - \alpha^*)T'_{\alpha^*} + M\frac{1}{2}(\hat{\alpha} - \alpha^*)^2 T''_{\alpha^*} + o_p(1) \tag{5}$$

The remainder term is $o_p(1)$ because $\sqrt{M}(\hat{\alpha} - \alpha^*)$ converges in distribution to a normal random variable by Theorem 1 and 2, which ensures that higher-order terms in the expansion vanish in probability. □

Proof of Theorem 7. Under Assumption 3 and H_1, $T_{\hat{\alpha}}$ converges to the population test statistics of T_{α_1} as $M \to \infty$. Since $F_{12}^{\alpha_1}$ (resp. $F_{12S}^{\alpha_1}$) does not satisfy Conditional Independence for the CI test (resp. Multivariate CI for the MCI test), with assumptions in Theorem 3 (resp. Theorem 5), we can show that the population T_{α_1} is a positive constant. Then, $MT_{\hat{\alpha}} \xrightarrow{p} \infty$. □

C.2 Mean and variance estimation for the tests without true mixture proportions

In this subsection, we explain how to estimate the mean and variance for the tests without true mixture proportions. Our approach utilizes the result of Lemma 3. We begin by analyzing the asymptotic behavior of each term in Equation 5.

C.2.1 Asymptotic behaviors of each term in the Taylor expansion of $MT_{\hat{\alpha}}$

By Theorem 3 and 5, MT_{α^*} converges in distribution to a sum of squared normal random variables. By Theorem 1 and 2, $\sqrt{M}(\hat{\alpha} - \alpha^*)$ converges to a normal distribution. The term T''_{α^*} converges to a constant in probability. The remaining term $\sqrt{M}T'_{\alpha^*}$ converges to a normal distribution, which is derived as follows.

T'_{α^*} is a V-statistic. For the CI test, using the same notations as in the proof of Theorem 4, it can be expressed as

$$T'_{\alpha^*} = \frac{1}{n^6 n'^6} \sum_{i_1,\dots,i_6=1}^{n} \sum_{q_1,\dots,q_6=1}^{n'} h'_{i_1,\dots,i_6,q_1,\dots,q_6}$$

where

$$h'_{i_1,\dots,i_6,q_1,\dots,q_6} := \frac{1}{6!6!} \sum_{(j_1,\dots,j_6)}^{(i_1,\dots,i_6)} \sum_{(r_1,\dots,r_6)}^{(q_1,\dots,q_6)} 2\Big\langle \Big(\tilde{\varphi}_{12}(x^{(j_1)}) - \tilde{\varphi}_{12}(x'^{(r_1)}) - \Big(\alpha^*\tilde{\varphi}_1(x_1^{(j_2)}) + (1-\alpha^*)\tilde{\varphi}_1(x_2'^{(r_2)})\Big)$$
$$\otimes \Big(\tilde{\varphi}_2(x_1^{(j_3)}) - \tilde{\varphi}_2(x_2'^{(r_3)})\Big) - \Big(\tilde{\varphi}_1(x_1^{(j_2)}) - \tilde{\varphi}_1(x_2'^{(r_2)})\Big) \otimes \Big(\alpha^*\tilde{\varphi}_2(x_1^{(j_3)}) + (1-\alpha^*)\,\tilde{\varphi}_2(x_2'^{(r_3)})\Big)\Big), \varphi_{j_4,\dots,j_6,r_4,\dots,r_6}\Big\rangle.$$

For the MCI test, using the same notations as in the proof of Theorem 6,

$$T'_{\alpha^*} = \frac{1}{n^2 n'^2} \sum_{i_1,i_2=1}^{n} \sum_{q_1,q_2=1}^{n'} f'_{i_1,i_2,q_1,q_2},$$

where

$$f'_{i_1,i_2,q_1,q_2} := \frac{1}{4} \sum_{(j_1,j_2)}^{(i_1,i_2)} \sum_{(r_1,r_2)}^{(q_1,q_2)} 2\left\langle \alpha^*\tilde{\varphi}_{12S}(x^{(j_1)}) + (1-\alpha^*)\,\tilde{\varphi}_{12S}(x'^{(r_1)}), \tilde{\varphi}_{12S}(x^{(j_2)}) - \tilde{\varphi}_{12S}(x'^{(r_2)})\right\rangle.$$

In both the CI and MCI cases, T'_{α^*} is non-degenerate, since in general,

$$E_{i_2,\dots,i_6,q_1,\dots,q_6}[h'_{i_1,\dots,i_6,q_1,\dots,q_6}] \neq 0, E_{i_1,\dots,i_6,q_2,\dots,q_6}[h'_{i_1,\dots,i_6,q_1,\dots,q_6}] \neq 0$$

and

$$E_{i_2,q_1,q_2}[f'_{i_1,i_2,q_1,q_2}] \neq 0, E_{i_1,i_2,q_1}[f'_{i_1,i_2,q_1,q_2}] \neq 0.$$

Since non-degenerate V-statistics has $\sqrt{M}$-asymptotic normality (Huang et al., 2023; Serfling, 1981), $\sqrt{M}T'_{\alpha^*}$ converges to a normal distribution.

C.2.2 Mean and Variance estimation of $T_{\hat{\alpha}}$ for the CI and MCI tests

We consider the mean and variance estimation for the CI test without true mixture proportions in this subsection. A similar derivation process can be applied to the MCI test. As with T_α, denote $\check{T}_{CI}$ as $\check{T}_\alpha$, specifying $\check{T}_{CI}$ is a function of α. Let $\check{T}'_\alpha$ and $\check{T}''_\alpha$ be the first and second order derivatives of $\check{T}_\alpha$ at α. Note that $\check{T}_{\alpha^*}$, $\check{T}'_{\alpha^*}$ and $\check{T}''_{\alpha^*}$ are V-statistics and we denote their corresponging U-statistics by $\check{T}_{U,\alpha^*}$, $\check{T}'_{U,\alpha^*}$ and $\check{T}''_{U,\alpha^*}$. We assume T''_{U,α^*} converges to a constant c_0 in probability. To estimate the expectation and variance of the right hand side of Equation 5, we simplify the equation by considering the asymptotic behavior of each term.

First, we can show $M(T_{\alpha^*} - \check{T}_{\alpha^*}) \xrightarrow{p} 0$, $\sqrt{M}(T'_{\alpha^*} - \check{T}'_{\alpha^*}) \xrightarrow{p} 0$ and $T''_{\alpha^*} - \check{T}''_{\alpha^*} \xrightarrow{p} 0$, following a similar analysis to that used to derive Theorem 3. Given the asymptotic equivalence of U-statistics and V-statistics (Lemma S5. in the supplement of Huang et al. (2023)), we obtain the following convergence results as $M \to \infty$,

$$\begin{aligned} M(T_{\alpha^*} - \check{T}_{U,\alpha^*}) &\xrightarrow{p} c_1, \\ \sqrt{M}(T'_{\alpha^*} - \check{T}'_{U,\alpha^*}) &\xrightarrow{p} 0, \\ T''_{\alpha^*} - \check{T}''_{U,\alpha^*} &\xrightarrow{p} 0, \end{aligned}$$

where c_1 is a constant.

In addition, following the procedure in the proof of Theorem 1,

$$\sqrt{M}\left((\hat{\alpha} - \alpha^*) - S_{\alpha^*}\right) \xrightarrow{p} 0,$$

where $S_{\alpha^*} := -\frac{1}{d_0}(\alpha^* E_{\hat{U}_{12}}[\tilde{g}_{12}] + (1-\alpha^*)E_{\hat{U}'_{12}}[\tilde{g}_{12}]) = \frac{1}{nn'}\sum_{i=1}^{n}\sum_{q=1}^{n'} l_{i,q}$ and $l_{i,q} := -\frac{1}{d_0}(\alpha^*\tilde{g}_{12}(x^{(i)}) + (1-\alpha^*)\tilde{g}_{12}(x'^{(q)}))$.

Combining these results, the Taylor expansion can be approximated by

$$M\{T_{\alpha^*} + (\hat{\alpha} - \alpha^*)T'_{\alpha^*} + \frac{1}{2}(\hat{\alpha} - \alpha^*)^2 T''_{\alpha^*}\} \simeq M\{\check{T}_{U,\alpha^*} + S_{\alpha^*}\check{T}'_{U,\alpha^*} + \frac{c_0}{2}S^2_{\alpha^*}\} + c_1 \tag{6}$$

Therefore, to perform the hypothesis test, we estimate the mean and variance of the right hand side of (6). Under mild conditions such as uniform integrability, the asymptotic means and variances of both sides are actually equivalent. Using the same notation as in the proof of Theorem 4, the U-statistics $\check{T}_{U,\alpha^*}$ and $\check{T}'_{U,\alpha^*}$ are expressed as

$$\begin{aligned} \check{T}_{U,\alpha^*} &= \frac{1}{(n)_2(n')_2}\sum_{i_1\neq i_2}\sum_{q_1\neq q_2}\check{h}_{i_1,i_2,q_1,q_2} \\ \check{T}'_{U,\alpha^*} &= \frac{1}{(n)_2(n')_2}\sum_{i_1\neq i_2}\sum_{q_1\neq q_2}\check{h}'_{i_1,i_2,q_1,q_2} \end{aligned}$$

where we define $\check{h}'_{i_1,i_2,q_1,q_2}$ as

$$\check{h}'_{i_1,i_2,q_1,q_2} := \frac{1}{2!2!}\sum_{(j_1,j_2)}^{(i_1,i_2)}\sum_{(r_1,r_2)}^{(q_1,q_2)} 2\langle \alpha^*\tilde{\varphi}_{12}(x^{(j_1)}) + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(r_1)}), \tilde{\varphi}_{12}(x^{(j_2)}) - \tilde{\varphi}_{12}(x'^{(r_2)})\rangle.$$

Mean Estimation: We next consider estimating the mean of (6), $M\{E[\check{T}_{U,\alpha^*}] + E[S_{\alpha^*}\check{T}'_{U,\alpha^*}] + \frac{c_0}{2}E[S^2_{\alpha^*}]\} + c_1$.

We can derive the asymptotic mean of each term as follows, as $M \to \infty$,

$$
\begin{aligned}
ME[\check{T}_{U,\alpha^*}] + c_1 = c_1 &= \nu {\alpha^*}^2 E_{i_1,i_2}[\tilde{k}_{12}(x^{(i_1)}, x^{(i_1)}) - \tilde{k}_{12}(x^{(i_1)}, x^{(i_2)})] \\
&\quad + \nu'(1-\alpha^*)^2 E_{q_1,q_2}[\tilde{k}_{12}(x'^{(q_1)}, x'^{(q_1)}) - \tilde{k}_{12}(x'^{(q_1)}, x'^{(q_2)})], \\
ME[S_{\alpha^*}\check{T}'_{U,\alpha^*}] &\to -\frac{2}{d_0}\left(\nu\alpha^* E_{i_1}\left[\tilde{g}_{12}(x^{(i_1)}) E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]\right] + \nu'(1-\alpha^*) E_{q_1}\left[\tilde{g}_{12}(x'^{(q_1)}) E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]\right]\right), \\
ME[S^2_{\alpha^*}] &\to \frac{1}{d_0^2}\left(\nu {\alpha^*}^2 V_i[\tilde{g}_{12}(x^{(i)})] + \nu'(1-\alpha^*)^2 V_q[\tilde{g}_{12}(x'^{(q)})]\right).
\end{aligned}
$$

Here, the limit of $ME[\check{T}_{U,\alpha^*}]$ follows from Theorem 4. The limit of $ME[S_{\alpha^*}\check{T}'_{U,\alpha^*}]$ is derived by considering only dominant terms of the multiplication $S_{\alpha^*}\check{T}'_{U,\alpha^*}$ and the fact that $E_{i_1,i_2,q_1,q_2}[h'_{i_1,i_2,q_1,q_2}] = 0$. The limit of $ME[S^2_{\alpha^*}]$ is equivalent to the asymptotic variance of $\hat{\alpha}_{CI}$.

Variance Estimation: Next, we consider estimating the variance of (6), which is written as

$$
M^2\{V[\check{T}_{U,\alpha^*}] + V[S_{\alpha^*}\check{T}'_{U,\alpha^*}] + \frac{c_0^2}{4}V[S^2_{\alpha^*}] + 2(Cov[\check{T}_{U,\alpha^*}, S_{\alpha^*}\check{T}'_{U,\alpha^*}] + \frac{c_0}{2}Cov[\check{T}_{U,\alpha^*}, S^2_{\alpha^*}] + \frac{c_0}{2}Cov[S_{\alpha^*}\check{T}'_{U,\alpha^*}, S^2_{\alpha^*}])\}.
$$

Similarly to the asymptotic mean calculation, we derive the limit of each term by considering only dominant terms whose expectations are nonzero. As $M \to \infty$, we have the following convergence

$$
\begin{aligned}
M^2 V\left[\check{T}_{U,\alpha^*}\right] &\to 2\nu^2 E_{i_1,i_2}\left[E^2_{q_1,q_2}\left[\check{h}_{i_1,i_2,q_1,q_2}\right]\right] + 2\nu'^2 E_{q_1,q_2}\left[E^2_{i_1,i_2}\left[\check{h}_{i_1,i_2,q_1,q_2}\right]\right] \\
&\quad + 16\nu\nu' E_{i_1,q_1}\left[E^2_{i_2,q_2}\left[\check{h}_{i_1,i_2,q_1,q_2}\right]\right] \left(= \lim_{M\to\infty} V\left[MT_{CI}\right]\right), \\
M^2 V\left[S_{\alpha^*}\check{T}'_{U,\alpha^*}\right] &= M^2 E\left[\left(S_{\alpha^*}\check{T}'_{U,\alpha^*}\right)^2\right] - M^2\left(E\left[S_{\alpha^*}\check{T}'_{U,\alpha^*}\right]\right)^2 \\
&\to 4\nu^2 E_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] E_{i_1}\left[E^2_{i_2,q_1,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] + 4\nu\nu' E_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] E_{q_1}\left[E^2_{i_1,i_2,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] \\
&\quad + 4\nu\nu' E_{q_1}\left[E^2_{i_1}\left[l_{i_1,q_1}\right]\right] E_{i_1}\left[E^2_{i_2,q_1,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] + 4\nu'^2 E_{q_1}\left[E^2_{i}\left[l_{i_1,q_1}\right]\right] E_{q_1}\left[E^2_{i_1,i_2,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] \\
&\quad + 8\nu^2 E^2_{i_1}\left[E_{q_1}\left[l_{i_1,q_1}\right] E_{i_2,q_1,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] \\
&\quad + 16\nu\nu' E_{i_1}\left[E_{q_1}\left[l_{i_1,q_1}\right] E_{i_2,q_1,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] E_{q_1}\left[E_{i_1}\left[l_{i_1,q_1}\right] E_{i_1,i_2,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] \\
&\quad + 8\nu'^2 E^2_{q_1}\left[E_{i}\left[l_{i_1,q_1}\right] E_{i_1,i_2,q_2}\left[\check{h}'_{i_1,i_2,q_1,q_2}\right]\right] - \lim_{M\to\infty} M^2 E^2\left[S_{\alpha^*}\check{T}'_{U,\alpha^*}\right] \\
&= 4\left(\nu E_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] + \nu' E_{q_1}\left[E^2_{i_1}[l_{i_1,q_1}]\right]\right) \\
&\quad \left(\nu E_{i_1}[E^2_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]] + \nu' E_{q_1}[E^2_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]]\right) \\
&\quad + 8\left(\nu E_{i_1}\left[E_{q_1}\left[l_{i_1,q_1}\right] E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]\right] + \nu' E_{q_1}\left[E_{i}\left[l_{i_1,q_1}\right] E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]\right]\right)^2 \\
&\quad - \lim_{M\to\infty} M^2 E^2\left[S_{\alpha^*}\check{T}'_{U,\alpha^*}\right], \\
M^2 V\left[S^2_{\alpha^*}\right] &= M^2 E\left[S^4_{\alpha^*}\right] - M^2 E^2\left[S^2_{\alpha^*}\right] \\
&\to 3\nu^2 E^2_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] + 6\nu\nu' E_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] E_{q_1}\left[E^2_{i_1}\left[l_{i_1,q_1}\right]\right] + 3\nu'^2 E^2_{q_1}\left[E^2_{i_1}\left[l_{i_1,q_1}\right]\right] \\
&\quad - \lim_{M\to\infty} M^2 E^2\left[S^2_{\alpha^*}\right] \\
&= 3\left(\nu E_{i_1}\left[E^2_{q_1}\left[l_{i_1,q_1}\right]\right] + \nu' E_{q_1}\left[E^2_{i_1}\left[l_{i_1,q_1}\right]\right]\right)^2 - \lim_{M\to\infty} M^2 E^2\left[S^2_{\alpha^*}\right],
\end{aligned}
$$

$$
\begin{aligned}
M^2 Cov[\check{T}_{U,\alpha^*}, S_{\alpha^*}\check{T}'_{U,\alpha^*}] &= M^2 E[\check{T}_{U,\alpha^*} S_{\alpha^*}\check{T}'_{U,\alpha^*}] - M^2 E[\check{T}_{U,\alpha^*}] E[S_{\alpha^*}\check{T}'_{U,\alpha^*}] \\
&\to 4\nu^2 E_{i_1,i_2}\left[E_{q_1,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] E_{q_1}[l_{i_2,q_1}]\right] \\
&\quad + 8\nu\nu' E_{i_1,q_1}\left[E_{i_2,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] E_{i_1}[l_{i_1,q_1}]\right] \\
&\quad + 8\nu\nu' E_{i_1,q_1}\left[E_{i_2,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] E_{q_1}[l_{i_1,q_1}]\right] \\
&\quad + 4\nu'^2 E_{q_1,q_2}\left[E_{i_1,i_2}[\check{h}_{i_1,i_2,q_1,q_2}] E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] E_{i_1}[l_{i_1,q_2}]\right], \\
M^2 Cov[\check{T}_{U,\alpha^*}, S^2_{\alpha^*}] &= M^2 E[\check{T}_{U,\alpha^*} S^2_{\alpha^*}] - M^2 E[\check{T}_{U,\alpha^*}] E[S^2_{\alpha^*}]
\end{aligned}
$$

$$
\begin{aligned}
&\to 2\nu^2 E_{i_1,i_2}\left[E_{q_1,q_2}[\check{h}_{i_1,i_2,q_1,q_2}]E_{q_1}[l_{i_1,q_1}]E_{q_1}[l_{i_2,q_1}]\right] \\
&+ 8\nu\nu' E_{i_1,q_1}\left[E_{i_2,q_2}[\check{h}_{i_1,i_2,q_1,q_2}]E_{q_1}[l_{i_1,q_1}]E_{i_1}[l_{i_1,q_1}]\right] \\
&+ 2\nu'^2 E_{q_1,q_2}\left[E_{i_1,i_2}[\check{h}_{i_1,i_2,q_1,q_2}]E_{i_1}[l_{i_1,q_1}]E_{i_1}[l_{i_1,q_2}]\right], \\
M^2 Cov[S_{\alpha^*}\check{T}'_{U,\alpha^*}, S^2_{\alpha^*}] &= M^2 E[S^3_{\alpha^*}\check{T}'_{U,\alpha^*}] - M^2 E[S_{\alpha^*}\check{T}'_{U,\alpha^*}]E[S^2_{\alpha^*}] \\
&\to 6\nu^2 E_{i_1}\left[E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{q_1}[l_{i_1,q_1}]\right] E_{i_1}\left[E^2_{q_1}[l_{i_1,q_1}]\right] \\
&+ 6\nu\nu' E_{i_1}\left[E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{q_1}[l_{i_1,q_1}]\right] E_{q_1}\left[E^2_{i_1}[l_{i_1,q_1}]\right] \\
&+ 6\nu\nu' E_{q_1}\left[E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{i_1}[l_{i_1,q_1}]\right] E_{i_1}\left[E^2_{q_1}[l_{i_1,q_1}]\right] \\
&+ 6\nu'^2 E_{q_1}\left[E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{i_1}[l_{i_1,q_1}]\right] E_{q_1}\left[E^2_{i_1}[l_{i_1,q_1}]\right] - \lim_{M\to\infty} M^2 E[S_{\alpha^*}\check{T}'_{U,\alpha^*}]E[S^2_{\alpha^*}] \\
&= 6\left(\nu E_{i_1}\left[E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{q_1}[l_{i_1,q_1}]\right] + \nu' E_{q_1}\left[E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}]E_{i_1}[l_{i_1,q_1}]\right]\right) \\
&\left(\nu E_{i_1}\left[E^2_{q_1}[l_{i_1,q_1}]\right] + \nu' E_{q_1}\left[E^2_{i_1}[l_{i_1,q_1}]\right]\right) - \lim_{M\to\infty} M^2 E[S_{\alpha^*}\check{T}'_{U,\alpha^*}]E[S^2_{\alpha^*}].
\end{aligned}
$$

Furthermore, the expactations of $\check{h}_{i_1,i_2,q_1,q_2}$ and $\check{h}'_{i_1,i_2,q_1,q_2}$ are written as

$$
\begin{aligned}
E_{i_1,i_2}[\check{h}_{i_1,i_2,q_1,q_2}] &= \left\langle \alpha^* E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})] + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(q_1)}), \alpha^* E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})] + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(q_2)}) \right\rangle, \\
E_{q_1,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] &= \left\langle \alpha^* \tilde{\varphi}_{12}(x^{(i_1)}) + (1-\alpha^*)E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})], \alpha^* \tilde{\varphi}_{12}(x^{(i_2)}) + (1-\alpha^*)E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle, \\
E_{i_2,q_2}[\check{h}_{i_1,i_2,q_1,q_2}] &= \frac{1}{2}\left\langle \alpha^* \tilde{\varphi}_{12}(x^{(i_1)}) + (1-\alpha^*)E_{q_2}[\tilde{\varphi}_{12}(x'^{(q_2)})], \alpha^* E_{i_2}[\tilde{\varphi}_{12}(x^{(i_2)})] + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(q_1)}) \right\rangle, \\
E_{i_1,i_2,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] &= \left\langle \alpha^* E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})] + (1-\alpha^*)\tilde{\varphi}_{12}(x'^{(q_1)}), E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})] - E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle \\
&= -\left\langle E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})], E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle + \left\langle E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})], \tilde{\varphi}_{12}(x'^{(q_1)}) \right\rangle, \\
E_{i_2,q_1,q_2}[\check{h}'_{i_1,i_2,q_1,q_2}] &= \left\langle \alpha^* \tilde{\varphi}_{12}(x^{(i_1)}) + (1-\alpha^*)E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})], E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})] - E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle \\
&= -\left\langle \tilde{\varphi}_{12}(x^{(i_1)}), E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle + \left\langle E_{i_1}[\tilde{\varphi}_{12}(x^{(i_1)})], E_{q_1}[\tilde{\varphi}_{12}(x'^{(q_1)})] \right\rangle.
\end{aligned}
$$

With these results, we have derived expressions for the asymptotic mean and variance of $MT_{\hat{\alpha}}$. In practice, each term in these expressions is estimated by replacing the population distributions U_{12}, U'_{12} and α^* with their empirical counterparts $\hat{U}_{12}$, $\hat{U}'_{12}$ and $\hat{\alpha}$.

For the MCI test, the asymptotic mean and variance of $T_{\hat{\alpha}}$ can be estimated similarly. This is done by replacing the terms from the CI test, namely $\tilde{g}_{12}$, $\tilde{k}_{12}$, $\check{h}_{i_1,i_2,q_1,q_2}$, $\check{h}'_{i_1,i_2,q_1,q_2}$, l_{i_1,q_1}, with $\tilde{g}_{12S}$, $\tilde{k}_{12S}$, f_{i_1,i_2,q_1,q_2}, f'_{i_1,i_2,q_1,q_2} and $l^{MCI}_{i_1,q_1}$, respectively. Here, we define

$$
l^{MCI}_{i_1,q_1} := -\frac{1}{d_0}(\alpha^* \tilde{g}_{12S}(x^{(i_1)}) + (1-\alpha^*)\tilde{g}_{12S}(x'^{(q_1)})).
$$

D EXPERIMENTS

D.1 Practical computation of test statistic

In this section, we explain how to calculate the test statistics in practice. Let $K_\tau \in \mathbb{R}^{M\times M}$ be the Gram matrix of x_τ with a kernel k_τ, where the entries are given by $(K_\tau)_{ij} = k_\tau(\mathbf{v}_{x_\tau,i}, \mathbf{v}_{x_\tau,j})$. Define $D_\alpha \in \mathbb{R}^{M\times M}$ as a diagonal matrix where $(D_\alpha)_{ii} = \alpha/n$ for $i \le n$ and $(D_\alpha)_{ii} = (1-\alpha)/n'$ for $i > n$. Let $\mathbf{1}$ be an $M \times 1$ vector of ones and define the centering matrix $H := (I - \mathbf{1}\mathbf{1}^T D_{\alpha^*})$. Then, T_{CI} can be computed in matrix form as

$$
T_{CI} = \mathrm{tr}(HK_1H^T D_{\alpha^*} HK_2H^T D_{\alpha^*}).
$$

Using centralized kernels $\tilde{k}_{1S}(x_{1S}, x'_{1S}) = \langle \varphi_1(x_1) - \mu_{X_1|X_S}(x_S), \varphi_1(x'_1) - \mu_{X_1|X_S}(x'_S)\rangle$ and $\tilde{k}_{2S}(x_{2S}, x'_{2S}) = \langle \varphi_2(x_2) - \mu_{X_2|X_S}(x_S), \varphi_2(x'_2) - \mu_{X_2|X_S}(x'_S)\rangle$,we can compute T_{MCI} as

$$T_{MCI} = \mathrm{tr}((\tilde{K}_{1S} \odot K_S) D_{\alpha^*} \tilde{K}_{2S} D_{\alpha^*})$$

where $\odot$ denotes the Hadamard product. $\tilde{K}_{1S} \in \mathbb{R}^{M \times M}$ and $\tilde{K}_{2S} \in \mathbb{R}^{M \times M}$ are the Gram matrices associated with $\tilde{k}_{1S}$ and $\tilde{k}_{2S}$, defined by $(\tilde{K}_{1S})_{ij} = \tilde{k}_{1S}(\mathbf{v}_{x_{1S},i}, \mathbf{v}_{x_{1S},j})$ and $(\tilde{K}_{2S})_{ij} = \tilde{k}_{2S}(\mathbf{v}_{x_{2S},i}, \mathbf{v}_{x_{2S},j})$.

For T_{MCI}, we need to estimate the centralized kernels $\tilde{k}_{1S}$ and $\tilde{k}_{2S}$. To do this, we consider the eigenvalue decomposition of the kernel matrix, $K_\tau = V_\tau^T \Lambda_\tau V_\tau$, which provides an empirical kernel map $\hat{\varphi}_\tau = [\hat{\varphi}_{\tau,1}(\mathbf{v}_{x_\tau}), ..., \hat{\varphi}_{\tau,n+n'}(\mathbf{v}_{x_\tau})] = V_\tau \Lambda_\tau^{1/2}$. In the original KCI test (Zhang et al., 2011), each feature map $\hat{\varphi}_{\tau,i}(\mathbf{v}_{x_\tau})$ is centralized as $\hat{\varphi}_{\tau,i}(\mathbf{v}_{x_\tau}) - E[\hat{\varphi}_{\tau,i}(\mathbf{v}_{x_\tau})|Z]$ by estimating the conditional expectation $E[\hat{\varphi}_{\tau,i}(\mathbf{v}_{x_\tau})|Z]$ using kernel ridge regression.

In our setting, KRR is performed in a weakly-supervised manner, similarly to the MCI MPE in Section 4. It is optimized by the first-order condition of non-convex loss. To reduce the computational costs, we implement KRR only for the eigenvectors corresponding to the top k eigenvalues and omit other eigenvectors when reconstructing the centralized Gram matrix $\tilde{K}_\tau$. In our experiments, we set $k = 5$, which we found to be sufficient to approximate the original Gram matrix accurately.

D.2 Experimental details

D.2.1 CI MPE with synthetic data

For the UCI datasets, the positive and negative classes were assigned as shown in Table 6. We set the search range for the mixture proportion to be $I_{\alpha^+} = \mathbb{R}_+$.

Table 6: Positive and Negative classes used for UCI dataset

	Positive	Negative
Shuttle	1	other classes
Wine	white	red
Dry Bean	DERMASON	other classes

D.2.2 CI MPE with real-world data

We used two datasets from the UCI repository: the Breast Cancer Wisconsin and Dry Bean datasets. For each dataset, we chose positive and negative classes and implemented the experiments, switching the classes. The procedure was as follows:

1. We first selected a candidate set of features X_i that were discriminative, satisfying $|E[X_i \mid Y=1] - E[X_i \mid Y=-1]| / \sqrt{V[X_i \mid Y=1]} > 0.5$, since a significant mean difference is essential for the efficient MPE.

2. We then applied the HSIC test to all pairs of features from this candidate set to identify those satisfying the CI condition, with a significance level 0.05.

3. For each detected CI feature pair, we ran our CI MPE method 10 times.

For the MPE task, we set $n = n' = 2000$ and used a Positive-Unlabeled (PU) setting with class priors $(\theta, \theta') = (1, 0.5)$.

D.2.3 MCI MPE with synthetic data

We used a regularization parameter $\lambda = 5 \times 10^{-4}$ and a Gaussian kernel with bandwidth $\sigma = 3.5$ for all MCI MPE experiments. The search ranges were set to $I_{\alpha_+} = [1.1, 1.5]$ and $I_{\alpha_-} = [-0.7, 0]$.

D.2.4 MCI MPE with real-world data

We used the Dry Bean dataset and set the positive and negative classes to SIRA and DERMASON, respectively. The procedure was as follows:

1. We searched for feature triplets (X_1, X_2, X_S) satisfying the MCI condition in the negative class by applying the KCI test (Zhang et al., 2011) to all possible triplets with a significance level 0.05. In the search, we constructed a candidate set of features that satisfies $|E[X_i \mid Y=1] - E[X_i \mid Y=-1]| / \sqrt{V[X_i \mid Y=1]} > 1$, similarly to D.2.2. Then we only used features in the set as the candidates for X_1 and X_2.

2. For each detected triplet, we ran our MCI MPE method 5 times and evaluated the estimation error for θ'.

For the MPE task, we set $n = n' = 1000$ and used a Positive-Unlabeled (PU) setting with class priors $(\theta, \theta') = (1, 0.5)$. We used a Gaussian kernel with bandwidth $\sigma = 1.0$ for KRR, set the regularization parameter to $\lambda = 10^{-3}$ and the search range $I_{\alpha_-} = [-1.25, -0.5]$. In this experiment, 48 triplets were detected and the resulting MAE of $\hat{\theta}'$ over all runs was 0.0312 ± 0.0304.

D.2.5 Hyperparameters for CI and MCI test

We used a Gaussian kernel for all CI and MCI test experiments, both with and without mixture proportions. For CI test of both cases, kernel bandwidth is set as $\sigma = 2.5$. For MCI test with mixture proportions, we set $\lambda = 5 \times 10^{-4}$ and $\sigma = 3.5$ for all kernels. For MCI test without mixture proportions, we set $\lambda = 5 \times 10^{-6}$ and $\sigma = 2.5$ for the test statistic, and $\lambda = 1 \times 10^{-2}$ and $\sigma = 3$ for the MCI MPE to estimate α^*. The search ranges for MPE are set as the same values in Section D.2.1 and D.2.3.

D.3 Additional experiments

D.3.1 Bias calculation of CI and MCI MPE

We conducted an additional experiment to investigate the relationship between MPE error and the degree of CI violation (i.e., correlation). We used the same Gaussian data generation process as in Section 6.2, where the CI or MCI assumption is only satisfied when $\sigma_{12} = 0$. Each experiment was repeated 100 times with sample sizes $n = n' = 2000$ and true class priors $(\theta, \theta') = (0.8, 0.2)$.

The results are presented in Tables 7 and 8. As shown, the MPE error remains small even when the CI or MCI assumption is weakly violated.

Table 7: Mean absolute error of $(\hat{\theta}, \hat{\theta}')$ with weakly-violated CI data

σ_{12}^2	MAE of $(\hat{\theta}, \hat{\theta}')$
0	$(0.026 \pm 0.018, 0.025 \pm 0.019)$
0.1	$(0.074 \pm 0.034, 0.030 \pm 0.021)$
0.2	$(0.132 \pm 0.025, 0.034 \pm 0.021)$

Table 8: Mean absolute error of $(\hat{\theta}, \hat{\theta}')$ with weakly-violated MCI data

σ_{12}^2	MAE of $(\hat{\theta}, \hat{\theta}')$
0	$(0.008 \pm 0.005, 0.015 \pm 0.010)$
0.1	$(0.017 \pm 0.016, 0.014 \pm 0.012)$
0.2	$(0.037 \pm 0.011, 0.010 \pm 0.011)$

D.3.2 Investigation of low power in the MCI test without true mixture proportions

To investigate the cause of the low test power, we compared our method to a test using the true null distribution (simulated with 1000 trials). We used the same setup as in Section 6.2 ($H_0 : \sigma_{12} = 0$, $H_1 : \sigma_{12} = 0.2$, $n = n' = 1000$). Table 9 summarizes the results.

Table 9: MCI test power: our null approximation vs. true null distribution

Method	Test power
Test with our null approximation	0.199
Test with true null distribution	0.293

As shown in Table 9, even with the true null distribution, the ideal test power remains low (0.293). This indicates that the poor separation between the true null and alternative distributions causes the low power in small samples.