Modeling by the MDL Principle

J. Rissanen

Helsinki Institute for Information Technology,

Technical Universities of Tampere and Helsinki, Finland,

and

University of London, Royal Holloway, UK

2/20/2005

Abstract

Data generated by a physical process incorporate regular features reflecting the generating machinery and noise. The objective of modeling is to learn the regularities and, for some applications, construct a smooth curve as the 'law' representing the restrictions forced by the data generating machinery leaving the rest of the data as 'noise'. The *MDL* (Minimum Description Length) principle seeks to minimize the code length of the data, given a class of models, which provides a formal way to measure three fundamental properties of data, the complexity, the amount of the regular features and the amount of noise. Quite naturally, the noise then gets defined, not as the high frequency part in the data as commonly done, but as the random incompressible part in light of the models considered.

In this talk I outline the recent status of the *MDL* theory with some applications.