BaSAR—A tool in R for frequency detection

Emma Granqvist\textsuperscript{a,b,*}, Matthew Hartley\textsuperscript{a}, Richard J. Morris\textsuperscript{a}

\textsuperscript{a}Department of Computational & Systems Biology, The John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK
\textsuperscript{b}Department of Cell & Developmental Biology, The John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

A R T I C L E   I N F O

Article history:
Received 11 May 2012
Received in revised form 7 July 2012
Accepted 31 July 2012

Keywords:
Frequency
Software
Bayesian
R

A B S T R A C T

Many biological processes are periodic, for example cell cycle expression, circadian rhythms and calcium oscillations. However, measured time series from these processes are commonly short and noisy, and finding frequencies in such data can be challenging. Here we present BaSAR, Bayesian Spectrum Analysis in R, a package for extracting frequency information from time series data. The software uses advanced techniques of Bayesian inference that are well suited for handling typical biological data. The core functions are designed for detecting a single key frequency, without the need for data pre-processing such as detrending. The package is freely available at CRAN – The Comprehensive R Archive Network: http://cran.r-project.org/web/packages/BaSAR.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Periodic phenomena are common in biology, over scales that range from fractions of a second to many years and from the molecular to the population level (Murray, 2003; Goldbeter, 1997). In cell biology, self-sustaining oscillations arise in many processes, including cytoskeleton dynamics, cell cycle gene expression, bacterial movements, auditory hair bundles, spatial protein distribution such as Min in \textit{Escherichia coli}, and gene expression following circadian clock rhythms (Kruse and Jülicher, 2005).

Much work in theoretical systems biology has been devoted to the derivation and study of equations that give rise to sustained periodicity. Similarly, on the data analysis side, the development of sophisticated pattern recognition techniques for frequency detection has received much attention. A common technique, the Fourier transform, is based on the representation of any integrable function by a sum of sines and cosines. These periodic basis functions can be employed to probe the strength of frequencies in a given time series. However, the underlying assumptions for optimality of this approach include uniformly sampled, long, stationary, harmonic signals that have either no or white noise. These conditions are rarely met in biology. Therefore, techniques for detrending and noise reduction are common, but these convolute the signal, causing information loss (MacKay, 2003; Jaynes and Bretthorst, 2003). An alternative approach is to take all known effects into consideration but to integrate over the unknowns in the system. Bayesian techniques provide the appropriate framework for carrying out such marginalizations over joint probability distributions.

The advantages of using the Bayesian approach in data analysis has been documented in a number of cases (see e.g. Kotyk et al., 1992; Baldi and Long, 2001; Sivia and Skilling, 2006; Huelsenbeck and Ronquist, 2001). Here we describe our implementation of Bayesian Spectrum Analysis (Bretthorst, 1988) that offers automated background model selection and local, high-resolution frequency detection, without the need for pre-processing the data.

1.1. Approach

By placing the problem of frequency detection in the framework of Bayesian inference, the known and well-documented problems of Fourier analysis (see e.g. Gibbs, 1899; Bracewell, 1978; Jaynes and Bretthorst, 2003) can be overcome. This idea was pioneered by Bretthorst (1988) and was applied with success to nuclear magnetic resonance data. We recently developed this approach further by combining it with nested sampling to calculate the evidences for model comparison (Sivia and Skilling, 2006). We also introduced the generation of local frequency information (Granqvist et al., 2011). We employed this approach for the analysis of circadian clock data and calcium oscillations. The software that we present here builds on these developments and makes them easily accessible to a wider community.

We summarise the main points of the methods and refer to Bretthorst (1988) and Granqvist et al. (2011) for further details. In the following, we assume that the data, \( D \), are given at \( N \) discrete time points, \( t_i, D = \{ d(t_1), \ldots, d(t_N) \} \). There is no requirement for these data to be equally spaced. The data can be modelled as...
a sum of the underlying signal, $s(t_i)$, a background trend, $g(t_i)$, and the noise present in the system, $\epsilon(t_i)$,

$$d(t_i) = s(t_i) + g(t_i) + \epsilon(t_i).$$

The signal can be approximated by a linear combination of $m_0$ model functions, $\psi_j$, parameterized by the angular frequency $\omega$:

$$s(t_i) = \sum_{j=1}^{m_0} a_j \psi_j(\omega, t_i).$$

in which $a = \{a_1, \ldots, a_{m_0}\}$ are the expansion coefficients. In the provided software, the harmonic functions $\sin(\omega t_i)$ and $\cos(\omega t_i)$ are used as default model functions. Similarly, any background functions that are present can be approximated by a set of trend functions that are independent of $\omega$. Legendre polynomials are used for this purpose.

Following Bayes’ rule, the posterior probability distribution over the angular frequency $\omega$ for a given model, $H$, is given by

$$P(\omega|D, H) \propto \frac{P(D|\omega, H)P(\omega|H)}{P(D|H)}.$$  

where $P(D|\omega, H)$ is the likelihood, $P(\omega|H)$ the prior distribution over $\omega$, and $P(D|H)$ the evidence. Eq. (3), combined with the model presented in Eqs. (1) and (2), gives the posterior probability distribution over $\omega$. After assigning priors, integrating out amplitudes and noise levels and calculating the likelihood function, the posterior has been shown (Brethorst, 1988) to be proportional to

$$P(\omega|D, H, I) \propto \left[1 - \frac{m \bar{h}^2}{N\bar{a}^2}\right]^{(m-N)/2},$$

in which $\bar{\bar{h}}^2$ is the mean-square of the data projected onto the orthonormal model functions, $\phi_j$, $\bar{\bar{h}}^2 = (\sum_{j=1}^{m} h_j^2)/m$, where $h_j = \sum_{i=1}^{N} d\phi_j(\omega, t_i)$. This set of model functions can include background functions in addition to the signal model functions in Eq. (2) (Brethorst, 1988). Periodic data and a good model will result in a high probability peak in the posterior distribution at the appropriate frequency.

To compare different models, we use posterior model ratios (Brethorst, 1988). A given model of the signal, $H_i$, can be compared to an alternative model, $H_j$, by calculating their model ratios,

$$\frac{P(H_j|D, I)}{P(H_j)} = \frac{P(H_j|D, I)}{P(H_j|D, I)}.$$

where $H_i$ represents the model with fewer parameters. When the ratio is above one, the simpler model is preferred (MacKay, 2003).

To calculate the evidence, the normalising component in Eq. (3), we use the method of nested sampling. This is a Bayesian variant of the Markov Chain Monte Carlo (MCMC) algorithm that concentrates its sampling efforts in high likelihood regions of parameter space (Sivia and Skilling, 2006). The algorithm focuses on the computation of the evidence, whilst at the same time generating samples from the posterior distribution. By transforming the problem to likelihood space, high-dimensional integration can be reduced to a sorting task. Random samples are taken from the prior and by rejecting the point with the worst likelihood, the algorithm iteratively contracts the spread of samples around high likelihood regions of parameter space. One of the remaining samples is chosen to generate a new sample by taking MCMC steps around it in search of a higher likelihood value. Then the new samples are again sorted and the worst point rejected. This process is iterated until a given number of posterior samples have been generated.

2. Software description

We have implemented the above methodology in R (R Development Core Team, 2008), this package is suitable for biologists who wish to determine whether their data contain periodic features. The package can deal with data with background trends, cases where the period changes over time, or when the data have nonuniform sampling intervals. It is also well suited for cases where a high resolution of the frequency is needed. The current version only deals with frequency searches in one dimension. The key functions are listed in Table 1, and presented briefly below. A tutorial of the package can be found in the supplementary material.
2.1. Key functions

BaSAR.post returns a normalised posterior probability distribution over the chosen range of frequency (ω). This is invoked in the manner:

\[
\text{BaSAR.post}(\text{data}, \text{start}, \text{stop}, \text{nsamples}, \text{nbackg}, \text{tpoints})
\]

where data is the time series as a 1D vector, start-stop is the range of the period that is of interest (in seconds), nsamples is the number of samples that will be calculated from the posterior, and tpoints is the vector of time points when the data were sampled (in seconds). The interval between the time points does not need to be uniform. BaSAR.nest calculates the evidence using nested sampling. Direct comparison of evidences can be used to evaluate models.

BaSAR.modelratio is a model comparison method that uses model ratios to allow the user to compare two models with different background functions. This procedure has been automated in BaSAR.auto. For time series in which the dominant frequency changes over time, BaSAR.local can be used to calculate the local frequency by windowing.

The outputs from all functions are the posterior probability distribution over ω. If the user wants to see the results over period instead, there is a helper-function for this called BaSAR.plotperiod.

2.2. Parameters and priors

The core model parameters that the user needs to specify are the period interval of interest, the number of samples over this interval, and the number of background functions. The period interval of interest might vary widely depending on the biological system, and must be decided from case to case. The choice of number of samples is a trade-off between frequency resolution and time of computation, but 100–500 samples is often a good starting point. The number of background functions needed can be estimated by using the automated model selection functions such as BaSAR.auto. The total number of model functions, \(m\) in Eq. (2), is two for the basic sine and cosine, plus any added background functions.

Model parameters that are not of principal interest, such as amplitudes and noise level, are assigned suitable priors (Jeffreys prior for noise levels, uniform prior for amplitudes) and then integrated over following 

\text{Brethorst (1988). At the moment they are not part of the software output.}
3. Application examples

Test cases were presented in Granqvist et al. (2011) on both real and simulated data. Here we introduce two further examples to illustrate the global and local frequency detection ability of BaSAR.

3.1. Cell cycle genes

We demonstrate BaSAR on gene expression data from cell cycle phases in fission yeast, *Schizosaccharomyces pombe*, taken from Rustici et al. (2004). These time series are short, and have long intervals between data points. Many genes peak in expression during specific phases of the cell cycle, resulting in cyclic expression profiles (Breeden, 2003). Fig. 1A shows one such gene's expression profile, *cdc15*. This gene is known to be expressed periodically and peaks in M phase, with a characterized role in the actin ring formation of cell division (Fankhauser et al., 1995; Zilahi et al., 2000). A clear peak can be seen in the posterior distribution output from *BaSAR_post* (Fig. 1B). It should be noted that only one time series for the gene is included here (from Elutriation 1), not the complete data set that the original study used. A Fourier period spectrum is shown for comparison (Fig. 1C). Whilst the Fourier spectrum performs well, the accuracy is limited by the sampling that is a function of the given time points, leading to a slight shift in frequency peak in this example. Furthermore, Fourier spectra require uniformly sampled data points. A simulated time series consisting of uniform random numbers between 1 and 2 is shown, along with its BaSAR results and its Fourier spectrum (Fig. 1D–F), to demonstrate a posterior distribution with no evidence for periodicity.

3.2. Calcium oscillations

In the symbiosis between legumes and nitrogen-fixing bacteria (rhizobia), calcium oscillations are induced in the plant root cells during the initial signalling stages. These calcium signals occur in and around the nucleus, and are essential for the symbiosis to establish (Oldroyd and Downie, 2006). The oscillations often start with a rapid period (of approximately 1 min) and slow down over time as the signal continues. The data also contain background trends due to fluorescence bleaching. An example time series is shown in Fig. 2A. In Fig. 2B, the result of analysing this time series with *BaSAR_local* is shown, demonstrating that this package copes with the background trend as well as shows the signal period changes over time, without need for pre-processing such as detrending. Good results can be achieved by a windowed Fourier spectra (Fig. 2C) but only after detrending, which was done using a moving average (Brockwell and Davis, 2002). BaSAR delivers superior results (Fig. 2B) and without the need for data pre-processing.

Acknowledgements

EG acknowledges PhD funding from the John Innes Foundation. RJM and MH are grateful for support from the BBSRC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biosystems.2012.07.004.

References


