RANDOM SIGNALS

BAYESIAN METHODS

P.R. WHITE

BAYES' THEOREM

BAYESIAN INFERENCE (PARAMETER ESTIMATION)

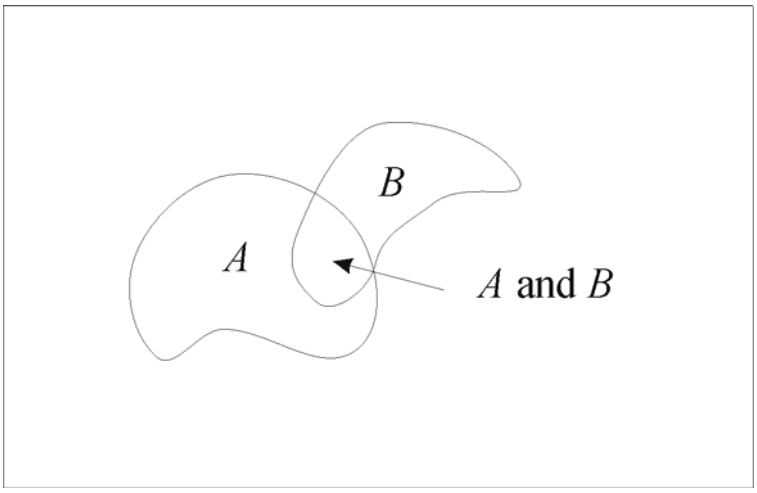
UNIFORM PRIOR DISTRIBUTIONS

NON-INFORMATIVE PRIORS

NUISANCE PARAMETERS

 © ISVR 2003

CONDITIONAL PROBABILITIES



The conditional probability $p(A|B)$ represents the probability of observing A given that B occurs.

Inspection of the above Venn diagram allows one to see that:

$$p(A \mid B) = \frac{p(A \text{ and } B)}{p(B)}$$ [1]

This is ratio of the area of the overlap between A and B to the total area of B.

By the same logic:

$$p(B \mid A) = \frac{p(A \text{ and } B)}{p(A)}$$

[1] $p(A \text{ and } B)$ is more formally written as $p(A \cap B)$

 © ISVR 2003

BAYES' THEOREM

The two previous two expressions can be re-arranged to read:

$$p(A \text{ and } B) = p(A \mid B)p(B) \quad \text{and} \quad p(A \text{ and } B) = p(A)p(B \mid A)$$

Equating these two expressions leads to Bayes' theorem

$$p(A \mid B)p(B) = p(B \mid A)p(A)$$

The form that we shall most frequently make use of is:

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

This expression allows us to "reverse" conditional probabilities, *i.e.* it tells us how to compute $p(A|B)$ from $p(B|A)$.

An alternative form for Bayes' theorem for multiple mutually independent events, A_j j=1,…,J is:

$$p(A_k \mid B) = \frac{p(B \mid A_k)\, p(A_k)}{\sum_{j=1}^{J} p(B \mid A_j)\, p(A_j)}$$

Note the above formula also requires that $A_1 \cup A_2 \cup \ldots \cup A_J = B$ (this is why putting J=1 in the above does not directly give the previous result).

 © ISVR 2003

MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

We have already considered the general problem of parameter estimation, namely given the data **x** can we compute estimates of the model parameters θ.

The principle adopted was that of Maximum Likelihood, which consisted of maximising the likelihood function, $l(\theta)$:

$$l(\theta) = p(\mathbf{x} \mid \theta)$$

Note it is commonly more useful to take logarithms and consider the problem of maximising:

$$L(\theta) = \log\{p(x \mid \theta)\}$$

It is important to realise that the conditional probability, $p(\mathbf{x} \mid \theta)$, is the "wrong way round". It is the probability of the data given a set of parameters.

The logical function to maximise is actually $p(\theta \mid \mathbf{x})$, *i.e.* the probability of the parameters given the data. In this case one finds the most likely parameters given the data.

 © ISVR 2003

BAYESIAN PARAMETER ESTIMATION

We can use Bayes' theorem to reverse the conditional probability.

Applying Bayes theorem leads to:

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta)p(\theta)}{p(\mathbf{x})} = \frac{l(\theta)p(\theta)}{p(\mathbf{x})}$$

The value of θ that maximises $p(\theta \mid \mathbf{x})$ is called the Maximum A Posteriori (MAP) estimate.

In the expression for $p(\theta \mid \mathbf{x})$ you should note that the term $p(\mathbf{x})$ is independent of the parameters θ.

Hence when maximising $p(\theta \mid \mathbf{x})$ the term $p(\mathbf{x})$ is a (positive) constant and so that maximising $p(\theta \mid \mathbf{x})$ gives the same answer as maximising just $l(\theta)p(\theta)$.

The term $p(\mathbf{x})$ is referred to as the *evidenc*e.

Whilst the term $p(\theta)$ is called the *prior*.

 © ISVR 2003

WHY USE THE POSTERIOR DENSITY?

What advantages are there to be gained by using $p(\theta \mid \mathbf{x})$ (the posterior density) in place of $p(\mathbf{x} \mid \theta)$ (the likelihood)?

1) It is the logical quantity to maximise.

It is more reasonable to be maximising the likelihood of the parameters given the data, rather than the likelihood of the data given the parameters.

2) Inclusion of prior information.

If one knows before an experiment that certain model parameters are likely to be distributed in some manner, then this information can be incorporated (in an appropriate manner) via the prior distribution $p(\theta)$.

3) One can remove parameters that are of no interest by integration.

For example consider the problem of estimating a (complex) sine wave's frequency. The model used is of the form:

$$x(t) = Ae^{2\pi ift}$$

The parameter *A* is of no direct interest. It can be removed using Bayesian methods (see later for details).

 © ISVR 2003

UNIFORM PRIOR DISTRIBUTIONS

The pdf $p(\theta)$ reflects ones (assumed) knowledge of the distribution parameters before the measurements **x** are made.

In many cases one may not have such prior information.

In the absence of such prior information it is natural to assume that all the parameters are equally likely, *i.e.*

$$p(\theta) = C$$

In this case the MAP estimate entails the maximisation of the function:

$$p(\theta \mid \mathbf{x}) \propto l(\theta) p(\theta) = C\, l(\theta)$$

Under these circumstances the MAP estimate corresponds to the maximum likelihood estimate.

 © ISVR 2003

PROBLEMS WITH UNIFORM PRIORS (I)

Consider a parameter that can take any value in the infinite interval $\theta \in [a,b]$. Consider the question as to what value C should take?

Since $p(\theta)$ is a pdf it must integrate to unity, *i.e.*

$$\int_a^b p(\theta)d\theta = 1 \quad \Rightarrow \quad \int_a^b Cd\theta = C(b-a) = 1 \quad \Rightarrow \quad C = \frac{1}{b-a}$$

As long as the region is finite this allows one to select C so that $p(\theta)$ is a proper pdf.

However if the interval is such that either a=-∞ and/or b=∞ then C=0, and the methodology falls down.

For example if one is estimating phase of a sine wave then one can form a uniform prior with $C = 1/(2\pi)$.
Alternatively if one is estimating the amplitude of a sine wave then any value is possible and so there is no natural choice for $C \neq 0$

One solution is to pick a value for C and overlook the fact that the implied prior $p(\theta)$ is not a true pdf (it is termed *improper*).

 © ISVR 2003

PROBLEMS WITH UNIFORM PRIORS (II)

The second problem associated with use a uniform prior comes if one can parameterise the same model in several ways.

For example consider the problem of estimating the parameters for a digital sine wave:

One might choose to parameterise the model in terms of frequency f hence use a uniform prior with $C = 2/f_s$ where f_s is the sampling frequency.

Alternatively one might choose to parameterise the model in terms of the period $T = 1/f$ and then it is natural to try a uniform prior for T.

However these are mutually incompatible, if there is a uniform prior on f then the distribution for T is NOT uniform.

So that, in general, assuming a uniform distribution for one parameterisation of a model will not be equivalent to assuming a uniform distribution for another parameterisation.

 © ISVR 2003

EXAMPLE OF USING PRIOR INFORMATION (I)

Consider the linear regression problem:

$$y_n = ax_n + b + w_n$$

Assuming a white Gaussian noise model (with variance σ^2) for w_n then the likelihood function for this problem is:

$$p(\mathbf{x} \mid a,b) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N} w_n^2} = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N} (y_n - ax_n - b)^2}$$

Let us know consider the case where we expect that the slope of our model, *a*, should be small. To impose this prior knowledge one might suppose a prior distribution for *a* of the form:

$$p(a) = \frac{1}{\sigma_a\sqrt{2\pi}} e^{-a^2/(2\sigma_a^2)}$$

That is *a* is modelled has having a Gaussian distribution with zero mean and a variance of σ_a^2.

If use Bayes' theorem to construct the posterior probability we obtain:

$$p(a,b \mid \mathbf{x}) \propto \frac{1}{(2\pi)^{(N+1)/2}\sigma_a\sigma^N} e^{-\left\{\frac{a^2}{2\sigma_a^2} + \frac{1}{2\sigma^2}\sum_{n=1}^{N} (y_n - ax_n - b)^2\right\}}$$

 © ISVR 2003

EXAMPLE OF USING PRIOR INFORMATION (II)

Taking logs leads to

$$\log\left\{ p\left(a,b\,|\,\mathbf{x}\right)\right\} = Const - \frac{a^2}{2\sigma_a{}^2} - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\ \left(y_n - ax_n - b\right)^2$$

So that maximising $\log\left\{ p\left(a,b\,|\,\mathbf{x}\right)\right\}$ is equivalent to minimising the cost function:

$$\Psi = \sum_{n=1}^{N}\ \left(y_n - ax_n - b\right)^2 + \frac{a^2\sigma^2}{\sigma_a{}^2}$$

The first term in Ψ is the least squares cost function and the second term penalises large values of *a*.

Solving the above for *a* gives: $a = \dfrac{\rho_{xy} - \bar{x}\,\bar{y}}{\rho_{xx} - \bar{x}^2 + \alpha}$

$$\rho_{xy} = \frac{1}{N}\sum_{n=1}^{N}\ x_n y_n,\ \rho_{xx} = \frac{1}{N}\sum_{n=1}^{N}\ x_n y_n,\ \bar{x} = \frac{1}{N}\sum_{n=1}^{N}\ x_n,\ \bar{y} = \frac{1}{N}\sum_{n=1}^{N}\ y_n,\ \alpha = \frac{\sigma^2}{N\sigma_a{}^2}.$$

The (standard) least squares solution is: $a = \dfrac{\rho_{xy} - \bar{x}\,\bar{y}}{\rho_{xx} - \bar{x}^2}$

Thus for large values of $\sigma_a{}^2$ the Bayes solution tends to zero, whilst for small values of $\sigma_a{}^2$ the Bayes solution tends to the LS solution.

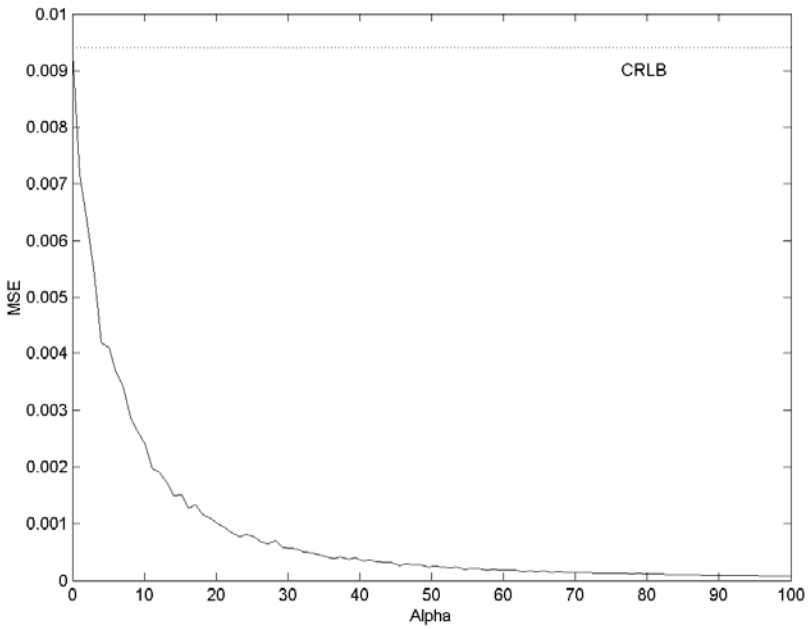
 © ISVR 2003

EXAMPLE OF USING PRIOR INFORMATION (III)

To test the performance of this estimator 1000 realisation of noisy data consisting 11 points generated from a line with a=0 were constructed.

The Bayes estimator was used with different values of α. The MSE was computed for the 1000 realisations.

The Cramer Rao Lower Bound (CRLB) was also found.



Note that the Bayes estimator actually exceeds the CRLB! This is possible since **accurate** prior information is available. In this case the value of a is assumed to be Gaussian with zero mean, when in fact it was always exactly zero!

 © ISVR 2003

EXAMPLE 2 (I)

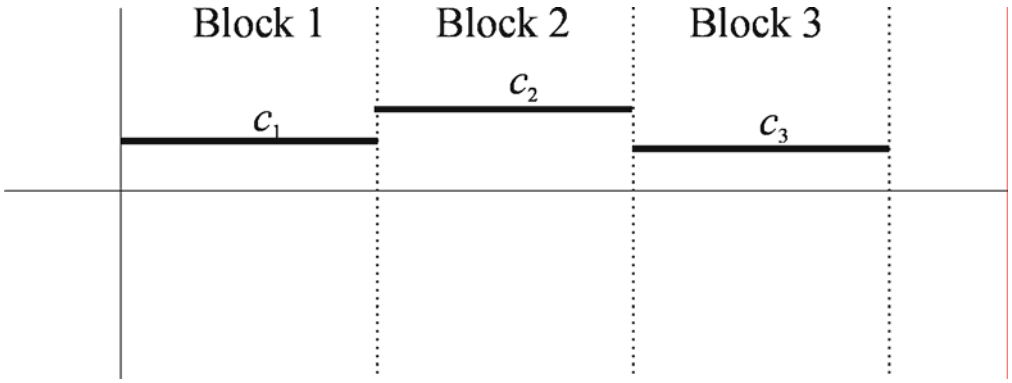
Consider the problem of estimating the mean of three connected data sets. We have a total of $3L$ samples in three blocks each of L samples.

The goal is to compute the means of the data within each of the 3 blocks.

The whole data is $x(1), \ldots, x(3L)$ with:

$$x_1(1), \ldots, x_1(L) = x(1), \ldots, x(L)$$
$$x_2(1), \ldots, x_2(L) = x(L+1), \ldots, x(2L)$$
$$x_3(1), \ldots, x_3(L) = x(L+1), \ldots, x(3L)$$

The 3 means we wish to estimate are c_1, c_2, and c_3.



The data is assumed to come from a slowly varying process, so that we also anticipate that the means should not be very different from one block to the next.

 © ISVR 2003

EXAMPLE 2 (II)

The data is assumed to be Gaussian so that, for example:

$$p(\mathbf{x}_1 \mid c_1) = \frac{1}{(2\pi)^{L/2}\sigma^L} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{L} (x_1(n)-c_1)^2}$$

The information with regards to the c_k's being similar is included via a prior condition. The difference in the means is assumed to also follow a Gaussian distribution with zero mean and variance σ_c e.g.

$$p(c_1 - c_2) = \frac{1}{\sqrt{2\pi}\,\sigma_c} e^{-(c_1-c_2)^2/\left(2\sigma_c{}^2\right)}$$

This leads to a posterior density of the form:

$$p(c_1, c_2, c_3 \mid \mathbf{x}) = p(\mathbf{x} \mid c_1)p(c_1 - c_2)p(\mathbf{x} \mid c_2)p(c_2 - c_3)p(\mathbf{x} \mid c_1)$$

$$\log\left\{ p(c_1, c_2, c_3 \mid x)\right\} \propto$$

$$\sum_{n=1}^{L} \left(x_1(n) - c_1\right)^2 + \sum_{n=1}^{L} \left(x_2(n) - c_2\right)^2 + \sum_{n=1}^{L} \left(x_3(n) - c_3\right)^2 + \frac{\sigma^2}{\sigma_c{}^2}(c_1 - c_2)^2 + \frac{\sigma^2}{\sigma_c{}^2}(c_2 - c_3)^2$$

Minimising this with respect to the means c_1, c_2, and c_3 leads to:

$$-2\sum_{n=1}^{L} x_1(n) + 2Lc_1 + \frac{\sigma^2}{\sigma_c{}^2} 2(c_1 - c_2) = 0$$

$$-2\sum_{n=1}^{L} x_2(n) + 2Lc_2 + \frac{\sigma^2}{\sigma_c{}^2} 2(c_2 - c_1) + \frac{\sigma^2}{\sigma_c{}^2} 2(c_2 - c_3) = 0$$

$$-2\sum_{n=1}^{L} x_1(n) + 2Lc_1 + \frac{\sigma^2}{\sigma_c{}^2} 2(c_3 - c_2) = 0$$

 © ISVR 2003

EXAMPLE 2 (III)

This can be written in matrix form as:

$$\begin{bmatrix} 1+\alpha & -\alpha & 0 \\ -\alpha & 1+2\alpha & -\alpha \\ 0 & -\alpha & 1+\alpha \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{bmatrix}$$

where $\bar{x}_k = \frac{1}{L}\sum_{n=1}^{L} x_k(n)$ and $\alpha = \frac{\sigma^2}{L\sigma_c^{\,2}}$

For large $\sigma_c^{\,2}$ (small α) then the solution is close to the individual means: $c_k = \bar{x}_k$

For small $\sigma_c^{\,2}$ (large α) then the solution is close to the overall mean,

$$c_1 = c_2 = c_3 = \frac{1}{3}\left(\bar{x}_1 + \bar{x}_2 + \bar{x}_3\right)$$

 © ISVR 2003

PHILOSOPHIES FOR CHOOSING PRIORS (I)

1) *Conjugate Priors*

These are so called c*onvenience* priors since they are designed to make life easy and not to satisfy any mathematical principles.

One chooses the prior distribution $p(\theta)$ so that the posterior distribution $p(\theta|\mathbf{x}) \propto l(\mathbf{x})\, p(\theta)$ is from the same family of distributions as $p(\theta)$.

Example

Consider a problem with a Poisson likelihood:

$$p(x|\theta) = \frac{e^{-\theta}\theta^{x}}{x!}$$

If one uses a Gamma distribution prior (with parameters α and β):

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$$

Then the posterior distribution is:

$$p(\theta|x) \propto \theta^{y+\alpha-1}e^{-\theta(1+1/\beta)}$$

The posterior is also a Gamma distribution (with different parameters)

 © ISVR 2003

PHILOSOPHIES FOR CHOOSING PRIORS (II)

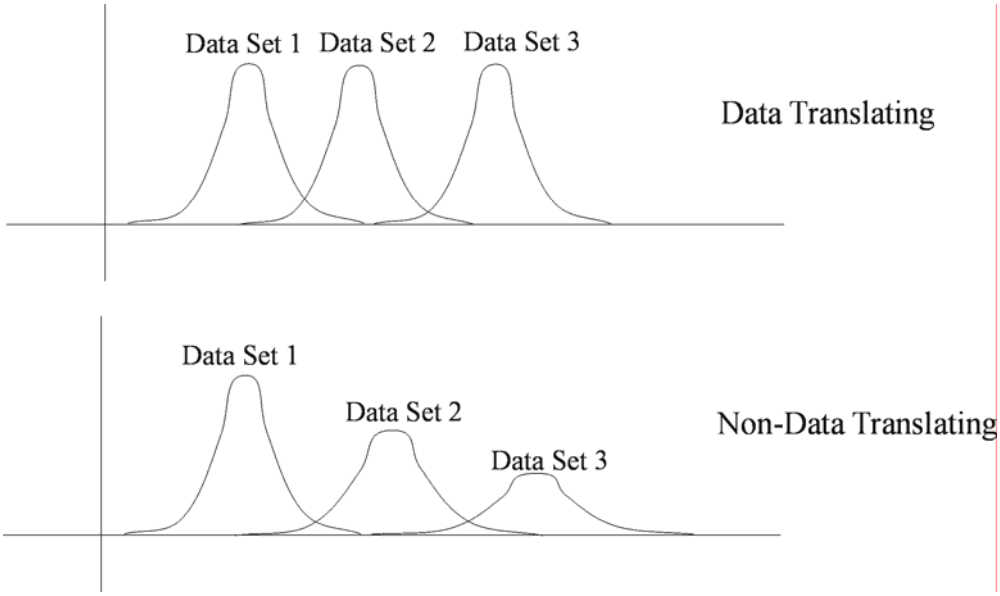
2) *Non-Informative Priors*

This is for applications where we wish to assume no prior information.

One seeks to overcome some of the problems associated with uniform priors. One seeks a prior that has minimum influence of relative to the likelihood.

One way to express this is to seek a prior that leads to a posterior distribution that is *data translated* in some transformed parameter space.

Data translated posterior one that has the same shape for each value of θ, the effect of the data **x** is to move the posterior along the θ-axis.



 © ISVR 2003

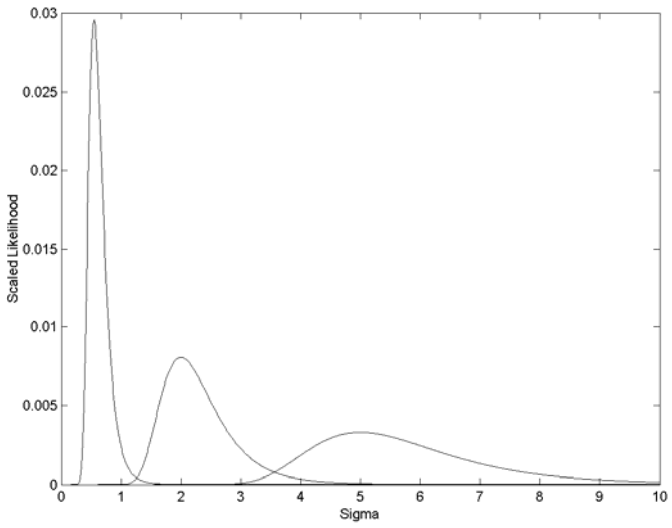
EXAMPLE OF A NON-INFORMATIVE PRIOR (I)

Consider the problem of estimating the standard deviation of zero Gaussian data.

The likelihood function for which is:

$$l(\sigma) \propto \frac{1}{\sigma^L} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{L} x(n)^2} = \frac{1}{\sigma^L} e^{-\frac{Ls}{2\sigma^2}}$$

where $s = \frac{1}{L}\sum_{n=1}^{L} x(n)^2$ is the only place that the data enters the likelihood (s is a *sufficient statistic*).



The above shows the (unnormalised) likelihood functions for different values of s. Clearly the likelihood is not data translating.

 © ISVR 2003

EXAMPLE OF A NON-INFORMATIVE PRIOR (II)

We can select a prior that will render the posterior density (approximately) data translating.

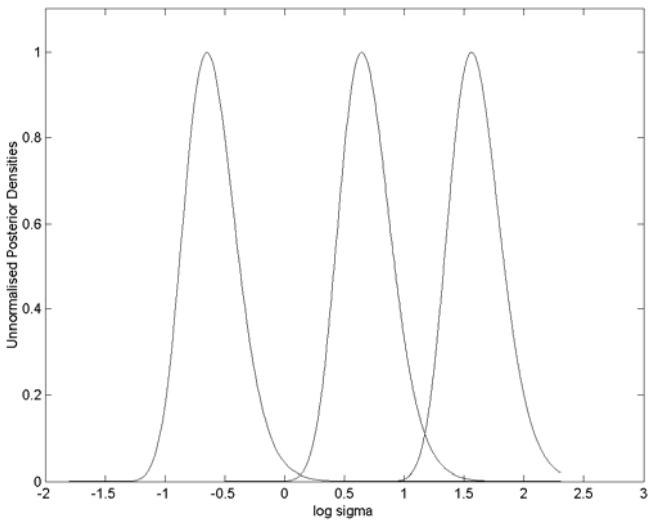
Specifically if we choose:

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

This is Jeffrey's prior and the resulting posterior density has the form:

$$p(\sigma \mid x) = \frac{(ns)^{n/2}}{2^{(n/2)-1}\Gamma(n/2)} \sigma^{-(n+1)} e^{-\frac{ns}{2\sigma^2}}$$

This produces a non-informative prior for $\log(\sigma)$.



 © ISVR 2003

NUISANCE PARAMETERS

Data models often have many parameters, some of which are of interest, others of which are not.

The parameters that are of no interest are called *nuisance parameters*.

Example

Consider a sine wave in noise:

$$x(t) = A\sin(2\pi ft + \phi) + \sigma w(t)$$

$w(t)$ is a unit variance white noise process (notice σ has been explicitly included in the model, it is the st. deviation of the noise).

If one seeks to estimate only frequency then the model parameters A, ϕ and σ are all nuisance parameters.

Such nuisance parameters are necessary in a mode to make it sufficiently general.

But using maximum likelihood they have to be estimated along with the parameters we are interested in.

So we have to estimate amplitude and phase to compute frequency!

 © ISVR 2003

REMOVING NUISANCE PARAMETERS

Consider a bivariate pdf $p(x, y)$. If one wants to obtain the pdf of x alone then you use:

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

We seek to use this to "remove" nuisance parameters.

For our example:

$$\begin{aligned} p(f \mid x) &= \iiint p(A, f, \phi, \sigma \mid x) dA d\phi d\sigma \\ &= \iiint p(A)\, p(\phi)\, p(\sigma)\, p(x \mid A, f, \phi, \sigma) dA d\phi d\sigma \end{aligned}$$

gives us an expression for the posterior distribution for f frequency alone.

This allows the estimation of frequency without explicitly estimating the nuisance parameters (A, ϕ and σ).

This is only possible when using posterior distributions – the likelihood, $p(x \mid A, f, \phi, \sigma)$, has to be multiplied by the priors.

 © ISVR 2003

EXAMPLE

Recall the sine wave parameter estimation problem:

$$x(n) = A\sin(2\pi fn\Delta + \phi) + w(n)$$

(note - we have returned to using σ implicitly and are employing a digital model)

The likelihood function is:

$$l(A,\phi,\sigma,f) = p(A,\phi,\sigma,f \mid x) \propto \sigma^{-L} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{L}\left(x(n)-A\sin(2\pi fn\Delta+\phi)\right)_2}$$

Assuming a uniform prior for the amplitude A and phase ϕ and integrating with respect to them leads to:

$$p(\sigma, f \mid x) \propto \sigma^{-(L+2)} e^{-\frac{1}{2\sigma^2}\left(\sigma_x{}^2 - 2|X(f)|^2/L\right)}$$

where $\sigma_x{}^2 = \frac{1}{L}\sum_{n=1}^{L} x(n)^2$.

If we now assume a Jeffrey's prior for σ and integrate with respect to it then:

$$p(f \mid x) \propto \left(1 - \frac{2|X(f)|^2}{L\sigma_x{}^2}\right)^{(2-L)/2}$$

This is essentially the result we saw using maximum likelihood!

 © ISVR 2003

SUMMARY

Bayesian methods allow us to convert likelihood functions into real pdfs.

This is powerful since it allows us to treat problems in a structured form and permits us to:

1) Include prior information in an appropriate manner.
   This allows one to improve performance of estimation algorithms (even potentially beating the CRLB!).

2) To “integrate out” unwanted nuisance parameters.
   This reduces the dimensionality of problems producing simpler and more robust inference schemes.

 © ISVR 2003