Mathematics and Statistics Undergraduate Handbook Supplement to the Handbook

Honour School of Mathematics and Statistics Syllabus and Synopses for Part C 2017–2018 for examination in 2018

Contents

SC4 is now Advanced Topics in Statistical Machine Learning.

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 July 2017 v.2

1 Honour School of Mathematics and Statistics

1.1 Units

See the current edition of the Examination Regulations at
http://www.admin.ox.ac.uk/examregs/
for the full regulations governing these examinations. The examination conventions can be found at http://www.stats.ox.ac.uk/current_students/bammath/examinations

In Part C,

- each candidate shall offer **six units** from the schedule of units for Part C
- and each candidate shall also offer **a dissertation** on a statistics project.

Of the six units from Part C, at least one unit should be from the schedule of 'Statistics' units.

Note: The dissertation is the equivalent of 2 units, so Part C is the equivalent of 8 units in total (6 from lecture courses, 2 from dissertation).

Units from the schedule of 'Mathematics Department units' for Part C of the Honour School of Mathematics are also available – see Section 3.

This booklet describes the 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

All of the units described in this booklet are "M-level".

We ask that you register by the end of week 10 Trinity Term 2017 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.

1.2 Language Classes

If spaces are available, Mathematics and Statistics students are also invited to apply to take classes in a foreign language. In 2016-2017, French and Spanish language classes will be offered. 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 a student's transcript. See https://www1.maths.ox.ac.uk/members/students/undergraduate-courses/teaching-and-learning/part-bc-students#PartC_Options
for further information.

1.3 Part C courses in future years

In any year, most courses available in Part C that year will normally also be available in Part C the following year. However, sometimes new options will be added or existing options may cease to run. The list of courses that will be available in Part C in any year will be published by the end of the preceding Trinity Term.

1.4 Course list by term

The list of 2017-2018 Part C courses by term is:

Michaelmas Term

SC1 Stochastic Models in Mathematical Genetics
SC2 Probability and Statistics for Network Analysis
SC6 Graphical Models

Hilary Term

SC4 Advanced Topics in Statistical Machine Learning
SC5 Advanced Simulation Methods
SC7 Bayes Methods
SC8 Four Topics in Computational Biology
C8.4 Probabilistic Combinatorics.

2. Statistics Units

2.1 **SC1 Stochastic Models in Mathematical Genetics –** 16 MT

Level: M-level
Method of Assessment: written examination
Weight: Unit

Recommended Prerequisites:
Part A A8 Probability.
SB3a Applied Probability would be helpful.

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 genealogical trees are studied in the first eight lectures. Diffusion processes and extensions to model additional key biological phenomena are studied in the second eight lectures.

Synopsis:
Evolutionary models in Mathematical Genetics:
The Wright-Fisher model. The Genealogical Markov chain describing the number ancestors back in time of a collection of DNA sequences.

The Coalescent process describing the stochastic behaviour of the ancestral tree of a collection of DNA sequences. Mutations on ancestral lineages in a coalescent tree. 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 DNA sequences 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.

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.

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 of a mutation. Genic selection. Extension of results from neutral to selection case. Behaviour of selected mutations.

Reading:

R. Durrett, *Probability Models for DNA Sequence Evolution*, Springer, 2008
A.Etheridge, Some Mathematical Models from Population Genetics. Ecole d'Eté de Probabilities de Saint-Flour XXXIX-2009, Lecture Notes in Mathematics, 2012
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'Eté de Probabilities de Saint-Flour XXXI - 2001*, Lecture Notes in Mathematics 1837, Springer, 2004

2.2 **SC2 Probability and Statistics for Network Analysis** – 16 MT

Level: M-level
Method of Assessment: Written examination
Weight: Unit

For this course, 2 lectures and 2 intercollegiate classes are replaced by 2 practical classes. (The total time for this course is the same as for other Part C courses.)

Recommended prerequisites: Part A A8 Probability and A9 Statistics

Aims and Objectives:

Many data come in the form of networks, for example friendship data and protein-protein interaction data. As the data usually cannot be modelled using simple independence assumptions, their statistical analysis provides many challenges. The course will give an introduction to the main problems and the main statistical techniques used in this field. The techniques are applicable to a wide range of complex problems. The statistical analysis benefits from insights which stem from probabilistic modelling, and the course will combine both aspects.

Synopsis:

Exploratory analysis of networks. The need for network summaries. Degree distribution, clustering coefficient, shortest path length. Motifs.

Probabilistic models: Bernoulli random graphs, geometric random graphs, preferential attachment models, small world networks, inhomogeneous random graphs, exponential random graphs.

Small subgraphs: Stein's method for normal and Poisson approximation. Branching process approximations, threshold behaviour, shortest path between two vertices.

Statistical analysis of networks: Sampling from networks. Parameter estimation for models. Inference from networks: vertex characteristics and missing edges. Nonparametric graph comparison: subgraph counts, subsampling schemes, MCMC methods. A brief look at community detection. Examples: protein interaction networks, social ego-networks.

Reading:
R. Durrett, *Random Graph Dynamics*, Cambridge University Press,2007
P. Grindrod, Mathematical Underpinnings of Analytics, Oxford University Press, 2015
R.v.d. Hofstad, Random Graphs and Complex Networks, Manuscript available at http://www.win.tue.nl/~rhofstad/
E.D Kolaczyk and G. Csádi, *Statistical Analysis of Network Data with R*, Springer, 2014
M. Newman, *Networks: An Introduction*. Oxford University Press, 2010

2.3 **SC4 Advanced Topics in Statistical Machine Learning** – 16 HT

Aims and Objectives:
Machine learning is widely used across many scientific and engineering disciplines, to construct methods to find interesting patterns and to predict accurately in large datasets.
This course introduces several widely used data machine learning techniques and describes their underpinning statistical principles and properties. The course studies both unsupervised and supervised learning and several advanced topics are covered in detail, including some state-of-the-art machine learning techniques. The course will also cover computational considerations of machine learning algorithms and how they can scale to large datasets.

Recommended prerequisites:
Part A A8 Probability and A9 Statistics.
Some material from SB2b Statistical Machine Learning will be used (which is mainly taught in the first two weeks of SB2b)

Synopsis:
Convex optimisation and support vector machines. Loss functions. Empirical risk minimisation.
Kernel methods and reproducing kernel Hilbert spaces. Representer theorem. Representation of probabilities in RKHS.
Nonlinear dimensionality reduction: kernel PCA, spectral clustering.
Probabilistic and Bayesian machine learning: mixture modelling, information theoretic fundamentals, EM algorithm, Probabilistic PCA. Variational Bayes. Laplace Approximation.
Collaborative filtering models, probabilistic matrix factorisation.
Gaussian processes for regression and classification. Bayesian optimisation.
(+Latent Dirichlet allocation [if time allows])

Software:
Knowledge of Python is not required for this course, but some examples may be done in Python. Students interested in learning Python are referred to the following free University IT online courses, which should ideally be taken before the beginning of this course:
https://courses.it.ox.ac.uk/detail/LY020
https://courses.it.ox.ac.uk/detail/LY021

Reading:
C. Bishop, Pattern Recognition and Machine Learning, Springer,2007
K. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, 2012

Further Reading:
T. Hastie, R. Tibshirani, J Friedman, Elements of Statistical Learning, Springer, 2009
Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011, http://scikit-learn.org/stable/tutorial/

2.4 **SC5 Advanced Simulation Methods** - 16 HT

Level: M-level
Methods of Assessment: This course is assessed by written examination.
Weight: Unit

Recommended Prerequisites:
Part A A8 Probability and A9 Statistics.
Part A A12 Simulation and Statistical Programming and SB3a Applied Probability would be an advantage but are not necessary.

Aims & Objectives:
The aim of the lectures is to introduce modern simulation methods.
This course concentrates on Markov chain Monte Carlo (MCMC) methods and Sequential Monte Carlo (SMC) methods. Examples of applications of these methods to complex inference problems will be given.

Synopsis:
Classical methods: inversion, rejection, composition.

Importance sampling. Monte Carlo estimation methods.

MCMC methods: elements of discrete-time general state-space Markov chains theory, Metropolis-Hastings algorithm.

Advanced MCMC methods: Gibbs sampling, slice sampling, tempering/annealing, reversible jump MCMC. pseudo-marginal MCMC.

Sequential importance sampling.

SMC methods: nonlinear filtering.

Reading:
C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd edition, Springer-Verlag, 2004

Further reading:
J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, 2001
B. D. Ripley, *Pattern Recognition and Neural Networks*, CUP, 1996

2.5 **SC6 Graphical Models** – 16 MT

Level: M-level
Method of Assessment: Written examination
Weight: Unit

Recommended Prerequisites:
Knowledge of Part A A8 Probability and A9 Statistics is assumed. SB1 Applied and Computational Statistics and SB2a Foundations of Statistical Inference would be useful, but not essential.

Aims and Objectives:
This course will give an overview of the use of graphical models as a tool for statistical inference. Graphical models relate the structure of a graph to the structure of a multivariate probability distribution, usually via conditional independence constraints. This has two broad uses: first, conditional independence can provide vast savings in computational effort, both in terms of the representation of large multivariate models and in performing inference with them; this makes graphical models very popular for dealing with big data problems. Second, conditional independence can be used as a tool to discover hidden structure in data, such as that relating to the direction of causality or to unobserved processes. As such, graphical models are widely used in genetics, medicine, epidemiology, statistical physics, economics, the social sciences and elsewhere.

Students will develop an understanding of the use of conditional independence and graphical structures for dealing with multivariate statistical models. They will appreciate how this is applied to causal modelling, and to computation in large-scale statistical problems.

Synopsis:
Independence, conditional independence, graphoid axioms
Exponential families, mean and canonical parameterisations, moment matching; contingency tables, log-linear models.
Undirected graphs, cliques, paths; factorisation and global Markov property, Hammersley-Clifford Theorem (statement only)
Trees, cycles, chords, decomposability, triangulation. Maximum likelihood in decomposable models, iterative proportional fitting.
The multivariate Gaussian distribution and Gaussian graphical models. The graphical Lasso.
Directed acyclic graphs, factorisation. Paths, d-separation, moralisation. Ancestral sets and sub-models. Decomposable models as intersection of directed and undirected models.
Running intersection property, Junction trees; message passing, computation of marginal and conditional probabilities, introduction of evidence.
Causal models, structural equations, interventions, constraint-based learning, faithfulness.
The Ising model, Gibbs sampling.

Reading:
S.L. Lauritzen, *Graphical Models*, Oxford University Press, 1996
T. Hastie, R Tibshirani and J.Friedman, Elements of Statistical Learning, 2nd edition, Springer 2009 (available for free at https://statweb/stanford.edu/~tibs/ElemStatLearn/) D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009

J. Pearl, *Causality,* 3rd edition, Cambridge University Press, 2013
M.J. Wainwright and M.I. Jordan, Graphical Models, Exponential Families, and Variational Inference, *Foundations and Trends in Machine Learning*, 2008 (available for free at https://people.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf)

2.6 **SC7 Bayes Methods** – 16 HT

Level: M-level
Method of Assessment: Written examination
Weight: Unit

Recommended prerequisites:
SB2a Foundations of Statistical Inference is desirable, of which 6 lectures on Bayesian inference, decision theory and hypothesis testing with loss functions are assumed knowledge. A12 Simulation and Statistical Programming desirable.

Synopsis:
Theory: Decision-theoretic foundations, Savage axioms. Prior elicitation, exchangeability. Bayesian Non-Parametric (BNP) methods, the Dirichlet process and the Chinese Restaurant Process. Asymptotics, information criteria and the Bernstein-von Mises approximation.

Computational methods: Bayesian inference via MCMC; Estimation of marginal likelihood; Approximate Bayesian Computation and intractable likelihoods; reversible jump MCMC.

Case Studies: extend understanding of prior elicitation, BNP methods and asymptotics through a small number of substantial examples. Examples to further illustrate building statistical models, model choice, model averaging and model assessment, and the use of Monte Carlo methods for inference.

Reading:
C.P. Robert*,The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd edition, Springer, 2001

Further Reading:
A. Gelman et al*, Bayesian Data Analysis*, 3rd edition, Boca Raton Florida: CRC Press, 2014
P Hoff*, A First Course in Bayesian Statistical Methods,* Springer, 2010
DeGroot, Morris H., *Optimal Statistical Decisions.* Wiley Classics Library. 2004.

2.7 **SC8 Topics in Computational Biology** – 16 HT

Level: M-level
Method of Assessment: Mini-project
Weight: Unit

Recommended Prerequisites:
Part A A8 Probability.
SB3a Applied Probability would be helpful.

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. This course focuses on four topics within this vast area: Molecular Dynamics, Molecule Enumeration, Comparative Biology and Advanced Phylogenetics.

Synopsis:
Advanced Phylogenetics – Inference of phylogenies is still an extremely important area, with a huge application domain. The lectures cover processes generating trees; tree enumeration; phylogenetic alignment; compatibility for multistate characters.

Molecular Dynamics (MD) – MD is another huge application area that describes the dynamics of molecules with few to thousands of atoms, for very short time periods like microseconds to nanoseconds. MD is bound to continue to grow for decades, and stochastic methods are central in exploring a large configuration space. The lectures include Hamiltonian Dynamics; integrators and comparison of MD trajectories; a few key applications.

Molecule Enumeration – How many molecules are possible with a given number of atoms, from the set of carbon/nitrogen/phosphorus/oxygen/sulphur (CNPOS)? This question is central in drug design and has many statistical problems embedded. Exhaustive enumeration is at present limited to molecules with 18 CNPOS atoms, and including one more atom expands the numbers about a factor of 10-100 at this point. Lectures cover Polya counting; reaction prediction; and a few applications.

Comparative Biology – Phylogenetics and comparative genomics have been the important areas of the last 15-20 years as a consequence of the growth to sequence data. However, there are other levels of data and biological organisation that are as least as interesting: protein structures, networks, shapes, movements The lectures include models of evolution of these data types; the
so-called COMPARATIVE MODEL; and simultaneous modelling of several levels.

Reading
Notes and slides will be available before each lecture. https://heingroupoxford.com/learning-resources/lectures/

Further Reading
M. Steel, Phylogeny: Discrete and Random Processes in *Evolution*, chapt 1-2, SIAM Press (2003).
T. Schlick, *Molecular Modeling and Simulation*. Chapt 13-14, Springer (2010).
M. Meringer "Structure Enumeration and Sampling" chapt. 8 in *Handbook in Chemoinformatics Algorithms* (eds Faulon) (2010). Chapman and Hall.
B.C. O'Meara Evolutionary Inferences from Phylogenies: A Review of Methods, *Annu. Rev. Ecol. Evol. Syst.* 2012. 43:267–85

2.8 **C8.4 Probabilistic Combinatorics -** 16 HT

Level: M-level
Method of Assessment: Written examination.
Weight: Unit

Recommended Prerequisites:
Part B Graph Theory and Part A A8 Probability. C8.3 Combinatorics is not an essential prerequisite for this course, though it is a natural companion for it.

Overview:
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.

Learning Outcomes:
The student will have developed an appreciation of probabilistic methods in discrete mathematics.

Synopsis:
First-moment method, with applications to Ramsey numbers, and to graphs of high girth and high chromatic number.
Second-moment method, threshold functions for random graphs.
Lovász Local Lemma, with applications to two-colourings of hypergraphs, and to Ramsey numbers.
Chernoff bounds, concentration of measure, Janson's inequality.
Branching processes and the phase transition in random graphs.
Clique and chromatic numbers of random graphs.

Reading:
N. Alon and J.H. Spencer, *The Probabilistic Method* , 3rd edition, Wiley, 2008

Further Reading:
B. Bollobás, *Random Graphs*, 2nd edition, Cambridge University Press, 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,* Cambridge University Press, 1995

3 Mathematics units

The Mathematics units that students may take are drawn from Part C of the Honour School of Mathematics. For full details of these units, see the Syllabus and Synopses for Part C of the Honour School of Mathematics, which are available on the web at https://www1.maths.ox.ac.uk/members/students/undergraduate-courses/teaching-and-learning/handbooks-synopses

The Mathematics units that are available are as follows:

C1.1: Model Theory	16 MT
C1.2: Godel's Incompleteness Theorems	16 HT
C1.3: Analytic Topology	16 MT
C1.4: Axiomatic Set Theory	16 HT
C2.1: Lie Algebras	16 MT
C2.2: Homological Algebra	16 MT
C2.3: Representation Theory of Semisimple Lie Algebras	16 HT
C2.4: Infinite Groups	16 HT
C2.5: Non-Commutative Rings	16 HT
C2.6: Introduction to Schemes	16 HT
C2.7: Category Theory	16 MT
C3.1: Algebraic Topology	16 MT
C3.2: Geometric Group Theory	16 MT
C3.3: Differentiable Manifolds	16 MT
C3.4: Algebraic Geometry	16 MT
C3.5: Lie Groups	16 HT
C3.6: Modular Forms	16 HT
C3.7: Elliptic Curves	16 HT
C3.8: Analytic Number Theory	16 HT
C3.9: Computational Algebraic Topology	16 HT
C4.1: Functional Analysis	16 MT
C4.2: Linear Operators	16 HT
C4.3:Functional Analytic Methods for PDEs	16 MT
C4.6: Fixed Point Methods for Nonlinear PDEs	16 HT
C4.8: Complex Analysis: Conformal Maps and Geometry	16 MT
C5.1: Solid Mechanics	16 MT
C5.2: Elasticity and Plasticity	16 HT
C5.3 Statistical Mechanics	16 MT
C5.4: Networks	16 HT
C5.5: Perturbation Methods	16 MT
C5.6: Applied Complex Variables	16 HT
C5.7: Topics in Fluid Mechanics	16 MT
C5.9 Mathematical Mechanical Biology	16 HT
C5.11: Mathematical Geoscience	16 MT
C5.12: Mathematical Physiology	16 MT
C6.1: Numerical Linear Algebra	16 MT
C6.2: Continuous Optimisation	16 HT
C6.3: Approximation of Functions	16 MT
C6.4: Finite Element Methods for PDEs	16 HT
C7.1 Theoretical Physics	24MT/16HT
C7.4 Introduction to Quantum Information	16 HT

C7.5: General Relativity I	16 MT
C7.6: General Relativity II	16 HT
C8.1: Stochastic Differential Equations	16 MT
C8.2: Stochastic Analysis and PDEs	16 HT
C8.3: Combinatorics	16 MT

4 Registration

We ask that students register in advance for the classes they wish to take, by the end of week 10 Trinity Term 2017, 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 2017-2018

SURNAME . . . . .FIRST NAME . . . .

COLLEGE . . . .

Note: As described in Section 1, you need to do a total of 6 units in Part C (in addition to doing a dissertation on a statistics project). At least one 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]

	SC1 Stochastic Models in Mathematical Genetics **–** 16 MT
	SC2 Probability and Statistics for Network Analysis – 16 equivalent MT
	SC4 Advanced Topics in Statistical Machine Learning – 16 HT
	SC5 Advanced Simulation Methods - 16 HT
	SC6 Graphical Models – 16 MT
	SC7 Bayes Methods – 16 HT
	SC8 Topics in Computational Biology – 16 HT
	C8.4: Probabilistic Combinatorics - 16 HT

For Mathematics units, please list the unit code and name:

. . . .

. . ..

. . . .
Please return this form to the Academic Administrator, Department of Statistics, 24-29 St Giles', by the end of week 10 Trinity Term 2017.