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

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

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 $\mathrm{d}x$ measures the infinitesimal domain of reference around x_n.