PhD opportunities in Statistics at St Andrews, 2019-2020

(Updated 9th November 2018.)

Applications are welcomed for students wishing to undertake a PhD in Statistics at St Andrews. Full funding (fees, plus stipend of approx. £14,700) is available for well-qualified students; we encourage applications as soon as possible to maximize your chances of being funded. UK, EU and other overseas students are all encouraged to apply. New PhD students would typically start in September 2019, but this is flexible.

Some general information about the division of statistics is given below, followed by a list of specific topics that are on offer this year. Finally, more information is given about how to apply.

Statistics at St Andrews

Statistics is a lively area of research at St Andrews. The Division of Statistics is one of three within the School of Mathematics and Statistics (https://www.mcs.st-and.ac.uk), and consists of 12 members of academic staff, 11 research staff and 9 PhD students. It was ranked first in Scotland in the 2008 Research Assessment Exercise (the last for which statistics was ranked separately from mathematics); the School of Mathematics and Statistics was ranked first in Scotland in the 2014 Research Assessment Framework, and is ranked 5th in the UK for 2019 by the Complete University Guide (source https://www.thecompleteuniversityguide.co.uk).

One major research strength is in the area of statistical ecology: contained within the School is the world-leading Centre for Research into Ecological and Environmental Modelling (CREEM; https://www.creem.st-and.ac.uk), which is housed in tailor-made facilities at the St Andrews Observatory on the edge of the town. We are a founding member of the National Centre for Statistical Ecology (https://www.ncse.org.uk), a multi-institution consortium that ensures regular intellectual exchange between researchers worldwide with similar interests. Several members of CREEM are also part of the university's multi-school Centre for Biological Diversity (CBD; http://synergy.st-andrews.ac.uk/research/cbd/).

A second research focus is in the area of Bayesian statistical inference and, relatedly, computer-intensive inference. Members of staff are also active in the fields of data mining, data smoothing, latent state models and experimental design. Lastly, we have a growing statistical medicine and molecular biology group (https://sites.google.com/view/smmb/home), led by a joint professorial appointment between the Schools of Mathematics and Statistics and the School of Medicine.

A brief summary of the research interest of each member of staff is given at the bottom of this section; more details can be found on the school and CREEM web sites.

New PhD students join a high-calibre but friendly research environment. Training is provided in the first year in the as part of St Andrews' participation in the Scottish Mathematical Sciences Training Centre (www.smstc.ac.uk) and Academy for Postgraduate Training in Statistics (www.apts.ac.uk), the latter consisting of four one-week residential courses. Students may get the opportunity to become involved in externally-funded research as part of CREEM's consultancy group (https://www.creem.st-andrews.ac.uk/consultancy/); they may also be able to assist on statistics training workshops delivered to professional scientists both in the UK and abroad. Some PhDs are supervised jointly with scientists from other institutions, and there may be opportunities for study at those places. PhD studies are expected to last approximately 3.5 years.

St Andrews is a small, vibrant university town. It is situated on the east coast of Scotland and framed by countryside, beaches and cliffs. The town has a rich cultural heritage, having once been at the centre of Scotland's political and religious life. Today it is known around the world as the Home of Golf and a bustling student town with a distinctively cosmopolitan feel, where students and university staff account for more than 30% of the local population. The university is the oldest in Scotland and third oldest in the English-speaking world. It is the top-rated university in Scotland for teaching quality and student satisfaction, and among the top rated in the UK for overall research; it regularly comes in the top few places in UK league tables compiled, for example, by broadsheet newspapers (e.g., 1st place 2018 Times and Sunday Times; 3rd place 2018 The Guardian). Its international reputation for delivering high quality teaching and research and student satisfaction make it one of the most sought-after destinations for prospective students from the UK, Europe and overseas.

More general information about postgraduate student life at St Andrews is given at the university web site https://www.st-andrews.ac.uk/subjects/study-options/pg/ and in the PhD prospectus https://www.st-andrews.ac.uk/study/prospectus/pg-prospectus/. School-specific information about applying is in this pdf: https://www.st-andrews.ac.uk/media/school-of-mathematics-and-statistics/documents/prospective-students/st-andrews-mathsstats-pgr-info.pdf

Brief summary of academic staff interests in the Division of Statistics

- Rosemary Bailey – design of experiments in agriculture, horticulture, ecology and medicine
- David Borchers – spatial capture-recapture, wildlife surveys, spatial modelling
- Carl Donovan – data mining, commercial statistics, multivariate statistics
- Andy Lynch – design or analysis of molecular biology experiments, especially applications of DNA/RNA sequencing to cancer research
- Giorgos Minas – statistics in molecular biology and medicine
- Michail Papathomas – Bayesian methods with application to genetics and biostatistics
- Valentin Popov – time series and hidden process models
- Len Thomas – wildlife (particularly acoustic) surveys, population dynamics modelling
- Hannah Worthington – mark-recapture analyses of wildlife survey data, particularly stopover models

Academic staff not taking PhD students in the coming academic year:

- Steve Buckland – biodiversity, sampling methods, computer-intensive methods
- Ian Goudie – mark-recapture, plant-capture, sequential inference
- Janine Illian – spatial statistics, biodiversity
- Monique Mackenzie – Random effects models, smoothing methods

Specific projects offered for 2019-20

We are currently looking for candidates for the following projects. In addition, prospective candidates with general interests related to those of staff members (see above) are welcome to contact them to discuss other possible projects.

1. Spatial capture-recapture methods for snow leopards.

Supervisors: David Borchers and Richard Glennie (University of St Andrews), and Koustubh Sharma (Snow Leopard Trust)

We do not know how many snow leopards are left in the world. Snow range country governments and scientists have launched an ambitious initiative to develop a robust assessment of the global snow leopard population within the next five years, and monitoring efforts will continue after that. Spatial capture-recapture (SCR) methods are central to these efforts. The very heterogeneous nature and massive range of suitable snow leopard habitat, the tiny fraction of the range that can be surveyed in any year, and the variety of data types (camera trap data, GPS tag data, genetic sampling data, prey survey data, environmental data) that ae available for informing estimates of abundance and density, raises methodological challenges for SCR analysis and for survey design. This PhD will address some of these problems, developing integrated open-population methods for using SCR and other data to assess and monitor snow leopard populations.

2. Modelling wildlife distribution in continuous space and time

Supervisor: David Borchers

So-called “occupancy models” are very widely used to monitor wildlife populations, and change in their distribution over time. This is because they have very undemanding data requirements and so occupancy surveys are cheap and easy to do. But current occupancy models require users to divide the region of potential occupancy into discrete cells, and the methods estimate the probability of these cells being occupied. This is not ideal because there is an infinite number of ways to divide space into cells, and interpretation of the occupancy estimates depends on how space was divided. This PhD aims to develop methods that model occupancy in continuous space, without the need to divide space into cells, and so free the methods from the subjectivity of user-defined cells. In a similar way that space is divided into cells, occupancy models usually divide time into “occasions”, and there is also potential to develop methods that use continuous time rather than discrete occasions.

3. Acoustic spatial capture-recapture methods.

Supervisor: David Borchers

Many species, like gibbons, frogs, kiwi and other bird species, are difficult to see but easy to hear. Using sounds rather than animals as the unit of capture, spatial capture-recapture (SCR) methods have great and largely unexplored potential to provide more reliable estimates of distribution, abundance and trend in populations than any methods currently used. But detecting sounds rather than animals introduces a range of challenges for SCR methods, including greater uncertainty in species recognition, difficult or impossible individual recognition, and the danger of “false positives” (mistaking some other species for the species of interest). This PhD will address these issue, in collaboration with scientists doing acoustic surveys, and researchers developing acoustic identification methods.

4. Object classification from mobile and static sensor feeds

Supervisors: Carl Donovan

The demand for video processing is rapidly increasing, driven by greater numbers of sensors with greater resolution, new types of sensors, new collection methods and an ever wider range of applications. For example, video surveillance, vehicle automation or wildlife monitoring, with data gathered in visual/infra-red spectra or SONAR, from multiple sensors being fixed or vehicle/drone-mounted.

This project will focus on a specific application – object (animal) extraction and classification from extremely high-resolution aerial video from moving platforms. Issues of data size, dynamic backgrounds, rapid platform and target movement and classification errors will all need to be resolved and propagated into the final goal – inferring the densities of target species.

The project will require solving substantive computational bottlenecks and creative programming e.g. GPU and distributed file systems. Elements can be found in Erichson & Donovan (2016), but is only a tiny fraction of what is required.

References:

Erichson, N. B. & Donovan, C. R. (2016) Randomized low-rank Dynamic Mode Decomposition for motion detection. *Computer Vision and Image Understanding*. Vol. 146 pp 40-50.

5. Estimating numbers of molecular ‘species’ in a tissue

Supervisors: Andy Lynch and Hannah Worthington

An experiment to profile e.g. proteins or messenger RNAs will capture a sample of those molecules. There will be a random element to the sample but, depending on the technology used, there will be characteristics that influence the chances of seeing a particular molecule. Most obviously - the more copies of a molecule present in a tissue, the more likely we are to detect it.

When multiple experiments have been run, the tendency in the literature has been simply to sum the number of different types of molecule seen across the experiments. This is clearly only a lower bound on the number of different types of molecule present, and could be improved upon. Better estimates may have an important role to play in experimental design and quality control. This project will look to improve upon current approaches, borrowing ideas from ecological statistics. Factors such as uncertainty in the individual experimental results, biological heterogeneity, knowledge of biological pathways and cross-platform correlations can then be incorporated.

6. Statistical underpinnings of mutational signature analyses.

Supervisors: Andy Lynch and Michail Papathomas

Mutational signatures have been one of the hot topics of cancer research for the past six years. The idea that one can take the mutational profile of a sample and infer the processes that have acted on the tumour over time is an extremely attractive one. One successful approach using non-negative matrix factorization has become extremely popular [1].

We can consider a generalized approach that proceeds as follows: Somatic mutations are identified in samples and sorted into c classes. An n x c contingency table (M) is then constructed by counting the numbers of mutations in each class for each of n patients. We then decompose the matrix M into W x S where S is a Signature matrix of dimension s x c, and W is a weights matrix of dimension n x s. The interpretation being that each column of S represents the mutation signature of a real physical process (e.g. Smoking, UV light exposure) and the row of W indicates how much weight that signature has for a particular patient.

While these methods have been hugely successful, there are a number of statistical questions such as 1) How do we incorporate uncertainty in the measurements M? 2) How do we estimate and report uncertainty about W and S? 3) How do we determine the power of a data set to estimate W and S? 4) How do we ensure a minimal distance between columns of S? 5) How do we incorporate prior information about S? 6) How do we deal with population structures? 7) Can we allow for non-additive contributions of signatures?

This project is to resolve some of the statistical underpinnings of this important methodology.

Reference:
[1] Alexandrov et al. Cell Rep. 2013 Jan 31; 3(1): 246–259.

7. Modelling the interface of metabolism, methylation and mitochondria in prostate cancer

Supervisor: Andy Lynch

The interface of metabolism, methylation and mitochondria is of key importance for understanding the aetiology, biology and prognosis of prostate cancer[1]. Each of these components individually is important and has potential for non-invasive monitoring, but their complex interactions may lead to a misleading signal, particularly when considering a bulk sample, as cell-to-cell heterogeneity may be great.

The prostate gland is a metabolically specialised organ with a primary function to secrete metabolites and proteins required to sustain sperm viability and reproductive efficacy.

Since hydrocarbon pools drawn from metabolic pathways are also used for histone and DNA modifications the implication is that epigenetic reprogramming is initimately connected to metabolic reprogramming and mitochondrial function.

Methylation changes are the most recurrent somatic events seen in prostate cancer (transcending both inter and intra heterogeneity), but whether these are a reflection of broader metabolic changes, or are driver events in and of themselves, has yet to be resolved. This work will allow for investigation of the ordering of events via the proposed model of the network.

The deconvolution of bulk tissue signals to determine the patterns of contribution from distinct tissue types has become commonplace in recent times, with applications to gene expression[e.g. 2-5], methylation[6] and ChIP[7] data. Existing deconvolution tools are generally characterised by a) having the entire genome in which to find signal, b) looking for large, discrete signals c) making little use of biological knowledge and d) generally operating on one dimension of data (transcription or methylation), meaning that we need to develop a new approach for our purposes.

We propose to develop methods that allow deconvolution of multiple combined data types for a well-defined network at the interface of methylation, mitochondria and metabolism. This will be predicated upon prior beliefs regarding a mathematical model of the network, combined with public data sets.

References:
1 Massie CE, Mills IG, Lynch AG (2017) PMID: 27117390
2 Quon G et al. (2013) PMID: 23537167
3 Newman AM et al. (2015) PMID: 25822800
4 Uruttia A et al. (2016) PMID: 27568558
5 Frishberg A et al. (2016) PMID: 27531105
6 Teschendorff A and Zheng SC (2017) PMID: 28517979
7 Rautio S et al. (2015) PMID: 26703974

8. Methods to model telomere length dynamics in a model organism.

Supervisor: Andy Lynch and Helder Ferreira (School of Biology)

Telomeres, the regions at the end of chromosomes, act as molecular clocks to limit cellular proliferation. Consequently, they are mis-regulated in a variety of diseases such as developmental disorders and cancers. Telomeres consist of long stretches of hexanucleotide repeats, making them difficult to study using Illumina DNA sequencing technologies.

The Lynch group have previously worked on estimating average telomere lengths from short-read sequencing and, independently, the Ferreira group have applied Nanopore sequencing to generate long (>4kb) reads from purified telomeres. This project aims to extend these efforts by combining Illumina sequencing with Nanopore sequencing to estimate precise, chromosome-specific telomere lengths from such data sets. This is crucial as it is the shortest individual telomere rather than the average telomere length that triggers cellular stress responses. Data generated specifically for this project will be from C. elegans, as it is a powerful animal model to look at developmental processes.

We will measure telomere lengths in different tissues during development as well as in mutants of conserved genes that cause human developmental disorders such as ATRX. We will look also at the (dis)regulation of gene-expression associated with telomere lengths, bringing in public data and prior knowledge of pathways to complement any data generated for the project.

9. Statistical design and inference for single cell gene expression data

Supervisors: Giorgos Minas and Andy Lynch

The advances in biotechnology over the last few decades provide great opportunities for a better understanding of how our body works and how to keep it healthy and, through this, how diseases work and how to overcome them. These advances have allowed us to observe gene activity in tissue samples, and have revealed much about disease biology despite these samples typically being a mix of many different cell types. More recently, it has become possible to observe gene activity at the single cell level – overcoming problems concerning the heterogeneity of cells in tissue, but possibly introducing new issues.

Such data can reveal complex biological interactions between genes, describe the biological impact of external conditions, and identify the presence of complex and rare cell populations (e.g. in oncology and developmental biology). However, these data are often very noisy and biased by various technological constraints. Therefore, their statistical analysis is challenging, and a number of statistical questions remain unanswered. This project will attempt to address some of these gaps. In particular, we will be interested to compare data from tissue samples to single-cell experiments, and develop methodology for the joint statistical analysis of the two data types.

These will be motivated by a number of case studies such as a) understanding the effects of cell-signalling and environment on gene-expression, and b) understanding the heterogeneity of gene expression present in cancer.

An interest in molecular biology is required but previous experience is not necessary. Background in statistics and/or computational biology will be beneficial.

10. Stochastic modelling of populations of interacting cells with complex underlying phenotypes

Supervisors: Giorgos Minas, Tomasso Lorenzi, Mark Chaplain

Individual based models describe the stochastic evolution of interacting individuals-cells with different phenotypes. Reaction networks on the other hand describe the stochastic interactions between different phenotypes-molecular species within a single cell. Current literature focuses either on the population level with individuals having simple phenotypes or to reaction networks of single cells ignoring the interactions between cells. This project will attempt a joint modelling of these two levels. We will be interested to see the advantages of this approach particularly in terms of analysing cellular decision making and its sensitivity to multidimensional changes of their external environment.

A strong interest in molecular and cellular biology is required but previous experience is not necessary. Background in stochastic processes and stochastic differential equations will be beneficial.

11. Bayesian identifiability for log-linear models.

Supervisor: Michail Papathomas

Log-linear modelling is the standard approach for investigating the full joint dependence structure between categorical variables such as phenotypes and SNPs. Complex dependence structures can be easily discerned using graphical log-linear models (Papathomas and Richardson, 2016). This can potentially lead to identifying functionally important pathways. The number of cells in the associated contingency table increases rapidly with the number of variables, creating sparse contingency tables with a number of zero cell counts, even for a large number of subjects. The presence of zero cell counts can potentially make some model parameters non-estimable, also referred to as non-identifiable. Non-identifiability is a major impediment to evaluating how risk factors interact, and understanding important biological mechanisms. Problems associated with identifiability are currently not sufficiently understood, and have not been addressed in a systematic manner. The aim of this project is to develop methods that will utilize information pertaining to the Bayesian identifiability of interaction parameters, towards choosing the best log-linear model given the data.

References:
Papathomas, M. and Richardson, S. (2016): Exploring dependence between categorical variables: benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. Journal of Statistical Planning and Inference. 173, 47-63

12. Modelling population dynamics from detection survey data

Supervisors: Len Thomas and Richard Glennie

Ecologists collect data on wild animal populations by a variety of survey methods and infer from these how the population changes over time through recruitment, survival, and movement processes. Conceptually, the population size is a hidden quantity that varies over space and time, and data are observed that depends on this hidden quantity and how it changes. Capture-recapture surveys (marked surveys), such as photo ID or camera trapping, can provide long-term data on individuals in the population; while count data or distance sampling surveys (unmarked surveys) provide information on the population level.

In this project, statistical methods would be developed toward the following aims:

1. Improve the modelling of population dynamics from spatial capture-recapture (SCR) data. Photo ID and camera trapping, in particular, can provide long-term data on individuals in a population. This aim would concentrate on developing open population SCR models suited to these survey methods allowing for animal movement, age-structured demographics, and spatially-varying demographics.

2. Integrate data across multiple sources. Data can be collected from breeding surveys, capture-recapture surveys, point counts, citizen science, and distance sampling. Integrated population models (IPMs) are a framework to allow for data from multiple sources to inform a common population process; however, the current methods to fit these models are limited by computation time. Here, the aim is to develop efficient computational methods using hidden Markov models to improve computation time and, therefore, improve model flexibility.

References:
Buckland, S.T., Newman, K.B., Fernández, C., Thomas, L. and Harwood, J., 2007.Embedding population dynamics models in inference. Statistical Science, pp.44-58.
Chandler, R.B., Hepinstall-Cymerman, J., Merker, S., Abernathy-Conners, H. and Cooper, R.J., 2018. Characterizing spatio-temporal variation in survival and recruitment with integrated population models. The Auk, 135(3), pp.409-426.
Ergon, T. and Gardner, B., 2014. Separating mortality and emigration: modelling space use, dispersal and survival with robust-design spatial capture–recapture data. Methods in Ecology and Evolution, 5(12), pp.1327-1336.
Glennie, R., Borchers, D.L., Murchie, M., Harmsen, B. and Foster, R., 2018. Open population maximum likelihood spatial capture-recapture. Pre-print, URI: http://hdl.handle.net/10023/11758.
Newman, K.B., Buckland, S.T., Morgan, B., King, R., Borchers, D.L., Cole, D., Besbeas, P., Gimenez, O. and Thomas, L., 2014. Modelling population dynamics. New York, NY, USA: Springer.

Centre for Biological Diversity-specific project

The following project is based within the Centre for Biological Diversity; note that the primary supervising school is not Mathematics and Statistics.

13. The influence of body condition on functional behavioural decisions of animals.

Supervisors: Nathan Bailey and Patrick Miller (Biology), Len Thomas (Statistics)

The goal of this project is to combine theoretical development with laboratory experiments with an appropriate model organism to predict and evaluate the role of body condition on behaviour. In animals, there is expected to be a fundamental trade-off between foraging and and anti-predator vigilance and behaviour. Energy-store body condition of individuals is predicted based upon existing theory to influence such functional behavioural decisions (Houston et al., 1993; Proc Roy Soc B). By more explicitly developing and testing fundamental theoretical predictions, this project will add value by providing more advanced tools to interpret the biological significance of measures of body condition made with free-ranging animals across a wide range of taxa. Thus, the project will improve our ability to monitor the health status of living animals, aiding in conservation applications.

Application procedure

Although there is no fixed deadline (unless noted otherwise for a particular topic), you are strongly encouraged to make your application as early as possible!

Many details of the general requirements and admissions procedure are given at the university web site https://www.st-andrews.ac.uk/study/pg/apply/research/

Applicants should have a good first degree (UK class 2:1 or better, or international equivalent) in mathematics, statistics or another scientific discipline with a substantial numerical component. Applicants with degrees in other subjects, such as biology, are invited to discuss their qualifications with the Postgraduate Officer (contact details below). A masters' level degree (MSc, etc.) is an advantage, as is any other relevant professional experience. Please note that our primary criterion for selection is academic excellence; most successful applicants (particularly those who are awarded scholarships) have a good to very good 1st class undergraduate degree and/or a distinction at MSc level. Those who do not have English as a first language, and who have not undertaken an undergraduate or graduate degree taught in English, should provide evidence of English proficiency (minimum IELTS 6.5 or equivalent).

Potential applicants are encouraged to contact the Postgraduate Officer responsible for PhDs in Statistics, in advance of making a formal application. He is: Len Thomas, email len.thomas@st-andrews.ac.uk, tel. 01334 461801.

To make a formal application, complete the appropriate online form at https://www.st-andrews.ac.uk/study/pg/apply/research/ (click on "Apply Now" on that page). You also need to provide the following supporting documentation: CV, evidence of qualifications and evidence of English language (if applicable). You are welcome to include a covering letter. You don't need to provide a research proposal unless you are proposing your own project, or sample of academic written work. You will need to ask two referees to provide academic references for you – once you fill in their name on the form, they will be sent emails asking them to upload their references. Please note that we give serious consideration to both the stature of your referees and the remarks that they make about you. More details about the application procedure are given at https://www.st-andrews.ac.uk/study/pg/apply/research/

Further School-specific information is on this page
https://www.st-andrews.ac.uk/maths/prospective/pg/phdprogrammes/
an in this pdf
https://www.st-andrews.ac.uk/media/school-of-mathematics-and-statistics/documents/prospective-students/st-andrews-mathsstats-pgr-info.pdf
which also contains some information about funding and scholarships. In addition to the scholarships mentioned there:

- The Centre of Research into Ecological and Environmental Modelling has a small scholarship fund; all students applying for School funding with an intended PhD topic in the field of statistical ecology are automatically considered.
- An up-to-date list of external scholarships is given at https://www.st-andrews.ac.uk/study/fees-and-funding/postgraduate/scholarships/research-scholarships/.

We look forward to hearing from you!