SUPAADA Advanced Data Analysis

Aims and Objectives.

At the end of the course students should be able to:

1) Describe qualitatively the theoretical foundations of the nature of probability, in the context of both a *frequentist* framework (explaining clearly the meaning of this term) and a Bayesian framework (i.e. as a logical system for *plausible reasoning*).

2) Define what is meant by a probability density function (pdf), and cumulative distribution function (cdf), as well as various descriptive statistics (e.g. mean, median, mode, moments, variance, covariance) used to characterize pdfs and cdfs.

3) Explain the importance of the Central Limit Theorem for the statistical properties of estimators, and the consequent importance of the Gaussian, or normal, pdf.

4) Apply the principle of least squares to formulate and solve simple line and curve fitting problems – using a matrix formulation where appropriate, and adapting the formulation to various cases and approximations (e.g. weighted least squares, correlated errors, non-linear problems).

5) Describe and apply the principle of maximum likelihood for frequentist parameter estimation, explaining the connection between maximum likelihood and least squares estimators for a Gaussian pdf.

6) Describe and apply the basic concepts of frequentist hypothesis testing, using the chi-squared goodness-of-fit test as an archetypal example.

7) Define in a Bayesian context the likelihood, prior and posterior distributions and their role in Bayesian inference and hypothesis testing, contrasting Bayesian and frequentist treatments of hypothesis testing.

8) Define the Fisher information matrix, its relation to the covariance matrix and its relevance to experimental design under the assumption of a Gaussian posterior.

9) Define the *evidence* and explain its role in Bayesian model selection, describing several numerical approximations to the evidence and their applicability.

10) Discuss general principles for assigning prior probabilities, including insufficient reason and maximum entropy.

11) Describe and apply data compression techniques for analysis of very large data sets, including singular value decomposition and principal component analysis.

12) Describe and apply efficient numerical techniques for generating random numbers and performing Monte Carlo simulations, including Markov Chain Monte Carlo methods for parameter estimation and model selection in problems of high dimensionality.

Learning Outcomes:

1) To acquire a working knowledge of advanced data analysis methods – i.e. to a level sufficient to permit their successful application to real data analysis problems, as might be encountered in students' own research projects.
2) To gain familiarity with the key differences between a frequentist and Bayesian approach to data analysis: the assumptions upon which each approach is founded and the circumstances in which each is applicable.
3) To develop awareness of the current literature on advanced data analysis for the physical sciences, and the software available to support its application to real problems.

Assessment Method:

1) **Multiple choice questions and problems (weighted 50%)**.
a) A series of 15 multiple choice questions, interspersed throughout the lectures.
b) A further series of 10 numerical problems, to be posted on my.SUPA after the course is completed.

2) Mock data challenge (weighted 50%). A mock data set, made available via my.SUPA at the end of the course, for which students have to write a short computer code (in a language / platform of their choice) to infer the embedded parameters (with errors / confidence regions), using the principles and practical ideas introduced during the course. Students will submit by email a copy of their source code and their results.

Martin Hendry
Jan 2011