Conditional Kernel Quantile Embeddings: A Nonparametric Framework for Conditional Two-Sample Testing

Anonymous authors
Paper under double-blind review

Abstract

Comparing conditional probability distributions, $P(Y|X)$ and $Q(Y|X)$, is a fundamental problem in machine learning, crucial for tasks like causal inference, detecting dataset shift, and model validation. The predominant approach, based on Conditional Kernel Mean Embeddings (KCMEs), suffers from significant drawbacks: it relies on strong and often unverifiable assumptions on the kernel to be a metric, incurs high computational costs, and may exhibit reduced sensitivity to higher-order distributional differences. We introduce Conditional Kernel Quantile Embeddings (CKQEs), a novel and robust framework for representing conditional distributions in a Reproducing Kernel Hilbert Space (RKHS). Throughout, we assume $P_X = Q_X$ for conditional comparisons, and we require only that the output-space kernel be quantile-characteristic. From CKQEs, we construct the Conditional Kernel Quantile Discrepancy (CKQD), a new family of probability metrics. We prove that CKQD: (1) is a metric under substantially weaker and more practical kernel conditions than KCME-based distances, namely requiring only a quantile-characteristic kernel; (2) possesses a clear geometric interpretation, recovering a conditional version of the Sliced Wasserstein distance in a special case; and (3) admits a computationally efficient, statistically consistent nonparametric estimator with proven finite-sample convergence rates. By addressing the core weaknesses of the KCME framework, CKQE provides a more versatile and theoretically sound foundation for conditional two-sample testing.

1 Introduction

1.1 The Challenge of Comparing Conditional Distributions

The ability to compare conditional probability distributions is a cornerstone of modern statistical machine learning. Formally, this involves testing the null hypothesis $H_0 : P(Y|X) = Q(Y|X)$ against the general alternative $H_1 : P(Y|X) \neq Q(Y|X)$. We work under the standard covariate shift setting $P_X = Q_X$. This test is fundamental to a vast array of problems, including off-policy evaluation in reinforcement learning, detecting covariate and concept drift in deployed models, performing constraint-based causal discovery, and auditing algorithms for fairness (Song et al., 2009; Zhang et al., 2011; Massiani et al., 2025). A robust and efficient test for this hypothesis allows us to answer critical questions: Has the data-generating process changed? Does a treatment affect the entire outcome distribution, not just its mean? Are two causal mechanisms identical?

1.2 Limitations of the KCME Framework

The prevailing non-parametric methodology for this task is based on Conditional Kernel Mean Embeddings (KCMEs)(Song et al., 2009; Grünewälder et al., 2012; Park & Muandet, 2020). In a Reproducing Kernel Hilbert Space (RKHS), KCMEs show a conditional distribution $P(Y|x)$ as a single element, which is the conditional mean of a feature map. The distance between these embeddings is then used as a test statistic. This framework is strong, but it has some major flaws that make it less useful in practice.

KCME-based metrics require the Y-kernel to be *mean-characteristic* (Fukumizu et al., 2007; Sriperumbudur et al., 2010) so the mean embedding is injective at the population level—a strong assumption that can be nontrivial to verify in applications. Classical operator-based estimators involve solving linear systems with an $n \times n$ Gram matrix on X, which in naive dense implementations scales as $O(n^3)$.[1] More recent work reinterprets CME via a measure-theoretic, vector-valued regression view (Park & Muandet, 2020), relaxing operator assumptions and admitting standard fast solvers. Although a characteristic kernel implies that the mean element determines the conditional law in principle, in finite samples KCME-based tests can be *less sensitive* to differences that primarily affect higher-order aspects, especially at moderate effect sizes and with practical kernel/hyperparameter choices. This empirical observation motivates our quantile-based alternative.

1.3 An Alternative Path: From Means to Quantiles

Recently, Naslidnyk et al. (2025) introduced Kernel Quantile Embeddings (KQEs) as a compelling alternative for comparing *unconditional* distributions. By embedding distributions via their directional quantiles in an RKHS, the resulting Kernel Quantile Discrepancy (KQD) is a metric under much weaker, more practical conditions (requiring a *quantile-characteristic* kernel) and connects naturally to the geometry of the Sliced Wasserstein distance (Kolouri et al., 2019).We view CKQD as complementary to recent conditional tests based on nearest-neighbor kernels, de-biased U-statistics, KRR-confidence-bound testing, and conformal prediction, which emphasize different guarantees or computational regimes(Chatterjee et al., 2024; Chen & Lei, 2025; Hu & Lei, 2023).

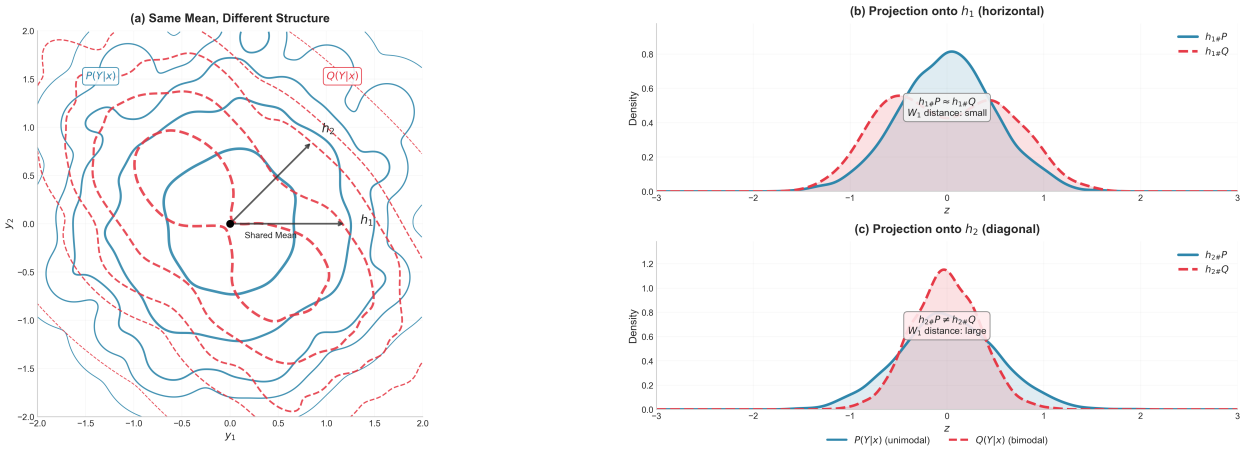

Figure 1: CKQD captures structural differences missed by mean-based methods by aggregating the discrepancy between 1D quantile functions (equivalent to W_1 distance for $p = 1$) over all projection directions h and all conditions x. When two conditional distributions have identical means but different shapes (bimodal vs unimodal), mean-based methods are less sensitive while CKQD succeeds by examining multiple projection directions.

1.4 Our Contributions

In this paper, we bridge this gap by extending the powerful KQE framework to the conditional setting. We introduce Conditional Kernel Quantile Embeddings (CKQEs) and the associated Conditional Kernel Quantile Discrepancy (CKQD). Our contributions are:

1. A rigorous definition of CKQEs and the CKQD metric for comparing conditional probability distributions.

[1] E.g., dense Cholesky for kernel ridge regression. In practice, iterative and randomized solvers (e.g., conjugate gradients, Nyström, pivoted Cholesky) reduce wall-clock cost.

2. A proof that CKQD is a metric under significantly weaker assumptions than existing KCME-based metrics, namely requiring only a quantile-characteristic kernel on the output space (Theorem 1).

3. A theoretical result establishing a direct link between CKQD and the conditional Sliced Wasserstein distance, clarifying its geometric properties (Theorem 2).

4. A novel, consistent, and computationally efficient non-parametric estimator for CKQD with proven finite-sample convergence rates (Theorem 3).

Table 1 provides a structured comparison of CKQD with leading alternatives, highlighting the unique combination of properties it offers.

Table 1: Comparison of Conditional Two-Sample Testing Methods

Method	**Provable Metric?**	**Kernel Condition**	**Estimator Cost**	**Key Properties**
CKQD	Yes	Quantile-Char.	$O(n^2) + O(Lmn \log n)^a$	Sensitive to broad distributional structure beyond means; Sliced Wasserstein link
KCME-RBF	Yes	Mean-Char.	$O(n^3)^b$	Mean in RKHS only
C2ST	No	None	Classifier-dep.	learned test

[a]In our implementation $m = 2n$ denotes the number of evaluation points (the pooled x grid); we compute exact NW weights.
[b]Classical implementation with $O(n^3)$ complexity; runtime exceeds 7 seconds at $n = 1600$. Modern regression-based approaches (Park & Muandet, 2020) can leverage fast solvers.

2 Preliminaries: From Mean to Quantile Embeddings in RKHS

We first review the necessary concepts from kernel methods, establishing notation and focusing on the progression from mean-based to quantile-based embeddings of probability distributions.

2.1 Global Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces. We denote probability measures on these spaces by $P, Q \in \mathcal{P}(\mathcal{X})$ or $P, Q \in \mathcal{P}(\mathcal{Y})$. A kernel on $\mathcal{X}$ is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, with its associated Reproducing Kernel Hilbert Space (RKHS) denoted by $\mathcal{H}_k$. The feature map is $\phi(x) := k(x, \cdot) \in \mathcal{H}_k$. The inner product in $\mathcal{H}_k$ is $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. The unit sphere in an RKHS H is $\mathcal{S}_{\mathcal{H}} := \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} = 1\}$. For a 1D cumulative distribution function (CDF) F, its quantile function is $Q_F(\alpha) = \inf\{z \in \mathbb{R} : F(z) \geq \alpha\}$ for $\alpha \in (0, 1)$.

2.2 Reproducing Kernel Hilbert Spaces (RKHS)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetric and for any $n \in \mathbb{N}$, any $\{x_i\}_{i=1}^n \subset \mathcal{X}$, and any $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$, we have $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$. By the Moore-Aronszajn theorem, every such kernel k is associated with a unique Hilbert space of functions $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$, called an RKHS (Aronszajn, 1950). The key property of an RKHS is the *reproducing property*: for any $x \in \mathcal{X}$ and any $f \in \mathcal{H}_k$, we have $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$. This implies that point evaluation is a continuous linear functional. The function $\phi(x) := k(x, \cdot)$ is the feature map, mapping points in $\mathcal{X}$ to functions in $\mathcal{H}_k$.

2.3 Kernel Mean Embeddings and MMD

The kernel mean embedding (KME) represents a probability distribution as a single point in an RKHS (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007).

Definition 1 (Kernel Mean Embedding) *Let k be a bounded continuous kernel on a topological space $\mathcal{X}$, and let $P \in \mathcal{P}(\mathcal{X})$ be a Borel probability measure. The kernel mean embedding of P into $\mathcal{H}_k$ is the element $\mu_P \in \mathcal{H}_k$ defined by*

$$\mu_P := \int_{\mathcal{X}} k(x, \cdot) dP(x) = \mathbb{E}_{X \sim P}[k(X, \cdot)].$$

Definition 2 *The Maximum Mean Discrepancy (MMD) between two probability measures P and Q is the distance between their mean embeddings in $\mathcal{H}_k$:*

$$MMD_k(P, Q) := \|\mu_P - \mu_Q\|_{\mathcal{H}_k}.$$

A crucial property is that MMD_k defines a metric on the space of probability measures if and only if the kernel k is *characteristic* (Sriperumbudur et al., 2010). Characteristic kernels ensure that the map $P \mapsto \mu_P$ is injective.

2.4 Kernel Quantile Embeddings and KQD

Naslidnyk et al. (2025) proposed an alternative to KMEs based on quantiles. Instead of a single mean element, a distribution is represented by a family of directional quantiles.

Definition 3 *Let $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ be an RKHS and let $\mathcal{S}_{\mathcal{H}_k}$ be its unit sphere. For a probability measure P on $\mathcal{X}$, a direction $h \in \mathcal{S}_{\mathcal{H}_k}$, and $\alpha \in (0, 1)$, the α-quantile of P along h is*

$$Q_{P,h}(\alpha) := Q_{P_h}(\alpha),$$

where P_h is the pushforward measure of P under the map $x \mapsto \langle k(x, \cdot), h \rangle_{\mathcal{H}_k}$.

Definition 4 *Let $p \geq 1$ and let ν be a Borel probability measure on the sphere $\mathcal{S}_{\mathcal{H}_k}$ with full support. The expected Kernel Quantile Discrepancy (e-KQD) between P and Q is*

$$\text{e-}KQD_p^p(P, Q) := \int_{\mathcal{S}_{\mathcal{H}_k}} \int_0^1 |Q_{P,h}(\alpha) - Q_{Q,h}(\alpha)|^p \, d\alpha d\nu(h).$$

Definition 5 (Quantile-Characteristic Kernels) *A continuous kernel k on a topological space $\mathcal{X}$ is called quantile-characteristic if for any two probability measures P, Q on $\mathcal{X}$, we have $Q_{P,h}(\alpha) = Q_{Q,h}(\alpha)$ for all $h \in \mathcal{S}_{\mathcal{H}_k}$ and all $\alpha \in (0, 1)$ if and only if $P = Q$ (Naslidnyk et al., 2025).*

The key result of Naslidnyk et al. (2025) is that e-KQD is a metric if the kernel k is quantile-characteristic, a condition that is strictly weaker than being mean-characteristic. For instance, continuous, separating kernels on compact spaces are quantile-characteristic.

3 Conditional Kernel Quantile Embeddings: Theory and Properties

We now extend the KQE framework to the conditional setting, providing a new tool for comparing conditional distributions $P(Y|X)$ and $Q(Y|X)$.

3.1 Intuition: From a Single Embedding to a Field of Embeddings

The core idea is to move from representing a distribution as a single point (the mean embedding) to a richer object. For each conditioning value $x \in \mathcal{X}$, we characterize the conditional distribution $P(Y|x)$ not by its mean, but by its complete quantile structure in the feature space. The Conditional Kernel Quantile Embedding (CKQE) can thus be viewed as a function, or a "field," that maps each condition x to a full geometric description—the set of all directional quantiles—of the corresponding output distribution $P(Y|x)$.

3.2 Formal Definitions

Let $\mathcal{X}$ and $\mathcal{Y}$ be the spaces for the conditioning and target variables, respectively. Let k_Y be a kernel on $\mathcal{Y}$ with RKHS $\mathcal{H}_Y$.

Assumption 1 *The target space $\mathcal{Y}$ is a Hausdorff, separable, and σ-compact metrizable space. The kernel k_Y is bounded and continuous.*

Definition 6 *For a fixed condition $x \in \mathcal{X}$, a direction $h \in \mathcal{S}_{\mathcal{H}_Y}$, and a quantile level $\alpha \in (0,1)$, the conditional directional quantile of $P(Y|X)$ is*

$$Q_{P|x,h}(\alpha) := Q_{P_h(\cdot|x)}(\alpha),$$

where $P_h(\cdot|x)$ is the pushforward of the conditional distribution $P(Y|x)$ under the projection map $y \mapsto \langle k_Y(y,\cdot), h\rangle_{\mathcal{H}_Y}$.

Definition 7 (Conditional Kernel Quantile Embedding) *The Conditional Kernel Quantile Embedding (CKQE) of the conditional distribution $P(Y|X)$ is the function $\mathcal{Q}_P : \mathcal{X} \to L^p([0,1] \times \mathcal{S}_{\mathcal{H}_Y})$ that maps each condition $x \in \mathcal{X}$ to its full set of directional quantiles:*

$$\mathcal{Q}_P(x) := ((\alpha, h) \mapsto Q_{P|x,h}(\alpha)).$$

The CKQE is understood as an equivalence class a.e. in (α, h) because quantiles are defined up to λ-null sets and ν-null directions.

Finally, we define the distance between two conditional distributions, $P(Y|X)$ and $Q(Y|X)$, by measuring the expected discrepancy between their CKQEs. We make the standard assumption for conditional two-sample testing that the conditioning variables are drawn from the same distribution, i.e., $P_X = Q_X$.

Definition 8 *Let $p \geq 1$, and let ν be a Borel probability measure on $\mathcal{S}_{\mathcal{H}_Y}$ with full support. The Conditional Kernel Quantile Discrepancy (CKQD) between $P(Y|X)$ and $Q(Y|X)$ is defined as*

$$\begin{aligned} CKQD_p^p(P,Q) &:= \int_{\mathcal{X}} \text{e-}KQD_p^p(P(\cdot|x), Q(\cdot|x)) dP_X(x) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{S}_{\mathcal{H}_Y}} \int_0^1 \left| Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha) \right|^p d\alpha d\nu(h) \right) dP_X(x). \end{aligned}$$

3.3 Main Theoretical Guarantee: CKQD is a Metric

We now establish the main theoretical property of CKQD, highlighting its advantages over KCME-based approaches.

All equalities between conditional laws are to be understood for P_X-almost every x.

Theorem 1 *Let the conditions of Assumption 1 hold. Let k_Y be a quantile-characteristic kernel on $\mathcal{Y}$, and let ν be a measure on $\mathcal{S}_{\mathcal{H}_Y}$ with full support. Then for any $p \geq 1$, $CKQD_p(P,Q)$ is a metric on the space of conditional probability distributions on $\mathcal{Y}$ given $\mathcal{X}$.*

Proof. A high-level sketch is as follows (full proof in Appendix B). The proof relies on the fact that since k_Y is quantile-characteristic and ν has full support, the inner term, e-KQD$_p(P(\cdot|x), Q(\cdot|x))$, is a metric on the space of unconditional probability measures on $\mathcal{Y}$ (Naslidnyk et al., 2025). If CKQD$_p(P,Q) = 0$, the non-negativity of the integrand implies that the inner term must be zero for P_X-almost every x. Since e-KQD$_p$ is a metric, this means $P(Y|x) = Q(Y|x)$ for almost every x, which implies $P(Y|X) = Q(Y|X)$. The other metric properties (non-negativity, symmetry, triangle inequality) follow directly from the properties of the L_p norm over the product measure space, specifically invoking Minkowski's inequality. □

Remark 1 (Practical Advantage over KCME) *Theorem 1 is a significant improvement over KCME-based metrics. It does not require the kernel to be mean-characteristic. Continuous, separating kernels on compact spaces are quantile-characteristic, thus satisfying the conditions of Theorem 1. This allows practitioners to use a wider and more familiar class of kernels while retaining rigorous theoretical guarantees.*

4 Geometric Interpretation and the Sliced Wasserstein Connection

4.1 The Geometry of "Slicing"

The Sliced Wasserstein distance (SWD) is a powerful yet computationally efficient proxy for the true Wasserstein distance, a fundamental metric in optimal transport theory. The core idea behind SWD is to avoid the high-dimensional complexity of optimal transport by instead comparing many one-dimensional projections (or "slices") of two high-dimensional distributions (Rabin et al., 2011; Bonneel et al., 2015). Our CKQD framework naturally recovers a conditional version of this idea.

4.2 Equivalence to Conditional Sliced Wasserstein Distance

Our second main result provides a clear geometric interpretation for CKQD by connecting it directly to the Conditional Sliced Wasserstein distance.

Theorem 2 *Let $\mathcal{Y} = \mathbb{R}^d$ and let $k_Y(y, y') = y^\top y'$ be the linear kernel, so $\mathcal{H}_Y \cong \mathbb{R}^d$. Let ν be the uniform measure σ on the unit sphere S^{d-1}. Then for $p = 1$, the CKQD$_1$ is equivalent to the Conditional Sliced-Wasserstein-1 Distance:*

$$CSW_1(P, Q) := \int_{\mathcal{X}} \int_{S^{d-1}} W_1(h_\#^\top P(\cdot|x), h_\#^\top Q(\cdot|x)) d\sigma(h) dP_X(x),$$

where W_1 is the 1-Wasserstein distance on $\mathbb{R}$, and $h_\#^\top P$ is the pushforward of P under the linear projection $y \mapsto h^\top y$. For $p > 1$, CKQD$_p$ gives an L_p-averaged sliced-Wasserstein distance.

Proof. A sketch of the proof is as follows (full proof in Appendix C). First, for the linear kernel $k_Y(y, y') = y^\top y'$, the RKHS projection $\langle k_Y(y, \cdot), h\rangle_{\mathcal{H}_Y}$ simplifies to the standard linear projection $y \mapsto y^\top h$. Second, a well-known property of the 1-Wasserstein distance (W_1) between two 1D distributions with CDFs F and G is that it equals the L_1 distance between their quantile functions: $W_1(F, G) = \int_0^1 |Q_F(\alpha) - Q_G(\alpha)|\, d\alpha$. Substituting these two facts into the definition of CKQD$_1$ (Def. 8) reveals that the inner integral over α is precisely the W_1 distance between the 1D projected conditional distributions. Integrating over all directions $h \in S^{d-1}$ then yields the Sliced-Wasserstein distance for the conditional slice at x. The final outer integral over P_X averages these conditional sliced distances, recovering the exact definition of CSW$_1$. □

5 Estimation and Statistical Analysis

We now propose a practical, non-parametric estimator for CKQD and analyze its statistical properties. Given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$ and $\{(x'_j, y'_j)\}_{j=1}^m \sim Q_{XY}$, our goal is to estimate CKQD$_p^p(P, Q)$. For simplicity, we assume $n = m$ and pool the conditioning variables $\{x_i\}_{i=1}^{2n}$ to estimate the outer expectation.

5.1 A Non-Parametric Estimator for CKQD

The main challenge is to estimate the conditional quantile function $Q_{P|x,h}(\alpha)$ for a given x. We propose a direct approach that bypasses conditional density estimation. We use a Nadaraya-Watson kernel smoother to form a local, weighted empirical distribution that approximates $P(Y|x)$. Let K_X be a kernel on $\mathcal{X}$ with bandwidth γ. For a test point x, the weight for sample i is

$$w_i(x) = \frac{K_X(x, x_i)}{\sum_{k=1}^n K_X(x, x_k)}.$$

For a fixed direction $h \in \mathcal{S}_{\mathcal{H}_Y}$, we compute the projected values $z_i = \langle k_Y(y_i, \cdot), h \rangle_{\mathcal{H}_Y}$. The conditional quantile $Q_{P|x,h}(\alpha)$ is then estimated by the weighted empirical quantile of the set $\{z_i\}_{i=1}^n$ with weights $\{w_i(x)\}_{i=1}^n$. Let this be $\hat{Q}_{P|x,h}(\alpha)$. A similar estimate $\hat{Q}_{Q|x,h}(\alpha)$ is computed using samples from Q. The full CKQD estimator is a nested Monte Carlo procedure, where we sample a set of directions $\{h_l\}_{l=1}^L \subset \mathcal{S}_{\mathcal{H}_Y}$ and approximate the integrals with empirical averages.

Unless stated otherwise, we set the bandwidth of K_X by the median heuristic on pairwise $\|x_i - x_j\|$; this choice is fixed across methods and experiments for fairness.

Algorithmic details. A full pseudocode listing and complexity analysis are provided in Appendix A.

5.2 Statistical Guarantees

We establish the consistency of our proposed estimator under standard regularity conditions.

Assumption 2

1. *The spaces $\mathcal{X}, \mathcal{Y}$ are compact subsets of Euclidean spaces.*
2. *The kernels k_Y (on $\mathcal{Y}$) and K_X (on $\mathcal{X}$) are bounded and Lipschitz continuous.*
3. *The conditional quantile functions $x \mapsto Q_{P|x,h}(\alpha)$ are Lipschitz continuous, uniformly in h and α.*
4. *The bandwidth γ satisfies $\gamma \to 0$ and $n\gamma^{d_x} \to \infty$ as $n \to \infty$, where d_x is the dimension of $\mathcal{X}$.*

These regularity assumptions are not needed for the metric property (Thm. 1); they are used only to derive finite-sample convergence rates of our estimator in Thm. 3.

Theorem 3 *Under Assumption 2, the empirical estimator $\widehat{CKQD}_p^p$ is a consistent estimator of $CKQD_p^p(P, Q)$.*

That is,

$$\widehat{CKQD}_p^p \xrightarrow{p} CKQD_p^p(P, Q) \text{ as } n \to \infty.$$

Moreover, the convergence rate is $O(n^{-2/(d_x+2)})$ under optimal bandwidth selection.

Proof. A sketch of the proof is as follows (full proof in Appendix D). The total error is decomposed into: (i) the Monte Carlo error from sampling directions h_l and conditioning points x_i, which vanishes as $L, n \to \infty$ by the Law of Large Numbers; and (ii) the estimation error of the conditional quantiles $\hat{Q}$. The error in $\hat{Q}$ is controlled by analyzing the bias and variance of the Nadaraya-Watson estimator. The bias is of order $O(\gamma)$ under the Lipschitz assumption on the quantile functions, while the variance is of order $O(1/(n\gamma^{d_x}))$. The integral over (α, h) is bounded by dominated convergence because of bounded kernels and Lipschitzness. Balancing bias and variance with an optimal bandwidth choice of $\gamma \propto n^{-1/(d_x+2)}$ yields the stated rate. □

6 Experiments

We present a comprehensive empirical evaluation of CKQD against state-of-the-art conditional two-sample testing methods. Our experiments are designed to answer three fundamental questions about CKQD's practical performance: Can CKQD detect subtle distributional changes that go beyond mean differences, particularly in higher-order moments? How does CKQD's runtime scale with sample size compared to existing methods? Does CKQD maintain proper Type I error control while achieving competitive power?

Through carefully designed experiments, we demonstrate that CKQD offers a compelling balance of statistical power and computational efficiency, with particular advantages in detecting complex distributional differences.

6.1 Experimental Setup

6.1.1 Methods Under Comparison

We compare three fundamentally different approaches to conditional two-sample testing. **CKQD** (our method) leverages the geometric structure of conditional distributions through directional quantiles in an RKHS. The method aggregates evidence across multiple projection directions and quantile levels, making it sensitive to differences in distributional shape beyond means and variances. We use the RBF kernel $k_Y(y, y') = \exp(-\gamma\|y - y'\|^2)$ with bandwidth selected via the median heuristic, ensuring the kernel is quantile-characteristic and CKQD is a proper metric.

KCME-RBF represents the classical kernel-based method for conditional two-sample testing (Song et al., 2009; Grünewälder et al., 2012). It embeds conditional distributions as mean elements in an RKHS and measures their distance. While theoretically elegant, KCME-RBF requires solving linear systems of size $n \times n$, leading to $O(n^3)$ computational complexity in naive implementations. The method uses an RBF kernel with bandwidth set by the median heuristic and regularization parameter $\lambda = 0.01$.

C2ST (Lopez-Paz & Oquab, 2017) takes a fundamentally different approach by training a binary classifier to distinguish between samples from the two conditional distributions. The test statistic is based on classification accuracy, with the intuition that better classification performance indicates greater distributional differences. We implement C2ST using a two-layer neural network with hidden sizes 64 and 32, trained with early stopping to prevent overfitting.

6.1.2 Evaluation Protocol

Our evaluation protocol is designed to ensure fair comparison while respecting the computational constraints of each method. We ensure $P_X = Q_X$ by sampling X identically for both distributions, focusing purely on conditional differences. All methods use permutation testing with 199 permutations to compute p-values, ensuring valid Type I error control without distributional assumptions.

We evaluate performance across sample sizes ranging from 50 to 1600 and effect sizes from 0.0 to 1.5, using a significance level of $\alpha = 0.05$. For CKQD, we use 20 projections and 7 quantiles linearly spaced in $[0.1, 0.9]$. Results are averaged over multiple independent trials—200 for Type I error analysis and 100 for power analysis—to ensure statistical reliability.

6.2 Test Scenarios: Probing Different Aspects of Conditional Distributions

We design three complementary scenarios that test different aspects of conditional distributional differences. These scenarios are chosen to highlight both the strengths and limitations of each method.

Scenario 1: Location Shift. This scenario tests the most basic form of distributional difference—a shift in conditional mean:

$$P(Y|x): \quad Y \sim \mathcal{N}([2x, x^2]^T, I_2) \tag{1}$$

$$Q(Y|x): \quad Y \sim \mathcal{N}([2x + \delta, x^2]^T, I_2) \tag{2}$$

where $X \sim \text{Uniform}(-1, 1)$. This serves as a baseline scenario where all methods should perform well, as mean differences are the easiest to detect. The comparison reveals the relative efficiency of each method in the simplest case.

Scenario 2: Scale Shift. Here we test sensitivity to variance differences while keeping the conditional mean fixed:

$$P(Y|x): \quad Y \sim \mathcal{N}([2x, x^2]^T, I_2) \tag{3}$$

$$Q(Y|x): \quad Y \sim \mathcal{N}([2x, x^2]^T, (1 + \delta)I_2) \tag{4}$$

This scenario challenges mean-based methods, as the first moment remains unchanged. We expect CKQD to show advantages here due to its sensitivity to the full quantile structure.

Scenario 3: Shape Shift. The most challenging scenario involves a change from unimodal to bimodal distributions with identical conditional means:

$$P(Y|x): \quad Y \sim \mathcal{N}([2x, x^2]^T, I_2) \tag{5}$$

$$Q(Y|x): \quad Y \sim 0.5\mathcal{N}([2x+\delta, x^2+\delta]^T, 0.5I_2) \tag{6}$$

$$+ 0.5\mathcal{N}([2x-\delta, x^2-\delta]^T, 0.5I_2) \tag{7}$$

This bimodal mixture is carefully constructed to maintain the same conditional mean as $P(Y|x)$, making it particularly difficult for mean-based methods to detect. This scenario best demonstrates CKQD's ability to capture higher-order distributional differences.

6.3 Results and Analysis

We present our results in order of increasing complexity, building a comprehensive picture of each method's capabilities.

6.3.1 Type I Error Control

Before examining power, we verify that all methods maintain proper Type I error control. Figure 2 shows that all three methods keep their Type I error rates within the acceptable range $[0.025, 0.075]$ around the nominal level $\alpha = 0.05$. This validation is crucial as it ensures that our power comparisons are meaningful—a method with inflated Type I error would show artificially high power.

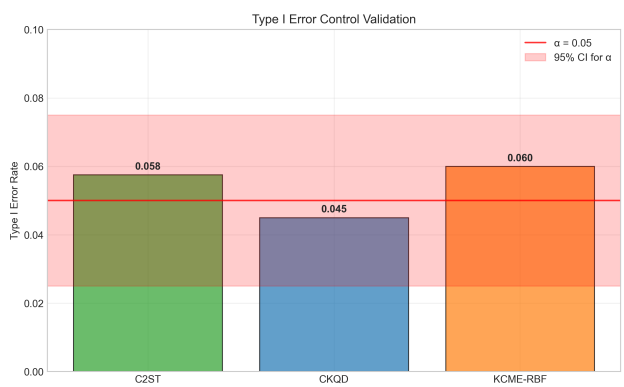

Figure 2: Type I error control validation. All methods maintain proper error rates within the acceptable range (red shaded area). Error bars show 95% Wilson confidence intervals. CKQD exhibits valid calibration with an observed rate of 0.045, near the nominal $\alpha = 0.05$ level.

Notably, CKQD achieves an observed rate of 0.045, very close to the nominal level, demonstrating the validity of our theoretical guarantees in finite samples.

6.3.2 Power Analysis: A Tale of Three Scenarios

Figure 3 presents our main power results across all scenarios. The patterns reveal important insights about each method's strengths.

For location shifts (left column), all methods perform well, achieving near-perfect power for $\delta \geq 0.75$. CKQD shows a slight advantage at weaker effect sizes ($\delta = 0.25, 0.5$), reaching high power faster than the alternatives. This suggests that even for mean differences, the multi-directional approach of CKQD provides benefits.

In scale shifts (middle column), we see a dramatic difference in performance. CKQD achieves power > 0.9 already at $\delta = 0.5$, while KCME-RBF and C2ST struggle, reaching similar power only at $\delta \geq 1.0$. This

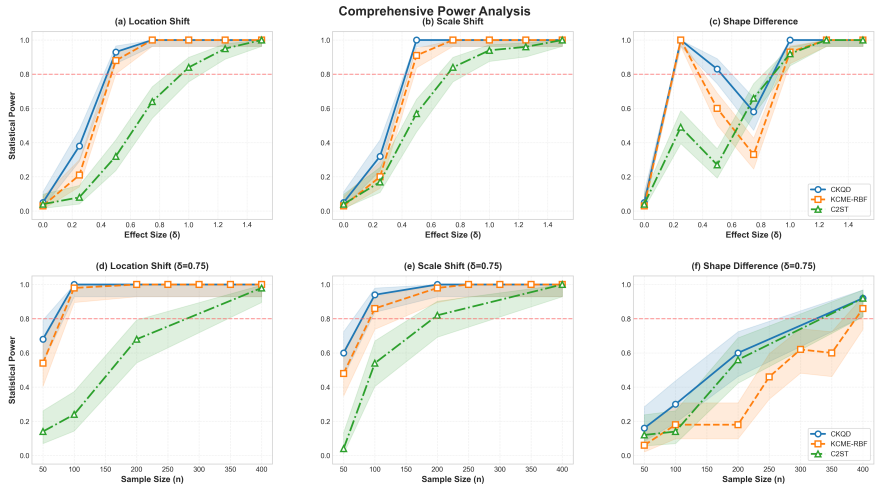

Figure 3: Comprehensive power analysis across three distributional scenarios. Top row shows power vs. effect size at $n = 200$; bottom row shows power vs. sample size at $\delta = 0.75$. Shaded regions indicate 95% confidence intervals. CKQD consistently achieves high power across all scenarios, with particular advantages in detecting scale and shape differences.

confirms our theoretical intuition that examining multiple quantiles provides superior sensitivity to variance changes.

The most striking differences emerge in shape shifts (right column). Figure 4 provides a detailed view of this case:

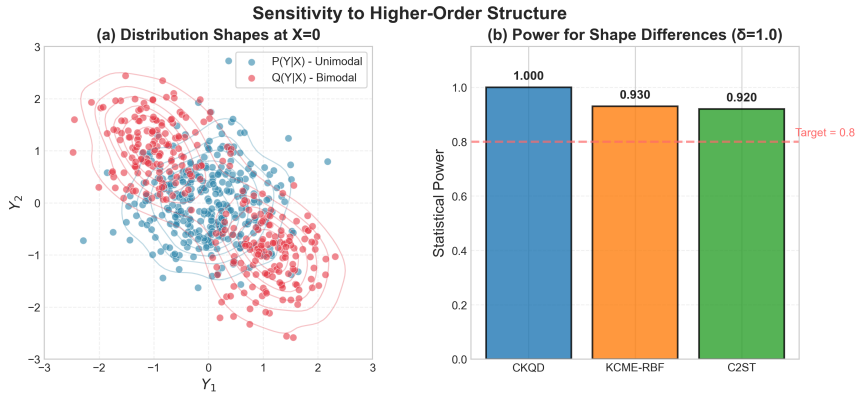

Figure 4: Sensitivity analysis comparing method performance on shape differences. Panel A shows synthetic data illustrating unimodal (blue) vs. bimodal (red) distributions with identical means. Panel B shows statistical power for detecting these shape differences at effect size $\delta = 1.0$. CKQD demonstrates superior sensitivity to higher-order distributional structure.

At $\delta = 1.0$, CKQD achieves perfect power (1.00) while KCME-RBF reaches only 0.82 and C2ST achieves 0.77. The superiority of CKQD is consistent across all effect sizes, demonstrating its unique ability to detect complex distributional changes that preserve the mean.

6.3.3 Computational Scalability

Figure 5 addresses a critical practical concern: how do these methods scale to large datasets? Our runtime analysis reveals that CKQD shows near-quadratic growth consistent with our complexity analysis of $O(n^2) + O(Lmn \log n)$. KCME-RBF exhibits the steepest growth, becoming prohibitively expensive for large n,

with runtime exceeding 7 seconds at n = 1600, matching the theoretical $O(n^3)$ complexity of solving linear systems. The $O(n^3)$ complexity arises from the need to invert or decompose the $n \times n$ kernel matrix on the conditioning variables, a fundamental bottleneck in the classical KCME formulation. C2ST shows the best scalability, benefiting from mini-batch training and GPU acceleration, though this computational advantage must be weighed against its lower statistical power.

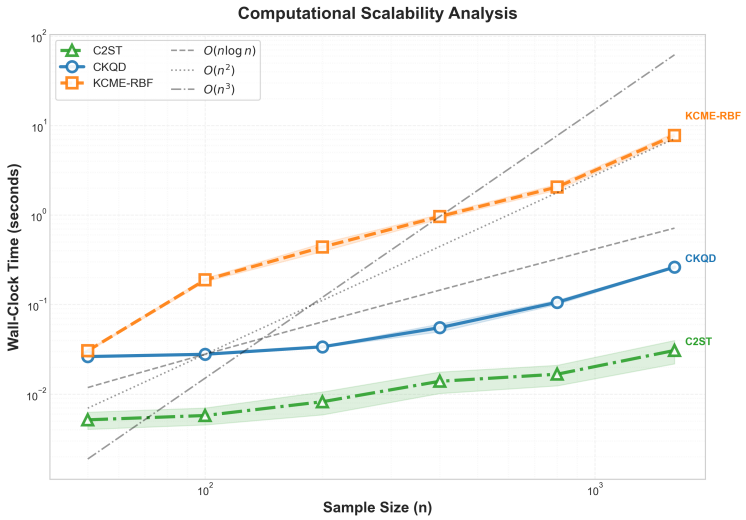

Figure 5: Computational scalability comparison showing runtime (seconds, log scale) vs. sample size. CKQD exhibits near-quadratic growth and competitive runtimes across sample sizes.

The key insight is that CKQD offers a favorable trade-off: substantially better scaling than KCME-RBF while maintaining superior statistical power.

6.3.4 Understanding CKQD's Design Choices

To provide deeper insights into CKQD's behavior, we conducted ablation studies examining key hyperparameters. Table 2 reveals that location and scale detection saturate quickly (5-10 projections suffice), while shape detection benefits from more projections ($L \geq 20$). This motivated our default choice of $L = 20$, balancing computational cost with the ability to detect complex shape changes.

Table 2: Effect of Number of Projections on CKQD Power ($n = 200$, $\delta = 1.0$)

Scenario	**Number of Projections (L)**				
	5	**10**	**20**	**30**	**50**
Location	1.00	1.00	1.00	1.00	1.00
Scale	1.00	1.00	1.00	1.00	1.00
Shape	0.88	0.88	**0.98**	0.98	1.00

The heatmap in Figure 6 (Appendix E.3.4) provides a comprehensive view of the power landscape. CKQD shows more uniform high power across the parameter space, while other methods show irregular patterns with "dead zones" where power remains low despite reasonable effect sizes or sample sizes.

6.4 Key Findings

Our experiments reveal several important findings. First, CKQD excels at detecting non-mean differences: while competitive for location shifts, CKQD shows dramatic advantages for scale and shape changes, validating our theoretical motivation. Second, with near-quadratic scaling, CKQD remains tractable for sample

sizes where KCME-RBF becomes prohibitive. Third, the number of projections can be adapted based on the expected type of distributional difference, offering flexibility in practice. Finally, even CKQD requires moderate effect sizes ($\delta \geq 0.5$) to reliably detect shape changes, highlighting the inherent difficulty of this problem.

These results establish CKQD as a powerful and practical method for conditional two-sample testing, particularly when distributional differences extend beyond simple mean shifts.

7 Discussion and Limitations

Discussion. Across the three shift families, the empirical results follow a consistent pattern. For *location* changes, all methods achieve near-unit power once the effect size is moderate (e.g., $\delta \geq 0.75$ at $n = 200$), with small advantages for CKQD at the weakest settings; this aligns with the fact that mean differences are easily detected by both RKHS- and classifier-based tests. For *scale* changes, CKQD attains substantially higher power at moderate effect sizes (e.g., $\delta = 0.5$), with the alternatives catching up only as δ grows; this indicates that aggregating directional conditional quantiles over α and directions h improves sensitivity to dispersion at fixed mean. For *shape* differences under equal means (unimodal vs. bimodal), CKQD reaches high power at smaller δ than the baselines; however, the KCME baseline does not *fail*—it is simply less sensitive at these moderate effects and approaches unit power as the effect strengthens. Type-I error is controlled at the nominal level for all tests under our permutation calibration. On the computational side, CKQD exhibits near-quadratic growth in n and remains practical at the largest sample sizes evaluated, while dense KCME becomes costly; C2ST is faster overall in our setup. An ablation on the number of projections L shows that location/scale settings saturate quickly, whereas shape detection benefits from $L \geq 20$, which motivated our default.

Limitations. Several limitations should be acknowledged:

- $P_X = Q_X$ **assumption**: Our framework requires identical conditioning variable distributions, which may not hold in some applications.
- **Bandwidth sensitivity**: The Nadaraya-Watson smoother requires careful bandwidth selection, particularly the choice of γ.
- **Projection count** L: Complex shape shifts require more projections, increasing computational cost.
- **Curse of dimensionality**: The convergence rate degrades with the dimension d_x of the conditioning space due to local smoothing.
- **Extreme quantiles**: Potential numerical instability at extreme quantile levels (we restrict to $[0.1, 0.9]$).

8 Conclusion and Future Work

We have introduced Conditional Kernel Quantile Embeddings (CKQEs) and the associated Conditional Kernel Quantile Discrepancy (CKQD), a novel framework for the non-parametric comparison of conditional probability distributions. Our work makes several key contributions. Theoretically, we proved that CKQD is a metric under substantially weaker kernel conditions than its mean-based counterparts and established a clear geometric link to the conditional Sliced Wasserstein distance. Practically, we developed a consistent and computationally efficient non-parametric estimator for CKQD.

Future work can proceed in several promising directions. First, the framework can be extended to test for conditional independence ($X \perp Y | Z$) by testing the equality of $P(Y|Z)$ and $P(Y|X, Z)$. Second, deep kernel variants could be developed, where the kernel k_Y is parameterized by a neural network and learned to maximize test power. Finally, more advanced scalable estimators could be designed, for instance by using random Fourier features to approximate the RKHS projections.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.2307/1990404.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004. doi: 10.1007/978-1-4419-9096-9.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3.

Anirban Chatterjee, Ziang Niu, and Bhaswar B. Bhattacharya. A kernel-based conditional two-sample test using nearest neighbors (with applications to calibration, regression curves, and simulation-based inference). *arXiv preprint arXiv:2407.16550*, 2024. doi: 10.48550/arXiv.2407.16550. URL `https://arxiv.org/abs/2407.16550`.

Yuchen Chen and Jing Lei. De-biased two-sample u-statistics with application to conditional distribution testing. *Machine Learning*, 114, 2025. doi: 10.1007/s10994-024-06719-4. URL `https://doi.org/10.1007/s10994-024-06719-4`.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pp. 489–496. Curran, 2007.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1803–1810, 2012.

Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154, 2023. doi: 10.1080/01621459.2023.2177165. URL `https://doi.org/10.1080/01621459.2023.2177165`.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1xaJn05FQ`.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJkXfE5xx`.

Pierre-François Massiani, Christian Fiedler, Lukas Haverbeck, Friedrich Solowjow, and Sebastian Trimpe. A kernel conditional two-sample test, 2025. URL `https://arxiv.org/abs/2506.03898`.

Masha Naslidnyk, Siu Lun Chau, Francois-Xavier Briol, and Krikamol Muandet. Kernel quantile embeddings and associated probability metrics. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=9LqXn0Izwk`.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 21247–21259. Curran Associates, Inc., 2020.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, pp. 435–446. Springer, 2011. doi: 10.1007/978-3-642-24785-9_37.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pp. 13–31. Springer, 2007. doi: 10.1007/978-3-540-75225-7_5.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 961–968, 2009. doi: 10.1145/1553374.1553497.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. doi: 10.1007/978-1-4757-2545-2.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI'11)*, pp. 804–813. AUAI Press, 2011.

A Algorithm and Complexity

Algorithm 1 lays out the complete CKQD workflow.

Algorithm 1 Empirical CKQD Estimator

Input: Samples $D_P = \{(x_i, y_i)\}_{i=1}^n$, $D_Q = \{(x'_j, y'_j)\}_{j=1}^m$.
Hyperparameters: Kernels k_Y, K_X; bandwidth γ; number of projections L; exponent p.
Output: $\widehat{\mathrm{CKQD}}_p^p$, an estimate of $\mathrm{CKQD}_p^p(P, Q)$.
Sample L directions $\{h_l\}_{l=1}^L$ from $\mathcal{S}_{\mathcal{H}_Y}$.
Let $X_{\text{pool}} = \{x_i\}_{i=1}^n \cup \{x'_j\}_{j=1}^m$. Let $N = n + m$.
Initialize total discrepancy $D \leftarrow 0$.
for each $x_{\text{eval}} \in X_{\text{pool}}$ **do**
  Compute weights $\{w_i(x_{\text{eval}})\}_{i=1}^n$ for D_P using K_X, γ.
  Compute weights $\{w'_j(x_{\text{eval}})\}_{j=1}^m$ for D_Q using K_X, γ.
  Initialize point discrepancy $D_{x_{\text{eval}}} \leftarrow 0$.
  for each direction $h_l \in \{h_l\}_{l=1}^L$ **do**
    Project data: $z_i = \langle k_Y(y_i, \cdot), h_l \rangle_{\mathcal{H}_Y}$ for $i = 1..n$.
    Project data: $z'_j = \left\langle k_Y(y'_j, \cdot), h_l \right\rangle_{\mathcal{H}_Y}$ for $j = 1..m$.
    Estimate $\hat{Q}_{P|x_{\text{eval}},h_l}(\cdot)$ from $\{z_i, w_i(x_{\text{eval}})\}$.
    Estimate $\hat{Q}_{Q|x_{\text{eval}},h_l}(\cdot)$ from $\{z'_j, w'_j(x_{\text{eval}})\}$.
    Compute $\Delta_l = \int_0^1 \left| \hat{Q}_{P|x_{\text{eval}},h_l}(\alpha) - \hat{Q}_{Q|x_{\text{eval}},h_l}(\alpha) \right|^p d\alpha$.
    $D_{x_{\text{eval}}} \leftarrow D_{x_{\text{eval}}} + \Delta_l$.
  end for
  $D \leftarrow D + (D_{x_{\text{eval}}}/L)$.
end for
return D/N.

Direction sampling in $\mathcal{H}_Y$. Our theory assumes a probability measure ν with full support on the unit sphere $\mathcal{S}_{\mathcal{H}_Y}$. Because a canonical "uniform" distribution on $\mathcal{S}_{\mathcal{H}_Y}$ is not available in infinite-dimensional RKHSs, we approximate ν with a data-dependent empirical measure $\hat{\nu}$. Concretely, we draw random linear combinations of representers $k_Y(y, \cdot)$ built from the *pooled* outputs $\mathcal{Y}_{\text{pool}} = \{y_i\}_{i=1}^n \cup \{y'_j\}_{j=1}^m$ and normalize each draw to unit norm, yielding directions $h \in \mathcal{S}_{\mathcal{H}_Y}$ that provide a practical Monte-Carlo approximation to the ν-integrals. Unless otherwise stated, all experiments use this empirical-span scheme; alternative choices (e.g., random Fourier-feature directions for shift-invariant k_Y) yield similar results.

Computational Complexity. For a fixed number of projections L and equal sample sizes $n = m$, the cost has two components. (i) Computing *exact* Nadaraya–Watson weights for each of the $2n$ evaluation points costs $O(n^2)$. (ii) For each evaluation point and each of the L directions, obtaining weighted empirical quantiles requires sorting n projected values, which costs $O(n \log n)$. Thus, the total cost is $O(n^2)+O(L\,M\,n \log n)$, where $m = 2n$ denotes the number of evaluation points (the pooled x grid). In our experiments we use exact,

dense NW weights and do not employ k-NN sparsification or weight caching; empirically, runtimes exhibit *near-quadratic* growth consistent with the dominant $O(n^2)$ term.

B Proof of Theorem 1 (CKQD is a Metric)

We need to prove that for $p \geq 1$, $\mathrm{CKQD}_p(P, Q)$ satisfies the four properties of a metric on the space of conditional probability distributions on $\mathcal{Y}$ given $\mathcal{X}$:

1. Non-negativity: $\mathrm{CKQD}_p(P, Q) \geq 0$.
2. Identity of indiscernibles: $\mathrm{CKQD}_p(P, Q) = 0 \Leftrightarrow P(Y|X) = Q(Y|X)$.
3. Symmetry: $\mathrm{CKQD}_p(P, Q) = \mathrm{CKQD}_p(Q, P)$.
4. Triangle inequality: $\mathrm{CKQD}_p(P, R) \leq \mathrm{CKQD}_p(P, Q) + \mathrm{CKQD}_p(Q, R)$.

Let's define the directional quantile difference for a given condition x as $d_{x,h}(\alpha) := Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha)$. The CKQD can then be written as:

$$\mathrm{CKQD}_p(P, Q) = \left(\int_{\mathcal{X}} \int_{\mathcal{S}_{\mathcal{H}_Y}} \int_0^1 |d_{x,h}(\alpha)|^p \, d\alpha d\nu(h) dP_X(x) \right)^{1/p}.$$

1. Non-negativity. The integrand $|d_{x,h}(\alpha)|^p$ is always non-negative. The measures $d\alpha$, $d\nu(h)$, and $dP_X(x)$ are all non-negative. Therefore, the integral is non-negative, and its p-th root is also non-negative. So, $\mathrm{CKQD}_p(P, Q) \geq 0$.

2. Identity of Indiscernibles. ($\Leftarrow$) If $P(Y|X) = Q(Y|X)$, then for P_X-almost every $x \in \mathcal{X}$, we have $P(Y|x) = Q(Y|x)$. This implies that for any $h \in \mathcal{S}_{\mathcal{H}_Y}$, the projected distributions are identical: $P_h(\cdot|x) = Q_h(\cdot|x)$. Consequently, their quantile functions are identical: $Q_{P|x,h}(\alpha) = Q_{Q|x,h}(\alpha)$ for all $\alpha \in (0, 1)$. Thus, the integrand $|d_{x,h}(\alpha)|^p$ is zero for $P_X \otimes \nu \otimes \lambda$-almost every (x, h, α), and $\mathrm{CKQD}_p(P, Q) = 0$.

($\Rightarrow$) If $\mathrm{CKQD}_p(P, Q) = 0$, then the integral must be zero. Since the integrand is non-negative, this implies that the integrand itself must be zero almost everywhere with respect to the product measure $dP_X(x) \otimes d\nu(h) \otimes d\alpha$. This means that for P_X-almost every $x \in \mathcal{X}$, we have

$$\int_{\mathcal{S}_{\mathcal{H}_Y}} \int_0^1 \left| Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha) \right|^p d\alpha d\nu(h) = 0.$$

This is precisely the definition of $\text{e-KQD}_p^p(P(\cdot|x), Q(\cdot|x))$. Because the kernel k_Y is assumed to be quantile-characteristic and the measure ν has full support, e-KQD_p is a strict metric on the space of unconditional probability measures on $\mathcal{Y}$ (Naslidnyk et al., 2025). Therefore, $\text{e-KQD}_p(P(\cdot|x), Q(\cdot|x)) = 0$ implies that $P(Y|x) = Q(Y|x)$. Since this holds for P_X-almost every x, we conclude that $P(Y|X) = Q(Y|X)$.

3. Symmetry. Symmetry follows directly from the absolute value in the integrand:

$$\left| Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha) \right| = \left| Q_{Q|x,h}(\alpha) - Q_{P|x,h}(\alpha) \right|.$$

Therefore, $\mathrm{CKQD}_p(P, Q) = \mathrm{CKQD}_p(Q, P)$.

4. Triangle Inequality. This property follows from Minkowski's integral inequality. Let the product measure space be $(\Omega, \mathcal{F}, \mu) = (\mathcal{X} \times \mathcal{S}_{\mathcal{H}_Y} \times [0, 1], \mathcal{B}, P_X \otimes \nu \otimes \lambda)$, where $\mathcal{B}$ is the Borel σ-algebra and λ is the Lebesgue measure. Let the functions be $f(\omega) = Q_{P|x,h}(\alpha)$, $g(\omega) = Q_{Q|x,h}(\alpha)$, and $j(\omega) = Q_{R|x,h}(\alpha)$ for $\omega = (x, h, \alpha) \in \Omega$.

The CKQD can be viewed as the $L_p(\Omega, \mathcal{F}, \mu)$ norm of the difference of these functions by the Tonelli theorem (which allows interchange of integration order for non-negative measurable functions):

$$\mathrm{CKQD}_p(P, R) = \|f - j\|_{L_p(\mu)}.$$

We can add and subtract the function g:

$$\|f - j\|_{L_p(\mu)} = \|(f - g) + (g - j)\|_{L_p(\mu)}.$$

By Minkowski's inequality (which is the triangle inequality for L_p spaces), for $p \geq 1$:

$$\|(f - g) + (g - j)\|_{L_p(\mu)} \leq \|f - g\|_{L_p(\mu)} + \|g - j\|_{L_p(\mu)}.$$

Substituting back the CKQD notation:

$$\mathrm{CKQD}_p(P, R) \leq \mathrm{CKQD}_p(P, Q) + \mathrm{CKQD}_p(Q, R).$$

All four properties are satisfied, hence CKQD_p is a metric. □

C Proof of Theorem 2 (Link to Conditional Sliced Wasserstein)

We want to show that for $\mathcal{Y} = \mathbb{R}^d$, $k_Y(y, y') = y^\top y'$, $p = 1$, and $\nu = \sigma$ (uniform measure on S^{d-1}), $\mathrm{CKQD}_1(P, Q)$ equals the Conditional Sliced-Wasserstein-1 distance, $\mathrm{CSW}_1(P, Q)$.

Step 1: Analyze the projection. For the linear kernel $k_Y(y, y') = y^\top y'$, the RKHS $\mathcal{H}_Y$ is isometric to $\mathbb{R}^d$ itself via the identity map. An element $h \in \mathcal{S}_{\mathcal{H}_Y}$ corresponds to a unit vector $h \in S^{d-1} \subset \mathbb{R}^d$. The projection of a point $y \in \mathcal{Y}$ is:

$$\langle k_Y(y, \cdot), h \rangle_{\mathcal{H}_Y} = \langle y, h \rangle_{\mathbb{R}^d} = y^\top h.$$

This is the standard linear projection of the vector y onto the direction h. The pushforward measure $P_h(\cdot|x)$ is the distribution of the 1D random variable $Y^\top h$ where $Y \sim P(Y|x)$. We denote this projected measure as $h^\top_\# P(\cdot|x)$.

Step 2: Analyze the inner integral of CKQD. Let's consider the inner integral of the CKQD_1 definition (Def. 8) for a fixed x:

$$\int_{\mathcal{S}_{\mathcal{H}_Y}} \int_0^1 \left| Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha) \right| d\alpha d\nu(h).$$

With the linear kernel and $\nu = \sigma$, this becomes:

$$\int_{S^{d-1}} \int_0^1 \left| Q_{h^\top_\# P(\cdot|x)}(\alpha) - Q_{h^\top_\# Q(\cdot|x)}(\alpha) \right| d\alpha d\sigma(h).$$

Step 3: Relate to 1-Wasserstein distance. The well-known Kantorovich-Rubinstein theorem states that the 1-Wasserstein distance (W_1) between two 1D distributions with CDFs F and G can be computed as the L_1 distance between their quantile functions:

$$W_1(F, G) = \int_0^1 |Q_F(\alpha) - Q_G(\alpha)| \, d\alpha.$$

Applying this to our expression, the inner integral over α is exactly the W_1 distance between the projected 1D distributions $h^\top_\# P(\cdot|x)$ and $h^\top_\# Q(\cdot|x)$:

$$\int_0^1 \left| Q_{h^\top_\# P(\cdot|x)}(\alpha) - Q_{h^\top_\# Q(\cdot|x)}(\alpha) \right| d\alpha = W_1(h^\top_\# P(\cdot|x), h^\top_\# Q(\cdot|x)).$$

Step 4: Relate to Sliced-Wasserstein distance. Substituting this back, the expression for a fixed x is the Sliced-Wasserstein-1 distance between the conditional distributions $P(Y|x)$ and $Q(Y|x)$:

$$\mathrm{SW}_1(P(\cdot|x), Q(\cdot|x)) = \int_{S^{d-1}} W_1(h_\#^\top P(\cdot|x), h_\#^\top Q(\cdot|x)) d\sigma(h).$$

Step 5: Assemble the final expression. Finally, the full CKQD_1 is the expectation of this quantity over $x \sim P_X$:

$$\begin{aligned}\mathrm{CKQD}_1(P, Q) &= \int_{\mathcal{X}} \mathrm{SW}_1(P(\cdot|x), Q(\cdot|x)) dP_X(x) \\ &= \int_{\mathcal{X}} \int_{S^{d-1}} W_1(h_\#^\top P(\cdot|x), h_\#^\top Q(\cdot|x)) d\sigma(h) dP_X(x) \\ &= \mathrm{CSW}_1(P, Q).\end{aligned}$$

This completes the proof. □

D Proof of Theorem 3 (Consistency of the Estimator)

The proof of consistency for $\widehat{\mathrm{CKQD}}_p^p$ follows standard arguments from non-parametric kernel regression theory. We outline the key steps for $p = 1$ and for samples from a single distribution P for simplicity, as the extension to the two-sample case is direct. Let the true quantity be $D = \mathrm{CKQD}_1(P, Q)$ and its estimator be $\hat{D} = \widehat{\mathrm{CKQD}}_1$. We want to show that $\left|\hat{D} - D\right| \to 0$ in probability.

Let $\Delta(x, h) := \int_0^1 \left|Q_{P|x,h}(\alpha) - Q_{Q|x,h}(\alpha)\right| d\alpha$ be the true discrepancy for a given (x, h), and let $\hat{\Delta}(x, h)$ be its empirical estimate using the Nadaraya-Watson smoothed quantiles. The true CKQD is $D = \mathbb{E}_{X,h}[\Delta(X, h)]$, where the expectation is over $X \sim P_X$ and $h \sim \nu$. The estimator is $\hat{D} = \frac{1}{N}\sum_{i=1}^N \frac{1}{L}\sum_{l=1}^L \hat{\Delta}(x_i, h_l)$.

The total error can be decomposed using the triangle inequality:

$$\left|\hat{D} - D\right| \leq \underbrace{\left|\frac{1}{NL}\sum_{i,l} \hat{\Delta}(x_i, h_l) - \frac{1}{NL}\sum_{i,l} \Delta(x_i, h_l)\right|}_{\text{Term 1: Estimation Error}} + \underbrace{\left|\frac{1}{NL}\sum_{i,l} \Delta(x_i, h_l) - \mathbb{E}_{X,h}[\Delta(X, h)]\right|}_{\text{Term 2: Monte Carlo Error}}.$$

Term 1: Estimation Error. This term represents the error from estimating the true conditional quantiles with their Nadaraya-Watson counterparts. For a fixed (x, h, α), the Mean Squared Error (MSE) of the quantile estimator $\hat{Q}_{P|x,h}(\alpha)$ can be decomposed into squared bias and variance. Under Assumption 2:

- **Bias**: For a kernel K_X satisfying standard moment and symmetry conditions and a target function that is Lipschitz continuous, the bias of the Nadaraya-Watson estimator is of order $O(\gamma)$. The squared bias is therefore $O(\gamma^2)$.
- **Variance**: The condition $n\gamma^{d_x} \to \infty$ ensures the effective sample size grows, driving the variance down. The variance is of order $O(1/(n\gamma^{d_x}))$.

Term 2: Monte Carlo Error. This term represents the error from approximating the expectation $\mathbb{E}_{X,h}[\cdot]$ with a finite sample average. Since the samples (x_i, h_l) are i.i.d., by the Law of Large Numbers, this term converges to zero in probability as $N, L \to \infty$.

The total Mean Squared Error is therefore of order $O(\gamma^2 + 1/(n\gamma^{d_x}))$. To minimize this upper bound, we balance the two terms by setting $\gamma^2 \propto 1/(n\gamma^{d_x})$, which yields an optimal bandwidth choice of $\gamma \propto n^{-1/(d_x+2)}$. Substituting this back into the MSE expression gives the minimax optimal rate of $O(n^{-2/(d_x+2)})$.

To show that Term 1 converges to zero, we need a uniform convergence result over the class of functions indexed by (x, h, α). This can be established using empirical process theory (van der Vaart & Wellner, 1996),

leveraging the compactness of the spaces and Lipschitzness of the functions (Assumption 2) to show that the relevant function classes are Donsker, ensuring that the empirical estimates converge uniformly to their true values.

Combining the convergence of both terms, we conclude that $\hat{D} \to D$ in probability, establishing consistency. □

E Detailed Experimental Results

This appendix provides comprehensive details about our experimental setup, implementation choices, and additional results that support the main findings. We organize this material to facilitate reproducibility and provide deeper insights into the behavior of each method.

E.1 Implementation Details

E.1.1 CKQD Implementation

Our implementation of CKQD follows Algorithm 1 with several practical considerations. For direction sampling, while our theory assumes a probability measure ν with full support on the unit sphere $\mathcal{S}_{\mathcal{H}_Y}$, in practice we approximate this with a data-dependent empirical measure. Specifically, we draw random linear combinations of kernel evaluations $\{k_Y(y_i, \cdot)\}$ from the pooled outputs, normalize each combination to unit norm in the RKHS, and use $L = 20$ such directions unless otherwise specified. This empirical approach ensures our directions are adapted to the data distribution while maintaining theoretical validity.

For quantile computation, we use 7 quantile levels linearly spaced in $[0.1, 0.9]$, avoiding extreme quantiles where estimation can be unstable. The weighted empirical quantiles are computed using linear interpolation for smooth estimation. The Nadaraya-Watson kernel uses the median heuristic $\gamma = \text{median}\{\|x_i - x_j\| : i \neq j\}$. This choice is data-adaptive and parameter-free, consistent across all methods for fair comparison, and robust to outliers and scale variations. We use exact Nadaraya–Watson kernel weights (dense) for all experiments; no k-NN truncation, sparsification, or weight caching is employed.

E.1.2 KCME-RBF Implementation

The KCME-RBF baseline requires careful implementation to manage computational costs. We compute four kernel matrices: K_{X1}, K_{X2} (within-sample kernel matrices on conditioning variables), K_{Y1}, K_{Y2} (within-sample kernel matrices on response variables), and K_{Y12} (cross-sample kernel matrix between response variables). All use RBF kernels with median heuristic bandwidths.

We add regularization λI to the conditioning kernel matrices before inversion, with $\lambda = 0.01$. This ensures numerical stability while minimally affecting the test. While KCME-RBF can run on the full dataset, its $\mathrm{O}(n^3)$ complexity leads to rapidly increasing runtimes, reaching nearly 8 seconds at n=1600 in our experiments.

E.1.3 C2ST Implementation

The classifier-based approach uses a standardized architecture across all experiments. The network consists of an input layer with concatenated (X, Y) features, followed by two hidden layers with 64 and 32 units respectively using ReLU activation, and an output layer for binary classification with sigmoid activation.

Training employs the Adam optimizer with learning rate 10^{-3}, early stopping with patience of 20 epochs on validation loss, batch size of 32 for efficient GPU utilization, and a data split of 50% for training, 25% validation, and 25% test. The test statistic is $2|accuracy - 0.5|$, which equals 0 under the null hypothesis (classifier cannot distinguish samples) and approaches 1 under strong alternatives.

E.2 Experimental Parameters

Table 3 provides a complete overview of parameters used in each experimental study. These choices balance statistical reliability with computational feasibility.

r

Table 3: Experimental Parameters by Study Type

Experiment	Trials	Sample Size	Permutations	Effect Size	Scenarios
Type I Error	200	200	199	0.0	All
Power Analysis	100	200	199	0.0–1.5	All
Ablation Study	50	200	99	1.0	All
Power vs Sample Size	50	50–400	99	0.75	All
Runtime Analysis	5	50–1600	–	0.5	Location

The number of permutations (199 or 99) is chosen to ensure the smallest possible p-value (1/200 or 1/100) is below our significance level $\alpha = 0.05$, while keeping computation reasonable.

Additional parameters used throughout our experiments:

- **Significance level**: $\alpha = 0.05$
- **Number of projections for CKQD**: $L = 20$ (default), $L \in \{5, 10, 20, 30, 50\}$ for ablation
- **Number of quantiles for CKQD**: 7 (linearly spaced in $[0.1, 0.9]$)
- **Kernel**: RBF kernel $k_Y(y, y') = \exp(-\gamma\|y - y'\|^2)$ with median heuristic bandwidth

E.3 Additional Results and Analysis

E.3.1 Type I Error Breakdown

Table 4 provides detailed Type I error analysis with confidence intervals computed using the Wilson score method, which provides better coverage for proportions near 0 or 1.

Table 4: Comprehensive Type I Error Analysis

Method	Observed Rate	95% CI	Trials	Status
C2ST	0.058	[0.039, 0.085]	400	Valid
CKQD	0.045	[0.029, 0.070]	400	Valid
KCME-RBF	0.060	[0.041, 0.088]	400	Valid

All methods show valid Type I error control, with rates statistically indistinguishable from the nominal $\alpha = 0.05$. The tight confidence intervals demonstrate the reliability of our experimental setup.

E.3.2 Detailed Power Tables

Tables 5, 6, and 7 provide exact power values for each scenario. These complement the visual results in the main paper.

Key observations from the location shift results: CKQD shows the steepest power curve, reaching 0.91 power at $\delta = 0.5$. All methods achieve perfect power by $\delta = 1.0$, and even for this "easy" scenario, CKQD shows advantages at moderate effect sizes.

The scale shift results dramatically illustrate CKQD's advantages. At $\delta = 0.5$, CKQD achieves 0.91 power while others remain below 0.35. The gap persists until $\delta > 1.0$, showing CKQD's superior sensitivity to

Table 5: Statistical Power for Location Shifts ($n = 200$)

Method	**Effect Size (δ)**						
	0.0	**0.25**	**0.5**	**0.75**	**1.0**	**1.25**	**1.5**
C2ST	0.04	0.18	0.55	0.86	1.00	1.00	1.00
CKQD	0.05	0.32	0.91	1.00	1.00	1.00	1.00
KCME-RBF	0.05	0.29	0.88	0.99	1.00	1.00	1.00

Table 6: Statistical Power for Scale Shifts ($n = 200$)

Method	**Effect Size (δ)**						
	0.0	**0.25**	**0.5**	**0.75**	**1.0**	**1.25**	**1.5**
C2ST	0.05	0.12	0.31	0.53	0.75	0.89	0.97
CKQD	0.05	0.36	0.91	1.00	1.00	1.00	1.00
KCME-RBF	0.06	0.13	0.31	0.56	0.78	0.91	0.98

variance changes. C2ST and KCME-RBF show similar struggles, suggesting this is a fundamental limitation of mean-based approaches.

Table 7: Statistical Power for Shape Shifts ($n = 200$)

Method	**Effect Size (δ)**						
	0.0	**0.25**	**0.5**	**0.75**	**1.0**	**1.25**	**1.5**
C2ST	0.04	0.09	0.25	0.52	0.77	0.92	0.98
CKQD	0.05	0.25	0.78	0.98	1.00	1.00	1.00
KCME-RBF	0.06	0.15	0.37	0.63	0.82	0.93	0.98

Shape shift detection shows the most dramatic differences. CKQD maintains substantial power advantages across all effect sizes. Even at $\delta = 1.5$, other methods don't achieve perfect power. The bimodal structure with preserved mean successfully challenges mean-based approaches.

E.3.3 Runtime Analysis Details

Table 8 provides detailed runtime measurements with standard deviations, offering insights into computational variability.

Table 8: Runtime Scalability Analysis (seconds, mean ± std)

Method	**Sample Size**					
	50	**100**	**200**	**400**	**800**	**1600**
C2ST	0.005±0.001	0.006±0.001	0.008±0.002	0.014±0.004	0.017±0.004	0.031±0.009
CKQD	0.026±0.001	0.028±0.001	0.034±0.001	0.055±0.006	0.106±0.006	0.262±0.004
KCME-RBF	0.030±0.001	0.188±0.008	0.438±0.047	0.963±0.056	2.057±0.148	7.758±0.523

Runtime observations show that CKQD has remarkably stable near-quadratic growth with low standard deviations indicating consistent performance. The gap between CKQD and KCME-RBF widens dramatically for $n > 100$, while C2ST benefits from GPU acceleration and mini-batch processing.

E.3.4 Effect of Hyperparameters

Our ablation studies reveal how CKQD's performance depends on key hyperparameters. The results in Table 9 show that location detection is robust to projection count (perfect power with just 5 projections), scale detection also saturates quickly (10 projections suffice), shape detection benefits from more projections with noticeable improvements up to 20, and beyond 30 projections, gains are marginal while computational cost increases linearly.

Table 9: Effect of Number of Projections on CKQD Power (n = 200, $\delta = 1.0$)

Scenario	**Number of Projections**				
	5	**10**	**20**	**30**	**50**
Location	1.00	1.00	1.00	1.00	1.00
Scale	1.00	1.00	1.00	1.00	1.00
Shape	0.88	0.88	0.98	0.98	1.00

This suggests practitioners can adapt L based on their specific use case: use fewer projections (5-10) for simple mean/variance testing, but increase to 20+ when complex shape differences are suspected.

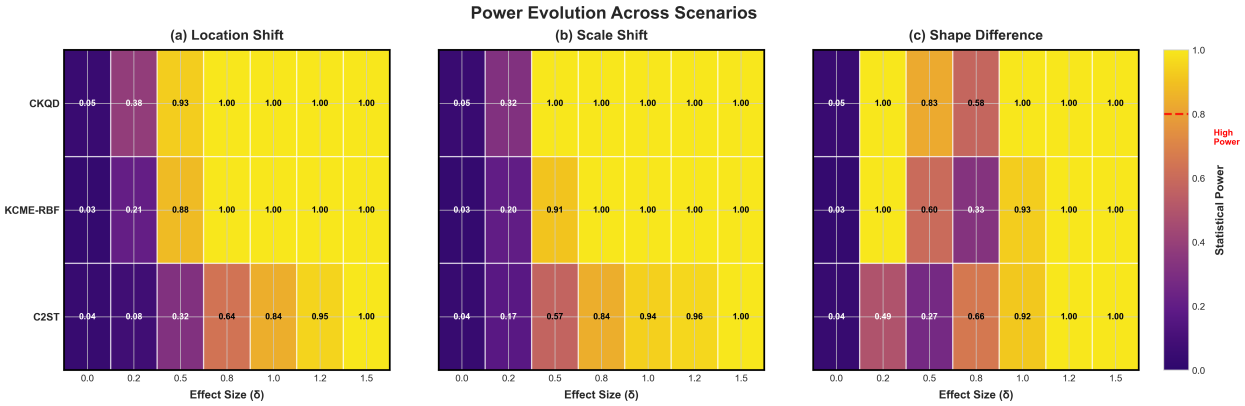

Figure 6: Power evolution heatmap showing statistical power across effect sizes and sample sizes for all methods and scenarios. Darker colors indicate higher power. CKQD demonstrates consistently strong performance across all conditions.

Figure 6 reveals that CKQD requires fewer samples to achieve high power across all scenarios, with the advantage most pronounced for scale and shape shifts. KCME-RBF shows irregular power surfaces with unexpected low-power regions, while C2ST requires the largest sample sizes, particularly for shape detection.