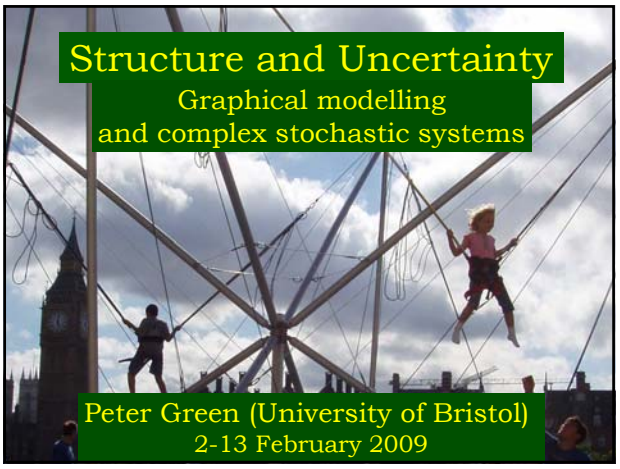


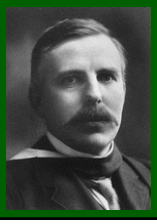
Structure and Uncertainty

Graphical modelling
and complex stochastic systems

Peter Green (University of Bristol)
2-13 February 2009

What has statistics to say about science and technology?

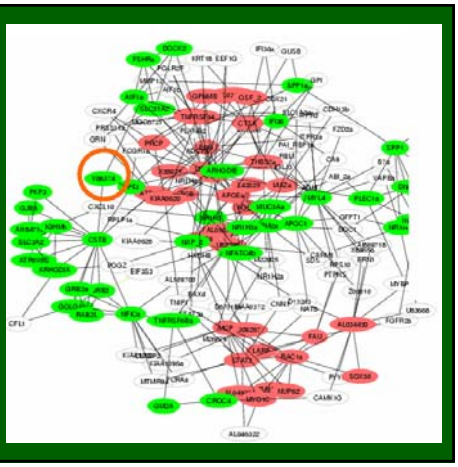
Statistics and science



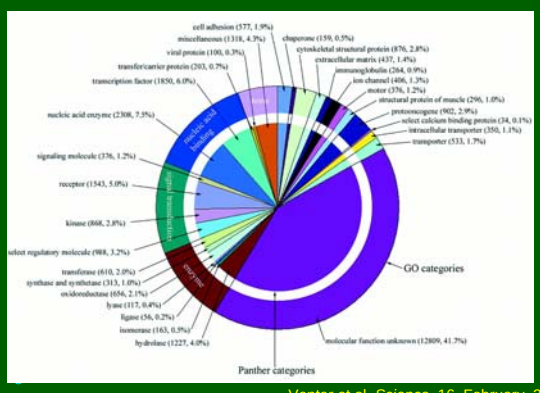
Ernest Rutherford (1871-1937)

What has statistics to say about the complexity of modern science?

Gene networks

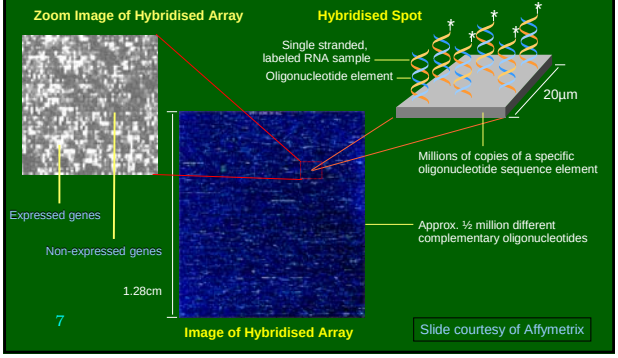


Functional categories of genes in the human genome



Venter et al, *Science*, 16 February, 2001

Gene expression using Affymetrix microarrays



Slide courtesy of Affymetrix

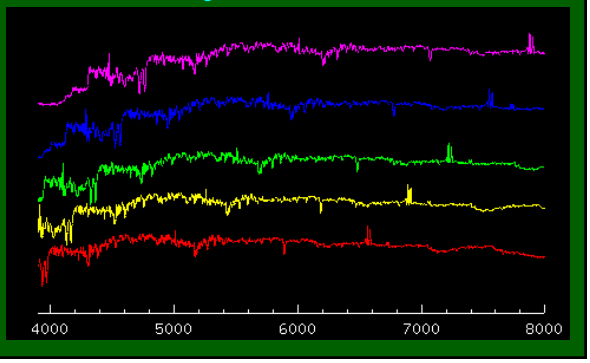
7

8

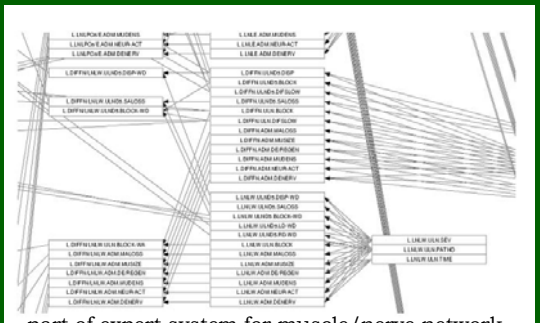
Velocity of recession determines 'colour' through redshift effect

9

Astronomy: redshifts



Probabilistic expert systems



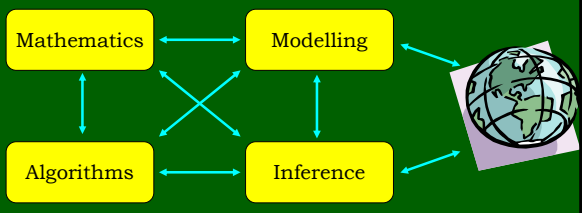
part of expert system for muscle/nerve network

Complex stochastic systems

Problems in these areas – and many others - have been successfully addressed in a modern statistical framework of structured stochastic modelling

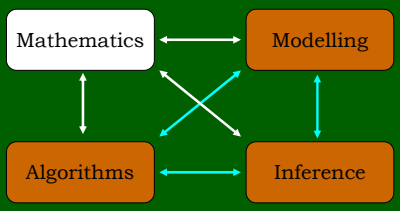
12

Graphical modelling



13

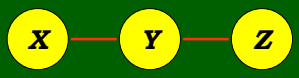
1. Mathematics



14

Conditional independence

- X and Z are conditionally independent given Y if, knowing Y, discovering Z tells you nothing more about X: $p(X \mid Y,Z) = p(X \mid Y)$
- $X \perp Z \mid Y$



15

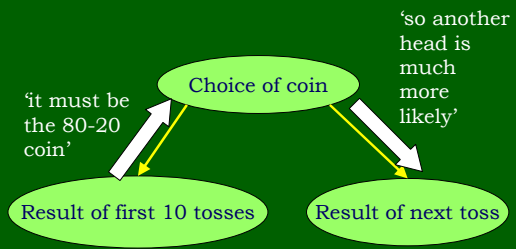
Coin-tossing

- You take a coin from your pocket, and toss it 10 times and get 10 heads
- What is the chance that the next toss gives head?

Now suppose there are two coins in your pocket – a 80-20 coin and a 20-80 coin – what is the chance now?

16

Coin-tossing



(conditionally independent given coin)
(the odds on a head are now 3.999986 to 1)

17

Conditional independence

as seen in data on perinatal mortality vs. ante-natal care....

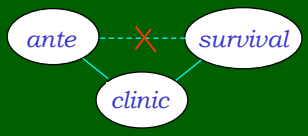
Clinic	Ante	Survived	Died	% died
A	less	176	3	1.7
	more	293	4	1.3
B	less	197	17	7.9
	more	23	2	8.0

Does survival depend on ante-natal care?

18

.... what if you know the clinic?

Conditional independence



survival and *clinic* are dependent

and *ante* and *clinic* are dependent

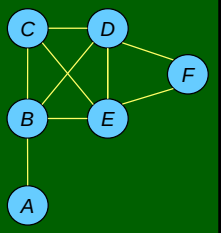
but *survival* and *ante* are CI given *clinic*

19

Graphical models

Use ideas from graph theory to

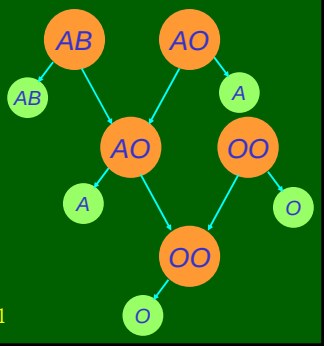
- represent structure of a joint probability distribution
- by encoding conditional independencies



20

Mendelian inheritance - a natural structured model

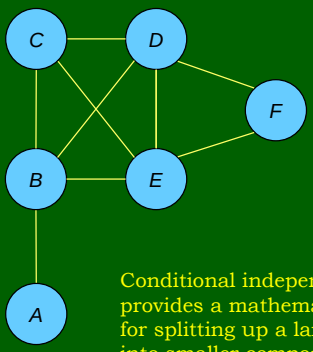
Mendel



Where does the graph come from?

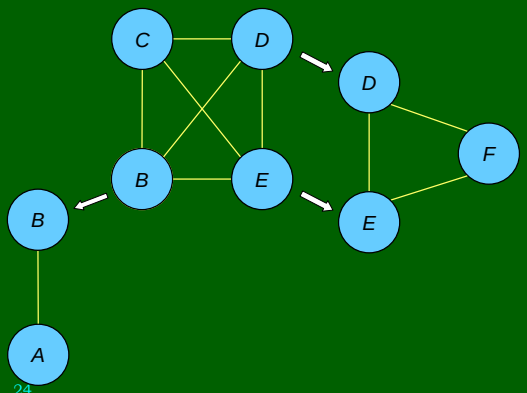
- Genetics
  - pedigree (family connections)
- Physical and biological systems
  - supposed causal effects
- Contingency tables
  - hypothesis tests on data
- Gaussian case
  - graph determined by non-zeroes in inverse variance matrix (i.e. non-zero partial correlations)

22



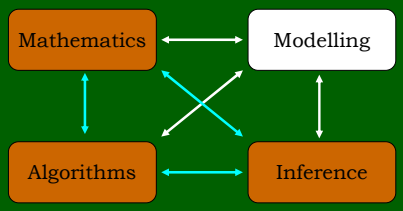
Conditional independence provides a mathematical basis for splitting up a large system into smaller components

23



24

2. Modelling



25

Structured systems

A framework for building models, especially probabilistic models, for empirical data

Key idea -

- understand complex system
- through global model
- built from small pieces
  - comprehensible
  - each with only a few variables
  - modular

26

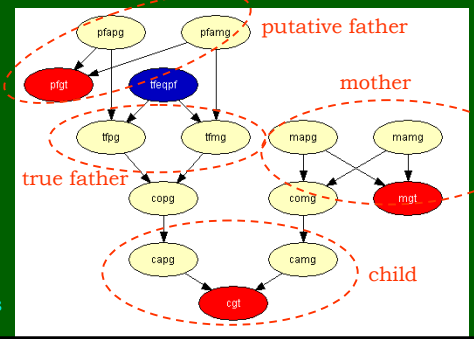
Modular structure

Basis for

- understanding the real system
- capturing important characteristics statistically
- defining appropriate methods
- computation
- inference and interpretation

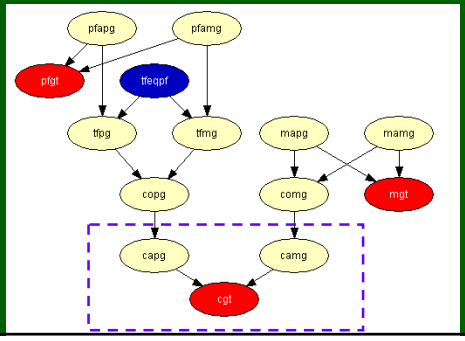
27

Building a model, for genetic testing of paternity using DNA probes



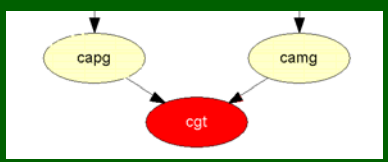
28

Building a model, for genetic testing of paternity



29

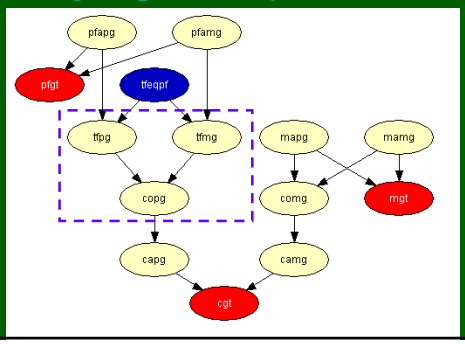
… genes determine genotype



e.g. if child's paternal gene is '10' and maternal gene is '12', then its genotype is '10-12'

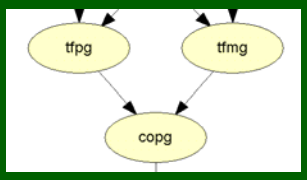
30

Building a model, for genetic testing of paternity



31

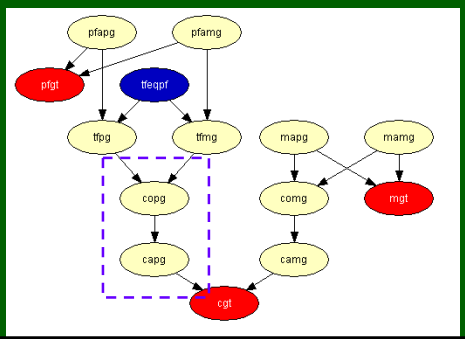
… Mendel's law



the gene that the child gets from the father is equally likely to have come from the father's father or mother

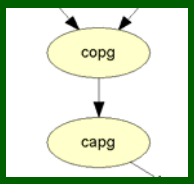
32

Building a model, for genetic testing of paternity



33

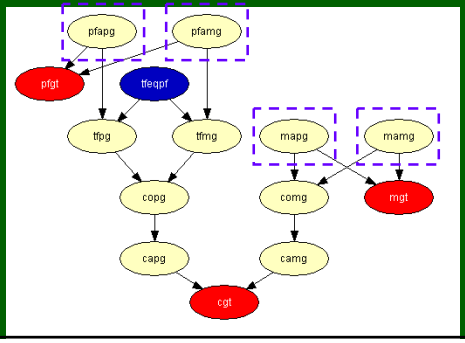
… with mutation



there is a small probability of a gene mutating

34

Building a model, for genetic testing of paternity



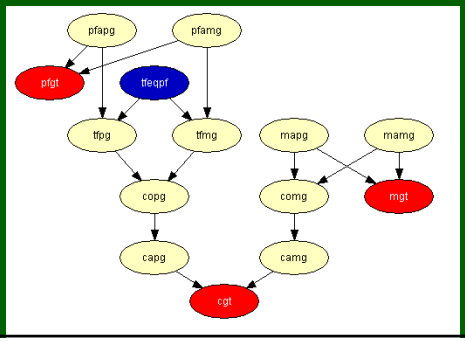
35

… using population data

we need gene frequencies relevant to assumed population for 'founder' nodes

36

Building a model, for genetic testing of paternity

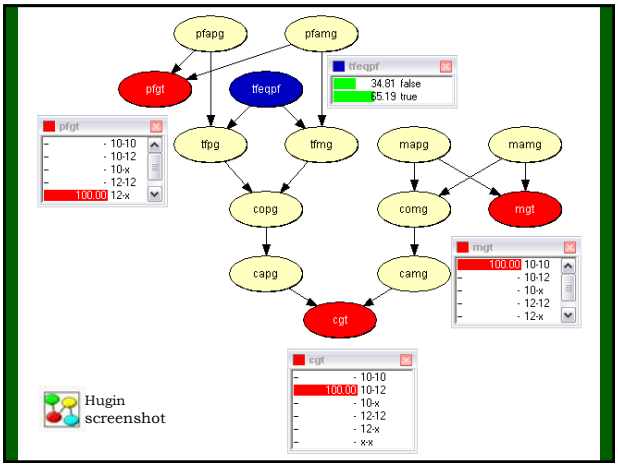


37

Building a model, for genetic testing of paternity

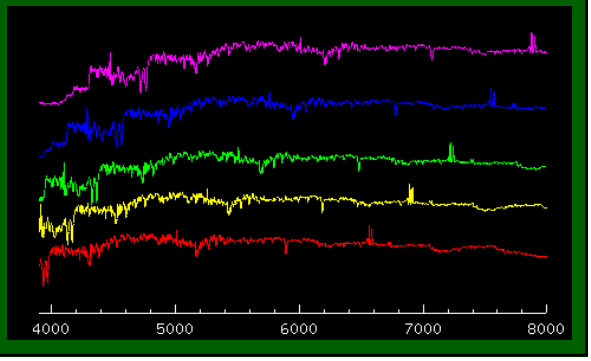
- Having established conditional probabilities within each of these local models….
- We can insert 'evidence' (data) and draw probabilistic inferences…

38



40

Photometric redshifts



Photometric redshifts

$$x_{ij} = a_j + b_i + M\left(\sum_{\ell=1}^{4} w_{i\ell} M^{-1}\{\psi_{\ell j}(z_i)\}\right) + \varepsilon_{ij}$$

$$\psi_{\ell j}(z) = \int h_j((1+z)u)\phi_\ell(u)du$$

$M(\cdot) = -2.5\log_{10}(\cdot)$
magnitude $= M$(flux)

42

Photometric redshifts

$$x_{ij} = a_j + b_i + M\left(\sum_{\ell=1}^{4} w_{i\ell} M^{-1}\{\psi_{\ell j}(z_i)\}\right) + \varepsilon_{ij}$$

Multiplicative model (on flux scale), involving an unknown mixture of templates

$$M(\cdot) = -2.5\log_{10}(\cdot)$$

magnitude $= M$(flux)

43

Photometric redshifts

redshift

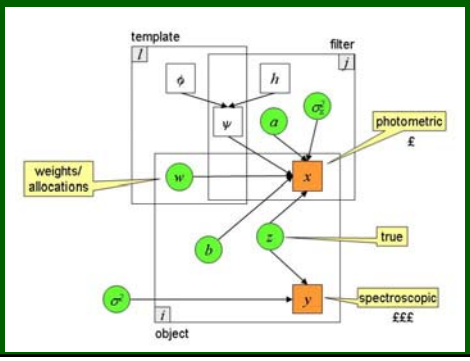
filter response

template

$$\psi_{\ell j}(z) = \int h_j((1+z)u)\phi_\ell(u)du$$

44

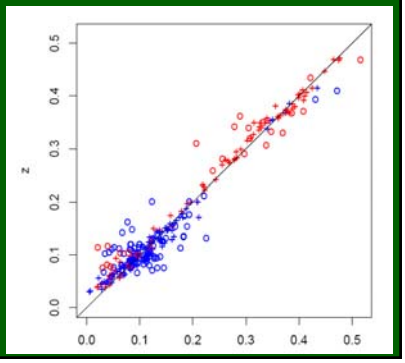
Photometric redshifts



45

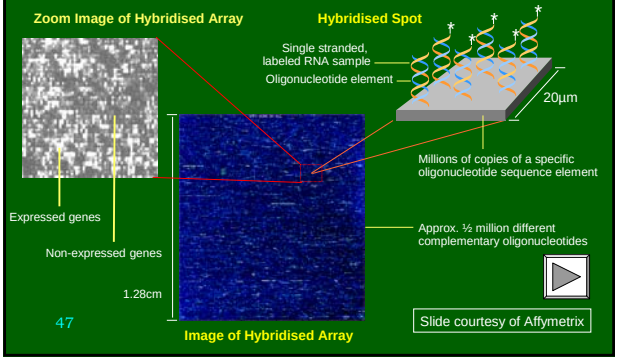
Photometric redshifts

good agreement with 'gold-standard' spectrographic measurement



46

Gene expression using Affymetrix microarrays



47

Variation and uncertainty

Gene expression data (e.g. Affymetrix™) is the result of multiple sources of variability

- condition/treatment
- biological
- array manufacture
- imaging
- technical
- within/between array variation
- gene-specific variability

Structured statistical modelling allows considering all uncertainty at once

48

Single array model: motivation

(Hein, Richardson, Causton, Ambler & G, 2004)

Key observations:		Conclusions:
PMs and MMs both increase with spike-in concentration (MMs slower than PMs)	→	MMs bind fraction of signal
Spread of PMs increase with level	→	Multiplicative (and additive) error; transformation needed
Considerable variability in PM (and MM) response within a probe set	→	Varying reliability in gene expression estimation for different genes
Probe effects approximately additive on log-scale	→	Estimate gene expression measure from PMs and MMs on log scale

49

BGX single array model

$PM_{gj} \sim N(S_{gj} + H_{gj}, \tau^2)$

$MM_{gj} \sim N(\Phi S_{gj} + H_{gj}, \tau^2)$

Background noise, additive

fraction

signal

Non-specific hybridisation

$\log(S_{gj}+1) \sim TN(\mu_g, \xi_g^2)$ $j=1,\ldots,J$

$\log(H_{gj}+1) \sim TN(\lambda, \eta^2)$

Gene expression index (BGX): θ_g=median(TN (μ_g, ξ_g^2))
"Pools" information over probes j=1,…,J

Gene specific error terms: exchangeable

Array-wide distribution

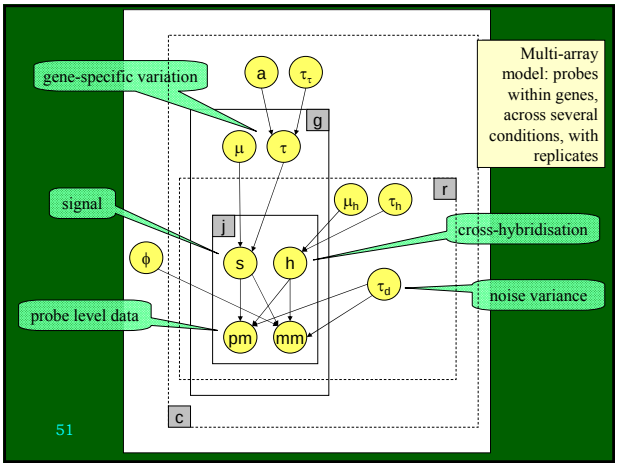
"Empirical Bayes"

$\log(\xi_g^2) \sim N(a, b^2)$

Priors: "vague" $\tau^2 \sim \Gamma(10^{-3}, 10^{-3})$ $\Phi \sim B(1,1)$, $\mu_g \sim U(0,15)$ $\eta^2 \sim \Gamma(10^{-3}, 10^{-3})$, $\lambda \sim N(0,10^3)$

g=1,…,G (thousands), j=1,…,J (11-20)

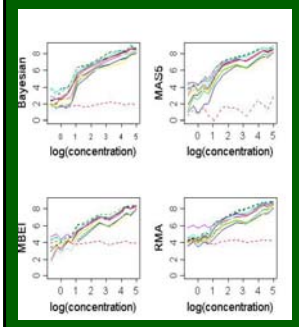
50

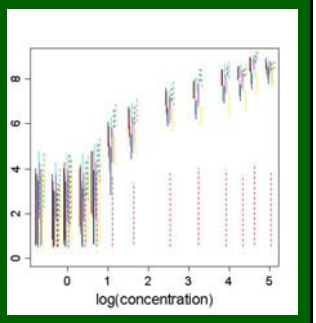


51

Single array model performance:
11 genes spiked in at 13 (increasing) concentrations

Hein, Richardson, Causton, Ambler & G, 2004

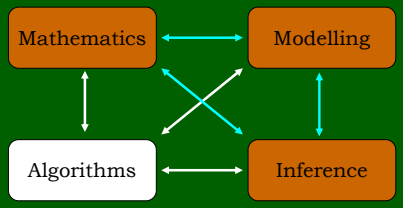




Indication of good resolution at low concentrations and probably reduced bias

Unlike competitors, method provides confidence intervals

3. Algorithms



53

Algorithms for probability and likelihood calculations

Exploiting graphical structure:

- Markov chain Monte Carlo
- Probability propagation (Bayes nets)
- Expectation-Maximisation
- Variational (mean-field) methods

Graph representation used in user interface, data structures and in controlling computation

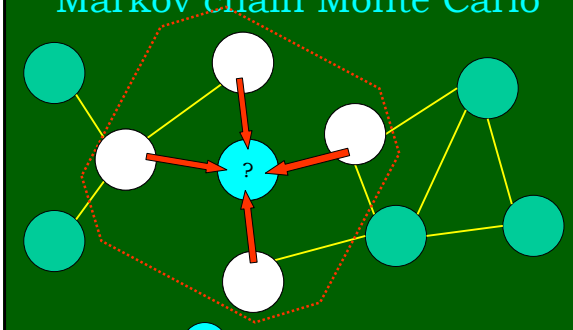
54

Markov chain Monte Carlo

- Subgroups of one or more variables updated randomly,
  - maintaining detailed balance with respect to target distribution
- Ensemble converges to equilibrium = target distribution (= Bayesian posterior, e.g.)

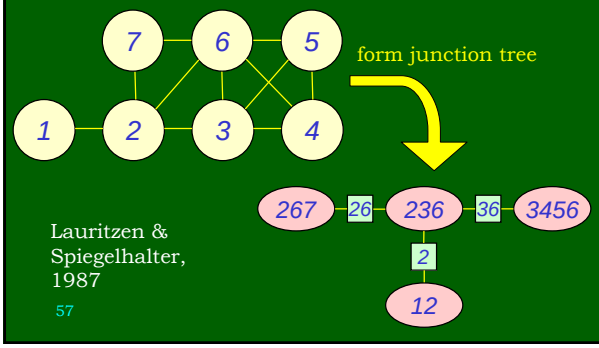
55

Markov chain Monte Carlo



56 Updating ? - need only look at neighbours

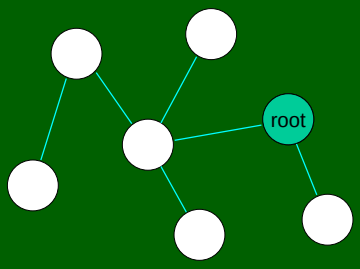
Probability propagation



Lauritzen & Spiegelhalter, 1987

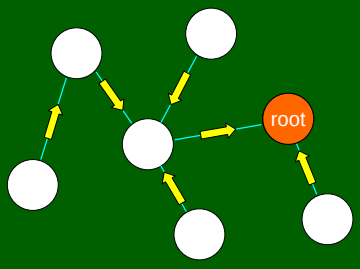
57

Message passing in junction tree



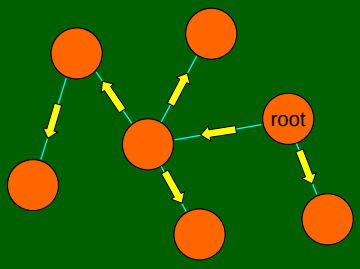
58

Message passing in junction tree - collect



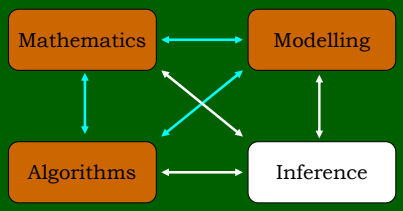
59

Message passing in junction tree - distribute



60

4. Inference



61

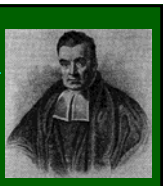
Bayesian

62

or non-Bayesian

63

Bayesian paradigm in structured modelling



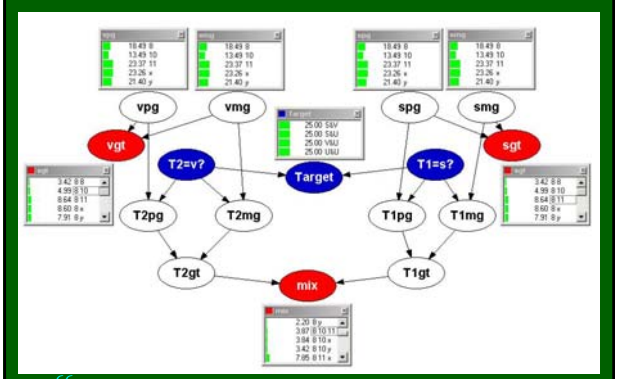
- 'borrowing strength'
- automatically integrates out all sources of uncertainty
- properly accounting for variability at all levels
- including, in principle, uncertainty in model itself
- avoids over-optimistic claims of certainty

64

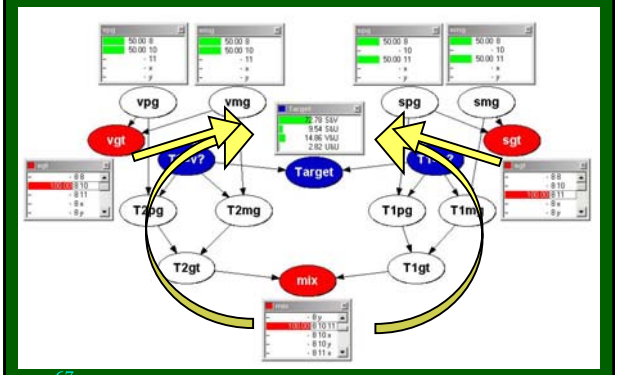
Bayesian structured modelling

- 'borrowing strength'
- automatically integrates out all sources of uncertainty
- … for example in forensic statistics with DNA probe data…..

65



66



67

Bayesian structured modelling

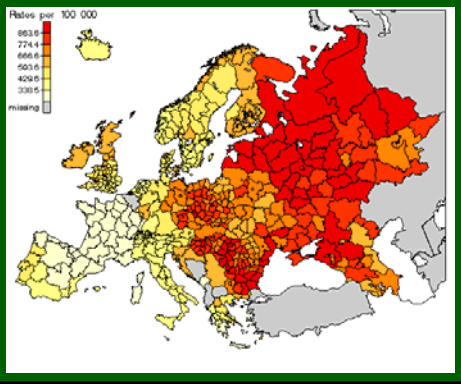
- 'borrowing strength'
- automatically integrates out all sources of uncertainty
- … for example in hidden Markov models for disease mapping

68

John Snow's 1855 map of cholera cases

69

Mortality for diseases of the circulatory system in males in 1990/1991



70

Disease mapping

- Observe counts y_i of cases of rare, non-infectious disease in regions i
- Standard model:

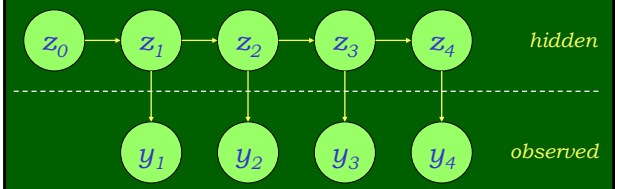
$$y_i \sim \text{Poisson}(\lambda_i E_i)$$

- where E_i are adjusted populations at risk
- Relative risks λ_i vary due to unmeasured risk factors, assumed spatially correlated
- Use space as surrogate to separate signal and noise

71

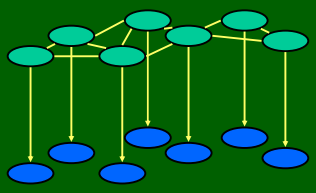
Hidden Markov models

e.g. Hidden Markov chain



72

Hidden Markov random fields

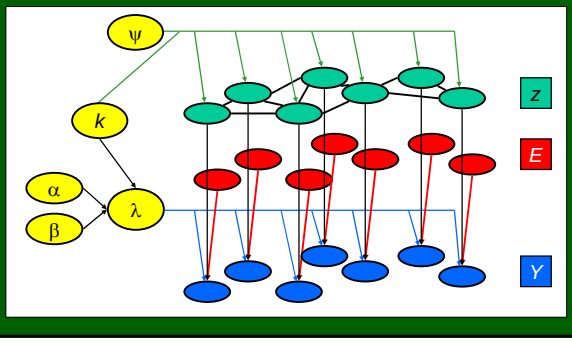


Unobserved dependent field

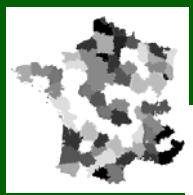
Observed conditionally-independent (e.g. discrete) field

73

Hierarchical mixture model for disease mapping – a hidden MRF

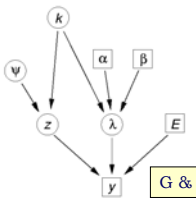


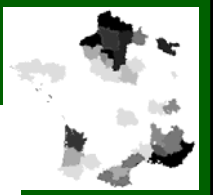
Mapping of rare diseases
using Hidden Markov model



Larynx cancer in females in France, 1986-1993 (standardised ratios)

75





Posterior probability of excess risk

G & Richardson, 2002

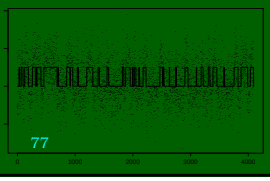
Bayesian structured modelling

- 'borrowing strength'
- automatically integrates out all sources of uncertainty
- … for example in modelling complex biomedical systems like ion channels…..

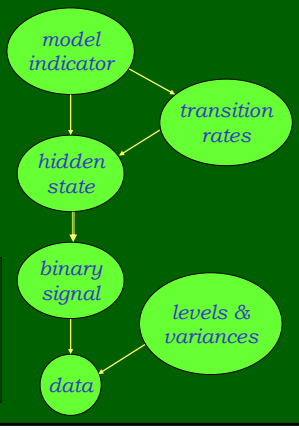
76

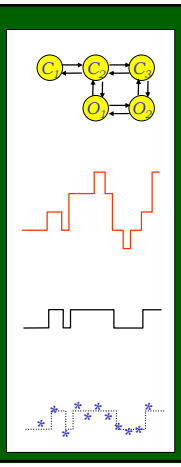
Ion channel model

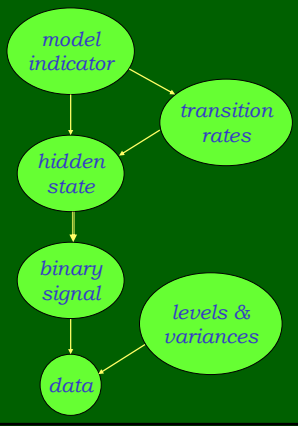
Hodgson and Green, Proc Roy Soc Lond A, 1999



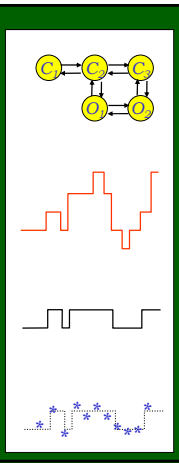
77







78



Unknown physiological states of channel, unknown connections

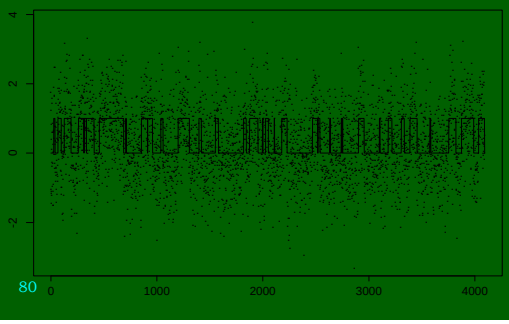
Continuous time Markov chain on this graph, with unknown transition rates

Only open/closed status of states is relevant to observation

We observe only in discrete time, with highly correlated noise

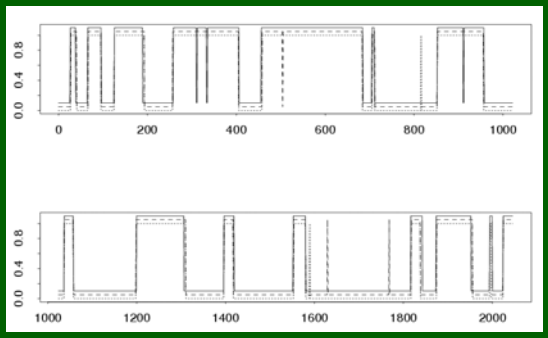
79

Truth and simulated data



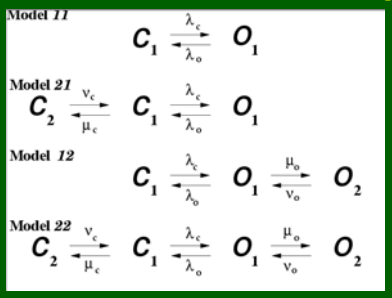
80

Truth and 2 restorations



Ion channel model choice

posterior probabilities



.405

.119

.369

.107

82

Structured systems' success stories include...

- Genomics & bioinformatics
  - DNA & protein sequencing, gene mapping, evolutionary genetics
- Spatial statistics
  - image analysis, environmetrics, geographical epidemiology, ecology
- Temporal problems
  - longitudinal data, financial time series, signal processing

83

Structured systems' challenges include...

- Very large/high-dimensional data sets
  - genomics, telecommunications, commercial data-mining...

84

Summary

Structured stochastic modelling (the 'HSSS' approach) provides a powerful and flexible approach to the challenges of complex statistical problems

- Applicable in many domains
- Allows exploiting scientific knowledge
- Built on rigorous mathematics
- Principled inferential methods

85

http://www.stats.bris.ac.uk/~peter

P.J.Green@bristol.ac.uk

86