Information and Complexity in Statistical Modeling. By Jorma Rissanen. Springer-Verlag, New York, 2007, viii+142 pp., ISBN 978-0-387-36610-4, $44.95.

Reviewed by **Ioannis Kontoyiannis**

In statistics – and in much of science – the central goal is to identify regularities in empirical observations. In classical mechanics, for example, physical laws allow the scientist to give succinct descriptions for complicated phenomena, like the motion of the planets, using a few simple equations. This description offers a "summary" or "explanation" of the existing empirical observations and can also be used to predict future observations. According to Occam's razor, the simplest such explanation is the one that should be favored. Similarly, the statistician fits a model to his or her data, in order to identify some of the statistical structural characteristics of the phenomenon that is being observed, and to make inferences about the underlying mechanism that produced the data. A (perhaps the) fundamental question in statistics is how this should be done, and it is a hard question to answer.

In statistical studies this question often takes the following form: we have some observable quantity of interest – the size of a cancerous tumor, for example – and there is a very large number of possible factors – like the genes in the patient's DNA – that may or may not influence this quantity. Typically, once we know which factors are relevant, it is fairly easy to come up with a statistical model that captures their influence. But, most of the time, the hard part is to figure out *which* of the factors to include. This is the fundamental problem of *model selection*, and it is one of the most active areas in current statistical research.

This book is an exposition of Rissanen's approach toward addressing this fundamental issue, based on the notion of *stochastic complexity* and the *Minimum Description Length* (MDL) *principle*.

Consider the following simple task: Given a binary string $x = (x_1, x_2, \ldots, x_n)$, what can be said about the bits x_i? Are they independent? What is the most likely value of the next bit x_{n+1}? How much information can we reliably extract about the structure of x, and what is that information? According to Rissanen (and to Kolmogorov and others), there is an essentially unique and fundamentally correct way to go about answering these questions. Fix a universal Turing machine U,[1] and define the *Kolmogorov complexity* $K(x)$ of x as the length, denoted length(p^*), of the shortest binary program p^* such that, when U runs p^*, it produces x. Formally,

$$K(x) = \min\{\text{length}(p) \text{ for programs } p \text{ such that } U(p) = x\}.$$

[1]Recall that a Turing machine is the simplest mathematical model of a digital computer, namely, a machine which can execute any well-defined algorithm, using finite but unlimited memory. A *universal* Turing machine is one which can emulate any other Turing machine. Except for the obvious limitation of finiteness, all our PC's are basically universal Turing machines. Mathematically, a Turing machine is described as a (special) map U from the set $\{0,1\}^*$ of finite-length binary strings to the set of all finite- or infinite-length binary strings.

Then x contains $K(x) = \text{length}(p^*)$ bits of information and is most economically described by p^*. This approach certainly conforms with the intuition that the simplest explanation (in this case the shortest description) should be favored, and, as was shown by Kolmogorov, the complexity $K(x)$ is essentially independent of the choice of the universal Turing machine U.

But, as was also shown by Kolmogorov, because of the vast richness of the space of all possible programs p, the complexity $K(\cdot)$ is *not a computable function*; that is, there is no algorithm that can effectively determine $K(x)$ for every string x. So, if anything concrete is to be said about the data, it is necessary to restrict the class of programs p that are allowed as possible descriptions. A natural and elegant way to do this is via a well-known result in information theory, the *Kraft correspondence*, which states that there is a precise and essentially one-to-one correspondence between probability distributions and binary descriptions.[2]

Kraft correspondence: Given a probability distribution Q on a discrete set A, there is a way to describe each element x of A without ambiguity, using approximately $-\log_2 Q(x)$ bits. In other words, there is a "nice" map C from A to the set $\{0,1\}^*$ of finite-length binary strings, such that $\text{length}(C(x)) \approx -\log_2 Q(x)$ bits, for every x in A.[3] Conversely, any such map C defines a probability distribution Q on A by reversing the above relationship: $Q(x) \approx 2^{-\text{length}(C(x))}$.

In this context, therefore, finding a good "description" or "explanation" for the data x, namely, finding a short binary program p that produces x, is equivalent to fitting a good statistical model to x, where a "good model" is one that leads to the shortest description for x. This thinking leads to a natural way to restrict the class of programs p we allow: we can fix a parametric model $\mathcal{M} = \{Q_\theta \,:\, \theta \in \Theta\}$, or more generally a collection $\{\mathcal{M}_\gamma\}$ of such models, and consider only programs that correspond to probability distributions from this collection.[4]

For the sake of concreteness, suppose our data string $x = (x_1, x_2, \ldots, x_n)$ consists of observations x_i taking values in a finite set A, so that $x \in A^n$, and let's choose and fix a single parametric model $\mathcal{M} = \{Q_\theta \,:\, \theta \in \Theta\}$ of distributions Q_θ over A^n. Further, for simplicity, suppose that Θ is a finite set, so that $\mathcal{M}$ contains only finitely many distributions. [The extension to the more usual case of real- or vector-valued data, and to models parametrized by open subsets of $\mathbb{R}^d$, requires little more than more involved notation.]

[2]The Kraft correspondence described here is really a simple consequence of an information-theoretic result usually referred to as the "Kraft inequality." The term "Kraft correspondence" is not standard, but it is more suggestive for the purposes of this discussion.

[3]For our purposes, it suffices to think of "nice" maps simply as those that are invertible. For the more curious reader, what is really required is that C has the prefix-free property: the string $C(x)$ is not the beginning of the string $C(y)$ for any $x \neq y$.

[4]Recall that a parametric model is simply a collection of distributions indexed by some parameter θ, usually taking values in some subset Θ of $\mathbb{R}^d$. For example, in the case of a binary string x, we may consider the model $\mathcal{M} = \{Q_\theta \,:\, \theta \in (0,1)\}$, where, according to distribution Q_θ, the successive bits x_i are independent, and $\Pr\{x_i = 1\} = \theta$.

Now we are in a position to define the central idea in Rissanen's approach:

The *stochastic complexity of the data x relative to the model $\mathcal{M}$* is the length of the shortest binary description of x that can be obtained from descriptions corresponding to probability distributions in $\mathcal{M}$. Accordingly, the *Minimum Description Length* (MDL) *principle* mandates that, among all distributions in $\mathcal{M}$, the one that most accurately captures the nature of the data x is the distribution which achieves this shortest description.

The above definitions, although intuitively satisfying, are too vague to translate into equations. In particular, they do not specify how we should compute the description length for x that corresponds to any given distribution Q_θ. For example, the obvious answer of $-\log_2 Q_\theta(x)$ bits suggested by the Kraft correspondence is erroneous! The reason is the following subtle but crucial observation: in order to use Q_θ to describe x in a truly invertible manner, we need also to describe the distribution Q_θ itself!

As it turns out, this "gap" in Rissanen's definition is intentional. The problem of optimally choosing a distribution from $\mathcal{M}$ to describe our data in as few bits as possible is exactly the topic of an area known as *universal data compression* in the information theory literature. This is why, after the Introduction in Chapter 1, the following two chapters in the book deal with the basic information-theoretic ideas of data compression and universal data compression.

There are many ways to fill the above gap – many universal data compression algorithms – several of which are considered in Chapter 5, where the notion of stochastic complexity is introduced and its properties are discussed. Here we will briefly describe two special cases: two-part codes, and the normalized maximum likelihood (NML) code.

Two-part coding: We can try to describe x in two steps: first describe a distribution Q_θ, and then describe x using Q_θ. To do this we need first to decide on a code for the parameters $\theta \in \Theta$. Again, this can be done in many ways, and Rissanen proposes a canonical way to design such a code; but let's suppose for now that we have such a code C that maps Θ to $\{0,1\}^*$. Then the stochastic complexity of the data is

$$SC(x) = \min\Big\{\text{length}(C(\theta)) + [-\log_2 Q_\theta(x)] \ : \ \theta \in \Theta\Big\},$$

and the MDL principle states that the distribution Q_θ that best captures the statistical properties of x is the one corresponding to the parameter θ^* that achieves the above minimum.[5]

[5]It is worth noting that, in view of the Kraft correspondence, choosing a code C for Θ is somewhat reminiscent of the Bayesian problem of choosing a prior distribution on Θ, although the ways in which the two are used are quite different.

Two remarks are in order here. First, note that so far *nothing* has been assumed about the data; there is no requirement that x be a sample from a "true" distribution, let alone that this distribution belongs to $\mathcal{M}$. Furthermore, the optimality properties of the MDL principle described in Chapter 5 hold in great generality, with minimal, if any, statistical assumptions on x (and for the case of the NML code this is true in an even stronger sense than for the two-part code). Second, we observe that, up to this point, our review could have just as well been of Rissanen's first book [3], published in 1989, which described the state of affairs at that time. The main point of departure from the "old MDL theory" is the idea that *any* universal code can be used for statistical inference, and this departure was prompted primarily by the point of view advanced in a 1998 paper by Barron, Rissanen, and Yu [1] and by the development of the NML code.

The normalized maximum likelihood (NML) *code*: As noted above, what we ideally would like to do is to use $\min_\theta[-\log_2 Q_\theta(x)]$ bits to describe x, but this does not correspond to a true codelength for the data. Note, by the way, that the $\hat{\theta}(x)$ that achieves this minimum is nothing but the classical maximum likelihood estimate (MLE) for the parameter θ. Nevertheless, we can use the "ideal" codelength of "$\ell^*(x) = -\log_2 Q_{\hat{\theta}(x)}(x)$ bits" as a yardstick against which we measure our performance. From this we can construct a minimax problem which identifies the best code for the worst possible distribution of the data, where now *every* possible code C and *every* possible distribution (not just those in $\mathcal{M}$) are allowed in the game, and the objective function that's being optimized is the expected value of the difference between the codelength of the code C and the ideal $\ell^*(x)$. The solution to this problem is the NML code, which turns out to be the code given by the Kraft correspondence for the distribution

$$Q_{NML}(x) = \frac{Q_{\hat{\theta}(x)}(x)}{\sum_{\text{all } y} Q_{\hat{\theta}(y)}(y)}.$$

What is this Q_{NML} distribution like? Several examples in the book illustrate its form, use, and properties. In a way, this distribution looks like a strange "perturbation" of the classical MLE, and it is natural to wonder why we should care about it. One answer is that it is well motivated by the above discussion; another is that, as the applications (Chapter 9) in the book illustrate, it performs well in various practical problems; also, in Chapter 5 it is shown to enjoy important optimality properties. But perhaps the most important feature of Q_{NML}, and, more generally, the models produced by any version of the MDL principle, is that they are *by design* good at avoiding "overfitting," which is one of the biggest problems in applied statistics. The reason is mathematically far from obvious, but intuitively fairly simple: Since, by construction, the MDL distributions lead to honest coding algorithms, they implicitly describe the distribution which is chosen as well as the data. It is, therefore, natural to expect that more complex distributions (and

more complex models as a whole) will be harder to describe and hence naturally "penalized."

It is interesting that this year also saw the publication of another beautiful book on MDL, Grünwald's text [2], which has a very different, much more leisurely style. In Chapter 17 of [2] there are many illustrations of the MDL principle for a range of realistic statistical inference problems, and the performance of MDL-based methods is carefully compared to that of many of the standard and commonly used methods and techniques.

Chapter 4 of the book under review gives an introduction to Kolmogorov complexity and related notions. In particular – and this is another major point of departure from the "old MDL" – Rissanen describes Kolmogorov's structure function and the Kolmogorov minimal sufficient statistic. Given a string x, suppose we focus on a special sub-class of programs p that produce x, namely, programs that first describe a set B of strings such that $x \in B$, and then identify x as an element of B using $\approx \log_2 |B|$ bits. Then the overall description of x takes $\approx K(B)+\log_2 |B|$ bits. For any $\alpha > 0$, the *structure function* $h_x(\alpha)$ is defined as

$$h_x(\alpha) = \min \Big\{ \log_2 |B| \ : \ \text{sets } B \text{ s.t. } x \in B \text{ and } K(B) \leq \alpha \Big\}.$$

The idea behind this definition is that the description of B contains all the "structure" in x, while the remaining $\log_2 |B|$ bits describe the "noise" in x. As it happens, $h_x(\alpha)$ is nonincreasing in α, and there is an optimal point $\bar{\alpha}$ after which it follows the line $K(x) - \alpha$. The *minimal sufficient statistic decomposition* of x is then

$$K(x) = \bar{\alpha} + h_x(\bar{\alpha}) = \min_{B \,:\, x \in B} \Big[K(B) + h_x(K(B)) \Big].$$

The interpretation of this relationship is that x contains $\bar{\alpha}$ bits of *learnable information* and $h_x(\bar{\alpha}) = K(x) - \bar{\alpha}$ bits of *noise*.

The rest of the material in the book is too complex – both mathematically and conceptually – to describe here in detail, so we give only an outline. In the same way that stochastic complexity is a statistical analog for the non-computable Kolmogorov complexity, in Chapter 6 we get corresponding "stochastic" or "statistical" analogs for the structure function and the minimal sufficient statistic. Chapter 7 is on "optimally distinguishable models." In a way, the question considered here is similar to the problem of choosing a code C for the parameters θ in the earlier two-part code example – is there an "optimal" choice? This issue is crucial when the parameter set Θ is an open subset of $\mathbb{R}^d$, since then we have to pick out a countable subset Θ' of Θ and assign finite-length descriptions only to the parameters θ in that subset. How should this set Θ' of "representatives" be chosen? The results of Chapter 6 are used to answer the question. Chapter 8 is a broad discussion of the MDL principle in full generality, and Chapter 9 illustrates its utility in a variety of statistical tasks.

Summary: Like Rissanen's first book [3], this book is part monograph, part manifesto, part advanced textbook. It is a short book about a long, complex, fascinating story. The style is both engaging and provocative. Whether or not the reader agrees with Rissanen's opinions about statistics – and they certainly are strong opinions – this is an interesting and evocative text.

Finally, we should point out the obvious: The MDL principle and the statistical ideas surrounding it have a fairly long history, and they were not all introduced by Rissanen himself. Some of the people that have played a role in this development are Christopher Wallace, David Boulton, A. Philip Dawid, Hirotsugu Akaike, Andrew Barron, Vijay Balasubramanian, Bin Yu, and Peter Grünwald, as well as others mentioned in the book.

Acknowledgments. This is an expanded version of a review I wrote for the American Mathematical Society's *Mathematical Reviews*, at the recommendation of Radu Zaharopol. I wish to thank Kevin Clancey, executive editor of Mathematical Reviews, for agreeing to allow this expanded version to be published separately.

References

1. A. R. Barron, J. Rissanen, and B. Yu, The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** (1998) 2743–2760.

2. P. Grünwald, *The Minimum Description Length Principle*, MIT Press, Cambridge, MA, 2007.

3. J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

Athens University of Economics and Business, Athens, Greece
yiannis@aueb.gr