International Symposium on Information Theory and its Applications, ISITA2008
Auckland, New Zealand, 7-10, December, 2008

Entropy Estimation Using MDL and Piecewise Constant Density Models

Gilles Menez, Maria-João Rendas and Eric Thierry

Lab. I3S, CNRS-UNSA
Les Algorithmes, 2000 rte des Lucioles
06903 Sophia Antipolis cedex, France
E-mail: {menez}{rendas}{et}@i3s.unice.fr

Abstract

We compare the performance of MDL-based estimators of the differential entropy of scalar random variables to several state-of-the-art entropy estimators. The estimators studied determine the entropy of a piecewise constant probability density function whose complexity – the number of intervals of constant density value, or "bins" – is automatically adjusted to the sample size. The estimators are based on an efficient implementation of the Maximum Likelihood Estimator that has been recently proposed in the literature. Simulation studies for several data distributions show that if the MDL penalty is conveniently defined, performance compares favorably with the state-of-the-art entropy estimators tested, while at the same time providing a compact model for the data set analyzed.

1. Introduction

Information theory tools and measures are increasingly used in many applications, in particular in the domain of machine learning. A fundamental problem in this framework is to estimate the (differential) Shannon entropy of a random variable[1] $X \sim f$ from a finite number N of statistically independent samples from f:

$$\{x_1, \ldots, x_N\} \xrightarrow{?} H(f) = \mathrm{E}_X\left[-\ln f(X)\right], \qquad x_i \sim f \ .$$

We restrict to *continuous random variables*, i.e. $x_i \in \mathcal{X} \subseteq \Re$.

1.1. Background

Entropy estimation from a finite sample is an old problem, for which a number of approaches have been proposed, see [1] for a review. Three broad classes are usually identified: *(i) plug-in estimators*, which estimate entropy in a 2-step procedure, building first an estimate $\hat{f}$ of the data probability density function (*pdf*), and assessing in a second step the problem of computing the expected value $\mathrm{E}_X[\cdot]$ in the definition of $H(\cdot)$; and *(ii) sample-spacing* [13], and *(iii) nearest-neighbor distances* based estimators, that estimate H directly, without resorting to explicit estimation of f [7].

In class *(i)*, popular choices for $\hat{f}$ are *kernel-based* density estimators (*kde*)

$$\hat{f}_N(x:h) = \frac{1}{Nh}\sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h}\right) \ , \qquad (1)$$

where $\mathcal{K}(\cdot)$ is a non-negative *kernel* function with suitable properties, and the parameter $h > 0$ is the kernel *width*. Computation of the expected value resorts either to numerical integration [5, 2], or, more often, to Monte-Carlo methods such as re-substitution [4], splitting-data [3] or cross-validation [4]. Under conditions on f (continuity, tail, peak values), weaker or stronger forms of consistency have been demonstrated for these estimators, see [1]. We concentrate here on the cross-validation (or leave-one-out) form of these estimators:

$$\hat{H}_N(f:h) = -\frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}_{N-1,i}(x_i:h) \ , \qquad (2)$$

where $\hat{f}_{N-1,i}$ is the estimator (1) based on the reduced sample $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$. $\hat{H}_N(\cdot)$ has been studied in [4], that establishes its root-N consistency under several conditions.

Choice of the kernel's width h critically dictates the performance of these estimators. For *density estimation* two main methodologies are used [8]: classical (amongst which cross-validation) methods and "plug-in" methods where h is chosen to minimize an estimate of the estimation bias. No clear supremacy of one single approach seems to hold, see [8]. In [4], the author suggests that *for entropy estimation* h be chosen by finding the value that leads to the *smallest* estimated entropy, i.e.

$$\hat{H}_{H\&M}(f) \stackrel{\triangle}{=} \min_{h>0} \hat{H}_N(f:h) \ . \qquad (3)$$

Nearest-neighbor estimators (*nne*) *(iii)* have been initially proposed by Kozachenko and Leonenko [7].

[1] $X \sim f$ indicates that f is the probability density function (*pdf*) of the random variable X.

They directly estimate H by (for scalar data)

$$\hat{H}_{\mathrm{K\&L}}(f:k) \triangleq \frac{1}{N}\sum_{i=1}^{N}\ln(\rho_{i,k}) + \ln\left(\frac{\sqrt{\pi}}{\Gamma(3/2)}\right) + \ln(N) - L_{k-1} + \gamma \ , \quad (4)$$

where $k \geq 1$ is the order of the neighborhood considered, $\Gamma(\cdot)$ is the gamma function, γ is the Euler constant, and $\rho_{i,k}$ is the distance from x_i to its k-th neighbor. The parameter k determines the degree of "smoothing" of the estimate – and controls thus the bias-variance trade-off – much like h for the kernel methods. Its determination is affected by the same difficulties. For $k = 1$, whith ρ_i the distance to the closest neighboor of x_i,

$$\hat{H}_{\mathrm{K\&L}}(f:1) = \frac{1}{N}\sum_{i=1}^{N}\ln(\rho_i) + \ln\left(\frac{\sqrt{\pi}}{\Gamma(3/2)}\right) + \ln(N-1) + \gamma \ .$$

Class *(ii)* has been proposed by Vasicek [13], and is based on the uniformity of the probability integral for continuous densities:

$$\hat{H}_{\mathrm{V}}(f:m) \triangleq \frac{1}{N-m}\sum_{i=1}^{N-m}\ln\left(x^{(i+m)} - x^{(i)}\right) \ , \quad (5)$$

where $\{x^{(1)}, \ldots, x^{(N)}\}$ is the order statistic of the sample sequence, and $m \geq 1$ is the order of the spacings considered. Again, determination of the best value of m is subject to the same difficulties as h (*kde*) and k (*nne*). Consistency of this estimator has been proved under the condition that $m \to \infty$ and $m/N \to 0$. A typical choice is $m = \sqrt{N}$. $\hat{H}_V$ is also known to be asymptotically efficient [1].

2. MDL estimators of entropy

We compare the performance of a plug-in estimator based on a simple *parametric* model $\hat{f}_{MLD}$ of the *pdf* of the data, that enables *analytic* computation of the expected value:

$$\hat{H}_{\mathrm{MDL}}(f) = \mathrm{E}_{\hat{f}_{MDL}}\left[-\ln\hat{f}_{MDL}(x)\right] \ , \quad (6)$$

with the performance of the estimators introduced in the previous section. In fact, we consider that $\hat{f}_{MDL}$ is a *piece-wise constant* density, with support coincident with the interval spanned by the data. This method can be related to sample-spacing methods, the value of m being locally adjusted to the data characteristics. Put in other words, $\hat{f}_{MDL}$ is a *variable bin histogram*. Let $\mathbf{x}^{(N)}$ denote the ordered data sequence, and let $I \triangleq \left[x^{(1)}, x^{(N)}\right]$. We consider estimates of the data *pdf* of the form

$$\hat{f}_{\mathrm{MDL}}(x:K) = \begin{cases} \alpha_i, & x \in I_i^K \qquad i = 1, \ldots, K \\ 0, & x \notin I \end{cases} \ , \quad (7)$$

where the collection $\{I_i^K\}_{i=1}^K$ is a partition of I. The parameter K – the complexity of the *pdf* model – is determined using the Minimum Description Length (MDL) [9].

According to MDL, when selecting from a set of competing models $\{\mathcal{M}_K, K = 1, 2, \ldots\}$, one should chose the model $\mathcal{M}_K$ for which one can find an associated universal code that gives the best compression of the observed data $\mathbf{x}^{(N)}$ (i.e., the shortest code). For the precise definition of the notion of universal code for a ensemble of probability distributions $\mathcal{M}$, we refer the reader to [9], pointing out here just their main property: *asymptotically* – i.e. for large N – they are able to compress the observed sequence $\mathbf{x}^{(N)}$ as well as the best code $f^{\star}_{\mathbf{x}^{(N)}} \in \mathcal{M}$ for $\mathbf{x}^{(N)}$. Note that this property is valid for *any* observed sequence, and does not rely on the assumption that the observed sequence is drawn for a member of $\mathcal{M}$. For our application, this means that we are free of assumptions about the data generating *pdf*. Originally proposed as a penalized likelihood technique – the penalty being justified in the framework of "two-part codes" – MDL can be formally based on the use of universal codes, the consistency of the derived estimators being inherited from its asymptotic equivalence to Maximum Likelihood. Performance for finite sample sizes is dependent on how efficient the universal code used is, i.e., of its *redundancy*, for each observed sequence, with respect to the optimal code $f^{\star}_{\mathbf{x}^{(N)}} \in \mathcal{M}$. It has been shown [11] that the code that minimizes the maximum redundancy is the Normalized Maximum Likelihood (NML) code, whose codelenght is given by

$$-\log f_{\mathrm{NML}}(\mathbf{x}^{(N)}) = -\log f^{\star}_{\mathbf{x}^{(N)}}(\mathbf{x}^{(N)}) - \log\int_{\mathcal{X}^N} f^{\star}_{y^{(n)}}(y^{(n)})\, dy^{(n)} \ , \quad (8)$$

when the integral in the second term exists, and where $f^{\star}_{x^{(n)}}(\cdot)$ is the Maximum Likelihood estimate of the data *pdf* for sequence $\mathbf{x}^{(N)}$ in model $\mathcal{M}$:

$$f^{\star}_{\mathbf{x}^{(N)}}(\cdot) = \arg\max_{f(\cdot)\in\mathcal{M}} f(\mathbf{x}^{(N)}) \ . \quad (9)$$

NML still leads, as equation (8) shows, to a penalized likelihood criterion. Its penalty is called the *parametric complexity* of the model $\mathcal{M}$, and we will denote it by $\mathcal{C}(\mathcal{M}, N)$. When considering a discrete set of models indexed by a natural parameter K,$\{\mathcal{M}_K\}$, we will use the shorter notation $\mathcal{C}(K, N) \triangleq \mathcal{C}(\mathcal{M}_K, N)$

2.1. Computing the parametric complexity

The application of the MDL principle to the identification of piece-wise constant density models (7) has been addressed in [6]. Considering a quantification of the original data to a fine grid the paper is able to derive expressions for the parametric complexity, and shows that determination of the optimal model can be efficiently solved using a Dynamic Programming algorithm. The parametric complexity that must be added to the neglikelihood is, accoding to these authors, defined recursively by the following expressions:

$$\begin{aligned} \mathcal{C}_{\mathrm{NML}}(K,N) &\triangleq \log\left(\mathcal{R}(K,N)\right) + \log\left(C_{K-1}^{N}\right) \qquad (10) \\ \mathcal{R}(K,N) &= \mathcal{R}(K,N-1) + \frac{N}{K-2}\mathcal{R}(K,N-2), \\ \mathcal{R}(K,2) &= \sum_{r=0}^{N} C_r^N \left(\frac{r}{N}\right)^r \left(\frac{N-r}{N}\right)^{N-r} \quad , \end{aligned}$$

where $C_r^N \triangleq N!/r!(N-r)!$, and $\mathcal{R}(K,1) = 1$. Besides the artificial need to consider discretization of the observed data, several issues are raised by the solution proposed in this reference. One of the puzzling points is the independence of the derived criterion on the fineness of data discretization, which seems unnatural to us. Another intriguing aspect concerns the fact that the parametric complexity is taken into account during search for the optimal code of a given complexity $f^{\star K}_{\mathbf{x}^{(N)}}$. More fundamentally, the NML complexity determined in the paper considers models where the partition I is fixed, and only the estimated amplitudes α_i are allowed to vary. However, estimation is done for models parametrized *simultaneously* by the set of cut-points and the *pdf* levels. This suggests that the complexity proposed by the paper is too small (because in the integration in the penalty term of (8) a sub-optimal $f^{\star K}_{\mathbf{x}^{(N)}}$ is in fact considered). This mismatch should – and in fact it does – lead to overly-complex models. We compared the expressions (10) for the parametric complexity to a corrected version of the classic BIC approximation:

$$\mathcal{C}_{\mathrm{BIC}}(K,N) \triangleq \frac{K}{\log}(N) + \log\left(C_{K-1}^{N}\right) \quad ,$$

where K is the number of bins: the BIC approximation is used for the the amplitudesα_i, with respect to which the corresponding model has a *continuous variation* – i.e, the resulting set $\mathcal{M}$ has the structure of a differentiable manifold – and the other structural parameters are accounted for by counting the number of ways that $K-1$ bin limits can be defined from N data points. The two estimators that are obtained using these two alternative formulations of the MDL penalty are designated by MDL/NML and MDL/BIC.

3. Performance analysis

We compare the performance of several differential entropy estimators to the performance of the "plug-in" estimators defined in the previous section (6)-(7), on data drawn from several univariate continuous distributions. Our results show that even if – unlike the kernel based estimators tested – both MDL versions lead to consistent estimators of the entropy (since the density estimators are themselves consistent), the MDL/NML entropy estimator displays a much larger negative bias than the MDL/BIC method, confirming the over-parametrization behavior foreseen. This demonstrates that important characteristics of the densities identified by the MDL/NML estimator proposed in [6], even if providing a compact model for the data that presents good visual agreement with the density, can significantly diverge from those of the generating *pdf*.

The set of estimators considered in this comparative study are:
• *(i) Histogram*: the plug-in estimator for the histogram density estimator using a constant bin width chosen according to the rule [10]: $W = 3.49\,\sigma(\mathbf{x}^{(N)})\,N^{-1/3}$;
• *(ii) kde*: the *kde* estimator with optimal AMISE (Asymptotic Mean Integrated Squared Error) bandwidth ([12], pages 45 and 47): $h_{AMISE} = 1.06\,N^{-1/5}$;
• *(iii) Hall*: the *kde* estimator (2) with bandwidth chosen as suggested in [4], eq. (3);
• *(iv) NN*: the *nne* estimator (4) with $k = 1$.

The first two (classic) estimators are considered for completeness, the last two being representative of state-of-the-art entropy estimators.
We consider four distinct distributions: *normal*, for which we expect that classic methods perform reasonably well, a *balanced mixture* of two normal densities, χ^2, and *Cauchy* – as an example of an infinite variance distribution. For each distribution, data sets of sizes $N = 500, 1000, 2000, 10000$ were considered. The following values have been used for the parameters of the four distributions considered.

- Normal : centered, $\sigma = 0.5$;
- (Balanced Normal) Mixture : $\mu_1 = -1$, $\sigma_1 = 0.5$, $\mu_2 = +1$, $\sigma_2 = 0.5$;
- χ^2 : 2 degrees of freedom;
- Pareto, $k = 1, x_m = 1$;
- Triangular: symmetric, in $[-1, 1]$.

3.1. Numerical Results

For all six distributions, 50 Monte-Carlo runs of all six estimators were conducted. Figures 2 to 7 summarize the results obtained. In these Figures, the horizontal black line indicates the true entropy value. The four groups of vertical bars correspond to growing sample sizes ($N = 50, 100, 500, 2000$, from left to right). Each vertical bar indicates the interval

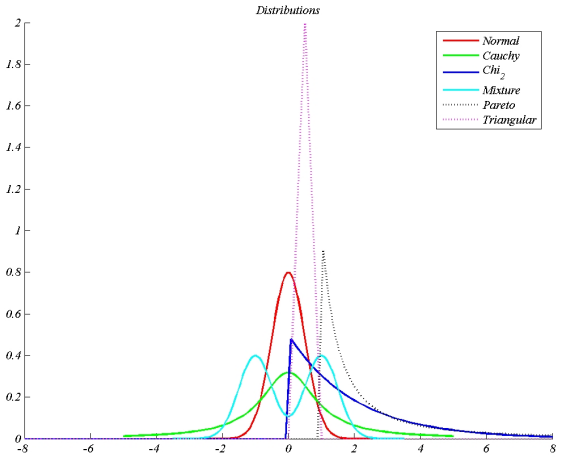

Figure 1: the 6 densities used in the numerical study

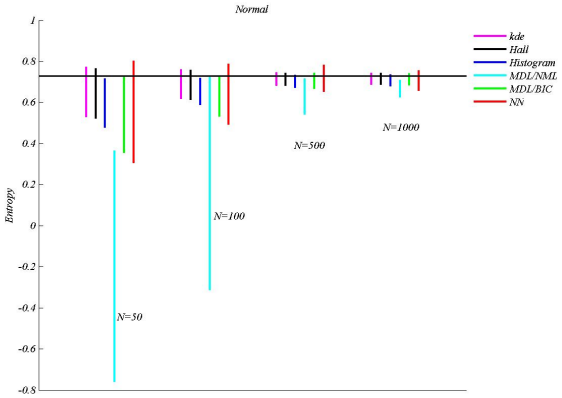

Figure 2: Normal distribution.

of one standard deviation (over the 50 Monte Carlo trials) around the estimator average. The cyan and green lines correspond to the MDL-based estimators: NML and BIC, respectively. For the other estimators, the following color code is used: magenta $\equiv$ *kde*; black $\equiv$ *Hall*; blue $\equiv$ *Histogram*; red $\equiv$ *NN*.

We verify that for large sample sizes the behavior of the MDL-based estimator is comparable to the behavior of the best of the four other estimators. In particular, its bias steadily decreases to zero, even for the heavy tail Cauchy and Pareto distributions, unlike the kernel based method proposed in [4], displaying for these distributions a behaviour comparable (but better) to the state-of-the-art *NN* method. Note that all other plug-in and *kde* estimators have, for these densities, large bias even for very large values of N.

Comparison of the two MDL estimators implemented confirms the analysis presented before. MDL/NML as proposed in [6] is under-penalized, lead-

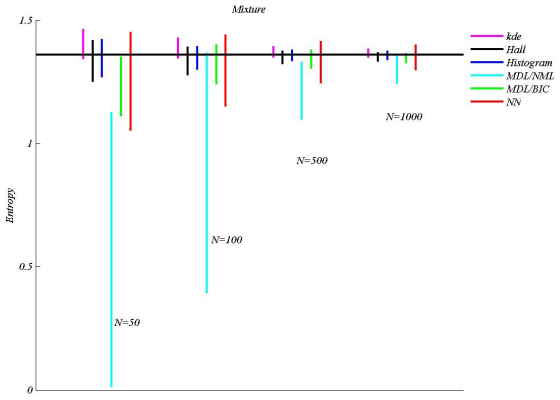

Figure 3: Mixture distribution.

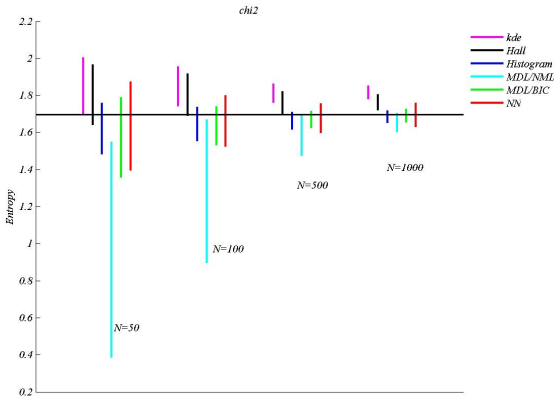

Figure 4: χ^2 distribution.

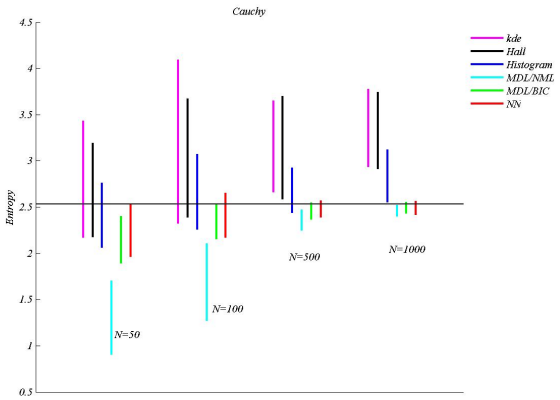

Figure 5: Cauchy distribution.

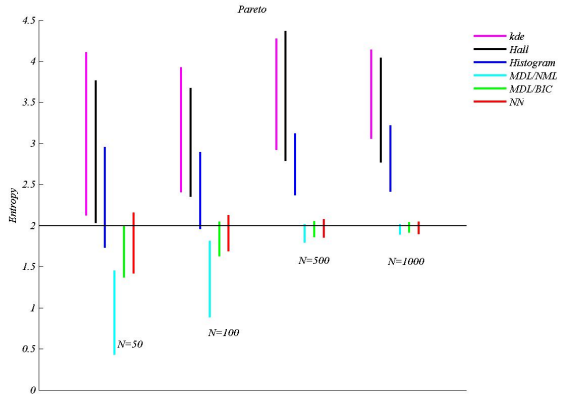

Figure 6: Pareto distribution.

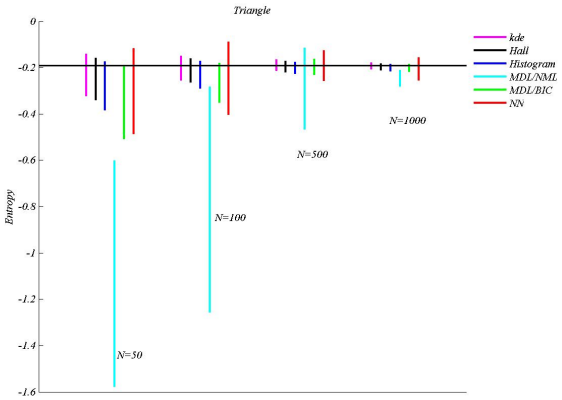

Figure 7: Triangular distribution.

ing to over-complex models that have an entropy much smaller than the law generating the data: this estimator presents a systematic large negative bias, as well as a large variance. Figures 8 to 13 display the histogram of the number of bins identified by MDL/NML and MDL/BIC in the 50 Monte-Carlo experiments for values of $N = 50, 100, 500, 1000$. Each Figure corresponds to one of the distributions considered. While MDL/BIC identifies models whose complexity steadily increases with the number of observations, and a moderate number of "bins", MDL/NML displays an erratic behavior, identifying models that rely on a fine partition of the data support.

4. Conclusions

We presented a comparison of the performance of two MDL plug-in estimators of the differential entropy of a scalar random variable. One of the estimators considers the NML penalty for variable constant bin models previously presented in the literature, while the other is proposed here for the first time, and is based on the standard BIC criterion. The numerical study conducted demonstrates the poor behavior of the MDL/NML estimator, which under-estimates

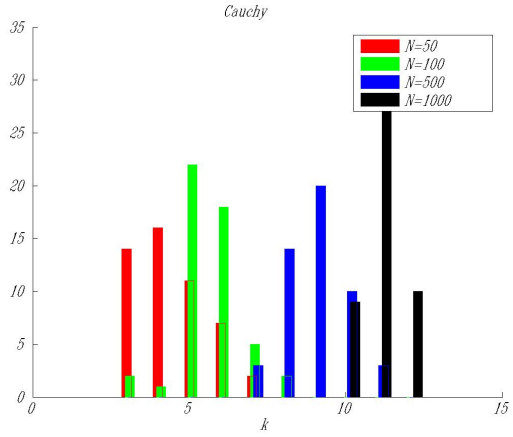

Figure 8: Histogram of k, 50 runs (MDL/BIC).

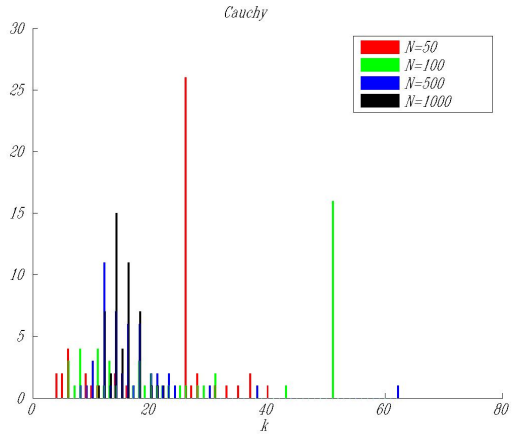

Figure 9: Histogram of k, 50 runs (MDL/KONT).

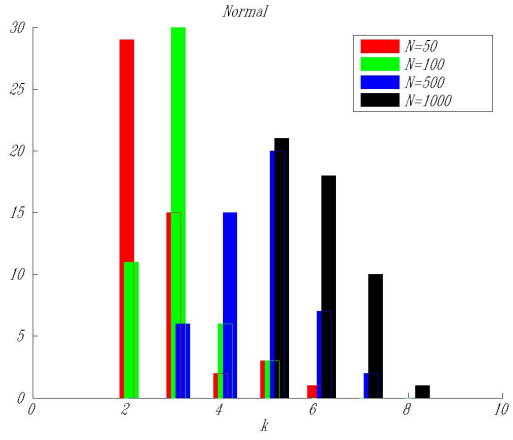

Figure 10: Histogram of k, 50 runs (MDL/BIC).

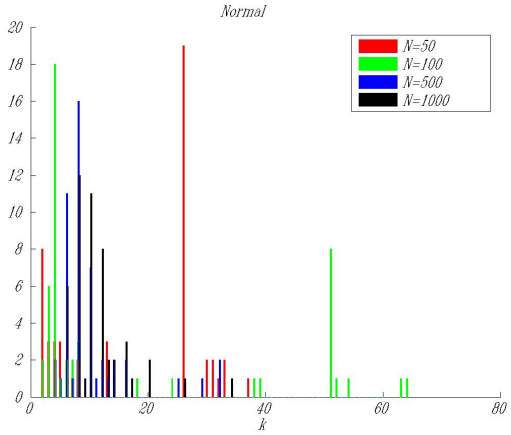

Figure 11: Histogram of k, 50 runs (MDL/KONT).

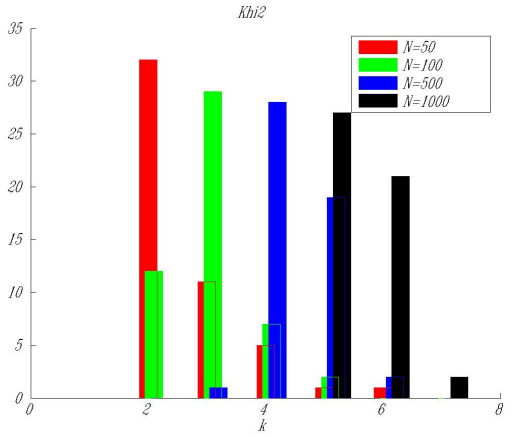

Figure 12: Histogram of k, 50 runs (MDL/BIC).

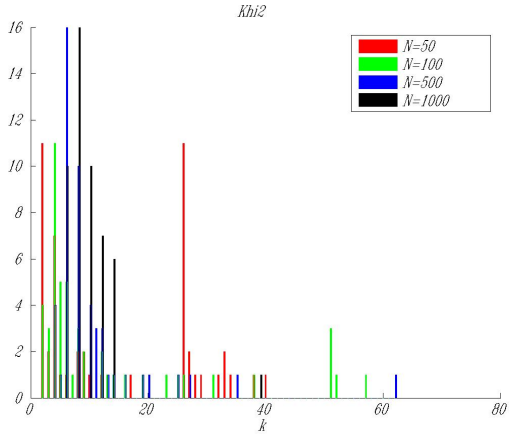

Figure 13: Histogram of k, 50 runs (MDL/KONT).

model complexity, leading to large negative entropy biases. The BIC-based penalty, on the contrary, leads to an estimator with small bias and variance, even for small N and heavy tailed ditributions, beating its closer competitor, the NN, under virtually all conditions analyzed.

Acknowledgments

Lack of space prevents us from detailing the exact expressions used for numerical determination of the parametric complexity, which, if naively coded, easily lead to complexity and numerical problems. We want to acknowledge the friendly collaboration of our colleague Jean Marc Fédou on the determination of alternative (computable) expressions.

References

[1] J. Beirlant, E.J. Dudewicz, L. Györfi, and E.C. van der Meulen. Nonparametric entropy estimation: An overview. *International J. Mathematical and Statistical Sciences*, 6:17–39, 1997.

[2] L. Györfi and E. C. van der Meulen. Density-free convergence properties of various estimators of entropy. *Comput. Statist. Data Anal.*, 5:425–436, 1987.

[3] L. Györfi and E. C. van der Muelen. *Nonparametric Functional Estimation and Related Topics*, chapter On nonparametric estimation of entropy functionals, pages 81–95. Kluwer Academic Press, 1990.

[4] P. Hall and S. C. Morton. On the estimation of entropy. *Ann. Inst. Statist. Math.*, 45:69–88, 1993.

[5] H. Joe. On the estimation of entropy and other functionals. *Ann. Inst. Statist. Math.*, 41:683–697, 1989.

[6] P. Kontkanen, P. Myllymäki. Mdl histogram density estimation. In *AISTATS*, 2007.

[7] L. F. Kozachenko and N. N. Leonenko. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23 (2):95–101, 1987.

[8] Clive R. Loader. Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, 27:415–438, 1999.

[9] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, Information Science & Statistics, 2007.

[10] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[11] Yu M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17, 1987.

[12] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[13] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:54–59, 1976.