COMPUTATIONAL STATISTICS

STATISTIQUE INFORMATIQUE

by Guy Mélard, gmelard@ulb.ac.be

http://homepages.ulb.ac.be/~gmelard/statinfo.html

October 2008

ECARES/IRS/Dept Math/Dept Sc Eco

0. Introduction

Objectives

- In what follows, "econometrics" can be used nearly always where "statistics" is expressed
- Computational statistics:

  not using computers for doing statistics (using a package)

  but implementing a new statistical method (for research purposes, development of a new module in a package, etc.)

Evolution

- faster and faster (computers)
- smaller and smaller (programming effort)
- better and better (tools)

  1. 3rd generation programming languages (Fortran(*), Pascal(*), C(*), C++(*) ...)
  2. general purpose software: Excel
  3. 4th generation programming languages:

     statistical/econometric packages: SAS, SPSS, RATS, ...

     mathematical/statistical environments: SPlus, R, MATLAB, Gauss

Constant features

- Never fast enough (computers, not tools)
- Never as ruled by theory (algorithms should not be an implementation of formulas)
- Always consider exceptions (causes: 1. The user, 2. The data)
- Old programming languages still alive and well (Fortran '77, '90, '95)
- Algorithmic structures (the most important!)
- Data structures have always been simple (for how long?)

Tools

1. 3rd generation programming languages (Fortran(*), Pascal(*), C(*), ...)
   + libraries (Numerical Recipes, Press et al. (*); NAG(*), IMSL(*))
2. General purpose software: Excel(*) (at least supplemented with macros: Visual Basic for Applications)
3. 4th generation programming languages + modules/tool kits/programs:
   statistical/econometric packages: SAS(*), SPSS (*) (only recently), RATS, ...
   mathematical/statistical environments:
      SPlus(*), R
      MATLAB(*) (matrix calculation, based on Linpack and Eispack renowned libraries)
      Gauss (optimisation)

(*) available at ULB

The most important things:

- to be familiar with the language (even if only a subset is used)
- to write as nicely as possible, to comment what is not obvious (program = poem)
- to prepare test runs (data, problem and solution) exercising most cases
- critical mind (for all statisticians but more in computational statistics)
- cross-check using other approaches (Excel, MATLAB)

Contents of the course

Statistique informatique/Computational Statistics (Prof. Guy Mélard)
(2ème licence sciences mathématiques, master en statistique, master en sciences économiques = 6 ECTS (théorie : 2, travaux personnels : 1)
2ème licence en sciences économiques = 3 ECTS)

0. Introduction (1 h)

The Statistician and computers, Use of computers in statistics
Equipment, Software
Access to a computer, Criticism, Cost, Common points
Some references

1. 3rd generation languages for statistics (7 h)

1.1 Basic algorithmic language
1.2 Fortran primer
1.3 Introduction to Fortran 90
1.4 Advanced study of Fortran 90 and elements of Fortran 95
1.5 Instructions of Fortran IV and Fortran 77 to be deciphered
1.6 The use of scientific libraries
1.7 Preparation of test data sets

2. 4th generation languages for statistics (4 h)

2.1 General concepts
2.2 Example: MATLAB

3. Main algorithms in statistics (12 h)

3.1 Computing variances and covariances
3.2 Probabilities and quantiles
3.3 Generation of pseudo-random numbers and variables
3.4 Monte Carlo method
3.5 Multiple linear regression
3.6 Introduction to non-linear regression
3.7 Resampling and the method of bootstrap
3.8 Case studies

Exam = personal project using partly Fortran and partly Matlab or R

Introduction

LE STATISTICIEN ET L'INFORMATIQUE (based on an article of Pierre DAGNELIE, Biométrie-Praximétrie, vol. 15, 1975)

Use of computers in statistics

- administrative statistics (surveys)
- statistics in research
- teaching of statistics
- research in statistics (simulation, ...)

Equipment

- scientific calculators with statistical functions (factorials, mean, variance, linear regression)
- computers with statistical functions: have disappeared !
- now general purpose computers with software: micro-computers (PCs), mini-computers, mainframes, super-computers

Software

- statistical software packages (since +/– 1975)

  Examples: SPSS, SAS, Minitab, Genstat, Glim, TSP, Troll

  Remark. Most of them were developed on mainframes and ported to (+/- completely) on minis then PCs
- personal software: specific to an application

  Big danger of personal programs: often not efficient and buggy, lack of portability (language, equipment)

  Better solution : use libraries and concentrate efforts on the main program and data sets

Access to a computer

- remote batch processing (“traitement par lots à distance”)
    - job preparation in an appropriate language
    - introduction of the job using an editor
    - submission of the job
    - output (error list): on screen or listing
    - interpretation
- interactive treatment

  dialog with the statistical package

Criticism

- remote batch processing is slower but ... leaves time for reflection
- interactive treatment allows for trial and error (various methods, plots) but ... don’t loose the objective

Cost

Both modes can cost a lot:

- remote batch processing: each time a part of the treatment is done again (reading data, doing transformations, ...)
- interactive treatment: does cost more except on a PC

Common points

- data entry should be done once and for all with careful checking
- keep order among the results

Some references

Books

Chambers, J. M., "Computational methods for data analysis", Wiley, New York, 1977.

(algorithms, very good but becoming old)

Griffiths, P. and Hill, I. D. (eds), "Applied Statistical Algorithms", Ellis Horwood, Chichester, 1985.

(Fortran algorithms from Applied Statistics, not interesting)

Gentle, J. E., "Numerical Linear Algebra for Applications in Statistics", Springer-Verlag, 1998 *(very good)*

Gentle, J. E., "Elements of Computational Statistics", Springer-Verlag, 2002 *(more up to date than Kennedy and Gentle)*

Kennedy, W. J. Jr, and Gentle, J. E., "Statistical Computing", Marcel Dekker, New York, 1980. *(algorithms, OK but not without errors)*

Martinez W. L., Martinez, A. R, "Computational Statistics Handbook with MATLAB", CRC Press, 2001. (*bad reviews on Amazon.com*)

Metcalf, Michael and Reid, John, "Fortran 90/95 Explained", Oxford University Press, 1996. (*a good book on Fortran 90*)

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T., "Numerical recipes - The art of scientific computing", Cambridge University Press, Cambridge, 1986.

(good, Fortran, C or Pascal programs on sale)

Thisted, Ronald A., "Elements of statistical Computing", Chapman & Hall, New York, 1988. *(algorithms, excellent)*

Periodicals

J. Roy. Statist. Soc. Ser. C, Applied Statistics (*algorithms*)
Computational Statistics and Simulation
Computational Statistics and Data Analysis
Computational Statistics Quarterly

Meetings :

COMPSTAT (International Association for Statistical Computing)
Computational statistics section (American Statistical Association)
Interface

Table of contents of Thisted’s book: "Elements of statistical Computing"

1/ Introduction to statistical computing

- Early, classical and modern concerns
- Computation in different areas of statistics
- Different kinds of computation in statistics
- Statistics in different areas of computer science
- Some notes on the history of statistical computing

2/ Basic numerical methods

- Floating point arithmetic, rounding error and error analysis
- Algorithms for moment computations
- Floating-point standards

3/ Numerical linear algebra

- Multiple linear regression (Householder, Gram-Schmidt, Givens)
- Solving linear systems
- The Cholesky factorization
- The SWEEP operator
- Colinearity and conditioning
- Regression diagnostics, regression updating
- Principal components and eigenproblems, solving eigenproblems
- Generalizations of least-squares regression (GLM, WLS, GLS, GLIM)
- Additional topics and further reading (regression, robust regression, subset regression)

4/ Nonlinear statistical methods

- Maximum likelihood estimation
- Solving scalar and the vector case
- Obtaining the Hessian matrix
- Optimization methods (grid search; linear search, step direction: Newton, steepest descent, Marquardt, ...; constrained optimization)
- Computer-intensive methods (projection-selection regression; projection-pursuit regression)
- Missing data: the EM algorithm
- Time series analysis

5/ Numerical integration and approximation

- Newton-Cotes methods; Improper integrals; Gaussian quadrature; Interpolating splines; Monte Carlo integration; Multiple integrals; Bayesian computations
- General approximation methods
  - series approximation; continued fractions;
  - polynomial approximation; rational approximation
- Tail-areas and inverse cdf's for normal, , *F*, *t* distributions

6/ Smoothing and density estimation

- Histograms and related density estimators
- Linear smoothers; spline smoothing; nonlinear smoothers
- Choosing the smoothing parameter
- Applications and extensions

Table of contents of Kennedy and Gentle’s book: "Statistical Computing"

1/ Introduction

2/ Computer organization

3/ Error in floating-point calculation

4/ Programming and statistical software

    Programming languages : introduction
    Components of programming languages
    (data types, data structures, syntax, control structures)
    Program development
    Statistical software

5/ Approximating probabilities and percentage points in selected probability distributions

    General methods in approximation
    The normal, Student's *t*, beta, *F*, distributions

6/ Random numbers: generation, tests and applications

    Generation of uniform random numbers
    Test of random number generators
    General techniques for generation of nonuniform random variates
    Generation of variates from specific distributions
    Application: the Monte Carlo method, sampling and randomization

7/ Selected computational methods in linear algebra

    Methods based on orthogonal transformations
    Gaussian elimination and the sweep operator
    Cholesky decomposition and rank-one update

8/ Computational methods for multiple linear regression analysis

    Basic computational methods
    Regression model building
    Multiple regression under linear restrictions

9/ Computational methods for classification models

- Fixed-effects models
- Analysis of covariance
- Computing expected mean squares

10/ Unconstrained optimization and nonlinear regression

- Methods for unconstrained optimization
- Computational methods in nonlinear regression

11/ Model fitting based on criteria other than least squares

- Minimum Lp-norm estimators
- Other robust estimators
- Biased estimations

12/ Selected multivariate methods

- Canonical correlations
- Principal components, factor analysis
- Multivariate analysis of variance

Table of contents of the book by Press, Flannery, Teukolsky, and Vetterling "Numerical recipes - The art of scientific computing"
Note : programs in FORTRAN, Pascal or C included

1/ Preliminaries

2/ Solution of linear algebraic equations

3/ Interpolation and extrapolation

4/ Integration of functions

5/ Evaluation of functions

6/ Special functions

Normal, Chi-square, Students's, and F-distributions

7/ Random numbers

Exponential and normal deviates

Gamma, Poisson, and binomial deviates

8/ Sorting

Heapsort, Quicksort

Indexing and ranking

9/ Root finding and nonlinear sets of equations

10/ Minimization and maximization of functions

11/ Eigensystems

12/ Fast Fourier Transform (=FFT)

13/ Fourier and spectral methods

Table of contents of the book by Press, Flannery, Teukolsky, and Vetterling: "Numerical recipes in Fortran 90 - The art of parallel scientific computing"

Table of contents of the book Elements of Computational Statistics by James E. Gentle (taken from Amazon, February 2003)

Review by statman13 (see more about me) from Princeton, NJ USA

At first I thought this was a revision of his excellent book with Kennedy on statistical computing. But after browsing it I discovered it was a book on a subject that is near and dear to my "computationally intensive statistical methods". I then discovered a whole chapter on bootstrap methods, a topic of have studied, taught and written about!

I concur with the editorial reviewer on the content of the book. So I will not go into a detailed description that would just be repetitious.

The distinction that Gentle chooses to make between statistical computing and computational statistics is interesting. He sees statistical computing as methods of calculation. So statistical computing encompasses numerical analysis methods, Monte Carlo integration etc. On the other hand computational statistics involves computer-intensive methods like bootstrap, jackknife, cross-validation, permutation or randomization tests, projection pursuit, function estimation, data mining, clustering and kernel methods. But Gentle includes some other tools that are not necessarily intensive such as transformations, parametric estimation and some graphical methods.

Where would you put the EM algorithm and Markov Chain Monte Carlo? These are computational algorithms and hence I think belong under statistical computing, but they also can be computationally intensive methods especially MCMC. What does Gentle say. Well Chapter 1 is on preliminaries and he includes a section on the role of optimization in statistical inference. Here the EM algorithm is well placed as well as many other computing techniques like iteratively reweighted least squares, Lagrange multipliers and quasi-Newton methods.

The bootstrap chapter provides a self-contained introduction to the topic supported by a good choice of references. Variance estimation and the various types of bootstrap confidence intervals for parameters are discussed. Independent samples are the main topic though section 4.4 briefly describes dependency cases such as in regression analysis and time series.

The book is up-to-date and authoritative and is a very good choice for anyone interested in computer-intensive methods and its connections to statistical computing. This is the way modern statistics is moving and so is worth looking at.

Table of contents of the book "Computational Statistics Handbook with MATLAB" by Wendy L. Martinez, Angel R. Martinez (taken from Amazon, February 2003)

Histograms
Kernel Density Estimation
Finite Mixtures
Generating Random Variables

STATISTICAL PATTERN RECOGNITION
Introduction
Bayes Classification
Evaluating the Classifier
Classification Trees
Clustering

NONPARAMETRIC REGRESSION
Introduction
Smoothing
Kernel Methods
Regression Trees

MARKOV CHAIN MONTE CARLO METHODS
Introduction
Background
Metropolis-Hastings Algorithms
The Gibbs Sampler
Convergence Monitoring

SPATIAL STATISTICS
Introduction
Visualizing Spatial Point Processes
Exploring First Order and Second Order Properties
Modeling Spatial Point Processes
Simulating Spatial Point Processes

APPENDICES
Introduction to MATLAB
Index of Notation
Projection Pursuit Indexes
MATLAB Code for Trees
List of MATLAB Statistics Toolbox Functions
List of Computational Statistics Toolbox Functions

Reviewer: A reader from London, United Kingdom
I think there would be a real interest in a book on "computational statistics" and related topics that showed details of analyses and algorithms using Matlab. This book is expensive and extremely disappointing.The explanations are sparse and very weak and the m.files are usually small add-ons to functions from the Stats Toolbox.
I think in any book on this topic there have to be detailed explanations of how methods work and what their limitations are.Otherwise the reader can find themselves in a lot of trouble very quickly. There is insufficient detail either for a student coming to the topics for the first time or for someone actually wanting to analyse data.
Other books that people might want to have a look at:
1)Statistical Pattern Recognition 2nd edition . Andrew Webb.This is not oriented to any particular language.Good introduction.
2)Netlab. Ian Nabney (this has excellent Matlab functions for neural networks)

3)Modern applied statistics with S 4th edition, Venables and Ripley. This uses a different language (but which will be relatively easy for Matlab users to learn), but learning S or R (free!) makes a huge number of tools available.
4)The recent data mining book by Hand et al. This offers clear and cogent explanations. It is good for someone who does not want overly mathematical descriptions.
I haven't looked properly at the recent Hastie,Friedman and Tibishirani book yet, but you can find reviews on the Amazon page for the book.

Reviewer: A reader from Northridge, CA United States
My major complain was that the authors, in general, did not present algorithms clearly. Limit selections of algorithms did not help either. As a result, you cannot use this book as a reference because it just does not contain enough material. You cannot learn much about computational statistics with this book because the statistic methods and algorithms are not adequately presented. you cannot even write codes for your own statistic analysis with the MATLAB examples shown in the book unless you have the Statistics Toolbox. The only persons that might be benefit from this book are those who don't want to read the Statistics Toolbox manual on line. Given that the Mathworks no longer ship printed manuals, this book may be used a companion of the Statistics Toolbox.

Reviewer: A reader from Atlanta, GA USA
I ordered this book assuming to get something useful, but i got the impression while reading the book that it is like a collection of notes from other books, wrapped with some matlab code. More worrying to me was that the mathematics makes a sloppy impression. For me that means I cannot grab the book to lookup something and use the code without having to be concerned on the validity. The bottomline being I will not use it for applications and the book is a waste of money.
9 of 10 people found the following review helpful:

Reviewer: James C Wrenholt (see more about me) from Lincoln, NE United States
As an independent student of probability and statistics this was a great find. You get a great overview of the useful algorithms of computational statistics. The chapter on Exploratory Data Analysis with its use of multidimensional graphing was very enlightening. It's wonderful that each topic is accompanied by source code (free on-line) that lets you see exactly how it's done. It's easy to tweak the code and explore your own data as you go along. You get just enough theory to understand the algorithm and lots of good common sense and rules-of-thumb on how to best apply it. Finally, the extensive bibliography and chapter-by-chapter annotations will point you straight to the best source for more in-depth study.