

**Department of Statistics, University of Leeds,  
2009/10**

**Math3752: Project in Statistics,  
Semester 1, 15 credits**

**Math3753: Project in Statistics  
Semester 2, 15 credits**

**Math5812: Assignment in Statistics,  
semesters 1 and 2, 35 credits**

**Topics for study**

The following staff have offered to supervise projects and assignments in Statistics for 2009/10:

Dr. A.J. Backzkowski, Dr. S. Barber, Dr. C.A. Gill, Dr. A. Gusnanto, Prof. J. T. Kent, Prof. W. R. Gilks (not available Semester 1), Prof. C. C. Taylor, Prof. A. Yu. Veretennikov (not available Semester 1), Dr. J. Voss

Some suggestions for topics are given in the following pages. (This list was last updated 16 September 2009.)

It is also possible to propose your own topic. Please contact me (Professor Kent, [j.t.kent@leeds.ac.uk](mailto:j.t.kent@leeds.ac.uk)) as soon as possible if you wish to take this module in order to discuss a possible topic and supervisor. In particular, there may be limits on the number of students allowed to take these modules, and it may be necessary to re-allocate some projects between supervisors in order to balance workloads.

In general I shall try to accommodate students' interests on a first-come first-served basis.

## **Supervisor — Dr. A.J. Baczkowski**

### I. Statistical clustering methods

This project will examine methods of clustering multivariate data. Work using SAS or R or another statistical package on some data will be undertaken. The student will be expected to learn the appropriate computer commands.

#### References:

1. Chatfield, C. and Collins, A.J. (1980) Introduction to Multivariate Analysis. Chapman-Hall.
2. Krzanowski, W.J. (1994) Multivariate Analysis. Arnold.

### II. Modelling the results of Mathematics examinations

As part of its monitoring of the examination marks the School of Mathematics runs a program to model examination marks as a mean level plus a module effect plus a student effect. This project could examine the model and assess whether it is valid or not (is it robust?), discussing any ways to improve the model (should Statistics modules be assessed differently to Pure modules?). Use of SAS or R or similar will be required for this project.

### III. Computer simulation methods

Supervisor — Dr. A.J. Baczkowski

This project will examine methods of simulation using the computer. The project may examine the use and generation of pseudo-random numbers, the use of simulation techniques in Monte-Carlo methods. Modern methods include Markov chain Monte-Carlo procedures. The project could be directed towards computer methods such as randomization tests and bootstrapping.

#### References:

1. Ripley, B.D. (1987) Stochastic Simulation. Wiley.
2. Ross, S. (1996) Simulation. Academic Press.

### IV. Testing randomness of a sequence of digits

Murier and Rousson (1998) recently performed eight tests of randomness on the first 200000000 digits of  $\pi$ . They found that the digits of  $\pi$  could be regarded as a random sequence. Indeed, they would regard transcendental numbers such as  $\pi$  or  $e$  as giving a definition of randomness!

This project will not attempt to replicate their results but will instead examine various tests of randomness which have been proposed over the years. The advantages and disadvantages of each test can be compared.

#### References:

1. Murier, T. and Rousson, V. (1998) On the randomness of the decimals of  $\pi$ . Student, 2, 237-246.

## V. Time series

The module MATH3802 “Time Series” is mainly devoted to the Box-Jenkins approach to time-series modelling. This module will examine other procedures for studying and forecasting time series data. Topics could include spectral methods, Bayesian methods, non-linear methods, and so on.

This project could be taken by someone not intending to take MATH3802. Equally, it could be taken by someone intending to take MATH3802 — it will not necessarily help you with the material in MATH 3802!

### References:

1. Chatfield, C. (1996) *The Analysis of Time Series: An Introduction* (5th edition). Chapman-Hall.
2. Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction*. Clarendon Press.
3. Tong, H. (1990) *Non-linear Time Series: a Dynamical System Approach*. Clarendon Press.

## VI. Circular data

It has long been known that honeybees recruit fellow bees to fly to sources of nectar by dancing. In a recent series of experiments researchers constructed a robot bee which danced and tried to send bees to a particular location. Around the hive they placed observers who counted how many bees went to that location. The observations from this experiment consisted of the number of bees travelling in different directions around the circle. How can we analyze such data?

Circular data arises in many other situations. Most of the standard statistical procedures you have already met are not directly applicable for such data, so different methods have to be developed. This project will entail the student constructing a series of notes which describe analysis of circular data. Some simple data sets may be obtained and analyzed.

### References:

1. Fisher, N. (1993) *Statistical Analysis of Circular Data*. Cambridge University Press. (You would be strongly advised to purchase this book if accepted to do this topic.)
2. Kirchner, W.H. and Towne, W.F. (1994) The Sensory Basis of the Honeybee’s Dance Language. *Scientific American*, 270 (June 1994), 52-59.
3. Mardia, K.V. (1972) *Statistics of Directional Data*. Academic Press.

Supervisor - Dr S. Barber

**Topics:**

Bootstrapping, generalized linear models, survival analysis are only available as 15 credit projects. Multivariate analysis, sequential clinical trials, and wavelet methods are available as either 15 credit projects or 35 credit assignments.

**Bootstrapping**

Supervisor - Dr S. Barber

We often get to see a sample drawn from some population. From a sample we can calculate a summary statistic, (call it  $\theta$ , say). Examples are the mean, median, and standard deviation. But in order to make use of these summary statistics, we need some idea of their variability. If we know the type of distribution that the data came from (such as normal), we might be able to work out the sampling distribution of  $\theta$ . This enables us to carry out hypothesis tests or find confidence intervals.

Sometimes we do not know the distribution that the data came from, or it is too complicated to work with. In this case, we need alternative methods for testing or constructing confidence intervals. One option is to use the bootstrap method invented by Efron. The basic idea is very simple: to get an idea of the distribution of  $\theta$ , we create a new sample by drawing at random from the original data set. From this, we can compute a new value of  $\theta$ . Once we have done this many times, we have a whole sample of  $\theta$  values which we can use to get some idea about the sampling distribution of  $\theta$ .

The bootstrap has been used in many different areas of statistics and this project could look at some or all of (a) the different types of bootstrap method, (b) the areas in which the bootstrap has been used, and (c) the theoretical justification behind the bootstrap.

**References**

Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.  
Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.  
Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance* **1** 180-182.

**Generalized linear models**

This topic CANNOT be chosen if MATH3715 or MATH5715 are taken  
Supervisor - Dr S. Barber

You have already met examples of what is called the normal linear model: linear regression (LR) and analysis of variance (ANOVA) both model a response variable  $Y$  in terms of one or model explanatory variables  $X_1, \dots, X_p$ . Specifically, LR and ANOVA both assume that the response variable  $Y$  has a normal distribution and relate the mean of  $Y$  to the observed values of  $X_1, \dots, X_p$ .

These are special cases of the generalized linear model (GLIM), which expands the normal linear model in a number of ways. Most importantly, GLIMs are not restricted to a normal response. This allows binomial and Poisson data to be modelled using the same tools as the normal linear model.

This project would involve looking at the properties of GLIMs, describing how GLIMs are built up from key components, and analysing several data sets to illustrate the range of data types that can be modelled using GLIMs.

#### References

- Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. Wiley.  
Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*, second edition. Chapman & Hall / CRC.  
McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman and Hall.

### Survival analysis

Supervisor - Dr S. Barber

A group of patients suffering from a disease receive a drug and their survival times  $T$  are monitored. The observations are the times of death of each patient. How can we estimate the proportion of patients who will survive beyond a time  $t$ , the survival function  $S(t) = P(T > t)$ ? A key feature of survival analysis is that some observations are censored; for some patients we only know that they survived longer than a certain time, not their exact time of death.

Several methods exist to estimate  $S(t)$  which do not require the data to come from any specified distribution. The Kaplan-Meier estimator (Kaplan & Meier, 1958) is but one of these non-parametric estimators. How is it constructed, and what are its properties? Can modelling lifetimes using parametric distributions, such as the exponential, gamma, or Weibull, give further insights into the problem?

Often we wish to compare the lifetimes of two groups of patients receiving different treatments. Here a regression procedure is often used, with the dependent variable  $Y$  being the survival time and the independent variable  $X$  representing the treatment groups. Other covariates such as age and sex can also be included and one method of analysis is Cox's proportional hazards model.

This project will involve studying a selection of these issues and methods in detail and using the methods to analyse some survival data using  $R$ .

#### References:

- Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall.  
Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.  
Collett, D. (1994). *Modelling survival data in medical research*. Chapman & Hall.

### Multivariate analysis

This topic CANNOT be chosen if MATH3772 or MATH5772 are taken  
Supervisor - Dr S. Barber

You have met the idea of *bivariate* distributions such as the bivariate normal, but these are just special cases of *multivariate* distributions which are used in multivariate analysis. Multivariate data arise when several different observations are made at once — examples might include several measurements on an individual, or times to run several different races. Here, each observation is a *vector* rather than being a single value.

Multivariate analysis can be divided into two areas. On the one hand, standard techniques such as *t*-tests have multivariate analogues to test whether the mean of two multivariate normal samples are significantly different. On the other hand, there are many techniques which are distinctly multivariate in nature such as principal components analysis (which looks for “interesting” combinations of the variables) and cluster analysis (which looks for natural groups in the data). This project will involve selecting some of these topics and discussing them in depth before applying them to real data.

**References:**

- Chatfield, C. & Collins, A.J. (1980). *Introduction to Multivariate Analysis*. Chapman & Hall.
- Manly, B.F.J. (2004). *Multivariate Statistical Methods: A Primer*, Chapman & Hall / CRC.
- Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press.

### Sequential clinical trials

Supervisor - Dr S. Barber

Medical regulatory authorities insist extensive clinical trials be carried out before any new drug or treatment is licensed for use. Most large clinical trials are now carried out in a sequential manner: the data are regularly analysed as results become available. In order to do this, special methods are required and this project will look at these “group sequential designs”.

Questions to be addressed could include: What modifications to standard methods are needed to cope with this situation and why are they necessary? What are the practical and ethical benefits of group sequential designs, and what are the drawbacks? Why are “error spending” designs so popular in practice?

The books below are specialised text covering this area in great detail, but many medical statistics books contain some information on designing group sequential clinical trials.

**References:**

- Jennison, C. & Turnbull, B.W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary, *Statistical Science*, **5**, 299-317.
- Jennison, C. & Turnbull, B.W. (1990). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley.

## Wavelet methods in statistics

Supervisor - Dr S. Barber

Wavelets are a special type of basis function that can be used to analyse other functions. They have been used in many areas of statistics including non-parametric regression, density estimation, survival analysis, time series analysis, image analysis (the FBI use wavelets to store their fingerprint data), and pattern recognition (recognising people by their iris patterns). Wavelets were also used in the animation process that produced the film “A Bug’s Life”.

Wavelets are useful in all these areas as they are “localised” – unlike sine waves or polynomials, they are only non-zero in a short interval. This means that they can be used to describe localised behaviour, so they are good for explaining jumps in functions or edges in images.

This assignment would involve looking into just what is required for a function to be a wavelet and the wide variety of wavelet functions that exist. The assignment would also look at some of the range of wavelet applications and illustrate the potential of wavelets through real and artificial examples.

### References

- Abramovich, F., Bailey, T.C., & Sapatinas, T. (2000). Wavelet analysis and its statistical applications, *Statistician* **49**, 1-29.
- Horgan, G.W. (1999). Wavelets for data smoothing: A review and some simulation results, *Journal of Applied Statistics*, **26**, 923-932.
- Mackenzie, D. (2001). Wavelets: Seeing the forest - and the trees, *National Academy of Sciences “Beyond Discovery” series*, <http://www.beyonddiscovery.org>.
- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley.
- Walden, A.T. & Percival, D. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

**University of Leeds**  
**School of Mathematics**  
**Suggested Project titles for MATH 3752/3**  
Supervisor: Dr. Arief Gusnanto

### 1. Ridge regression for highly correlated data

The standard least-squares regression (LS) is useful for modelling if the explanatory variables  $X$  of size  $n \times p$  are approximately independent and its parameters can only be estimated if the number of variables  $p$  is greater than  $n$ . However, this 'ideal' situation may not exist in certain experiments. In chemometrics, calibration of near-infrared instruments produces data with thousands of variables, but from limited number of samples. This is exacerbated with the fact that the variables are highly correlated with correlation coefficient ranges from 0.96 to 1. With these characteristics, LS either fails due to  $n < p$ , or unstable due to high variance of the estimates. Ridge regression (RR) is one of several methods that can be employed to deal with the problem. RR works in such situation due to 'regularisation' of parameter estimation compared to that of LS. This project will explore the use of RR with applications to chemometrics data.

#### References

- Miller, J., and Miller, J. (2005). *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall
- Brereton, R. (2007). *Applied chemometrics for scientists*, John Wiley & Sons
- Hoerl, A., and Kennard, R. (1970). Ridge Regression: Biased Estimation for Non orthogonal Problems, *Technometrics*, **12**: 55–67

### 2. Partial Least Squares Regression

In standard least-squares linear regression (LS), we try to explain a relationship between a vector of response  $y$  with explanatory variables in matrix  $X$  of size  $n \times p$ . LS would be estimable if  $n > p$ . However, experiments in chemometrics and microarray produces data with a characteristic of  $n \ll p$ . In dealing with the problem of  $n \ll p$ , partial least squares regression (PLS) constructs new explanatory variables, often called factors or components, that have maximal covariance with  $y$ . We then regress  $y$  on the components. Interestingly, if  $y$  is binary, it can be shown that PLS can be used for discrimination. This project will explore the use of PLS with applications in chemometrics and microarray data.

#### References

- Miller, J., and Miller, J. (2005). *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall
- Brereton, R. (2007). *Applied chemometrics for scientists*, John Wiley & Sons

Geladi, P., and Kowalski, B. (1986). Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, **185**: 1–17

Barker, M., and Rayens, W. (2003). Partial Least Squares for discrimination, *Journal of Chemometrics*, **17**(3):166–173

### 3. Multiple testing and false discovery rate in microarray data analysis

Microarray technology enables us to obtain expressions of thousands of genes simultaneously, although usually from a limited number of samples. The main objective of the experiments is usually to identify genes that are differentially expressed between two conditions or groups, such as treatment versus control. So, a test is carried out for each genes with the null hypothesis that the gene's expressions are the same between the two groups, i.e. they are 'equally expressed'. However, with thousands of tests done simultaneously, getting false positives is inevitable. To deal with this, a multiple testing adjustment has been proposed, by controlling the probability of family-wise type-I error, or family-wise error rate (FWER). However, using FWER control may be too restrictive for microarray data, where we generate biological hypotheses rather than testing them. Therefore, controlling for false discovery rate (FDR) has been seen as a reasonable method, where we control for the proportion of false positive among significant results. This project will explore the different false positives control methods with applications to microarray data.

#### References

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., Ploner, A. (2005). False discovery rate, sensitivity, and sample size for microarray studies, *Bioinformatics*, **21**, 3017–3024

Dudoit, S., Shaffer, J.P., and Boldrick J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**(1): 71–103

Gusnanto, A., Calza, S., Pawitan, Y. (2007). Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology*, **18**(2):187-193 (available from the first author)

## **Supervisor — Prof. J.T. Kent**

### I. Shape analysis

Shape analysis involves the study of geometric properties of objects which are invariant under changes in translation, rotation and scale (e.g. two similar triangles are said to have the same shape). Applications range from the shapes of leaves and faces to the arrangement of neolithic standing stones. This project will focus on statistical methodology for finding the average shape and summarizing the variability in shape for a group of objects.

References:

1. Dryden, I.L. and Mardia, K.V. (1998) *Statistical Shape Analysis*
2. Small, C.G. (1996) *The Statistical Theory of Shape*.

### II. Robustness

The usual arithmetic mean as an estimator of the centre of a collection of data is optimal under an assumption of normality, but can behave very badly in the presence of outliers. In general the subject of robustness seeks to construct estimators which perform reasonably well when the model assumptions are true, but offer protection in case some rogue data values occur. The median is a simple example of a robust estimator of location. Other important applications include regression and multivariate analysis.

References:

1. Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*. Wiley.
2. Hampel, F.R. Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: the Approach Based on Influence Functions*. Wiley.
3. A.C. Atkinson (1984) *Plots, Transformations and Regression*. Clarendon Press.

### III. Benford's law

Benford's Law describes the distribution of first digits in tables of numbers such as populations of countries, logarithms, etc. The frequencies of the possible digits  $\{1, 2, \dots, 9\}$  are not equal: "1" occurs considerably more often than "9". The purpose of this project is to investigate the types of datasets for which Benford's Law holds and to explore continuous distributions for which the first digit follows Benford's Law either exactly or approximately.

References:

1. Leemis, L.M., Schmeiser, B.W., and Evans, D.L. (2000) *The American Statistician*, 54, 236—241.

### IV. An analysis of the US presidential election of 2000

The US presidential election was extremely close, especially in the deciding state of Florida, where eventually it was decided Bush beat Gore by only a few hundred votes. There have been arguments that a poor design of ballot (the "hanging chads") in Palm

Beach County led to many votes being spoiled and that if the votes had been counted as intended, then Gore would have won Florida and hence the presidency.

This project will look at the statistical evidence behind these claims.

References:

1. R.L. Smith (2002), A statistical assessment of Buchanan's vote in Palm Beach county. *Statistical Science* 17, 441-457.

## Comparing DNA sequence alignments.

*Supervisors: Wally Gilks and Kerstin Hommola.*

Bioinformaticians often need to align DNA sequences which are related through evolution. For example, the following three sequences: CCCAATGAC, CCCAAGGAAT, ACAGTTAAAT could be aligned like this:

```
CCCAATGA--C
CCCAAGGAA-T
AC-AGTTAAAT
```

(where a '-' denotes a gap), or like this:

```
CCCAATG-A-C
CCCAAGG-AAT
-ACAGTTAAAT
```

There are many different methods for sequence alignment, each potentially producing a different result. It would be useful to be able to assess how different these alignments are from each other, and perhaps to be able to derive a 'consensus' alignment. This requires an alignment *metric* to measure the distance between any pair of alignments. We have developed such a metric, but we do not yet have a method to produce a consensus alignment from it.

This project would aim to develop a method for deriving a consensus alignment using the metric, to try it out with some real DNA sequence data and alignment methods, and to find a way to display the results, for example by drawing a 'map' in which each alignment would appear as a point, with the consensus alignment located at a point somewhere in the middle of all the other points.

## The distribution of proline residues in protein sequences.

*Supervisor: Wally Gilks*

Considerable research has been done on the problem of protein folding and structure prediction. A vast wealth of information on amino-acid sequences of proteins and their corresponding three-dimensional structures is now available on-line in databases such as Swissprot. Yet the goal of being able to take the sequence of amino acids in any given protein and from this quickly determine the three-dimensional structure of the molecule still seems far off – current methods are both time-consuming and unreliable.

The amino acid proline may play an important role in the mechanics of protein folding, due to its unique form of interaction with the protein backbone. In this project we will use statistical methods to investigate how the presence of proline in an amino-acid sequence affects the structure of the protein in which it is found.

The first and simplest question we will examine is whether proline is distributed at

random among proteins. If proline plays an important role in the folding of proteins, its appearance in proteins might be expected to appear more consistently than suggested by a null hypothesis of randomness. We will explore this in a sample of protein sequences from Swissprot, using statistical randomisation tests.

By extending this approach, we will explore whether proline is distributed at random within protein sequences, or if it forms patterns in protein sequence space. For example, it may be that proline residues are more regularly spaced within a protein than would be expected by chance. Its spatial relation to other amino acids could similarly be examined – for example its spatial relation to the simplest amino acid glycine. This is of particular interest since it is known that intrinsically disordered proteins (which do not form compact globular structures like most other proteins) are proline-glycine rich. A further extension would be to compare results within and between structural, functional and evolutionary classes of the SCOP protein hierarchy.

.

## **Supervisor — Prof C.C. Taylor**

### I. Nonparametric density estimation

The histogram (in some guise) is widely used at nearly all levels of school, and is often confused with a bar chart. If the data are of a continuous nature, and the histogram is rescaled so that the area under the bars sums to unity, then this is a simple example of a nonparametric estimation of the probability distribution for the data.

In the simple case where the class intervals are all the same width there are two user-choices: the width of the intervals, and the location (a translation parameter). These choices can be critical to both, the appearance of the histogram and to its performance as an estimator of the probability density function. In constructing such a histogram, the class intervals need not be of equal width; in this case there are further user-based choices.

This project will start with a review of histogram and a simple refinement, known as the Average Shifted Histogram (ASH). The second part of the project will focus on an alternative method which makes use of kernel functions. Both parts will involve a mixture of theory — in particular considering sampling properties of the estimators — as well as illustrations using the computer package R.

#### **References:**

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. London : Chapman and Hall.

### II. Statistical analysis of point patterns

Data can often be recorded simply as “events” which happen at a particular location, or at a particular time. Examples of the latter would include arrivals of buses at a bus stop (in which case the data are the arrival times). In this case we may be interested to model the distribution of the arrival times by considering the inter-arrival periods. Simple models would include an exponential distribution for arrivals which follow a Poisson process, or a normal distribution for arrivals which approximate a timetable. Events which are recorded at a location differ in two respects. Firstly, the data are typically 2- dimensional — rather than one-dimensional in time — and secondly, there is no natural ordering of the data. Nevertheless, it is still of interest to explore the distribution of the point locations. Corresponding to the Poisson process for arrival times, there is a point pattern version in which points can be thought of as occurring at random (uniformly distributed) and independently in a given region, where the number of points in the region will have a Poisson distribution. Examples of point pattern data include locations of birds nests, locations of trees and incidences of leukemia.

Statistical methods to analyze such patterns include consideration of the nearest-neighbour distances, as well as attempts to measure variations in intensity.

#### **References:**

Diggle, P.J. (1983). *Statistical analysis of spatial point patterns*. London : Academic Press.

### III. Statistical pattern recognition

A very common type of data is where each observation consists of a number of descriptive measurements (attributes) and a given class. An example would be a bank's database which consists of a description of each customer who has applied for credit (income, sex, age, marital status, home ownership) together with the manager's decision about the application (give credit, refuse credit, interview). In such databases, the attributes may be easy to measure, but the class will often require expertise and experience.

In analyzing such data the goal is to learn or estimate some function which can predict the class given only the attributes. In a two-class problem, logistic regression is one method to provide such a function. More generally, functions can take the form of a decision tree (in which the attributes can be real-values or categorical) or a linear discriminant function. In assessing the accuracy of such a method, we could compare the predicted class with a previously known class.

This project will focus on one of three possible methods to predict classes: classification and regression trees, neural networks, and classical statistical discrimination techniques and will include a blend of theory and computing.

**References:**

Hastie, T., Tibshirani, R., Friedman, J. (2001). *Elements of statistical learning : data mining, inference, and prediction* .  
New York ; London : Springer.

## Supervisor — Prof. A.Yu. Veretennikov

### I. Extreme values theory

In many practical situations we are interested in knowing about the largest value which can occur rather than about the average value. For example, if you live on the coast you will be much more interested in knowing about the likelihood of a high tide breaching the sea-wall outside your front door than in knowing about the average height of the tide along the shore.

The traditional way of tackling such data is through extreme value distributions. We are here interested in the distribution of the maximum value.

Other approaches exist which could be explored. For example, we could regard the height of the sea-wall as a threshold. What is the distribution of the amount by which the sea-level exceeds the height of the sea-wall?

References:

1. Falk, M., Husler, J. e al. (1994) Laws of Small Numbers: Ezremes and Rare Events. Birkhäuser.
- 2 .Gumbel, E.J. (1954) Saisical Theory of Extreme Values and Some Practical Applications. US Government Printing Office.
3. Kinnison, R.R. (1985) Applied Extreme Value Statistics. Macmillan.
4. Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983) Extremes and Related Properties of Random Sequences and Series. Springer.

### Four topics in Statistics and Probability

II. Studying various “measures of quality” for parametric estimators: via Cramer-Rao, Bahadur and using “moderate deviation” approach.

One is estimating a parameter. How to measure the quality of an estimator? In different settings it can be measured by a variance, by values of probabilities of “undesirable events”, etc. The goal is to study various approaches on some simple parametric models.

III. Studying of methods of prediction of random sequences of different nature.

Suppose one is observing a random sequence with values 0 and 1 and the goal is to predict future values. Suppose the sequence is not really random — at least, not really independent and identically distributed — but is constructed by some simple “quasi-random” algorithm. How to reconstruct this algorithm? A problem is such a reconstruction for certain simple random and non-random models. A more advanced problem could be how to proceed if one does not know the exact model.

IV. Studying of the rate of convergence to an equilibrium for a Markov chain.

One is observing a Markov process with unknown transition probabilities. The goal is to estimate the rate of convergence to an equilibrium distribution or/and the spectral gap for the transition probability matrix via observations for certain (simple!) models.

V. Kalman-Bucy filtering.

This will examine such topics as Gaussian vectors, conditional distributions, normal correlation, Kalman-Bucy linear filter. There are a lot of books available.

References

1. A.N. Shiriyayev (1984) Probabilitiy (translated by R.P. Boas). Springer-Verlag.

III. Rigorous mathematical statistics: parameter estimation; density estimation; spectral density estimation; hypothesis testing

Several different topics are combined by one idea: to give a rigorous mathematical presentation of the subject. While applying statistical methods, it is useful (if not indispensable!) to realize constraints and requirements (assumptions) of methods involved. One way to do it is to use rigorous probability theory.

References:

1. Le Cam, L. and Yang, G.L. (2000) Asymptotics in Statistics: Some Basic Concepts. Springer.
2. Borovkov, A.A. (1998) Mathematical Statistics. Gordon and Breach.
3. Ibragimov, I.A. and Khasminskii, R.Z. (1981) Statistical Estimation — Asymptotic Theory. Springer.
4. Bahadur, R.R. (1971) Some Limit Theorems in Statistics. SIAM.
5. Zurbenko, I.G. (1986) The Spectral Analysis of Time Series. North-Holland.

Further topics (with references and abstracts available upon request)

VI. Markov chains as discretization methods for solving partial differential equations

VII. Markov chain stability and applications to MCMC (Markov Chain Monte Carlo)

VIII. Some methods of tracking a signal under a random noise

IX. What are Lyapunov exponents in stochastic systems

X. Linear algebra methods for analysing ergodicity of Markov chains

# Supervisor: Dr. Jochen Voss

For each topic there is a brief sketch of some of the ideas that could be covered.

## 1) *Random Walks on Finite Graphs*

- understand and summarise results about discrete random walks up to the result that hitting probabilities are harmonic functions
- write computer code to simulate paths of a random walk on some simple finite graph (e.g. a rectangle in  $\mathbb{Z}^2$ )
- connect the two, i.e. for some hitting probability compute the result both analytically and numerically and show that both numbers "coincide".

## 2) *Random Number Generation* This topic cannot be chosen if MATH5835 is taken.

- summarise methods for random number generation (e.g. linear congruential generators, thermal noise in semiconductors, dice) and comment on the differences between random and "pseudo random" numbers.
- Implement some random number generators on a computer and do some simple statistical tests on the output to measure quality.
- illustrate that randomly chosen generators are typically very bad by finding an example where things go wrong

## 3) *Ising Model*

- understand and implement a method to simulate states from the Ising model
- generate simulations from the Ising model
- investigate phase transition, i.e. how the number of stationary distributions (one or two) depends on temperature.

## 4) *Mathematics of Juggling [no probability in here]*

- understand and summarise the "siteswap" notation for juggling patterns ( <http://en.wikipedia.org/wiki/Siteswap> )
- prove that the number of balls in a pattern equals the Cesaro limit of the siteswap sequence