MULTIVARIATE EDGEWORTH-BASED ENTROPY ESTIMATION

Marc M. Van Hulle

K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg, Herestraat 49, B-3000 Leuven, BELGIUM
E-mail: marc@neuro.kuleuven.ac.be

ABSTRACT

We develop the general, multivariate case of the Edgeworth approximation of differential entropy, and introduce an approximate formula for Gaussian mixture densities. We use these entropy approximations in a new algorithm for selecting the optimal number of clusters in a data set, and in a new mutual information test with which one can statistically decide whether a distribution can be factorized along a given set of axes.

1. INTRODUCTION

The approximation of the one-dimensional differential entropy [1] based on the one-dimensional Edgeworth expansion of a Gaussian became popular ever since it was introduced in Independent Component Analysis (ICA) [2, 3] and projection pursuit [4]. However, concerns have been raised against the Edgeworth expansion: it would not estimate well the structure near the centroid of the density [5], and it would be sensitive to outliers [6]. What happens when we would consider the general case of multi-dimensional entropy estimation, based on the multi-dimensional Edgeworth expansion?

In this article, we will develop the multivariate Edgeworth approximation of differential entropy. We will compare its performance for a number of popular unimodal densities and entropy estimation techniques, and introduce an approximate formula for Gaussian mixture densities. Instead of Gaussian mixtures we can also consider clustered data: we will develop a new algorithm for selecting the optimal number of clusters in a data set, based on Edgeworth approximations of entropy. Finally, we will develop a mutual information test with which one can statistically decide whether a distribution can be factorized along its coordinate axes, or any other given set of axes.

The author is supported by research grants received from the Belgian Fund for Scientific Research – Flanders (G.0248.03 and G.0234.04), the Interuniversity Attraction Poles Programme – Belgian Science Policy (IUAP P5/04), the Flemish Regional Ministry of Education (Belgium) (GOA 2000/11), and the European Commission (IST-2002-001917, NEST-2003-012963).

2. EDGEWORTH EXPANSION

The Edgeworth expansion of the density $p(\mathbf{v})$, $\mathbf{v} = [v_1, \ldots, v_d] \in V \subseteq \mathcal{R}^d$, up to the fifth order about its best normal estimate ϕ_p (*i.e.*, with the same mean and covariance matrix as p) is given by [7] (also called Gram-Charlier A series):

$$p(\mathbf{v}) \approx \phi_p(\mathbf{v}) \left(1 + \frac{1}{3!} \sum_{i,j,k} \kappa^{i,j,k} h_{ijk}(\mathbf{v}) + \kappa \right), \qquad (1)$$

with h_{ijk} the ijk-th Hermite polynomial and $\kappa^{i,j,k} = \frac{\kappa^{ijk}}{\sqrt{s_i^2 s_j^2 s_k^2}}$, where κ^{ijk} is the sample third cumulant over input dimensions i, j, k, and s_i the sample second central moment over input dimension i. The term κ collects the terms in h_{ijkl} and h_{ijklpq} that will drop out in the first-order entropy approximation developed next.

Since the differential entropy $H(p) = H(\phi_p) - J(p)$, with the latter the negentropy, we obtain the following approximation:

$$\begin{aligned} H(p) &= H(\phi_p) - \int_V p(\mathbf{v}) \log \frac{p(\mathbf{v})}{\phi_p(\mathbf{v})} d\mathbf{v} \\ &\approx H(\phi_p) - \int_V \phi_p(\mathbf{v})(1 + Z(\mathbf{v})) \times \\ &\quad \log(1 + Z(\mathbf{v})) d\mathbf{v} \\ &\approx H(\phi_p) - \int_V \phi_p(\mathbf{v})(Z(\mathbf{v}) + 0.5 Z(\mathbf{v})^2) d\mathbf{v} \\ &= H(\phi_p) - \frac{1}{12} \left(\sum_{i=1}^{d} (\kappa^{i,i,i})^2 + 3 \sum_{i,j=1, i \neq j}^{d} (\kappa^{i,i,j})^2 \right. \\ &\quad \left. + \frac{1}{6} \sum_{i,j,k=1, i<j<k}^{d} (\kappa^{i,j,k})^2 \right) \qquad (2) \end{aligned}$$

which converges on the order of $\mathcal{O}(N^{-2})$, with N the number of data points in the sample, with $Z(\mathbf{v}) = \frac{1}{3!} \times \sum_{i,j,k} \kappa^{i,j,k} h_{ijk}(\mathbf{v})$, which is obtained after retaining the dominant terms and using $\int_V \phi_p(\mathbf{v}) Z(\mathbf{v}) d\mathbf{v} = 0$ and the orthogonality properties of the Hermite polynomials. The

term $H(\phi_p)$ is the familiar expression for the d-dimensional entropy: $H(\phi_p) = 0.5 \log |\Sigma| + \frac{d}{2} \log 2\pi + \frac{d}{2}$, where $|.|$ stands for determinant.

As a 1D example, we consider the standard normal distribution and the exponential distribution with $\lambda = 1$, so that the two distributions have the same unit variance, but their entropies differ: 1.419 and 1 (nats), respectively. We estimate the differential entropy using 1-spacings estimates [8], nearest-neighbor estimates [9], and Edgeworth estimates. The result is shown in Fig. 1A. We observe that the spacings estimate converges the slowest. We observe that, for the exponential, the Edgeworth-based estimates are biased, a logical consequence of truncating the Edgeworth series beyond the third cumulants. Furthermore, we consider entropy estimation as a function of the input dimensionality d. We consider the same distributions as in the 1D case, but now independently along each dimension d, and take $N = 1000$. We perform 1000 runs and plot the average ratio $H(p)/d$ (Fig. 1B). We observe for the Edgeworth-based estimates of the multivariate exponential distribution that the bias in $H(p)/d$ stays constant. More importantly, we see that the nearest-neighbor result for the exponential distribution diverges towards the Gaussian case (which is reached around $d \approx 30$). We examined this further and also tried a multivariate Gamma distribution ($\alpha = 2$, result not shown): also in this case, the nearest-neighbor result diverges. This suggests that the divergence of the multivariate exponential distribution is not due to the discontinuity at its origin but rather its tails, which become more important in higher dimensions.

Finally, we compare the time-complexities of the latter two methods. Since the nearest-neighbor method requires, for every data point, a search for the closest other data point in the data set, its complexity in the one-dimensional case is on the order of $\mathcal{O}(N^2)$; the Edgeworth method's complexity is on the order of $\mathcal{O}(N)$. In the d-dimensional case, we will have $\mathcal{O}(N^2 d)$ and $\mathcal{O}(N d^3)$, respectively.

2.1. Gaussian mixture density

The Edgeworth approximation we developed can only be reasonably applied to unimodal input densities. However, we can develop an Edgeworth approximation specifically for Gaussian mixture densities, $\tilde{p}(\mathbf{v}) = \frac{1}{N_c} \sum_{i=1}^{N_c} \phi_i(\mathbf{v})$, when the mixtures are far enough apart since then $\sum_i \phi_i \approx \phi_i(1 + Z)$, in the area where the ith (truncated) component in the mixture is the largest. Hence, we have that:

$$\begin{aligned} H(\tilde{p}) &= -\int_V \tilde{p}(\mathbf{v}) \log \tilde{p}(\mathbf{v}) d\mathbf{v} \\ &\approx -\frac{1}{N_c} \sum_i \int_{V_i} \phi_i(\mathbf{v}) \log \frac{\phi_i(\mathbf{v})}{N_c} (1 + Z_{\tilde{p}}) d\mathbf{v} \end{aligned}$$

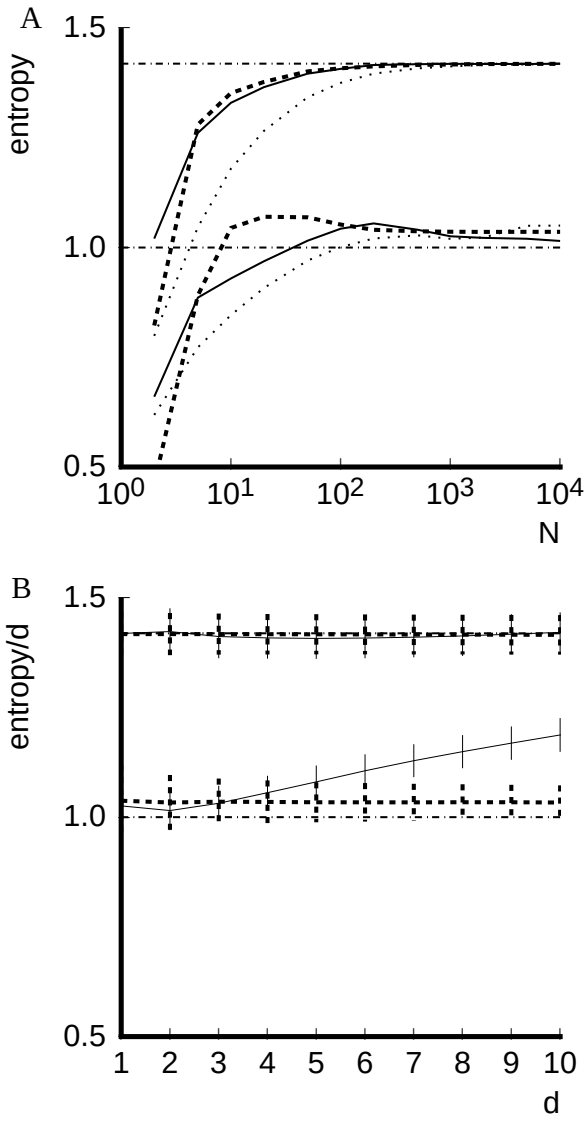

Fig. 1. (A) Differential entropy estimation as a function of sample size N for the 1D case (in nats), using 1-spacings (dotted lines), nearest-neighbor (solid lines), and Edgeworth estimates (dashed lines) for Gaussian and exponential densities (top and bottom curves). Each curve is an average over 1000 runs. The theoretical values are indicated by the stippled lines. The standard deviations are on the order of 0.05 for $N = 10000$, and are not shown for clarity's sake. (B) Differential entropy estimation as a function of the dimension d given $N = 1000$. Same line conventions as in (A). The vertical bars represent standard deviations.

$$\approx \frac{1}{N_c}\sum_i H(\phi_i)_{V_i} + \log N_c + \frac{1}{12N_c}\sum_k \left(\sum_{i=1}^{d} (\tilde{\kappa}^{i,i,i})^2 + 3 \sum_{i,j=1,i\neq j}^{d} (\tilde{\kappa}^{i,i,j})^2 + \frac{1}{6} \sum_{i,j,k=1,i<j<k}^{d} (\tilde{\kappa}^{i,j,k})^2 \right)_{V_k}, \quad (3)$$

where we define the domain of integration as $V_i = \{\mathbf{v}|\, i = \arg\min_j \phi_j(\mathbf{v})\}$. (We have taken a homogeneous mixture, but it can be readily extended to a heterogeneous mixture, see further.) The accuracy of this adapted entropy estimate is shown in Fig. 2 for $N = 1000$ data points generated from $\tilde{p}$ which consists of N_c unit variance Gaussians positioned along the real line and separated by a given distance β. The adapted entropy estimate is also compared to our original ("unimodal") Edgeworth approximation. We observe that the unimodal approximation increases without bound when β increases. (This can be understood by observing that the variance of the distribution becomes larger as the interkernel distance β becomes larger, and this variance is the largest contribution in the expansion.) The adapted version converges to the correct solution for an infinite interkernel distance β, $H(\tilde{p}) = \frac{1}{N_c}\sum_i H(\phi_i) + \log N_c$, and becomes better for smaller β as N_c increases. As an heuristic, one could compute the Edgeworth approximation both ways and take the minimum value. As a result of this, the error with respect to the correct entropy will be minimal, and the computationally expensive calculation of $H(\tilde{p})$ using the nearest-neighbor method ($\mathcal{O}(N^2 d)$ *vs.* $\mathcal{O}(Nd^3)$), or the slow convergence using plug-in estimates [10] ($\mathcal{O}(N^{-\frac{1}{2}})$ *vs.* $\mathcal{O}(N^{-2})$), can be avoided.

3. CLUSTERING

Consider the selection of the optimal number of clusters in a data set, an important issue in exploratory data analysis, since many clustering algorithms require such prior knowledge [11, 12, 13, 14]. Assume that the clusters are determined with an algorithm that assigns data points to single clusters. As observed in Fig. 2 for the two kernel case, for a kernel center distance β larger than about 3, the Edgeworth entropy approximation for the whole data set becomes larger than the clustered data approximation eq. (3). (Note that, theoretically, the border between the unimodal and bimodal cases is exactly at $\beta = 2$, for univariate Gaussians with equal variances.)

Hence, we develop the following clustering strategy. Assuming that we have k clusters ($k > 1$), we determine, for each pair of clusters i and j, the Edgeworth approximation for the set of data points belonging to the two clusters, H_{ij},

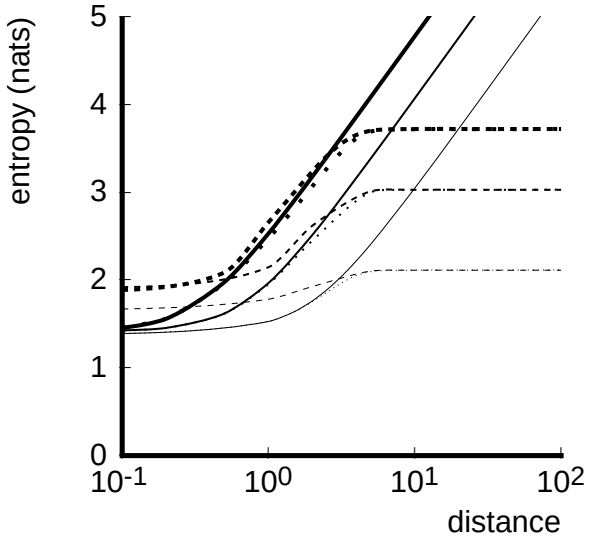

Fig. 2. Accuracy of the Edgeworth approximation of the differential entropy (full lines) as a function of the distance between 2, 5 and 10 univariate Gaussian kernels (increasing line thicknesses). Also shown are the adapted Edgeworth approximation (dashed lines) and the nearest-neighbor estimate (dotted lines). The entropy computed by numerical integration practically coincides with the nearest-neighbor estimate and is not shown. A unit distance corresponds to the unit standard deviation of the Gaussian kernels.

and the entropy according to eq. (3), which we need to modify as follows, since the clusters do not necessarily contain the same number of data points:

$$H^c_{ij}(\tilde{p}) \approx \sum_{k=i,j} \frac{N_k}{N_i+N_j}\left(H(\phi_k)_{V_k} - \log\frac{N_k}{N_i+N_j}\right) + \frac{1}{12}\sum_{k=i,j}\frac{N_k}{N_i+N_j} \times \left(\sum_{i=1}^{d} (\tilde{\kappa}^{i,i,i})^2 + 3\sum_{i,j=1,i\neq j}^{d} (\tilde{\kappa}^{i,i,j})^2 + \frac{1}{6}\sum_{i,j,k=1,i<j<k}^{d} (\tilde{\kappa}^{i,j,k})^2 \right)_{V_k},$$

with V_i the data points that belong to cluster i and N_i their number. As soon as there is a cluster pair i, j for which the entropy H_{ij} is not larger than the entropy H^c_{ij}, we have only one cluster, and conclude that k cannot be the true number of clusters. Hence, we consider the largest k-value, that is not rejected, as the true number of clusters; when there is more than one k-value accepted, we probably have a hierarchical cluster structure. The $k = 1$ case is inferred from the rejection of the $k = 2$ case. Note that covering the $k = 1$

case is an important issue for clustering algorithms (for a discussion, see [12]).

We take two real-world examples. Both are available from the UCI Machine Learning Repository (`http://www-1.ics.uci.edu/~mlearn/MLRepository.html`). The first example is the IRIS plants data set which consists of 3 classes of 50 instances with four attributes (*i.e.*, a 4 dimensional data set). Each class refers to a type of iris plant. One class is linearly separable from the other two; the latter two are not linearly separable from each other. The second example is the Wisconsin breast cancer data set, which consists of nine-dimensional measurements of 699 patients from two classes (benign and malignant). We assume that the clusters are determined with the k-means clustering algorithm. We perform 20 runs of the clustering algorithm, using different (prototype vector) initializations, and plot the probability that k is accepted (Fig. 3). We take the largest k value above chance level (accept/reject k) and observe that, for both examples, this corresponds to the correct number of clusters.

Finally, a note on our cluster selection procedure: since we rely on Gaussian mixtures, for obtaining our entropy estimates, it could be that our approach favors k-means clustering since there exists a formal connection between the EM algorithm for Gaussian mixtures and k-means clustering (distortion error *vs.* maximum likelihood, *e.g.*, see [15]). Hence, it could be that our procedure performs less well for other clustering techniques that assign data points to single clusters.

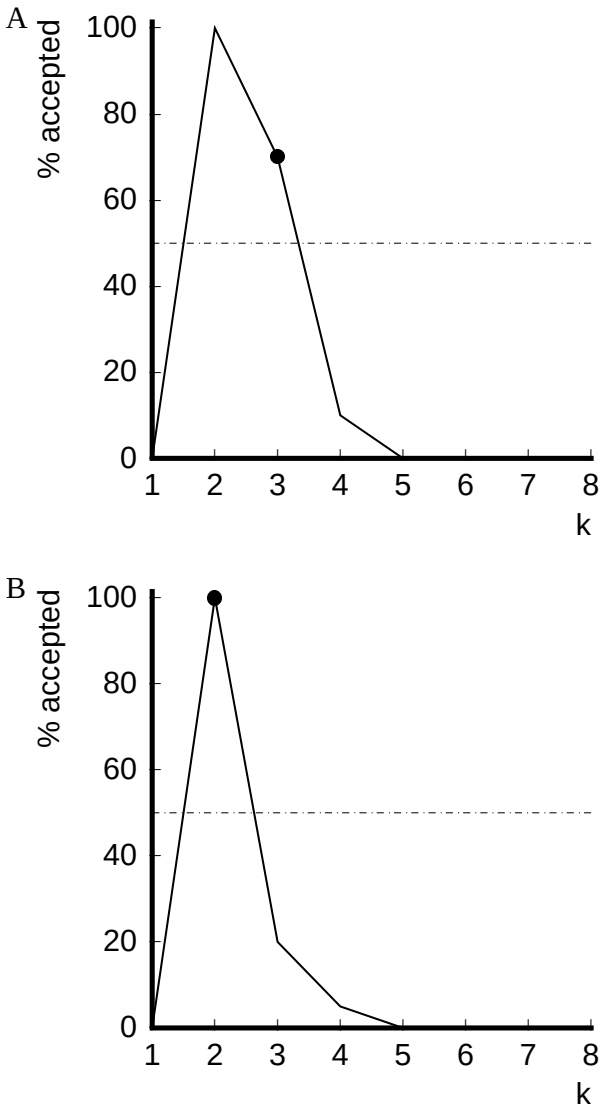

Fig. 3. Determining the optimal number of clusters for the IRIS (A) and Wisconsin breast cancer data sets (B). The filled circle indicates the correct number of clusters. The stippled line indicates chance level.

4. MUTUAL INFORMATION TESTING

We will estimate the multivariate entropy and use it in a mutual information test with which one can statistically decide whether a multivariate distribution can be factorized along its coordinate axes, $P(\mathbf{v}) \stackrel{?}{=} \Pi_i P_i(v_i)$, or any other given set of axes.

Consider a homogeneous mixture of three d-dimensional Gaussians with standard deviations 0.1, and centered at three different corners of a flat square with vertex length β. When $\beta = 0$, we have only one Gaussian, and the distribution is factorizable. In all other cases, the distribution is, in theory, not factorizable. We compute both the sum of the one-dimensional marginal entropies $\sum_i H_i(v_i)$ and the joint entropy $H(\mathbf{v})$, and take the difference. This is then the mutual information $\mathrm{MI}(\mathbf{v})$. For the joint entropy estimate, we take the minimum of the (unimodal) Edgeworth entropy and that of the Gaussian mixture eq. (3) (*cf.*, the heuristic in section 2.1). We repeat this for 999 Monte Carlo generated data sets, rank the MI's, and verify the rank of the zero value (*i.e.*, theoretically factorizable). As a significance level we take $\alpha = 10\%$ (two-sided). We take for the joint density's dimension $d = 2$ and 10. The results are shown in

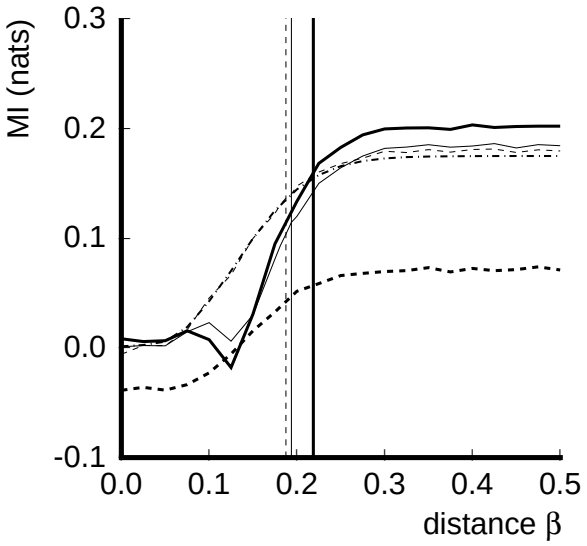

Fig. 4. Mutual information as a function of the distance β between three Gaussian densities with standard deviations 0.1 and located at three corners of a flat square. Plotted are averages over 999 runs obtained with our method, for the $d = 2$ (thin continuous line), and $d = 10$ dimensional cases (thick continuous line), and with the nearest-neighbor method, also for the $d = 2$ (thin dashed line) and $d = 10$ cases (thick dashed line); the standard deviations of our method are on the order of 0.08 and 0.14, for $d = 2$ and $d = 10$, and for the nearest-neighbor method 0.08 and 0.16, respectively. The stippled line indicates the theoretical result. The vertical lines correspond to the mutual information plots and indicate the $\alpha = 10\%$ rank test boundary, but which is not reached for the $d = 10$ case in the nearest-neighbor method.

Fig. 4 (continuous lines). We see a gradual increase of MI as the vertex length β increases until a saturation level is reached. The $\alpha = 10\%$ boundaries are also shown (vertical lines with corresponding line thicknesses). For comparison's sake, we have also plotted the $d = 2$ and 10 results for the nearest-neighbor method. The $d = 2$ result (thin dashed line) corresponds quite closely to the theoretical result (stippled line), which was determined by numerical integration. The $\alpha = 10\%$ boundary (vertical dashed line) is close to our method's (vertical thin continuous line). However, the $d = 10$ result (thick dashed line) is completely wrong and the test fails to reject the null hypothesis over the entire β range shown (meaning that the joint distribution is thought to be factorizable everywhere). This clearly shows that the (modified) Edgeworth approximation is not necessarily inferior in the higher-dimensional case. Furthermore, the computational complexity in N is far less than that of the nearest-neighbor method.

5. CONCLUSION

We have developed the general, multivariate case of the Edgeworth approximation of differential entropy, and introduced an approximation for Gaussian mixtures. We have used both entropy estimates for selecting the optimal number of clusters, and for deciding whether a distribution can be factorized along a given set of axes. As an example application of the latter, when a topographic map is used for mapping the mixture distribution of signals with sub-Gaussian source densities in blind source separation, the map's density should be factorizable along the (rectangular) lattice coordinates into statistically independent components [16, 17, 18]. We have provided here a test to decide whether this is indeed the case. If not, the map could not be well trained, or the number of independent sources could be different from the lattice dimensionality.

6. REFERENCES

[1] C.E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379-423, 1948.

[2] P. Comon, "Independent component analysis – a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287-314, 1994.

[3] S.-I. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation", in *Advances in Neural Processing Systems*, David S. Touretzky, Michael C. Mozer and Michael E. Hasselmo, Eds. 1996, vol. 8, pp. 757-763, The MIT Press.

[4] M. Jones and R. Sibson, "What is projection pursuit?" *J. of the Royal Statistical Society A*, vol. 150, pp. 1-36, 1987.

[5] J. Friedman, "Exploratory projection pursuit," *J. of the American Statistical Association*, vol. 82, no. 397, pp. 249-266, 1987.

[6] P. Huber, "Projection pursuit", *The Annals of Statistics*, vol. 13, no. 2, pp. 435-475, 1985.

[7] O.E. Barndorff-Nielsen and D.R. Cox, *Inference and Asymptotics*, Chapman and Hall: London, 1989.

[8] P. Hall, "Limit theorems for sums of general functions of m-spacings," *Math. Proc. Camb. Phil. Soc.*, vol. 96, pp. 517-532, 1984.

[9] L.F. Kozachenko and N.N. Leonenko, "Sample estimate of entropy of a random vector," *Problems of Information Transmission*, vol. 23, pp. 95-101, 1987.

[10] I.A. Ahmad and P.E. Lin, “A nonparametric estimation of the entropy for absolutely continuous distributions”, *IEEE Trans. Information Theory*, vol. 22, pp. 372-375, 1976.

[11] G. Milligan and M. Cooper, “An examination of procedures for determining the number of clusters in a data set”, *Psychometrika*, vol. 50, no. 2, pp. 159-179, 1985.

[12] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a dataset via the Gap statistic,” *J. Royal Statist. Soc. B*, vol. 63, pp. 411-423, 2001.

[13] S. Dudoit and J. Fridlyand, “A prediction-based resampling method to estimate the number of clusters in a dataset,” *Genome Biology*, vol. 3, no. 7, pp. 0036.1-0036.21, 2002.

[14] C. Sugar and G. James, “Finding the number of clusters in a dataset: An information-theoretic approach,” *J. Am. Stst. Assoc.*, vol. 98(463), pp. 750-763, 2003.

[15] M.M. Van Hulle, “Mixture density modeling, Kullback-Leibler divergence, and differential log-likelihood,” *Neural Computation*, vol. 17(3), pp. 503-513.

[16] P. Pajunen, A. Hyvärinen, and J. Karhunen, “Nonlinear Blind Source Separation by Self-Organizing Maps,” in *Progress in Neural Information Processing. Proceedings of the International Conference on Neural Information Processing*, S.-I. Amari, L. Xu, L.-W. Chan, I. King, and K.-S. Leung, Eds., vol. 2, pp. 1207–1210, Springer-Verlag, Singapore, 1996.

[17] J.K. Lin, D.G. Grier, and J.D. Cowan, “Faithful representations of separable distributions,” *Neural Computation*, vol. 9, pp. 1305-1320, 1997.

[18] M.M. Van Hulle, *Faithful representations and topographic maps: From distortion- to information-based self-organization*, Wiley, New York, 2000.