DIRECT ENSEMBLE ESTIMATION OF DENSITY FUNCTIONALS

Alan Wisler[1]*, Kevin Moon*[2]*, Visar Berisha*[1]

[1]Schools of ECEE and SHS, Arizona State University
[2]Genetics and Applied Math Departments, Yale University

ABSTRACT

Estimating density functionals of analog sources is an important problem in statistical signal processing and information theory. Traditionally, estimating these quantities requires either making parametric assumptions about the underlying distributions or using non-parametric density estimation followed by integration. In this paper we introduce a direct nonparametric approach which bypasses the need for density estimation by using the error rates of k-NN classifiers as "data-driven" basis functions that can be combined to estimate a range of density functionals. However, this method is subject to a non-trivial bias that dramatically slows the rate of convergence in higher dimensions. To overcome this limitation, we develop an ensemble method for estimating the value of the basis function which, under some minor constraints on the smoothness of the underlying distributions, achieves the parametric rate of convergence regardless of data dimension.

Index Terms— divergence estimation, k-nearest neighbor classifier, ensemble estimation, non-parametric

1. INTRODUCTION

Density functionals that map probability density functions (PDFs) to $\mathbb{R}$ have been used in many signal processing applications involving classification [1], segmentation [2], source separation [3], clustering [4], and other domains. Traditional estimation of these quantities typically relies on assuming a parametric model for the underlying PDFs, and calculating the desired functional from the estimated parameters of that model. Parametric methods offer good mean squared error (MSE) convergence rates $\mathcal{O}(N^{-1})$ (N represents the number of samples) when an accurate parametric model is known, but become asymptotically biased if the data do not fit the assumed model. To guarantee an asymptotically consistent estimator in these scenarios, two general classes of estimators exist: 1) non-parametric plug-in estimators and 2) graph-based direct estimators [5].

Non-parametric plug-in estimators have been used for estimating functionals of both discrete and analog distributions [6–8], with a particular focus on entropy estimation [9, 10]. While non-parametric plug-in estimators don't require a parametric model, they generally have high variance, are sensitive to outliers, and scale poorly with dimension [5]. Alternatively, graph-based estimators, exploit the asymptotic properties of minimal graphs to *directly* estimate density functionals without estimating the underlying distributions. These methods have been used to estimate density functionals such as entropy [11], the α-divergence [5], and the D_p-divergence [12]. Graph based methods bypass the complication of fine tuning parameters such as kernel bandwidth or histogram bin size and can offer faster convergence rates in some scenarios [5]. In this paper, we attempt to overcome two of the fundamental limitations of previously derived graph-based estimators: their specificity (their inability to estimate a broad variety of density functionals) and their convergence rates in higher dimensions.

A limitation of graph-based estimators is their specificity to the density functional being estimated. Whereas plug-in estimation methods can use the same general procedure to estimate a wide range of density functionals, most graph-based estimators can only be used to estimate a specific density functional. To overcome this, we use a set of density functionals, which can be estimated by a corresponding set of graph-based estimators, as data-driven basis functions. We can then use linear combinations of these data-driven basis functions to estimate unknown density functionals that cannot be estimated via graph-based methods directly. This strategy was originally introduced in [13] using a new basis set similar to the deterministic Bernstein polynomial basis. This paper extends that work by re-interpreting the k-nearest neighbor (k-NN) classifier error rate as a data-driven basis function.

In contrast to the Bernstein-like basis in [13], much is already known about the finite-sample convergence properties of the k-NN error rate. The k-NN error rate converges in MSE to its asymptotic value at the parametric rate when $d \leq 4$, but slows to $\mathcal{O}(N^{-\frac{4}{d}})$ at higher dimensions [14]. To overcome the slow convergence rate, we generalize theory previously developed for ensemble estimation of density functionals [15–17] to develop ensemble estimators of the k-NN error rate that can guarantee a $\mathcal{O}(N^{-1})$ rate of convergence independent of dimension. In [15–17], ensembles were used in the context of density estimation by varying the bandwidth parameter. We generalize this approach through an ensemble formed by *varying the sample size* instead; we show that this approach can yield the same parametric convergence rate. This approach shares similarities to other resampling methods, such as bootstrap and jackknife estimators, which can use subsets of data to measure and correct for the bias [18–21]. While jackknife estimators can in theory be used to eliminate several bias terms and significantly improve bias convergence [22], their practical utility is generally limited by their high variance [19]. The proposed method offers a principled way of reducing lower-order bias terms while controlling the variance of the estimator. By improving the convergence rate in our estimates of the basis functions, we also improve the convergence rate of the resulting density functional estimators. We verify this empirically by ap-

plying the proposed method to estimate the Hellinger distance directly from data drawn from two distributions.

2. NEAREST NEIGHBOR BASIS FUNCTIONS

Consider the problem in which we are given a set of data $[\mathbf{X}_N, \mathbf{y}_N]$ containing N instances, where each instance is represented by a d-dimensional feature vector $\mathbf{x}_i$ and a binary label y_i. Suppose that this data is sampled from an unknown underlying distribution, $f_{\mathbf{x}}(\mathbf{x})$, where $f_{\mathbf{x}}(\mathbf{x}) = p_0 f_0(\mathbf{x}) + p_1 f_1(\mathbf{x})$ consists of the two conditional class distributions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ for classes 0 and 1, with prior probabilities p_0 and p_1 respectively. The posterior likelihood of class 1, $\eta(\mathbf{x})$, evaluated at a point $\mathbf{x} = \mathbf{x}^*$, is

$$\eta(\mathbf{x}^*) = P[y = 1|\mathbf{x} = \mathbf{x}^*] = \frac{p_1 f_1(\mathbf{x}^*)}{f_{\mathbf{x}}(\mathbf{x}^*)}. \quad (1)$$

Many density functionals, including all f-divergences, can be expressed in terms of $\eta(\mathbf{x})$ as

$$G(f_0, f_1) = \mathbb{E}_f[g(\eta)] = \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \quad (2)$$

where $g(\eta)$ is some function corresponding to the functional $G(f_0, f_1)$. Additionally, Cover and Hart [23] showed that the error rate of a k-NN classifier trained on N samples converges to

$$\lim_{N\to\infty} R_k(\mathbf{X}_N) = R_k(\mathbf{X}_\infty) = E_f[r_k(\eta)], \quad (3)$$

where

$$r_k(\eta) = \sum_{i=\lceil \frac{k}{2} \rceil}^{k} \binom{k}{i} \left(\eta^i (1-\eta)^{k-i+1} + (1-\eta)^i \eta^{k-i+1} \right). \quad (4)$$

Rather than use these error rates to measure the performance of a k-NN classifier, as is most common, we exploit our knowledge of the asymptotic properties of these error rates to estimate other desirable quantities. We thus re-interpret the error rates of k-NN classifiers for a range of k values $k \in \mathcal{K} = [1, ..., K]$ as basis functions. Suppose that we wish to estimate some $G(f_0, f_1)$ in a regime for which traditional plug-in estimators under-perform. In our previous work [13], we identified a fitting criterion for identifying weights $\boldsymbol{\alpha} = \alpha_1, ..., \alpha_K$ that minimize the residual

$$\left| g(\eta) - \sum_{k\in\mathcal{K}} \alpha_k r_k(\eta) \right|^2, \quad (5)$$

when both the basis functions and the density functional can be expressed in terms of the posterior (as in Eqns. (2) and (4)). Using this approach, we can approximate the density functional $G(f_0, f_1)$ by

$$\hat{G}(f_0, f_1) = \sum_{k\in\mathcal{K}} \alpha_k \mathbb{E}_f[r_k(\eta)] \approx G(f_0, f_1). \quad (6)$$

Because the convergence of (3) depends only on the continuity of $f_{\mathbf{x}}(\mathbf{x})$, we can estimate $\hat{G}(f_0, f_1)$ non-parametrically with minimal assumptions on $f_{\mathbf{x}}(\mathbf{x})$. Unfortunately, reliably quantifying the assymptotic risk of a k-nearest neighbor classifier with only a finite number of samples for both training and evaluation poses several challenges. Most prominent is the finite sample bias for k-NN classifiers with limited training data which leads to glacially slow convergence at higher dimensions. To overcome this bias, the following Section investigates ensemble estimation of the k-NN asymptotic error rate to achieve fast convergence in higher dimensions.

3. ASYMPTOTIC RISK ESTIMATION

Error rate estimation is a well-studied problem and numerous solutions exist such as cross validation and bootstrapping [20, 21]; however we use a hold-out estimate in order to simplify the proofs later in this Section. We define the hold-out estimate of a k-NN classifier trained on a training set $\mathbf{X}_t$ and evaluated on a hold-out set $\mathbf{X}_h$ as

$$\hat{R}_k(\mathbf{X}_t, \mathbf{X}_h) = \frac{1}{|\mathbf{X}_h|} \sum_{\mathbf{x}_i \in \mathbf{X}_h} \left| g(\mathbf{X}_t, \mathbf{x}_i) - y_i \right| \quad (7)$$

where $g(\mathbf{X}_t; \mathbf{x}_i)$ represents the output of the k-NN classifier trained on $\mathbf{X}_t$ and evaluated on $\mathbf{x}_i$. If we assume that densities $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ 1) are absolutely continuous over $\mathbb{R}^n$, 2) have compact support, 3) have s continuous derivatives, and 4) vanish close to the boundaries; then the bias of the k-NN error rate from a classifier trained on N samples is

$$\mathbb{B}[R_k(\mathbf{X}_N)] = \sum_{j=2}^{s-1} c_j N^{-\frac{j}{d}} + \mathcal{O}(N^{-\frac{s}{d}}), \quad (8)$$

where d represents the intrinsic dimension of the distribution and the expansion constants c_j depend generally upon k, the Euclidean metric used to generate the k-NN graph, and the underlying distributions [14]. Asymptotically the first term in the sum will dominate, and the bias of $R_k(N)$ approaches zero at a rate of $\mathcal{O}(N^{-\frac{2}{d}})$. For low dimensional problems ($d \leq 4$) the squared bias matches the MSE parametric rate $\mathcal{O}(N^{-1})$. However the bias converges glacially slow at higher dimensions.

To improve the rate of convergence, we use an ensemble of finite sample estimators to cancel the lower order bias terms. First, we partition our data set into $N/2$ training samples $\mathbf{X}_t$ and $N/2$ holdout samples $\mathbf{X}_h$. We then train L k-NN classifiers on L subsets $\mathbf{X}_i$, where $\mathbf{X}_i$ represents $M_i = l_i N$ samples from $\mathbf{X}_t$ and $l_i \leq 0.5$ and $i \in \mathcal{I} = [1, \ldots, L]$. Additionally, assume that each subsample is evaluated on N held-out data points. The ensemble estimate of the error rate is

$$\Phi_k(\mathbf{X}_N, \mathbf{w}^*) = \sum_{i\in\mathcal{I}} w_i \hat{R}_k(\mathbf{X}_i, \mathbf{X}_h) \quad (9)$$

where $\mathbf{w}^* = [w_1, w_2, ..., w_L]$. If $\sum_{i\in\mathcal{I}} w_i = 1$, then the bias of this ensemble is equal to

$$\mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}^*)] = \mathbb{E}\left[\sum_{i\in\mathcal{I}} w_i \hat{R}_k(\mathbf{X}_i, \mathbf{X}_h) \right] - R_k(\infty) \quad (10)$$

$$= \sum_{i\in\mathcal{I}} w_i \left[\sum_{j=2}^{s-1} c_j M_i^{-\frac{j}{d}} + \mathcal{O}(N^{-\frac{s}{d}}) \right], \quad (11)$$

since $\hat{R}_k(\mathbf{X}_i, \mathbf{X}_h)$ is an unbiased estimate of $R_k(\mathbf{X}_i)$. If $s \geq \lceil \frac{d}{2} \rceil$, we can ensure that the bias convergence rate is of order $\mathcal{O}(N^{-\frac{1}{2}})$ by selecting weights which set all terms of $j < \lceil \frac{d}{2} \rceil$ to zero:

$$w_1, ..., w_L = \underset{w_1,\ldots,w_L}{\operatorname{argmin}} \sum_{i\in\mathcal{I}} w_i^2$$

$$\text{Subject to } \sum_{i\in\mathcal{I}} w_i = 1 \quad (12)$$

$$\sum_{i\in\mathcal{I}} w_i l_i^{-\frac{j}{d}} = 0, j \in \mathcal{J}$$

where $\mathcal{J} = [2, ..., \lceil d/2 - 1 \rceil]$. Note that as long as $L \geq \lceil d/2 - 1 \rceil$ and all l_i are assigned unique values, (12) is guaranteed to have at least one solution [15]. Additionally, since the variance of each of the subsample estimators converges at a rate of $\mathcal{O}(N^{-1})$ the variance of a linear combination of these estimators will converge at $\mathcal{O}(N^{-1})$ [15]. Thus the MSE of $\Phi_k(\mathbf{X}_N, \mathbf{w}_0)$ will converge at rate $\mathcal{O}(N^{-1})$.

This optimization criterion imposes no constraints on the magnitude of the weights and empirically leads to excessively high variance estimates, making it impractical for small N [15–17]. A suggested solution in [15–17] is to relax the optimization criteria so that, rather than setting the bias terms to zero, they are bounded by a bias threshold ϵ_1 scaled by $N^{-\frac{1}{2}}$, which is minimized subject to a fixed variance threshold which we will call ϵ_2. However, when $\epsilon_2 \leq ||w_0||_2^2$, the bias threshold ϵ_1 can become functionally dependent on N, which could potentially slow the rate of convergence. As an alternative, we propose setting the bias and variance thresholds equal to each other ($\epsilon_1 = \epsilon_2 = \epsilon$), thus allowing the variance threshold to scale with the bias threshold. This ensures that we maintain the desired rate of convergence asymptotically while still controlling the variance at lower N. The resulting fitting routine is

$$\begin{aligned} \mathbf{w}_r = w_1, ..., w_L = \underset{w_1,...,w_L}{\operatorname{argmin}} \quad & \epsilon \\ \text{Subject to } & \sum_{i\in\mathcal{I}} w_i = 1 \\ & \sum_{i\in\mathcal{I}} w_i l_i^{-\frac{j}{d}} \leq \epsilon N^{\frac{j}{d}-\frac{1}{2}} \quad , j \in \mathcal{J} \\ & \sum_{i\in\mathcal{I}} w_i^2 \leq \lambda\epsilon, \end{aligned} \tag{13}$$

where λ is a tuning parameter used to control the trade-off between minimizing the bias and the variance. In theorem 1 we show that the MSE convergence rate of this new estimator is of order $\mathcal{O}(N^{-1})$.

Theorem 1. *If there exists a set of weights* $\mathbf{w}_0$ *satisfying the constraints of* (12)*, then*

$$\mathbb{E}\Big[\big(\Phi_k(\mathbf{X}_N, \mathbf{w}_r) - R_k(\infty)\big)^2\Big] = \mathcal{O}(N^{-1}). \tag{14}$$

Proof: *Lemma 1:* If there exists a set of weights $\mathbf{w}_0$ satisfying the constraints of (12), then ϵ in (13) is bounded by

$$\epsilon \leq \frac{\epsilon^*}{\lambda} = \frac{||\mathbf{w}_0||_2^2}{\lambda}$$

Proof of Lemma 1: Suppose that $\lambda\epsilon > \epsilon^*$. Since $\mathbf{w}_0$ satisfies the constraints of (13), ϵ^* violates the minimality of ϵ. Therefore by contradiction $\epsilon \leq \epsilon^*$.

Now we can define the bias of $\Phi_k(\mathbf{X}_N, \mathbf{w}_r)$ as

$$\begin{aligned} \mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)] &= \sum_{j=2}^{\lceil \frac{d}{2}-1 \rceil} c_j \sum_{i\in\mathcal{I}} w_i M_i^{-\frac{j}{d}} + \mathcal{O}(N^{-\frac{1}{2}}) \\ &\leq \sum_{j=2}^{\lceil \frac{d}{2}-1 \rceil} c_j \epsilon N^{-\frac{1}{2}} + \mathcal{O}(N^{-\frac{1}{2}}) \end{aligned}$$

Using Lemma 1, this can be upper bounded by

$$\mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)] \leq \sum_{j=2}^{\lceil \frac{d}{2}-1 \rceil} c_j \epsilon^* N^{-\frac{1}{2}} + \mathcal{O}(N^{-\frac{1}{2}}), \tag{15}$$

thus ensuring that $\mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)] = \mathcal{O}(N^{-\frac{1}{2}})$. Similarly, we can express the variance as

$$\begin{aligned} \mathbb{V}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)] &= \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} w_i w_j \text{cov}(\hat{R}_k(\mathbf{X}_i, \mathbf{X}_h), \hat{R}_k(\mathbf{X}_j, \mathbf{X}_h)) \\ &\leq \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} w_i w_j \sqrt{\mathbb{V}[\hat{R}_k(\mathbf{X}_i, \mathbf{X}_h)]\mathbb{V}[\hat{R}_k(\mathbf{X}_j, \mathbf{X}_h)]}. \end{aligned} \tag{16}$$

Since $\hat{R}_k(\mathbf{X}_i, \mathbf{X}_h)$ is the average of $N/2$ i.i.d. Bernouli random variables, its variance can be expressed as

$$\mathbb{V}[R_k(\mathbf{X}_N, l_i)] = \frac{2R_k(\mathbf{X}_i)\big(1 - R_k(\mathbf{X}_i)\big)}{N} \leq \frac{1}{2N}, \tag{17}$$

and the total variance can be bounded by

$$\mathbb{V}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)] \leq \frac{1}{2N}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} w_i w_j = \frac{1}{2N}. \tag{18}$$

Since both $\mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)]^2$ and $\mathbb{V}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)]$ are bounded by rate $\mathcal{O}(N^{-1})$, the MSE must be bounded by this rate as well.

It is worth noting that while Theorem 1 guarantees $\mathcal{O}(N^{-1})$ convergence asymptotically, it may converge more slowly for specific ranges of N. Returning to the original problem of estimating an unknown density functional, we estimate $G(f_0, f_1)$ using a linear combination of these ensemble estimators

$$\hat{G}(\mathbf{X}, \boldsymbol{\alpha}) = \sum_{k\in\mathcal{K}} \alpha_k \Phi_k(\mathbf{X}_N, \mathbf{w}_r). \tag{19}$$

If $\hat{G}(\mathbf{X}, \boldsymbol{\alpha})$ is an asymptotically consistent estimate of $G(f_0, f_1)$, then the bias is

$$\mathbb{B}[\hat{G}] = \sum_{k\in\mathcal{K}} \alpha_k \mathbb{B}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)], \tag{20}$$

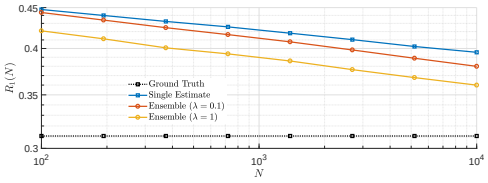
and we can bound the variance by

$$\mathbb{V}[\hat{G}] \leq \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}} \alpha_j \alpha_k \sqrt{\mathbb{V}[\Phi_j(\mathbf{X}_N, \mathbf{w}_r)]\mathbb{V}[\Phi_k(\mathbf{X}_N, \mathbf{w}_r)]}. \tag{21}$$

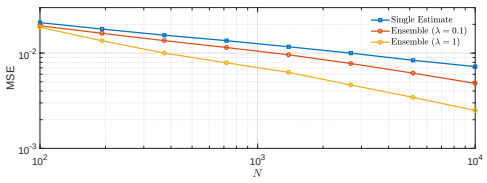
Since all of the basis functions comprising $\hat{G}(\mathbf{X}, \boldsymbol{\alpha})$ achieve an MSE convergence rate of $\mathcal{O}(N^{-1})$, the combined estimator also converges at rate $\mathcal{O}(N^{-1})$.

4. NUMERICAL RESULTS

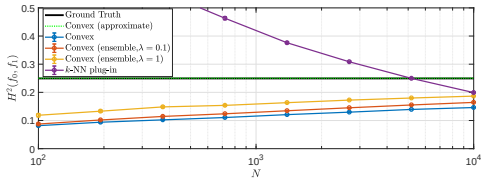
To evaluate the efficacy of the method proposed in this paper we consider the problem of estimating the Hellinger distance between two multivariate normal distributions. In this experiment each class PDF is distributed according to $f(\mathbf{x}) \sim N(\mu\mathbf{1}_d, \boldsymbol{\Sigma}_d)$, where the i, j term in $\boldsymbol{\Sigma}_d$ is defined by $\sigma_{i,j} = \beta^{|i-j|}$ for all $i, j \in [1, \ldots, d]$. We set $\mu = 0$ and $\beta = 0.8$ for distribution $f_0(\mathbf{x})$, $\mu = 1$ and $\beta = 0.9$ for distribution $f_1(\mathbf{x})$, and $d = 10$. This set of parameters offers each distribution a unique elliptical covariance structure and was identified as particularly challenging for direct estimation in [13].



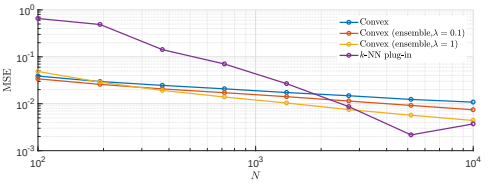
(a) Estimates of R_1



(b) MSE of R_1 estimates



(c) Estimates of $H^2(f_0, f_1)$



(d) MSE of $H^2(f_0, f_1)$ estimates

Fig. 1: (a) & (c) display the mean and standard error across sample size for each estimator of the 1-NN classifier error rate and Hellinger-squared distance respectively, while (b) & (d) display the corresponding MSE results for these estimators. These results reflect the average of 1000 Monte Carlo trials where the 12 l_i values are logarithmically spaced between 0.05 and 0.5 and $\mathbf{k} = [1, 3, 5, 7, 9]$.

At each sample size N, we compare 3 methods of estimating R_k for all $k \in \mathbf{k}$. The first method is to simply evaluate the error rate of a k-NN classifier on the held-out portion of the data. We then generate ensemble estimates using the relaxed method described in Section 3 using $\lambda = 0.1$ and $\lambda = 1$. For the $k = 1$ basis, the resulting estimates are displayed relative to the ground truth in Figure 1a, and the corresponding MSE is displayed in Figure 1b. We see that while all three estimates contain some finite-sample bias, the bias is significantly reduced using the proposed method, particularly when λ is set to prioritize minimization of the bias more heavily. As a result of the reduced bias, we see a significant improvement in the rate at which the MSE decays with N in Figure 1b. It is worth noting that, despite the improvement, the slope of the MSE for the ensemble methods here is not indicative of the $\mathcal{O}(N^{-1})$ rate guaranteed by the proposed ensemble method since the theorem does not guarantee that it converges consistently at this rate across all N. As we increase the sample size, we expect that the rate would eventually reach the expected $\mathcal{O}(N^{-1})$.

Next we evaluate the ensemble method when using $R_k(N)$ as a basis set for estimating density functionals like the squared Hellinger distance:

$$H^2(f_0, f_1) = \frac{1}{2} \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \tag{22}$$

where $g(\eta) = (\sqrt{\eta} - \sqrt{1-\eta})^2$. Estimates of the Hellinger distance are generated using the optimal fitting weights, $\boldsymbol{\alpha}$, for estimating $g(\eta)$ (using the fitting criterion in [13]) combined with the three estimates of the basis functions used in Figure 1a and 1b. Since the basis weights, $\boldsymbol{\alpha}$, are estimated using the convex optimization procedure outlined in [13], we refer to this as the convex method. In addition to the three estimates acquired in this manner, we compare against a non-parametric density (k-NN) estimation+plug-in strategy for estimating the Hellinger distance which is calculated using the universal divergence estimation approach described in [24] and implemented in the ITE toolbox [25].

Figure 1c displays the predicted values of each method relative to the ground truth. In addition to the results presented, we compare with a jackknife estimator based on equation 4.17 in [22]; however due to the magnitude of the variance in these estimates, they fall outside the range of the plots presented. These results confirm that the bias reduction achieved in the basis set translates to a similar improvement in the combined estimator. While the plug-in estimator appears to exhibit faster convegence here, observing this estimator at higher N, shows that it crosses the true value and is significantly biased. In the MSE results, we see that at lower N selecting the smaller λ value yields better performance, while the higher λ performs best for larger N. This matches our expectation that bias reduction becomes a higher priority as N increases, and a corresponding increase in λ is appropriate.

5. CONCLUSION

In this paper we consider using a weighted combination of k-NN error rates in order to estimate unknown density functionals. To improve the convergence rate of this approach, which slows dramatically at higher dimensions, we develop an ensemble estimate of the k-NN error rate which can guarantee $\mathcal{O}(N^{-1})$ convergence regardless of dimension, when the densities are sufficiently smooth. We evaluate the efficacy of this approach by estimating the Hellinger distance for a pair of multivariate Gaussian distributions on 10-dimensional data. In this scenario, our approach generally outperformed a plug-in estimator that first requires non-parametric density estimation.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge Dennis Wei and Karthikeyan Ramamurthy at IBM Watson Research Labs for their help in discussing the ideas presented in this paper. This research was supported in part by Office of Naval Research grants N000141410722 (Berisha) and N000141712826 (Berisha).

7. REFERENCES

[1] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, 2003.

[2] 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*. Springer, 2003, pp. 147–163.

[3] K. E. Hild, D. Erdogmus, and J. C. Principe, "Blind source separation using Renyi's mutual information," *Signal Processing Letters, IEEE*, vol. 8, no. 6, pp. 174–176, 2001.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[5] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

[6] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[7] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[8] Q. Wang, S. R. Kulkarni, S. Verdú *et al.*, "Universal estimation of information measures for analog sources," *Foundations and Trends® in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

[9] G. Valiant and P. Valiant, "A clt and tight lower bounds for estimating entropy." in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, no. 179, 2010, pp. 1–1.

[10] ——, "Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.

[11] A. O. Hero III and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics and Photonics, 1998, pp. 250–261.

[12] V. Berisha, A. Wisler, A. O. Hero III, and A. Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *Signal Processing, IEEE Transactions on*, vol. 64, no. 3, pp. 580–591, 2016.

[13] A. Wisler, V. Berisha, A. Spanias, and A. O. Hero, "A data-driven basis for direct estimation of functionals of distributions," *arXiv preprint arXiv:1702.06516*, 2017.

[14] D. Psaltis, R. Snapp, and S. S. Venkatesh, "On the finite sample performance of the nearest neighbor classifier," *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 820–837, 1994.

[15] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE transactions on information theory*, vol. 59, no. 7, pp. 4374–4388, 2013.

[16] K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.

[17] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Improving convergence of divergence functional ensemble estimators," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1133–1137.

[18] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.

[19] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[20] ——, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[21] A. K. Jain, R. C. Dubes, and C.-C. Chen, "Bootstrap techniques for error estimation," *IEEE transactions on pattern analysis and machine intelligence*, no. 5, pp. 628–633, 1987.

[22] W. Schucany, H. Gray, and D. Owen, "On bias reduction in estimation," *Journal of the American Statistical Association*, vol. 66, no. 335, pp. 524–533, 1971.

[23] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[24] D. J. Sutherland, L. Xiong, B. Póczos, and J. Schneider, "Kernels on sample sets via nonparametric divergence estimates," *arXiv preprint arXiv:1202.0302*, 2012.

[25] Z. Szabó, "Information theoretical estimators toolbox," *Journal of Machine Learning Research*, vol. 15, pp. 283–287, 2014.