Theory and application of entropic-graph based I-divergence estimation

Olivier Michel[†,*], Alfred 0. Hero III[⋆]
† Université de Nice Sophia-Antipolis,
Labo. d'Astrophysique (UMR 6525 CNRS),
Parc Valrose, 06108 Nice Cedex 02, France
⋆The University of Michigan,
Dept. of EECS,Ann Arbor, MI 48109-2122, USA
omichel@unice.fr,hero@eecs.umich.edu

ABSTRACT

This paper addresses the problem of robust classification of mixture densities by using an entropic-graph information divergence estimate; this provides a means to robustly estimate I-divergence without using any explicit probability density function estimation procedure. We previously applied entropic-graph methods to clustering and classification for mixture densities having uniform contamination density. This paper describes an extension of our previous methods to mixture densities with arbitrary contamination density. Under the assumption that at least one of the pdf's can be estimated from a training sample, a binary hypothesis test is proposed for testing whether an independent target sample has identical distribution as the training sample. This test is based on thresholding an entropic-graph I-divergence estimate constructed from the Minimal Spanning Tree (MST) spanning the target sample on a transformed data space.

1 Introduction

The problem of estimating the I-divergence arises in a very large class of density classification problems for clustering and pattern recognition [1, 2]. In these problems one applies a threshold test to an estimate of $I_\nu(f,g)$ in order to decide whether f is equal to g. I-divergence estimation also arises in image registration where the I-divergence can be directly related to mutual information between two images f and g [3, 4]. For an overview of entropy and I-divergence estimation applications the reader can refer to [5] and [1]. In this paper we present a methodology for robust estimation of $I_\nu(f,g)$ for unknown f and g where g is an arbitrary dominating density. It is assumed that an independent identically distributed (i.i.d.) training sample from g is available. The proposed divergence estimator performs a non-linear transformation on the data sample $\mathcal{X}_n$, producing a transformed data sample $\mathcal{Y}_n$, and constructs a graph, called the k-minimal spanning tree (k-MST), on a minimal k-point subset $\mathcal{Y}_{n,k}$ of the transformed data. Here k plays the same role as the parameter α in the α-trimmed mean estimator of the ensemble mean of a sample: when there are outliers k can be selected to ensure outlier resistance. The log length of the k-MST gives an estimator of the Rényi information divergence.

As contrasted with density plug-in techniques, graph-based entropy estimators enjoy the following properties: they can have faster asymptotic convergence rates, especially for non-smooth densities and for low dimensional feature spaces; they completely bypass the complication of choosing and fine tuning parameters such as histogram bin size, density kernel width, complexity, and adaptation speed; the α parameter in the α-entropy function is varied by varying the interpoint distance measure used to compute the weight of the minimal graph. On the other hand, the need for combinatorial optimization is a bottleneck for large numbers of feature samples. This has motivated the development of greedy minimal graph approximations that preserve advantages such as robustness against outliers [6, 7]. This paper presents an extension of [7] to the case of unknown outlier densities.

After recalling basic definitions and properties of minimal graphs in Section 2, the I-divergence estimation problem is developed in Section 3. Finally, an application of robust classification and mixture detection is presented, and ROC curves are presented which illustrate the effectiveness of our approach.

2 MST and k-MST based entropy estimators

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 $f(x_i)$ supported on $[0,1]^d$.

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)

This work was supported in part by a NATO Collaborative Linkage Grant PST.CLG.977010 ans ASFOR MURI Grant F49620-97-0028

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 a subset $\mathcal{X}^*_{n,k}$ defined by

$$L(\mathcal{X}^*_{n,k}) = \min_{\mathcal{X}_{n,k}} L(\mathcal{X}_{n,k})$$

The planar k-MST problem was shown to be NP-complete in [8]. Ravi *et al* proposed a greedy polynomial time algorithm for the planar k-MST problem with approximation ratio $O(k^{\frac{1}{4}})$.

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

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

where $\beta_{\nu,d}$ is a constant equal to the Rényi entropy of parameter ν

$$H_\nu(f) = \frac{1}{1-\nu} \ln \int f^\nu(x) dx \qquad (2)$$

for $f(x)$ equals the uniform density on $[0,1]^d$. In [9] Hero and Michel presented a d-dimensional extension of the planar k-MST approximation of Ravi et al, called the greedy k-MST approximation. In that paper we proved that when $k = \alpha n$, $\alpha \in [0,1]$, and the length $L(\mathcal{X}^*_{n,k})$ of this approximation is substituted into (1) one obtains a strongly consistent and robust estimator of the Rényi entropy (2):

$$\hat{H}_\nu(\mathcal{X}^*_{n,k}) \rightarrow \min_{A: P(A) \geq \alpha} \frac{1}{1-\nu} \ln \int_A f^\nu(x) dx \quad (a.s.)$$

where the minimization is performed over all d-dimensional Borel subsets of $[0,1]^d$ having probability $P(A) = \int_A f(x)dx \geq \alpha$. This result was used in [10] to specify robust estimators of Rényi entropy which perform outlier rejection for the case that f is a mixture density of the form

$$f = (1-\varepsilon)g + \varepsilon h \qquad (3)$$

where $\varepsilon \in [01]$, g is the uniform probability density function (pdf) over $[0,1]^d$, f, h are pdf's.

3 Extension: I-Divergence Estimation

Let $g(x)$ be a reference density on $\mathbb{R}^d$ which dominates the density $f(x)$ of a sample point $x = [x^1, \ldots, x^d]^T$ in the sense that for all x such that $g(x) = 0$ we have $f(x) = 0$. A related quantity is the α-divergence between two feature densities f and g of order $\alpha \in (0,1)$ [11, 2, 1]

$$D_\alpha(f\|g) = \frac{1}{\alpha - 1} \ln \int f^\alpha(z) g^{1-\alpha}(z) dz. \qquad (4)$$

$D_\alpha(f\|g)$ is a measure of similarity or closeness of f and g in the sense that $D_\alpha(f\|g) \geq 0$ with equality iff $f = g$ almost everywhere (a.e.). When $\alpha \rightarrow 1$ the α-divergence converges to the Kullback-Leibler divergence

$$\mathrm{KL}(f\|g) = \int_{\mathcal{Z}} g(z) \ln \frac{g(z)}{f(z)} dz$$

On the other hand, $D_{\frac{1}{2}}(f\|g)$ is the Hellinger affinity between f and g [12].

For any x such that $g(x) > 0$ let $g(x)$ have the product representation

$$g(x) = g(x^1) g(x^2|x^1) \ldots g(x^d|x^{d-1}, \ldots, x^1)$$

where $g(x^k|x^{k-1}, \ldots, x^1)$ denotes the conditional density associated with $g(x)$ of the k-th component. In what follows we will ignore the set $\{x : g(x) = 0\}$ since, as $f(x) = 0$ over this set, it has probability zero. Now consider generating the vector $y = [y^1, \ldots, y^d]^T \in \mathbb{R}^d$ by the following vector transformation

$$\begin{aligned} y^1 &= G(x^1) \qquad (5) \\ y^2 &= G(x^2|x^1) \\ &\vdots \quad \vdots \\ y^d &= G(x^d|x^{d-1}, \ldots, x^1) \end{aligned}$$

where

$$G(x^k|x^{k-1}, \ldots, x^1) = \int_{-\infty}^{x^k} g(\tilde{x}^k|x^{k-1}, \ldots, x^1) d\tilde{x}^k$$

is the cumulative conditional distribution of the k-th component, which is monotone increasing except on the zero probability set $\{x : g(x) = 0\}$. Thus, except for this probability zero set, the conditional distribution has an inverse

$$\begin{aligned} x^k &= G^{-1}(y^k|x^{k-1}, \ldots, x^1) \\ &= G^{-1}(y^k|y^{k-1}, \ldots, y^1) \end{aligned}$$

and it can be shown (via the standard Jacobian formula for transformation of variables) that the induced joint density, $h(y)$, of the vector y takes the form:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))} \qquad (6)$$

Let $L(\mathcal{Y}^*_{n,k})$ denote the length of the greedy approximation to the k-MST constructed on the transformed random variables y, where $\mathcal{Y}^*_{n,k}$ is the set of k points spanned by this k-MST approximation. Then, from the results of [9] cited in the previous section, we know that

$$\hat{H}_\nu(\mathcal{Y}^*_{n,k}) \rightarrow \frac{1}{1-\nu} \ln \int h^\nu(y) dy \quad (a.s.) \qquad (7)$$

Making the inverse transformation $y \rightarrow x$ specified by (5) in the above integral, noting that, by the Jacobian

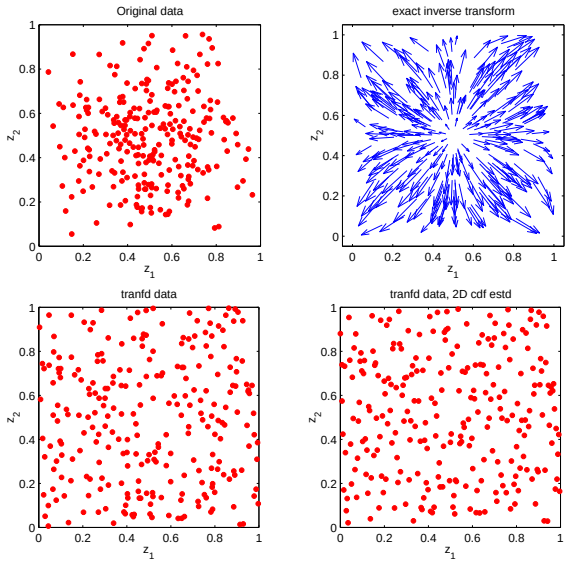

Figure 1: *Top left: a sample from a separable triangular p.d.f. over the unit square. Top right: a vector field indicating the action of the exact separable inverse transformation of coordinates on each sample point in Top right. Bottom left: same sample points as in Top left after applying transformation indicated in Top right. Bottom right same as Bottom left except that estimated transformation of coordinates was implemented using k-nearest-neighbor density estimators for each of the marginals.*

formula, $dy = g(x)dx$, and using the expression (6) for h, it easy to see that the integral in the right hand side of (7) is equivalent to the Rényi information divergence of $f(x)$ with respect to $g(x)$

$$\frac{1}{1-\nu}\ln\int h^{\nu}(y)dy = \frac{1}{1-\nu}\ln\int\left(\frac{f(x)}{g(x)}\right)^{\nu} g(x)dx.$$

Hence we have established that $\hat{H}_{\nu}(\mathcal{Y}^{*}_{n,k})$ is a strongly consistent estimator of the Rényi information divergence above. Thus the length $L(\mathcal{Y}_n)$ of the MST constructed on the transformed random variables $\mathcal{Y}_n$ can be used in place of the length $L(\mathcal{X}_n)$ in (1) to give a consistent estimate of the divergence (4) of f relative to a known reference g:

$$\hat{D}_{\alpha}(f\|g) = \frac{1}{1-\alpha}\left[\ln L(\mathcal{Y}_n)/n^{\alpha} - \ln\beta_{\nu,d}\right]. \qquad (8)$$

An example of this procedure is shown in Figure 1 for a 2D separable triangular reference density g over $[0,1]^2$ which in this case equals the actual marginal density f of the observed i.i.d. points $\mathcal{X}_n$. Thus for this example the true divergence is zero. By triangular density we mean: $g(x) = (2-4|x_1-\frac{1}{2}|)(2-4|x_2-\frac{1}{2}|)$, $x=(x_1,x_2)$. A random sample of $n=100$ points was generated from g. The uniformizing transformation in this case is separable too, with each component transformation equal to the marginal cumulative density function $G(x) = \int_0^x (2-4|z-\frac{1}{2}|)dz$ of the 1D triangular density.

We investigated both exact uniformizing transformations and approximated transformations formed from estimates of the one dimensional component density functions. The transformed sample is essentially uniform both for the exact and the estimated transformations. Therefore, as $n\to\infty$ it is expected that $L(\mathcal{Y}_n)/n^{\alpha}$ will converge to $\beta_{\nu,d}$ and the estimated divergence (8) will converge to zero as desired.

The results of [9] can thus be easily be extended to classification against any *arbitrary* distribution g, and not just the uniform distribution studied in [10]. In many practical problems occasional spurrious feature vectors may appear due to noise, false alarms, or small unimportant shifts and deformations during the image formation process. In such situations we are interested in robust entropy or divergence estimators which are resistant to these spurrious outliers. This problem is related to robust clustering for which it is common to adopt a finite mixture model to capture the incidence of points arising from different distributions [13].

4 Aplication: Robust clustering and classification

Here we apply the k-MST to robustly cluster and classify a triangular vs. uniform density. 256 samples were simulated from a uniform-triangular mixture density 3 where $g=1$ is a uniform density and h is the separable triangular shaped product density, introduced in the previous section, both supported on the unit square. Note that, unlike in the usual situation, the "outlier" distribution h has lower entropy than the target distribution g which makes the problem of clustering the realizations from g more challenging [14].

The α-divergence $D_{\alpha}(f,h)$ was estimated by $\hat{H}_{\alpha}(\mathcal{Y}_n)$ for $\alpha=\frac{1}{2}$ ($\gamma=1$) using the MST estimator. $\mathcal{Y}_n$ was obtained by applying the "uniformizing" coordinate transformation to $\mathcal{X}_n$ derived in the preceding section. In this sequence of experiments the estimate $\hat{H}_{\alpha}(\mathcal{Y}_n)$ was thresholded to decide between the hypotheses $H_0:\epsilon=0$ vs. $H_1:\epsilon\neq 0$. Simulations were performed to generate the receiver operating characteristic (ROC) curves indicated in Figures 2 for various values of ϵ, over the range $\epsilon\in\{0.1,0.3,0.5,0.7,0.9\}$.

The first plot (upper-left) gives the ROC curve obtained under the assumption that the analytical expression of the density function h is known, and is namely a triangular separable 2D distribution over $[0,1]^2$. Thus, the "uniformization" applied on $\mathcal{X}_n$ is exact. The other ROC curves were obtained for a priori unknown -but separable- distribution h, and with the hypothesis that a set of realizations with h alone is available (i.e. one can observe the case $\varepsilon=1$). For each experiment a set of 100 tests is performed, and an observation of a set of $N=256$ realizations of a process with pdf h is used to construct the uniformization function from its observed cdf. Note that, as expected, in each case the detection

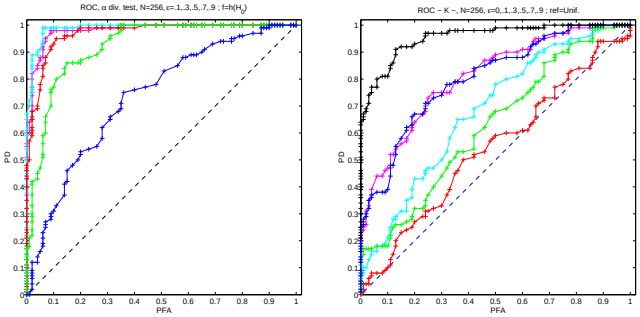

Figure 2: *TOP : ROC curves for known reference distribution (top left) or estimated reference distribution (Top right) -see text-*

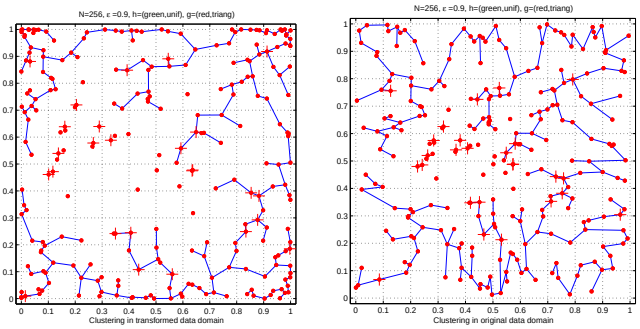

Figure 3: *Scatterplot of of a uniform-triangular mixture after applying the uniformizing coordinate transformation. Labels 'o' and '+' mark the transformed realizations from the uniform and triangular densities, respectively. Superimposed is the k-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$; Right: same as left except displayed in the original data domain.*

performance improves as the difference, indexed by ϵ, between the assumed H_0 and H_1 densities increases.

In a second sequence of experiments we selected two realizations of the triangular-uniform mixture model for the value $\epsilon = 0.1$. The k-MST procedure ($k = 90$) was implemented on $\mathcal{Y}_n$ as a robust algorithm to cluster data points from the uniform density. The cluster of points are defined as those points connected by the k-MST graph. The k-MST length can thus be used as a robust estimate $\hat{H}_\alpha(\mathcal{Y}_{n,k})$ of the uncontaminated divergence $D_\alpha(g,h)$. Figure 3 illustrates the effectiveness of this clustering method: within the cluster defined by the vertices of the k-MST the proportion of contaminating points from h has dropped from the original 10% to less than 4%.

5 Conclusion

A new approach to the I-divergence based mixture detection problem has been proposed. One of its most attractive feature is that it does not require any density estimates, at least when a training sample of the contamination density is available and when this multivariate density has independant components. For the case of more general non-separable contamination densities we are currently investigating a whitening procedure which is applied on the training data as a preprocessing step.

References

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

[2] I.Csizár, "Information-type measure of divergence of probability distribution and indirect observations," *Studia Sci. Math. Hung.*, vol.2, pp.299-318, 1967.

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

[4] B.Ma, A.O.Hero, J.Gorman, O.Michel, "Image registration with minimal spanning tree algorithm," in *Proc. of IEEE Conf. on Image Processing (ICIP'2000)*, Vancouver, CA, Oct. 2000.

[5] J.Beirlant, E.J. Dudewicz, L. Györfi and E. van der Meulen, "Nonparametric entropy estimation: an overview," Intern. J. Math. Stat. Sci., vol.16, no.1, pp.17-39, june 1997.

[6] O.Michel and A.O.Hero, "Pruned MST's for entropy estimation and outlier rejection," *Proc. of IEEE-IT workshop on DECI, Detection, Classification and Imaging*, Santa-Fe, NM, USA., Feb. 99.

[7] A.O.Hero and O.Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," *Proc. of IEEE SP Workshop on higher Order statistics*, Ceasarea, Israel, 1999.

[8] R. Ravi, M. Marathe, D. Rosenkrantz and S. Ravi, "Spanning trees - short or small," *SIAM Journal on Discrete Math.*, vol.9, pp.178-200, 1996.

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

[10] A.O.Hero and O.Michel, "Robust entropy estimation strategies based on edge weighted random graphs ,"in *Proc. International Symposium on Optical Science, Engineering and Instrumentation (SPIE)*, San Diego, CA, July 1998.

[11] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, vol.1, pp.547-561, 1961.

[12] L. Birgé and P. Massart, "Estimation of integral functions of a density," *Annals of stat.*, vol.23, pp.11-29, 1995.

[13] G.L.McLachlan and K.E.Basford, *Mixture models and Inference: application to clustering*, Marcel Dekker, New York, NY, 1988.

[14] A.O. Hero, B.Ma, O. Michel and J.D. Gorman, "Applications of Entropic Spanning Graphs," submitted to *IEEE Signal Processing Magazine*, Dec. 2001.