Physica A 388 (2009) 3969–3984

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

The minimum description length principle for probability density estimation by regular histograms

François Chapeau-Blondeau *, David Rousseau

Laboratoire d'Ingénierie des Systèmes Automatisés (LISA), Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers, France

ARTICLE INFO

Article history:
Received 17 December 2008
Received in revised form 15 May 2009
Available online 16 June 2009

PACS:
05.45.Tp
05.40.-a

Keywords:
Statistical information processing
Data analysis
Random signal analysis
Probability density estimation
Minimum description length
Statistical information theory

ABSTRACT

The minimum description length principle is a general methodology for statistical modeling and inference that selects the best explanation for observed data as the one allowing the shortest description of them. Application of this principle to the important task of probability density estimation by histograms was previously proposed. We review this approach and provide additional illustrative examples and an application to real-world data, with a presentation emphasizing intuition and concrete arguments. We also consider alternative ways of measuring the description lengths, that can be found to be more suited in this context. We explicitly exhibit, analyze and compare, the complete forms of the description lengths with formulas involving the information entropy and redundancy of the data, and not given elsewhere. Histogram estimation as performed here naturally extends to multidimensional data, and offers for them flexible and optimal subquantization schemes. The framework can be very useful for modeling and reduction of complexity of observed data, based on a general principle from statistical information theory, and placed within a unifying informational perspective.

1. Introduction

In statistical information processing, probability density estimation is a ubiquitous and very useful process. For probability density estimation from observed data, a much common approach proceeds through the construction of an empirical histogram with regular (equal width) bins. When a fixed number of bins is imposed, the construction of a histogram is a rather straightforward operation. However, the number of bins in itself has a major impact on the quality of the estimation realized by the histogram for the underlying probability density. For a given number N of observed data points, if the number of bins is too small, the resolution of the histogram is poor and leads to a very raw estimate of the probability density. On the contrary, if the number of bins is too large, the counts of data points in the bins fluctuate strongly to yield a very jerky histogram as a poor estimate for the probability density. This points to an optimal number of bins between these two extremes that will lead to an optimal histogram for estimating the probability density. Any approach aiming at determining an optimal number of bins needs necessarily to rely on a definite criterion to measure optimality in this context with histograms. A specially interesting approach of this type is based on the principle of minimum description length.

The minimum description length (MDL) principle provides a general approach for statistical modeling and inference from observed data [1–3]. Briefly stated, this principle amounts to choosing for data, among a class of possible models, the model that allows the shortest description of the data. The MDL approach is rooted in the Kolmogorov theory of complexity [4]. Since its formal introduction some thirty years ago [5], the MDL principle has developed along both theoretical and practical directions. The theoretical foundations of the MDL principle have been investigated to great depth in statistics, and new theoretical aspects are still being explored [1,3,6]. At the same time, the MDL principle has been considered to provide solutions to a large variety of problems, including nonlinear time series modeling [7,8], Markov-

* Corresponding author.
E-mail address: chapeau@univ-angers.fr (F. Chapeau-Blondeau).

0378-4371/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2009.05.044

process order estimation [9,10], data clustering [11,12], signal denoising [13,14], image segmentation [15,16], curve fitting [17,18], analysis of chaotic systems [19,20], genomic sequencing [21,22], neural networks [23,24]. Novel applications also are still emerging [6]. We believe that the MDL approach still holds many potentialities relevant to scientific investigation. A specifically interesting aspect is that the MDL principle offers a unifying thread for approaching many distinct tasks of signal and data processing that otherwise would stand as separate problems. Furthermore, the unified view which is provided is formulated as a information-theoretic framework, and this may be specially relevant to advance an information point of view in science [25–27].

Application of the MDL principle to probability density estimation by histograms was introduced in Ref. [28]. Part of the present paper consists in reviewing this approach of Ref. [28], and also in providing additional illustrative examples, through a presentation emphasizing intuitive and concrete arguments. Implementation of the MDL principle critically relies on definite specifications for measuring the description lengths. As another part of the present paper, we also consider alternative ways of measuring the description lengths, which differ from the choice made in Ref. [28], and which arguably can be found more suited in this context of probability density estimation by histograms. We also explicitly exhibit here the complete forms of the description lengths that arise from the various choices, through formulas involving the information entropy and redundancy of the data, and which are not given in other studies. And we analyze and compare these formulas for the description lengths. We also provide an application to measured data, in the line of a presentation emphasizing concrete and physical appreciation of the MDL approach. In this way, for a part the present paper has a pedagogical and illustrative intent as it proposes a detailed and illustrated review emphasizing concrete interpretations and intuition, on the MDL principle for probability density estimation by histograms. For another part, the paper provides additional results and insight with comparison of alternative choices and complementary analyses.

Minimum description length is often associated with another comparable approach identified as minimum stochastic complexity. These are two distinct, although related, approaches. In particular, stochastic complexity is usually based on the introduction, for the parameters of the model, of a specific prior probability distribution, upon which the subsequent results depend. A uniform prior can be used as in Ref. [28], or the so-called Jeffreys prior as in Ref. [3]. Both description length and stochastic complexity are examined in Ref. [28] for probability density estimation by histograms. Ref. [29] concentrates on stochastic complexity with uniform prior for probability density estimation by histograms. These two notions of description length and stochastic complexity can be defined as distinct notions, as it emerges from Refs. [28,29,3]. However, some other studies imply the terminologies "description length" and "stochastic complexity" as synonyms to designate a same underlying notion. Ref. [30] uses the terminologies "description length" and "stochastic complexity" essentially as synonymous, although there is a single underlying notion which is description length as we understand it here, and not stochastic complexity as in Refs. [28,29,3]. Ref. [30] provides detailed mathematical proofs concerning asymptotic properties and a general theoretical bound, through the introduction of an index of resolvability, for the statistical accuracy and efficacy of probability density estimation by any type of estimators, not necessarily histograms. Further refinements and improvements on these theoretical properties are given in Ref. [31]. Two asymptotic theorems are also proved in Ref. [28], and two theorems concerning upper bounds are established in Ref. [29]. Ref. [32] confronts, for histogram estimation, several forms of penalized maximum-likelihood methods that include the MDL and stochastic complexity based approaches of Ref. [28]. Refs. [33,34] present another form of MDL for histogram density estimation, as they define stochastic complexity by means of the notion of normalized maximum likelihood to avoid a specific prior and in order to obtain a minimax optimality, and then complement this stochastic complexity by a measure of the description length of the parameters to form the criterion to be minimized. In our present paper, for probability density estimation by histograms, we concentrate on the minimum description length, as in Ref. [28] and Ref. [30], and not on the minimum stochastic complexity as considered in Refs. [28,29] with uniform prior, or in Ref. [3] with Jeffreys' prior, or in Refs. [33,34] via normalized maximum likelihood. We see this minimum description length endowed with the advantage of a simple and concrete informational interpretation which is not shared by the minimum stochastic complexity. We review, illustrate and complement the MDL approach here. So far, MDL for probability density estimation by histograms has mainly been discussed in the literature connected to information theory and statistics. Formal proofs have been established for important mathematical properties of the approach. As a complement, we propose here to discuss the MDL methodology in a more physically-oriented presentation, leaning on concrete intuition and illustrative examples. Such a relation between information theory and statistical physics seems interesting to us to promote for the potentialities of mutual enrichment, as for instance illustrated by the recent studies of Refs. [35–38].

2. A histogram model for probability density

One disposes of N observed data points x_n forming the data set

$$\mathbf{x} = \{x_n, n = 1, \ldots N\}. \tag{1}$$

These N data points x_n are assumed to be N independent realizations of a random variable X distributed according to the probability density function $f(x)$. The probability $P(\mathbf{x})$ of observing a given data set $\mathbf{x}$ is therefore expressible as

$$P(\mathbf{x}) = \mathrm{d}x^N \prod_{n=1}^{N} f(x_n), \tag{2}$$

where dx measures the infinitesimal domain of reference around x_n.

One seeks to estimate the probability density $f(x)$ from the N data points x_n of Eq. (1). For this purpose, a histogram model is introduced for the unknown density $f(x)$ under the common form of an approximation by a piecewise constant function. This histogram model is denoted $\mathcal{M}$ and is defined as follows. The density $f(x)$ is modeled by K constant plateaus of value f_k, for $k = 1$ to K, each of these plateaus being defined in the abscissa between $x_{\min}$ and $x_{\max}$ over a regular bin of width

$$\delta x = \frac{x_{\max} - x_{\min}}{K} = \frac{\Delta x}{K}, \tag{3}$$

with $x_{\min}$ and $x_{\max}$ respectively the minimum and maximum values of the x_n's over the data set $\mathbf{x}$ of Eq. (1). Especially, consistency of the probability density model imposes

$$\sum_{k=1}^{K} f_k \delta x = 1. \tag{4}$$

The probability $P(\mathbf{x})$ of Eq. (2), based on the histogram model $\mathcal{M}$ for the density $f(x)$, is expressible as

$$P(\mathbf{x}) = \mathrm{d}x^N \prod_{k=1}^{K} f_k^{N_k}, \tag{5}$$

where N_k is the number of data points x_n of the data set $\mathbf{x}$ that fall within bin number k, verifying $\sum_{k=1}^{K} N_k = N$.

3. Maximum-likelihood histogram estimation

When the number of bins K is fixed, the density model $\mathcal{M}$ is specified by the K parameters f_k for $k = 1$ to K. To determine these parameters from the data, a standard approach is the maximum-likelihood method [39] which consists in selecting those values of the parameters f_k that maximize the probability $P(\mathbf{x})$ in Eq. (5) of the observed data set $\mathbf{x}$. Maximizing $P(\mathbf{x})$ of Eq. (5) under the constraint of Eq. (4) is achieved by the well-known maximum-likelihood solution

$$\widehat{f}_k = \frac{N_k}{N \delta x}, \quad k = 1, \ldots K. \tag{6}$$

The maximum-likelihood solution of Eq. (6) completely specifies, for the probability density $f(x)$, the histogram model with a fixed number K of regular bins.

4. Minimum description length

Another point of view can be adopted to arrive at the solution of Eq. (6). Information theory stipulates that to code data x_n appearing with probability $P(x_n)$, the optimal code assigns a codeword with length $-\log P(x_n)$. To code the whole data set $\mathbf{x}$ of Eq. (1), the optimal code assigns a length $-\log P(\mathbf{x})$, which by the probability model of Eq. (5) is

$$L_{\text{data}} = -\log P(\mathbf{x}) = -\log(\mathrm{d}x^N) - \sum_{k=1}^{K} N_k \log(f_k). \tag{7}$$

The maximum-likelihood solution of Eq. (6) maximizes the likelihood $P(\mathbf{x})$ of Eq. (5) and equivalently the loglikelihood $\log P(\mathbf{x})$. Therefore, the solution of Eq. (6) also minimizes the code length $L_{\text{data}} = -\log P(\mathbf{x})$ of Eq. (7). The solution of Eq. (6) selects from the data, the K parameters f_k of the probability density model $\mathcal{M}$, so that the optimal code designed for the data from this density model, achieves the minimal code length. This is the rationale of the MDL principle: to select the parameters of the model that allow the shortest coding of the complete data. This guarantees that the selected model is the best (within its class) at capturing the structures and regularities in the data.

We can add here, that the minimum of the description length (7) achieved by the solution of Eq. (6) can be expressed as

$$L_{\min} = NH(\{\widehat{p}_k\}) - N \log(K) + N \log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{8}$$

where we have introduced the entropy

$$H(\{\widehat{p}_k\}) = -\sum_{k=1}^{K} \widehat{p}_k \log(\widehat{p}_k) \tag{9}$$

of the empirical probabilities $\widehat{p}_k = \widehat{f}_k \delta x = N_k/N$ deduced from Eq. (6).

Here, when the number of bins K of the histogram model is fixed in an a priori way, the MDL solution coincides with the maximum-likelihood solution of Eq. (6). However, the MDL principle can be extended to also optimally select the number of bins K of the model from the data, along with the K parameter values f_k for $k = 1$ to K. This extension proceeds in the

following way. The complete coding of the data should here include two parts. The first part is the coding of the data based on a definite probability density model to assign the code lengths. For a given data set $\mathbf{x}$, the description length needed by this first part is L_{data} of Eq. (7), that we can also write $L_{\text{data}} \triangleq L(\mathbf{x}|\mathcal{M})$, the description length of the data given a definite model $\mathcal{M}$ of probability density. The second part needed for a complete coding of the data is the description of the parameters that completely specify the underlying probability density model $\mathcal{M}$. These parameters include the number of bins K along with the K values f_k for $k = 1$ to K. The description length needed by this second part in charge of coding the parameters of the model $\mathcal{M}$ is denoted $L_{\text{model}} \triangleq L(\mathcal{M})$; and we shall soon see how to explicitly quantify this description length $L(\mathcal{M})$. Now the complete coding of the data set $\mathbf{x}$ has a total description length L_{total} which sums up the two parts as

$$L_{\text{total}} \triangleq L(\mathbf{x}|\mathcal{M}) + L(\mathcal{M}), \tag{10}$$

signifying that the total description length of the data is the description length of the data given the model plus the description length of the model.

For a given data set $\mathbf{x}$, the MDL principle then dictates to select the model parameters $\{K; f_k, k = 1, \ldots K\}$ so as to minimize the total description length L_{total} of Eq. (10), i.e.

$$\{\widehat{K}; \widehat{f}_k, k = 1, \ldots \widehat{K}\} = \arg\min_{\{K;f_k\}} L_{\text{total}} = \arg\min_{\{K;f_k\}} [L(\mathbf{x}|\mathcal{M}) + L(\mathcal{M})]. \tag{11}$$

This is an optimization principle based on optimal coding and information theory. In a prescribed class of models (histograms with regular bins here), the best model for the data is the model that, when known, enables the most efficient (shortest) coding of these data.

5. Description length for the data

As already stated, the description length $L(\mathbf{x}|\mathcal{M})$ for the data given the model is supplied by Eq. (7). The term $-\log(\mathrm{d}x^N)$ in Eq. (7) is a constant common to all models. For the purpose of discriminating among models, it is often chosen to omit this constant $-\log(\mathrm{d}x^N)$ in the description length, with no impact on the final result concerning the model choice. However here, we prefer to maintain this term, in order to keep track of the complete value of the description length, and convey some additional insight into the modeling process beyond the choice of the model itself. So equivalently, the description length of Eq. (7) for the data given the model is written as

$$L(\mathbf{x}|\mathcal{M}) = -\sum_{k=1}^{K} N_k \log(f_k \mathrm{d}x). \tag{12}$$

Next, we have to address the quantification of the description length $L(\mathcal{M})$ for the model.

6. Description length for the model parameters as independent real variables

To quantify the description length $L(\mathcal{M})$ of the model, a possibility is to use a procedure derived from Ref. [28]. The approach from Ref. [28] to quantify the description length $L(\mathcal{M})$ of the model, considers the K model parameters f_k as K independent real (continuously-valued) variables, which need to be quantized to finite precision in order to allow their coding. The histogram model for the density of the data assigns a probability $p_k = f_k\delta x$ to bin k with width δx. Under this model also, the number N_k of data points falling in bin k has expected value $\mathrm{E}(N_k) = Np_k = Nf_k\delta x$ and standard deviation $\sigma(N_k) = [Nf_k\delta x(1-f_k\delta x)]^{1/2}$, according to the properties of the binomial distribution [40]. Therefore, since $f_k = \mathrm{E}(N_k)/(N\delta x)$, for all k, estimating f_k is equivalent to estimating the mean $\mathrm{E}(N_k)$ of random variable N_k with standard deviation $\sigma(N_k)$. The value $\sigma(f_k) = \sigma(N_k)/(N\delta x) = [f_k(1-f_k\delta x)/(N\delta x)]^{1/2}$ fixes a natural precision with which f_k can be estimated and need to be coded. This determines $\sigma(f_k)$ as the quantization step relevant for coding the model parameters f_k. One has the probability $p_k \in [0, 1]$ and the density $f_k = p_k\delta x^{-1} \in [0, \delta x^{-1}]$. The parameter f_k therefore can take its values in the interval $[0, \delta x^{-1}]$ and is estimated and quantized with the precision $\sigma(f_k)$. Accordingly, a total number $\delta x^{-1}/\sigma(f_k)$ of different values for f_k can be distinguished and need to be coded separately, at a code length $\log[\delta x^{-1}/\sigma(f_k)]$. For the K parameters f_k the code length results as

$$L(\{f_k\}) = \sum_{k=1}^{K} \log\left[\frac{\delta x^{-1}}{\sigma(f_k)}\right] = \frac{K}{2}\log(N) - \frac{1}{2}\sum_{k=1}^{K} \log[f_k\delta x(1-f_k\delta x)]. \tag{13}$$

An alternative, comparable, approach to quantify the cost of coding continuously-valued parameters is described in Ref. [1], based on a slightly more involved mathematical formulation. It turns out that quantifying the coding cost of continuously-valued model parameters is an important and recurrent step when applying the MDL principle. We review this alternative approach from Ref. [1] in the Appendix, for better appreciation of different existing variants for applying the MDL principle. With the present approach derived from Ref. [28] and proceeding through Eq. (13), the description length

for the model is $L(\mathcal{M}) = L(\{f_k\})$, which is then added to the description of the data given the model $L(\mathbf{x}|\mathcal{M})$ of Eq. (12). The total description length $L_{\text{total}} = L(\mathbf{x}|\mathcal{M}) + L(\mathcal{M})$ of Eq. (10) then results as

$$L_{\text{total}} = -\sum_{k=1}^{K}\left(N_k \log(f_k \mathrm{d}x) + \frac{1}{2}\log[f_k\delta x(1-f_k\delta x)]\right) + \frac{K}{2}\log(N). \tag{14}$$

The model parameters $\{K; f_k\}$ are then determined by minimizing the total length L_{total} of Eq. (14), under the constraint of Eq. (4). To simplify this minimization, it is possible to use an approximation as in Ref. [28]. In Eq. (14), the quantity $f_k\delta x$ is the probability p_k of bin k under the histogram model of the probability density. The number of bins K can often be expected to be sufficiently large to assume this probability $f_k\delta x \ll 1$, authorizing the approximation

$$\log(1-f_k\delta x) \approx -f_k\delta x. \tag{15}$$

Under this approximation, the code length of Eq. (13) for the model parameters reduces to (in nats)

$$L(\{f_k\}) = \frac{K}{2}\log(N) + \frac{1}{2} - \frac{1}{2}\sum_{k=1}^{K}\log(f_k\delta x), \tag{16}$$

and the minimization of L_{total} of Eq. (14) can be performed in two steps. First, at given K, the solution for the f_k's realizing, under the constraint of Eq. (4), the minimum of L_{total}, is accessible in closed form as

$$\widehat{f}_k = \frac{N_k + 1/2}{N + K/2}\frac{1}{\delta x}, \quad k = 1, \ldots K. \tag{17}$$

Then, when the $\widehat{f}_k$'s of Eq. (17) are plugged back into L_{total} of Eq. (14), one obtains

$$\begin{aligned} L_{\text{total}} = &-\sum_{k=1}^{K}\left[\left(N_k+\frac{1}{2}\right)\log\left(N_k+\frac{1}{2}\right) + \frac{1}{2}\log\left(1-\frac{N_k+1/2}{N+K/2}\right)\right] \\ &+ \left(N+\frac{K}{2}\right)\log\left(N+\frac{K}{2}\right) + \frac{K}{2}\log(N) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right). \end{aligned} \tag{18}$$

A useful equivalent expression of Eq. (18) is

$$L_{\text{total}} = \left(N+\frac{K}{2}\right)H(\{\widehat{p}_k\}) - \frac{1}{2}\sum_{k=1}^{K}\log\left(1-\frac{N_k+1/2}{N+K/2}\right) + \frac{K}{2}\log(N) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{19}$$

where the entropy $H(\cdot)$ as in Eq. (9) is with the empirical probabilities $\widehat{p}_k = \widehat{f}_k\delta x = (N_k + 1/2)/(N + K/2)$ deduced from Eq. (17). Moreover, in the conditions of the approximation of Eq. (15), the sum over k in Eq. (19) evaluates to -1 nat, so as to yield for Eq. (19),

$$L_{\text{total}} = \left(N+\frac{K}{2}\right)H(\{\widehat{p}_k\}) + \frac{1}{2} + \frac{K}{2}\log(N) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right). \tag{20}$$

Eq. (18), or Eq. (19) or (20), defines a function $L_{\text{total}} = L_{\text{total}}(K)$ of the sole (unknown) variable K, whose minimum can be numerically found to determine the minimizer $\widehat{K}$. Together this $\widehat{K}$ and the $\widehat{f}_k$'s of Eq. (17) form the minimum description length solution to the density estimation problem according to the approach proposed in Ref. [28]. It is to be noted that Ref. [28] rather chooses to estimate the bin probabilities p_k rather than the density values f_k as we do in this Section 6, and so the specific formulas may differ between both places; but the philosophy is the same, as far as we understand it in Ref. [28].

An important aspect should be emphasized concerning the approach of this Section 6 to quantify the description length $L(\mathcal{M})$ of the model. The approach codes the model parameters f_k, for $k = 1$ to K, as if they were independent and real (continuously-valued) parameters. Because of the constraint of Eq. (4), the parameters f_k are not independent. Furthermore, any effective estimation of the f_k's will be performed from the integers N_k, which form a minimal sufficient statistic here. Since the K nonnegative integers N_k sum to N, there are only a finite number of feasible configurations for the N_k's, and accordingly only a finite number of possible values for the f_k's (instead of a continuum of values as would suggest their being considered as real variables). By taking these two features (dependency and discreteness) into account, a more efficient coding could be envisaged. Also, the coding of the model parameters in this Section 6 takes the form of a lossy coding, in connection with Eq. (13), based on the quantization at a finite precision $\sigma(f_k)$ of the f_k's treated as continuously-valued parameters. Instead, a lossless coding could be envisaged. This we address now, by considering another way of quantifying the description length $L(\mathcal{M})$ of the model.

7. Description length for the model with joint parameters

As in the previous Section 6, the aim is to code the parameter values f_k, for $k = 1$ to K, that instantiate the histogram model for the probability density of an observed data set $\mathbf{x}$ of Eq. (1). These f_k's to be coded are matched to the data set $\mathbf{x}$ and in actuality are estimated from this data set $\mathbf{x}$. Any effective estimation of the f_k's from $\mathbf{x}$ must be based on the counts N_k of data points per bin, which form a minimal sufficient statistic here. Therefore, coding the model parameters f_k amounts to coding the integers N_k, for $k = 1$ to K. Each integer N_k can assume $N + 1$ distinct values, between 0 and N. As a simple proposal then, lossless coding of an N_k can be realized with a code length of $\log(N + 1)$. There are K of these integers N_k, however they always sum to N, so only $K - 1$ of them need be coded explicitly, the last one being recoverable since N is assumed known by the decoder (at a coding cost common to all models and not included in the description length $L(\mathbf{x})$). So the code length $L(\{f_k\})$ to code the K parameters f_k can be taken as

$$L(\{f_k\}) = (K - 1)\log(N + 1). \tag{21}$$

If one forms the description length of the model as $L(\mathcal{M}) = L(\{f_k\})$ and then adds it to the description $L(\mathbf{x}|\mathcal{M})$ of the data given the model in Eq. (12), the total description length $L_{\text{total}} = L(\mathbf{x}|\mathcal{M}) + L(\mathcal{M})$ of Eq. (10) follows as

$$L_{\text{total}} = -\sum_{k=1}^{K} N_k \log(f_k \mathrm{d}x) + (K - 1)\log(N + 1). \tag{22}$$

The minimization of Eq. (22) according to Eq. (11) can be solved (analytically) first for the f_k's, with the solution again given by Eq. (6), which is also

$$\widehat{f}_k = \frac{N_k K}{N \Delta x}, \quad k = 1, \ldots K. \tag{23}$$

When these $\widehat{f}_k$'s of Eq. (23) are plugged back into Eq. (22), one obtains the description length

$$L_{\text{total}} = NH(\{\widehat{p}_k\}) + (K - 1)\log(N + 1) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{24}$$

with the probabilities $\widehat{p}_k = N_k/N$ deduced from Eq. (23) used in the entropy $H(\cdot)$ of Eq. (9). Eq. (24) is seen as a function $L_{\text{total}}(K)$ of K alone, to be minimized (numerically) to obtain the minimizer $\widehat{K}$. This $\widehat{K}$ together with the $\widehat{f}_k$'s of Eq. (23) provide a complete solution to the problem of optimal probability density estimation by a histogram with regular bins.

It is possible to suggest an improvement for the code length $L(\{f_k\})$ of Eq. (21) for the K parameters f_k. Eq. (21) is based on a separate coding of $K - 1$ values of N_k, but it does not fully exploit the dependence between the N_k's. The fact that the N_k's sum to N, for instance restrains configurations with several N_k's simultaneously close to N. Because of the dependence of the N_k's a global coding of the K values N_k can be achieved, which is more efficient than a separate coding. For K integers N_k ranging between 0 and N and verifying $\sum_{k=1}^{K} N_k = N$, there is a number $A_{N,K}$ of distinct possible configurations given by Ref. [40, p. 38] as

$$A_{N,K} = \frac{(N + K - 1)!}{N!(K - 1)!}. \tag{25}$$

Then, a lossless coding of the K values N_k is feasible by coding one among the $A_{N,K}$ distinct possible configurations. This is achievable with a code length of $\log(A_{N,K})$, leading to the replacement of Eq. (21) by

$$L(\{f_k\}) = \log(A_{N,K}) = \log\left[\frac{(N + K - 1)!}{N!(K - 1)!}\right], \tag{26}$$

which can be verified to be indeed a more efficient (shorter) code length than Eq. (21).

With this description length $L(\{f_k\}) = L(\mathcal{M})$ for the model, added to the description length $L(\mathbf{x}|\mathcal{M})$ of the data given the model in Eq. (12), one obtains the total description length $L_{\text{total}} = L(\mathbf{x}|\mathcal{M}) + L(\mathcal{M})$ of Eq. (10) as

$$L_{\text{total}} = -\sum_{k=1}^{K} N_k \log(f_k\, \mathrm{d}x) + \log(A_{N,K}). \tag{27}$$

At fixed K, the description length L_{total} of Eq. (27) is again minimized by the f_k's given in Eq. (23). When these values are plugged back into L_{total} of Eq. (27), one obtains a total length expressible as

$$L_{\text{total}} = NH(\{\widehat{p}_k\}) + \log(A_{N,K}) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{28}$$

with again the probabilities $\widehat{p}_k = N_k/N$ deduced from Eq. (23) used in the entropy $H(\cdot)$ of Eq. (9). Eq. (28) is seen as a function $L_{\text{total}}(K)$ of K alone, to be minimized to obtain the minimizer $\widehat{K}$, this providing, in conjunction with the $\widehat{f}_k$'s of Eq. (23), a complete solution to the problem of optimal probability density estimation by a histogram with regular bins.

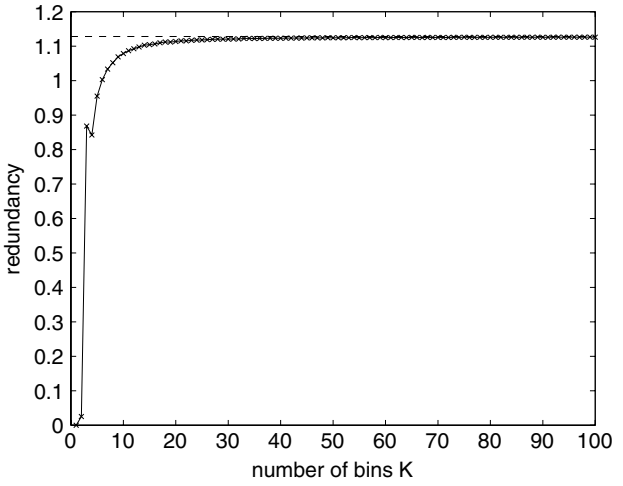

Fig. 1. Redundancy $H_{\max}(K) - H(\{\widehat{p}_k\})$ in bits of the data represented over K bins, as a function of the number of bins K. The data set is formed by $N = 10^5$ points x_n drawn from a Gaussian probability density $f(\cdot)$ with standard deviation $\sigma = \Delta x/9$. The dashed line is the saturation level $\log(\Delta x) - H_{\text{diff}}[f]$ which for the Gaussian density here is $\log[(2\pi \mathrm{e})^{-1/2}\Delta x/\sigma]$.

8. Analysis of the total description lengths

It is now interesting to analyze and compare the various total description lengths L_{total} obtained with the different possible coding strategies, and which are provided by Eqs. (8), (20), (24) and (28). These total lengths show in common a term $N\log(\Delta x/\mathrm{d}x)$. In this term, $\mathrm{d}x$ is the precision or resolution with which the data points x_n are measured or defined, while Δx is the total range over which the data points x_n take their values. For instance, a typical situation could be data represented with 16 binary digits, for which $\Delta x/\mathrm{d}x = 2^{16}$. This amounts to $N\log(\Delta x/\mathrm{d}x) = 16N$ bits which represents the description length associated with direct fixed-length coding of the N data points, with no attempt of optimizing the coding based on a probability model for the data. This common term can thus be understood as the initial description length L_{initial} prior to any optimized coding:

$$L_{\text{initial}} = N\log\left(\frac{\Delta x}{\mathrm{d}x}\right). \tag{29}$$

In the total description lengths, another common term is $N\log(K)$. With K bins to distribute the data points, $\log(K)$ can be interpreted as the maximum entropy $H_{\max}(K) = \log(K)$ achieved with uniform probability over the K bins. In these conditions, the total description length of Eq. (8) takes the form

$$L_{\text{total}} \equiv L_1 = L_{\text{initial}} - N[H_{\max}(K) - H(\{\widehat{p}_k\})]. \tag{30}$$

We recall that this total length L_1 of Eq. (30), is the optimal coding length $L(\mathbf{x}|\mathcal{M}) = -\log P(\mathbf{x}|\mathcal{M})$ for the data set $\mathbf{x}$ based on the histogram model $\mathcal{M}$ of probability density, but when we omit to include any coding cost $L(\mathcal{M})$ for the model itself. In Eq. (30), $H(\{\widehat{p}_k\})$ is the entropy of the empirical probability distribution $\{\widehat{p}_k\}$ estimated from the data over the K bins. It cannot be above the maximum entropy $H_{\max}(K)$, so that in Eq. (30) the difference $H_{\max}(K) - H(\{\widehat{p}_k\})$ is nonnegative, and it measures the information redundancy of the data represented over K bins. The nonnegative redundancy usually expresses in Eq. (30), a reduction of the initial coding length L_{initial} which is afforded by nonuniform (variable-length) coding based on a probability model for the data. This reduction is possible except when the data are distributed with uniform probability, in which case the empirical entropy $H(\{\widehat{p}_k\})$ matches the maximum entropy $H_{\max}(K)$, the redundancy vanishes and no gain on L_{initial} is achieved in Eq. (30) because the uniform (fixed-length) coding in L_{initial} already corresponds to the optimal coding. On the contrary, if the departure of the data from uniform probability is strong, then $H(\{\widehat{p}_k\})$ is much less than $H_{\max}(K)$, the nonnegative redundancy $H_{\max}(K) - H(\{\widehat{p}_k\})$ is large, and a large reduction on L_{initial} is accessible in Eq. (30).

Also in Eq. (30), when the number of bins K is increased, entropy $H_{\max}(K) = \log(K)$ increases as well; in addition, provided K grows while adhering to $K \ll N$, the empirical entropy $H(\{\widehat{p}_k\})$ usually increases as a consequence of this increased resolution K. Moreover, the redundancy $H_{\max}(K) - H(\{\widehat{p}_k\})$ usually also increases with K. This redundancy grows from 0 at $K = 1$ to $\log(\Delta x) - H_{\text{diff}}[f]$ at large K, where $H_{\text{diff}}[f] = -\int f\log f$ is the differential entropy associated with the probability density $f(\cdot)$ of the data. A typical evolution with K of the redundancy $H_{\max}(K) - H(\{\widehat{p}_k\})$ is shown in Fig. 1 when $f(\cdot)$ is a Gaussian density. With $H_{\max}(K) - H(\{\widehat{p}_k\})$ an increasing function of K, the coding length L_1 in Eq. (30) decreases with K, since L_{initial} is invariant with K. Larger K implies higher accuracy in modeling the probabilities of the data, which in turn implies more efficiency in the nonuniform coding based on these probabilities; whence the decreasing coding length L_1 in Eq. (30) as K grows.

However, complete coding of the data, as understood by the MDL principle, implies to count also the coding cost $L(\mathcal{M})$ of the model. This is achieved by the total lengths in Eqs. (20), (24) and (28), yet with different ways of quantifying $L(\mathcal{M})$. Eq. (24) results from a simple yet exact (lossless) coding of the model parameters $\{f_k\}$ according to Eq. (21). The resulting total description length of Eq. (24) can be expressed as

$$L_{\text{total}} \equiv L_3 = L_{\text{initial}} + L_{\text{model}} - N[H_{\max}(K) - H(\{\widehat{p}_k\})], \tag{31}$$

with $L_{\text{model}} = L(\mathcal{M})$ the coding length for the model, which here, from Eq. (21), is

$$L_{\text{model}} = (K-1)\log(N+1). \tag{32}$$

Eq. (28) results from a more efficient lossless coding of the model parameters $\{f_k\}$ according to Eq. (26). The resulting total description length of Eq. (28) can be expressed in a similar form

$$L_{\text{total}} \equiv L_4 = L_{\text{initial}} + L_{\text{model}} - N[H_{\max}(K) - H(\{\widehat{p}_k\})], \tag{33}$$

with now for the model, from Eq. (26),

$$L_{\text{model}} = \log(A_{N,K}). \tag{34}$$

Eq. (20) results from an approximate (lossy) coding of the model parameters $\{f_k\}$ treated as independent real variables quantized to a finite precision. The resulting total description length of Eq. (20) can also be expressed in a rather similar form as

$$L_{\text{total}} \equiv L_2 = L_{\text{initial}} + L_{\text{model}} - \left[NH_{\max}(K) - \left(N + \frac{K}{2}\right) H(\{\widehat{p}_k\})\right], \tag{35}$$

with now for the model, from Eq. (20),

$$L_{\text{model}} = \frac{1}{2} + \frac{K}{2}\log(N). \tag{36}$$

The total description lengths L_{total} of Eqs. (31), (33) and (35) are similar in form with Eq. (30) yet with the visible difference that they explicitly include the coding cost L_{model} of the model, measured in one form or another according to Eqs. (32), (34) and (36). In each case, the final step to the histogram estimation problem rests in solving

$$\widehat{K} = \arg\min_K L_{\text{total}}(K). \tag{37}$$

The total description lengths L_{total} of Eqs. (31), (33) and (35) have in common to include the initial data length L_{initial} which is independent of K, plus the model length L_{model} which usually increases with the number K of model parameters, minus the redundancy which usually is an increasing function of K (see Fig. 1). $L_{\text{total}}(K)$ is then usually formed by an increasing function of K (i.e. $L_{\text{initial}} + L_{\text{model}}$) plus a decreasing function of K (minus the redundancy). One expects then to have a unique minimum and minimizer $\widehat{K}$ for $L_{\text{total}}(K)$ in Eq. (37). This is at least the expected, overall behavior, for a typical data set. In practice, local fluctuations of the estimated empirical entropy $H(\{\widehat{p}_k\})$ can induce small local increase of minus the redundancy with K, as visible in the figures of Section 9. But this does not affect the methodology for solving the problem of histogram estimation.

Before explicitly solving the problem of Eq. (37) in several illustrative examples, it is interesting to compare the coding length L_{model} assigned to the model by the various coding strategies of Eqs. (32), (34) and (36). These three coding lengths are presented in Fig. 2 as a function of the model size K.

Fig. 2 illustrates the behavior that generally holds in the ranges of interest for N and K. In general, the coding length L_{model} from Eq. (36) is the shortest, as it is associated with a lossy coding of the K model parameters. The coding length L_{model} from Eq. (32) is the longest, as it is associated with a lossless coding of the K model parameters yet without taking full advantage of the dependence among these parameters. The coding length L_{model} from Eq. (34) is intermediate, as it is associated with a lossless coding of the K model parameters taking full advantage of the dependence among these parameters. Further insight can be obtained in the condition $1 \leq K \ll N$ which may frequently hold in practice; then the coding length of Eq. (36) becomes $L_{\text{model}} \approx (K/2)\log(N)$, while that of Eq. (32) becomes $L_{\text{model}} \approx (K-1)\log(N)$, and that of Eq. (34) becomes $L_{\text{model}} \approx (K-1)\log(N) - \log[(K-1)!]$. If in addition one considers the condition $1 \ll K \ll N$, Eq. (34) further gives $L_{\text{model}} \approx (K-1)[\log_b(N) - \log_b(K-1) + 1/\ln(b)]$, where b is the logarithm base. These approximate expressions for L_{model} are also shown in Fig. 2, while we keep in mind that it is the dependence in K of L_{model} at given N which is relevant to solve the problem of Eq. (37).

The model description length L_{model} is thus shorter with the lossy coding associated with the total length $L_{\text{total}} = L_2$ of Eq. (35), and by comparison the model description length L_{model} is longer with the lossless coding associated with the total length $L_{\text{total}} = L_4$ of Eq. (33) (and still even longer with $L_{\text{total}} = L_3$ of Eq. (31)).

The other term in the total description length L_{total} formed by minus the redundancy behaves in the opposite way. The entropy $H(\{\widehat{p}_k\})$ in Eqs. (31) and (33) is based on the empirical probabilities $\widehat{p}_k = N_k/N$, while in Eq. (35) entropy $H(\{\widehat{p}_k\})$ is based on the probabilities $\widehat{p}_k = (N_k + 0.5)/(N + K/2)$. In general, for a given set of counts $\{N_k\}$, the probability distribution $\{(N_k + 0.5)/(N + K/2)\}$ in Eq. (35) is closer to uniformity than the distribution $\{N_k/N\}$ in Eqs. (31) and (33). As a result, the entropy $H(\{\widehat{p}_k\})$ is closer to the maximum $H_{\max}(K)$ in Eq. (35) than in Eq. (33). The redundancy $NH_{\max}(K) - (N+K/2)H(\{\widehat{p}_k\})$

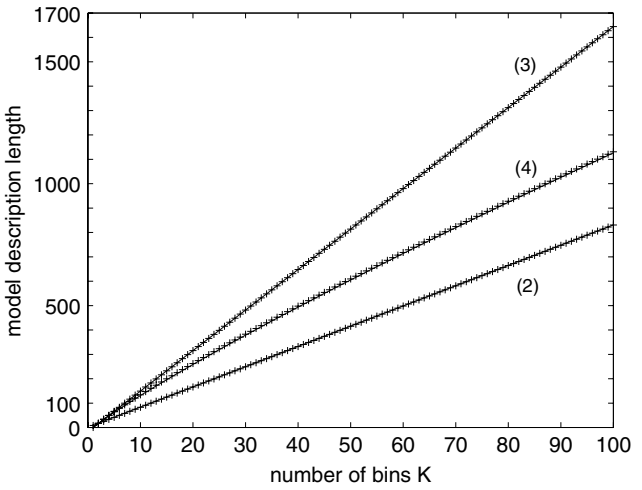

Fig. 2. Description length L_{model} in bits for coding the model, as a function of the number of bins K, for a data set with $N = 10^5$ points. Solid lines: (2) lossy coding of Eq. (36), (3) lossless coding of Eq. (32), (4) lossless coding of Eq. (34). Crosses (+): approximate expressions $L_{\text{model}} \approx (K/2)\log(N)$ for (2), $L_{\text{model}} \approx (K-1)\log(N)$ for (3), $L_{\text{model}} \approx (K-1)[\log_b(N) - \log_b(K-1) + 1/\ln(b)]$ for (4).

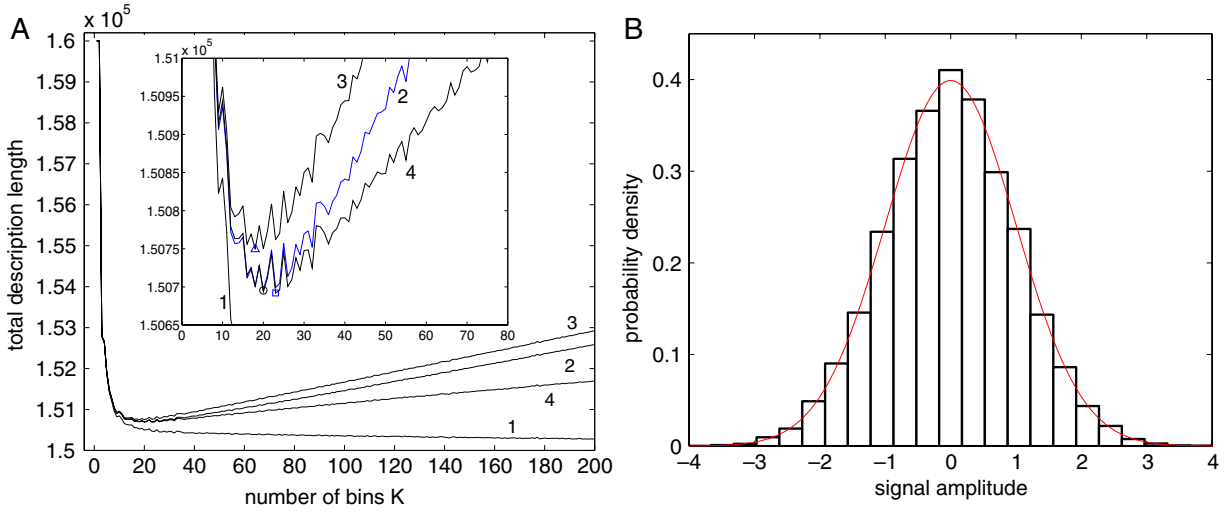

Fig. 3. Panel A: Total description length L_{total} in bits, as a function of the number of bins K, for a data set with $N = 10^4$ points, and $L_{\text{initial}} = 16N = 160\,000$ bits, drawn from probability density $f(\cdot)$ which is the Gaussian $\mathcal{N}(0, \sigma = 1)$ with zero mean and standard deviation $\sigma = 1$: (1) L_1 from Eq. (30) with no model coding, (2) L_2 from Eq. (35) with lossy coding of the model, (3) L_3 from Eq. (31) with lossless coding of the model, (4) L_4 from Eq. (33) with more efficient lossless coding of the model. The inset magnifies the region where the minimum of L_{total} is shown by marker (○): ($\widehat{K} = 20$, $L_2(\widehat{K}) = 150\,695$ bits), (△): ($\widehat{K} = 18$, $L_3(\widehat{K}) = 150\,749$ bits), (□): ($\widehat{K} = 23$, $L_4(\widehat{K}) = 150\,692$ bits). Panel B: Histogram model at the optimum number of bins $\widehat{K} = 23$ minimizing L_4 of Eq. (33), superimposed to the true Gaussian probability density $f(\cdot)$.

in Eq. (35) is therefore smaller than the redundancy $NH_{\max}(K) - NH(\{\widehat{p}_k\})$ in Eq. (33); the effect in this direction is even accentuated by the prefactor $(N + K/2)$ which is stronger in Eq. (35) than the prefactor N in Eq. (33), contributing to the smaller redundancy in Eq. (35).

To summarize, both the model length L_{model} and the redundancy increase with K and are smaller for the lossy coding of Eq. (35) than for the lossless codings of Eqs. (31) and (33). The difference of these two functions of K controls the total description length L_{total}, and one can expect well-defined minimum and minimizer $\widehat{K}$ for $L_{\text{total}}(K)$ in Eq. (37). We now explicitly solve the minimization of Eq. (37) in several illustrative examples.

9. Examples

As a first example, $N = 10^4$ data points x_n were drawn from the probability density $f(\cdot)$ taken as the Gaussian density $\mathcal{N}(0, \sigma = 1)$ with zero mean and standard deviation $\sigma = 1$. The total description length L_{total} has been computed according to the four strategies compared in Section 8, and is shown in Fig. 3A as a function of the number of bins K.

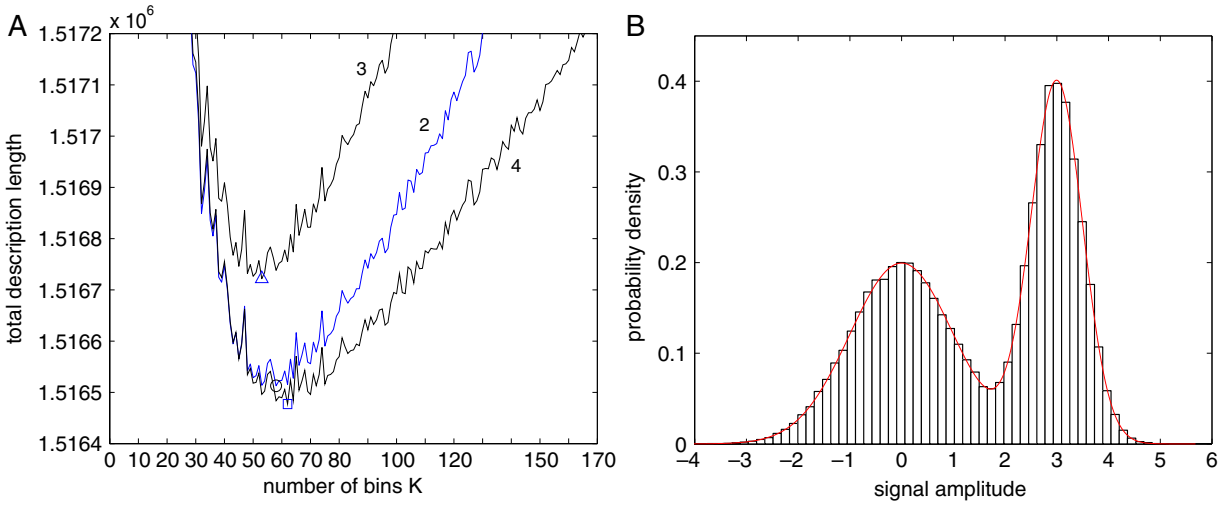

Fig. 4. Panel A: Total description length L_{total} in bits, as a function of the number of bins K, for a data set with $N = 10^5$ points, and $L_{\text{initial}} = 16N = 1\,600\,000$ bits, drawn from probability density $f(\cdot)$ which is the bi-Gaussian mixture $0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(3, 0.5)$: (2) L_2 from Eq. (35) with lossy coding of the model, (3) L_3 from Eq. (31) with lossless coding of the model, (4) L_4 from Eq. (33) with more efficient lossless coding of the model. The minimum of L_{total} is shown by marker (○): ($\widehat{K} = 58$, $L_2(\widehat{K}) = 1\,516\,513$ bits), (△): ($\widehat{K} = 53$, $L_3(\widehat{K}) = 1\,516\,722$ bits), (□): ($\widehat{K} = 62$, $L_4(\widehat{K}) = 1\,516\,477$ bits). Panel B: Histogram model at the optimum number of bins $\widehat{K} = 62$ minimizing L_4 of Eq. (33), superimposed to the true bi-Gaussian probability density $f(\cdot)$.

In Fig. 3A, the total description length L_1 from Eq. (30), steadily decreases with K, as announced, because L_1 does not incorporate the model description length. On the contrary in Fig. 3A, the total description lengths L_2, L_3 and L_4, which incorporate the model description length, exhibit a minimum for an optimal value of K. In Fig. 3A, the length L_3, which comes from a relatively poor coding strategy for the parameters, is, as a rule, always larger than the lengths L_2 and L_4. In the region of the minimum in Fig. 3A, the lengths L_2 and L_4, although they are based on distinct coding strategies, assume very close values. L_2 from Eq. (35) is based on a lossy approximate coding of the parameters: this provides a shorter code length for the parameters associated with a less accurate (longer) coding for the data. On the contrary, L_4 from Eq. (33) is based on an exact lossless coding of the parameters: this costs a longer code length for the parameters associated with a more accurate (shorter) coding for the data. These two complementary situations of L_2 and L_4 tend to compensate in the region of the minimum in Fig. 3A, to lead to close values of the total description length. However, there is a slight superiority of L_4 over L_2 in Fig. 3A, in the sense that L_4, at the optimal setting ($\widehat{K} = 23$, $L_4(\widehat{K}) = 150\,692$ bits), achieves a slightly shorter minimal total length $L_4(\widehat{K}) = 150\,692$ bits and at the same time a higher resolution in the histogram definition with an optimal number of bins $\widehat{K} = 23$. Fig. 3B shows the optimal histogram model estimated for the probability density $f(\cdot)$ of the data set, at $\widehat{K} = 23$.

A second example is presented in Fig. 4, for data points drawn from a Gaussian mixture density. A similar overall behavior is observed in Fig. 4 for the total description lengths L_2, L_3 and L_4 as in Fig. 3. The length L_3 is always larger, while L_2 and L_4 take close values in the region of the minimum. Also in Fig. 4, the shortest description length and at the same time the highest histogram resolution $\widehat{K}$, are achieved by L_4 at the optimal setting ($\widehat{K} = 62$, $L_4(\widehat{K}) = 1\,516\,477$ bits). This is a double benefit associated with L_4: shortest minimal code length and at the same time highest optimal resolution $\widehat{K}$. Although the length L_2 is close to L_4 in the region of the minimum, and both L_2 and L_4 fluctuate in these regions from one data set to another with same size N, this double benefit observed with L_4 in Fig. 4, was never exchanged between L_4 and L_2. This was the rule for all the configurations we tested, for all the densities in this Section 9.

We also tested probability densities that accept a very small number of regular bins for accurate estimation. For uniform densities for which a single bin is adequate, the estimation based on the total lengths L_2, L_3 and L_4, all generally yield the optimal number of bins $\widehat{K} = 1$, with in general the shortest code length afforded by $L_4(\widehat{K})$. Comparable conditions are presented in Fig. 5 with a density which is constant over two separate intervals of equal width, separated by an interval with zero probability.

As visible in Fig. 5, the total lengths L_2, L_3 and L_4, all yield the appropriate number of bins $\widehat{K} = 3$, while the shortest code length is afforded by $L_4(\widehat{K}) = 15\,434$ bits.

Fig. 6 presents the example of a density which is constant over two separate intervals of unequal widths, separated by an interval with zero probability. The total lengths L_2, L_3 and L_4, all yield the appropriate number of regular bins $\widehat{K} = 4$, while the shortest code length is afforded by $L_4(\widehat{K}) = 15\,531$ bits.

10. Application to measured data

This section presents an application of histogram estimation by MDL on measured data. The data x_n are formed by the intensities of gray-level images with size $N = 512 \times 512$ pixels. These intensities are initially measured over 256 levels,

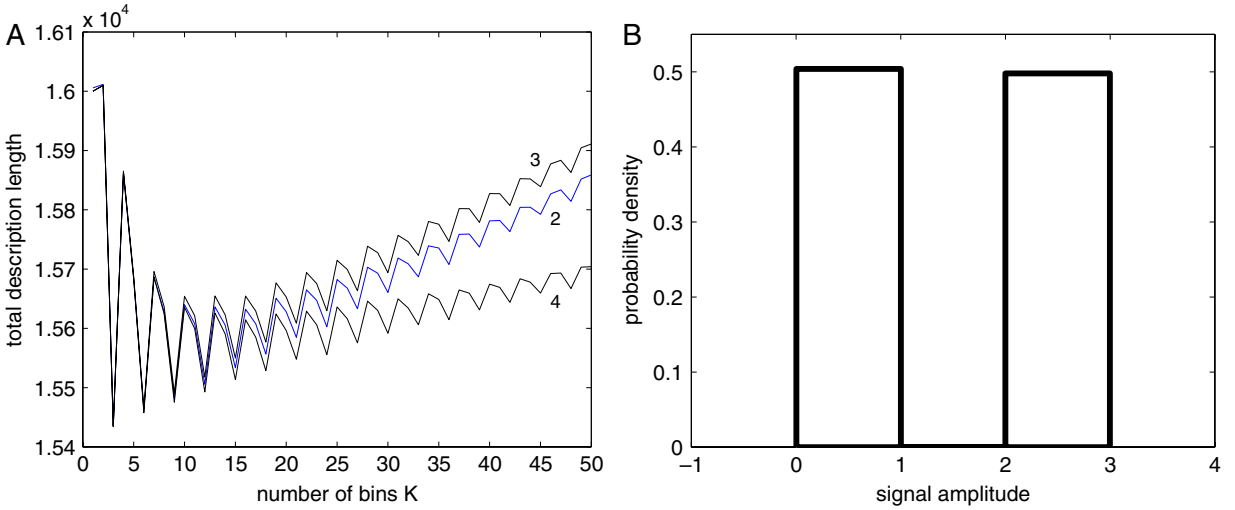

Fig. 5. Panel A: Total description length L_{total} in bits, as a function of the number of bins K, for a data set with $N = 10^3$ points, and $L_{\text{initial}} = 16N = 16\,000$ bits, drawn from probability density $f(\cdot) \sim 0.5\mathcal{U}([0, 1]) + 0.5\mathcal{U}([2, 3])$ which is the mixture of two uniform densities over $[0, 1]$ and $[2, 3]$: (2) L_2 from Eq. (35) with lossy coding of the model, (3) L_3 from Eq. (31) with lossless coding of the model, (4) L_4 from Eq. (33) with more efficient lossless coding of the model. The minimum of L_{total} is, for (2): ($\widehat{K} = 3$, $L_2(\widehat{K}) = 15\,438$ bits), for (3): ($\widehat{K} = 3$, $L_3(\widehat{K}) = 15\,435$ bits), for (4): ($\widehat{K} = 3$, $L_4(\widehat{K}) = 15\,434$ bits). Panel B: Histogram model at the optimum number of bins $\widehat{K} = 3$.

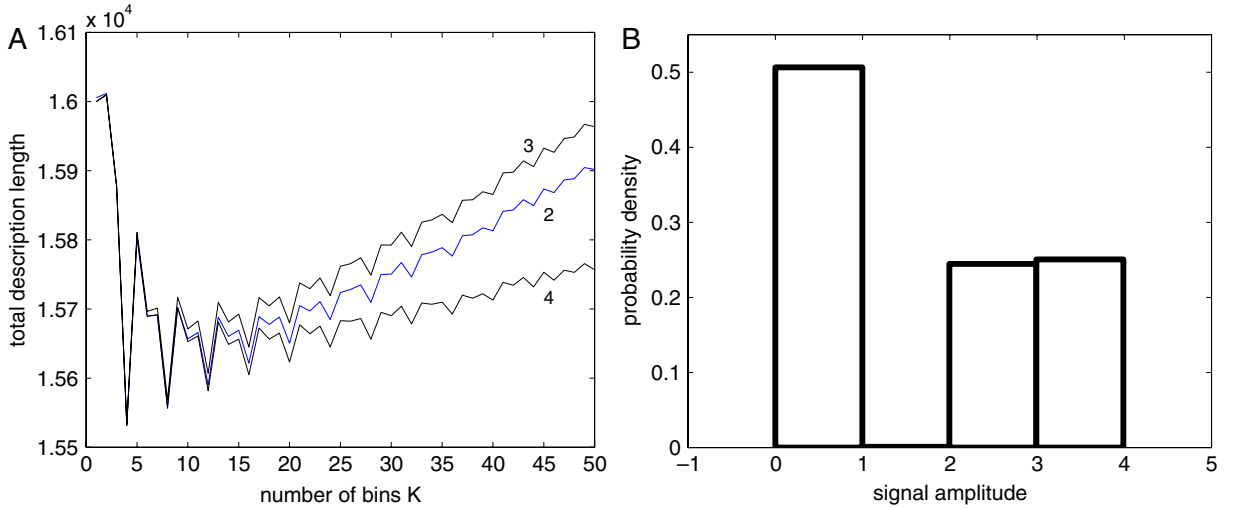

Fig. 6. Panel A: Total description length L_{total} in bits, as a function of the number of bins K, for a data set with $N = 10^3$ points, and $L_{\text{initial}} = 16N = 16\,000$ bits, drawn from probability density $f(\cdot) \sim 0.5\mathcal{U}([0, 1]) + 0.5\mathcal{U}([2, 4])$ which is the mixture of two uniform densities over $[0, 1]$ and $[2, 4]$: (2) L_2 from Eq. (35) with lossy coding of the model, (3) L_3 from Eq. (31) with lossless coding of the model, (4) L_4 from Eq. (33) with more efficient lossless coding of the model. The minimum of L_{total} is, for (2): ($\widehat{K} = 4$, $L_2(\widehat{K}) = 15\,532$ bits), for (3): ($\widehat{K} = 4$, $L_3(\widehat{K}) = 15\,534$ bits), for (4): ($\widehat{K} = 4$, $L_4(\widehat{K}) = 15\,531$ bits). Panel B: Histogram model at the optimum number of bins $\widehat{K} = 4$.

across a range from $x_{\min} = 0$ to $x_{\max} = 1$, with measurement resolution $\mathrm{d}x = 1/256$. For two standard images, minimization of the description length $L_{\text{total}}(K)$ from Eq. (28) leads to the optimal MDL histograms shown in Fig. 7.

The results of Fig. 7 show that the optimal trade-off between accuracy and parsimony according to the MDL principle, is achieved by histograms that employ a number of bins $\widehat{K}$ which is less than the initial 256 levels over which the intensities are initially measured. These optimal values of $\widehat{K}$ derived from an information-theoretic principle, are also consistent with the qualitative appreciation resulting from visual inspection: Image "Lena" displays comparatively less variability and richness of details across the gray levels, and consistently can be adequately represented over a relatively small number $\widehat{K} = 83$ of levels. Meanwhile, image "Boats" displays more variability and richness of details across the gray levels, and consistently requires a larger number $\widehat{K} = 160$ of levels for adequate representation.

In addition, the optimal MDL histograms of Fig. 7 realize what can be viewed as an automatic subquantization of the intensities of the initial images. This subquantization is optimal in an information-theoretic sense expressed by MDL. At the same time, by visual inspection of the images at the optimal subquantization in Fig. 7, no essential features and details concerning the informational content of the images appear to be lost. We have here two properties simultaneously registered

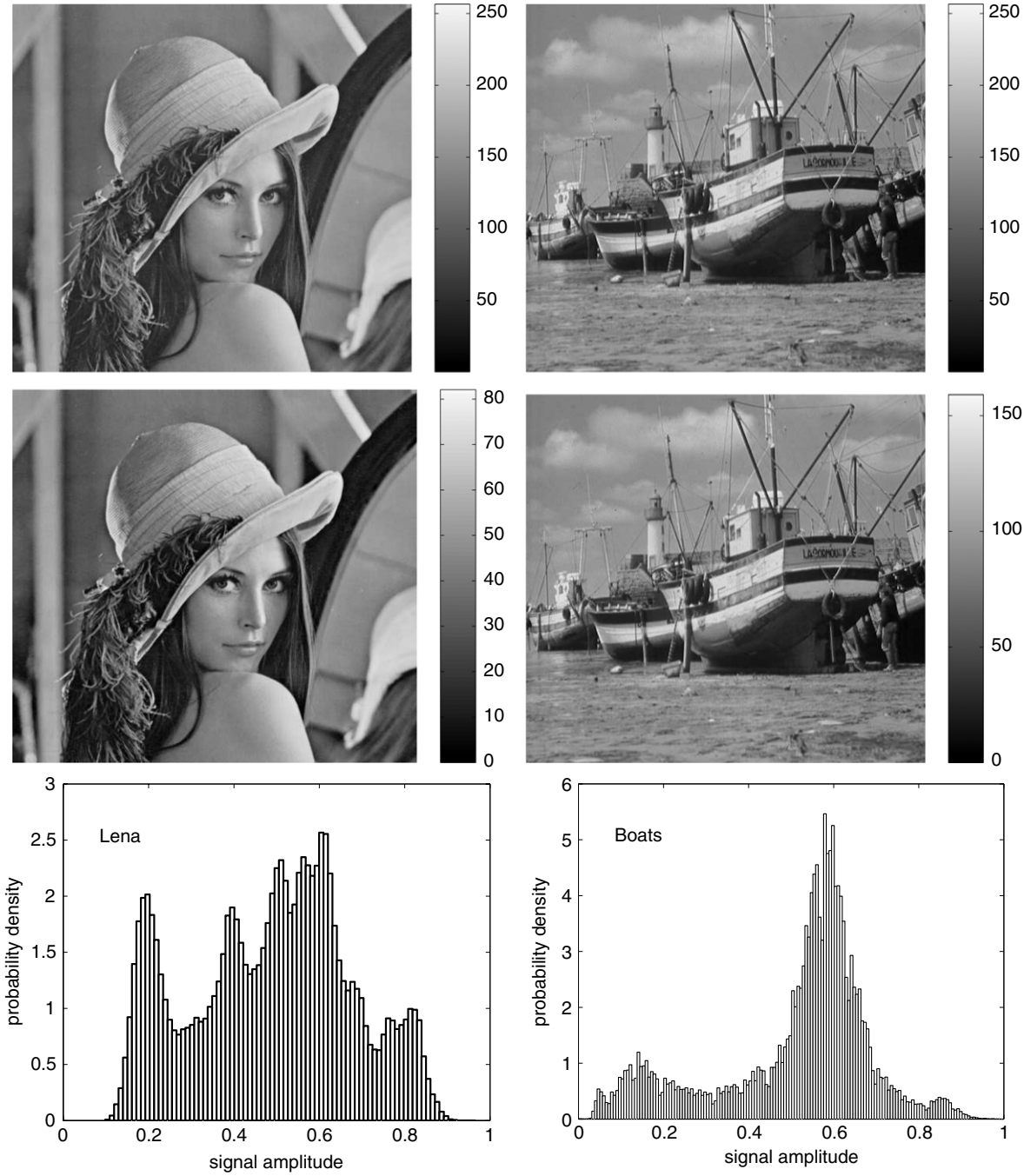

Fig. 7. Top row: Two gray-level images with size $N = 512 \times 512$ pixels initially measured over 256 levels of intensity. Middle row: Images subquantized over the optimal number of bins $\widehat{K}$ minimizing the description length $L_{\text{total}}(K)$ of Eq. (28): $\widehat{K} = 83$ (Lena), $\widehat{K} = 160$ (Boats). Bottom row: Optimal MDL histograms over the $\widehat{K}$ bins.

at two distinct levels (optimal MDL subquantization and visual perception). No connection is explicitly introduced by the MDL procedure between these two properties. However, their simultaneous occurrence could be a mark of some deeper connection originating in the fact that both properties have in common to relate to the informational content of the images.

11. Discussion

As we already mentioned, the approach of Section 6 is based on Ref. [28], and it treats the K model parameters as continuously-valued independent variables, which are approximated to a finite precision and coded separately through a

lossy coding. By contrast, the approach of Section 7 treats the K model parameters as discrete dependent variables, which are jointly coded through an exact lossless coding. These two distinct approaches are best represented by the total description lengths L_2 from Eq. (35) and L_4 from Eq. (33). It is remarkable to observe, based on the examples of Section 9, that these two distinct approaches lead nevertheless to results which are close for the optimal histogram models. This may be interpreted as a mark of robustness of the optimal solutions resulting from the MDL principle, which are not strongly affected by the specific ways used to describe or code the data, provided reasonable and efficient coding methods are confronted. This contributes to confirm that an essential significance of this principle is at a general informational level, and for a part it transcends the quantitative details of the descriptions. From the results of Section 9, a slight superiority though can be granted to the approach via L_4 of Eq. (33) when, for a shorter minimum description length, it affords at the same time a larger resolution of the histogram. There is however another important aspect relevant for a differentiated assessment of the two approaches.

A specificity of the approach from Ref. [28] and Section 6 is that it uses for the model parameters $\{f_k\}$, a code which is not decodable by the receiver. The reason is that this approach is based on a coding procedure, as described in Section 6, which arranges for each parameter f_k a code length which is dependent upon the value of this parameter, instantiated at $\widehat{f}_k$ from the data. This is expressed by the parameter code length of Eq. (13), or under Eq. (15), the approximation of Eq. (16), both bearing explicit dependence on the parameters $\{f_k\}$. The coder, as it knows the data, knows the parameter values $\widehat{f}_k$ estimated from the data, and can therefore arrange the code for these parameters, which is a variable-length code as implied by Eq. (13) or Eq. (16). The receiver receives first the coded parameters, and it needs to decode these parameters to be in a position then to decode the data coded with the variable-length coding based on the probability model specified by the decoded parameters. Therefore, when the coder uses for the parameters a code which depends on the values of these parameters as established by the data, the receiver is unable to decode the parameters since it does not know the data yet. Such a nondecodable coding procedure, however, can still be employed as a benchmark for probability density estimation: It provides a definite coding strategy for which the best achievable coding parsimony (minimum description length) serves to determine an optimal model for the probability density. The approach can thus be felt adequate, because the problem which is tackled at the root is the estimation of a probability density not the actual transmission of data to a putative receiver.

Yet, if the code for the complete data is decodable only by a receiver which already knows the data, one can feel that an adequate model for the data has not been obtained through the coding process. The alternative approach of Section 7 is not limited in this way. It provides a code for the complete data which is perfectly decodable by a receiver which knows nothing about the data. This is obtained based on a coding of the model parameters $\{f_k\}$, which is independent of the actual values of the parameters being coded, as expressed by the parameter code length of Eq. (26) or Eq. (21). In this respect, the approach based on L_4 of Eq. (33) can be preferred as a more appropriate way of applying the MDL principle to probability density estimation by regular histograms.

We add that in principle, the optimal value of $\widehat{K}$ selected by the MDL process should also be coded. This would incur a small additional cost to the total description length L_{total}. An estimate for the code length of $\widehat{K}$ is of order $\log(\widehat{K})$. This, as soon as the number N of data points is not too small, becomes negligible when compared to the parameters code length $L(\{f_k\})$ or the data code length $L(\mathbf{x}|\mathcal{M})$. As a consequence, this additional cost is not included here, with no sensible impact. Finally, for a thorough coding of the complete data set, a few additional informations may also need to be coded, like the number N of data points or the two limits $x_{\min}$ and $x_{\max}$ of the histogram. This would incur another small extra cost to the total description, but this cost is fixed and common to all models so it plays no role in the model selection and is therefore omitted in the MDL process.

The MDL principle for probability density estimation by histograms, can be extended in several directions. A possible direction can consider a wider model class of parametric histograms, consisting of nonregular histograms with a variable number K of bins of unequal widths δx_k, for $k = 1$ to K. At a general level the MDL principle still applies, with all the different widths δx_k which need to be coded at a cost to be included in the description length $L(\mathcal{M})$ of the model. The resulting total description length L_{total} then has to be minimized also as a function of the adjustable variables δx_k. In practice, this leads to a much more computationally demanding multivariate minimization process, in comparison to the approach with equal-width bins which ultimately requires only a minimization according to the single variable K according to Eq. (37). Nonregular histograms are considered in this way in Refs. [29,33,34].

In another direction, the MDL principle for histogram estimation, especially under the form of Section 7 and Eqs. (26)–(28), can be easily extended to histograms in higher dimensionality. For instance in three dimensions, a single data point x_n in Eq. (1) is replaced by a triplet coordinate (x_n, y_n, z_n) representing a joint realization of the random variables (X, Y, Z) distributed according to the three-dimensional probability density $f(x, y, z)$ that one seeks to estimate by a three-dimensional histogram. Each coordinate axis can be divided into, respectively, K_x, K_y and K_z bins with three distinct widths δx, δy and δz. The total number of bins (cells) in three dimensions is $K = K_x K_y K_z$. With this K, the code length for the K model parameters (the K constant values for the density over the K cells) is as before $L(\mathcal{M}) = \log(A_{N,K})$ as in Eq. (26). The total description length L_{total} is given by an expression similar to Eq. (28), controlled by an entropy $H(\{\widehat{p}_k\})$ which is now the entropy in three dimensions estimated from the empirical probabilities $\widehat{p}_k = N_k/N$ over the K three-dimensional cells. One then scans the bin numbers (K_x, K_y, K_z) and associated empirical entropy, searching for the minimum of L_{total} as in Eq. (28). The same process can be performed in arbitrary dimension D. Of course, the search process for the minimization of L_{total} will get more computationally demanding as the dimensionality D of the data space increases, but the principle of the method remains the same. Alternatively, all D coordinate axes can be divided into a unique similar number K_0 of bins, the same for

each axis, leading to a total number $K = K_0^D$ of cells in D dimensions. The search for the minimization of L_{total} is then now reduced to a one-dimensional search in K_0. A much simpler minimization process results, at the price of reduced resolution in the D-dimensional histograms. Other strategies can be envisaged to organize the division of the coordinate axes associated with the MDL criterion. In this same direction of multidimensional histograms with MDL, a recent study [41] introduces a new family of multidimensional histograms in the context of data mining for query answering. A local parametric model is employed to describe each multidimensional bin of the histogram, and local application of MDL is performed to optimize the parameterization. This approach of Ref. [41] is reminiscent of kernel methods, which form a different methodology for probability density estimation based on a definite choice of parametric kernel functions, and which although quite distinct from histograms, can also be handled with MDL for optimal parameterization. Beyond histogram estimation, such extensions of the MDL principle naturally offer, as suggested by the examples of Fig. 7, flexible and optimal subquantizations of multidimensional data based on an informational principle. This can be very useful for reduction of the complexity of multidimensional data sets, which are more and more pervasive in many areas of observational sciences, and with a control of the procedure obtained in an informational framework.

Appendix. Quantization of continuously-valued parameters

In this Appendix, we describe an alternative but comparable approach to Section 6, for quantifying the cost of coding continuously-valued model parameters, based on Ref. [1, p. 55]. The K model parameters f_k take their values in $[0, \delta x^{-1}]$. When considered as continuously-valued, each parameter f_k has to be quantized to a finite precision to make its coding possible. A quantization step or precision h_k is assumed for the quantization of parameter f_k, for $k = 1$ to K. It results in a total number $\delta x^{-1}/h_k$ of different values for f_k which can be distinguished and need to be coded separately, at a code length $\log(\delta x^{-1}/h_k)$. For the K parameters f_k the code length results as

$$L(\{f_k\}) = \sum_{k=1}^{K} \log\left(\frac{\delta x^{-1}}{h_k}\right) = K \log(\delta x^{-1}) - \sum_{k=1}^{K} \log(h_k). \tag{A.1}$$

Given the model parameters $\{f_k\}$, the data $\mathbf{x}$ are coded as in Eq. (7) at the cost $-\log P(\mathbf{x}|\{f_k\})$, which is added to the model cost of Eq. (A.1) to yield the total description length

$$L_{\text{total}}(\{f_k\}) = -\log P(\mathbf{x}|\{f_k\}) + L(\{f_k\}). \tag{A.2}$$

In Eq. (A.2), the f_k's that minimize $L_{\text{total}}(\{f_k\})$ also minimize $-\log P(\mathbf{x}|\{f_k\})$ and are given by the $\widehat{f_k}$'s of Eq. (6). This is so because in Eq. (A.2), the model code length $L(\{f_k\})$ does not (cannot) depend on the f_k's, since these are not known to the decoder. However, it is not the exact $\widehat{f_k}$'s of Eq. (6) which are used to code the data $\mathbf{x}$ at the minimum cost $-\log P(\mathbf{x}|\{\widehat{f_k}\})$. Instead, it is a set $\{f_k\}$, which is close to $\{\widehat{f_k}\}$, and which results from quantization of the $\widehat{f_k}$'s at the finite precisions h_k. This induces a total description length $L_{\text{total}}(\{f_k\})$ slightly longer than the minimum $L_{\text{total}}(\{\widehat{f_k}\})$. Since h_k measures the maximum deviation of f_k from $\widehat{f_k}$, the maximum of the overcost $L_{\text{total}}(\{f_k\})$ above $L_{\text{total}}(\{\widehat{f_k}\})$ can be obtained through the Taylor expansion

$$L_{\text{total}}(\{f_k\}) = -\log P(\mathbf{x}|\{\widehat{f_k}\}) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} J_{ij}(\{\widehat{f_k}\})h_i h_j + L(\{f_k\}), \tag{A.3}$$

with

$$J_{ij}(\{f_k\}) = \frac{\partial^2}{\partial f_i \partial f_j} - \log P(\mathbf{x}|\{f_k\}), \tag{A.4}$$

and no contribution of the first derivatives since $L_{\text{total}}(\{f_k\})$ and $-\log P(\mathbf{x}|\{f_k\})$ are at a minimum in $\{f_k\} = \{\widehat{f_k}\}$. It is then useful to express the total length of Eq. (A.3) as $L_{\text{total}}(\{f_k\}) = -\log P(\mathbf{x}|\{\widehat{f_k}\}) + \Phi(\{h_k\})$. This new function $\Phi(\{h_k\})$, defined from Eq. (A.3), captures the dependence of $L_{\text{total}}(\{f_k\})$ with the precisions $\{h_k\}$ for coding the K parameters f_k. It is then natural to select the precisions $\{h_k\}$ in order to minimize the cost expressed by $\Phi(\{h_k\})$. If the quantization steps $\{h_k\}$ are small, long code lengths are entailed for the parameters $\{f_k\}$, but also high precision is obtained in describing the probabilities of the data, allowing to come close to the minimum code length $-\log P(\mathbf{x}|\{\widehat{f_k}\})$. On the contrary, larger quantization steps $\{h_k\}$ entail shorter code lengths for the parameters $\{f_k\}$, but also less accuracy in describing the probabilities of the data with a coding performance further away from the minimum $-\log P(\mathbf{x}|\{\widehat{f_k}\})$. One can thus expect an optimal configuration for the precisions $\{h_k\}$ that will minimize the total description length of Eq. (A.3), and that will come by nullifying the derivatives

$$\frac{\partial \Phi}{\partial h_k} = \frac{\partial}{\partial h_k}\left[\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} J_{ij}(\{\widehat{f_k}\})h_i h_j - \sum_{k=1}^{K}\log(h_k)\right]. \tag{A.5}$$

Due to the symmetry $J_{ij} = J_{ji}$, it follows from Eq. (A.5),

$$\sum_{i=1}^{K} J_{ik}(\{\widehat{f_k}\})h_i - \frac{1}{h_k} = 0, \tag{A.6}$$

for all $k = 1$ to K.

To make the J_{ij}'s explicit, one has from Eq. (5), $\partial[-\log P(\mathbf{x}|\{f_k\})]/\partial f_i = -N_i/f_i$, and then $J_{ij}(\{f_k\}) = N_i/f_i^2$ if $i = j$, and $J_{ij}(\{f_k\}) = 0$ if $i \neq j$. The optimal precisions $\{h_k\}$ then follow from Eq. (A.6) as

$$h_k = \frac{\widehat{f}_k}{\sqrt{N_k}} = \frac{\sqrt{N_k}}{N}\delta x^{-1}, \tag{A.7}$$

for all $k = 1$ to K. The optimal precisions $\{h_k\}$ of Eq. (A.7) are dependent upon the optimal parameter values $\{\widehat{f}_k\}$. For consistency of their derivation, the optimal coding precisions h_k's of Eq. (A.7) are determined by the $\widehat{f}_k$'s but would not vary for coding other f_k's that would deviate from the $\widehat{f}_k$'s.

At the optimal $\{h_k\}$ of Eq. (A.7) one finds $\sum_i \sum_j J_{ij}(\{\widehat{f}_k\})h_i h_j = K$ and a minimum for $\Phi(\{h_k\})$ which is $K/2 + K\log(N) - \sum_{k=1}^{K}\log(\sqrt{N_k})$, leading to a parameter code length in Eq. (A.1) as

$$L(\{\widehat{f}_k\}) = \frac{K}{2}\log(N) - \frac{1}{2}\sum_{k=1}^{K}\log\left(\frac{N_k}{N}\right). \tag{A.8}$$

Also, in Eq. (A.3), the data code length $-\log P(\mathbf{x}|\{\widehat{f}_k\})$ is provided by Eq. (8). This leads, at the optimal $\{h_k\}$ of Eq. (A.7), for the total description length of Eq. (A.2), to the minimum

$$L_{\text{total}}(\{\widehat{f}_k\}) = -\sum_{k=1}^{K}\left(N_k + \frac{1}{2}\right)\log(N_k) + \frac{K}{2} + \left(N + \frac{K}{2}\right)\log(N) + \frac{K}{2}\log(N) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{A.9}$$

or equivalently

$$L_{\text{total}}(\{\widehat{f}_k\}) = \left(N + \frac{K}{2}\right)H(\{\widehat{p}_k\}) + \frac{K}{2} + \frac{K}{2}\log(N) - N\log(K) + N\log\left(\frac{\Delta x}{\mathrm{d}x}\right), \tag{A.10}$$

with this time the entropy estimator

$$H(\{\widehat{p}_k\}) = -\sum_{k=1}^{K}\frac{N_k + 1/2}{N + K/2}\log\left(\frac{N_k}{N}\right). \tag{A.11}$$

Eqs. (A.8), (A.9) and (A.10) are close, respectively, to Eqs. (16), (18) and (20) from Section 6. However, this approach based on Eqs. (A.8)–(A.10) suffers from the same limitation as the approach of Section 6, as explained in Section 11: the code it uses for the model parameters $\{\widehat{f}_k\}$ is not decodable by the receiver. This is so because Eq. (A.8) arranges for the parameters $\{\widehat{f}_k\}$ a code length which depends on the data through the N_k's. Since these counts N_k are not known to the receiver when it starts its decoding task, this first step of decoding the parameters cannot take place. To circumvent this limitation, a simpler approach [1] would quantify the code length for K real independent parameters as $K\log(N)/2$, specially at large N, to replace Eq. (A.8) or Eq. (16). But for the present histogram parameters, as we mentioned, K independent real (continuously-valued) independent parameters is not a natural assumption, and this is not needed by the alternative approach of Section 7.

References

[1] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, Singapore, 1989.
[2] A.R. Baron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Transactions on Information Theory 44 (1998) 2743–2760.
[3] P.D. Grünwald, The Minimum Description Length Principle, MIT Press, Cambridge, MA, 2007.
[4] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, Springer, Berlin, 1997.
[5] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.
[6] P. Grünwald, I.J. Myung, M. Pitt, Advances in Minimum Description Length: Theory and Applications, MIT Press, Cambridge, MA, 2005.
[7] K. Judd, A. Mees, On selecting models for nonlinear time series, Physica D 82 (1995) 426–444.
[8] L. Diambra, Maximum entropy approach to nonlinear modeling, Physica A 278 (2000) 140–149.
[9] E.J. Hannan, B.G. Quinn, The determination of the order of an autoregression, Journal of the Royal Statistical Society B 41 (1979) 190–195.
[10] O. Alata, C. Olivier, Choice of a 2-D causal autoregressive texture model using information criteria, Pattern Recognition Letters 24 (2003) 1191–1201.
[11] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, H. Tirri, An MDL framework for data clustering, in: P. Grünwald, I.J. Myung, M. Pitt (Eds.), Advances in Minimum Description Length: Theory and Applications, MIT Press, Cambridge, MA, 2005, pp. 323–353.
[12] D.J. Navarro, M.D. Lee, An application of minimum description length clustering to partitioning learning curves, in: Proceedings IEEE International Symposium on Information Theory, Adelaide, Australia, 4–9 Sept. 2005, pp. 587–591.
[13] T. Hediger, A. Passamante, M.E. Farrell, Characterizing attractors using local intrinsic dimensions calculated by singular-value decomposition and information-theoretic criteria, Physical Review A 41 (1990) 5325–5332.
[14] J. Rissanen, MDL denoising, IEEE Transactions on Information Theory 46 (2000) 2537–2543.
[15] S.C. Zhu, A. Yuille, Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1998) 884–900.
[16] F. Galland, N. Bertaux, P. Réfrégier, Minimum description length synthetic aperture radar image segmentation, IEEE Transactions on Image Processing 12 (2003) 995–1006.
[17] T.C.M. Lee, Regression spline smoothing using the minimum description length principle, Statistics and Probability Letters 48 (2000) 71–81.
[18] F. Chapeau-Blondeau, Le principe de longueur de description minimale pour la modélisation des données, ou la théorie statistique de l'information pour bien exploiter les mesures, Le Bup - Bulletin de l'Union des Professeurs de Physique et de Chimie 100 (889 (2)) (2006) 145–155.
[19] R. Brown, N.F. Rulkov, E.R. Tracy, Modeling and synchronizing chaotic systems from experimental data, Physics Letters A 194 (1994) 71–76.

[20] S.S. Chen, L.F. Chen, Y.T. Wu, Y.Z. Wu, P.L. Lee, T.C. Yeh, J.C. Hsieh, Detection of synchronization between chaotic signals: An adaptive similarity-based approach, Physical Review E 76 (2007) pp. 066208, 1–11.
[21] W. Szpankowski, W. Ren, L. Szpankowski, An optimal DNA segmentation based on the MDL principle, International Journal of Bioinformatics Research and Applications 1 (2005) 3–17.
[22] J.S. Conery, Aligning sequences by minimum description length, EURASIP Journal on Bioinformatics and Systems Biology 72936 (2007) 1–14.
[23] R. Meir, J.F. Fontanari, Data compression and prediction in neural networks, Physica A 200 (1993) 644–654.
[24] M. Small, C.K. Tse, Minimum description length neural networks for time series prediction, Physical Review E 66 (2002) pp. 066701, 1–12.
[25] D.G. Luenberger, Information Science, Princeton University Press, Princeton, 2006.
[26] F.A. Bais, J.D. Farmer, The physics of information, in: P. Adriaans, J. van Benthem (Eds.), Philosophy of Information, in: Handbook of the Philosophy of Science, North Holland, Amsterdam, 2008 (Chapter 5b). Also arXiv:0708.2837v2.
[27] M. Mézard, A. Montanari, Information, Physics, and Computation, Oxford University Press, Oxford, 2009.
[28] P. Hall, E.J. Hannan, On stochastic complexity and nonparametric density estimation, Biometrika 75 (1988) 705–714.
[29] J. Rissanen, Density estimation by stochastic complexity, IEEE Transactions on Information Theory 38 (1992) 315–323.
[30] A.R. Baron, T.M. Cover, Minimum complexity density estimation, IEEE Transactions on Information Theory 37 (1991) 1034–1054.
[31] T. Zhang, From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation, The Annals of Statistics 34 (2006) 2180–2210.
[32] L. Birgé, Y. Rozenholc, How many bins should be put in a regular histogram, European Series in Applied and Industrial Mathematics: Probability and Statistics 10 (2006) 24–45.
[33] P. Kontkanen, P. Myllymäki, Information-theoretically optimal histogram density estimation, Technical Report 2006-2 Helsinki Institute for Information Technology, Finland, 17 March 2006.
[34] P. Kontkanen, P. Myllymäki, MDL histogram density estimation, in: Proceedings 11th International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico, 21–24 March 2007.
[35] J.W. Lee, J.B. Park, H.H. Jo, J.S. Yang, H.T. Moon, Minimum entropy density method for the time series analysis, Physica A 388 (2009) 137–144.
[36] G. Nicolis, Equality governing nonequilibrium fluctuations and its information theory and thermodynamic interpretations, Physical Review E 79 (2009) pp. 011106, 1–8.
[37] A. Carbone, H.E. Stanley, Scaling properties and entropy of long-range correlated time series, Physica A 384 (2007) 21–24.
[38] S. Martinez, A. Plastino, B.H. Soffer, Information and thermodynamics first law, Physica A 356 (2005) 167–171.
[39] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1991.
[40] W. Feller, An Introduction to Probability Theory and Its Applications, vol. I, Wiley, New York, 1968.
[41] H. Wang, K.C. Sevcik, Histograms based on the minimum description length principle, The VLDB (Very Large Data Base) Journal 17 (2008) 419–442.