GLOBAL PERFORMANCE PREDICTION FOR DIVERGENCE-BASED IMAGE REGISTRATION CRITERIA

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
{kksredddy,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

Divergence measures find application in many areas of statistics, signal processing and machine learning, thus necessitating the need for good estimators of divergence measures. While several estimators of divergence measures have been proposed in literature, the performance of these estimators is not known. We propose a simple kNN density estimation based plug-in estimator for estimation of divergence measures. Based on the properties of kNN density estimates, we derive 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. Based on these results, we specify the optimal choice of tuning parameters for minimum mean square error. We also present results on convergence in distribution of the proposed estimator. These results will establish a basis for analyzing the performance of image registration methods that maximize divergence.

Index Terms— divergence estimation, performance characterization, plug-in estimators, kNN density estimators

1. INTRODUCTION

Nonparametric estimation of divergence from finite number of samples is an important tool in domains such as statistics, signal processing and machine learning. For example, the Jensen difference [1] and mutual information [2] are used as similarity measures in image registration and other pattern recognition problems.

Several estimators of divergence have been proposed in literature including estimators based on entropic graphs [1], gap estimators [3] and nearest neighbor distances [4]. The estimators proposed by these authors are shown to be asymptotically consistent. However, unlike the results provided in this article, theoretical expressions for bias, variance and confidence intervals are not available.

Theoretical performance approximations are essential for assessing the quality of different divergence estimators and for optimizing these estimators over tuning parameters. Relevant tuning parameters may include kernel width of the density estimator used to approximate the information divergence, the choice of divergence parameters, e.g, α in the Rényi divergence, and the partitioned sample size used in cross-validation.

While histogram estimators have been widely used in divergence estimation, they are inadequate when the feature dimension exceeds two or three (the standard histogram estimator uses one dimensional features). In many applications which require divergence estimation, high dimensional features have been shown to be advantageous. For example, high dimensional features have been observed to empirically improve image registration performance [1]. However, if plug-in divergence estimation is desired, high dimensional density estimation is necessary. The results of this paper will be useful in these cases.

Our method is in general applicable to all divergence measures. However, in order to illustrate our method, we restrict our attention to the Shannon mutual information (MI) measure. Shannon MI has been widely adopted as a medical image registration criterion [2]. While the primary motivation of this paper is image registration, here we focus on the estimation of the image registration criterion. We derive the mean square error and asymptotic distribution of the plug-in estimator of MI. This leads to a central limit theorem (CLT) that enables us to obtain confidence intervals on the MI estimate. Currently the only known method for obtaining such confidence intervals is to perform empirical averaging, e.g., the bootstrap. However, unlike the analysis performed in this paper, the bootstrap does not provide analytical expressions for bias, variance or confidence intervals and is therefore not useful for general performance prediction.

The rest of this paper is organized as follows. The kNN density plug-in estimator for MI estimation is introduced in Section 2. The mean square error and the asymptotic distribution of these estimators is discussed in Section 3. Based on the results in Section 3, optimal selection of bandwidth parameters and optimal partitioning of sample space is shown in Section 4. We show simulations validating our theory in Section 5 and give conclusions in Section 6.

[*] **Acknowledgement**: This work is partially funded by the National Institutes of Health, grant number 2P01CA087634-06A2 and the Air Force Office of Scientific Research, grant number FA9550-09-1-0471.

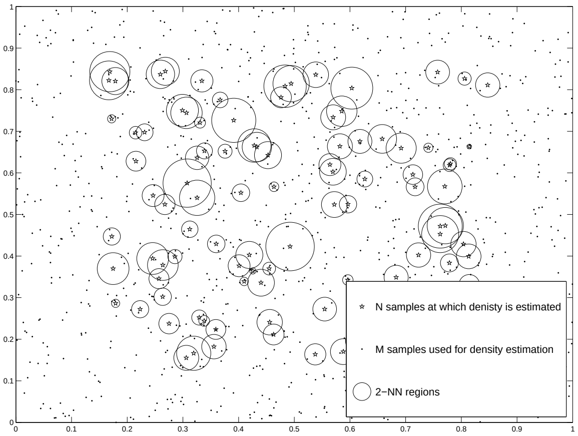

Fig. 1. Partitioning of sample space.

For more detail on the theory developed here, the reader is referred to the technical report [5]. In this paper, bold face type will be used to indicate random variables and random vectors.

2. kNN PLUG-IN ESTIMATORS

Let μ be the standard Lebesgue measure. The Shannon entropy of a random vector $\mathbf{X}$ with density function f_X is given by

$$H(\mathbf{X}) = -\int f_X \log(f_X) d\mu. \tag{1}$$

The joint entropy of random vectors $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$H(\mathbf{X},\mathbf{Y}) = -\int f_{XY} \log(f_{XY}) d\mu, \tag{2}$$

where f_{XY} is the joint density of $\mathbf{X}$ and $\mathbf{Y}$. The Shannon MI between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is then given by

$$I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}). \tag{3}$$

We use a classic plug-in estimator to estimate MI from $N+M$ d-dimensional i.i.d samples $\{(\mathbf{X_i},\mathbf{Y_i}); i = 1,\ldots,N+M\}$ of the underlying joint density f_{XY}. We estimate the Shannon MI by estimating the individual entropies. We estimate the joint Shannon entropy $H(\mathbf{X},\mathbf{Y})$ from samples using the *plug-in* estimate

$$\hat{\mathbf{H}}(\mathbf{X},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} -\log(\hat{\mathbf{f}}_{\mathbf{XY}}(\mathbf{X_i},\mathbf{Y_i})), \tag{4}$$

where $\hat{\mathbf{f}}_{\mathbf{XY}}$ is a k nearest neighbor density estimate (kNN) estimated using the remaining M samples.

The kNN density estimate [6] is given by

$$\hat{\mathbf{f}}_{\mathbf{XY}}(X,Y) = \frac{k-1}{M\mathbf{V_k}(X,Y)}, \tag{5}$$

where $\mathbf{V_k}(X,Y)$ is the volume corresponding to the kth nearest neighbor distance between the point of density estimation (X,Y) and the M i.i.d samples $\{(\mathbf{X_i},\mathbf{Y_i}); i = N+1,\ldots,N+M\}$. This partitioning of the samples is illustrated in Fig. 1.

We estimate the marginal entropies by first obtaining estimates of the marginal density using kNN density estimates

$$\hat{\mathbf{f}}_{\mathbf{X}}(X) = \frac{k-1}{M\mathbf{V_k}(X)}, \tag{6}$$

where $\mathbf{V_k}(X)$ is the volume corresponding to the kth nearest neighbor distance between the point of density estimation X and the M i.i.d samples $\{\mathbf{X_i}; i = N+1,\ldots,N+M\}$, and then plugging the estimated marginals into Eq. 7.

$$\hat{\mathbf{H}}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} -\log(\hat{\mathbf{f}}_{\mathbf{X}}(\mathbf{X_i})). \tag{7}$$

Denote the estimated MI by $\hat{\mathbf{I}}$.

$$\hat{\mathbf{I}} = \hat{\mathbf{H}}(\mathbf{X}) + \hat{\mathbf{H}}(\mathbf{Y}) - \hat{\mathbf{H}}(\mathbf{X},\mathbf{Y}). \tag{8}$$

3. PROPERTIES OF kNN PLUG-IN ESTIMATORS

We make the assumption that f_{XY} is two times continuously differentiable and is bounded away from 0. Under this assumption, we have established the following theorems [5] concerning kNN plug-in estimates :

3.1. Mean square error

Theorem 3.1. *The bias of the plug-in estimator $\hat{\mathbf{I}}$ is given by*

$$\textit{Bias}(\hat{\mathbf{I}}) = c_{b1}\left(\frac{k}{M}\right)^{2/d} + c_{b2}\left(\frac{1}{k}\right) + o\left(\left(\frac{k}{M}\right)^{2/d} + \frac{1}{k}\right),$$

where

$$c_{b1} = \mathbb{E}\Big[-c_d f_{XY}^{-(d+2)/d}(\mathbf{X},\mathbf{Y}) tr[\nabla^2(f_{XY}(\mathbf{X},\mathbf{Y}))]\Big],$$
$$c_{b2} = 0.5,$$

are constants which depend on the underlying density f_{XY} and the constant $c_d = (\Gamma^{(2/d)}((d+2)/2))/(\pi(d+2))$.

Theorem 3.2. *The variance of the plug-in estimator $\hat{\mathbf{I}}$ is given by*

$$\textit{Var}(\hat{\mathbf{I}}) = c_v\left(\frac{1}{N}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where

$$c_v = Var\left[\log\left(\frac{f_X(\mathbf{X})f_Y(\mathbf{Y})}{f_{XY}(\mathbf{X},\mathbf{Y})}\right)\right],$$

is a constant which depends on the underlying density.f_{XY}.

The mean square error (ignoring the higher order terms) is then given by

$$\text{MSE}(\hat{\mathbf{I}}) = \left(c_{b1} \left(\frac{k}{M} \right)^{2/d} + c_{b2} \left(\frac{1}{k} \right) \right)^2 + c_v \left(\frac{1}{N} \right). \quad (9)$$

3.2. Central Limit Theorem

Theorem 3.3. *Let* $\mathbf{Z}$ *be a standard normal random variable. Then,*

$$\lim_{N,M\to\infty} Pr \left(\frac{\sqrt{N}(\hat{\mathbf{I}} - \mathbb{E}[\hat{\mathbf{I}}])}{\sqrt{c_v}} \leq \alpha \right) = Pr(\mathbf{Z} \leq \alpha).$$

4. OPTIMIZATION OF PARAMETERS

Using these theorems, we can tune the kernel width k for a given sample size $N + M$ and select the optimal partitioning of the sample space for minimum mean square error.

4.1. Optimization of kernel width k

Minimizing the M.S.E. over k is equivalent to minimizing the square of the bias (and equivalently, the absolute value of the bias) over k. We observe that the constants c_{b1} and c_{b2} can possibly have opposite signs and therefore optimizing the value of k w.r.t the absolute value of bias will require higher order terms. Instead, we optimize the upper bound on the absolute value of the bias. This upper bound is given by

$$|\text{Bias}|(\hat{\mathbf{I}}) \leq |c_{b1}| \left(\frac{k}{M} \right)^{2/d} + |c_{b2}| \left(\frac{1}{k} \right). \quad (10)$$

The optimal value of k w.r.t this bound is then given by

$$k_{opt} \quad = \quad \lfloor k_0 M^{\frac{2}{2+d}} \rfloor. \quad (11)$$

where $\lfloor x \rfloor$ is the closest integer to x and the constant $k_0 = (|c_{b2}|d/2|c_{b1}|)^{\frac{d}{d+2}}$.

4.2. Optimal partitioning of sample space

The M.S.E at the optimal value of k is then given by

$$\text{MSE}(\hat{\mathbf{I}}) = b_0^2 M^{\frac{-4}{2+d}} + c_v \left(\frac{1}{N} \right). \quad (12)$$

Under the constarint that the total number of samples $T = N + M$ is fixed, the optimal choice of N as a function of M is then given by

$$N_{opt} = \lfloor N_0 M^{\frac{6+d}{2(2+d)}} \rfloor, \quad (13)$$

where the constant $N_0 = \frac{\sqrt{c_v(2+d)}}{2b_0}$.

For the optimal choices of k and N, the M.S.E in terms of M is given by

$$\text{MSE}(\hat{\mathbf{I}}) = b_0^2 M^{\frac{-4}{2+d}} + \frac{c_v}{N_0} M^{\frac{-(6+d)}{2(2+d)}}. \quad (14)$$

4.3. Discussion

1. **Choice of partition**: The optimal choice of N (the number of samples used for estimating entropy) grows at a smaller rate as compared to M (the number of samples used for density estimation). This agrees with our intuition that in higher dimensions, density estimation is the more difficult problem as compared to the problem of entropy estimation when the density is known, and therefore a greater fraction of the total realizations available should be used for estimation of the density.

2. **Choice of kernel width parameter**: The optimal k grows at a smaller rate as compared to the total number of samples M used for density estimation and this rate decreases as the dimension d increases. This can be explained by observing that the choice of k primarily controls the bias of the entropy estimator. For a fixed choice of k and M $(k < M)$, we expect the bias in the density estimates (and correspondingly in the entropy estimates) to increase as the dimension increases. For fixed M, to ensure optimal bias, we would therefore require that the density estimates are based on realizations which lie in smaller neighborhoods as the dimension increases. This in turn corresponds to choosing a smaller k relative to M as the dimension d grows.

5. SIMULATIONS

We estimated the Shannon MI of a 2 dimensional beta distribution with parameters $\alpha = 2$, $\beta = 2$ and compared our theoretical predictions with the observed bias and variance. In the first experiment, we fixed N to be 1000 and varied M. For each value of M, we optimized the kernel width k according to Eq.11. The variation of the bias of the estimator with changing M is shown in Fig. 2. In the next experiment, we fixed M to be 10000, chose the corresponding optimal value of k and varied N. The variation of the variance of the estimator against N is shown in Fig. 3. The proximity of the theoretical and emperical curves in these experiments validates our theory.

We performed the Kolmogorov-Smirnov test on the estimated MI, which resulted in the null hypothesis that the MI estimate could have the normal distribution. We generated a Q-Q plot of the MI estimate against the normal distribution. The resulting plot shown in Fig. 4 is linear, validating our theory on the asymptotic normal distribution of the plug-in estimates.

In the final experiment, we consider a mixture density $f_m = pf_\beta + (1-p)f_u$, where f_β is a beta distribution with parameters $\alpha = 2$, $\beta = 2$, f_u is a uniform density and p is the mixing ratio. We vary the mixing ratio p and evaluate the MI. The variation of the true MI and estimated MI with p is shown in Fig. 5 along with the 95% confidence intervals using Theorem 3.3. We find the estimated MI to lie within the confidence interval predicted by our theory.

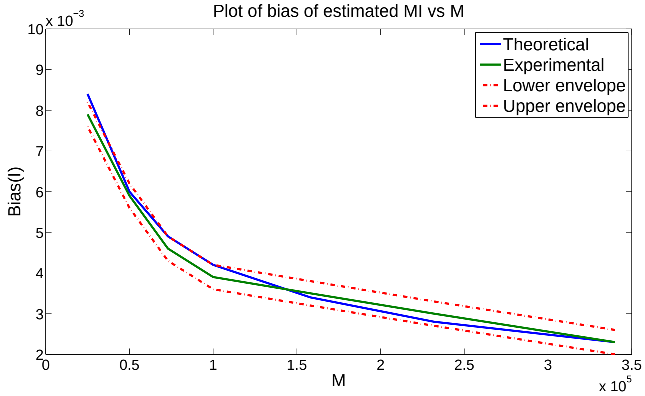

Fig. 2. Variation of bias of estimated MI vs M for fixed N = 1000 with $\pm 95\%$ confidence envelopes.

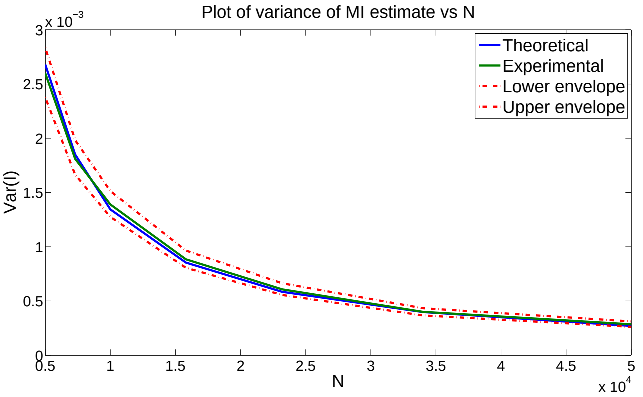

Fig. 3. Variation of variance of estimated MI vs N for fixed M = 10000 and bandwidth k = 411 with $\pm 95\%$ confidence envelopes.

6. CONCLUSION

We have obtained analytic approximations to the mean square error of the MI estimate $\hat{\mathbf{I}}$ and have shown that the estimator has an asymptotic normal distribution. The development and analysis of these kNN plug-in estimates facilitate characterization of error involved in divergence estimation in terms of the bandwidth parameters, sample size and the underlying densities (in the form of the constants $\{c_{b1}, c_{b2}, c_v\}$). As a consequence, we can specify the necessary sample size required to obtain requisite accuracy. This is not possible using current divergence estimation methods and underlines the significance of the results established in this work.

7. REFERENCES

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

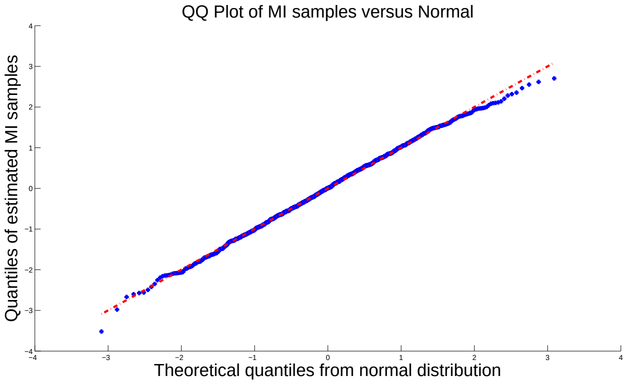

Fig. 4. Q-Q plot of normalized MI estimate and standard normal distribution.

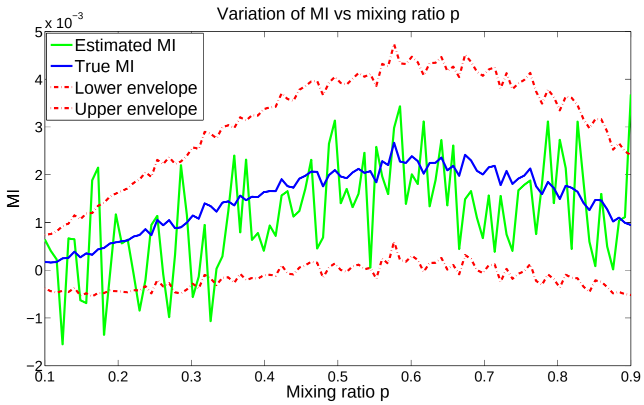

Fig. 5. Variation of MI with mixing ratio p with $\pm 95\%$ confidence envelopes.

[2] P. Viola and W.M. Wells, "Alignment by maximization of mutual information," *Proc. of 5th Int. Conf. on Computer Vision, MIT*, vol. 1, pp. 16–23, 1995.

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

[4] 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.

[5] K. Sricharan, R. Raich, and A. O. Hero, "Plug-in estimators for non-linear functionals of densities," *Technical Report, Communications and Signal Processing Laboratory, The University of Michigan*, July 2009, (To appear).

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