Construction of Irregular Histograms by Penalized Maximum Likelihood: a Comparative Study

Panu Luosto*, Ciprian Doru Giurcăneanu*†‡ and Petri Kontkanen*†
*Department of Computer Science, University of Helsinki, Finland
†Helsinki Institute for Information Technology, HIIT, Finland
‡ Department of Statistics, University of Auckland, New Zealand
Email: panu.luosto@cs.helsinki.fi, c.giurcaneanu@auckland.ac.nz, petri.kontkanen@cs.helsinki.fi

***Abstract*—Theoretical advances of the last decade have led to novel methodologies for probability density estimation by irregular histograms and penalized maximum likelihood. Here we consider two of them: the first one is based on the idea of minimizing the excess risk, while the second one employs the concept of the normalized maximum likelihood (NML). Apparently, the previous literature does not contain any comparison of the two approaches. To fill the gap, we provide in this paper theoretical and empirical results for clarifying the relationship between the two methodologies. Additionally, we introduce a new variant of the NML histogram. For the sake of completeness, we consider also a more advanced NML-based method that uses the measurements to approximate the unknown density by a mixture of densities selected from a predefined family.**

I. Introduction

The regular histogram, in which all the bins are equally wide, is generally considered to be the simplest probability density estimator. A more advanced option is the so-called irregular histogram (IH), where the bins are allowed to have different widths. In both cases, the unknown density is approximated by a piecewise constant density model. However, in most of the practical situations, the number and the borders of the bins are not known a priori, and they should be chosen from a predefined collection of histogram models.

Significant efforts have been dedicated to defining selection rules based on penalized maximum likelihood (PML). One of the first attempts was to apply the Akaike criterion [1] to histogram density estimation (see [2] and the references therein). During the last decade, two new methodologies have been proposed. The first one, which is more popular in the community of statisticians, is based on the idea of choosing the penalty so as to minimize the excess risk [3], [4]. The second methodology has its grounds in information theory and employs the normalized maximum likelihood (NML) as a criterion for selecting the histogram model. The interested reader can find in [5] the definition of the NML as well as strong theoretical results on its properties. However, the computation of the NML poses troubles, and it seems that the first practical solution for IH selection by NML was introduced in [6].

Surprisingly, the methods from [4] and [6] have not been compared up to now, and this is mainly because the NML criterion was misinterpreted in some of the previous studies. As we aim to clarify the relationship between the two methodologies, the main contributions of this work are: (i) Show how the theoretical results from [4] can be applied to the selection rule from [6]; (ii) Introduce a new variant of the NML criterion; (iii) Conduct experiments with a large family of distributions for evaluating the capabilities of five PML criteria whose formulas are based either on [4] or on [5]. The most relevant examples are discussed in detail for providing guidance to the practitioners.

The rest of the paper is organized as follows. We give in the next section a more formal description of the density estimation problem. Then, in Section III, we analyze the NML criterion [6] by resorting to the findings from [4]. We introduce a new variant of NML histogram in Section IV, where we also present all the other selection rules whose performance is evaluated in our experiments. Section V is devoted to the description of experiments and the interpretation of their outcome. Section VI concludes the paper.

II. Problem formulation

Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with common law P on a measurable space $(\mathcal{Z}, \mathcal{T})$. Under the hypothesis that P admits a density s_* with respect to the probability measure μ, or equivalently, $s_* = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, we want to estimate s_* from the measurements $\xi_1, \ldots, \xi_n$. We take μ to be the Lebesgue measure on $\mathcal{Z}$.

The main definitions as well as most of the notations which we use are akin to those from [3], [4, Ch. 7], [7, Ch. 5]. So, for a measurable function f on $\mathcal{Z}$, we have $P(f) = \mathbb{E}[f(\xi)]$, where $\mathbb{E}[\cdot]$ denotes the expectation operator and ξ stands for a generic random variable of law P on $(\mathcal{Z}, \mathcal{T})$. Moreover, $P_n(f) = n^{-1} \sum_{i=1}^{n} f(\xi_i)$ is the empirical distribution associated to the samples $\xi_1, \ldots, \xi_n$.

Now we are prepared to introduce the well-known definition of the histogram. It is convenient to assume that $\mathcal{Z}$ is a compact interval of $\mathbb{R}$. Let $\Lambda_M = \bigcup_{j=1}^{D_M} \mathcal{I}_j$ be a finite partition of $\mathcal{Z}$, with the property that $\mu(\mathcal{I}) > 0$ for all $\mathcal{I} \in \Lambda_M$. Furthermore, we consider the set of piecewise constant functions with respect to Λ_M, namely $\tilde{M} = \{s = \sum_{\mathcal{I} \in \Lambda_M} \beta_{\mathcal{I}} \mathbf{1}_{\mathcal{I}} \ : \ (\beta_{\mathcal{I}})_{\mathcal{I} \in \Lambda_M} \in \mathbb{R}^{D_M}\}$, where $\mathbf{1}_{\mathcal{A}}$ denotes the indicator function of a set $\mathcal{A}$. Let M be the subset of functions in $\tilde{M}$ that are densities with respect to Λ_M: $M = \{s \in \tilde{M} \ : \ s \geq 0, \int_{\mathcal{Z}} s \mathrm{d}\mu = 1\}$. Given the model M, the maximum likelihood (ML) estimator is that function $s \in$

M which minimizes $P_n(-\log s) = n^{-1}\sum_{i=1}^{n}[-\log s(\xi_i)]$ ($\log(\cdot)$ stands for the natural logarithm). It can be easily shown that the expression of the ML estimator is [7, Ch. 5]: $\hat{s}_n(M) = \sum_{\mathcal{I}\in\Lambda_M}\frac{P_n(\mathcal{I})}{\mu(\mathcal{I})}\mathbf{1}_{\mathcal{I}} = \frac{1}{n}\sum_{\mathcal{I}\in\Lambda_M}\frac{1}{\mu(\mathcal{I})}\left[\sum_{i=1}^{n}\mathbf{1}_{\mathcal{I}}(\xi_i)\right]\mathbf{1}_{\mathcal{I}}$.

The problem addressed in this paper consists in selecting, from a predefined collection $\mathcal{M}_n$ of IH models, the model $\hat{M}$ which satisfies the condition:

$$\hat{M} = \arg\min_{M\in\mathcal{M}_n}\{P_n(-\log\hat{s}_n(M)) + \mathrm{pen}(M)\}, \quad (1)$$

where $\mathrm{pen}(M)$ is a nonnegative penalty. In the next section, we discuss two important methods for choosing the penalty term.

III. Risk bounds and the penalty term

Conventionally, the risk associated with an arbitrary density function s is $P(-\log s)$, and this leads to the following definition of the excess risk [7]: $P(-\log s) - P(-\log s_*) = \int_{\mathcal{Z}} s_*\log\frac{s_*}{s}\mathrm{d}\mu$. It is easy to notice that the excess risk equals the Kullback-Leibler (KL) divergence $D_{KL}(s_*, s)$.

Among all densities s which belong to a given model M, the one which minimizes $D_{KL}(s_*, s)$ is $s_M = \sum_{\mathcal{I}\in\Lambda_M}\frac{P(\mathcal{I})}{\mu(\mathcal{I})}\mathbf{1}_{\mathcal{I}} = \sum_{\mathcal{I}\in\Lambda_M}\frac{1}{\mu(\mathcal{I})}\left[\int_{\mathcal{Z}} s_*\mathbf{1}_{\mathcal{I}}\mathrm{d}\mu\right]\mathbf{1}_{\mathcal{I}}$ [7]. For obvious reasons, s_M is called the KL projection of s_* onto M. An oracle who has full knowledge on the density s_* will choose the model $M \in \mathcal{M}_n$ so as to minimize $D_{KL}(s_*, s_M)$. Hence, the performance of the model selection criterion in (1) can be evaluated by comparing $\mathbb{E}[D_{KL}(s_*, \hat{s}_n(\hat{M}))]$ with $\inf_{M\in\mathcal{M}_n}[D_{KL}(s_*, s_M)]$. The main drawback of this approach comes from the fact that the KL divergence is infinite for all intervals $\mathcal{I} \in \Lambda_{\hat{M}}$ where $s_*(\mathcal{I}) > 0$ and $\hat{s}_n(\hat{M})$ is identically zero on $\mathcal{I}$ [3, Sec. 2.2].

A more suitable candidate for the loss function is the squared Hellinger distance (SHD) $h^2(s_*, s) = (1/2)\int_{\mathcal{Z}}\left(\sqrt{s_*} - \sqrt{s}\right)^2\mathrm{d}\mu$, where s_* and s have the same significance as above. The performance of the selection criterion can be evaluated by finding upper bounds for $\mathbb{E}[h^2(s_*, \hat{s}_n(\hat{M}))]$ [3], [4]. We analyze next the NML criterion from [6] by using [4, Th. 7.9], which we reproduce below:

Theorem 1 (Th. 7.9 in [4]). *Assumptions:* **(A1)** $\mathcal{Z} = [0, 1]$*;* **(A2)** *For some positive real number ρ, $s_* \geq \rho$ almost everywhere, and $\int_{\mathcal{Z}} s_*(\log s_*)^2 d\mu \leq L < \infty$;* **(A3)** *Consider on $\mathcal{Z}$ the grid $\mathcal{G} = \{q/N_n \ : \ q = 0, 1, \ldots, N_n\}$, where N_n is a positive integer which satisfies the inequality*

$$N_n - 1 \leq n/(\log n)^2. \quad (2)$$

Let $\mathcal{M}_n$ be a collection of histogram models such that, for any $M \in \mathcal{M}_n$, the cut points of the partition Λ_M belong to $\mathcal{G}$. Additionally, Σ is a positive constant which does not depend on n, and $(x_M)_{M\in\mathcal{M}_n}$ is a family of nonnegative weights such that

$$\sum_{M\in\mathcal{M}_n} e^{-x_M} \quad \leq \quad \Sigma. \quad (3)$$

Let $c_1 > 1/2$ and $c_2 = 2(1 + c_1^{-1})$. If the following inequality holds for the penalty term in (1),

$$n\cdot\mathrm{pen}(M) \geq c_1\left(\sqrt{D_M - 1} + \sqrt{c_2 x_M}\right)^2 \forall M \in \mathcal{M}_n, \quad (4)$$

then there is a constant $C(c_1, \rho, L, \Sigma)$ such that

$$\mathbb{E}[h^2(s_*, \hat{s}_n(\hat{M}))] \leq \frac{\inf_{M\in\mathcal{M}_n}\{D_{KL}(s_*, s_M) + \mathrm{pen}(M)\}}{1 - (2c_1)^{-1/5}} + \frac{C(c_1, \rho, L, \Sigma)}{n}.$$

According to [6], the NML penalty term is

$$n\cdot\mathrm{pen}_{\mathrm{NML}}(M) = \log\binom{N_n - 1}{D_M - 1} + \log\mathcal{C}(D_M, n), \quad (5)$$

where $\mathcal{C}(D_M, n) = \sum_{\substack{\nu_1+\cdots+\nu_{D_M}=n,\\ \nu_1,\ldots,\nu_{D_m}\geq 0}}\frac{n!}{\nu_1!\cdots\nu_{D_M}!}\prod_{i=1}^{D_M}\left(\frac{\nu_i}{n}\right)^{\nu_i}$.

Observe that the use of the penalty (5) requires a grid like in assumption **(A3)** of Theorem 1. We will clarify later if the step size of the grid should be chosen so as (2) is satisfied, or if other options are also possible.

For the analysis of the formula in (5), we resort to an approximation. More precisely, by employing a result from [8], the penalty term of the NML criterion can be written as follows [9, p. 14]:

$$n\cdot\mathrm{pen}_{\mathrm{NML}}(M) = \log\binom{N_n - 1}{D_M - 1} + \frac{D_M - 1}{2}\log\frac{n}{2\pi} + \log\frac{\pi^{D_M/2}}{\Gamma(D_M/2)} + o(1). \quad (6)$$

The well-known Stirling formula for the Euler integral of the second kind leads to the identity $-\log\Gamma(D_M/2) = -[(D_M - 1)/2]\log(D_M/2) + D_M/2 - \gamma/(6D_M) - (1/2)\log(2\pi)$, where $\gamma \in (0, 1)$. After dropping the terms which do not depend on D_M and ignoring the term $-\gamma/(6D_M)$, the expression in (6) becomes

$$n\cdot\mathrm{pen}_{\mathrm{NML}}(M) \approx \log\binom{N_n - 1}{D_M - 1} \quad (7)$$

$$+ \frac{D_M}{2}\log n \quad (8)$$

$$+ \frac{D_M}{2} + \frac{1 - D_M}{2}\log D_M.$$

It is worth mentioning that the term in (8) is the same as the penalty of the well-known Bayesian Information Criterion (BIC) [10]. More importantly, the quantities from (7) and (8) are the dominant ones within the expression of $\mathrm{pen}_{\mathrm{NML}}(M)$. This makes us to choose, in Theorem 1, the weights $(x_M)_{M\in\mathcal{M}_n}$ so as the formula of $c_1c_2x_M$ from (4) to be given by the sum of the two terms from (7)-(8). So, $x_M = \frac{1}{c}\log\left[\binom{N_n-1}{D_M-1}n^{D_M/2}\right]\ \forall M \in \mathcal{M}_n$, where $c = c_1c_2$. It is easy to notice that the condition $c_1 > 1/2$ leads to $c > 3$.

Furthermore, we focus on verifying the condition from (3):

$$\sum_{M\in\mathcal{M}_n} e^{-x_M} \quad = \quad \sum_{D=1}^{N_n}\left[\binom{N_n - 1}{D - 1}\right]^{1-1/c} n^{-D/(2c)}$$

$$\leq 1 + \sum_{D=2}^{N_n}\left[\frac{(N_n - 1)\,e}{D - 1}\right]^{(D-1)(1-1/c)} n^{-D/(2c)}. \quad (9)$$

The key point is to prove the convergence for the series in (9) when $n \to \infty$ and $N_n \to \infty$. To gain more insight, let us assume that $c = 4$, which is equivalent to $c_1 = 1$. Our choice is mainly motivated by [4, p. 236], where it is mentioned that the optimal value for c_1 is one. If additionally we have that $N_n - 1 < n^{1/6}$, then the convergence can be easily demonstrated. Unfortunately, in most of the practical situations, the condition is not fulfilled.

By using a well-known inequality, we get $\sum_{M \in \mathcal{M}_n} e^{-x_M} \geq LB(n, N_n)$, where $LB(n, N_n) = \sum_{D=2}^{N_n} \left[\frac{N_n - 1}{D-1}\right]^{(D-1)(1-1/c)} n^{-D/(2c)}$. It is easy to verify numerically that $LB(n, N_n)$ is an increasing function of n when $N_n = \lfloor n/(\log n)^2 + 1 \rfloor$ and $n \in [10^2, 10^4]$. For instance, when $c = 4$ ($c_1 = 1$ and $c_2 = 4$), $LB\left(n, \lfloor n/(\log n)^2 + 1 \rfloor\right)$ is approximately 1.64 for $n = 10^2$, but it increases to the value of 2902.11 for $n = 10^4$. It seems, the condition in (2) does not guarantee the convergence of the series (3) evaluated for the weights $(x_M)_{M \in \mathcal{M}_n}$ which correspond to the penalty given by (7)-(8).

The analysis outlined above shows that, for small and moderate samples sizes, the NML criterion cannot be expressed in a form that allows to apply Theorem 1 for evaluating its performance. This makes it necessary to compare the methods from [4] and [6] by Monte Carlo simulations. Apparently, the only published attempt of using the theoretical results from [4] in formulating selection rules for IH is [11]. In the next section, we discuss briefly the criteria from [6], [11] as well as other criteria included in our empirical tests.

IV. Estimation methods

Method NML-1: Hereafter, the name NML-1 will be employed for the method involving the use of the regular grid $\mathcal{G}$ and the penalty in (5).

Method NML-2: Remark in the formula of $\mathcal{C}(D_M, n)$ from (5) that some of the bins might be empty, in the sense that $\nu_i = 0$ for some of the indexes $i \in \{1, \ldots, D_M\}$. An interesting alternative is to optimize the choice of k non-empty bins, or equivalently, to consider only those partitions of $\mathcal{Z} = [0, 1]$ which are defined as sequences of intervals $(\overline{\mathcal{I}}_1, \underline{\mathcal{I}}_1, \ldots, \overline{\mathcal{I}}_{k-1}, \underline{\mathcal{I}}_{k-1}, \overline{\mathcal{I}}_k)$. Note that the non-empty bins $\overline{\mathcal{I}}_j$ ($1 \leq j \leq k$) alternate with the empty bins $\underline{\mathcal{I}}_j$ ($1 \leq j \leq k-1$). The convention that the first and the last bins are non-empty allows us to apply the same scheme when the grid $\mathcal{G}$ is not defined for the entire $\mathcal{Z}$ but for the interval $[\min_{1 \leq i \leq n} \xi_i, \max_{1 \leq i \leq n} \xi_i]$. For an arbitrary non-empty bin, we take $w_j = N_n \cdot \mu(\overline{\mathcal{I}}_j) > 0$. Similarly, $e_j = N_n \cdot \mu(\underline{\mathcal{I}}_j) \geq 0$. Because the cut points are forced to be on the grid $\mathcal{G}$, the following identity holds: $\sum_{j=1}^{k-1}(w_j + e_j) + w_k = N_n$. The modified NML penalty term is given by

$$n \cdot \mathrm{pen}'_{\mathrm{NML}}(M) = \log \min\{n, N_n\} + \log \binom{N_n + k - 2}{2k - 2} + \log \mathcal{C}'(k, n), \quad (10)$$

where $\mathcal{C}'(k, n) = \sum_{\substack{\nu_1 + \cdots + \nu_k = n, \\ \nu_1, \ldots, \nu_k \geq 1}} \frac{n!}{\nu! \cdots \nu_k!} \prod_{i=1}^{k} \left(\frac{\nu_i}{n}\right)^{\nu_i}$. The first term in (10) comes from a uniform distribution for the number of non-empty bins in the set $\{1, 2, \ldots, \min\{n, N_n\}\}$. The expression of the second term is based on two facts: (i) All possible choices of bins are assumed to be equiprobable; (ii) The identity $w_1 + (e_1 + 1) + \cdots + w_{k-1} + (e_{k-1} + 1) + w_k = N_n + k - 1$ defines a composition of the number $N_n + k - 1$ into $2k - 1$ parts. The third term can be calculated efficiently by using the recurrence $\mathcal{C}'(k + 2, n) + 2\mathcal{C}'(k + 1, n) = (n/k - 1)\mathcal{C}'(k, n)$ [12]. For numerical stability, it may be advisable to start the recurrence with $\mathcal{C}'(n, n) = n!/n^n$ and $\mathcal{C}'(n - 1, n) = 2(n - 1)\mathcal{C}'(n, n)$.

Method CG (clustgram): The clustgram models have been recently introduced in [13] as an extension of the IH. This novel density estimator is promising, but so far it has only been applied to a very limited number of examples. In the case of CG, the cut points of the partitions belong to a grid of type $\mathcal{G}$. More importantly, the estimator is a mixture of densities selected from the following family: uniform, shifted exponential, Laplace, normal, shifted half-normal. The selection process is based on a variant of NML [13].

Method RMG (Rozenholc-Mildenberger-Gather): Relying on the theoretical results from [4], the authors of [11] derived a PML criterion for which the penalty term has the expression: $n \cdot \mathrm{pen}_{\mathrm{RMG}}(M) = \log \binom{n-1}{D_M - 1} + (D_M - 1) + (\log D_M)^{5/2}$. The extensive experiments from [11] led to the recommendation of restricting $\mathcal{M}_n$ to partitions with the cut points on the grid defined by the measurements, instead of taking $\mathcal{M}_n$ like in Theorem 1. It has been also observed experimentally in [11] that AIC and BIC yield modest results, and this behavior has been explained by the fact that the two criteria "do not account for multiple partitions with the same number of bins".

Method MRT (Menez-Rendas-Thierry): A "corrected" variant of BIC from [14] that solves the drawback noticed in [11]. The formula of its penalty term is $n \cdot \mathrm{pen}_{\mathrm{MRT}}(M) = \log \binom{n}{D_M - 1} + D_M \log n$. Like in RMG, the definition of $\mathcal{M}_n$ for MRT is based on the data-dependent grid $\mathcal{G}'$.

V. Empirical comparison of the methods

We compare empirically the performance of the estimation methods described in the previous section. Like in [6], [11], our implementations are based on the dynamic programming algorithm. If the optimal number of bins (or non-empty bins in the case of NML-2 and CG) is selected from the set $\{1, 2, \ldots, K\}$, then the time complexity of the algorithm is $O(Kn^2)$. In our settings, $K = \min\{100, \lceil(\max_{1 \leq i \leq n} \xi_i - \min_{1 \leq i \leq n} \xi_i)/\delta\rceil\}$, where δ is the step size of the grid.

In the case of NML-1, the most difficult part is the computation of $\mathcal{C}(D_M, n)$ from (5). In our implementation, the recursive formula from [15] is applied when $n < 10^3$, while for larger values of n we approximate $\mathcal{C}(D_M, n)$ by using [16, Eq. (9)]. Another important aspect is related to the fact that at most n of the intervals of the regular grid can contain measurements. Hence, the computational effort is lowered if each block of consecutive empty intervals is handled as a single large interval. In our settings, the step size of the regular grid is $\delta = 2 \cdot 10^{-2}$. Additionally, we noticed in most of the

experiments that the step size values $2 \cdot 10^{-4}$ or $2 \cdot 10^{-6}$ lead to slightly worse estimation results.

In comparison with NML-1, the method NML-2 uses less computational resources because we need to maintain information only about the non-empty intervals of the regular grid. Moreover, since we wanted the method not to involve arbitrarily chosen parameters, we have selected the step size of the grid for NML-2 from the set $\{2^0, 2^{-1}, \ldots, 2^{-19}\}$ by minimizing the criterion whose penalty is given in (10).

Following the recommendations from [11], we used the grid $\mathcal{G}'$ for RMG and MRT. To have a safeguard against extremely narrow bins, we limited the smallest possible bin width to $\delta = 2 \cdot 10^{-6}$. The values of parameters used in the implementation of CG are: $\epsilon = 0.001$, $a = 100$, $b = 16$ (see [13] for more clarifications on the significance and selection of the parameters).

We tested the estimation methods by using 50 different source distributions, including simple distributions and finite mixtures. For all sample sizes $n \in \{50, 100, 200, 400, 800, 1600, 3200\}$ we generated 100 random samples. Remark that not all the distributions that were used in the tests satisfy the assumptions from Theorem 1. For each realization, we defined the grid $\mathcal{G}$ on the interval $[\min_{1 \le i \le n} \xi_i, \max_{1 \le i \le n} \xi_i]$. Given s_* and the sample size n, $\mathbb{E}[h^2(s_*, \hat{s}_n(\hat{M}))]$ was approximated by replacing the expectation operator with an average over the 100 realizations. Furthermore, like in [14], we employed the estimated distribution to get an estimate of the entropy. The results of four typical test cases are plotted in Fig. 1.

When the distribution was a mixture with clearly separated components resembling the types in the palette of CG (Examples 1 and 2), CG was the best in density estimation, but not necessarily in entropy estimation. In Example 1 we have a mixture of six normals. For the largest sample size ($n = 3200$), CG selected the correct number and types of components in 75% of the cases. Whenever the source distribution had strongly overlapping components, CG was not superior to the other methods. The richness of the family of distributions from which the components are selected was a pitfall for CG in Example 3: When $400 \le n \le 3200$, CG modelled the measurements with exactly one normal distribution, making the estimation performance to be moderate. Only when we increased the sample size further to $n = 6400$, CG started to include more components in the model.

In the evaluation of the IH methods, a direct comparison of the effect of the penalty itself is possible only between RMG and MRT, because, in both cases, the grid as well as the search procedures are the same. In our tests, RMG was consistently slightly better in density estimation than MRT, but MRT estimated the entropy of unimodal distributions usually better (Examples 1–2 vs. Example 3). The grid used by RMG and MRT ensures that there are no empty bins in the resulting histogram. In the case of three well separated uniform components (not shown in Fig. 1) the consequence was that RMG and MRT overestimated the entropy while NML-1, NML-2 and CG were somewhat more accurate and prone to underestimation.

It is worth mentioning that NML-2 worked often well with ragged multimodal source distributions, which turned out to be difficult for the other methods when the sample size was small (Examples 1 and 4). For a better understanding of this behavior, we provide some more statistics concerning Example 4, where s_* is a mixture of five uniform distributions: When $n = 50$, NML-2 used systematically a coarse grid and the average number of non-empty bins in the resulting IH was about seven, whereas NML-1 selected the IH model with one single bin in most of the cases.

VI. Conclusion

We showed that the risk bound analysis from [4] is of little practical value for the NML-1 penalty term. However, in our simulations the performance of NML-1 in terms of SHD was similar to that of RMG which has been specially designed to minimize the statistical risk. Like RMG and MRT, the novel NML-2 is a fully automatic method that seems to be suitable for density and entropy estimation of ragged multimodal distributions. In some cases, CG achieves a significantly lower statistical risk than the IH methods.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, pp. 716–723, 1974.

[2] P. Hall, "Akaike's information criterion and Kullback-Leibler loss for histogram density estimation," *Probab. Th. Rel. Fields*, vol. 85, pp. 449–467, 1990.

[3] G. Castellan, "Modified Akaike's criterion for histogram density estimation," Univ. de Paris-Sud, France, Technical Report #99.61, 1999.

[4] P. Massart, *Concentration inequalities and model selection*. Springer Verlag, 2007.

[5] J. Rissanen, *Information and Complexity in Statistical Modeling*. New York: Springer Verlag, 2007.

[6] P. Kontkanen and P. Myllymäki, "MDL histogram density estimation," in *Proc. 11th Int. Workshop on Artificial Intelligence and Statistics*, 2007.

[7] A. Saumard, "Estimation par minimum de contraste régulier et heuristique de pente en sélection de modèles," Ph.D. dissertation, Univ. de Rennes 1, France, Oct. 2010.

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[9] P. Kontkanen, "Computationally efficient methods for MDL-optimal density estimation and data clustering," Ph.D. dissertation, Dept. Computer Science, Univ. of Helsinki, Finland, Nov. 2009.

[10] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, March 1978.

[11] Y. Rozenholc, T. Mildenberger, and U. Gather, "Combining regular and irregular histograms by penalized likelihood," *Comput. Stat. Data An.*, vol. 54, pp. 3313–3323, 2010.

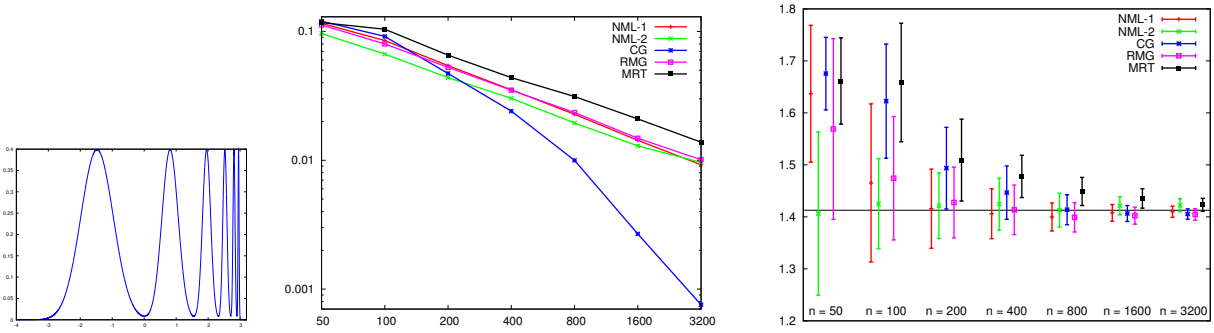
[12] P. Luosto and P. Kontkanen, "The normalized maximum likelihood distribution of the multinomial model class with positive maximum likelihood parameters," University of Helsinki, Department of Computer Science, Tech. Rep. C-2012-5, 2012.

[13] ——, "Clustgrams: an extension to histogram densities based on the minimum description length principle," *Cent. Eur. J. Comp. Sci.*, vol. 1, pp. 466–481, 2011.

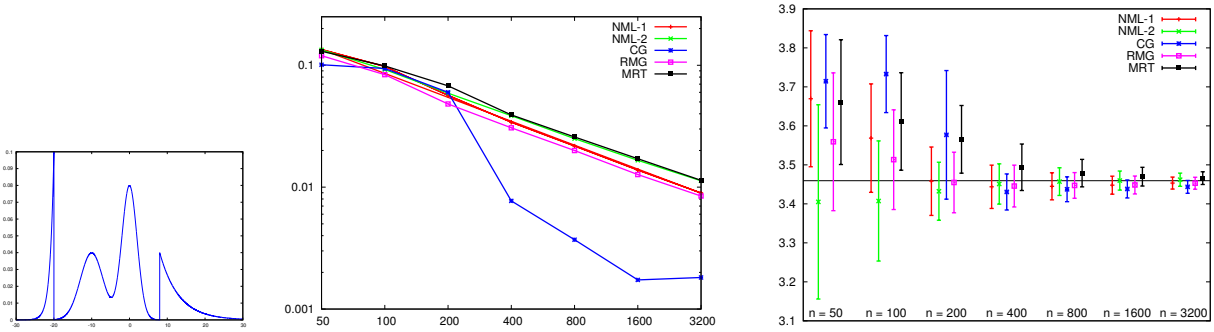
[14] G. Menez, M.-J. Rendas, and E. Thierry, "Entropy estimation using MDL and piecewise constant density models," in *Proc. Int. Symp. on Information Theory and its Applications*, Auckland, New-Zealand, Dec. 2008, pp. 3966–3969.

[15] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Inform. Process. Lett.*, vol. 103, no. 6, pp. 227–233, 2007.

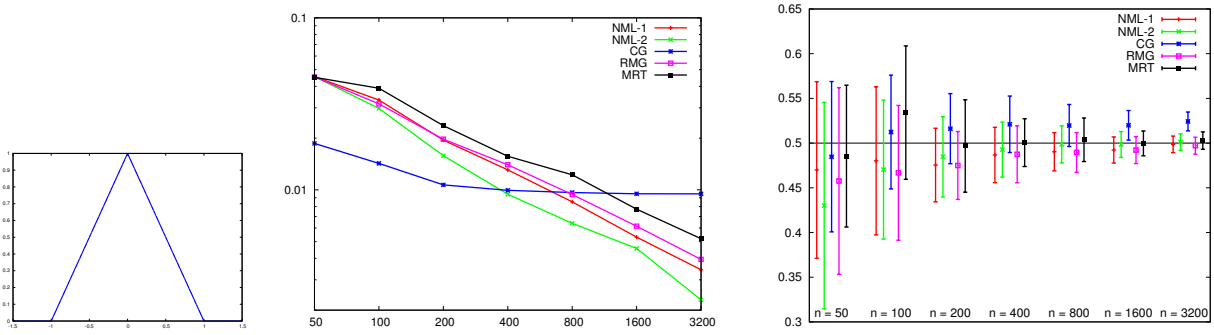
[16] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Probl. Inf. Transm.*, vol. 34, no. 2, pp. 142–146, 1998.



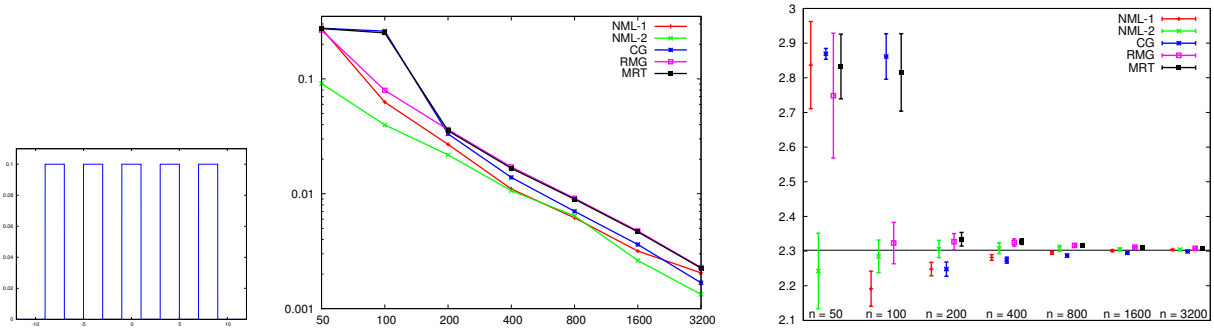
Example 1. Mixture of 6 normal distributions: $\mathcal{Z} = \mathbb{R}$, $s_* = \frac{32}{63}\varphi(-\frac{31}{21}, \frac{32^2}{63^2}) + \frac{16}{63}\varphi(\frac{17}{21}, \frac{16^2}{63^2}) + \frac{8}{63}\varphi(\frac{41}{21}, \frac{8^2}{63^2}) + \frac{4}{63}\varphi(\frac{53}{21}, \frac{4^2}{63^2}) + \frac{2}{63}\varphi(\frac{59}{21}, \frac{2^2}{63^2}) + \frac{1}{63}\varphi(\frac{62}{21}, \frac{1^2}{63^2})$, where $\varphi(\overline{\mu}, \sigma^2)$ is the density function of a normal distribution with mean $\overline{\mu}$ and variance σ^2.



Example 2. Mixture of 2 shifted exponential and 2 normal distributions: $\mathcal{Z} = \mathbb{R}$, $s_* = 0.1f_1 + 0.3\,\varphi(-10, 3^2) + 0.4\,\varphi(0, 2^2) + 0.2f_2$ where $f_1(z) = \exp(-(-20 - z))\,\mathbf{1}_{]-\infty,-20]}(z)$ and $f_2(z) = 0.2\exp(-(z-8)/5)\,\mathbf{1}_{[8,\infty[}(z)$.



Example 3. Triangular distribution: $\mathcal{Z} = [-1, 1]$, $s_*(z) = z + 1$ if $z \in [-1, 0]$ and $s_*(z) = 1 - z$ if $z \in]0, 1]$.



Example 4. Mixture of 5 uniform distributions: $\mathcal{Z} = [-9, 9]$, $s_* = (1/5)\sum_{m=1}^{5} U(\mathcal{I}_m)$, where $\mathcal{I}_m = [-13 + 4m, -11 + 4m]$ for all $m \in \{1, \ldots, 5\}$ and $U(\mathcal{I})$ is the density function of a uniform distribution on the interval $\mathcal{I}$.

Fig. 1. Comparison of the five methods in density and entropy estimation. The source distributions are shown in the leftmost column. In the middle column, the average squared Hellinger distances to the true distribution are presented on a logarithmic scale. The estimated entropies in nats are plotted as error bars in the rightmost column, the centre point of a bar indicating the average and the total height of a bar corresponding to two sample standard deviations. The horizontal line indicates the true value of entropy in nats.