PERFORMANCE-DRIVEN ENTROPIC INFORMATION FUSION

Kumar Sricharan[1], Raviv Raich[2], Alfred O. Hero III[1]*

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109
[2] School of EECS, Oregon State University, Corvallis, OR 97331
{kksreddy,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

Advances in technology have resulted in acquisition and subsequent fusion of data from multiple sensors of possibly different modalities. Fusing data acquired from different sensors occurs near the front end of sensing systems and therefore can become a critical bottleneck. It is therefore crucial to quantify the performance of sensor fusion. Information fusion involves estimating and optimizing an information criterion over a transformation that maps data from from one sensor data to another. It is crucial to the task of fusion to estimate divergence to a high degree of accuracy and to quantify error in the estimate. To this end, we propose a class of plug-in estimators based on k-nearest neighbor (k-NN) graphs for estimating divergence. For this class of estimators, we derive a large sample theory for the bias and variance and develop a joint central limit theorem for the distribution of the estimators over the domain of the transformation space. In this paper, we apply our theory to two applications: (i) detection of anomalies in wireless sensor networks and (ii) fusion of hyperspectral images of geographic images using intrinsic dimension.

Index Terms— Information fusion, dimension estimation, entropy estimation, k-NN density estimation, plug-in estimation, central limit theorem, confidence intervals

1. INTRODUCTION

Advances in technology have promoted acquisition and subsequent fusion of data from multiple sensors of possibly different modalities. Fusing data acquired from different sensors occurs near the front end and therefore can become a critical bottleneck. In this paper, we are specifically concerned about information-level fusion - i.e., we seek to combine the most informative features of the sensor data. Due to inherent uncertainty in information-level fusion problems, data from sensors are modeled as random variables (probability based models). Information fusion involves estimating and optimizing an information criterion over a transformation from one sensor data to another. Some important applications of information-level fusion include registration, dimension analysis and multi-modal anomaly detection.

One of the crucial components of information-level fusion is the choice of objective function. Because the data is modeled as a set of random variables, a natural choice for an objective function is information divergence. Information divergence is the distance between probability distributions of different random variables. Information divergence can be interpreted as the relative entropy of one distribution w.r.t. the other distribution. An inherent problem is that we do not have access to the probability distributions, but rather, have access only to realizations from the distributions. It is therefore crucial to the task of fusion to estimate divergence to a high degree of accuracy and to quantify estimation error. The quantification and prediction of estimator error is arguably one of the most critical parts of a successful fusion algorithm. To this end, we have derived a large sample theory for the bias and variance and developed a joint central limit theorem for this class of estimators. This large sample theory has led to the development of performance-driven algorithms for determining the true intrinsic dimension of data, detecting anomalies in a given sample, selecting weights for k-NN graphs to accurately estimate entropy and perform classification in high-dimensional data and in finding structure in data.

For concreteness, we will focus on two applications of information fusion in this paper: (i) fusing wireless sensor network data; and (ii) fusing AVIRIS hyperspectral image data. The fusion applications are described briefly in Section 2. In Section 3, we present our large sample theory for Shannon entropy fusion criteria. We apply our theory to the problem of anomaly detection in wireless sensor networks using entropic fusion. In Section 4, we extend our theory on entropy estimation to the problem of intrinsic dimension estimation in high dimensional data. We use our theory to propose a weighted k-NN graph estimator of dimension which has a parametric rate of convergence. We apply our weighted k-NN dimension estimator to fuse hyperspectral images of a given geographical location. Finally, we give our conclusions in Section 5.

2. FUSION TASKS

We fuse two different types of data in this paper. We briefly describe this data and the corresponding fusion tasks below.

2.1. Fusion of wireless sensor network data across space and time

The experiment was set up on a Mica2 platform, which consists of 14 sensor nodes randomly deployed inside and outside a lab room. Wireless sensors communicate with each other by broadcasting and the received signal strength (RSS), defined as the voltage measured by a receiver's received signal strength indicator circuit (RSSI), was recorded for each pair of transmitting and receiving nodes. There were $14 \times 13 = 182$ pairs of RSSI measurements over a 30 minute period, and each sample was acquired every 0.5 sec. During the measuring period, students walked into and out of lab at random times, which caused anomaly patterns in the RSSI measurements. Finally, a web camera was employed to record activity for ground truth. In Section 3, we will apply our theory on entropy estimation to detect these anomalies.

2.2. Fusion of AVIRIS hyperspectral image data across space and wavelength

AVIRIS is a proven instrument used for Earth Remote Sensing. It is a unique optical sensor that delivers calibrated images of the upwelling spectral radiance in 224 contiguous spectral channels (bands) with wavelengths from 400 to 2500 nanometers. The AVIRIS sensor collects data that can be used for characterization of the Earth's surface and atmosphere from geometrically coherent spectroradiometric measurements. With proper calibration, and correction for atmospheric effects, the measurements can be converted to ground reflectance data which can then be used for quantitative characterization of surface features.

In this paper, we look at AVIRIS data of Mofett field, which is a joint civil-military airport located between northern Mountain View and northern Sunnyvale, California, USA. The scanner type is nadir-viewing, whiskbroom. Moffett field is shown in the visible band in 1. In the figure, the right quarter of the image is comprised of urban areas and vegetation while the remaining plain looking regions are water bodies.

The hyperspectral response of the location at wavelengths 10, 50, 100, 160 of Mofett field is shown in 2.2. The data matrix is therefore of dimension 128x128 (pixels) x 224 wavelengths. In Section 4, we estimate the local dimension at each pixel location to classify different regions of Moffett field.

Fig. 1. Moffett field (visible band).
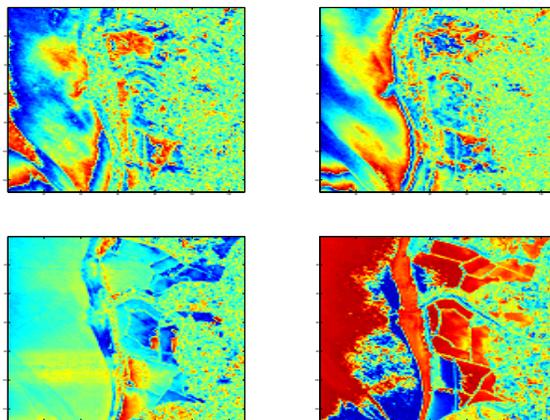
Fig. 2. Hyperspectral images at wavelengths 10, 50, 100, 160 of Mofett field.

3. ENTROPY ESTIMATION

Shannon entropy ($-\int \log f(x)f(x)dx$) arises in applications of machine learning, signal processing and statistical estimation. Entropy based applications for image matching, image registration and texture classification are developed in [1, 2]. Entropy functional estimation is fundamental to independent component analysis in signal processing [3]. Entropy has also been used in Internet anomaly detection [4] and data and image compression applications [5]. Several entropy based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [6]. Parameter estimation methods based on entropy have been developed in [7].

In these applications, the entropy must be estimated empirically from sample realizations of the underlying densities. This problem has received significant attention in the mathematical statistics community. Several estimators of Shannon entropy have been proposed for general multivariate densities f. These include consistent estimators based on entropic

graphs [8], gap estimators [9], nearest neighbor distances [10, 11], Edgeworth approximations [12], convex risk minimization [13] and kernel density estimates [14]. However, general results on rates of convergence of estimators are unavailable. Since the rate of convergence relates the number of samples to the performance of the estimator, convergence rates have great practical utility.

The results stated in this section improve upon existing results on k-NN estimators available in literature. Goria et.al. [10] shows that the estimator they propose is asymptotically unbiased and consistent. Liitiäinen et.al. [11] provides rates of convergence of the bias of these k-NN estimators. Evans et.al. [15] establish an upper bound on the rates of decay of the variance, while the authors of [8, 16] provide upper bounds on the ℓ_1 rate of convergence.

Our analysis improves on this work by establishing exact rates of decay of the bias and variance of *data-split* versions of the estimator proposed by Goria et.al. Our analysis exploits a close relation between density estimation and the geometry of proximity neighborhoods in the data sample. Finally, our theory establishes a CLT for the proposed k-NN estimators. We apply these results to derive confidence intervals for Shannon entropy.

The reminder of the section is organized as follows. Section 3.1 formulates the problem and introduces the data-split plug-in estimator. The main results concerning the bias, variance and asymptotic distribution of these estimators are stated in Section 3.3 and the consequences of these results are discussed. We validate our theory with simulations in Section 3.4. In Section 3.5, we use our theory to detect anomalies in wireless sensor networks at specified false alarm rate. Additional details on proofs and results are given in our technical report [17].

3.1. Preliminaries

Notation

We will use bold face type to indicate random variables and random vectors and regular type face for constants. We denote the expectation operator by the symbol $\mathbb{E}$ and the variance operator as $\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$. We denote the bias of an estimator by $\mathbb{B}$.

3.2. Plug-in estimators

We are interested in estimating the Shannon entropy $H(f)$ of density f, with compact support $\mathcal{S}$ which is constrained to lie on a smooth $\mathbb{C}_\infty$ manifold $\mathcal{M}$ of intrinsic dimension d. We assume that the manifold is embedded in $\mathbb{R}^D$, for some $d < D$. We assume that the support $\mathcal{S}$ does not have any boundaries.

The Shannon entropy $H(f)$ has the form

$$H(f) = \int -\log(f(x))f(x)d\mu(x) = \mathbb{E}[-\log(f(x))].$$

Here, μ denotes the Lebesgue measure and $\mathbb{E}$ denotes statistical expectation w.r.t density f. We require that the density f be uniformly bounded away from 0 and finite on the support $\mathcal{S}$, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty \ \forall x \in \mathcal{S}$. Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ be T independent and identically distributed sample realizations in $\mathbb{R}^D$ distributed according to density f. The random vectors in $\mathcal{M}$ are constrained to lie on a d-dimensional Riemannian submanifold M of $\mathbb{R}^D$ ($d < D$).

The plug-in estimator is constructed using a data splitting approach as follows. The data sample is randomly subdivided into two parts $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ of N and M points respectively. In the first stage, we estimate the k-NN density estimator $\hat{\mathbf{f}}$ at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M realizations $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$. Subsequently, we use the N samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to approximate the functional $H(f)$ to obtain the plug-in estimator:

$$\hat{\mathbf{H}}_{k,M,N}(f) = \frac{1}{N} \sum_{i=1}^{N} -\log(\hat{\mathbf{f}}(\mathbf{X}_i)).$$

Let $d(X, Y)$ denote the Euclidean distance between points X and Y and $\mathbf{d}_X^{(k)}$ denote the Euclidean distance between a point X and its k-th nearest neighbor amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. The k-NN region is $\mathbf{S}_k(X) = \{Y : d(X, Y) \leq \mathbf{d}_X^{(k)}\}$ and the volume of the k-NN region is $\mathbf{V}_k(X) = \int_{\mathbf{S}_k(X)} dZ$. The standard k-NN density estimator [18] is defined as $\hat{\mathbf{f}}(X) = \frac{k-1}{M\mathbf{V}_k(X)}$.

Define $\tilde{\mathbf{H}}_{k,M,N}(f) = \hat{\mathbf{H}}_{k,M,N}(f) + [\log(k-1) - \Psi(k-1)]$. We note that the estimators $\tilde{\mathbf{H}}$ corresponds to data-split versions of the Shannon entropy estimator of Goria et.al. [10].

3.3. Main results and consequences

We now state the main theorems corresponding to the bias, variance and asymptotic distribution of $\tilde{\mathbf{H}}_{k,M,N}(f) = \tilde{\mathbf{H}}$. We assume that k grows logarithmically in M, i.e. $k = \Theta(\log(M))$. We assume that the density f has continuous partial derivatives of order d.

The bias of $\tilde{\mathbf{H}}$ was previously derived by Liitiäinen et.al. [11]. Because $\Psi(k-1) - \log(k-1) \to 0$ as $k \to \infty$, the estimator $\tilde{\mathbf{H}}$ will have identical variance up to leading terms as $\hat{\mathbf{H}}$. Likewise, $\tilde{\mathbf{H}}$, when suitably normalized, will converge to the same distribution as the estimator $\hat{\mathbf{H}}$.

3.3.1. Bias and Variance

Theorem 3.1. *The bias of the plug-in estimator $\tilde{\mathbf{H}}_{k,M,N}(f)$ is given by*

$$\mathbb{B}(\tilde{\mathbf{H}}_{k,M,N}(f)) = \sum_{i=2}^{d} c_i \left(\frac{k}{M}\right)^{i/d} + o\left(\frac{k}{M}\right),$$

where c_i are constants which depend on the underlying density f.

Theorem 3.2. *The variance of the plug-in estimator $\tilde{\mathbf{H}}_{k,M,N}(f)$ is given by*

$$\mathbb{V}(\tilde{\mathbf{H}}_{k,M,N}(f)) = v_1\left(\frac{1}{N}\right) + v_2\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where $v_1 = \mathbb{V}[\log(f(\mathbf{Y}), \mathbf{Y})]$ and $v_2 = \mathbb{V}[\log(\mathbf{Y})g'(f(\mathbf{Y}), \mathbf{Y})]$.

Our result is an improvement on the results of Evans et.al. in that we are able to provide the exact leading terms for the variance.

3.3.2. Central limit theorem

In addition to the results on bias and variance shown in the previous section, we show that our plug-in estimator, appropriately normalized, weakly converges to the normal distribution. We study the asymptotic behavior of the plug-in estimates under the following limiting conditions: (a) $k/M \to 0$, (b) $k \to \infty$, and (c) $N \to \infty$. As shorthand, we will collectively denote the above limiting assumptions by $\Delta \to 0$.

Theorem 3.3. *The asymptotic distribution of the normalized plug-in estimator $\tilde{\mathbf{H}}_{k,M,N}(f)$ is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\tilde{\mathbf{H}}_{k,M,N}(f) - \mathbb{E}[\tilde{\mathbf{H}}_{k,M,N}(f)]}{\sqrt{\mathbb{V}[\log(f(\mathbf{Y}), \mathbf{Y})]}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ is a standard normal random variable.

3.4. Simulations

We validate our theory using using the 2 dimensional mixture density $f_m = pf_\beta + (1-p)f_u$; f_β: Beta density with parameters a=4,b=4; f_u: Uniform density; Mixing ratio $p = 0.8$. Constants c_i; $i = 1, 2..5$ are estimated using Monte-Carlo methods [19].

We show the Q-Q plot of the normalized Shannon entropy estimate and the standard normal distribution in Fig. 3. The linear Q-Q plot validates Theorem 3.3 on asymptotic normality of the plug-in estimator. Using the CLT, we plot the 95% confidence intervals for the entropy functional as a function of sample size in Fig. 4.

3.5. Anomaly detection in networks

We apply our theory to the problem of anomaly detection in wireless sensor networks. The mission of this experiment is to use the 182 RSS sequences to detect any intruders (anomalies). To remove the temperature drifts of receivers we preprocess the data by removing their local mean values. Let $y_i[n]$ be the pre-processed n-th sample of the i-th signal and denote $y[n] = (y_1[n], \ldots, y_{182}[n])'$.
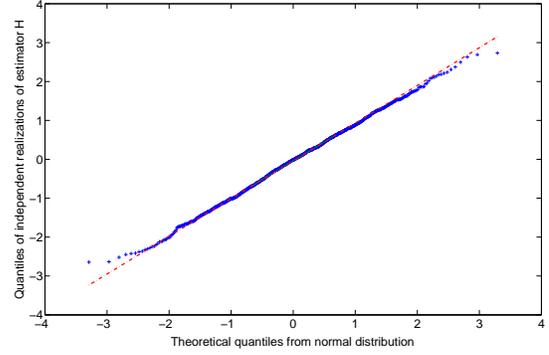
Fig. 3. q-q comparing independent realizations of the normalized Shannon estimator (L.H.S. of Theorem 3.3) on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points validates the central limit theorem.
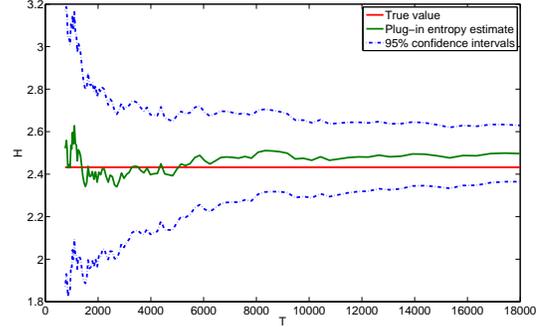
Fig. 4. Predicted confidence intervals on Shannon entropy for varying sample size T using the central limit theorem 3.3. The confidence intervals decrease with sample size as expected.

We now estimate the Shannon entropy for each 1-dimensional, 182 sample sequence $y[n]$ using the estimator $\tilde{\mathbf{H}}$. We detect anomalies by thresholding the entropy estimate $\tilde{H}[n]$. A time sample n is regarded to be anomalous if the entropy estimate $\tilde{H}[n]$ exceeds a specified threshold. We seek to choose the threshold appropriately for achieving a desired false alarm rate.

To this end, we estimate the entropies $\tilde{H}[n]$ for the time instants $n = 1, \ldots, 50$ when no anomalies were known to have occurred and subsequently estimate the mean μ and variance σ^2 of the entropy estimates for this nominal time interval $n \in [1, 50]$. Using these estimates of the mean and variance, we use the central limit theorem 3.3 to set the threshold t_α for a given false alarm rate α as $t_\alpha = \mu + z_{\alpha/2}\sigma$ where $z_{\alpha/2}$ is the z-score corresponding to coverage $1 - \alpha$. This threshold t_α is then used to detect anomalies at time instants $n > 50$.

The desired and corresponding observed false alarm rates

Desired and observed false alarm rates						
Desired	.20	.10	.05	.02	.01	.005
Observed	.269	.111	.062	.026	.015	.009
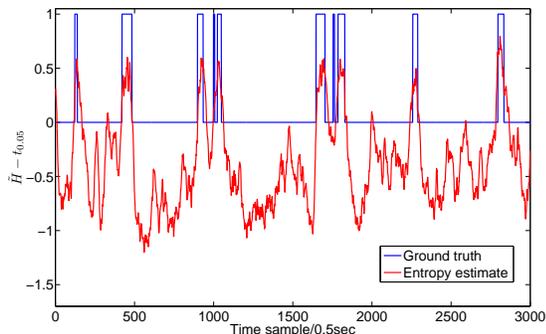
Fig. 5. Entropy estimator $\tilde{H}$ implemented as a scan statistic over time for anomaly detection in wireless ad hoc sensor network experiment. Ground truth indicator function (in blue) indicates when anomalous activity occurred. The entropy estimator detects these anomalies whenever the entropy estimate crosses the level $\alpha = 0.05$ threshold $t_{0.05}$ analytically determined by the CLT in Theorem 3.3.

are shown in the table above. The slightly higher observed false alarm rates can be attributed to the temporal dependence between the RSS sequences at successive time samples. This dependence results in marginally higher entropy estimates at non-anomalous time instants immediately preceding and succeeding anomalous time intervals as compared to entropy estimates at nominal time instants farther away from anomalous activity. This is corroborated by Fig. 5, which shows the ground truth and the normalized entropy estimator response ($\tilde{H}[n] - t_\alpha$ with false alarm rate $\alpha = 0.05$) as a function of time.

ROC curves corresponding to the entropy estimator are shown in Fig. 6 in addition to the ROC curves using the subspace method of Lakhina et.al. [4] and the covariance based estimator of Chen et.al. [20]. It is clear that the detection performance using the entropy estimator is marginally better than the subspace and covariance based methods of Lakhina et.al. and Chen et.al. respectively. The Area under the ROC curves were found to be 0.9784, 0.9722 and 0.9645 for the entropy, covariance and subspace based anomaly detection methods respectively.

4. DIMENSION ESTIMATION

Intrinsic dimensionality is an important concept in high dimensional datasets whose principal modes of variation lie on a subspace of substantially lower dimension, the intrinsic dimension d. In such cases dimensionality reduction can be accomplished without loss of information. An accurate estim-
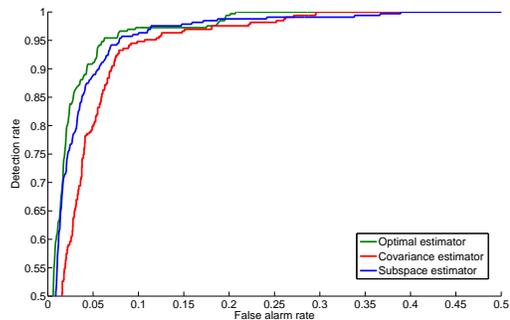
Fig. 6. ROC curves for entropy, covariance and subspace based anomaly detection. The performance of the entropy based method is the best as measured by area under the curve (0.9784 and compared to 0.9722 and 0.9645).

ator of intrinsic dimension is a prerequisite for setting the embedding dimension of DR algorithms such as principal components analysis (PCA), ISOMAP, and Laplacian eigenmaps. Until recently the most common method for selecting an embedding dimension for these algorithms was to detect a knee in a residual error curve, e.g., scree plots of sorted eigenvalues.

In this section we introduce a new dimensionality estimator that is based on fluctuations of the sizes of nearest neighbor balls centered at a subset of the data points. In this respect it is similar to Costa's k-nearest neighbor (kNN) graph dimension estimator [21], to Farahmand's dimension estimator based on nearest neighbor distances [22] and to Bickel et.al. 's maximum likelihood estimator [23]. The estimator can also be related to the Leonenko's Rényi entropy estimator [24]. However, unlike these estimators, our new dimension estimator is derived directly from a mean squared error (M.S.E.) optimality condition for partitioned k-NN estimators of multivariate density functionals. This guarantees that our estimator has a parametric M.S.E. convergence rate of $O(1/T)$ where T is the number of samples.

The section is organized as follows. We first introduce the the general form of the new dimension estimator. We then show that the estimator is related to the k-NN plug-in entropy estimator defined in the previous section. We use the theory established in the previous section to obtain expressions for the asymptotic bias and variance of the new dimension estimator. The analytical expressions for bias and variance allow us to propose a weighted k-NN dimension estimator which has a faster convergence rate of $O(1/T)$ as compared to the convergence rate of $O(1/T)^{1/d}$ in the case of the dimension estimators of Costa et.al. , Farahmand et.al. and Bickel et.al.

4.1. Problem formulation

We are interested in estimating the intrinsic dimension d, given i.i.d samples $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ in $\mathbb{R}^D$ distributed

according to density f.

4.1.1. Log-length statistic

Define the k-log-length statistic to be

$$\mathbf{L_k}(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\log\left(r_k(\mathbf{X}_i)\right),$$

where $r_k(\mathbf{X}_i)$ is the k-nearest neighbor (k-NN) distance from target sample X_i to the M reference samples $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}$ For the rest of this section, let $N = \lfloor T/2 \rfloor$ and $M = T - N$.

4.1.2. Relation to Shannon entropy

We can write the following relation

$$
\begin{aligned}
\tilde{\mathbf{H}}_{k,M,N}(f) &= \frac{1}{N}\sum_{i=1}^{N}\psi(k) - \log(c_d M) - d\log(r_k(\mathbf{X}_i)) \\
&= \psi(k) - \log(c_d M) - d\mathbf{L_k}(\mathcal{X}).
\end{aligned}
$$

Assuming $\tilde{\mathbf{H}}_{k,M,N}(f) \approx H(f)$, where $H(f)$ is the Shannon entropy, we can estimate the dimension using the following simple slope based estimator:

4.1.3. Intrinsic dimension estimate based on varying bandwidth k

Let k_1 and k_2 be two different choices of bandwidth parameters. Let $\mathbf{L_{k_1}}(\mathcal{X})$ and $\mathbf{L_{k_2}}(\mathcal{Z})$ be the length statistics evaluated at bandwidths k_1 and k_2 using data $\mathcal{X}$ and $\mathcal{Z}$ respectively.

$$\hat{\mathbf{d}} = \frac{\psi(k_2 - 1) - \psi(k_1 - 1)}{\mathbf{L_{k_2}}(\mathcal{X}) - \mathbf{L_{k_1}}(\mathcal{X})}$$

(1)

From our analysis in the previous section, we note that the bias and variance of the dimension estimator are of order $O(1/T)^{1/d}$ and $O(1/T)$ respectively.

4.1.4. Weighted dimension estimator

In this section, we improve on the previous slope based estimator by using the following weighted estimator

$$\hat{\mathbf{d}}_w = \frac{\sum_{l=1}^{k} w_l \psi(l)}{\mathbf{L}_w(\mathcal{X})}$$

where

$$\mathbf{L}_w(\mathcal{X}) = \sum_{l=1}^{k} w_l \mathbf{L_l}(\mathcal{X}),$$

and w_l, $l = 1, \ldots, k$ are the k-NN weights chosen according to the non-trivial solution $\sum_i |w_i| \neq 0$ to the optimization problem

$$
\begin{aligned}
\underset{w}{\text{minimize}} \quad & ||w||_1 \\
\text{subject to} \quad & \gamma_w(0) = 0, \; i \in 0, 1, \ldots, d.
\end{aligned}
$$

with $\gamma_w(u) = \sum_{l=1}^{k} w_l l^{u/d}$. Using this different choice of weights, we can immediately see that the new dimension estimator converges at the parametric rate of $O(1/T)$ and is therefore an immediate improvement over the estimators of Costa et.al. [21], Bickel and Levina [25] and Farahmand et.al. [22], which converge at the rate of $(1/T)^{1/d}$. The central limit theorem follows as well.
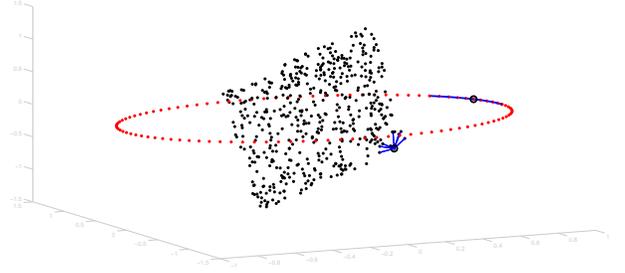
Fig. 7. Illustration of data in a sample belonging to a mixture of manifolds. The black points on the plane have intrinsic dimension 2 while the red points on the circle have intrinsic dimension 1. The blue lines depict points in the neighborhood.

4.2. Local dimension estimation

Many high-dimensional datasets of practical interest exhibit a varying complexity in different parts of the data space. This is the case, for example, of data bases of images containing many samples of a few textures of different complexity. Such phenomena can be modeled by assuming that the data lies on a collection of manifolds with different intrinsic dimensions.

Carter et. al. [26] introduce a method to estimate the local dimensionality associated with each point in a dataset, without any prior information about the manifolds, their quantity and their sampling distributions. Their proposed method uses a global dimensionality estimator based together with an algorithm for computing neighborhoods in the data with similar topological properties. They define the local dimension estimate of each point to be the intrinsic dimension of a small neighborhood sample centered at each point. Intuitively, their method takes advantage of the fact that the local neighborhoods of each point hug the respective manifolds to which the corresponding point belongs. This is illustrated in Fig. 7. The neighborhoods in the above illustration are depicted by the blue lines. It is clear from the illustration that the neighborhoods 'hug' the respective manifolds in each case.

In their work, they use the estimator of Costa et.al. to estimate the intrinsic dimension of each sample. We modify their local dimension algorithm by replacing the intrinsic dimension estimator they use with our propose weighted k-NN graph estimator.
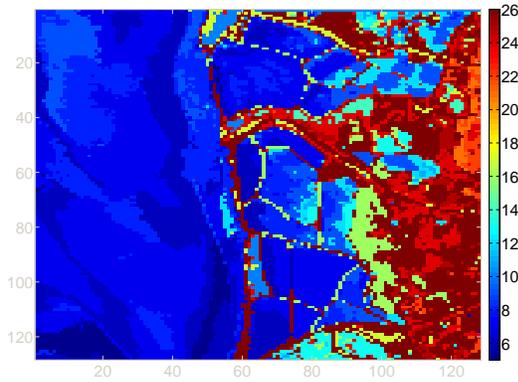
Fig. 8. Dimension estimate of Moffett field.

Fig. 9. Segmented image of Moffett field.

4.3. Dimension based image fusion

We apply local dimension estimation to do dimension based image fusion of the 224 different hyperspectral images of Moffett field. The data matrix is of dimension 128x128 (pixels) x 224 wavelengths. The local dimension estimate at each of the 128 x 128 pixel locations is shown in Fig. 8.

From the figure, we see that the urban and vegetation areas to the right have much higher intrinsic dimension as compared to the intrinsic dimension of the uniform water bodies to the left. This is in agreement with our intuition that urban and vegetation areas have a relatively complicated texture in comparison to water bodies with plain texture. The local dimension estimate can therefore successfully be used to fuse information from hyperspectral images of varying wavelengths to reveal complexity across a geographic image.

We use this image of the dimension estimate to subsequently segment the image using a simple 'canny' edge detector. This is shown in Fig. 9. The segmented image

successfully demarcates the water bodies from the urban and vegetation regions.

5. CONCLUSION

We focused on a class of data-split k-NN density plug-in estimators for estimating Shannon entropy. We derived the bias, variance and mean square error of the estimator in terms of the sample size, the dimension of the samples and the underlying probability distribution. In addition, we developed a central limit theorem for these estimators and used our theory to specify confidence intervals on the entropy. We used our entropy estimator to perform anomaly detection in wireless sensor networks and used our asymptotic theory to set thresholds appropriately to achieve specified false alarm rates.

Next, we applied our theory to the problem of dimension estimation. Dimension estimators defined in literature suffer from high bias due to curse of dimensionality. We address this problem by proposing a weighted k-NN intrinsic dimension estimator, where the optimal weights are chosen such that the resulting estimator has parametric convergence rate of $O(1/T)$. We applied our weighted k-NN dimension estimator to the problem of fusing hyperspectral images of Moffett field, which enabled us to successfully segment regions of Moffett field with different complexity.

Using the theory presented in the paper, one can therefore optimize fusion algorithms and also specify the minimum necessary sample size required to obtain requisite accuracy in fusion applications like structure discovery in graphical models and dimension estimation for support sets of low intrinsic dimension. See [17] for more details on these applications.

6. REFERENCES

[1] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, no. 5, pp. 85 – 95, sep 2002.

[2] H. Neemuchwala and A. O. Hero, "Image registration in high dimensional feature space," *Proc. of SPIE Conference on Electronic Imaging, San Jose*, January 2005.

[3] E. G. Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *Proc. 4th Intl. Symp. on ICA and BSS*, pp. 1047–1052, 2003.

[4] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *In ACM SIGCOMM*, 2005, pp. 217–228.

[5] A.K. Jain, "Image data compression: A review," *Proceedings of the IEEE*, vol. 69, no. 3, pp. 349 – 389, March 1981.

[6] O. Vasicek, "A test for normality based on sample entropy.," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 38, pp. 54–59, 1976.

[7] B. Ranneby, "The maximum spacing method. an estimation method related to the maximum likelihood method.," *Scandinavian Journal of Statistics*, vol. 11, pp. 93–112, 1984.

[8] A. O. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and alpha-entropy," *Technical Report CSPL-334 Communications and Signal Processing Laboratory, The University of Michigan*, March 2003.

[9] B. van Es, "Estimating functionals related to a density by class of statistics based on spacing," *Scandinavian Journal of Statistics*, 1992.

[10] V. V. Mergel M. N. Goria, N. N. Leonenko and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Nonparametric Statistics*, 2004.

[11] E. Liitiäinen, A. Lendasse, and F. Corona, "A boundary corrected expansion of the moments of nearest neighbor distributions," *Random Structures and Algorithms*, vol. 37, pp. 223–247, September 2010.

[12] Marc M. Van Hulle, "Edgeworth approximation of multivariate differential entropy," *Neural Computation*, vol. 17, no. 9, pp. 1903–1910, 2005.

[13] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847 –5861, November 2010.

[14] P. B. Eggermont and V. N. LaRiccia, "Best asymptotic normality of the kernel density entropy estimator for smooth densities," *Information Theory, IEEE Transactions on*, vol. 45, no. 4, pp. 1321 –1326, May 1999.

[15] D. Evans, "A law of large numbers for nearest neighbor statistics," *Proceedings of the Royal Society A*, vol. 464, pp. 3175–3192, 2008.

[16] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of R\'enyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *ArXiv e-prints*, Mar. 2010.

[17] K. Sricharan, R. Raich, and A. O. Hero III, "Empirical estimation of entropy functionals with confidence," *ArXiv e-prints*, Dec. 2010.

[18] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, 1965.

[19] V. C. Raykar and R. Duraiswami, "Fast optimal bandwidth selection for kernel density estimation," in *Proceedings of the sixth SIAM International Conference on Data Mining*, J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, Eds., 2006, pp. 524–528.

[20] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," submitted to IEEE Trans. on Signal Process., preprint available in arXiv:1009.5331.

[21] J.A. Costa, A. Girotra, and A.O. Hero, "Estimating local intrinsic dimension with k-nearest neighbor graphs," in *2005 IEEE/SP 13th Workshop on Statistical Signal Processing*, 2005, pp. 417–422.

[22] A.M. Farahmand, C. Sepesvari, and J-Y Audibert, "Manifold-adaptive dimension estimation," *Proc of 24th Intl Conf on Machine Learning*, pp. 265–272, 2007.

[23] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Advances in Neural Information Processing Systems 17*, vol. 48109, no. C, pp. 777–784, 2004.

[24] N. Leonenko, L. Prozanto, and V. Savani, "A class of rényi information estimators for multidimensional densities," *Annals of Statistics*, vol. 36, pp. 2153–2182, 2008.

[25] P. J. Bickel and Y. Ritov, "Estimating integrated squared density derivatives: Sharp best order of convergence estimates," *Sankhya: The Indian Journal of Statistics*, vol. 50, pp. 381–393, October 1988.

[26] K. M. Carter, R. Raich, and A. O. Hero, "On local intrinsic dimension estimation and its applications," *Trans. Sig. Proc.*, vol. 58, pp. 650–663, February 2010.