Assignment 3

You are required to write a report and give a presentation on *one* of the following topics. The report should address each of the specific questions asked in the topic description below, but need not be limited to these specific points. You should also give enough background to the topic for your answers to make sense to a reader who has not met the topic before — your target audience is an imaginary student on your MSc programme who has not met the topic you are reporting on.

One introductory reference is given for each topic, but you will have to identify others for yourself.

Some of these topics overlap with modules you may have taken here at Leeds. Since the purpose of this project is to give you a chance to practise finding and synthesising information for yourself — which you will need to do for your summer project — you cannot take a project which overlaps with a module you have already taken. Where there is an exclusion, it is noted with the project description.

You must hand in a written report (12–15 pages) at your session in week 11 (two paper copies and a signed "declaration of academic integrity" form please). Also, at your session in week 11, you will have to give a short presentation (8 minutes) on *some aspect* of your topic — the presentation need not cover all the material in your report.

You must also submit .pdf copies of your report (deadline 5pm 8th May) and slides (deadline 9am on the day of your presentation) via the module area on the VLE. This electronic copy will be used to check for any possible instances of plagiarism in your work.

Please understand that while you are expected to work independently on the project, you are free to consult your supervisor to discuss any problems you might have. The supervisor for each project is listed after the project title.

Please email me with your first and second preferences for topics by 9AM on Monday 2nd March 2015.

Topics

1. **Classification and Regression Trees (CART)**

   Consider data $y_i, \boldsymbol{x}_i, i = 1, \ldots, n$ in which either $y_i \in \mathbb{R}$ (regression) or y_i categorical (classification) and each $x_{ji}, j = 1, \ldots, p$ is either categorical or real-valued. CART is a *recursive partitioning* method which grows a tree that be viewed as a fitted "model". It can be used for predicting unknown values of y given new data $\boldsymbol{x}$ (in a similar way to ordinary linear regression models). A recent review is given by Loh (2014).

   Specific questions to address: Describe how CART works, and some inherent limitations. Describe some methods which determine how to split (partition) the data. Investigate how to obtain a tree of the right size.

2. **Models for circular data**

   Most data lie in euclidean space, so distances between points are easily defined. When data are expressed as angles (on the circle or sphere), then more care is required. For example, consider the arithmetic mean of 1° and 359°. A useful introduction is given by Jammalamadaka & Sengupta (2001)

   Specific questions to address: What are the circular counter-parts to the usual summary statistics, or the normal distribution? What about the central limit theorem, and other routes to inference? Can we obtain suitable regression models for the case when the response is circular, but the explanatory variable is linear, or when both are circular, etc.?

3. **Principal variables for dimension reduction**

   For many large-scale datasets, it is necessary to reduce dimensionality of the variables under consideration to the point where further exploration and realistic analysis can be performed. Many dimension reduction techniques aim to transform the variables to give a reduced set, which may not be easy to interpret. Recently, dimension reduction through principal variables has been advocated to aid interpretability. A description is given in Cumming & Wooff (2007).

   Specific questions to address: How does this method relate to the competing dimension reduction techniques of principal component and factor analysis? What are the benefits of the principal variable approach? What are the roles of eigenvalues and eigenvectors in identifying the principal variables? Where have the successful applications of the method been and why were they successful?

4. **Models for compositional data**

   In statistical analyses, compositions are vectors whose components are the proportion of some whole and are subject to a constant sum constraint. For example, data on the percentage of different metals in a rock might be in compositional form (e.g., 20% iron, 40% zinc, 40% lead). Standard multivariate analyses have been shown to be poor at analysing such data and the often-used Dirichlet distribution has many inadequacies. Some alternative models for compositional data are described in Aitchison (1982).

Specific questions to address: What are the difficulties underlying the statistical analysis of compositional data? Why is the Dirichlet distribution often found to be inadequate when modelling compositional data? What other models are available for such data?

5. **Multiple testing corrections**
(exclusions: MATH2735, MATH3880, MATH5880)

   Imagine carrying out a collection of m hypothesis tests, all at say a 5% level of significance, when all the null hypotheses are true. On average some of the tests would appear to be significant by chance. Multiple testing corrections attempt to allow for this problem by making the individual tests more stringent. An overview of the problem is given by Shaffer (1995).

   Specific questions to address: What are the experimentwise error rate and false discovery rate? Which methods control which of these rates? How do the Bonferroni, Holm, and Benjamini-Hochberg methods work? How can these methods be extended to adjusting p-values rather than just deciding which hypothesis to reject at a given significance level?

6. **Causal identifiability** (Peter Thwaites)
(exclusions: MATH5820M)

   Under what conditions is it possible to determine the probability distribution of the effects of a manipulation of a system without doing an experiment? An introduction to this idea is given in Pearl (1995).

   Specific questions to address: Explain Pearl's Back Door and Front Door theorems.

7. **Markov properties on undirected graphs** (Peter Thwaites)

   A probabilistic graph is a network consisting of nodes representing variables and edges representing associations between these variables. Markov properties on such graphs are statements of the conditional independence relationships of these variables. An introduction to probabilistic graphical modelling is given in Lauritzen (1996).

   Specific questions to address: Explain the difference between *pairwise*, *local* and *global* Markov properties. Under what circumstances are these properties equivalent?

8. **Factorising probability distributions over undirected graphs** (Peter Thwaites)

   A probabilistic graph is a network consisting of nodes representing variables and edges representing associations between these variables. There are various results governing whether or not a probability distibution factorises over such a graph. An introduction to probabilistic graphical modelling is given in Lauritzen (1996).

Specific questions to address: Explain the meanings of the words *clique* and *separator* in connection with undirected graphs. How is the decomposability of a graph related to the question of factorisability?

9. **Spatial autocorrelation using Black-White Maps** (Peter Thwaites)
(exclusions MATH2740)

One way of determining the spatial autocorrelation of neighbouring regions is through the idea of Black-White Maps. Regions are coloured black or white depending on the outcome of some binary indicator such as *greater or less than the median*. Analysis is done on the *join count statistics* of these graphs. An introduction to spatial autocorrelation can be found in Cliff & Ord (1973).

Specific questions to address: Derive the moments of the join count statistics under Binomial and Hypergeometric sampling.

10. **Kernel-based density estimation** (Peter Thwaites)
(exclusions: MATH5714M)

When a density function $f(x)$ does not follow any standard distribution, we can attempt to estimate it from the data using a kernel-based approach. An introduction to this idea is given in Silverman (1986).

Specific questions to address: What properties is it advisable for a kernel to have? Describe the different kernels used for this purpose. Derive the expected value for the kernel density $\hat{f}_h(x)$ for a general kernel.

11. **Copula representations of joint probability distributions**

Instead of defining a probability distribution over multiple variables using some joint probability structure, it is possible to break the task down into smaller parts by specifying marginal distributions for each variable separately and then specifying their correlation through a copula. The use of copula representations in distribution specification is discussed in Clemen & Reilly (1999).

Specific questions to address: How is the copula form of a joint probability distribution related to its joint density? What are the commonly used copulae and what are the consequences of choosing different forms? Can we demonstrate the utility of using the copula form in fitting a distribution to some freely available data?

12. **Local linear regression**
(exclusions: MATH5714M)

Given a set of data $(x_i, y_i), i = 1, \ldots, n$, linear regression models are often used to investigate relationships. In the case that the relationship is not linear (and no simple transformation can make it so), then a *local* fit provides an alternative approach. The simplest version is to take a local average (Nadaraya-Watson), but this can be generalized (Fan & Gijbels, 1996)

Specific questions to address: How can the quality of the estimator be assessed? What is the role of the smoothing parameter, and how can it be chosen? How does the local linear estimator compare to the local constant, and under what conditions will one perform better?

13. **Missing data**

    Many data sets suffer from missing data. How they are dealt with is a complex issue, requiring careful thought. The simplest solution — deleting all cases with any missing data — is rarely the best option. One solution is *imputation*, discussed by Sterne *et al.* (2009).

    Specific questions to address: What do the acronyms MAR, MCAR, MNAR stand for any what is their significance? What is wrong with complete case analysis? What is mean (or median) imputation, and why is it (generally) a bad idea? What are Rubin's rules, and how do you use them?

14. **Mixture-based clustering**
    (exclusion: MATH5772)

    Cluster analysis aims to separate observations into suggested natural groups when we have no group membership information. Historically, hierarchical agglomerative clustering has been popular (see any text on multivariate statistics), but a more recent approach is to use clustering algorithms based on mixture distributions, as used by Wehrens *et al.* (2004).

    Specific questions to address: Explain how hierarchical agglomerative clustering works, noting the effects of choice of distance measure and linkage rule. What limitations of hierarchical clustering are overcome by mixture-based clustering? And what are the drawbacks of mixture-based clustering? How, in mathematical terms, does one fit a mixture-based clustering method?

15. **Model-based geostatistics**

    Diggle *et al.* (1998) present a methodology for extending a conventional geostatistical approach for analysing spatially located data to the case where a Gaussian assumption is inappropriate.

    Specific questions to address: What is geostatistics? What is kriging? How is the model-based approach implemented and how is the resulting fitted model assessed?

16. **Minimal surfaces** (John Wood)

    Minimal surfaces are ones which minimize area at least locally. They may be given as parametrised surfaces or as graphs of functions. They satisfy partial differential equations in each case, and are given by a Weierstrass formula.

Specific questions to address: What is the Weierstrass formula? What is Bernstein's theorem for minimal surfaces? Find some generalizations of it more recent than the ones in Osserman (1969).

17. **Gauss maps of surfaces** (John Wood)

    The Gauss map of a surface S in $\mathbb{R}^3$, described by Hoffman & Osserman (1985, sec 1–3), gives the normal to the surface, and thus defines a map from S to the unit 2-sphere. It reflects many properties of the surface.

    Specific questions to address: What properties does it reflect and how? Under what conditions is the Gauss map harmonic? Under what conditions is a map from S the Gauss map of a surface in $\mathbb{R}^3$?

18. **Quaternions and octonians in geometry** (John Wood)

    The quaternions and octonians are number systems which extend the complex numbers; they are described by, for example, Conway & Smith (2003). They have many applications in geometry.

    Specific questions to address: How do quaternions represent rotations in $\mathbb{R}^3$? How does this generalize to $\mathbb{R}^4$? How do octonians allow us to define the exceptional Lie group G_2?

19. **Lie symmetries of differential equations** (Oleg Chalykh)

    Many important ordinary and partial differential equations have a non-trivial group of symmetries which are certain transformations preserving the equation. Knowing these symmetries is very important since this considerably helps solving the equations. An introduction into this subject is given in Olver (1993) and Ibragimov (1999).

    Specific questions to address: What is the Lie bracket of vector fields? How a vector field defines a one-parametric group of transformations? What is the link between Lie groups and Lie algebras? How to define a symmetry of an algebraic equation? What is a symmetry of a differential equation? Give examples of differential equations (either ordinary or partial) with non-trivial Lie symmetries. By using a particular example, describe the general method for determining all Lie symmetries of a given differential equation.

20. **Soliton cellular automata** (Oleg Chalykh)

    Cellular automata became a popular tool for describing systems with complex behaviour since J Conway's celebrated Game of Life. More recently, interesting examples of cellular automata have been found which exhibit some striking patterns similar to those observed in water waves (solitons). Some early works on this subject are Takahashi & Satsuma (1990) and Torii, Takahashi & Satsuma (1996). Since then, further interesting generalisations and links have been found.

Specific questions to address: What is a cellular automaton? What is the box-ball system? Why is it called a soliton automaton? What are the conserved quantities for the (non-periodic) bax-ball system? How to encode them combinatorially? What other versions of the bax-ball system are known?

21. **Complex algebraic curves** (Oleg Chalykh)

    Real algebraic curves have been studied for more than two thousands years ever since ancient Greeks discovered quadrics and their beautiful properties. With the invention of algebra it became clear that it is more natural to study algebraic curves over complex numbers. This theory is at the foundation of algebraic geometry and it has numerous links with other areas, e.g., number theory, topology and knot theory, complex analysis and differential equations. A nice introduction is given in Kirwan (1992).

    Specific questions to address: Why studying algebraic curves is more natural over complex numbers? How to check whether a given algebraic curve is smooth (or regular)? Give examples of singular curves. How one compactifies an algebraic curve? What is the statement of Bezout's theorem? What is the relation between algebraic curves and real 2-dimensional surfaces? What is the genus of a curve?

22. **Stability of point vortex configurations** (Stephen Griffiths)

    A point vortex is a mathematical idealisation of vortices observed in the atmosphere and ocean. The co-evolution of a set of point vortices can be expressed as coupled nonlinear ordinary differential equations, which exhibit a range of interesting behaviour depending upon the number of vortices and the initial conditions. The paper by Acheson (2000) describes some behaviour that arises in a simple system of four vortices of equal strength.

    Specific questions to address: Try to derive some general properties of the equations of motion (i.e., seek conserved quantities, such as analogues of energy, centre of mass, etc.). Derive some of the basic results given in Acheson for the motion of leapfrogging pairs (e.g., speed of translation, and the leapfrogging condition of Love). Write a short computer program to integrate the equations of motion and thus recover the results given in Acheson. Is it possible to prove (analytically) the stability results demonstrated numerically by Acheson? What other similar vortex configurations have been investigated in the past? There is a long history (over 100 years) of studies in this area. What is the effect of adding boundaries to these problems (perhaps with just one or two point vortices)?

References

Acheson, D.J. (2000). Instability of vortex leapfrogging *European Journal of Physics* **21** 269-273.

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Statist. Soc. B*, **44**, 139-77.

Clemen, R.T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, **45**, 208-24

Cliff, A.D. and Ord, J.K. (1973). *Spatial autocorrelation*, Pion.

Conway, J.H. & Smith, D.A. (2003). *On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry*. Natick: A K Peters, Ltd.

Cumming, J.A. and Wooff, D.A. (2007). Dimension reduction via principal variables. *Computational Statistics and Data Analysis*, **52**, 550-65.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299-350.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66*. CRC Press.

Goutis, C. & Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician* **53** 217-224.

Hoffman, D.A. &Osserman, R. (1985). The Gauss map of surfaces in $\mathbb{R}^3$ and $\mathbb{R}^4$. *Proc. London Math. Soc.* **3-50** 27-56.

Ibragimov, N.H. (1999). *Elementary Lie Group Analysis and Ordinary Differential Equations*. Chichester: Wiley.

Jammalamadaka, S.R & Sengupta, A. (2001). *Topics in circular statistics*. London : World Scientific.

Kirwan, F. (1992). *Complex Algebraic Curves*. Cambridge: Cambridge University Press.

Lauritzen, S.L. (1996). *Graphical Models*. Oxford.

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, **82**, 329–348.

Olver, P. (1993). *Applications of Lie Groups to Differential Equations*, 2nd edition. New York: Springer.

Osserman, R. (1969) *A Survey of Minimal Surfaces*. New York, London: Van Nostrand Reinhold.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82** 669–710.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561-576.

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

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., & Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* **338**, 2393-2397.

Takahashi, D. & Satsuma, J. (1990). A soliton cellular automaton. *J. Phys. Soc. Jpn.* **59** 3514-3519.

Torii, M., Takahashi, D. & Satsuma, J., (1996). Combinatorial representation of invariants of a soliton cellular automaton. *Phys. D* **92** 209-220.

Wehrens, R., Buydens, L.M.C., Fraley, C., & Raftery, A.E. (2004). Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification* **21**, 231-253.