

## Information Theory and Statistics, Part II

## Part II: The principle of Minimum Description Length

### Caveats about model selection

- In statistical applications, we are frequently forced to compare competing models or descriptions of a data set
- Through a cycle of hypothesis tests, computational and graphical diagnostics, we construct a view of the data and ultimately settle on (one or more) fits
- To help guide this process, we often resort to model selection criteria to help us order or prioritize the descriptions

### Model selection

- Selection criteria became more important as computational methods (and the available computing power) made it possible to fit many different models
- As researchers started examining the theoretical properties of these criteria (or rather the procedure of using these criteria to select a single model) it became clear that there are different ways to judge their performance
- For example, are we interested in “consistency” (in the sense that we choose the “true” model with high probability) or “prediction accuracy” (the selected model has small mean squared error performance)

## Model selection

- We believe the true value of a selection criterion is the insights it provides into real problems
- As Mallows (1973) put it “The greatest value of the device [model selection] is that it helps the statistician to examine some aspects of the structure of his data and helps him to recognize the ambiguities in front of him.”
- In general, we should apply any selection criterion with care, examining the structure of several good-fitting models rather than restricting our attention to a single “best”

## The Principle of Minimum Description Length

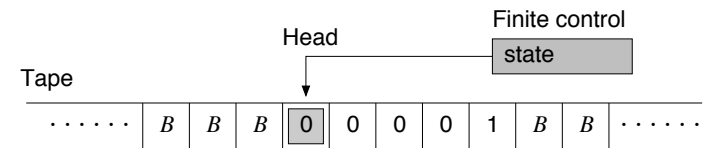
*Choose the model that gives the shortest description of the data*

## MDL

- MDL has been developed by Jorma Rissanen in a series of papers beginning in the late 1970s
- MDL has its philosophical roots in notions of complexity put forward by Komogorov, Solomonoff and Chaitin in the 1960s
- These researchers proposed that the complexity of an object (in their case, a binary string) be measured by the length of the shortest program needed to “print” this string on a universal computer

## Kolmogorov complexity

- Taking the Universal Turing Machine as their starting point, the idea was to see how many bits were required to exhibit a given string
- If the string had a lot of regularity, it could be expressed by a simple program; while a “random” string might require incorporating the string itself into the program



## Kolmogorov complexity

- One can imagine loading a Turing machine with a program or “model” for the data followed by some compressed representation or encoding of the data which the program would use to create the printout, to exhibit the string
- It is possible to push these ideas quite far; what emerges are notions of universal probability  $P_U(x)$  (the probability of a string  $x$  is the chance that random inputs fed to the Turing machine will exhibit  $x$ ) and the Shannon-style relationship

$$K(x) = \log \frac{1}{P_U(x)}$$

## But alas...

- Unfortunately, very little about Kolmogorov complexity is actually computable
- Rissanen made these ideas practical, cutting right to information-theoretic principles and actual coding schemes

## Making things a little more precise...

- Suppose we are given a data string  $x = (x_1, \dots, x_n)$  and a model  $P(x)$ ; we can use this distribution to encode the data string in essentially

$$\log \frac{1}{P(x)}$$

bits using one of the codes described this morning; we know the best we can do is code with the “true” or “data generating” distribution

Note that this is simply minus the log-likelihood

## Multiple models

- Now suppose we have several different possible models  $P_1, \dots, P_K$  for our data string
- To remain consistent with our coding story, we cannot simply choose the model with the smallest

$$\log \frac{1}{P_1(x)}, \dots, \log \frac{1}{P_K(x)}$$

- If we were to encode data based on the best model in this collection, the person on the receiving end of our bit stream would be at a loss as to how to decode the data

## Multiple models

- To decode the data, the receiver needs to know what codebook to use; that means we have to somehow transmit our model choice together with the encoded data
- Now, our “best” models are small in terms of the combined code lengths

$$\text{“bits to specify model } k\text{”} + \log \frac{1}{P_k(x)}$$

- If we only have  $K$  models, we can transmit the first choice in  $\log K$  bits

## Multiple models

- In realistic applications, however, we are rarely choosing between a handful of models; instead, we usually have several model classes competing for our attention

$$\mathcal{M} = \{P_\theta^n(x), \theta \in R^K\}$$

- In general, given a  $K$ -dimensional parametric family, what price do we have to pay in bits to communicate our choice of  $\theta$  ?

## Multiple models

- Suppose further that we have nested model classes,  $\mathcal{M}_0 \subset \mathcal{M}_1$
- For example,  $\mathcal{M}_1$  might be a normal linear model with one extra predictor than is included in  $\mathcal{M}_0$
- Even if we were to finesse the communication issue, should we rely on code length to tell us which model to prefer?

## Rissanen’s lower bound

- When our receiver only knows that we are encoding the data using some member of the model class, we have to spend roughly

$$\frac{K}{2} \log n$$

- bits to encode our choice of parameter
- While we have already seen coding costs associated with using a fixed distribution, this result tells us how much more we pay for making a choice

## Multiple models

- The shift from one to many models/codes takes us from the coding examples this morning (Shannon-Fano, Huffman, etc.) to something known as universal coding
- While MDL is a broad and far-reaching principle for estimation, it hinges on being able to encode your model choice in a “fair” way; those descriptions that achieve Rissanen’s lower bound are, in this sense, valid

## Two-stage coding

- If from our model class we select the parameter  $\theta$ , communicate that choice and then use  $\theta$  to encode the data, we have an overall code length of

$$\log \frac{1}{P_{\hat{\theta}}^n(x)} + \frac{K}{2} \log n$$

- We can then compare this for all choices of  $\theta$ ; obviously, the choice providing the shortest code length is the MLE  $\hat{\theta}$  (MDL implies Maximum Likelihood in a given model class)

## Two-stage coding

- The simplest universal code based on the model class  $\mathcal{M}$  has a code length of

$$\log \frac{1}{P_{\hat{\theta}}^n(x)} + \frac{K}{2} \log n$$

- This expression is known in different quarters as BIC, the Bayesian information criterion of Schwarz (1978)

## Ok, but...

- That seems like a lot of machinery to derive BIC; the strength of MDL comes from the fact that we can be somewhat creative about our choice of codes (I suppose depending on your disposition, this is either a blessing or a curse)
- There are other coding schemes that are also valid, but are very different in character

## Predictive MDL

- We can write any given probability distribution for a string  $x^n = (x_1, \dots, x_n)$  in its so-called predictive form

$$P(x^n) = \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1})$$

- In many parametric models, the separate conditionals all depend on the same parameter  $\theta$ ; predictive MDL gets its name by using observations  $(x_1, \dots, x_i)$  to form the MLE  $\hat{\theta}_i$

- Then the following joint distribution is free of unknown parameters

$$P(x^n) = \prod_{i=1}^n P_{\hat{\theta}_{i-1}}(x_i | x_1, \dots, x_{i-1})$$

## Predictive MDL

- The code length corresponding to this procedure is then

$$-\sum_{i=1}^n \log P_{\hat{\theta}_{i-1}}(x_i | x_1, \dots, x_{i-1})$$

- This approach to model selection is identical to the so-called prequential method of Dawid

## Mixture MDL

- The word “mixture” here, says it all; we build a code for our string using a mixture of all the models in the class; formally, we can write

$$m(x^n) = \int P_{\theta}^n(x^n) w(\theta) d\theta$$

- We then build a code using the mixture distribution, giving us a code length of

$$\log \frac{1}{m(x^n)}$$

## Mixture MDL

- Rissanen based his Stochastic Information Complexity on a simple second order expansion of this mixture form

$$SIC(x^n) = \log \frac{1}{P_{\hat{\theta}}(x^n)} + \frac{1}{2} \log \det \hat{\Sigma}_n$$

- where  $\hat{\Sigma}_n$  is the Hessian of minus the log-likelihood evaluated at the MLE

## Normalized maximized likelihood

- Finally, a relatively new form of MDL is based on a code using the following distribution

$$\tilde{P}(x^n) = \frac{P_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} P_{\hat{\theta}(y^n)}(y^n)}$$

- Originally studied by Shtarkov (1987), this code also “fairly” represents the model class; although there are implementation difficulties when we are working with models for continuous data
- This new code so impressed Rissanen that he redefined Stochastic Complexity in terms of NML

## An umbrella

- Under the MDL framework, we have seen big statistical ideas, from both Bayesians and frequentists emerge, emerge through simple considerations based on coding
- One attractive aspect of MDL is that it puts a new face on old procedures, allowing them to be compared under a single framework
- In addition, it provides rough and ready guidance for developing usable model selection techniques in non-standard settings

## This is just the beginning

- Time does not really permit much more in terms of examples, but you now have ample background to dig into the literature
- Review papers by Hansen and Yu (2001), and Barron, Rissanen and Yu (1998) are excellent next steps; so too is the text by Rissanen in 1989

## Information theory and statistics

- Today we have introduced you to a few important concepts from information theory; entropy, divergence, redundancy, I-projection
- We have illustrated how they have influenced both the theory and practice of statistics; or perhaps more loosely, what these ideas say about statistical methodology
- Our goal was merely to promote the idea that interdisciplinary work can be inspiring on many levels (a case that hopefully doesn't need that much promotion among statisticians)