

Mathematics and Statistics Undergraduate Handbook

Supplement to the Handbook

Honour School of Mathematics and Statistics Syllabus and Synopses for Part C 2010–2011 for examination in 2011

Contents

1	Honour School of Mathematics and Statistics	2
2	Statistics units and half units	3
2.1	MS1a Graphical Models and Inference – 16 MT	3
2.2	MS1b: Statistical Data Mining - 12HT	4
2.3	MS2a: Bioinformatics and Computational Biology - 16MT	5
2.4	MS2b: Stochastic Models in Mathematical Genetics - 16HT	6
2.5	MS4b/C11.1b: Probabilistic Combinatorics - 16HT	7
3	Mathematics units and half units	9
4	Registration	11

Every effort is made to ensure that the list of courses offered is accurate at the time of going online. However, students are advised to check the up-to-date version of this document on the Department of Statistics website.

Notice of misprints or errors of any kind, and suggestions for improvements in this booklet should be addressed to the Academic Administrator in the Department of Statistics.

Updated October 2010

1 Honour School of Mathematics and Statistics

[See the current edition of the Examination Regulations for the full regulations governing these examinations.]

Students staying on to take the four-year course will take 2 units from Part C in their fourth year, and will also offer a dissertation on a statistics project. Of the 2 units from Part C, at least half a unit will be from the schedule of 'Statistics' units for Part C.

This booklet describes the half-units available in Part C. Information about dissertations/ statistics projects is available on the Department of Statistics website at http://www.stats.ox.ac.uk/current_students/bammath/projects

We ask that you register by the end of week 9 Trinity Term 2010 for classes for the Mathematics/ Statistics courses that you wish to take. A registration form is attached to these synopses. Some combinations of subjects are not advised and lectures in these subjects may clash. However, when timetabling lectures we will aim to keep clashes to a minimum.

Language Classes: Mathematics and Statistics students are also invited to apply to take classes in a foreign language. In 2010-2011, French and (provisionally) German language classes will be run in MT and HT. Students' performance in these classes will not contribute to the degree classification in Mathematics and Statistics. However successful completion of the course may be recorded on student transcripts. See <http://www.maths.ox.ac.uk/current-students/undergraduates/handbooks-synopses> for further information.

2 Statistics units and half units

2.1 MS1a: Graphical Models and Inference 16MT

Recommended Prerequisites

BS1 Applied statistics and BS2a Foundations of Statistical Inference would be helpful but not essential.

Aims & Objectives

Graphical models have become increasingly important in many areas where statistics play a role. They enable the description and analysis of complex stochastic systems via their natural modularity, expressed in terms of (mathematical) graphs which encode conditional independence structure. The modules correspond typically to well-understood, classical models. This course builds upon and develops the specific theory and computational tools needed in the analysis of graphical models for categorical and multivariate Gaussian data as well as Bayesian graphical models for complex stochastic systems.

Synopsis

Topics include:

1. Conditional independence and Markov properties.
2. Log-linear graphical models for categorical data.
3. Gaussian graphical models.
4. Graphical models for complex stochastic systems

Method of Examination

Written examination

Reading

1. D. Edwards, *Introduction to Graphical Models* (2nd ed.), Springer-Verlag, New York (2002).
2. S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford (1996).
3. P. J. Green, N. L. Hjort and S. Richardson, eds. *Highly Structured Stochastic Systems*, Oxford University Press, Oxford (2003).

2.2 **MS1b: Statistical Data Mining** - 12HT plus 4 1-hour computer practical classes

Recommended Prerequisites

Part A Probability and Statistics. BS1 *Applied Statistics* would be an advantage.

Aims & Objectives

'Data mining' is now widely used to find interesting patterns in large databases, for example in insurance, in marketing and in many scientific fields. With large amounts of data we can search for quite subtle patterns. This course concentrates on the statistical tools used to identify patterns, and then to identify those which are interesting not just the result of chance associations.

Synopsis

Fundamentals of pattern recognition, machine learning and data mining.

Exploratory methods: principal components analysis, biplots, independent component analysis, multidimensional scaling.

Cluster Analysis: K-means, hierarchical methods, vector quantisation, self-organising maps.

Linear discriminant analysis, logistic discrimination, linear separation.

Feed-forward neural networks, Classification trees, ensemble methods, V-fold cross-validation. trees, tree ensembles

Method of Assessment

This course is assessed by mini-project.

Reading

C. Bishop, *Neural Networks for Pattern Recognition*, Oxford UP (1995).

D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press (2001).

T Hastie, R Tibshirani, J Friedman, *Elements of Statistical Learning*, Springer (2009)

Further Reading

B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge UP (1996).

2.3 MS2a: Bioinformatics and Computational Biology- 16MT

Recommended Prerequisites

None. In particular, no previous knowledge of Genetics will be necessary.

Aims & Objectives

Modern molecular biology generates large amounts of data, such as sequences, structures and expression data, that needs different forms of statistical analysis and modelling to be properly interpreted. The field of Computational Biology is viewed as the study of the models, statistical methodology and algorithms needed to do bioinformatics analysis. This course aims to present core topics of these fields with an emphasis on modelling and computation.

Synopsis

Stochastic Models of Sequence and Genome Evolution including models of single nucleotide/amino acid/codon evolution.

Phylogenies: enumerating phylogenies, the probability of sequences related by a specified phylogeny, the minimal number of events needed to explain a data set (Parsimony).

Alignment Algorithms. Comparing 2 strings, an arbitrary number of strings, find segments of high similarity in 2 strings. The analogous algorithms for probabilistic models of insertion-deletions [statistical alignment].

Genome annotation – how are proteins and RNA genes predicted for a DNA sequence? Common patterns in a set of sequences. How are identical patterns found in independent strings? How are conserved patterns found by evolutionary analysis (footprinting)?

Network Inference and Network Evolution. Networks can be inferred since they govern observable phenomena like expression levels, concentrations of metabolites. Network evolve over time in analogy with sequences and their evolution must be modelled.

Comparative Biology of Structures, Patterns, Shapes and Dynamical Systems. Any biological object that can be called homologous has an evolutionary model that needs stochastic modelling and this is not confined to sequences and networks.

Integrative Genomics – models analysing multiple types of high-throughput data simultaneously.

Method of Assessment

This course is assessed by mini-project.

Reading

Z Yang, *Computational Molecular Evolution*, Oxford University Press (2004).

C. Semple and M. Steel, *Phylogenetics*, Oxford University Press (2003).

Durbin et al., *Biological Sequence Analysis*, Cambridge University Press (1998).

T. Jiang et al., (editors) *Current Topics in Computational Biology*, MIT Press, (2003).

M.S. Waterman et al., *Computational Genome Analysis: An Introduction*, Springer (2004).

2.4 MS2b: Stochastic Models in Mathematical Genetics - 16HT

Aims & Objectives

The aim of the lectures is to introduce modern Stochastic models in Mathematical Population Genetics and give examples of real world applications of these models. Stochastic and Graph theoretic properties of coalescent and gene trees are studied in the first eight lectures. Extensions to model additional key biological phenomena, and applications, are studied in the second eight lectures.

Synopsis

Evolutionary models in Mathematical Genetics:

The Wright-Fisher model. The Genealogical Markov chain describing the number of ancestors back in time of a collection of genes.

The Coalescent process describing the stochastic behaviour of the ancestral tree of a collection of genes. Mutations on ancestral lineages in a coalescent tree. Inferring the time to the most recent common ancestor in a sample of genes from the number of mutations occurring to the genes. Models with a variable population size.

The frequency spectrum and age of a mutation. Ewens' sampling formula for the probability distribution of the allele configuration of genes in a sample in the infinitely-many-alleles model. Hoppe's urn model for the infinitely-many-alleles model.

The infinitely-many-sites model of mutations on DNA sequences. Gene trees as perfect phylogenies describing the mutation history of a sample of DNA sequences. Graph theoretic constructions and characterizations of gene trees from DNA sequence variation. Gusfield's construction algorithm of a tree from DNA sequences. Examples of gene trees from data. The probability distribution of a gene tree.

Modelling biological forces in Population Genetics:

Recombination. The effect of recombination on genealogies. Detecting recombination events under the infinitely-many-sites model. Hudson's algorithm. Haplotype bounds on recombination events.

Modelling recombination in the Wright-Fisher model. The coalescent process with recombination: the ancestral recombination graph. Properties of the ancestral recombination graph. Applications of coalescent-based methods to the estimation of historical recombination rate.

Introduction to diffusion theory. Tracking mutations forward in time in the Wright-Fisher model. Modelling the frequency of a neutral mutation in the population via a diffusion process limit. The generator of a diffusion process with two allelic types. The probability of fixation and expected time to loss or fixation of a mutation. The frequency spectrum of a mutation.

Genic selection. Extension of results from neutral to selection case. Behaviour of selected mutations. Brief discussion of modern approaches to detecting selection from variation data.

Method of Assessment
Written examination

Reading

R. Durrett, *Probability Models for DNA Sequence Evolution*, Springer (2008).
W. J. Ewens, *Mathematical Population Genetics*, 2nd ed, Springer (2004).
J. R. Norris, *Markov Chains*, Cambridge University Press (1999).
M. Slatkin and M. Veuille, *Modern Developments in Theoretical Population Genetics*, Oxford Biology (2002).
S. Tavaré and O. Zeitouni, *Lectures on Probability Theory and Statistics, Ecole d'Été de Probabilités de Saint-Flour XXXI - 2001*, Lecture Notes in Mathematics 1837. Springer (2004).

2.5 MS4b/C11.1b: **Probabilistic Combinatorics** - 16HT

[In the synopses booklet for Mathematics Part C, this course appears in the Mathematics Department Units section; in this booklet it is in the Statistics section. For any Mathematics and Statistics student taking this half-unit, it will count as a Statistics half-unit. Note that the prerequisite is C11.1a Graph Theory, which is available to Mathematics and Statistics students in Section 3 – C11.1a counts as a Mathematics half-unit for Mathematics and Statistics students.]

Recommended Prerequisites

C11.1a Graph Theory. Part A Probability.

Learning outcomes

To develop an appreciation of probabilistic methods in discrete mathematics.

Aims and objectives

Probabilistic combinatorics is a very active field of mathematics, with connections to other areas such as computer science and statistical physics. Probabilistic methods are essential for the study of random discrete structures and for the analysis of algorithms, but they can also provide a powerful and beautiful approach for answering deterministic questions. The aim of this course is to introduce some fundamental probabilistic tools and present a few applications.

Synopsis

Spaces of random graphs. Threshold functions.

First and second moment methods. Chernoff bounds. Applications to Ramsey numbers and random graphs.

Lovasz Local Lemma. Two-colourings of hypergraphs (property B).

Poisson approximation, and application to the distribution of small subgraphs. Janson's inequality.

Concentration of measure. Martingales and the Azuma–Hoeffding inequality.

Chromatic number of random graphs.

Talagrand's inequality.

Method of Assessment

Written examination

Reading

N. Alon and J.H. Spencer. *The Probabilistic Method*, Second edition, Wiley, 2000.

Further reading:

B. Bollobas, *Random Graphs*, second edition, CUP, 2001.

M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, ed., *Probabilistic Methods for Algorithmic Discrete Mathematics* (Springer, 1998).

S.Janson, T. Luczak and A.Rucinski, *Random Graphs*, John Wiley and Sons, 2000.

M. Mitzenmacher and E. Upfal. *Probability and Computing : Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, New York (NY), 2005.

M. Molloy and B. Reed, *Graph Colouring and the Probabilistic Method* (Springer, 2002).

R. Motwani and P. Raghavan, *Randomized Algorithms* (CUP, 1995).

3 Mathematics units and half units

The Mathematics units and half units that students may take are drawn from Part C of the Honour School of Mathematics. For full details of these units and half-units, see the Syllabus and Synopses for Part C of the Honour School of Mathematics, which are available on the web at

<http://www.maths.ox.ac.uk/current-students/undergraduates/handbooks-synopses>

The Mathematics units and half-units that are available are as follows:

C1.1a Gödel's Incompleteness Theorems

C1.1b Model Theory

C1.2a Analytic Topology

C1.2b Axiomatic Set Theory

C2.1a Lie Algebras

C2.2a Finite Group Theory

C2.2b Building Infinite Groups

C3.1a Algebraic Geometry

C3.1b Algebraic Topology

C4.1a Functional Analysis

C4.1b Banach and C^* Algebras

C5.1a Methods of Functional Analysis for PDEs

[Only available to students who have not offered C5.1a Methods of Functional Analysis for PDEs at Part B]

C5.1b Fixed Point Methods for Nonlinear PDEs

C5.2b Calculus for Variations

C6.1a Solid Mechanics

C6.1b Elasticity and Plasticity

C6.3a Perturbation Methods

C6.3b Applied Complex Variables

C6.4a Topics in Fluid Mechanics

C7.1b Quantum Theory and Quantum Computers

C7.2a General Relativity I

C8.1a Mathematics and the Environment

C8.1b Mathematical Physiology

C9.1a Analytic Number Theory

C9.1b Elliptic Curves

C10.1a Stochastic Differential Equations

C10.1b Brownian Motion in Complex Analysis

C11.1a Graph Theory

C12.1a Numerical Linear Algebra

C12.1b Continuous Optimization

C12.2a Approximation of Functions

C12.2b Finite Element Methods for Partial Differential Equations.

4 Registration

We ask that students register in advance for the classes they wish to take, by the end of week 9 Trinity Term 2010, using the form overleaf.

Because of the large number of options which are available in Part C, some lectures will clash. See the Syllabus and Synopses for Part C of the Honour School of Mathematics for information on which lectures may clash.

FHS MATHEMATICS AND STATISTICS
REGISTRATION FORM: PART C CLASSES 2010-2011

SURNAMEFIRST NAME

EMAIL ADDRESS

COLLEGE

Note: As described in Section 1, you need to do a total of 2 units in Part C (in addition to doing a dissertation on a statistics project). At least half a unit will be from the schedule of 'Statistics' units for Part C

Please give details of the subjects in which you wish to take classes.
I wish to take classes in the following subjects: [Please Tick]

- MS1a Graphical Models and Inference
- MS1b Statistical Data Mining
- MS2a Bioinformatics and Computational Biology
- MS2b Stochastic Models in Mathematical Genetics
- MS4b Probabilistic Combinatorics

For Mathematics units or half-units, please list the unit or half-unit code and name:
Unit code Unit name

.....

.....

.....

Please return this form to the Academic Administrator, Department of Statistics, 1 South Parks Road, by the end of week 9 Trinity Term 2010.