A DIMENSION-INDEPENDENT DISCRIMINANT BETWEEN DISTRIBUTIONS

Salimeh Yasaei Sekeh, Brandon Oselio, Alfred O. Hero III

Department of Electrical Engineering and Computer Science
University of Michigan
1301 Beal Ave, Ann Arbor, MI, 48109, USA

ABSTRACT

Henze-Penrose divergence is a non-parametric divergence measure that can be used to estimate a bound on the Bayes error in a binary classification problem. In this paper, we show that a cross-match statistic based on optimal weighted matching can be used to directly estimate Henze-Penrose divergence. Unlike an earlier approach based on the Friedman-Rafsky minimal spanning tree statistic, the proposed method is dimension-independent. The new approach is evaluated using simulation and applied to real datasets to obtain Bayes error estimates.

Index Terms— Bayes error rate, classification, Henze-Penrose divergence, Cross-match test statistic, Optimal weighted matching, Friedman-Rafsky statistic.

1. INTRODUCTION

Many information theoretic measures have been applied to measure the discrimination between probability density functions. They have been used in various applications in signal processing, classification, image registration, clustering and structure learning, see [1, 2, 3, 4]. A special class of divergence measures, called f-divergences have the property that the divergence functional f is convex and $f(1) = 0$. Among the different divergence functions belonging to the f-divergence family, [5, 6] the Henze-Penrose (HP) divergence has been of great interest due to its application to binary classification, in particular to bound the Bayes error rate.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathcal{R}^d$ be realizations of random vector $\mathbf{X}$ and class labels $y \in \{0, 1\}$, with prior probabilities $c_0 = P(y = 0)$ and $c_1 = P(y = 1)$, such that $c_0 + c_1 = 1$. Given conditional distributions $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$, the Bayes error rate is given by

$$\epsilon = \int_{\mathcal{R}^d} \min\left\{c_0 p_0(\mathbf{x}), c_1 p_1(\mathbf{x})\right\} \mathrm{d}\mathbf{x}. \quad (1)$$

The Bayes error rate is the expected risk for the Bayes classifier, which assigns a given feature vector $\mathbf{x}$ to the class with the highest posterior probability, and is the lowest possible error rate of any classifier for a particular joint distribution. It is thus a reasonable measure for assessing the intrinsic difficulty of a particular classification problem. By estimating and bounding this value, we can then have a better understanding of the problem difficulty, which allows the user to make more informed decisions.

We define the HP-divergence between p_0 and p_1, $D_c(p_0, p_1)$ by

$$\frac{1}{4c_0c_1}\left[\int_{\mathcal{R}^d} \frac{\left(c_0 p_0(\mathbf{x}) - c_1 p_1(\mathbf{x})\right)^2}{c_0 p_0(\mathbf{x}) + c_1 p_1(\mathbf{x})}\, \mathrm{d}\mathbf{x} - (c_0 - c_1)^2\right]. \quad (2)$$

Note that for all c_0 and c_1, $0 \leq D_c(p_0, p_1) \leq 1$ and when $p_0 = p_1$ the HP-divergence becomes zero.

The authors of [7] showed that HP-divergence yields tighter bounds on the Bayes error rate ϵ, given in (1), than those based on the Bhattacharya distance, [8]. In particular, the following bound on the Bayes error rate holds:

$$\frac{1}{2} - \frac{1}{2}\sqrt{u_c(p_0, p_1)} \leq \epsilon \leq \frac{1}{2} - \frac{1}{2} u_c(p_0, p_1), \quad (3)$$

where $u_c(p_0, p_1) = 4c_0c_1 D_c(p_0, p_1) + (c_0 - c_1)^2$.

In this paper we propose a new direct estimator for HP-divergence using a statistic based on optimal weighted matching [9]. Matching for general graphs is a combinatorial optimization problem that can be solved in polynomial time. In [9], the optimal weighted matching was used to find a statistical test for equal posterior distributions using the cross match statistic. We demonstrate that the same statistic described in that series of papers can be utilized to estimate HP-divergence. We emphasize that the proposed weighted matching estimator is completely different from weighted K-NN graph estimators.

The rest of the paper is organized as follows. Section 2 briefly describes related work on HP-divergence and optimal weighted matching. Section 3 defines the cross-match statistic, and in Section 4 we prove that the cross-match statistic approximately tends to the HP-divergence when samples sizes of two classes increases simultaneously in a specific regime. Section 5 shows sets of simulations for our proposed method and compares the Friedman-Rafsky (FR) and cross-match estimators experimentally, and we estimate the Bayes error rate on a few real datasets. Finally, Section 6 concludes the paper.

2. RELATED WORK

Several estimators for HP-divergence have been proposed in the literature: Plug-in estimates were introduced in [10] and later have been studied more in [11, 12, 13]. Plug-in approaches estimate the underlying distribution function and then plug this value into the divergence function. The drawback with the plug-in estimates is that these methods are not accurate near support boundaries and are also more computationally complex. There have been a number of attempts to non-parametrically approximate divergence measures using graph-based algorithms such as minimal spanning tree (MST), [14, 15] and k-nearest neighbors graphs (k-NNG), [16].

One of the most common direct estimators is based on Friedman-Rafsky (FR) multivariate test statistic [17]. This approach is constructed from the MST on the concatenated data set drawn from sufficiently smooth probability densities. Henze and Penrose [18] showed that the FR test is consistent against all alternatives. Therefore, the HP-divergence has the appealing property that there exists

This work was partially supported by ARO grant W911NF-15-1-0479.

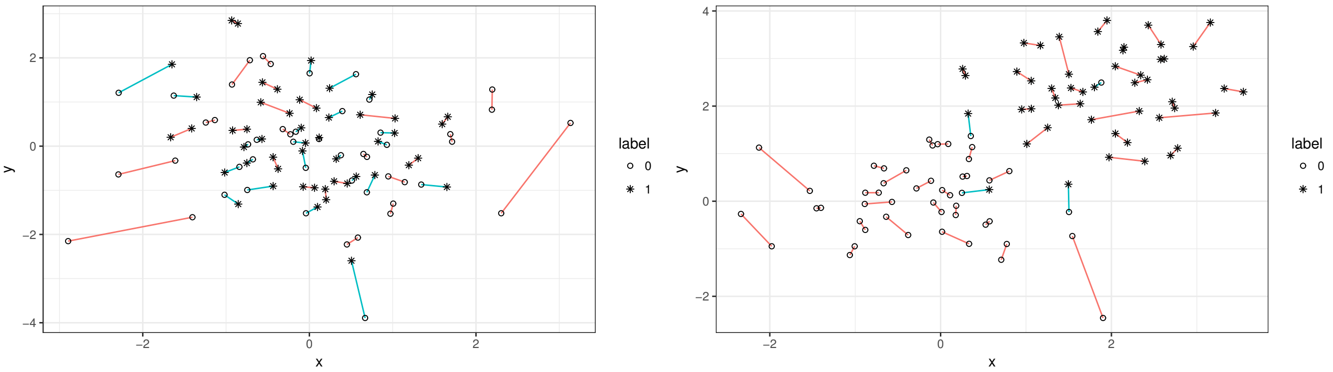

Fig. 1. An example of the cross-match statistics for two cases $p_0 = p_1$ (left-generated from standard Gaussian distributions) and $p_0 \neq p_1$ (right-Generated from Gaussian distributions with means $[0, 0]$, $[2, 2]$). The total number of blue edges is the cross match statistics.

an asymptotically consistent direct estimator in terms of the FR test statistic, see [18, 19, 7]. The variance of the FR test statistic under the assumption of equal distributions depends on the dimension of the data d, which may be unknown, especially when the support of the densities is a common but unknown lower dimensional manifold.

Optimal weighted matching is a well studied combinatorial optimization problem [20]. It has been used extensively in operations engineering. Previous statistical work using weighted matching have derived useful applications of the cross-match test statistic in fields like biological networks [9, 21].

3. THE CROSS-MATCH TEST STATISTIC

Consider N i.i.d. samples $\mathcal{X}_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathcal{R}^d$ and corresponding labels $y_i \in \{0, 1\}$. Define $\mathbf{y} = (y_1, \ldots, y_N)$, and further $m = \sum\limits_{l=1}^{N} y_l$, and $n = N - m$, so that m is the number of samples in $\mathbf{x}$ with class 1, and n is the number with class label 0. Further, we create D, a $N \times N$ Euclidean distance matrix, with $D_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$. Without loss of generality, we assume N is even, as we can always add a 'ghost point' $\mathbf{x}_{N+1}$, where $D_{iN+1} = 0, \forall i$. In the following, we consider a complete weighted graph $G = (V, E, D)$, with the vertices $V = 1, \ldots, N$ representing the sample points $\mathbf{x}_1, \ldots, \mathbf{x}_N$, edges $E = \{\{i, j\},\ i, j \in V\}$, and weights for each edge $\{i, j\}$ as D_{ij}.

A complete matching $M \subset E$ on a weighted graph is a set of edges such that no two edges in M share a common vertex, and every vertex is used in the matching. The complete minimum weighted matching M^* is defined as the matching on G such that $M^* = \arg\min_M \sum_{i,j \in M} D_{ij}$. We note that this is similar to the FR test [17], which uses the same matrix D to find the minimal spanning tree. The FR test statistic is the total number of edges in the D-based MST connecting different labeled nodes.

Using this matching, we find the *cross-match statistic*, $\mathcal{A}(\mathcal{X}_N)$ which is the number of edges that match dichotomous samples, i.e. samples with different class labels, that is

$$\mathcal{A}(\mathcal{X}_N) = \sum_{\{i,j\} \in M^*} \Big(y_i(1 - y_j) + (1 - y_i)y_j \Big). \tag{4}$$

In Figure 1 we show two numerical examples. The left plot shows samples from two equal distributions, and right plot shows samples from differing distributions. Qualitatively, we see that $\mathcal{A}$ is much greater for the equivalent distributions than for the differing distributions, because the optimal matching tries to reduce long distances, which will reduce the number of edges between differing distributions.

In Proposition 1 in [9], under the assumption of equal distributions, the expectation and variance of $\mathcal{A}(\mathbf{x})$ are derived:

$$E[\mathcal{A}] = \frac{mn}{N - 1}, \quad Var[\mathcal{A}] = \frac{2n(n-1)m(m-1)}{(N-3)(N-1)^2}. \tag{5}$$

We note that the mean and variance of the cross-match statistic under equal distributions are dimension-independent, but this is not true for the FR statistic, whose variance is dependent on the degrees of the MST. The maximal degrees of the MST is in fact dependent on the dimension d of the underlying samples, e.g., the MST has maximal degree 4 in $d = 2$ dimensions while its maximal degree is known to be between 13 or 14 in 3 dimensions [22]. This dependence causes the FR statistic to perform poorly in higher dimensions. In Section 5 we perform a set of experiments where dimension varies to demonstrate the advantage of the cross-match statistic over the FR statistic.

4. HP-DIVERGENCE ESTIMATION

Here we introduce the cross-match statistic as an estimate of the HP-divergence given in (2). Assume that we have two sets of samples $\mathcal{X}_m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\mathcal{U}_n = \{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ with two different labels. In order to show asymptotic convergence to HP-divergence, we make the following assumption regarding the cross-match statistic (similar to Lemma 1 in [18]).

Assumption 1: For disjoint sets $\mathcal{X}_m, \mathcal{U}_n$ and $\{s, t\}$ we have

$$\Big| \mathcal{A}(\mathcal{X}_m \cup \{s, t\} \cup \mathcal{U}_n) - \mathcal{A}(\mathcal{X}_m \cup \mathcal{U}_n) \Big| \leq k_d. \tag{6}$$

where k_d is a constant that may depend on d. This means that even if the optimal matching changes a great deal, the number of edges that are between the two samples is still approximately the same.

We empirically check this assumption in Figure 2. We generate two sets of d-dimensional samples from standard Gaussian with mean $\mu_0 = [0]_d$, $\mu_1 = [1]_d$ and $\Sigma_0 = \Sigma_1 = I_d$ for $d = 2, 4, 6, 8$. We plot the difference in cross-match statistic when adding two points (labeled by $\mathcal{A}_{\text{diff}}$), and perform this test over varying sample size. We see that $\mathcal{A}$ does not vary significantly when adding a new sample in the tested cases.

Lemma 1 *Let $g : \mathcal{R}^d \times \mathcal{R}^d \to [0,1]$ be a symmetric and measurable function, such that for almost every $\mathbf{x} \in \mathcal{R}^d$, $g(\mathbf{x}, .)$ is measurable with $\mathbf{x}$ a Lebesgue point of the functions $p(.)g(\mathbf{x}, .)$ and $p(.)$. For each N, let $\mathbf{Z}_1^N, \mathbf{Z}_2^N, \dots, \mathbf{Z}_N^N$ be independent d-dimensional variables with common density function p_N convergent to p as $N \to \infty$ and set $\mathcal{Z}_N = \{\mathbf{Z}_1^N, \dots, \mathbf{Z}_N^N\}$. Consider the complete minimum weighted matching M^* on $\mathcal{Z}_N$. Then*

$$\lim_{N\to\infty} N^{-1} E \sum\sum_{1\le i<j\le N} g(\mathbf{Z}_i^N, \mathbf{Z}_j^N) \mathbf{1}\{(\mathbf{Z}_i^N, \mathbf{Z}_j^N) \in M^*(\mathcal{Z}_N)\}$$

$$= \frac{1}{2} \int_{\mathcal{R}^d} g(\mathbf{x}, \mathbf{x})\, p(\mathbf{x}). \tag{7}$$

Proof: For given $\mathbf{x}$ in a subset $\mathcal{S} \in \mathcal{R}^d$, the degree of vertex $\mathbf{x}$ in $M^*(\mathcal{S})$ is one. Let $\mathbf{x}$ be a Lebesgue point of $p(.)$ and $p(.)g(\mathbf{x}, .)$ and $\mathcal{Z}_N^{\mathbf{x}}$ be the point process $\{\mathbf{x}, \mathbf{Z}_2^N, \mathbf{Z}_3^N, \dots, \mathbf{Z}_N^N\}$. Let $\mathcal{B}(\mathbf{x}, r) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \le r\}$. Therefore, we can write

$$\begin{aligned} E\sum_{j=2}^{N} \big|g(\mathbf{x}, \mathbf{Z}_j^N) - g(\mathbf{x}, \mathbf{x})\big| \mathbf{1}\{\mathbf{Z}_j^N \in \mathcal{B}(\mathbf{x}, N^{-1/d})\} \\ = (N-1) \int_{\mathcal{B}(\mathbf{x}, N^{-1/d})} \big|g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, \mathbf{x})\big| p_N(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ = (N-1) \int_{\mathcal{B}(\mathbf{x}, N^{-1/d})} \big|g(\mathbf{x}, \mathbf{y}) p_N(\mathbf{y}) - h(\mathbf{x}, \mathbf{x}) p_N(\mathbf{x}) \\ + g(\mathbf{x}, \mathbf{x})(p_N(\mathbf{x}) - p_N(\mathbf{y})\big|\, \mathrm{d}\mathbf{y}, \end{aligned} \tag{8}$$

Since $\mathbf{x}$ is a Lebesgue point of p_N and $g(\mathbf{x}, .)P_N(.)$ then (8) tends to zero. Note that the degree of vertex in $M^*(\mathcal{Z}_N^{\mathbf{x}})$ is one. For almost all $\mathbf{x}$,

$$E\sum_{j=2}^{N} g(\mathbf{x}, \mathbf{Z}_j^N) \mathbf{1}\{(\mathbf{x}, \mathbf{Z}_j^N) \in M^*(\mathcal{Z}_N^x)\} = g(\mathbf{x}, \mathbf{x}) + o(1). \tag{9}$$

The function g has range $[0,1]$ so the left hand side of (9) is bounded by one. By the dominated convergence theorem

$$\begin{aligned} N^{-1} E \textstyle\sum\sum_{1\le i<j\le N} g(\mathbf{Z}_i^N, \mathbf{Z}_j^N) \mathbf{1}\{(\mathbf{Z}_i^N, \mathbf{Z}_j^N) \in M^*(\mathbf{Z}_N)\} \\ = \frac{1}{2} E \sum_{j=2}^{N} g(\mathbf{Z}_1^N, \mathbf{Z}_j^N) \mathbf{1}\{(\mathbf{Z}_1^N, \mathbf{Z}_j^N) \in M^*(\mathbf{Z}_N)\} \\ = \frac{1}{2} \int_{\mathbf{x}} p_N(\mathbf{x}) E \sum_{j=2}^{N} g(\mathbf{x}, \mathbf{Z}_i^N) \mathbf{1}\{(\mathbf{x}, \mathbf{Z}_j^N) \in M^*(\mathbf{Z}_N)\}. \end{aligned} \tag{10}$$

The last line in (10) tends to right hand side of (8). □

The following theorem proves the direct estimate of HP-divergence based on $\mathcal{A}(\mathcal{X}_N)$. Due to space limitations only an outline of the proof is given.

Theorem 1 *As $m \to \infty$ and $n \to \infty$ such that $m/N \to c_1$ and $n/N \to c_0$, where $N = m + n$. Denote $\mathcal{A}_{m,n} := \mathcal{A}(\mathcal{X}_m \cup \mathcal{U}_n)$ the cross-match statistic given by the optimal weighted matching over $\mathcal{X}_m$ and $\mathcal{U}_n$. Then under Assumption 1 we have*

$$1 - \left(\frac{N}{m\ n}\right) \mathcal{A}_{m,n} \to D_c(p_0, p_1), \quad a.s. \tag{11}$$

Proof: The proof shares some similarity with the FR convergence proof of the HP-divergence in [18]. The primary difference lies in handling the difference between the cross-match statistic when nodes are added, i.e. (6). We use Lemma 1 and Poissonization to prove (11).

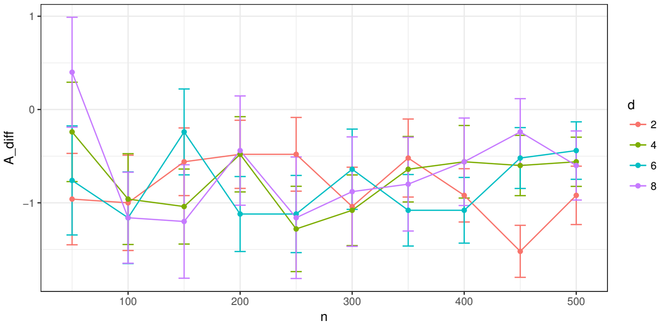

Fig. 2. The cross-match statistics difference with error bars at the standard deviation from 50 trials for the Gaussian samples by adding two points.

Let M_m and N_n be Poisson variables with mean m and n such that $m+n$ is even and independent of one another and of $\mathbf{X}_i$ and $\mathbf{U}_j$. Let $\mathcal{X}'_m$ and $\mathcal{U}'_n$ be the Poisson processes $\{\mathbf{X}_1, \dots, \mathbf{X}_{M_m}\}$ and $\{\mathbf{U}_1, \dots, \mathbf{U}_{N_n}\}$, respectively. Set $\mathcal{A}'_{m.n} = \mathcal{A}(\mathcal{X}'_m \cup \mathcal{U}'_n)$, the cross-match statistic. By (6), we have

$$\Big|\mathcal{A}'_{m,n} - \mathcal{A}_{m,n}\Big| \le k_d \left(|M_m - m| + |N_n - n|\right). \tag{12}$$

Note that $(m+n)^{-1} E\big|\mathcal{A}'_{m,n} - \mathcal{A}_{m,n}\big| \to 0$. Poissonization makes the identities of the points of $\mathcal{X}'_m \cup \mathcal{U}'_n$ conditionally independent, given their positions. For each m and n let $\mathbf{Z}_1^{m,n}, \mathbf{Z}_2^{m,n}, \dots$ be independent discrete variables with common density $p_{m,n}(\mathbf{x}) = (mp_0(\mathbf{x}) + np_1(\mathbf{x}))/(m+n)$. Let $W_{m,n}$ be an independent Poisson variable with even valued mean $(m+n)$. Let $\mathcal{Z}'_{m,n} = \{\mathbf{Z}_1^{m,n}, \dots, \mathbf{Z}_{W_{m,n}}^{m,n}\}$ be a non-homogeneous Poisson process of rate $mp_0 + np_1$. Following the same arguments in [18], assign a mark from the set $\{1,2\}$ to each point of $\mathcal{Z}'_{m,n}$. Specifically, a point $\mathbf{x}$ is assigned mark 1 with probability $mp_0(\mathbf{x})/\big(mp_0(\mathbf{x}) + np_1(\mathbf{x})\big)$ and mark 2 otherwise. Let $\widetilde{\mathcal{X}}_m$ and $\widetilde{\mathcal{U}}_n$ be the set of points of $\mathcal{Z}'_{m,n}$ marked 1 and 2 respectively. Also denote $\widetilde{\mathcal{A}}_{m,n}$ the cross match statistic given from optimal weighted matching over $\widetilde{\mathcal{X}}_m \cup \widetilde{\mathcal{U}}_n$. Define the probability of two points in $\mathcal{Z}'_{m,n}$ having different marks by $g_{m,n}(\mathbf{x}, \mathbf{y})$:

$$g_{m,n}(\mathbf{x}, \mathbf{y}) = \frac{mp_0(\mathbf{x})np_1(\mathbf{y}) + np_1(\mathbf{x})mp_0(\mathbf{y})}{(mp_0(\mathbf{x}) + np_1(\mathbf{x}))(mp_0(\mathbf{y}) + np_1(\mathbf{y}))}. \tag{13}$$

We know that $m/N \to c_0$ and $n/N \to c_1$, hence $g_{m,n}(\mathbf{x}, \mathbf{y}) \to g(\mathbf{x}, \mathbf{y})$ where

$$g(\mathbf{x}, \mathbf{y}) = \frac{c_0 c_1 \big(p_0(\mathbf{x})p_1(\mathbf{y}) + p_1(\mathbf{x})p_0(\mathbf{y})\big)}{\big(c_0 p_0(\mathbf{x}) + c_1 p_1(\mathbf{x})\big)\big(c_0 p_1(\mathbf{y}) + c_1 p_1(\mathbf{y})\big)}. \tag{14}$$

So, the conditional expectation $E\big[\widetilde{\mathcal{A}}_{m,n} | \mathcal{Z}'_{m,n}\big]$ becomes:

$$\sum\sum_{1\le i<j\le W_{m,n}} g_{m,n}(\mathbf{Z}_i^{m,n}, \mathbf{Z}_j^{m,n}) \mathbf{1}\{(\mathbf{Z}_i^{m,n}, \mathbf{Z}_j^{m,n}) \in M^*(\mathcal{Z}'_{m,n})\}. \tag{15}$$

By taking expectations in (15), one yields $E\big[\widetilde{\mathcal{A}}_{m,n}\big]$.

Let $\mathcal{Z}_{m,n} := \{\mathbf{Z}_1^{m,n}, \mathbf{Z}_2^{m,n}, \dots, \mathbf{Z}_{m,n}^{(m+n)}\}$ be the original non-Poissonized set of points. By the fact that

$$E\big[|M_m + N_n - (m+n)|\big] = o(m+n),$$

the Poissonized limit of $E\big[\tilde{\mathcal{A}}_{m,n}\big]$. Set $p(\mathbf{x}) = c_0 p_0(\mathbf{x}) + c_1 p_1(\mathbf{x})$, then $p_{m,n}(\mathbf{x}) \to p(\mathbf{x})$. Using Lemma 1, we get

$$\begin{aligned} \frac{E\big[\widetilde{\mathcal{A}}_{m,n}\big]}{(m+n)} &\to \frac{1}{2}\int_{\mathcal{R}^d} g(\mathbf{x},\mathbf{x})p(\mathbf{x}) \\ &= c_0\ c_1 \int_{\mathcal{R}^d} \frac{p_0(\mathbf{x})p_1(\mathbf{x})}{c_0 p_0(\mathbf{x}) + c_1 p_1(\mathbf{x})}. \end{aligned} \tag{16}$$

This completes the proof of Theorem 1. □

5. EXPERIMENTS

We perform multiple experiments to demonstrate the utility of the proposed direct estimator of HP-divergence in terms of dimension and sample size. We subsequently apply our estimator to determine empirical bounds on the Bayes error rate for various datasets.

For the following simulations, the sample sizes for each class were equal ($m = n$). Each simulation used a multivariate Normal distribution for each class.

We first analyze the estimator's performance as the sample size $N = m + n$ increases. For each value of N, the simulation was run 50 times, and the results were averaged. Samples from each class were i.i.d. 2-dimensional Normal random variables, with $\mu_0 = [0, 0]$ and $\mu_1 = [1, 1]$, $\Sigma_0 = \Sigma_1 = I_2$.

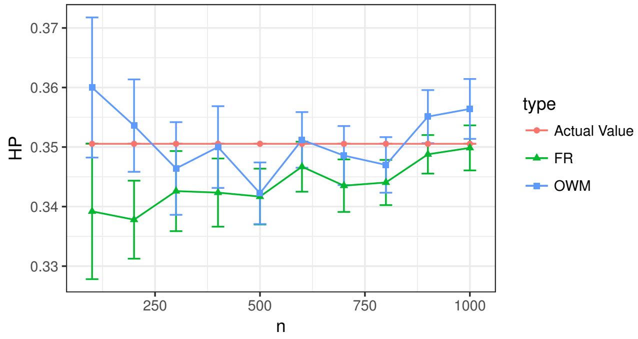

Fig. 3. HP-divergence estimation vs. sample size n. Error bars denote the standard deviation over 50 trials. The proposed estimator and the FR estimator perform approximately equivalently over this range of sample sizes.

We see that as N increases the performance of the FR estimator and our proposed estimator (labeled OWM) are comparable for N up to 1000. The observed variance of our estimators are slightly higher than the FR estimator. For dimension $d = 2$ this is not surprising as we would expect the FR estimator to perform the best in this case.

Figure 4 (top) shows the averaged estimates of the HP-divergences over increasing dimension. Here we see that the proposed cross-matching estimator shows improvement with respect to the FR estimator, as expected. For each dimension evaluated in Figure 4, $N = 1000$, and $\mu_0 = [0]_d$ and $\mu_1 = [0.5]_d$, $\Sigma_0 = \Sigma_1 = I_d$. The proposed cross-matching estimator is slightly less biased as dimension increases, and as shown in Figure 4 (bottom) we improve in empirical MSE.

Next we show the results of applying the HP-divergence estimator to 4 different real data sets. Table 1 shows the cross match statistics and estimated upper bounds for Bayes Error (denoted by the column labeled ϵ).

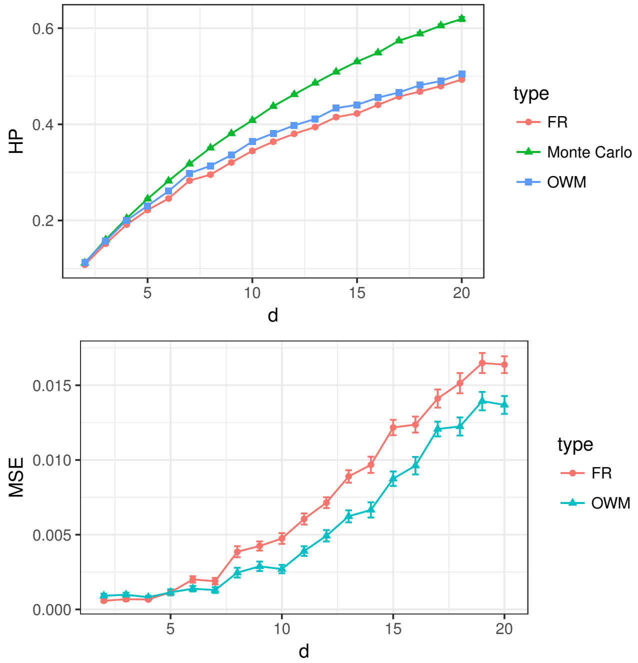

Fig. 4. HP-divergence (top) and empirical MSE (bottom) vs. dimension. The empirical MSE of both estimators increases for larger dimensional data sets. The MSE is better for the proposed (OWM) estimator.

Bayes Error Bounds					
Data set	$\mathcal{A}(\mathcal{X}_N)$	$\widehat{D}_c$	n	m	ϵ
Breast cancer [23]	33	0.791	488	241	0.093
Mines vs. Rocks [24]	7	0.864	97	111	0.067
Pima diabetes [24]	67	0.641	549	283	0.161
Hyper thyroid [24]	37	0.743	3012	151	0.023

Table 1. $\mathcal{A}(\mathcal{X}_N)$, $\widehat{D}_c$, n, m and ϵ are the cross-match statistics, HP-divergence estimates using $\mathcal{A}(\mathcal{X}_N)$, sample sizes and upper bounds for Bayes Error respectively.

6. CONCLUSION

We proposed a new dimension-independent direct estimator of HP-divergence using a statistic derived from optimal weighted matching. The estimator is more accurate than the FR approach and its variance is independent of the dimension of the support of the distributions. This translates to improved MSE performance as compared to other HP-divergence estimation methods, especially for high dimension. We validated our proposed estimator using simulations, and illustrated the approach for the meta-learning problem of estimating Bayes classification error for four real-world data sets.

7. ACKNOWLEDGEMENTS

We thank the UCI machine learning repository for the use of their various datasets used in this paper [24].

8. REFERENCES

[1] Xuan Guorong, Chai Peiqi, and Wu Minhui, "Bhattacharyya distance feature selection," in *In Pattern Recognition, Proceedings of the 13th International Conference on IEEE*, 1996, vol. 2, pp. 195–199.

[2] Paul Viola and William M Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[3] A B Hamza and H Krim, "Image registration and segmentation by maximizing the jensen-rényi divergence," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2003, pp. 147–163.

[4] Kevin R Moon, Morteza Noshad, Salimeh Yasaei Sekeh, and Alfred O Hero, "Information theoretic structure learning with confidence," in *in Proc. IEEE Int. Conf. Acoust Speech Signal Process*, 2017.

[5] S Ali and S D Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Statist. Soc. Ser. B (Methodology.)*, pp. 131–142, 1996.

[6] I Csiszár and P C Shields, "Information theory and statistics: A tutorial," *J. Royal Statist. Soc. Ser. B (Methodology.)*, vol. 1, no. 4, pp. 417–528, 2004.

[7] Visar Berisha, Alan Wisler, Alfred O. Hero, and Andreas Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Trans. on Signal Process.*, vol. 64, no. 3, pp. 580–591, 2016.

[8] A Battacharyya, "On a measure of divergence between two multinomial populations," *Sankhy ā: The Indian Journal of Statistics*, pp. 401–406, 1946.

[9] Paul R Rosenbaum, "An exact distribution-free test comparing two multivariate distributions based on adjacency," *Journal of Royal Statistics Society B*, vol. 67, no. 4, pp. 515–530, 2005.

[10] Kumar Scricharan, Raviv Raich, and Alfred O Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4135–4159, 2012.

[11] Kevin R Moon and Alfred O Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.

[12] Kevin R Moon and Alfred O Hero, "Ensemble estimation of multivariate f-divergence," in *IEEE International Symposium on Information Theory*, 2016, pp. 356–360.

[13] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero, "Non-parametric ensemble estimation of distributional functionals," *arXiv preprint arXiv:1601.06884v2*.

[14] J E Yukish, *Probability theory of classical Euclidean optimization*, Vol. 1675 of lecture notes in Mathematics, Springer-Verlag, Berlin, 1998.

[15] D Aldous and J M Steele, "Asymptotic for euclidean minimal spanning trees on random points," *Probab. Theory Related Fields*, vol. 92, pp. 247–258, 1992.

[16] Jillian Beardwood, J H Halton, and J M Hammersley, "The shortest path through many points," in *Mathematical Proceedings of the Cambridge Philosophical Society*, 1959, pp. 299–327.

[17] J H Friedman and L C Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *Ann. Statist.*, pp. 697–717, 1979.

[18] Norbert Henze and Mathew D Penrose, "On the multivariate runs test," *Ann. Statist.*, vol. 27, no. 1, pp. 290–298, 1999.

[19] Visar Berisha and Alfred O Hero, "Empirical non-parametric estimation of the fisher information," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 988–992, 2015.

[20] C H Papadimitriou and K Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Englewood Cliffs, Prentice Hall, 1982.

[21] Bo Lu and Paul R Rosenbaum, "Optimal pair matching with two control groups," *Journal of Computational and Graphical Statistics*, vol. 13, no. 2, pp. 422–434, 2004.

[22] Gabriel Robins and Jeffrey S Salowe, "On the maximum degree of minimum spanning trees," in *Proceedings of the tenth annual symposium on Computational geometry*. ACM, 1994, pp. 250–258.

[23] W H Wolberg and O L Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, pp. 9193–9196, 1990.

[24] M. Lichman, "UCI machine learning repository," 2013.