

AN INTRODUCTION TO  
BAYESIAN INFERENCE APPLIED TO SIGNAL AND DATA  
PROCESSING

W J FITZGERALD

Signal Processing Laboratory, Department of Engineering,  
University of Cambridge,  
Cambridge CB2 1PZ, UK.  
`wjf@eng.cam.ac.uk`

16/5/2001



UNIVERSITY OF  
CAMBRIDGE

## OVERVIEW

- Introduction to the Main Points of the Talk
- The Nature of Experimental Data
- Problems with Experimental Science
- Bayesian Statistical Modelling
- Scientific Inference
- Bayesian Treatment of Model Uncertainty
  - *M-closed*, *M-completed* and *M-open* Models

- Parameter Estimation and Model Selection
- Model Selection and Evidence
  - A simple example which demonstrates Ockham's Razor
- Decision Theory and Utility
- Estimation of a DC level in Noise
- Reliabilities of Estimators and Examples

- Parameter Estimation, Model Selection and Marginal Distributions
- Specifying Priors (Conjugate, Uninformative and Improper Priors)
- Sequential and Block-based Data analysis
- Parameter Estimation
- The General Linear Model
- Frequency Estimation, Spectral Analysis and the DFT

- Generalised Changepoint Detection
  - The Step
  - The Ramp
  - Applications to Communications - BPSK and QPSK
  - Mixed Models
  
- Model Selection for L.I.T.P. models
  - Evidence for AR model order (Uniform, Gaussian and Smoothness Priors)
  - Evidence for Polynomial Models
  - A few words about SVD, KLT, Complexity Reduction and Classification

- Bayesian Numerical Methods
  - Basic Introduction to Monte Carlo Methods
  - Importance Sampling
  - Bayesian Linear Regression
  - Analytic expressions
  - The Gibbs Sampler
  - Applications of the Gibbs Sampler to Linear Regression
  
- Markov Chain Monte Carlo (MCMC) methods
  - Metropolis Hastings (Gibbs sampler a special case)
  - Reversible Jump (and Model Selection)
  - Statistical Mechanics and MCMC - Hybrid Monte Carlo

- Applications
  - Interpolation of Missing Samples
  - Ion-Channel Changepoint Detection
  - Spatial Beamforming
  - Decaying Exponentials
  - High Resolution Spectral Estimation
  - Image Scratch Detection and Removal
  - Cyclostationarity and Communications
  - $\alpha$ -stable Distributions and Mixture Models
  - MCMC in Extreme Value Statistics
  - Tracking and Sequential MCMC
  - Sequential methods in Communications

## INTRODUCTION TO THE MAIN POINTS OF THE TALK

- **Parameter Estimation**

At the first level it is assumed that one of the models within a chosen set, is the correct (or appropriate) model with which to interpret the data and the problem of inference at this level consists of extracting values for the free parameters of the model, given the observed data.

Bayes' theorem may be used to express the posterior probability of the parameters as:

$$P(\boldsymbol{\theta}_k | \mathbf{D}, \mathcal{M}_k) = \frac{P(\mathbf{D} | \boldsymbol{\theta}_k, \mathcal{M}_k) P(\boldsymbol{\theta}_k | \mathcal{M}_k)}{P(\mathbf{D} | \mathcal{M}_k)} \quad (1)$$

where  $\mathcal{M}_k$  represents the  $k^{th}$  model,  $\mathbf{D}$  is a vector of observed data,  $\boldsymbol{\theta}_k$  is a vector of model parameters.  $P(\boldsymbol{\theta}_k | \mathbf{D}, \mathcal{M}_k)$  is the *Posterior probability*

and  $P(\mathbf{D}|\boldsymbol{\theta}_k, \mathcal{M}_k)$  is the *Likelihood*.  $P(\boldsymbol{\theta}_k | \mathcal{M}_k)$  is the *Prior probability* of the parameters  $\boldsymbol{\theta}_k$  before the data were observed and  $P(\mathbf{D} | \mathcal{M}_k)$  is a quantity called the *Evidence* which at this level is a constant.

- **Marginalisation**

In many problems it may not be required to estimate all elements of the model parameter vector  $\boldsymbol{\theta}_k$  and parameters which are of no interest are known as *Nuisance parameters*. A powerful feature of the Bayesian framework is that the *Nuisance parameters* may be removed from consideration by integration, a process called *marginalisation*.

The parameter vector may be partitioned into 2 sets of parameters:  $\boldsymbol{\theta}_{k,1}$ , which are the parameters of interest and  $\boldsymbol{\theta}_{k,2}$ , which are the *Nuisance parameters*.

The posterior probability can now be rewritten as:

$$P(\boldsymbol{\theta}_k | \mathbf{D}, \mathcal{M}_k) \equiv P(\boldsymbol{\theta}_{k,1}, \boldsymbol{\theta}_{k,2} | \mathbf{D}, \mathcal{M}_k)$$

And the posterior marginal density of the parameters of interest,  $\boldsymbol{\theta}_{k,1}$ , may be obtained from

$$P(\boldsymbol{\theta}_{k,1}|\mathbf{D}, \mathcal{M}_k) = \int_{\mathcal{R}_{\parallel}} P(\boldsymbol{\theta}_{k,1}, \boldsymbol{\theta}_{k,2}|\mathbf{D}, \mathcal{M}_k) d\boldsymbol{\theta}_{k,2}$$

## • General Linear Model

Suppose the observed data may be described by a model of the form:

$$d(n) = \sum_{q=1}^Q b_q g_q(n) + e(n) \quad \text{for} \quad 1 \leq n \leq N$$

where  $g_q(n)$  is the value of a time dependent model function  $g_q(t)$  evaluated at time  $t_n$ , represented by integers,  $n$ , for uniform sampling.

This can be written in the form of a matrix equation:

$$\mathbf{D} = \mathbf{G} \mathbf{b} + \mathbf{e} \tag{2}$$

where:  $\mathbf{D}$  is an  $N \times 1$  vector of data points,  $\mathbf{e}$  is an  $N \times 1$  vector of noise samples,  $\mathbf{G}$  is an  $N \times Q$  matrix whose columns are the basis functions

evaluated at each point in the time series and  $\mathbf{b}$  is a  $Q \times 1$  linear coefficient vector.

Many of the *standard* signal processing model structures can be represented with the General Linear Model: the Sinusoidal Model, the Autoregressive (AR) Model, the Autoregressive with External Input (ARX) Model, the Nonlinear Autoregressive (NAR) Model, the Volterra Model, the Radial Basis Function Model etc.

- Marginal Posterior Distribution

$$P(\boldsymbol{\theta}_Q | \mathbf{D}, \mathcal{M}_Q) \propto \frac{\left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{\frac{-(N-Q)}{2}}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}} \quad (3)$$

## MODEL SELECTION

- There are numerous statistical inference problems that require model selection with respect to a set of competing models,  $\{\mathcal{M}_k\}$ ,  $k = 1, \dots, K$ . This is the basis of the second level of inference.
- In many cases, the optimal model choice is taken as the one that maximises  $\arg \max_k P(\mathcal{M}_k | \mathbf{D})$ .

Since

$$P(\mathcal{M}_k | \mathbf{D}) \propto P(\mathbf{D} | \mathcal{M}_k)P(\mathcal{M}_k), \quad (4)$$

one can readily see that the central element of the Bayesian model selection procedure is the evaluation of the quantity  $P(\mathbf{D} | \mathcal{M}_k)$ , referred to as the marginal likelihood, integrated likelihood or *model evidence*.

- Denoting  $\boldsymbol{\theta}_k$  to be a vector of parameters under model  $\mathcal{M}_k$ ,  $P(\mathbf{D} | \mathcal{M}_k)$  can be written as

$$P(\mathbf{D} | \mathcal{M}_k) = \int P(\mathbf{D} | \boldsymbol{\theta}_k, \mathcal{M}_k) P(\boldsymbol{\theta}_k | \mathcal{M}_k) d\boldsymbol{\theta}_k. \quad (5)$$

- The Bayes factor  $BF_{a,b}(\mathbf{D})$  for model  $\mathcal{M}_a$  against model  $\mathcal{M}_b$ , which is used extensively in Bayesian hypothesis testing and model selection, is defined to be the ratio of the respective evidences of the two models, namely,

$$BF_{a,b}(\mathbf{D}) \triangleq \frac{P(\mathbf{D} | \mathcal{M}_a)}{P(\mathbf{D} | \mathcal{M}_b)} \quad (6)$$

The Bayes factor can be interpreted as providing a measure of how much the data  $\mathbf{D}$  have increased or decreased the odds on  $\mathcal{M}_a$  relative to  $\mathcal{M}_b$ , and is

given by,

$$BF_{a,b}(\mathbf{D}) = \frac{P(\mathcal{M}_a | \mathbf{D})}{P(\mathcal{M}_b | \mathbf{D})} / \frac{P(\mathcal{M}_a)}{P(\mathcal{M}_b)}. \quad (7)$$

$BF_{a,b}(\mathbf{D}) > 1$  signifies that in relative terms model  $\mathcal{M}_a$  is more plausible in the light of  $\mathbf{D}$ . The posterior odds ratio  $P(\mathcal{M}_a | \mathbf{D})/P(\mathcal{M}_b | \mathbf{D})$  provides a summary of the evidence for  $\mathcal{M}_a$  against  $\mathcal{M}_b$ .

## INTRODUCTION TO NUMERICAL METHODS

- In the last sections it has been shown how to conduct inference in situations where the integrals required to perform marginalisation have been analytically tractable. However, in more realistic situations the integrations have to be performed numerically and this requirement has led to a full scale investigation of such integration methods.
- Three areas of Bayesian analysis that require evaluation of integrals are:

$$P(\mathbf{D} \mid \mathcal{M}_k) = \int P(\mathbf{D} \mid \boldsymbol{\theta}_k, \mathcal{M}_k)P(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k. \quad (8)$$

$$P(\boldsymbol{\theta}_{k,1}|\mathbf{D}, \mathcal{M}_k) = \int_{\mathcal{R}_{\parallel}} P(\boldsymbol{\theta}_{k,1}, \boldsymbol{\theta}_{k,2}|\mathbf{D}, \mathcal{M}_k) d\boldsymbol{\theta}_{k,2}$$

$$\mathbb{E}_{\pi}(f) = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

## NUMERICAL INTEGRATION TECHNIQUES

- Numerical Quadrature methods
- Saddlepoint approximations and Edgeworth series
- Laplaces approximation
- Monte Carlo methods
  - Static methods
    - ◇ Rejection method
    - ◇ Importance sampling
  - Dynamic methods
    - ◇ Markov Chain Monte Carlo (MCMC)

◇ Metropolis-Hastings, Gibbs sampler, Reversible jump MCMC ..

- Sequential MCMC and Particle Filters

Many problems of interest can be viewed as a hidden dynamical system observed sequentially by a second uncertain process:

○

$$x_k = f(x_{k-1}, u_k, \Delta t_k)$$

$x_k$  is the hidden state at iteration  $k$ ,  $f(., .)$  is the transition function,  $\Delta t_k$  is the time between iterates  $k$  and  $k - 1$ ,  $u_k$  is process noise.

○

$$y_k = h(x_k, v_k)$$

$y_k$  is an observation at iteration  $k$ ,  $h(., .)$  is the measurement function,  $v_k$  is measurement noise.

- The objective is to specify and update the posterior distribution  $p(x_{0:k}|y_{1:k})$ . Using Bayes rule we can express this sequentially as:

$$p(x_{0:k}|y_{1:k}) = \frac{p(y_k|x_k)p(x_k|x_{k-1})}{p(y_k|y_{1:k-1})}p(x_{0:k-1}|y_{1:k-1})$$

Suppose  $\theta^1, \dots, \theta^n$  are an i.i.d. sample set from  $\pi$  and that we evaluate

$$\hat{f} = \frac{1}{n} \sum_{i=1}^n f(\theta^i). \quad (9)$$

Then  $\hat{f}$  is said to be a Monte Carlo estimator of  $\mathbb{E}_\pi(f)$ . Since

$$\mathbb{E}\hat{f} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}f(\theta^i) = \mathbb{E}_\pi(f), \quad (10)$$

$\hat{f}$  is clearly an unbiased estimator. and the variance is given by

$$\text{var}(\hat{f}) = \frac{1}{n} \int [f(\boldsymbol{\theta}) - \mathbb{E}_\pi(f)]^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (11)$$

## THE NATURE OF EXPERIMENTAL DATA

All scientific investigations of the so-called real world essentially involve three related parts, after deciding what it is that one wishes to investigate;

- The design of the experimental measuring apparatus.
- The measurement process itself.
- The analysis of the gathered data - (If it isn't archived and forgotten !!)

We desire to make propositions about the real world, which we believe to be either true or false. (It is sometimes the case that data are analysed without thought ever being given to the underlying hypothesis that one is actually trying to infer!) - (Some data are VERY expensive to gather and one shouldn't be afraid of computer time !)

## PROBLEMS WITH EXPERIMENTAL SCIENCE

One of the major problems with experimental science is that very rarely can a series of logical deductions be made from data to hypothesis. This is because of experimental uncertainties, usually in the form of added noise.

In these situations we have to reason as best we can in situation of incomplete information - this is called **Scientific Inference**.

R.T. Cox, in 1946, showed that any method of inference which satisfies simple rules for a) *Logical* and b) *Consistent Reasoning* must be equivalent to the use of ordinary probability theory, as originally formulated by Bernoulli, Bayes and Laplace.

$$p(H|\mathbf{D}, I) = \frac{p(H|I)p(\mathbf{D}|H, I)}{p(\mathbf{D}|I)}$$

## BAYESIAN STATISTICAL MODELLING

- Most of what is referred to as *Statistical Modelling* is encapsulated in the *Likelihood* function. This provides the probability of making a set of observations under a fixed model given some parameter values.
- In the **Bayesian Framework** the *Likelihood* is combined with *Prior* information to produce a *Posterior* distribution.
- The *Prior* information generally takes the form of probabilities and probability distributions expressing prior belief in parameter values and model hypotheses.
- The *Posterior* distribution is the modified expression of these *Conditional* on the observed data.

- When the data are sufficiently informative and are in reasonable agreement with the Priors, the Posterior distribution is not sensitive to the details of the prior.
- With sparse data the prior dominates and vice versa.

## SCIENTIFIC INFERENCE

- Accordingly, the *conditional probability density function* (*p.d.f.*),  $p(H|\mathbf{D}, \mathbf{I})$  summarises our inference about the hypothesis  $H$  given the data  $\mathbf{D}$  and our prior knowledge  $I$  about  $H$  and the experimental setup. Since the numerical value of the probability assigned to any particular  $H$  is a measure of how much we believe that it is the true hypothesis, a natural estimate is given by that  $H$  which maximises  $p(H|\mathbf{D}, \mathbf{I})$ .
- The width, or spread, of this p.d.f. about the maximum tells us the reliability of the estimate: if the p.d.f. is sharply-peaked then we are confident of our prediction, but if it is broad then we are fairly uncertain about the true hypothesis.

## BAYESIAN TREATMENT OF MODEL UNCERTAINTY

- Many scientific disciplines attempt to find a model that satisfactorily predicts the available (and possibly future) observations.
- In a conventional parametric representation, a model  $\mathcal{M}$  in the Bayesian sense has two components, a likelihood  $p(\mathbf{D} \mid \boldsymbol{\theta}, \mathcal{M})$  and a prior  $p(\boldsymbol{\theta} \mid \mathcal{M})$ , where  $\mathbf{D}$  denotes observed data and  $\boldsymbol{\theta}$  denotes a vector of parameters under  $\mathcal{M}$ .
- Together they constitute one's predictive beliefs for sequences of observables, namely,

$$p(\mathbf{D} \mid \mathcal{M}) = \int p(\mathbf{D} \mid \boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta} \mid \mathcal{M}) d\boldsymbol{\theta}. \quad (12)$$

- Conditional upon the adequacy of the model, Bayes' theorem provides a mechanism for posterior or predictive inferences from the data.

## MODEL UNCERTAINTY

- A general inference problem involves two levels of inference, one accounting for model uncertainty and the other accounting for uncertainty about parameters in a particular model formulation. The Bayesian paradigm offers a framework within which both aspects can be considered in a coherent fashion.
- Statistical models are in many cases nothing but simplifications of complicated phenomena, and as such may not necessarily represent the true underlying process.
- From a subjectivist Bayesian standpoint, a model embodies one's understanding of the underlying process in such a way that is useful for some purpose, such as a parsimonious description of the salient features, or the ability to predict future events.

- When there is uncertainty about whether a particular model best serves a specific purpose, or from a decision-theoretical point of view yields the greatest expected utility for a specific utility function, a range of possible models should to be considered.
- Once these models have been specified, one can then proceed to perform model criticism on these models, in terms of their explanatory or predictive capacity and adequacy for the intended purpose, given the data.
- It is necessary to draw clear distinctions between different perspectives on those possible models in the presence of model uncertainty - there are three alternative ways of viewing them.
  - The first, which are called the *M-closed* view, corresponds to believing that one of the models is true, though without the explicit knowledge of which one is true. This will be appropriate if one knows for certain that the real world mechanism is amongst a finite set of specified models.

- Reality is however rarely quite as straightforward. Nature typically does not offer an exhaustive list of possible mechanisms together with a guarantee that one of them is true.
  - The second alternative is are called the *M-completed* view, corresponding to the case where although an individual has separately formulated a concrete belief model, due to intractability of analysis or other factors, one elects to use other models, which are to be evaluated on the basis of the former.
  - The third alternative, called the *M-open* view, acknowledges that the considered models are merely a range of specified models available for comparison. In this case, however, there is no separate actual belief model. The *M-open* view is perhaps the closest to what one typically adopts in practice when one lacks the necessary knowledge or expertise to be able to accept a ‘true’ model.
- 
- From both the *M-completed* perspective and the *M-open* perspective,

assigning the prior probabilities  $\{p(\mathcal{M}_k)\}$  does not, strictly speaking, make sense. A non-decision-theoretic Bayesian approach to model comparison will therefore be problematic under conditions in which these two perspectives are appropriate.

- Issues arising from model uncertainty are best formulated within a decision-theoretic framework.

## THE BAYESIAN APPROACH

- A Bayesian model is described by the prior distribution  $p(\theta)$  of a random parameter  $\theta \in \Theta$  and by the likelihood  $p(\mathbf{y}|\theta)$  of the observations  $\mathbf{y}$ .
- In this framework, all information on  $\theta$  based on the observations  $\mathbf{y}$  is included in the posterior distribution  $p(\theta|\mathbf{y})$  which one can obtain using Bayes' theorem

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}$$

where the normalizing constant  $p(\mathbf{y})$  is obtained by integration

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\theta) p(\theta) d\theta$$

- The information given by the posterior distribution might be too complicated to analyse directly when one is confronted with a decision problem. This can however be naturally handled in a Bayesian framework.

## DECISION THEORY AND UTILITY (COST) FUNCTIONS

- For some applications, the desired result is the complete posterior distribution or some probability statement concerning a subset of parameters or function of parameters - In these cases a *Point Estimate* is not required.
- However, in some cases a *Point Estimate* is required - To derive Bayesian *Point Estimates* of unknown random parameters one must specify a *Loss Function* - which measures the loss caused by an estimation error as a function of the parameter estimate and the 'true' (unknown) value - (Quadratic, Absolute, Zero-one loss etc).

$$\text{choose } \hat{\theta} : \min E (L(\hat{\theta}, \theta)) = \int L(\hat{\theta}, \theta) p(\theta | D, I) d\theta$$

## ESTIMATION OF A DC LEVEL IN ADDITIVE (IID) GAUSSIAN NOISE

○

$$d_i = \theta + n_i$$

- Assume the noise is zero mean with variance  $\sigma_n^2$ .
- Assume that the prior for  $\theta$  is Gaussian, centered around  $m$  with variance (hyperparameter)  $\sigma_\theta^2$ .
- Bayes' theorem gives

$$p(\theta|\mathbf{D}) = \frac{p(\mathbf{D}|\theta)p(\theta)}{p(\mathbf{D})}$$

- The Likelihood function is given by

$$p(\mathbf{D}|\theta) = \prod_{i=1}^N p(d_i|\theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(d_i - \theta)^2}{2\sigma_n^2}\right)$$

- The Prior is given by

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{(\theta - m)^2}{2\sigma_\theta^2}\right)$$

- Therefore

$$p(\theta|\mathbf{D}) = \frac{1}{(2\pi)^{\frac{(N+1)}{2}} p(\mathbf{D}) \sigma_n^N \sigma_\theta} \\ \times \exp\left(-\frac{1}{2}\left(\sum_{i=1}^N \frac{(d_i - \theta)^2}{\sigma_n^2} + \frac{(\theta - m)^2}{\sigma_\theta^2}\right)\right)$$

- This may be written as

$$p(\theta|\mathbf{D}) \propto \exp\left(-\frac{(\theta - \bar{\theta})^2}{2\sigma_m^2}\right)$$

which is again a normal probability density function with mean  $\bar{\theta}$  and variance  $\sigma_m^2$  given by

$$\bar{\theta} = \frac{\frac{m}{\sigma_{\theta}^2} + \frac{Nd}{\sigma_n^2}}{\frac{1}{\sigma_{\theta}^2} + \frac{N}{\sigma_n^2}}$$

and

$$\sigma_m^2 = \frac{1}{\frac{1}{\sigma_{\theta}^2} + \frac{N}{\sigma_n^2}}$$

- Writing (for the case of zero prior mean, for convenience)

$$\sigma_m^2 = \frac{\sigma_\theta^2 \sigma_n^2}{N\sigma_\theta^2 + \sigma_n^2}$$

- We obtain

$$p(\theta|\mathbf{D}) = k(d) \exp\left(-\frac{1}{2\sigma_m^2}\left(\theta - \frac{\sigma_m^2}{\sigma_n^2} \sum_{i=1}^N d_i\right)^2\right)$$

- This is another Gaussian density and therefore

$$\hat{\theta}_{MAP}(\mathbf{D}) = \frac{\sigma_m^2}{\sigma_n^2} \sum_{i=1}^N d_i = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \frac{\sigma_n^2}{N}} \left( \frac{1}{N} \sum_{i=1}^N d_i \right)$$

- If  $\sigma_\theta$  is large,

$$\hat{\theta}_{MAP}(\mathbf{D}) \approx \frac{1}{N} \sum_{i=1}^N d_i$$

- It is seen that the posterior mean  $\bar{\theta}$  is a weighted average of the prior mean and the sample mean with weighting factors of  $1/\sigma_{\theta}^2$  and  $N/\sigma_n^2$  respectively. As the prior variance grows and the prior pdf becomes spread out, the posterior mean approaches the sample mean. Also, as  $N$  grows, the same limit is obtained, as common sense would have us believe.

## RELIABILITIES OF ESTIMATORS

- Denoting the quantity of interest by  $\theta$ . The MAP estimate,  $\theta_0$ , is such that

$$\frac{\partial p(\theta|\mathbf{D}, I)}{\partial \theta} \Big|_{\theta_0} = 0$$

(Should also check the sign of the 2nd derivative )

- What is the width of the Posterior distribution ?

- Define

$$L = \ln p(\theta | \mathbf{D}, I)$$

and expand around the maximum using a Taylor series.

- We obtain

$$L = L(\theta_0) + \frac{1}{2} \frac{\partial^2 L}{\partial \theta^2} \Big|_{\theta_0} (\theta - \theta_0)^2 + \dots$$

- There is obviously no linear term since  $\frac{\partial p(\theta|\mathbf{D}, I)}{\partial \theta}|_{\theta_0} = 0$ .
- Therefore we may write

$$p(\theta|\mathbf{D}, I) \approx A \exp\left(\frac{1}{2} \frac{\partial^2 L}{\partial \theta^2}|_{\theta_0} (\theta - \theta_0)^2\right)$$

- This is a Gaussian form and we may write

$$\theta = \theta_0 \pm \sigma$$

where  $\sigma = \left(\frac{-\partial^2 L}{\partial \theta^2}|_{\theta_0}\right)^{-1/2}$ . The generalisation to the multivariate case is straightforward.

## EXAMPLE OF RELIABILITY ESTIMATES

- If we observe  $R$  heads in a coin toss (or number of cases of a disease, etc) and  $N-R$  tails (or number of cases without the disease etc) then

$$p(H|\mathbf{D}, I) \propto H^R(1 - H)^{N-R}$$

where  $0 \leq H \leq 1$  and assuming a uniform prior.

$$L = C + R \ln H + (N - R) \ln(1 - H)$$

- What is the ‘best’ estimate and ‘reliability’ of  $H$  ?

$$\frac{dL}{dH} = \frac{R}{H} - \frac{N - R}{1 - H}$$

$$\frac{d^2L}{dH^2} = -\frac{R}{H^2} - \frac{N - R}{(1 - H)^2}$$

- Therefore

$$H_0 = \frac{R}{N}$$

$$\sigma = \sqrt{\frac{H_0(1 - H_0)}{N}}$$

- If our estimate of  $H_0$  is good, then  $H_0(1 - H_0)$  is constant and

$$\sigma \propto \frac{1}{\sqrt{N}}$$

## PARAMETER ESTIMATION AND MODEL SELECTION

- Posterior Density

$$p(\mathbf{w}|\mathbf{D}, M_i) = \frac{p(\mathbf{D}|\mathbf{w}, M_i) p(\mathbf{w}|M_i)}{p(\mathbf{D}|M_i)}$$

- 

$$p(\mathbf{D}|M_i) = \int p(\mathbf{D}|\mathbf{w}, M_i) p(\mathbf{w}|M_i) d\mathbf{w}$$

- Marginal Density

Partition  $\mathbf{w}$  into 2 sets of parameters:

$\mathbf{w}_1$  - parameters of interest

$\mathbf{w}_2$  - Nuisance parameters

$$p(\mathbf{w}|\mathbf{D}) \equiv p(\mathbf{w}_1, \mathbf{w}_2|\mathbf{D})$$

$$p(\mathbf{w}_1|\mathbf{D}) = \int_{\mathbf{w}_2} p(\mathbf{w}_1, \mathbf{w}_2|\mathbf{D}) d\mathbf{w}_2$$

## MODEL SELECTION AND EVIDENCE

$$p(M_i|\mathbf{D}) = \frac{p(\mathbf{D}|M_i)p(M_i)}{p(\mathbf{D})}$$

- Are there three Decaying Functions or two ? - Chemical kinetics, NMR, etc.
- Are there three Frequency Components or two ? - Radar, Sonar, Speech etc.
- Are there three Diffraction Peaks or two ? - Material science, X-rays etc.
- Are there three Classes or two ? - Classification, Image Segmentation etc.
- Are there three Significant Singular Values or two ? - Data Compression, Chaos etc.

In many cases we can assume

$$p(M_i|\mathbf{D}) \propto p(\mathbf{D}|M_i) = \int p(\mathbf{D}|\mathbf{w}, M_i) p(\mathbf{w}|M_i) d\mathbf{w}$$

and

$$p(M_i|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{w}_{\text{map}}, M_i) p(\mathbf{w}_{\text{map}}|M_i) \Delta\mathbf{w}$$

This gives rise to William of Ockham's razor (d. 1349). (Actually, the first formulation was given by John Ponce of Cork in 1639)

## A SIMPLE EXAMPLE WHICH DEMONSTRATES OCKHAM'S RAZOR

- Consider 2 scientists, Mr A and Mrs B.
- Mr A has a theory which has a certain number of parameters in the model.
- Mrs B also has a theory but with one extra parameter  $\lambda$  say.
- **If both scientists do an experiment together and measure data  $D$ , whose theory should we prefer given the observed data ?**

- Let  $A$  and  $B$  represent the two theories (hypotheses). Then what we need to calculate is

$$\frac{p(A|D)}{p(B|D)} = \frac{p(A)}{p(B)} \times \frac{p(D|A)}{p(D|B)}$$

- If this is  $> 1$  we prefer Mr A's theory, if  $< 1$  we prefer Mrs B's theory.
- The priors  $p(A)$  and  $p(B)$  come into the calculation. We could take the 'track record' of the scientists into account !

- Mrs B has an extra unknown parameter. We can therefore marginalise to obtain

$$p(D|B) = \int p(D, \lambda|B)d\lambda = \int p(D|\lambda, B)p(\lambda|B)d\lambda$$

- We need an approximation to enable us to carry out this integral.
- Mrs B says that  $\lambda$  lies between  $\lambda_{min}$  and  $\lambda_{max}$ , and assigns a uniform prior.
- Also, the best fit value of  $\lambda$  is found to be  $\lambda_0 \pm \delta\lambda$ .

- We can therefore write

$$p(D|B) = \frac{1}{\lambda_{max} - \lambda_{min}} \int p(D|\lambda, B) d\lambda$$

$$= \frac{\delta\lambda}{\lambda_{max} - \lambda_{min}} p(D|\lambda_0, B)$$

- Therefore we have

$$\frac{p(A|D)}{p(B|D)} = \frac{p(A)}{p(B)} \times \frac{p(D|A)}{p(D|\lambda_0, B)} \times \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda}$$

- This expression has terms which express;
  - ◇ Prior preference for A and B.
  - ◇ Best predictions in terms of the Likelihood ratio - this favours B because of the extra parameter.
  - ◇ A penalisation term - only a fraction of B's probabilistic investment will yield a good fit. This is the 'Ockham factor'.

## MODEL SELECTION

- What is it one wants out of a Model ?
  - ‘Physicists’ - some think there are real models ‘out there’ !
  - ‘Engineers’ - consider a model in terms of it’s predictive ability.
- Prediction of data from a set of realizations of measured data.**

- Data fitting
- Time series model selection
- Control and Filtering (Sequential Model selection and Model change)
- Feature selection in classification
- Complexity reduction in ‘Neural Networks’
- For many model structures the ‘evidence’,  $E = p(\mathbf{D}|M_i) \propto p(M_i|\mathbf{D})$ , can be written

$$\ln E = \frac{P - N}{2} \ln(2\pi\sigma_e^2) - \frac{1}{2} \ln(\det(G^t G)) - \frac{N - P}{2} - P \log(2\beta) + p \ln(\phi(\alpha))$$

## SPECIFYING PRIORS

- The fundamental difference between Bayesian and classical statistics is that in Bayesian statistics unknown parameters are treated as random variables, and that the use of Bayes' theorem requires the specification of prior distributions for these parameters.
- Whilst this facilitates the inclusion of genuine prior belief about parameters, the choice of prior distribution cannot be made blindly; considerable care is needed and there are some very substantial issues involved.

## CONJUGATE PRIORS

- Computational difficulties arise in using Bayes' Theorem when it is necessary to evaluate the normalizing constant in the denominator,

$$\int p(\theta)p(x|\theta)d\theta.$$

- For example, suppose  $X_1, \dots, X_n$  are independent Poisson  $P(\theta)$  variables, and our beliefs about  $\theta$  are that it *definitely* lies in the range  $[0, 1]$ , but that all values within that range are equally likely: thus,  $p(\theta) = 1; 0 \leq \theta \leq 1$ .
- Then the normalizing constant is

$$\int_0^1 \exp(-n\theta)\theta^{\sum x_i} d\theta,$$

and this integral, an incomplete gamma function, can only be evaluated numerically.

- So, even simple choices of priors can lead to awkward numerical problems.
- Can we identify a prior distribution for which the posterior distribution is in the same family of distributions as the prior ?
- Such priors are called *conjugate priors*.
- Such priors can make the mathematics much easier.

## USE OF CONJUGATE PRIORS

The use of conjugate priors should be seen for what it is: a convenient mathematical device. However, expression of one's prior beliefs as a parametric distribution is always an approximation. In many situations the richness of the conjugate family is great enough for a conjugate prior to be found which is sufficiently close to one's beliefs for this extra level of approximation to be acceptable. However, if this is not the case, they should not be used just because they made the mathematics easier !

## OBTAINING CONJUGATE PRIORS

- Provided they are not in direct conflict with our prior beliefs, and provided such a family can be found, the simplicity induced by using a conjugate prior is compelling.
- But in what situations can a conjugate family be obtained?
- The only case where conjugates can be easily obtained is for data models within the *exponential family*.

That is,

$$p(x|\theta) = h(x)g(\theta) \exp\{t(x)c(\theta)\}$$

for functions  $h, g, t$  and  $c$  such that

$$\int p(x|\theta)dx = g(\theta) \int h(x) \exp\{t(x)c(\theta)\}dx = 1$$

- This might seem restrictive, but in fact includes the exponential distribution, the Poisson distribution, the one-parameter Gamma distribution, the Binomial distribution and the Normal distribution (with known variance).
- Then, with a prior of  $p(\theta)$ ,

$$\begin{aligned} p(\theta|x) &\propto p(\theta)l(\theta;x) \\ &= p(\theta) \prod_{i=1}^n \{h(x_i)\} g(\theta)^n \exp\left\{\sum_{i=1}^n t(x_i)c(\theta)\right\} \\ &\propto p(\theta)g(\theta)^n \exp\left\{\sum_{i=1}^n t(x_i)c(\theta)\right\}. \end{aligned}$$

- Thus if we choose

$$p(\theta) \propto g(\theta)^d \exp\{bc(\theta)\},$$

- we obtain

$$\begin{aligned} p(\theta|x) &\propto g(\theta)^{n+d} \exp\{c(\theta) [\sum_{i=1}^n t(x_i) + b]\} \\ &= g(\theta)^{\tilde{d}} \exp\{\tilde{b}c(\theta)\}, \end{aligned}$$

giving a posterior in the same family as the prior, but with modified parameters.

## CONJUGATE PRIORS

Likelihood	Prior	Posterior
$x \sim B(n, \theta)$	$Be(p, q)$	$Be(p + x, q + n - x)$
$x_1, \dots, x_n \sim Po(\theta)$	$Ga(p, q)$	$Ga(p + \sum_{i=1}^n x_i, q + n)$
$x_1, \dots, x_n \sim N(\theta, \tau^{-1}), (\tau \text{ known})$	$N(b, c^{-1})$	$N(\frac{cb + n\tau\bar{x}}{c + n\tau}, \frac{1}{c + n\tau})$
$x_1, \dots, x_n \sim Ga(k, \theta) (k \text{ known})$	$Ga(p, q)$	$Ga(p + nk, q + \sum_{i=1}^n x_i)$
$x_1, \dots, x_n \sim Ge(\theta)$	$Be(p, q)$	$Be(p + n, q + \sum_{i=1}^n x_i - n)$
$x \sim NeB(r, \theta)$	$Be(p, q)$	$Be(p + r, q + x - r)$

Table 1: Some Standard Conjugate relationships.

## UNINFORMATIVE PRIOR DISTRIBUTIONS

- Another way of constructing a prior distribution is to use an uninformative or 'flat' prior distribution,  $p(\theta) = \text{constant}$  for all  $\theta$ . Such priors can usually be obtained as limiting cases of conjugate priors.
- Uninformative priors are the usual way of representing ignorance about  $\theta$ .
- It can be argued that they are more objective than a subjectively assessed prior distribution since the later may contain personal bias as well as background information.
- In some applications the amount of prior information available is far less than the information contained in the data. In this situation, there is not much point in worrying about the precise specification of the prior.
- An uninformative prior also has the advantage of keeping the mathematics relatively simple.

- However, there are a number of problems associated with the use of uninformative priors.
- Firstly, if the range of values of  $\theta$  is infinite, then the uninformative prior pdf. does not integrate to 1, - it is said to be an improper prior.

$$p(\theta) \propto 1; \quad \theta \in \mathcal{R}$$

but this cannot be a valid probability density since then

$$\int_{\mathcal{R}} p(\theta) d\theta = \infty.$$

## IMPROPER PRIORS

Is it valid to use a posterior distribution obtained by specifying an improper prior to reflect vague knowledge?

Although there are some further difficulties involved, generally the use of improper prior distributions is considered to be acceptable.

The point really is that if we chose  $c$  (inverse variance, say) to be any value other than zero, we would have obtained a perfectly proper prior and there would have been no problems about the subsequent analysis.

Therefore, we could choose  $c$  arbitrarily close to zero and obtain a posterior arbitrarily close to the one we actually obtained by using the improper prior  $p(\theta) \propto 1$ .

## REPRESENTATIONS OF IGNORANCE

- Attempting to represent ignorance within the standard conjugate analysis of a Normal mean leads to the concept of improper priors.
- But there are more fundamental problems as well.
- If, say, we specify a prior for  $\theta$  of the form  $p(\theta) \propto 1$ , and consider the parameter  $\phi = \theta^2$ , then

$$\begin{aligned} p(\phi) &= p(\theta^2) \times \left| \frac{d\theta}{d\phi} \right| \\ &\propto \sqrt{\phi} \end{aligned}$$

- On the other hand, if we were ignorant about  $\theta$ , we are surely equally ignorant about  $\phi$ , and so might equally have made the specification  $p(\phi) \propto 1$ .

- Thus, prior ignorance as represented by uniformity of belief, does not translate across scales.
- One particular point of view is that specification of prior ignorance *should* be consistent across 1—1 parameter transformations.
- This leads to the concept of ‘Jeffreys’ priors’, based on the concept of Fisher information:

$$I(\theta) = -E \left\{ \frac{d^2 \log p(x|\theta)}{d\theta^2} \right\} = E \left\{ \left( \frac{d \log p(x|\theta)}{d\theta} \right)^2 \right\}.$$

- Then, the Jeffreys’ prior is defined as

$$p_0(\theta) \propto |I(\theta)|^{1/2}.$$

- The consistency is verified in the following sense.
- Suppose,  $\phi = g(\theta)$  is a 1—1 transformation of  $\theta$ . Then, by the change of variable rule

$$\begin{aligned} p(\phi) &\propto p_0(\theta) \times \left| \frac{d\theta}{d\phi} \right| \\ &= I(\theta)^{1/2} \times \left| \frac{d\theta}{d\phi} \right| \end{aligned}$$

But, by definition,  $I(\phi) = I(\theta) \times (d\theta/d\phi)^2$ , so

$$p(\phi) \propto I(\phi)^{1/2}.$$

- Thus, the Jeffreys prior for  $\theta$  has transformed naturally to the Jeffreys prior for  $\phi$ .

- Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ , with  $\mu$  known.
- The likelihood is

$$l = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

and the Fisher Information is

$$\begin{aligned} I_\sigma &= -E\left[\frac{\partial^2 l}{\partial \sigma^2}\right] = -E\left[\frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum (x_i - \mu)^2\right] \\ &= -\frac{n}{\sigma^2} + \frac{3n}{\sigma^2} \end{aligned}$$

- Therefore

$$I_{\sigma} = \left(\frac{2n}{\sigma^2}\right)^{1/2}$$

Hence the Jeffreys prior is  $p(\sigma) \propto 1/\sigma$  and the prior distribution for  $\ln(\sigma)$  is flat.

## SEQUENTIAL OR 'BLOCK-BASED' DATA ANALYSIS ?

- A Data set is denoted by  $\mathbf{D}_k$
- The posterior probability for a hypothesis  $H$  is

$$p(H|\mathbf{D}_k, I) \propto p(\mathbf{D}_k|H, I)p(H|I)$$

- This is a 'one-step' process and all the data are considered collectively.

- Alternatively, we can think of the data sequentially,  $\mathbf{D}_1, \mathbf{D}_2, \dots$
- The posterior  $p(H|\mathbf{D}_1, I)$  can be used as the prior for the next inference ..

$$p(H|\mathbf{D}_1, \mathbf{D}_2, I) \propto p(\mathbf{D}_1, \mathbf{D}_2|H, I)p(H|I)$$

OR, if we condition on  $\mathbf{D}_1$ , we have

$$p(H|\mathbf{D}_1, \mathbf{D}_2, I) \propto p(\mathbf{D}_2|H, \mathbf{D}_1, I)p(H|\mathbf{D}_1, I)$$

- If the data blocks are independent,

$$p(\mathbf{D}_2|H, \mathbf{D}_1, I) \equiv p(\mathbf{D}_2|H, I)$$

Therefore

$$p(H|\mathbf{D}_1, \mathbf{D}_2, I) \propto p(\mathbf{D}_2|H, I)p(H|\mathbf{D}_1, I)$$

And in general (- this is how we can carry out Multi-sensor Data Fusion)

$$p(H|\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, I) \propto p(\mathbf{D}_3|H, I)p(H|\mathbf{D}_1, \mathbf{D}_2, I)$$

## PARAMETER ESTIMATION

(We will first deal with Parameter Estimation (for a fixed Model) and then consider the more difficult task of Model Selection.)

- In an estimation problem one assumes that the model is true for some unknown values of the model parameters, and one explores the constraints imposed on the parameters by the data, using Bayes' theorem.
- The hypothesis space for an estimation problem is therefore the set of possible values of the parameter vector  $\mathbf{w}$ , and it is this vector that will form the *hypothesis* that will be used in Bayes' theorem.
- The data form the sample space, and both the hypothesis space and the sample space may be either discrete or continuous.
- In many real-world problems the observed data may be represented as:

$$d(n) = f(n) + e(n)$$

where  $d(n)$  is a particular observed data value and  $f(n)$  is a parameterised model for the data.

- The term  $e(n)$  is a random term which may be considered as being due to measurement error when observing the data or as error arising from incorrect modelling of the data. In the absence of any explicit knowledge to the contrary, the error term is usually assumed to be a zero-mean white Gaussian process.
- In general the data model  $f(n)$  will be a nonlinear function of the model parameters.

## THE GENERAL LINEAR MODEL

- Any data which may be described in terms of a linear combination of basis functions with an additive noise component satisfies the *General Linear Model*.
- Suppose the observed data may be described by a model of the form:

$$d(n) = \sum_{q=1}^Q b_q g_q(n) + e(n) \quad \text{for} \quad 1 \leq n \leq N$$

where  $g_q(n)$  is the value of a time dependent model function  $g_q(t)$  evaluated at time  $t_n$ .

- This can be written in the form of a matrix equation:

$$\mathbf{D} = \mathbf{G} \mathbf{b} + \mathbf{e}$$

where:

$\mathbf{D}$  is an  $N \times 1$  matrix of data points

$\mathbf{e}$  is an  $N \times 1$  vector of noise samples

$\mathbf{G}$  is an  $N \times Q$  matrix whose columns are the basis functions evaluated at each point in the time series

$\mathbf{b}$  is a  $Q \times 1$  linear coefficient vector.

- Many of the *standard* signal processing structures are representatives of the general linear model structure and some examples are:

- Sinusoidal Model

$$d(n) = \sum_{q=1}^Q \{a_q \sin(\omega_q t_n) + b_q \cos(\omega_q t_n)\} + e(n)$$

- Autoregressive (AR) Model

$$d(n) = \sum_{q=1}^Q a_q d(n - q) + e(n)$$

- Autoregressive with External Input (ARX) Model

$$d(n) = \sum_{q=1}^Q a_q d(n-q) + \sum_{q=1}^Q b_q u(n-q) + e(n)$$

where  $u(n)$  is a known system input.

- Nonlinear Autoregressive (NAR) Model

$$\begin{aligned} d(n) &= \sum_{q=1}^Q a_q d(n-q) \\ &+ \sum_{q_1=1}^{Q_1} \sum_{q_2=1}^{Q_2} a_{q_1 q_2} d(n-q_1) d(n-q_2) \\ &+ \dots + e(n) \end{aligned}$$

- Volterra Model

$$\begin{aligned}d(n) &= \sum_{q=0}^Q a_q u(n - q) \\ &+ \sum_{q_1=0}^{Q_1} \sum_{q_2=0}^{Q_2} a_{q_1 q_2} u(n - q_1) u(n - q_2) \\ &+ \dots + e(n)\end{aligned}$$

- The error term or innovations process can sometimes be assumed to be a zero-mean white Gaussian process  $N(0, \sigma^2)$  defined by the probability density:

$$p(\mathbf{e}) = (2 \pi \sigma^2)^{-\frac{N}{2}} \exp \left[ -\frac{\mathbf{e}^T \mathbf{e}}{2 \sigma^2} \right]$$

- The data likelihood may be written as:

$$p(\mathbf{D} \mid \mathbf{w}_m, \sigma, \mathbf{b}, M) = (2 \pi \sigma^2)^{-\frac{N}{2}} \exp \left[ -\frac{\mathbf{e}^T \mathbf{e}}{2 \sigma^2} \right]$$

where  $\mathbf{w}_m$  denotes the parameters of the matrix of basis functions  $\mathbf{G}$ .

- Substituting for the general linear model gives:

$$\therefore p(\mathbf{D}|\mathbf{w}_m, \mathbf{b}, \sigma, M) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{(\mathbf{D} - \mathbf{G}\mathbf{b})^T(\mathbf{D} - \mathbf{G}\mathbf{b})}{2\sigma^2}\right]$$

- Application of Bayes' theorem gives the joint posterior probability density function for the various parameters as:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma|\mathbf{D}, M) = \frac{p(\mathbf{D}|\mathbf{w}_m, \mathbf{b}, \sigma, M) p(\mathbf{w}_m, \mathbf{b}, \sigma|M)}{p(\mathbf{D}|M)}$$

- Assuming that  $\mathbf{w}_m$ ,  $\mathbf{b}$  and  $\sigma$  are statistically independent:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma | M) = p(\mathbf{w}_m | M) p(\mathbf{b} | M) p(\sigma | M)$$

and the joint posterior density becomes:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma | \mathbf{D}, M) = \frac{p(\mathbf{D} | \mathbf{w}_m, \mathbf{b}, \sigma, M) p(\mathbf{w}_m | M) p(\mathbf{b} | M) p(\sigma | M)}{p(\mathbf{D} | M)}$$

- Often there will be little prior knowledge concerning the noise variance  $\sigma^2$  and it is then appropriate to use the Jeffreys' prior:

$$p(\sigma | M) = \frac{1}{\sigma}$$

- Similarly, in the absence of any prior knowledge concerning the model parameters uniform priors may be used.

$$p(\mathbf{w}_m|M) = k_w \quad \text{and} \quad p(\mathbf{b}|M) = k_b$$

- Thus the posterior probability becomes:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma|\mathbf{D}, M) = \frac{k_w k_b p(\mathbf{D}|\mathbf{w}_m, \mathbf{b}, \sigma, M)}{\sigma p(\mathbf{D}|M)}$$

- It should be noted that both the Jeffreys' prior and the uniform prior are *improper* in the sense that they are not normalised and this can raise both philosophical and practical problems.

- If it is assumed that the solution required to the parameter estimation problem is the maximum of the posterior density function (i.e. MAP estimate) then the constant scaling terms and the evidence term  $p(\mathbf{D}|M)$  are independent of the parameter values and the posterior may conveniently be written as:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma | \mathbf{D}, M) = k p(\mathbf{D} | \mathbf{w}_m, \mathbf{b}, \sigma, M)$$

where:

$$k = \frac{k_w k_b}{\sigma p(\mathbf{D}|M)}$$

and its actual value is not required for the MAP solution.

- Substituting for the data likelihood from above gives:

$$p(\mathbf{w}_m, \mathbf{b}, \sigma | \mathbf{D}, M) = k (2 \pi \sigma^2)^{-\frac{N}{2}} \times \exp \left[ -\frac{(\mathbf{D} - \mathbf{G}\mathbf{b})^T (\mathbf{D} - \mathbf{G}\mathbf{b})}{2 \sigma^2} \right]$$

- The procedure now depends on the particular problem under consideration.
- In some cases the matrix  $\mathbf{G}$  is completely specified so the parameter values  $\mathbf{w}_m$  are known. Since  $\mathbf{w}_m$  are no longer considered as random parameters they do not appear in the argument for the posterior density.

$$p(\mathbf{w}_m, \mathbf{b}, \sigma | \mathbf{D}, M) \equiv p(\mathbf{b}, \sigma | \mathbf{D}, M)$$

- The objective, in this case, is to estimate the coefficient vector  $\mathbf{b}$ . The noise standard deviation  $\sigma$  can sometimes be considered to be a nuisance parameter which may be integrated out

$$p(\mathbf{b}|\mathbf{D}, M) = \int_0^\infty p(\mathbf{b}, \sigma|\mathbf{D}, M) \frac{d\sigma}{\sigma}$$

using the Gamma integral definition:

$$\int_0^\infty x^{\alpha-1} \exp(-Qx) dx = \frac{\Gamma(\alpha)}{Q^\alpha}$$

we obtain:

$$p(\mathbf{b}|\mathbf{D}, M) = \frac{1}{2} \Gamma\left(\frac{N-1}{2}\right) \left[ \frac{1}{2} (\mathbf{D} - \mathbf{G}\mathbf{b})^T (\mathbf{D} - \mathbf{G}\mathbf{b}) \right]^{-\frac{N-1}{2}}$$

- The MAP estimate for the parameter vector  $\mathbf{b}$  may be obtained by maximising the above to give:

$$\hat{\mathbf{b}} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D}$$

This is the familiar least squares (minimum variance) estimate that is obtained from linear algebra.

- A second application of the posterior probability, is in determining the MAP estimate of the elements  $\mathbf{w}_m$  of the model matrix  $\mathbf{G}$  without inferring values for the coefficient vector  $\mathbf{b}$  and the noise standard deviation  $\sigma$ .

- The coefficient vector  $\mathbf{b}$  may be marginalised by integrating the posterior probability density, with respect to  $\mathbf{b}$ .

$$p(\sigma, \mathbf{w}_m | \mathbf{D}, M) = \int_{\mathcal{R}^Q} p(\{\mathbf{b}, \mathbf{w}_m, \sigma\} | \mathbf{D}, M) d\mathbf{b}$$

where  $\mathcal{R}^Q$  is an  $Q$  dimensional space of real numbers.

- This may be achieved by use of the following standard integral:

$$\int_{\mathfrak{R}^Q} \exp \left[ -\frac{1}{2\sigma^2} (\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{b} + c) \right] d\mathbf{x}$$

$$= \frac{(2\pi\sigma^2)^{Q/2}}{\sqrt{\det(\mathbf{A})}} \exp \left[ -\frac{1}{2\sigma^2} \left( c - \frac{\mathbf{b}^T \mathbf{A}^{-1} \mathbf{b}}{4} \right) \right]$$

$$p(\sigma, \mathbf{w}_m | \mathbf{D}, \mathbf{M}) \propto \frac{(2\pi\sigma^2)^{-\left(\frac{N-Q}{2}\right)}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}} \exp \left[ -\left( \frac{\mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D}}{2\sigma^2} \right) \right]$$

- The standard deviation may again be integrated out as a gamma integral to give:

$$p(\mathbf{w}_m | \mathbf{D}, M) \propto \frac{\left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{\frac{-(N-Q)}{2}}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}}$$

which is in the form of a student-t distribution.

- Note that this is a function of  $\mathbf{w}_m$  only. This means that there is no need to know about the standard deviation nor the values of the linear parameters in order to infer the values of  $\mathbf{w}_m$ . Here the integrals have been done analytically. In most large problems involving many parameters and Non-Gaussian distributions, these integrations have to be performed numerically, (MCMC etc).

## FREQUENCY ESTIMATION

- As an example of the application of the general linear model, consider the detection of a single frequency.

$$f(t) = A \cos(\omega t) + B \sin(\omega t)$$

The data are assumed to consist of samples from the signal  $f(t)$  corrupted with independent white zero mean Gaussian noise samples with standard deviation  $\sigma$ .

- This signal model belongs to the general linear model family and the structure of the  $\mathbf{G}$  matrix is:

$$\mathbf{G} = \begin{bmatrix} \cos(\omega t_1) & \sin(\omega t_1) \\ \cos(\omega t_2) & \sin(\omega t_2) \\ \cos(\omega t_3) & \sin(\omega t_3) \\ \vdots & \vdots \\ \cos(\omega t_N) & \sin(\omega t_N) \end{bmatrix}$$

and the linear coefficient vector is:

$$\mathbf{b} = \begin{bmatrix} A \\ B \end{bmatrix}$$

- The general expression for the marginalised posterior probability density for the parameters  $\mathbf{w}_m$  of the model matrix  $\mathbf{G}$  may be used to express the probability density for the angular frequency  $\omega$  as:

$$p(\omega \mid \mathbf{D}, \mathbf{I}) \propto \frac{\left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{\frac{(2-N)}{2}}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}}$$

- The columns of the  $\mathbf{G}$  matrix are nearly orthogonal. This may be used to simplify the marginal posterior for  $\omega$ . The matrix  $\mathbf{G}^T \mathbf{G}$  is approximately equal to  $N/2$  times the identity matrix.

- It is easy to show that:

$$\mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \approx \sum_{i=1}^N d_i^2 - \frac{2}{N} \left( \sum_{i=1}^N d_i \cos(\omega t_i) + \sum_{i=1}^N d_i \sin(\omega t_i) \right)^2$$

- Define

$$I(\omega) = \left( \sum_{i=1}^N d_i \cos(\omega t_i) \right)$$

$$Q(\omega) = \left( \sum_{i=1}^N d_i \sin(\omega t_i) \right)$$

- Further approximations may be made to obtain:

$$\mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \approx \sum_{i=1}^N d_i^2 - \frac{2}{N} I(\omega)^2 - \frac{2}{N} Q(\omega)^2$$

- The Schuster periodogram is defined as

$$C(\omega) = \frac{1}{N} (I^2(\omega) + Q^2(\omega))$$

so that the marginal density for the angular frequency may be expressed in terms of the Schuster periodogram as:

$$p(\omega \mid \mathbf{D}, \mathbf{I}) \propto \left[ 1 - \frac{2C(\omega)}{\sum_{i=1}^N d_i^2} \right]^{\frac{2-N}{2}}$$

- Note that the term inside the square brackets is small (thus implying that the marginal density is large) if  $2C(\omega) \approx \sum_{i=1}^N d_i^2$ . This occurs if most of the data energy is concentrated around a single frequency  $\omega$ .

- The Schuster periodogram, (and hence the Discrete Fourier transform), is designed to determine the value of a single frequency in white Gaussian zero mean noise. Therefore it should really only be used on data that satisfies the single frequency model, and is not really designed for use in resolving two or more closely spaced frequencies, nor should it be used when the data is corrupted by non-Gaussian noise.

- It is interesting to compare the discrete Fourier transform power spectrum and the Bayesian marginal density for a single frequency. If the frequency bin positions correspond with the sample points in the time series then the two procedures will be exactly equivalent. But what happens if the data contain multiple frequencies ?

If the number of data points,  $N$ , is large and the frequencies are such that

$$|\omega_i - \omega_j| \gg \frac{2\pi}{N} \quad i \neq j$$

then the single model function method given above is sufficient.

- However, if the condition is not satisfied, the model function  $\mathbf{G}$  matrix must be written, for the case of two frequencies, as

$$\mathbf{G} = \begin{bmatrix} \cos(\omega_1 t_1) & \sin(\omega_1 t_1) & \cos(\omega_2 t_1) & \sin(\omega_2 t_1) \\ \cos(\omega_1 t_2) & \sin(\omega_1 t_2) & \cos(\omega_2 t_1) & \sin(\omega_2 t_1) \\ \cos(\omega_1 t_3) & \sin(\omega_1 t_3) & \cos(\omega_2 t_1) & \sin(\omega_2 t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(\omega_1 t_N) & \sin(\omega_1 t_N) & \cos(\omega_2 t_1) & \sin(\omega_2 t_1) \end{bmatrix}$$

The analysis then continues as before and it is necessary to determine the maximum of  $p(\omega_1, \omega_2 | \mathbf{d}, M)$  with respect to the two frequencies. The extension to the general case of  $Q$  frequencies is evident.

- Any data analysis problem can be formulated in the same way. We define the model function  $\mathbf{G}$  matrix and off we go with exactly the same analysis - this gives us the **Generalised Periodogram**. The removal of the linear (amplitude) parameters and the noise variance still leaves us with a multidimensional parameter space which either has to be searched for the maximum or plotted as a function of the parameters and the maximum obtained.
- Including the normalisation constants, the full expression that needs to be searched for a maximum as a function of  $\{\omega\}$  is

(with  $q = \frac{1}{2} (2\pi)^{-(N-M)/2}$ )

$$p(\{\omega\} | \mathbf{D}) = q \frac{\Gamma(\frac{N-M}{2})}{\sqrt{\det \mathbf{G}^T \mathbf{G}}} \left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{-(N-M)/2}$$

- Ignoring additive constants the logarithm of this function may be written in the form

$$E(\{\omega\} | \mathbf{D}) = -\log p(\{\omega\} | \mathbf{D}) = \frac{N - M}{2} \log [\mathbf{D}^T \mathbf{D} - \mathbf{f}^T \mathbf{f}]$$

$$+\frac{1}{2} \log \det \mathbf{G}^T \mathbf{G}$$

where

$$\mathbf{f} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}$$

Let  $\mathbf{b}$  be defined as

$$\mathbf{b} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}$$

- Our aim is now to calculate  $\frac{\partial E}{\partial \alpha}$  where  $\alpha$  is a scalar component of  $\omega$ .

- Two fundamental identities from vector calculus are required to compute this derivative:

1.

$$\frac{\partial}{\partial \alpha} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \left( \frac{\partial}{\partial \alpha} \mathbf{A} \right) \mathbf{A}^{-1}$$

2.

$$\frac{\partial}{\partial \alpha} \log \det \mathbf{A} = \text{Trace} \left( \mathbf{A}^{-1} \frac{\partial}{\partial \alpha} \mathbf{A} \right)$$

- Differentiation above yields

$$\frac{\partial E}{\partial \alpha} = \frac{N - M}{2} [\mathbf{D}^T \mathbf{D} - \mathbf{f}^T \mathbf{f}]^{-1} \left[ -2 \mathbf{f}^T \frac{\partial \mathbf{f}}{\partial \alpha} \right]$$

$$+ \frac{1}{2} \text{Trace} \left\{ (\mathbf{G}^T \mathbf{G})^{-1} [\mathbf{G}^T \mathbf{L} + \mathbf{L}^T \mathbf{G}] \right\}$$

where

$$\mathbf{L} = \frac{\partial \mathbf{G}}{\partial \alpha}$$

$$\frac{\partial \mathbf{f}}{\partial \alpha} = \mathbf{L}(\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} + \mathbf{G}(\mathbf{G}^T \mathbf{G})^{-1} \mathbf{L}^T \mathbf{D}$$
$$- \mathbf{G}(\mathbf{G}^T \mathbf{G})^{-1} [\mathbf{G}^T \mathbf{L} + \mathbf{L}^T \mathbf{G}] (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D}$$

- After some algebraic manipulation one can show that:

$$\mathbf{f}^T \frac{\partial \mathbf{f}}{\partial \alpha} = \mathbf{b}^T \mathbf{L}^T (\mathbf{D} - \mathbf{f}) = (\mathbf{D} - \mathbf{f})^T \mathbf{L} \mathbf{b}$$

- Finally one obtains

$$\frac{\partial E}{\partial \alpha} = -\frac{1}{\sigma^2}(\mathbf{D} - \mathbf{G}\mathbf{b})^T \mathbf{L}\mathbf{b} + \text{Trace} \left\{ (\mathbf{G}^T \mathbf{G})^{-1} [\mathbf{G}^T \mathbf{L}] \right\}$$

where

$$\sigma^2 = \frac{1}{N - M} [\mathbf{D}^T \mathbf{D} - \mathbf{f}^T \mathbf{f}]$$

- This gradient information can be used in a numerical optimisation algorithm that tries to search for the global maximum of the posterior density.

## REMINDER CONCERNING FOURIER TRANSFORMS

- The basic steps are as follows;

$$x_s(t) = x_a(t)s(t)$$

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

$$x_s(t) = \sum_{n=-\infty}^{\infty} x(n)\delta(t - nT)$$

- The Fourier Transform of this is

$$X_s(\omega) = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(n)\delta(t - nT) \exp(-j\omega t) dt$$

$$X_s(\omega) = \sum_{n=-\infty}^{\infty} x(n) \exp(-j\omega nT)$$

$$= \sum_{n=-\infty}^{\infty} x(n) \exp(-j\Omega n)$$

where

$$\Omega = \omega T = 2\pi f T = 2\pi \left( \frac{f}{f_s} \right)$$

- This is related to the Fourier Transform of the analogue signal by

$$X_s(\omega) = \frac{1}{T} \sum_{n=-\infty}^{\infty} X_a(\omega - \omega_0)$$

where  $\omega_0 = 2\pi/T$ . (This introduces the idea of aliasing etc).

- The DFT is introduced because in general we will only have a finite number of data sample, say  $0 \leq n \leq N - 1$  and we restrict frequency to the discrete set

$$\left\{0, \frac{1}{T}, \frac{2}{T}, \dots, \frac{N-1}{T}\right\}$$

$$f = \frac{k}{NT} = \frac{k}{N} f_s$$

Therefore

$$X(k) = \{X_s(\omega)\} = \sum_{n=0}^{N-1} x(n) \exp(-j2\pi \frac{kn}{N})$$

This is the DFT.

## CHANGEPOINT DETECTION

- Consider  $N$  samples of data from a piecewise constant input with independent additive Gaussian noise.

$$d_i = \begin{cases} \mu_1 + e_i & \text{if } i \leq m \\ \mu_2 + e_i & \text{otherwise} \end{cases}$$

- The likelihood function is:

$$p(\mathbf{D}|\{\mu_1 \mu_2 \sigma m\}, M) = \prod_{i=1}^N p(e_i) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^N e_i^2 \right) \right]$$

where

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^m (d_i - \mu_1)^2 + \sum_{i=m+1}^N (d_i - \mu_2)^2$$

- The Posterior probability is given by:

$$p(\{\mu_1 \mu_2 \sigma m\} | \mathbf{D}, M) = \frac{p(\mathbf{D} | \{\mu_1 \mu_2 \sigma m\}, M) p(\{\mu_1 \mu_2 \sigma m\} | M)}{p(\mathbf{D} | M)}$$

- Assuming that the parameters are independent, assign the following priors:

$$p(\mu_1|M) = k_1$$

$$p(\mu_2|M) = k_2$$

$$p(\sigma|M) = 1/\sigma$$

$$p(m|M) = k_3$$

- The posterior probability becomes:

$$p(\{\mu_1 \mu_2 \sigma m\}|\mathbf{D}, M) \propto \frac{1}{\sigma} p(\mathbf{D}|\{\mu_1 \mu_2 \sigma m\}, M)$$

- The parameters  $\{\mu_1 \mu_2 \sigma\}$  can be considered as *nuisance* parameters and may be integrated out as:

$$p(\{m\}|\mathbf{D}, M) = \int_0^\infty d\sigma \int_{-\infty}^\infty d\mu_1 \int_{-\infty}^\infty d\mu_2 p(\{\mu_1 \mu_2 \sigma m\}|\mathbf{D}, M)$$

- This integral may be performed analytically to give:

$$p(\{m\}|\mathbf{D}, M) \propto \frac{1}{\sqrt{m(N-m)}} \left[ \sum_{i=1}^N d_i^2 - \frac{1}{m} S_l^2 - \frac{1}{N-m} S_r^2 \right]^{-\left(\frac{N}{2}-1\right)}$$

where  $S_l = \sum_{i=1}^m d_i$  and  $S_r = \sum_{i=m+1}^N d_i$

This is a function of only the changepoint  $m$  and the data  $\mathbf{D}$ .

- The changepoint problem may also be formulated for the general linear model as:

$$d_i = \begin{cases} \sum_{p=1}^P \alpha_p g_p(i) + e_i & \text{if } i \leq m \\ \sum_{p=1}^P \beta_p g_p(i) + e_i & \text{otherwise} \end{cases}$$

where  $g_p(i)$  is the value of a time-dependent model function evaluated at time  $t_i$ .

- This may be expressed in the general linear model form:

$$\mathbf{D} = \mathbf{G} \mathbf{b} + \mathbf{e}$$

and the likelihood becomes:

$$p(\mathbf{D} | \{m\}, \sigma, \mathbf{b}, M) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left[ -\frac{\mathbf{e}^T \mathbf{e}}{2\sigma^2} \right]$$

- Assigning uniform priors to the linear parameters  $\mathbf{b}$  and the Jeffreys' prior to  $\sigma$ , the posterior density may be marginalised over the *nuisance* parameters to give:

$$p(m|\mathbf{D}, M) \propto \frac{\left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{\frac{-(N-2)}{2}}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}}$$

## APPLICATIONS OF CHANGEPOINTS TO COMMUNICATIONS

- The changepoint framework can be used to detect discrete phase changes of a sinusoidal carrier wave, as one would obtain in a BPSK or QPSK signal, without prior knowledge of the carrier phase.
- In Binary Phase Shift Keying (BPSK) a carrier sinusoid is modulated with digital information by means of 180 degree phase changes in the carrier.
- In order to model the  $180^\circ$  phase change, the changepoint model includes a change in the basis functions at the point where the presence of a  $180^\circ$  changepoint is being tested. The changepoint is modelled as a  $180^\circ$  rotation of the axes. This method of rotation of the axes is also used in the changepoint models for detecting  $90^\circ$  phase changes.

$$\mathbf{G} = \begin{bmatrix} \sin(\omega t_1) & \cos(\omega t_1) \\ \sin(\omega t_2) & \cos(\omega t_2) \\ \sin(\omega t_3) & \cos(\omega t_3) \\ \vdots & \vdots \\ \sin(\omega t_m) & \cos(\omega t_m) \\ -\sin(\omega t_{m+1}) & -\cos(\omega t_{m+1}) \\ \vdots & \vdots \\ -\sin(\omega t_N) & -\cos(\omega t_N) \end{bmatrix} \quad (13)$$

- A block of 80 samples of a noisy BPSK signal was simulated using a sine wave with a period of 20 samples. A phase change of  $180^\circ$  occurred at sample number 25. White Gaussian noise was added to the signal to produce a signal to noise ratio of 7.2 dB. The most probable position for the phase change based on the data block was at sample number 23.

Figure 1:  $180^\circ$  phase change: noisy data (solid line) and clean signal (dash-dot line).

Figure 2: Plot of unnormalised marginal probability of a  $180^\circ$  phase change.

## SIMPLE STEP DETECTOR IN GENERALIZED MATRIX FORM

- The matrix  $\mathbf{G}$  will consist of two columns. The first column will consist of  $m$  rows of 1's and  $N - m$  rows of 0's thereafter. On the other hand, the second column will consist of  $m$  rows of 0's and  $N - m$  rows of 1's thereafter:

$$\mathbf{G}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

This implies

$$\mathbf{G}^T \mathbf{G} = \begin{bmatrix} m & 0 \\ 0 & N - m \end{bmatrix}$$

$$(\mathbf{G}^T \mathbf{G})^{-1} = \frac{1}{m(N-m)} \begin{bmatrix} N-m & 0 \\ 0 & m \end{bmatrix}$$

$$\mathbf{D}^T \mathbf{G} = \begin{bmatrix} \sum_{i=1}^m d_i & \sum_{i=m+1}^N d_i \end{bmatrix}$$

$$\mathbf{D}^T \mathbf{D} = \sum_{i=1}^N d_i^2$$

Putting all these terms together the posterior density becomes:

$$p(m|\mathbf{D}, M) \propto \frac{\left[ \mathbf{D}^T \mathbf{D} - \mathbf{D}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{D} \right]^{\frac{-(N-2)}{2}}}{\sqrt{\det(\mathbf{G}^T \mathbf{G})}}$$

$$= \frac{1}{\sqrt{m(N-m)}} \left[ \sum_{i=1}^N d_i^2 - \frac{1}{m} \left( \sum_{i=1}^m d_i \right)^2 - \frac{1}{N-m} \left( \sum_{i=m+1}^N d_i \right)^2 \right]^{\frac{-(N-2)}{2}}$$

as found before.

## THE RAMP

- For this model, we write :

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_m \\ d_{m+1} \\ d_{m+2} \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N - m \end{bmatrix} \begin{bmatrix} \mu \\ r \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_m \\ e_{m+1} \\ e_{m+2} \\ \vdots \\ e_N \end{bmatrix}$$

- Therefore :

$$\mathbf{G}^T \mathbf{G} = \begin{bmatrix} N & \sum_{m+1}^N (i - m) \\ \sum_{m+1}^N (i - m) & \sum_{m+1}^N (i - m)^2 \end{bmatrix}$$

$$(\mathbf{G}^T \mathbf{G})^{-1} = \frac{1}{N \sum_{m+1}^N (i - m)^2 - \left( \sum_{m+1}^N (i - m) \right)^2}$$

$$\times \begin{bmatrix} \sum_{m+1}^N (i - m)^2 & \sum_{m+1}^N (m - i) \\ \sum_{m+1}^N (m - i) & N \end{bmatrix}$$

$$\mathbf{d}^T \mathbf{G} = \begin{bmatrix} \sum_{i=1}^N d_i & \sum_{i=m+1}^N d_i(i-m) \end{bmatrix}$$

$$\mathbf{d}^T \mathbf{d} = \sum_{i=1}^N d_i^2$$

## CHANGEPOINTS IN POLYNOMIAL MODELS

- The flexibility of the general linear model allows us to detect changepoints in polynomials and other models where the basis functions are non-linear, but the models are linear in their coefficients.

$$d_i = \begin{cases} \sum_{j=0}^k \alpha_j x_i^j + e_i & \text{for } i \leq m \\ \sum_{j=0}^k \beta_j x_i^j + e_i & \text{for } m < i \leq N \end{cases}$$

Thus, we define the matrix of basis functions,  $\mathbf{G}$ , for the case when  $k = 2$  as:

$$\mathbf{G} = \begin{bmatrix} 1 & x_1 & x_1^2 & 0 & 0 & 0 \\ 1 & x_2 & x_2^2 & 0 & 0 & 0 \\ 1 & x_3 & x_3^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m-1} & x_{m-1}^2 & 0 & 0 & 0 \\ 1 & x_m & x_m^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{m+1} & x_{m+1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_N & x_N^2 \end{bmatrix}$$

- The general linear model accounts for other non-linear basis functions in a similar manner.

## MIXED MODELS

- For example, at a changepoint, the model order of an AR process may increase from a two-pole model to a four-pole model. An AR process may become an ARMA process. Moreover, changepoints may divide the sequence into segments with signals (of whatever model) and segments with only noise (of a completely different structure than the signal).
- In the case of a changepoint,  $m$ , that divides a two-pole AR model and a three-pole AR model, we have:

$$d_i = \begin{cases} \sum_{j=1}^2 \alpha_j d_{i-j} + e_i & \text{for } i \leq m \\ \sum_{j=1}^3 \beta_j d_{i-j} + e_i & \text{for } m < i \leq N \end{cases}$$

And hence :

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{m-1} \\ d_m \\ d_{m+1} \\ \vdots \\ d_N \end{bmatrix} = \mathbf{G} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{m-1} \\ e_m \\ e_{m+1} \\ \vdots \\ e_N \end{bmatrix}$$

From above, we have:

$$\mathbf{G} = \begin{bmatrix} d_0 & d_{-1} & 0 & 0 & 0 \\ d_1 & d_0 & 0 & 0 & 0 \\ d_2 & d_1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{m-2} & d_{m-3} & 0 & 0 & 0 \\ d_{m-1} & d_{m-2} & 0 & 0 & 0 \\ 0 & 0 & d_m & d_{m-1} & d_{m-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & d_{N-1} & d_{N-2} & d_{N-3} \end{bmatrix}$$

with which we are able to find the posterior probability by evaluating the expression given previously for all  $m = 1, \dots, N$ .

## MODEL SELECTION AND EVIDENCE

$$p(M_i|\mathbf{D}) = \frac{p(\mathbf{D}|M_i)p(M_i)}{p(\mathbf{D})}$$

- Are there three Decaying Functions or two ? - Chemical kinetics, NMR, etc.
- Are there three Frequency Components or two ? - Radar, Sonar, Speech etc.
- Are there three Diffraction Peaks or two ? - Material science, X-rays etc.
- Are there three Classes or two ? - Classification, Image Segmentation etc.

- Are there three Significant Singular Values or two ? - Data Compression, Chaos etc.

In many cases we can assume

$$p(M_i|\mathbf{D}) \propto p(\mathbf{D}|M_i) = \int p(\mathbf{D}|\mathbf{w}, M_i) p(\mathbf{w}|M_i) d\mathbf{w}$$

and

$$p(M_i|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{w}_{\text{MAP}}, M_i) p(\mathbf{w}_{\text{MAP}}|M_i) \Delta\mathbf{w}$$

Ockham's razor etc.

## DANGEROUS EVIDENCE CALCULATIONS FOR LIPT MODELS !!

•

$$p(e(1), e(2), \dots, e(N)) = \left(\frac{1}{2\pi\sigma_e^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma_e^2} \mathbf{e}^t \cdot \mathbf{e}\right)$$

•

$$p(x(1), x(2), \dots, x(N) \mid \underline{a}, \mathcal{M}) = \left(\frac{1}{2\pi\sigma_e^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma_e^2} (\underline{x}^t \cdot \underline{x} - 2\underline{x}^t G \underline{a} + \underline{a}^t G^t G \underline{a})\right)$$

- Uniform Priors (Do not attempt this at home !)

$$p(\underline{x} | \mathcal{M}) = \int .. \int p(\underline{x} | \underline{a}, \mathcal{M}) d\underline{a}$$

$$= \left(\frac{1}{2\pi\sigma_e^2}\right)^{N/2} \int .. \int \exp(-(\underline{a}^t A \underline{a} + B^t \underline{a} + C)) d\underline{a}$$

where

$$C = \frac{\underline{x}^t \underline{x}}{2\sigma_e^2}, \quad B = \frac{-\underline{x}^t G}{\sigma_e^2}, \quad A = \frac{G^t G}{2\sigma_e^2}$$

Using the identity

$$\underline{a}^t A \underline{a} + B^t \underline{a} + C = \left( \underline{a}^t + \frac{B^t A^{-1}}{2} \right) A \left( \underline{a} + \frac{A^{-1} B}{2} \right) - \frac{B^t A^{-1} B}{4} + C$$

and writing

$$\underline{z} = A^{1/2} \left( \underline{a} + \frac{A^{-1}B}{2} \right)$$

we have

$$d\underline{z} = |A|^{1/2} d\underline{a}$$

Therefore

$$p(\underline{x} | \mathcal{M}) = \left(\frac{1}{2\pi\sigma_e^2}\right)^{N/2} \exp\left(\frac{B^t A^{-1} B}{4} - C\right) |A|^{-1/2} \int_{-\infty}^{\infty} \exp(-\underline{z}^t \cdot \underline{z}) d\underline{z}$$

- Uniform Window (from  $-\beta$  to  $\beta$  and height  $1/2\beta$ )

We get

$$p(\underline{x} | \mathcal{M}) = \left(\frac{1}{2\pi\sigma_e^2}\right)^{(P-N)/2} \exp\left(\frac{B^t A^{-1} B}{4} - C\right) |A|^{-1/2} \frac{1}{(2\beta)^p} \int_{-\alpha}^{\alpha} \exp(-\underline{z}^t \cdot \underline{z}) d\underline{z}$$

where

$$\beta = \max |a_i|, \quad \underline{z} = A^{1/2} \left( \underline{a} + \frac{A^{-1} B}{2} \right)$$

and

$$\alpha = A^{1/2} \left( \underline{\beta} + \frac{A^{-1} B}{2} \right), \quad d\underline{z} = |A|^{1/2} d\underline{a}$$

Therefore

$$p(\underline{x} \mid \mathcal{M}) = (2\pi\sigma_e^2)^{(P-N)/2} \exp\left(-\frac{(N-P)}{2}\right) (2\sigma_e^2)^{P/2}$$

$$|G^t G|^{-1/2} (2\beta)^{-P} (\sqrt{\pi}\phi(\alpha))^P$$

Therefore

$$\ln E = \frac{P-N}{2} \ln(2\pi\sigma_e^2) - \frac{1}{2} \ln(\det(G^t G)) - \frac{N-P}{2} - P \log(2\beta) + p \ln(\phi(\alpha))$$

- Smoothness Priors

$$p(\underline{a}) = (2\pi)^{-P/2} |R|^{-1/2} \exp\left(-\frac{1}{2}\underline{a}^t R^{-1} \underline{a}\right)$$

Everything continues as before.

We now have a ‘regularization’ parameter to tune - or we can use a MAP estimate etc.

(Applications to KL transform, SVD, Classification, etc ....)

## BAYESIAN NUMERICAL METHODS

### Basic Introduction to Monte Carlo Methods

- For most statistical applications, the numerical aspect of the problem can be cast in the form of computing the expected value of some function of interest  $f(\boldsymbol{\theta})$ , with respect to a ‘target’ probability density  $\pi(\boldsymbol{\theta})$ ,

$$\mathbb{E}_{\pi}(f) = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (14)$$

where  $\boldsymbol{\theta}$  is a vector of parameters of interest.

- In Bayesian inference,  $\pi$  is typically taken as a posterior distribution. For scalar  $\boldsymbol{\theta}$ , the usual way to evaluate an analytically intractable integral will

be via *deterministic* numerical integration. As the dimensionality increases, however, problems associated with applying deterministic techniques increase.

- In contrast, Monte Carlo integration is as easy in 10 dimensions as in 1 dimension and hence becomes the preferred method.

Suppose  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$  are a sample set from  $\pi$  and form

$$\hat{f} = \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{\theta}_i). \quad (15)$$

Then  $\hat{f}$  is said to be a Monte Carlo estimator of  $\mathbb{E}_\pi(f)$ .

Since

$$\mathbb{E}\hat{f} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}f(\boldsymbol{\theta}_i) = \mathbb{E}_{\pi}(f), \quad (16)$$

$\hat{f}$  is clearly an unbiased estimator. The variance is given by

$$\text{var}(\hat{f}) = \frac{1}{n} \int [f(\boldsymbol{\theta}) - \mathbb{E}_{\pi}(f)]^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (17)$$

## IMPORTANCE SAMPLING

- Suppose one wishes to evaluate the expectation of some function  $f$  under  $\tilde{\pi}$ , having already obtained a sample  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}\}$  from  $\pi$ .
- If both  $\pi$  and  $\tilde{\pi}$  are normalised densities, an importance sampling estimator of  $\mathbb{E}_{\tilde{\pi}}[f(\boldsymbol{\theta})]$  is given by

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}}[f(\boldsymbol{\theta})] &= \int f(\boldsymbol{\theta}) \frac{\tilde{\pi}(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{\theta}^{(i)}) \frac{\tilde{\pi}(\boldsymbol{\theta}^{(i)})}{\pi(\boldsymbol{\theta}^{(i)})},\end{aligned}\tag{18}$$

provided that  $\pi$  dominates  $\tilde{\pi}$ , i.e., the support of  $\pi$  covers that of  $\tilde{\pi}$ . This is the principle of importance sampling.

## BAYESIAN LINEAR REGRESSION

- The Bayesian analysis of the problem of making a linear fit to data containing definite but unknown independent Gaussian noise of variance  $\sigma^2$ , starts off by writing down the direct probability density function for the data as

$$p(\mathbf{D}|\mathbf{x}, c, m, \sigma) \propto \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (d_i - c - mx_i)^2\right]$$

- Using Bayes' theorem and assuming uniform priors for  $c$  and  $m$  and a Jefferys' prior for  $\sigma$  we obtain

$$p(c, m, \sigma | \mathbf{D}, \mathbf{x}) \propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (d_i - c - mx_i)^2\right]$$

- We now introduce the quantities  $\nu = N - 2$  and

$$\bar{x} = \frac{1}{N} \sum_i x_i \quad \bar{d} = \frac{1}{N} \sum_i d_i$$

and

$$\bar{c} = \bar{d} - \bar{m}\bar{x} \quad \bar{m} = \frac{\sum_i (x_i - \bar{x})(d_i - \bar{d})}{\sum_i (x_i - \bar{x})^2}$$

$$s^2 = \frac{1}{\nu} \sum_i (d_i - \bar{c} - \bar{m}x_i)^2$$

- We rarely know anything at all about  $\sigma^2$ , so we can integrate it out of the problem (treating it as a nuisance parameter).

- As a result, we can calculate analytic expressions for the Marginal posterior distributions for the parameters  $a$  and  $b$  as follows

$$p(c|\mathbf{D}, \mathbf{x}) \propto \left[ \nu + \frac{\sum_i (x_i - \bar{x})^2}{s^2 \sum_i x_i^2 / N} (c - \bar{c})^2 \right]^{-(\nu+1)/2}$$

$$p(m|\mathbf{D}, \mathbf{x}) \propto \left[ \nu + \frac{\sum_i (x_i - \bar{x})^2}{s^2} (m - \bar{m})^2 \right]^{-(\nu+1)/2}$$

## THE GIBBS SAMPLER

- **The Hammersley-Clifford Theorem**

- ◇ The conditional distributions  $p_i(a_i \mid a_{i \neq j}, \mathbf{D})$  contain sufficient information to produce samples from  $p(a_1, a_2, \dots, a_k \mid \mathbf{D})$
  - ◇ This is like maximizing an objective function successively in each direction of a given basis - might end up in a saddlepoint.
  - ◇ Quite easy to show that the joint density can be directly derived from the conditional densities.
- Assume that the parameter space consists of  $k$  components  $\{a_1, a_2, a_3 \dots a_k\}$ . The components are initialized to starting values  $\{a_1^0, a_2^0, a_3^0 \dots a_k^0\}$  and the Gibbs sampler proceeds by drawing random variates from conditional densities in a cyclical iterative pattern as follows:

First iteration:

$$\begin{aligned} a_1^1 &\leftarrow p(a_1 \mid a_2^0 a_3^0 \dots a_{k-1}^0 a_k^0, \mathbf{D}) \\ a_2^1 &\leftarrow p(a_2 \mid a_3^0 a_4^0 \dots a_k^0 a_1^1, \mathbf{D}) \\ a_3^1 &\leftarrow p(a_3 \mid a_4^0 a_5^0 \dots a_1^1 a_2^1, \mathbf{D}) \\ &\vdots \\ a_k^1 &\leftarrow p(a_k \mid a_1^1 a_2^1 \dots a_{k-2}^1 a_{k-1}^1, \mathbf{D}) \end{aligned}$$

**Second iteration:**

$$\begin{aligned}
 a_1^2 &\leftarrow p(a_1 \mid a_2^1 a_3^1 \dots a_{k-1}^1 a_k^1, \mathbf{D}) \\
 a_2^2 &\leftarrow p(a_2 \mid a_3^1 a_4^1 \dots a_k^1 a_1^2, \mathbf{D}) \\
 a_3^2 &\leftarrow p(a_3 \mid a_4^1 a_5^1 \dots a_1^2 a_2^2, \mathbf{D}) \\
 &\vdots \\
 a_k^2 &\leftarrow p(a_k \mid a_2^2 a_1^2 \dots a_{k-2}^2 a_{k-1}^2, \mathbf{D})
 \end{aligned}$$

**$n^{\text{th}}$  iteration:**

$$\begin{aligned}
 a_1^n &\leftarrow p(a_1 \mid a_2^{n-1} a_3^{n-1} \dots a_{k-1}^{n-1} a_k^{n-1}, \mathbf{D}) \\
 &\vdots
 \end{aligned}$$

- As soon as a variate is drawn, then it is inserted immediately into the conditional probability density function, and it remains there until it is substituted in the next iteration.
- At the end of the  $j^{\text{th}}$  iteration the sample  $\{a_1^j, a_2^j, a_3^j \dots a_{k-1}^j, a_k^j\}$  is a sample from the joint density. In common with other Markov chain approaches the Gibbs sampler requires an initial transient period to converge to equilibrium.
- How much of the initial series is affected by the initial state is difficult to ascertain, but some literature is available on the subject. This initial period of length  $M$  is known as the “burn in” and it varies in length depending on the problem. One should always discard the first  $M$  samples.

## THE STRAIGHT LINE AND THE GIBBS SAMPLER

- The Gibbs sampler, for the case of a straight line model, works as follows:
- A joint distribution  $p(m, c, \sigma | \mathbf{D}, I)$  is sought. Initial values  $(m_0, c_0, \sigma_0)$  for the random variables  $m, c$  and  $\sigma$  are selected at random or based on some prior estimates.
- One then keeps all but one of the variables constant and draw a sample from the conditional distribution for the remaining variable. The current variable is then fixed at the value of the sample just drawn.
- A new estimate for the next variable is then drawn while keeping all other variables fixed. This is shown below: (the  $\leftarrow$  symbol means “is drawn from the distribution”)

$$\begin{aligned}
m_1 &\leftarrow p(m|c_0, \sigma_0, \mathbf{D}, I) \\
c_1 &\leftarrow p(c|m_1, \sigma_0, \mathbf{D}, I) \\
\sigma_1 &\leftarrow p(\sigma|m_1, c_1, \mathbf{D}, I) \\
m_2 &\leftarrow p(m|c_1, \sigma_1, \mathbf{D}, I) \\
c_2 &\leftarrow p(c|m_2, \sigma_1, \mathbf{D}, I) \\
\sigma_2 &\leftarrow p(\sigma|m_2, c_2, \mathbf{D}, I) \\
&\vdots \quad \vdots \quad \vdots \\
m_n &\leftarrow p(m|c_{n-1}, \sigma_{n-1}, \mathbf{D}, I) \\
c_n &\leftarrow p(c|m_n, \sigma_{n-1}, \mathbf{D}, I) \\
\sigma_n &\leftarrow p(\sigma|m_n, c_n, \mathbf{D}, I)
\end{aligned}$$

- A model of the form  $d_i = mx_i + c$  is used. The Gibbs sampler will be used to obtain estimates of the joint density for the slope,  $m$ , and intercept,  $c$ , of the line and the noise variance,  $\sigma$ .

- The likelihood function for Gaussian noise is:

$$p(\mathbf{D}|m, c, \sigma, I) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \times \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^N (d_i - mx_i - c)^2 \right]$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \times \exp \left[ -\frac{1}{2\sigma^2} (D_2 - 2cD_1 - 2mX_D + 2mcX_1 + m^2X_2 + Nc^2) \right]$$

where

$$D_1 = \sum_{i=1}^N d_i, D_2 = \sum_{i=1}^N d_i^2, X_D = \sum_{i=1}^N x_i d_i, X_1 = \sum_{i=1}^N x_i, X_2 = \sum_{i=1}^N x_i^2$$

- We need to find the distribution  $p(m, c, \sigma | \mathbf{D}, I)$  and this can be obtained from Bayes' Theorem as follows:

$$p(m, c, \sigma | \mathbf{D}, I) = \frac{p(\mathbf{D} | m, c, \sigma, I)p(m|I)p(c|I)p(\sigma|I)}{p(\mathbf{D} | I)}$$

There is no reason to assume that the intercept, slope and variance are not independent, so one can assign uniform priors to the slope and intercept and a Jeffreys prior to the variance.

$$\begin{aligned} p(m|I) &= k_1 \\ p(c|I) &= k_2 \\ p(\sigma|I) &= \frac{k_3}{\sigma} \end{aligned}$$

- $p(\mathbf{D}|I)$  is a constant (for a fixed model) and ignoring all constants:

$$p(m, c, \sigma | \mathbf{D}, I) \propto \sigma^{-1} p(\mathbf{D} | m, c, \sigma, I)$$

$$\propto \sigma^{-(N+1)} \exp \left[ -\frac{1}{2\sigma^2} (D_2 - 2cD_1 - 2mX_D + 2mcX_1 + m^2X_2 + Nc^2) \right]$$

- The distribution that will be used for drawing samples for  $m$  is now derived. The current values for  $c$  and  $\sigma$  are held constant when drawing a new sample for  $m$  and so:

$$p(m | c, \sigma, \mathbf{D}, I) \propto \exp \left[ -\frac{1}{2\sigma^2} (m^2X_2 - 2m(X_D - cX_1)) \right]$$

- Hence we arrive at the distribution for  $m$ :

$$p(m|c, \sigma, \mathbf{D}, I) \propto \exp \left[ -\frac{1}{2} \left( \frac{\sqrt{X_2}}{\sigma} \right)^2 \left( m - \frac{X_D - cX_1}{X_2} \right)^2 \right]$$

- Thus  $p(m|c, \sigma, \mathbf{D}, I)$  is a Gaussian distribution with mean =  $\frac{X_D - cX_1}{X_2}$  and variance =  $\frac{\sigma}{\sqrt{X_2}}$ .
- Using a similar derivation, the distribution for the intercept,  $c$ , is

$$p(c|m, \sigma, \mathbf{D}, I) \propto \exp \left[ -\frac{1}{2} \left( \frac{\sqrt{N}}{\sigma} \right)^2 \left( c - \frac{D_1 - mX_1}{N} \right)^2 \right]$$

Another Gaussian, mean =  $\frac{D_1 - mX_1}{N}$  and variance =  $\frac{\sigma}{\sqrt{N}}$ .

- The distribution for  $\sigma$  is obtained from:

$$p(\sigma|m, c, \mathbf{D}, I) \propto \sigma^{-(N+1)} \exp(-A)$$

where

$$A = \frac{1}{2\sigma^2} (m^2 X_2 - 2m (X_D - cX_1) + Nc^2 + D_2 - 2cD_1)$$

- This has the form of the ‘Square-root Inverted Gamma ( $\text{Ga}^{-\frac{1}{2}}$ )’ distribution:

$$p(x) = cx^{-(2\alpha+1)} \exp\left[-\frac{\beta}{x^2}\right] \quad \alpha > 0, \beta > 0 \text{ and } x > 0$$

with  $\alpha = \frac{N}{2}$ ,  $\beta = \frac{1}{2} \sum_{i=1}^N (d_i - mx_i - c)^2$  and  $c = \frac{2\beta^\alpha}{\Gamma(\alpha)}$ .

- Variates from this distribution may be generated by using a transformation applied to variates drawn from a Gamma distribution with  $\frac{N}{2}$  degrees of freedom.
- After an initial transient the distributions for the different variables will converge to their invariant distributions as the Markov chain converges. This method can be a very powerful numerical tool for use on non-Gaussian processes.

## THE INTERPOLATION OF MISSING SAMPLES

- The aim of this section is to outline a method for restoring missing samples in digital audio signals.
- The section of audio signal in question is modelled as a stationary autoregressive process, and missing samples are imputed using the Gibbs sampler.
- Clicks are a familiar problem in audio gramophone signals, and take the form of sudden unexpected bursts of impulsive noise with random but finite duration.
- These bursts of noise have numerous causes such as dirt, electrical interference or mechanical damage to the storage medium. The original signal is often effectively lost.
- Several methods of detecting clicks have been devised, with the best approaches being model based.
- Once a click has been detected the *suspect* samples are removed and

replaced by interpolation.

Missing (ie. unknown) data:

$$[x(m) \ x(m+1) \ \dots \ x(m+L-1)]^T = \mathbf{x}_u$$

Observed (ie. known) data

$$[x(1) \ x(2) \ \dots \ x(m-1)]^T = \mathbf{x}_{k1}$$

$$[x(m+L) \ x(m+L+1) \ \dots \ x(N)]^T = \mathbf{x}_{k2}$$

Define the Augmented data vector containing both observed and missing data:

$$\mathbf{x} = [\mathbf{x}_{k1}^T \ \mathbf{x}_u^T \ \mathbf{x}_{k2}^T]^T$$

Audio signals are, in general, well-modelled as Autoregressive processes:

$$x(n) = \sum_{p=1}^Q a_p x(n-p) + e(n)$$

This expression may be written in matrix form in two ways as follows:

$$\mathbf{e} = \mathbf{G}_a \mathbf{x} \quad (19)$$

or

$$\mathbf{e} = \mathbf{x} - \mathbf{G}_x \mathbf{a} \quad (20)$$

In equation 19 the matrix contains the AR model parameters and the data appears as a vector whereas in equation 20 the data appears in the matrix

and the AR parameters are in vector form. The reason for using the two forms is algebraic convenience in what follows.

Consider the excitation energy  $\mathbf{e}^T \mathbf{e}$ .

From equation 19:

$$\mathbf{e}^T \mathbf{e} = \mathbf{x}^T \mathbf{G}_a^T \mathbf{G}_a \mathbf{x}$$

$$= [\mathbf{x}_{k1}^T \ \mathbf{x}_u^T \ \mathbf{x}_{k2}^T] \mathbf{G}_a^T \mathbf{G}_a [\mathbf{x}_{k1}^T \ \mathbf{x}_u^T \ \mathbf{x}_{k2}^T]^T$$

Partitioning the matrix and combining the observed data  $\mathbf{x}_{k1}$  and  $\mathbf{x}_{k2}$  into a single vector  $\mathbf{x}_k$  gives:

$$\mathbf{e}^T \mathbf{e} = [\mathbf{x}_k^T \ \mathbf{x}_u^T] \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_u \end{bmatrix}$$

where  $\mathbf{x}_k^T = [\mathbf{x}_{k1}^T \ \mathbf{x}_{k2}^T]$ .

$$\mathbf{e}^T \mathbf{e} = \mathbf{x}_k^T \mathbf{A} \mathbf{x}_k + 2\mathbf{x}_k^T \mathbf{B} \mathbf{x}_u + \mathbf{x}_u^T \mathbf{D} \mathbf{x}_u \quad (21)$$

which is a quadratic function of the unknown data vector  $\mathbf{x}_u$ .

An alternative expression for the excitation energy can be obtained from equation 20 as follows:

$$\mathbf{e}^T \mathbf{e} = (\mathbf{x} - \mathbf{G}_x \mathbf{a})^T (\mathbf{x} - \mathbf{G}_x \mathbf{a})$$

$$\mathbf{e}^T \mathbf{e} = \mathbf{x}^T \mathbf{x} + 2\mathbf{x}^T \mathbf{G}_x \mathbf{a} + \mathbf{a}^T \mathbf{G}_x^T \mathbf{G}_x \mathbf{a} \quad (22)$$

which is a quadratic function of the AR parameter vector  $\mathbf{a}$ .

Assume that the AR excitation  $e(n)$  is a zero-mean i.i.d. Gaussian process

$N(0, \sigma^2)$ , then:

$$p(\mathbf{e}) = \prod_{n=1}^N p(e(n))$$

$$p(\mathbf{e}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\mathbf{e}^T \mathbf{e}}{2\sigma^2}\right)$$

$$p(\mathbf{x} | \mathbf{a}, \sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\mathbf{e}^T \mathbf{e}}{2\sigma^2}\right) \quad (23)$$

which is the likelihood function for the augmented data.

If the missing data  $\mathbf{x}_u$  are fixed then one may maximize the likelihood with respect to the AR parameter vector  $\mathbf{a}$  by means of equations 23 and 22 to give:

$$\hat{\mathbf{a}} = (\mathbf{G}_x^T \mathbf{G}_x)^{-1} \mathbf{G}_x^T \mathbf{x}$$

If the AR parameters are fixed then one may maximize the likelihood with respect to the unknown data  $\mathbf{x}_u$  by means of equations 23 and 21 to give:

$$\hat{\mathbf{x}}_u = -\mathbf{D}^{-T} \mathbf{B}^T \mathbf{x}_k$$

Vaseghi, [10], describes a method for jointly estimating the AR parameters and the missing data by successively maximizing the likelihood with respect to  $\mathbf{a}$  and then with respect to  $\mathbf{x}_u$ . The procedure is iterated until

the values converge at a maximum of the likelihood function  $p(\mathbf{x} \mid \mathbf{a}, \sigma)$ . The EM algorithm may also be used although it works in a slightly different way from the above method. Upon each iteration, the interpolant  $\hat{\mathbf{x}}_u(i+1)$  may be computed *linearly* from the previous iterate  $\hat{\mathbf{x}}_u(i)$ , essentially leapfrogging the AR step. The resultant interpolant maximizes the predictive density  $p(\mathbf{x}_u \mid \mathbf{x}_k)$  – not the likelihood – so the results of using the EM algorithm are slightly different from those of the ML procedure. The problem may also be considered in a Gibbs Sampling framework which is, as discussed earlier, a Markov chain based Monte Carlo sampling scheme for simulating jointly distributed random variates and which may be used for sampling the joint density  $p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k)$ . The joint probability for the unknown variables may be written as:

$$p(\mathbf{x}_u, \mathbf{x}_k \mid \mathbf{a}, \sigma) \equiv p(\mathbf{x} \mid \mathbf{a}, \sigma)$$

$$p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k) = \frac{p(\mathbf{x} \mid \mathbf{a}, \sigma) p(\mathbf{a}, \sigma)}{p(\mathbf{x}_k)}$$

The Gibbs sampler requires the individual conditional densities.

$$p(\mathbf{x}_u \mid \mathbf{x}_k, \mathbf{a}, \sigma) = \frac{p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k)}{p(\mathbf{a}, \sigma \mid \mathbf{x}_k)}$$

$$= \frac{p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k)}{\int p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k) \mathbf{d}\mathbf{x}_k}$$

which is a multivariate Gaussian density in  $\mathbf{x}_u$ .

$$p(\mathbf{a} \mid \sigma, \mathbf{x}_k, \mathbf{x}_u) = \frac{p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k)}{\int p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k) \mathbf{d}\mathbf{a}}$$

which is a multivariate Gaussian density in  $\mathbf{a}$ .

$$p(\sigma \mid \mathbf{x}_k, \mathbf{x}_u, \mathbf{a}) = \frac{p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k)}{\int p(\mathbf{x}_u, \mathbf{a}, \sigma \mid \mathbf{x}_k) d\sigma}$$

which leads to a square-root inverted gamma density in  $\sigma$ .

It is straightforward to sample from these density functions so the Gibbs' Sampler may be used to sample from the joint posterior density.



structure of the matrices is taken into account. Matrix operations include the use of Levinson's algorithm, band LU decomposition and the Gaxpy-Cholesky decomposition.

There is an important difference between the results obtained from the sampling approach and approaches based on Maximum Likelihood (ML). The sampling method provides an estimate of the missing audio data which is based on a *typical* section of AR model excitation  $\mathbf{e}$  whereas the ML method provides a method which is based on the lowest excitation energy  $\mathbf{e}^T \mathbf{e}$ . The implication of this is that for relatively long sections of missing data, the ML methods provide a signal estimate which decreases in amplitude towards the centre of the missing data gap. Another way of looking at this is that the excitation is assumed to be Gaussian and the “most probable” Gaussian signal is zero but this is certainly not typical of what one would observe. The result is that, for long sections of missing data, the sampling approach provides estimates which are perceptually superior to those obtained from the ML approach.

## DIFFERENCES BETWEEN ‘PHYSICISTS’ AND ‘ENGINEERS’ AND MO

- The purpose of choosing a model must be kept in mind. Selecting a single model ignores model uncertainty and so will underestimate uncertainty about quantities of interest.
- Clearly better to take into account model uncertainty if one can. There is a Bayesian way of dealing with this first introduced in 1978 by Leamer.
- One can incorporate our uncertainty in model selection by forming a weighted average over the finite set of models using as the weighting the posterior model probabilities, given the data.
- If we have a set of candidate models  $M = M_1, \dots, M_k$  and if  $\Delta$  is the quantity of interest which could be a parameter, a future observation or the utility of a course of action, then the posterior distribution of  $\Delta$  given the data is

$$p(\Delta|D) = \sum_{k=1}^K p(\Delta|M_k, D)p(M_k|D)$$

and this is just the average of the posterior distributions under each model weighted by the corresponding posterior model probabilities.

- In the above expression we have;

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{l=1}^K p(\mathbf{D}|M_l)p(M_l)}$$

and

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|\theta_k, M_k)p(\theta_k|M_k)d\theta_k.$$

is the marginal likelihood of model  $M_k$  and  $\theta_k$  is the vector of parameters of model  $M_k$ .

- In general, the above model averaging is difficult to compute since the integrals are difficult and the number of terms can be large.
- For a problem with  $p$  covariates, the number of models in the finite sum is  $K = 2^p$ . In practice, only a small number of these models will have much support from the data.

Several approaches can be taken to address the uncertainty problem and they are based on either Discrete or Continuous model uncertainty.

Those based upon Discrete uncertainty are;

- The use of the Ockham window (Madigan and Raftery) which firstly works by considering if a particular model predicts the data worse than a model

which predicts the data well. If this is the case, the model is no longer considered. Secondly, one excludes all models that receive less support from the data than any of their sub models. The method cuts down the number of models to be averaged over by only including the more likely models.

- Markov Chain Monte Carlo Model Composition  $MC^3$ , (Raftery, Madigan and Hoeting) which approximates a complete Bayesian approach by generating a stochastic process which moves through model space. The method involves constructing a Markov chain with neighbourhoods for each model in the model space. These neighbourhoods consist of the model itself and all models with one variable less or one variable more than the given model. The method also cuts down the number of models to be averaged over as not all models will be visited by the Markov chain.

## PRIOR PROBABILITIES

The prior distribution  $p(\theta)$ , quantifies information about  $\theta$  prior to any (further) data being gathered. Sometimes  $p(\theta)$  can be constructed on the basis of past data. For example, if a quality inspection program has been running for some time, the distribution of the number of defectives in past batches can be used as the prior distribution for the number of defectives in future batches.

## CONJUGATE PRIOR DISTRIBUTIONS

- This is a prior distribution which when combined with the likelihood function, produces a posterior distribution which is in the same family as the prior.
- If we find a conjugate prior distribution which adequately fits our prior beliefs regarding  $p(\theta)$  we should use it since it will usually simplify calculations. The task of quantifying prior knowledge then amounts to specifying the parameters of the prior distribution.

## DECONVOLUTION

- Firstly, we assume that the peak positions,  $c_i$ , lie on a continuous one-dimensional space denoted by the closed interval  $[a, b]$  such that,

$$a \leq c_i \leq b,$$

where  $c_i$  is the actual position of the  $i^{th}$  peak in the data.

- Next we define the  $N$  sample positions of the discretized data set,

$$x[i] = \frac{L(i-1)}{(N-1)} + a,$$

where  $L = b - a$ ,  $i = 1, 2, \dots, N$ .

- The peak model is represented as,

$$y_p[i] = \sum_{j=1}^{k_s} a'_{j,k_s} \frac{1}{\sqrt{2\pi v_{j,k_s}}} \exp\left(-\frac{(x[i] - c_{j,k_s})^2}{2v_{j,k_s}}\right), \quad k_s \geq 1 \quad (24)$$

and the spline model as,

$$y_s[i] = \sum_{j=1}^{k_n+2} S(\xi_{i,k_n})_{i,j} a''_{j,(k_n+2)}, \quad k_n \geq 2 \quad (25)$$

- Therefore the joint model is,

$$\begin{aligned}
 \mathcal{M}_0 : y[i] &= n[i], & k_s = 0, k_n = 0 \text{ or } 1, \\
 \mathcal{M}_k : y[i] &= \left( \frac{1}{\sqrt{2\pi v_{sp}}} \exp \left( -\frac{(x[i])^2}{2v_{sp}} \right) \right) \otimes (y_p[i] + y_s[i]) + n_k[i], \\
 n_k[i] &\underset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)
 \end{aligned} \tag{26}$$

where,

$\mathcal{M}_k$	is the joint model order $k = \{k_s, k_n\}$ .
$k_s$	is the number of Gaussian peaks.
$k_n$	is the number of knots in the splines.
$x[i]$	is the discretized data space.
$y[i]$	is the intensity/energy values corresponding to $x[i]$ .
$y_p[i]$	is the part of the data which consists of only Gaussian peaks.
$y_s[i]$	is the part of the data which consists of only the spline model.
$a'_{j,k_s}$	is the amplitude of peak $j$ , in a $k_s$ -peak model.
$v_{j,k_s}$	is the variance of the peak $j$ , in a $k_s$ -peak model, ( $v_{j,k_s} > 0$ ).
$c_{j,k_s}$	is the location of peak $j$ , in a $k_s$ -peak model, ( $0 < c_{j,k_s} < L$ ).
$S(\xi_{i,k_n})_{i,j}$	is the normalised cubic B-spline coefficient, $\xi_{i,k_n}$ , ( $0 < \xi_{i,k_n} < L$ ).
$a''_{j,(k_n+2)}$	is the amplitude of peak $j$ , in a $k_n$ -knot cubic B-spline model.
$n_k[i]$	is the noise at each point, $i$ , in the data space, assumed to be AWGN.
$\sigma_k^2$	is the noise variance of $n_k[i]$ .
$v_{sp}$	is the variance of the Gaussian spread function.
$\otimes$	is the convolution operator.

- In the cubic spline model, for  $k_n$  knots,  $k_n + 2$  modal amplitudes are needed []. The matrices are defined such that their subscripts represent their associated model order, knots( $k_n$ ), peaks( $k_s$ ) or joint( $k$ ) models. As for the column vectors, their indices represent their size.
- We define,

$$k' = k_s + k_n + 2,$$

to simplify some of the expressions.

- Representing in a vector matrix form, we have,

$$\begin{aligned}
 \mathbf{y} &= \mathbf{H}_N \left( \mathbf{D}_{k_s}(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}) \mathbf{a}'_{k_s} + \mathbf{E}_{k_n}(\xi_{k_n}) \mathbf{a}''_{(k_n+2)} \right) + \mathbf{n}_k \\
 &= \mathbf{H}_N \left[ \mathbf{D}_{k_s}(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}) \quad \mathbf{E}_{k_n}(\xi_{k_n}) \right] \mathbf{a}_{k'} + \mathbf{n}_k \\
 &= \mathbf{H}_N \mathbf{G}_k(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}, \xi_{k_n}) \mathbf{a}_{k'} + \mathbf{n}_k
 \end{aligned} \tag{27}$$

where,

$$\begin{array}{ll}
 \mathbf{y} & \triangleq [y(1) \quad y(2) \quad \dots \quad y(N)]^T & \in \mathbb{R}^N \\
 \mathbf{x} & \triangleq [x(1) \quad x(2) \quad \dots \quad x(N)]^T & \in \mathbb{R}^N \\
 \mathbf{c}_{k_s} & \triangleq [c_{1,k_s} \quad c_{2,k_s} \quad \dots \quad c_{k_s,k_s}]^T & \in \mathbb{R}^{+k_s} \\
 \mathbf{v}_{k_s} & \triangleq [v_{1,k_s} \quad v_{2,k_s} \quad \dots \quad v_{k_s,k_s}]^T & \in \mathbb{R}^{+k_s} \\
 \mathbf{a}_{k'} & \triangleq [a'_{1,k_s} \quad a'_{2,k_s} \quad \dots \quad a''_{(k_n+1),(k_n+2)} \quad a''_{(k_n+2),(k_n+2)}]^T & \in \mathbb{R}^{k'} \\
 \mathbf{n}_k & \triangleq [n(1) \quad n(2) \quad \dots \quad n(N)]^T & \in \mathbb{R}^N \\
 \xi_{k_n} & \triangleq [\xi_{1,k_n} \quad \xi_{2,k_n} \quad \dots \quad \xi_{k_n,k_n}]^T & \in \mathbb{R}^{k_n} \\
 & \mathbf{D}_{k_s}(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}) & \in \mathbb{R}^{N \times k_s} \\
 & \mathbf{E}_{k_n}(\xi_{k_n}) & \in \mathbb{R}^{N \times (k_n+2)} \\
 & \mathbf{G}_k(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}, \xi_{k_n}) & \in \mathbb{R}^{N \times k'} \\
 & \mathbf{H}_N & \in \mathbb{R}^{N \times N}
 \end{array}$$

where,

$$[\mathbf{D}_{k_s}(\mathbf{c}_{k_s}, \mathbf{v}_{k_s})]_{i,j} = \frac{1}{\sqrt{2\pi v_{j,k_s}}} \exp\left(-\frac{(x[i] - c_{j,k_s})^2}{2v_{j,k_s}}\right) \quad (28)$$

$$[\mathbf{E}_{k_n+2}(\xi_{k_n})]_{ij} = S(\xi_{i,k_n})_{ij} \quad (29)$$

$$[\mathbf{H}]_{ij} = \frac{1}{\sqrt{2\pi v_{sp}}} \exp\left(-\frac{(x[j] - x[i])^2}{2v_{sp}}\right) \quad (30)$$

The final form (27) is that of a general linear model commonly found throughout the Gibbs sampler MCMC and numerical Bayesian estimation literature.

○

## PRIOR DISTRIBUTIONS

- We define the set of parameters associated with the joint model of order  $k$  as,

$$\theta_k \triangleq [\mathbf{c}_{k_s}, \mathbf{v}_{k_s}, \xi_{k_n}, \mathbf{a}_{k'}, \sigma_k^2].$$

The entire parameter space  $P$  can be written as a finite union of subspaces:

$$P = \bigcup_{k_s=0}^{k_{s\max}} \bigcup_{k_n=0}^{k_{n\max}} \{k_s \times k_n\} \times \Theta_k \times \Psi, \quad (31)$$

where  $\Theta_k$  corresponds to the space in which the vector  $\theta_k$  lies.  $\Psi$  is the

space in which the hyperparameters,  $\psi$ , lie.

$$\Theta_{\{0,i\}} \triangleq (\text{noise variance})$$

$$\triangleq \mathbb{R}^+ \quad \text{for } i = (0, 1),$$

$$\Theta_k \triangleq (\text{positions}) \times (\text{variances}) \times (\text{knots}) \times (\text{amplitudes}) \times (\text{noise variance})$$

$$\triangleq \mathbb{C}_{k_s} \times \mathbb{R}^{+k_s} \times \mathbb{C}_{k_n} \times \mathbb{R}^{k'} \times \mathbb{R}^+ \quad \text{for } k_s \in \{1, \dots, k_{s_{\max}}\}, k_n \in \{2, \dots, k_{n_{\max}}\}$$

where,

$$\mathbb{C}_{k_s} \triangleq (0, L)^{k_s},$$

$$\mathbb{C}_{k_n} \triangleq (0, L)^{k_n}.$$

- The structure of the joint prior for all the parameters is assumed to be,

$$\begin{aligned}
 p(\mathbf{c}_{k_s}, \mathbf{v}_{k_s}, \xi_{k_n}, \mathbf{a}_{k'}, \sigma_k^2, k, \psi) &= p(\mathbf{c}_{k_s} | k_s) p(\mathbf{v}_{k_s} | k_s) p(\xi_{k_n} | k_n) \\
 &\times p(\mathbf{a}_{k'} | \mathbf{c}_{k_s}, \mathbf{v}_{k_s}, \xi_{k_n}, k, \sigma_k^2) p(\sigma_k^2) p(k | \psi) p(\psi)
 \end{aligned}
 \tag{32}$$

- This is a hierarchical structure for the joint prior distribution, where the independence between parameters have been chosen for convenience. We assign the following distributions to the prior conditional densities for the parameters and the model dimension prior,

-

## BAYESIAN MODEL AND POSTERIOR DISTRIBUTION

A Bayesian model is described by the prior distribution  $p(\theta)$  of a random parameter  $\theta \in \Theta$  and by the likelihood  $p(\mathbf{y}|\theta)$  of the observations  $\mathbf{y}$ .

In this framework, all information on  $\theta$  based on the observations  $\mathbf{y}$  is included in the posterior distribution  $p(\theta|\mathbf{y})$  which one can obtain using Bayes' theorem

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}$$

where the normalizing constant  $p(\mathbf{y})$  is obtained by integration

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\theta) p(\theta) d\theta$$

The information given by the posterior distribution (??) might be too complicated to analyse directly when one is confronted with a decision problem. This can however be naturally handled in a Bayesian framework.

## BAYESIAN DECISION

When one wants to make a decision, the main goal is to minimise the risk of being wrong. In order to quantify the degree of error committed one introduces a cost function that penalises the values of  $\theta$  which we think are not satisfactory. This technique, well known in the orthodox statistical community, takes the following form in a Bayesian framework :

Let  $L(\cdot, \cdot) : \Theta^2 \rightarrow \mathbb{R}^+$  be a cost function and  $p(\theta | \mathbf{y})$  be the posterior distribution of  $\theta$ . The expected posterior cost function is defined as

$$\rho(\theta_*) \triangleq \int_{\Theta} L(\theta_*, \theta) p(\theta | \mathbf{y}) d\theta = \mathbb{E}_{p(\theta | \mathbf{y})} (L(\theta_*, \theta))$$

Given a cost function  $L(\cdot, \cdot) : \Theta^2 \rightarrow \mathbb{R}^+$ , then the associated Bayesian estimator is defined as

$$\hat{\theta}(\mathbf{y}) \triangleq \arg \theta_* \in \Theta \min \rho(\theta_*)$$

Obtaining a Bayesian estimator is thus in general a joint integration/optimisation problem, which takes simpler forms in the following two very important cases:

- the quadratic cost function,  $L_1(\theta_*, \theta) = (\theta_* - \theta)^T \mathbf{Q}(\theta_* - \theta)$ , for any positive definite matrix  $\mathbf{Q}$  leads to the posterior mean, also known as the MMSE estimate for  $\theta$

$$\hat{\theta}_{MMSE}(\mathbf{y}) = \int_{\Theta} \theta p(\theta | \mathbf{y}) d\theta$$

which is an integration problem.

- the cost function  $L_2(\theta_* - \theta; \delta) = 1 - \mathbb{I}_{\{\theta; \|\theta_* - \theta\|_2 \leq \delta\}}$  ( $\theta_* - \theta$ ) which gives as  $\delta \rightarrow 0$  the MAP estimator, that is

$$\hat{\theta}_{MAP}(\mathbf{y}) = \arg \theta \in \Theta \max p(\theta | \mathbf{y}) p(\theta)$$

which is an optimisation problem.

Assume here that  $\theta = (\theta_1, \dots, \theta_{n_\theta}) \in \Theta \subset \mathbb{R}^{n_\theta}$ . Then the evaluation of marginal distributions, *e.g.*

$$p(\theta_i | \mathbf{y}) = \int p(\theta | \mathbf{y}) d\theta_1 \dots d\theta_{i-1} d\theta_{i+1} \dots, d\theta_{n_\theta}$$

for  $i = 1, \dots, n_\theta$ , also requires integration. Similarly the evaluation of any marginal estimator (*e.g.*  $\mathbb{E}[\theta_i | \mathbf{y}]$  or  $\arg \theta_i \max p(\theta_i | \mathbf{y})$  for  $i = 1, \dots, n_\theta$ )

involves extra-integration steps over the parameters that one wants to integrate out. Other quantities can be of interest in order to qualify the estimator: these include, for example, the conditional covariance

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})} (\boldsymbol{\theta}\boldsymbol{\theta}^T) - \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})} [\boldsymbol{\theta}] \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})} [\boldsymbol{\theta}^T]$$

which again involves integrations.

## MODEL CHOICE

Assume that we are analyzing data  $\mathbf{y}$  and we believe that the data arise from one of a set of possible models  $\mathcal{M}_0, \dots, \mathcal{M}_{k_{\max}}$ , where under model  $\mathcal{M}_i$ ,  $\mathbf{y}$  has density  $p_i(\mathbf{y}|\theta_i)$ , conditional on  $\theta_i \in \Theta_i$ . The parameter vectors  $\theta_i$  are unknown and are typically of different dimension. Let  $p_i(\theta_i)$  denote the prior density for  $\theta_i$ , and let  $p_i$  denote the prior probability of the model  $\mathcal{M}_i$ . For the sake of convenience, we introduce a random variable  $k \in \{0, \dots, k_{\max}\}$  such that  $\Pr(k = i) = \Pr(\mathcal{M}_i) = p_i$ . The prior probability distribution for the random parameters  $(k, \theta)$  is defined on a space of the form  $\Theta \triangleq \bigcup_{i=0}^{k_{\max}} \{i\} \times \Theta_i$  and can be written

$$p(k, d\theta) = \sum_{i=0}^{k_{\max}} p_i(i, d\theta_i) \mathbb{I}_{\{i\} \times \Theta_i}(k, \theta),$$

where

$$p_i(i, d\theta_i) = p_i(\theta_i) d\theta_i$$

(we assume that  $p_i(d\theta_i)$  admits a dominating measure  $d\theta_i$ , usually the Lebesgue measure) and

$$\mathbb{I}_{\{i\} \times \Theta_i}(k, \theta) = \begin{cases} 1, & \text{if } (k, \theta) \in \{i\} \times \Theta_i \\ 0, & \text{otherwise} \end{cases},$$

*i.e.*  $(k, \theta)$  is in one of the spaces  $\{i\} \times \Theta_i$ , and the prior probability of  $k$  being equal to  $i$  and for  $\theta$  being in an infinitesimal set centered around  $\theta_i$  is  $p_i(i, \theta_i) d\theta_i$ .

After observing  $\mathbf{y}$ , one obtains the posterior distribution using Bayes' theorem

$$p(k, d\theta | \mathbf{y}) = \sum_{i=0}^{k_{\max}} p(i | \mathbf{y}) p_i(d\theta_i | \mathbf{y}) \mathbb{I}_{\{i\} \times \Theta_i}(k, \theta)$$

where  $p_i(d\theta_i | \mathbf{y}) p(i | \mathbf{y})$  is the posterior probability of model  $\mathcal{M}_i$  and is given by

$$p(i | \mathbf{y}) \triangleq p(\mathcal{M}_i | \mathbf{y}) = \frac{m_i(\mathbf{y}) p_i}{\sum_{i=0}^{k_{\max}} m_i(\mathbf{y}) p_i},$$

where

$$m_i(\mathbf{y}) \triangleq p(\mathbf{y}|i) = \int_{\Theta_i} p_i(\mathbf{y}|\theta_i) p_i(\theta_i) d\theta_i$$

is called the *marginal* distribution of  $\mathbf{y}$  under model  $\mathcal{M}_i$ . Assuming  $\mathcal{M}_i$  is the *true* model,  $p(\mathbf{y}|i)$  is the density according to which  $\mathbf{y}$  will actually occur. For this reason,  $m_i(\cdot)$  is also called the *predictive* density of  $\mathbf{y}$ . Under a 0-1 loss function, the optimal model is that  $\mathcal{M}_i$  which maximizes the posterior model probability  $p(i|\mathbf{y})$ ,  $i = 1, \dots, k_{\max}$ .

Note that  $p(i|\mathbf{y})$  can be written as

$$p(i|\mathbf{y}) = \left( 1 + \sum_{j \neq i} \frac{p_j}{p_i} B_{ji} \right)^{-1},$$

where the factor

$$B_{ji} = \frac{m_j(\mathbf{y})}{m_i(\mathbf{y})}$$

is called the *Bayes factor* of model  $\mathcal{M}_j$  against  $\mathcal{M}_i$ . Intuitively, the Bayes factor can be interpreted as the odds of  $\mathcal{M}_j$  against  $\mathcal{M}_i$  given by the

observations. Note that Bayes factors can be used to summarize the analysis independently of the model prior beliefs,  $p_i$ .

The Bayesian approach to model selection can be applied to a wide variety of problems, including multiple comparisons and the testing of non-nested hypotheses. The results are easily interpreted (as opposed to frequentist P-values) and automatically penalize overparametrizations.

## DISCUSSION

Bayesian statistics involves integration and/or optimisation steps, see expressions (??), (??), (??) and (??) for example. Except in certain special cases, Bayesian inference cannot be performed analytically, and this will be illustrated on several applications in Section ??. The ability to integrate and/or maximize complex multi-dimensional functions is thus extremely important in Bayesian statistics. This problem has severely limited the development of the Bayesian approach in statistics and related fields. Monte Carlo methods are a set of powerful numerical methods which allow to partly solve it.