

Deep Into the Domain Shift: Transfer Learning Through Dependence Regularization

Shumin Ma¹, Zhiri Yuan², Qi Wu³, Yiyang Huang, Xixu Hu, Cheuk Hang Leung, Dongdong Wang, and Zhixiang Huang

Abstract—Classical domain adaptation methods acquire transferability by regularizing the overall distributional discrepancies between features in the source domain (labeled) and features in the target domain (unlabeled). They often do not differentiate whether the domain differences come from the marginals or the dependence structures. In many business and financial applications, the labeling function usually has different sensitivities to the changes in the marginals versus changes in the dependence structures. Measuring the overall distributional differences will not be discriminative enough in acquiring transferability. Without the needed structural resolution, the learned transfer is less optimal. This article proposes a new domain adaptation approach in which one can measure the differences in the internal dependence structure separately from those in the marginals. By optimizing the relative weights among them, the new regularization strategy greatly relaxes the rigidity of the existing approaches. It allows a learning machine to pay special attention to places where the differences matter the most. Experiments on three real-world datasets show that the improvements are quite notable and robust compared to various benchmark domain adaptation models.

Index Terms—Copula, domain adaptation, domain divergence, regularization.

Manuscript received 22 May 2022; revised 12 October 2022, 5 February 2023, and 17 March 2023; accepted 17 May 2023. This work was supported in part by the InnoHK initiative, in part by the Government of the Hong Kong Special Administrative Region, and in part by Laboratory for AI-Powered Financial Technologies Ltd. The work of Shumin Ma was supported in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, in part by the BNU-HKBU United International College under Grant 2022B1212010006, and in part by the Guangdong Higher Education Upgrading Plan (2021–2025) under Grant UIC R0400001-22 and Grant UIC R0700019-22. The work of Qi Wu was supported in part by the Hong Kong Research Grants Council through the General Research Fund under Grant 14206117, Grant 11219420, and Grant 11200219; in part by the CityU SRG-Fd Fund under Grant 7005300; and in part by the CityU-JD Digits Laboratory in Financial Technology and Engineering, HK Institute of Data Science. (Corresponding author: Qi Wu.)

Shumin Ma is with the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science and the Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519087, China.

Zhiri Yuan, Yiyang Huang, Xixu Hu, and Cheuk Hang Leung are with the CityU-JD Digits Joint Laboratory in Financial Technology and Engineering and the School of Data Science, City University of Hong Kong, Hong Kong.

Qi Wu is with the School of Data Science, the CityU-JD Digits Joint Laboratory in Financial Technology and Engineering, and the Institute of Data Science, City University of Hong Kong, Hong Kong, and also with Laboratory for AI-Powered Financial Technologies Ltd., Hong Kong (e-mail: qiwu55@cityu.edu.hk).

Dongdong Wang and Zhixiang Huang are with JD Digits Technology, Beijing 100176, China.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TNNLS.2023.3279099>.

Digital Object Identifier 10.1109/TNNLS.2023.3279099

I. INTRODUCTION

UNSUPERVISED domain adaptation emerges when one estimates a prediction function in a given target domain without any labeled samples by exploiting the knowledge available from a source domain where labels are known. The critical step in the transfer is to extract feature representations that are invariant across domains. A large body of work learns the domain-invariant feature representations by minimizing various metrics on the feature distributions between domains: Proxy- \mathcal{A} distance [1], total variation distance [2], [3], [4], maximum mean discrepancy (MMD) [5], [6], [7], Wasserstein distance [8], [9], and so on. It is worth noting that in most of the literature, the domain invariance is measured on the overall feature distributions between the source and the target domains.

However, the overall feature distribution difference encodes both the marginals' distinctions and the dependence difference into one metric value, making it hard to identify whether the domain difference comes from the marginals or the dependence structure. As a motivation example, we use Fig. 1 to clarify this point. In Fig. 1, there are three random vectors (\mathbf{X} , \mathbf{Y} , and \mathbf{Z}) that follow three different Gaussian distributions ($P^{\mathbf{X}}$, $P^{\mathbf{Y}}$, and $P^{\mathbf{Z}}$, with the details shown in the caption). It can be observed that $P^{\mathbf{X}}$ and $P^{\mathbf{Y}}$ only differ in the 2nd marginal distribution where $P_2^{\mathbf{X}}$ is $N(1, 1)$ and $P_2^{\mathbf{Y}}$ is $N(0, 1)$. Namely, the overall distribution difference is completely caused by the marginal distinction. However, for $P^{\mathbf{Y}}$ and $P^{\mathbf{Z}}$, it is well observed that the marginals are the same, while the covariance matrix differs. Namely, the overall distribution difference over the latter two distributions is controlled by the dependence difference. In general cases, any two distributions can differ in the marginal distributions and the dependence structure simultaneously. One can check that the Kullback–Leibler (KL) divergence between $P^{\mathbf{X}}$ and $P^{\mathbf{Y}}$ is the same as that of $P^{\mathbf{Z}}$ and $P^{\mathbf{Y}}$, which is 1/2 in the example, that is to say, a single divergence value cannot distinguish between marginal difference and dependence difference. Furthermore, the overall feature distributions' divergence sums the marginals' and dependence differences up together in a relatively fixed manner. Such rigidity is undesirable when the prediction function has different sensitivities to the changes in the marginals versus changes in the dependence structure. Especially in many financial and business applications, the prediction function may depend heavily on the dependence

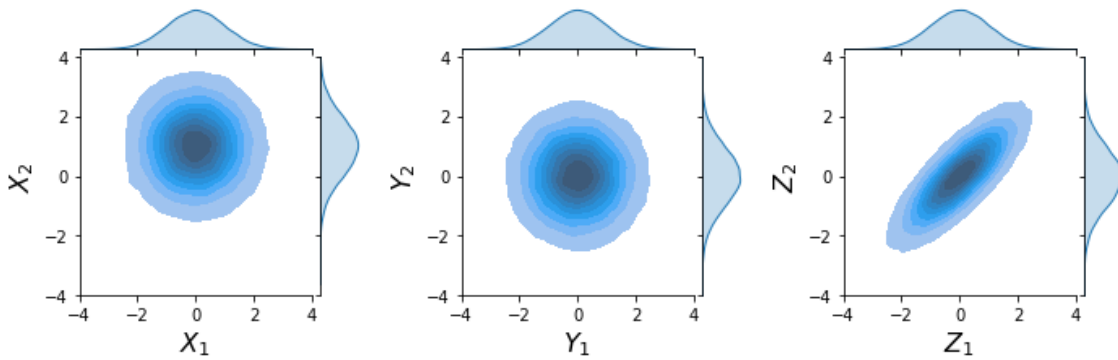


Fig. 1. Visualization of three 2-D Gaussian distributions: $(P^X) N\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $(P^Y) N\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $(P^Z) N\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{1-e^{-1}} \\ \sqrt{1-e^{-1}} & 1 \end{bmatrix}\right)$. P^X and P^Y differ only in the 2nd marginal distribution, while P^Z and P^Y differ in the dependence structure. One can check that the KL divergence between P^X and P^Y is $1/2$ and it is the same as the KL divergence between P^Z and P^Y , that is, one single divergence value ($1/2$ in this example) cannot distinguish whether the distribution difference comes from the marginals or the dependence structure.

structure of the features. For example, during the market crash, the stocks show remarkably synchronous co-movement, implying a stronger dependence than at regular times. Such observations motivate us to relax the binding between the marginals' and dependence differences in one metric value and pay attention to the difference term that matters the most in the transfer.

We propose a new domain adaptation model to regularize the domain differences in the dependence structure via the copula distance (CD), separately from the marginal divergence (MD). The idea is inspired by Sklar's theorem [10], which states that any multivariate distribution can be decomposed as the product of marginal distributions and a copula function and vice versa. It explicitly shows that the copula function, together with the marginal distributions, is sufficient to recover the original multivariate distribution. The efficacy and versatility of our approach are demonstrated with real-world classification and regression problems.

The contributions of this article are summarized as follows: 1) we propose a novel deep domain adaptation framework that allows more flexibility to combine the marginal difference and the dependence difference into a regularizer; 2) we explore the structural properties of the CD that guarantee the algorithm convergence of our approach; and 3) our proposed model proves its efficiency on two novel datasets (a large-scale retail credit dataset and an intraday equity price dataset) and one standard UCI dataset.

II. RELATED WORK

Due to the ability of deep neural nets to learn rich feature representations, deep domain adaptation models have focused on using these networks to learn invariant representations, i.e., intermediate features whose distributions are the same in the source and the target domains, while at the same time achieving a small prediction error on the source domain. The hope is that the learned representation, together with the hypothesis learned from the source domain, can generalize to the target domain. There are many ways to measure the domain invariance [11]. Duan et al. [12] used the L_2 norm to directly align features of different domains. Long et al. [6] proposed a deep adaptation network (DAN) architecture that embeds the feature representations in a reproducing kernel Hilbert space (RKHS) and reduces the domain discrepancy through

multikernel MMD. Ganin et al. [1] proposed a domain adversarial neural network (DANN) to learn the domain-invariant features with a min-max formulation. Zhao et al. [13] characterized a fundamental tradeoff between learning invariant representations and achieving a small joint error on both domains when the marginal label distributions differ from the source to the target. Furthermore, Chen et al. [14] tackled the knowledge transfer problem under the generalized covariate shift condition by the Bregman divergence. Wang et al. [15] proposed a novel neural embedding matching method by enforcing consistent classwise cross-domain instance distributions in the embedding space. Zhang et al. [16] proposed a two-stage progressive training strategy to learn invariant, discriminative, and domain-agnostic subspace. Another line of work proposes to reduce the domain discrepancy through minimizing the optimal transport loss between the source and target distributions [8], [9]. For example, Shen et al. [9] minimized the empirical Wasserstein distance between the source and target samples. However, these methods focus on the overall distribution discrepancy and often do not differentiate whether the domain differences come from the marginals or the dependence structure.

To explicitly encode the dependence difference in the domain adaptation framework, Sun et al. [17] proposed a correlation alignment (CORAL) model to measure the dissimilarity by the Frobenius norm of the covariance matrices from the two domains. Sun and Saenko [18] further combined CORAL with deep neural networks and verified its effectiveness through extensive experiments on the standard benchmark datasets. Zhang et al. [19] matched distributions by aligning the RKHS covariance matrices across domains. Chen et al. [20] integrated the MMD and CORAL into a unified framework and exploited the higher order statistics for domain alignment. Our approach is more general in that it separates the marginals' divergence and the dependence difference and integrates them into one regularizer. It facilitates us to detect the changes in the marginals and the dependence structure simultaneously.

Our work is closely related to copulas. Copulas have been successfully used in many deep learning methods. Zhao and Udell [21] performed the missing value imputation by developing a semiparametric algorithm to estimate copula parameters from incomplete mixed data. Marti et al. [22]

and Tran et al. [23] proposed to summarize and measure the pairwise correlations between variables, which is shown to well capture the various dependence patterns. By assuming the underlying features that have a specific structure, Lopez-Paz et al. [24] adopted nonparametric vine copula for semisupervised domain adaptation problems. Letizia and Tonello [25] used copula to generate dependent data in a segmented way. Xuan et al. [26] incorporated copula into a doubly nonparametric sparse nonnegative matrix factorization framework. Quan et al. [27] well documented the literature that takes advantage of copulas to model the correlation in multivariate data in a smart grid. We extend the idea of copulas and construct a copula-based divergence measure to quantify the dependence difference. Our proposed measure incorporates more statistical information than the commonly used covariance matrices that only capture the second-order statistics.

III. COPULA DISTANCE

Suppose that a d -dimensional random variable $\mathbf{X} = [X_1, \dots, X_d]$ is characterized by the cumulative distribution function (cdf) P and its density function is denoted by p . Sklar's theorem [10] states that there exists a copula C such that $P(x_1, \dots, x_d) = C(P_1(x_1), \dots, P_d(x_d))$, where $P_i(\cdot)$ is the marginal cdf. Furthermore, any continuous density function $p(x_1, \dots, x_d)$ can be written in terms of univariate marginal density functions $\{p_i(x_i)\}_{i=1}^d$ and a unique copula density function $c : [0, 1]^d \rightarrow \mathbb{R}$, which characterizes the dependence structure

$$p(x_1, \dots, x_d) = c(u_1, \dots, u_d) \times \prod_{i=1}^d p_i(x_i) \quad (1)$$

where $u_i := P_i(x_i)$, $\forall 1 \leq i \leq d$.

One can use the copula function to extract a clean quantification of the dependence strength between any components $[X_i, X_j]$ in the random vector \mathbf{X} . For example, the mutual information between X_i and X_j , a well-known dependence measure in information theory, is equivalent to the negative entropy of the copula function, namely

$$\mathcal{H}_{\text{KL}}(P_{ij}, P_i P_j) = \int_0^1 \int_0^1 c_{ij}(u_i, u_j) \log c_{ij}(u_i, u_j) du_i du_j.$$

Here, \mathcal{H}_{KL} denotes the KL divergence, P_{ij} is the joint distribution of $[X_i, X_j]$, and P_i and P_j represent the marginal distributions. $c_{ij}(u_i, u_j)$ is short for the density function $c(1, \dots, 1, u_i, 1, \dots, 1, u_j, 1, \dots, 1)$, where the i th and the j th arguments are u_i and u_j , respectively, and the other arguments are all 1. It should be noted that mutual information is a special case of a more general dependence measurement framework [22] that computes the distance $\mathcal{H}(P_{ij}, P_i P_j)$ with any divergence measure \mathcal{H} . We list a few examples in the following, where \mathcal{H} takes χ^2 distance [28], Hellinger distance [28], and α -divergence [29] (the detailed derivation can be found in the Appendix).

- 1) $\mathcal{H}_{\chi^2}(P_{ij}, P_i P_j) = \int_0^1 \int_0^1 (c_{ij}^2(u_i, u_j) - 1) du_i du_j.$
- 2) $\mathcal{H}_H(P_{ij}, P_i P_j) = \int_0^1 \int_0^1 [\sqrt{c_{ij}(u_i, u_j)} - 1]^2 du_i du_j.$

$$3) \mathcal{H}_\alpha(P_{ij}, P_i P_j) = (1/1 - \alpha^2) \int_0^1 \int_0^1 [1 - c_{ij}(u_i, u_j)^{-(\alpha+1)/2}] c_{ij}(u_i, u_j) du_i du_j.$$

For most of the divergence measures \mathcal{H} , it can be proved that $\mathcal{H}(P_{ij}, P_i P_j)$ is a function of the copula densities (see the Appendix). For a given random vector \mathbf{X} , if $\mathcal{H}(P_{ij}, P_i P_j)$ provides a clean measure of any of its component pairs $[X_i, X_j]$, one can run through the indexes, give them different weights β_{ij} , and sum over all the weighted \mathcal{H} s. The resulting quantity would allow one to look into the complete internal structure of \mathbf{X} along the direction of any user-defined attention angle using any measure \mathcal{H} . This observation motivates us to use such a measure to quantify the difference in internal dependencies between two random vectors \mathbf{X} and \mathbf{Y} . Its precise definition is given as follows.

Definition 1 (CD): Let $\mathbf{X} = [X_1, \dots, X_d] \in \mathbb{R}^d$ and $\mathbf{Y} = [Y_1, \dots, Y_d] \in \mathbb{R}^d$ be two random vectors. Let $P_{ij}^{\mathbf{X}}$ and $P_{ij}^{\mathbf{Y}}$ be the cumulative joint distributions of any of their component pairs $[X_i, X_j]$ and $[Y_i, Y_j]$, respectively. Let $\mathcal{H}(\cdot, \cdot)$ be a measure of distribution difference. We define the CD between $[X_i, X_j]$ and $[Y_i, Y_j]$ ($\forall 1 \leq i < j \leq d$) as

$$\text{CD}_{\mathcal{H}}([X_i, X_j], [Y_i, Y_j]) = |\mathcal{H}(P_{ij}^{\mathbf{X}}, P_i^{\mathbf{X}} P_j^{\mathbf{X}}) - \mathcal{H}(P_{ij}^{\mathbf{Y}}, P_i^{\mathbf{Y}} P_j^{\mathbf{Y}})|.$$

The CD between random vectors \mathbf{X} and \mathbf{Y} with respect to the positive weights $\boldsymbol{\beta} = \{\beta_{ij}\}_{1 \leq i < j \leq d}$ is defined as

$$\text{CD}_{\mathcal{H}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\beta}) = \sum_{1 \leq i < j \leq d} \beta_{ij} \text{CD}_{\mathcal{H}}([X_i, X_j], [Y_i, Y_j]). \quad (2)$$

As explained before, although the CD seems irrelevant to the copulas at first glance, it is actually a function of the copula densities. In the following context, we will use a set of d terms $\{\mathcal{H}(P_i^{\mathbf{X}}, P_i^{\mathbf{Y}})\}_{i=1}^d$ to measure the marginals' distinctions and the CD $\text{CD}_{\mathcal{H}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\beta})$ to measure the dependence difference of any two random vectors $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^d$. To explain the concepts in more detail, we use the distributions in Fig. 1. We record the marginals' distinctions and dependence difference between each two of the three distributions under KL divergence, Jensen-Shannon (JS) divergence, and MMD in Table I.

When the copula density takes some specific function form, it can be proved that the CD has two nice properties: bounded and monotonic. These two properties guarantee the convergence of our proposed algorithms in Section IV. There are multiple parametric copula density functions (Gaussian copula, t copula, Gumbel copula, and so on) that can be effectively incorporated into our model framework. In this work, we take the copula density in the form of Gaussian copula. Gaussian copula has been widely used in the financial literature because it is convenient to capture the dependence embedded in the random vector [30], [31], [32]. A random vector $\mathbf{X} \in \mathbb{R}^d$ is said to have Gaussian copula with parameter $\Sigma \in \mathbb{R}^{d \times d}$ if the copula density function $c(u_1, \dots, u_d) = |\Sigma|^{-(1/2)} \exp(-(1/2)\mathbf{x}^T(\Sigma^{-1} - \mathbf{I})\mathbf{x})$, where $\mathbf{x} := [\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)]^T$ with Φ being the cumulative density function for the standard normal distribution. We state the boundedness and monotonicity under the Gaussian copula here and defer the proof to the Appendix.

Proposition 1 (Boundedness): The CD defined in (2) is bounded when the divergence metric \mathcal{H} is taken to be MMD

TABLE I

FOR ANY TWO ADJACENT DISTRIBUTIONS IN FIG. 1, THE OVERALL DISTRIBUTION DIVERGENCE AND THE 1ST-D AND THE 2ND-D MARGINAL DISTINCTION TOGETHER WITH THE CD ARE RECORDED. SPECIFICALLY, WE TAKE THE KL DIVERGENCE, JS DIVERGENCE, AND MMD AS THE DISTRIBUTION DIVERGENCE MEASURE FOR ILLUSTRATION PURPOSES

\mathcal{H}	$P^{\mathbf{X}}$ vs. $P^{\mathbf{Y}}$				$P^{\mathbf{Z}}$ vs. $P^{\mathbf{Y}}$			
	$\mathcal{H}(P^{\mathbf{X}}, P^{\mathbf{Y}})$	$\mathcal{H}(P_1^{\mathbf{X}}, P_1^{\mathbf{Y}})$	$\mathcal{H}(P_2^{\mathbf{X}}, P_2^{\mathbf{Y}})$	CD	$\mathcal{H}(P^{\mathbf{Z}}, P^{\mathbf{Y}})$	$\mathcal{H}(P_1^{\mathbf{Z}}, P_1^{\mathbf{Y}})$	$\mathcal{H}(P_2^{\mathbf{Z}}, P_2^{\mathbf{Y}})$	CD
KL	0.5	0	0.5	0	0.5	0	0	0.5
JS	0.5	0	0.5	0	0.859	0	0	0.859
MMD	0.107	0	0.175	0	0.042	0	0	0.042

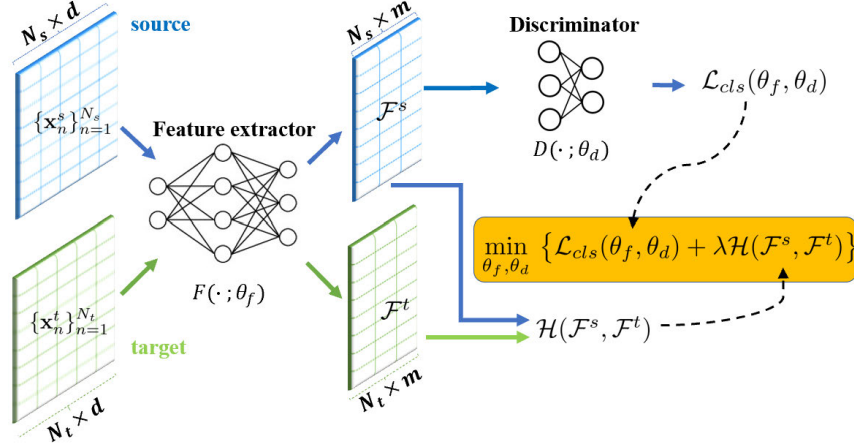


Fig. 2. Network flow in deep domain adaptation models. All samples, no matter from the source domain $\{\mathbf{x}_n^s\}_{n=1}^{N_s}$ (blue) or the target domain $\{\mathbf{x}_n^t\}_{n=1}^{N_t}$ (green), are fed into a feature extractor $F(\cdot; \theta_f)$ to extract features that are both discriminative and domain-invariant. The discriminative features are achieved by training a discriminator $D(\cdot; \theta_d)$ to minimize the loss function $\mathcal{L}_{cls}(\theta_f, \theta_d)$, while the domain invariance is measured by $\mathcal{H}(\mathcal{F}^s, \mathcal{F}^t)$.

distance, Wasserstein-2 distance, or some of the ϕ -divergence (including JS distance, Hellinger distance, and total variation distance).

Proposition 2 (Monotonicity): Let $\Sigma^{\mathbf{X}}$ and $\Sigma^{\mathbf{Y}}$ be the Gaussian copula parameters for the random vectors $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^d$, respectively. Given all of the other entries in $\Sigma^{\mathbf{X}}$ and $\Sigma^{\mathbf{Y}}$ fixed, the CD $\text{CD}_{\mathcal{H}}([X_i, X_j], [Y_i, Y_j])$ is monotonically increasing with $|(\Sigma_{ij}^{\mathbf{X}})^2 - (\Sigma_{ij}^{\mathbf{Y}})^2|$, $\forall 1 \leq i < j \leq d$. The monotonicity is satisfied by general probability distribution divergence measures, including MMD distance, Wasserstein-2 distance, and most of the commonly used ϕ -divergence (including KL divergence, χ^2 distance, Hellinger distance, and so on).

IV. UNSUPERVISED DOMAIN ADAPTATION

Notations: In unsupervised domain adaptation, we are given a *source* domain $\mathcal{D}_s = \{(\mathbf{x}_n^s, y_n^s)\}_{n=1}^{N_s}$ with N_s labeled examples and a *target* domain $\mathcal{D}_t = \{\mathbf{x}_n^t\}_{n=1}^{N_t}$ with N_t unlabeled examples. It is assumed that the two domains are characterized by different probability distributions, while they share the same feature space. In a classification task, the goal is to learn a transferable classifier to minimize the classification error on the target domain using all the given data.

Deep domain adaptation methods begin with a feature extractor that can be implemented by a neural network. The feature extractor is supposed to learn the domain-invariant feature representations from both domains. Specifically, the feature extractor learns a function $F(\mathbf{x}; \theta_f) : \mathbb{R}^d \rightarrow \mathbb{R}^m$ that maps an instance to an m -dimensional representation with the network parameters θ_f (as shown in Fig. 2).

For simplicity, we denote the feature representation of a source instance \mathbf{x}_n^s as $\mathbf{F}_n^s := F(\mathbf{x}_n^s; \theta_f)$ and a target instance \mathbf{x}_n^t as $\mathbf{F}_n^t := F(\mathbf{x}_n^t; \theta_f)$. We define the source feature set $\mathcal{F}^s := \{\mathbf{F}_n^s\}_{n=1}^{N_s}$ and the target feature set $\mathcal{F}^t := \{\mathbf{F}_n^t\}_{n=1}^{N_t}$.

When domain adaptation is applied to a classification problem, the feature extractor is followed by a discriminator trained with samples from the source domain \mathcal{D}_s . Given the feature representations \mathcal{F}^s computed by the feature extractor on the source domain, together with the labels $\{y_n^s\}_{n=1}^{N_s}$ ($y_n^s \in \{1, 2, \dots, l\}$), we can train a discriminator $D(\cdot; \theta_d) : \mathbb{R}^m \rightarrow \mathbb{R}^l$ that is characterized by the parameters θ_d . The discriminator loss function is defined as the cross entropy between the predicted probabilistic distribution and the one-hot encoding of the class labels

$$\mathcal{L}_{cls}(\theta_f, \theta_d) := -\frac{1}{N_s} \sum_{n=1}^{N_s} \sum_{i=1}^l \mathbf{1}(y_n^s = i) \cdot \log D(\mathbf{F}_n^s; \theta_d)_i \quad (3)$$

where $\mathbf{1}(\cdot)$ is the indicator function and $D(\cdot; \theta_d)_i$ represents the i th element in the predicted distribution $D(\cdot; \theta_d)$.

Besides the discriminator that is trained to learn the discriminative features, the feature extractor is also followed by a discrepancy term that measures the difference between the source features \mathcal{F}^s and the target features \mathcal{F}^t to learn domain-invariant feature representations. To be specific, for a given discrepancy measure \mathcal{H} , the empirical discrepancy between the source feature distribution and the target feature distribution is given by $\mathcal{H}(\mathcal{F}^s, \mathcal{F}^t)$. In the literature, there have been multiple choices of the discrepancy measures \mathcal{H} , such as MMD distance [6], Wasserstein distance [9], JS distance [33],

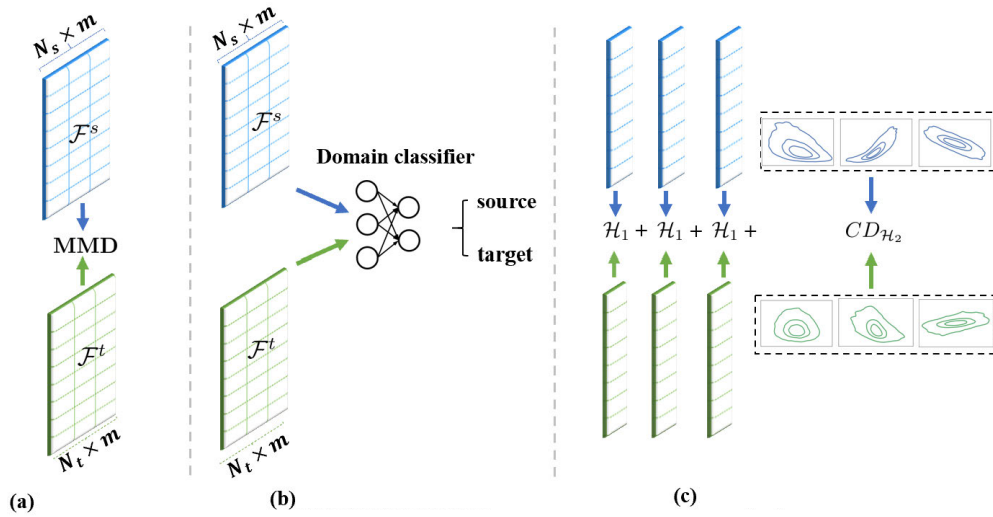


Fig. 3. There can be various choices of \mathcal{H} to measure the feature difference in (4), such as (a) MMD distance and (b) Wasserstein distance (adversarial distance), JS distance, Proxy- \mathcal{A} distance, and so on. (c) Our proposed model looks into the detailed feature structures and regularizes the source error by dynamically adjusting the MD and the CD.

and Proxy- \mathcal{A} distance [34]. We call the latter few measures as adversarial distance because they are implemented with a domain classifier to quantify the invariance between \mathcal{F}^s and \mathcal{F}^t . We illustrate the commonly chosen discrepancy measures in domain adaptation models in Fig. 3(a) and (b).

In summary, the detailed objective function to train a deep domain adaptation network is

$$\min_{\theta_f, \theta_d} \left\{ \mathcal{L}_{\text{cls}}(\theta_f, \theta_d) + \lambda \mathcal{H}(\mathcal{F}^s, \mathcal{F}^t) \right\} \quad (4)$$

where the coefficient λ controls the tradeoff between discriminative and transferrable feature learning.

A. Copula-Based Domain Adaptation Networks

In this work, we propose a copula-based domain adaptation network (CDAN) to allow more flexibility and emphasis on the dependence structure. To be specific, on the source domain, we split the distribution of the source features $\mathcal{F}^s \subseteq \mathbb{R}^{N_s \times m}$ into m marginal distributions and the copula between the marginals. We do the same on the target domain. This facilitates us to evaluate the feature differences between the source domain and the target domain by the sum of two terms: 1) the sum of marginal feature differences $\{\mathcal{H}(\mathcal{F}_i^s, \mathcal{F}_i^t)\}_{i=1}^m$ and 2) the CD between the source features and the target features $CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta)$. Here, for $* \in \{s, t\}$, $\mathcal{F}_i^* := \{\mathbf{F}_{n,i}^*\}_{n=1}^{N_*} \subseteq \mathbb{R}^{N_* \times 1}$ with $\mathbf{F}_{n,i}^*$ being the i th dimensional value of \mathbf{F}_n^* . In the later context, we call the MD and call the CD. By adding the MD (with hyperparameters $\{\alpha_i\}_{i=1}^m$) and the CD (with hyperparameters β) as regularization terms, we arrive at the objective function to train a CDAN, namely

$$\min_{\theta_f, \theta_d} \left\{ \mathcal{L}_{\text{cls}}(\theta_f, \theta_d) + \sum_{i=1}^m \alpha_i \mathcal{H}_1(\mathcal{F}_i^s, \mathcal{F}_i^t) + CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta) \right\}. \quad (5)$$

The detailed divergence framework is shown in Fig. 3(c). Notice that in (5), we differentiate the divergence metric \mathcal{H}_1 that is used to calculate the MD with \mathcal{H}_2 that is for the CD calculation. It is worth noting that our model can well

accommodate the commonly used divergence measures in the literature.

Our proposed model has three advantages. On one hand, rather than encoding the divergence of each marginal feature and the dependence difference in a single value as in (4), we split the divergence of the joint feature distributions. Thus, we can identify to what extent the marginal feature differences and the CD contribute to the target risk respectively. Moreover, using hyperparameters $\{\alpha_i\}_{i=1}^m$ and β to separately control the MD and CD allows us to dynamically adjust the hyperparameters in a data-driven manner. It implicitly shows that there can be a tradeoff between the MD and the CD. Finally, different from using one distance metric to measure both the marginal feature differences and the dependence difference, our proposed model provides a more convenient and elaborate way to detect the changes in the marginal distributions and the dependence structure.

It is straightforward to generalize our CDAN model to the regression tasks. Similar to the classification setting, a domain adaptation model for a regression task combines the feature extractor $F(\cdot; \theta_f)$ together with a regressor network $D(\cdot; \theta_d)$ to form the basis of a supervised learning network. We denote the predicted value for a sample \mathbf{x}_n^s as $\hat{y}_n^s := D(F(\mathbf{x}_n^s; \theta_f); \theta_d)$. The regressor loss function is defined as the mean squared error between the predicted values $\{\hat{y}_n^s\}_{n=1}^{N_s}$ and the ground-truth values $\{y_n^s\}_{n=1}^{N_s}$ on the source domain: $\mathcal{L}_{\text{rgr}}(\theta_f, \theta_d) := \sum_{n=1}^{N_s} (\hat{y}_n^s - y_n^s)^2 / N_s = \sum_{n=1}^{N_s} (D(F(\mathbf{x}_n^s; \theta_f); \theta_d) - y_n^s)^2 / N_s$. Adding the MD and the CD as regularizer, we obtain the objective function to train a CDAN for a regression task

$$\min_{\theta_f, \theta_d} \left\{ \mathcal{L}_{\text{rgr}}(\theta_f, \theta_d) + \sum_{i=1}^m \alpha_i \mathcal{H}_1(\mathcal{F}_i^s, \mathcal{F}_i^t) + CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta) \right\}.$$

B. Algorithm

The complete process of CDAN algorithm is presented in Algorithm 1. In particular, we provide a detailed description

Algorithm 1 CDAN Algorithm

Input : Source data $\mathcal{D}_s = \{(\mathbf{x}_n^s, y_n^s)\}_{n=1}^{N_s}$, Target data $\mathcal{D}_t = \{(\mathbf{x}_n^t, y_n^t)\}_{n=1}^{N_t}$, Maximum training epoch S , Divergence metric $\mathcal{H}_1, \mathcal{H}_2$, Parameters $\{\alpha_i\}_{i=1}^m, \beta$.

Output: Optimal model parameters θ_f, θ_d .

```

1 for epoch = 1 to S do
2    $\mathcal{F}^s \leftarrow \{F(\mathbf{x}_n^s; \theta_f)\}_{n=1}^{N_s}$ ;
3    $\mathcal{F}^t \leftarrow \{F(\mathbf{x}_n^t; \theta_f)\}_{n=1}^{N_t}$ ;
4    $MD(\theta_f) \leftarrow \sum_{i=1}^m \alpha_i \mathcal{H}_1(\mathcal{F}_i^s, \mathcal{F}_i^t)$ ;
5    $CD(\theta_f) \leftarrow CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta)$  as calculated in
      Eq. (2);
6   Get the source error  $\mathcal{L}_{cls}(\theta_f, \theta_d)$  with Eq. (3);
7    $Loss \leftarrow \mathcal{L}_{cls}(\theta_f, \theta_d) + MD(\theta_f) + CD(\theta_f)$ ;
8    $Loss.backward()$ 
```

on how to learn the Gaussian copula parameter Σ and how to update the model parameters θ_f and θ_d .

1) *Learning the Gaussian Copula Parameter Σ :* Ruppert and Matteson [35] proposed a moment-matching approach to learn the Gaussian copula parameter Σ through Kendall's tau. Denote $\rho_\tau(X_i, X_j)$ as Kendall's tau between random variables X_i and X_j , that is, $\rho_\tau(X_i, X_j) := \mathbb{E}[\text{sign}((X_i - \tilde{X}_i)(X_j - \tilde{X}_j))]$, where $[\tilde{X}_i, \tilde{X}_j]$ is an independent copy of $[X_i, X_j]$. Then, it can be proved that $\Sigma_{ij} = \sin(\pi/2)\rho_\tau(X_i, X_j)$. However, computing Kendall's tau by definition incurs a complexity of $O(N^2)$ when the sample size is N , making it expensive to train deep neural networks. Moreover, gradient vanishing occurs during training of the neural network because of the sign function.

To address the gradient vanishing issue, we propose to replace the sign function with the tanh function with parameter a , namely, $\rho_\tau(X_i, X_j) \approx \rho(X_i, X_j; a) := \mathbb{E}[\tanh(a(X_i - \tilde{X}_i)(X_j - \tilde{X}_j))]$. We prove that $\lim_{a \rightarrow \infty} \rho(X_i, X_j; a) = \rho_\tau(X_i, X_j)$ (with the proof attached in the Appendix.)

Proposition 3: For two random variables X_1 and X_2 , it holds that $\lim_{a \rightarrow \infty} \rho(X_1, X_2; a) = \rho_\tau(X_1, X_2)$.

To further reduce the computational complexity, we adopt the unbiased estimate of Kendall's tau that can be computed with linear complexity. More specifically, given $\{[x_{n,1}, x_{n,2}]\}_{n=1}^N \subseteq \mathbb{R}^{N \times 2}$ as N realizations of $[X_i, X_j]$, an unbiased estimator for $\rho_\tau(X_i, X_j)$ is $\rho_\tau(X_i, X_j) = (2/N) \sum_{n=1}^{N/2} \tanh(a(x_{2n-1,1} - x_{2n,1})(x_{2n-1,2} - x_{2n,2}))$. It significantly reduces the computational cost of Kendall's tau from $O(N^2)$ to $O(N)$.

2) *Learning the Model Parameters θ_f and θ_d :* For illustration convenience, we conduct the calculation in the classification setting. It can be similarly done for the regression problem, and thus, we omit here. From (5), we have

$$\begin{aligned} \nabla_{\theta_d} &= \partial_{\theta_d} \mathcal{L}_{cls}(\theta_f, \theta_d) \\ &= - \sum_{n=1}^{N_s} \sum_{i=1}^l \mathbf{1}(y_n^s = i) \cdot \partial_{\theta_d} D(\mathbf{F}_n^s; \theta_d)_i / (N_s D(\mathbf{F}_n^s; \theta_d)_i) \end{aligned}$$

$$\nabla_{\theta_f} = \sum_{i=1}^m \alpha_i \partial_{\theta_f} \mathcal{H}_1(\mathcal{F}_i^s, \mathcal{F}_i^t) + \partial_{\theta_f} CD_{\mathcal{H}_2} + \partial_{\theta_f} \mathcal{L}_{cls}(\theta_f, \theta_d)$$

where $CD_{\mathcal{H}_2}$ is the abbreviation for $CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta)$.

By the chain rule

$$\begin{aligned} \partial_{\theta_f} \mathcal{L}_{cls}(\theta_f, \theta_d) \\ = - \sum_{n=1}^{N_s} \sum_{i=1}^l \frac{\mathbf{1}(y_n^s = i) \cdot \partial_{\theta_f} D(\mathbf{F}_n^s; \theta_d)_i \cdot \partial_{\theta_f} F(\mathbf{x}_n^s; \theta_f)}{N_s D(\mathbf{F}_n^s; \theta_d)_i}. \end{aligned}$$

The derivatives $\partial_{\theta_f} \mathcal{H}_1(\mathcal{F}_i^s, \mathcal{F}_i^t)$ and $\partial_{\theta_f} CD_{\mathcal{H}_2}(\mathcal{F}^s, \mathcal{F}^t; \beta)$ depend on the choice of \mathcal{H}_1 and \mathcal{H}_2 . In the experiment, we will take \mathcal{H}_1 as the MMD distance and \mathcal{H}_2 as the KL divergence. Specifically, if \mathcal{H}_1 is MMD distance with the characteristic kernel function k , the unbiased estimate of squared MMD distance between \mathcal{F}_i^s and \mathcal{F}_i^t is given as [6]

$$\begin{aligned} \mathcal{H}_{\text{MMD}}^2(\mathcal{F}_i^s, \mathcal{F}_i^t) := & \sum_{n,n'=1}^{N_s} \frac{k(\mathbf{F}_{n,i}^s, \mathbf{F}_{n',i}^s)}{N_s^2} + \sum_{n,n'=1}^{N_t} \frac{k(\mathbf{F}_{n,i}^t, \mathbf{F}_{n',i}^t)}{N_t^2} \\ & - \sum_{n=1}^{N_s} \sum_{n'=1}^{N_t} \frac{2k(\mathbf{F}_{n,i}^s, \mathbf{F}_{n',i}^t)}{N_s N_t}. \end{aligned}$$

Thus,

$$\begin{aligned} \partial_{\theta_f} \mathcal{H}_{\text{MMD}}(\mathcal{F}_i^s, \mathcal{F}_i^t) \\ = \left[\sum_{n,n'=1}^{N_s} \frac{\partial_{\theta_f} k(\mathbf{F}_{n,i}^s, \mathbf{F}_{n',i}^s)}{N_s^2} + \sum_{n,n'=1}^{N_t} \frac{\partial_{\theta_f} k(\mathbf{F}_{n,i}^t, \mathbf{F}_{n',i}^t)}{N_t^2} \right. \\ \left. - \sum_{n=1}^{N_s} \sum_{n'=1}^{N_t} \frac{2\partial_{\theta_f} k(\mathbf{F}_{n,i}^s, \mathbf{F}_{n',i}^t)}{N_s N_t} \right] / (2\mathcal{H}_{\text{MMD}}(\mathcal{F}^s, \mathcal{F}^t)). \end{aligned}$$

When \mathcal{H}_2 is the KL divergence, then

$$\partial_{\theta_f} CD_{\mathcal{H}_{\text{KL}}} = \sum_{i < j} \frac{\beta_{ij} \partial_{\theta_f} |\log(1 - (\Sigma_{ij}^s)^2) / (1 - (\Sigma_{ij}^t)^2)|}{2}$$

where Σ_{ij}^s (Σ_{ij}^t) is the Gaussian copula parameter of $\{[\mathbf{F}_{n,i}^s, \mathbf{F}_{n,j}^s]\}_{n=1}^{N_s}$ ($\{[\mathbf{F}_{n,i}^t, \mathbf{F}_{n,j}^t]\}_{n=1}^{N_t}$).

V. EXPERIMENTS

To prove the efficacy of our proposed model, we test it on one toy dataset and three real-world datasets, with the details of the three real-world datasets summarized in Table II. All experiments in this section are run on Dell 7920 with Intel¹ Xeon¹ Gold 6250 CPU at 3.90 GHz and a set of NVIDIA Quadro RTX 6000 GPU. The code is available at https://github.com/yzR1991/Deep_into_The_Domain_Shift. We run our python code on the anaconda virtual environment with Microsoft Windows Server 2019 Standard as OS. The code can also be run in other OS, device, or environment, with Pytorch version 1.7 or above.

¹Registered trademark.

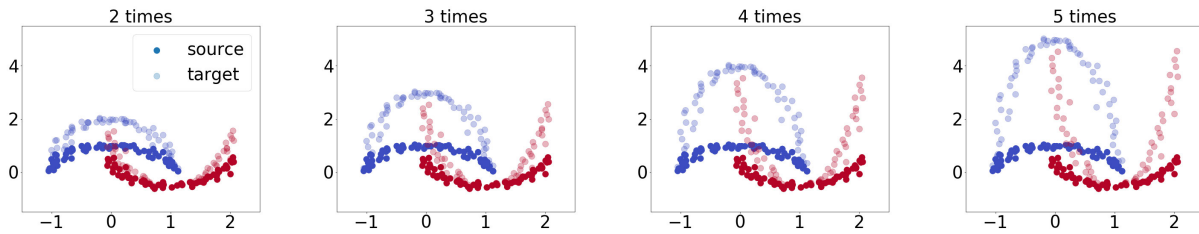


Fig. 4. Illustration of the four transfer tasks on the synthetic dataset. The two classes of the source samples are blue and red, and points that are more transparent represent the target samples.

TABLE II

INFORMATION OF THE REAL-WORLD DATASETS USED IN THIS ARTICLE

Dataset	#Instances	#Features	Task
Retail credit data	1100000	69	Classification
Equity price data	71242	22	Regression
UCI wine quality	6197	12	Regression

TABLE III

ACCURACY FOR THE INTERTWINING MOONS DATASET

	2 times	3 times	4 times	5 times
MLP	88.99	80.88	78.73	77.10
CORAL	97.45	94.40	92.32	90.26
DAN	97.04	94.17	91.18	91.04
CDAN	97.91	94.42	93.17	91.54

A. Toy Problem: Two Intertwining Moons

The source domain considered here is the classical binary problem with two intertwining moons, each class corresponding to one moon. Specifically, for the blue moon in Fig. 4, the points roughly falls on the upper half circle of $y = (1 - x^2)^{1/2}$, with points for the red moon falling on lower half circle of $y = 0.5 - (1 - (1 - x^2)^{1/2})$. We then consider four different target domains by stretching the circle into ellipses where the length of the major axis can be two, three, four, and five times that of the minor axis. Such stretching from source domain to target domain strongly affects the relationship between the vertical ordinates and the horizontal ordinates of each point, thus making changes to the internal dependence structure of the data. For each domain, we generate 1024 instances (512 of each class).

For each transfer task, we compare the CDAN model with DAN [6], CORAL [17], and the no-adaptation baseline [multi-layer perceptron (MLP)]. Each trial of the models is repeated ten times, and we report the average accuracy in Table III. We remark that the larger the length of the major axis, the more difficult the problem becomes, as all of the four models unanimously show a weaker adaptation ability. Our CDAN provides the best performance in all of the four transfer tasks, indicating that CDAN actually captures the dependence difference precisely.

B. Retail Credit Classification

In this section, we apply our methods to improve the classification accuracy of a credit risk model on a novel real-world anonymous dataset. The dataset is kindly provided by one of the largest global technology firms that operates in both the e-commerce business and the lending business. It records the monthly credit status of roughly one-half million customers, spanning from January 2016 to June 2020. The credit status

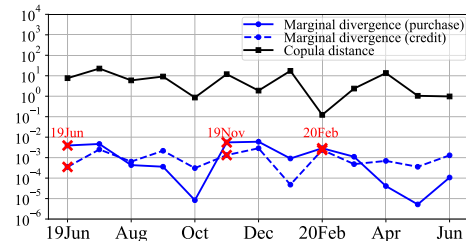


Fig. 5. Distribution shift in the raw data.

shows whether a customer is in default. In addition to the credit status, customers' monthly shopping, purchasing, and loan history are also included in the dataset in detail. We collect 69 features for each customer in each month from the raw dataset and normalize them to the range $[0, 1]$. A binary classification model is constructed to distinguish customers who default (labeled as 0) from those who have paid off all the debts on time (labeled as 1).

Domain adaptation is needed when one forecasts customers' credit risk with the classification model trained with the past data because significant distribution shifts exist between months, especially between the off-season and the peak season, between before-COVID times and after-COVID times. We record the distribution shifts between two consecutive months in Fig. 5. Specifically, we collect two features that show great importance for the classification: each customer's monthly total purchase and his monthly credit ratio (available credit amount/total credit limit). To illustrate the distribution shifts, we take the three points on 19Jun in Fig. 5 as an example. The blue (resp., blue dashed) point records the MMD distance of monthly purchase (resp., credit ratio) distribution between May and June, and the black point records the CD between May and June. From Fig. 5, we can identify three peak periods circled by a red box (19Jun, 19Nov, and 20Feb) that show substantial marginal differences, verifying the necessity of transfer learning. What is more, the CD remains high over the whole year, suggesting that special attention to the dependence difference is in urgent need.

We first evaluate our methods on the transfer between the off-seasons and the peak seasons (specifically, the sales seasons June and November) and build two transfer tasks: 19May→Jun and 19Oct→Nov. We further investigate the COVID impact on transferring the classification models and include the evaluation on six more transfer tasks: 19Dec→20Jan, 20Jan→Feb, 20Feb→Mar, 20Mar→Apr, 20Apr→May, and 20May→Jun. In each transfer task, the source domain consists of samples from the former month and the target domain samples are from the latter month. The sample size for each domain is in the magnitude of 10^5 customers.

TABLE IV
PERFORMANCE OF RETAIL CREDIT CLASSIFICATION IN EIGHT TRANSFER TASKS

	19May→Jun	19Oct→Nov	19Dec→20Jan	20Jan→Feb	20Feb→Mar	20Mar→Apr	20Apr→May	20May→Jun
MLP	73.85 ± 0.05	56.11 ± 0.05	90.66 ± 0.005	75.60 ± 0.03	76.55 ± 0.04	90.78 ± 0.009	91.61 ± 0.005	88.06 ± 0.018
DAN	79.00 ± 0.03	79.31 ± 0.05	92.09 ± 0.004	87.41 ± 0.01	91.04 ± 0.01	91.88 ± 0.008	92.78 ± 0.002	90.88 ± 0.009
CORAL	77.86 ± 0.04	76.23 ± 0.05	92.12 ± 0.003	87.13 ± 0.01	89.15 ± 0.01	91.94 ± 0.008	92.79 ± 0.003	90.94 ± 0.009
AFN	77.53 ± 0.02	80.01 ± 0.02	83.73 ± 0.02	82.72 ± 0.02	83.32 ± 0.01	84.52 ± 0.01	86.24 ± 0.01	84.71 ± 0.02
MCD	75.30 ± 0.10	71.08 ± 0.11	82.60 ± 0.08	78.40 ± 0.09	76.94 ± 0.08	86.48 ± 0.04	86.00 ± 0.04	84.85 ± 0.05
CDAN	80.44 ± 0.04	80.79 ± 0.06	91.93 ± 0.005	88.20 ± 0.02	91.51 ± 0.01	92.46 ± 0.010	92.74 ± 0.003	91.39 ± 0.010

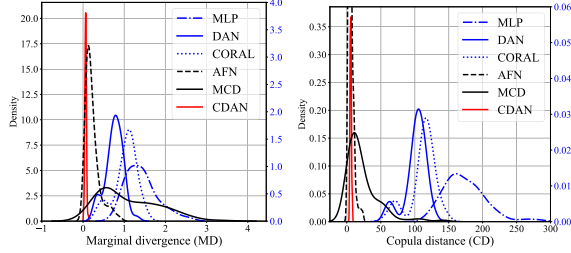


Fig. 6. Distribution of the learned features' MD (left) and CD (right) over 100 trials for each model.

We mainly follow the standard evaluation protocol [36] for unsupervised domain adaptation and use all source samples with binary labels and all target samples without labels [6]. We compare our CDAN model to four classical domain adaptation models: DAN [6], CORAL [17], AFN [37], MCD [38], as well as the no-adaptation baseline (MLP), which is a fully connected neural network with multiple hidden layers. For our model CDAN, we set \mathcal{H}_1 to be the MMD distance [6] and set \mathcal{H}_2 to be the KL divergence [39]. We set $\alpha_i = \alpha (\forall i)$ and $\beta_{ij} = \beta (\forall i, j)$ and select the hyperparameter pair (α, β) by grid search. The detailed implementation procedures are summarized in the Appendix.

In Table IV, we record the averages and standard errors of area under curve (AUC) over 100 randomized trials for each model in each task. CDAN outperforms the other models in six transfer tasks. It deserves attention that the outperformance is significant in that the increase in the AUC score is far larger than the standard deviation. As an illustration, we plot the density of the learned features' MD and CD in Fig. 6 for each model. The density is estimated from the 100 trials for each model in a specific transfer task (20Jan→Feb). We find that CDAN again shows superiority in terms of contracting both the marginal and the dependence differences. Such observation, together with the optimized hyperparameters, explains why CDAN outperforms other domain adaptation models.

C. Intraday Equity Price Regression

We collect the intraday 5-min asset prices of 22 stocks selected from HKEX according to the market cap and daily turnover. The 22 stocks cover eight industries (according to the Hang Seng Industry Classification System) that include information technology, financials, consumer discretionary, and so on. The data span from December 1, 2014 to December 31, 2020, and consist of 71 242 observations with information of the first half-hour in each trading day excluded. We divide the observations into two domains according to the Hang Seng Index (HSI) daily returns. Specifically, the target domain

TABLE V
MEAN α -QUANTILES $Q_\alpha^s (Q_\alpha^t)$ OVER ALL 22 STOCKS AND THE MEAN QUANTILE DEPENDENCE $\tau_\alpha^s (\tau_\alpha^t)$ OVER ALL STOCK PAIRS IN THE SOURCE (TARGET) DOMAIN. VALUES IN THE BRACKETS RECORD THE CORRESPONDING MINIMUM AND MAXIMUM

	$\alpha = 0.03$	$\alpha = 0.10$	$\alpha = 0.25$
Q_α^s	78.5 (37.7, 114.8)	87.4 (39.8, 147.6)	103.9 (44.3, 198.3)
Q_α^t	76.6 (36.7, 131.3)	84.2 (38.7, 149.0)	97.7 (43.5, 158.7)
τ_α^s	28.9 (4.3e-2, 88.0)	36.8 (1.6e-1, 89.6)	48.5 (8.4e-1, 98.6)
τ_α^t	42.6 (1.7e-1, 100.0)	47.0 (2.2, 93.8)	58.4 (2.4, 99.7)

includes observations of the days when the daily return is less than the 0.1-quantile of the whole six-year return series, and the source domain includes the remaining observations. The goal is to forecast the next 5-min price of the 22 stocks with their last-hour price as the input.

To illustrate the distribution shift of the two domains, we record the quantiles and the quantile dependence of each domain in Table V. We denote $X_i^s (X_i^t)$ as the price series for stock i in the source (target) domain and define its whole price series as $X_i := X_i^s \cup X_i^t$. Each X_i is normalized by its first price. For a given quantile level α and $* \in \{s, t\}$ indicating the domain, the average α -quantile of domain $*$ is defined by $Q_\alpha^* := \sum_{i=1}^{22} Q_{\alpha,i}^* / 22$, where $Q_{\alpha,i}^*$ is the α -quantile of the price series X_i^* . The quantile dependence is given by $\tau_\alpha^* := \sum_{1 \leq i < j \leq 22} \tau_{\alpha,[i,j]}^* / 231$, where $\tau_{\alpha,[i,j]}^* := 100 \times \Pr(X_j^* \leq Q_{\alpha,j} | X_i^* \leq Q_{\alpha,i})$ and $Q_{\alpha,i}$ is the α -quantile of X_i . We see that the marginal quantiles of either domain do not differ that much, but as α increases, the marginals' differences between the two domains get larger. Furthermore, the quantile dependence significantly differs irrespective of the α value.

We compare the following five models: RNN, long-short term memory (LSTM), DANN [1], CORAL [18], and DAN [6]. Specifically, RNN and LSTM serve as the no-adaptation benchmarks and only utilize the source samples to do the training. CORAL and DAN are the same as in Section V-B, except that they use an LSTM as a feature extractor in this section. Though AFN and MCD are SOTA models, they are not originally designed for sequential data and perform not well in this case, so we do not list the corresponding results. For more implementation details, see the Appendix.

In Table VI, we summarize the experimental results of the six models. There are seven performance metric columns in Table VI. The 1st column represents the mean and the standard

TABLE VI
EFFICIENCY OF EQUITY PRICE FORECAST FOR THE SIX MODELS. QUANTITIES ARE AVERAGED OVER 100 RUNS

	RMSE	RE(%)	Q1 RE(%)	Q2 RE(%)	Q3 RE(%)	LRE(%)	SRE(%)
RNN	0.0507 \pm 3.4e-3	3.33 \pm 0.22	3.23	3.35	3.47	9.15 \pm 0.90	1.02 \pm 0.12
LSTM	0.0446 \pm 1.5e-3	3.04 \pm 0.12	2.96	3.02	3.12	8.62 \pm 0.70	0.94 \pm 0.10
DANN	0.0475 \pm 9.6e-3	3.00 \pm 0.56	2.59	2.90	3.32	8.44 \pm 1.60	0.98 \pm 0.22
CORAL	0.0449 \pm 1.5e-3	3.06 \pm 0.14	2.97	3.06	3.15	8.68 \pm 0.87	0.96 \pm 0.09
DAN	0.0255 \pm 2.5e-3	1.89 \pm 0.21	1.75	1.89	2.05	6.44 \pm 1.55	0.45 \pm 0.10
CDAN	0.0235 \pm 2.3e-3	1.77 \pm 0.21	1.63	1.75	1.89	6.06 \pm 1.53	0.43 \pm 0.09

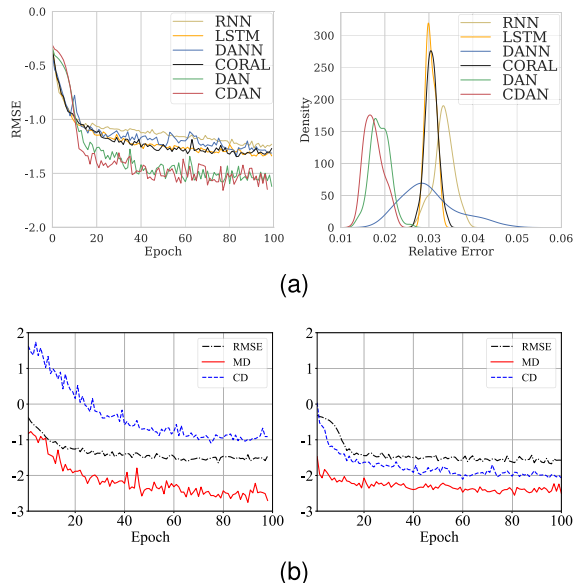


Fig. 7. Performance of forecasting the equity price in a particular (typical) trial. All values are in the base 10 logarithms. (a) Comparison of test RMSE (left) and RE distribution (right). (b) DAN model (left) versus CDAN model (right).

deviation of the RMSE over 100 trials. The 2nd–5th columns record the detailed information of the relative errors (REs) over 100 trials. Specifically, the 2nd column records the mean and the standard deviation. The 3rd–5th columns record the 0.25-quantile, 0.5-quantile, and 0.75-quantile of the RE over 100 trials. In addition to the RE over 100 trials, we also record the maximal (in the 6th column LRE) and the minimal RE (in the 7th column SRE) among the 22 stocks. We find that CDAN achieves the best performance. Moreover, its Q2 RE is close to its mean RE, and the standard deviation of RE is quite small, showing that the CDAN model is pretty stable in terms of model performance. We plot additional visualizing pictures in Fig. 7 to have a better understanding of the results. In Fig. 7(a), CDAN converges fast in terms of the test RMSE, and it is efficient in controlling the REs. Diving deeper into the training details, from Fig. 7(b), we observe that the CD of CDAN decreases more significantly (to around 10^{-2}) than that of DAN (to around 10^{-1}). It shows that CDAN does capture the dependence difference, which explains and contributes to the outperformance of CDAN.

D. Wine Quality Regression

The UCI wine quality dataset [40] contains records of red and white vinho verde wine samples from the north of Portugal, with a sample size of 1599 and 4598, respectively. Each

TABLE VII
RMSE, R2 SCORE (R2), AND RE FOR WINE QUALITY PREDICTION

	W→R			R→W		
	RMSE	R2	RE	RMSE	R2	RE
MLP	0.125	0.131	0.110	0.143	0.067	0.115
AFN	0.129	0.087	0.119	0.145	0.032	0.118
MCD	0.125	0.137	0.109	0.144	0.042	0.116
DANN	0.127	0.104	0.115	0.147	0.006	0.119
DAN	0.122	0.175	0.109	0.138	0.123	0.112
CORAL	0.125	0.136	0.109	0.144	0.054	0.115
CDAN	0.120	0.201	0.108	0.133	0.177	0.109

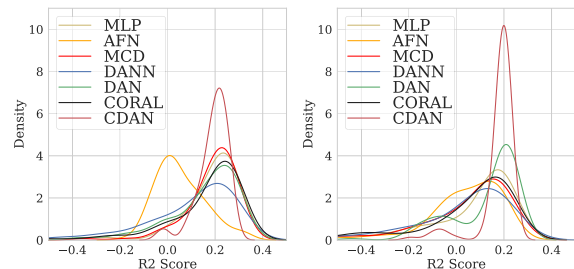


Fig. 8. R2 score distributions over 100 runs for the transfer task W→R (left) and R→W (right).

record has 12 features, such as pH, alcohol, and quality. The red wine and white wine samples differ in feature distributions. Our goal is to predict the wine quality with two transfer tasks, from white wine to red wine (W→R) and from red wine to white wine (R→W).

We compare CDAN to six neural network baselines, namely, MLP, AFN [37], MCD [38], DANN [1], CORAL [18], and DAN [6]. Each neural network model has two hidden layers and each hidden layer has eight units. For each model, we run 100 trials and record the RMSE, R2 scores, and REs.

The results are summarized in Table VII. We conclude that the CDAN model outperforms the other benchmarks by achieving the highest R2 score, the smallest RE, and the smallest RMSE. Furthermore, we plot the R2 score distribution over the 100 runs for the two transfer tasks in Fig. 8. We see that the R2 score for CDAN model is more concentrated than its competitors, showing that the outperformance is quite stable.

E. Parameter Sensitivity and Ablation Study

To further look into the sensitivity of the parameters α and β , we compare the model performance on the retail credit dataset and the equity price dataset. For the retail dataset, we list the model performance under various combinations of α and β in the eight transfer tasks in Table VIII. Among the

TABLE VIII
PARAMETER SENSITIVITY FOR THE CDAN MODEL IN EIGHT TRANSFER TASKS ON THE RETAIL CREDIT DATASET

(α, β)	19May \rightarrow Jun	19Oct \rightarrow Nov	19Dec \rightarrow Jan	20Jan \rightarrow Feb	20Feb \rightarrow Mar	20Mar \rightarrow Apr	20Apr \rightarrow May	20May \rightarrow Jun
(0.01,0.01)	80.02 \pm 0.04	79.33 \pm 0.06	92.23\pm0.005	88.04 \pm 0.02	91.14 \pm 0.01	92.28 \pm 0.01	92.75 \pm 0.003	91.12 \pm 0.01
(0.01,0.1)	79.97 \pm 0.04	80.95 \pm 0.04	91.89 \pm 0.005	87.14 \pm 0.02	90.71 \pm 0.01	91.65 \pm 0.01	92.67 \pm 0.004	90.71 \pm 0.01
(0.01, 1)	79.84 \pm 0.03	81.02\pm0.03	91.84 \pm 0.005	86.82 \pm 0.02	90.36 \pm 0.02	91.63 \pm 0.01	92.54 \pm 0.003	91.15 \pm 0.01
(0.1,0.01)	78.98 \pm 0.04	78.90 \pm 0.06	92.18 \pm 0.004	87.80 \pm 0.02	91.40 \pm 0.01	92.44 \pm 0.01	92.78\pm0.003	91.20 \pm 0.01
(0.1,0.1)	79.79 \pm 0.03	79.95 \pm 0.05	91.85 \pm 0.006	86.90 \pm 0.02	90.94 \pm 0.01	91.85 \pm 0.01	92.68 \pm 0.003	90.91 \pm 0.01
(0.1, 1)	79.95 \pm 0.03	80.87 \pm 0.04	91.86 \pm 0.005	87.17 \pm 0.02	90.54 \pm 0.01	91.50 \pm 0.02	92.52 \pm 0.004	91.02 \pm 0.01
(1,0.01)	80.44\pm0.04	80.79 \pm 0.06	91.93 \pm 0.005	88.20\pm0.02	91.51\pm0.01	92.46\pm0.01	92.74 \pm 0.003	91.39\pm0.01
(1,0.1)	80.26 \pm 0.04	79.66 \pm 0.06	91.90 \pm 0.006	87.64 \pm 0.02	90.91 \pm 0.01	91.97 \pm 0.01	92.76 \pm 0.003	91.07 \pm 0.01

TABLE IX
PARAMETER SENSITIVITY FOR THE CDAN MODEL ON THE EQUITY PRICE DATASET

(α, β)	RMSE	RE(%)	Q1 RE(%)	Q2 RE(%)	Q3 RE(%)	LRE(%)	SRE(%)
(0.01,0.01)	0.0235 \pm 2.3e-3	1.77 \pm 0.21	1.63	1.75	1.89	6.06 \pm 1.53	0.43 \pm 0.09
(0.01, 0.1)	0.0237 \pm 1.9e-3	1.80 \pm 0.19	1.66	1.79	1.92	6.14 \pm 1.58	0.43 \pm 0.08
(0.1,0.01)	0.0243 \pm 2.6e-3	1.83 \pm 0.26	1.66	1.79	1.98	6.52 \pm 2.00	0.44 \pm 0.10
(0.1, 0.1)	0.0243 \pm 2.6e-3	1.84 \pm 0.25	1.66	1.80	2.00	6.14 \pm 1.51	0.44 \pm 0.09
(0.1, 1)	0.0249 \pm 2.5e-3	1.87 \pm 0.26	1.70	1.86	2.03	6.40 \pm 1.80	0.46 \pm 0.09
(1, 0.1)	0.0247 \pm 2.3e-3	1.87 \pm 0.23	1.72	1.87	1.98	6.64 \pm 1.86	0.44 \pm 0.10
(1, 1)	0.0259 \pm 2.4e-3	1.94 \pm 0.23	1.80	1.95	2.06	6.68 \pm 1.67	0.48 \pm 0.09
(1, 10)	0.0306 \pm 4.7e-3	2.18 \pm 0.35	1.94	2.09	2.31	6.88 \pm 2.01	0.59 \pm 0.13
(10, 1)	0.0276 \pm 3.2e-3	2.04 \pm 0.27	1.83	2.01	2.20	6.87 \pm 1.85	0.50 \pm 0.11

TABLE X
ABLATION STUDY ON WINE QUALITY PREDICTION

(α, β)	W \rightarrow R R2	R \rightarrow W R2
(0, 0)	0.131	0.067
(0, 0.1)	0.145	0.012
(0, 1)	0.130	0.069
(0, 10)	0.121	0.070
(0.1, 0)	0.151	0.137
(1, 0)	0.181	0.158
(10, 0)	0.093	0.153
(1, 1)	0.201	0.177

TABLE XI
COMPARISON OF DIFFERENT DIVERGENCE MEASURES FOR WINE QUALITY PREDICTION

$\mathcal{H}_1 + \mathcal{H}_2$	W \rightarrow R			R \rightarrow W		
	RMSE	R2	RE	RMSE	R2	RE
MMD + KL	0.120	0.201	0.108	0.133	0.177	0.109
MMD + W1	0.135	0.140	0.124	0.146	0.178	0.115
MMD + χ^2	0.135	0.143	0.124	0.147	0.176	0.114
KL + KL	0.128	0.175	0.112	0.150	0.036	0.118
KL + W1	0.126	0.210	0.112	0.149	0.150	0.117
KL + χ^2	0.133	0.140	0.115	0.147	0.122	0.118
W1 + KL	0.124	0.191	0.110	0.145	0.112	0.115
W1 + W1	0.126	0.208	0.113	0.144	0.179	0.114
W1 + χ^2	0.125	0.200	0.110	0.144	0.151	0.114

best performances in each transfer task, we find that the coefficient of MD is in most cases larger than that of CD, indicating that the marginal differences and the dependence difference weigh differently in measuring the overall domain divergence. For the equity price dataset, we test nine candidate pairs of hyperparameters (α, β) and record the corresponding model performances in Table IX. It shows that as the coefficients increase, the model performance tends to get worse. Also, the change in the CD parameter β can significantly affect the model performance. It thus confirms the motivation of learning deep features by jointly adapting MD and CD since a good tradeoff between them could enhance feature transferability.

To evaluate the efficiency of taking MD and CD separately into the regularizer, we run experiments on the wine quality dataset. We compare the results of either $\alpha = 0$ or $\beta = 0$ and summarize them in Table X. The ablation study shows that MD and CD are both essential in terms of good model performance.

F. Comparison of Different Divergence Measures

As we have mentioned, there can be multiple choices over the divergence measures \mathcal{H}_1 and \mathcal{H}_2 . In this section, we investigate the performance difference caused by various divergence measures. Specifically, the candidate divergence

measures for \mathcal{H}_1 include KL divergence, W1 (abbreviated for Wasserstein-1) distance, and MMD. Also, the candidate divergence measures for \mathcal{H}_2 include KL divergence, χ^2 (abbreviated for Pearson χ^2) divergence, and W1 distance. In Table XI, we record the model performance of CDAN with the various combinations of \mathcal{H}_1 and \mathcal{H}_2 on the UCI wine quality dataset. From Table XI, we see that the CDAN model with \mathcal{H}_1 taking MMD distance and \mathcal{H}_2 taking KL divergence performs the best in terms of RMSE and RE. Also, it should be noted that the performance difference caused by the divergence measures can be as large as that brought by different models. This reminds us to be prudent in choosing suitable divergence measures.

VI. CONCLUSION

This work proposes a new domain adaptation framework that facilitates a user to detect whether the domain difference in a transfer task comes from the marginals' differences or the dependence difference. Specifically, we quantify the dependence difference with CD, a difference measure endowed with boundedness and monotonicity to guarantee the algorithm convergence. By optimizing the relative weights between the MD and the CD, we can acquire transferability across domains

in a more flexible way. Experiments on real-world datasets demonstrate the efficacy and robustness of our approach compared to a variety of existing domain adaptation models.

PROOF OF PROPOSITION 1 AND 2

Proof: We begin with the analysis over the explicit form of the bivariate CD $CD_{\mathcal{H}}(\mathbf{X}, \mathbf{Y})$ when \mathcal{H} is taken to be different divergence measures. Note that the CD between multivariate distributions is defined in terms of bivariate subdistributions. Thus, it is enough to prove the boundedness and monotonicity of $CD_{\mathcal{H}}(\mathbf{X}, \mathbf{Y})$ between any two bivariate random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^2$. Suppose that the Gaussian copula parameters for \mathbf{X} and \mathbf{Y} are $\Sigma^{\mathbf{X}}$ and $\Sigma^{\mathbf{Y}}$, respectively. Their copula density functions are

$$\begin{aligned} c^{\mathbf{X}}(u_1, u_2) &= |\Sigma^{\mathbf{X}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T((\Sigma^{\mathbf{X}})^{-1} - \mathbf{I})\mathbf{x}\right) \\ c^{\mathbf{Y}}(u_1, u_2) &= |\Sigma^{\mathbf{Y}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T((\Sigma^{\mathbf{Y}})^{-1} - \mathbf{I})\mathbf{x}\right) \end{aligned} \quad (6)$$

where $\mathbf{x} := [x_1, x_2]^T = [\Phi^{-1}(u_1), \Phi^{-1}(u_2)]^T$ with Φ being the cdf of the standard normal distribution.

The first divergence class is ϕ -divergence (see [41] for the detailed descriptions of the ϕ -divergence family). Given a convex function $\phi(x)$ such that $\phi(1) = 0$, the ϕ divergence between two distributions $P^{\mathbf{X}}$ and $P^{\mathbf{Y}}$ is defined by $\mathcal{H}_{\phi}(P^{\mathbf{X}}, P^{\mathbf{Y}}) = \int \phi(dP^{\mathbf{X}}/dP^{\mathbf{Y}})dP^{\mathbf{Y}}$. With the following proposition, we prove that the CD between bivariate random vectors \mathbf{X} and \mathbf{Y} , $CD_{\mathcal{H}_{\phi}}(\mathbf{X}, \mathbf{Y})$, can be fully characterized by the copula density functions $c^{\mathbf{X}}$ and $c^{\mathbf{Y}}$.

Proposition 4: For any bivariate random vector $\mathbf{X} \in \mathbb{R}^2$, the ϕ -divergence between the probability distribution $P^{\mathbf{X}}$ and the product of marginal distributions $P_1^{\mathbf{X}}P_2^{\mathbf{X}}$ is

$$\mathcal{H}_{\phi}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}}) = \int_0^1 \int_0^1 \phi(c^{\mathbf{X}}(u_1, u_2))du_1du_2.$$

For any two bivariate random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^2$, the CD between \mathbf{X} and \mathbf{Y} when \mathcal{H} takes ϕ -divergence is

$$\begin{aligned} CD_{\mathcal{H}_{\phi}}(\mathbf{X}, \mathbf{Y}) &= \left| \int_0^1 \int_0^1 \phi(c^{\mathbf{X}}(u_1, u_2)) - \phi(c^{\mathbf{Y}}(u_1, u_2))du_1du_2 \right|. \end{aligned}$$

Proof: We denote the two marginal density functions for the bivariate random vector \mathbf{X} as $p_1^{\mathbf{X}}(\cdot)$ and $p_2^{\mathbf{X}}(\cdot)$. By the definition of copula density function, the ϕ -divergence between $P^{\mathbf{X}}$ and $P_1^{\mathbf{X}}P_2^{\mathbf{X}}$ is

$$\begin{aligned} \mathcal{H}_{\phi}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}}) &= \int \phi\left(\frac{p_1^{\mathbf{X}}(x_1)p_2^{\mathbf{X}}(x_2)c^{\mathbf{X}}(P_1^{\mathbf{X}}(x_1), P_2^{\mathbf{X}}(x_2))}{p_1^{\mathbf{X}}(x_1)p_2^{\mathbf{X}}(x_2)}\right)dP_1^{\mathbf{X}}(x_1)dP_2^{\mathbf{X}}(x_2) \\ &= \int \phi(c^{\mathbf{X}}(P_1^{\mathbf{X}}(x_1), P_2^{\mathbf{X}}(x_2)))dP_1^{\mathbf{X}}(x_1)dP_2^{\mathbf{X}}(x_2). \end{aligned}$$

With change of variables $u_1 = P_1^{\mathbf{X}}(x_1)$ and $u_2 = P_2^{\mathbf{X}}(x_2)$, we finally have $\mathcal{H}_{\phi}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}}) = \int_0^1 \int_0^1 \phi(c^{\mathbf{X}}(u_1, u_2))du_1du_2$. For the bivariate random vector \mathbf{Y} , we similarly have $\mathcal{H}_{\phi}(P^{\mathbf{Y}}, P_1^{\mathbf{Y}}P_2^{\mathbf{Y}}) = \int_0^1 \int_0^1 \phi(c^{\mathbf{Y}}(u_1, u_2))du_1du_2$. The CD between \mathbf{X} and \mathbf{Y}

is defined as the absolute difference between $\mathcal{H}_{\phi}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}})$ and $\mathcal{H}_{\phi}(P^{\mathbf{Y}}, P_1^{\mathbf{Y}}P_2^{\mathbf{Y}})$. That completes the proof. \square

With Proposition 4, we can directly obtain the results in Section III.

- 1) When $\phi(x) = x^2 - 1$, the resulting ϕ -divergence is a χ^2 distance. Thus, $\mathcal{H}_{\chi^2}(P_{ij}, P_iP_j) = \int_0^1 \int_0^1 (c_{ij}^2(u_i, u_j) - 1)du_idu_j$.
- 2) When $\phi(x) = (\sqrt{x} - 1)^2$, it corresponds to Hellinger distance. Thus, we have $\mathcal{H}_H(P_{ij}, P_iP_j) = \int_0^1 \int_0^1 [\sqrt{c_{ij}(u_i, u_j)} - 1]^2 du_idu_j$.
- 3) When $\phi(x) = (x(1 - x^{-(\alpha+1)/2}))/ (1 - \alpha^2)$, it results in α -divergence. Thus, $\mathcal{H}_{\alpha}(P_{ij}, P_iP_j) = (1/(1 - \alpha^2)) \int_0^1 \int_0^1 [1 - c_{ij}(u_i, u_j)^{-(\alpha+1)/2}] c_{ij}(u_i, u_j) du_idu_j$.

Proposition 4 states that the CD defined by the ϕ -divergence is a function of the copula densities. Also, it proves that the ϕ -divergence between the joint distribution and the product of marginals is purely a function of the copula densities, that is to say, when the divergence metric is taken to be ϕ -divergence, the calculation of the CD has nothing to do with the marginal distributions. Thus, in the following proofs, when calculating the CD between any two random vectors \mathbf{X} and \mathbf{Y} , we will assume that the marginals are both standard normal distributions that have a mean of 1 and a variance of 0. This will greatly simplify our calculation. We can just calculate the CD between two bivariate Gaussian vectors with copula densities $c^{\mathbf{X}}(\mathbf{u})$ and $c^{\mathbf{Y}}(\mathbf{u})$. Furthermore, given that \mathbf{X} is Gaussian with standard normal marginals, we know that their Gaussian copula parameter $\Sigma^{\mathbf{X}}$ is exactly the correlation matrix [42]. In the following proof, we will write $\Sigma^{\mathbf{X}} := \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, with $\rho \in [-1, 1]$. Now, we are ready to provide the explicit forms of the CD for various choices of divergence measures.

KL Divergence: When $\phi(x) = x \log x$, the corresponding ϕ -divergence is KL divergence. By definition, we have

$$\mathcal{H}_{\text{KL}}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}}) = \iint p^{\mathbf{X}}(x_1, x_2) \log \frac{p^{\mathbf{X}}(x_1, x_2)}{p_1^{\mathbf{X}}(x_1)p_2^{\mathbf{X}}(x_2)} dx_1 dx_2.$$

Given that $p^{\mathbf{X}}(x_1, x_2) = p^{\mathbf{X}}(x_1)p^{\mathbf{X}}(x_2)c^{\mathbf{X}}(u_1, u_2)$, with (6), we have

$$\begin{aligned} \mathcal{H}_{\text{KL}}(P^{\mathbf{X}}, P_1^{\mathbf{X}}P_2^{\mathbf{X}}) &= \iint \frac{p^{\mathbf{X}}(x_1, x_2)([x_1, x_2](\mathbf{I} - (\Sigma^{\mathbf{X}})^{-1})[x_1, x_2]^T - \log |\Sigma^{\mathbf{X}}|)}{2} \\ &\quad \times dx_1 dx_2 \\ &= \frac{1}{2} \mathbb{E}_{p^{\mathbf{X}}}([x_1, x_2](\mathbf{I} - (\Sigma^{\mathbf{X}})^{-1})[x_1, x_2]^T - \log |\Sigma^{\mathbf{X}}|) \\ &= \frac{1}{2} (-\log |\Sigma^{\mathbf{X}}| + 2 - 2) \\ &= -\frac{1}{2} \log |\Sigma^{\mathbf{X}}|. \end{aligned}$$

The third equality comes from [43], where it proves that $\mathbb{E}_{p^{\mathbf{X}}}([x_1, x_2](\Sigma^{\mathbf{X}})^{-1}[x_1, x_2]^T) = 2$. Finally, with the definition of the CD, we have

$$CD_{\mathcal{H}_{\text{KL}}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{2} |\log |\Sigma^{\mathbf{X}}| - \log |\Sigma^{\mathbf{Y}}||.$$

χ^2 Distance: When $\phi(x) = x^2 - 1$, the resulting ϕ -divergence is a χ^2 distance. By definition, we have

$$\mathcal{H}_{\chi^2}(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) = \iint \left(\frac{p^{\mathbf{X}}(x_1, x_2)}{p_1^{\mathbf{X}}(x_1) p_2^{\mathbf{X}}(x_2)} \right)^2 p_1^{\mathbf{X}}(x_1) p_2^{\mathbf{X}}(x_2) dx_1 dx_2 - 1.$$

Since $(p^{\mathbf{X}}(x_1, x_2))/(p_1^{\mathbf{X}}(x_1) p_2^{\mathbf{X}}(x_2)) = c^{\mathbf{X}}(u_1, u_2)$ and $p_1^{\mathbf{X}}(x) = p_2^{\mathbf{X}}(x) = (1/(2\pi)^{1/2}) \exp(-x^2/2)$, we can further simplify the calculation as

$$\begin{aligned} \mathcal{H}_{\chi^2}(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) &= \iint \frac{\exp\left([x_1, x_2] \left(\frac{I}{2} - (\Sigma^{\mathbf{X}})^{-1}\right) [x_1, x_2]^T\right)}{2\pi |\Sigma^{\mathbf{X}}|} dx_1 dx_2 - 1 \\ &= |\Sigma^{\mathbf{X}}|^{-1} - 1. \end{aligned}$$

The last equality comes from the following fact: $2(\Sigma^{\mathbf{X}})^{-1} - I$ is positive definite with determinant 1. This gives: $\iint \exp([x_1, x_2]((I/2) - (\Sigma^{\mathbf{X}})^{-1})[x_1, x_2]^T) dx_1 dx_2 = 2\pi$. Finally, we have

$$\text{CD}_{\mathcal{H}_{\chi^2}}(\mathbf{X}, \mathbf{Y}) = ||\Sigma^{\mathbf{X}}|^{-1} - |\Sigma^{\mathbf{Y}}|^{-1}|.$$

It is not hard to derive more results about the CD for other ϕ -divergence. Thus, we omit them here and turn to the derivation for Wasserstein-2 distance and MMD distance.

Wasserstein-2 Distance: Assume that the marginal distributions of \mathbf{X}, \mathbf{Y} are standard normals. By directly applying the conclusion in [44, Proposition 7] we have

$$\mathcal{H}_W^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) = 4 - 2\text{Tr}((\Sigma^{\mathbf{X}})^{\frac{1}{2}}) = 4 - 2\sqrt{2 + 2\sqrt{|\Sigma^{\mathbf{X}}|}}.$$

Thus,

$$\begin{aligned} \text{CD}_{\mathcal{H}_W}(\mathbf{X}, \mathbf{Y}) &= \left| \sqrt{4 - 2\sqrt{2 + 2\sqrt{|\Sigma^{\mathbf{X}}|}}} - \sqrt{4 - 2\sqrt{2 + 2\sqrt{|\Sigma^{\mathbf{Y}}|}}} \right|. \end{aligned}$$

Gaussian MMD Distance: Assume that the marginal distributions of X and Y are standard normals. For ease of calculation, we take the simplest kernel function $k(\mathbf{X}, \mathbf{Y}) = e^{-||\mathbf{X} - \mathbf{Y}||_2^2}$. However, we emphasize that the following calculation applies to all Gaussian kernels. Using the kernel trick, the squared MMD distance can be computed as the expectation of kernel functions

$$\begin{aligned} \mathcal{H}_{\text{MMD}}^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) &= \mathbb{E}_{\mathbf{X}, \mathbf{X}} k(\mathbf{X}, \mathbf{X}) + \mathbb{E}_{\dot{\mathbf{X}}, \dot{\mathbf{X}}} k(\dot{\mathbf{X}}, \dot{\mathbf{X}}) - 2\mathbb{E}_{\mathbf{X}, \dot{\mathbf{X}}} k(\mathbf{X}, \dot{\mathbf{X}}) \quad (7) \end{aligned}$$

where $\dot{\mathbf{X}} \in \mathbb{R}^2$ is a random Gaussian vector with cdf $P^{\dot{\mathbf{X}}}(x_1, x_2) = P_1^{\dot{\mathbf{X}}}(x_1) P_2^{\dot{\mathbf{X}}}(x_2)$ and the Gaussian copula parameter $\Sigma^{\dot{\mathbf{X}}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

From (7), we know that to calculate the squared MMD distance $\mathcal{H}_{\text{MMD}}^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}})$, we need to calculate the three expectations on the right-hand side of this equation. We begin with the calculation of $\mathbb{E}_{\mathbf{X}, \dot{\mathbf{X}}} k(\mathbf{X}, \dot{\mathbf{X}})$. By definition, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \dot{\mathbf{X}}} k(\mathbf{X}, \dot{\mathbf{X}}) &= \iint \frac{k(\mathbf{X}, \mathbf{Y}) \exp\left(-\frac{1}{2}\mathbf{X}^T (\Sigma^{\mathbf{X}})^{-1} \mathbf{X} - \frac{1}{2}\mathbf{Y}^T (\Sigma^{\dot{\mathbf{X}}})^{-1} \mathbf{Y}\right)}{4\pi^2 \sqrt{|\Sigma^{\mathbf{X}} \Sigma^{\dot{\mathbf{X}}}|}} \end{aligned}$$

$$\begin{aligned} &\times d\mathbf{X} d\mathbf{Y} \\ &= \iiint \frac{\exp\left(-\frac{1}{2}[x_1, x_2, y_1, y_2] A [x_1, x_2, y_1, y_2]^T\right)}{4\pi^2 \sqrt{|\Sigma^{\mathbf{X}} \Sigma^{\dot{\mathbf{X}}}|}} \\ &\quad \times dx_1 dx_2 dy_1 dy_2 \\ &= \frac{1}{4\pi^2 \sqrt{|\Sigma^{\mathbf{X}} \Sigma^{\dot{\mathbf{X}}}|}} \times (2\pi)^2 |A|^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{|2\Sigma^{\mathbf{X}} + 2\Sigma^{\dot{\mathbf{X}}} + I|}} = \frac{1}{\sqrt{21 + 4|\Sigma^{\mathbf{X}}|}}. \end{aligned}$$

Here, the matrix $A := \begin{pmatrix} (\Sigma^{\mathbf{X}})^{-1} + 2I & -2I \\ -2I & (\Sigma^{\dot{\mathbf{X}}})^{-1} + 2I \end{pmatrix} \in \mathbb{R}^{4 \times 4}$ is positive semidefinite with determinant $(2\Sigma^{\mathbf{X}} + 2\Sigma^{\dot{\mathbf{X}}} + I_2)/(|\Sigma^{\mathbf{X}} \Sigma^{\dot{\mathbf{X}}}|)$. Similarly, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \mathbf{X}} k(\mathbf{X}, \mathbf{X}) &= \frac{1}{\sqrt{|4\Sigma^{\mathbf{X}} + I|}} = \frac{1}{\sqrt{9 + 16|\Sigma^{\mathbf{X}}|}} \\ \mathbb{E}_{\dot{\mathbf{X}}, \dot{\mathbf{X}}} k(\dot{\mathbf{X}}, \dot{\mathbf{X}}) &= \frac{1}{\sqrt{|4\Sigma^{\dot{\mathbf{X}}} + I|}} = \frac{1}{5}. \end{aligned}$$

Organizing the three terms together, we have the squared MMD distance

$$\mathcal{H}_{\text{MMD}}^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) = \frac{1}{\sqrt{9 + 16|\Sigma^{\mathbf{X}}|}} + \frac{1}{5} - \frac{2}{\sqrt{21 + 4|\Sigma^{\mathbf{X}}|}}$$

and the CD

$$\begin{aligned} \text{CD}_{\mathcal{H}_{\text{MMD}}}(\mathbf{X}, \mathbf{Y}) &= \left| \sqrt{\frac{1}{\sqrt{9 + 16|\Sigma^{\mathbf{X}}|}} + \frac{1}{5} - \frac{2}{\sqrt{21 + 4|\Sigma^{\mathbf{X}}|}}} \right. \\ &\quad \left. - \sqrt{\frac{1}{\sqrt{9 + 16|\Sigma^{\mathbf{Y}}|}} + \frac{1}{5} - \frac{2}{\sqrt{21 + 4|\Sigma^{\mathbf{Y}}|}}} \right|. \end{aligned}$$

Boundedness: Given that $\Sigma^{\mathbf{X}}$ and $\Sigma^{\mathbf{Y}}$ for Gaussian random vectors are in essence the correlation matrix, we know that $|\Sigma^{\mathbf{X}}| \leq 1$ and $|\Sigma^{\mathbf{Y}}| \leq 1$. Thus, it is easy to verify that when \mathcal{H} is Wasserstein-2 distance or Gaussian MMD distance, the CD is bounded. Furthermore, we know that the divergence measures (including the total variation distance, Hellinger distance, and JS divergence) are bounded by definition. Consequently, the corresponding CD is bounded.

Monotonicity: We fix the Gaussian copula parameter $\Sigma^{\mathbf{X}}$ and express $\text{CD}_{\mathcal{H}}(\mathbf{X}, \mathbf{Y})$ as a function of $\Sigma_{12}^{\mathbf{X}} = \rho$. A simple observation is that, if a function $f(x)$ is monotonically increasing with respect to x , then given y fixed, $|f(x) - f(y)|$ is monotonically increasing with respect to $|x - y|$. Thus, if $\mathcal{H}(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}})$ is increasing with respect to ρ^2 , we can conclude that the corresponding CD is monotonically increasing with $|(\Sigma_{12}^{\mathbf{X}})^2 - (\Sigma_{12}^{\mathbf{Y}})^2|$. We check them one by one.

- 1) $\mathcal{H}_{\text{KL}}(P_{12}, P_1 P_2) = -(1/2)(1 - \rho^2)$ monotonically increases with ρ^2 .
- 2) $\mathcal{H}_{\chi^2}(P_{12}, P_1 P_2) = (1/(1 - \rho^2)) - 1$ monotonically increases with ρ^2 .
- 3) $\mathcal{H}_W^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) = 4 - 2\sqrt{2 + 2\sqrt{1 - \rho^2}}$ monotonically increases with ρ^2 .
- 4) $\mathcal{H}_{\text{MMD}}^2(P^{\mathbf{X}}, P_1^{\mathbf{X}} P_2^{\mathbf{X}}) = (1/((25 - 16\rho^2)^{1/2})) + (1/5) - (2/((25 - 4\rho^2)^{1/2}))$.

From the above calculations, we conclude that $\mathcal{H}(P^X, P_1^X P_2^X)$ is increasing with respect to ρ^2 when \mathcal{H} is KL divergence, χ^2 distance, and Wasserstein-2 distance. For Gaussian MMD distance, we consider the function $f(x) = (1/((25 - 16x)^{1/2})) - (2/((25 - 4x)^{1/2}))$, $x \in [0, 1]$. The first derivative $f'(x) = 8(25 - 16x)^{-3/2} - 4(25 - 4x)^{-3/2} > 0$, suggesting that $\mathcal{H}_{\text{MMD}}(P^X, P_1^X P_2^X)$ is increasing with respect to ρ^2 . \square

PROOF OF PROPOSITION 3

Proof: Without loss of generality, we assume that the probability density function $p(x_1, x_2)$ of $[X_1, X_2]$ is continuous. Consider the area $A_\delta = \{(x_1, x_2, \tilde{x}_1, \tilde{x}_2) : |x_1 - \tilde{x}_1| \leq \delta \text{ or } |x_2 - \tilde{x}_2| \leq \delta\} \subset \mathbb{R}^4$. Define $c_\delta := \iiint\limits_{A_\delta} p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2$. It always holds that

$$\begin{aligned} \rho(X_1, X_2; a) &= \mathbb{E}[\tanh(a(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))] \\ &= \left(\iiint\limits_{A_\delta} + \iiint\limits_{\mathbb{R}^4 - A_\delta} \right) \tanh(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)) \\ &\quad \times p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) d\tilde{x}_1 d\tilde{x}_2 dx_1 dx_2. \end{aligned}$$

Outside A_δ , we have $|a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)| \geq a\delta^2$ and

$$\begin{aligned} &\lim_{a \rightarrow \infty} \iiint\limits_{\mathbb{R}^4 - A_\delta} \tanh(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)) \\ &\quad \times p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2 \\ &= \lim_{a \rightarrow \infty} \iiint\limits_{\mathbb{R}^4 - A_\delta} \text{sign}(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)) \\ &\quad \times p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2 \\ &\quad \text{(dominated convergence theorem)} \\ &= \rho_\tau(X_1, X_2) - \lim_{a \rightarrow \infty} \iiint\limits_{A_\delta} \text{sign}(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)) \\ &\quad \times p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2. \end{aligned}$$

Notice that inside A_δ , $|\tanh(x)| \leq 1$. Thus, $\forall \delta$

$$\begin{aligned} &|\lim_{a \rightarrow \infty} \rho(X_1, X_2; a) - \rho_\tau(X_1, X_2)| \\ &\leq \lim_{a \rightarrow \infty} \iiint\limits_{A_\delta} (|\text{sign}(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2))| \\ &\quad + |\tanh(a(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2))|) \\ &\quad \times p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2 \\ &\leq 2 \iiint\limits_{A_\delta} p(x_1, x_2) p(\tilde{x}_1, \tilde{x}_2) dx_1 dx_2 d\tilde{x}_1 d\tilde{x}_2 = 2 c_\delta. \end{aligned}$$

Indeed, it is easy to verify that $c_0 = 0$, and c_δ is finite and is continuous with respect to δ . Thus, we let δ approach zero and get the desired result that $\lim_{a \rightarrow \infty} \rho(X_1, X_2; a) = \rho_\tau(X_1, X_2)$. \square

IMPLEMENTATION DETAILS

All experimental models were trained using the Adam optimizer implemented by Pytorch with an initial learning rate of 0.01. All activation functions were taken to be the ReLU.

A. Toy Problem

The MLP model is a simple neural network with two hidden layers of eight and four neurons.

In the DAN model, the feature extractor is a neural network with two hidden layers of eight and four neurons. The domain discrepancy is evaluated by the MMD distance and the discriminator is the final output layer.

The CORAL model has the same architecture as the DAN model, except for that the domain discrepancy is measured by the Frobenius norm.

The CDAN model differs from the CORAL model only in the measurement of the domain discrepancy.

B. Retail Credit Classification

The MLP model is a simple neural network with two hidden layers of 128 and 64 neurons.

In the DAN model, the feature extractor is a neural network with three hidden layers of 64, 32, and 16 neurons. The domain discrepancy is evaluated by the MMD distance of the 16-D feature representations between the source and the target domains. The discriminator is the final output layer.

The AFN model contains three hidden layers with 128, 64, and 64 neurons. The output tensors of the second hidden layer are aligned to a scale vector, which is predetermined according to [37].

In the MCD model, the neural network consists of three hidden layers with 128, 64, and 64 neurons. For output tensors of the second hidden layer from the target domain, two additional networks are used to construct the regularization term mentioned in [38].

The CORAL model is essentially the same model as the DAN model, except that the domain discrepancy is evaluated by the Frobenius norm of the covariance matrices of the 16-D feature representations.

In the CDAN model, the feature extractor is a neural network with six hidden layers of 128, 128, 128, 128, 64, and eight neurons. The MD is calculated by the Gaussian MMD distance of each dimensional representation among the 8-D features. The CD is with respect to the KL divergence. We run the model for 100 trials, and each trial costs about 2–3 min.

C. Intraday Equity Price Regression

After separating the historical prices of 22 stocks into two domains, we slice the data into pieces with length 12 and package them into batches of size 1024. For each stock in either domain, we use MinMaxScaler to normalize its price. For each model, we use an LSTM layer as the feature extractor and conduct the batch normalization for each linear layer. We train each model for at most 100 epochs, and the early stopping threshold is set to be 20 epochs. We tune the hyperparameters by grid search and also fine-tune the network parameters (including but not limited to the number of layers, the number of units, and so on).

The LSTM (resp. RNN) model consists of one LSTM (resp. RNN) layer of hidden size 64 and two linear layers of size 64 and 32.

The DANN model consists of three neural networks, namely, a feature extractor, a discriminator, and a regressor. The feature extractor contains one LSTM layer of hidden size 64 and a linear layer of size also 64. The discriminator is a binary classifier, which consists of three linear layers of size 64, 32, and 16. The regressor consists of two linear layers of size 64 and 32.

The CORAL model consists of one LSTM layer of hidden size 64 and three linear layers of size 64, 32, and 16. After extracting the features by LSTM, we calculate the regularization term according to [17].

The DAN model consists of one LSTM layer of hidden size 64 and three linear layers of size 64, 32, and 16. After extracting the features by LSTM, we calculate the MMD with a two-Gaussian-kernel function.

Our model CDAN consists of one LSTM layer of hidden size 64 and two linear layers of size 64 and 32. After extracting the features by LSTM, we calculate the divergence between marginal distributions by a two-Gaussian-kernel MMD and calculate the CD with respect to the KL divergence. We run the model for 100 trials, and each trial costs about 10–20 min.

REFERENCES

- [1] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Apr. 2016.
- [2] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.
- [3] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, “Adversarial multiple source domain adaptation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 8559–8570.
- [4] H. Liu, M. Long, J. Wang, and M. Jordan, “Transferable adversarial training: A general approach to adapting deep classifiers,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4013–4022.
- [5] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.
- [6] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
- [7] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.
- [8] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.
- [9] J. Shen, Y. Qu, W. Zhang, and Y. Yu, “Wasserstein distance guided representation learning for domain adaptation,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
- [10] M. Sklar, “Fonctions de repartition an dimensions et Leurs Marges,” *Publ. Inst. Statist. Univ. Paris*, vol. 8, no. 3, pp. 229–231, 1959.
- [11] S. Zhao et al., “A review of single-source deep unsupervised visual domain adaptation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.
- [12] L. Duan, D. Xu, and I. W. Tsang, “Domain adaptation from multiple sources: A domain-dependent regularization approach,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.
- [13] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, “On learning invariant representations for domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7523–7532.
- [14] S. Chen, L. Han, X. Liu, Z. He, and X. Yang, “Subspace distribution adaptation frameworks for domain adaptation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5204–5218, Dec. 2020.
- [15] Z. Wang, B. Du, and Y. Guo, “Domain adaptation with neural embedding matching,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2387–2397, Jul. 2020.
- [16] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. L. P. Chen, “Guide subspace learning for unsupervised domain adaptation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3374–3388, Sep. 2020.
- [17] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1–11.
- [18] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–450.
- [19] Z. Zhang, M. Wang, Y. Huang, and A. Nehorai, “Aligning infinite-dimensional covariance matrices in reproducing kernel Hilbert spaces for domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3437–3445.
- [20] C. Chen et al., “HoMM: Higher-order moment matching for unsupervised domain adaptation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3422–3429.
- [21] Y. Zhao and M. Udell, “Missing value imputation for mixed data via Gaussian copula,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 636–646.
- [22] G. Marti, S. Andler, F. Nielsen, and P. Donnat, “Exploring and measuring non-linear correlations: Copulas, lightspeed transportation and clustering,” in *Proc. NIPS Time Workshop*, 2017, pp. 59–69.
- [23] C. D. Tran, O. O. Rudovic, and V. Pavlovic, “Unsupervised domain adaptation with copula models,” in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
- [24] D. Lopez-Paz, J. Hernandez-Lobato, and B. Schölkopf, “Semi-supervised domain adaptation with copulas,” in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 674–682.
- [25] N. A. Letizia and A. M. Tonello, “Segmented generative networks: Data generation in the uniform probability space,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1338–1347, Mar. 2022.
- [26] J. Xuan, J. Lu, G. Zhang, R. Y. D. Xu, and X. Luo, “Doubly non-parametric sparse nonnegative matrix factorization based on dependent Indian buffet processes,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1835–1849, May 2018.
- [27] H. Quan, A. Khosravi, D. Yang, and D. Srinivasan, “A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4582–4599, Nov. 2020.
- [28] S. Nowozin, B. Cseke, and R. Tomioka, “f-GAN: Training generative neural samplers using variational divergence minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 1–137. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/file/cedeb6e872f539bef8c3f919874e9d7-Paper.pdf>
- [29] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. A. Katsoulakis, “Cumulant GAN,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 6, 2022, doi: [10.1109/TNNLS.2022.3161127](https://doi.org/10.1109/TNNLS.2022.3161127).
- [30] C. Donnelly and P. Embrechts, “The devil is in the tails: Actuarial mathematics and the subprime mortgage crisis,” *ASTIN Bull.*, vol. 40, no. 1, pp. 1–33, 2010.
- [31] Y. Fang and L. Madsen, “Modified Gaussian pseudo-copula: Applications in insurance and finance,” *Insurance, Math. Econ.*, vol. 53, no. 1, pp. 292–301, Jul. 2013.
- [32] W. Y. Wang and Z. Hua, “A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls,” in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, Mar. 2014, pp. 1155–1165.
- [33] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–10.
- [34] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, p. 137.
- [35] D. Ruppert and D. S. Matteson, *Statistics and Data Analysis for Financial Engineering*, vol. 13. New York, NY, USA: Springer, 2011.
- [36] W. Xiao, Z. Ding, and H. Liu, “Implicit semantic response alignment for partial domain adaptation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13820–13833.
- [37] R. Xu, G. Li, J. Yang, and L. Lin, “Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1426–1435.
- [38] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum classifier discrepancy for unsupervised domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

- [39] D. Belov and R. Armstrong, "Distributions of the Kullback–Leibler divergence with applications," *Brit. J. Math. Stat. Psychol.*, vol. 64, pp. 291–309, May 2011.
- [40] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decis. Support Syst.*, vol. 47, no. 4, pp. 547–553, Nov. 2009.
- [41] I. Sason and S. Verdú, " f -divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
- [42] L. D. Valle, "Bayesian copulae distributions, with application to operational risk management," *Methodol. Comput. Appl. Probab.*, vol. 11, no. 1, pp. 95–115, Mar. 2009.
- [43] K. B. Petersen and M. S. Pedersen. (Nov. 2012). *The Matrix Cookbook*. [Online]. Available: <http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html>
- [44] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 31, no. 2, pp. 231–240, Jan. 1984, doi: [10.1307/mmj/1029003026](https://doi.org/10.1307/mmj/1029003026).



Shumin Ma received the B.S. degree in mathematics from Fudan University, Shanghai, China, in 2013, and the Ph.D. degree in financial mathematics from The Chinese University of Hong Kong, Hong Kong, in 2018.

She is currently an Assistant Professor with the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science and the Department of Mathematical Science, BNU-HKBU United International College (UIC), Zhuhai, China. Prior to joining UIC, she held a

post-doctoral position at the School of Data Science, City University of Hong Kong, Hong Kong. She was a Visiting Scholar with the Columbia Business School, New York, NY, USA.



Zhiri Yuan received the B.S. degree in mathematics from Fudan University, Shanghai, China, in 2013, and the Ph.D. degree in mathematics from The Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2020.

He specialized in algebraic geometry and representation theory and worked as a Teaching Assistant at CUHK during his Ph.D. degree. After graduation, he worked at Ortec Finance, Hong Kong, as a Data Scientist for one year, where he specialized in the real estate sector and used machine learning to

forecast house prices. Later, he returned to academia and served as a Post-Doctoral Researcher, focusing on deep learning applications for financial markets, at the JD Digital Lab, School of Data Science, City University of Hong Kong, Hong Kong, from 2020 to 2022.



Qi Wu received the Ph.D. degree in applied mathematics from Columbia University, New York, NY, USA, in 2013.

He is currently an Associate Professor with the School of Data Science, City University of Hong Kong (CityU), Hong Kong. He is also the Program Leader at Laboratory for AI-Powered Financial Technologies Ltd., Hong Kong. Prior to CityU, he was an Assistant Professor at The Chinese University of Hong Kong and a Finance Practitioner at Lehman Brothers, London, U.K.; UBS, Stamford,

CT, USA; and DTCC, New York.



Yiyan Huang received the bachelor's degree in finance from the Department of Statistics and Finance, University of Science and Technology of China (USTC), Hefei, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Data Science, City University of Hong Kong (CityU), Hong Kong.

He is also a Visiting Research Fellow with the Department of Economics, Brown University, Providence, RI, USA. His research interests lie at the intersection of economics and finance and machine

learning, with a special focus on causal inference.



Xixu Hu received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Data Science, City University of Hong Kong, Hong Kong, under the supervision of Prof. Qi Wu.

Her research interests include domain adaptation, domain generalization, and adversarial robustness.



Cheuk Hang Leung received the B.Sc. degree in mathematics from The University of Hong Kong, Hong Kong, in 2009, the M.Sc. degree from The Hong Kong University of Science and Technology, Hong Kong, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2018.

He is currently working at the JD Digits Joint Laboratory in Financial Technology and Engineering, City University of Hong Kong, Hong Kong.



Dongdong Wang received the master's degree in electrical and information engineering from Tianjin University (TJU), Tianjin, China, in 2018.

He is currently an AI Researcher at JD Digits Technology, Beijing, China, specializing in scientific decision-making and financial risk management research. He has a particular interest in deep learning and causal inference.



Zhixiang Huang is currently the Director of the Financial AI Team, JD Digits Technology, Beijing, China, where he leads efforts to develop cutting-edge artificial intelligence (AI) models for financial products. His team's focus areas include optimizing user growth, credit scoring, fraud and money laundering detection, and post-loan management. His expertise spans a wide range of AI technologies, encompassing multimodal data mining, scorecard technology, causal inference, graph computing, and more, which he applies to drive innovation and excellence in his team's work.