City Research Online

City St George's, University of London

Citation: Cowell, R., Lauritzen, S. L. & Mortera, J. (2007). Identification and separation of DNA mixtures using peak area information. pp. 28-34. doi: 10.1016/j.forsciint.2006.03.021

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2369/

Link to published version: https://doi.org/10.1016/j.forsciint.2006.03.021

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Cass means business

Cass Business School
City of London

Faculty of Actuarial Science and Statistics

Identification and Separation of DNA Mixtures Using Peak Area Information

R.G. Cowell, S.L. Lauritzen
and
J. Mortera.

Statistical Research Paper No. 25

November 2004

ISBN 1-901615-82-0

Cass Business School
106 Bunhill Row
London EC1Y 8TZ
T +44 (0)20 7040 8470
www.cass.city.ac.uk

“Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission”.

Identification and Separation of DNA Mixtures using Peak Area Information

R. G. Cowell*
Cass Business School

S. L. Lauritzen
University of Oxford

J. Mortera
Università di Roma Tre

November 26, 2004

Abstract

We show how probabilistic expert systems can be used to analyse forensic identification problems involving DNA mixture traces using quantitative peak area information. Peak area is modelled with conditional Gaussian distributions. The expert system can be used for ascertaining whether individuals, whose profiles have been measured, have contributed to the mixture, but also to predict DNA profiles of unknown contributors by separating the mixture into its individual components. The potential of our methodology is illustrated on case data examples and compared with alternative approaches. The advantages are that identification and separation issues can be handled in a unified way within a single network model and the uncertainty associated with the analysis is quantified.

Some key words and phrases: Bayesian network, conditional Gaussian distributions, DNA mixture, DNA profile, forensic identification, mixture separation, probabilistic expert system, peak weight.

*Faculty of Actuarial Science and Statistics, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ, UK. email: rgc@city.ac.uk

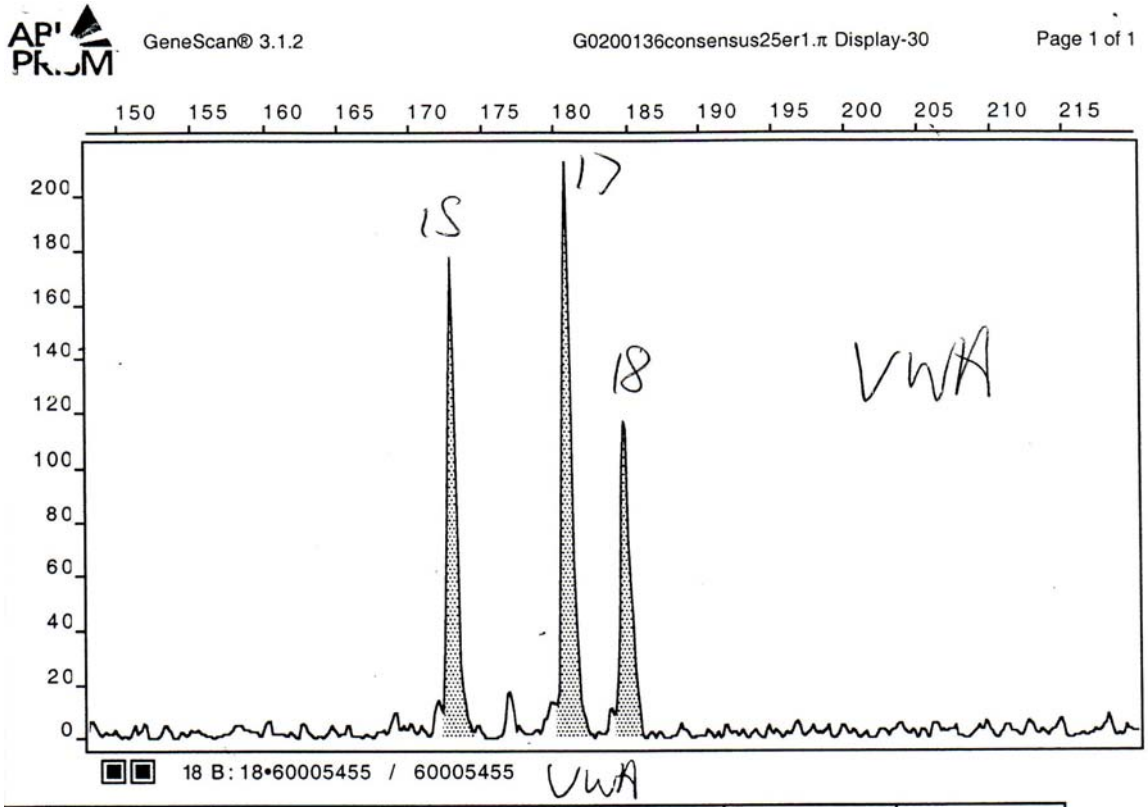
1 Introduction

Recent work has demonstrated the potential of using probabilistic expert systems (PES) for evaluating DNA evidence (Dawid *et al.* 2002). This article is concerned with the analysis of *mixed traces* where several individuals may have contributed to a DNA sample left on a scene of crime. Mortera *et al.* (2003) showed how to construct a PES using information about which alleles were present in the mixture, and we refer to this article for a general description of the problem and for genetic background information.

However the results of a DNA analysis are usually represented as an *electropherogram* (EPG) measuring responses in *relative fluorescence units* (RFU) and the alleles in the mixture correspond to peaks with a given height and area around each allele, see Figure 1. The band intensity around each allele in the relative fluorescence units represented, for example, through their *peak areas*, contains important information about the composition of the mixture. This information was ignored by Mortera *et al.* (2003) but exploited by e.g. Evett *et al.* (1998) who emphasized and discussed its rôle in evidential calculations. Perlin and Szabady (2001) and Wang *et al.* (2002) used numerical methods known as *Linear Mixture Analysis* (LMA) and *Least Square Deconvolution* (LSD) for separating mixture profiles using peak area information.

In this article we build a PES for mixture traces based on conditional Gaussian distributions for the peak areas, given the composition of the true DNA mixtures; see Chapter 7 of Cowell *et al.* (1999) as well as Lauritzen and Jensen (2001). Such a network enables us to perform evidential calculation and separation of DNA mixtures in a unified way, yielding a natural quantification of any possible uncertainty associated with the analysis.

For the sake of simplicity we only consider the case where the mixed trace contains DNA from exactly two contributors, but we emphasize that the flexibility and modularity of the PES approach enables extension and modification of our network to include complications such as an unknown number of contributors, indirect evidence, etc. along the lines given in Mortera *et al.* (2003). An analysis of a mixed trace can have different purposes, several of which can be relevant simultaneously, making a unified approach particularly suitable. However, for the sake of exposition we consider the issues separately. The first focus of our analysis will be that of *evidential calculation*, detailed in § 4. Here a *suspect* with known genotype is held and we want to determine the likelihood ratio for the hypothesis that the suspect has contributed to the mixture vs. the hypothesis that the contributor is a randomly chosen individual. We distinguish two cases: the other contributor could be a *victim* with a known genotype or a *contaminator* with an



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
18B, 6	55.39	172.83	15 181	1116	2077
18B, 7	57.33	180.75	17 219	1503	2150
18B, 8	58.32	184.76	18 120	869	2187

Figure 1: An electropherogram (EPG) of marker VWA from a mixture. Peaks represent alleles at 15, 17 and 18 and the areas and height of the peaks express the quantities of each. Since the peak around 17 is high, this indicates two alleles with repeat number 17. This image is supplied courtesy of LGC Limited, 2004.

unknown genotype, possibly without a direct relation to the crime. In § 5, we show how to use our network for *separation of profiles,* i.e. identifying the genotype of each of the possibly unknown contributors to the mixture, the evidential calculation playing a secondary role. Finally, in § 6 we will contrast our approach with existing alternatives and discuss perspectives for further developments.

2 Basic model assumptions

We assume the usual Mendelian genetic model for the allele composition of the mixture traces with gene frequencies of single STR alleles being those reported in Evett *et al.* (1998) and Butler *et al.* (2003) for U.S. Caucasians, the latter being used for data taken from Perlin and Szabady (2001)[1] and Wang *et al.* (2002). The peak area information is included in the analysis through the relative peak weight. The (absolute) *peak weight* w_a of an allele with *repeat number* a is defined by scaling the peak area with the repeat number as

$$w_a = a\alpha_a,$$

where α_a is the peak area around allele a. Multiplying the area with the repeat number is a crude way of correcting for the fact that alleles with a high repeat number tend to be less amplified than alleles with a low repeat number. The *relative peak weight* r_a is obtained by scaling with the total weight

$$r_a = w_a/w_+, \quad w_+ = \sum_a w_a.$$

Our simple model for the observed peak weight, R_a, uses a normal distribution

$$R_a \sim \mathcal{N}(\mu_a, \tau^2), \quad \mu_a = \{\theta n_a^1 + (1-\theta)n_a^2\}/2, \tag{1}$$

where θ is the fraction of DNA in the mixture originating from the first contributor, n_a^i is the number of alleles with repeat number a possessed by person i, and $\tau^2 = \sigma^2\mu_a(1-\mu_a)+\omega^2$, where σ^2 and ω^2 are variance factors for the contributions to the variation from the amplification and measurement processes. Unless stated otherwise, we have used $\sigma^2 = 0.01$ and $\omega^2 = 0.001$, corresponding approximately to a standard deviation for the observed relative weight of about

$$\sqrt{0.01/4 + 0.001} = 0.06$$

[1]This dataset has an observed allele 25.2 of the marker FGA. As none of the 302 subjects in Butler *et al.* (2003) had this allele, we chose to use 1/604=0.00166 as the allele frequency.

for $\mu_a = 0.5$. This is about the variability in peak imbalance that has been reported in the literature when amplifying DNA from one heterozygous individual, for which $\mu_a = 0.5$.

Our model ignores the obvious correlation between weights due to the fact that they must add up to unity. The model can be seen as a second order approximation to a more sophisticated model based on gamma distributions for the absolute scaled peak weights to be discussed elsewhere. The simple model above seems in any case sufficiently accurate and adequate for the purposes of the present paper. In general the variance factors may depend on the marker and on the amount of DNA analysed, but here we use the values above.

3 Bayesian networks for DNA mixtures with peak weights

3.1 Object-Oriented Networks

Object-oriented Bayesian networks (Koller and Pfeffer 1997; Dawid 2003) have a hierarchical structure where any node itself can represent a (object-oriented) network containing several *instances* of other generic *classes* of networks. This framework is particularly suited for an application area such as the present because we can exploit the similarity between elements of the networks in a modular and flexible construction, making the networks more and more complex by simply adding new objects which perform different tasks.

Instances have interface *input* and *output* nodes as well as ordinary nodes. Instances of a particular class have identical conditional probability tables for non-input nodes. Instances are connected by arrows from output nodes to input nodes. These arrows represent identity links whereas arrows between ordinary nodes represent probabilistic dependence (Cowell *et al.* 1999). Implementation of object-oriented Bayesian networks is supported by the program HUGIN 6[2].

3.2 Description of the network classes

Below we describe the different objects used in our construction of the network. In what follows, **bold** will indicate a network class, and `teletype` will indicate a node. In the figures, instances of a certain class are represented by

[2]See `www.hugin.com`

a rounded rectangle, discrete nodes have a single outline, whereas continuous nodes have a double outline. Interface nodes are represented with a grey ring; input nodes having a dotted outline and output nodes having a solid outline. Also, dark grey nodes will indicate where possible evidence might be inserted and black nodes are target nodes or nodes of interest where results will be read.

3.2.1 The founder class

The network class **founder** of Figure 2 contains a single node `founder` with the relevant repertory of alleles as its states, and an associated probability table describing their frequencies.

For illustration, we show marker FGA having observed alleles coded A to C and the aggregation of all unobserved alleles coded as x. The probability table is shown in Table 1.

Table 1: Gene frequencies for marker FGA as reported in Evett *et al.* (1998).

Allele	A	B	C	x
Frequency	0.187	0.165	0.139	0.509

3.2.2 The genotype class

The network class **gt** in Figure 3 represents an individual's genotype `gt`, which is given by the unordered pair of paternal and maternal genes, {`pg`, `mg`}. Input nodes `pg` and `mg` are copies of node `founder` of class **founder**. The paternal and maternal genes, `pg` and `mg`, are chosen independently from the same population whose allele frequencies are assumed known. Output node `gt` is the logical combination of `pg` and `mg`.

3.2.3 The query class

The network class **whichgt** of Figure 4 describes the selection between two genotypes.

If the Boolean node `query?` has the value *true*, then the output node, `outgt`, will have identical genotype to the input node, `ingt`; otherwise it will be identical to input node `othergt`. This is written in the HUGIN expression language as: `outgt` := if(`query?` == *true*, `ingt`, `othergt`).[3]

[3]The function if(A, x, y) takes the value x if condition A is satisfied, otherwise y.

Figure 2: Network **founder** for founder gene

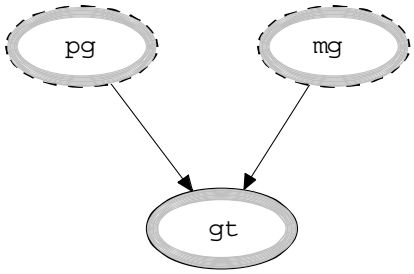


Figure 3: Network **gt** for genotype

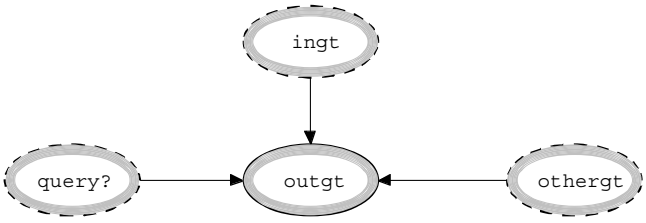


Figure 4: Network **whichgt** for selecting a genotype

3.2.4 The joint genotype

The network class **jointgt** of Figure 5 represents the combined genotype of two individuals, p1 and p2. Node `p1gt&p2gt` is simply the logical combination of the two input genotypes in `p1gt` and `p2gt`.

3.2.5 The number of alleles

The network class **nalleles** shown in Figure 6 counts the number, varying from 0 to 2, of a certain allelic type in a genotype. Thus, for allele A say, `nA` := if(`gt` == AA, 2, if (or (`gt` == AB, `gt` == AC, `gt` == Ax), 1, 0)).

3.2.6 The weight of an allele in the mixture

The network class **alleleinmix** shown in Figure 7 computes whether a certain allelic type is in the mixture and its mean contribution to the peak area of the mixture.

Input nodes `p1gt` and `p2gt`, the genotypes of the two people, p1 and p2, contributing to the mixture, have identity links to the input node `gt` in the two instances of class **nalleles**, `n1A` and `n2A`. The Boolean node `Ainmix?` is *true* if at least one of the two contributors has allele A. Thus, `Ainmix?` := if(and (`n1A_nA` == 0, `n2A_nA` == 0), *false*, *true*), where `n1A_nA` and `n2A_nA` refer to the output nodes of the two instances of class **nalleles**, `n1A` and `n2A`.

Input node `frac` represents the fraction of DNA contributed by p1, denoted by θ in § 2. The states of node `frac` are on a discrete scale ranging from $[0, 5]$ for convenience. The scale of node `frac` can easily be modified to a finer grid. Output node `meanA`:= `n1A_nA` * frac + `n2A_nA` * (5 - frac). This differs from the expression for the mean in (1) by a scale factor of 10 which is appropriately corrected for throughout.

3.2.7 The peak weight

The network class **peakweight** shown in Figure 8 represents the observable peak weight.

The input node `mean` is identified, for example, with output node `meanA` of class **alleleinmix**. The intermediate continuous node `weight`, with discrete parent `mean`, has a mean given by the values of node `mean` and variance equal to $0.01 * \texttt{mean} * (10 - \texttt{mean})$, representing variations in the amplification process. The observed peak weight is modelled by the continuous node `weightobs` to allow for additional measurement error of the true weight, in node `weight`. Thus this network class is representing the Gaussian part of the model (1).

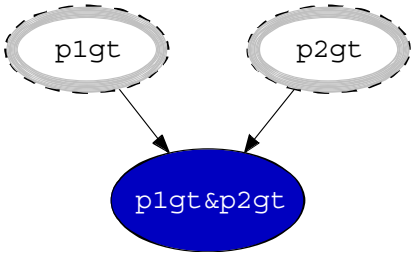


Figure 5: Network **jointgt** for genotype pairs

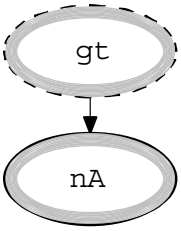


Figure 6: Network **nalleles** for counting the number of alleles

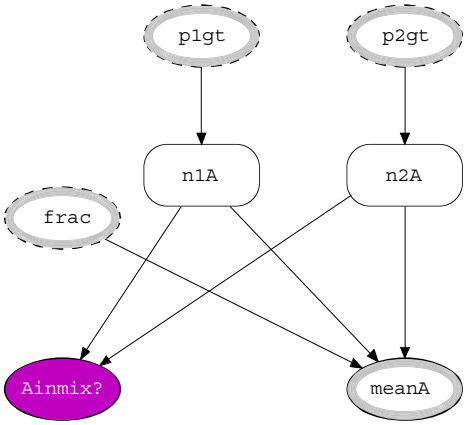


Figure 7: Network **alleleinmix** for alleles in mixture

3.2.8 The target class

The network class **target** shown in Figure 9 has two Boolean output nodes `p1=s?` (`p2=v?`) with *true*, *false* states, representing whether contributor p1 (p2) is the *suspect* (*victim*) or not. The black `target` node is the logical conjunction of the nodes `p1=s?` and `p2=v?`. These nodes are given a uniform prior distribution so that `target` node has a uniform prior distribution over its states. This enables the computation of the likelihood ratio as described in Mortera *et al.* (2003).

3.2.9 The marker class

The network class **marker** in Figure 10 is an upper level network containing several instances of the classes defined above. This class is made to represent information related to a particular marker. Here it is illustrated for a marker having three alleles represented in the mixture. Input nodes `spg`, `smg`, `u1pg`, `u1mg`, `vpg`, `vmg`, `u2pg` and `u2mg` are all copies of node `founder` of class **founder**; u1 and u2 are two unspecified individuals. Input nodes `p1=s?` and `p2=v?` are identified with the corresponding output nodes of class **target**. The nodes `sgt`, `u1gt`, `vgt` and `u2gt` are all instances of class **gt**; `p1gt` and `p2gt` are instances of **whichgt** and when `p1=s?` is *true* (*false*), `p1gt` will be identical to `sgt` (`u1gt`). A similar relationship holds between nodes `p2=v?`, `p2gt`, `vgt` and `u2gt`. The node `jointgt` is an instance of **jointgt**; `Amean`, `Bmean`, `Cmean` and `xmean` are instances of class **alleleinmix**; `Aweight`, `Bweight`, `Cweight` and `xweight` are instances of class **peakweight**. Input node `frac` is identified with the corresponding node in the master network described below.

3.2.10 The master network

Figure 11 gives the master network used for both identification and separation of DNA mixtures from two contributors. It refers to the data from Evett *et al.* (1998) shown in Table 2.

`D8`, `D18`, `FGA`, and `THO1` are all instances of network class **marker**; `D21` and `VWA` are instances of a simple modification of network class `marker` and the other network classes it calls, in order to account for 4 observed alleles. `D8`, `D18`, `FGA`, `THO1`, `D21` and `VWA` each have 8 `founder` instances with their appropriate gene frequencies as input to the 8 input nodes of class `marker`. The `frac` node is connected to all the markers showing their dependence via this quantity. `target`, an instance of class **target** is connected to each marker via its output nodes `p1=s?` and `p2=v?`. Once constructed, the master network can be used to insert and propagate case evidence in the appropriate

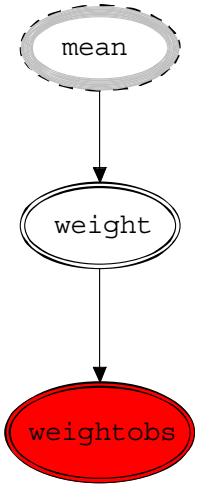


Figure 8: Network **peakweight** for peak weight

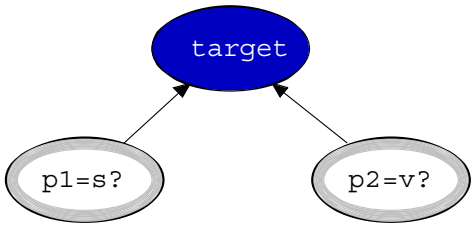


Figure 9: Network **target**

Table 2: *Evett* data showing mixture composition, peak areas and relative weights from a 10:1 mixture of two individuals, with suspect's genotype specified.

Marker	Alleles	Peak area	Relative weight	Suspect
D8	10	6416	0.4347	10
	11	383	0.0285	
	14	5659	0.5368	14
D18	13	38985	0.8871	13
	16	1914	0.0536	
	17	1991	0.0592	
D21	59	1226	0.0525	
	65	1434	0.0676	
	67	8816	0.4284	67
	70	8894	0.4515	70
FGA	21	16099	0.5699	21
	22	10538	0.3908	22
	23	1014	0.0393	
THO1	8	17441	0.4015	8
	9.3	22368	0.5985	9.3
VWA	16	4669	0.4170	16
	17	931	0.0884	
	18	4724	0.4747	18
	19	188	0.0199	

internal nodes, and the marginal posterior probabilities of the quantities of interest can be read from the corresponding nodes.

3.2.11 Amelogenin marker

In the analysis of DNA mixtures the determination of the sex of the two contributors, based on the *amelogenin* marker, is extremely important. To build a network for amelogenin one needs to make the following changes to the previous classes. No founder class is needed and the genotype class has a single output node `gt` with states XX for female and XY for male, with equal prior probabilities. The **query** and **jointgt** classes only need trivial modifications to reduce their state spaces. The allele counting class **nalleles**, for a male contributor, `gt`== XY, (for a female contributor, `gt`== XX) has `nX`==1 (2) and `nY`== 1 (0). The network class that gives the weight of an allele in the mixture is similar to **alleleinmix** of § 3.2.6, except `Xinmix?` is always set to *true*. All other network classes remain unchanged.

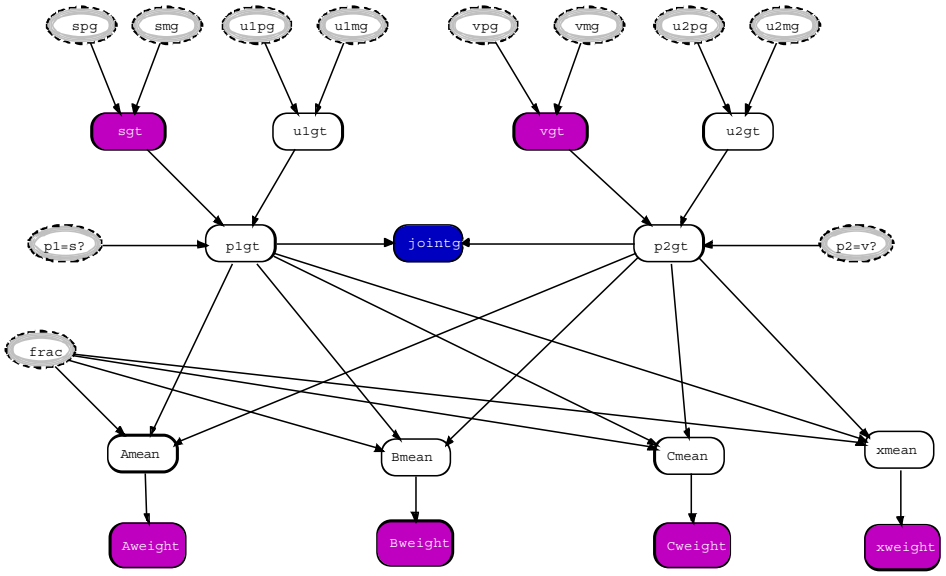


Figure 10: Network **marker**

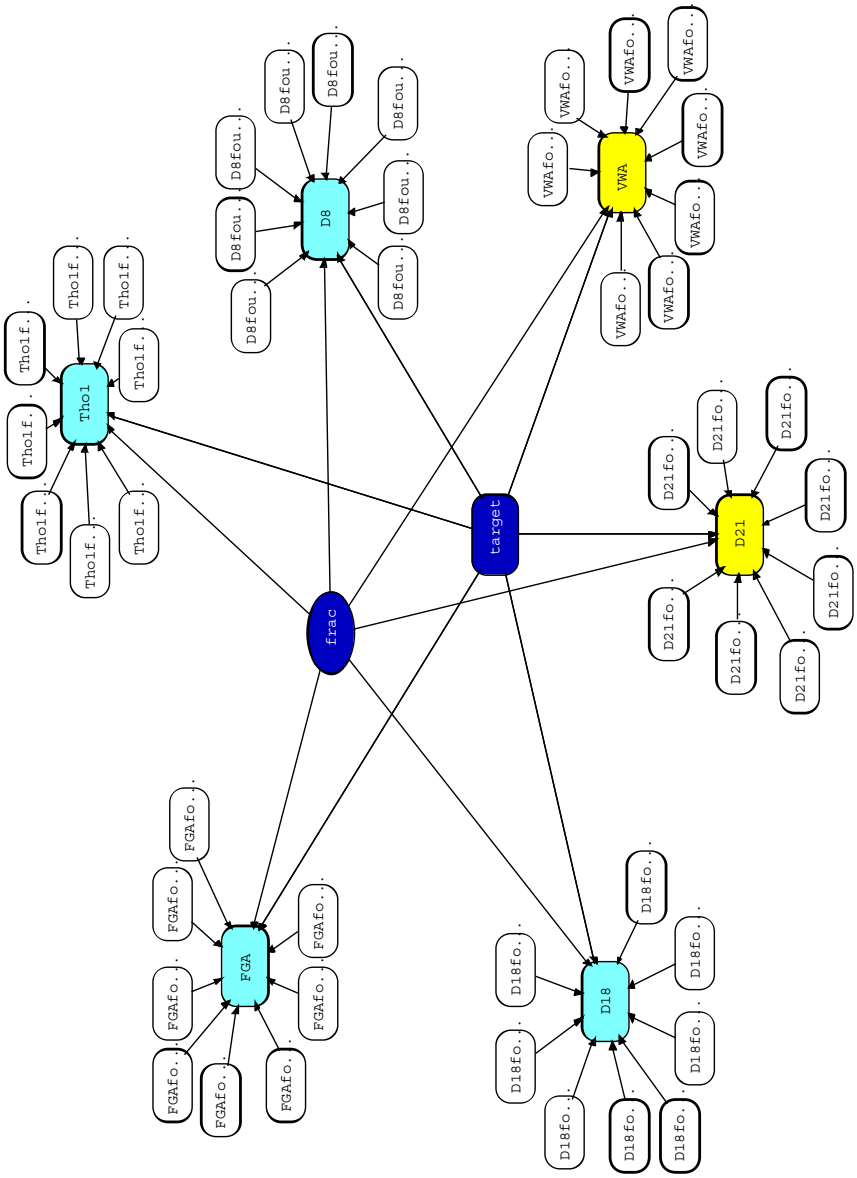


Figure 11: Master network for identification and separation of mixtures

3.3 MAIES: An expert system for analysing mixed traces

In parallel to the development of the object-oriented networks described in § 3.2 an alternative computer program MAIES— Mixture Analysis In Expert Systems — was developed to provide an independent check of the calculations as well as a flexible environment for specification and input or output of data to allow for experimentation with different methods.

In MAIES, information concerning allele frequencies of markers, and genotypic information (if available) about the potential contributors, is entered via simple popup dialog boxes activated by menu selections. The main program interface displays three user-selectable tabbed windows. One window is used for entering the peak area measurements obtained from the EPG. A second tabbed window displays the graphical model generated on the basis of the marker data, which can be used for entering or retracting evidence on individual nodes (if desired) and for showing the marginal distributions on nodes as bar charts or density curves. The third window is a simple text area that is used to display the marginal distributions of the nodes in the network - the contents of this window may be saved to a file, or cut-and-pasted to other applications, if desired.

After entering peak area information and available genetic profiles on people, the user constructs — by mouse clicking — a single Bayesian network on which the probability calculations are performed. In constructing the Bayesian network the user may specify the scale σ^2 of the amplification error variance, the measurement error variance ω^2, and the granularity in the discrete states of a node that represents the true fraction of DNA originating from individual 1 (again through a popup dialog box). Sensitivity analysis may be performed in a simple, straightforward manner by varying these three inputs. Peak areas are automatically converted to normalized weights by the program, and entered as evidence in the relevant nodes.

The user can temporarily retract or reinstate evidence on the two potential contributors to the mixture by use of menu selections, thus allowing an evidential calculation to be converted to one of deconvoluting a mixture arising from two unknown contributors, or vice versa.

The Bayesian network generated by MAIES may be considered equivalent to an "unfolded" version of the object-oriented networks described earlier § 3.2. An example of a network generated for a single marker with two alleles observed in the mixture is shown in Figure 12. The structure is similar to the network shown in Figure 10, and like the object-oriented network described earlier there are several distinct modules of repetition that can be seen in the figure: indeed it is this repetitive structure that makes it possible for MAIES

to create the much larger Bayesian networks required to analyse mixtures on several markers. We now describe these various structures and how they interrelate.

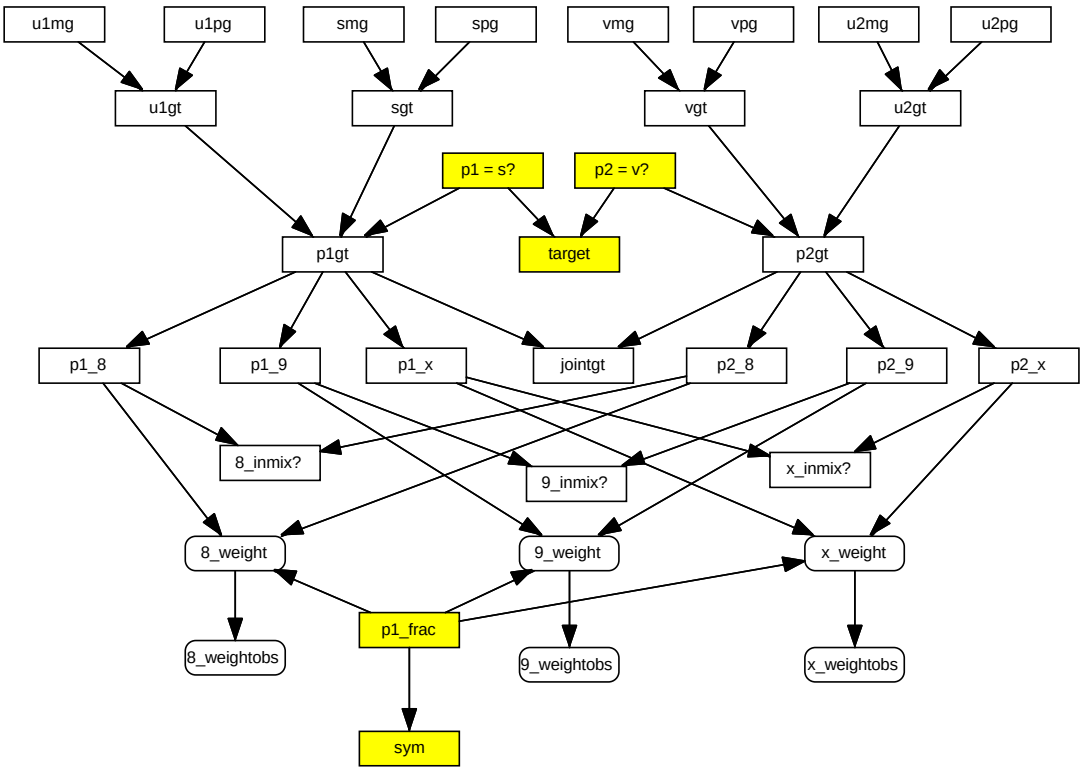


Figure 12: The structure of a Bayesian network generated by MAIES for a single marker, in which two allele peaks (8 and 9) were observed.

3.3.1 Founding people

MAIES currently assumes that DNA from two individuals are in the mixture. Thus it sets up nodes for four founding individuals who are paired up, prefixed by `s` (for suspect), `v` (for victim), and `u1` and `u2` representing two unspecified persons from the population. Corresponding to each of these individuals is a triple of nodes representing their genotype on the marker, and the individuals' paternal and maternal genes. They are joined up as in Figure 3 and their function is the same. The probability tables associated with the maternal and paternal genes contain the allele frequencies of the observed alleles, whilst

the conditional probability table associated with the genotype node is the logical combination of the maternal and paternal gene.

3.3.2 Actual contributors to the mixture

The genotypes on the marker of the two individuals p1 and p2 whose DNA is in the mixture are the nodes labelled `p1gt` and `p2gt`. Node `p1gt` has incoming arrows from nodes `u1gt`, `sgt` and a `(yes,no)` valued binary node labelled `p1 = s?`. The function of this latter node is similar to the `query?` node of Figure 4, namely to set the genotype of node `p1gt` to be that of `sgt` if `p1 = s?` takes the value `yes`, otherwise set the genotype of node `p1gt` to be that of `u1gt`. An equivalent relationship holds between the genotype nodes `p2gt`, `vgt`, `u2gt` and `p2 = v?`.

The node labelled `target` represents the four possible combinations of values of the two nodes `p1 = s?` and `p2 = v?` as in Figure 9 and described in § 3.2.8.

The network also has a node representing the joint genotypes of individuals p1 and p2, which is labelled `jointgt`, with incoming arrows from `p1gt` and `p2gt`; the function of this part of the network is equivalent to the object shown in Figure 5.

3.3.3 Allele counting nodes

On the level below the genotype nodes for p1 and p2 is a set of nodes representing the number of alleles (taking the value of 0, 1 or 2) of a certain type in each individual. Thus, for example, the node `p1_8` counts the number of alleles of repeat number **8** in the genotype of individual p1 for the given marker: this value only depends upon the genotype of the individual p1 and hence there is an arrow from `p1gt` to `p1_8`. These nodes model the n_a^i variables introduced in (1).

3.3.4 Repeat number nodes

On the level below the allele counting nodes are the repeat number nodes, labelled `8_inmix?`, `9_inmix?` and `x_inmix?`. These are `(yes,no)` binary valued nodes representing whether or not the particular alleles are present in the mixture: thus for example allele 8 is present in the mixture if either of the allele counting nodes `p1_8` or `p2_8` takes a non-zero value. For the node `x_inmix?` the **x** refers to all of the alleles in the marker that are not observed. When using repeat number information as evidence in calculations, this will be given the value `no`, the other repeat number nodes will be given the value `yes`.

3.3.5 True and observed weight nodes

These nodes are represented by the rounded rectangle shapes. The nodes `8_weight`, `9_weight` and `x_weight` represent the true relative peak weights r_8, r_9 and r_x respectively of the alleles 8, 9 and `x` in the amplified DNA sample; the nodes `8_weightobs`, `9_weightobs` and `x_weightobs` represent the measured weights. The observed weight is given a conditional-Gaussian distribution with mean the true weight, and variance ω^2. Each true-weight node is given a conditional-Gaussian distribution with mean $\mu_a = \{\theta n_a^1 + (1-\theta)n_a^2\}/2$, where the fraction θ of DNA from p1 in the mixture is modelled in the network by a discrete distribution in the node labelled `p1_frac`. The variance is taken to be $\sigma^2\mu_a(1-\mu_a)$, as specified in § 2. The `sym` node is only used for separating a mixture of two unknown contributors, and will be described later in § 5.2.

3.3.6 Networks with more than one marker

The network displayed in Figure 12 generated by MAIES is for a single marker; for mixture problems involving several markers the structure is similar but more complex because the number of nodes grows with the number of markers (in the *Graham* example, see § 4.1 below, there are 325 nodes). In such a network the nodes shaded in Figure 12 occur only once. The unshaded nodes are replicated once for each marker, with each node having text in their labels to identify the marker that the allele or genotype nodes refer to. There will also be extra repeat number, allele counting and allele weight nodes in each marker having more than two observed alleles in the mixture, extending the pattern for the one-marker network in the obvious manner.

4 Evidence calculations

All evidence calculations reported have been made using MAIES, combined with strategic independent checks using HUGIN. The variable describing the mixture ratio has been discretized to having 101 states $0, 0.01, 0.02, \ldots, 0.99, 1$, but experiments indicate very low sensitivity to the discretization as long as it is not far too rough and 10-20 states would probably be fully appropriate.

4.1 Genotype of suspect and victim available

This example is taken from Wang *et al.* (2002), stating P. Graham of the Texas Department of Public Safety as the data source. Table 3 displays the measured peak area, the relative weight, and DNA mixture composition on

9 markers, together with the genotypes of two potential contributors, here named suspect and victim. We will in the following refer to this data as the *Graham* data.

The evidence in this table is now entered into the network and the information propagated. Taking appropriate ratios in the posterior probabilities associated with the target node yields the likelihood ratio in favour of the hypothesis that the victim and suspect vs. that of the victim and a random individual being the contributors to the mixture. Table 4, column "Areas" displays the logarithm of this likelihood ratio, and column "Alleles" the corresponding ratio when only the evidence on the repeat number of the alleles is used. The last columns show the log-likelihood ratio when the mixture ratio θ is assumed known at given values.

The posterior distribution of the mixture ratio θ is displayed in Figure 13. Note that the likelihood ratio is essentially constant in the region $0.3 < \theta < 0.4$ which is the plausible region in the light of the data. Note also that this posterior distribution has its maximum around 0.34, close to the value reported in Wang *et al.* (2002).

The inclusion of area information is indeed strengthening the evidence against the suspect, increasing the logarithm of the likelihood ratio from 12.93 to 14.48, approximately corresponding to a factor 36. This is a modest increase and reflects the fact that when information about the genotype of the victim is available, peak area does not make much difference to the likelihood ratio as genotypes themselves are very informative.

4.2 Only genotype of suspect available

Our next example is taken from Evett *et al.* (1998) and has only information of the genotype from one potential contributor, here named the *suspect*, whereas the other contributor is a *contaminator*. The data is displayed in Table 2 and is henceforth referred to as the *Evett* data. Table 5 displays the logarithm of this likelihood ratio together with the corresponding ratio when peak weights are ignored, and the ratios when the mixture ratio θ is assumed known at given values.

Note that the strengthening of evidence against the suspect is more dramatic when information on the contaminator is absent: the logarithm of the likelihood ratio changes from 4.4 to 8.23, corresponding to an additional factor around 6000, as compared to a factor 36 above.

Also here the likelihood ratio is essentially constant over a region which completely covers the plausible $0.85 < \theta < 0.95$. The posterior distribution of the mixture ratio θ is displayed in Figure 14. The maximum occurs around the value 0.89 which is a little off the true 10:1 mixture proportion.

Table 3: *Graham* data showing mixture composition, peak areas, relative weights, suspect's and victim's profiles.

Marker	Alleles	Peak area	Relative weight	Suspect	Victim
D3	15	1242	0.3361		15
	16	657	0.1897	16	
	17	1546	0.4742	17	17
D5	7	486	0.0999	7	
	12	512	0.1804	12	
	13	1886	0.7198		13
D7	10	614	0.3232	10	
	11	1169	0.6768		11
D8	12	1842	0.6166		12
	13	490	0.1777	13	
	16	461	0.2057	16	
D13	8	734	0.3128		8
	9	1068	0.5120	9	9
	11	299	0.1752	11	
D18	12	440	0.1724	12	
	13	1503	0.6380		13
	15	387	0.1896	15	
D21	30	842	0.3087		30
	30.2	490	0.1808	30.2	
	31.2	509	0.1941	31.2	
	32.2	804	0.3164		32.2
FGA	22	850	0.3483		22
	23	468	0.2005	23	
	24	681	0.3045		24
	25	315	0.1467	25	
VWA	16	616	0.1900	16	
	17	2021	0.6625		17
	18	425	0.1475	18	

Table 4: Logarithm of the likelihood ratios of s&v vs. u&v for the *Graham* data.

	Areas	Alleles	Assumed known mixture ratio								
θ			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Log_{10}LR	14.48	12.93	11.07	13.65	14.48	14.47	11.10	4.5	-5.51	-22.84	-59.52

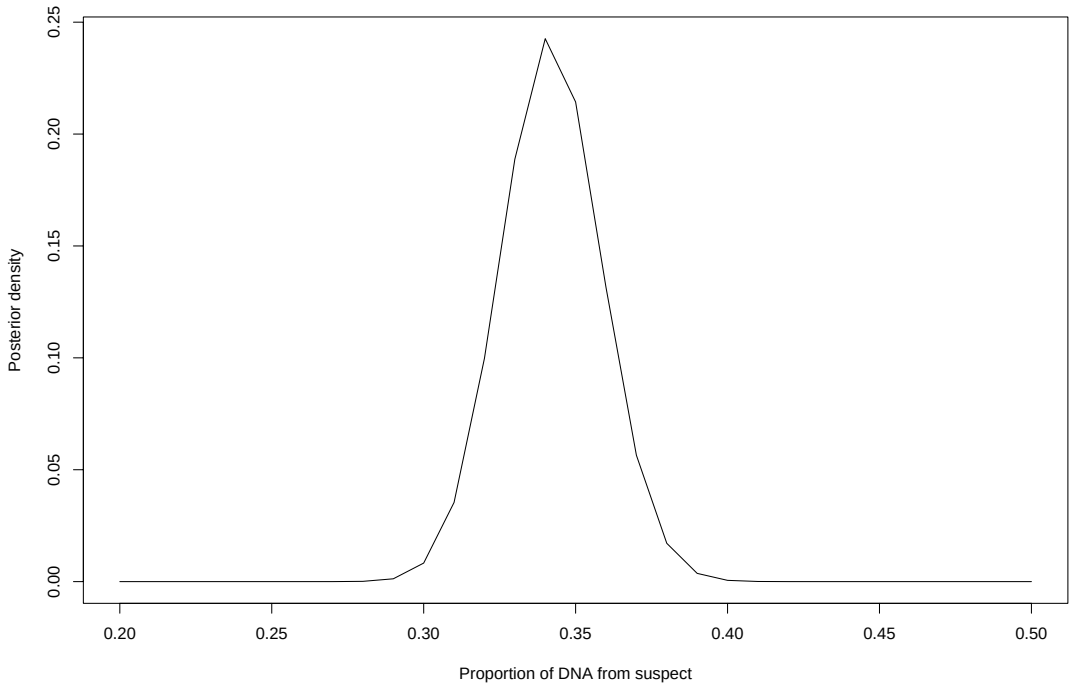


Figure 13: Posterior distribution of the mixture ratio for the *Graham* data using both the suspect's and the victim's genotype.

Table 5: Logarithm of the likelihood ratios of s&u vs. 2u for the *Evett* data.

	Areas	Alleles	Assumed known mixture ratio								
θ			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Log_{10}LR	8.23	4.40	-272.78	-167.52	-97.42	-41.88	5.07	8.03	8.53	8.53	8.53

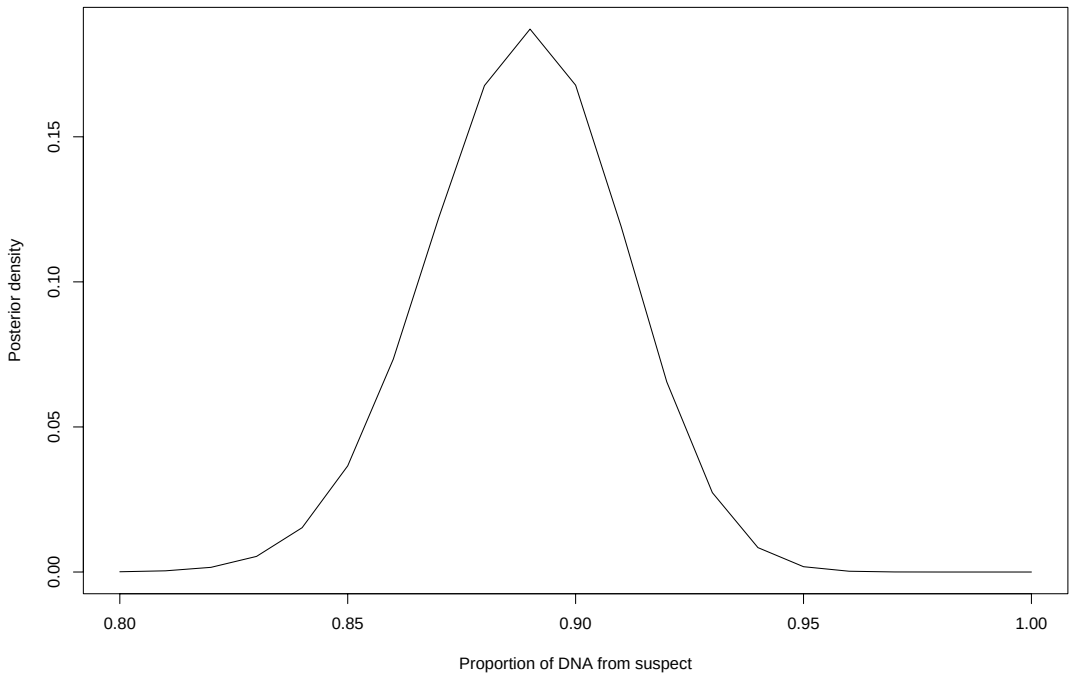


Figure 14: Posterior distribution of the mixture ratio for the *Evett* data using the suspect's genotype.

The absolute value of the likelihood ratios are slightly different from those given by Evett *et al.* (1998), who report a logarithm of the likelihood ratio of 7.3. This discrepancy is most likely due to slight differences between our model and the model used by Evett *et al.* (1998). On the other hand, they report a likelihood ratio based on allele presence alone of 5800, whereas we find a ratio around 25000, and this appears to be somewhat strange, as we have used the gene frequencies reported in their paper.

5 Separation of mixtures

Deconvolution of mixtures or separating a mixed DNA profile into its components has been studied by Perlin and Szabady (2001) and Wang *et al.* (2002), among others. A mixed DNA profile has been collected and the genotypes of one or more unknown individuals who have contributed to the mixture is desired, for example with the purpose of searching for a potential perpetrator among an existing database of DNA profiles.

The easiest case to consider is clearly that of separation of a single unknown profile, i.e. when the genotype of one of the contributors to the mixture is known. The case when both contributors are unknown is more difficult. In the latter situation this is only possible to a reasonable accuracy when the contributions to the DNA mixture has taken place in very different proportions.

We have chosen to show two alternative methods for predicting the genotype of the unknown contributor(s). In the first method we report the most probable genotype (or pair of genotypes) of the unknown contributor(s) for each marker separately. This result is obtained directly from the standard propagation method in the probabilistic expert system, known as sum-propagation. Note that this genotype is not necessarily the jointly most probable across markers. We therefore also report the joint probability of the genotypes chosen in this way. If this happens to be larger than 0.5, the most probable genotype has clearly been identified.

The second method calculates, by a method termed *semimax*-propagation, the most likely joint configuration of all unobserved discrete nodes, given the evidence available, and reports the genotypes of the unknown contributor(s) associated with this configuration. The semimax propagation first integrates over all unobserved continuous variables and then performs max-propagation as described in Cowell *et al.* (1999), Section 6.4.1, to identify the most probable configuration. Note again that this may not be the most probable genotype across markers. There is no general efficient method for calculating the latter, but identifying the two configurations above and reporting their joint

probabilities would be fully satisfactory for most purposes as they are most interesting when their joint probability is high.

The two methods generally give results that agree quite closely, the difference largely being due to correlations between the markers originating from the fact the fraction of DNA supplied by each contributor is unknown. When this fraction is well determined by the evidence, the markers are close to being independent. In such cases the two methods tend to given identical results. It then also holds that the joint posterior probability of the genotypes of the unknown contributors is approximately equal to the product of those probabilities for each marker separately.

It would generally seem appropriate to report a list of probable genotypes for the unknown contributor(s), with their associated probabilities, but this would demand a slightly more sophisticated calculation and is beyond the scope of this particular paper.

5.1 Separating a single unknown profile

Our next example is using data from Perlin and Szabady (2001), henceforth referred to as the *Perlin* data, displayed in Table 6.

The two individuals contributing to the mixture are here named *suspect* and *victim* and Table 7 displays the predicted genotype of the suspect, using information from the victim alone.

As in Perlin and Szabady (2001) the genotype of the unknown contributor is essentially determined exactly and the posterior distribution of the mixture ratio concentrates around the true value of 0.7, as displayed in Figure 15. For comparison we have also made a similar calculation for the other two examples. The results are displayed in Table 8 and Table 9.

Here the situation for the *Graham* data is similar to the *Perlin* data: all markers are correctly identified, with probabilities very close to 1 in all cases. Analysis of the *Evett* data also yield probabilities close to 1 on all markers, but not so close as the for *Perlin* and *Graham* data. Evett *et al.* (1998) does not contain the genotype of the second contributor so we do not know whether there are classification errors for this example. Figure 14 and Figure 16 display the posterior distribution of the mixture ratio for these two cases.

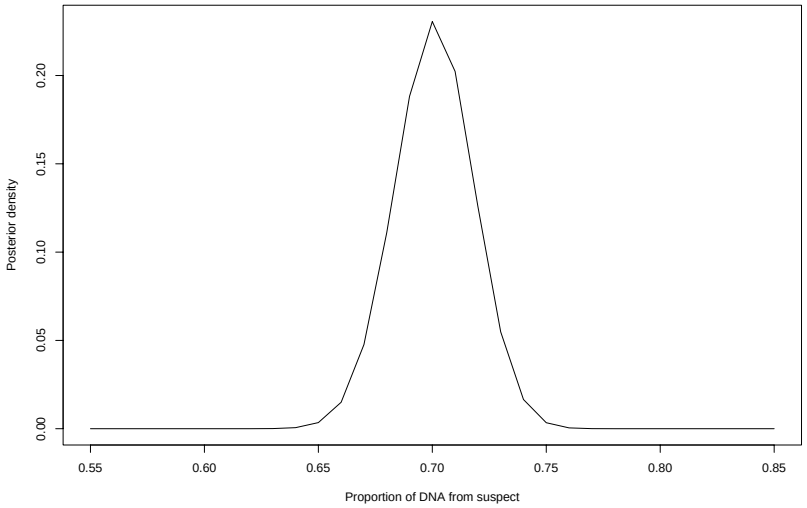


Figure 15: Posterior distribution of the mixture ratio for the *Perlin* data, using genotypic information on the victim only.

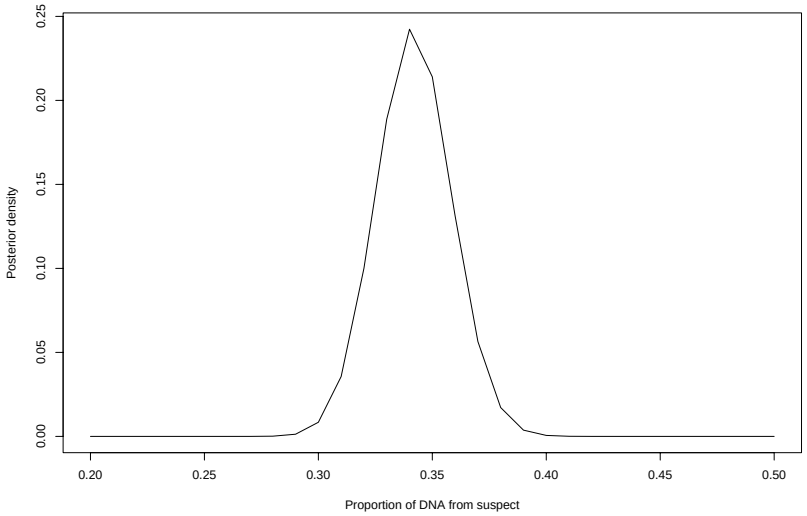


Figure 16: Posterior distribution of mixture ratio for *Graham* data using genotypic information from the victim only.

Table 6: *Perlin* data showing mixture composition, peak areas, relative weights, suspect's and victim's genotypes from a 7:3 mixture of the two individuals.

Marker	Alleles	Peak area	Relative Weight	Suspect	Victim
D2	16	0.3190	0.1339		16
	18	0.6339	0.2992	18	
	20	0.3713	0.1947		20
	21	0.6758	0.3722	21	
D3	14	1.0365	0.5010	14	14
	15	0.9635	0.4990	15	15
D8	9	0.7279	0.2832	9	
	12	0.2749	0.1426		12
	13	0.6813	0.3829	13	
	14	0.3160	0.1913		14
D16	11	1.4452	0.6801	11	
	13	0.2889	0.1607		13
	14	0.2660	0.1593		14
D18	12	0.3443	0.1504		12
	13	0.6952	0.3290	13	
	14	0.6755	0.3443	14	
	17	0.2850	0.1764		17
D19	12.2	0.6991	0.3109	12.2	
	14	0.6060	0.3092		14
	15	0.6949	0.3799	15	
D21	27	0.2787	0.1289		27
	29	0.7876	0.3913	29	
	30	0.9337	0.4798	30	30
FGA	19	1.0580	0.4621	19	19
	24	0.2830	0.1561		24
	25.2	0.6589	0.3817	25.2	
THO1	6	0.3178	0.1268		6
	7	1.0074	0.4691	7	7
	9	0.6749	0.4041	9	
VWA	17	1.4755	0.7265	17	
	18	0.5245	0.2735		18

Table 7: Predicted genotype of suspect for the *Perlin* data, using genotype information for victim only. All markers are correctly identified by both sum and semi-max propagation.

Marker	Genotype	Probability
D2	18 21	1
D3	14 15	1
D8	9 13	1
D16	11 11	1
D18	13 14	1
D19	12.2 15	1
D21	29 30	1
FGA	19 25.2	1
TH01	7 9	1
VWA	17 17	1

Table 8: Predicted genotype of suspect for *Graham* data, using genotype for victim only. All markers are correctly identified by both sum and semi-max propagation.

Marker	Genotype	Probability
D3	16 17	0.997638
D5	7 12	1
D7	10 10	0.999464
D8	13 16	1
D13	9 11	0.999607
D18	12 15	1
D21	30.2 31.2	1
FGA	23 25	1
VWA	16 18	1

Table 9: Predicted genotype of contaminator for *Evett* data, using information from suspect. Identical results are obtained using sum and semi-max propagation. The number in brackets is the product of individual marker probabilities.

Marker	Genotype	Probability
D8	11 14	0.904607
D18	16 17	1
D21	59 65	1
FGA	21 23	0.922646
TH01	9.3 9.3	0.926062
VWA	17 19	1
joint	0.7757173	(0.772921)

5.2 Separating two unknown profiles

We now turn to the problem of separating a mixture into two components, using peak area and repeat number information but no information regarding the two contributors to the mixture. Using only this information will lead to an identifiability problem in assigning genotype combinations to each person, because of the symmetry between the individuals p1 and p2 in the network of Figure 12 or in the equivalent object-oriented network Figure 10.

To remove this problem it is sufficient to enter evidence that the proportion of DNA in the sample from individual p1 is at least one half of the total DNA in the sample. (The alternative, that individual p1 contributes at most half of the DNA to the mixture sample could as equally well be used to break the symmetry.) Using HUGIN this symmetry breaking may be achieved by entering likelihood evidence directly into the fraction node; in MAIES direct entering of likelihood evidence is not possible, so instead this is achieved by entering evidence on the `sym` node mentioned in § 3.3.5. The node `sym` has two possible states, $\theta \geq 0.5$ and $\theta \leq 0.5$. Selecting one state as evidence breaks the symmetry (the user does this via a menu selection).

Our first example uses the *Evett* data, ignoring the information on the suspect. The posterior distribution of the mixture ratio θ is displayed as the solid curve in Figure 17. The distribution is similar in shape to that in Figure 14, which uses the suspect genotype information. The broken curve in Figure 17 shows the posterior using the larger variance factor $\sigma^2 = 0.1$. We note that this change of variance has a notable effect on the posterior distribution of mixture ratio.

The predicted genotypes of the two contributors are shown in Table 10, with the suspect's profile being predicted correctly for both choices of variance even though the probability of the chosen genotype is strongly reduced.

Our next example uses the *Perlin* data. The posterior distribution for θ is shown as the solid curve in Figure 18, with the mode at 0.69 very close to the value reported of 0.7. The predicted genotypes of the two contributors is shown in Table 11, with all classifications correct. Still, the joint probability of the chosen genotype indicates that other plausible explanations are available, essentially due to uncertainty about the genotype for marker VWA.

Increasing σ^2 by a factor of 10 to $\sigma^2 = 0.1$ yields the posterior distribution shown by the broken line Figure 18. In this case the effect of choosing an inflated variance factor is dramatic, also yielding reduced genotype probabilites and several classification errors as shown in Table 12. Note also that here there is a large discrepancy between probability of the joint genotype and the product of the probabilities for each marker.

Similar behaviour occurs in our final example that uses the *Graham* data.

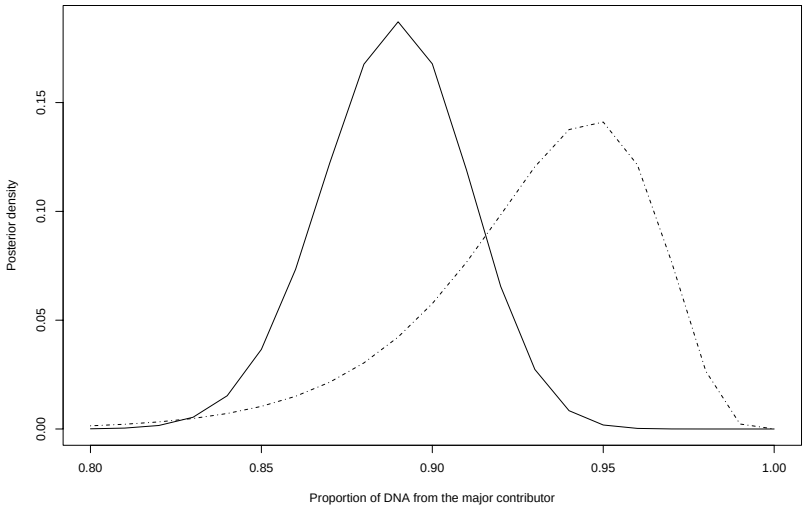


Figure 17: Posterior distribution of mixture ratio from *Evett* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

Table 10: Predicted genotypes of both contributors for *Evett* data with $\sigma^2 = 0.01$ and $\sigma^2 = 0.1$. Identical results are obtained using sum and semi-max propagation, with suspect (p1) correct on every marker. The number in brackets is the product of individual marker probabilities.

	$\sigma^2 = 0.01$			$\sigma^2 = 0.1$		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D8	10 14	11 14	0.904607	10 14	11 14	0.669688
D18	13 13	16 17	1	13 13	16 17	0.995388
D21	67 70	59 65	1	67 70	59 65	0.999967
FGA	21 22	21 23	0.922646	21 22	21 23	0.517756
TH01	8 9.3	9.3 9.3	0.926062	8 9.3	9.3 9.3	0.606364
VWA	16 18	17 19	1	16 18	17 19	0.999964
joint	0.7757173		(0.772921)	0.210368		(0.209264)

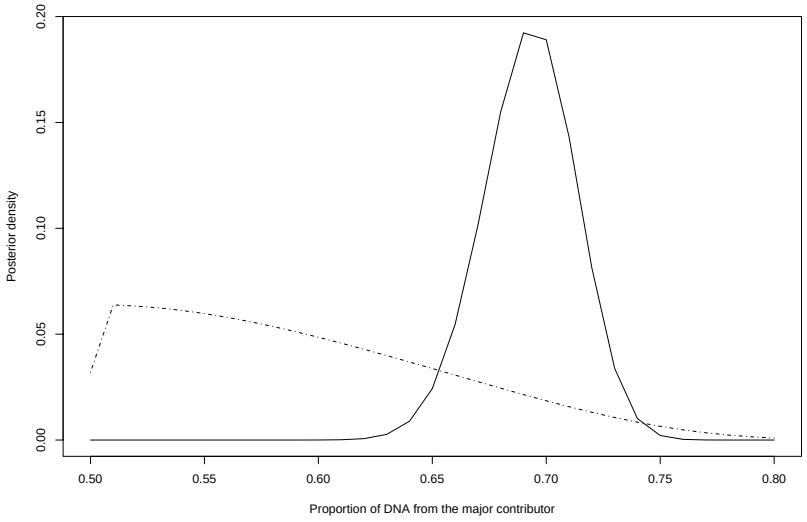


Figure 18: Posterior distribution of mixture ratio from *Perlin* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

Table 11: Predicted genotypes of both contributors for *Perlin* data with $\sigma^2 = 0.01$. The number in brackets is the product of individual marker probabilities. All classifications are correct but the marker VWA is uncertain.

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D2	18 21	16 20	0.999072	18 21	16 20	0.999072
D3	14 15	14 15	0.998342	14 15	14 15	0.998342
D8	9 13	12 14	0.997656	9 13	12 14	0.997656
D16	11 11	13 14	0.999805	11 11	13 14	0.999805
D18	13 14	12 17	0.999928	13 14	12 17	0.999928
D19	12.2 15	14 14	0.828718	12.2 15	14 14	0.828718
D21	29 30	27 30	0.987463	29 30	27 30	0.987463
FGA	19 25.2	19 24	0.990340	19 25.2	19 24	0.990340
TH01	7 9	6 7	0.969079	7 9	6 7	0.969079
VWA	17 17	18 18	0.537648	17 17	18 18	0.537648
joint	0.4342251		(0.420059)	0.4342251		(0.420059)

Table 12: Predicted genotypes of both contributors for *Perlin* data with $\sigma^2 = 0.1$. The number in brackets is the product of individual marker probabilities. There are classification errors in markers D3, FGA and VWA (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D2	18 21	16 20	0.349986	18 21	16 20	0.349986
D3	14 15	14 15	0.601826	14 15	*15 15*	0.149452
D8	9 13	12 14	0.339805	9 13	12 14	0.339805
D16	11 11	13 14	0.371007	11 11	13 14	0.371007
D18	13 14	12 17	0.361439	13 14	12 17	0.361439
D19	12.2 15	14 14	0.214513	12.2 15	14 14	0.214513
D21	29 30	27 30	0.438172	29 30	27 30	0.438172
FGA	19 25.2	19 24	0.444306	19 25.2	*24 24*	0.114614
TH01	7 9	6 7	0.413837	7 9	6 7	0.413837
VWA	*17 18*	*17 17*	0.432392	17 17	18 18	0.0819453
joint	0.000510		(7.1723e-05)	0.000107		(8.7074e-07)

The posterior distribution of θ is shown as the solid curve in Figure 19, with a maximum around 0.65; the predicted profiles are shown in Table 13, with one classification error. However note for this classification error (in D7, using sum-propagation) the probability assigned to the genotype pair is around 0.68, with the correct classification (picked out out by the semi-max method) has a probability of around 0.32. Note that the two chosen genotypes together account for essentially all of the probability mass.

Increasing the variance factor σ^2 to 0.1 yields more classification errors and also probabilities much lower, as shown in Table 14. The corresponding posterior distribution of θ is plotted as the broken line in Figure 19.

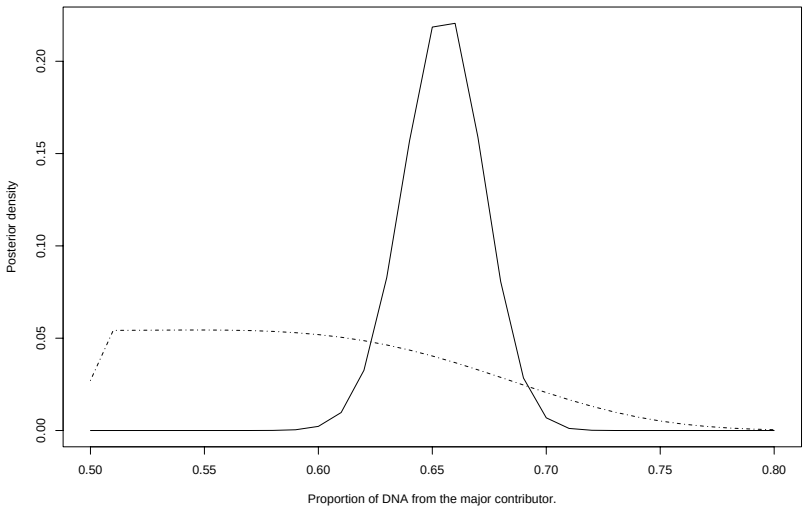


Figure 19: Posterior distribution of mixture ratio from *Graham* data using no genotypic information: solid line $\sigma^2 = 0.01$, broken line $\sigma^2 = 0.1$.

6 Discussion

In the previous sections we have demonstrated how a probabilistic expert system can be used for analysing DNA mixtures using peak area information, yielding a coherent way of predicting genotypes of unknown contributors and assessing evidence for particular individuals having contributed to the mixture. We emphasize the advantages of using a model-based approach as ours over numerical separation techniques such as Linear Mixture Analysis (LMA) (Perlin and Szabady 2001) and Least Square Deconvolution (LSD)

Table 13: Prediction of two unknown genotypes for *Graham* data, with $\sigma^2 = 0.01$. The number in brackets is the product of individual marker probabilities. There is a classification error in marker D7 (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D3	16 17	15 17	0.993063	16 17	15 17	0.993063
D5	7 12	13 13	0.999836	7 12	13 13	0.999836
D7	*11 11*	*10 11*	0.677898	10 10	11 11	0.321722
D8	13 16	12 12	0.952498	13 16	12 12	0.952498
D13	9 11	8 9	0.997447	9 11	8 9	0.997447
D18	12 15	13 13	0.990162	12 15	13 13	0.990162
D21	30.2 31.2	30 32.2	0.994523	30.2 31.2	30 32.2	0.994523
FGA	23 25	22 24	0.995630	23 25	22 24	0.995630
VWA	16 18	17 17	0.997152	16 18	17 17	0.997152
joint	0.6242217		(0.6251788)	0.2979028		(0.2967021)

Table 14: Prediction of two unknown genotypes for *Graham* data, using $\sigma^2 = 0.1$. There are now classification errors in three markers (italicized).

	sum prop			semi-max		
Marker	Genotype p1	Genotype p2	Probability	Genotype p1	Genotype p2	Probability
D3	16 17	15 17	0.339087	16 17	15 17	0.339087
D5	7 12	13 13	0.425175	7 12	13 13	0.425175
D7	*11 11*	*10 11*	0.375857	10 10	11 11	0.155175
D8	*12 13*	*12 16*	0.327032	13 16	12 12	0.279082
D13	9 11	8 9	0.317089	*11 11*	8 9	0.107251
D18	12 15	13 13	0.310149	12 15	13 13	0.310149
D21	30.2 31.2	30 32.2	0.309763	30.2 31.2	30 32.2	0.309763
FGA	23 25	22 24	0.340326	23 25	22 24	0.340326
VWA	16 18	17 17	0.350301	16 18	17 17	0.350301
joint	1.74778e-04		(6.4358e-05)	7.2491e-05		(7.6696e-06)

(Wang *et al.* 2002). Since the results are model based there is a natural and directly interpretable quantification of all uncertainties associated with the analysis, and the analysis is extendable to similar but different situations using the modularity and flexibility of the PES approach.

None of the data used as examples had information on the amelogenin marker, but this marker can clearly be very informative concering the unknown fraction of DNA for each contributor when these are of different sex, in particular when genotypes (including sex) is known for one of the contributors.

The examples considered have also demonstrated that there are issues which need further consideration. In particular it appears that the performance of the system is sensitive to the scaling factors used to model the variation in the amplification and measurement processes. This is a serious problem which needs attention. Preliminary investigations seem to indicate that this factor depends critically on the total *amount* of DNA available for analysis. As this necessarily is varying from case to case, a calibration study should be performed to take this properly into account. In any case we find it comforting that the system itself would warn against trusting an uncertain prediction, by yielding an associated low classification probability.

Methods for diagnostic checking and validation of the model should be developed based upon comparing observed weights to those predicted when genotypes are assumed correct. Such methods could also be useful for calibrating the variance parameters σ^2 and ω^2. To indicate a possible way ahead we note that the network can itself be used for predicting peak weight given a hypothesised composition of the mixture and of the two contributors. Table 15 gives the predicted peak weights for the *Perlin* data based on the repeat numbers in the mixture composition, the true mixture composition, and on the suspect's and victim's genotype. The last two columns show the limits of the 95% predictive interval $[\mu_a - 1.96\tau, \mu_a + 1.96\tau]$ for the weight. For a 95% predictive interval we might expect about one of the areas of the table to lie outside of its predicted interval, as 21 of the 31 intervals are independent (the weights for each marker must necessarily add to one); all expected areas are within their intervals, indicating that the variance at least is not too small.

The predicted peak weights are also useful for identifying measurement errors. For example, if the predicted weight is of the same order of magnitude as the cut-off threshold, the peak is likely to be missed.

Another issue to be further investigated is the possibility of using a model based on gamma distributed absolute peak weights, thereby treating the correlation between individual peak areas in a proper way and avoiding the somewhat unfortunate fact that Gaussian distributions can take negative

Table 15: Prediction of relative peak weight for *Perlin* data, using the mixture, the suspect's and the victim's DNA composition.

Marker	Allele	Relative Weight	Predicted relative weight	
			$\mu_a - 1.96\tau$	$\mu_a + 1.96\tau$
D2	16	0.1339	0.0565	0.2435
	18	0.2992	0.2378	0.4622
	20	0.1947	0.0565	0.2435
	21	0.3722	0.2378	0.4622
D3	14	0.5010	0.3840	0.6160
	15	0.4990	0.3840	0.6160
D8	9	0.2832	0.2378	0.4622
	12	0.1426	0.0565	0.2435
	13	0.3829	0.2378	0.4622
	14	0.1913	0.0565	0.2435
D16	11	0.6801	0.5909	0.8091
	13	0.1607	0.0565	0.2435
	14	0.1593	0.0565	0.2435
D18	12	0.1504	0.0565	0.2435
	13	0.3290	0.2378	0.4622
	14	0.3443	0.2378	0.4622
	17	0.1764	0.0565	0.2435
D19	12.2	0.3109	0.2378	0.4622
	14	0.3092	0.1909	0.4091
	15	0.3799	0.2378	0.4622
D21	27	0.1289	0.0565	0.2435
	29	0.3913	0.2378	0.4622
	30	0.4798	0.3840	0.6160
FGA	19	0.4621	0.3840	0.6160
	24	0.1561	0.0565	0.2435
	25.2	0.3817	0.2378	0.4622
THO1	6	0.1268	0.0565	0.2435
	7	0.4691	0.3840	0.6160
	9	0.4041	0.2378	0.4622
VWA	17	0.7265	0.5909	0.8091
	18	0.2735	0.1909	0.4091

values. Ideally the method should be generalised to deal with higher complexity such as the simultaneous analysis of several traces, an unknown but large number of contributors, etc., and we have not yet made a proper investigation of the computational complexity issues associated. Finally we emphasize that for the moment we have not dealt with incorporating artifacts such as stutter, pull-up, allelic dropout, etc., but we hope to pursue this and other aspects in the future.

Acknowledgement

This research was supported by a Research Interchange Grant from the Leverhulme Trust. We are indebted to participants in the above grant and to Sue Pope and Niels Morling for constructive discussions. We thank Caryn Saunders for supplying the EPG image used in Figure 1.

References

Butler, J. M., Schoske, R., Vallone, P. M., Redman, J. W., and Kline, M. C. (2003). Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American and Hispanic populations. *Journal of Forensic Sciences*, **48**, (4). Available online at www.astm.org.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Dawid, A. P. (2003). An object-oriented Bayesian network for estimating mutation rates. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Jan 3–6 2003, Key West, Florida*, (ed. C. M. Bishop and B. J. Frey).

Dawid, A. P., Mortera, J., Pascali, V. L., and van Boxel, D. W. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, 577–95.

Evett, I., Gill, P., and Lambert, J. (1998). Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences*, **43**, 62–9.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI)*, pp. 302–13.

Lauritzen, S. L. and Jensen, F. (2001). Stable local computation with conditional Gaussian distributions. *Statistical Computing*, **11**, 191–203.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.

Perlin, M. and Szabady, B. (2001). Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, **46**, 1372–8.

Wang, T., Xue, N., and Wickenheiser, R. (2002). Least square deconvolution (LSD): A new way of resolving STR/DNA mixture samples. Presentation at the 13th International Symposium on Human Identification, October 7–10, 2002, Phoenix, AZ.

Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND STATISTICS

Actuarial Research Papers since 2001

135. Renshaw A. E. and Haberman S. On the Forecasting of Mortality Reduction Factors. February 2001.
ISBN 1 901615 56 1

136. Haberman S., Butt Z. & Rickayzen B. D. Multiple State Models, Simulation and Insurer Insolvency. February 2001. 27 pages.
ISBN 1 901615 57 X

137. Khorasanee M.Z. A Cash-Flow Approach to Pension Funding. September 2001. 34 pages.
ISBN 1 901615 58 8

138. England P.D. Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". November 2001. 17 pages.
ISBN 1 901615 59 6

139. Verrall R.J. A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. November 2001. 10 pages.
ISBN 1 901615 62 6

140. Renshaw A.E. and Haberman. S. Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. January 2002. 38 pages.
ISBN 1 901615 63 4

141. Ballotta L. and Haberman S. Valuation of Guaranteed Annuity Conversion Options. January 2002. 25 pages.
ISBN 1 901615 64 2

142. Butt Z. and Haberman S. Application of Frailty-Based Mortality Models to Insurance Data. April 2002. 65 pages.
ISBN 1 901615 65 0

143. Gerrard R.J. and Glass C.A. Optimal Premium Pricing in Motor Insurance: A Discrete Approximation. (Will be available 2003).

144. Mayhew, L. The Neighbourhood Health Economy. A systematic approach to the examination of health and social risks at neighbourhood level. December 2002. 43 pages.
ISBN 1 901615 66 9

145. Ballotta L. and Haberman S. The Fair Valuation Problem of Guaranteed Annuity Options: The Stochastic Mortality Environment Case. January 2003. 25 pages.
ISBN 1 901615 67 7

146. Haberman S., Ballotta L. and Wang N. Modelling and Valuation of Guarantees in With-Profit and Unitised With-Profit Life Insurance Contracts. February 2003. 26 pages.
ISBN 1 901615 68 5

147. Ignatov Z.G., Kaishev V.K and Krachunov R.S. Optimal Retention Levels, Given the Joint Survival of Cedent and Reinsurer. March 2003. 36 pages.
ISBN 1 901615 69 3

148. Owadally M.I. Efficient Asset Valuation Methods for Pension Plans. March 2003. 20 pages.
ISBN 1 901615 70 7

149. Owadally M.I. Pension Funding and the Actuarial Assumption Concerning Investment Returns. March 2003. 32 pages.

ISBN 1 901615 71 5

150. Dimitrova D, Ignatov Z. and Kaishev V. Finite time Ruin Probabilities for Continuous Claims Severities. Will be available in August 2004.

151. Iyer S. Application of Stochastic Methods in the Valuation of Social Security Pension Schemes. August 2004. 40 pages.

ISBN 1 901615 72 3

152. Ballotta L., Haberman S. and Wang N. Guarantees in with-profit and Unitized with profit Life Insurance Contracts; Fair Valuation Problem in Presence of the Default Option[1]. October 2003. 28 pages.

ISBN 1-901615-73-1

153. Renshaw A. and Haberman. S. Lee-Carter Mortality Forecasting Incorporating Bivariate Time Series. December 2003. 33 pages.

ISBN 1-901615-75-8

154. Cowell R.G., Khuen Y.Y. and Verrall R.J. Modelling Operational Risk with Bayesian Networks. March 2004. 37 pages.

ISBN 1-901615-76-6

155. Gerrard R.G., Haberman S., Hojgaard B. and Vigna E. The Income Drawdown Option: Quadratic Loss. March 2004. 31 pages.

ISBN 1-901615-77-4

156. Karlsson, M., Mayhew L., Plumb R, and Rickayzen B.D**.** An International Comparison of Long-Term Care Arrangements. An Investigation into the Equity, Efficiency and sustainability of the Long-Term Care Systems in Germany, Japan, Sweden, the United Kingdom and the United States. April 2004. 131 pages.

ISBN 1 901615 78 2

157. Ballotta Laura. Alternative Framework for the Fair Valuation of Participating Life Insurance Contracts. June 2004. 33 pages.

ISBN 1-901615-79-0

158. Wang Nan. An Asset Allocation Strategy for a Risk Reserve considering both Risk and Profit. July 2004. 13 pages.

ISBN 1 901615-80-4

Statistical Research Papers

1. Sebastiani P. Some Results on the Derivatives of Matrix Functions. December 1995. 17 Pages.

ISBN 1 874 770 83 2

2. Dawid A.P. and Sebastiani P. Coherent Criteria for Optimal Experimental Design. March 1996. 35 Pages.

ISBN 1 874 770 86 7

3. Sebastiani P. and Wynn H.P. Maximum Entropy Sampling and Optimal Bayesian Experimental Design. March 1996. 22 Pages.

ISBN 1 874 770 87 5

4. Sebastiani P. and Settimi R. A Note on D-optimal Designs for a Logistic Regression Model. May 1996. 12 Pages.

ISBN 1 874 770 92 1

5. Sebastiani P. and Settimi R. First-order Optimal Designs for Non Linear Models. August 1996. 28 Pages.

ISBN 1 874 770 95 6

6. Newby M. A Business Process Approach to Maintenance: Measurement, Decision and Control. September 1996. 12 Pages.

ISBN 1 874 770 96 4

7. Newby M. Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. September 1996. 16 Pages.
ISBN 1 874 770 97 2

8. Cowell R.G. Mixture Reduction via Predictive Scores. November 1996. 17 Pages.
ISBN 1 874 770 98 0

9. Sebastiani P. and Ramoni M. Robust Parameter Learning in Bayesian Networks with Missing Data. March 1997. 9 Pages.
ISBN 1 901615 00 6

10. Newby M.J. and Coolen F.P.A. Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. March 1997. 9 Pages.
ISBN 1 901615 01 4.

11. Newby M.J. Approximations for the Absorption Distribution and its Negative Binomial Analogue. March 1997. 6 Pages.
ISBN 1 901615 02 2

12. Ramoni M. and Sebastiani P. The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. June 1997. 11 Pages.
ISBN 1 901615 10 3

13. Ramoni M. and Sebastiani P. Learning Bayesian Networks from Incomplete Databases. June 1997. 14 Pages.
ISBN 1 901615 11 1

14. Sebastiani P. and Wynn H.P. Risk Based Optimal Designs. June 1997. 10 Pages.
ISBN 1 901615 13 8

15. Cowell R. Sampling without Replacement in Junction Trees. June 1997. 10 Pages.
ISBN 1 901615 14 6

16. Dagg R.A. and Newby M.J. Optimal Overhaul Intervals with Imperfect Inspection and Repair. July 1997. 11 Pages. ISBN 1 901615 15 4

17. Sebastiani P. and Wynn H.P. Bayesian Experimental Design and Shannon Information. October 1997. 11 Pages. ISBN 1 901615 17 0

18. Wolstenholme L.C. A Characterisation of Phase Type Distributions. November 1997. 11 Pages. ISBN 1 901615 18 9

19. Wolstenholme L.C. A Comparison of Models for Probability of Detection (POD) Curves. December 1997. 23 Pages. ISBN 1 901615 21 9

20. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. February 1999. 19 Pages. ISBN 1 901615 37 5

21. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. November 1999. 12 Pages ISBN 1 901615 40 5

22. Cowell R.G. FINEX : Forensic Identification by Network Expert Systems. March 2001. 10 pages.
ISBN 1 901615 60X

23. Cowell R.G. When Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric. March 2001. 11 pages.
ISBN 1 901615 61 8

24. Kaishev, V.K., Dimitrova, D.S., Haberman S., and Verrall R.J. Autumatic, Computer Aided Geometric Design of Free-Knot, Regression Splines. August 2004. 37 pages.
ISBN 1-901615-81-2

25. Cowell R.G., Lauritzen S.L., and Mortera, J. Identification and Separation of DNA Mixtures Using Peak Area Information. December 2004. 39 pages.
ISBN 1-901615-82-0

Faculty of Actuarial Science and Statistics

Actuarial Research Club

The support of the corporate members

CGNU Assurance
Computer Sciences Corporation
English Matthews Brockman
Government Actuary's Department
Watson Wyatt Partners

is gratefully acknowledged.

ISBN 1-901615-82-0