2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

THE EFFECT OF MISSING DATA ON ROBUST BAYESIAN SPECTRAL ANALYSIS

Jacqueline Christmas

Department of Computer Science, University of Exeter, Exeter, EX4 4QF, UK.

ABSTRACT

We investigate the effects of missing observations on the robust Bayesian model for spectral analysis introduced by Christmas [2013]. The model assumes Student-t distributed noise and uses an automatic relevance determination prior on the precisions of the amplitudes of the component sinusoids and it is not obvious what their effect will be when some of the otherwise temporally uniformly sampled data is missing.

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

1. INTRODUCTION

In [1] we introduce a Bayesian model for spectral analysis that is made robust to outliers by assuming that the observation noise is distributed according to the heavy-tailed Student-t distribution. We learn posterior distributions for the noise variables, as well as the amplitudes and phases of the sinusoids that make up the spectral components. In this paper we explore the effects of missing observations on the model.

In general our observations are discrete samples of some (unknown) continuous function. If the sampling interval, f_s, is uniform then there is an infinite number of other continuous functions that fit the data, but there will only be one limited to the Nyquist frequency ($\frac{1}{2}f_s$) [2, 3]. Where the true function includes frequencies above the Nyquist limit, unavoidable *aliasing* occurs. With uniform sampling the effects of aliasing are predictable; [4] shows that where the spacings are unequal, spectral analysis produces comparable results to those from uniformly-spaced data, but that the effects of aliasing must be analysed in light of the actual time intervals recorded.

With real datasets it is often the intention to sample uniformly in time, but errors or failures ensure that some observations are missing. In these cases the missing values are often interpolated (see, for example, [5]) or the data are resampled to a uniform time interval (so-called *gridding*; see, for example, [6]) so that standard methods of analysis may be utilised. A method that operates on the irregularly-sampled directly is *normalised convolution* [7], which associates a degree of certainty with each sample and uses these certainties to interpolate the missing data.

Our model also operates directly on the incomplete data, but aims to learn the degree of uncertainty in the resulting spectral decomposition through the estimation of posterior probability distributions. It is not clear what the effects are of the Student-t noise assumption and the use of the *automatic relevance determination* priors [8, 9] on the precisions of the component amplitudes (which tend to suppress components for which there is no evidence in the data, by constraining the amplitudes to be close to zero).

A brief overview of the model is described in section 2. Results from synthetic and real data are shown in section 3 and conclusions drawn in section 4.

2. THE MODEL

Our model deconstructs the set of N observations into a sum of C sinusoidal components with varying amplitudes and phases, so that the nth observation at time t_n is defined as:

$$y_n = \left[\sum_{c=1}^{C} a_c \cos(\phi_c - \omega_c t_n)\right] + \epsilon_n \tag{1}$$

where a_c, ϕ_c and ω_c are respectively the amplitude, phase and angular frequency of the cth component and ϵ_n is the observation noise. Noise is often assumed to be Gaussian distributed, but the Gaussian is known to be badly affected by outliers, so if the true distribution is significantly heavier-tailed than Gaussian then any model built upon this assumption is likely to perform badly. We assume that the noise is distributed according to the heavy-tailed Student-t with zero mean:

$$\begin{aligned} \mathrm{p}(\epsilon_n) &= \mathcal{S}(\epsilon_n \,|\, 0, \lambda, d) & (2)\\ &= \frac{\Gamma\left(\frac{(d+1)}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \left(\frac{\lambda}{\pi d}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda \epsilon_n^2}{d}\right)^{-\frac{d+1}{2}} & (3)\\ &= \int_0^\infty \mathcal{N}(\epsilon_n \,|\, 0, (\lambda z)^{-1})\, \mathcal{G}(z \,|\, \frac{d}{2}, \frac{d}{2})\, \mathrm{d}z & (4) \end{aligned}$$

where λ is known as the precision, d is the degrees of freedom, $\Gamma(\cdot)$ is the gamma function, $\mathcal{N}(\cdot)$ a Gaussian distribution and $\mathcal{G}(\cdot)$ a Gamma distribution[1]. 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.

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

978-1-4799-1180-6/13/$31.00 ©2013 IEEE

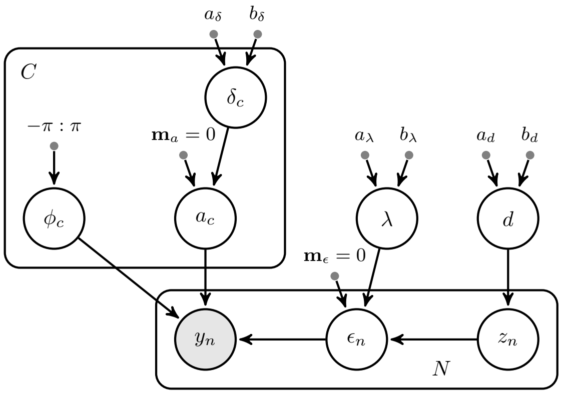


Fig. 1 A graphical representation of the model priors.

The ω_c and t_n are parameters whose values we specify. We treat as variables the a_c, ϕ_c, λ, d and the latent z_n (from (4)) and infer posterior distributions for them. The integrals required for exact Bayesian inference are intractable, so, instead of using computationally expensive Monte Carlo methods, we use variational approximation.

2.1. Priors

The Student-t observation noise assumption leads to the following likelihood (using vector notation):

$$\mathrm{p}(y_n \mid \mathbf{a}, \boldsymbol{\phi}, \lambda, d; \boldsymbol{\omega}, t_n) = \mathcal{S}\big(y_n \mid \mathbf{a}^{\mathrm{T}} \cos(\boldsymbol{\phi} - \boldsymbol{\omega} t_n), \lambda, d\big) \tag{5}$$

In this model we consider the component sinusoids to be independent of one another, so to each amplitude we assign a Gaussian prior, with precision δ_c, and to each δ_c a conjugate Gamma distribution:

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

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

The latter is the *automatic relevance determination* prior [8, 9]. With no prior information about the component phases ϕ_c, we model each of them as Uniformly distributed over the full range of values, i.e. any 2π range:

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

The observation noise precision λ is assigned a conjugate Gamma prior, as is the degrees of freedom d and each element of the latent variable $\mathbf{z}$ (from (4)):

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

$$\mathrm{p}(d) = \mathcal{G}(d \mid a_d, b_d) \tag{10}$$

$$\mathrm{p}(z_n) = \mathcal{G}(z_n \mid d/2, d/2) \tag{11}$$

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

2.2. Posteriors

The integrals required to calculate the evidence in exact Bayesian inference are intractable. Instead we use the variational approximation technique which minimises the Kullback-Leibler (KL) divergence [10, 11] between the approximate posterior distribution $\mathrm{q}(\cdot)$ and the true posterior $\mathrm{p}(\cdot \mid \mathbf{y})$ (for tutorials see [12, 13]). We use mean field approximation for the minimisation [14, 15] (see also [16, 17]), which exploits an assumed factorisation of the posteriors (conditioning of the true posteriors on $\mathbf{y}$ has been omitted for clarity):

$$\mathrm{p}(\mathbf{a} \mid \boldsymbol{\delta})\,\mathrm{p}(\boldsymbol{\delta})\,\mathrm{p}(\boldsymbol{\phi})\,\mathrm{p}(\lambda)\,\mathrm{p}(\mathbf{z} \mid d)\,\mathrm{p}(d) \approx \left[\prod_c \mathrm{q}(a_c)\,\mathrm{q}(\delta_c)\,\mathrm{q}(\phi_c)\right]\left[\prod_n \mathrm{q}(z_n)\right]\mathrm{q}(\lambda)\,\mathrm{q}(d) \tag{12}$$

With $\langle\cdot\rangle$ denoting a posterior expectation and, for convenience, defining $u_{c,n} = \cos(\phi_c - \omega_c t_n)$, we get the following definitions for the approximate posteriors. For the amplitudes $\mathrm{q}(a_c)$ is the Gaussian $\mathcal{N}(a_c \mid \mu_c, \sigma_c^2)$, where

$$\sigma_c^2 = \left(\langle\delta_c\rangle + \langle\lambda\rangle \sum_{n=1}^{N} \langle z_n\rangle \langle u_{c,n}^2\rangle\right)^{-1} \tag{13}$$

$$\mu_c = \sigma_c^2 \langle\lambda\rangle \sum_{n=1}^{N} \langle z_n\rangle \langle u_{c,n}\rangle \left(\sum_{\substack{i=1\\ i\neq c}}^{C} \langle a_i\rangle \langle u_{i,n}\rangle - y_n\right) \tag{14}$$

This allows the possibility of negative amplitudes; to avoid oscillations around zero, and the consequent effect on the phases and on convergence generally, we use the absolute value of $\langle a_c\rangle$ in the later posterior expressions.

For the precision of each amplitude, $\mathrm{q}(\delta_c)$ is the Gamma $\mathcal{G}(\delta_c \mid \alpha_{\delta_c}, \beta_{\delta_c})$, where

$$\alpha_{\delta_c} = a_\delta + 1/2 \tag{15}$$

$$\beta_{\delta_c} = b_\delta + \langle a_c^2\rangle/2 \tag{16}$$

For each phase, $\mathrm{q}(\phi_c)$ is a circular Generalised Von Mises distribution of order 2. The potential bimodality of this distribution is problematic, but a good approximation (see [1]) is the standard, unimodal von Mises $\mathcal{M}(\phi_c \mid \mu_c, \kappa_c) \propto \exp(\kappa_c \cos(\phi_c - \mu_c))$, where:

$$\mu_c = \mathrm{atan2}(\beta_{c,1}, \alpha_{c,1}) \tag{17}$$

$$\kappa_c = \alpha_{c,1}/\cos(\mu_c) \tag{18}$$

$$\alpha_{c,1} = \langle a_c\rangle\langle\lambda\rangle \sum_{n=1}^{N} \langle z_n\rangle \cos(\omega_c t_n) \langle g_{c,n}\rangle \tag{19}$$

$$\beta_{c,1} = \langle a_c\rangle\langle\lambda\rangle \sum_{n=1}^{N} \langle z_n\rangle \sin(\omega_c t_n) \langle g_{c,n}\rangle \tag{20}$$

$$\langle g_{c,n}\rangle = y_n - \sum_{\substack{i=1\\ i\neq c}}^{C} \langle a_i\rangle \langle u_{i,n}\rangle \tag{21}$$

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

$$\alpha_\lambda = a_\lambda + N/2 \tag{22}$$

$$\beta_\lambda = b_\lambda + \frac{1}{2}\sum_{n=1}^{N}\langle z_n\rangle\Bigg[y_n^2 - 2y_n\sum_{c=1}^{C}\langle a_c\rangle\langle u_{c,n}\rangle + \sum_{c=1}^{C}\Bigg(\langle a_c^2\rangle\langle u_{c,n}^2\rangle + \langle a_c\rangle\langle u_{c,n}\rangle\sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle\Bigg)\Bigg] \tag{23}$$

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

$$\alpha_z = (\langle d\rangle + 1)/2 \tag{24}$$

$$\beta_{z_n} = \frac{1}{2}\Bigg\{\langle d\rangle + \langle\lambda\rangle\Bigg[y_n^2 - 2y_n\sum_{c=1}^{C}\langle a_c\rangle\langle u_{c,n}\rangle + \sum_{c=1}^{C}\Bigg(\langle a_c^2\rangle\langle u_{c,n}^2\rangle + \langle a_c\rangle\langle u_{c,n}\rangle\sum_{\substack{i=1\\ i\neq c}}^{C}\langle a_i\rangle\langle u_{i,n}\rangle\Bigg)\Bigg]\Bigg\} \tag{25}$$

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

$$\alpha_d = a_d + N/2 \tag{26}$$

$$\beta_d = b_d - \frac{1}{2}\left[N + \sum_{n=1}^{N}\Bigg(\langle\log(z_n)\rangle - \langle z_n\rangle\Bigg)\right] \tag{27}$$

The expectations $\langle\cos(\phi_c)\rangle$ and $\langle\sin(\phi_c)\rangle$ required for $\langle u_{c,n}\rangle$ are calculated numerically.

In general the hyperparameters for one posterior are dependent on the hyperparameters of one or more of the other posteriors. [15] show that evaluating the each posterior in turn and then iterating over the whole set converges to a local minimum of the KL divergence.

3. RESULTS

In all of these tests we start with N original observations with a uniform unit time interval between them (i.e. $\Delta t = 1$); M of these observations are then removed.

The effect on the model of missing observations depends on how the missing observations are arranged and which angular frequencies (the values in $\boldsymbol{\omega}$) are selected to train the model on. We will demonstrate the two extremes of arrangements: (A) values missing at random and (B) a contiguous block of missing values which we locate at the centre of the observation period. If M is very small then we might be happy to construct $\boldsymbol{\omega}$ as if none of the observations are missing, since Δt is still 1 in the majority of cases (even though we are trying to estimate distributions for a total of N amplitude and phase variables given fewer than N observations); we shall refer to this as $\boldsymbol{\omega}_f$. However, with more observations missing it would perhaps be better to construct $\boldsymbol{\omega}$ based on the $N - M$ observations; we shall refer to this as $\boldsymbol{\omega}_p$. For comparison, as well as the four cases so far described (arrangements A and B, and the two sets of frequencies $\boldsymbol{\omega}_f$ and $\boldsymbol{\omega}_p$), we will also show the results for $N - M$ observations uniformly distributed in time using the frequencies in $\boldsymbol{\omega}_p$.

To construct one test set, a signal is constructed from a single significant component, with amplitude 10 and phase $\pi/6$. To avoid the contaminating effects of leakage, the angular frequency of this significant component, ω_c, is set to the smallest value in whichever of $\boldsymbol{\omega}_f$ and $\boldsymbol{\omega}_p$ is being tested that is greater than or equal to 0.1. For arrangements A and B, $N = 100$ observations are recorded at 1Hz; for arrangement A, M observations are randomly selected and removed; for arrangement B the centre M observations are removed. For the comparison set, $N - M$ observations are uniformly-distributed across the observation period. Noise is then added to the observations: a single set of M samples is taken from a zero-mean Student-t distribution, with precision 1 and degrees of freedom 1.5, and this same set of noise samples is added to the observations in each of the five test cases.

For each test set, the five different cases were trained for M in the range of 0 to 50, at intervals of 5, until convergence. This was repeated for 100 test sets. As in [1] we use uninformative priors, with hyperparameters set as follows:

$$a_\lambda = b_\lambda = a_\delta = b_\delta = 10^{-6}, \quad a_d = 100, \quad b_d = 10 \tag{28}$$

Figure 2 shows how the estimate of the amplitude at the significant frequency (in black) falls, and the mean amplitude at all other frequencies (in grey) rises, as the proportion of missing observations increases. Figure 3 shows how the uncertainty (in the form of the posterior standard deviation) increases, both for the significant component and the others. The comparison case shows a small rise in the uncertainty as the number of observations decreases.

Figure 4 shows how the posterior concentration hyperparameter κ_c (18) for the phase of the significant component changes as the missing proportion increases. As in figures 2 and 3, the shape of the plots does not depend on the choice of angular frequency set, so henceforth only those for $\boldsymbol{\omega}_p$ are shown. For arrangement A there is a clear relationship between the proportion of missing values and the concentration parameter: as M increases, so κ_c, and therefore certainty, decreases. For arrangement B the effect is less marked, and there is clearly a link between the sudden drop in the estimate of the significant component's amplitude (see figure 2d) and that of the significant phase concentration. Again, the comparison case shows a higher uncertainty as the number of observations decreases.

Figure 5 shows how the posterior distribution of the phase changes as the missing proportion increases for one arbitrarily selected test set. For clarity, only results for the 0%, 15%, 30% and 50% missing values are shown. For arrangement

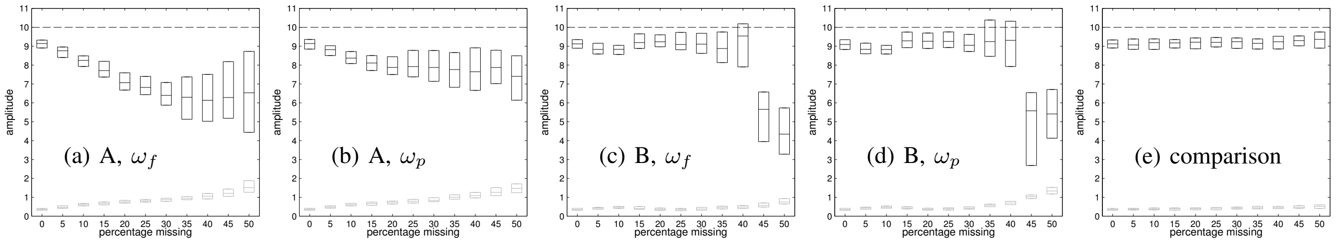


Fig. 2 As the proportion of missing observations increases, the estimate of the amplitude at the significant frequency (in black) decreases and the mean estimate of all other frequencies (in grey) increases. The horizontal dashed line marks the true significant amplitude. The subcaptions refer to the missing value arrangement (A=missing at random, B=missing a single block in the centre of the observations) and which set of angular frequencies is used. The boxes showing the median and 25th and 75th percentiles for each proportion of missing values.

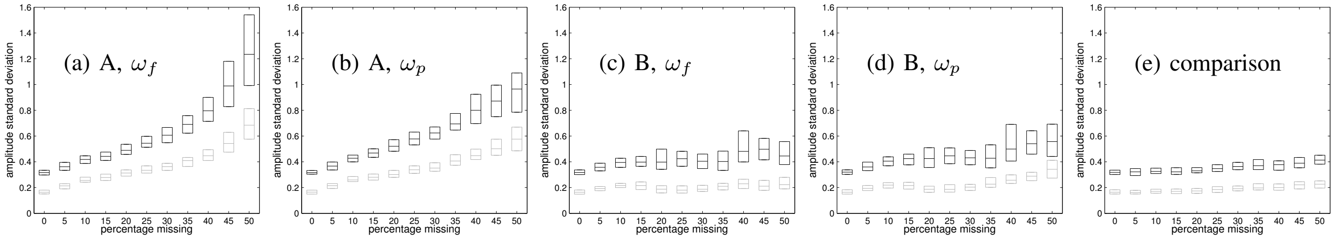
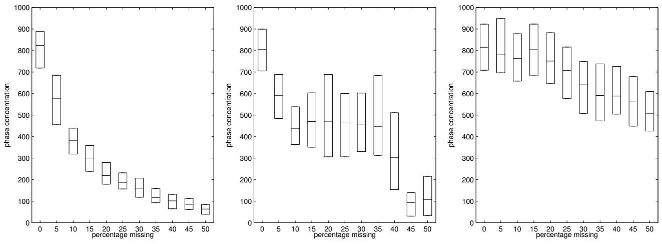


Fig. 3 As the proportion of missing observations increases, the uncertainty (posterior standard deviation) in the amplitude at the significant frequency (in black) and at all other frequencies (in grey) increases. Plot format is as per figure 2.



(a) A, ω_p (b) B, ω_p (c) comparison

Fig. 4 The change in the posterior concentration κ_c (18) of the significant phase as the proportion of missing observations increases. Plot format is as per figure 2.

A, as expected, the uncertainty (spread) in the phase estimation increases as the number of missing values increases. Although the expected value becomes less accurate, the true value (marked as a thin, vertical, black line) lies well within the body of the distribution. For arrangement B, in this example, we get an accurate phase estimate only where there is no missing data, and we see the pattern shown in figure 4b whereby the concentration does not change markedly until we reach the 50% missing mark. The comparison case continues to accurately estimate the phase and the concentration does not change significantly.

Using the posterior expectations for the amplitude and phase variables, a_c and ϕ_c, we may use (1), without the noise term, to reconstruct the wave form associated with the observations. Ideally the model will capture the noise separately and so the reconstruction will be of the underlying signal and not the observations themselves. For each test, two mean squared errors are calculated: (a) between the reconstruction and the underlying signal at the observation times, and (b) between the full reconstruction (i.e. for all N observation times) and the full signal. Figure 6 shows, for each of the five test cases, how these three error measures vary as the proportion of missing values increases.

As before, we see that the results for arrangements A and B do not appear to be highly dependent on the choice of set of angular frequencies. In both cases a better reconstruction of the signal is achieved for arrangement B than in A. The comparison case is almost unperturbed by missing data.

Figure 7 shows an example of the original signal overlaid with the observations and signal reconstruction for one of the tests with 25% missing values. Here we can see that for arrangement B the model appears to be better at absorbing the gross outliers into the noise distribution.

3.1. Real data

The top plot in figure 8 shows the monthly precipitation recorded by the Rakkestad weather station in Norway between January 1990 and December 2012 [18]. Many of the weather stations are truly missing data, but in this case the data is complete and a block of 26 observations (approximately 10%) have been artificially removed prior to training the model. The model was trained for 500 iterations on the mean-centred data and the results compared with those from the discrete Fourier transform (DFT) (where the missing values were set to zero). This plot shows the true data in grey,

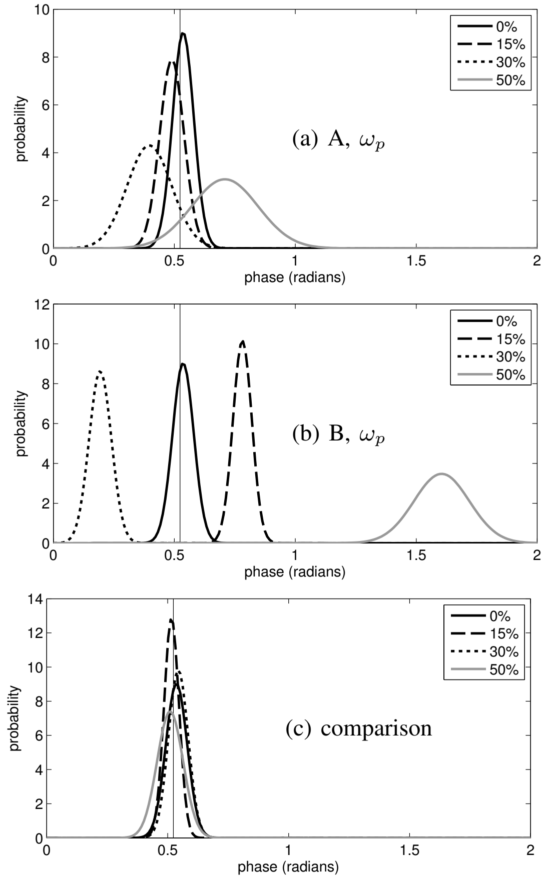
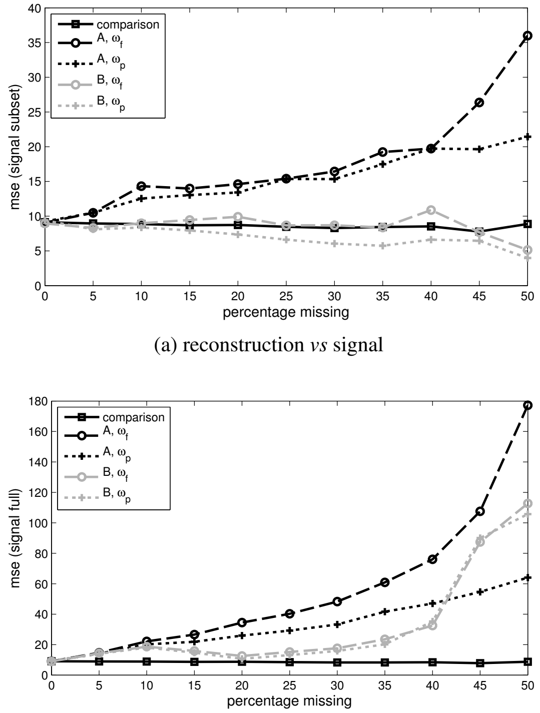


Fig. 5 The change in the posterior distribution of the significant phase as the proportion of missing observations increases. Subcaption format is as per figure 2. (a) Shows how, for arrangement A, the spread of the distribution increases as the missing proportion increases, but the expectation remains close to the true value (vertical black line). (b) Shows that arrangement B (for this example) leads to less accurate expectations where data is missing, but the spread is less affected. The comparison case, in (c), is largely unaffected by the reduction in the number of observations.

overlaid with the reconstruction from the trained model; for the missing data this is poor, but for the remaining data it is generally very good. More significantly, the bottom plot compares the amplitude spectrum from the model with those from DFT. The spectra are very similar, but note how many components have been "switched off" by automatic relevance determination in the new model (amplitudes of less that 0.1mm are marked by black circles), showing that the model is more parsimonious in its representation of the original data. Despite the missing data, the annual precipitation cycle is still clearly identified.

4. CONCLUSIONS

In conclusion, the effect on the model of missing observations depends on how the missing values are distributed, with a lesser effect observed where a contiguous block is missing than where observations are randomly missing. There seems to be little difference between the two sets of angular frequencies, the $N/2$ values in $\boldsymbol{\omega}_f$ and $(N-M)/2$ values in $\boldsymbol{\omega}_p$, even where a significant proportion of observations are missing.



(a) reconstruction *vs* signal

(b) full reconstruction *vs* signal

Fig. 6 For each test, two mean squared errors are calculated: (a) between the reconstruction and the underlying signal at the observation times, and (b) between the full reconstruction (i.e. for all N observation times) and the full signal. The mean values over each of the 100 test sets are shown for each proportion of missing values. For arrangement B, the performance of the reconstructions at the observed timestamps are indistinguishable from the comparison case.

5. REFERENCES

[1] J. Christmas, "Robust Bayesian spectral analysis," *Submitted to IEEE Transactions on Signal Processing*, 2013.

[2] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the IEEE*, vol. 90, no. 2, pp. 280–305, 2002, reprinted from Transactions of the AIEE, pp 617-644, Feb 1928.

[3] C.E. Shannon, "Communications in the presence of noise," *Proceedings of the IRE*, vol. 37, pp. 10–21, January 1949.

[4] T.J. Deeming, "Fourier analysis with unequally-spaced data," *Astrophysics and Space Science*, vol. 36, no. 1, pp. 137–158, 1974.

[5] H.-M. Adorph, "Interpolation of irregularly sampled data series - a survey," *Astronomical Data Analysis Software and Systems IV, ASP Conference Series*, vol. 77, pp. 460–463, 1995.

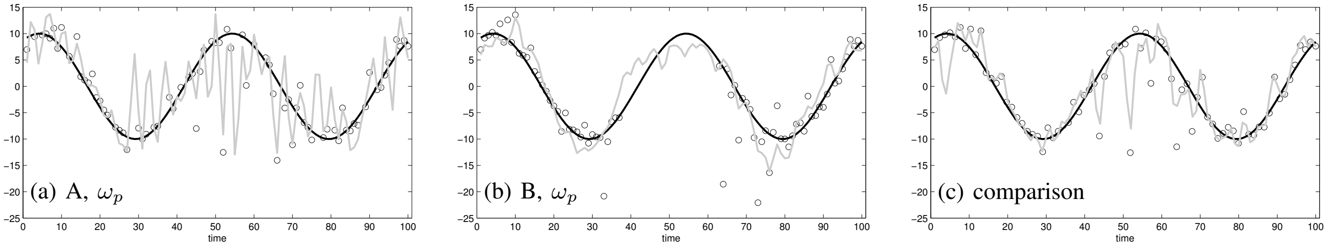


Fig. 7 With 25% missing values, the original signal (black line), observations (black circles) and the reconstruction of the signal (grey line) from the trained model. Subcaption format is as per figure 2.

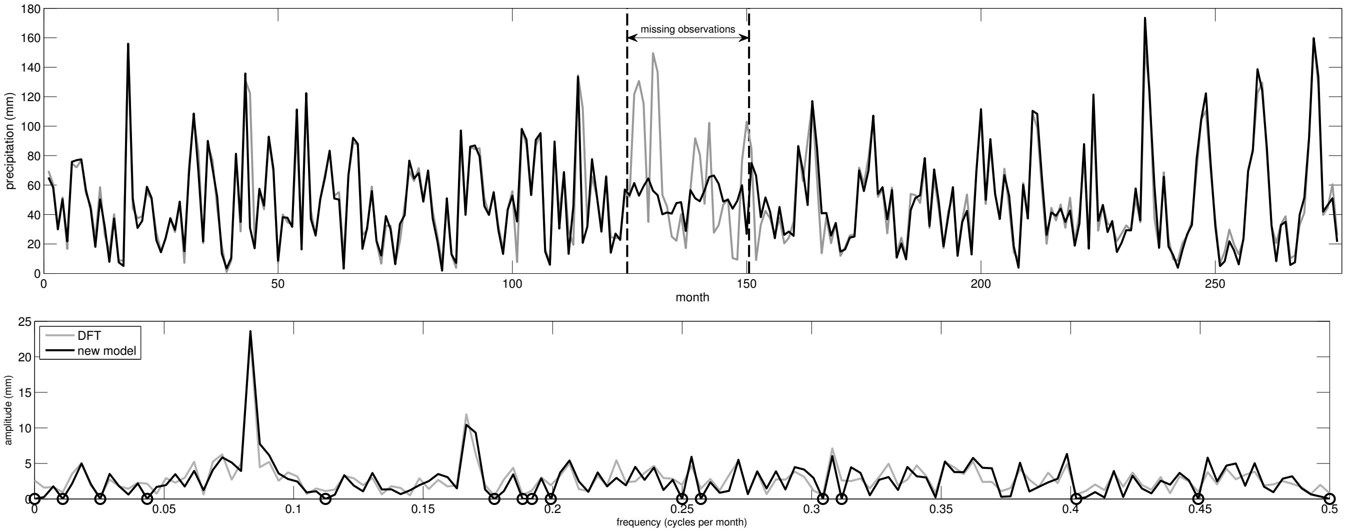


Fig. 8 The top plot shows the original monthly precipitation data (grey) overlaid with the model's reconstruction (black). The reconstruction of the missing values is poor, but that for the remaining observations it is generally good. A more significant result is shown below: the grey line is the amplitude spectrum from the DFT, while the black line is that from the model. Black circles identify where the amplitude is smaller than 0.1mm, i.e. where components have been "switched off" by automatic relevance determination. Note that the model is more parsimonious than DFT in its representation of the data. The maximum peak occurs at a period of 12 months, the annual precipitation cycle.

[6] M. Rauth, *Gridding Geophysical Potential Fields from Noisy Scattered Data*, Ph.D. thesis, University of Vienna, 1998.

[7] H. Knutsson and C.-F. Westin, "Normalized and differential convolution: Methods for interpolation and filtering of incomplete and uncertain data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1993, pp. 515–523.

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

[9] R.M. Neal, *Bayesian Learning for Neural Networks*, Ph.D. thesis, University of Toronto, Canada, 1995.

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

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

[12] 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.

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

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

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

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

[17] M. Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, University College London, 2003.

[18] (Norwegian) Meteorologisk institutt, "eKlima," `http://eklima.met.no`, last accessed 20th April 2013.