Rényi Information Divergence via Measure Transformations on Minimal Spanning Trees

Alfred O. Hero[1]
Dept. EECS
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122 USA
hero@eecs.umich.edu

Olivier J.J. Michel
Laboratoire de Physique,
URA-1325 CNRS,
École Normale Supérieure de Lyon,
46 allée d'Italie,
69364 Lyon Cedex 07, France
omichel@physique.ens-lyon.fr

Abstract — **We apply the results of [2] to estimation of Rényi I-divergence between an unknown distribution and a known reference distribution using power weighted pruned minimal graphs spanning a random sample of n points from the unknown distribution. In particular we establish that the weight of a minimal graph connecting the points converges a.s. in n to the I-divergence after a suitable change of measure.**

I. Introduction

Let $\mathcal{X}_n = \{x_1, x_2, \ldots, x_n\}$ denote a sample of i.i.d. data points in R^d having unknown Lebesgue multivariate density $f(x)$ supported on $[0,1]^d$. Define the order ν Rényi I-divergence [1] with respect to a dominating reference density $f_o(x)$

$$I_\nu(f, f_o) = \frac{1}{\nu - 1} \ln \int \left(\frac{f(x)}{f_o(x)}\right)^\nu f_o(x) dx \qquad (1)$$

The I-divergence takes on its minimum value (equals zero) if and only if $f = f_o$ (a.e.). $I_\nu(f, f_o)$ reduces to the Rényi entropy $H_\nu(f)$ when f_o is equal to a uniform density over $[0,1]^d$. Special cases of interest are obtained for $\nu = \frac{1}{2}$ for which one obtains the log Hellinger distance squared and for $\nu \to 1$ for which one obtains the Kullback-Liebler divergence.

II. MST's and Entropy Estimation

A spanning tree $\mathcal{T}$ through the sample $\mathcal{X}_n$ is a connected acyclic graph which passes through all the n points $\{x_i\}_i$ in the sample. $\mathcal{T}$ is specified by an ordered list of edge (Euclidean) lengths e_{ij} connecting certain pairs (x_i, x_j), $i \neq j$, along with a list of edge adjacency relations. The power weighted length of the tree $\mathcal{T}$ is the sum of all edge lengths raised to a power $\gamma \in (0, d)$, denoted by: $\sum_{e \in \mathcal{T}} |e|^\gamma$. The minimal spanning tree (MST) is the tree which has the minimal length $L(\mathcal{X}_n) = \min_{\mathcal{T}} \sum_{e \in \mathcal{T}} |e|^\gamma$. For any subset $\mathcal{X}_{n,k}$ of k points in $\mathcal{X}_n$ define $\mathcal{T}_{\mathcal{X}_{n,k}}$ the k-point MST which spans $\mathcal{X}_{n,k}$. The k-MST is defined as that k-point MST which has minimum length. Thus the k-MST spans the densest k-dimensional subset $\mathcal{X}^*_{n,k}$ of $\mathcal{X}_n$. The k-MST computation is NP complete. In [2] we presented asymptotic results for a d-dimensional extension of the planar k-MST approximation of Ravi et al, called the greedy k-MST approximation, which runs in polynomial time.

Let $\nu \in (0,1)$ be defined by $\nu = (d - \gamma)/d$ and define the statistic

$$\hat{H}_\nu(\mathcal{X}^*_{n,k}) = \frac{1}{1-\nu} \ln \left(n^{-\nu} L(\mathcal{X}^*_{n,k}) \right) + \beta(\nu, d) \qquad (2)$$

where β is a constant equal to the ν-th order Rényi entropy of the uniform density on $[0,1]^d$. Let $G(x)$ be the coordinate transformation on $[0,1]^d$ which maps the reference distribution f_o to a uniform distribution and define the transformed data sample $\mathcal{Y}_n = G(\mathcal{X}_n)$. Then using the results of [2] it can be shown that $\hat{H}_\nu(\mathcal{Y}^*_{n,n})$ is an a.s. consistent estimator of the I-divergence (1). Furthermore, with $\alpha = k/n$, $\hat{H}_\nu(\mathcal{Y}^*_{n,k})$ is an α-trimmed estimator of I-divergence in the sense that

$$\hat{H}_\nu(\mathcal{Y}^*_{n,k}) \to \min_{A: P(A) \geq \alpha} \frac{1}{1-\nu} \ln \int_A \left(\frac{f(x)}{f_o(x)}\right)^\nu f_o(x) dx \quad (a.s.) \qquad (3)$$

where the minimization is performed over all d-dimensional Borel subsets of $[0,1]^d$ having probability $P(A) = \int_A f_o(x) dx \geq \alpha$.

Let f follow the mixture model

$$f = (1 - \epsilon) f_1 + \epsilon f_o, \qquad (4)$$

where f_o is a known outlier density and f_1, $\epsilon \in [0,1]$ are unknown. Then for small ϵ and α close to one it can easily be shown that the right hand side of (3), which is $I_\nu(f, f_o)$, is to a close approximation $I_\nu(f_1, f_o)$. Thus $\hat{H}_\nu(\mathcal{Y}^*_{n,k})$ is a robust estimator of $I_\nu(f_1, f_o)$.

Note the following: the estimator $\hat{H}_\nu(\mathcal{Y}^*_{n,k})$ does not require performing the difficult step of density estimation; estimates of various orders ν of I_ν can be obtained by varying the edge power exponent; the sequence of trees $\mathcal{Y}_{n,2}, \ldots \mathcal{Y}_{n,n} = \mathcal{Y}_n$ provides a natural extension of rank order statistics for multidimensional data. Here k plays the same role as the parameter α in the α-trimmed mean estimator for 1-dimensional data.

References

[1] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, 1989.

[2] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[1] This research was supported in part by AFOSR under MURI grant F49620-97-0028.