Estimating Parameters of Sinusoids from Noisy Data Using Bayesian Inference with Simulated Annealing

DURSUN ÜSTÜNDAĞ, MEHMET CEVRİ
The Department of Mathematics, The Faculty of Science and Letters,
The University of Marmara, Göztepe Campus 34722, Istanbul, TURKEY.
e-mails: dustundag@marmara.edu.tr, mcevri@gmail.com

Abstract: In this paper, we consider Bayesian analysis proposed by Bretthorst for estimating parameters of the corrupted signals and incorporate it with a simulated annealing algorithm to obtain a global maximum of the posterior probability density of the parameters. Thus, this analysis offers different approach to find parameter values through a directed, but random, search of the parameter space. For this purpose, we developed a Mathematica code of this Bayesian approach and used it for recovering sinusoids corrupted by random noise. The simulation results support the effectiveness of the method.

Key-Words: Bayesian Statistical Inference, Simulated Annealing, Parameter Estimations, Optimization, Spectral Analysis, Signal Processing.

1 Introduction

The ultimate goal of collecting data is to gain meaningful information about a physical system. However, in many situations, the quantities that we would like to determine are different from the ones which we are able to have measured. If the data we measured depends on the quantities we want, it contains at least some information about them. Therefore, extracting this information from data is subject to this paper.

In many experiments, a discrete data set $\mathbf{D}=\{d_1,d_2,...,d_N\}^T$ is recorded. These values are sampled from an unknown function $y(t)$ at discrete times $\{t_1,....,t_N\}^T$ and also denoted as the output of the physical system that we want to be modeled:

$$d_i = y(t_i) = f(t_i,\mathbf{\theta}) + e_i,\ (i=1,2,...,N), \qquad (1)$$

where $f(t,\mathbf{\theta})$ is the model signal function, which is to be determined, and $\{e_i\}$ represents a vector of errors that are independent and are drawn from the zero mean Gaussian distribution with the variance of σ^2. Different models correspond to different choices of the model signal function $f(t,\mathbf{\theta})$. According to Bretthorst, the most general model may be given in the following form:

$$f(t,\mathbf{\theta}) = \sum_{j=1}^{m} \mathbf{B}_j \mathbf{G}_j(t,\{\omega\}) \qquad (2)$$

The model signals $\mathbf{G}_j(t,\{\omega\})$ are functions of parameters $\{\omega_1,...,\omega_\rho\}$ $(m=2\rho)$ that we label collectively $\{\omega\}$. The $\mathbf{B}_j$ represents the amplitude corresponding to the j th model function $\mathbf{G}_j(t,\{\omega\})$. The goal of data analysis is usually to infer the parameters $\mathbf{\theta}=\{\{\omega_1,B_{11},B_{21}\},...,\{\omega_\rho,B_{1\rho},B_{2\rho}\}\}$ by using the observed data values $\mathbf{D}$.

Besides estimating the parameters, there are two additional important problems. The one we will postpone to discuss here is to assess whether or not the model is appropriate for explaining the data. The other is to obtain an indication of the uncertainties in the parameter values, i.e. some measures of how far they are away from the true parameters.

In this paper, we address the problem of estimation of sinusoids in white Gaussian noise within a Bayesian framework. This problem is of great interest in many fields of science, including seismology, nuclear magnetic resonance and radar. Under an assumption of a known number of sinusoids, several algorithms have already been applied to the spectral analysis and parameter estimation problems, such as least-square fitting [11], discrete Fourier transform [6], and the periodogram [23]. With least square fitting, the model is completely defined and the question remaining is to find the parameters by minimizing the sum of squared residuals. This approach is usually too restrictive and assumes more about the problem than is really known. Generally, we do not know which model functions will fit the data best, or how many parameters are required to

fix the model functions. Since Cooley and Tukey[6] introduced the fast Fourier transform (FFT) technique in 1965, followed the rapid development of computers, the discrete Fourier transform has been a very powerful tool in spectral analysis. In a typical situation in spectral analysis an attempt is made to gain information about an underlying signal in a time series which has sampled discretely over a finite time, and which is corrupted by a random noise. In this case, the Fourier transform consists of the signal transform plus a noise transform. If the peak of signal transform is larger than the noise transform, the added noise does not change the location of the peak very much. One can then estimate the frequency from the location of the peak of the data transform, as intuition suggests. Unfortunately, this technique does not work well when the signal-to-noise ratio in the data is small. In 1987, Jaynes derived periodogram directly from the principles of Bayesian inference. After his work, researchers have given much attention to the relationship between Bayesian inference and parameter estimation. Bretthorst and some other researchers [1, 2, 6, 8,27 , 28] have done excellent works in this area for the last fifteen years.

In this study, we consider combining a Bayesian approach, suggested by Bretthorst[3,4,5] with a very different optimization algorithm, called simulated annealing[7]. It explores the entire surface of the posterior probability density of the frequencies and tries to optimize it while moving both uphill and downhill. Thus, it is largely independent of the starting values, often a critical input in conventional optimization algorithms. Furthermore, it can escape from local optima and go on to find the global optimum by the uphill and downhill moves. So a computer programme for the implementation of this method using *Mathematica* programming language is developed and then applied to some computer simulated results based on the single and multiple harmonic frequency models.

2 Bayesian Parameter Estimation

Let us now reconsider the problem given in Equation (1) within the Bayesian framework, which combines likelihood with prior information to produce a posterior distribution. The prior information generally takes the form of probabilities and probability distributions expressing our prior belief in parameter values and model hypothesis. If $\boldsymbol{\theta}$ is a set of parameters of interest, then their joint posterior probability density of the parameters given the observed data $\mathbf{D}$ and $\mathbf{I}$ is

$$P(\boldsymbol{\theta},\sigma^2 \mid \mathbf{D},\mathbf{I}) = \frac{P(\boldsymbol{\theta},\sigma^2 \mid \mathbf{I})P(\mathbf{D} \mid \boldsymbol{\theta},\sigma^2,\mathbf{I})}{P(\mathbf{D} \mid \mathbf{I})}. \qquad (3)$$

The quantity $P(\boldsymbol{\theta},\sigma^2 \mid \mathbf{I})$ is called the prior distribution; it represents prior knowledge of the parameters $\{\boldsymbol{\theta},\sigma^2\}$ given the information $\mathbf{I}$. The sampling distribution $P(\mathbf{D} \mid \boldsymbol{\theta},\sigma^2,\mathbf{I})$ is the likelihood of the data $\mathbf{D}$ given the model parameters. The probability function $P(\mathbf{D} \mid \mathbf{I})$ is a normalization constant of the posterior probability density. Therefore, Equation (3) can also be written in the following form:

$$P(\boldsymbol{\theta},\sigma^2 \mid \mathbf{D},\mathbf{I}) \propto P(\boldsymbol{\theta},\sigma^2 \mid \mathbf{I})P(\mathbf{D} \mid \boldsymbol{\theta},\sigma^2,\mathbf{I}). \qquad (4)$$

In words, the posterior probability density of the parameters is proportional to the product of the likelihood and the prior probability density of the parameters.

A key component in Bayes theorem is the likelihood function $P(\mathbf{D} \mid \boldsymbol{\theta},\sigma^2,\mathbf{I})$, which is proportional to the probability density of the noise. If the variance of the noise σ^2 is known, then the likelihood function takes on the following form:

$$P(\mathbf{D} \mid \boldsymbol{\theta},\sigma^2,\mathbf{I}) = (2\pi\sigma^2)^{\frac{-N}{2}} \exp\left[-\frac{NQ}{2\sigma^2}\right], \qquad (5)$$

where the exponent Q is defined as follows

$$Q = \sum_{i=1}^{N}\left(d_i - \sum_{j=1}^{m} B_j G_j(t_i,\{\omega\})\right)^2. \qquad (6)$$

This is equivalent to

$$Q = \overline{d^2} - \frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N} d_i B_j G_j(\{\omega\},t_i) + \frac{1}{N}\sum_{j=1}^{m}\sum_{k=1}^{m}\Omega_{jk} B_j B_k, \qquad (7)$$

where

$$\Omega_{jk} = \sum_{i=1}^{N} G_j(\{\omega\},t_i) \otimes G_k(\{\omega\},t_i), (j,k = 1,...,m). \qquad (8)$$

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 *marginalization.* In order to obtain the probability density that depends only on the

frequencies $\{\omega\}$, therefore we need to integrate out the *nuisance parameters* $\{\mathbf{B}\}$ and σ dependence,

$$P(\{\omega\}\,|\,\mathbf{D},\mathbf{I}) = \int P(\boldsymbol{\theta},\sigma\,|\,\mathbf{D},\mathbf{I})d\{\mathbf{B}\}d\sigma \ , \qquad (9)$$

which implies what the data and prior information have to tell us about the frequencies $\{\omega\}$ regardless of the values of the amplitudes $\{\mathbf{B}\}$ and the deviation of the noise σ. To do this, it is necessary to make the matrix $\Omega_{jk}(\omega)$ to be diagonal, effectively by introducing new orthogonal model functions:

$$H_j(t) = \frac{1}{\sqrt{\lambda_j}}\sum_{k=1}^{m} a_{jk}G_k(\{\omega\},t),\ (j=1,...,m), \quad (10)$$

where

$$\sum_{i=1}^{N} H_j(t_i)H_k(t_i) = \delta_{jk} = \begin{cases} 1 & j=k \\ 0 & j\neq k \end{cases} \ . \qquad (11)$$

This diagonalization process gives a new expression for the model function,

$$f(t) = \sum_{k=1}^{m} A_k H_k(t), \qquad (12)$$

where the new amplitude A_k is related to the old amplitude B_j by

$$A_k = \sqrt{\lambda_k}\sum_{j=1}^{m} B_j a_{kj}, (k=1,...,m) \qquad (13)$$

and where a_{kj} represents the j th component of the k th normalized eigenvector of Ω_{jk}, with λ_j as the corresponding eigenvalue. Substituting this into Equation (12) and defining

$$h_j = \sum_{i=1}^{N} d_i H_j(t_i,\{\omega\}), \qquad (14)$$

to be the projection of the data onto model functions $H_j(t,\{\omega\})$, we can then proceed to perform the m integration over A_j's analytically. If σ^2 is known, then the problem is completed, provided we have no additional prior information. The joint posterior probability density of the $\{\omega\}$ parameters, conditional on the data and our knowledge of σ is then given by

$$P(\{\omega\}\,|\,\mathbf{D},\sigma^2,\mathbf{I}) \propto \exp\left(\frac{m\overline{\mathbf{h}^2}}{2\sigma^2}\right). \qquad (15)$$

But if there is no prior information available about the noise, then σ^2 can be taken as a nuisance parameter and eliminated as it was done with the amplitudes. Often there will be little prior knowledge concerning the noise variance and it is then appropriate to use the Jeffreys' prior:

$$P(\sigma) = \frac{1}{\sigma} \qquad (16)$$

By using this prior and integrating over σ parameter the posterior probability density of the frequencies in Equation (15) turns into the following form:

$$P(\{\omega\}\,|\mathbf{D},I) \propto \left(1-\frac{m\,\overline{\mathbf{h}^2}}{N\overline{\mathbf{d}^2}}\right)^{\frac{m-N}{2}}. \qquad (17)$$

This has the form of the "*Student t- distribution*" with $(N-m)$ degree of freedom. As well as determining the $\{\omega\}$ parameters for which the posterior probability is a maximum, it is also desirable to compute the variances associated with them. This is because one always requires knowing how good the estimate is. Therefore, let $\{\hat{\omega}\}$ represent the estimated values of $\{\omega\}$. Then we can expand the function $\overline{\mathbf{h}^2}$ in a Taylor series at the point $\{\hat{\omega}\}$, such that

$$P(\{\omega\}\,|\,\mathbf{D},\sigma^2,I)\,\alpha\exp\left[-\sum_{j,k=1}^{\rho}\frac{b_{jk}(\hat{\omega}_j-\omega_j)(\hat{\omega}_k-\omega_k)}{2\langle\sigma^2\rangle}\right] \qquad (18)$$

with b_{jk} defined as

$$b_{jk} = -\frac{m}{2}\frac{\partial^2\overline{\mathbf{h}^2}}{\partial\omega_j\partial\omega_k}\bigg|_{\substack{\omega_j=\hat{\omega}_j \\ \omega_k=\hat{\omega}_k}} \ . \qquad (19)$$

For an arbitrary model the matrix b_{jk} cannot be calculated analytically; however, it can be evaluated numerically. Using Equation (18) we can make the (*mean*) $\pm$ (*standard deviation*) approximations for the $\{\omega\}$ parameters. This can be done by using orthogonal transformation to convert the multiple integrals into the Gaussian types. It can then be performed multiple integrations analytically with respect to the frequencies

just as it was done with the amplitudes. Let ν_j and u_{kj} represent the jth eigenvalue and eigenvectors of the matrix b_{jk}, respectively. Then the new orthogonal variables are given by

$$s_j = \sqrt{\nu_j}\sum_{k=1}^{\rho}\Delta_k u_{kj}, \quad \Delta_j = \sum_{k=1}^{\rho}\frac{s_k u_{jk}}{\sqrt{\nu_k}} \ . \tag{20}$$

where $\Delta_k = \hat{\omega}_k - \omega_k$. Using these orthogonal variables, the estimate variance for ω_k is simply defined in the following form:

$$\sigma_{\omega_k}{}^2 = \left\langle \sigma^2 \right\rangle \sum_{j=1}^{\rho}\frac{u_{jk}^2}{\nu_j} \ . \tag{21}$$

Thus the expected values of the $\boldsymbol{\omega}$ parameters can easily be implemented by

$$\left\langle \boldsymbol{\omega} \right\rangle = \hat{\boldsymbol{\omega}} \pm \boldsymbol{\sigma}_{\boldsymbol{\omega}} . \tag{22}$$

One can also show that the expected values of the parameters $\{A\}$ are given by $\langle A_j \rangle = h_j$, which when transformed to the parameters $\{B\}$ becomes

$$\langle B_k \rangle = \sum_{j=1}^{m}\frac{1}{\sqrt{\lambda_j}} h_j a_{jk} \ . \tag{23}$$

The uncertainty in the A_j is $\pm\sigma$, so that the corresponding variance in the $\{B_k\}$ is

$$\sigma^2{}_{B_k} = \sigma^2 \sum_{j=1}^{m}\frac{1}{\sqrt{\lambda_j}} a_{jk} a_{jl} \ . \tag{24}$$

3 Global Optimization Algorithm

The approach summarized above, introduced by Bretthorst, and requires analytic or numerical approximation of integrals which is not given here. As a result, the Bayesian parameter estimation problem turns into the maximization of the posterior probability density given in Equations (15) or (17). Unfortunately, conventional algorithms[13] based on the gradient direction fail to converge for this problem. Even when they converge, there is no assurance that they have found the global, rather than a local maximum. This is because the log of the probability density function is so sharply peaked and highly nonlinear function of the frequencies. Bretthorst used a pattern search algorithm described by Hook-Jevees to overcome this problem. However, its efficiency decisively depends on the choice of the initial step size. When the step sizes are chosen too small the algorithm requires much iteration to appropriately increase the length of search direction vector. When the step sizes are chosen too large they have to be reduced before the algorithm can advance towards the optimum. Finally, the choice of the starting point and the initial step sizes strictly determine which optimum will be found. It may be a local or a global optimum. To deal with this kind of problem, we consider here simulated annealing algorithm, which has been applied to continuous global optimization problems. Details of this algorithm can be found in the references [7, 10, 15, 16].

The essential starting parameters to maximize the posterior probability density function $P(\{\omega\} \mid \mathbf{D}, \sigma, I)$ are Γ which by analogy with original application is known as the system temperature irrespective of the objective function involved and $\mathbf{W} = \{\omega_1, \omega_2, ..., \omega_\rho\}$, the starting vector of the frequencies. A function evaluation is made at the starting point $\mathbf{W}$ and the value of $P(\mathbf{W} \mid \mathbf{D}, \sigma, I)$ is recorded. Next, $\mathbf{W}^*$ is obtained from $\mathbf{W}$ by changing only jth element of $\mathbf{W}$,

$$\omega^*{}_j = \omega_j + \delta\omega_j \ , \tag{25}$$

where

$$\delta\omega_j = \beta\sqrt{\Upsilon(\omega_j)}\,\mathrm{N}(0,1) \tag{26}$$

is recalculated until $0 \le \omega^*{}_j \le 2\pi$ for all over j's. The step size $\delta\omega_j$ is proportional with the squared root of the Cramer-Rao lower bound $\Upsilon(\omega_j)$ [11], defined by

$$\Upsilon(\omega_j) = \sqrt{\frac{48\sigma^2}{N^3(\hat{B}_{1j}{}^2 + \hat{B}_{2j}{}^2)}} \ (j = 1, 2, ..., \rho) \tag{27}$$

for each frequency ω_j and β is an appropriate constant for scaling. $\Upsilon(\omega_j)$ is a lower limit to the variance for a measurement of the frequency ω_j, and so its squared root generates a natural scale size of the search space around the estimated value of ω_j. It is expected that better solutions lie close to solutions that are already good and so normally distributed step sizes can then be used[14].

If $P(\mathbf{W}^* \mid \mathbf{D}, \sigma, I)$ is larger than $P(\mathbf{W} \mid \mathbf{D}, \sigma, I)$, $\mathbf{W}^*$ and its probability density value are recorded since this is the best current values of the optimum frequency

vector denoted as $\mathbf{W}_{opt}$. Otherwise, the Metropolis criterion [21] decides on acceptance. For this purpose, the value

$$p = \exp[-\frac{\delta P(\mathbf{W}|\mathbf{D}, I)}{\Gamma}] \quad (28)$$

is computed and compared with a random number chosen from uniform distribution in the interval [0, 1]. If p is greater than this number, the new point is accepted and $\mathbf{W}$ is updated with $\mathbf{W}^*$ and the algorithm moves downhill. Otherwise, $\mathbf{W}^*$ is rejected. From an optimization point of view, an iterative search accepting only new point with highest function value is like rapidly quenching a physical system at zero temperature: It is very likely to get stuck in a metastable, local maximum. On the contrary, the simulated annealing permits downhill moves under the control of the temperature parameter Γ. At higher temperature only the gross behavior of the cost function is relevant to the search. As temperature decreases, finer details can be developed to get a good final point. While the optimality of the final point cannot be guaranteed, the method is able to proceed toward better maxima even in the presence of many local maxima. From Equation (28) there are two factors, lower temperatures and larger differences in the function's value, which decrease the probability of a downhill move. After n_Γ iterations the Γ temperature is reduced by the annealing schedule defined as

$$\Gamma_k = \Gamma_0 \exp[-r_\Gamma k^{\frac{1}{m}}], \quad (29)$$

for some $0 \le r_\Gamma < 1$ and the initial temperature Γ_0. A lower temperature makes a given downhill move less likely, so the number of Γ rejections increase and the step lengths decline. In addition, the first point tried at the new temperature is the current optimum. The smaller steps starting at the current optimum focuses attention on the most promising area. The algorithm ends by comparing the last n_ε values of the largest function values from the end of each temperature reduction with the most recent one. If all these differences at the k th iteration

$$\begin{aligned} &\left\| P(\mathbf{W}^*_k|\mathbf{D},\mathbf{I}) - P(\mathbf{W}^*_{k-l}|\mathbf{D},\mathbf{I}) \right\| \le \varepsilon, (l = 1,2,..,n_\varepsilon) \\ &\left\| P(\mathbf{W}^*_k|\mathbf{D},\mathbf{I}) - P(\mathbf{W}^*_{opt}|\mathbf{D},\mathbf{I}) \right\| \le \varepsilon \end{aligned} \quad (30)$$

are less than a pre-assigned number $\varepsilon(>0)$ then the algorithm terminates. This criterion helps to ensure that the global maximum is reached.

4 Power Spectral Density

Before we discuss the computer simulated examples, there is something we need to say about how to display the results. The usual way the result from a spectral analysis is displayed is in the form of a power spectral density, which shows the strength of the variation (energy) as a function of frequency. In Fourier transform spectroscopy, this is typically taken as the squared magnitude of the discrete Fourier transform of the data. In order to display our results in the form of a power spectral density, it is necessary to give an attention to its definition that shows how much power is contained in a unit frequency. According to Bretthorst [3], the Bayesian power spectral density[27], which is defined as the expected value of the power of the signals over the joint posterior probability density, is implemented in the following form:

$$p(\omega) = \frac{N}{2} \int (\sum_{j=1}^{\rho} \mathbf{B}^2{}_j) p(\{\omega\}, \{\mathbf{B}\} | D, \sigma^2, I) d\mathbf{B}_1 ... d\mathbf{B}_\rho \quad (31)$$

Performing integrals over $\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_\rho$ by using Equations (18) and (19), the power spectral density can be approximated as

$$p(\omega) = 2(\sigma^2 + \sum_{j=1}^{\rho} h^2(\hat{\omega}_j)) \sum_{k=1}^{\rho} (\frac{b_{kk}}{2\pi\sigma^2})^{\frac{1}{2}} \exp(-\frac{b_{kk}(\hat{\omega}_k - \omega)^2}{2\sigma^2}). \quad (32)$$

This function stresses information about the total energy carried by the signal and about the accuracy of each line. In the next section, we will present some numerical examples how well this technique works.

5 Computer Simulated Examples

To verify the performance of the algorithm, we generated data vector from one, two and five sinusoids and then added the noise vector generated from a zero mean Gaussian distribution with a deviation of σ to this vector. Here the time t runs over the symmetric time interval $-T$ to T in $(2T+1) = 512$ integer steps. To ensure that the data have zero mean value the average value of the data is computed and subtracted from each data point. After obtaining simulated data, we carried out its Bayesian analysis, assuming that we know the mathematical form of the model but, not the values of the parameters. We first gave starting values to the list of the frequencies to begin a multidimensional search for finding a global maximum of the posterior probability density given in Equations (15) or (17). As an initial estimate of the frequencies for the maximization procedure, we chose the locations of the peaks in the power spectrum density, which is a

squared magnitude of the fast Fourier transform (FFT) of data. Once, the frequencies are estimated, we carry on calculating the amplitudes and parameter errors approximately by using Equations (21), (23) and (24).

In the first example, the noisy data set is generated from the following equation:

$$d_i = 0.001 + 0.5403\cos(0.3t_i) - 0.8415\sin(0.3t_i) + e_i \quad (i = 1,...,512) \tag{33}$$

where $e_i \square N(0,1)$. For this data, the computer programme we developed was run on the workstation in two cases. In the first case, we assumed that the standard deviation of noise is actually known and used the probability density given in Equation (15) to maximize. The estimated values associated with the estimated deviations are tabulated in Table 1. In the second case, we also suppose that the standard deviation of noise σ is unknown and repeated calculations using the probability density given in Equation (17). The results are similar to those shown in Table 1. It can be seen that a single frequency and amplitudes are recovered very well. As we can see, when the noise level is unknown Bayesian inference can give very accurate results together with an estimate of the noise level and the signal-to-noise ratio (sTn)[3,4,5] shown in Table 1. The usual way of displaying the results in the spectral analysis is in the form of a power spectral density shown in Fig. 2. The Fourier Spectral density contains spurious side lobes, but these do not appear in the plot of the Bayesian spectral density.

In the second example, we consider a model signal with two close harmonic frequencies:

$$d_i = 0.5403\cos(0.3t_i) - 0.8415\sin(0.3t_i) - 0.4161\cos(0.31t_i) - 0.9093\sin(0.31t_i) + e_i \tag{34}$$

In a similar way, we produced the same size data corrupted by the zero mean Gaussian noise with $\sigma = 1$. Using this data shown in Figure 3, we run our *Mathematica* code again in the case where the deviation of noise is unknown. The time interval is the same as the first example and Table 2 contains the estimated values associated with the deviations for the parameters. Here again the computed values are all very close to the true values and the two frequency lines are clearly resolved. This is also clearly evident from the Bayesian power spectral density shown in Fig. 4. It can be seen that the Fourier power spectral density shows only a single peak in between the two frequencies. Actually with the Fourier spectral density when the separation of two frequencies is less than the Nyquist step, $|\omega_1 - \omega_2| < 2\pi / N$, the two frequencies are indistinguishable. This is simply because there are no sample points in between the two frequencies in the frequency domain. If $|\omega_1 - \omega_2| > 2\pi / N$ theoretically the two frequencies can then be distinguished. If $|\omega_1 - \omega_2|$ is not large enough, the resolution will be very poor. Therefore, it is hard to tell where the two frequencies are located. This is just the inherent problem of the discrete Fourier power spectral density. In this example the two frequencies are separated by 0.01, which is less than the Nyquist step size. There is no way by using Fourier power spectral density that one can resolve the two frequencies, however Bayesian power spectral density gives us very good results with high accuracy. Furthermore, Figs. 1-3 show the power of the method for recovering the signal from the noisy data using the estimated values of the parameters of sinusoids.

In general, let us consider a multiple harmonic frequency model signal:

$$d_i = \cos(0.1\,t_i + 1) + 2\cos(0.15t_i + 2) + 5\cos(0.3t_i + 3) + 2\cos(0.31t_i + 4) + 3\cos(t_i + 5) + e_i \tag{35}$$

The time series with a sample size of 512, including the zero mean Gaussian noise with $\sigma = 1$, is shown in Figure 5. The best estimates and the marginal deviations for all the parameters are tabulated in Table 3. Once again, all the frequencies have been well resolved, even the third and fourth frequencies which are too closed are not to be separated by the Fourier power spectral density shown in Fig. 5. This is why Fourier Power spectrum shows only four peaks. The Bayesian power spectral density plot for this example shows five sharp peaks. These results we obtained are also similar to that of Bretthorst [3] and they also demonstrate the advantages of Bayesian inference using together with simulated annealing algorithm.

Moreover, we initially assumed that the values of the random noise in data were drawn from the zero mean Gaussian normal density. Fig. 6 shows the exact and estimate probability densities of the random noise in data. It is seen that the estimated (dotted) probability density is closer to the true (solid) probability density and the histogram of the data is much closer to the true probability density. The histogram is also known as a nonparametric estimator of the probability density because it does not depend on specified parameters.

6 Conclusions

In this work we have partially developed and applied Bayesian inference with simulated annealing to the spectral analysis and parameter estimation problem. The results presented here show that Bayesian statistical inference used together with a simulated annealing provides rational approach for estimating, in an optimal way, values of parameters of sinusoids corrupted by random noise. Both frequency and amplitudes can be recovered from the experimental data and the prior information with high accuracy, especially the frequency, which is the most important parameter in spectral analysis. A significant advantage of the Bayesian approach comes from the very large posterior probabilities, which are sharply peaked in the neighborhood of the best fit. This helps to simplify the problem of choosing starting values and performing the random search for the best parameters. Although simulated annealing spend large consumption of CPU time, it is competitive when compared to the multiple runs often used with conventional algorithms to test different starting values.

We have not addressed the problem of model selection, which is a big part of spectral analysis. In analyzing experimental data, it is not always known which model functions apply. In general one has enough prior information in a given experiment to select the best model among a finite set of model functions. Bayesian inference helps us to accomplish this. This work and a comparison with the most recent methods will deserve further investigations. In addition, the computer simulations show that Mathematica is powerful system for doing mathematics by computer and it has steadily grown in breadth and depth to become today an unparalleled platform for all forms of computations.

7 Acknowledgments

This work is a part of two projects, whose names are "Bayesian Model Selection and Parameter Estimation" with a number **FEN - DKR - 200407 - 0082** and "Bayesian Spectrum Estimation of Harmonic Signals" with a number **FEN-BGS-060907-0191**, supported by the University of Marmara, Istanbul, Turkey.

References:

[1] J.M. Bernardo and A.F.M. Smith, Bayesian *Theory*, Willey Series in Probability and Statistics, New York, 2000.

[2] G.E.P. Box and G.C. Tiao, Bayesian Inference in Statistical Analysis, Wiley, New York, 1992.

[3] G.L. Bretthorst, An introduction to Parameter Estimation Using Bayesian Probability Theory, Maximum *Entropy and Bayesian Methods edited* Fougère P.F., Kluwer Academic Publishers, The Netherlands, 1990, pp. 53-79.

[4] G.L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics, Springer-Verlag Berlin Heidelberg New York, 1997.

[5] G.L. Bretthorst, W.C. Hutton, J.R. Garbow and J. J.H. Ackerman, Exponential Model Selection (in NMR) using Bayesian probability theory, *Magnetic Resonance, Vol. 2*,2005, pp. 55-63.

[6] J.W. Cooley and J.W. Tukey, an Algorithm for the Machine Calculation of Complex Fourier Series, *Mathematics of Computation, Vol. 19*, 1965, pp. 297-301.

[7] A. Corana, M. Marchesi, C. Martini and S. Ridella, Minimizing Multimodal Functions of Continuous Variables with the Simulated Annealing Algorithm, *ACM Transactions on Mathematical Software*, Vol. 13, 1987, pp. 262-280.

[8] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*; John Wiley & Sons, 1989.

[9] P. Gregory, *Bayesian Logical Data Analysis for the Physical Science*, Cambridge University Press, 2005.

[10] W.L Goffe, G.D. Ferier and J. Rogers, Global Optimization of Statistical Functions with Simulated Annealing, *Journal of Econometrics, Vol. 60*, 1994, pp 65-100.

[11] P. Handel , Cramer-Rao Bounds and Non-Linear Least Squares for a Seven Parameter Dual Channel Sinewave Model, *IMCT2007-IEEE Instrumentation and Measurement Technology Conference*, Warsaw, Poland, 2007, pp. 1-3.

[12] H.L. Harney, *Bayesian Inference: Parameter Estimation and Decisions*, Springer-Verlag Berlin Heidelberg, 2003.

[13] M.H. Hayes, *Statistical Digital Signals Processing and Modeling*, John Wiley & Sons. Inc., New York, 1996.

[14] T.R. Hooke and T.A. Jevees, Direct Search Solution of Numerical and Statistical Problems, *Journal of Association of Computer Machinery*, 1962, pp. 212-229.

[15] J. Ireland, Simulated Annealing and Bayesian Posterior Distribution Analysis Applied to Spectral Emission Line Fitting, *Journal of Solar Physics*, Springer Netherlands, 2005, pp. 237-252.

[16] L. Ingbe, Simulated Annealing: Practice Versus Theory, *Mathematical Computation and Modelling*, Vol. 18, 1993, pp.29–57.

[17] E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, D. Reidel, Dordrecht-Holland, 1983.

[18] E.T. Jaynes, Bayesian Spectrum and Chirp Analysis, *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems edited by C.R.* Smith and G.J. Erickson, D. Reidel, Dordrecht-Holland, 1987, pp. 1-37.

[19] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1989.

[20] E.T. Jaynes, *Probability theory: the logic of science*, edited by G.L. Bretthorst, Cambridge University Press, 2003.

[21] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, Equation of States Calculations by Fast Computing Machines, *Journal of Chemical Physics*, Vol. 21, 1953, pp. 1087-1092.

[22] W.H. Press, B.P. Flannery, S.A. Teukolshy and W.T. Vetterling, *Numerical Recipes in C: The Art of Computing*, Second Edition, Cambridge University Press, 1995.

[23] A. Schuster, The Periodogram and its Optical Analogy, *Proceedings of the Royal Society of London, Vol. 77*, 1905, pp. 136.

[24] D. Üstündağ, N.M. Queen, G.K. Skinner, and J.E. Bowcock, Two New Methods and Comparison with the Cambridge MaxEnt Package, Maximum Entropy and Bayesian Methods edited by Grandy W. T. and Schick L.H., *Kluwer Academic Publishers*, Netherlands, 1991, pp. 295-301.

[25] D. Üstündağ, N.M. Queen and J.E. Bowcock, A Method for Retrieving Images From Noisy, Incomplete Data, Proceeding of The Fifth Alvey Vision Conference, University of Reading, 1989, pp. 239–245.

[26] P. Weillin, R. Gayllord and S. Kamin, *an Introduction to Programming with Mathematica*, Cambridge University Press, 2005.

[27] P.M. Djurić, H.T. Li, Bayesian Spectrum Estimation of Harmonic Signals, *IEEE Signal Processing Letters*, Vol. 2, 1995, pp. 213-215.

[28] A. Ouldali, S. Sadoudi and Z. Messaoudi, Extended Kalman Filter for Parameter Estimation of Wideband polynomial Phase Signals in Sensors Arrays, Proceedings of the 7th Wseas International Conference on Signal Processing (SIP'08), Istanbul, Turkey, 2008, pp. 19-24.

[29] M. Malyutov, R.S. Protassov, D. Golan and R. Mirchev, Bayesian Approach to Parameter Estimation in Individual Protein Molecule Dynamics Model, Wseas Transactions on Systems, Vol. 1, 2002, pp. 204-209.

[30] J.A. Ruiz-Cruz, J.R. Montejo-Garai and J.M. Rebollar, Application of the Simulated Annealing for Waveguide Filters with Complicated Frequency Response, *Wseas Transactions on Mathematics*, Vol. 3, 200, pp. 807-812.

Table 1. Computer simulations for a single harmonic frequency model.

	Bayesian Parameter estimation	
σ : Known N : 512 m : 3 *sTn* : 0.725 σ : 1.	Parameters	Estimated Values
	w_1	0.299852 ± 0.0005
	B_1	0.603579 ± 0.05
	B_2	-0.817479± 0.06

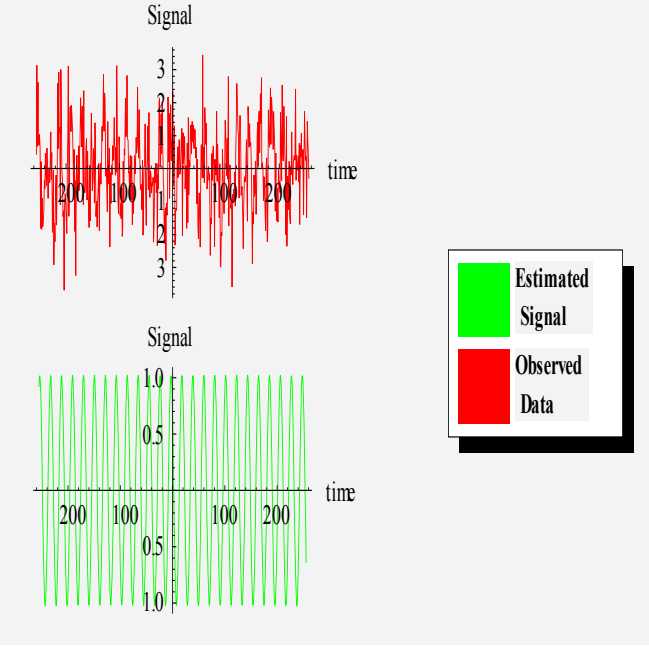

Fig. 1. Recovering signal from noisy data produced from single harmonic frequency model

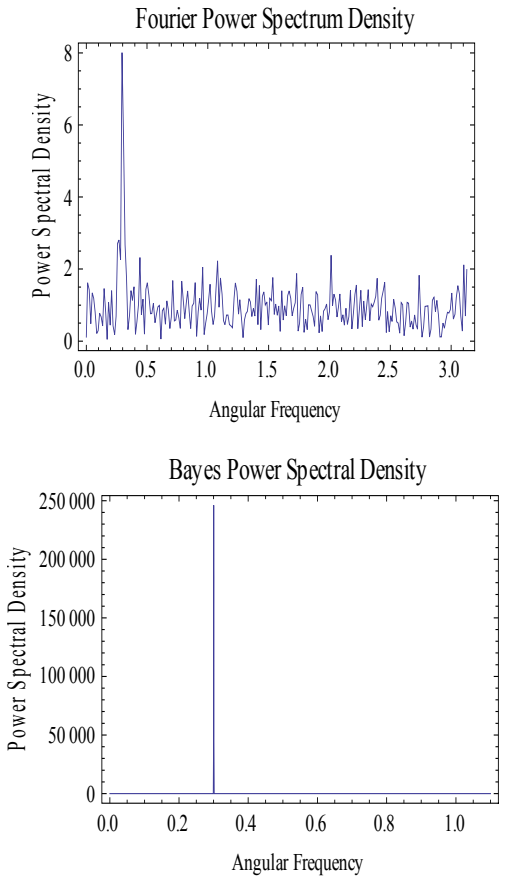

Fig. 2. Spectral analysis of one frequency model.

Table 2. Computer simulations for two closed frequency model

	Bayesian Parameter estimation	
	Parameters	Estimated Values
σ: Unknown N: 512 m : 6 *sTn* 1.029 σ: 0.9963	w_1	0.300082± 0.0009
	w_2	0.310574± 0.0001
	B_1	0.520649± 0.063
	B_2	-0.869805± 0.063
	B_3	-0.381937± 0.063
	B_4	-0.855687± 0.063

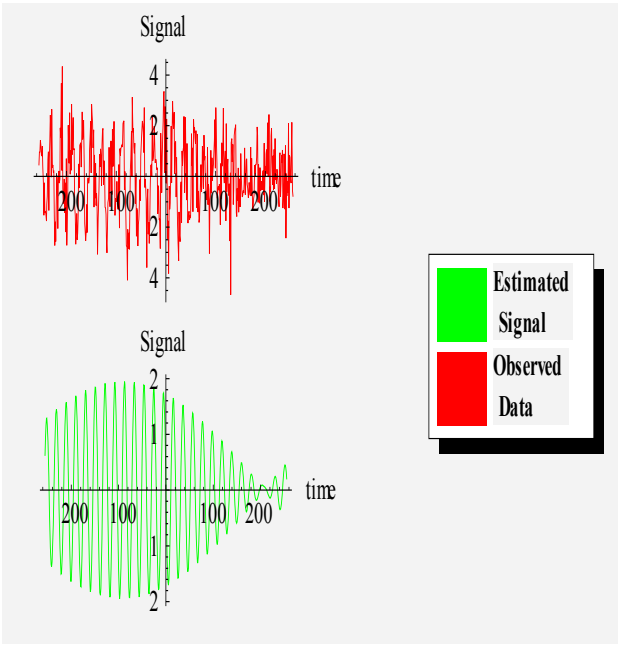

Fig. 3. Recovering signals from noisy data produced from two closed harmonic frequency model.

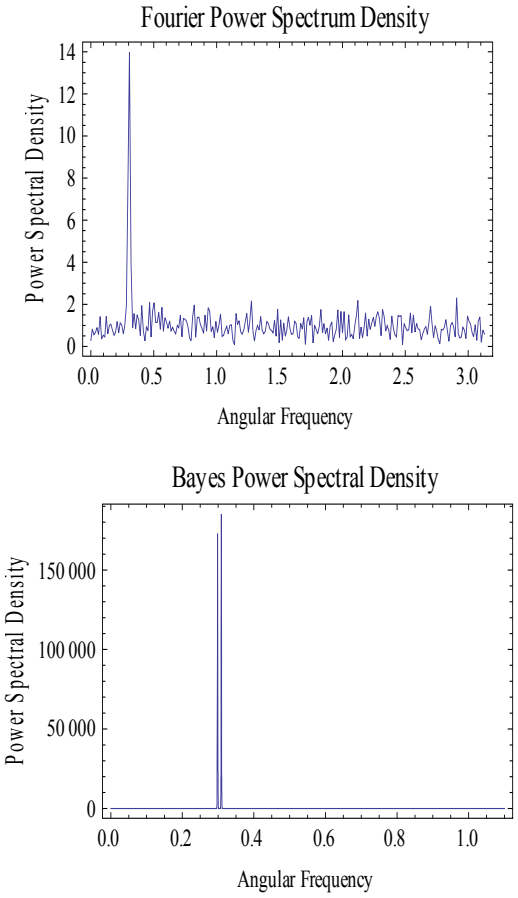

Fig. 4. Spectral analysis of two frequencies model.

Table 3. Computer simulations for a multiple harmonic frequency model[1].

	Parameters	Estimated Values
σ : Unknown N : 512 m : 15 sTn 1.029 $\hat{\sigma}$: 0.9963	w_1	0.0994±0.0004
	w_2	0.1501±0.0002
	w_3	0.2998±0.0001
	w_4	0.3100±0.0003
	w_5	1.0000±0.0001
	A_1	0.94±0.08
	A_2	2.07±0.08
	A_3	4.97±0.09
	A_4	1.98±0.09
	A_5	2.98±0.08

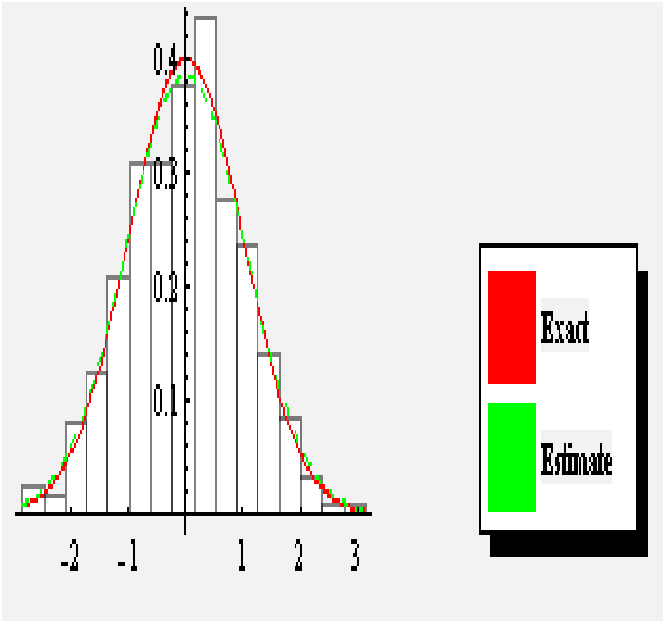

Fig. 6. Comparison of exact and estimate probability densities of noise in data.

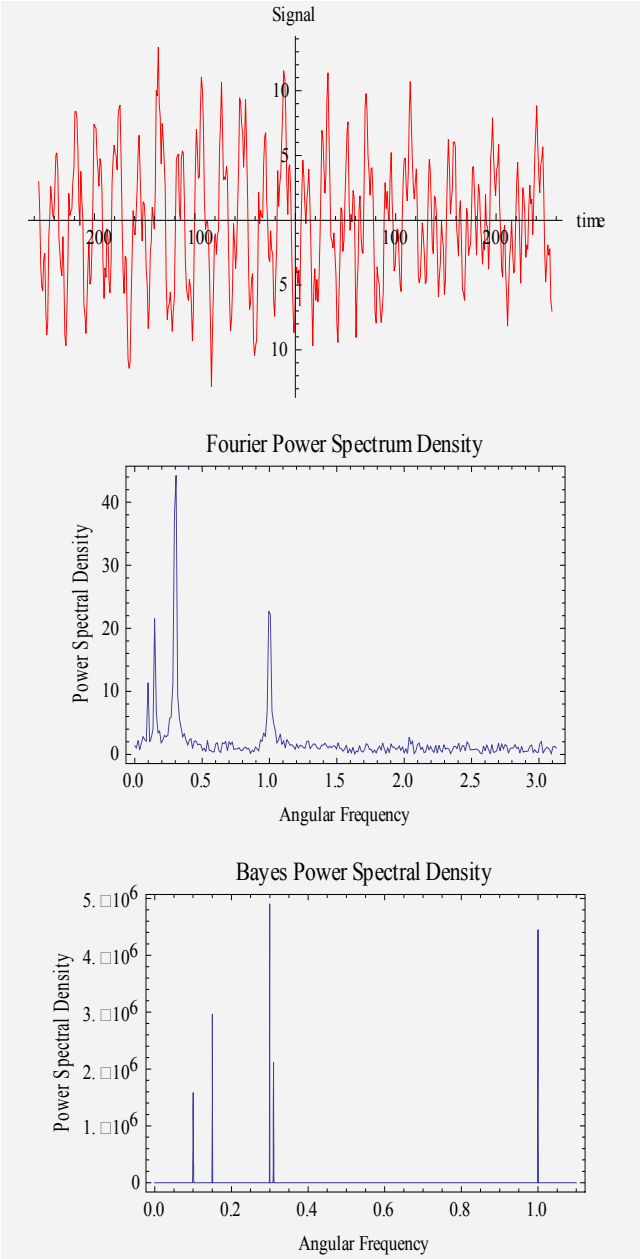

Fig. 5. Spectral analysis of multiple frequency model.

[1] In order to compare the results with that of Bretthorst in this example we converted

$$A_i = \sqrt{B_i^2 + B^2_{i+\rho}}, (i = 1, \ldots, \rho).$$