1. Proposal for a 4-month Newton Institute Research Programme entitled

Probability and Statistics in Forensic Science

Authors

Professor David Balding, UCL Genetics Institute, University College London

Professor Norman Fenton. School of Electronic Engineering and Computer Science, Queen Mary University of London

Professor Richard D. Gill, Mathematical Institute, Leiden University, Netherlands

Dr David Lagnado, Cognitive, Perceptual and Brain Sciences Department, University College London

Professor Leila Schneps, Institut de Math'ematiques de Jussieu, Universit'e Pierre et Marie Curie, Paris, France

2. Scientific Case for the Proposal

2.1. Mathematical/Scientific Background

The research programme will build on the work of various specialists – many of whom will attend the programme – on the main applications of probability, statistics and Bayesian methods to forensic science. One of the main goals of the proposed research semester is for the participants to produce a *consensual set of guidelines* (extending the recent Practitioner Guides issued by the Royal Statistical Society [47-50][1]) to be made public, specifying conditions under which particular techniques can be used to provide results and reliability estimates that are sufficiently certain to be presented in court without the risk of being challenged on appeal.

Some of the main research topics in this area are the use of Bayesian networks for evidence evaluation (Aitken, Biedermann, Dawid, Evett, Fenton, Kadane, Lagnado, Neil, Overill, Taroni, Thompson) and statistical methods for DNA analysis, particularly in the difficult situations that arise in actual cases of forensic detection: mixed, low template or degraded DNA samples or rare Y-haplotypes (Balding, Buckleton, Cowell, Gill, Lauritzen, Mortera). Statistical databases play an essential role in other types of trace evidence such as fibre analysis, soil analysis, or drug traces on banknotes. Other topics for example concern cell phone evidence, spatial statistics in modelling of fingerprints (Evett, Neumann, Skerrett), and, particularly in the U.S., statistical methods applied to discrimination cases or to truth detection (Faigman, Fienberg, Stern).

[1] A full list of relevant references to support this proposal can be found at www.eecs.qmul.ac.uk/~norman/projects/Newton_Ins_Refs.pdf, and the numbered references used here refer to those.

There are fundamental mathematical, statistical and algorithmic challenges in developing these methods to increase their reliability, and in ensuring consistency and applicability of the methods to forensic evaluation. Furthermore, these subjects suffer in general from a lack of consensus on methodology amongst the forensic community, and even more from the problem of conflicting, controversial and widely misinterpreted court authorities on the application and communication of the methods. Indeed, recent reports [2,8,10,40] have highlighted the difficulties faced by the criminal justice system in responding adequately to the dramatic increase in the amount and complexity of forensic science, especially concerning evidence that has a statistical component such as forensic match evidence. The increasing need for sophisticated mathematical methods in analysing this type of evidence in court, and for these methods to be sufficiently understood to be considered generally acceptable in the courtroom, indicates that a proposal aimed at addressing these problems is not just timely, but even urgent.

The major barriers facing the optimal use of mathematics in the courtroom of these are on three levels: *scientific, cultural*, and *communication*. As explained below in section 5 (outline of programme structure), the programme would attempt to address all three types of barriers in an interdisciplinary manner, with the scientific aspect concerning the whole span of the programme, and the problems of introducing the latest scientific knowledge to members of the legal profession and the task of communicating these ideas to the widest possible public occupying the final two months.

2.2. Main aims of the proposal/Future directions and developments

Aim 1: Unlock the potential of Bayes and Bayesian Networks for analysis and presentation of forensic evidence and legal arguments.

It is increasingly widely acknowledged that Bayesian networks (BNs) and Bayesian methods generally have the potential to solve problems in several of the topics above. A BN is a graphical model of uncertainty that is especially well-suited to legal arguments. It enables us to visualise and model dependencies between multiple hypotheses and pieces of evidence and to calculate the revised probability beliefs about all uncertain factors when any piece of new evidence is presented. BNs have been used for the analysis of DNA evidence specifically, but have also recently been applied to support much broader evidence combination and evaluation. This includes explicit modelling of the uncertainty associated with forensic process and testing errors (because most of the likelihoods required are not easily available, these are often simply ignored), and explicit modelling of crime linkage (wherein the BN may clarify what inferences can legitimately be drawn in linked crimes, as well as highlighting the dangers of the selective use of similarity evidence). BNs also have the potential to incorporate and model ***absence of evidence*** as evidence in its own right. This is a critical, but typically ignored, aspect for much forensic evidence (for example, we have been involved in cases where extremely limited DNA evidence has been presented as highly probative for the prosecution argument, whereas in reality the absence of more significant DNA evidence may have been far more probative in favour of the defence).

Bayesian methods can also potentially tackle head on the fallacy (as in RvT ruling [2]) that there is a clear notion of forensic databases being statistically sound or not. The core issue is that ***all*** existing and potential forensic databases have fundamental weaknesses, irrespective of the type of forensic evidence. All are based on approaches to sampling and data analysis that cannot handle the concept of systematic differences in non-homogeneous populations, nor can they properly handle issues associated with systematic sampling and measurement

errors. Hence fundamentally flawed conclusions may be drawn even when a database with a so-called 'firm scientific base' is used. Ideally the conclusions that can be validly drawn must take full account of the inherent uncertainty associated with database itself. Bayes may help with this (meta-analysis) but has generally been ignored in this context.

While the above show the immense promise and potential of Bayesian methods to address critical problems in forensics and the law, there are a number of practical and theoretical limitations that impede their potential for much more widespread practical use. Here we highlight some of the challenges along with proposed methods to address them:

- Difficulty of identifying acceptable variables and mutually exclusive state spaces to represent the relevant hypotheses, evidence and dependencies between them. This is necessary for coherent arguments before doing any Bayesian (or other) probabilistic inference calculations. This also addresses the many limitations arising from relying on the likelihood ratio as the primary means of evaluating the probative value of evidence [26]. To address this challenge we propose that the work on BN legal idioms [21] provide a starting point.
- Difficulty of properly and efficiently modelling the problems of inherent asymmetry of dependence relationships of the prosecution and defence cases (often chosen by lawyers because they think it is the one with the most compelling supporting argument, rather than what they believe) which result in mutually exclusive causal paths and inevitable redundancy; similarly, the difficulty of being able to reconcile fundamentally different views of experts about related evidence. Current BN theory and tools are neither able to identify redundancy, nor provide alternative structuring. To address this challenge we feel an extended BN theory and graphical notation is required (possibly based on the notion of chain event graphs [42] and extended notions of soft evidence [25]) along with mechanisms that enable easy prototyping and changes.
- Difficulty of specifying prior (and conditional prior) probabilities in a way that can be ultimately acceptable to the legal community. A potential method to address this would be to extend current (mainly manual) approaches to sensitivity analysis (using extended inference algorithms), which could also minimize the amount of probability elicitation required (and hence reduce dependence on subjective priors). It may be possible to identify those variables where fairly wide differences in the prior probabilities have little impact on model predictions.
- Addressing the limitations claimed in [11] of the capability of BNs in handling continuous variables. To address this challenge we propose building on the dynamic discretization work [33], since the capability of handling continuous variables is critical for more accurate analysis and presentation of forensic evidence.

Aim 2: Address the challenge offered by new DNA methodologies

In recent years the principal challenge in the area of DNA evidence has been the evaluation of low–template DNA profiles. Here "template" roughly means the amount of DNA measured in units of mass, and refinements of DNA profiling protocols means that usable profile information can be obtained from just a few tens of picograms of DNA, which is the

DNA content of a handful of human cells. The other relevant factor is degradation: progress has also been made in profiling very small and degraded samples. These both generate stochastic effects that complicate the evaluation of evidential weight, and forensic scientists have in the past developed short-cuts and approximations of doubtful validity. More rigorous statistical evaluation of low-template/degraded profiles has recently made great advances and several software packages are now available. More work is needed to compare different modelling approaches and software, and issues surrounding the evaluation of hypotheses. Our programme will contribute to establishing best practice, and hence encourage courts to adopt these superior but more complex methods of evaluation.

Other issues are now coming to the fore in the forensic analysis of biological samples. The short tandem repeat (STR) genetic markers developed around 20 years ago are still the most widely used, but in other areas of genetics SNP markers and DNA sequencing have replaced STRs. If we were to start again now, it is unlikely that STR technology would be adopted, because of well-known problems with experimental artefacts such as stutter. The advantages of the newer technologies have not so far been sufficient to justify the cost of switching, but is seems likely that at least for some forensic applications DNA sequencing will become prominent in the near future, and there is a need to develop novel statistical methods to evaluate evidential weight in the presence of, for example, occasional sequencing errors.

Other new genomics technologies are rapidly gaining a foothold in forensics, including RNA and methylation-based assays. These can both be used to investigate the tissue from which a small biological sample has been obtained: in certain scenarios it is important to know whether the profiled DNA has come, for example, from semen or from menstrual or non-menstrual blood. Methylation data have recently been used to identify the age of the donor of a biological sample. There will be many statistical issues connected with the evaluation of such evidence: it is difficult to be precise about which questions because these technologies and their forensic applications are evolving rapidly, but there has to date been almost no investigation of relevant statistical issues. An over-arching theme is prediction of characteristics of the person from whom a biological sample is derived: for example various visual phenotypes relating to face, hair and eyes, as well as height and age. We will evaluate the many statistical tools for prediction in this setting, and for example discuss combining different types of biological markers into prediction algorithms.

Aim 3: Solve the "fundamental problem of forensic statistics".

When using Y-chromosome DNA profiles, it frequently happens that a DNA profile found on a crime scene and matching a suspect's profile does not appear in the relevant database. This creates a big challenge to the analyst, who is required to supply a likelihood ratio (LR) or match-probability in order to quantify the evidential value of the match. Sensible estimation of the LR seems to rely on sensible estimation of the population frequency of this previously unseen haplotype. Recent notable cases when this occurred are the Amanda Knox case in Italy, and the Tamara Wolvers case in the Netherlands: both cases are controversial.

This apparently rather specialised problem is actually a paradigmatic example of a broad class of problems in forensic statistics. New technologies supply new opportunities for "individualisation" (the aim of identifying the perpetrator of a crime through a match between a crime scene trace and a sample from a suspect). Each new technology starts with a rather restricted knowledge base; in particular databases are small. New cases will frequently

exhibit "features" which were previously unknown or very rare. The rare haplotype problem is a mathematically clean and extreme version of problems occurring all over forensic statistics. If we can solve the problem in its most simple manifestation, we can hope to make progress when we are faced with the same problem in more complex contexts.

For the rare haplotype problem, there are already exist three approaches of quite different nature: Roewer et al. (2000) [51], based on Bayesian estimation of the haplotype frequency with a beta prior; Brenner (2010) [52], based on the number of singletons observed in the database; and Andersen et al. (2013) [53] using a mixture of independent discrete Laplace distributions as a parametric approximation of the distribution of allelic frequencies, justified by branching diffusion process models of mutation and population growth. Recently Cereda and Gill (2014) [54] have added two further proposals: one is similar to Brenner's, and like Brenner's is strongly related to the classic Good-Turing estimator. A second method is based on Anevski et al's study [55] of a non-parametric maximum-likelihood estimator. It is somehow intermediate between the parametric approach of Andersen and non-parametric methods based on Good-Turing estimators. It might avoid the disadvantages of those while moreover providing a supplementary means of evaluating their accuracy.

For any method it is imperative to assess two more levels of uncertainty, beyond the uncertainty about which hypothesis is true given the evidence, which would hold if we knew everything about the population probability distribution. LR is a ratio of probabilities which are usually based on a model which is at best only a good approximation to the truth. Moreover we only estimate parameters of that model by fitting it to the data in our database.

On top of this proliferation of models and techniques, confusion reigns in the literature about "the likelihood ratio". Of course it is agreed that the likelihood ratio means $\Pr(E \mid H_p, B) / \Pr(E \mid H_d, B)$ where E = Evidence, B = Background, and H_p and H_d are hypotheses of prosecution and of defence. What is not agreed is what we should take as E, what we should take as B, and what is "Pr". The analyst has a totality D of data at their disposal. This could for instance consist of a data-base of DNA profiles from past cases, samples, or whatever; together with the specific DNA profiles in the case at hand. Both E and B are parts of D, mathematically they are both functions of D. There are many choices involved – what should we take as E and what shall we take as B? Is the pair (E, B) simply a one-to-one transformation of D, or do they together form a reduction of the data D? Reducing the data means that information has been discarded, but in a situation with many nuisance parameters it can be wise to do this: one discards (or conditions on) the part of the data which primarily tells us about the nuisance parameters, and only indirectly about the primary question of interest (prosecution versus defence).

The previously listed "solutions" of the rare haplotype problem actually all correspond to different choices of data reduction, different choices of E and B, as well as to different specifications of "Pr". They are not just different estimators of the same likelihood ratio: they are different estimators of different likelihood ratios. If we were God and knew all probabilities we would simply look at $\Pr(D \mid H_p) / \Pr(D \mid H_d)$. In practice we don't know everything; in particular we don't know Pr. We do not report to the court *the* likelihood ratio but we report to the court *an* estimate/surrogate/guess for *a* likelihood ratio. It could for instance be very wise to *reduce* the data D to a much smaller amount of information (E, B) because *the* likelihood ratio based on the data reduction is much easier to estimate than the

likelihood ratio based on all the data. And under "easy to estimate" we can distinguish two dimensions: bias and variance. A "wrong" model with fewer unknown parameters introduces bias but can dramatically reduce variance, and could thereby well finally give better predictions, i.e., more reliable likelihood ratios.

As we move information from *E* to *B* we may make the statistical task of the analyst easier, and the information which he or she gives to the court more reliable, but at the same time, Bayesian base rates may be changing, in particular, if the background information *B* delimits the population of potential perpetrators. Altogether, the rare haplotype problem embodies just about every challenge in forensic statistics in a pure form.

2.3 Workshops

One opening workshop and two international conferences are planned, with the latter aimed at bringing academic experts in the domain who are not participating in the programme to the Newton Institute. The latter part of the programme will also host several panels and one-day workshops aimed at collaborating with and communicating with members of the legal profession.

The opening workshop, near the beginning of the semester, entitled *The nature of questions arising in court that can be addressed via probability and statistical methods*, will provide a structural framework for the main goals of the research programme.

The first international conference, *Bayesian networks in evidence analysis*, will take place early in the second month of the programme, and will be devoted to investigating the use of Bayesian networks in evidence analysis models for actual criminal cases ranging from the simplest (with very little evidence) to more complex ones. This workshop could be structured around producing satisfactory BN models for some key well-known cases, for example those of Sally Clark and Barry George (use of probability for considering probative value of specific individual pieces of evidence and combinations of evidence), since in such cases it is known that all previous attempts to properly model these have not been complete.

The second international conference, *Statistics in biological and physical evidence analysis*, planned for early in the third month of the programme, will be devoted the use of cutting-edge statistical methods in all the trace analysis situations that cause problems in court: physical evidence problems such as fibre or soil analysis, and DNA problems such as mixed samples, LCN samples, degraded samples and rare Y-haplotypes. Again, the investigations will be supported by analysis of key well-known cases such as the Amanda Knox trial (low template and mixed DNA and rare Y-haplotype).

Because it is difficult for members of the legal profession to be free of their professional obligations long enough to attend a lengthy conference, the latter part of the programme, devoted to *sharing the methods developed during the programme with the wider legal community*, will be based on a series of panels, open days and 1-day workshops. The third month will be devoted to involving key members of the legal profession experienced in the problems of evidence analysis in the task of setting up the guidelines, and the fourth month to communicating the results to the widest possible legal public via high-impact days organised not only within the Newton Institute but also in London and possibly other cities; far-flung participants can also be reachable via video linking.

2.4. Suitability for the Newton Institute and benefit to UK

Unlike most mathematical domains, the mathematics of forensic science suffers not only from the difficulty of sharing and communicating new developments amongst mathematicians but, more severely, that of communicating progress to the wider forensic community in such a way as to make a real difference in situations of crime detection and trials. A significant part of the problem arises from the fact that as various individuals are working independently on some of the important questions described above, there is no absolute consensus on certain basic tenets that need to be respected in order for mathematical arguments to be correctly applied, which makes any mathematical argument used in court vulnerable to challenge. Beyond being an opportunity for joint research and teaching activities, an important goal of the meeting would be to unanimously set forth conditions guaranteeing that any mathematical argument used in court that respect them fully must be unimpeachable. The Newton Institute seems to be one of the only institutes in the world with the infrastructure to host a long-term joint project involving a hundred people.

The UK is in the forefront of some of the most important recent advances in probabilistic and statistical methods in forensic science; it is also one of the countries in which mathematics has suffered most in court, with a series of high profile cases in some of which defendants have been wrongly convicted due to faulty mathematics, whereas in others correct mathematical arguments have been rejected on appeal on the grounds that such reasoning encroaches on the traditional province of the jury, or rulings were made against the use of mathematics at all in certain situations. Any new research or general conclusions produced from the research semester will be liable to have a direct impact on the practice of forensic science in the UK and on its presentation in the courtroom.

3. Programme organisers and Scientific Advisory Committee

The research semester will be planned conjointly by:

Organisers:

David Balding, Professor of Statistical Genetics, University College London

Norman Fenton, Professor of Risk Information Management, Queen Mary University London

Richard Gill, Professor of Mathematical Statistics, Leiden University, Netherlands

David Lagnado, Senior Lecturer in Cognitive and Decision Sciences, University College London

Leila Schneps, Research Director in Mathematics, Université Pierre et Marie Curie, Paris, France

4. Outline of programme structure

One or more organisers will be present throughout the semester, and each will be responsible for some lectures or introductory courses. Research activities will be carried out for the entire fourth months duration. During the first two months, a series of short courses given by academic experts will address foundational issues with the purpose of establishing a common knowledge base for participants who are specialists in particular areas and for young researchers, who may be participants in the research programme or come from outside.

Furthermore, regular expository lectures will illustrate the various research topics being carried on during the semester.

1st month: The opening month of the workshop will host an initial workshop entitled:

The nature of questions arising in court that can be addressed via probability and statistical methods. The speakers will be for the most part mathematicians and statisticians who have first-hand experience of acting as expert witnesses in court. Case analyses will be an essential part of the content.

There will also be 6-hour introductory short courses on the following topics:

Bayesian networks and associated software, case analysis, and statistical frameworks for legal reasoning.

2nd month: During the second month of the workshop, the knowledge base developed in the first month will be shared and consolidated through an international conference uniting participants in the semester with specialists from outside, on the subject:

Bayesian networks in evidence analysis

There will also be 6-hour introductory short courses on the following topics:

Statistical methods in DNA analysis, analysis of trace evidence

3rd month: The knowledge base developed around DNA and trace analysis during the second month will be shared and consolidated through a second international conference, to be entitled:

Statistics in biological and physical evidence analysis

The specific goal of the third month of the semester will be to begin work on the task of bringing the scientific research out to the legal community. Throughout the month, panels and working groups bringing together evidence experts from academia and the law will work on prioritising the research ideas with respect to their feasibility and impact in a legal context, and incorporating them into guidelines establishing their proper use and limitations in legal situations.

4th month: The final month of the programme will be devoted to bringing the fruits of the previous months to a much wider audience, so as to present a genuine consensus to the legal community. This dissemination of results will be accomplished through a series of high-impact "days" targeted at members of the legal profession; some of these can take place at the Newton Institute and others in London or even in both or other places via video linking. One of the main goals of the semester is to produce and distribute a set of written guidelines to help regulate the use of the most advanced mathematical techniques within the courtroom.

5. Recent or forthcoming short meetings on similar topics

Quantitative Methods in the Legal Evaluation of Evidence – challenges and prospects, 10 April 2013, Lund University, Sweden

Spatial Statistics 2013: Revealing intricacies in spatial and spatio-temporal data with statistics, 4-7 June 2013, Columbus, Ohio, USA

Evaluation of Forensic Evidence Workshop, Forensic Science Society, 8 & 9 July 2013, University of Central Lancashire

Forensic Horizons 2013 - supporting research and development & delivering best practice for the justice system. Forensic Science Society Annual Conference, Manchester, 6-8 November 2013

Australasian Bayesian Network Modelling Society (ABNMS) Hobart, Tasmania, 27-28 November 2013

First Winchester Conference on Trust, Risk, Information and the Law, Winchester University, 29 April 2014

Forensic Europe Expo – Evaluation of Forensic Evidence, 29 – 30 April 2014, Olympia, London, www.forensicseuropeexpo.com/page.cfm

International conference on Behavioral Legal Studies: Cognition, Motivation, and Moral Judgments, June 10-12 2014, Jerusalem and Ramat-Gan, Israel.

DNA in Forensics 2014 – 9th international Y-chromosome meeting, 6th international EMPOP meeting, Brussels, 14-16 May, 2014.

ICFIS2014, 9th International Conference on Forensic Inference and Statistics, 19-22 August, 2014, Leiden University, Netherlands

20th Meeting of the International Association of Forensic Sciences (IAFS), 13 – 17 October 2014, COEX, Seoul, Korea, http://www.iafs-2014.org/about.php