entropy

MDPI

Article

Measuring Statistical Dependence via Characteristic Function IPM

Povilas Daniušis [1,2,*], Shubham Juneja [3], Lukas Kuzma [3] and Virginijus Marcinkevičius [3]

[1] Neurotechnology, Laisvės av. 125A, 06118 Vilnius, Lithuania
[2] Research Institute of Natural and Technological Sciences, Vytautas Magnus University, Universiteto Str. 10, Akademija, 53361 Kaunas, Lithuania
[3] Institute of Data Science and Digital Technologies, Vilnius University, Akademijos Str. 4, 08412 Vilnius, Lithuania
* Correspondence: povilas.daniusis@vdu.lt

Abstract

We study statistical dependence in the frequency domain using the integral probability metric (IPM) framework. We propose the uniform Fourier dependence measure (UFDM) defined as the uniform norm of the difference between the joint and product-marginal characteristic functions. We provide a theoretical analysis, highlighting key properties, such as invariances, monotonicity in linear dimension reduction, and a concentration bound. For the estimation of the UFDM, we propose a gradient-based algorithm with singular value decomposition (SVD) warm-up and show that this warm-up is essential for stable performance. The empirical estimator of UFDM is differentiable, and it can be integrated into modern machine learning pipelines. In experiments with synthetic and real-world data, we compare UFDM with distance correlation (DCOR), Hilbert–Schmidt independence criterion (HSIC), and matrix-based Rényi's α-entropy functional (MEF) in permutation-based statistical independence testing and supervised feature extraction. Independence test experiments showed the effectiveness of UFDM at detecting some sparse geometric dependencies in a diverse set of patterns that span different linear and nonlinear interactions, including copulas and geometric structures. In feature extraction experiments across 16 OpenML datasets, we conducted 160 pairwise comparisons: UFDM statistically significantly outperformed other baselines in 20 cases and was outperformed in 13.

Keywords: statistical dependence; IPM; characteristic functions; uniform norm; independence testing; supervised feature extraction

Academic Editor: Boris Ryabko

Received: 6 November 2025
Revised: 3 December 2025
Accepted: 5 December 2025
Published: 12 December 2025

Citation: Daniušis, P.; Juneja, S.; Kuzma, L.; Marcinkevičius, V. Measuring Statistical Dependence via Characteristic Function IPM. *Entropy* **2025**, *27*, 1254. https://doi.org/10.3390/e27121254

1. Introduction

The estimation of statistical dependence plays an important role in various statistical and machine learning methods (e.g., hypothesis testing [1], feature selection and extraction [2,3], causal inference [4], self-supervised learning [5], representation learning [6], interpretation of neural models [7], among others). In recent years, various authors (e.g., [1,8–14]) have suggested different approaches to measuring statistical dependence.

In this paper, we focus on the estimation of statistical dependence using characteristic functions (CFs) and integral probability metric (IPM) framework. We propose and investigate a novel IPM-based statistical dependence measure, defined as the uniform norm of the difference between the joint and product-marginal CFs. After introducing core concepts, we conduct a short review of the previous work (Section 2). In Section 3, we formulate the proposed measure and its empirical estimator and perform their theoretical analysis. Section 4 is devoted to empirical investigation. Finally, in Section 5 we discuss results,

limitations, and future work. Appendix A contains technical details, such as mathematical proofs, and auxiliary tables. The main contributions of this paper are the following:

- **Theoretical and methodological contributions.** We propose a new IPM-based statistical dependence measure (UFDM) and derive its properties. The main theoretical result of this paper is the structural characterisation of UFDM, which includes invariance under linear transformations and augmentation with independent noise, monotonicity under linear dimension reduction, vanishing under independence, and a concentration bound for its empirical estimator. We additionally propose a gradient-based estimation algorithm with an SVD warm-up to ensure numerical stability.
- **Empirical analysis.** We conduct an empirical study demonstrating the practical effectiveness of UFDM in permutation-based independence testing across diverse linear, nonlinear, and geometrically structured patterns, as well as in supervised feature-extraction tasks on real datasets.

In addition, we provide the accompanying code repository https://github.com/povidanius/UFDM (accessed on 4 December 2025).

1.1. IPM Framework

In the context of estimation of statistical dependence, the IPM is a class of metrics between two probability distributions $P_{X,Y}$ and $P_X P_Y$, defined for a function class $\mathcal{F}$:

$$\mathrm{IPM}(P_{X,Y}, P_X P_Y | \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{E}_U f(U) - \mathbb{E}_V f(V)|, \tag{1}$$

where $U \sim P_{X,Y}$, and $V \sim P_X P_Y$ [15].

1.2. Characteristic Functions

Let $X \in \mathbb{R}^{d_X}$, $Y \in \mathbb{R}^{d_Y}$, and $(X^T, Y^T)^T \in \mathbb{R}^{d_X + d_Y}$ be random vectors defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us recall that their characteristic functions are given by

$$\phi(\alpha) := \mathbb{E}_X e^{i\alpha^T X}, \ \phi(\beta) := \mathbb{E}_Y e^{i\beta^T Y}, \text{ and } \phi(\alpha, \beta) := \mathbb{E}_{X,Y} e^{i(\alpha^T X + \beta^T Y)}, \tag{2}$$

where $i^2 = -1$, $\alpha \in \mathbb{R}^{d_X}$, and $\beta \in \mathbb{R}^{d_Y}$. Having n i.i.d. realisations $(x_i, y_i)_{i=1}^n$, the corresponding empirical characteristic functions (ECFs) are given by

$$\phi_n(\alpha) := \frac{1}{n} \sum_{j=1}^n e^{i\alpha^T x_j}, \ \phi_n(\beta) := \frac{1}{n} \sum_{j=1}^n e^{i\beta^T y_j}, \text{ and } \phi_n(\alpha, \beta) := \frac{1}{n} \sum_{j=1}^n e^{i(\alpha^T x_j + \beta^T y_j)}. \tag{3}$$

The uniqueness theorem states that X and Y have the same distribution if and only if their CFs are identical [16]. Therefore, CFs can be considered a description of a distribution. Alternatively, a CF ϕ can be represented as a real vector $(\Re\phi, \Im\phi) \in \mathbb{R}^2$, where $\Re$ and $\Im$ denote real and imaginary components [17]. This viewpoint avoids explicit reliance on the imaginary unit i and makes the geometric structure of CFs more transparent.

For convenience, let us define $\gamma = (\alpha^T, \beta^T)^T$, $\psi(\gamma) = \phi(\alpha)\phi(\beta)$ and let $\psi_n(\gamma) = \psi_n(\alpha, \beta)$ be its empirical counterpart. In our study, we will utilise IPM framework for investigation of the statistical dependence via

$$\Delta(\gamma) = \phi(\gamma) - \psi(\gamma) \tag{4}$$

and its empirical counterpart

$$\Delta_n(\gamma) = \phi_n(\gamma) - \psi_n(\gamma). \tag{5}$$

2. Previous Work

Various theoretical instruments have been employed for statistical dependence estimation. For example, weighted L^2 spaces and CFs (e.g., distance correlation, [13]), reproducing kernel Hilbert spaces (RKHS) (HSIC [1], DIME [18]), information theory (mutual information [19], and generalisations such as MEF [20,21]) and copula theory ([10,22]), among others. Since our work is rooted in the CF-based line of research and IPM framework, and it is empirically evaluated for independence testing and representation learning, let us consider DCOR, HSIC, and MEF, because these three measures form the compact set of high-performing baselines that span CFs, IPMs, and information-theoretic methods, which are widely used in representation learning tasks.

Distance correlation. DCOR [13] is defined as

$$\mathrm{DCOR}(X,Y) = \frac{\mathrm{DCOV}(X,Y)}{\sqrt{\mathrm{DCOV}(X,X)\,\mathrm{DCOV}(Y,Y)}},$$

where the distance covariance (DCOV) is given by

$$\mathrm{DCOV}^2(X,Y) = \int_{\mathbb{R}^{d_X+d_Y}} |\Delta(\gamma)|^2 w(\gamma) d\gamma, \tag{6}$$

with weighting function $w(\gamma) = w(\alpha,\beta) = (c_{d_X} c_{d_Y} ||\alpha||^{1+d_X} ||\beta||^{1+d_Y})^{-1}$, where $c_{d_X} = \pi^{(1+d_X)/2}/\Gamma((1+d_X)/2)$, and $c_{d_Y} = \pi^{(1+d_Y)/2}/\Gamma((1+d_Y)/2)$, and $\Gamma(.)$ is the gamma function. This weighting function allows one to avoid the direct estimation of the integral, expressing it in terms of the covariance of distances between data points [13]. The later result of [23] generalises the distance correlation to multiple random vectors. Given the i.i.d. sample pairs (x_i, y_i), $i = 1,\ldots,n$, the empirical unbiased estimator of the squared distance covariances [24] is defined as

$$\mathrm{DCOV}_n^2(X,Y) = \frac{1}{n(n-3)} \sum_{i \neq j} A_{ij} B_{ij}, \tag{7}$$

where matrices $A = (A_{ij})$, $B = (B_{ij})$ are given by

$$A_{ij} = a_{ij} - \frac{1}{n-2}\sum_{k=1}^{n} a_{ik} - \frac{1}{n-2}\sum_{k=1}^{n} a_{kj} + \frac{1}{(n-1)(n-2)} \sum_{k,\ell=1}^{n} a_{k\ell},$$

with Euclidean distance $a_{ij} = \|x_i - x_j\|$. The matrix B is defined analogously using distances $b_{ij} = \|y_i - y_j\|$. The empirical DCOR is then obtained as follows:

$$\mathrm{DCOR}_n(X,Y) = \frac{\mathrm{DCOV}_n(X,Y)}{\sqrt{\mathrm{DCOV}_n(X,X)\,\mathrm{DCOV}_n(Y,Y)}}.$$

Note that the biased version of the empirical distance-based estimator Equation (7) is equivalent to the ECF-based estimator of Equation (6) (Theorem 1, [13]). While consistency is established for the biased estimator under the moment condition $\mathbb{E}(\|X\| + \|Y\|) < \infty$ (Theorem 2, [13]), the unbiased estimator Equation (7) differs only by a finite-sample correction and converges to the same population quantity Equation (6) [24], implying consistency under the same moment condition.

HSIC. For reproducing kernel Hilbert spaces (RKHS) $\mathcal{F}$ and $\mathcal{G}$ with kernels k and l, it is defined as

$$\mathrm{HSIC}(X,Y) = \|\mathbb{E}_{XY} k(X,\cdot) \otimes l(Y,\cdot) - \mathbb{E}_X k(X,\cdot) \otimes \mathbb{E}_Y l(Y,\cdot)\|^2_{\mathrm{HS}},$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm, and $\otimes$ is the tensor product [1]. Taking a product kernel $\kappa((x,y),(x',y')) = k(x,x')l(y,y')$, HSIC is equal to the squared maximum mean discrepancy, which is an instance of an IPM with function class $\mathcal{F} = \{f : ||f||_{H_\kappa} \leq 1\}$, where H_κ is RKHS generated by κ [25]. Having a sample of paired n i.i.d. observations, the empirical estimator is

$$\mathrm{HSIC}_n(X,Y) = \frac{1}{(n-1)^2}\mathrm{tr}(KHLH)$$

with kernel matrices $K_{ij} = k(x_i,x_j)$, $L_{ij} = l(y_i,y_j)$, and centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. When both kernels k and l are translation-invariant (i.e., $k(x,x') = k_0(x-x')$ on $\mathbb{R}^{d_X}$ and $l(y,y') = l_0(y-y')$ on $\mathbb{R}^{d_Y}$, with k_0, l_0 positive definite functions such as the Gaussian $k_0(v) = \exp(-\|v\|^2/(2\sigma^2))$ with $\sigma > 0$), the product kernel $\kappa((x,y),(x',y')) = k(x,x')l(y,y') = k_0(x-x')l_0(y-y')$ is also translation-invariant on $\mathbb{R}^{d_X+d_Y}$. In this case, $\kappa(u,v) = \kappa_0(u-v)$ for some positive definite function κ_0 on $\mathbb{R}^{d_X+d_Y}$, and HSIC can be expressed in the frequency domain as

$$\mathrm{HSIC}(X,Y) = \int_{\mathbb{R}^{d_X+d_Y}} |\Delta(\gamma)|^2 \, F^{-1}\kappa_0(\gamma)\, d\gamma, \tag{8}$$

where $\gamma = (\alpha^T,\beta^T)^T$, and $F^{-1}\kappa_0$ denotes the inverse Fourier transform of κ_0. Therefore, for translation-invariant kernels, HSIC is structurally analogous to distance covariance, since it also corresponds to the squared L^2 norm of Δ (Equation (4)), with weighting determined by κ.

MEF. Shannon mutual information is defined by $\mathrm{MI}(X,Y) = \mathbb{E}_{X,Y}\log\frac{p(X,Y)}{p(X)p(Y)}$ [19]. The neural estimation of mutual information (MINE, [26]) uses its variational (Donsker–Varadhan) representation $\mathrm{MI}(X,Y) \approx max_\theta \mathbb{E}_{X,Y} f(x,y|\theta) - \log(\mathbb{E}_X\mathbb{E}_Y e^{f(x,y|\theta)})$, since it allows avoiding density estimation (here $f(x,y|\theta)$ is a neural network with parameters θ). In this case, the optimisation is performed over the space of neural network parameters, which often leads to unstable training and biased estimates due to the unboundedness of the objective and the difficulty of balancing the exponential term. The matrix-based Rényi's α-order entropy functional (MEF) [20,21,27] provides a kernel version of mutual information that avoids both density estimation and neural optimization. For random variables X and Y with distributions P_X, P_Y, and P_{XY}, it is defined as

$$\mathrm{MEF}_\alpha(X,Y) = \mathbf{S}_\alpha(P_X) + \mathbf{S}_\alpha(P_Y) - \mathbf{S}_\alpha(P_{XY}), \tag{9}$$

where $\mathbf{S}_\alpha(P_X) = \frac{1}{1-\alpha}\log_2(\mathrm{tr}(T_X^\alpha))$ and T_X is the normalised kernel integral operator on $L^2(P_X)$ [27]. Given i.i.d. samples $\{(x_i,y_i)\}_{i=1}^n$ with Gram matrices $K_{ij} = k(x_i,x_j)$ and $L_{ij} = l(y_i,y_j)$, the empirical estimator is

$$\mathrm{MEF}_{\alpha,n}(X,Y) = \mathbf{S}_{\alpha,n}\Big(\frac{K}{\mathrm{tr}(K)}\Big) + \mathbf{S}_{\alpha,n}\Big(\frac{L}{\mathrm{tr}(L)}\Big) - \mathbf{S}_{\alpha,n}\Big(\frac{K\odot L}{\mathrm{tr}(K\odot L)}\Big), \tag{10}$$

where $\odot$ denotes the element-wise product, $\mathbf{S}_{\alpha,n}(A) = \frac{1}{1-\alpha}\log_2(\sum_i \lambda_i(A)^\alpha)$, and λ_i are eigenvalues of $n \times n$ matrix A.

Motivation

The motivation of our work stems from the theoretical observation that applying the L^∞ norm to Δ Equation (4) yields a novel, structurally simple IPM with some advantageous properties, such as the ability to detect arbitrary statistical dependencies, invariance under full-rank linear transformations and coordinate augmentation with independent noise, and monotonicity under linear dimension reduction (Theorem 1).

Since the L^∞ norm isolates the most informative frequencies where dependence concentrates, we hypothesise that its empirical estimator could extract important structure from Δ that may be diluted by weighted L^2 or other global approaches such as DCOV, HSIC, and MEF.

3. Proposed Measure

Given two random vectors X and Y of dimensions d_X and d_Y, and assuming possibly unknown joint distribution $P_{X,Y}$, we define our measure via IPM with function class $\mathcal{F} = \{f : f(z) = e^{i\gamma^T z}; \gamma, z \in \mathbb{R}^{d_X+d_Y}, i^2 = -1\}$, which corresponds to the following.

Definition 1. *Uniform Fourier Dependence Measure.*

$$\mathrm{UFDM}(X,Y) = ||\Delta||_{L_\infty} = \sup_\gamma |\Delta(\gamma)|. \tag{11}$$

Since CF is a Fourier transform of a probability distribution, and the norm in L^∞ is called a uniform norm, we refer to it as Uniform Fourier Dependence Measure (UFDM).

Theorem 1. UFDM *has the following properties:*

1. $0 \le \mathrm{UFDM}(X,Y) \le 1$.
2. $\mathrm{UFDM}(X,Y) = \mathrm{UFDM}(Y,X)$.
3. $\mathrm{UFDM}(X,Y) = 0$ *if and only if* $X \perp Y$ ($\perp$ *denotes statistical independence).*
4. *For Gaussian random vectors* $X \sim N(0,\Sigma_X)$, $Y \sim N(0,\Sigma_Y)$ *with cross-covariance matrix* $\Sigma_{X,Y}$ *we have* $\mathrm{UFDM}(X,Y) = \sup_{\alpha,\beta} e^{-\frac{1}{2}(\alpha^T\Sigma_X\alpha+\beta^T\Sigma_Y\beta)}|e^{-\alpha^T\Sigma_{X,Y}\beta} - 1|$.
5. *Invariance under full-rank linear transformation:* $\mathrm{UFDM}(AX+a, BY+b) = \mathrm{UFDM}(X,Y)$ *for any full-rank matrices* $A \in \mathbb{R}^{d_X\times d_X}$, $B \in \mathbb{R}^{d_Y\times d_Y}$ *and vectors* $a \in \mathbb{R}^{d_X}$, $b \in \mathbb{R}^{d_Y}$.
6. *Linear dimension reduction does not increase* $\mathrm{UFDM}(X,Y)$.
7. *If* $X \perp \mathcal{E}$, *for any continuous function* $f : \mathbb{R}^{d_X} \to \mathbb{R}^{d_Y}$, $\lim_{\lambda\to\infty} \mathrm{UFDM}(X, f(X)+\lambda\mathcal{E}) = 0$, *if* $\mathcal{E}$ *has a density.*
8. *If* X *and* Y *have densities, then* $\mathrm{UFDM}(X,Y) \le \min\{1, \sqrt{2\mathrm{MI}(X,Y)}\}$, *where* $\mathrm{MI}(X,Y)$ *is mutual information.*
9. *Invariance to augmentation with independent noise: let* X, Y, Z *be random vectors such that* $Z \perp (X^\top, Y^\top)^\top$. *Then* $\mathrm{UFDM}((X^\top, Z^\top)^\top, Y) = \mathrm{UFDM}(X,Y)$.

Proof. See Appendix A.1. □

Interpretation of UFDM via canonical correlation analysis (CCA). In the Gaussian case, the UFDM objective reduces analytically to CCA via a closed-form expression (Theorem 1, Property 4): after whitening (setting $u = \Sigma_X^{1/2}\alpha$ and $v = \Sigma_Y^{1/2}\beta$), it becomes $\max_{u,v} e^{-\frac{1}{2}(|u|^2+|v|^2)}(1 - e^{-u^\top Kv})$, where $K = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}$. By von Neumann's inequality, the maximizers (u,v) align with the leading singular vectors of K, corresponding to the top CCA pair. Note that since Gaussian independence is equivalent to the vanishing of the leading canonical correlation ρ_1 (as all remaining correlations $0 \le \rho_j \le \rho_1$, $j > 1$ must also vanish), UFDM's focus on the leading canonical correlation entails no loss of discriminatory power.

Interpretation of UFDM via cumulants. Let us recall that $\gamma = (\alpha^T, \beta^T)^T$, $\phi(\gamma) = \phi(\alpha, \beta)$, $\psi(\gamma) = \phi(\alpha)\phi(\beta)$. For general distributions, writing $\Delta(\gamma) = \psi(\gamma)(\exp(C(\gamma)) - 1)$ offers a cumulant-series factorization, with $C(\gamma) = \log \frac{\phi(\gamma)}{\psi(\gamma)} = \sum_{p,q \geq 1} \frac{i^{p+q}}{p!q!} \langle \kappa_{p,q}, \alpha^{\otimes p} \otimes \beta^{\otimes q} \rangle$, where $\kappa_{p,q}$ are cross-cumulants and $\alpha^{\otimes p} \otimes \beta^{\otimes q}$ are the $(p+q)$-order tensors formed by the tensor product of p copies of α and q copies of β. The leading term, corresponding to $p = q = 1$, is $\frac{i^2}{1!,1!} \langle \kappa_{1,1}, \alpha \otimes \beta \rangle = -\alpha^\top \Sigma_{XY} \beta$ (with $\kappa_{1,1} = \Sigma_{XY}$ for centered variables), which aligns with the CCA interpretation, while higher-order $\kappa_{p,q}$ terms capture non-Gaussian deviations, interpreting UFDM as a frequency-domain approach that aligns (α, β) with cross-cumulant directions under marginal damping by $\psi(\gamma)$.

Remark on the representations of CFs. Since $\mathrm{UFDM}(X, Y) = \sup_\gamma \|(\Re\Delta(\gamma), \Im\Delta(\gamma))\|_2$, the UFDM objective naturally operates on the real two-dimensional vector formed by the real and imaginary parts of $\Delta(\gamma)$. This aligns with recent work on real-vector representations of characteristic functions [17] and shows that UFDM does not rely on any special algebraic role of the imaginary unit.

3.1. Estimation

Having i.i.d. observations $(X^n, Y^n) = (x_j, y_j) \sim P_{X,Y}$, $j = 1, 2, \ldots, n$, we define and discuss empirical estimators of UFDM. Recall that (Section 1.2) that $\gamma = (\alpha^T, \beta^T)^T$ and let $\phi(\alpha), \phi(\beta)$, and $\phi(\gamma)$ be CFs of X, Y, and (X, Y), respectively ($\alpha \in \mathbb{R}^{d_X}$, $\beta \in \mathbb{R}^{d_Y}$, and $\gamma \in \mathbb{R}^{d_X + d_Y}$). Let us also denote norms $||f||^t_{L_\infty} = \sup_{||\tau|| < t} |f(\tau)|$, $||f||_{L_\infty} = \sup_\tau |f(\tau)|$, for $t > 0$ and multivariate τ.

Empirical estimator. Let us define the empirical estimator of UFDM for a fixed $t > 0$:

$$\mathrm{UFDM}^t_n(X^n, Y^n) = ||\Delta_n||^t_{L_\infty}. \tag{12}$$

3.2. Estimator Convergence

The ECF is a uniformly consistent estimator of CF in each bounded subset [28] (i.e., $\lim_{n\to\infty} \sup_{||\gamma|| < t} |\phi(\gamma) - \phi_n(\gamma)| = 0$ almost surely for any fixed $t > 0$) [28]. By the triangle inequality, this implies the following:

Proposition 1. *For a fixed $t > 0$, $\lim_{n\to\infty} ||\Delta_n - \Delta||^t_{L_\infty} = 0$, almost surely.*

Theorem 2 ([29])**.** *If $t_n \to \infty$ and $\frac{\log t_n}{n} \to 0$, as $n \to \infty$, then $\lim_{n\to\infty} \sup_{||\gamma|| < t_n} |\xi(\gamma) - \xi_n(\gamma)| = 0$ almost surely for any* CF *$\xi(\gamma)$ and corresponding* ECF *$\xi_n(\gamma)$.*

This implies the convergence of the empirical estimator Equation (12):

Proposition 2. *If $t_n \to \infty$ and $\frac{\log t_n}{n} \to 0$, as $n \to \infty$, then $\lim_{n\to\infty} ||\Delta_n||^{t_n}_{L_\infty} = \mathrm{UFDM}(X, Y)$, almost surely.*

Proof. See Appendix A.1. □

Note that ECF does not converge to CF [28,29] uniformly in the entire space. Therefore, to ensure the convergence of the empirical estimator of UFDM, we need to bound the norm by slowly growing balls as in Theorem 2. The finite–sample analysis of the convergence of empirical UFDM Equation (12) to its truncated population counterpart ($\mathrm{UFDM}^t(X, Y) = ||\Delta||^t_{L_\infty}$) yields the following concentration inequality.

Theorem 3. *Let us assume that* $\mathbb{E}||X||^2 < \infty$, $\mathbb{E}||Y||^2 < \infty$. *Let us define* $d = d_X + d_Y$, $Z = (X^T, Y^T)^T$, *and* $W = ||X|| + ||Y|| + ||Z||$. *Then there exists a constant* C, *such that for every fixed* $\varepsilon > \frac{1}{n}$, $t > 0$:

$$\Pr(\,|\mathrm{UFDM}_n^t(X^n, Y^n) - \mathrm{UFDM}^t(X, Y)| > \varepsilon) \;\leq\; 2\Big(\tfrac{Ct}{\varepsilon}\Big)^d \exp\Big(-\tfrac{n}{18}\,(\tfrac{\varepsilon}{2} - \tfrac{1}{n})^2\Big) + \frac{\sigma^2}{nL^2},$$

where $L = \mathbb{E}W$, *and* $\sigma^2 = \mathbb{E}(W - L)^2$.

Proof. See Appendix A.2. □

3.3. Estimator Computation

In practice, UFDM can be estimated iteratively using Algorithm 1. Since it depends on initial parameters α and β, the complementary Algorithm 2 is designed for their data-driven initialisation. According to our experience with UFDM applications, Algorithm 2 is very important, since without it we often encountered stability issues, and initially had to rely on various heuristics, such as parameter normalisation to the unit sphere. In our opinion, this is because Δ_n is a highly nonlinear optimisation surface (especially in larger dimensions), which complicates the finding of the corresponding maxima.

Algorithm 1 UFDM estimation

Require: Number of iterations N, batch size n_b, initial $\alpha \in \mathbb{R}^{d_X}, \beta \in \mathbb{R}^{d_Y}$.
  for *iteration* $= 1$ to N **do**
    Sample batch $(X^{n_b}, Y^{n_b}) = (x_i, y_i)_{i=1}^{n_b}$.
    Standardise (X^{n_b}, Y^{n_b}) to zero mean and unit variance.
    $\alpha, \beta \leftarrow \mathrm{AdamW}([\alpha, \beta], -|\Delta_{n_b}(\alpha, \beta)|)$.
  end for
  return $\Delta(\alpha, \beta)$, α, β

Algorithm 2 SVD warm-up

Require: Batch size n_b.
  Sample batch $(X^{n_b}, Y^{n_b}) = (x_i, y_i)_{i=1}^{n_b}$.
  Compute cross-covariance $C = (X^{n_b})^\top Y^{n_b}/n_b$.
  Decompose: $[U, \Sigma, V^H] = \mathrm{SVD}(C)$.
  $\alpha \leftarrow U_{:,1}, \quad \beta \leftarrow V_{1,:}^{H,\top}$.
  return α, β

The computational complexity of Algorithm 2 consists of cross-covariance computation and finding its SVD a complexity of $O(n_b d_X d_Y + d_X d_Y \min(d_X, d_Y))$. Having initialisation of α and β, the complexity of Algorithm 1 is $O(N\, n_b\, (d_X + d_Y))$. Hence, the total computational complexity of the sequential application of Algorithm 2 and Algorithm 1 is $O(n_b d_X d_Y + d_X d_Y \min(d_X, d_Y) + N\, n_b\, (d_X + d_Y))$. Finally, having the optimal α^* and β^* computed by Algorithm 1, the evaluation of empirical UFDM has computational complexity linear in sample size.

4. Experiments

For UFDM, we used SVD warm-up (Algorithm 2) for parameter initialisation and fixed truncation parameter t to 25.0. For kernel measures, HSIC and MEF, we used Gaussian kernels for both X and Y, with a bandwidth selected using median heuristics [30]. For MEF measure α was set to 1.01, as in [21].

4.1. Permutation Tests

Permutation tests with UFDM. We compared UFDM, DCOR, HSIC, and MEF in permutation-based statistical independence testing ($H_0 : X \perp Y$ versus the alternative $H_1 : X \not\perp Y$) using a set of multivariate distributions. We investigated scenarios with a sample size of $n = 750$ and data dimensions $d \in \{5, 15, 25\}$ ($d_X = d_Y = d$). To ensure valid finite-sample calibration, permutation p-values were computed with the Phipson–Smyth correction [31].

Hyperparameters. We used 500 permutations per p-value. The number of iterations in UFDM estimation Algorithm 1 was set to 100. The batch size equaled the sample size ($n = 750$). We used a learning rate of 0.025. Due to the high computation time (permutation tests took $\approx$ 6.3 days on five machines with Intel i7 CPU, 16GB of RAM, and Nvidia GeForce RTX 2060 12 GB GPU), we relied on 500 p-values for each test in the H_0 scenario and on 100 p-values for each test in the H_1 scenario.

Distributions analysed. In the H_0 case, X was sampled from multivariate uniform, Gaussian, and Student $t(3)$ distributions (corresponding to no-tail, light-tail, and heavy-tail scenarios, respectively), and Y was independently sampled from the same set of distributions. Afterwards, we examined the uniformity of the p-values obtained from permutation tests using different statistical measures, through QQ-plots and Kolmogorov–Smirnov (KS) tests.

In the H_1 case, X and Y were related through statistical dependencies described in Table A2. These dependencies include structured dependence patterns, where X was sampled from the same set of distributions (multivariate uniform, Gaussian, and Student $t(3)$), and Y was generated as $Y = f(X) + 0.1\epsilon$, with ϵ denoting additive Gaussian noise independent of X. We also examined more complex dependencies (Table A2), where the relationship between X and Y was modeled using copulas, bimodal, circular, and other nonlinear patterns. Using this setup, we evaluated the empirical power of the permutation tests based on the same collection of statistical measures.

Results for H_0. As shown in Figure 1, UFDM, DCOR, HSIC, and MEF exhibited approximately uniform permutation p-values across all distribution pairs and dimensions, with empirical false rejection rates (FRR) remaining close to the nominal 0.05 level. Isolated low KS p-values below 0.05 occurred in only two cases: one for MEF in the Gaussian/Gaussian pair at dimension 5 (p-value of 0.01) and one for UFDM in the Gaussian/Student-t pair at dimension 5 (p-value of 0.03), suggesting minor sampling variability rather than systematic deviations from uniformity. These results show that UFDM remained comparably stable to DCOR, HSIC and MEF, in terms of type-I error control under H_0.

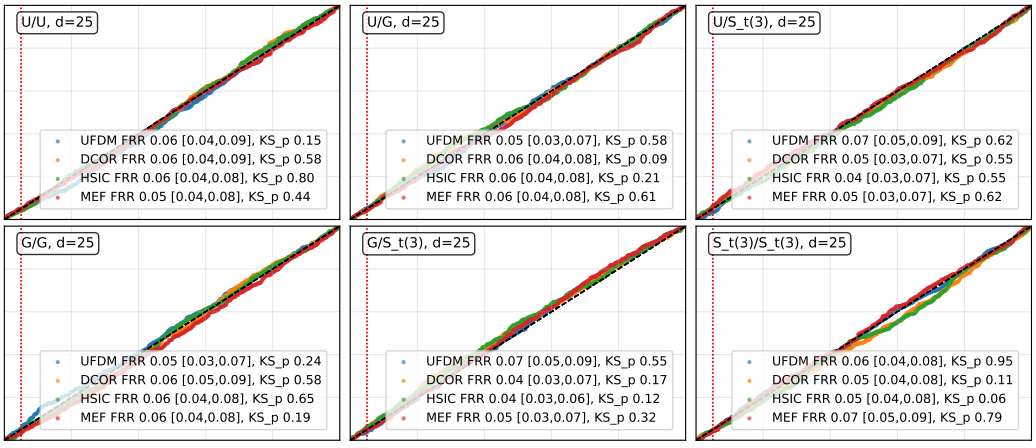

Figure 1. *Cont.*

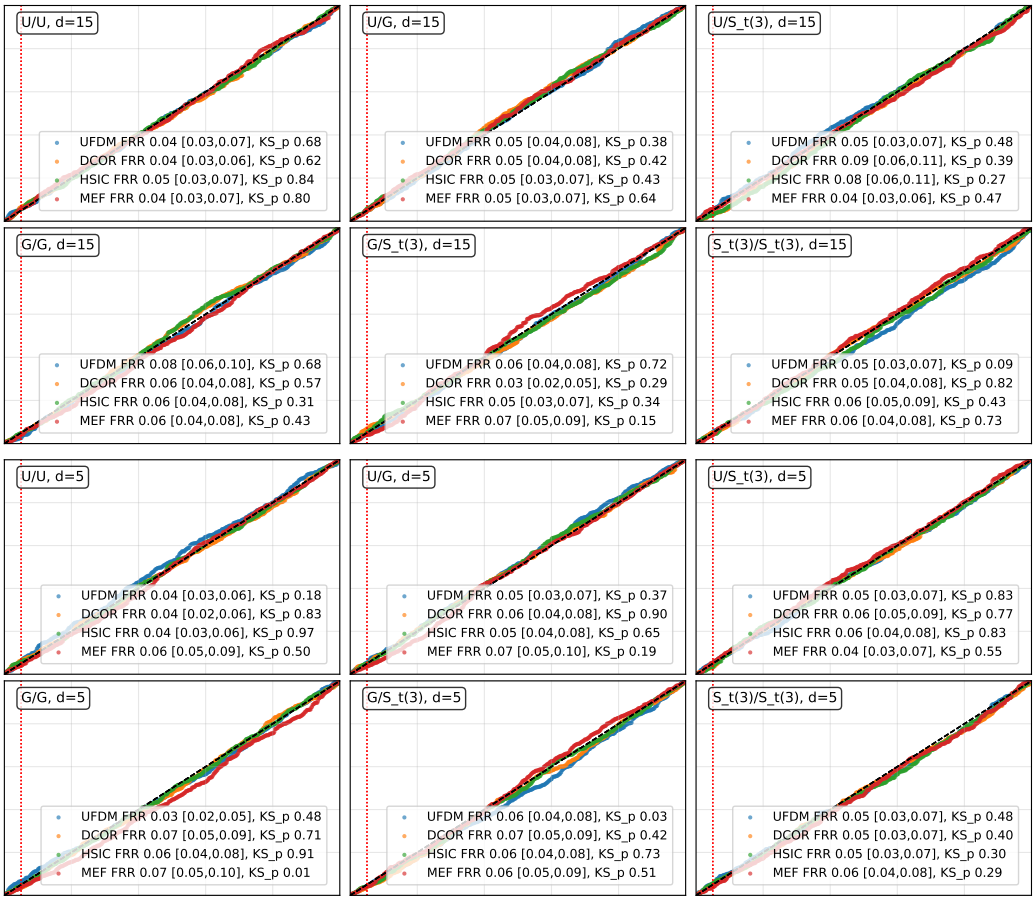

Figure 1. Empirical QQ-plots of p-values under H_0. The dashed vertical line corresponds to the nominal significance level 0.05. The empirical FRR and its Wilson confidence interval, p-values of KS test are reported in the legend.

Results for H_1. The empirical power and its 0.95-Wilson confidence intervals (CIs) are presented in Tables 1 and 2. These results show that, in most cases, the empirical power of UFDM, DCOR, HSIC, and MEF was approximately equal to 1.00. However, Table 2 also reveals that for the sparse *Circular* and *Interleaved Moons* patterns ($d \geq 15$), MEF exhibited a noticeable decrease in empirical power. We conjecture that this reduction may stem from MEF's comparatively higher sensitivity to kernel bandwidth selection in these specific, geometrically structured patterns. On the other hand, UFDM's robustness in these settings may also be explained by its *invariance to augmentation with independent noise* (Theorem 1, Property 9), which helps to preserve the detectability of sparse geometric dependencies embedded within high-dimensional noise coordinates.

Ablation experiment. The necessity of the SVD warm-up (Algorithm 2) is empirically demonstrated in Table A1, where the p-values obtained without SVD warm-up systematically fail to reveal dependence in many nonlinear patterns.

Remark on the stability of the estimator. Since the UFDM objective is non-convex, different random initialisations may potentially lead to distinct local optima. To assess the impact of this issue, we investigated the numerical stability of the UFDM estimator. We computed the mean and standard deviation of the statistic across 50 independent runs for each distribution pattern and dimension (Tables 1 and 2), as well as for the corresponding permuted patterns in which dependence is destroyed, as reported in Table 3. The obtained results align with the permutation test findings. While a slight upward shift is observed under independent (permuted) data, the proposed estimator retained consistent separa-

tion between dependent and independent settings and exhibited stable behaviour across random restarts.

Table 1. Empirical power and Wilson CIs for the dependent data (structured dependence patterns) at $\alpha = 0.05$.

Distribution of Y	UFDM	DCOR	HSIC	MEF		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 5$	
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 15$	$X \sim U[0,1]^d$
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 25$	
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 5$	
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 15$	$X \sim \mathcal{N}(0, I_d)$
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	0.97 [0.92, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 25$	
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		

Table 1. *Cont.*

Distribution of Y	UFDM	DCOR	HSIC	MEF		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 5$	$X \sim$ Student's $t(3)$
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	0.98 [0.93, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Polynomial	0.96 [0.90, 0.98]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 15$	
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	0.98 [0.93, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	0.99 [0.95, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.99 [0.95, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Linear (1.0)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 25$	
Linear (0.3)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Logarithmic	0.97 [0.92, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Quadratic	0.98 [0.93, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Polynomial	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
LRSO (0.05)	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		
Heteroscedastic	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]		

Table 2. Empirical power with 95% Wilson confidence intervals for dependent data (complex dependence patterns) at $\alpha = 0.05$.

Pattern	UFDM	DCOR	HSIC	MEF	
Mixture Bimodal Marginal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 5$
Mixture Bimodal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Circular	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Gaussian Copula	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Clayton Copula	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Interleaved Moons	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Mixture Bimodal Marginal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 15$
Mixture Bimodal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Circular	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.87 [0.79, 0.92]	
Gaussian Copula	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Clayton Copula	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Interleaved Moons	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.49 [0.39, 0.59]	
Mixture Bimodal Marginal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	$d = 25$
Mixture Bimodal	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Circular	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.52 [0.42, 0.62]	
Gaussian Copula	0.98 [0.93, 0.99]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Clayton Copula	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	
Interleaved Moons	1.00 [0.96, 1.00]	0.97 [0.92, 0.99]	0.96 [0.90, 0.98]	0.27 [0.19, 0.36]	

Table 3. UFDM statistic (mean ± std) under true dependence/permuted independence.

Dependence Pattern	$d = 5$	$d = 15$	$d = 25$	
Linear (strong)	$0.191 \pm 0.035/0.023 \pm 0.010$	$0.208 \pm 0.012/0.045 \pm 0.009$	$0.231 \pm 0.011/0.083 \pm 0.015$	$X \sim U[0,1]^d$
Linear (weak)	$0.123 \pm 0.035/0.023 \pm 0.009$	$0.143 \pm 0.014/0.044 \pm 0.011$	$0.166 \pm 0.012/0.081 \pm 0.015$	
Logarithmic	$0.172 \pm 0.048/0.023 \pm 0.009$	$0.189 \pm 0.022/0.042 \pm 0.008$	$0.195 \pm 0.014/0.087 \pm 0.015$	
Quadratic	$0.199 \pm 0.048/0.025 \pm 0.013$	$0.200 \pm 0.020/0.045 \pm 0.009$	$0.212 \pm 0.017/0.082 \pm 0.017$	
Polynomial	$0.185 \pm 0.047/0.023 \pm 0.010$	$0.195 \pm 0.026/0.047 \pm 0.013$	$0.208 \pm 0.015/0.083 \pm 0.016$	
Contaminated sine	$0.059 \pm 0.009/0.006 \pm 0.002$	$0.080 \pm 0.008/0.009 \pm 0.004$	$0.116 \pm 0.007/0.015 \pm 0.006$	
Conditional variance	$0.102 \pm 0.024/0.023 \pm 0.010$	$0.142 \pm 0.016/0.043 \pm 0.010$	$0.173 \pm 0.015/0.080 \pm 0.016$	
Linear (strong)	$0.240 \pm 0.013/0.042 \pm 0.011$	$0.239 \pm 0.011/0.077 \pm 0.014$	$0.250 \pm 0.010/0.101 \pm 0.009$	$X \sim \mathcal{N}(0, I_d)$
Linear (weak)	$0.230 \pm 0.011/0.044 \pm 0.014$	$0.235 \pm 0.013/0.077 \pm 0.013$	$0.244 \pm 0.012/0.104 \pm 0.013$	
Logarithmic	$0.254 \pm 0.031/0.034 \pm 0.010$	$0.184 \pm 0.031/0.051 \pm 0.013$	$0.136 \pm 0.023/0.079 \pm 0.014$	
Quadratic	$0.212 \pm 0.035/0.028 \pm 0.013$	$0.176 \pm 0.025/0.049 \pm 0.010$	$0.146 \pm 0.022/0.080 \pm 0.014$	
Polynomial	$0.190 \pm 0.041/0.027 \pm 0.009$	$0.176 \pm 0.027/0.048 \pm 0.011$	$0.174 \pm 0.020/0.073 \pm 0.009$	
Contaminated sine	$0.059 \pm 0.009/0.006 \pm 0.002$	$0.082 \pm 0.010/0.011 \pm 0.005$	$0.114 \pm 0.009/0.014 \pm 0.005$	
Conditional variance	$0.184 \pm 0.014/0.038 \pm 0.012$	$0.208 \pm 0.013/0.077 \pm 0.012$	$0.218 \pm 0.012/0.102 \pm 0.013$	
Linear (strong)	$0.173 \pm 0.016/0.031 \pm 0.013$	$0.181 \pm 0.017/0.053 \pm 0.012$	$0.207 \pm 0.013/0.080 \pm 0.013$	$X \sim$ Student's $t(3)$
Linear (weak)	$0.165 \pm 0.018/0.030 \pm 0.012$	$0.182 \pm 0.018/0.059 \pm 0.014$	$0.205 \pm 0.013/0.082 \pm 0.012$	
Logarithmic	$0.150 \pm 0.041/0.024 \pm 0.011$	$0.096 \pm 0.019/0.041 \pm 0.011$	$0.121 \pm 0.026/0.064 \pm 0.014$	
Quadratic	$0.082 \pm 0.037/0.014 \pm 0.007$	$0.078 \pm 0.020/0.029 \pm 0.010$	$0.097 \pm 0.022/0.048 \pm 0.012$	
Polynomial	$0.037 \pm 0.022/0.009 \pm 0.004$	$0.050 \pm 0.023/0.016 \pm 0.008$	$0.085 \pm 0.020/0.033 \pm 0.012$	
Contaminated sine	$0.057 \pm 0.008/0.006 \pm 0.002$	$0.078 \pm 0.009/0.011 \pm 0.004$	$0.115 \pm 0.008/0.014 \pm 0.004$	
Conditional variance	$0.124 \pm 0.018/0.027 \pm 0.009$	$0.158 \pm 0.014/0.054 \pm 0.011$	$0.180 \pm 0.011/0.079 \pm 0.013$	
Mixture bimodal marginal	$0.496 \pm 0.007/0.048 \pm 0.011$	$0.500 \pm 0.008/0.083 \pm 0.012$	$0.500 \pm 0.008/0.101 \pm 0.012$	Complex patterns
Mixture bimodal	$0.883 \pm 0.006/0.036 \pm 0.013$	$0.935 \pm 0.006/0.047 \pm 0.016$	$0.972 \pm 0.005/0.058 \pm 0.016$	
Circular	$0.277 \pm 0.023/0.048 \pm 0.010$	$0.259 \pm 0.022/0.089 \pm 0.012$	$0.231 \pm 0.032/0.114 \pm 0.011$	
Gaussian copula	$0.241 \pm 0.011/0.038 \pm 0.016$	$0.248 \pm 0.013/0.049 \pm 0.013$	$0.254 \pm 0.010/0.056 \pm 0.013$	
Clayton copula	$0.284 \pm 0.013/0.038 \pm 0.013$	$0.287 \pm 0.013/0.047 \pm 0.014$	$0.290 \pm 0.015/0.060 \pm 0.014$	
Interleaved moons	$0.418 \pm 0.017/0.020 \pm 0.008$	$0.384 \pm 0.024/0.052 \pm 0.013$	$0.339 \pm 0.041/0.095 \pm 0.014$	

4.2. Supervised Feature Extraction

Feature construction is often a key initial step in machine learning with tabular data. These methods can be roughly classified into feature selection and feature extraction. Feature selection identifies a subset of relevant inputs, either incrementally (e.g., via univariate filters) or through other strategies, and feature extraction transforms inputs into lower-dimensional, informative representations. In our experiments, we used the latter approach because of its computational effectiveness. The total computational time for these experiments was ≈ 94.3 h on single Intel i7 CPU, 16GB of RAM, and Nvidia GeForce RTX 2060 12 GB GPU machine.

Let $(x_i, y_i)_{i=1}^n$ be a classification dataset consisting of n pairs of d_X-dimensional inputs x_i, and d_Y-dimensional one-hot encoded outputs y_i. In our experiments, we used a collection of OpenML classification datasets [32], which cover different domains, input and output dimensionalities. We randomly split the data into training, validation, and test sets using the proportions $(0.5, 0.1, 0.4)$, respectively. We followed the dependence maximisation scheme (e.g., [3,33]) by seeking

$$W^* = arg\max_W \text{DEP}(Wx, y) - \lambda \text{tr}((W^T W - I)^T (W^T W - I)), \tag{13}$$

where DEP $\in \{\text{UFDM}, \text{DCOR}, \text{HSIC}, \text{MEF}\}$. To evaluate the obtained features $f(x) = W^* x$, we used logistic regression's [34] accuracy, measured on the test set. For each baseline method, we selected the dimensions of the features that correspond to the maximal validation accuracy of the investigated method, checking all dimensions starting from 1 with a step of 10% of d_X. Similarly, we selected $\lambda \in \{0.1, 1.0, 10.0\}$. The feature extraction loss Equation (13) was optimised via Algorithm 1 for 100 epochs, with the learning rate set to 0.025, as in permutation testing experiments (Section 4.1).

Baselines. We compared the following baselines: unmodified inputs (denoted as RAW); and Equation (13) scheme with dependence measures: UFDM, DCOR, MEF, and HSIC.

We also included the neighbourhood component analysis (NCA) [35] baseline, which is specially tailored for classification.

Evaluation metrics. Let us denote $a_{r,p}(b, b'|d) = 1$, if for r runs on the dataset d the average test set accuracy of baseline b is statistically significantly higher than that of b' with p-value threshold p. For statistical significance assessment, we used Wilcoxon's signed-rank test [36]. We computed the win ranking (WR) and loss ranking (LR) as

$$\mathrm{WR}(b) = \sum_{d} \sum_{b' \neq b} a_{25,0.05}(b, b'|d) \text{ and } \mathrm{LR}(b) = \sum_{d} \sum_{b' \neq b} a_{25,0.05}(b', b|d). \tag{14}$$

Based on these metrics, Table 4 includes full information on how many cases each baseline method statistically significantly outperformed the other method.

Results. Using 18 datasets, we conducted 80 feature efficiency evaluations (excluding the RAW baseline) and 160 feature efficiency comparisons, of which 97 (∼60%) were statistically different. The results of the feature extraction experiments are presented in Tables 4 and 5. They reveal that, although MEF showed best WR, UFDM also performed comparable to other measures: it statistically significantly outperformed them in $6 + 4 + 5 + 5 = 20$ cases (listed in Table 6), and was outperformed in $2 + 4 + 2 + 5 = 13$ cases (Table 4).

In addition to pairwise statistical comparisons using Wilcoxon's test, we also conducted statistical analysis to clarify whether some method is globally better or worse over multiple datasets using the methodology described in [37]. In this analysis, the Friedman/Iman–Davenport test ($\alpha = 0.05$) showed a global significant difference between the five methods. The Nemenyi post hoc test ($\alpha = 0.05$, critical difference 1.884) revealed that RAW was significantly outperformed by the other methods; however, it also showed the absence of a global best-performing method.

Table 4. Pairwise wins matrix: entry (i, j) is the number of cases where the method in row i outperformed the method in column j (Wilcoxon's signed-rank test, 25 runs, p-value threshold 0.05).

	UFDM	**DCOR**	**MEF**	**HSIC**	**NCA**
UFDM	0	6	4	5	5
DCOR	2	0	3	4	3
MEF	4	8	0	9	7
HSIC	2	4	2	0	3
NCA	5	7	6	8	0

Table 5. Classification accuracy comparison. n denotes dataset size, d_X is input dimensionality, and n_c is the number of classes. Best-performing method that is also statistically significant when compared with all other methods (Wilcoxon's signed-rank test, 25 runs, p-value threshold 0.05) is indicated in bold (otherwise, best-performing method is underlined).

Dataset	(n, d_X, n_c)	**RAW**	**UFDM**	**DCOR**	**MEF**	**HSIC**	**NCA**
Australian	(690, 14, 2)	0.710	0.853	0.846	0.850	0.824	0.844
Collins	(500, 22, 2)	0.840	0.926	0.906	0.941	0.927	**0.949**
Heart-statlog	(270, 13, 2)	0.621	0.824	0.823	0.826	0.816	0.817
Mfeat-factors	(2000, 216, 10)	0.783	0.968	0.970	0.968	0.968	0.969
Mfeat-pixel	(2000, 240, 10)	0.946	0.956	0.948	0.957	0.951	**0.959**
Mfeat-zernike	(2000, 47, 10)	0.741	0.812	0.810	0.814	0.811	0.804
Micro-mass	(360, 1300, 10)	0.874	0.925	0.919	0.931	0.923	0.882
Optdigits	(5620, 64, 10)	0.949	0.964	0.961	0.960	0.957	0.963
Parkinsons	(195, 22, 2)	0.756	0.827	0.828	0.850	0.836	0.837
Scene	(2407, 299, 2)	0.886	0.987	0.988	0.953	0.988	0.962
Segment	(2310, 19, 7)	0.760	0.912	0.911	**0.943**	0.936	0.941
Sonar	(208, 60, 2)	0.685	0.745	0.733	0.757	0.734	0.770
Spectf	(349, 44, 2)	0.729	0.737	0.739	0.738	0.739	0.750
USPS	(9298, 256, 10)	0.924	**0.944**	0.941	0.934	0.936	0.940
Wdbc	(569, 30, 2)	0.699	0.948	0.951	0.938	0.900	**0.968**
Wine	(178, 13, 3)	0.552	0.945	0.917	**0.954**	0.947	0.936
WR(b)			20	12	28	11	26
LR(b)			13	25	15	26	18

Table 6. Twenty cases (Measures Outperformed) where UFDM outperformed the other baselines.

Dataset	n	d_X	Measures Outperformed
Australian	690	14	DCOR, HSIC, NCA
Collins	500	22	DCOR
Micro-mass	360	1300	NCA
Mfeat-pixel	2000	240	DCOR, HSIC
Mfeat-zernike	2000	47	NCA
Optdigits	5620	64	DCOR, MEF, HSIC
Scene	2407	299	MEF, NCA
USPS	9298	256	DCOR, MEF, HSIC, NCA
Wdbc	569	30	MEF, HSIC
Wine	178	13	DCOR

5. Conclusions

Results. We proposed and analysed an IPM-based statistical dependence measure, UFDM, defined as the L^∞ norm of the difference between the joint and product-marginal characteristic functions. UFDM applies to pairs of random vectors of possibly different dimensions and can be integrated into modern machine learning pipelines. In contrast to global measures (e.g., DCOR, HSIC, MEF), which aggregate information across the entire frequency domain, UFDM identifies spectrally localised dependencies by highlighting frequencies where the discrepancy is maximised, thereby offering potentially interpretable insights into the structure of dependence. We theoretically established key properties of UFDM, such as invariance under linear transformations and augmentation with independent noise, monotonicity under dimension reduction, and vanishing under independence. We also showed that UFDM's objective aligns with the vectorial representation of CFs. In addition, we investigated the consistency of the empirical estimator and derived a finite-sample concentration bound. For practical estimation, we proposed a gradient-based estimation algorithm with SVD warm-up, and this warm-up was found to be essential for stable convergence.

We evaluated UFDM on simulated and real data in permutation-based independence testing and supervised feature extraction. The permutation test experiments ($n = 750$, $d \in \{5, 15, 25\}$) indicated that in this regime UFDM performed comparably to established baseline measures, exhibiting similar empirical power and calibration across diverse dependence structures. Notably, UFDM maintained high power on the *Circular* and *Interleaved Moons* datasets, where some other measures displayed reduced sensitivity under these geometrically structured dependencies. These findings suggest that UFDM provides a complementary addition to the family of widely used dependence measures (DCOR, HSIC, and MEF).

Further experiments with real data demonstrated that, in dependence-based supervised feature extraction, UFDM often performed on par with the well-established alternatives (HSIC, DCOR, MEF) and with NCA, which is specifically designed for classification. Across 16 datasets and 160 pairwise comparisons, UFDM statistically significantly outperformed other baselines in 20 cases and was outperformed in 13. To facilitate reproducibility, we provide an open-source repository.

Limitations. Computing UFDM requires maximising a highly nonlinear objective, which makes the estimator sensitive to initialisation and optimisation settings. Although the proposed SVD warm-up substantially improves numerical stability, estimation may still become more challenging as dimensionality d increases or sample size n decreases. From the perspective of the effective (n, d), our empirical evaluation covers two different tasks. First, in independence testing with synthetic data and $n = 750$ and $d \in \{5, 15, 25\}$, UFDM maintained effectiveness across diverse dependence structures. Our preliminary experiments with $n = 375, d \in \{5, 15, 25\}$, and $n = 750, d = 50$ indicate a reduction in power for several dependency patterns, whereas DCOR, HSIC, and MEF remained comparatively stable. Nonetheless, UFDM preserved its performance for sparse geometrically structured depen-

dencies (e.g., *Interleaved Moons*), where alternative measures often show more pronounced loss of sensitivity. Due to the high computational cost of UFDM permutation tests, we omitted systematic exploration of these regimes, leaving it to future work. On the other hand, in supervised feature extraction on real datasets, we examined substantially broader (n, d) ranges, including high-dimensional settings such as USPS ($n = 9298$, $d = 256$), MICRO-MASS ($n = 360$, $d = 1300$), and SCENE ($n = 2407$, $d = 299$). UFDM outperformed one or more baselines on several such datasets (Table 6), suggesting that it may be effective in some larger-dimensional machine learning tasks.

Future work and potential applications. Identifying the limit distribution of the empirical UFDM could enable faster alternatives to permutation-based statistical tests, which would also facilitate the systematic analysis of previously mentioned (n, d) settings. However, since the empirical UFDM is not a U- or V-statistic like HSIC or distance correlation, this would require a non-trivial analysis of the extrema of empirical processes. Possible extensions of UFDM include multivariate generalisations [23] and weighted or normalised variants to enhance empirical stability. From an application perspective, UFDM may prove useful in causality, regularisation, representation learning, and other areas of modern machine learning where statistical dependence serves as an optimisation criterion.

Author Contributions: Conceptualization, P.D.; methodology, P.D.; software, P.D., S.J., L.K., V.M.; validation, P.D., and V.M.; formal analysis, P.D.; writing—original draft preparation, P.D., and V.M.; writing—review and editing, P.D.; funding acquisition, P.D., and V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Vytautas Magnus University and Vilnius University.

Data Availability Statement: All synthetic data were generated as described in the manuscript; real datasets were obtained from OpenML (https://www.openml.org (accessed on 4 December 2025)). Code to reproduce the experiments is available at https://github.com/povidanius/UFDM (accessed on 4 December 2025). No additional unpublished data were used.

Acknowledgments: We sincerely thank Dominik Janzing for pointing out the possible theoretical connection between UFDM and HSIC, Marijus Radavičius for a remark that the convergence of empirical UFDM in the entire space requires special investigation, and Iosif Pinelis for [38]. We also acknowledge Pranas Vaitkus, Mindaugas Bloznelis, Linas Petkevičius, Aleksandras Voicikas, Osvaldas Putkis, and colleagues from Neurotechnology for discussions. We feel grateful to Neurotechnology, Vytautas Magnus University, and the Institute of Data Science and Digital Technologies, Vilnius University, for supporting this research. We also thank the anonymous reviewers for their valuable feedback.

Conflicts of Interest: Povilas Daniušis is employee of Neurotechnology. The paper reflects the views of the scientists and not the company. The other authors declare no conflict of interest.

Appendix A

Appendix A.1. Proofs

In the proofs, we interchangeably abbreviate $\phi_X(\alpha)$ with $\phi(\alpha)$, $\phi_Y(\beta)$ with $\phi(\beta)$, and $\phi_{X,Y}(\alpha, \beta)$ with $\phi(\gamma)$, where $\gamma = (\alpha^T, \beta^T)^T$.

Proof of Theorem 1. Property 1. By Cauchy–Schwarz inequality.

$$|\Delta(\alpha,\beta)|^2 = |\mathbb{E}_{X,Y}(e^{i\alpha^T X} - \phi_X(\alpha))(e^{i\beta^T Y} - \phi_Y(\beta))|^2 \leq \\ \leq \mathbb{E}_X|(e^{i\alpha^T X} - \phi_X(\alpha))|^2 \mathbb{E}_Y|(e^{i\beta^T Y} - \phi_Y(\beta))|^2. \quad \text{(A1)}$$

Recall that for complex numbers z and z' we have $|z-z'|^2 = |z|^2 - z\overline{z'} - \overline{z}z' + |z'|^2$, where $\overline{z}$ is complex conjugate of z. Therefore by plugging $z = e^{i\alpha^T X}$ and $z' = \phi_X(\alpha)$ from the definition of CF we obtain

$$\mathbb{E}_X|e^{i\alpha^T X} - \phi_X(\alpha)|^2 \;=\; 1 - \phi_X(\alpha)\overline{\phi_X(\alpha)} - \overline{\phi_X(\alpha)}\phi_X(\alpha) + |\phi_X(\alpha)|^2 \;=\; 1 - |\phi_X(\alpha)|^2,$$

and similarly $\mathbb{E}_Y|(e^{i\beta^T Y} - \phi_Y(\beta))|^2 = 1 - |\phi_Y(\beta)|^2$. Since the absolute value of CF is bounded by 1, we have that Equation (A1) is also bounded by 1.

Property 2.

$$\begin{aligned}\mathrm{UFDM}(X,Y) &= \sup_{\alpha,\beta} |\mathbb{E}_{X,Y} e^{i(\alpha^T X + \beta_T Y)} - \mathbb{E}_X e^{i\alpha^T X}\mathbb{E}_Y e^{i\beta^T Y}| \\ &= \sup_{\beta,\alpha} |\mathbb{E}_{Y,X} e^{i(\beta^T Y + \alpha^T X)} - \mathbb{E}_Y e^{i\beta^T Y}\mathbb{E}_X e^{i\alpha^T X}| = \mathrm{UFDM}(Y,X).\end{aligned}$$

Property 3. Let us assume that $X \perp Y$. Then $\phi_{X,Y}(\alpha,\beta) = \mathbb{E}_{X,Y} e^{i(\alpha^T X + \beta^T Y)} = \mathbb{E}_X\mathbb{E}_Y e^{i(\alpha^T X + \beta^T Y)} = \phi_X(\alpha)\phi_Y(\beta)$. Therefore, $\mathrm{UFDM}(X,Y) = 0$. On the other hand, if $\mathrm{UFDM}(X,Y) = 0$ then $\phi_{X,Y}(\alpha,\beta) = \phi_X(\alpha)\phi_Y(\beta)$ for all $\alpha \in \mathbb{R}^{d_X}$,$\beta \in \mathbb{R}^{d_Y}$. Let $\tilde{X}$ and $\tilde{Y}$ be two independent random vectors, having the same distributions as X and Y, respectively. Therefore $\phi_{X,Y}(\alpha,\beta) = \phi_X(\alpha)\phi_Y(\beta) = \phi_{\tilde{X}}(\alpha)\phi_{\tilde{Y}}(\beta) = \phi_{\tilde{X},\tilde{Y}}(\alpha,\beta)$. The uniqueness of CF [16] implies that distributions of (X,Y) and $(\tilde{X},\tilde{Y})$ coincide, from what directly follows that $X \perp Y$.

Property 4. Let $\Sigma_{X,Y}$ be cross-covariance matrix of X and Y. Since X and Y are Gaussian, we have $\phi_X(\alpha) = e^{-\frac{1}{2}\alpha^T\Sigma_X\alpha}$, $\phi_Y(\beta) = e^{-\frac{1}{2}\beta^T\Sigma_Y\beta}$, $\phi_{X,Y}(\alpha,\beta) = e^{-\frac{1}{2}(\alpha^T\Sigma_X\alpha + \beta^T\Sigma_Y\beta + 2\alpha^T\Sigma_{X,Y}\beta)}$. Therefore, by Equation (11)

$$\mathrm{UFDM}(X,Y) = \sup_{\alpha,\beta} e^{-\frac{1}{2}(\alpha^T\Sigma_X\alpha + \beta^T\Sigma_Y\beta)}|e^{-\alpha^T\Sigma_{X,Y}\beta} - 1|. \tag{A2}$$

Property 5. Since $\phi_{AX+a,BY+b}(\alpha,\beta) = e^{i\alpha^T a + i\beta^T b}\phi_{X,Y}(A^T\alpha, B^T\beta)$, and $\phi_{AX+a}(\alpha) = e^{i\alpha^T a}\phi_X(A^T\alpha)$, $\phi_{BY+b}(\beta) = e^{i\beta^T b}\phi_Y(B^T\beta)$, we have

$$\begin{aligned}\mathrm{UFDM}(AX+a,BY+b) &= \sup_{\alpha,\beta} |\phi_{AX+a,BY+b}(\alpha,\beta) - \phi_{Ax+a}(\alpha)\phi_{BY+b}(\beta)| = \\ &= \sup_{\alpha,\beta} |e^{i\alpha^T a + i\beta^T b}||\Delta(A^T\alpha, B^T\beta)| = \sup_{\alpha,\beta} |\Delta(A^T\alpha, B^T\beta)|.\end{aligned}$$

Since both A and B are full-rank matrices, and $A \in \mathbb{R}^{d_X \times d_X}$, $B \in \mathbb{R}^{d_Y \times d_Y}$, the maximization of the last equation is equivalent to the maximization of $|\Delta(\alpha,\beta)|$, which by definition is $\mathrm{UFDM}(X,Y)$.

Property 6. If $A' \in \mathbb{R}^{d_{X'} \times d_X}$, $B' \in \mathbb{R}^{d_{Y'} \times d_Y}$, $a' \in \mathbb{R}^{d_{X'}}$, $b' \in \mathbb{R}^{d_{Y'}}$ are parameters of linear dimension reduction, where $d_{X'} < d_X$, and $d_{Y'} < d_Y$, we have

$$\mathrm{UFDM}(A'X + a', B'Y + b') \leq \mathrm{UFDM}(AX + a, BY + b), \tag{A3}$$

for any A, B, a, b of the same dimensions (defined as in Property 5), because maximisation of LHS is conducted in smaller space than that of RHS. By Property 5, it follows that $\mathrm{UFDM}(AX + a, BY + b) = \mathrm{UFDM}(X,Y)$.

Property 7. The independence of X and $\mathcal{E}$ implies that

$$\text{UFDM}(X, f(X) + \lambda\mathcal{E}) = \sup_{\alpha,\beta} |\mathbb{E}e^{i(\alpha^T X + \beta^T f(X))}\phi_{\mathcal{E}}(\lambda\beta) - \phi_X(\alpha)\phi_{f(X)}(\beta)\phi_{\mathcal{E}}(\lambda\beta)|,$$

which converges to 0, since by multivariate Riemann–Lebesgue lemma [39] the common term $|\phi_{\mathcal{E}}(\lambda\beta)| \to 0$, when $\lambda \to \infty$. The multivariate Riemann–Lebesgue lemma can be applied since $\mathcal{E}$ has a density.

Property 8. Recall that the total variation distance between joint probability measure $P_{X,Y}$ and product measure $P_X P_Y$ is given by

$$\text{TV}(P_{X,Y}, P_X P_Y) = \frac{1}{2}\int |p_{X,Y}(x,y) - p_X(x)p_Y(y)|dxdy,$$

where $p_{X,Y}(x,y)$ is joint density, and $p_X(x)$, $p_Y(y)$ are marginal ones. Recall that Pinsker's inequality for total variation states that $\text{TV}(P_{X,Y}, P_X P_Y) \leq \sqrt{\frac{1}{2}\text{MI}(X,Y)}$, where $\text{MI}(X,Y)$ is mutual information between X and Y. Therefore,

$$\begin{aligned}|\Delta(\alpha,\beta)| &= \left|\int e^{i(\alpha^T x + \beta^T y)}(p_{X,Y}(x,y) - p_X(x)p_Y(y))dxdy\right| \\ &\leq \int |p_{X,Y}(x,y) - p_X(x)p_Y(y)|dxdy = 2\text{TV}(P_{X,Y}, P_X P_Y).\end{aligned}$$

By taking the supremum we have $\text{UFDM}(X,Y) \leq \min\{1, 2\text{TV}(P_{X,Y}, P_X P_Y)\} \leq \min\{1, \sqrt{2\text{MI}(X,Y)}\}$ by Property 1 and Pinsker's inequality.

Property 9. Independence condition $Z \perp (X^\top, Y^\top)^\top$ gives

$$\Delta_{(X^\top, Z^\top)^\top, Y}(\alpha_X, \alpha_Z, \beta) = \varphi_Z(\alpha_Z)\,\Delta_{X,Y}(\alpha_X, \beta).$$

Since $|\varphi_Z(\alpha_Z)| \leq 1$ and $|\varphi_Z(0)| = 1$, we have $\sup_{\alpha_X,\alpha_Z,\beta}|\Delta_{(X^\top,Z^\top)^\top,Y}| = \sup_{\alpha_X,\beta}|\Delta_{X,Y}(\alpha_X,\beta)|$. Therefore, $\text{UFDM}((X^\top, Z^\top)^\top, Y) = \text{UFDM}(X,Y)$. □

Proof of Proposition 2. Let $\epsilon > 0$. Since ECF is CF, and a product of two CFs also is CF, by Theorem 2 and triangle inequality, we can find natural number n_0 such that $\forall n > n_0$: $||\Delta - \Delta_n||^{t_n}_{L_\infty} = \sup_{||\gamma||<t_n}|\Delta(\gamma) - \Delta_n(\gamma)| = \sup_{||\gamma||<t_n}|\phi(\gamma) - \psi(\gamma) - \phi_n(\gamma) + \psi_n(\gamma)| = \sup_{||\gamma||<t_n}|\phi(\gamma) - \phi_n(\gamma) + \psi_n(\gamma) - \psi(\gamma)| \leq \sup_{||\gamma||<t_n}|\phi(\gamma) - \phi_n(\gamma)| + \sup_{||\gamma||<t_n}|\psi(\gamma) - \psi_n(\gamma)| \leq \epsilon$, almost surely. From the inverse triangle inequality for norms we have $|||\Delta||^{t_n}_{L_\infty} - ||\Delta_n||^{t_n}_{L_\infty}| \leq ||\Delta - \Delta_n||^{t_n}_{L_\infty} \leq \epsilon$, almost surely. On the other hand, along with the definition of $\text{UFDM}(X,Y) = \lim_{n\to\infty}||\Delta(\gamma)||^{t_n}_{L_\infty}$, this implies that $|\text{UFDM}(X,Y) - ||\Delta_n||^{t_n}_{L_\infty}| \leq |\text{UFDM}(X,Y) - ||\Delta||^{t_n}_{L_\infty}| + |||\Delta||^{t_n}_{L_\infty} - ||\Delta_n||^{t_n}_{L_\infty}|$ will be arbitrarily small almost surely, when n is sufficiently large. □

Appendix A.2. Proof of Theorem 3

Proof of Theorem 3. Recall that $Z = (X^T, Y^T)^T$, $\gamma = (\alpha^T, \beta^T)^T \in \mathbb{R}^d$ with $d = d_X + d_Y$ and

$$\Delta(\gamma) = \phi(\gamma) - \psi(\gamma), \qquad \Delta_n(\gamma) = \phi_n(\gamma) - \psi_n(\gamma).$$

Step 1. Lipschitz continuity. First, we will prove that $\Delta(\gamma)$ and $\Delta_n(\gamma)$ are Lipschitz continuous. For the population version, consider

$$|\Delta(\gamma) - \Delta(\gamma')| \leq |\phi(\gamma) - \phi(\gamma')| + |\psi(\gamma) - \psi(\gamma')|.$$

Since $\phi(\gamma) = \mathbb{E}\exp(i\gamma^T Z)$, by inequality $|e^{ia} - e^{ib}| \le |a-b|$, $a, b \in \mathbb{R}$

$$|\phi(\gamma) - \phi(\gamma')| \le \mathbb{E}|\exp(i\gamma^T Z) - \exp(i\gamma'^T Z)| \le \mathbb{E}|(\gamma-\gamma')^T Z| \le \|\gamma-\gamma'\|\mathbb{E}\|Z\|.$$

Similarly,

$$\begin{aligned}|\psi(\gamma) - \psi(\gamma')| = |\phi(\alpha)\phi(\beta) - \phi(\alpha')\phi(\beta')| &= |\phi(\alpha)||\phi(\beta) - \phi(\beta')| + |\phi(\beta')||(\phi(\alpha) - \phi(\alpha')| \\ &\le |\phi(\alpha) - \phi(\alpha')| + |\phi(\beta) - \phi(\beta')|,\end{aligned}$$

since $|\phi(\alpha)| \le 1$, $|\phi(\beta)| \le 1$. Therefore,

$$|\phi(\alpha) - \phi(\alpha')| \le \mathbb{E}\|X\|\|\alpha - \alpha'\|, \quad |\phi(\beta) - \phi(\beta')| \le \mathbb{E}\|Y\|\|\beta - \beta'\|.$$

Thus,

$$|\psi(\gamma) - \psi(\gamma')| \le (\mathbb{E}\|X\| + \mathbb{E}\|Y\|)\|\gamma - \gamma'\|,$$

so $\Delta(\gamma)$ is Lipschitz with constant $L = \mathbb{E}\|Z\| + \mathbb{E}\|X\| + \mathbb{E}\|Y\| < \infty$. For the empirical version,

$$|\Delta_n(\gamma) - \Delta_n(\gamma')| \le |\phi_n(\gamma) - \phi_n(\gamma')| + |\psi_n(\gamma) - \psi_n(\gamma')|,$$

where

$$|\phi_n(\gamma) - \phi_n(\gamma')| \le \frac{1}{n}\sum_{j=1}^{n} |\gamma^T Z_j - \gamma'^T Z_j| \le \left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|\right)\|\gamma - \gamma'\|,$$

and

$$|\psi_n(\gamma) - \psi_n(\gamma')| \le |\phi_n(\alpha) - \phi_n(\alpha')| + |\phi_n(\beta) - \phi_n(\beta')| \le \frac{1}{n}\left(\sum_{j=1}^{n}\|X_j\| + \sum_{j=1}^{n}\|Y_j\|\right)\|\gamma - \gamma'\|.$$

Define $L_n = \frac{1}{n}\sum_{j=1}^{n}(\|Z_j\| + \|X_j\| + \|Y_j\|)$, so $\Delta_n(\gamma)$ is Lipschitz with random constant L_n. Recall that $\mathbb{E}L_n = L$, $\mathbb{E}(L_n - L)^2 = \sigma^2/n$ are finite because of bounded second moment assumption. L_n concentrates around L, and by Cantelli's inequality, we have

$$\Pr(L_n \ge 2L) = \Pr(L_n - L \ge L) \le \frac{1}{1 + n(L/\sigma)^2} \le \frac{\sigma^2}{nL^2}. \tag{A4}$$

Step 2. Construct a δ-net and bound the deviation on the δ-net. For $B_t = \{\gamma : \|\gamma\| < t\}$, construct a δ-net $\{\gamma_1, \ldots, \gamma_{N(t,\delta)}\}$ such that every $\gamma \in B_t$ is within δ of some γ_k. The cardinality satisfies $N(t,\delta) \le (3t/\delta)^d$ [40].

For fixed γ_k, bound $|\Delta_n(\gamma_k) - \Delta(\gamma_k)|$. Changing one Z_j to Z_j' alters $\phi_n(\gamma_k)$ by at most $2/n$, $\phi_n(\alpha_k)$ and $\phi_n(\beta_k)$ by at most $2/n$ each, and $\psi_n(\gamma_k)$ by at most $4/n$. Thus, $|\Delta_n(\gamma_k) - \Delta_n'(\gamma_k)| \le 6/n$. By McDiarmid's inequality,

$$\Pr(|\Delta_n(\gamma_k) - \mathbb{E}\Delta_n(\gamma_k)| > u) \le 2\exp\left(-\frac{nu^2}{18}\right).$$

Compute the bias: $\mathbb{E}\phi_n(\gamma_k) = \phi(\gamma_k)$, and $\mathbb{E}\psi_n(\gamma_k) = \frac{1}{n}\phi(\gamma_k) + \left(1 - \frac{1}{n}\right)\psi(\gamma_k)$, so

$$\mathbb{E}\Delta_n(\gamma_k) = \left(1 - \frac{1}{n}\right)\Delta(\gamma_k), \quad |\mathbb{E}\Delta_n(\gamma_k) - \Delta(\gamma_k)| \le \frac{1}{n}.$$

Thus,

$$\Pr(|\Delta_n(\gamma_k) - \Delta(\gamma_k)| > \varepsilon) \le 2\exp\left(-\frac{n}{18}\left(\varepsilon - \frac{1}{n}\right)^2\right), \quad \varepsilon > \frac{1}{n}.$$

Step 3. Extend to the entire frequency ball. For any $\gamma \in B_t$, choose γ_k with $\|\gamma - \gamma_k\| \leq \delta$. Then we have

$$|\Delta_n(\gamma) - \Delta(\gamma)| \leq |\Delta_n(\gamma) - \Delta_n(\gamma_k)| + |\Delta_n(\gamma_k) - \Delta(\gamma_k)| + |\Delta(\gamma_k) - \Delta(\gamma)| \tag{A5}$$

$$\leq L_n\delta + |\Delta_n(\gamma_k) - \Delta(\gamma_k)| + L\delta. \tag{A6}$$

Thus, $\sup_{\gamma\in B_t} |\Delta_n(\gamma) - \Delta(\gamma)| \leq (L_n + L)\delta + \max_k |\Delta_n(\gamma_k) - \Delta(\gamma_k)|$. Then by union bound

$$\Pr\left(\sup_{\gamma\in B_t} |\Delta_n(\gamma) - \Delta(\gamma)| > \varepsilon\right) \leq \Pr\left((L_n + L)\delta > \frac{\varepsilon}{2}\right) + \Pr\left(\max_k |\Delta_n(\gamma_k) - \Delta(\gamma_k)| > \frac{\varepsilon}{2}\right). \tag{A7}$$

Recall that in Equation (A4) we showed that $\Pr(L_n > 2L) \leq \frac{\sigma^2}{nL^2}$. Choosing $\delta = \frac{\varepsilon}{6L}$ implies

$$\Pr((L_n + L)\delta > \frac{\varepsilon}{2}) = \Pr(L_n > 2L) \leq \frac{\sigma^2}{nL^2}. \tag{A8}$$

For the max term, by the union bound,

$$\Pr(\max_k |\Delta_n(\gamma_k) - \Delta(\gamma_k)| > \frac{\varepsilon}{2}) \leq 2N(t,\delta) \exp\left(-\frac{n}{18}\left(\frac{\varepsilon}{2} - \frac{1}{n}\right)^2\right), \tag{A9}$$

where $N(t,\delta) \leq (3t/\delta)^d = \left(\frac{18tL}{\varepsilon}\right)^d$.

Step 4: Final bound. Plugging Equation (A8) and Equation (A9) into Equation (A7) we have

$$\Pr(\sup_{\gamma\in B_t} |\Delta_n(\gamma) - \Delta(\gamma)| > \varepsilon) \leq 2\left(\frac{Ct}{\varepsilon}\right)^d \exp\left(-\frac{n}{18}\left(\frac{\varepsilon}{2} - \frac{1}{n}\right)^2\right) + \frac{\sigma^2}{nL^2}.$$

Finally, the stated bound follows from the inverse triangle inequality for norms. □

Appendix A.3. Ablation Experiment on SVD Warm-Up

Table A1. p-value means and standard deviations of the analysed dependence patterns in permutation tests for UFDM without SVD warm-up (Algorithm 2), uniformly initialising parameters α and β from $[-1, 1]$ interval. Here $X \sim \mathcal{N}(0, I_d)$.

Distribution of Y	$d = 5$	$d = 15$	$d = 25$
Linear (1.0)	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
Linear (0.3)	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
Logarithmic	0.035 ± 0.097	0.192 ± 0.251	0.387 ± 0.297
Quadratic	0.023 ± 0.061	0.298 ± 0.291	0.285 ± 0.145
Polynomial	0.002 ± 0.000	0.062 ± 0.134	0.056 ± 0.078
LRSO (0.05)	0.002 ± 0.001	0.041 ± 0.066	0.026 ± 0.040
Heteroscedastic	0.004 ± 0.006	0.002 ± 0.001	0.003* ± 0.003

Appendix A.4. Dependency Patterns

Table A2. Dependence structures. $Lin[a, b]$ denotes uniform linear spacing over given interval $[a, b]$, $a < b$, $X \perp \mathcal{E} \sim \mathcal{N}(0, I)$, and d is dimension. Fixed parameters $k = 6$, $\rho = 0.85$, $\theta = 5.0$. By $\odot$ we denote element-wise product.

Type	Formula
Structured dependence patterns ($X \sim \{\mathcal{N}(0, I_d), U[0,1]^d, \text{Student } t_3(0, I_d)\}$)	
Linear(p)	$Y = pWX + 0.1\mathcal{E}, p \in \mathbb{R}$
Logarithmic	$Y = \log(1.0 + WX \odot WX) + 0.1\mathcal{E}$
Quadratic	$Y = WX \odot WX + 0.1\mathcal{E}$
Cubic	$Y = 0.5(WX \odot WX \odot WX) - WX \odot WX + 0.1\mathcal{E}$
LRSO(p)	$X_0 \sim P_X, \quad Y_0 = \sin(k(w^T X_0))\mathbf{1}_d + 0.1\mathcal{E}$ (proportion $1-p$) $X_1 \perp Y_1 \sim \mathcal{N}(0, 25^2 I_d)$ (proportion p), $(X, Y) = \text{random-shuffle}(X_0 \cup X_1, Y_0 \cup Y_1)$
Heteroscedastic	$Y = (1.0 + \mathcal{E}_1)WX + 0.1\mathcal{E}, \mathcal{E}_1 \sim \mathcal{N}(0, I)$
Complex dependence patterns	
Bimodal	$S \sim \text{Uniform}(\{-1, 1\})$ $\mu_X = 2\mathbf{1}_{d_X}, \quad \mu_Y = 2\mathbf{1}_{d_Y}$ $X \sim \mathcal{N}(S\mu_X, I_{d_X})$ $Y \sim \mathcal{N}(S\mu_Y, I_{d_Y})$
Sparse bimodal	$X \sim 0.5\mathcal{N}(\mu, I_d) + 0.5\mathcal{N}(-\mu, I_d), \quad \mu = (2, 0, \ldots, 0)$
Sparse circular	$T \sim Lin[0, 2\pi], \quad R \sim \mathcal{N}(1, 0.2^2)$ $X = (R\cos T, R\sin T, \eta), \quad \eta \sim \mathcal{N}(0, I_{d-2})$ $Y = (R\cos(T+\delta), R\sin(T+\delta), \zeta) + 0.1\mathcal{E}, \quad \delta \sim \mathcal{N}(0,1),\ \zeta \sim \mathcal{N}(0, I_{d-2})$
Gaussian copula	Marginals $\sim \mathcal{N}(0, \rho\mathbf{1}_{d\times d} + (1-\rho)I_d)$.
Clayton copula	Parameter θ and standard normal marginals for each component.
Interleaved Moons	$(X_0, L_X) = \text{make\_moons}(), \quad (Y_0, L_Y) = \text{make\_moons}()$ For each sample i: $X_i^{(1,2)} = (X_0)_i$ $Y_i^{(1,2)} \sim \text{Uniform}\{(Y_0)_j \mid (L_Y)_j \neq (L_X)_i\}$ $X_i^{(3:d)}, Y_i^{(3:d)} \sim \mathcal{N}(0, I_{d-2})$

We used `sklearn.datasets.make_moons`.

References

1. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings of the 16th International Conference on Algorithmic Learning Theory (ALT), Singapore, 8–11 October 2005.
2. Daniušis, P.; Vaitkus, P.; Petkevičius, L. Hilbert–Schmidt component analysis. *Lith. Math. J.* **2016**, *57* , 7–11. [CrossRef]
3. Daniušis, P.; Vaitkus, P. Supervised feature extraction using Hilbert-Schmidt norms. In Proceedings of the 10th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Burgos, Spain, 23–26 September 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 25–33.
4. Hoyer, P.; Janzing, D.; Mooij, J.M.; Peters, J.; Schölkopf, B. Nonlinear causal discovery with additive noise models. In Proceedings of the Advances in Neural Information Processing Systems 21 (NeurIPS 2008), Vancouver, BC, Canada, 8–11 December 2008.
5. Li, Y.; Pogodin, R.; Sutherland, D.J.; Gretton, A. Self-Supervised Learning with Kernel Dependence Maximization. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021.
6. Ragonesi, R.; Volpi, R.; Cavazza, J.; Murino, V. Learning unbiased representations via mutual information backpropagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Virtual, 19–25 June 2021; pp. 2723–2732.
7. Zhen, X.; Meng, Z.; Chakraborty, R.; Singh, V. On the Versatile Uses of Partial Distance Correlation in Deep Learning. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
8. Chatterjee, S. A New Coefficient of Correlation. *J. Am. Stat. Assoc.* **2021**, *116*, 2009–2022. [CrossRef]
9. Feuerverger, A. A consistent test for bivariate dependence. *Int. Stat. Rev.* **1993**, *61*, 419–433. [CrossRef]

10. Póczos, B.; Ghahramani, Z.; Schneider, J.G. Copula-based kernel dependency measures. *arXiv* **2012**, arXiv:1206.4682. [CrossRef]
11. Puccetti, G. Measuring linear correlation between random vectors. *Inf. Sci.* **2022**, *607*, 1328–1347. [CrossRef]
12. Shen, C.; Priebe, C.E.; Vogelstein, J.T. From Distance Correlation to Multiscale Graph Correlation. *J. Am. Stat. Assoc.* **2020**, *115*, 280–291. [CrossRef]
13. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
14. Tsur, D.; Goldfeld, Z.; Greenewald, K. Max-Sliced Mutual Information. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; Curran Associates, Inc.: Red Hook, NY, USA, 2023; pp. 80338–80351.
15. Sriperumbudur, B.K.; Fukumizu, K.; Gretton, A.; Schölkopf, B.; Lanckriet, G.R.G. On the empirical estimation of integral probability metrics. *Electron. J. Stat.* **2012**, *6*, 1550–1599. [CrossRef]
16. Jacod, J.; Protter, P. *Probability Essentials*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2003.
17. Richter, W.-D. On the vector representation of characteristic functions. *Stats* **2023**, *6*, 1072–1081. [CrossRef]
18. Zhang, W.; Gao, W.; Ng, H.K.T. Multivariate tests of independence based on a new class of measures of independence in Reproducing Kernel Hilbert Space. *J. Multivar. Anal.* **2023**, *195*, 105144. [CrossRef]
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
20. Yu, S.; Giraldo, L.G.S.; Jenssen, R.; Príncipe, J.C. Multivariate Extension of Matrix-Based Rényi's α-Order Entropy Functional. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2960–2966. [CrossRef]
21. Yu, S.; Alesiani, F.; Yu, X.; Jenssen, R.; Príncipe, J.C. Measuring Dependence with Matrix-based Entropy Functional. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual, 2–9 February 2021; pp. 10781–10789.
22. Lopez-Paz, D.; Hennig, P.; Schölkopf, B. The Randomized Dependence Coefficient. In Proceedings of the Advances in Neural Information Processing Systems 26 (NeurIPS 2013), Lake Tahoe, NV, USA, 5–8 December 2013; Curran Associates, Inc.: Red Hook, NY, USA, 2013.
23. Böttcher, B.; Keller-Ressel, M.; Schilling, R. Distance multivariance: New dependence measures for random vectors. *arXiv* **2018**, arXiv:1711.07775. [CrossRef]
24. Székely, G.J.; Rizzo, M.L. Partial distance correlation with methods for dissimilarities. *Ann. Stat.* **2014**, *42*, 2382–2412. [CrossRef]
25. Schölkopf, B.; Smola, A.J.; Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2018.
26. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 10–15 July 2018 ; pp. 531–540.
27. Sanchez Giraldo, L.G.; Rao, M.; Principe, J.C. Measures of Entropy From Data Using Infinitely Divisible Kernels. *IEEE Trans. Inf. Theory* **2015**, *61*, 535–548. [CrossRef]
28. Ushakov, N.G. *Selected Topics in Characteristic Functions*; De Gruyter: Berlin, Germany, 2011.
29. Csörgő, S.; Totik, V. On how long interval is the empirical characteristic function uniformly consistent. *Acta Sci. Math. (Szeged)* **1983**, *45*, 141–149.
30. Garreau, D.; Jitkrittum, W.; Kanagawa, M. Large sample analysis of the median heuristic. *arXiv* **2017**, arXiv:1707.07269.
31. Phipson, B.; Smyth, G.K. Permutation *p*-values should never be zero: Calculating exact *p*-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **2010**, *9*, 39. [CrossRef]
32. Vanschoren, J.; van Rijn, J.N.; Bischl, B.; Torgo, L. OpenML: Networked Science in Machine Learning. *SIGKDD Explor.* **2013**, *15*, 49–60. [CrossRef]
33. Zhang, Y.; Zhou, Z.H. Multilabel Dimensionality Reduction via Dependence Maximization. *ACM Trans. Knowl. Discov. Data* **2010**, *4*, 14:1–14:21. [CrossRef]
34. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series, Chapman & Hall; Routledge: Oxfordshire, UK, 1989.
35. Goldberger, J.; Hinton, G.E.; Roweis, S.; Salakhutdinov, R.R. Neighbourhood components analysis. In Proceedings of the Advances in Neural Information Processing Systems 17 (NeurIPS 2004), Vancouver, BC, Canada, 13–18 December 2004.
36. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
37. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
38. Euclidean Norm of Sub-Exponential Random Vector is Sub-Exponential? MathOverflow. Version: 2025-05-06. Available online: https://mathoverflow.net/q/492045 (accessed on 10 April 2025).

39. Bochner, S.; Chandrasekharan, K. *Fourier Transforms (AM-19)*; Princeton University Press: Princeton, NJ, USA, 1949.
40. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2018. [CrossRef]