Measuring the Discrepancy between Conditional Distributions: Methods, Properties and Applications

Shujian Yu[1*], **Ammar Shaker**[1], **Francesco Alesiani**[1] and **Jose Principe**[2]
[1]NEC Labs Europe, 69115 Heidelberg, Germany
[2]University of Florida, Gainesville, FL 32611, USA
{Shujian.Yu, Ammar.Shaker, Francesco.Alesiani}@neclab.eu, principe@cnel.ufl.edu

Abstract

We propose a simple yet powerful test statistic to quantify the discrepancy between two conditional distributions. The new statistic avoids the explicit estimation of the underlying distributions in high-dimensional space and it operates on the cone of symmetric positive semidefinite (SPS) matrix using the Bregman matrix divergence. Moreover, it inherits the merits of the correntropy function to explicitly incorporate high-order statistics in the data. We present the properties of our new statistic and illustrate its connections to prior art. We finally show the applications of our new statistic on three different machine learning problems, namely the multi-task learning over graphs, the concept drift detection, and the information-theoretic feature selection, to demonstrate its utility and advantage. Code of our statistic is available at https://bit.ly/BregmanCorrentropy.

1 Introduction

Measuring the discrepancy or divergence between two conditional distribution functions plays a leading role in numerous real-world machine learning problems. One vivid example is the modeling of the seasonal effects on consumer preferences, in which the statistical analyst needs to distinguish the changes on the distributions of the merchandise sales conditioning on the explanatory variables such as the amount of money spent on advertising, the promotions being run, etc.

Despite substantial efforts have been made on specifying the discrepancy for unconditional distribution (density) functions (see [Anderson *et al.*, 1994; Gretton *et al.*, 2012b; Pardo, 2005] and the references therein), methods on quantifying the discrepancy of regression models or statistical tests for conditional distributions are scarce.

Prior art falls into two categories. The first relies heavily on the precise estimation of the underlying distribution functions using different density estimators, such as the k-Nearest Neighbor (kNN) estimator [Wang *et al.*, 2009] and the kernel density estimator (KDE) [Lee and Park, 2006]. However, density estimation is notoriously difficult for high-dimensional data. Moreover, existing conditional tests (e.g., [Zheng, 2000; Fan *et al.*, 2006]) are always one-sample based, which means that they are designed to test if the observations are generated by a conditional distribution $p(y|x)$ in a particular parametric family with parameter θ, rather than distinguishing $p_1(y|x)$ from $p_2(y|x)$. Another category defines a distance metric through the embedding of probability measures in another space (typically the reproducing kernel Hilbert space or RKHS). A notable example is the Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012b] which has attracted much attention in recent years due to its solid mathematical foundation. However, MMD always requires high computational burden and carefully hyper-parameter (e.g., the kernel width) tuning [Gretton *et al.*, 2012a]. Moreover, computing the distance of the embeddings of two conditional distributions in RKHS still remains a challenging problem [Ren *et al.*, 2016].

Different from previous efforts, we propose a simple statistic to quantify the discrepancy between two conditional distributions. It directly operates on the cone of symmetric positive semidefinite (SPS) matrix to avoid the estimation of the underlying distributions. To strengthen the discriminative power of our statistic, we make use of the correntropy function [Santamaría *et al.*, 2006], which has demonstrated its effectiveness in non-Gaussian signal processing [Liu *et al.*, 2007], to explicitly incorporate higher order information in the data. We demonstrate the power of our statistic and establish its connections to prior art. Three solid examples of machine learning applications are presented to demonstrate the effectiveness and the superiority of our statistic.

2 Background Knowledge

2.1 Bregman Matrix Divergence and Its Computation

A symmetric matrix is positive semidefinite (SPS) if all its eigenvalues are non-negative. We denote $\mathcal{S}_+^n$ the set of all $n \times n$ SPS matrices, i.e., $\mathcal{S}_+^n = \{A \in \mathbb{R}^{n\times n} | A = A^T, A \succcurlyeq 0\}$. To measure the nearness between two SPS matrices, a reliable choice is the Bregman matrix divergence [Kulis *et al.*, 2009]. Specifically, given a strictly convex, differentiable function φ that maps matrices to the extended real numbers, the Bregman divergence from the matrix ρ to the matrix σ is

*Contact Author

defined as:

$$D_{\varphi,B}(\sigma\|\rho) = \varphi(\sigma) - \varphi(\rho) - \mathrm{tr}((\nabla\varphi(\rho))^T(\sigma-\rho)), \quad (1)$$

where $\mathrm{tr}(A)$ denotes the trace of matrix A.

When $\varphi(\sigma) = \mathrm{tr}(\sigma\log\sigma - \sigma)$, where $\log\sigma$ is the matrix logarithm, the resulting Bregman divergence is:

$$D_{vN}(\sigma\|\rho) = \mathrm{tr}(\sigma\log\sigma - \sigma\log\rho - \sigma + \rho), \quad (2)$$

which is also referred to von Neumann divergence in quantum information theory [Nielsen and Chuang, 2011]. Another important matrix divergence arises by taking $\varphi(\sigma) = -\log\det\sigma$, in which the resulting Bregman divergence reduces to:

$$D_{\ell D}(\sigma\|\rho) = \mathrm{tr}(\rho^{-1}\sigma) + \log_2\frac{|\rho|}{|\sigma|} - n, \quad (3)$$

and is commonly called the LogDet divergence.

2.2 Correntropy Function: A Generalized Correlation Measure

The correntropy function of two random variables x and y is defined as [Santamaría *et al.*, 2006]:

$$V(x,y) = \mathbb{E}[\kappa(x,y)] = \iint \kappa(x,y)dF_{X,Y}(x,y), \quad (4)$$

where $\mathbb{E}$ denotes mathematical expectation, κ is a positive definite kernel function, and $F_{X,Y}(x,y)$ is the joint distribution of (X,Y). One widely used kernel function is the Gaussian kernel given by:

$$\kappa(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-y)^2}{2\sigma^2}\}. \quad (5)$$

Taking Taylor series expansion of the Gaussian kernel, we have:

$$V_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n n!}\mathbb{E}[\frac{(x-y)^{2n}}{\sigma^{2n}}]. \quad (6)$$

Therefore, correntropy involves all the even moments[1] of random variable $e = x - y$. Furthermore, increasing the kernel size σ makes correntropy tends to the correlation of x and y [Santamaría *et al.*, 2006].

A similar quantity to the correntropy is the centered correntropy:

$$\begin{aligned} U(x,y) &= \mathbb{E}[\kappa(x,y)] - \mathbb{E}_x\mathbb{E}_y[\kappa(x,y)] \\ &= \iint \kappa(x,y)(dF_{X,Y}(x,y) - dF_X(x)dF_Y(y)), \end{aligned} \quad (7)$$

where $F_X(x)$ and $F_Y(y)$ are the marginal distributions of X and Y, respectively.

The centered correntropy can be interprested as a nonlinear counterpart of the covariance in RKHS [Chen *et al.*, 2017]. Moreover, the following property serves as the basis of our test statistic that will be introduced later.

[1]A different kernel would yield a different expansion, for instance the sigmoid kernel $\kappa(x,y) = \tanh(\langle x,y\rangle + \theta)$ admits an expansion in terms of the odd moments of its argument.

Property 2.1. *Given n random variables $x_1, x_2, \cdots, x_n$ and any set of real numbers $\alpha_1, \alpha_2, \cdots, \alpha_n$, for any symmetric positive definite kernel $\kappa(x,y)$, the centered correntropy matrix C defined as $C(i,j) = U(x_i,x_j)$ is always positive semidefinite, i.e., $C \in \mathcal{S}_+^n$.*

Proof. By Property 2 in [Rao *et al.*, 2011], $U(x,y)$ is a symmetric positive semidefinite function, from which it follows that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j U(x_i,x_j) \geq 0. \quad (8)$$

C is obviously symmetric. This concludes our proof. □

In practice, data distributions are unknown and only a finite number of observations $\{x_i,y_i\}_{i=1}^N$ are available, which leads to the sample estimator of centered correntropy[2] [Rao *et al.*, 2011]:

$$\hat{U}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_i,y_i) - \frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N}\kappa_\sigma(x_i,y_j). \quad (9)$$

3 Methods

3.1 Problem Formulation

We have two groups of observations $\{(x_i^1,y_i^1)\}_{i=1}^{N_1}$ and $\{(x_i^2,y_i^2)\}_{i=1}^{N_2}$ that are assumed to be independently and identically distributed ($i.i.d.$) with density functions $p_1(x,y)$ and $p_2(x,y)$, respectively. y is a dependent variable that takes values in $\mathbb{R}$, and x is a vector of explanatory variables that takes values in $\mathbb{R}^p$. Typically, the conditional distribution $p(y|x)$ are unknown and unspecified. The aim of this paper is to suggest a test statistic to measure the nearness between $p_1(y|x)$ and $p_2(y|x)$.

$$\begin{aligned} H_0 &: \mathsf{Pr}\left(p_1(y|x) = p_2(y|x)\right) = 1 \\ H_1 &: \mathsf{Pr}\left(p_1(y|x) = p_2(y|x)\right) < 1 \end{aligned}$$

3.2 Our Statistic and the Conditional Test

We define the divergence from $p_1(y|x)$ to $p_2(y|x)$ as:

$$\begin{aligned} D_{\varphi,B}(p_1(y|x)\|p_2(y|x)) = \quad & D_{\varphi,B}(C_{xy}^1\|C_{xy}^2) \\ & -D_{\varphi,B}(C_x^1\|C_x^2), \end{aligned} \quad (10)$$

where $C_{xy} \in \mathcal{S}_+^{p+1}$ denotes the centered correntropy matrix of the random vector concatenated by x and y, and $C_x \in \mathcal{S}_+^p$ denotes the centered correntropy matrix of x. Obviously, C_x is a submatrix of C_{xy} by removing the row and column associated with y.

Although Eq. (10) is assymetic itself, one can easily achieve symmetry by taking the form:

$$\begin{aligned} D_{\varphi,B}(p_1(y|x) : p_2(y|x)) = \frac{1}{2}(&D_{\varphi,B}(p_1(y|x)\|p_2(y|x)) \\ &+ D_{\varphi,B}(p_2(y|x)\|p_1(y|x))). \end{aligned} \quad (11)$$

[2]Throughout this work, we determine kernel width σ with the Silverman's rule of thumb [Silverman, 1986].

Algorithm 1 Test the conditional distribution divergence (CDD) based on the matrix Bregman divergence

Input: Two groups of observations $S^1 = \{(x_i^1, y_i^1)\}_{i=1}^{N_1}$ and $S^2 = \{(x_i^2, y_i^2)\}_{i=1}^{N_2}$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$; $D_{\varphi,B}$; Permutation number P; Significant rate η.
Output: Test *decision* (Is $H_0 : \mathsf{Pr}\,(p_1(y|x) = p_2(y|x)) = 1$ $True$ or $False$?).

```
Measure CDD d_0 on S^1 and S^2 with Eq. (11).
for t = 1 to P do
    (S_t^1, S_t^2) ← random split of S^1 ⋃ S^2.
    Measure CDD d_t on S_t^1 and S_t^2 with Eq. (11).
end for
if (1+Σ_{t=1}^{P} 1[d_0 ≤ d_t]) / (1+P) ≤ η then
    decision←False
else
    decision←True
end if
return decision
```

Given Eq. (10) and Eq. (11), we design a simple permutation test to distinguish $p_1(y|x)$ from $p_2(y|x)$. Our test methodology is shown in Algorithm 1, where $\mathbf{1}$ indicates an indicator function. The intuition behind this scheme is that if there is no difference on the underlying conditional distributions, the test statistic on the ordered split (i.e., d_0) should not deviate too much from that of the shuffled splits (i.e., $\{d_t\}_{t=1}^{P}$).

4 Conditional Divergence Properties

We present useful properties of our statistic (i.e., Eq. (10) or Eq. (11)) based on the LogDet divergence. In particular we show it is non-negative (when evaluated on covariance matrices) which reduces to zero when two data sets share the same linear regression function. We also perform Monte Carlo simulations to investigate its detection power and establish its connection to prior art.

Property 4.1. $D_{\ell d}(\Sigma_{xy}^1||\Sigma_{xy}^2) - D_{\ell d}(\Sigma_x^1||\Sigma_x^2) \geq 0$.

Proof. The proof follows immediately from the fact that $D_{\ell d}(\Sigma_{yx}^1||\Sigma_{yx}^2) - D_{\ell d}(\Sigma_x^1||\Sigma_x^2)$ is the KL divergence between $p_1(y|x)$ and $p_2(y|x)$ for Gaussian distributions with zero mean and of covariances $\Sigma_{yx}^1, \Sigma_{yx}^2$ (see Eq. (15)) and $D_{KL}(p_1(y|x)||p_2(y|x)) \geq 0$. □

Property 4.2. *Let $x_1 \sim N(\mu_x^1, \Sigma_x^1)$ and $x_2 \sim N(\mu_x^2, \Sigma_x^2)$ be two input processes of full rank covariance matrices of size $p \times p$. For a common linear system, defined by a full rank matrix W of size $r \times p$ such that $y = Wx$, we have:*

$$D_{\ell d}(\Sigma_{xy}^1||\Sigma_{xy}^2) - D_{\ell d}(\Sigma_x^1||\Sigma_x^2) = 0$$

Proof. Denote $M = \left| \begin{array}{c} I_p \\ W \end{array} \right|$, where I_p denotes an identity matrix, we have $\Sigma_{xy}^i = W\Sigma_x^i W^T$, from which we have:

$$\begin{aligned} D_{\ell d}(\Sigma_{xy}^1||\Sigma_{xy}^2) - D_{\ell d}(\Sigma_x^1||\Sigma_x^2) &= \\ D_{\ell d}(M\Sigma_x^1 M^T||M\Sigma_x^2 M^T) - D_{\ell d}(\Sigma_x^1||\Sigma_x^2) &= 0, \end{aligned}$$

where we used Property 12 and Lemma 5 of [Kulis *et al.*, 2009], since $\mathrm{range}(\Sigma_{xy}^i) = p \leq r + p$. □

4.1 Power Test

Our aim here is to examine if our statistic is really suitable for quantifying the discrepancy between two conditional distributions. Motivated by [Zheng, 2000; Fan *et al.*, 2006], we generate four groups of data that have distinct conditional distributions. Specifically, in model (a), the dependent variable y is generated by $y = 1 + \sum_{i=1}^{p} x_i + \epsilon$, where ϵ denotes standard normal distribution. In model (b), $y = 1 + \sum_{i=1}^{p} x_i + \psi$, where ψ denotes the standard Logistic distribution. In model (c), $y = 1 + \sum_{i=1}^{p} \log x_i + \epsilon$. In model (d), $y = 1 + \sum_{i=1}^{p} \log x_i + \psi$. For each model, the input distribution is an isotropic Gaussian.

To evaluate the detection power of our statistic on any two models, we randomly generate 500 samples from each model which has the same dimensionality m on explanatory variable x. We then use Algorithm 1 ($P = 500, \eta = 0.1$) to test if our statistic can distinguish these two data sets. We repeat this procedure with 100 independent runs and use the percentage of success as the detection power. The conditional KL divergence (see Eq. (14)) estimated with an adaptive kNN estimator [Wang *et al.*, 2009] is implemented as a baseline competitor. Table 1 summarizes the power test results when $p = 3$ and $p = 30$. Although all methods perform good for $p = 3$, our test statistic is significantly more powerful than conditional KL divergence in high-dimensional space.

We also depict in Fig. 1 the power of our statistic in case I: model (a) against model (b) and case II: model (c) against model (d), with respect to different kernel widths. For case I (the first row), larger width is preferred. This is because the second-order information dominates in the linear model. For case II (the second row), we need smaller width to capture more higher order information in highly nonlinear model.

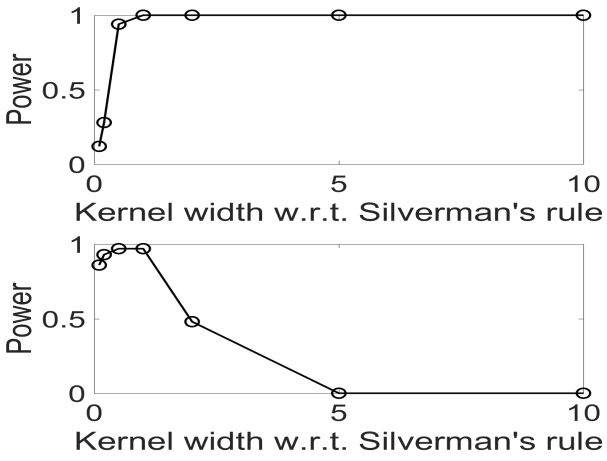


Figure 1: Power of our statistics with respect to the kernel width. The x-axis denotes the ratio of our used kernel width with respect to the one selected with Silverman's rule of thumb.

	Conditional KL				von Neumann (C)				LogDet (C)			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
$p = 3$												
(a)	0.09	1	1	1	0.04	1	1	1	0.06	1	1	1
(b)	1	0.07	1	1	1	0.07	1	0.96	1	0.09	1	0.92
(c)	1	1	0.12	1	1	1	0.13	0.97	1	1	0.12	0.91
(d)	1	1	1	0.07	1	0.96	0.97	0.09	1	0.94	0.97	0.07
$p = 30$												
(a)	0.12	0.67	1	1	0.04	1	1	1	0.10	1	1	1
(b)	0.60	0.14	1	1	1	0.10	1	1	1	0.10	1	1
(c)	1	1	0.09	0.34	1	1	0.11	0.67	1	1	0.06	0.60
(d)	1	1	0.34	0.07	1	1	0.58	0.14	1	1	0.50	0.13

Table 1: Power test for conditional KL divergence and our statistics implemented with von Neumann and LogDet divergences on centered correntropy matrix C

4.2 Relation to Previous Efforts

Gaussian Data

As mentioned earlier, the centered correntropy matrix C evaluated with a Gaussian kernel encloses all even higher order statistics of pairwise dimensions of data. If we replace C with its second-order counterpart (i.e., the covariance matrix Σ), we have:

$$D_{\varphi,B}(p_1(y|x)\|p_2(y|x)) = D_{\varphi,B}(\Sigma^1_{xy}\|\Sigma^2_{xy}) - D_{\varphi,B}(\Sigma^1_x\|\Sigma^2_x). \tag{12}$$

Taking Eq. (3) into Eq. (12), we obtain:

$$\begin{aligned} D_{\ell D}(p_1(y|x)\|p_2(y|x)) &= \mathrm{tr}((\Sigma^2_{xy})^{-1}\Sigma^1_{xy}) + \log\frac{|\Sigma^2_{xy}|}{|\Sigma^1_{xy}|} \\ &\quad -(p+1) - \mathrm{tr}((\Sigma^2_x)^{-1}\Sigma^1_x) - \log\frac{|\Sigma^2_x|}{|\Sigma^1_x|} + p. \end{aligned} \tag{13}$$

On the other hand, the KL divergence between two conditional distributions on a pair of random variables satisfies the following chain rule (see proof in Chapter 2 of [Cover and Thomas, 2012]):

$$\begin{aligned} D_{KL}(p_1(y|x)\|p_2(y|x)) &= D_{KL}(p_1(x,y)\|p_2(x,y)) \\ &\quad - D_{KL}(p_1(x)\|p_2(x)). \end{aligned} \tag{14}$$

Suppose the data is Gaussian distributed, i.e., $p_1(x) \sim \mathcal{N}(\mu^1_x, \Sigma^1_x)$, $p_2(x) \sim \mathcal{N}(\mu^2_x, \Sigma^2_x)$, $p_1(x,y) \sim \mathcal{N}(\mu^1_{xy}, \Sigma^1_{xy})$, $p_2(x,y) \sim \mathcal{N}(\mu^2_{xy}, \Sigma^2_{xy})$, Eq. (14) has a closed-form expression (see proof in [Duchi, 2007]):

$$\begin{aligned} &D_{KL}(p_1(y|x)\|p_2(y|x)) \\ &= \frac{1}{2}\{\mathrm{tr}((\Sigma^2_{xy})^{-1}\Sigma^1_{xy}) + \log_2\frac{|\Sigma^2_{xy}|}{|\Sigma^1_{xy}|} - (p+1) \\ &- \mathrm{tr}((\Sigma^2_x)^{-1}\Sigma^1_x) - \log\frac{|\Sigma^2_x|}{|\Sigma^1_x|} + p \\ &+ (\mu^2_{xy} - \mu^1_{xy})^T(\Sigma^2_{xy})^{-1}(\mu^2_{xy} - \mu^1_{xy}) \\ &- (\mu^2_x - \mu^1_x)^T(\Sigma^2_x)^{-1}(\mu^2_x - \mu^1_x)\}. \end{aligned} \tag{15}$$

Comparing Eq. (13) with Eq. (15), it is easy to find that our baseline variant reduces to the conditional KL divergence under Gaussian assumption. The only difference is that the conditional KL divergence contains a Mahalanobis Distance term on the mean, which can be interpreted as the first-order information of data.

Beyond Gaussian Data

Gaussian assumption is always over-optimistic. If we stick to the conditional KL divergence, the probability estimation becomes inevitable, which is notoriously difficult in high-dimensional space [Nagler and Czado, 2016]. By making use the correntropy function, our statistic avoids the estimation of the underlying distribution, but it explicitly incorporates the higher order information which was lost in Eq. (12).

5 Machine Learning Applications

We present three solid examples on machine learning applications to demonstrate the performance improvement in the state-of-the-art (SOTA) methodologies gained by our conditional divergence statistic.

5.1 Multitask Learning

Consider an input set $\mathcal{X}$ and an output set $\mathcal{Y}$ and for simplicity that $\mathcal{X} \in \mathbb{R}^p$, $\mathcal{Y} \in \mathbb{R}$. Tasks can be viewed as T functions $f_t, t = 1, \cdots, T$, to be learned from given data $\{(x_{ti}, y_{ti}) : i = 1, \cdots, N_t, t = 1, \cdots, T\} \subseteq \mathcal{X} \times \mathcal{Y}$, where N_t is the number of samples in the t-th task. These tasks may be viewed as drawn from an unknown joint distribution of tasks, which is the source of the bias that relates the tasks. Multitask learning is the paradigm that aims at improving the generalization performance of multiple prediction problems (tasks) by exploiting potential useful information between related tasks.

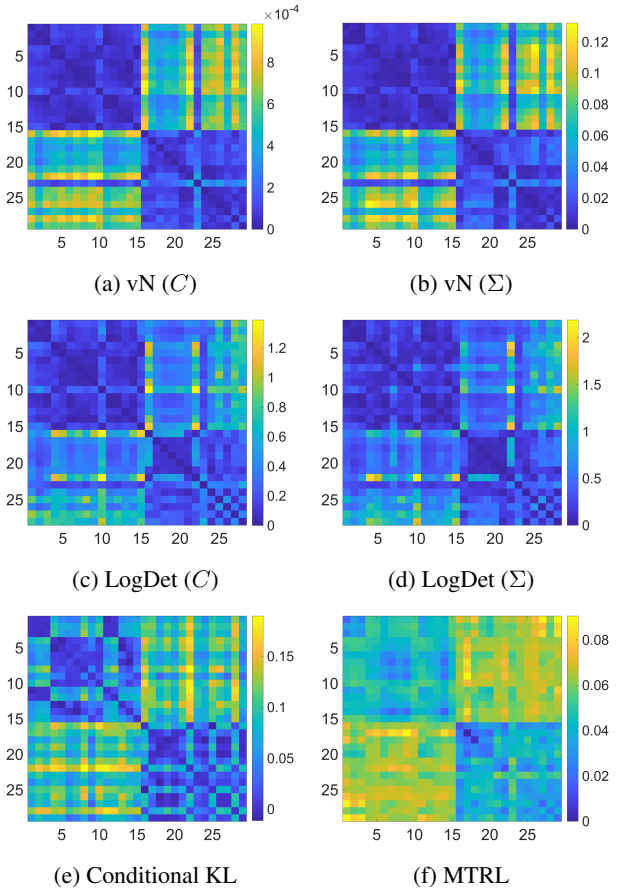
Visualizing Task-relatedness

We first test if our proposed statistic is able to reveal the relatedness among multiple tasks. To this end, we select data from 29 tasks that are collected from various landmine fields[3]. Each object in a given data set is represented by a

[3] http://www.ee.duke.edu/~lcarin/LandmineData.zip.

9-dimensional feature vector and the corresponding binary label (1 for landmine and 0 for clutter). The landmine detection problem is thus modeled as a binary classification problem.

Among these 29 data sets, 1-15 correspond to regions that are relatively highly foliated and 16-29 correspond to regions that are bare earth or desert. We measure the task-relatedness with the conditional divergence $D_{\varphi,B}(p_1(y|x) : p_2(y|x))$ and demonstrate their pairwise relationships in a task-relatedness matrix. Thus we expect that there are approximately two clusters in the task-relatedness matrix corresponding to two classes of ground surface condition. Fig. 2 demonstrates the visualization results generated by our proposed statistic and the conditional KL divergence estimated with an adaptive kNN estimator [Wang *et al.*, 2009]. We also compare our method with MTRL [Zhang and Yeung, 2010], a widely used objective to learn multiple tasks and their relationships in a convex manner. Our vN (C) and vN (Σ) clearly demonstrate two clusters of tasks. By contrast, both MTRL and conditional KL divergence indicate a few misleading relations (i.e., tasks in the same group surface condition have high divergence values).



(a) vN (C) (b) vN (Σ)

(c) LogDet (C) (d) LogDet (Σ)

(e) Conditional KL (f) MTRL

Figure 2: Visualize task-relatedness in landmine detection data set. "vN" refers to von Neumann, C denotes centered correntropy matrix, and Σ denotes covariance matrix. For example, vN (C) denotes our statistic implemented with von Neumann divergence on centered correntropy matrix.

Multitask Learning over Graph Structure

In the second example, we demonstrate how our statistic improves the performance of the Convex Clustering Multi-Task Learning (CCMTL) [He *et al.*, 2019], a SOTA method for learning multiple regression tasks. The learning objective of CCMTL is given by:

$$\min_{W} \frac{1}{2}\sum_{t=1}^{T} ||w_t^T x_t - y_t||_2^2 + \frac{\lambda}{2}\sum_{i,j\in G} ||w_i - w_j||_2, \quad (16)$$

where $W = [w_1^T, w_2^T, \cdots, w_T^T] \in \mathbb{R}^{T\times p}$ is the weight matrix constitutes of the learned linear regression coefficients in each task, G is the graph of relations over all tasks (if two tasks are related, then there is an edge to connect them), and λ is a regularization parameter.

CCMTL requires an initialization of the graph structure G_0. In the original paper, the authors set G_0 as a kNN graph on the prediction models learned independently for each task. In this sense, the task-relatedness between two tasks is modeled as the ℓ_2 distance of their independently learned linear regression coefficients.

We replace the naïve ℓ_2 distance with our proposed statistic to reconstruct the initial kNN graph for CCMTL and test its performance on a real-world Parkinson's disease data set [Tsanas *et al.*, 2009]. We have 42 tasks and $5,875$ observations, where each task and observation correspond to a prediction of the symptom score (motor UPDRS) for a patient and a record of a patient, respectively. Fig. 3 depicts the prediction root mean square errors (RMSE) under different train/test ratios. To highlight the superiority of CCMTL and our improvement, we also compare it with MSSL [Gonçalves *et al.*, 2016], a recently proposed alternative objective to MTRL that learns multiple tasks and their relationships with a Gaussian graphical model. The performance of MTRL is relatively poor, and thus omitted here. Obviously, our statistic improves the performance of CCMTL with a large margin, especially when training samples are limited. Note that, we did not observe performance difference between centered correntropy matrix C and covariance matrix Σ. This is probably because the higher order information is weak or because the Silverman's rule of thumb is not optimal to tune kernel width here.

5.2 Concept Drift Detection

One important assumption underlying common classification models is the stationarity of the training data. However, in real-world data stream applications, the joint distribution $p(x, y)$ between the predictor x and response variable y is not stationary but drifting over time. Concept drift detection approaches aim to detect such drifts and adapt the model so as to mitigate the deteriorating classification performance over time.

Formally, the concept drift between time instants t_0 and t_1 can be defined as $p_{t_1}(x, y) \neq p_{t_2}(x, y)$ [Gama *et al.*, 2014]. From a Bayesian perspective, concept drifts can manifest two fundamental forms of changes: 1) a change in the marginal probability $p_t(x)$ or $p_t(y)$; and 2) a change in the posterior probability $p_t(y|x)$. Although any two-sample test (e.g., [Gretton *et al.*, 2012b]) on $p_t(x)$ or $p_t(y)$ is an option to

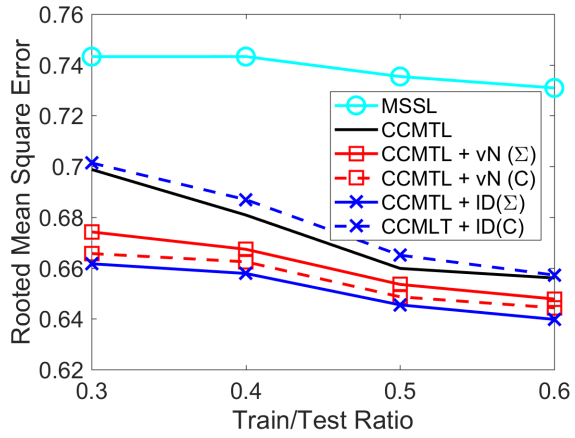


Figure 3: Root mean square errors (RMSE) of different multitask learning methodologies with respect to varying train/test ratios on the Parkinson's data set.

detect concept drift, existing studies tend to prioritize detecting posterior distribution change, because it clearly indicates the optimal decision rule [Dries and Rückert, 2009].

Error-based methods constitute a main category of existing concept drift detectors. These methods keep track of the online classification error or error-related metrics of a baseline classifier. A significant increase or decrease of the monitored statistics may suggests a possible concept drift. Unfortunately, the performance of existing manually designed error-related statistics depends on the baseline classifier, which makes them either perform poorly across different data distributions or difficult to be extended to the multi-class classification scenario.

Different from prevailing error-based methods, our statistic operates directly on the conditional divergence $D_{\varphi,B}(p_{t_1}(y|x) : p_{t_2}(y|x))$, which makes it possible to fundamentally solve the problem of concept drift detection (without any classifier). To this end, we test the conditional distribution divergence in two consecutive sliding windows (using Algorithm 1) at each time instant t. A reject of the null hypothesis indicates the existence of a concept drift. Note that the same permutation test procedure has been theoretically and practically investigated in PERM [Harel *et al.*, 2014], in which the authors use classification error as the test statistic. Interested readers can refer to [Harel *et al.*, 2014; Yu and Abraham, 2017] for more details on concept drift detection with permutation test.

We evaluate the performance of our method against four SOTA error-based concept drift detectors (i.e., DDM [Gama *et al.*, 2004], EDDM [Baena-Garcıa *et al.*, 2006], HDDM [Frías-Blanco *et al.*, 2014], and PERM) on two real-world data streams, namely the Digits08 [Sethi and Kantardzic, 2017] and the AbruptInsects [dos Reis *et al.*, 2016]. Among the selected competitors, HDDM represents one of the best-performing detectors, whereas PERM is the most similar one to ours. The concept drift detection results and the stream classification results over 30 independent runs are summarized in Table 2. Our method always enjoys the highest recall

Method	Precision	Recall	Delay	Accuracy (%)
DDM	0.49	0.50	50	89.22
EDDM	0.69	0.82	230	92.60
HDDM	**1**	0.83	133	97.47
PERM	0.81	0.88	99	**97.81**
vN (Σ)	0.77	**1**	**43**	92.82
LD (Σ)	0.83	**1**	113	93.43
vN (C)	0.80	**1**	60	90.07
LD (C)	0.77	**1**	53	92.23

(a) Digits08

Method	Precision	Recall	Delay	Accuracy (%)
DDM	0.83	0.83	25	67.47
EDDM	0.47	**1**	46	63.23
HDDM	**1**	**1**	15	76.94
PERM	0.50	**1**	31	71.98
vN (Σ)	0.50	**1**	**11**	72.11
LD (Σ)	0.47	**1**	21	75.94
vN (C)	0.50	**1**	**11**	72.52
LD (C)	0.49	**1**	23	**77.02**

(b) Abrupt Insect

Table 2: Quantitative metrics on real-world data sets. The Precision, Recall and Delay denote the concept drift detection precision value, recall value and detection delay, whereas the Accuracy denotes the classification accuracy in the testing set (%).

values and the shortest detection delay. Note that, the classification accuracy is not as high as we expected. One possible reason is that the short detection delay makes us do not have sufficient number of samples to retrain the classifier.

5.3 Feature Selection

Our final application is information-theoretic feature selection. Given a set of variables $S = \{X_1, X_2, \cdots, X_n\}$, feature selection refers to seeking a small subset of informative variables $S^\star \subset S$, such that $S^\star$ contains the most relevant yet least redundant information about a desired variable Y. From an information-theoretic perspective, this amounts to maximize the mutual information term $\mathbf{I}(y; S^\star)$. Suppose we are now given a set of "useless" features $\tilde{S}$ that has the same size as $S^\star$ but has no predictive power to y, Eq. (17) suggests that instead of maximizing $\mathbf{I}(y; S^\star)$, one can resort to maximize $D_{KL}(P(y|S^\star)\|P(y|\tilde{S}))$ as an alternative.

$$\begin{aligned}\mathbf{I}(y; S^\star) &= \iint P(y, S^\star) \log \frac{P(y, S^\star)}{P(y)P(S^\star)} \\ &= \iint \left(P(y|S^\star) \log \frac{P(y|S^\star)}{P(y)} \right) P(S^\star) \\ &= \mathbb{E}_S[D_{KL}(P(y|S^\star)||P(y))] \\ &= \mathbb{E}_S[D_{KL}(P(y|S^\star)||P(y|\tilde{S}))],\end{aligned} \tag{17}$$

the last line is by our assumption that $\tilde{S}$ has no predictive power to y such that $P(y|\tilde{S}) = P(y)$.

Motivated by the generic equivalence theorem between K-L divergence and the Bregman divergence [Banerjee *et al.*, 2005], we optimize our statistic in a greedy forward search

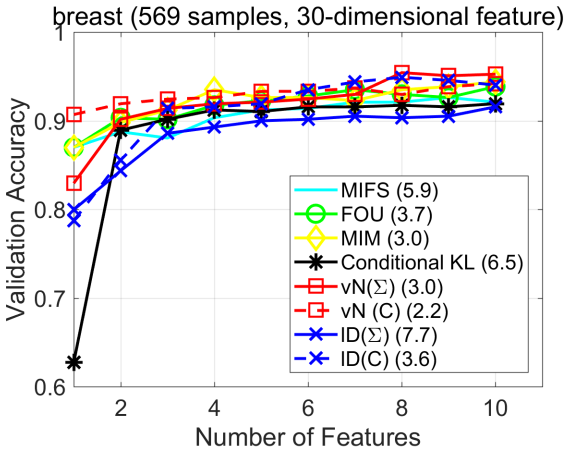
manner (i.e., adding the best feature at each round) and generate "useless" feature set $\tilde{S}$ by randomly permutating $S^\star$ as conducted in [François *et al.*, 2007].

We perform feature selection on two benchmark data sets [Brown *et al.*, 2012]. For both data sets, we select 10 features and use the linear Support Vector Machine (SVM) as the baseline classifier. We compare our method with three popular information-theoretic feature selection methods that target maximizing $\mathbf{I}(y; S^\star)$, namely MIFS [Battiti, 1994], FOU [Brown *et al.*, 2012], and MIM [Lewis, 1992]. The classification accuracies with respect to different number of selected features (averaged over 10 fold cross-validation) are presented in Fig. 4. As can be seen, methods on maximizing conditional divergence always achieve comparable performances to those mutual information based competitors. This result confirms Eq. (17). If we look deeper, it is easy to see that our statistic implemented with von Neumann divergence achieves the best performance on both data sets. Moreover, by incorporating higher order information, centered correntropy performs better than its covariance based counterpart.

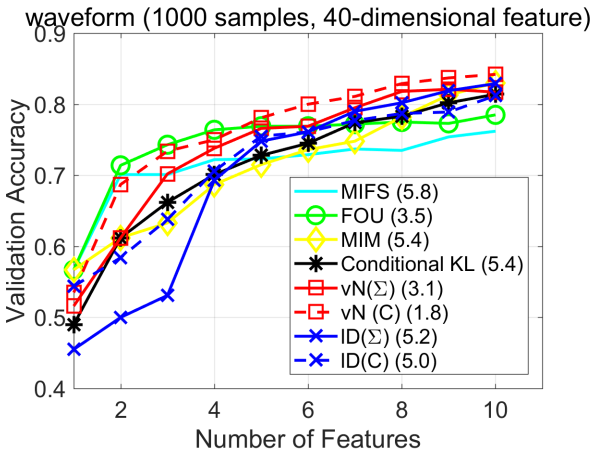
6 Conclusions

We propose a simple statistic to quantify conditional divergence. We also establish its connections to prior art and illustrate some of its fundamental properties. A natural extension to incorporate higher order information of data is also presented. Three solid examples suggest that our statistic can offer a remarkable performance gain to SOTA learning algorithms (e.g., CCMTL and PERM). Moreover, our statistic enables the development of alternative solutions to classical machine learning problems (e.g., classifier-free concept drift detection and feature selection by maximizing conditional divergence) in a fundamental manner.

Future work is twofold. We will investigate more fundamental properties of our statistic. We will also apply our statistic to more challenging problems, such as the continual learning [Kirkpatrick *et al.*, 2017] which also requires the knowledge of task-relatedness.



(a) breast



(b) waveform

Figure 4: Validation accuracy on (a) breast and (b) waveform data sets. The number of samples and the feature dimensionality for each data set are listed in the title. The value beside each method in the legend indicates the average rank in that data set.

References

[Anderson *et al.*, 1994] Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

[Baena-Garcıa *et al.*, 2006] Manuel Baena-Garcıa, José del Campo-Ávila, Raúl Fidalgo, Albert Bifet, R Gavalda, and R Morales-Bueno. Early drift detection method. In *Fourth international workshop on knowledge discovery from data streams*, volume 6, pages 77–86, 2006.

[Banerjee *et al.*, 2005] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *JMLR*, 6(Oct):1705–1749, 2005.

[Battiti, 1994] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE TNNLS*, 5(4):537–550, 1994.

[Brown *et al.*, 2012] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *JMLR*, 13(Jan):27–66, 2012.

[Chen *et al.*, 2017] Badong Chen, , et al. Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering. *IEEE TSP*, 65(11):2888–2901, 2017.

[Cover and Thomas, 2012] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[dos Reis *et al.*, 2016] Denis Moreira dos Reis, Peter Flach, Stan Matwin, and Gustavo Batista. Fast unsupervised online drift detection using incremental kolmogorov-smirnov test. In *SIGKDD*, pages 1545–1554. ACM, 2016.

[Dries and Rückert, 2009] Anton Dries and Ulrich Rückert. Adaptive concept drift detection. In *SDM*, pages 235–246. SIAM, 2009.

[Duchi, 2007] John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3, 2007.

[Fan *et al.*, 2006] Yanqin Fan, Qi Li, and Insik Min. A nonparametric bootstrap test of conditional distributions. *Econometric Theory*, 22(4):587–613, 2006.

[François *et al.*, 2007] Damien François, Fabrice Rossi, Vincent Wertz, and Michel Verleysen. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing*, 70(7-9):1276–1288, 2007.

[Frías-Blanco *et al.*, 2014] Isvani Frías-Blanco, José del Campo-Ávila, Gonzalo Ramos-Jimenez, Rafael Morales-Bueno, Agustín Ortiz-Díaz, and Yailé Caballero-Mota. Online and non-parametric drift detection methods based on hoeffding bounds. *IEEE TKDE*, 27(3):810–823, 2014.

[Gama *et al.*, 2004] Joao Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Brazilian symposium on artificial intelligence*, pages 286–295. Springer, 2004.

[Gama *et al.*, 2014] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.

[Gonçalves *et al.*, 2016] André R Gonçalves, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning with gaussian copula models. *JMLR*, 17(1):1205–1234, 2016.

[Gretton *et al.*, 2012a] Arthur Gretton, , et al. Optimal kernel choice for large-scale two-sample tests. In *NeurIPS*, pages 1205–1213, 2012.

[Gretton *et al.*, 2012b] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(Mar):723–773, 2012.

[Harel *et al.*, 2014] Maayan Harel, Shie Mannor, Ran El-Yaniv, and Koby Crammer. Concept drift detection through resampling. In *ICML*, pages 1009–1017, 2014.

[He *et al.*, 2019] Xiao He, Francesco Alesiani, and Ammar Shaker. Efficient and scalable multi-task regression on massive number of tasks. In *AAAI*, volume 33, pages 3763–3770, 2019.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, , et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

[Kulis *et al.*, 2009] Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-rank kernel learning with bregman matrix divergences. *JMLR*, 10(Feb):341–376, 2009.

[Lee and Park, 2006] Young Kyung Lee and Byeong U Park. Estimation of kullback–leibler divergence by local likelihood. *Annals of the Institute of Statistical Mathematics*, 58(2):327–340, 2006.

[Lewis, 1992] David D Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics, 1992.

[Liu *et al.*, 2007] Weifeng Liu, Puskal P Pokharel, and José C Príncipe. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE TSP*, 55(11):5286–5298, 2007.

[Nagler and Czado, 2016] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[Nielsen and Chuang, 2011] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th edition, 2011.

[Pardo, 2005] Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2005.

[Rao *et al.*, 2011] Murali Rao, Sohan Seth, Jianwu Xu, Yunmei Chen, Hemant Tagare, and José C Príncipe. A test of independence based on a generalized correlation function. *Signal Processing*, 91(1):15–27, 2011.

[Ren *et al.*, 2016] Yong Ren, Jun Zhu, Jialian Li, and Yucen Luo. Conditional generative moment-matching networks. In *NeurIPS*, pages 2928–2936, 2016.

[Santamaría *et al.*, 2006] Ignacio Santamaría, Puskal P Pokharel, and José Carlos Principe. Generalized correlation function: definition, properties, and application to blind equalization. *IEEE TSP*, 54(6):2187–2197, 2006.

[Sethi and Kantardzic, 2017] Tegjyot Singh Sethi and Mehmed Kantardzic. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications*, 82:77–99, 2017.

[Silverman, 1986] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[Tsanas *et al.*, 2009] Athanasios Tsanas, Max A Little, Patrick E McSharry, and Lorraine O Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE TBME*, 57(4):884–893, 2009.

[Wang *et al.*, 2009] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE TIT*, 55(5):2392–2405, 2009.

[Yu and Abraham, 2017] Shujian Yu and Zubin Abraham. Concept drift detection with hierarchical hypothesis testing. In *SDM*, pages 768–776. SIAM, 2017.

[Zhang and Yeung, 2010] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *UAI*, pages 733–742, 2010.

[Zheng, 2000] John Xu Zheng. A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5):667–691, 2000.