The Perturbed Variation

Maayan Harel
Department of Electrical Engineering
Technion, Haifa, Israel
maayanga@tx.technion.ac.il

Shie Mannor
Department of Electrical Engineering
Technion, Haifa, Israel
shie@ee.technion.ac.il

Abstract

We introduce a new discrepancy score between two distributions that gives an indication on their *similarity*. While much research has been done to determine if two samples come from exactly the same distribution, much less research considered the problem of determining if two finite samples come from similar distributions. The new score gives an intuitive interpretation of similarity; it optimally perturbs the distributions so that they best fit each other. The score is defined between distributions, and can be efficiently estimated from samples. We provide convergence bounds of the estimated score, and develop hypothesis testing procedures that test if two data sets come from similar distributions. The statistical power of this procedures is presented in simulations. We also compare the score's capacity to detect similarity with that of other known measures on real data.

1 Introduction

The question of similarity between two sets of examples is common to many fields, including statistics, data mining, machine learning and computer vision. For example, in machine learning, a standard assumption is that the training and test data are generated from the same distribution. However, in some scenarios, such as Domain Adaptation (DA), this is not the case and the distributions are only assumed similar. It is quite intuitive to denote when two inputs are similar in nature, yet the following question remains open: given two sets of examples, how do we test whether or not they were generated by similar distributions? The main focus of this work is providing a similarity score and a corresponding statistical procedure that gives one possible answer to this question.

Discrepancy between distributions has been studied for decades, and a wide variety of distance scores have been proposed. However, not all proposed scores can be used for testing similarity. The main difficulty is that most scores have not been designed for statistical testing of similarity but equality, known as the Two-Sample Problem (TSP). Formally, let P and Q be the generating distributions of the data; the TSP tests the null hypothesis $H_0 : P = Q$ against the general alternative $H_1 : P \neq Q$. This is one of the classical problems in statistics. However, sometimes, like in DA, the interesting question is with regards to similarity rather than equality. By design, most equality tests may not be transformed to test similarity; see Section 3 for a review of representative works.

In this work, we quantify similarity using a new score, the Perturbed Variation (PV). We propose that similarity is related to some predefined value of permitted variations. Consider the gait of two male subjects as an example. If their physical characteristics are similar, we expect their walk to be similar, and thus assume the examples representing the two are from similar distributions. This intuition applies when the distribution of our measurements only endures small changes for people with similar characteristics. Put more generally, similarity depends on what "small changes" are in a given application, and implies that similarity is domain specific. The PV, as hinted by its name, measures the discrepancy between two distributions while allowing for some perturbation of each distribution; that is, it allows small differences between the distributions. What accounts for small differences is a parameter of the PV, and may be defined by the user with regard to a specific domain.

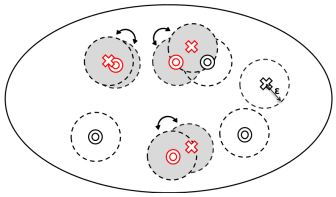


Figure 1: X and O identify samples from two distributions, doted circles denote allowed perturbations. Samples marked in red are matched with neighbors, while the unmatched samples indicate the PV discrepancy.

Figure 1 illustrates the PV. Note that, like perceptual similarity, the PV turns a blind eye to variations of some rate.

2 The Perturbed Variation

The PV on continuous distributions is defined as follows:

Definition 1. *Let P and Q be two distributions on a Banach space $\mathcal{X}$, and let $M(P,Q)$ be the set of all joint distributions on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q. The PV, with respect to a distance function $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and ϵ, is defined by*

$$PV(P, Q, \epsilon, \mathbf{d}) \doteq \inf_{\mu \in M(P,Q)} \mathbb{P}_\mu[d(X, Y) > \epsilon], \tag{1}$$

over all pairs $(X, Y) \sim \mu$, such that the marginal of X is P and the marginal of Y is Q.

Put into words, Equation (1) defines the joint distribution μ that couples the two distributions such that the probability of the event of a pair $(X, Y) \sim \mu$ being within a distance grater than ϵ is minimized.

The solution to (1) is a special case of the classical mass transport problem of Monge [1] and its version by Kantorovich: $\inf_{\mu \in M(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\mu(x, y)$, where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable cost function. When c is a metric, the problem describes the 1^{st} Wasserstein metric. Problem (1) may be rephrased as the optimal mass transport problem with the cost function $c(x, y) = 1_{[d(x,y)>\epsilon]}$, and may be rewritten as $\inf_\mu \iint 1_{[d(x,y)>\epsilon]} \mu(y|x) dy \, P(x) dx$. The probability $\mu(y|x)$ defines the transportation plan of x to y. The PV optimal transportation plan is obtained by perturbing the mass of each point x in its ϵ neighborhood so that it redistributes to the distribution of Q. These small perturbations do not add any cost, while transportation of mass to further areas is equally costly. Note that when $P = Q$ the PV is zero as the optimal plan is simply the identity mapping. Due to its cost function, the PV it is not a metric, as it is symmetric but does not comply with the triangle inequality and may be zero for distributions $P \neq Q$. Despite this limitation, this cost function fully quantifies the intuition that small variations should not be penalized when similarity is considered. In this sense, similarity is not unique by definition, as more than one distribution can be similar to a reference distribution.

The PV is also closely related to the Total Variation distance (TV) that may be written, using a coupling characterization, as $TV(P, Q) = \inf_{\mu \in M(P,Q)} \mathbb{P}_\mu [X \neq Y]$ [2]. This formulation argues that any transportation plan, even to a close neighbor, is costly. Due to this property, the TV is known to be an overly sensitive measure that overestimates the distance between distributions. For example, consider two distributions defined by the dirac delta functions $\delta(a)$ and $\delta(a + \epsilon)$. For any ϵ, the TV between the two distributions is 1, while they are intuitively similar. The PV resolves this problem by adding perturbations, and therefore is a natural extension of the TV. Notice, however, that the ϵ used to compute the PV need not be infinitesimal, and is defined by the user.

The PV can be seen as a conciliatory between the Wasserstein distance and the TV. As explained, it relaxes the sensitivity of the TV; however, it does not "over optimize" the transportation plan. Specifically, distances larger than the allowed perturbation are discarded. This aspect also contributes to the efficiency of estimation of the PV from samples; see Section 2.2.

$PV(\mu_1, \mu_2, \epsilon) = \frac{1}{2}$:

$a_1 = 0, a_2 = 1, a_3 = 2, a_4 = 2.1$
$w_1 = w_2 = \frac{1}{4}, w_3 = w_4 = 0$
$v_4 = \frac{1}{2}, v_1 = v_2 = v_3 = 0$

$$Z = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

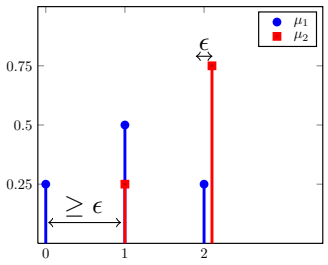


Figure 2.1: Illustration of the PV score between discrete distributions.

2.1 The Perturbed Variation on Discrete Distributions

It can be shown that for two discrete distributions Problem (1) is equivalent to the following problem.

Definition 2. *Let μ_1 and μ_2 be two discrete distributions on the unified support $\{a_1, ..., a_N\}$. Define the neighborhood of a_i as $ng(a_i, \epsilon) = \{z\,; d(z, a_i) \leq \epsilon\}$. The PV$(\mu_1, \mu_2, \epsilon, \mathbf{d})$ between the two distributions is:*

$$\min_{w_i \geq 0, v_i \geq 0, Z_{ij} \geq 0} \frac{1}{2}\sum_{i=1}^{N} w_i + \frac{1}{2}\sum_{j=1}^{N} v_j \tag{2}$$

$$\text{s.t.} \sum_{a_j \in ng(a_i, \epsilon)} Z_{ij} + w_i = \mu_1(a_i),\ \forall i$$

$$\sum_{a_i \in ng(a_j, \epsilon)} Z_{ij} + v_j = \mu_2(a_j),\ \forall j$$

$$Z_{ij} = 0\,, \quad \forall (i,j) \notin ng(a_i, \epsilon).$$

Each row in the matrix $Z \in \mathbb{R}^{N \times N}$ corresponds to a point mass in μ_1, and each column to a point mass in μ_2. For each i, $Z(i,:)$ is zero in columns corresponding to non neighboring elements, and non-zero only for columns j for which transportation between $\mu_2(a_j) \to \mu_1(a_i)$ is performed. The discrepancies between the distributions are depicted by the scalars w_i and v_i that count the "leftover" mass in $\mu_1(a_i)$ and $\mu_2(a_j)$. The objective is to minimize these discrepancies, therefore matrix Z describes the optimal transportation plan constrained to ϵ-perturbations. An example of an optimal plan is presented in Figure 2.1.

2.2 Estimation of the Perturbed Variation

Typically, we are given samples from which we would like to estimate the PV. Given two samples $S_1 = \{x_1, ..., x_n\}$ and $S_2 = \{y_1, ..., y_m\}$, generated by distributions P and Q respectively, $\widehat{\text{PV}}(S_1, S_2, \epsilon, d)$ is:

$$\min_{w_i \geq 0, v_i \geq 0, Z_{ij} \geq 0} \frac{1}{2n}\sum_{i=1}^{n} w_i + \frac{1}{2m}\sum_{j=1}^{m} v_j \tag{3}$$

$$\text{s.t.} \sum_{y_j \in \text{ng}(x_i, \epsilon)} Z_{ij} + w_i = 1, \quad \sum_{x_i \in \text{ng}(y_j, \epsilon)} Z_{ij} + v_j = 1, \quad \forall i, j$$

$$Z_{ij} = 0\,, \qquad \forall (i,j) \notin \text{ng}(x_i, \epsilon),$$

where $Z \in \mathbb{R}^{n \times m}$. When $n = m$, the optimization in (3) is identical to (2), as in this case the samples define a discrete distribution. However, when $n \neq m$ Problem (3) also accounts for the difference in the size of the two samples.

Problem (3) is a linear program with constraints that may be written as a totally unimodular matrix. It follows that one of the optimal solutions of (3) is integral [3]; that is, the mass of each sample is transferred as a whole. This solution may be found by solving the optimal assignment on an appropriate bipartite graph [3]. Let $G = (V = (A, B), E)$ define this graph, with $A = \{x_i, w_i\,; i = 1, ..., n\}$ and $B = \{y_j, v_j\,; j = 1, ..., m\}$ as its bipartite partition. The vertices $x_i \in A$ are linked

Algorithm 1 *Compute* $\widehat{PV}(S_1, S_2, \epsilon, \mathbf{d})$

Input: $S_1 = \{x_1, ..., x_n\}$ and $S_2 = \{y_1, ..., y_m\}$, ϵ rate, and distance measure $\mathbf{d}$.
1. Define $\hat{G} = (\hat{V} = (\hat{A}, \hat{B}), \hat{E})$: $\hat{A} = \{x_i \in S_1\}$, $\hat{B} = \{y_j \in S_2\}$,
   Connect an edge $e_{ij} \in \hat{E}$ if $d(x_i, y_j) \leq \epsilon$.
2. Compute the maximum matching on $\hat{G}$.
3. Define S_w and S_v as number of unmatched edges in sets S_1 and S_2 respectively.

Output: $\widehat{PV}(S_1, S_2, \epsilon, d) = \frac{1}{2}(\frac{S_w}{n} + \frac{S_v}{m})$.

with edge weight zero to $y_j \in \text{ng}(x_i)$ and with weight ∞ to $y_j \notin \text{ng}(x_i)$. In addition, every vertex x_i (y_j) is linked with weight 1 to w_i (v_j). To make the graph complete, assign zero cost edges between all vertices x_i and w_k for $k \neq i$ (and vertices y_j and v_k for $k \neq j$).

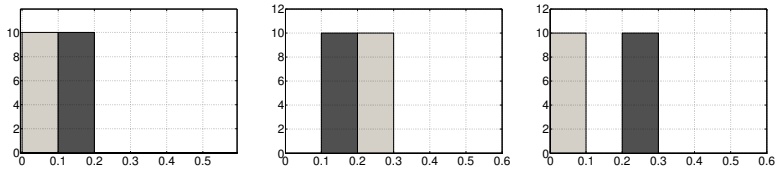
We note that the Earth Mover Distance (EMD) [4], a sampled version of the transportation problem, is also formulated by a linear program that may be solved by optimal assignment. For the EMD and other typical assignment problems, the computational complexity is more demanding, for example using the Hungarian algorithm it has an $O(N^3)$ complexity, where $N = n+m$ is the number of vertices [5]. Contrarily, graph G, which describes $\widehat{\text{PV}}$, is a simple bipartite graph for which maximum cardinality matching, a much simpler problem, can be applied to find the optimal assignment. To find the optimal assignment, first solve the maximum matching on the partial graph between vertices x_i, y_j that have zero weight edges (corresponding to neighboring vertices). Then, assign vertices x_i and y_j for whom a match was not found with w_i and v_j respectively; see Algorithm 1 and Figure 1 for an illustration of a matching. It is easy to see that the solution obtained solves the assignment problem associated with $\widehat{\text{PV}}$.

The complexity of Algorithm 1 amounts to the complexity of the maximal matching step and of setting up the graph, i.e., additional $O(nm)$ complexity of computing distances between all points. Let k be the average number of neighbors of a sample, then the average number of edges in the bipartite graph $\hat{G}$ is $|\hat{E}| = n \times k$. The maximal cardinality matching of this graph is obtained in $O(kn\sqrt{(n+m)})$ steps, in the worst case [5].

3 Related Work

Many scores have been defined for testing discrepancy between distributions. We focus on representative works for nonparametric tests that are most related to our work. First, we consider statistics for the Two Sample Problem (TSP), i.e., equality testing, that are based on the asymptotic distribution of the statistic conditioned on the equality. Among these tests is the well known Kolmogorov-Smirnov test (for one dimensional distributions), and its generalization to higher dimensions by minimal spanning trees [6]. A different statistic is defined by the portion of k-nearest neighbors of each sample that belongs to different distributions; larger portions mean the distributions are closer [7]. These scores are well known in the statistical literature but cannot be easily changed to test similarity, as their analysis relies on testing equality.

As discussed earlier, the 1^{st} Wasserstein metric and the TV metric have some relation to the PV. The EMD and histogram based L_1 distance are the sample based estimates of these metrics respectively. In both cases, the distance is not estimated directly on the samples, but on a higher level partition of the space: histogram bins or signatures (cluster centers). It is impractical to use the EMD to estimate the Wasserstein metric between the *continuous* distributions, as convergence would require the number of bins to be exponentially dependent on the dimension. As a result, it is commonly used to rate distances and not for statistical testing. Contrarily, the PV is estimated directly on the samples and converges to its value between the underlying continuous distributions. We note that after a good choice of signatures, the EMD captures perceptual similarity, similar to that of the PV. It is possible to consider the PV as a refinement of the EMD notion of similarity; instead of clustering the data to signatures and moving the signatures, it perturbs each sample. In this manner, it captures a finer notion of similarity better suited for statistical testing.



(a) PV($\epsilon = 0.1) = 0$ (b) PV($\epsilon = 0.1) = 0$ (c) PV($\epsilon = 0.1) = 1$

Figure 2: Two distributions on $\mathbb{R}$: The PV captures the perceptual similarity of (a),(b) against the disimilarity in (c). The $L_1^1 = 1$ on $I_1 = \{(0, 0.1), (0.1, 0.2), ...\}$ for all cases; on $I_2 = \{(0, 0.2), (0.2, 0.4), ...\}$ it is $L_1^2(P_a, Q_a) = 0, L_1^2(P_b, Q_b) = 1, L_1^2(P_c, Q_c) = 1$; and on $I_3 = \{(0, 0.3), (0.3, 0.6), ...\}$ it is $L_1^3(P_a, Q_a) = 0, L_1^3(P_b, Q_b) = 0, L_1^3(P_c, Q_c) = 0$.

The partition of the support to bins allows some relaxation of the TV notion. Therefore, instead of the TV, it may be interesting to consider the L_1 as a similarity distance on the measures after discretization. The example in Figure (2) shows that this relaxation is quite rigid and that there is no single partition that captures the perceptual similarity. In general, the problem would remain even if bins with varying width were permitted. Namely, the problem is the choice of a single partition to measure similarity of a reference distribution to multiple distributions, while choosing multiple partitions would make the distances incomparable. Also note that defining a "good" partition is a difficult task, which is exasperated in higher dimensions.

The last group of statistics are scores established in machine learning: the d_A distance presented by Kifer et al. that is based on the maximum discrepancy on a chosen subset of the support [8], and Maximum Mean Discrepancy (MMD) by Gretton et al., which define discrepancy after embeddings the distributions to a Reproducing Kernel Hilbert Space (RKHS)[9]. These scores have corresponding statistical tests for the TSP; however, since their analysis is based on finite convergence bounds, in principle they may be modified to test similarity. The d_A captures some intuitive notion of similarity, however, to our knowledge, it is not known how to compute it for a general subset class [1]. The MMD captures the distance between the samples in some RKHS. The MMD may be used to define a similarity test, yet this would require defining two parameters, σ and the similarity rate, whose dependency is not intuitive. Namely, for any similarity rate the result of the test is highly dependent on the choice of σ, but it is not clear how it should be made. Contrarily, the PV's parameter ϵ is related to the data's input domain and may be chosen accordingly.

4 Analysis

We present sample rate convergence analysis of the PV. The proofs of the theorems are provided in the supplementary material. When no clarity is lost, we omit $\mathbf{d}$ from the notation. Our main theorem is stated as follows:

Theorem 3. *Suppose we are given two i.i.d. samples $S_1 = \{x_1, ..., x_n\} \in \mathbb{R}^d$ and $S_2 = \{y_1, ..., y_m\} \in \mathbb{R}^d$ generated by distributions P and Q, respectively. Let the ground distance be $\mathbf{d} = \|\cdot\|_\infty$ and let $\mathcal{N}(\epsilon)$ be the cardinality of a disjoint cover of the distributions' support. Then, for any $\delta \in (0, 1)$, $N = \min(n, m)$, and $\eta = \sqrt{\frac{2(\log(2(2^{\mathcal{N}(\epsilon)} - 2)) + \log(1/\delta))}{N}}$ we have that*

$$\mathbb{P}\left(\left|\widehat{PV}(S_1, S_2, \epsilon) - PV(P, Q, \epsilon)\right| \leq \eta\right) \geq 1 - \delta.$$

The theorem is defined using $\|\cdot\|_\infty$, but can be rewritten for other metrics (with a slight change of constants). The proof of the theorem exploits the form of the optimization Problem 3. We use the bound of Theorem 3 construct hypothesis tests. A weakness of this bound is its strong dependency on the dimension. Specifically, it is dependent on $\mathcal{N}(\epsilon)$, which for $\|\cdot\|_\infty$ is $O((1/\epsilon)^d)$: the number of disjoint boxes of volume ϵ^d that cover the support. Unfortunately, this convergence rate is inherent; namely, without making any further assumptions on the distribution, this rate is unavoidable and is an instance of the "curse of dimensionality". In the following theorem, we present a lower bound on the convergence rate.

[1] Most work with the d_A has been with the subset of characteristic functions, and approximated by the error of a classifier.

Theorem 4. *Let $P = Q$ be the uniform distribution on $\mathbb{S}^{d-1}$, a unit $(d-1)$–dimensional hypersphere. Let $S_1 = \{x_1, ..., x_N\} \sim P$ and $S_2 = \{y_1, ..., y_N\} \sim Q$ be two i.i.d. samples. For any $\epsilon, \epsilon', \delta \in (0,1)$, $0 \leq \eta < 2/3$ and sample size $\frac{\log(1/\delta)}{2(1-3\eta/2)^2} \leq N \leq \frac{\eta}{2} e^{d(1-\frac{\epsilon^2}{2})/2}$, we have $PV(P, Q, \epsilon') = 0$ and*

$$\mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq 1 - \delta. \tag{4}$$

For example, for $\delta = 0.01, \eta = 0.5$, for any $37 \leq N \leq 0.25 e^{d(1-\frac{\epsilon^2}{2})/2}$ we have that $\widehat{\text{PV}} > 0.5$ with probability at least 0.99. The theorem shows that, for this choice of distributions, for a sample size that is smaller than $O(e^d)$, there is a high probability that the value of $\widehat{\text{PV}}$ is far form PV.

It can be observed that the empirical estimate $\widehat{\text{PV}}$ is stable, that is, it is almost identical for two data sets differing on one sample. Due to its stability, applying McDiarmid inequality yields the following.

Theorem 5. *Let $S_1 = \{x_1, ..., x_n\} \sim P$ and $S_2 = \{y_1, ..., y_m\} \sim Q$ be two i.i.d. samples. Let $n \geq m$, then for any $\eta > 0$*

$$\mathbb{P}\left(|\widehat{PV}(S_1, S_2, \epsilon) - \mathbb{E}[\widehat{PV}(n, m, \epsilon)]| \geq \eta\right) \leq e^{-\eta^2 m^2 / 4n},$$

where $\mathbb{E}[\widehat{PV}(n, m, \epsilon)]$ is the expectation of $\widehat{PV}$ for a given sample size.

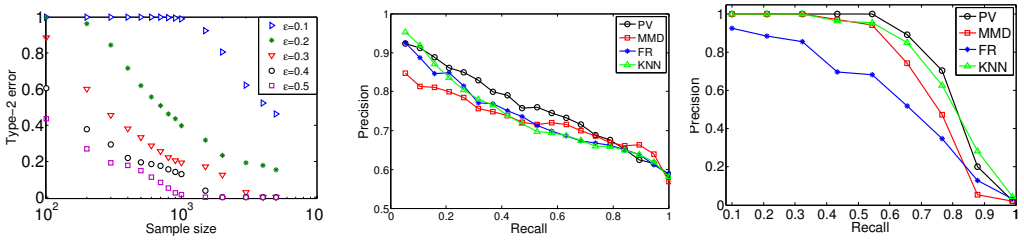
This theorem shows that the sample estimate of the PV converges to its expectation without dependence on the dimension. By combining this result with Theorem 3 it may be deduced that only the convergence of the bias – the difference $|\mathbb{E}[\widehat{\text{PV}}(n, m, \epsilon)] - \text{PV}(P, Q, \epsilon)|$ – may be exponential in the dimension. This convergence is distribution dependent. However, intuitively, slow convergence is not always the case, for example when the support of the distributions lies in a lower dimensional manifold of the space. To remedy this dependency we propose a bootstrapping bias correcting technique, presented in Section 5. A different possibility is to project the data to one dimension; due to space limitations, this extension of the PV is left out of the scope of this paper and presented in Appendix A.2 in the supplementary material.

5 Statistical Inference

We construct two types of complementary procedures for hypothesis testing of similarity and dissimilarity[2]. In the first type of procedures, given $0 \leq \theta < 1$, we distinguish between the null hypothesis $\mathcal{H}_0^{(1)} : \text{PV}(P, Q, \epsilon, \mathbf{d}) \leq \theta$, which implies similarity, and the alternative hypothesis $\mathcal{H}_1^{(1)} : \text{PV}(P, Q, \epsilon, \mathbf{d}) > \theta$. Notice that when $\theta = 0$, this test is a relaxed version of the TSP. Using $\text{PV}(P, Q) = 0$ instead of $P = Q$ as the null, allows for some distinction between the distributions, which gives the needed relaxation to capture similarity. In the second type of procedures, we test whether two distributions are similar. To do so, we flip the role of the null and the alternative. Note that there isn't an equivalent of this form for the TSP, therefore we can not infer similarity using the TSP test, but only reject equality. Our hypothesis tests are based on the finite sample analysis presented in Section 4; see Appendix A.1 in the supplementary material for the procedures.

To provide further inference on the PV, we apply bootstrapping for approximations of Confidence Intervals (CI). The idea of bootstrapping for estimating CIs is based on a two step procedure: approximation of the sampling distribution of the statistic by resampling with replacement from the initial sample – the bootstrap stage – following, a computation of the CI based on the resulting distribution. We propose to estimate the CI by Bootstrap Bias-Corrected accelerated (BCa) interval, which adjusts the simple percentile method to correct for bias and skewness [10]. The BCa is known for its high accuracy; particularly, it can be shown, that the BCa interval converges to the theoretical CI with rate $O(N^{-1})$, where N is the sample size. Using the CI, a hypothesis test may be formed: the null $\mathcal{H}_0^{(1)}$ is rejected with significance α if the range $[0, \theta] \not\subset [\underline{CI}, \overline{CI}]$. Also, for the second test, we apply the principle of CI inclusion [11], which states that if $[\underline{CI}, \overline{CI}] \subset [0, \theta]$, dissimilarity is rejected and similarity deduced.

[2]The two procedures are distinct, as, in general, lacking evidence to reject similarity is not sufficient to infer dissimilarity, and vice versa.



(a) The Type-2 error for varying perturbation sizes and ϵ values.

(b) Precision-Recall: Gait data.

(c) Precision-Recall: Video clips.

6 Experiments

6.1 Synthetic Simulations

In our first experiment, we examine the effect of the choice of ϵ on the statistical power of the test. For this purpose, we apply significance testing for similarity on two univariate uniform distributions: $P \sim U[0,1]$ and $Q \sim U[\Delta(\epsilon), 1+\Delta(\epsilon)]$, where $\Delta(\epsilon)$ is a varying size of perturbation. We considered values of $\epsilon = [0.1, 0.2, 0.3, 0.4, 0.5]$ and sample sizes up to 5000 samples from each distribution. For each value ϵ', we test the null hypothesis $\mathcal{H}_0^{(1)} : PV(P,Q,\epsilon') = 0$ for ten equally spaced values of $\Delta(\epsilon')$ in the range $[0, 2\epsilon']$. In this manner, we test the ability of the PV to detect similarity for different sizes of perturbations. The percentage of times the null hypothesis was falsely rejected, i.e. the type-1 error, was kept at a significance level $\alpha = 0.05$. The percentage of times the null hypothesis was correctly rejected, the power of the test, was estimated as a function of the sample size and averaged over 500 repetitions. We repeated the simulation using the tests based on the bounds as well as using BCa confidence intervals.

The results in Figure (3(a)) show the type-2 error of the bound based simulations. As expected, the power of the test increases as the sample size grows. Also, when finer perturbations need to be detected, more samples are needed to gain statistical power. For the BCa CI we obtained type-1 and type-2 errors smaller than 0.05 for all the sample sizes. This shows that the convergence of the estimated PV to its value is clearly faster than the bounds. Note that, given a sufficient sample size, any statistic for the TSP would have rejected similarity for any $\Delta > 0$.

6.2 Comparing Distance Measures

Next, we test the ability of the PV to measure similarity on real data. To this end, we test the ranking performance of the PV score against other known distributional distances. We compare the PV to the multivariate extension of the Wald-Wolfowitz score of Friedman & Rafsky (FR) [6] , Schilling's nearest neighbors score (KNN) [7], and the Maximum Mean Discrepancy score of Gretton et al. [9] (MMD)[3]. We rank similarity for the applications of video retrieval and gait recognition.

The ranking performance of the methods was measured by precision-recall curves, and the Mean Average Precision (MAP). Let r be the number of samples similar to a query sample. For each $1 \leq i \leq r$ of these observations, define $r_i \in [1, T-1]$ as its similarity rank, where T is the total number of observations. The Average Precision is: $AP = 1/r \sum_i i/r_i$, and the MAP is the average of the AP over the queries. The tuning parameter for the methods – k for the KNN, σ for the MMD (with RBF kernel), and ϵ for the PV – were chosen by cross-validation. The Euclidian distance was used in all methods.

In our first experiment, we tested raking for video-clip retrieval. The data we used was collected and generated by [12], and includes 1,083 videos of commercials, each of about 1,500 frames (25 fps). Twenty unique videos were selected as query videos, each of which has one similar clip in

[3]Note that the statistical tests of these measures test equality while the PV tests similarity and therefore our experiments are not of statistical power but of ranking similarity. Even in the case of the distances that may be transformed for similarity, like the MMD, there is no known function between the PV similarity to other forms of similarity. As a result, there is no basis on which to compare which similarity test has better performance.

Table 1: MAP for Auslan, Video, and Gait data sets. Average MAP ($\pm$ standard deviation) computed on a random selection of 75% of the queries, repeated 100 times.

DATA SET	$\widehat{\text{PV}}$	KNN	MMD	FR
VIDEO	**0.758** ± 0.009	**0.741** ± 0.014	0.689 ± 0.008	0.563 ± 0.019
GAIT	**0.792** ± 0.021	0.736 ± 0.014	0.722 ± 0.017	0.698 ± 0.017
GAIT-F	**0.844** ± 0.017	0.750 ± 0.015	0.729 ± 0.017	0.666 ± 0.016
GAIT-M	0.679 ± 0.024	0.712 ± 0.017	0.716 ± 0.031	**0.799** ± 0.016

the collection, to which 8 more similar clips were generated by different transformations: brightness increased/decreased, saturation increased/decreased, borders cropped, logo inserted, randomly dropped frames, and added noise frames. Lastly, each frame of a video was transformed to a 32-RGB representation. We computed the similarity rate for each query video to all videos in the set, and ranked the position of each video. The results show that the PV and the KNN score are invariant to most of the transformations, and outperform the FR and MMD methods (Table 1 and Figure 3(c)). We found that brightness changes were most problematic for the PV. For this type of distortion, the simple RGB representation is not sufficient to capture the similarity.

We also tested gait similarity of female and male subjects; same gender samples are assumed similar. We used gait data that was recorded by a mobile phone, available at [13]. The data consists of two sets of 15min walks of 20 individuals, 10 women and 10 men. As features we used the magnitude of the triaxial accelerometer.We cut the raw data to intervals of approximately 0.5secs, without identification of gait cycles. In this manner, each walk is represented by a collection of about 1500 intervals. An initial scaling to [0,1] was performed once for the whole set. The comparison was done by ranking by gender the 39 samples with respect to a reference walk.

The precision-recall curves in Figure 3(b) show that the PV is able to retrieve with higher precision in the mid-recall range. For the early recall points the PV did not show optimal performance; Interestingly, we found that with a smaller ϵ, the PV had better performance on early recall points. This behavior reflects the flexibility of the PV: smaller ϵ should be chosen when the goal is to find very similar instances, and larger when the goal is to find higher level similarity. The MAP results presented in Table 1 show that the PV had better performance on the female subjects. From examination of the subject information sheet we found that the range of weight and hight within the female group is 50-77Kg and 1.6-1.8m, while within the male group it is 47-100Kg and 1.65-1.93m; that is, there is much more variability in the male group. This information provides a reasonable explanation to the PV results, as it appears that a subject from the male group may have a gait that is as dissimilar to the gait of a female subject as it is to a different male. In the female group the subjects are more similar and therefore the precision is higher.

7 Discussion

We proposed a new score that measures the similarity between two multivariate distributions, and assigns to it a value in the range [0,1]. The sensitivity of the score, reflected by the parameter ϵ, allows for flexibility that is essential for quantifying the notion of similarity. The PV is efficiently estimated from samples. Its low computational complexity relies on its simple binary classification of points as neighbors or non-neighbor points, such that optimization of distances of faraway points is not needed. In this manner, the PV captures only the essential information to describe similarity. Although it is not a metric, our experiments show that it captures the distance between similar distributions as well as well known distributional distances. Our work also includes convergence analysis of the PV. Based on this analysis we provide hypothesis tests that give statistical significance to the resulting score. While our bounds are dependent on the dimension, when the intrinsic dimension of the data is smaller than the domains dimension, statistical power can be gained by bootstrapping. In addition, the PV has an intuitive interpretation that makes it an attractive score for a meaningful statistical testing of similarity. Lastly, an added value of the PV is that its computation also gives insight to the areas of discrepancy; namely, the areas of the unmatched samples. In future work we plan to further explore this information, which may be valuable on its own merits.

Acknowledgements

This Research was supported in part by the Israel Science Foundation (grant No. 920/12).

References

[1] G. Monge. Mémoire sur la théorie des déblais et de remblais. Histoire de l'Academie Royale des Sciences de Paris, avec les Memoires de Mathematique et de Physique pour la meme annee, 1781.

[2] L. Rüschendorf. Monge–kantorovich transportation problem and optimal couplings. *Jahresbericht der DMV*, 3:113–137, 2007.

[3] A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons Inc, 1998.

[4] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998.

[5] R.K. Ahuja, L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*, chapter 12, pages 469–473. Prentice Hall, 1993.

[6] J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

[7] M.F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, pages 799–806, 1986.

[8] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases*, pages 180–191. VLDB Endowment, 2004.

[9] A. Gretton, K. Borgwardt, B. Schölkopf, M. Rasch, and E. Smola. A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems 19*, 2007.

[10] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, chapter 14, pages 178–188. Chapman & Hall/CRC, 1993.

[11] S. Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority; 2nd edition*. Chapman and Hall/CRC, 2010.

[12] J. Shao, Z. Huang, H. Shen, J. Shen, and X. Zhou. Distribution-based similarity measures for multi-dimensional point set retrieval applications. In *Proceeding of the 16th ACM international conference on Multimedia MM 08*, 2008.

[13] J. Frank, S. Mannor, and D. Precup. Data sets: Mobile phone gait recognition data, 2010.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*, chapter 5, pages 258–261. Cambridge University Press, New York, NY, USA, 2004.

[15] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M.J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.