

Title: High-throughput Data Analysis

Weight: Half Unit (16 lectures)

Method of Assessment: 1½-hour examination

Prerequisites: Part A (second-year) courses in Statistics and Probability.

Lecturer Dr Peter Clifford – 16 HT

Learning outcomes: Understanding the problems of working with high-throughput, high-frequency data-streams in applications such as finance, engineering and environmental monitoring. Understanding the limitations of established statistical techniques in dealing with such data. Development of working knowledge and experience with conventional and simulation-based methods for online statistical inference. Understanding how to construct informative ‘data sketches’ from transiently observed data-streams. Experience in the use of such sketches.

Synopsis:

Illustrations of the problems of high-throughput, high-frequency data analysis in finance, internet monitoring and object-tracking. Defining the objective, namely the construction of efficient statistical algorithms that operate with limited storage and process with one pass over the data.

Online data analysis. Introduction to dynamic statistical models. Gaussian and non-Gaussian models; Kalman filter, extended Kalman filter. Estimating volatility and cross-correlation in high-frequency financial time series. Problems of asynchronicity and scale.

Simulation based online filtering: concept of particle filters, design, optimisation, performance characteristics, connections with the Kalman filter. Applications to object tracking and parameter estimation in mathematical models of finance.

Data-sketching. Probabilistic counting. Estimation of frequency moments by stable-law projection. Comparison of data streams via data sketches. Estimation of entropy and cross-entropy. Application to MCMC convergence diagnostics and monitoring of internet traffic.

Reading

A. Doucet, N. De Freitas and N.J. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

A. Doucet and A.M. Johansen, *A tutorial on particle filtering and smoothing: fifteen years later*, Technical report, Department of Statistics, University of British Columbia, http://www.cs.ubc.ca/~arnaud/doucet_johansen_tutorialPF.pdf, 2008.

J.P. Nolan, *Information on Stable Distributions*, <http://academic2.american.edu/~jpnolan/stable/stable.html>

Selected readings of papers on data sketching: as indicated in the lectures.

Background reading: financial applications

P. Lequeux (Editor), *Financial Markets Tick by Tick*, Wiley, 1999.

Dacorogna, M. M., Gençay, R., Müller, U.A., Olsen, R. B. and Pictet, O.V., *An Introduction to High-Frequency Finance*, Academic Press, 2001.