ROBUST ESTIMATION OF POINT PROCESS INTENSITY FEATURES USING K-MINIMAL SPANNING TREES

Alfred O. Hero

Department of EECS
University of Michigan
Ann Arbor, MI 48109, USA
hero@eecs.umich.edu

Olivier Michel

Ecole Normale Superieure de Lyon
69394, LYON cedex 07, FRANCE
omichel@ens-lyon.fr

ABSTRACT

Graph representations of multi-dimensional random processes have previously been proposed for pattern recognition [1], exploratory non-linear regression [2], clustering [3], and testing for randomness [4]. In [2] a method of robust non-linear regression was introduced which was based on pruning a minimal spanning tree (MST) in such a way as to reveal structure in a background of uniformly distributed noise. However, the pruning technique in [2] is ad hoc. In this paper we present a robust version of the MST to estimate complexity features of a point process intensity function under an epsilon contaminated model for the intensity. The principal feature considered is the Renyi entropy of the mixture and a strongly consistent entropy estimator is given which depends on the data only through the total length of the MST passing through the data points. Robustification of the MST estimator is achieved by applying the theory of k-minimum MST's [5].

1. SUMMARY

Let $dN(x)$, $x \in \mathcal{X}$, be a multi-dimensional Poisson point process having normalized multivariate intensity $\lambda(x)$ which is the mixture: $\lambda(x) = (1-\epsilon)\lambda_s(x) + \epsilon\lambda_o(x)$, $0 \leq \epsilon < 1$. In this work we apply minimal spanning trees to estimate invariant features of the "signal component" λ_s when the "noise component" λ_o is a uniform intensity and ϵ is unknown. This is closely related to the problem of extracting features of a probability density function belonging to a family of mixture distributions. A (normalized) spanning tree is a connected acyclic graph which passes through all coordinates associated with the point cloud generated by the process. It consists of an ordered list of normalized edge lengths along with a list of edge adjacency relations. The total length of the tree is defined as the sum of all edge lengths. The minimal spanning tree (MST) is the spanning tree which posesses minimal total length. It is evident that the MST representation of a point process is naturally translation, scale and rotation invariant. This invariance property is important for image processing applications where patterns of interest may be articulated at arbitrary orientations and spatial positions, e.g. as occurs in automated pattern matching, word spotting, radar target detection, and industrial inspection.

In this paper we will focus on a specific set of features which are related to intrinsic complexity of the intensity function λ_s. These features are defined in terms of the Renyi entropy $H_\alpha(dN)$ of fractional order α, $0 < \alpha \leq 1$, associated with λ_s

$$H_\alpha(\lambda_s) = \frac{1}{1-\alpha} \ln \int_{\mathcal{X}} \lambda_s^\alpha(x) dx. \tag{1}$$

The Renyi entropy equals zero for $\alpha = 0$ and converges to the Shannon entropy $-\int_{\mathcal{X}} \lambda_s \ln \lambda_s$ as α approaches 1. For any fractional order α the Renyi entropy is maximized for a uniform intensity and minimized for an intensity concentrated at a single point.

The objective of this paper is to specify robust estimates of the Renyi entropy (1) of λ_s which are insensitive to small amounts of contaminating noise. Standard techniques using histogram or adaptive kernel methods of nonparametric density estimation applied to λ are not robust to the noise contamination component λ_o of the intensity. The contribution of this paper is the application of k-minimum MST's to give robust estimates of H_α which do not require performing nonparametric density estimation.

Assume that a MST has been constructed on the n points lying in an observed point cloud in $\mathbb{R}^p$ and define $\{l_i\}_{i=1}^{n-1}$ as the set of lengths of the $n-1$ edges or vertices of the MST. We next define the fractional power weighted length of these edge lengths

$$L_n^\alpha = \sum_{i=1}^{n-1} l_i^{(1-\alpha)p}, \tag{2}$$

where as above $0 < \alpha < 1$ and p is the dimension of the measurement space. For example when $\alpha = \frac{p-1}{p}$, $L_n^\alpha = L_n$ is the total Euclidean length of the MST, and more generally when $\alpha = \frac{p-d}{p}$, d a positive integer, $(L_n^\alpha)^{\frac{1}{d}}$ is the Euclidean l_d norm of the lengths of the

[1]This research was supported in part by AFOSR grant F49620-96-0028

MST vertices. By convention, when $(1-\alpha)p$ is rational $l_i^{(1-\alpha)p} = \left(l_i^{\frac{1}{q}}\right)^d$, d, q positive integers, $l_i^{\frac{1}{q}}$ denotes the positive q-th root of l_i.

The basis for our work is the following estimation formula for the Renyi entropy (1) of fractional order α, $0 < \alpha < 1$,

$$\hat{H}_\alpha(\lambda) = \frac{1}{1-\alpha} \ln\left(n^{-\alpha} L_n^\alpha\right) + \beta(\alpha, p), \qquad (3)$$

where β is a constant independent of λ. It follows directly from the results of Steele [6] that the MST complexity estimate $\hat{H}_\alpha$ defined by (3) is a strongly consistent estimator of H_α, i.e. $\hat{H}_\alpha$ converges to H_α with probability one as the average number of observed points n goes to infinity.

When a point cloud has a total of n points, the k-minimum spanning tree is the minimum length MST among those that any k of the n points - thus the standard MST is equivalent to the n-minimum MST. The k-minimum tree achieves robustness to uniform noise since such noise will tend to produce points whose neighborhoods of "nearest neighbors" are larger than the corresponding neighborhoods of signal points. Thus exchanging a noise point with a signal point will typically reduce the total MST length. By looking for the knee in the k-minimum MST total edge length curve, plotted as a function of k, we are able to identify the best estimate of the number of offending noise points $N-k$; the corresponding k-minimum MST effectively eliminates these from the point cloud. As compared to the length L_n^α of the original MST, the length $L_{n,k}^\alpha$ of this k-minimum MST gives a robust estimate of the Renyi entropy of λ_s when used in (3).

2. NUMERICAL EXAMPLES

In Fig. 1 is a set of 100 random points drawn from the uniform distribution $f_o(x)$ over a rectangle in the plane $[0, 15] \times [0, 15]$. Figure 2 shows a graph of the Euclidean lengths ($\alpha = \frac{p-1}{p}$) of the k-minimum MST for $k \in \{0, 1, \ldots, 75\}$ computed from these points. Note that this curve is well approximated by a straight line with constant slope (regression line not shown). Figure 3 shows a set of 100 random points on the plane drawn from the mixture distribution $(1-\epsilon)f_1(x) + \epsilon f_o(x)$ where, $\epsilon = 1/2$, f_o is uniform as above and $f_1(x)$ is a density whose constant contours are annuli centered at the point $(7.5, 7.5)$ with mode occurring at radius 4 and of standard width unity about the mode (see FIg. 4. Roughly half of the points in Fig. 3 arise from f_1 and the other half from f_o. Figure 5 is a plot of the length of the k-MST as a function of k over the range $k \in \{0, 1, \ldots, 75\}$. Note the presence of a breakpoint or knee in this curve which occurs in the vicinity of $k = 34$. In Fig. 7 we show the corresponding 34 node spanning tree. Observe that when k is chosen as the location of the knee of the length curve the k-minimum tree effectively extracts the points arising from the annular density f_1 and eliminates most of those arising from the uniform noise background f_o.

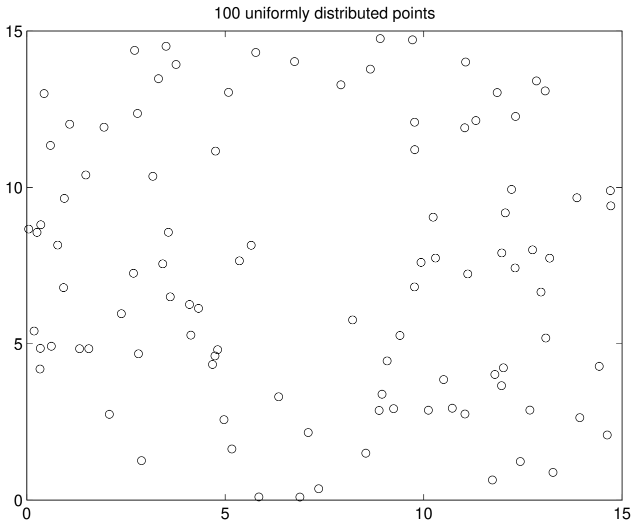

Figure 1: 100 points drawn from the uniform distribution on the plane.

The Renyi entropy of any fractional order α of the uniform distribution is zero while that of the annular distribution is given in Table ?? for fractional orders $\alpha = 1/2, 1/3, 2/3, 1/4, 3/4$. By performing multiple (50) experiments we generated estimates of the Renyi entropy of these fractional orders from the power weighted length of the full MST. The estimator means and variances are shown in Figs ? for a range of mixture parameters $\epsilon \in [0, 1]$. Fig ? shows corresponding robust estimator curves obtained by using the k-minimum MST and a visual "breakpoint selection" rule explained above. Thus a more or less unbiased estimate results from appying the k-minimum MST Renyi entropy estimation method, however at the expense of an increase in variance due to reduction in sample statistics from elimination of from 25 to 35 percent of the nodes.

3. APPROXIMATE LINEAR-TIME K-MINIMUM MST

Solving the k-minimum MST problem is closely related to solving the traveling salesman problem and

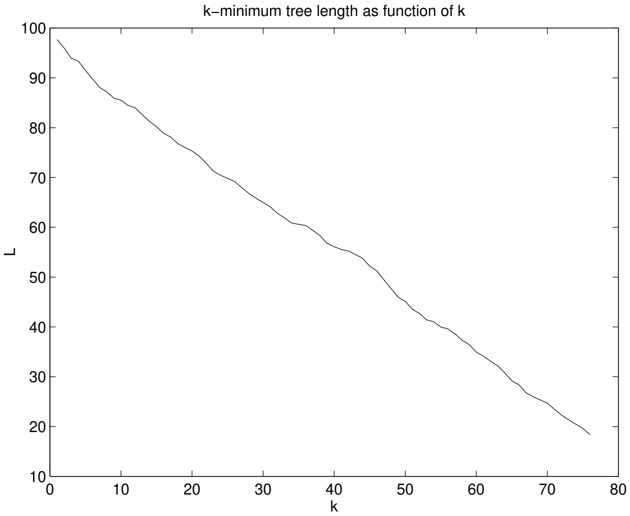

Figure 2: Length of k-minimum MST plotted as a function of k for points shown in Fig. 1 drawn from uniform distribution.

Figure 4: Annular density f_1 centered at $(7.5, 7.5)$ with mode at radius 4 and standard width equal to unity.

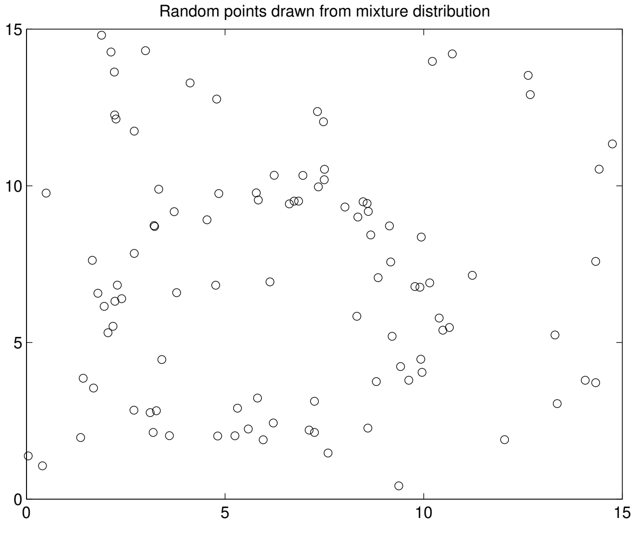

Figure 3: 100 points drawn from mixture density $\frac{1}{2}f_1 + \frac{1}{2}f_o$ on the plane. Here f_o is uniform and f_1 is as shown in figure Fig. 4

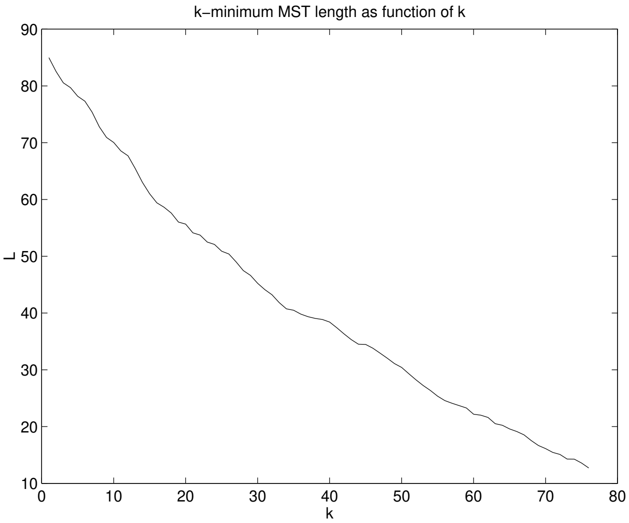

Figure 5: Length of the k-minimum MST plotted as a function of k for mixture data.

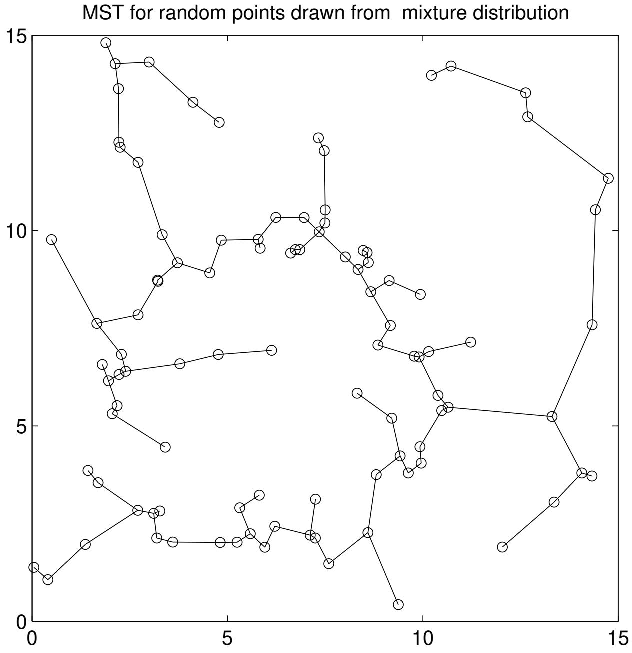

Figure 6: The minimal spanning tree for points shown in Fig. 3 drawn from mixture distribution.

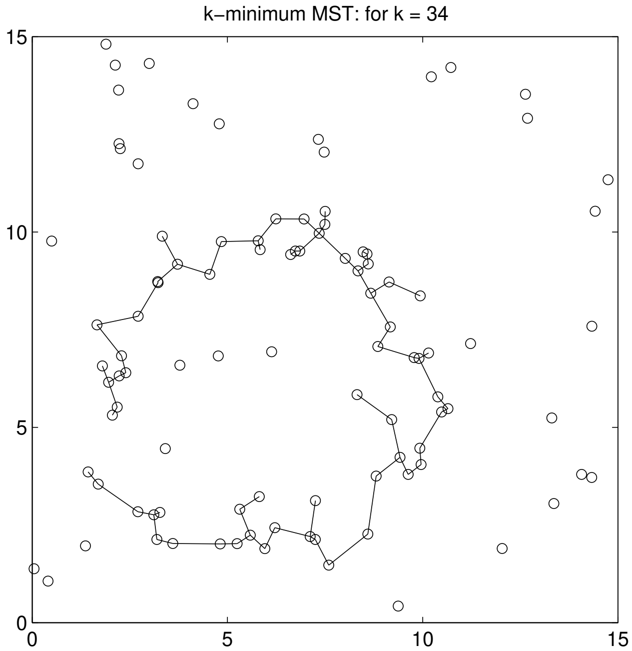

Figure 7: The k-minimum MST for $k = 34$ for points shown in Fig. 3 drawn from mixture distribution.

then selectively removing the $n - k$ customers which most reduce the total travel distance of the salesman. Like the traveling salesman problem, the k-MST problem is known to be NP-complete(?) and therefore the exact solution is very time consuming for even moderately large numbers of points. To circumvent this difficulty we implement approximate k-minimum algorithms which run in linear time. We next implement the algorithm presented in [7] which has a run time which is approximately linear in n and k, and has a quantifiable error.

4. CONCLUSION

Some problems which remain. We need to develop an objective method to find the breakpoint of the length curve in order to select the best pruning of the MST from the k-minimal MST's. A piecewise linear least squares fit of this curve does not give a very good estimate of this breakpoint. We expect a picewise exponential fit to provide better results. It would be very useful to characterize the statistics of the L_n^α curve for various values of α. This is a topic for future work.

5. REFERENCES

[1] G. Toussaint, "The relative neighborhood graph of a finite planar set," *Pattern Recognition*, pp. 261–268, 1980.

[2] P. Banks, "The minimal spanning tree for nonparametric regression and structure discovery," in *Book of Abstracts of the 1996 Meeting of the Classification Society of North America*, p. 54, 1996.

[3] C. Zahn, "Graph-theoretical methods for detecting and describing Gestalt clusters," *IEEE Trans. on Computers*, pp. 68–86, 1971.

[4] R. Hoffman and A. K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognition Letters*, pp. 175–180, 1983.

[5] R. Ravi, R. Sundaram, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees short and small,", pp. 546–555, 1994.

[6] J. M. Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *Ann. Probab.*, pp. 1767–1787, 1988.

[7] D. Eppstein, "Faster geometric k-point MST approximation," Technical Report 95-13, Dept. of Information and Computer Science, University of California, Irvine, mar 1995.