Bayesian Spectral Analysis with Student-t Noise

Jacqueline Christmas, *Member, IEEE*

***Abstract*—We introduce a Bayesian spectral analysis model for one-dimensional signals where the observation noise is assumed to be Student-t distributed, for robustness to outliers, and we estimate the posterior distributions of the Student-t hyperparameters, as well as the amplitudes and phases of the component sinusoids. The integrals required for exact Bayesian inference are intractable, so we use variational approximation. We show that the approximate phase posteriors are Generalised von Mises distributions of order 2 and that their spread increases as the signal to noise ratio decreases. The model is demonstrated against synthetic data, and real GPS and Wolf's sunspot data.**

***Index Terms*—Bayesian methods, Fourier series, discrete Fourier transforms, parameter estimation, amplitude estimation, phase estimation.**

I. Introduction

We introduce a Bayesian spectral analysis model which assumes that the observation noise is distributed according to the heavy-tailed Student-t distribution, to provide robustness in the presence of outliers (that is, a small fraction of extreme observations that are unrepresentative of the rest of the sample), and learns posterior distributions for the noise variables, as well as the amplitudes and phases of the component sinusoids. Bayesian modelling allows us to incorporate any prior knowledge we may have regarding the likely distributions of the model variables and avoids the overfitting known to arise in maximum likelihood estimation by evaluating over all possible values of those variables.

Analysis of periodic data often depends upon spectral analysis, whereby the set of observations is the decomposed into a sum of sinusoidal components with varying frequencies, amplitudes and phases. This transformation from temporal to frequency space may identify key characteristics of the underlying periodicity.

Schuster [1] introduces the periodogram as a means of identifying periodicity in data. The Fourier Transform is a deterministic method for decomposing a continuous function into component sinusoids. Where the function is described only by a number of discrete observations, this decomposition is referred to as the Discrete Fourier Transform (DFT). The periodogram is determined by the DFT if the time intervals between observations are equal. The DFT became a popular tool with the development of the computationally efficient Fast discrete Fourier Transform (FFT) [2] technique.

Jaynes [3,4] applies Bayesian inference to the problem, demonstrating formal support for the periodogram. He assumes that the observations are samples of a continuous function with added Gaussian–distributed noise (of known variance) and that the underlying signal is a single sinusoid of unknown amplitude, phase and frequency. Bretthorst [5] extends this analysis to calculate the probability that multiple frequencies are present in the signal.

Dou and Hodgson [6] introduce a Bayesian model for estimating the period, phase and amplitude of multiple sinusoids, using Markov chain Monte Carlo (MCMC) to avoid the intractable integrals that commonly arise in Bayesian inference. Andrieu and Doucet [7] include model order (i.e. the number of component sinusoids in the signal) as a variable in their Bayesian model and use Reversible Jump MCMC [8] to estimate its posterior distribution. This model is extended to dynamic systems by Nielsen *et al* [9,10].

Bayesian models that assume the noise to be Gaussian are known to be badly affected by the presence of outliers. There are many different methods of dealing with outliers (e.g. [11, 12]); there is a whole field devoted to identifying and removing them from the data before further processing. In the context of signal processing, Ruggeri [13] and Zoubir *et al* [14] review robustness in Bayesian and non-Bayesian models respectively.

Rather than removing outliers, we accomodate them within the probabilistic model. Chave *et al* [15] combine maximum likelihood estimators with section averaging spectral analysis techniques to obtain robust models. Ahdesmäki *et al* [16] (extending [17]) use a g-statistic combined with multiple testing to produce a spectral anslysis model that is robust both to outliers and other commonly occurring problems with the observations. Roberts and Penny [18] represent the noise with a finite mixture of Gaussians, enabling the outliers to be absorbed into one or more Gaussians with high variance. This leads to improved performance over a Gaussian noise model, but the tails still decay exponentially and the variance is always finite. Lange *et al* [19] show that replacing Gaussian assumptions with the Student-t provides more robust statistical inference for a variety of real datasets (see, e.g., [20]). For a univariate variable, x, the Student-t distribution is defined as:

$$\mathrm{p}(x) = \mathcal{S}(x \,|\, \mu, \lambda, d) \tag{1}$$

$$= \frac{\Gamma\left(\frac{(d+1)}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \left(\frac{\lambda}{\pi d}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{d}\right)^{-\frac{d+1}{2}} \tag{2}$$

where μ is the mean, λ is known as the precision and d is the degrees of freedom. As $d \to \infty$ the distribution tends to a Gaussian, at $d = 2$ it is the Cauchy and when $d < 2$ the variance becomes effectively infinite. An alternative definition reveals the Student-t as an infinite mixture of Gaussians ($\mathcal{N}(\cdot)$) with a shared mean and precisions sampled from a Gamma

The author is with the College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK (email: J.T.Christmas@exeter.ac.uk).

distribution[1] based on the degrees of freedom:

$$\mathcal{S}(x \,|\, \mu, \lambda, d) = \int_0^\infty \mathcal{N}(x \,|\, \mu, (\lambda z)^{-1})\, \mathcal{G}(z \,|\, \frac{d}{2}, \frac{d}{2})\, \mathrm{d}z \quad (3)$$

As an alternative to introducing a variable to represent the model order, we use the technique of *automatic relevance determination* [21,22] to automatically "switch off" component sinusoids for which there is no evidence in the data by constraining their amplitudes to be close to zero.

We treat the amplitudes and phases of the component sinusoids, and the precision and degrees of freedom of the Student-t noise distribution as variables whose posterior distributions we wish to estimate. The frequencies of the components are treated as parameters which we specify, as are the timestamps associated with each of the observations. The integrals required to perform exact Bayesian inference are intractable, so we use variational approximation to estimate the posterior distributions. The posteriors for the phases turn out to be potentially bimodal and/or asymmetric Generalised von Mises distributions of order 2 [23,24].

In section II we introduce the model and the variational Bayesian method for approximating the inference of the posterior distributions. In section II-C we take a closer look at the Generalised Von Mises distribution. We demonstrate the model both on synthetic data (III) and on real GPS and sunspots data (IV). Conclusions are drawn in section V.

II. The model

We express each observation (of N, one-dimensional data), y_n, as the sum of its C component sinusoids, plus a noise term ϵ_n:

$$y_n = \mathbf{a}^{\mathrm{T}} \cos\left(\boldsymbol{\phi} - \boldsymbol{\omega} t_n\right) + \epsilon_n \quad (4)$$

where $\boldsymbol{\omega}$, $\mathbf{a}$ and $\boldsymbol{\phi}$ are the vectors of angular frequencies, amplitudes and phases of the components respectively, and t_n is the timestamp associated with the nth observation. We assume that the observation noise is Student-t distributed with zero mean, precision λ and degrees of freedom d, which leads to the following likelihood:

$$\mathrm{p}(y_n \,|\, \mathbf{a}, \boldsymbol{\phi}, \lambda, d; \boldsymbol{\omega}, t_n) = \mathcal{S}\left(y_n \,|\, \mathbf{a}^{\mathrm{T}} \cos(\boldsymbol{\phi} - \boldsymbol{\omega} t_n), \lambda, d\right) \quad (5)$$

which we may re-express using (3), introducing a new latent variable, $\mathbf{z}$.

Thus the model consists of two parameters, the C-dimensional $\boldsymbol{\omega}$ and the N-dimensional $\mathbf{t}$, and the variables $\mathbf{a}$ and $\boldsymbol{\phi}$ (both C-dimensional), $\mathbf{z}$ (N-dimensional), λ and d. For the variables we aim to calculate their posterior distributions, which provide us with both estimates of their values and a measure of the uncertainty in those estimates.

A. Priors

In this simple model we consider the component sinusoids to be independent of one another. For each amplitude we assign a Gaussian prior, with precision δ_c:

$$\mathrm{p}(a_c \,|\, \delta_c) = \mathcal{N}(a_c \,|\, 0, \delta_c^{-1}) \quad (6)$$

[1] defined as $\mathcal{G}(x \,|\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$

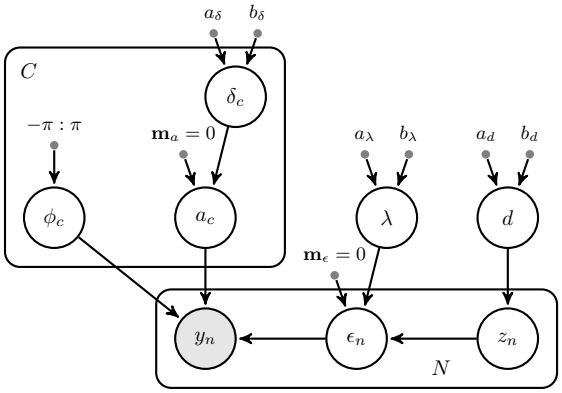


Figure 1 A graphical representation of the model priors.

and we assign δ_c a conjugate Gamma distribution:

$$\mathrm{p}(\delta_c) = \mathcal{G}(\delta_c \,|\, a_\delta, b_\delta) \quad (7)$$

This so-called *automatic relevance determination* prior [21,22] causes a_c to be constrained to be close to zero where there is no evidence in the data for a contribution at that frequency.

With no prior information about the component phases ϕ_c, we model each of them with a Uniform distribution over the full range of values, i.e. any 2π range; we choose $-\pi$ to π:

$$\mathrm{p}(\phi_c) = \mathcal{U}(\phi_c \,|\, -\pi, \pi) \quad (8)$$

Were prior information available we would instead specify a von Mises distribution for each phase. The Uniform distribution is a special case of the von Mises and we use it here to maintain the clarity of the implementation of the model by minimising the number of parameters.

The observation noise precision λ is assigned a conjugate Gamma prior:

$$\mathrm{p}(\lambda) = \mathcal{G}(\lambda \,|\, a_\lambda, b_\lambda) \quad (9)$$

Using the definition for the Student-t distribution from (3), we introduce a latent variable, $\mathbf{z}$, with the Gamma prior for each element defined as

$$\mathrm{p}(z_n) = \mathcal{G}(z_n \,|\, d/2, d/2) \quad (10)$$

and, finally, for the degrees of freedom we also specify a Gamma prior:

$$\mathrm{p}(d) = \mathcal{G}(d \,|\, a_d, b_d) \quad (11)$$

A graphical representation of the model priors is shown in figure 1.

B. Posteriors

In an exact Bayesian model we would need to calculate the evidence by performing the following integration:

$$\int \mathrm{p}(\mathbf{y} \,|\, \mathbf{a}, \boldsymbol{\phi}, \lambda, d)\, \mathrm{p}(\mathbf{a} \,|\, \boldsymbol{\delta})\, \mathrm{p}(\boldsymbol{\delta})\, \mathrm{p}(\boldsymbol{\phi})\, \mathrm{p}(\lambda)\, \mathrm{p}(\mathbf{z} \,|\, d)\, \mathrm{p}(d)$$
$$d\mathbf{a}\, d\boldsymbol{\delta}\, d\boldsymbol{\phi}\, d\lambda\, d\mathbf{z}\, dd \quad (12)$$

As is often the case, this is intractable, so we resort to an approximation scheme. Rather than using a computationally

expensive Monte Carlo method, we use variational approximation. If Ω is the set of all variables, i.e. $\Omega = \{\mathbf{a}, \boldsymbol{\delta}, \boldsymbol{\phi}, \lambda, \mathbf{z}, d\}$, and $\mathrm{q}(\Omega) \approx \mathrm{p}(\Omega \,|\, \mathbf{y})$ denotes the approximate posterior distribution of Ω, then the variational Bayesian inference technique minimises the Kullback-Leibler (KL) divergence [25,26] between $\mathrm{q}(\Omega)$ and $\mathrm{p}(\Omega \,|\, \mathbf{y})$ (for tutorials see [27,28] and [29, chapter 10]). The divergence is greater than or equal to zero and is zero only where the two distributions are identical. There is nothing inherently approximate about variational methods; the approximations come about because we make assumptions about the factorisation of the posteriors:

$$\mathrm{q}(\Omega) = \prod_i \mathrm{q}_i(\Omega_i) \tag{13}$$

where Ω_i is a group of variables (possibly just a single one) in Ω. Attias [30] (see also [31,32]) exploits the assumed factorisation to find a general expression for the minimisation of the KL divergence in a mean field sense:

$$\log\Big(\mathrm{q}(\Omega_i)\Big) = \mathbb{E}_{/\Omega_i}\left[\log\left(\prod_j \mathrm{p}(\mathbf{y}, \Omega_j)\right)\right] \tag{14}$$

where $\mathbb{E}_{/x}[f(x)]$ denotes the posterior expectation of $f(x)$ with respect to all variables except x. If conjugate priors are chosen for each group, then the approximate posterior turns out to have the same functional form as the prior [30,33] and the variational approximations may thus be found by evaluating (14) for each group in turn. In general the hyperparameters for one posterior are dependent on the posterior hyperparameters of one or more of the other groups. [33] show that evaluating the each group in turn and then iterating over the whole set converges to a local minimum.

To maintain the clarity of this paper, we define (13) to be a full factorisation of the posteriors:

$$\left[\prod_c \mathrm{q}(a_c)\,\mathrm{q}(\delta_c)\,\mathrm{q}(\phi_c)\right]\left[\prod_n \mathrm{q}(z_n)\right]\mathrm{q}(\lambda)\,\mathrm{q}(d) \tag{15}$$

and, for convenience, define

$$u_{c,n} = \cos(\phi_c - \omega_c t_n) \tag{16}$$

We now use the factorised variational approximation technique for each variable in turn, starting with each a_c, the amplitude of the cth component sinusoid. From (14) and (15):

$$\log(\mathrm{q}(a_c)) = \mathbb{E}_{/a_c}\left[\log(\mathrm{p}(\mathbf{y} \,|\, a_c, \phi_c, \lambda, \mathbf{z})\,\mathrm{p}(a_c \,|\, \delta_c))\right] + const \tag{17}$$

All terms not dependent on a_c have been absorbed into the constant term ($const$). Expanding this and absorbing further terms into the constant, we end up with

$$\begin{aligned}
&\log(\mathrm{q}(a_c)) \\
&= -\frac{1}{2}\left[a_c^2\left(\langle\delta_c\rangle + \langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\langle u_{c,n}^2\rangle\right)\right. \\
&\qquad \left. -2a_c\left(\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\langle u_{c,n}\rangle\sum_{\substack{i=1\\ i\neq c}}^{C}\Big(\langle a_i\rangle\langle u_{i,n}\rangle - y_n\Big)\right)\right] \\
&\qquad + const
\end{aligned} \tag{18}$$

where $\langle\cdot\rangle$ denotes the posterior expectation. Since this is quadratic in a_c, it can be seen that $\mathrm{q}(a_c)$ is the Gaussian $\mathrm{q}(a_c) = \mathcal{N}(a_c \,|\, \mu_c, \sigma_c^2)$, where

$$\sigma_c^2 = \left(\langle\delta_c\rangle + \langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\langle u_{c,n}^2\rangle\right)^{-1} \tag{19}$$

$$\mu_c = \sigma_c^2\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\langle u_{c,n}\rangle\left(\sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle - y_n\right) \tag{20}$$

Following a similar procedure for the precision of each amplitude, δ_c, we obtain the Gamma posterior $\mathrm{q}(\delta_c) = \mathcal{G}(\delta_c \,|\, \alpha_{\delta_c}, \beta_{\delta_c})$, where

$$\alpha_{\delta_c} = a_\delta + 1/2 \tag{21}$$

$$\beta_{\delta_c} = b_\delta + \langle a_c^2\rangle/2 \tag{22}$$

For each phase we obtain a circular Generalised Von Mises distribution of order 2, $\mathrm{q}(\phi_c) = \mathcal{GvM}(\phi_c \,|\, \boldsymbol{\alpha}_c, \boldsymbol{\beta}_c)$ (see section II-C), where:

$$\alpha_{c,1} = \langle a_c\rangle\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\cos(\omega_c t_n)\langle g_{c,n}\rangle \tag{23}$$

$$\beta_{c,1} = \langle a_c\rangle\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\sin(\omega_c t_n)\langle g_{c,n}\rangle \tag{24}$$

$$\alpha_{c,2} = -\frac{\langle a_c^2\rangle}{4}\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\cos(2\omega_c t_n) \tag{25}$$

$$\beta_{c,2} = -\frac{\langle a_n^2\rangle}{4}\langle\lambda\rangle\sum_{n=1}^{N}\langle z_n\rangle\sin(2\omega_c t_n) \tag{26}$$

with

$$\langle g_{c,n}\rangle = y_n - \sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle \tag{27}$$

For the observation noise precision λ we get the Gamma $\mathrm{q}(\lambda) = \mathcal{G}(\lambda \,|\, \alpha_\lambda, \beta_\lambda)$, where

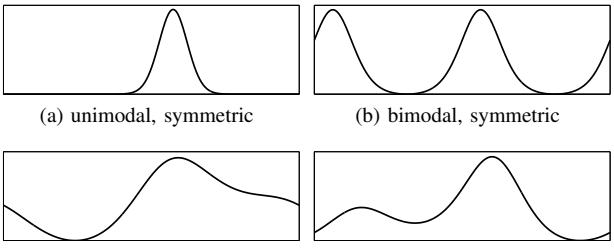
$$\alpha_\lambda = a_\lambda + N/2 \tag{28}$$

$$\begin{aligned}
\beta_\lambda = b_\lambda + \frac{1}{2}\sum_{n=1}^{N}\langle z_n\rangle\Bigg[y_n^2 - 2y_n\sum_{c=1}^{C}\langle a_c\rangle\langle u_{c,n}\rangle \\
+ \sum_{c=1}^{C}\left(\langle a_c^2\rangle\langle u_{c,n}^2\rangle + \langle a_c\rangle\langle u_{c,n}\rangle\sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle\right)\Bigg]
\end{aligned} \tag{29}$$

For each latent variable z_n we get the Gamma $\mathrm{q}(z_n) = \mathcal{G}(z_n \,|\, \alpha_z, \beta_{z_n})$, where

$$\alpha_z = (\langle d\rangle + 1)/2 \tag{30}$$

$$\begin{aligned}
\beta_{z_n} = \frac{1}{2}\Bigg\{\langle d\rangle + \langle\lambda\rangle\Bigg[y_n^2 - 2y_n\sum_{c=1}^{C}\langle a_c\rangle\langle u_{c,n}\rangle + \\
\sum_{c=1}^{C}\left(\langle a_c^2\rangle\langle u_{c,n}^2\rangle + \langle a_c\rangle\langle u_{c,n}\rangle\sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle\right)\Bigg]\Bigg\}
\end{aligned} \tag{31}$$



(a) unimodal, symmetric (b) bimodal, symmetric

(c) unimodal, asymmetric (d) bimodal, asymmetric

Figure 2 Some examples of Generalised von Mises distributions of order 2. (a) $\alpha_1 = 10$, $\beta_1 = 5$, $\alpha_2 = 0$, $\beta_2 = 0$, (b) $\alpha_1 = 0$, $\beta_1 = 0$, $\alpha_2 = 1$, $\beta_2 = 1$, (c) $\alpha_1 = 0.1689$, $\beta_1 = 0.4500$, $\alpha_2 = 0.1846$, $\beta_2 = 0.0556$, (d) $\alpha_1 = 0.5$, $\beta_1 = 0$, $\alpha_2 = 0$, $\beta_2 = 0.5$.

and, finally, for the degrees of freedom d we get the Gamma $\mathrm{q}(d) = \mathcal{G}(d \,|\, \alpha_d, \beta_d)$, where

$$\alpha_d = a_d + N/2 \tag{32}$$

$$\beta_d = b_d - \frac{1}{2}\left[N + \sum_{n=1}^{N}\Big(\langle \log(z_n)\rangle - \langle z_n\rangle\Big)\right] \tag{33}$$

Here we have used Stirling's approximation[2] for the $\log(\Gamma(d/2))$ term, as per [20].

Thus we end up with a set of expressions which define the approximate posteriors for each of the model variables, but each is dependent on the expected values of one or more of the other variables. Each of the expressions is reassessed iteratively until convergence.

The expectations associated with the Gaussian posteriors for each of the a_c and the Gamma posteriors for λ, d and each of the δ_c and z_n are the standard results:

$$\langle a_c\rangle = \mu_c \qquad \langle a_c^2\rangle = \sigma_c^2 + \mu_c^2 \tag{34}$$

$$\langle \delta_c\rangle = \alpha_{\delta_c}/\beta_{\delta_c} \qquad \langle \lambda\rangle = \alpha_\lambda/\beta_\lambda \tag{35}$$

$$\langle z_n\rangle = \alpha_z/\beta_{z_n} \qquad \langle \log(z_n)\rangle = \psi(\alpha_z) - \log(\beta_{z_n}) \tag{36}$$

$$\langle d\rangle = \alpha_d/\beta_d \tag{37}$$

This leaves us with the expectations $\langle u_{c,n}\rangle$ and $\langle u_{c,n}^2\rangle$ to consider, for which we need to take a closer look at the Generalised von Mises distribution.

C. Generalised von Mises distribution

The Generalised von Mises (GvM) distribution of order K is a circular distribution defined as [23,24]:

$$\mathrm{p}(\theta) \propto \sum_{k=1}^{K}\Big(\alpha_k \cos(k\theta) + \beta_k \sin(k\theta)\Big) \tag{38}$$

The first order distribution is the standard, symmetric, unimodal von Mises. If the distribution is of order 2, as it is in posteriors for the ϕ_c, then the result is an asymmetric, bimodal distribution, or rather one that varies between symmetric and asymmetric, and unimodal and bimodal according to the values in α_k and β_k. Figure 2 shows the shapes of some example GvM distributions of order 2.

[2]Stirling's first order approximation for $\log(\Gamma(a))$ is $(a - \frac{1}{2})\log(a) - a$

Algorithm 1 Process for estimating the posterior distributions for each of the model variables.

```
define the frequencies, ω
define the values of the prior hyperparameters
randomly initialise each a_c, φ_c, δ_c, z_n, λ and d
while not converged do
    calculate each ⟨cos(φ_c − ω_c t_n)⟩ and ⟨cos²(φ_c − ω_c t_n)⟩ (see section II-C)
    update the posteriors for each a_c using (19–20)
    update the posteriors for each δ_c using (21–22)
    update the posteriors for each φ_c using (23–26)
    update the posterior for λ using (28–29)
    update the posteriors for each z_n using (30–31)
    update the posteriors for d using (32–33)
end while
calculate each μ_c and k_c using (41–42)
```

The potential bimodality of this distribution may be considered to be problematic. Note that $g_{c,n}$ (27) represents the magnitude of the contribution of the cth component sinusoid to the nth observation. If there is no contribution from the component with angular frequency ω_c then $\langle g_{c,n}\rangle$ and $\langle a_c\rangle$ will be approximately zero, the terms in $2\omega_c$ will dominate the GvM and the distribution will, therefore, be bimodal; but if there is no contribution then the phase is arbitrary. If, however, there is a contribution from this component then $\langle g_{c,n}\rangle$ will have significant magnitude and hence the ω_c terms will dominate, leading to a unimodal distribution that is close to a standard von Mises:

$$\mathcal{G}v\mathcal{M}(\phi_c \,|\, \boldsymbol{\alpha}_c, \boldsymbol{\beta}_c) \approx \mathcal{M}(\phi_c \,|\, \alpha_{c,1}, \beta_{c,1}) \tag{39}$$

For the iterative procedure we do not need the expectation $\langle\phi_c\rangle$, only $\langle\cos(\phi_c - \omega_c t_n)\rangle$ and $\langle\cos(2\phi_c - 2\omega_c t_n)\rangle$, which we calculate numerically. Once convergence has been achieved we may calculate $\langle\phi_c\rangle$ by rewriting the von Mises in (39) in its alternative form:

$$\mathcal{M}(\phi_c \,|\, \alpha_{c,1}, \beta_{c,1}) \propto \exp\Big(k_c \cos(\phi_c - \mu_c)\Big) \tag{40}$$

where

$$\mu_c = \mathrm{atan2}(\beta_{c,1}/\alpha_{c,1}) \tag{41}$$

$$k_c = \alpha_{c,1}/\cos(\mu_c) \tag{42}$$

In this form $\langle\phi_c\rangle = \mu_c$ and k_c is a concentration parameter that acts like a precision; when k_c is large the variable is tightly distributed around the expected value; as it becomes smaller the distribution becomes flatter, tending towards the Uniform distribution $\mathcal{U}(-\pi, \pi)$ as k_c tends to zero.

D. Summary of algorithm

A summary of the algorithm is shown in algorithm 1. With Δt as the interval between each of the N observations in $\mathbf{y}$, we calculate $\mathrm{integer}(N/2) + 1$ angular frequencies uniformly distributed across the range 0 to $\pi/\Delta t$ inclusive, giving the vector $\boldsymbol{\omega}$ and the number of components C. The time vector $\mathbf{t}$ is the vector of elapsed time since the first observation. The expected values of the amplitudes and phases are initialised to

random values within appropriate ranges and the parameters of the prior distribution for the observation noise are set. The process then iteratively recalculates the posterior distributions for each variable in turn, until they converge. Finally the mean and concentration parameters for each of the phases are calculated.

The amplitude distributions are Gaussian, which admits the possibility of negative values. This is not in itself a problem as a component with amplitude $-a$ and phase ϕ is equivalent to a component with amplitude $+a$ and phase $\phi+\pi$. However, amplitudes with small magnitudes may oscillate between negative and positive values during the iterative process, causing big jumps in their associated phases and the model does not converge properly. To prevent this we use the absolute value of $\langle a_c \rangle$ in place of $\langle a_c \rangle$ in each of the posterior expressions.

We have chosen to initialise the variables randomly to demonstrate that the model converges to a good solution. We could converge more quickly if we chose to initialise the amplitudes and phases from the FFT results.

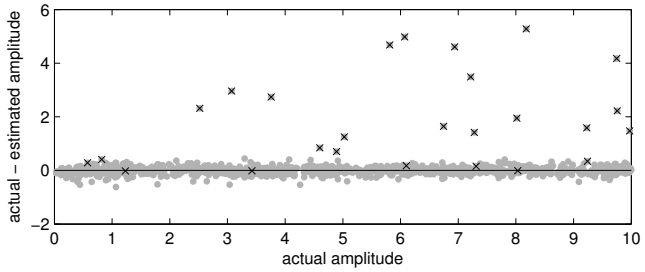
III. Illustration: synthetic data

In this section we describe results obtained from synthetic data for which we know the actual values of each of the variables. In each case there are 200 observations, at 1 second intervals, giving the set of 101 Fourier angular frequencies $\boldsymbol{\omega}$ in the range 0 to π. We start with 1,000 signals each generated from a single sinusoid with added Student-t noise. Each sinusoid has an amplitude drawn from $\mathcal{U}(0.1, 10)$, phase from $\mathcal{U}(-\pi, \pi)$ and angular frequency uniformly randomly selected from $\boldsymbol{\omega}$ (to avoid leakage). The noise distribution is Student-t, with the degrees of freedom drawn from $\mathcal{U}(1, 10)$ and precision fixed at 1. Uninformative priors are used, with

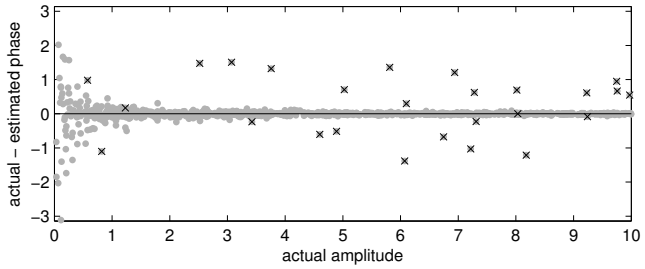
$$a_\lambda = b_\lambda = a_\delta = b_\delta = a_d = b_d = 10^{-6} \tag{43}$$

Figure 3a shows a plot of the actual amplitudes *vs* residual amplitudes for the 1,000 tests as grey dots, overlaid with black crosses where the angular frequencies are 0 or π (the lowest and highest frequencies in $\boldsymbol{\omega}$). The residuals are generally low; larger magnitudes are associated with the marked frequencies. A similar result is obtained for the phases (figure 3b), but here we can also see the effect of the noise where the true amplitude is less than the noise standard deviation.

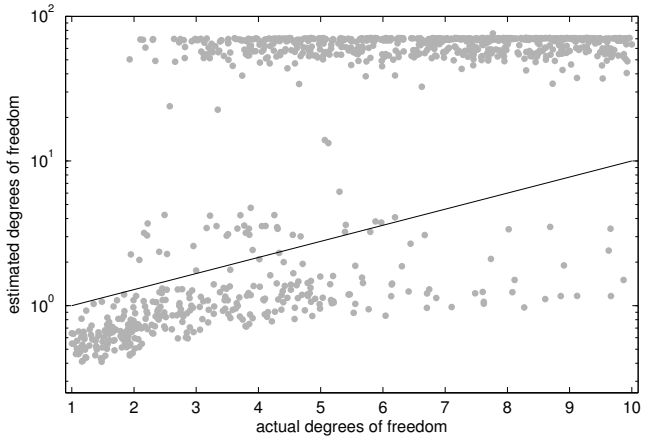
Figure 3c compares the actual and estimated degrees of freedom; here we can see that where the actual degrees of freedom are small, and hence there are more likely to be extreme noise values, the estimated noise distribution is very heavy-tailed to absorb them. With only 200 observations, a Student-t distribution with degrees of freedom 5 may be visually indistinguishable from a Gaussian and we can see that there is a transition at about the $d = 4$ mark from heavy-tailed to Gaussian noise distribution. As Bretthorst points out [5], with the full set of Fourier frequencies the noise can be captured either in the noise distribution or in the component sinusoids (particularly those with high frequencies). It seems that in this model, with uninformative priors, the outliers are being absorbed into the noise distribution while the remainder of the noise is being captured in the sinusoids. It will be shown



(a) amplitude



(b) phase



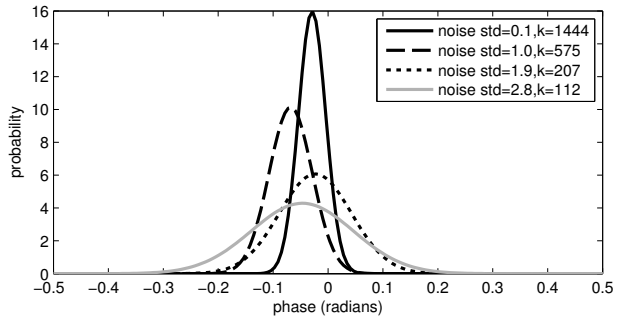
(c) degrees of freedom

Figure 3 The actual vs estimated (a) amplitudes, (b) phases and (c) degrees of freedom for 1,000 synthetic datasets; the dashed lines indicate equality. In (a) and (b) the crosses mark tests where the frequency is either 0 or π.

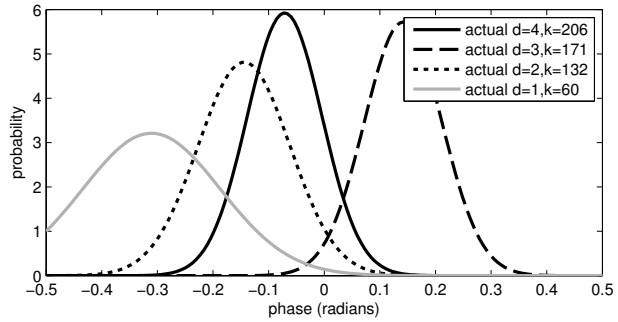
in section IV-B that this balance of power can be manipulated by changing the prior over the δ_c.

For a second set of 1,000 tests, the amplitudes and phases of single sinusoids are fixed at 1 and 0 respectively, while the degrees of freedom are drawn from $\mathcal{U}(1, 10)$ and noise standard deviations from $\mathcal{U}(0, 2)$. There is a linear correlation between the actual and estimated noise standard deviation (0.56), which becomes stronger as the noise becomes more Gaussian-like, but the precions are significantly over-estimated.

The algorithm is easily amended to reflect Gaussian noise assumptions. One method is to give the degrees of freedom d a very tight prior around a high mean, since the Student-t distribution tends to a Gaussian as d tends to infinity. However, if each z_n initialised to 1, and z_n and d are not updated, then the noise model is Gaussian (this has the additional benefit of faster run times). If the observation noise is Gaussian then

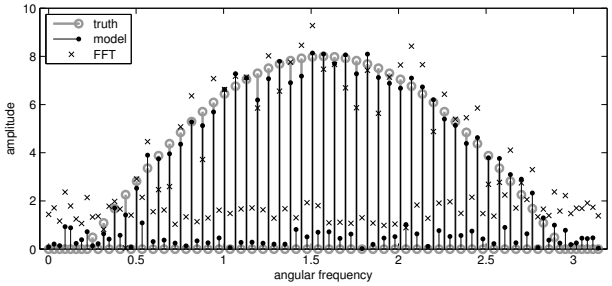


(a) fixed degrees of freedom

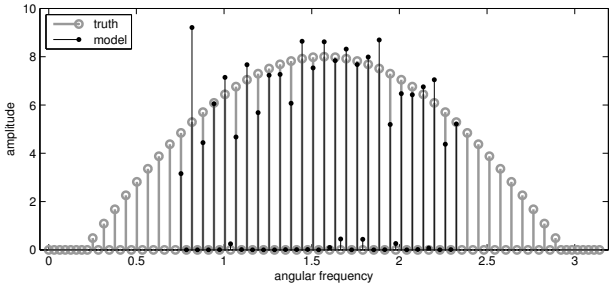


(b) fixed precision

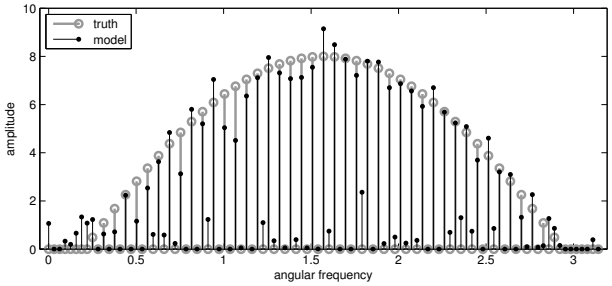
Figure 4 How phase distribution varies with noise. In (a) the degrees of freedom are fixed at $d = 1.5$ (with $a = 5$, $\phi = 0$, $\omega = 0.5969$) and the precision λ varied as shown in the legend. In (b) the precision is fixed at $\lambda = 1/3^2$ (with $a = 5$, $\phi = 0$, $\omega = 0.1257$) and the degrees of freedom d varied as shown in the legend. In both cases, as the noise variance increases, the certainty decreases.



(a) full $\boldsymbol{\omega}$, evenly-spaced observations



(b) partial $\boldsymbol{\omega}$, evenly-spaced observations



(c) full $\boldsymbol{\omega}$, unevenly-spaced observations

Figure 5 The true (grey stems) and estimated (black stems) amplitude spectra for three different scenarios (see subcaptions). Figure (a) is overlaid with the result from FFT (black crosses).

the model produces almost identical results with either noise assumption. If the observation noise is Student-t (or rather, if there are significant outliers) then the model with Student-t noise assumptions absorbs the outliers and gives a truer representation of the spectrum.

The model is competent at estimated the expected values of the variables, but the benefit of the Bayesian formulation comes from the quantification of uncertainty in those estimates. A single sinusoid (amplitude 5, phase 0, angular frequency 0.2827) is used to generate four signals, with different Student-t noise distributions. In each case the degrees of freedom are set to 1.5, and the noise standard deviation is set respectively to 0.1, 1.0, 1.9 and 2.8. The model is trained for each of the four resulting signals. Figure 4a compares the posterior distributions of the phases at the selected frequency; as expected, as the signal-to-noise ratio decreases, the phase distribution becomes flatter showing increasing uncertainty. In a second set of four signals the noise standard deviation is fixed at 3 and the degrees of freedom set respectively to 1, 2, 3 and 4. Figure 4b shows that as the degrees of freedom decrease, once again the phase distribution becomes flatter showing increasing uncertainty.

So far the model has been demonstrated on signals constructed from single sinusoids. Figure 5 shows three amplitude spectra with multiple components. In each case the truth is shown as grey stems topped with circles, the model's expectations from 200 observations with added Student-t noise ($\lambda = 1$, $d = 1.5$) as black stems topped with points and, in the top plot only, the FFT spectrum calculated from the noise observations as black crosses. Figure 5a shows the result from using the full set of Fourier frequencies and evenly-spaced observations. The effect of the noise can be seen in the FFT estimates, especially where the true amplitude is zero. The model is much less affected, with the amplitudes for the zero-components constrained to be at or close to zero. Figure 5b shows the results from training the model on a subset of the Fourier frequencies and the same noisy observations. For the selected frequencies the results are very similar, with the zero-components particularly well estimated. Figure 5c shows the results obtained using the full set of frequencies, but with irregular periods between observations. Some zero-components are completely switched off, but the estimates are in general poorer than those shown previously. The root mean squared (RMS) errors of the amplitude spectra are 0.48, 0.91 and 0.67 respectively, while that for FFT is 1.40. Running the same three tests but using a Gaussian noise model instead of the Student-t (against the same noisy observations) gives RMS errors of 1.39, 0.88 and 3.86 respectively.

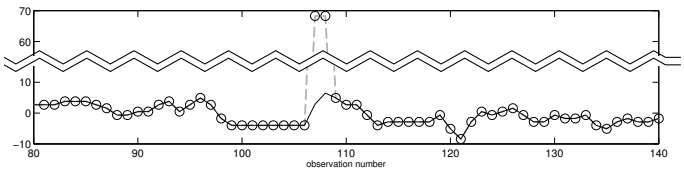


Figure 6 GPS data: the observations are shown as grey dashed lines marked with black circles; the black line is the model's reconstruction of the signal. The reconstructed signal is very close to the observations, except in the region of the spike where the excessive magnitude has been absorbed into the noise distribution and a more plausible signal is suggested.

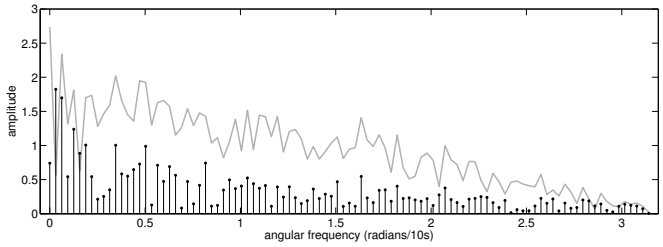


Figure 7 GPS data: the black stems mark the posterior expected value of the amplitude; the grey line is the result from FFT. The outliers have noticeably affected the FFT spectrum, whereas the model has absorbed them into the noise distribution.

IV. Results: real data

We demonstrate the model on two sets of real data: Global Positioning System (GPS) data recorded by a device on board a ship which was tied up alongside a jetty, and the traditional Wolf's sunspot data used by [34] and others.

A. GPS data

The top plot in figure 6 shows observations of latitude (converted to an offset in metres from some reference point) recorded by the GPS device at 10 second intervals. Note the large (and erroneous) spike at observations 107–8. The margin for error given for GPS in normal operation is 7.8 metres [35, table 3.4-1]; this spike is significantly greater than that and it is impossible for the device to have made such a movement within the time interval. The model was trained for 100 iterations using the priors shown in (43) and the signal reconstructed from the posteriors of the model variables. The result is shown in the bottom plot of figure 6; the grey dashed line and black circles mark the observations; the black line is the reconstruction. The reconstructed signal is very close to the observations, except in the region of the spike where the excessive magnitude has been absorbed into the noise distribution and a more plausible "truth" is suggested.

The amplitude spectra from the model and, for comparison, from FFT are shown in figure 7. The expected noise precision, $\langle\lambda\rangle$, is 0.616, and the degrees of freedom, $\langle d\rangle$, is 5.203.

B. Wolf's sunspot data

Wolf's annual sunspot data have been recorded since 1700 and are available from [36] up to and including 2009, giving 310 observations in total. Rather than using uninformative

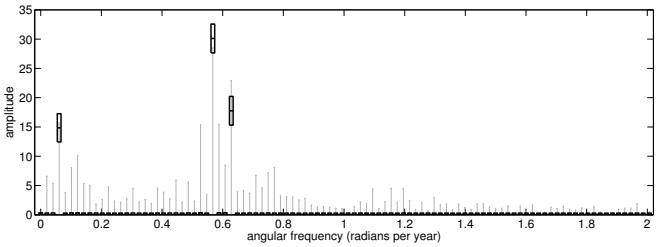


Figure 8 Amplitude spectra for Wolf's annual sunspots data for frequencies up to 2 rad/year. The grey stems show the FFT spectrum. The black boxes mark the posterior expected value of the amplitude and ± 2 standard deviations. Using the prior $\mathcal{G}(\delta_c \,|\, 5, 0.1)$ the model is constraining all but three components to have zero amplitude.

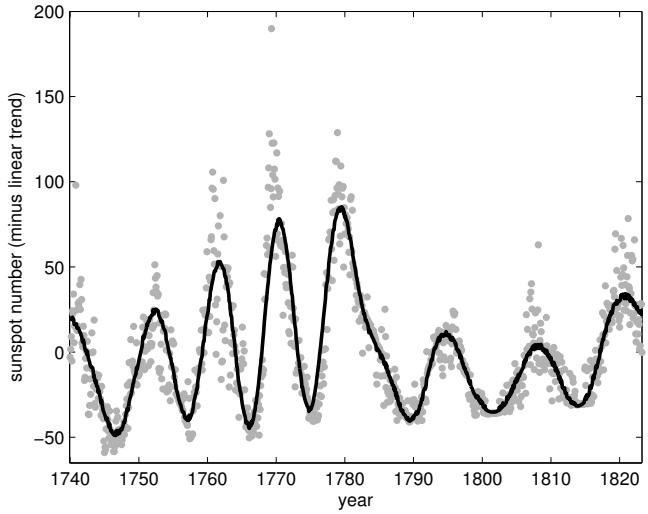


Figure 9 Wolf's monthly sunspot data, with linear trend removed, are shown as grey points. The black line is the reconstruction from the model using the $\mathcal{G}(\delta_c \,|\, 5, 0.1)$ prior over the amplitude precisions.

priors, the model was trained (after removal of the linear trend) with the prior on the amplitude precisions set to $\mathcal{G}(\delta_c \,|\, 5, 0.1)$. This has the effect of providing a small amount of pressure on the amplitudes to be "switched off". The resulting amplitude spectrum is shown in figure 8; the period of the peak amplitude is 11.07 years which is very close to the 11 years previously estimated by [34] and all but three frequencies have been switched off. The posterior distributions for the phases of the three peaks (not plotted here due to space constraints) have μ_c (mean) values of -2.89, -3.09 and 0.03 respectively and k_c (precision) values of 151, 599 and 209 respectively; the model is most certain about the phase of the main (11.07 year) peak.

Using the same prior over the δ_c, the model was trained against the first 1,000 observations (January 1740 to April 1832 inclusive) of the monthly sunspot data. In the resulting spectrum only 13 frequencies have posterior amplitudes that are more than two standard deviations away from zero (the remainder have amplitudes less than 0.26). The effect of this is shown in figure 9; notice how the model has produced a smooth reconstruction that appears to identify the underlying shape of the observations, absorbing the deviations into the noise distribution (the heavy-tailed Student-t distribution with $\langle\lambda\rangle = 0.022$ and $\langle d\rangle = 1.14$).

V. Conclusions

The iterative procedure required to estimate the model's posterior distributions means that it cannot compete with FFT in terms of speed or computational complexity, being quadratic in the number of frequencies. However, for data contaminated with outliers it is effective in separating the outliers from the underlying signal, and automatic relevance determination automatically "switches off" component frequencies across the spectrum. The selection of a subset of frequencies and the choice of priors cause more of the observed signal to be absorbed into the noise distribution. For an analysis of the effects of missing data on this model, see [37].

A similar model is obtained if the observations are multidimensional. In this case the noise precision λ may be assigned a Wishart prior, Wishart distributions may be assigned to the amplitude precisions (the δ_c) and multidimensional von Mises distributions for the phases.

The connection between the power spectrum and the coefficients in an autoregression model has long been recognised [38]. An autoregressive model that corresponds to this Bayesian, Student-t spectral analysis model has been described by [20], who also show the utility of automatic relevance determination to suppress unwarranted coefficients and thus automatically estimate the model order.

References

[1] A. Schuster, "On lunar and solar periodicities of earthquakes," in *Proceedings of the Royal Society*, vol. 61, 1897, pp. 455–465.

[2] J. Cooley and J. Tukey, "An algorithm for the machine computation of complex fourier series," *Mathematical Computation*, vol. 19, pp. 297–301, 1965.

[3] E. Jaynes, "Bayesian spectrum and chirp analysis," in *Proceedings of the Third Workshop on Maximum Entropy and Bayesian Method*, Laramie, Wyoming, August 1-4 1983.

[4] ——, *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. Smith and G. Erickson, Eds. D. Reidel Publishing Co., 1987.

[5] G. Bretthorst, "Excerpts from Bayesian spectrum analysis and parameter estimation," *Maximum Entropy and Bayesian Methods in Science and Engineering*, vol. 1, pp. 75–145, 1988.

[6] L. Dou and R. Hodgson, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation. I," *Inverse Problems*, vol. 11, no. 5, pp. 1069–1085, 1995.

[7] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via Reversible Jump MCMC," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2667–2676, 1999.

[8] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[9] J. Nielsen, "Sinusoidal parameter estimation - a Bayesian approach," Master's thesis, Aalborg University, 2009.

[10] J. Nielsen, M. Christensen, A. Cemgil, S. Godsill, and S. Jensen, "Bayesian interpolation and parameter estimation in a dynamic sinusoidal model," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 7, pp. 1986–1998, 2011.

[11] P. Huber, "Robust statistics: A review," *The Annals of Mathematical Statistics*, vol. 43, no. 4, pp. 1041–1067, 1972.

[12] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons Ltd, Chichester, 1994.

[13] F. Ruggeri, "Bayesian robustness," *Newsletter, European Working Group for Multiple Criteria Decision Aiding*, vol. 3, no. 17, pp. 6–10, 2008.

[14] A. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[15] A. Chave, D. Thomson, and M. Ander, "On the robust estimation of power spectra, coherences, and transfer functions," *Journal of Geophysical Research*, vol. 92, no. B1, pp. 633–648, 1987.

[16] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, vol. 6, no. 117, 2005.

[17] R. Pearson, H. Lähdesmäki, H. Huttunen, and O. Yli-Harja, "Detecting periodicity in nonideal datasets," in *Proc of the SIAM International Conference on Data Mining*, San Francisco, CA, USA, 13 May 2003.

[18] S. Roberts and W. Penny, "Variational bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, September 2002.

[19] K. Lange, R. Little, and J. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, pp. 881–896, 1989.

[20] J. Christmas and R. Everson, "Robust autoregression: Student-t innovations using variational Bayes," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 48–57, Jan 2011.

[21] D. Mackay, "Bayesian non-linear modelling for the prediction competition," *ASHRAE Transactions*, vol. 100, no. 2, pp. 1053–1062, 1994.

[22] R. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, University of Toronto, Canada, 1995.

[23] V. Maksimov, "Necessary and sufficient statistics for the family of shifts of probability distributions on continuous bicompact groups," *Theory of Probability and its Applications*, vol. 12, pp. 267–280, 1967, (translated by A.R. Kraiman).

[24] R. Gatto and S. Jammalamadaka, "The generalized von Mises distribution," *Statistical Methodology*, vol. 4, pp. 341–353, 2007.

[25] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[26] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc, 1991.

[27] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[28] H. Lappalainen and J. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*. Berlin: Springer-Verlag, 2000, pp. 75–92.

[29] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[30] H. Attias, "A variational Bayesian framework for graphical models," *Advances in Neural Information Processing Systems*, vol. 12, pp. 209–215, 2000.

[31] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics*, vol. 7. Oxford University Press, 2002, pp. 453–464.

[32] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, 2003.

[33] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2001, pp. 507–513.

[34] G. Bretthorst, "Bayesian analysis. I. Parameter estimation using quadrature NMR models," *Journal of Magnetic Resonance*, vol. 88, pp. 533–551, 1990.

[35] "Global Positioning System standard positioning service performance standard," U.S. DoD, Tech. Rep. 4th edition, September 2008.

[36] National Geophysical Data Center, http://www.ngdc.noaa.gov/nndc/struts/results?t=102827&s=5&d=8,430,9, last accessed 22 Mar 2013.

[37] J. Christmas, "The effect of missing data on robust Bayesian spectral analysis," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, UK, 2013.

[38] H. Akaike, "Power spectrum estimation through autoregressive model fitting," *Annals of the Institute of Statistical Mathematics*, vol. 21, no. 1, pp. 407–419, December 1969.

Jacqueline Christmas received a BSc in Computer Science from the University of Exeter in 1987. After many years in industry, she returned to Exeter where she was awarded an MSc in Applied Artificial Intelligence in 2007 and a PhD in the field of statistical pattern recognition in 2011. She worked on statistical models for tracking insects in noisy videos and is now involved in modelling sea waves. Current research interests are statistical modelling, pattern recognition and video object tracking.