Bayesian Data Analysis

W.J.Fitzgerald
Department of Engineering
University of Cambridge
Cambridge U.K.

Abstract

In this paper, methods of Bayesian inference will be introduced, and applied to the problem of the estimation of model parameters that are to be inferred given data and any relevant prior knowledge. The methods are also applied to the selection of the best model, within a set, to account for the observed data, and the problem of signal detection is also addressed.

1 Introduction

In data analysis one is interested to know whether the model function that is being used to describe the data, is actually appropriate for the particular problem under investigation, and if so, to extract values for the parameters of the model. We therefore need a way to choose between several possible models.

To solve this problem, one must enumerate the possible models and realise that in terms of real data the correct model may not be within the set choosen. All that we can do is compare various models within a set that we have defined to see which models are more plausible.

This, and the problem of extracting values for model parameters, is a problem of Scientific Inference, and to carry out consistent reasoning and inference, one should use Bayesian methods.

The structure of the paper is as follows. An introduction to Bayesian inference will be given in the next section, followed by the application to parameter estimation, model order selection and signal detection. Results are given for the determination of the number of decaying sinusoids present in a data sequence together with the estimation of the parameters, and conclusions are given in the final section.

2 Bayesian Inference

Bayesian inference is a mode of scientific inductive reasoning that has been applied to a vast number of scientific disciplines, Laplace (1812), Jaynes (1989). A distinctive feature of

the method is that it allows investigators to use both sample and explicit prior information in a logically consistent manner in making inferences. Bayes' theorem, sometimes also referred to as the principle of inverse probability, serves as a fundamental learning model in the Bayesian approach. Initial or prior information concerning a hypothesis H, written as $P(H|I)$, is combined with current sample information, given by the likelihood of the data D, written as $P(D|HI)$, to produce a 'posterior' distribution, $P(H|DI)$ that incorporates both prior and sample information, together with a very important normalisation factor $P(D|I)$, which will later be referred to as the 'evidence'. Bayes' theorem follows from some simple rules of probability theory, and can be written

$$P(H|DI) = P(H|I)\frac{P(D|HI)}{P(D|I)}$$

R.T.Cox in 1946 developed a rigorous mathematical theory which considers probability as a measure of degree of plausibility. The theory emerged from a set of qualitative desiderata for a measure of plausibility; a) that degrees of plausibility can be represented by real numbers, and b) that the system must be internally consistent and consistent with common sense. The desiderata make no reference to frequencies, random variables, ensembles or imaginary experiments and they only refer to the plausibility of propositions.

One is, of course, free to use the resulting theory to consider propositions concerning frequencies in repeated experiments, but this emerges from the theory, frequency not being the basis for probability theory.

3 Parameter Estimation

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 $\{\omega\}$, 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.

Bayes' theorem for our particular problem may be written as

$$P(\{\omega\}|DI) = P(\{\omega\}|I)\frac{P(D|\{\omega\}I)}{P(D|I)}$$

which tells us the posterior probability of a hypothesis after data have been aquired in terms of the prior probability of the hypothesis and the likelihood of the data.

The term in the denominator is independent of $\{\omega\}$, and given the prior and the likelihood it may be calculated from the axioms of probability theory using

$$P(D|I) = \sum_i P(\omega_i|I)P(D|\omega_i I)$$

This term is therefore a normalisation term, also known as the marginal likelihood, and it plays a very important role in model comparison.

If only a subset of model parameters are of interest for a particular application, then the remaining parameters can be considered to be nuisance parameters, and may be eliminated from consideration by a process of marginalisation. For example, if the particular problem has two parameters α and β and we are only interested in α, then we may calculate $P(\alpha|DI)$ from the full posterior $P(\alpha\beta|DI)$ by the same method as above for the calculation of $P(D|I)$, and we can write

$$P(\alpha|DI) = \int d\beta P(\alpha\beta|DI)$$

This process of marginalisation is of both practical and theoretical interest, since it can significantly reduce the dimension of the problem being addressed. The condition under which one is justified in using marginalisation and the full implications in terms of possible bias in the estimated parameters, are discussed at length by Cox and Reid (1989).

From a considerations of the error statistics arising from a mismatch between the model and the data, the likelihood function for a general model function given by $\sum_{j=1}^{m} B_j G_j(t_i, \{\omega\}) = \bar{B}\bar{G}$ can be written

$$P(\bar{D}|HI) = (2\pi)^{-N/2}\sigma^{-N}\exp\left[-\frac{1}{2\sigma^2}(\bar{D} - \bar{B}\bar{G})^T(\bar{D} - \bar{B}\bar{G}))\right]$$

where $\bar{D}$ is a data vector of length N, $\bar{B}$ is a model amplitude vector of length m and $\bar{G}$ is a matrix of model functions of order $N \times m$.

To consider the marginal probabilities for the parameters independent of amplitudes, the following integrals have to be undertaken,

$$P(\{\omega\}|\bar{D}I) \propto \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} dB_1 \ldots dB_m P(\bar{D}|HI)P(B_1, B_2, \ldots, B_m, \{\omega\}|I)$$

Making the substitutions $\mathcal{P} = \bar{G}^T\bar{D}$ for the model functions projected onto the data, and $\mathcal{Q} = \bar{G}^T\bar{G}$ for the model functions projected onto themselves, and carrying out the integration, the required marginal posterior probability may be written

$$P(\{\omega\}|\bar{D}I) = \frac{P(\{\omega\}|I)}{P(D|I)}(2\pi)^{(m-N)/2}\frac{\sigma^{m-N}}{|\mathcal{Q}|^{1/2}}\exp\left[-\frac{1}{2\sigma^2}(\bar{D}^2 - \mathcal{P}^T\mathcal{Q}^{-1}\mathcal{P})\right]$$

It should be noted that the function $\mathcal{P}$ can be called a generalised periodogram, since when the model function is a sinusoid, $\mathcal{P}$ is the periodogram.

If the variance is unknown, it too can be treated as a nuisaince parameter and integrated out of the problem. In this case, a Jeffreys prior should be used, since the variance is a scale parameter (see Jeffreys 1939). However, for only a small number of data points, the final difference between the assumption of a uniform prior and a Jeffreys prior is very small. Integrating the noise variance between 0 and ∞ with a Jeffreys prior $1/\sigma$ gives

$$P(\{\omega\}|\bar{D}I) = \frac{P(\{\omega\}|I)}{P(D|I)} \frac{\pi^{(m-N)/2}}{|Q|^{1/2}} \frac{\Gamma[(N-m)/2]}{2} (\bar{D}^2 - \mathcal{P}^T Q^{-1} \mathcal{P})^{(m-N)/2}$$

which is of the form of the Student t-distribution, as has been noted many times in the statistical literature (Zellner (1971), West and Harrison (1989), Duijndam (1988), Broemeling (1985) and Bretthorst (1989)).

The formalism can also readily deal with coloured noise by replacing the noise variance above by the coloured noise covariance matrix . This has the effect of renormalising the model functions and the data, Fitzgerald (1991), Whalen (1971).

4 Model Comparison

It has been assumed in the formalism given above for the parameter estimation problem that the model under consideration was the correct model with which to interpret the data. This is obviously an assumption, and it is of interest to test this assumption, since the Bayesian framework allows for such a comparison by calculating the relative probabilities of candidate models within a given set. Also from the framework emerges a quantification of Ockham's razor, which states that simpler models are to be preferred unless a more complicated model provides a significantly better fit to the data.

To use Bayes' theorem for model comparison, the prior information I asserts that one of the models within the set is correct, and the probability of model k can be written

$$P(k|DI) = P(k|I) \frac{P(D|kI)}{P(D|I)}$$

If we have no prior information preferring one model over the others, we can take the prior $P(k|I) = 1/m$, where m is the number of models within the set.

The likelihood $P(D|kI)$, which may also be written as $P(D|I_k)$, can be calculated from

$$P(D|kI) \equiv P(D|I_k) = \int d\omega_k P(\omega_k|I_k) P(D|\omega_k I_k)$$

where I_k is the only relevant information concerning model k.

The marginal likelihood is given by

$$P(D|I) = \sum_k P(k|I) P(D|kI)$$

In the model comparison problem, if attention is focused on the ratio of probabilities of models, the so-called 'odds', we can write the odds in favour of model k over model j, say, as

$$O_{kj} \equiv \frac{P(k|DI)}{P(j|DI)}$$

which from above, may be written as

$$O_{kj} = \left[\frac{P(k|I)}{P(j|I)}\right] \frac{\int d\omega_k P(\omega_k|I_k)P(D|\omega_k I_k)}{\int d\omega_j P(\omega_j|I_j)P(D|\omega_j I_j)}$$

$$\equiv \left[\frac{P(k|I)}{P(j|I)}\right] B_{kj}$$

where the factor in brackets is called the prior odds, and by assumption that our prior knowledge does not prefer one model to any other in the set, this factor is unity. The factor B_{kj} is called the Bayes' factor, (or the weight of evidence), Jefferys (1939), Good (1983), and can be equivalently written as

$$B_{kj} = \frac{P(D|I_k)}{P(D|I_j)}$$

which is the ratio of the prior predictive probabilities. This will be used in the results section to determine the number of decaying sinusoids present in data, for which case the various hypotheses will be model functions containing different numbers of terms.

5 Bayesian Detection of Signals

Let us consider the Bayesian detection of a known signal model which is buried in noise, and which has unknown parameters. This type of problem arises frequently in the context of radar and sonar, where the form of the signal one is trying to detect maybe known, but the parameters are unknown and possibly time varying. It may also be the case that the signal one is trying to detect may not be present.

The simplest detection problem involves choosing between two possible hypothesis, say, H_0 and H_1, referring to the null hypothesis and the 'signal present' hypothesis respectively. The decision rule for the binary detection problem is determined by the ratio of the posterior probabilities of hypothesis H_1 to hypothesis H_0, given the data and the prior information. This can be put on a decibel scale, a positive dB rating favouring H_1, and a zero dB rating favouring neither hypothesis.

The decision criterion can thus be written as

$$K = 10\log_{10}\left[\frac{P(H_1|DI)}{P(H_0|DI)}\right]$$

where the posterior probabilities of the two hypothesis are given as before, by

$$P(H_0|DI) = \frac{P(\omega|I)}{P(D|I)}\pi^{(m-N)/2}\frac{\Gamma[(N-m)/2]}{2}(D^2)^{(m-N)/2}$$

and

$$P(H_1|DI) = \frac{P(\omega|I)}{P(D|I)}\frac{\pi^{(m-N)/2}}{|Q|^{1/2}}\frac{\Gamma[(N-m)/2]}{2}(D^2 - \mathcal{P}^T Q^{-1}\mathcal{P})^{(m-N)/2}$$

Results obtained using this method can be found in Fitzgerald (1991).

6 Results

The data shown in Fig 1. consists of seven decaying sinusoids in the presence of a very small amount of noise, (standard deviation of 10^{-5}), with parameters shown in Table 1. Although this is synthetic data, the parameters have been choosen to correspond with realistic acoustic or vibration situations.

The Bayesian model order selection method was applied to this data, by evaluating the posterior probabilities of models containing between one and nine decaying sinusoidal functions. It was found that the evidence was overwhelmingly in favour of model order seven, and the estimated parameters for this model are shown in Table 2. It is seen that all of the model parameters are in very close agreement with the correct values.

As the noise level increases, so the parameter estimates start to deviate from the true values, as one would expect from common sense, and eventually at a certain noise level the model order selection breaks down, giving a high probability for the wrong model. It is interesting to notice that in this case, the variance of the estimated parameters is large, telling us not to give a great deal of plausibilty to the results. If one observes the error between the fitted model and the data, then for the original data with only a small amount of added noise, the error 'looks' the same to the eye for model orders greater than four, and hence any model order selection technique relying on residual error, would fail whereas the Bayesian inference method yields very plausible results, gracefully degrading as the observation noise increases.

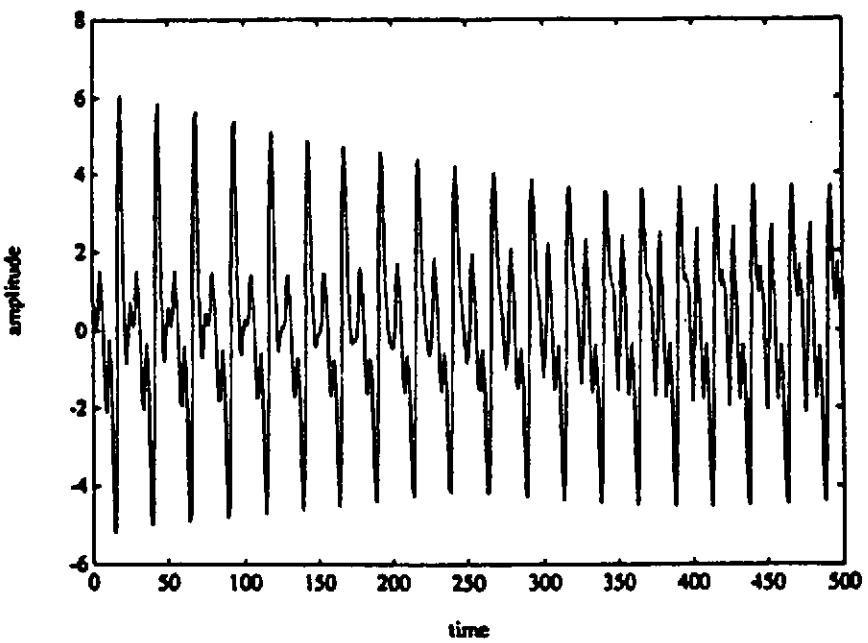

Fig 1. Simulated data consisting of 7 decaying sinusoids.

7 Conclusions

In this paper, an outline of the methodology used to conduct Bayesian inference in signal processing and data analysis has been given. The ideas associated with the removal of nuisance parameters have been described and the reduction in the dimensionality of the resulting parameter space has been pointed out. The intergration of the nuisance parameters can be carried out either analytically, for certain model functions, or numerically using purpose built fast and efficient numerical integrators.

It is clear that the methods can be adapted for the cases where the observation noise, and hence the form of the likelihood function, is non-Gaussian. For these cases it may not be possible to marginalise over the nuisance parameters analytically. One may require numerical integration or approximations, such as the expansion of the noise probability density function in terms of Hermite polynomials, to obtain the posterior probabilities of the parameters. This work is in progress.

It has been shown that the Bayesian paradigm gives a consistent framework with which to conduct scientific inference, but the full potential of the methods still remain to be explored in many engineering and applied science applications. It is hoped that this article will stimulate sufficient interest to enable the Bayesian methodology to be applied to signal processing in the very broadest sense.

The computer programs, written in C, Fortran and MATLAB, used for the analysis of the results presented in this paper are available upon request.

Frequency(Hz)	*Decay(s^{-1})*	*Amplitude*
200	4.0×10^{-4}	1.0
201	4.02×10^{-4}	1.0
401	8.02×10^{-4}	2.0
599	1.198×10^{-4}	1.0
602	1.204×10^{-4}	1.2
795	1.59×10^{-4}	0.5
803	1.606×10^{-4}	1.5

Table 1: Parameters of the data

Frequency(Hz)	*Decay(s^{-1})*	*Amplitude*
199.99	4.015×10^{-4}	0.996
200.99	4.218×10^{-4}	1.0
401.00	8.019×10^{-4}	1.999
598.99	1.198×10^{-4}	0.999
602.00	1.204×10^{-4}	1.199
795.00	1.590×10^{-4}	0.499
802.99	1.606×10^{-4}	1.499

Table 2: Inferred Parameters

8 References

1. Bretthorst, L. Bayesian Spectrum Analysis and Parameter Estimation, Lecture notes in statistics, Springer Verlag 1989.
2. Broemeling, L.D., Bayesian Analysis of Linear Models, Marcel Dekker, Inc, 1985.
3. Cox, D.R., and Reid, N., The Canadian Journal of Statistics, 17, 229, 1989.
4. Cox, R.T., Am.J.Phys, 17, 1, 1946.
5. Duijndam, A.J.W., Geophysical Prospecting, 36,878, 1988.
6. Fitzgerald, W.J., Bayesian Inference in Signal Processing, Cambridge University, Engineering Department Technical Report, CUED/F INFENG/T80, 1991.
7. Good, I.G., "Good Thinking" University of Minnesota Press, 1983.
8. Jaynes, E.T. Papers on Probability, Statistics and Statistical Physics, a reprint collection, ed. R.D.Rosenkrantz, Kluwer 1989.
9. Jefferys, H., Theory of Probability, Oxford University Press, 1939, reprinted 1985.
10. Laplace, P.S., Theorie Analytique des Probabilities, Paris, Courcier, 1812.
11. West, M., and Harrison, J., Bayesian Forcasting and Dynamic Models, Springer Verlag, 1989.
12. Whalen, A.D., Detection of signals in noise, p379, Academic Press, 1971
13. Zellner, A., An introduction to Bayesian Inference in Econometrics, New York, John Wiley and Sons, 1971.