Bayesian inference: an introduction

Peter Green

School of Mathematics
University of Bristol

8/9 September 2011 / MLSS 2011, Bordeaux

Outline

Outline

Summary: key ideas

- Inference: the process of discovering from data
  - about mechanisms that may have caused or generated that data
  - or at least explain it
- Goals are varied
  - perhaps simply predicting future data
  - more ambitiously, learning about scientific or societal truths
- In applied mathematics language, these are inverse problems
- Bayesian inference is about using probability to do all this
- A key strength: all sources of uncertainty are simultaneously and coherently considered
- It is model-based
  - in the language of machine learning, these are generative models
  - we can use Bayesian methods to choose and criticise our models

Summary: contents

- motivation
- probability
- basic principles and concepts
- modelling in principle and practice
- computing Bayesian inferences
- subjective and objective theories
- sensitivity to assumptions
- some more substantial applications

Connections

There are connections with all other topics covered in this summer school, but especially with

- Arnaud Doucet: Monte Carlo methods (yesterday & today)
- Yee Whye Teh: Bayesian nonparametrics (next Monday/Tuesday)
- Martin Wainwright: Graphical models (next Thursday/Friday)

so in my lectures, these topics will be under-played.

Machine Learning (As Explained to a Statistician)

Michael Jordan

- A loose confederation of themes in statistical inference (and decision-making)
- A focus on prediction and exploratory data analysis - not much worry about “coverage”
- A focus on computational methodology and empirical evaluation, with a dollop of empirical process theory - lots of nonparametrics, but not much asymptotics
- Sometimes Bayesian and sometimes frequentist - not much interplay

Third generation machine intelligence

Chris Bishop

General theme: deep integration of domain knowledge and statistical learning

- Bayesian framework
- Probabilistic graphical models
- Fast inference using local message-passing

Origins: Bayesian networks, decision theory, HMMs, Kalman filters, MRFs, mean field theory, ...

Statistics and machine learning

- Origins in different communities and with different traditions, but rapidly converging
- Machine learning
  - has been more ambitious in reach and scale
  - big problems, but structure within observations, not between them
  - focus on prediction, evaluated by cross-validation, etc
- Statistics
  - more emphasis on model-building, more reliance on models
  - more aim at scientific understanding, less concern with throughput
  - evaluation through models
- Still a lot to learn from each other

Bayesian and frequentist statistics

There are different paradigms for statistical inference – not just these two, by the way. Historically, philosophical debates have been interesting, sometimes distracting or destructive. Nowadays, there's more understanding and flexibility, less 'theology'.

Bayesian and frequentist statistics

Bayesian

- methods only come from models
- inferences should be made conditional on the current data
- focus on coherence
- natural in the setting of a long-term project with a domain expert?
- philosophically compelling but can be hard to do

Frequentist

- methods can come from anywhere
- inferential methods should give good answers in repeated use
- focus on calibration
- natural when writing software that will be used by many people with many data sets?

Bayesian inference: some other issues

- defending a Bayesian analysis: why should you have to?
- directness of inference
- flexibility of inference - ranking, selection
- borrowing strength (one thing is always informative about another)
- really pays off in complex, high-dimensional problems
- ubiquitous cheap computing has really allowed Bayesian analysis to blossom
- evaluating Bayesian methods by their frequentist performance

Outline

1. Motivation and key ideas

2. Basics
   - Basic principles and concepts
   - Priors
   - Hierarchical modelling
   - Exchangeability
   - Hidden Markov models and State space models

Probability

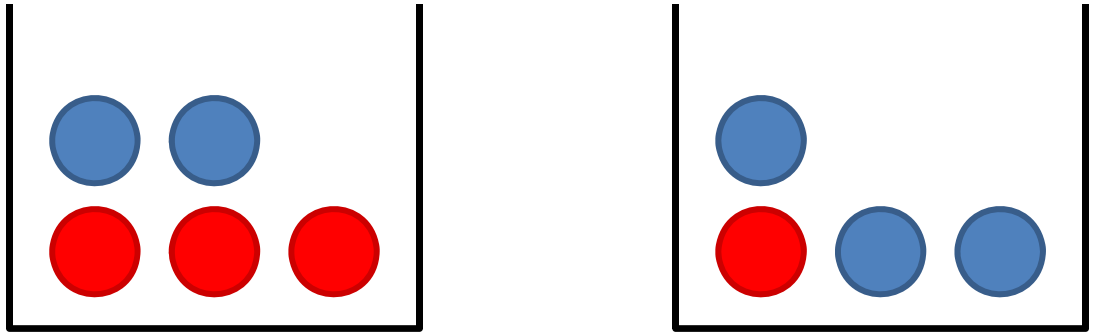
"nothing but common sense reduced to calculus"

- measures uncertainty on a $[0, 1]$ scale (with obvious interpretations of 0 and 1)
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$ (the chance of A or B is the sum of the individual chances, if A and B cannot both occur together)

This (or rather, this with the 2nd rule extended to countably infinite collections of events) is all you need for the entire theory.

Most of the time, we only need work with random variables – numerical-valued random outcomes – and their distributions.

Bayes theorem

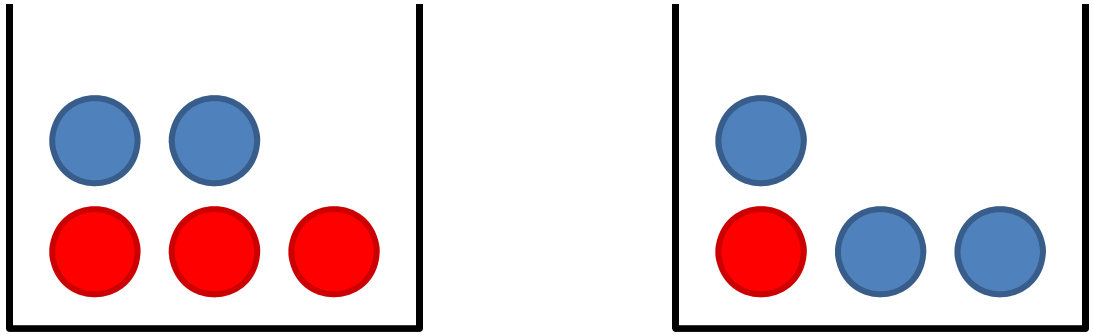


To escape the firing squad, you can draw two balls from the urns – if you get a red ball, you are freed. You choose an urn at random, and the ball you draw first is blue!

If you stick with the same urn for the 2nd draw, what is your chance of escape?

Would you do better to switch urns after the first draw?

Bayes theorem



A silly problem - but captures the ideas of

- inference: which urn did you draw from?
- prediction: what will the next ball be?
- decision: which is the better strategy?

– all determined by rules of probability.

Bayes theorem

Let the random quantities of interest by U (urn), $B1$ (first ball drawn) and $B2$ (second ball drawn). Then the inference question needs us to evaluate

$$P(U = \text{left}|B1 = \text{blue}) = P(U = \text{left} \cap B1 = \text{blue})/P(B1 = \text{blue})$$

But $P(U = \text{left} \cap B1 = \text{blue}) = P(U = \text{left}) \times P(B1 = \text{blue}|U = \text{left}) = 1/2 \times 2/5 = 1/5$

Also $P(B1 = \text{blue}) = P(U = \text{left} \cap B1 = \text{blue}) + P(U = \text{right} \cap B1 = \text{blue}) = 1/5 + 1/2 \times 3/4 = 23/40$

So our answer is $1/5 \div 23/40 = 8/23$.

Bayes theorem

For a binary variable like U it's cleaner to work with odds:

$$\frac{P(U = \text{left}|B1 = \text{blue})}{P(U = \text{right}|B1 = \text{blue})} = \frac{P(U = \text{left})}{P(U = \text{right})} \times \frac{P(B1 = \text{blue}|U = \text{left})}{P(B1 = \text{blue}|U = \text{right})}$$
$$= \frac{1/2}{1/2} \times \frac{2/5}{3/4} = \frac{8}{15}$$

That solves the inference question. We see that, given the data that $B1 =$ blue we now think it is more likely that we are drawing from the urn on the right.

But we are not certain about that – and the rules of probability correctly use that uncertainty in the prediction and decision questions.

Bayes theorem – prediction

For the prediction question we want $P(B2 = \text{red}|B1 = \text{blue})$, or equivalently the odds

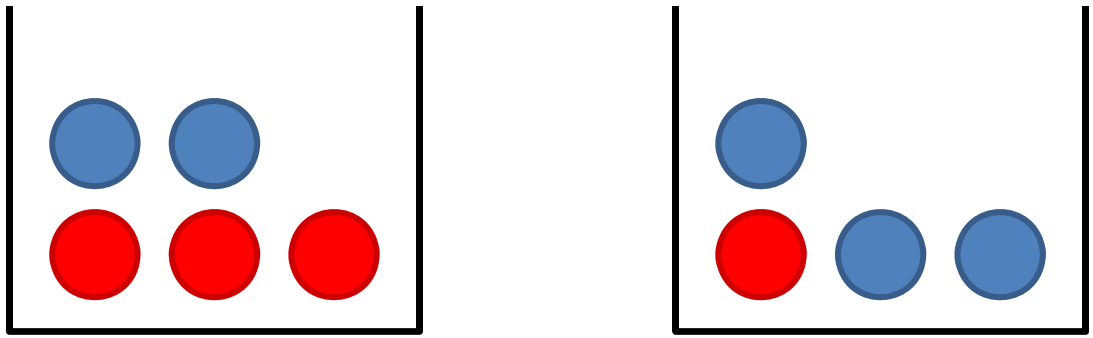
$$\frac{P(B2 = \text{red}|B1 = \text{blue})}{P(B2 = \text{blue}|B1 = \text{blue})} = \frac{P(B1 = \text{blue} \cap B2 = \text{red})}{P(B1 = \text{blue} \cap B2 = \text{blue})}$$

and to evaluate both numerator and denominator we go back to conditioning on U, e.g. $P(B1 = \text{blue} \cap B2 = \text{red})$
$= P(U = \text{left})P(B1 = \text{blue}|U = \text{left})P(B2 = \text{red}|U = \text{left}, B1 = \text{blue})$
$+P(U = \text{right})P(B1 = \text{blue}|U = \text{right})P(B2 = \text{red}|U = \text{right}, B1 = \text{blue})$
$= 1/2 \times 2/5 \times 3/4 + 1/2 \times 3/4 \times 1/3 = 11/40.$

Similarly
$P(B1 = \text{blue} \cap B2 = \text{blue}) = 1/2 \times 2/5 \times 1/4 + 1/2 \times 3/4 \times 2/3 = 12/40,$
and finally the odds that $B2 = \text{red}$ given $B1 = \text{blue}$ are $11/12$.

Bayes theorem – decision

For the decision question we want simply to compare these odds that $B2 =$ red given $B1 =$ blue with the same thing calculated assuming you switch urns after the first draw.



What do you think? Stick or switch?

Bayes theorem

Inference

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)\mathrm{d}\theta}$$

Prediction

$$p(y^+|y) = \frac{\int p(\theta)p(y|\theta)p(y^+|\theta,y)\mathrm{d}\theta}{\int p(\theta)p(y|\theta)\mathrm{d}\theta} = \int p(\theta|y)p(y^+|\theta,y)\mathrm{d}\theta$$

Consistent use of probability to quantify uncertainty

- All variables in the system as modelled as random variables – whatever their role: parameters, data, latent, observable, observed, unobservable, unobserved, future . . . – we do not blur these distinctions philosophically, just treat them together mathematically.
- The randomness of these quantities can be of different kinds – notably epistemological and aleatory
- epistemological uncertainty: concerning lack of knowledge about unique and potentially verifiable events, essentially a degree of ignorance
- aleatory uncertainty: concerned with essentially random phenomena

Epistemological and aleatory

Thanks to David Spiegelhalter

An experiment I carry out in front of school audiences helps to distinguish these 2 concepts.

I hold a coin and ask, "What is the chance this will come up heads?" They cheerfully say something like "50%" or "half-and-half."

I then toss the coin, catch it, flip it onto the back of my hand without revealing it, and ask, "What is the probability this is heads?" Pause. Then someone, less confidently, mumbles "50%."

I reveal the coin to myself, but not to them, and ask, "What is your probability that this is heads?" Very grudgingly they might eventually admit "50%."

Epistemological and aleatory

Thanks to David Spiegelhalter

In this experiment I have gone from pure aleatory uncertainty to pure epistemological uncertainty, showing

- epistemological uncertainty is "in the eye of the beholder" (my probability was eventually 0% or 100%, whereas theirs was still 50%),
- that the language of probability applied to both forms, and
- that these different types of uncertainty may be perceived differently.

Consistent use of probability to quantify uncertainty

- We work throughout with a single joint probability distribution – that of all the variables, whatever their status
- This joint distribution is arrived at ('assessed') through scientific judgement, guided by the laws of probability
- We model the process of observation by conditioning on observed values (may be philosophically controversial)
- We can then perform inference by looking at $P(\text{unobserved}|\text{observed})$ (=$P(\text{hidden}|\text{visible})$)

Likelihood and prior

- In a simple standard setting, suppose we have parameters θ and data y. Having observed y, we make inference about θ using the posterior $p(\theta|y)$.
- $p(\theta|y) = p(\theta, y)/p(y) \propto p(\theta, y)$ (proportional in θ)
- Almost always, the joint distribution $p(\theta, y)$ is created from
  - the marginal distribution (or prior) for θ, $p(\theta)$, and
  - the conditional distribution (or likelihood) for y given θ, $p(y|\theta)$ (a generative model),
  - multiplied to give $p(\theta, y) = p(\theta) \times p(y|\theta)$
- We will talk more about priors later, but quite commonly $p(\theta)$ and $p(y|\theta)$ are different in nature – the prior $p(\theta)$ is often entirely subjective, perhaps based on unquantified scientific judgement, while the likelihood $p(y|\theta)$ may be open to empirical verification.

Sequential acquisition of data

One of the many free benefits of using probability consistently in inference can be seen in the neat way that inferences are updated as data are acquired, in a sequential setting.
Suppose we acquire $y_1, y_2, \ldots$ in sequence, and wish to update our inference about θ after each observation. In the simple case where $y_1, y_2, \ldots$ are conditionally independent given θ, we have

$$p(\theta, y_1, y_2, \ldots, y_n) = p(\theta) \prod_{i=1}^{n} p(y_i|\theta) \quad \text{, so}$$

$$p(\theta|y_1, y_2, \ldots, y_n) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta) \propto p(\theta|y_1, y_2, \ldots, y_{n-1}) \times p(y_n|\theta)$$

. . . the prior for the n^{th} datum is the posterior after the $(n-1)^{\text{th}}$.

Utility and loss

Reading many accounts of Bayesian analysis, you'd think that it is only about combining prior and likelihood – what you thought about θ before you had any data, and what you learnt about θ by acquiring the data – together with some computational and presentational work.

But a third important ingredient is needed, always present though often implicit. It is essential to think about it explicitly if

- you are taking decisions
- you are testing hypotheses
- you are being thoughtful about estimation

This approach is called decision theory.

Utility and loss

Having modelled your system, observed data and obtained the posterior distribution $p(\theta|y)$ – what do you do with it?

Utility theory is very general and comprehensive, and allows you to analyse the value (for example in money terms) of acquiring data in the presence of uncertainty. For inference, it is enough to think about the simpler idea of loss functions.

- we observe y,
- make a decision $d = \delta(y)$, and
- incur a loss $L(d, \theta)$ if the true value of 'nature' is θ,

– 'how bad' is it to decide $d = \delta(y)$ when θ is true?

Loss functions and hypothesis testing

A hypothesis is a statement about θ. In the simplest case, imagine we know (or can assume) that either θ lies in a set Ω_0 or a set Ω_1, with $\Omega_0 \cap \Omega_1 = \emptyset$. We want to use data y to decide which of the statements $\theta \in \Omega_0$ and $\theta \in \Omega_1$ is true – that is we take a decision $d = d_0$ or d_1 where d_i is 'I think $\theta \in \Omega_i$'.

If we get the right answer, that's fine. Otherwise we incur a loss:

$$L(d_i, \theta) = \begin{cases} 0 & \text{if } \theta \in \Omega_i, \\ a_i & \text{otherwise,} \end{cases}$$

where $a_0, a_1 > 0$.

Now, you observe y – what should you decide?

Loss functions and hypothesis testing

If you knew θ, it's trivial to decide d_0 or d_1 – but you only know y.
An intuitively natural principle (justified by axiomatic utility theory) is to choose d to minimise the (posterior) expected loss.

$$E(L(d,\theta)|y) = \int L(d,\theta)p(\theta|y)\mathrm{d}\theta$$

In our simple testing problem, $E(L(d_i,\theta)|y) = a_i p(\theta \notin \Omega_i|y)$. This amounts to choosing d_0 if $p(\theta \in \Omega_0|y) > a_0/(a_0+a_1)$ – thresholds the posterior probability, taking into account the possibly different costs of false positives and false negatives.

Note that in this (and every) case, the optimal decision is naturally a function of y.

Loss functions and estimation

Estimation is the process of giving a single value of θ as a 'best guess' given the data – so is covered by decision theory with 'decision' that θ has a particular value $\widehat{\theta}$. The posterior expected loss if you assert that $\theta = \widehat{\theta}$ is then

$$E(L(\widehat{\theta}, \theta)|y) = \int L(\widehat{\theta}, \theta) p(\theta|y) \mathrm{d}\theta$$

A common choice is quadratic loss, $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ (for a single parameter θ), and we can then simplify $E(L(\widehat{\theta}, \theta)|y) = E((\widehat{\theta} - \theta)^2|y) = (\widehat{\theta} - E(\theta|y))^2 + \mathrm{var}(\theta|y)$. The best we can do is therefore to set $\widehat{\theta} = E(\theta|y)$ – i.e. use the posterior expectation. The same is true for a vector parameter, using $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^T A(\widehat{\theta} - \theta)$ for any positive definite A.

Loss functions and estimation

But the decision theory approach allows more flexibility than that – perhaps the cost of over-estimation is greater than that of underestimation, then you would want to assume that $L(\theta + c, \theta) > L(\theta - c, \theta)$ for all $c > 0$.
For example, suppose

$$L(\widehat{\theta}, \theta) = \begin{cases} \tau|\widehat{\theta} - \theta| & \widehat{\theta} < \theta \\ (1-\tau)|\widehat{\theta} - \theta| & \widehat{\theta} > \theta \end{cases}$$

then the posterior expected loss turns out to be minimised when $\widehat{\theta}$ is the $100\tau\%$ percentile of the posterior distribution $p(\theta|y)$.

In summary, optimal Bayesian estimation needs you to choose your loss function, and the choice of quadratic loss/posterior mean estimator is not inevitable.

Loss functions and frequentist inference

Loss functions also play a role in frequentist theory, though one that is less prominent.

But frequentists are interested in the expectation of the loss function under the distribution of the data not the parameter:

$$R(\delta, \theta) = \int L(\delta(y), \theta) p(y|\theta) \mathrm{d}y$$

For example, the minimax decision rule δ is that which minimises $\max_\theta R(\delta, \theta)$ over δ.

A decision rule is admissible if there is no other rule that is better for all θ: it turns out that, loosely speaking, every admissible rule is a Bayes rule!

Conjugacy

For some standard models $p(y|\theta)$ for data, a particular choice of prior $p(\theta)$ yields particular algebraic and computational advantages.

A prior $p(\theta)$ is conjugate for a particular likelihood $p(y|\theta)$ is the resulting posterior $p(\theta|y)$ has the same algebraic form.

For example, if $p(y|\theta)$ is the Poisson distribution $\exp(-\theta)\theta^y/y!$ and we assume a Gamma prior $p(\theta) = \beta^\alpha \theta^{\alpha-1} \exp(-\beta\theta)/\Gamma(\alpha)$, then

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) \propto \exp(-\theta)\theta^y \times \theta^{\alpha-1} \exp(-\beta\theta)$$
$$\propto \theta^{\alpha+y-1} \exp(-(\beta+1)\theta)$$

so $p(\theta|y)$ is another Gamma distribution, but with parameters (α, β) updated to $(\alpha + y, \beta + 1)$.

Conjugacy

A second observation z would further update this to Gamma$(\alpha + y + z, \beta + 2)$: so we can interpret the Gamma prior as equivalent to 'prior data': β observations whose total is α; conjugate priors can always be interpreted in this way.

Conjugacy used to be a supremely important factor in choosing a prior, but with the complexity of models typically now used, and the development of computational methods that make algebraic tractability less important, this is no longer true.

However, you still often see conditionally conjugate priors assumed even when MCMC computation is being used, because of small computational advantages. It's a mistake to let these considerations seduce you into adopting a prior that does not reflect the scientific judgements you wish to bring to the analysis.

Subjective and Objective Bayes

Subjective Bayesians take the view that once we have worked hard to understand the system under study, so that we can model everything coherently (with the aid of Bayes' theorem), then all probabilities properly represent our degrees of belief, and cannot be challenged.

The objective Bayes view is that achieving this understanding is too difficult, especially in complex models, and we will be inevitably be forced into simplifying assumptions (lots of independence for example) that will conflict with our true judgements. Conditional probabilities are taken not as representing judgements, but as quantifying the extent to what one event logically determines another. The emphasis switches to choosing priors to have minimal impact on posterior inference.

Uninformative priors

There have been various attempts to define 'objective' priors that represent complete prior ignorance in a logically consistent and realistic way.

Jeffreys' prior for a given likelihood is proportional to $|I(\theta)|^{1/2}$, $I(\theta)$ being Fisher's information matrix. It has the nice property of being equivariant under smooth transformations of θ.

Other ideas for representing ignorance are based on entropy, including maximum entropy priors, and Bernardo's reference priors, which have the property that they maximise the information that will be gained asymptotically under replications of the experiment.

Both Jeffreys' and reference priors have the unattractive feature that they depend on the form of the data that will be collected, not only on the intrinsic character of the 'state of nature' θ itself.

Improper priors

A distribution is improper if it integrates (or sums) to ∞. It is quite possible for a prior to be improper but the corresponding posterior

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)\mathrm{d}\theta}$$

to be perfectly proper – e.g. the previous example with $\alpha = \beta = 0$. It's tempting to rely on this when you have little or no genuine prior information. Is it safe to do so?

Yes – but only when the prior and the resulting posterior are both limits of corresponding proper (prior, posterior) pairs: $p_n(\theta) \to p(\theta)$, $p_n(\theta|y) \propto p_n(\theta)p(y|\theta) \to p(\theta|y)$.

Uninformative priors are often improper.

Some principles of Bayesian modelling

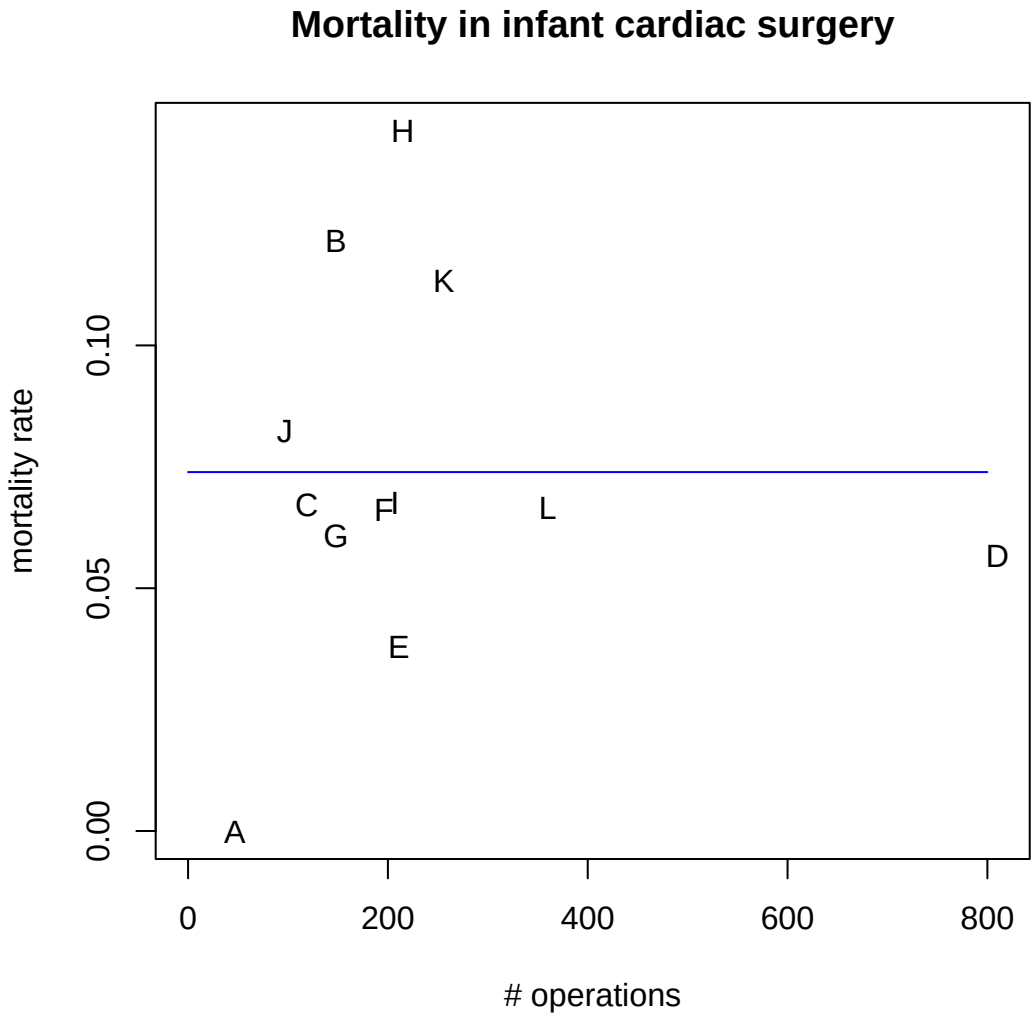
- Hierarchical models
- Exchangeability
- HMMs and state-space models

Motivation for hierarchical modelling

How to make inference on multiple parameters $\{\theta_1, \ldots, \theta_I\}$ measured on I units (persons, centres, areas, ...) *which are related or connected by the structure of the problem ?*

The 'surgical' example

In 12 hospitals carrying out cardiac surgery on babies, the numbers of operations performed and mortality rates are recorded. What are the best and worst hospitals? Are the differences more than can be attributable to chance? What rate do you expect in the 13th hospital? Or in the 12th hospital, in a different year?



The 'surgical' example

In this example, θ_i is the true mortality rate in the ith hospital. Let Y_i and n_i be the number of deaths and the number of operations, in the ith hospital. We might assume $Y_i \sim \text{Binomial}(n_i, \theta_i)$.
We can identify three different assumptions:

1. Identical parameters: All the θ's are identical, in which case all the data can be pooled and the individual units ignored.
2. Independent parameters: All the θ's are entirely unrelated, in which case the results from each unit can be analysed independently (for example using a fully specified prior distribution within each unit)
   $\rightarrow$ individual estimates of θ_i are likely to be highly variable (unless very large sample sizes)
3. Exchangeable parameters: The θ's are assumed to be 'similar' in the sense that the 'labels' convey no information

The 'surgical' example

In the 12 hospitals, the 'raw' mortality rates vary between 0/47 (hospital A) and 31/215=0.1442 (H); the aggregated rate is 208/2814=0.0739. What are the 'true' rates in hospitals A and H?

Non-Bayesian answer 1. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, \theta)$. The maximum likelihood estimator of θ is $(\sum_i Y_i)/(\sum_i n_i) = 0.0739$, which applies to both A and H.

Non-Bayesian answer 2. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, \theta_i)$, independently. The maximum likelihood estimator of θ_i is $Y_i/n_i = 0$ for A and 0.1442 for H.

Could the θ_i all be equal? If θ is 0.0739, the chance that Y_{H} is as big or bigger than 31 is 0.000284. So, no!

The 'surgical' example

Bayesian answer 1. Assume in addition that *a priori*, $\theta \sim \text{Beta}(\alpha, \beta)$ where α and β are say 4 and 46. (This gives a mean and variance for the Beta distribution roughly comparable to the sample mean and variance of the raw mortality rates). Then we get the posterior mean $= (\sum_i Y_i + \alpha)/(\sum_i n_i + \alpha + \beta) = 0.0740$ (for both A and H).

Bayesian answer 2. Making a similar prior assumption on each θ_i, the posterior mean of θ_i is $(Y_i + \alpha)/(n_i + \alpha + \beta)$ = 0.0412, 0.1321 for A and H.

The 'surgical' example

Which is best? Note that the Bayesian estimates are 'shrunk' towards the prior mean $\alpha/(\alpha+\beta) = 0.08$, to an extent depending on the 'denominator' n_i or n. This eliminates ridiculous conclusions like $\theta_{\mathrm{A}} = 0$. However, it is still the case that only the data from hospital i is used in estimating θ_i. Surely the other hospitals' data carries information too? (For example, suppose that Y_{H} was missing: would you be able to guess its value better after having observed the other data?)

The ‘surgical’ example

Our initial model 1 (Bayesian or non-Bayesian) revealed difficulty with the assumption that there was a common mortality rate θ in every hospital; we asked:

- Does this model adequately describe the random variation in outcomes for each hospital?
- Are the hospital failure rates more variable than our model assumes?

and concluded ’no’ and ’yes’, respectively.

Modelling the excess variation

Let's look at Bayesian model 2 above in more detail: we have modified model 1 to allow for a *different* failure probability, θ_i for each hospital i:

$$(y_i \mid \theta_i) \sim \text{Binomial}(n_i, \theta_i) \quad \text{where} \quad \theta_i \sim \text{Beta}(\alpha, \beta)$$

Interpretation:

- $\{\theta_i\}$, the 'true' surgical failure rate in the hospitals are viewed as a random sample from a common *population distribution*
  - $\Rightarrow$ hospital failure rates are assumed to be **similar** but not identical
  - Beta(α, β) prior describes the distribution of surgical failure rates amongst the 'population' of hospitals

How would you specify values for α and β?

Approximate 'empirical Bayes' approach

- Calculate crude failure rates y_i/n_i
- Calculate the observed mean and variance of the 12 values y_i/n_i
- Solve for $\widehat{\alpha}$ and $\widehat{\beta}$ to obtain a beta distribution with this mean and variance
- Using Beta($\widehat{\alpha}$, $\widehat{\beta}$) as a prior, apply Bayes theorem to obtain posteriors for true failure rates θ_i, $p(\theta_i|\widehat{\alpha}, \widehat{\beta}, y_1, y_2, \ldots, y_I)$

Potential problems with this approach:

- We are using the data twice:
  - Once to estimate the prior
  - Again to estimate θ_i for each hospital

  $\Rightarrow$ overestimate precision of our inference
- Using any point estimate for α and β ignores some posterior uncertainty about the population distribution of the θ_i's

Bayesian hierarchical models

The methods discussed here will allow us to do better, because we will be able to assume in advance that the true mortality rates across the hospitals are different (because the circumstances, patients, doctors, ... are different), but similar (because the operations, disease, ... are the same). The effect we will see is that the raw estimates are shrunk *towards each other*.

To do this, we need to deal with more than two sorts of variable – the parameters and data of ordinary Bayesian models. The hospitals problem has 3 levels of uncertainty – the hazard of this type of operation, the variability between hospitals, and chance factors in an individual patients' operation. Such models are called *hierarchical*.

Full hierarchical Bayes approach

- Assume a *joint probability model* for the entire set of parameters $(\theta_1, \theta_2, \ldots, \theta_I, \alpha, \beta)$
  – requires us to assign known prior distributions to α, β, e.g.

$$\alpha \sim \text{Exponential}(1) \quad \text{and} \quad \beta \sim \text{Exponential}(1)$$

- Apply Bayes theorem to calculate the joint posterior distribution of all the unknown quantities simultaneously.

Level 1: $y_i \sim \text{Binomial}(n_i, \theta_i)$, independently for each i
Level 2: $\theta_i \sim \text{Beta}(\alpha, \beta)$, independently for each i
Level 3: Prior for α, β

Advantages of this approach

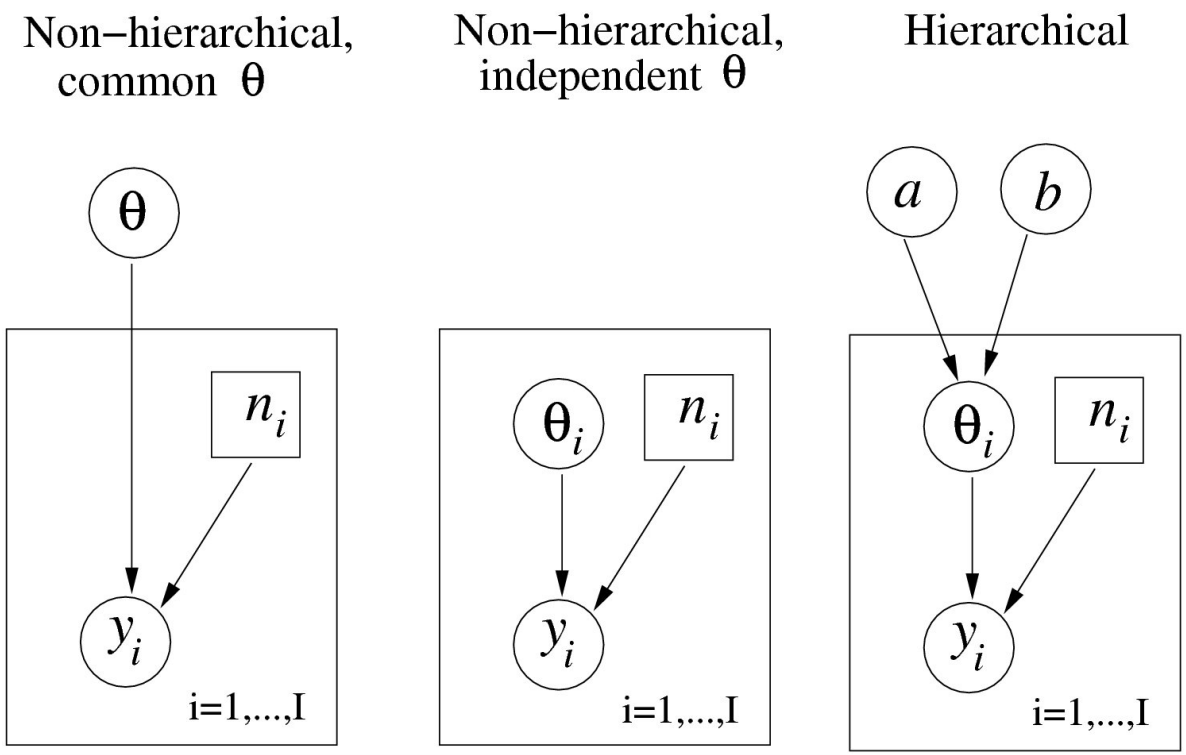
The posterior distribution for each θ_i

- *'borrows strength'* from the likelihood contributions for *all* hospitals, via their joint influence on the estimate of the unknown population (prior) parameters α and β
- reflects our full uncertainty about the true values of α and β

Such models are also called *Random effects* or *Multilevel* models.

Graphical models for surgical example

Directed acyclic graphs (DAGs):



Shrinkage and hierarchical models

To take a different example, suppose in each unit we observe a response x_i assumed to have a Normal likelihood

$$x_i \sim N(\theta_i, \tau_i^2)$$

Unit means θ_i are assumed to be exchangeable, and to have a Normal distribution

$$\theta_i \sim N(\mu, \sigma^2)$$

where μ and σ^2 are 'hyper-parameters', for the moment assumed known, as are τ_i^2.
It can be shown that, after observing x_i, Bayes' theorem gives

$$\theta_i | x_i \sim N(w_i \mu + (1 - w_i) x_i, (1 - w_i)\tau_i^2)$$

where $w_i = \tau_i^2/(\tau_i^2 + \sigma^2) \in (0, 1)$ is the weight given to the prior mean.

Shrinkage and hierarchical models

A Bayesian model therefore leads to inferences for each θ_i giving intervals that are *narrower* than in the non-Bayesian approach, but *shrunk* towards the prior mean response. w_i controls both the 'shrinkage', and the reduction in the width of the interval: it depends on precision of the individual unit i relative to the variability between units. When $\{\tau_i^2\}$ are also given a prior, the same principles apply, although the solution is less explicit.

In a hierarchical model, μ and σ^2 are random, and the effect of this is more complicated again, and *best seen numerically*; the amount of shrinkage is not determined in advance – it is discovered from the data (an automatic consequence of Bayes' theorem). μ will also be shrunk towards the data in its posterior distribution, so that the θ_i are now shrunk towards a "typical" x value.

Exchangeability and de Finetti's theorem

'Exchangeability' is a formal expression of the idea that we find no systematic reason to distinguish the individual random variables $\theta_1, ..., \theta_I$ – a *judgement* that they are 'similar' but not identical.

It is often an important ingredient in prior modelling.

An infinite sequence of 0/1 random variables $\theta_1, \theta_2, \ldots$ is called (infinitely) exchangeable if any finite subset has a joint distribution that is the same whatever the order in which the variables are written. E.g. $p(\theta_4, \theta_7, \theta_9) = p(\theta_7, \theta_9, \theta_4)$.

Exchangeability and de Finetti's theorem

If the variables are independent Bernoulli(ϕ), they are obviously exchangeable. This remains true if ϕ is random (as in the coin-tossing example, with two biased coins), since e.g.

$$p(\theta_4, \theta_7, \theta_9) = \int_0^1 p(\phi)\phi^{\theta_4}(1-\phi)^{1-\theta_4}\phi^{\theta_7}(1-\phi)^{1-\theta_7}\phi^{\theta_9}(1-\phi)^{1-\theta_9}d\phi$$

(in the case ϕ has a continuous distribution), and this obviously only depends on $\{\theta_4, \theta_7, \theta_9\}$ (in fact only their sum), not the order they appear. The remarkable thing is that the converse of this is true – the only way to get infinitely exchangeable 0/1 random variables is by Bernoulli trials with a fixed or random ϕ. This is (a form of) de Finetti's theorem. There are more general versions of the theorem, not just for 0/1 variables.

Exchangeability and de Finetti's theorem

It gives mathematical support for using hierarchical models: if your prior beliefs about a set of parameters (e.g. the hospital mortality rates $\{\theta_i\}$) are exchangeable (really just a symmetry assumption), then without loss of generality you can model them as i.i.d. from some distribution given ϕ, and then make ϕ random.

$$p(\theta_1, \theta_2, \ldots, \theta_I) = \int p(\phi) \prod_{i=1}^{I} p(\theta_i|\phi) d\phi$$

Thus, under broad conditions an assumption of exchangeable units is mathematically equivalent to assuming the θ's are drawn at random from some population distribution.

What else do hierarchical models address?

Real data about real systems are complex: classic statistical methods are not enough. Among the features that real data might have that we could begin to handle are:

- repeated measures,
- heterogeneity between individuals,
- explanatory variables at individual and group level,
- measurement errors, multiple instruments,
- missing data, informative censoring,
- spatial or temporal structure.

Summary: why hierarchical?

Many interlinked arguments to favour the use of hierarchical models:

- by breaking down the problem in layers, able to separate structural judgments on observables, on parameters and subjective information
- reduces the arbitrariness of hyperparameter choice $\rightarrow$ "robustify" the inference
- natural structure for expressing dependence, prior correlations, ... in a plausible way (see next lectures)
- through shrinkage and borrowing of strength, parameter estimates are stabilised
- by de Finetti, if our beliefs are exchangeable, then they can be expressed mathematically by a hierarchical model.

Hidden Markov models and State space models

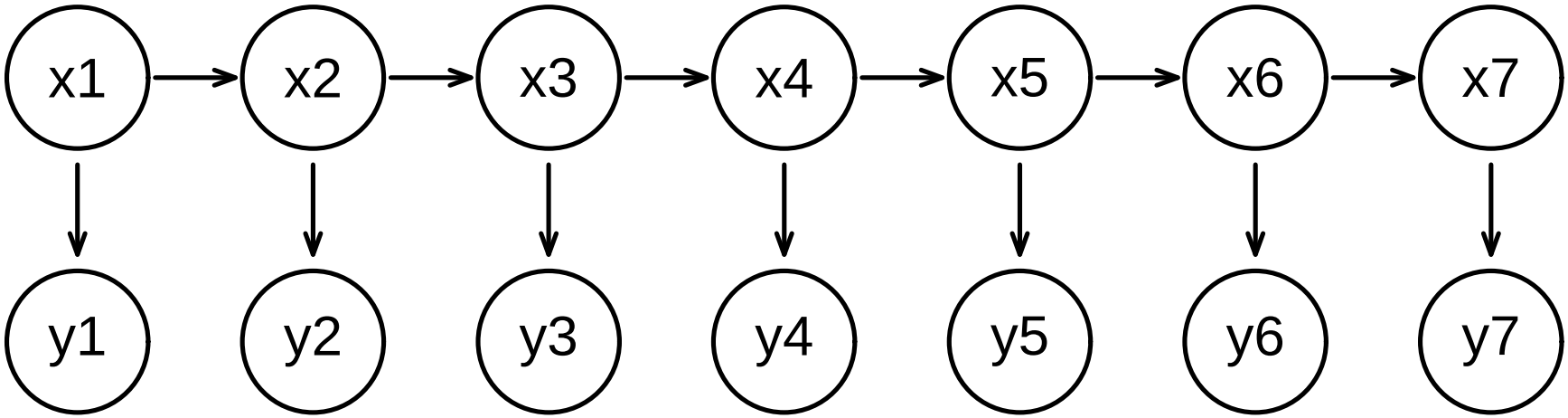
These classes of models, which are often treated as distinct, are really two flavours of a similar idea, and they collectively provide a flexible way of modelling dependent random systems evolving in 'time', where 'time' may be time, or linear position, or location in a genome, or ... The linear structure means that short cuts can often be taken in computing inferences.

The key idea is that there are two sequences – a 'hidden' one $\{x_t, t = 0, 1, 2, \ldots\}$ and an observed one $\{y_t, t = 0, 1, 2, \ldots\}$. The structure in the system is provided by assuming that $\{x_t\}$ is a Markov chain, and the simplicity from assuming that only x_t has a *direct* influence of y_t; loosely, y_t is a 'noisy version' of x_t.

What this means essentially is that the dependence is in the system, rather than the observation process, and this is realistic in very many applications – and often convenient computationally.

Hidden Markov models and State space models

In the language of graphical modelling, these models are represented by this generic directed acyclic graph:



Hidden Markov models

In HMM's as usually defined, the distinctive feature is that x_t has a finite state space. Sometimes the states are known, but not always. Usually, the transition probabilities between the states are unknown. In a *finite* HMM, the values of y_t are also in a finite set, so that everything is discrete. In a *normal* HMM, the distribution of y_t given x_t is normal.

Examples.

- communication channels
- DNA and protein sequencing
- ion channels
- speech recognition

State space models

As usually defined, the term state space models covers cases where the process x_t is continuous-valued. The classic version is the gaussian linear state space model

$$\begin{aligned} x_{t+1} &= ax_t + ru_t \\ y_t &= bx_t + sv_t \end{aligned}$$

where a, r, b, s are constants (known or unknown) and u_t and v_t are independent sequences of i.i.d. normal random variables. In many applications, these quantities are all vectors and matrices.

State space models

Examples.

- automatic control, signal processing
- time series, econometrics
- tracking

Filtering, smoothing and prediction

In many applications, where t represents actual time, we will wish to make *online* inference, that is to report what we know about the x process immediately after each y_t is observed. *Filtering* refers to estimating x_t, given $y_{\leq t} \equiv y_t, y_{t-1}, y_{t-2}, \ldots$ *Smoothing* refers to estimating x_s for some $s < t$, given $y_{\leq t}$. *Prediction* refers to estimating x_s for some $s > t$, given $y_{\leq t}$.

Even when there is no requirement to do inference online, it may still be an attractive option since it may be much cheaper to compute (say) $p(x_t|y_{\leq t})$ (filtering) than $p(x_t|\text{all } y)$, although this means throwing information away.

Kalman filtering

For the gaussian linear state space model, and vector generalisations of it, there is a well-known and long-standing algorithm called the Kalman filter for computing $p(x_t|y_{\leq t})$; because in a multivariate normal (gaussian) distribution, all conditional distributions are also normal, all that the algorithm needs to do is compute the mean and variance of the filtering distribution, that is $m_t = E(x_t|y_{\leq t})$ and $w_t = \text{var}(x_t|y_{\leq t})$.

These can be calculated by the following recursion:

$$\begin{aligned} m_t &= \frac{s^2 a m_{t-1} + (a^2 w_{t-1} + r^2) b y_t}{s^2 + (a^2 w_{t-1} + r^2) b^2} \\ w_t &= \frac{s^2 (a^2 w_{t-1} + r^2)}{s^2 + (a^2 w_{t-1} + r^2) b^2} \end{aligned}$$

Kalman filtering and variants

There are many different equivalent ways of writing this, and of course in the vector case the expressions involve matrices and look more complicated.

The Kalman filter is probably one of the most-often used algorithms in the whole of electronic engineering.

If the state-space model is not gaussian, and/or not linear, there is no general recursive formula for $p(x_t|y_{\leq t})$. Various adaptations of the idea have been devised to solve the filtering problem approximately. In recent years, the idea of *particle filtering*, where the distributions are represented by large random samples, and the calculations are all done by simulation, has become very popular.

Forwards/backwards recursions

We can read off the joint distribution of all variables from the DAG: letting $x = (x_0, x_1, \ldots, x_T)$ and $y = (y_1, y_2, \ldots, y_T)$ (note that we begin x at $t = 0$), we have

$$p(x, y) = p(x_0) \prod_{t=1}^{T} [p(x_t|x_{t-1})p(y_t|x_t)]$$

(proportional as a function of x), assuming there are no unknown parameters. Once the data are observed, they are fixed, so let us remove them from the notation, and abbreviate: $g_1(x_0, x_1) = p(x_0)p(x_1|x_0)p(y_1|x_1)$ and $g_t(x_{t-1}, x_t) = p(x_t|x_{t-1})p(y_t|x_t)$, for $t = 2, \ldots, T$, then

$$p(x, y) = \prod_{t=1}^{T} g_t(x_{t-1}, x_t)$$

Forwards/backwards recursions

To calculate $p(x_t|y) = p(x_t, y)/p(y)$ we need to sum this over all values of $x_0, x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T$, i.e.

$$p(x_t, y) = \sum_{x_0} \cdots \sum_{x_{t-1}} \sum_{x_{t+1}} \cdots \sum_{x_T} \prod_{t=1}^{T} g_t(x_{t-1}, x_t)$$

We can permute the order of the sums and products to find that the right hand side is the same as $r_t(x_t)s_t(x_t)$ where

$$r_t(x_t) = \sum_{x_{t-1}} g_t(x_{t-1}, x_t) \sum_{x_{t-2}} g_{t-1}(x_{t-2}, x_{t-1}) \ldots$$

and

$$s_t(x_t) = \sum_{x_{t+1}} g_{t+1}(x_t, x_{t+1}) \sum_{x_{t+2}} g_{t+2}(x_{t+1}, x_{t+2}) \ldots$$

Forwards/backwards recursions

But note the recursive structure:

$$r_t(x_t) = \sum_{x_{t-1}} g_t(x_{t-1}, x_t) r_{t-1}(x_{t-1}) \quad \text{and}$$

$$s_t(x_t) = \sum_{x_{t+1}} g_{t+1}(x_t, x_{t+1}) s_{t+1}(x_{t+1})$$

So we can make an enormous saving of computing effort by performing these two recursions, starting from $r_0(x_0) \equiv 1$ and $s_T(x_T) \equiv 1$. Having found all the r_t and s_t functions, you then just set

$$p(x_t|y) = \frac{r_t(x_t)s_t(x_t)}{\sum_{x_t} r_t(x_t)s_t(x_t)}$$

Forwards/backwards recursions

This argument can be easily modified for specific filtering, smoothing or predicting tasks, for example,

$$p(x_t|y_{\leq t}) = \frac{r_t(x_t)}{\sum_{x_t} r_t(x_t)}.$$

There are also modifications in the same spirit to deal with other (static) parameters, and with maximising or sampling rather than marginalising.

There are also analogous algorithms for exact probability calculations on certain other graphs with discrete variables – trees and junction trees, more general than linear chains.

More general lessons

HMMs and state space models are classic examples of a principle with much wider application – possibilities for flexible modelling of dependence in data through

- a (hidden) latent dependent process, indexed in time, space, . . . , with detailed structure chosen to facilitate inference
- data distributed as conditionally independent given the latent process, reflecting appropriate distributional assumptions for the context

– generates powerful classes of dependent mixture models useful in many domains.