STAT-F-408 Statistique informatique/Computational Statistics (M. Guy Mélard)

6 ECTS (théorie : 2, travaux personnels : 4)

Master en statistique (à option)

Master en sciences économiques (à option)

N.B. L'année académique 2009-2010 est la dernière année où le titulaire actuel enseignera ce cours.

0. Introduction (1 h) p. 0.2

LE STATISTICIEN ET L'INFORMATIQUE p. 0.2
Emplois de l'informatique en statistique
Matériels, Logiciels
Accès à l'ordinateur, Critique, Coût, Points communs
Références p. 0.3
Extraits des tables des matières de quelques ouvrages p. 0.5

1. Langages de 3e génération pour applications statistiques (7 h) p. 1

In 2009-2010 : le langage Fortran 90

1.1 Algorithmique p. 3
1.2 Rudiments de Fortran p. 34
1.3 Introduction au Fortran 90 p. 44
1.4 Approfondissement du Fortran 90 et éléments de Fortran 95 p. 61
1.5 Les instructions de Fortran IV et de Fortran 77 à pouvoir déchiffrer p. 72
1.6 Utilisation de librairies scientifiques p. 80
1.7 Elaboration de jeux d'essais p. 82

2. Langages de 4e génération adaptés à la statistique (4 h) p. 84

In 2009-2010 : Matlab ou R

2.1 Principes généraux p. 84
2.2 Exemple: MATLAB p. 84
Tableau comparatif Algorithmes-Fortran-MATLAB p. 85
MATLAB Primer p. 89
Le contenu de la boîte à outils statistique p. 129
Le contenu de la librairie GKSLIB p. 131

3. Principaux algorithmes en statistique (12 h) p. 133

3.1 Calcul de variances et covariances p. 134
3.2 Probabilités et quantiles p. 139
3.3 Génération de nombres et de variables pseudo-aléatoires p. 151
3.4 Méthode de Monte Carlo p. 158
3.5 Régression linéaire multiple p. 164
3.6 Introduction à la régression non linéaire p. 180
3.7 Rééchantillonnage et méthode du bootstrap p. 190
3.8 Compléments et études de cas p. 196

Examen : Travail noté, partiellement sur Fortran (6 points) et partiellement sur Matlab ou R (14 points). Cela va de soi, mais pour éviter des situations lamentables, **je précise que tout emprunt ou référence doit impérativement être citée sous peine de nullité du travail**. Voir http://www.bib.ulb.ac.be/fr/aide/eviter-le-plagiat/index.html

CHAPITRE 0 INTRODUCTION

LE STATISTICIEN ET L'INFORMATIQUE (basé sur un article de Pierre DAGNELIE, Biométrie-Praximétrie, vol. 15, 1975)

Emplois de l'informatique en statistique

- statistique administrative (recensements, enquêtes)
- statistique de recherche
- enseignement de la statistique
- recherche statistique (simulation, ...)

Matériels

- calculatrices scientifiques à fonctions statistiques (factorielles, moyenne, variance, regression linéaire)
- ordinateurs à fonctions statistiques: disparus !

Logiciels

- progiciels statistiques sur micro-ordinateurs, mini-ordinateurs, ordinateurs centraux, super-ordinateurs
  Exemples: SPSS, BMDP, SAS, Minitab, Genstat, Glim, TSP, Throll
  Remarque. La plupart ont été développés pour grosses machines et ont été transférées +/- complètement sur mini puis sur micro-ordinateurs. Problème de choix : consulter Francis (1981)
- logiciels personnels spécifiques à une application
  Danger des programmes personnels : souvent peu efficaces et incorrects, manque de portabilité (langage et matériel). Meilleure solution : utiliser les librairies et concentrer ses efforts sur le programme principal et les jeux d'essais. En voie de disparition

Accès à l'ordinateur

- traitement par lots à distance ("remote batch processing")
  - préparation du travail dans un langage approprié
  - introduction du travail à l'aide d'un éditeur
  - soumission du travail
  - récupération des résultats (liste d'erreurs): écran de visualisation ou listage
  - interprétation
- traitement interactif
  par dialogue avec le progiciel statistique

Critique

- le traitement par lots est plus lent mais ... permet de réfléchir
- le traitement interactif permet des essais (variations de méthodes, graphiques) mais ... on peut perdre la démarche

Coût

Les deux modes sont coûteux :

- traitement par lots : on doit reprendre une partie du traitement (lecture des données, transformations, ...)
- traitement interactif : coût plus élevé sauf sur micro-ordinateur

Points communs

- il faut effectuer la saisie des données une seule fois et la vérifier avec soin

- il faut conserver les résultats en ordre

Références :

Livres

Biran, A. and Breiner, M., "MATLAB for engineers", Addison-Wesley, Wokingham, 1995.
(bien pour MATLAB, mais très peu de statistique)

Chambers, J. M., "Computational methods for data analysis", Wiley, New York, 1977.
(algorithmes, très bien mais un peu vieux)

Chapman, S., Introduction to FORTRAN 90/95, McGraw-Hill, 1997.

Chapman, S., FORTRAN 90/95 for Scientists and Engineers, McGraw-Hill, 1997.

Efron, B. and Tibshirani, R. J., "An Introduction to the Bootstrap", Chapman & Hall, New York, 1993.
(excellent pour le bootstrap, algorithmes en S Plus)

Gentle , J.E., "Random Number Generation and Monte Carlo Methods", Springer-Verlag, 1998.

Gentle , J.E., "Numerical Linear Algebra for Applications in Statistics", Springer-Verlag, 1998.

Gentle, J. E., "Elements of Computational Statistics", Springer-Verlag, 2002.
(plus moderne que Kennedy et Gentle)

Griffiths, P. and Hill, I. D. (eds), "Applied Statistical Algorithms", Ellis Horwood, Chichester, 1985.
(programmes Fortran pour la statistique)

Kennedy, W. J. Jr, and Gentle, J. E., "Statistical Computing", Marcel Dekker, New York, 1980.
(algorithmes, bien mais contient quelques erreurs)

Klinke , S., Data Structures for Computational Statistics, Physica-Verlag, 1997.

Lignelet, P., "Fortran 90 et Fortran 95", Masson , Paris, 1996.
(le livre de Fortran 90 en français)

Martinez W. L., Martinez, A. R, "Computational Statistics Handbook with MATLAB", CRC Press, 2001.
(mauvaise appréciation sur Amazon.com)

Metcalf, Michael and Reid, John, "Fortran 90/95 Explained", Oxford University Press, 1996.
(un bon livre de 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.
(très bien, disquette avec les programmes en Fortran, C ou Pascal disponible)

Press, William H., Teukolsky, Saul A., Vetterling, William T. & Flannery, Brian P., Numerical Recipes in Fortran 77 and Fortran 90 IBM Diskette IBM 3.5 inch diskette, Cambridge University Press, Cambridge, 1996.

Press, William H., Teukolsky, Saul A., Vetterling, William T., Flannery, Brian P. and Metcalf, Michael, Numerical Recipes in Fortran 90, The Art of Parallel Scientific Computing, 2nd edition, Volume 2, Cambridge University Press, Cambridge, 1998.

Press, William H., Teukolsky, Saul A., Vetterling, William T. & Flannery, Brian P., Numerical Recipes in FORTRAN 77, The Art of Scientific Computing, 2nd edition, 1993.

Rao, C.R. (Editor), Computational Statistics (Handbook of Statistics, Vol 9), North-Holland, 1993.

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

Wagener, Jerrold L., "Principles of Fortran 77 Programming", Wiley, 1980.
(excellent livre sur le Fortran 77)

Ward, Tim, and Bromhead, Eddie, "Fortran and the Art of PC programming", Wiley, New York, 1989.
(pas un manuel de Fortran, ouvrage de spécialisation en Fortran sur PC)

Périodiques

J. Roy. Statist. Soc. Ser. C Applied Statistics (algorithmes)

Computational Statistics and Simulation
Computational Statistics and Data Analysis
Computational Statistics Quarterly

Comptes rendus de colloques :

COMPSTAT (International Association for Statistical Computing)
Computational Statistics Section (American Statistical Association)
Interface

Sites : voir le fichier en annexe

Table des matières de l'ouvrage de Thisted: "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 des matières de l'ouvrage de Kennedy et Gentle: "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 des matières de l'ouvrage de Press, Flannery, Teukolsky, and Vetterling: "Numerical recipes in Fortran 77 - The art of scientific computing" =

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

14/ Statistical description of data

Moments of a distribution
Efficient search for the median
Linear correlation, rank correlation

15/ Modeling of data

General linear least squares
Nonlinear models
Confidence limits on estimated model parameters
Robust estimation

16/ Integration of ordinary differential equations

17/ Two point boundary value problems

18/ Integral equations and inverse theory

19/ Partial differential equations

20/ Less-numerical algorithms

References for Volume 1

Index of programs and dependencies (Vol. 1)

Table des matières de l'ouvrage de Press, Flannery, Teukolsky, and Vetterling: "Numerical recipes in Fortran 90 - The art of parallel scientific computing"

21/ Introduction to Fortran 90 language features

22/ Introduction to parallel programming

23/ Numerical recipes utilities for Fortran

Fortran 90 code chapters (B1 to B20)

References for Volume 2

Appendices

Listing of utility modules (nrtype and nrutil)
Listing of explicit interface
Index of programs and dependencies (Vol. 2)

Table des matières de l'ouvrage "Elements of Computational Statistics" par James E. Gentle (pris sur Amazon, février 2003)

Preliminaries

Monte Carlo Methods for Inference

Randomization and Data Partitioning

Bootstrap Methods

Tools for Identification of Structure in Data

Estimation of Functions

Graphical Methods in Computational Statistics

Estimation of Probability Density Functions Using Parametric Models

Nonparametric Estimation of Probability Density Functions

Structure in Data

Statistical Models of Dependencies

Appendices

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 des matières de l'ouvrage "Computational Statistics Handbook with MATLAB" par Wendy L. Martinez, Angel R. Martinez (pris sur Amazon, février 2003)

PREFACE

INTRODUTION

- What is Computational Statistics?
- An Overview of the Book
- MATLAB Code

PROBABILITY CONCEPTS

- Introduction
- Probability
- Conditional Probability and Independence
- Expectation
- Common Distributions
- MATLAB Code

SAMPLING CONCEPTS

- Introduction

Sampling Terminology and Concepts
Sampling Distributions
Parameter Estimation
Empirical Distribution Function
MATLAB Code

GENERATING RANDOM VARIABLES

Introduction
General Techniques for Generating Random Variables
Generating Continuous Random Variable
Generating Discrete Random Variables

EXPLORATORY DATA ANALYSIS

Introduction
Exploring Univariate Data
Exploring Bivariate and Trivariate Data
Exploring Multi-Dimensional Data

MONTE CARLO METHODS FOR INFERENTIAL STATISTICS

Introduction
Classical Inferential Statistics
Monte Carlo Methods for Inferential Statistics
Bootstrap Methods
Assessing Estimates of Functions

DATA PARTITIONING

Introduction
Cross-Validation
Jackknife
Better Bootstrap Confidence Intervals
Jackknife-After-Bootstrap

PROBABILITY DENSITY ESTIMATION

Introduction
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.