Front. Electr. Electron. Eng. China 2010, 5(3): 274–280
DOI 10.1007/s11460-010-0104-0

RESEARCH ARTICLE

Jorma RISSANEN

Basics of estimation

Abstract This paper outlines a theory of estimation, where optimality is defined for all sizes of data — not only asymptotically. Also one principle is needed to cover estimation of both real-valued parameters and their number. To achieve this we have to abandon the traditional assumption that the observed data have been generated by a "true" distribution, and that the objective of estimation is to recover this from the data. Instead, the objective in this theory is to fit 'models' as distributions to the data in order to find the regular statistical features. The performance of the fitted models is measured by the probability they assign to the data: a large probability means a good fit and a small probability a bad fit. Equivalently, the negative logarithm of the probability should be minimized, which has the interpretation of code length. There are three equivalent characterizations of optimal estimators, the first defined by *estimation capacity*, the second to satisfy necessary conditions for optimality for all data, and the third by the *complete* Minimum Description Length (MDL) principle.

Keywords models, optimal estimation, estimation capacity, complete Minimum Description Length (MDL) principle

1 Introduction

In traditional estimation the observed data sequence is thought to have been generated by a "true" distribution, originally assumed to be specified by parameters, which were unknown except for their number, and the central problem is to estimate this distribution. Since the "true" parameter value so-to-speak is hiding among the set of all parameter values considered the problem amounts to estimation of distributions from their samples. An example is the two-parameter family of normal density functions. If we select a distance measure between the estimated and the "true" parameters, its minimization gives us an estimation algorithm. Again an example is the least squares estimator for the mean of a normal density function with a known variance.

A vast generalization of this technique, known already to Gauss, is the familiar maximum likelihood (ML) estimator introduced by Fisher. It amounts to the following: If we have a set $\{f(\cdot;\theta) : \theta \in \Omega^k\}$ of density functions, one for each parameter value in the range Ω^k, and evaluate them all at the fixed observed data sequence $x^n = x_1, \ldots, x_n$, then the resulting map $\theta \mapsto f(x^n;\theta)$ is called the *likelihood function*. The maximizing parameter value $\hat{\theta}(x^n)$ then is taken as the estimate of the "true" parameter. Although the justification of the maximum likelihood estimator was to a large extent intuitive, it was possible to show that under certain conditions the estimates converge in probability to the "true" parameter θ whatever it is, and moreover the limit of the covariance $\lim_n E_\theta\big(\hat{\theta}(x^n) - \theta\big)\big(\hat{\theta}(x^n) - \theta\big)'$ is the smallest possible by the Cramér-Rao inequality.

Such a success led to the further more general assumption that the observed data have been generated by a "nonparametric" density function $f(x^n)$ of some type like continuous, once differentiable, and so on. Guided by the assumed general properties of the "true" data generating density function one could again devise estimation algorithms with good asymptotic properties. Although such estimators are not regarded as "parametric" they actually involve estimation of increasing numbers of parameters, because to fit any density function to data we need an algorithm to calculate its values, and a good way to do it is in terms of lots of parameters.

Everything seemed to be under control except for some disturbing consequences. First, since the recovery of a distribution from its samples is not possible from finite amounts of data the basic problem formulation seems ill posed; by wanting the "truth" we are asking too much. One unfortunate implication is that the obvious optimal performance is asymptotic and hence of little

Received March 1, 2010; accepted March 19, 2010

Jorma RISSANEN (✉)
Helsinki Institute for Information Technology, Tampere University of Technology, Tampere 33720, Finland
E-mail: jrrissanen@yahoo.com

relevance unless the data sequence is 'sufficiently' large, leaving open the question of how large is 'sufficiently' large. Moreover, what perhaps is a bit embarrassing, the asymptotically estimated smallest covariance implies by the Central Limit Theorem that the "truth" cannot be found even asymptotically unless it is normally distributed. Further, a most serious shortcoming is that the estimation of the number of parameters cannot be done by the same maximum likelihood technique. Their estimation called for totally different criteria inspired by the imagined assumptions about the "truth", some good and some less good, without any theory to explain why some techniques were good and for which data they were good. In fact, the word "estimation" is generally applied only to the estimation of the real-valued parameters, while the estimation of their number is called "model selection" with the illogical implication that only the sets of density functions $f(x^n;\theta,k)$ with each fixed number of parameters are "models" but the individual members are not. Finally, although in the philosophical sense the existence of the "true" data generating distribution is neither here nor there, the assumption focuses attention exclusively to the estimation of just one model irrespective of others nearby; after all, the "truth" has no neighbors we are interested in. This could explain the surprising lack of progress in estimation in general and especially in hypothesis testing, which is designed with the crude and lopsided philosophy that 'either this hypothesis is true or something else far away is true'.

We conclude this brief survey of traditional estimation with a few words about the popular Bayesian techniques, which lead to a different family of estimators. We credit the Bayesians, as the appliers of these techniques are called, for the important relaxation of the use of probability, namely, that it can be applied to events such as parameters, even when repeated data of their occurrence do not exist. However, the utility of the foundational assumption that the model class defined by the likelihood function also must include distributions for the parameters, called *priors*, is less clear. These were originally meant to represent prior knowledge about the "true" parameter values, knowledge which must not and cannot be extracted from the given data sequence. Since the priors affect the estimation in an essential manner this of course poses problems when little or no prior knowledge about the probabilities of the events that various parameter values are "true" exists, which is the case in nearly all applications. The procedure then is to look for so-called noninformative priors, which in effect eliminates the need for priors altogether and seems to suggest that something in the foundations, as far as estimation is concerned, is amiss. After all, why create a problem for the overwhelming majority of the applications and then work hard to remove it.

In this paper, instead of searching for the "truth" we regard the objective of estimation to be to learn statistical regular features in an observed data set. For this we need a set of models as distributions to represent the properties we want to learn. This set of models includes both parametric models with a fixed number of real-valued parameters and models where the number of parameters is also to be estimated. Since a data set in general has many properties this view of estimation amounts in effect to asking if the data have this particular property and to what extent. We then need a means to assess the goodness of each estimated model that we fit to the data.

We measure the goodness by the probability or density a distribution defined by the estimated parameters assigns to the observed data: a large probability means a good fit, and, conversely, a small probability a bad fit. Equivalently this can be measured by the negative logarithm of the probability, which admits the additional concrete interpretation as code length. because we can encode both data with repeated occurrence and single isolated data points like parameter values there is no difficulty in the interpretation of probabilities: they are equivalent with code lengths. The fact that encoding data with the shortest code length cannot be done unless we take advantage of the regular features in data suggested the Minimum Description Length (MDL) principle to estimation [1], which in turn seemed to give the impression that estimation and data compression are equivalent. However, they are not because of a subtle but important difference: we can take advantage of the regular features for data compression implicitly without specifying them, while for estimation we must spell them out explicitly.

We view estimation as analogous to measuring a physical property of an object, like its temperature, length, or weight, by a yardstick as the devise or instrument appropriate to the property and capable of providing accurate measurements. The property and the selection of the instrument is guided by physics. The object in our case is the data sequence, and the property we are interested in is its regular statistical feature expressed by the 'theory' provided by the set or models. Actually, the object is a machinery that provides the data, which can be complex consisting of many physical objects such as an economic system. Since nearly all we can learn about the behavior of the machinery is in the data we simply identify the object with the data.

The fundamental concepts needed are then just three: the data set, the set of models, and the set of estimators, from which we construct an optimal estimator in terms of the maximum probability or density assigned to the data. We describe three different but equivalent ways to define the optimal estimator, all of which admit intuitively appealing interpretations. In the resulting theory of estimation nothing is lost by dropping the assumption

of "truth". By contrast, the theory covers virtually all aspects of estimation, and optimality can be defined not only asymptotically but for all amounts of data. As a side result we get a formal justification of the ML estimator on data sets of any size, which hitherto has escaped any attempts.

To conclude this introduction we mention the most important problem in estimation or 'modeling', the selection of the model class itself. The selection of the 'best' model class with the probability assignment as the criterion is non-computable even when the models are restricted to the algorithmically defined ones. This is where prior knowledge is needed rather than just as a distribution on parameter values. For these reasons, the theory of optimal estimation here will be restricted to each somehow given class of models, the goodness of which can be assessed in a comparative sense rather than absolutely.

2 Modeling problem

The modeling problem we are concerned with begins with a set of observed data $Y = \{y_t : t = 1, 2, \ldots, n\}$, or Y observed jointly together with other so-called *explanatory* data $Y, X = \{(y_t, x_{1,t}, x_{2,t}, \ldots)\}$. The objective is to learn statistical properties in Y which are expressed by a set of distributions as *models*

$$\{f(Y|X_s; \theta, s)\}.$$

Here s is a structure parameter and $\theta = \theta_1, \ldots, \theta_{k(s)}$ denotes real-valued parameters, whose number depends on the structure. The *structure* is simply a subset of the models, typically used to indicate the variables in X that affect most strongly the behavior of Y. Usually each model is defined either for a single datum $f(y_t|x_{1,t}, x_{2,t}, \ldots, x_{k(s),t}, s)$ and extended to sequences by independence, or if $x_{1,t}, x_{2,t}, \ldots, x_{k(s),t}$ themselves are defined by the $k(s)$ values of $y_{t-1}, \ldots, y_{t-k(s)}$, the extension is done by the product of the conditionals.

To simplify the notations we write both the data Y and the data Y, X simply as $\mathbf{x}$ without indicating the explanatory data at all. It is understood that the models are distributions on Y or conditional distributions on Y given the explanatory data X if they too are observed, and whether they are distributions or conditional distributions is completely irrelevant to the theory. For the same reason we also consider only structures which are determined by the number of parameters, so that the order is fixed $\theta^k = \theta_1, \ldots, \theta_k$, also written as θ if k is understood and not estimated. We then have the two model classes,

$$\mathcal{M}_k = \left\{ f(\mathbf{x}; \theta, k) : \theta \in \Omega^k \subset R^k \right\},\ k \leqslant n,$$

and

$$\mathcal{M} = \bigcup_{k=0}^{K} \mathcal{M}_k,\ K \leqslant n,$$

depending on whether we are considering the number of parameters fixed or if it, too, is to be estimated. We also write $N_K = \{1, 2, \ldots, K\}$. The class of models $\mathcal{M}$ includes many of the so-called nonparametric models such as histograms. In fact, if nonparametric models can be fitted to data at all they must be defined by an algorithm — itself defined by parameters.

There is a natural generalization of the parametric model classes to catch trends in the data. We may want to make the number of parameters to depend on the size of the data thus $k(\alpha, n)$, where α consists of further parameters. An example is $k = \alpha_1 n^{\alpha_2}$, $\alpha_2 < 1$. Such a generalization causes no difficulties to our treatment of the estimation problem.

Corresponding to the two model classes we have sets $\mathcal{F}_k$ and $\mathcal{F}$ of *estimator* functions:

$$\mathcal{F}_k = \left\{ \bar{\theta}(\cdot) : \mathbf{x} \mapsto \bar{\theta}(\mathbf{x}) \in \Omega^k \right\},$$

$$\mathcal{F} = \left\{ \bar{\theta}(\cdot), \bar{k}(\cdot) : \mathbf{x} \mapsto \bar{\theta}(\mathbf{x}), \bar{k}(\mathbf{x}) \in \bigcup_k \Omega^k \times N_K \right\}.$$

In order to exclude pathological estimators, which define no meaningful learnable statistical properties, we assume, first, for discrete data that $\mathcal{F}_k$ excludes estimators for which $f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right) = 1$ for some data sequence $\mathbf{x}$, and $f\left(\mathbf{y}; \bar{\theta}(\mathbf{y}), k\right) = 0$ for $\mathbf{y} \neq \mathbf{x}$. For data ranging over a continuum, we restrict both the model classes and estimator functions so that they define probabilities for quantized data. For this it is enough if the models in $\mathcal{M}_k$ are continuous functions of the parameters for all data, and the estimator functions in $\mathcal{F}_k$ are continuous. Clearly, these conditions can be weakened but we are not interested in doing so. The only measurable sets we need are finite and open sets together with their closures.

3 Estimators and yardsticks

We begin by defining a family of yardsticks, each defined by an estimator, among which a special one will be selected. First, each estimator of the real valued parameters for fixed k defines a partition of the data set X^n by the equivalence classes $\mathbf{x} \equiv \mathbf{y}$ if $\bar{\theta}(\mathbf{x}) = \bar{\theta}(\mathbf{y})$, together with the model

$$\bar{f}(\mathbf{x}; k) = \frac{f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right)}{\bar{C}_k} = f\left(\mathbf{x}|\bar{\theta}(\mathbf{x}), k\right) \times \frac{\bar{g}\left(\bar{\theta}(\mathbf{x}), k\right)}{\bar{C}_k}, \quad (1)$$

$$f\left(\mathbf{x}|\bar{\theta}(\mathbf{x}), k\right) = \frac{f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right)}{\bar{g}\left(\bar{\theta}(\mathbf{x}), k\right)} = \frac{f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right)}{\bar{g}\left(\bar{\theta}(\mathbf{x}); \bar{\theta}(\mathbf{x}), k\right)},$$

$$\bar{g}\left(\bar{\theta}(\mathbf{x});\theta,k\right)=\int_{\bar{\theta}(\mathbf{y})=\bar{\theta}(\mathbf{x})} f\left(\mathbf{y};\theta,k\right)\mathrm{d}\mathbf{y}, \tag{2}$$

$$\bar{C}_k=\int f\left(\mathbf{x};\bar{\theta}(\mathbf{x}),k\right)\mathrm{d}\mathbf{x}$$

$$=\int_{\Omega^k}\bar{g}\left(\bar{\theta},k\right)\mathrm{d}\bar{\theta}<\infty. \tag{3}$$

There is a way to handle the case where the normalizing integral is infinite, but we do not go into it here.

Similarly, if we also consider estimators for the number of parameters we define a further finite partition of the equivalence classes: $\mathbf{x}\equiv\mathbf{y}$ if $\bar{\theta}(\mathbf{x}),\bar{k}(\mathbf{x})=\bar{\theta}(\mathbf{y}),\bar{k}(\mathbf{y})$ together with the model

$$\bar{f}(\mathbf{x})=\frac{\bar{f}\left(\mathbf{x};\bar{k}(\mathbf{x})\right)}{\bar{C}}=\frac{\bar{f}\left(\mathbf{x};\bar{k}(\mathbf{x})\right)}{\bar{P}\left(\bar{k}(\mathbf{x})\right)}\times\frac{\bar{P}\left(\bar{k}(\mathbf{x})\right)}{\bar{C}}, \tag{4}$$

$$\bar{P}(k)=\int_{\bar{k}(\mathbf{y})=k}\bar{f}\left(\mathbf{y};k\right)\mathrm{d}\mathbf{y}, \tag{5}$$

$$\bar{C}=\sum_k\bar{P}(k). \tag{6}$$

4 Necessary conditions

There is a fundamental difficulty with the maximization of the universal criterion we have selected, namely, the probability or density an estimator assigns to the data by the associated distributions (1) and (4), for the two model classes, respectively. It is because no distribution exists that assigns the maximum probability to all data: we can 'tailor' a distribution to some data point to get a large probability on the cost of sacrificing the performance on other data by assigning small probability to them. After all, there is just so much of the probability mass available. We already did exclude extreme 'tailoring' from the estimator class $\mathcal{F}_k$ but weaker forms of it are still possible which we must deal with.

We describe three equivalent ways to select a special estimator and the yardstick it determines, each of which admits an interpretation illuminating different intuitively attractive properties. We start by determining a necessary condition for optimality of estimators $\bar{\theta}(\cdot)$ at some fixed $\mathbf{x}$ for discrete data. It is as follows: For $\bar{\theta}(\cdot)\in\mathcal{F}_k$ to maximize $\bar{f}(\mathbf{x};k)$ at a point $\mathbf{x}$ it is necessary that

$$\bar{\theta}(\mathbf{x})=\hat{\theta}(\mathbf{x}). \tag{7}$$

This is equivalent to the statement that unless the equality holds a better estimator can be found, or as the condition

$$\bar{\theta}(\mathbf{x})\neq\hat{\theta}(\mathbf{x})\Longleftrightarrow\bar{f}(\mathbf{x};k)<f^*(\mathbf{x};k), \tag{8}$$

where $\bar{\theta}(\cdot)\in\mathcal{F}_k$, $f^*(\mathbf{x};k)=f\big(\mathbf{x};\hat{\theta}(\mathbf{x}),k\big)/C_k^*$, and

$$C_k^*=\bar{C}_k+f\big(\mathbf{x};\hat{\theta}(\mathbf{x}),k\big)-f\left(\mathbf{x};\bar{\theta}(\mathbf{x}),k\right).$$

This is the justification of the ML estimator: if the numerator of the ratio $\bar{f}(\mathbf{x};k)$ at $\mathbf{x}$ is not maximized neither is the ratio itself, which is the probability criterion we have selected as the universal criterion. It then follows that $\hat{\theta}(\cdot)$ is the only estimator which satisfies the necessary condition for all $\mathbf{x}$, and this clearly is the strongest achievable justification.

Similarly, the necessary condition for optimality of estimators $\bar{\theta}(\cdot),\bar{k}(\cdot)$ at $\mathbf{x}$ is

$$\bar{\theta}(\mathbf{x}),\bar{k}(\mathbf{x})=\hat{\theta}(\mathbf{x}),\hat{k}(\mathbf{x}). \tag{9}$$

The estimator $\hat{\theta}(\cdot),\hat{k}(\cdot)$ is the only estimator which satisfies the necessary condition for all $\mathbf{x}$. Clearly, this too is the most we can ask for.

We write the optimal yardstick distribution as follows:

$$\hat{C}_k=\int_{\Omega^k}\mathrm{d}\theta\int_{\hat{\theta}(\mathbf{y})=\theta}f\big(\mathbf{y};\hat{\theta}(\mathbf{y}),k\big)\mathrm{d}\mathbf{y}$$

$$=\int_{\Omega^k}\hat{g}\big(\hat{\theta},k\big)\mathrm{d}\hat{\theta}, \tag{10}$$

$$\hat{f}(\mathbf{y};k)=\frac{f\big(\mathbf{y};\hat{\theta}(\mathbf{y}),k\big)}{\hat{C}_k}, \tag{11}$$

$$\hat{f}(\mathbf{y})=\frac{\max_k f\big(\mathbf{y};\hat{\theta}(\mathbf{y}),k\big)/\hat{C}_k}{\hat{C}}, \tag{12}$$

$$\hat{C}=\sum_k\int_{\hat{k}(\mathbf{y})=k}f\big(\mathbf{y};\hat{\theta}(\mathbf{y}),k\big)\mathrm{d}\mathbf{y}/\hat{C}_k, \tag{13}$$

where in the last equation the structure (number of parameters) that maximizes $\hat{f}(\mathbf{y};k)$ rather than the likelihood $f\big(\mathbf{y};\hat{\theta}(\mathbf{y}),k\big)$ is written as $\hat{k}(\mathbf{y})$. In fact, the number of the parameters that maximizes the likelihood for all data is the maximum possible, K, allowed, and we would not be able to estimate k at all. We see that if we consider the negative logarithms of (1) and (4) as criteria, they are minimized by the two estimators $\hat{\theta}(\mathbf{x})$ and $\hat{k}(\mathbf{x})$, respectively. The model $\hat{f}(\mathbf{y};k)$ was introduced by Shtarkov [2] as a universal model for data compression; we discuss it further below. The second model $\hat{f}(\mathbf{y})$ was introduced in my 2009 Shannon lecture [3]. We do not consider these as universal models applicable in general. Rather, we regard $f\big(\mathbf{y};\hat{\theta}(\mathbf{x}),k\big)$ and $f\big(\mathbf{y};\hat{\theta}(\mathbf{x}),\hat{k}(\mathbf{x})\big)$ for the two model classes, respectively, as the models applicable to all data generated by the same physical machinery, whose statistical behavior we hope was captured by these models.

If we take the requirement for optimality that an estimator must satisfy the necessary conditions for all data, then because of the uniqueness the estimator $\hat{\theta}(\cdot),\hat{k}(\cdot)$ could be said to satisfy also sufficient conditions for optimality. Consider a class of estimators 'fair', if it does not 'favor' any data on cost of reducing the performance on other data. This is difficult to pin down formally. All the traditional estimators for the real-valued parameters, such as the least squares, the method of the moments,

and the traditional ML estimators are 'fair'. The same is true of the estimators of the number of parameters resulting from optimization of the multitude of the existing criteria. Clearly, the estimator $\hat{\theta}(\cdot), \hat{k}(\cdot)$ is supremely 'fair', because for all data its performance is the best and cannot be beaten without changing the criterion.

5 Estimation capacity and index of separation

We next describe another way to arrive at the same two optimal estimators, which illustrates another property of theirs. Define the *estimation capacities* $\log \hat{C}_k$ and $\log \hat{C}$, all n, for the two model classes, respectively, where

$$\hat{C}_k = \max_{\bar{\theta}(\cdot)} \bar{C}_k = \max_{\bar{\theta}(\cdot)} \int f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right) \mathrm{d}\mathbf{x}, \tag{14}$$

$$\hat{C} = \max_{\bar{\theta}(\cdot), \bar{k}(\cdot)} \bar{C} = \max_{\bar{\theta}(\cdot), \bar{k}(\cdot)} \sum_k \int_{\bar{k}(\mathbf{x})=k} f\left(\mathbf{x}; \bar{\theta}(\mathbf{x}), k\right) \mathrm{d}\mathbf{x} / \bar{C}_k. \tag{15}$$

Notice that $\hat{C}$ determines $\hat{C}_k$ and the estimators so that the definition of the first capacity is in fact superfluous for the definition of the optimal estimator. However, it serves as the relevant measure of complexity of the estimation of the models in $\mathcal{M}_k$. We'll see later that the estimation of the number of the parameters is less complex as measured by the capacity $\log \hat{C}$, which does not exceed $\log K$.

A normalized form of the capacity, the *index of separation*, gives a particularly attractive property of the optimal estimators. It is defined as follows:

$$\mu\left(\hat{\theta}, k\right) = \frac{\hat{C}_k}{|\Omega^k|}$$

for a fixed k, and

$$\mu\left(\hat{k}\right) = \frac{\hat{C}}{K}$$

when k is being estimated, where $|\Omega^k|$ denotes the volume of the k-dimensional parameter space. We see that it is maximized by the same estimator that achieves the capacity. Hence, while the capacity measures the difficulty of the estimation job the index of separation measures how well the estimation can be done. This is clear for a finite number of models, like $\mu\left(\hat{k}\right)$, where the set of models is $\left\{\hat{f}(\cdot, k) : k = 1, 2, \ldots, K\right\}$. It is also a quite natural measure even when there are a continuum of models and it seems that there is nothing to separate. However, to each $\hat{\theta}$ there corresponds an equivalence class of data points, namely $\left\{\mathbf{y} : \hat{\theta}(\mathbf{y}) = \hat{\theta}\right\}$, which have the density $f\left(\mathbf{y}; \hat{\theta}; k\right) / \hat{g}\left(\hat{\theta}; k\right)$. This becomes clearer if we obtain it from a finite set of models by a simple limiting process: Partition the parameter space into hypercubes Δ of equal side length. Denote by $|\Delta|$ the volume, and there are $|\Omega^k|/|\Delta|$ hyper cubes including some portions of cubes on the boundary, which does not affect the limit. Let $\hat{\theta}_i$ be the center of the ith hyper cube in some enumeration, and we have

$$\mu\left(\hat{\theta}, k\right) = \lim_{|\Delta| \to 0} \frac{\sum_i \hat{g}\left(\hat{\theta}_i; k\right)}{|\Omega^k|/|\Delta|} = \frac{\hat{C}_k}{|\Omega^k|}.$$

The index of separation is between zero and unity, the unity corresponding to a perfect separation: the estimator never makes a mistake. In our model classes it cannot be reached except perhaps in the limit since all the models have a common support. It is zero when all the distinct models are equal and cannot be separated.

6 Complete MDL principle

The original MDL principle can be stated thus [1,4], (for an earlier related but more primitive version see Ref. [5])

- "*Find a model with which the observed data and the model can be encoded with the shortest code length*":

$$\min_{\theta, k} \left[\log \frac{1}{f(\mathbf{x}; \theta, k)} + L(\theta, k)\right].$$

We left the selection of the code length for the parameters $L(\theta, k)$ vague: anything decodable was acceptable. We know from coding theory that for decoding the so-called prefix requirement, the generalized Kraft inequality, is needed:

$$\sum_k \int 2^{-L(\theta, k)} \mathrm{d}\theta \leqslant 1.$$

This version, which we might call the *general* MDL principle, is very broad since almost anything can be encoded. It has, however, two shortcomings, the selection of the code length $L(\theta, k)$ and the fact that even when the generalized Kraft inequality holds with equality the resulting code,

$$F(\mathbf{x}) = f\left(\mathbf{x}; \hat{\theta}(\mathbf{x}), \hat{k}(\mathbf{x})\right) \times 2^{-L\left(\hat{\theta}(\mathbf{x}), \hat{k}(\mathbf{x})\right)},$$

is incomplete and hence non-optimal. A complete code is one defined by the negative logarithm of a distribution which integrates to unity. For discrete data it means that all the leaves of the code tree are codewords of some data and conversely.

Over the years two types of complete codes were constructed. The first is a Bayes type of mixture code:

$$f_w(\mathbf{x}) = \sum_k \int f(\mathbf{x}; \theta, k) w(\theta, k) \mathrm{d}\theta,$$

where $w(\theta, k)$ is a prior. The mixture could be maximized over k, which gives a criterion for order estimation. Although the result depends on the prior, the dependence disappears asymptotically, because the integrand peaks at $f\left(\mathbf{x}; \hat{\theta}(\mathbf{x}), k\right) w\left(\hat{\theta}(\mathbf{x}), k\right)$ and becomes independent of the prior as $n \to \infty$. There is a way to define an optimal prior even non-asymptotically although

the sense of optimality has little to do with estimation. The way is by Shannon's channel capacity and since it actually suggested the estimation capacity we describe it.

Start with the familiar minmax problem in universal coding for $\mathcal{M}_k$:

$$\min_q \max_\theta D\left(f_{\theta,k} \| q\right),$$

where q ranges over all density functions, and

$$D(f_{\theta,k}\|q) = \int f_{\theta,k}(\mathbf{x}) \log \frac{f_{\theta,k}(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x}$$

is the Kullback-Leibler (KL) distance between the density $f(\mathbf{x};\theta,k)$, written as $f_{\theta,k}$, and q. This turned out to be very difficult to solve with no neat solution, and the problem was changed to

$$\max_w \min_q \int w(\theta) D(f_{\theta,k}\|q)\mathrm{d}\theta. \tag{16}$$

This is more tractable since by McMillan theorem, also called the noise-free coding theorem, the minimizing q for all w is the mixture model

$$q_w(\mathbf{x}) = \int f(\mathbf{x};\theta,k) w(\theta)\mathrm{d}\theta.$$

To see this we have

$$\begin{aligned}\int w(\theta)\mathrm{d}\theta \int f(\mathbf{x};\theta,k) \ln \frac{f(\mathbf{x};\theta,k)}{q(\mathbf{x})}\mathrm{d}\mathbf{x} \\ = \int q_w(\mathbf{x}) \ln \frac{1}{q(\mathbf{x})}\mathrm{d}\mathbf{x} \\ + \int w(\theta)\mathrm{d}\theta \int f(\mathbf{x},\theta,k) \ln f(\mathbf{x},\theta,k)\mathrm{d}\mathbf{x},\end{aligned}$$

where in the left-hand side of the equation we switched the order of the two integrations. The second term in the right-hand side of the equation does not depend on q, and by the noise-free coding theorem the minimizing q in the first term is $q_w(\mathbf{x})$.

Although the maximizing prior w^* is still difficult to find the maxmin value, call it $\log C_{w^*}$, is formally Shannon's *channel capacity*. In communication there are no parameters and the role of w is to alter the probabilities of the input symbols to the communication channel in order to increase speed and reduce errors. Asymptotically $\log C_{w^*}$, which also can be shown to agree with the minmax value, is called the 'regret', and the maximizing prior converges to the famous Jeffrey's prior in Bayesian statistics. We see that even when $\log C_{w^*}$ asymptotically behaves like the estimation capacity the maxmin problem solved is appropriate for data compression rather than estimation, where there is no reason to 'regret' optimal behavior.

The second type of complete code, defined by the estimation capacity $\log \hat{C}_k$, is the Normalized Maximum Likelihood (NML) code, originally suggested by Shtarkov [2], for data compression. He defined it by asking for a universal code with code length closest to the *idealized* code length $\log\left[1/f\left(\mathbf{x};\hat{\theta}(\mathbf{x}),k\right)\right]$ as the solution $\hat{f}(\mathbf{x};k) = f\left(\mathbf{x};\hat{\theta}(\mathbf{x}),k\right)/\hat{C}_k$ to the minmax problem

$$\min_q \max_{\mathbf{x}} \log \frac{f\left(\mathbf{x};\hat{\theta}(\mathbf{x}),k\right)}{q(\mathbf{x})} = \log \hat{C}_k.$$

The minmax procedure actually fails to provide a formalized justification for the ML estimator or the code it defines, because the minmax argument applies even to codes defined by estimators $\bar{\theta}(\cdot) \neq \hat{\theta}(\cdot)$. Since their minmax value $\log \bar{C}_k$ is smaller than $\log \hat{C}_k$, why prefer the NML code?

The real justification is (7) for a fixed k, and in the general case (9), for we can interpret them in the spirit of the *complete* MDL principle: First, since both are determined by the model class their description or code length is common to all data and can be ignored. Further, both codes $f\left(\mathbf{x};\hat{\theta}(\mathbf{x}),k\right)/\hat{C}_k$ and $f\left(\mathbf{x};\hat{\theta}(\mathbf{x}),\hat{k}(\mathbf{x})\right)/\hat{C}$ are complete, and finally the necessary conditions ensure that the code length defined by the negative logarithms of these, respectively, is the shortest for all data just as required by the MDL principle.

To strengthen the optimality of the ML estimator $\hat{\theta}(\cdot)$ we can prove that it solves the minmax problem

$$\min_{\bar{\theta}(\cdot):\bar{C}_k>1} \max_\theta D(f_{\theta,k}\|\bar{f}_k) \tag{17}$$

for all θ and k. Similarly the MDL estimator $\hat{\theta}(\cdot), \hat{k}(\cdot)$ is the solution to

$$\min_{\bar{\theta}(\cdot),\bar{k}(\cdot)} \max_\theta D(f_{\theta,k}\|\bar{f}), \tag{18}$$

where we used the notations

$$\begin{aligned} f_{\theta,k} \ \text{ for } \ f_{\theta,k}(\mathbf{y}) &= f(\mathbf{y};\theta,k), \\ \bar{f}_k \ \text{ for } \ \bar{f}_k(\mathbf{y}) &= \frac{f\left(\mathbf{y};\bar{\theta}(\mathbf{y}),k\right)}{\bar{C}_k}, \\ \bar{f} \ \text{ for } \ \bar{f}(\mathbf{y};\mathcal{M}) &= \frac{f\left(\mathbf{y};\bar{\theta}(\mathbf{y}),\bar{k}(\mathbf{y})\right)/\bar{C}_{\bar{k}(\mathbf{y})}}{\bar{C}}. \end{aligned}$$

For computational purposes, for large amounts of data we give an asymptotic formula for the capacity [6]:

$$\ln \hat{C}_k = \frac{k}{2}\ln\frac{n}{2\pi} + \ln \int |J(\theta)|^{1/2}\mathrm{d}\theta + o(1), \tag{19}$$

where $J(\theta)$ is the Fisher information matrix:

$$J(\theta) = \lim n^{-1} E_\theta \left\{ \frac{\partial^2 \log[1/f(\mathbf{x};\theta,k)]}{\partial\theta_i \partial\theta_j} \right\}.$$

Extensive studies of the general MDL principle without separation of estimation from data compression

appear in Refs. [4] and [7].

7 Consistency

The ML estimator has been shown to be consistent; i.e., for all "true" parameters θ,

$$\hat{\theta}(\mathbf{x}) \to \theta$$

in probability. Many other estimators exist that are consistent in probability. However, the ML estimator is distinguished from all the other estimators in that the limit of the covariance matrix, the inverse of the Fisher information matrix $J(\theta)$,

$$E_\theta\big(\hat{\theta}(\mathbf{x})-\theta\big)\big(\hat{\theta}(\mathbf{x})-\theta\big)' \to J^{-1}(\theta),$$

is the smallest possible, which is a fine result [8], albeit asymptotic. Consistency in probability of the estimators for the number of parameters has also been shown for a multitude of criteria, although to minimize the covariance is not a meaningful one.

The question of consistency is still valid in our case except that instead of just covariance we compare estimators by the KL distance,

$$D\left(f_{\theta,k}\|\bar{f}\right) = \int f_{\theta,k}(\mathbf{x}) \log \frac{f_{\theta,k}(\mathbf{x})}{\bar{f}(\mathbf{x})} \mathrm{d}\mathbf{x},$$

between data generating models $f_{\theta,k}(\mathbf{x}) = f(\mathbf{x};\theta,k)$ and yardstick distributions defined by estimators $\bar{f}(\mathbf{x})$. Consistency then means the convergence for all parameters $\theta = \theta_1,\ldots,\theta_k$ and all k,

$$D\left(f_{\theta,k}\|\bar{f}\right)/n \to 0 \tag{20}$$

as $n \to \infty$, which applies both to the estimators of the real-valued parameters and of their number. We also regard it as a property of the estimators and the model classes without assuming any particular "true" parameter.

In Refs. [4] and [9] we proved a theorem that together with (20) gives the optimal convergence rate of estimators $\bar{\theta}(\mathbf{x}), \bar{k}(\mathbf{x})$ as follows:

Theorem 1 *For all $\bar{\theta}(\cdot), \bar{k}(\cdot)$ such that $\bar{\theta}(\mathbf{x})$ is consistent in probability for all θ, k,*

$$D(f_{\theta,k}\|\bar{f})/n \geqslant \frac{k-\epsilon}{2n}\ln n$$

for all ϵ as $n \to \infty$. The inequality holds for all k and θ except some in a set $A_{\theta,k,\epsilon}$ whose volume goes to zero as n grows. The bound $(k/2)\ln n$ is reached by $\hat{\theta}(\cdot), \hat{k}(\cdot)$.

This clearly plays a similar role to the estimators of the number of parameters as the Cramér-Rao inequality for the estimators of the real-valued parameters. To our knowledge no other optimal consistency rate for the estimation of the number of parameters exists.

References

1. Rissanen J. Modeling by shortest data description. Automatica, 1978, 14(5): 465–471
2. Shtarkov Yu M. Universal sequential coding of single messages. Translated from Problems of Information Transmission, 1987, 23(3): 3–17
3. Rissanen J. Optimal estimation. IEEE Information Theory Society Newsletter, 2009, 59(3): 1
4. Rissanen J. Information and Complexity in Statistical Modeling. New York, NY: Springer Verlag, 2007
5. Wallace C S, Boulton D M. An information measure for classification. Computer Journal, 1968, 11(2): 185–194
6. Rissanen J. Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 1996, 42(1): 40–47
7. Grünwald P D. The Minimum Description Length Principle. Cambridge, MA: MIT Press, 2007
8. Cramér H. Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press, 1946
9. Rissanen J. Stochastic complexity and modeling. Annals of Statistics, 1986, 14(3): 1080–1100

Jorma Rissanen was born on October 20, 1932 in Pielisjarvi, Finland. He received the Licentiate and the Doctor degrees of Technology in control theory and mathematics from the Technical University of Helsinki in 1960 and 1965, respectively. He is currently a fellow of Helsinki Institute for Information Technology, and Professor Emeritus in Technical University of Tampere, Finland. He has published over 100 papers and the Springer Verlag book *Information and Complexity in Statistical Modeling*.

He is the recipient of the IBM Corporate Award in 1991 for MDL principle and stochastic complexity; IEEE 1993 Richard W. Hamming Medal "For fundamental contributions to information theory, statistical inference, control theory, and the theory of complexity"; in 1998 an IEEE Information Theory Society Golden Jubilee Award for Technological Innovation for the invention of arithmetic coding; the 2007 Kolmogorov Medal from Computer Learning Research Center of University of London; and the 2009 Shannon Award from the IEEE Information Theory Society. He has also received two best paper awards from International Federation of Automatic Control in 1981 and the IEEE Information Theory Society in 1986, respectively.

He received Honorary Doctorate from the Technical University of Tampere, Finland, in 1992. He is a foreign member of Finland's Academy of Science and Letters. He is an IEEE Fellow.