

Bayesian methods applied to LISA source identification

Graham Woan, University of Glasgow

Outline of talk

- Part I: why use Bayesian methods?
 - Fundamentals
 - Usefulness (and inevitability) of priors
 - Model selection and Occam factors
- Part II: application to LISA data analysis
 - Toy problem
 - Markov Chain Monte Carlo methods
 - Extensions and global analysis

Why Bayesian methods?

Orthodox statistical methods are concerned solely with deductions following experiments with populations:



"The trouble is that what we [statisticians] call modern statistics was developed under strong pressure on the part of biologists. As a result, there is practically nothing done by us which is directly applicable to problems of astronomy."

Jerzy Neyman, founder of frequentist hypothesis testing.

Why Bayesian methods?

- Bayesian methods are the algebra of inductive reasoning
- probability is used as a measure of a state of knowledge, or “degree of belief”
- extension of boolean logic to work with values between 0 and 1, representing the above (though it’s not fuzzy logic, which is not a probabilistic framework)

Why Bayesian methods?

Deductive reasoning:

Given: “all pulsars are neutron stars” and “our target is a pulsar”

Deduce: “our target is a neutron star”

Inductive reasoning:

Given: “all pulsars are neutron stars” and “our target is a neutron star”

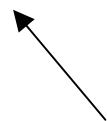
Infer: “it is *more probable* our target is a pulsar”

(how much more depends on the alternatives, within our world-view)

Why Bayesian methods?

Bayesian methods explore the joint probability space of data and hypotheses within some global model, quantifying their joint uncertainty and consistency as a scalar function:

$$p(\text{data, hypotheses} \mid \text{world view})$$



means "given"

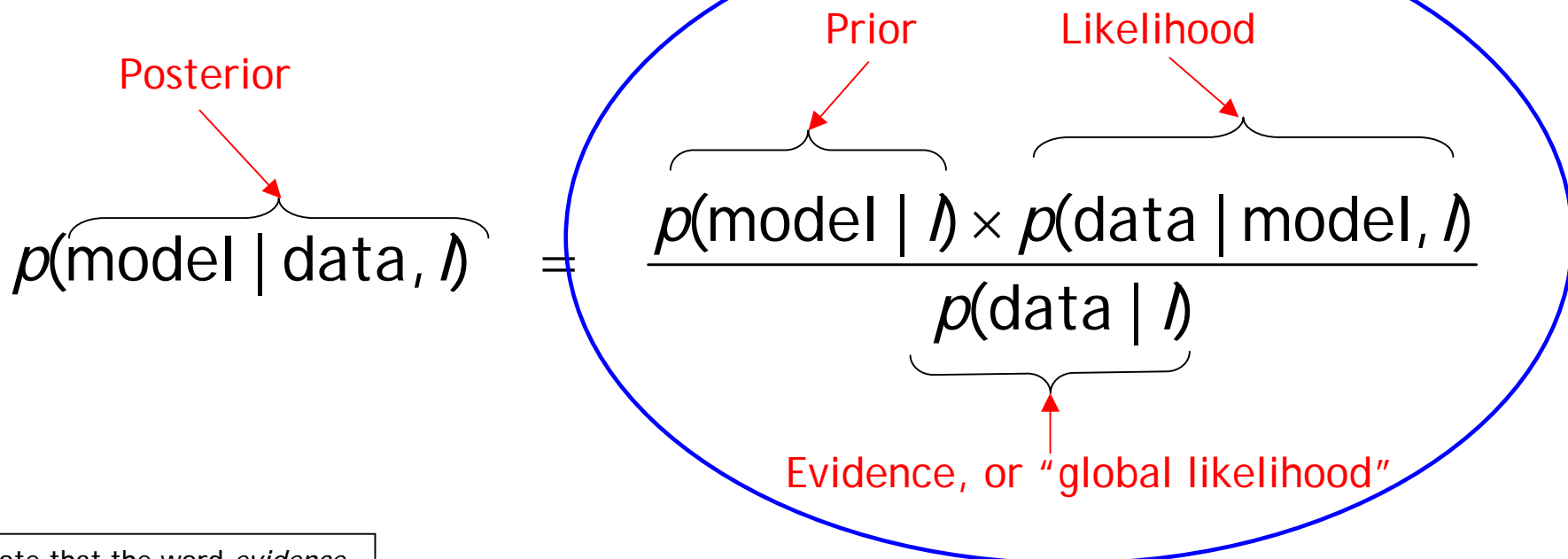
Polya, Cox and Jaynes showed that there is only one algebra consistent with this idea (and some further, very reasonable, constraints), which leads to (amongst other things) the *product rule*:

$$p(d, h \mid w) = p(d \mid w)p(h \mid d, w) = p(h \mid w)p(d \mid h, w)$$

Why Bayesian methods?

This leads to Bayes' theorem: the appropriate rule for updating our degree of belief (in one of several hypotheses within some world view) when we have new data:

$$p(h | d, w) = \frac{p(h | w) \times p(d | h, w)}{p(d | w)}$$



[note that the word *evidence* is sometimes used for something else (the 'log odds'). We will stick to the $p(d|I)$ definition here.]

We can usually calculate all these terms

Marginalisation

- We can also deduce the *marginal probabilities*. If X and Y are propositions that can take on values drawn from $X \in \{x_1, x_2, \dots, x_n\}$, and $Y \in \{y_1, y_2, \dots, y_m\}$, then

$$p(x_i) = p(x_i) \underbrace{\sum_{j=1..m} p(y_j | x_i)}_{=1} = \sum_{j=1..m} p(x_i) p(y_j | x_i) = \sum_{j=1..m} p(x_i, y_j)$$

this gives use the probability of X when we don't care about Y . In these circumstances, Y is known as a *nuisance parameter*.

- All these relationships can be smoothly extended from discrete probabilities to probability densities, e.g.

$$p(x) = \int p(x, y) dy$$

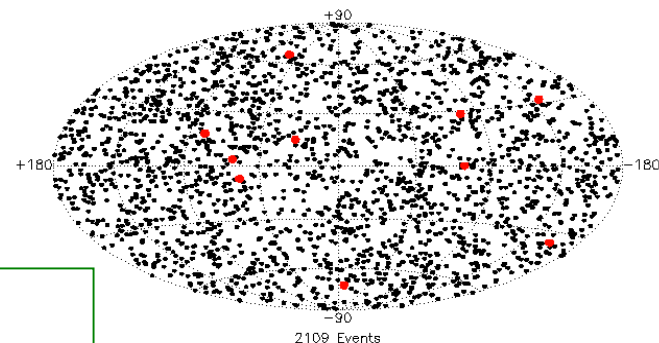
where " $p(y)dy$ " is the probability that y lies in the range y to $y+dy$.

Bayesian parameter estimation

Zeroth-order astronomical example:

Flux density of a GRB

Take Gamma Ray Bursts to be equally luminous events, distributed homogeneously in the Universe. We see **three** gamma ray photons from a GRB in an interval of **1 s**. What is the flux of the source, F ?



• The seat-of-the-pants answer is $F=3$ photons/s, with an uncertainty of about $\sqrt{3}$, but we can do better than that by including our prior information on luminosity and homogeneity (similar to correcting for a Malmquist bias). Call this background information I :

Homogeneity implies that the probability the source is in any particular volume of space is proportional to the volume, so the prior probability that the source is in a thin shell of radius r is

$$p(r | I) dr \propto 4\pi r^2 dr$$

Bayesian parameter estimation

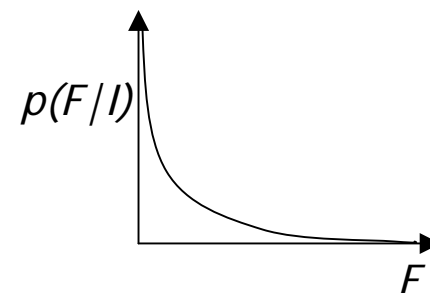
- But the sources have a fixed luminosity, L , so r and F are directly related by

$$F = \frac{L}{4\pi r^2}$$

hence $\frac{dF}{dr} \propto -r^{-3}$

- The prior on F is therefore

$$p(F | l) \propto p(r | l) \left| \frac{dr}{dF} \right| \propto F^{-5/2}$$



Interpretation: low flux sources are intrinsically more probable, as there is more space for them to sit in.

- We now apply Bayes' theorem to determine the posterior for F after seeing n photons:

$$p(F | n, l) \propto p(F | l) p(n | F, l)$$

Bayesian parameter estimation

- The Likelihood for F comes from the Poisson nature of photons:

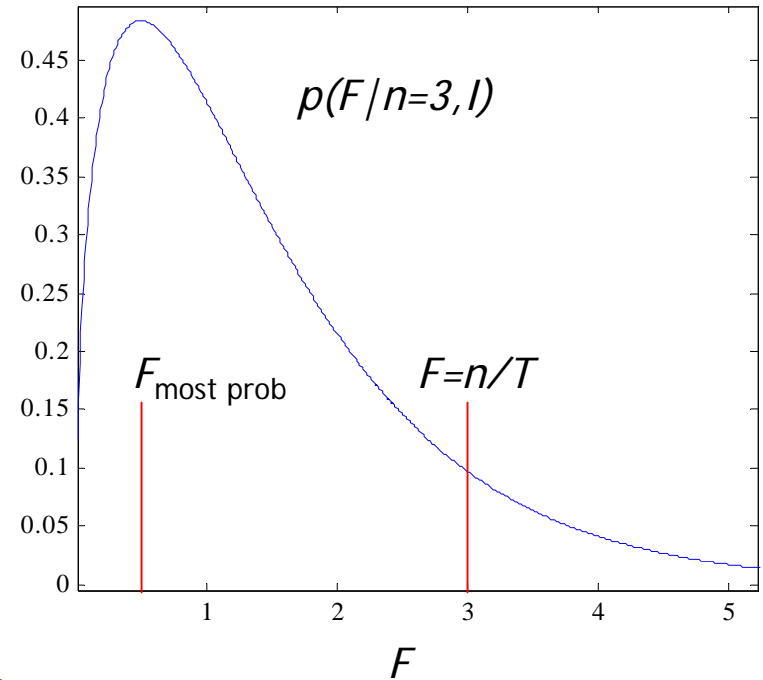
$$p(n | F, l) = F^n \exp(-F) / n!$$

so finally,

$$p(F | n, l) \propto F^{n-5/2} \exp(-F)$$

For $n=3$ we get this \longrightarrow

with the most probable value of F equalling 0.5 photons/sec.



- Clearly it is more probable this is a distant source from which we have seen an unusually high number of photons than it is an unusually nearby source from which we have seen an expected number of photons. (The most probable value of F is $n-5/2$, approaching n for $n \gg 1$)

Bayesian hypothesis testing

- For hypothesis spaces that are not finite, or which are hard to fully explore, we can work with the odds ratio of two competing hypotheses. This can be divided into the **prior odds** and the **Bayes' factor**

$$O_{12} = \frac{\text{prob}(H_1 | d, I)}{\text{prob}(H_2 | d, I)} = \underbrace{\frac{\text{prob}(H_1 | I)}{\text{prob}(H_2 | I)}}_{\text{prior odds}} \times \underbrace{\frac{\text{prob}(d | H_1, I)}{\text{prob}(d | H_2, I)}}_{\text{Bayes' factor}}$$

- The prior odds simply express our prior preference for H_1 over H_2 , and is set to 1 if you are indifferent.
- The Bayes' factor is just the ratio of the evidences. For a hypothesis that depends on a parameter a , we know that

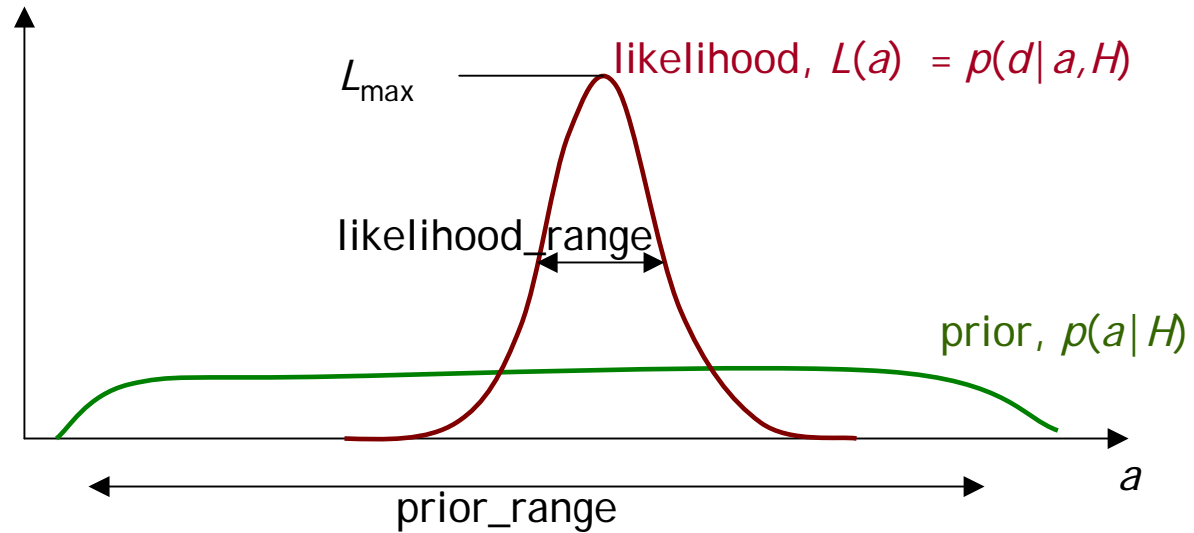
$$p(a | d, H_1, I) = \frac{p(a | H_1, I)p(d | a, H_1, I)}{p(d | H_1, I)}$$

so the evidence is simply the joint probability of the parameter(s) and the data, marginalised over all hypothesis parameter values:

$$p(d | H_1, I) = \int p(a | H_1, I)p(d | a, H_1, I)da$$

Bayesian hypothesis testing

- To look at this a bit more generally, we can split the evidence into two approximate parts, the maximum of the likelihood and an “Occam factor”:



$$p(d | H) = \int p(a | H) p(d | a, H) da \approx L_{\max} \underbrace{\frac{\text{likelihood_range}}{\text{prior_range}}}_{\text{the 'Occam factor'}}$$

i.e., **evidence = maximum likelihood x Occam factor**

The Occam factor penalises models that include wasted parameter space, even if they show a good ML fit.

Occam's Razor



William of Occam
(1288 - 1348)

"Frustra fit per plura, quod fieri potest per pauciora."

"It is vain to do with more what can be done with less."

Everything else being equal, we favour models which are *simple*.

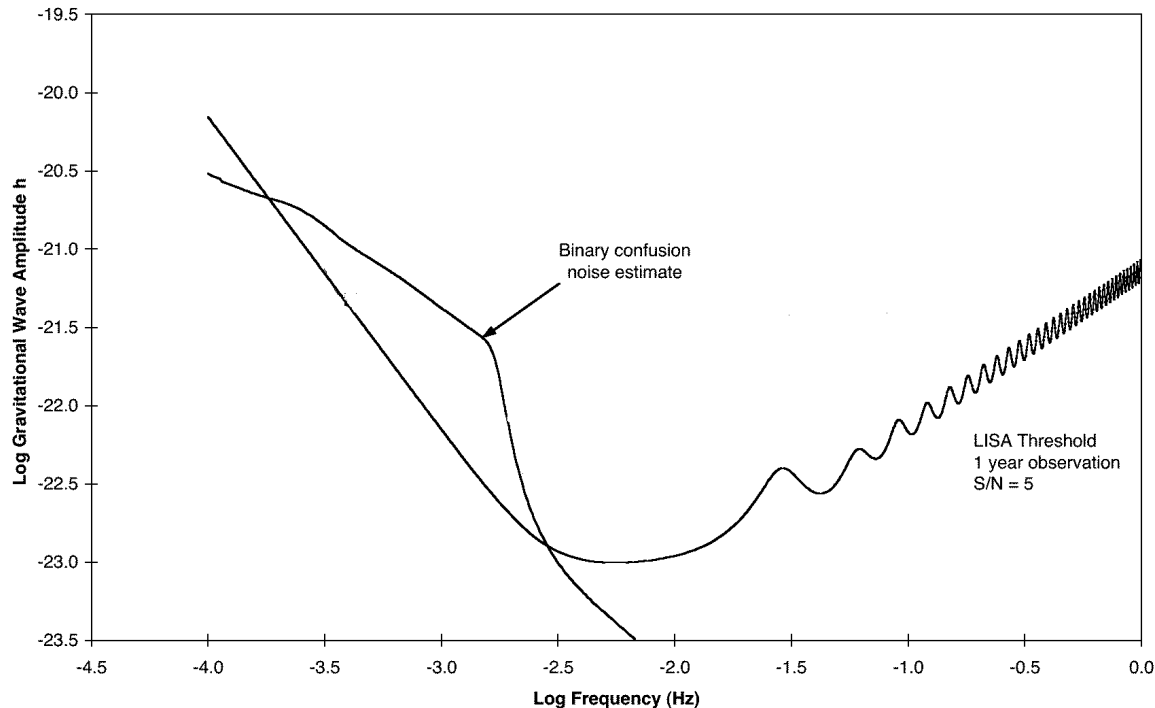
Occam factors penalise wasteful models with priors that allow lots of parameter values excluded by the data.

Part II

application to LISA data analysis

LISA source identification

- This has implications for the analysis of LISA data, which is expected to contain many (perhaps 50,000) signals from white dwarf binaries. The data will contain resolvable binaries and binaries that just contribute to the overall noise (either because they are faint or because their frequencies are too close together).



Bayes can sort these out without having to introduce ad hoc acceptance and rejection criteria, and without needing to know the “true noise level” (whatever that means!).

Things that are not generally true

- *“A time series of length T has a frequency resolution of $1/T$.”*

Frequency resolution also depends on signal-to-noise ratio. We know the period of PSR 1913+16 to 1e-13 Hz, but haven't been observing it for 3e5 years.

In fact frequency resolution is $\approx \frac{1}{T \times \text{snr}}$.

- *“you can subtract sources piece-wise from data.”*

Only true if the source signals are orthogonal over the observation period.

- *“frequency confusion sets a fundamental limit for low-frequency LISA.”*

This limit is set by *parameter* confusion, which includes sky location and other relevant parameters (with a precision dependent on snr).

LISA source identification

- Toy (zeroth-order LISA) problem:

You are given a time series of 1000 data points comprising a number of sinusoids embedded in gaussian noise. Determine the number of sinusoids, their amplitudes, phases and frequencies and the standard deviation of the noise.

- We could think of this as comparing hypotheses H_m that there are m sinusoids in the data, with m ranging from 0 to m_{\max} . Equivalently, we could consider this a parameter fitting problem, with m an unknown parameter within the global model.

signal
$$s^{(m)}(t_j, \mathbf{a}_m) = \sum_{i=1}^m \left[A_i^{(m)} \cos(2\pi f_i^{(m)} t_j) + B_i^{(m)} \sin(2\pi f_i^{(m)} t_j) \right],$$

parameterised by
$$\mathbf{a}_m = [A_1^{(m)}, B_1^{(m)}, f_1^{(m)}, \dots, A_m^{(m)}, B_m^{(m)}, f_m^{(m)}, \sigma_m^2]$$

giving data
$$d_j = s^{(m)}(t_j, \mathbf{a}_m) + \epsilon_j^{(m)}, \quad \text{for } j = 1, \dots, N$$

and a likelihood
$$p(\mathbf{d}|m, \mathbf{a}_m) \propto \frac{1}{\sigma_m^N} \exp \left\{ -\frac{1}{2\sigma_m^2} \sum_{j=1}^N \left[d_j - s^{(m)}(t_j, \mathbf{a}_m) \right]^2 \right\}$$

LISA source identification

- With suitably chosen priors on m and \mathbf{a}_m we can write down the full posterior pdf of the model

$$p(m, \mathbf{a}_m | \mathbf{d}) = \frac{p(m, \mathbf{a}_m) p(\mathbf{d} | m, \mathbf{a}_m)}{p(\mathbf{d})}$$

But this is $(3m+2)$ dimensional, with $m \sim 100$ in our toy problem, so the direct evaluation of marginal pdfs for, say, m or σ_m or to extract the pdf of a component amplitude, is unfeasible.

- In a toy problem some of these integrals may be analytic, but they are not strictly analytic in the real LISA problem.
- The rest of this talk is *one way* to explore this space using a modified Markov Chain Monte Carlo technique...

Application to the LISA confusion problem

PHYSICAL REVIEW D **72**, 022001 (2005)

Bayesian modeling of source confusion in LISA data

Richard Umstätter,^{1,*} Nelson Christensen,^{2,†} Martin Hendry,^{3,‡} Renate Meyer,^{1,§} Vimal Simha,^{3,||} John Veitch,^{3,¶}
Sarah Vigeland,^{2,**} and Graham Woan^{3,††}

¹*Department of Statistics, University of Auckland, Auckland, New Zealand*

²*Physics and Astronomy, Carleton College, Northfield, Minnesota 55057, USA*

³*Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*

(Received 9 June 2005; published 7 July 2005)

One of the greatest data analysis challenges for the Laser Interferometer Space Antenna (LISA) is the need to account for a large number of gravitational wave signals from compact binary systems expected to be present in the data. We introduce the basis of a Bayesian method that we believe can address this challenge and demonstrate its effectiveness on a simplified problem involving 100 synthetic sinusoidal signals in noise. We use a reversible jump Markov chain Monte Carlo technique to infer simultaneously the number of signals present, the parameters of each identified signal, and the noise level. Our approach therefore tackles the detection and parameter estimation problems simultaneously, without the need to evaluate formal model selection criteria, such as the Akaike Information Criterion or explicit Bayes factors. The method does not require a stopping criterion to determine the number of signals and produces results which compare very favorably with classical spectral techniques.

DOI: [10.1103/PhysRevD.72.022001](https://doi.org/10.1103/PhysRevD.72.022001)

PACS numbers: 04.80.Nn, 02.70.Rr, 06.20.Dk

Markov Chain Monte Carlo methods

- We need to be able to evaluate marginal integrals of the form

$$p(x) = \int \dots \int p(x, x_1, \dots, x_n) dx_1 \dots dx_n$$

- The approach is to sample in the (x, x_1, \dots, x_n) space so that the **density of samples reflects the posterior probability** $p(x, x_1, \dots, x_n)$.
- MCMC algorithms perform random walks in the parameter space so that the probability of being in a hypervolume dV is $\propto p(x, x_1, \dots, x_n) dV$.
- The random walk is a Markov chain: the transition probability of making a step depends on the proposed location, a'_k and the current location a_k

Metropolis-Hastings Algorithm

- We want to explore $p(a)$. Let the current location be a_t .

- 1) Choose a candidate state a'_t using a *proposal distribution* $q(a'_t | a_t)$.

- 2) Compute the Metropolis ratio

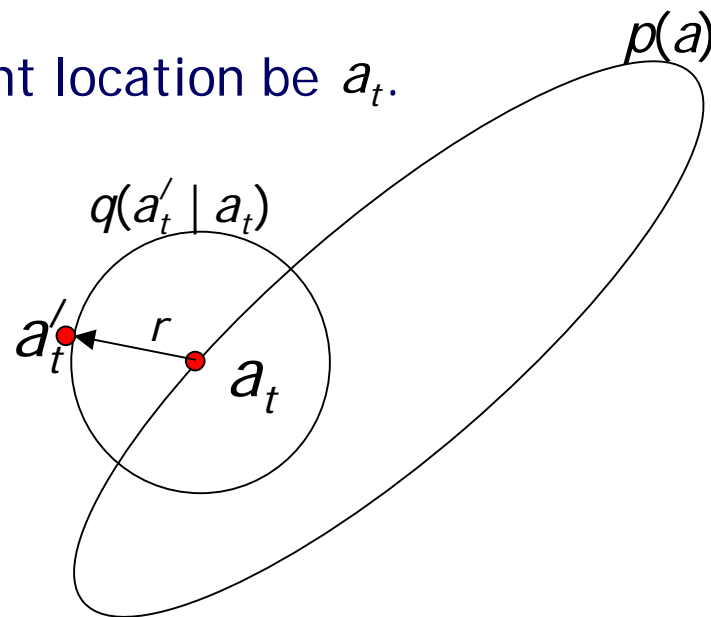
$$R = \frac{p(a'_t)q(a_t | a'_t)}{p(a_t)q(a'_t | a_t)}$$

- 3) If $R > 1$ then make the step (i.e., $a_{t+1} = a'_t$)
if $R < 1$ then make the step with probability R , otherwise set $a_{t+1} = a_t$,
so that the location is repeated.

i.e., make the step with an *acceptance probability*

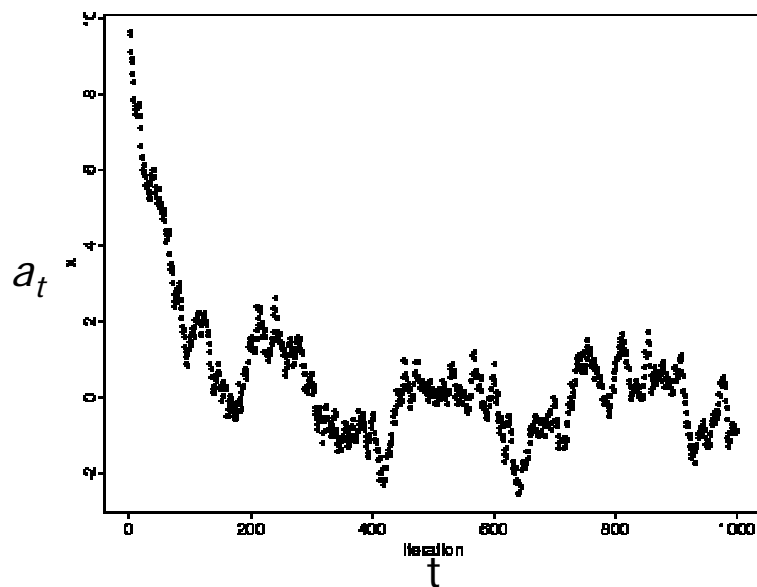
$$\alpha(a'_t | a_t) = \min \left[1, \frac{p(a'_t)q(a_t | a'_t)}{p(a_t)q(a'_t | a_t)} \right].$$

- 4) Choose the next candidate based on the (new) current position...

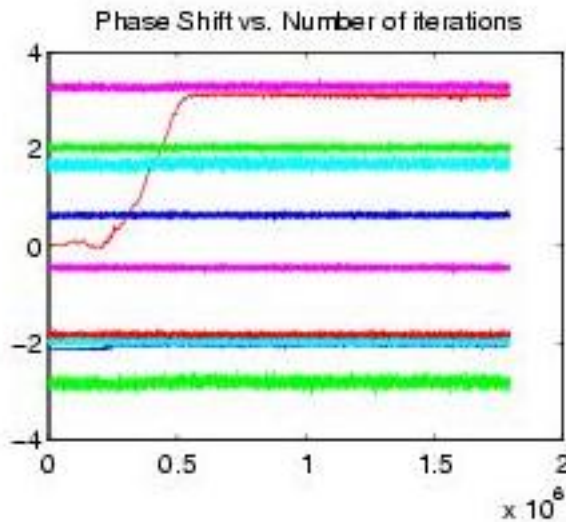
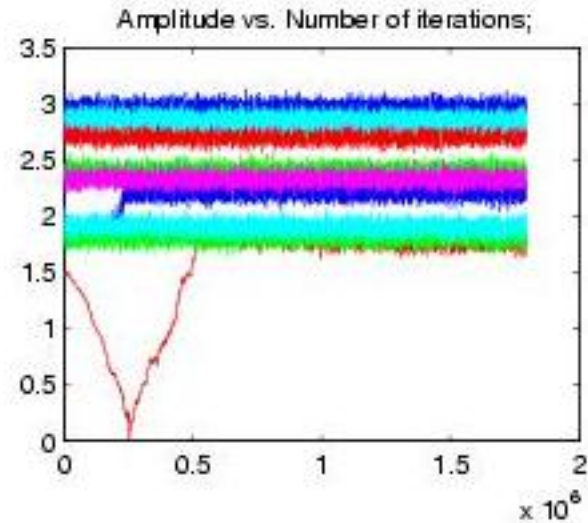
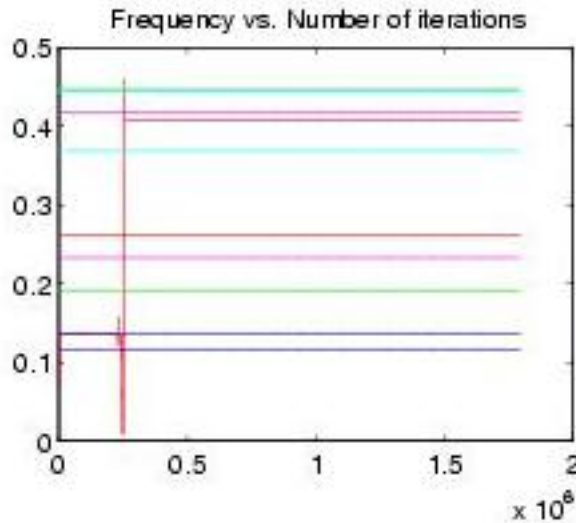


Metropolis-Hastings Algorithm

- $\{a_t\}$ form a Markov chain drawn from $p(a)$, so a histogram of $\{a_t\}$, or any of its components, approximates the (joint) pdf of those components.
- The form of the acceptance probability guarantees reversibility even for proposal distributions that are asymmetric.
- There is a burn-in period before the equilibrium distribution is reached:



Multiple chains



Histograms of the post convergence chains provide estimates of the posterior pdfs

Reversible Jump MCMC

- Trans-dimensional moves (changing m) cannot be performed in conventional MCMC. We need to make jumps from k to k' dimensions $k' \in \{k - 1, k + 1\}$
- Reversibility is guaranteed if the acceptance probability for an upward transition is

$$\alpha_{k \mapsto k'}(\mathbf{a}'_{k'} | \mathbf{a}_k) = \min \left\{ 1, \frac{p(\mathbf{a}'_{k'}, k') p(\mathbf{d} | \mathbf{a}'_{k'}, k') p_{k \mapsto k'}}{p(\mathbf{a}_k, k) p(\mathbf{d} | \mathbf{a}_k, k) q(\mathbf{r}) p_{k' \mapsto k}} |J_{k \mapsto k'}| \right\}$$

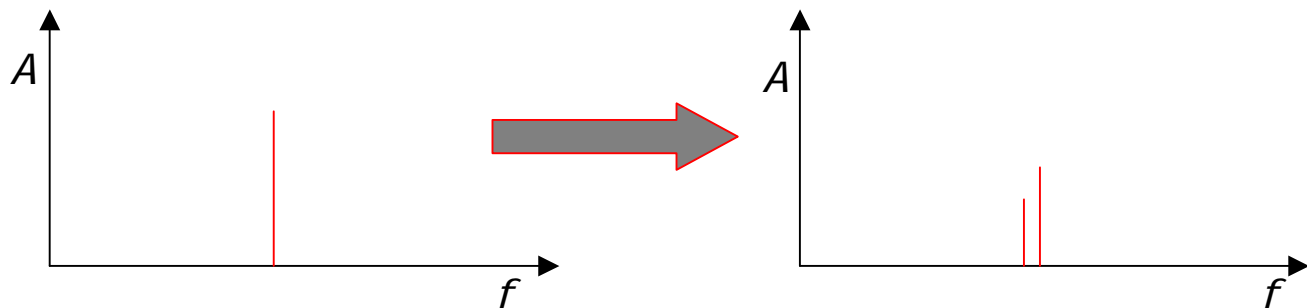
where $|J_{k \mapsto k'}| = \left| \frac{\partial \mathbf{t}(\mathbf{a}'_{k'}, \mathbf{r})}{\partial (\mathbf{a}_k, \mathbf{r})} \right|$ is the Jacobian determinant of the transformation of the old parameters [and proposal random vector \mathbf{r} drawn from $q(\mathbf{r})$] to the new set of parameters, i.e. $\mathbf{a}'_{k'} = \mathbf{t}(\mathbf{a}_k, \mathbf{r})$.

- We use two sorts of trans-dimensional moves:
 - 'split and merge' involving adjacent signals
 - 'birth and death' involving single signals

Trans-dimensional split-and-merge transitions

- A split transition takes the parameter subvector $\mathbf{a}_{(i)} = (A_i^{(k)}, B_i^{(k)}, f_i^{(k)})$ from a_k and splits it into two components of similar frequency but about half the amplitude:

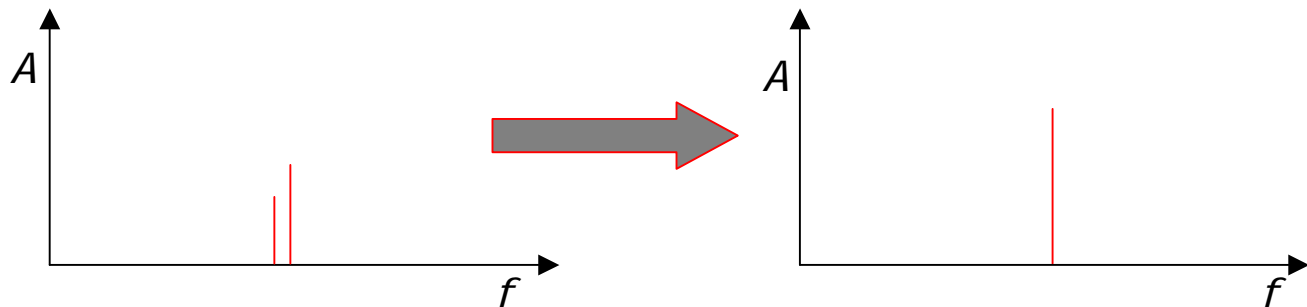
$$\mathbf{t}_{k \rightarrow k'}(\mathbf{a}_{(i)}, \mathbf{r}) = \begin{pmatrix} \frac{1}{2}A_i^{(k)} + r_A \\ \frac{1}{2}B_i^{(k)} + r_B \\ f_i^{(k)} + r_f \\ \frac{1}{2}A_i^{(k)} - r_A \\ \frac{1}{2}B_i^{(k)} - r_B \\ f_i^{(k)} - r_f \end{pmatrix} = \begin{pmatrix} A_{i_1}^{(k')} \\ B_{i_1}^{(k')} \\ f_{i_1}^{(k')} \\ A_{i_2}^{(k')} \\ B_{i_2}^{(k')} \\ f_{i_2}^{(k')} \end{pmatrix}$$



Trans-dimensional split-and-merge transitions

- A merge transition takes two parameter subvectors and merges them to their mean:

$$t_{k' \mapsto k}(a'_{(i_1)}, a'_{(i_2)}) = \begin{pmatrix} A_{i_1}^{(k')} + A_{i_2}^{(k')} \\ B_{i_1}^{(k')} + B_{i_2}^{(k')} \\ \frac{1}{2}f_{i_1}^{(k')} + \frac{1}{2}f_{i_2}^{(k')} \\ \frac{1}{2}(A_{i_1}^{(k')} - A_{i_2}^{(k')}) \\ \frac{1}{2}(B_{i_1}^{(k')} - B_{i_2}^{(k')}) \\ \frac{1}{2}(f_{i_1}^{(k')} - f_{i_2}^{(k')}) \end{pmatrix} = \begin{pmatrix} A_i^{(k)} \\ B_i^{(k)} \\ f_i^{(k)} \\ r_A^{(k)} \\ r_B^{(k)} \\ r_f^{(k)} \end{pmatrix}$$



Delayed rejection

- Sampling can be improved (beyond Metropolis Hastings) if a second proposal is made following, and based on, an initial rejected proposal. The initial proposal is only rejected if this second proposal is also rejected.
- Acceptance probability of the second stage has to be chosen to preserve reversibility (detailed balance):

acceptance probability for 1st stage:

$$\alpha_1(\mathbf{a}'|\mathbf{a}) = \min \left[1, \frac{p(\mathbf{a}')p(\mathbf{d}|\mathbf{a}')q_1(\mathbf{a}|\mathbf{a}')}{p(\mathbf{a})p(\mathbf{d}|\mathbf{a})q_1(\mathbf{a}'|\mathbf{a})} \right]$$

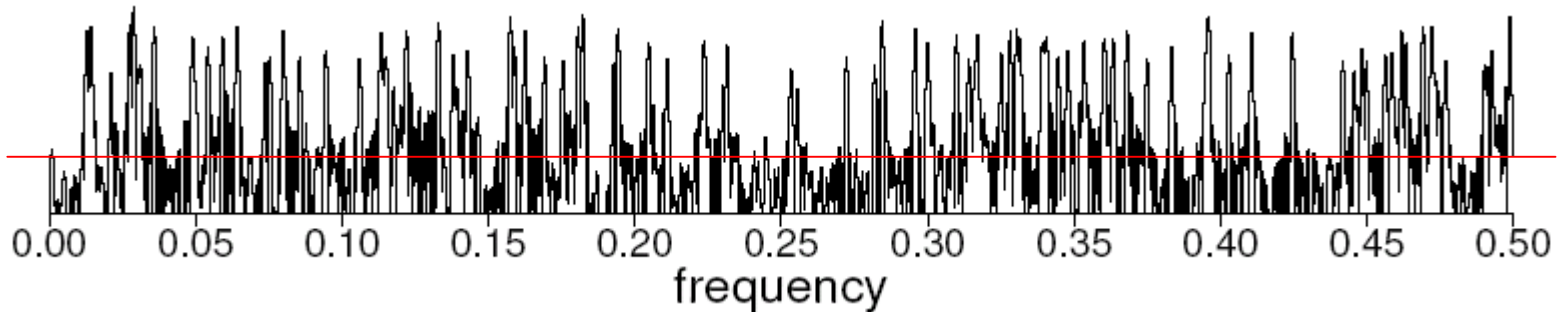
and for the 2nd stage:

$$\alpha_2(\mathbf{a}''|\mathbf{a}', \mathbf{a}) = \min \left\{ 1, \frac{p(\mathbf{a}'')p(\mathbf{d}|\mathbf{a}'')q_1(\mathbf{a}'|\mathbf{a}'')q_2(\mathbf{a}|\mathbf{a}', \mathbf{a}'')[1 - \alpha_1(\mathbf{a}'|\mathbf{a}'')]}{p(\mathbf{a})p(\mathbf{d}|\mathbf{a})q_1(\mathbf{a}'|\mathbf{a})q_2(\mathbf{a}''|\mathbf{a}, \mathbf{a}') [1 - \alpha_1(\mathbf{a}'|\mathbf{a})]} \right\}$$

➔ Delayed Rejection Reversible Jump Markov Chain Monte Carlo method 'DRRJMCMC' Green & Mira (2001) Biometrika 88 1035-1053.

Initial values

- A good initial choice of parameters greatly decreases the length of the 'burn-in' period to reach convergence (equilibrium). For simplicity we use a thresholded FFT:



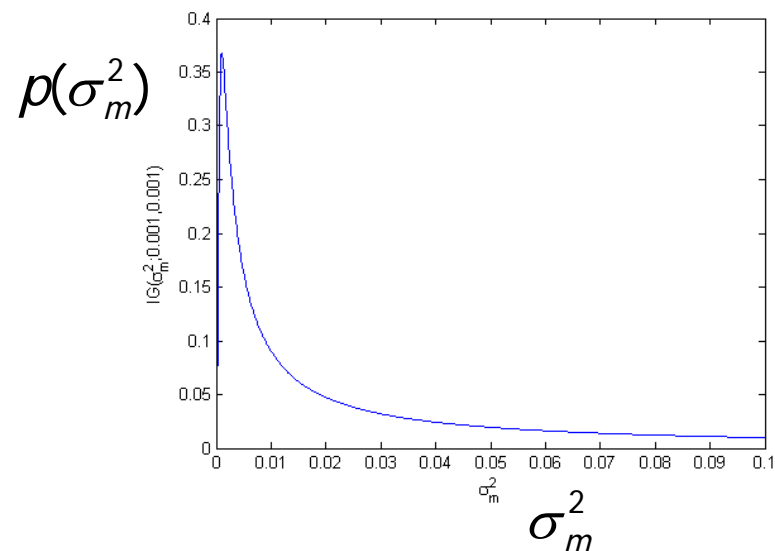
- The threshold is set low, as it is easier to destroy bad signals than to create good ones.

Simulations

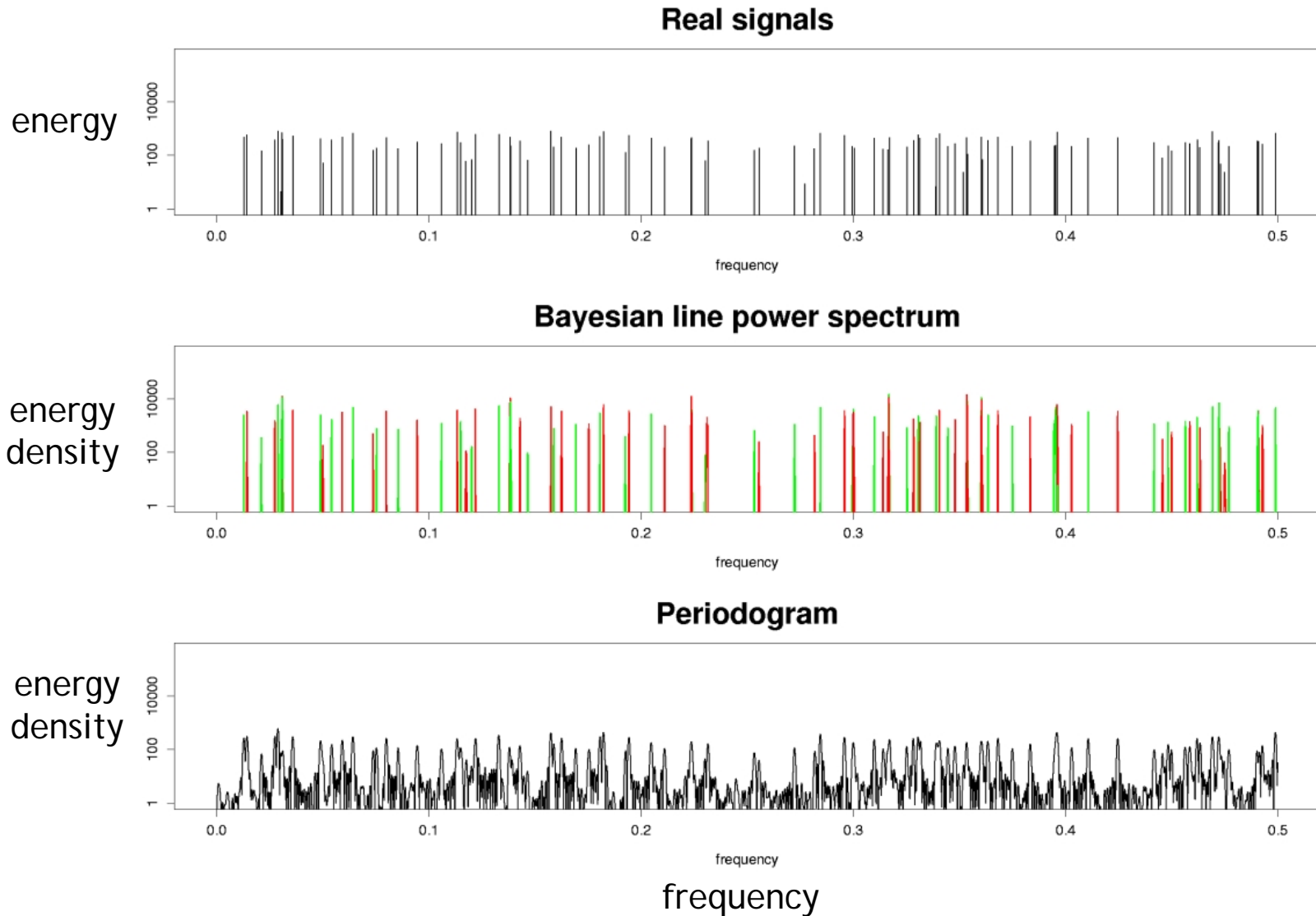
- 1000 time samples with Gaussian noise
- 100 embedded sinusoids of form $A\cos 2\pi ft + B\sin 2\pi ft$
- A s and B s chosen randomly in $[-1 \dots 1]$
- f s chosen randomly in $[0 \dots 0.5]$
- Noise $\sigma = 1$

Priors

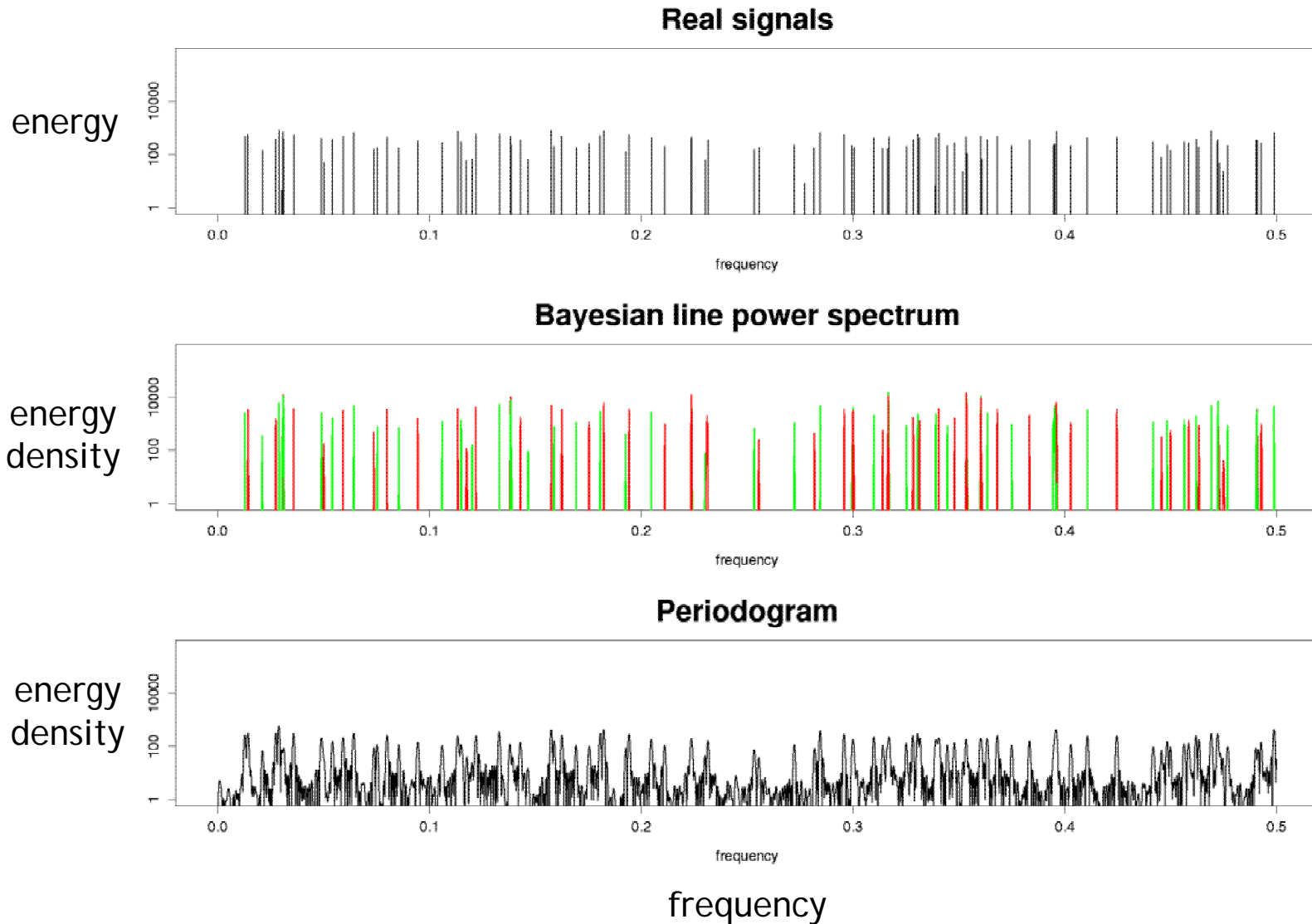
- A_m, B_m uniform over $[-5 \dots 5]$
- f_m uniform over $[0 \dots 0.5]$
- σ_m^2 has a standard vague inverse-gamma prior $IG(\sigma_m^2; 0.001, 0.001)$



results teaser (spectral density)

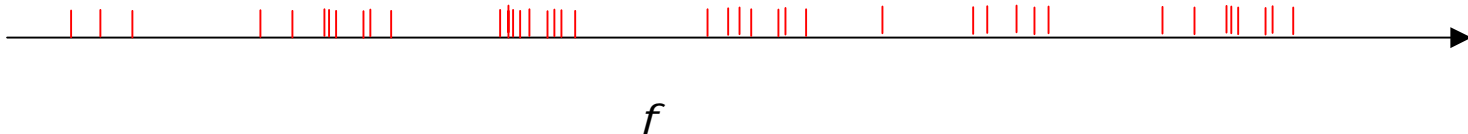


results teaser (spectral density)

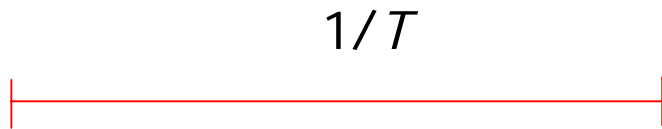
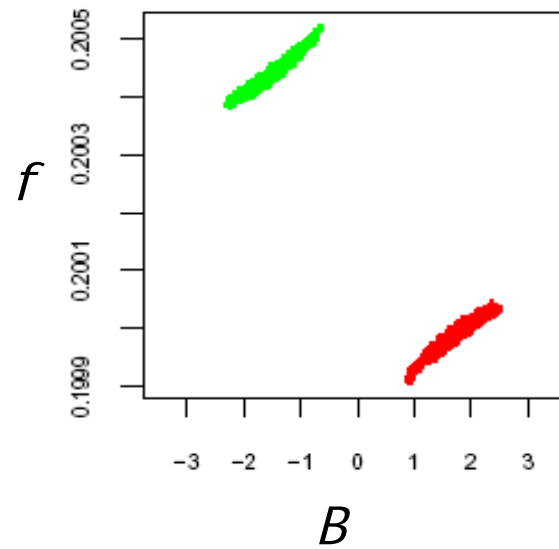
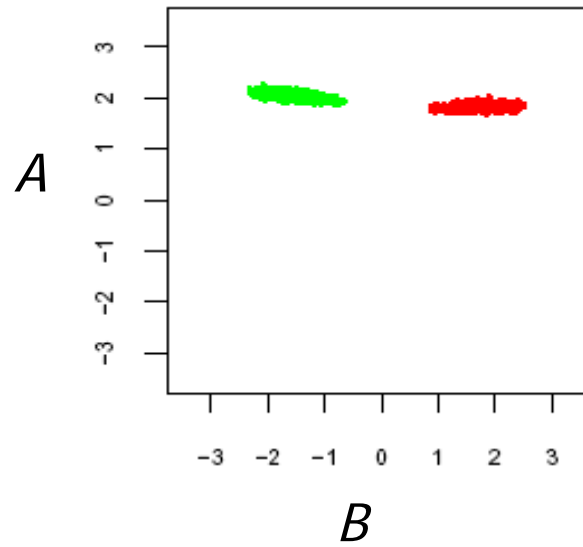
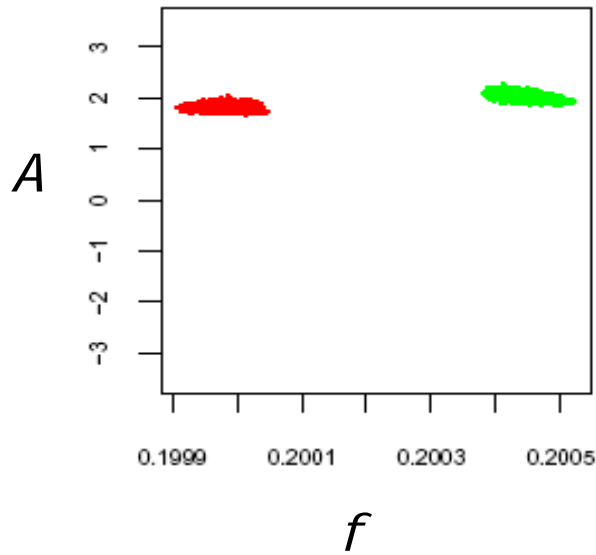


Label-switching

- As set up, the posterior is invariant under signal renumbering – we have not specified what we mean by ‘signal 1’.
- Break the symmetry by ordering in frequency:
 1. Fix m at the most probable number of signals, containing n MCMC steps.
 2. Order the nm MCMC parameter triples (A, B, \hat{f}) in frequency.
 3. Perform a rough density estimate to divide the samples into m blocks.
 4. Perform an iterative minimum variance cluster analysis on these blocks.
 5. Merge clusters to get exactly m signals.
 6. Tag the parameter triples in each cluster.

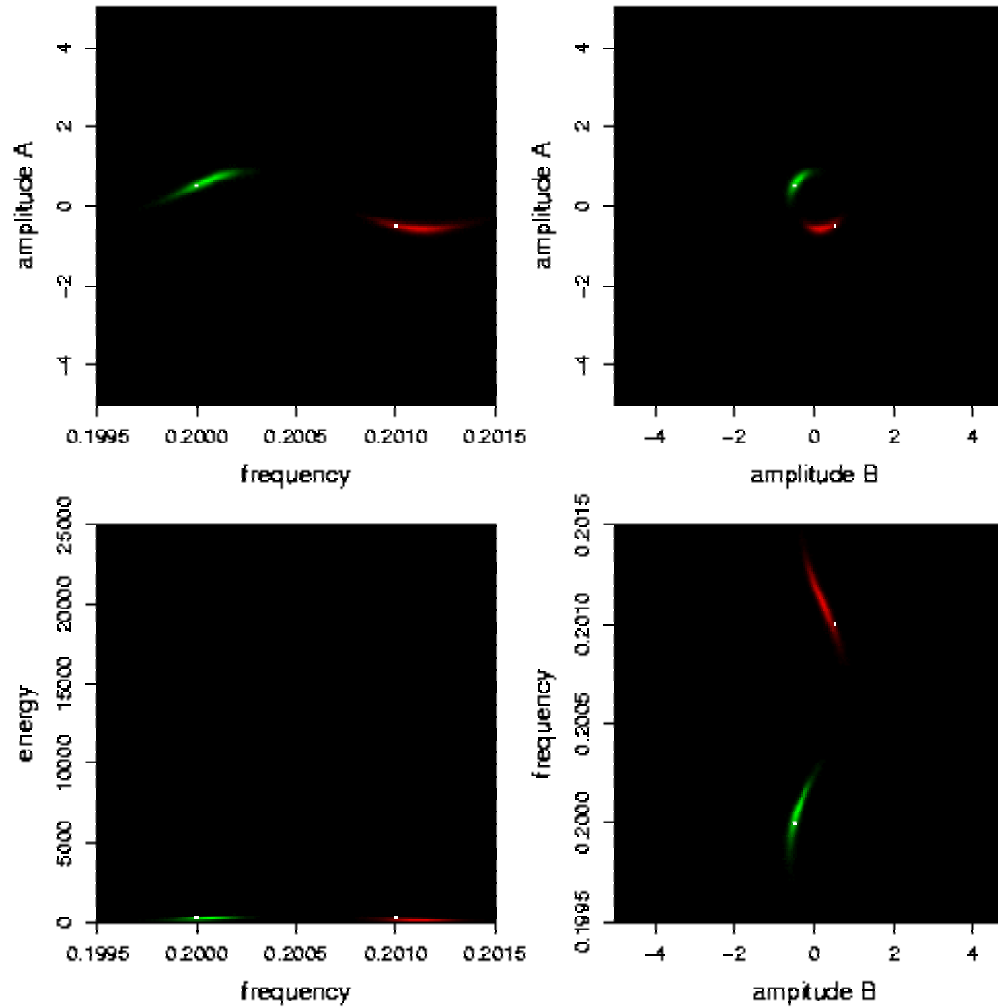


Strong, close signals



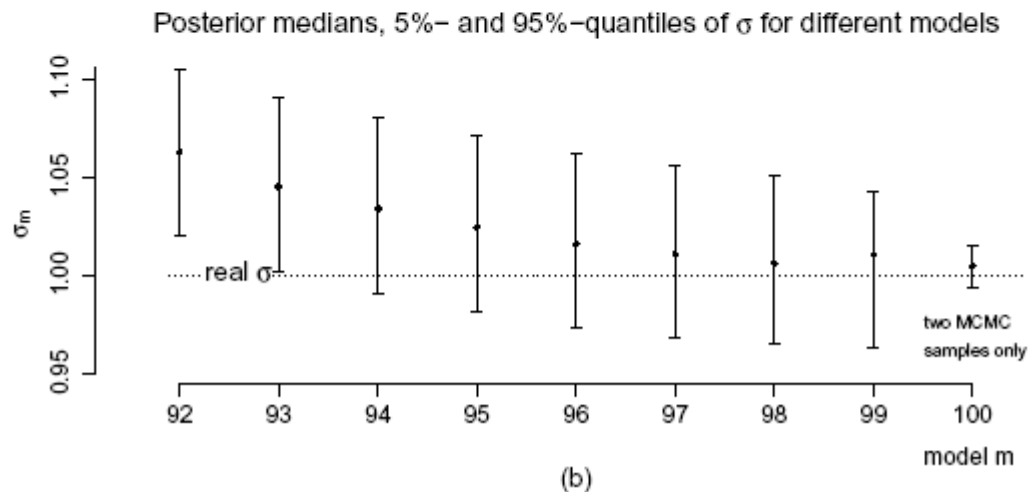
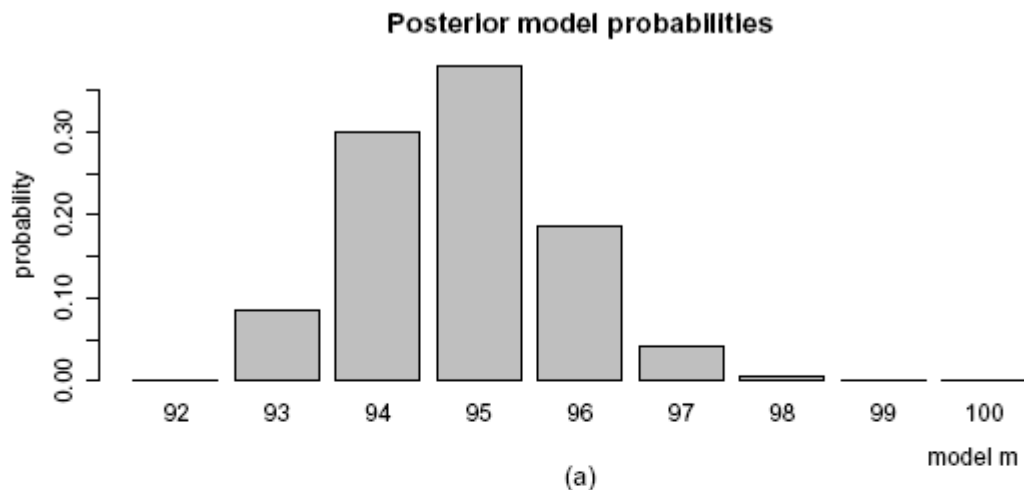
Signal mixing

- Two signals (red and green) approaching in frequency:

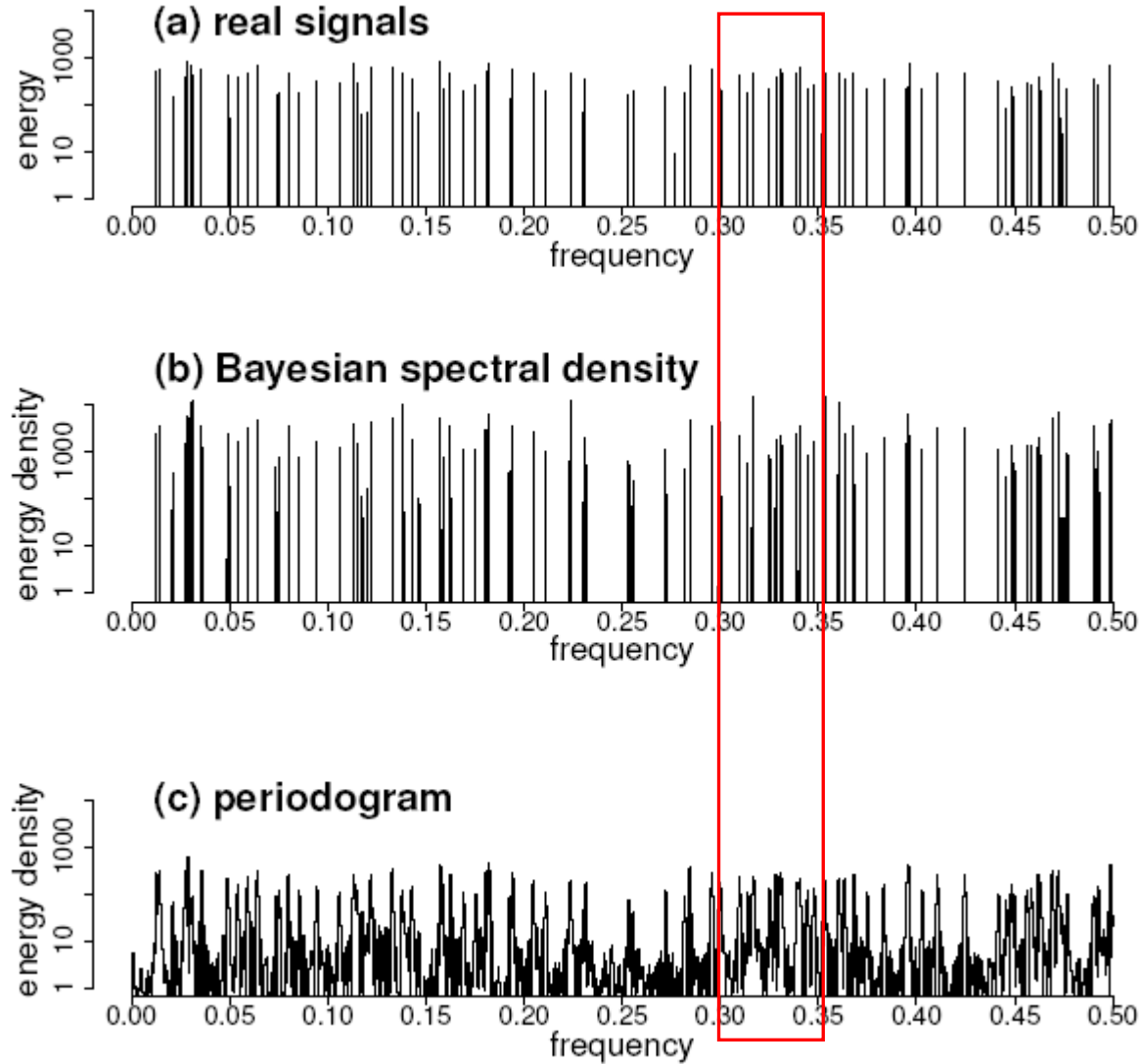


Marginal pdfs for m and σ

$$p(m | \{D\}, I) \propto \iiint_{\{A\}_m, \{f\}_m, \sigma_m} p(m, \{A\}_m, \{B\}_m, \{f\}_m, \sigma | \{D\}, I) d\{A\}_m d\{B\}_m d\{f\}_m d\sigma_m$$

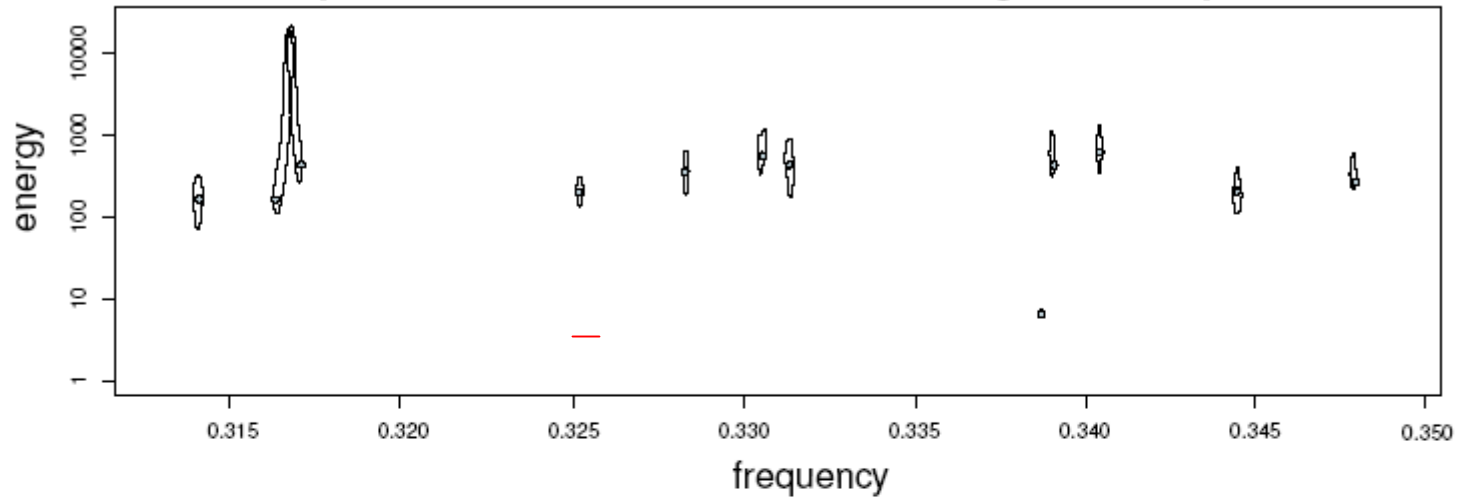


Spectral density estimates

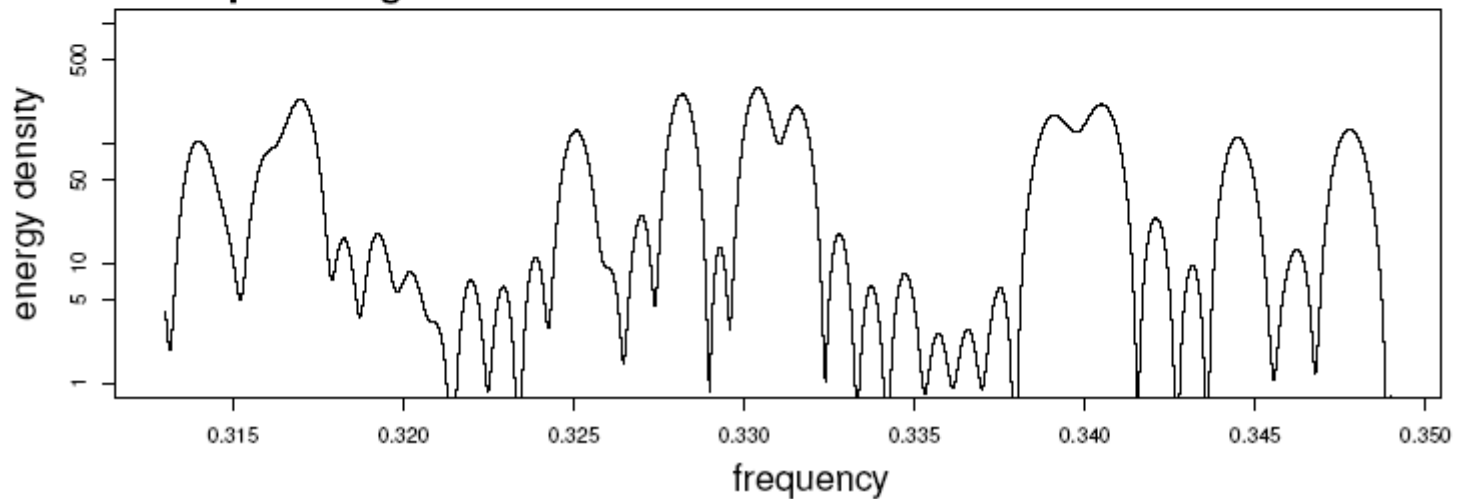


Joint energy/frequency posterior

95% posterior confidence areas of energy and frequency

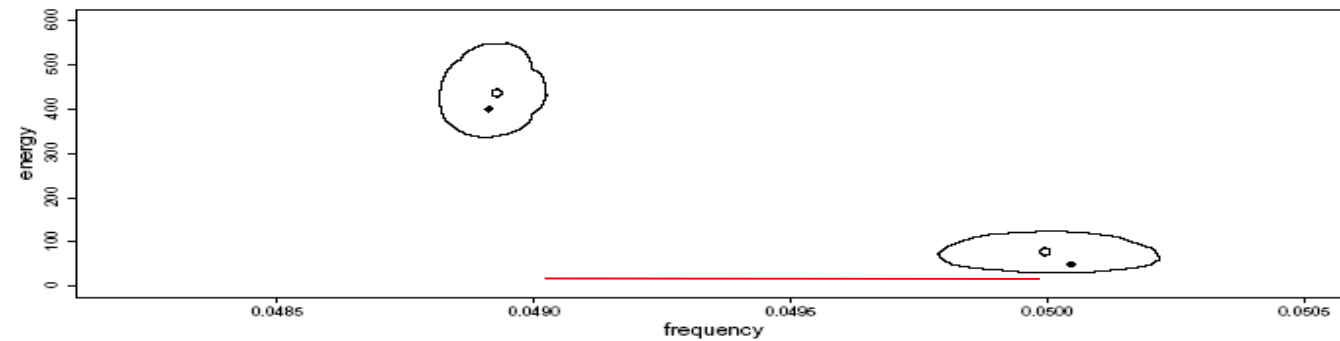
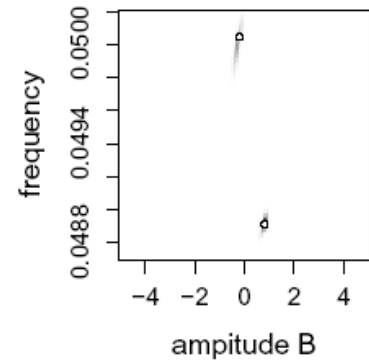
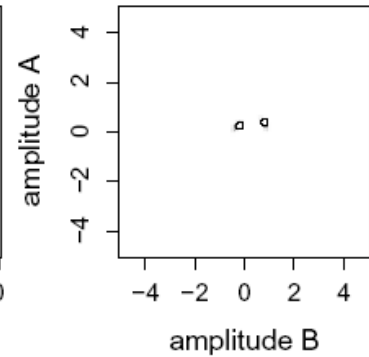
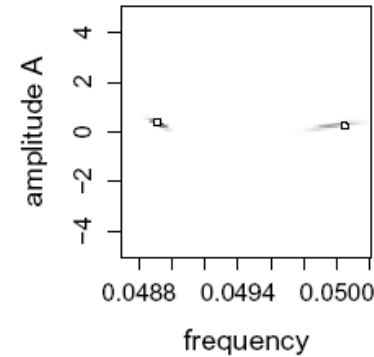


periodogram

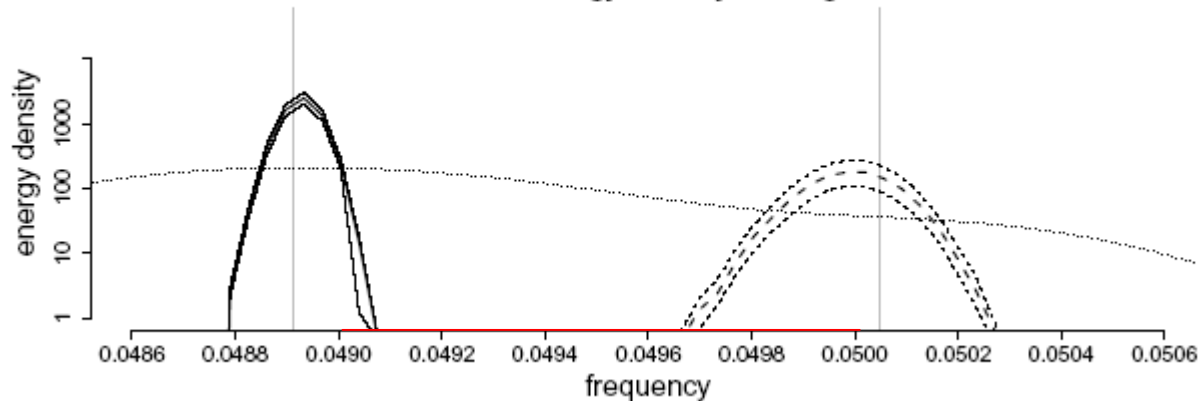


Well-separated signals ($\sim 1/T$)

These signals (separated by ~ 1 Nyquist step) can be easily distinguished and parameterised.

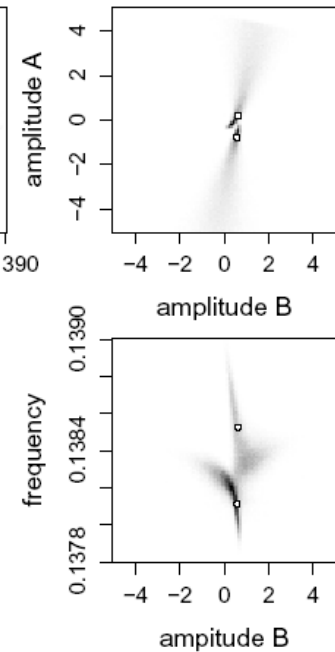
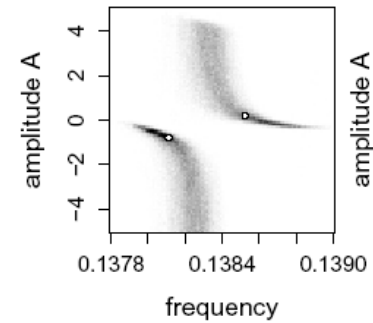
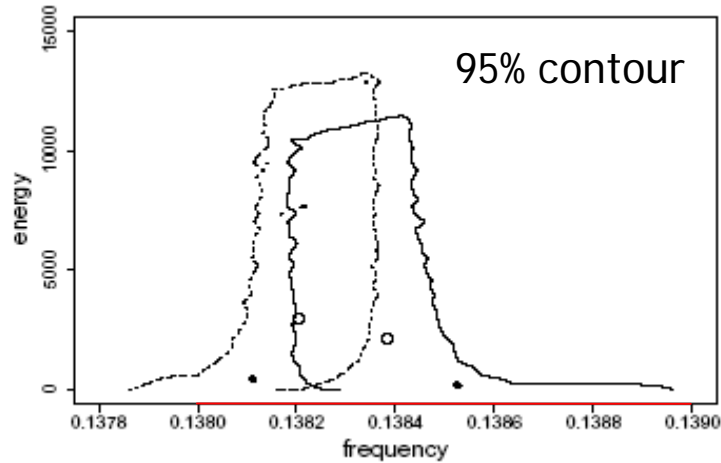


Estimated energy density for Region-1



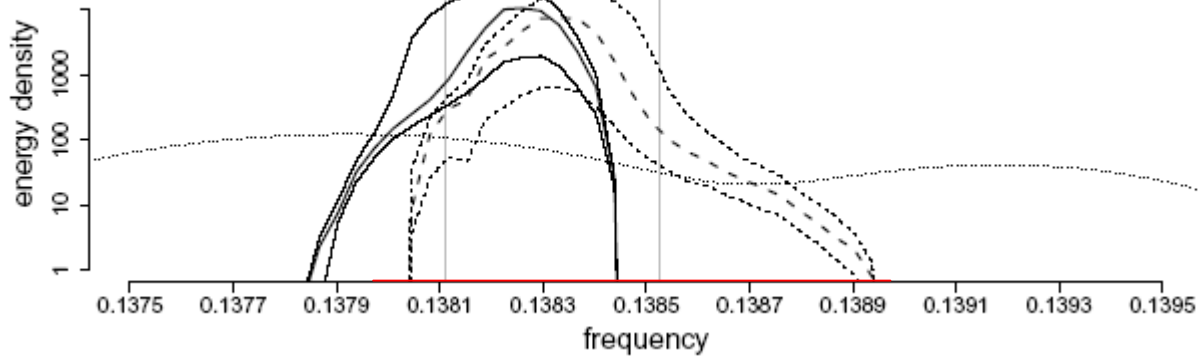
Closely-spaced signals

Signals can be distinguished, but parameter estimation is difficult.



Joint posterior distribution of amplitudes and frequency for

Estimated energy density for Region-2



Conclusions and Extensions to full LISA

- A Bayesian framework is the natural way to address LISA data analysis.
- We have implemented an MCMC method of extracting useful information from zeroth-order LISA data under difficult conditions.
- Extension to orbital Doppler/source location information should **improve** source identification (clustering is in a higher dimensional space).
- Extension to TDI variables is straightforward. Raw Doppler measurements could also be used, with a suitable data covariance matrix.
- There is nothing special about WD-WD signals here. Similar analyses could be performed for BH mergers, EMRI sources etc...
- -> a hierarchical scheme to solve the global LISA data analysis challenge.