Bayesian modeling of source confusion in LISA data

Nelson Christensen, Sarah Vigeland
Carleton College

Richard Umstätter, Renate Meyer
University of Auckland

Graham Woan, Vimal Simha, John Veitch,
Martin Hendry
University of Glasgow

Source Confusion for LISA

- LISA sensitivity range 10^{-5} - 0.1 Hz
- Most abundant source will be close-by White Dwarf Binaries: 0.1 mHz to 3 mHz
- Source confusion below 1 mHz
- Resolvable sources above 5 mHz
- 1 mHz to 5 mHz band presents a data analysis challenge
- 50 000 to 100 000 potential signals

Markov chain Monte Carlo

- Developed a toy problem - N sinusoidal signals within noise
- Apply a numerical Bayesian technique, a Metropolis-Hastings algorithm
- Estimate parameters associated with each sinusoid, signal number N and noise level
- Bayesian – Occam's Razor preference for smaller N

Extraction of Parameters for Signal

- We assume a simple signal that consists of N superposed sinusoidal signals:

$$\mathrm{f}(t) = \Sigma_{i=1}^{N}[A_i^{(1)} \cos(2\pi f_i t) + A_i^{(2)} \sin(2\pi f_i t)]$$

We fit this model to the data:

$$d_t = \mathrm{f}(t, \boldsymbol{a}, N) + \epsilon$$

with noise $\epsilon \propto N(0, \sigma^2)$

$$\boldsymbol{a} = [A_1^{(1)}, A_1^{(2)}, f_1, \cdots, A_N^{(1)}, A_N^{(2)}, f_N]$$

N is also an unknown parameter

Likelihood and Bayes

The joint likelihood that these data
$\boldsymbol{d} = \{d_t\}$ arise from the parameter vector $\boldsymbol{a}$
is given as
$p(\boldsymbol{d}|\boldsymbol{a}) \propto \exp\left(-\frac{1}{2\sigma^2}\Sigma_t \left(d_t - \mathrm{f}(t, \boldsymbol{a}, N)\right)^2\right)$

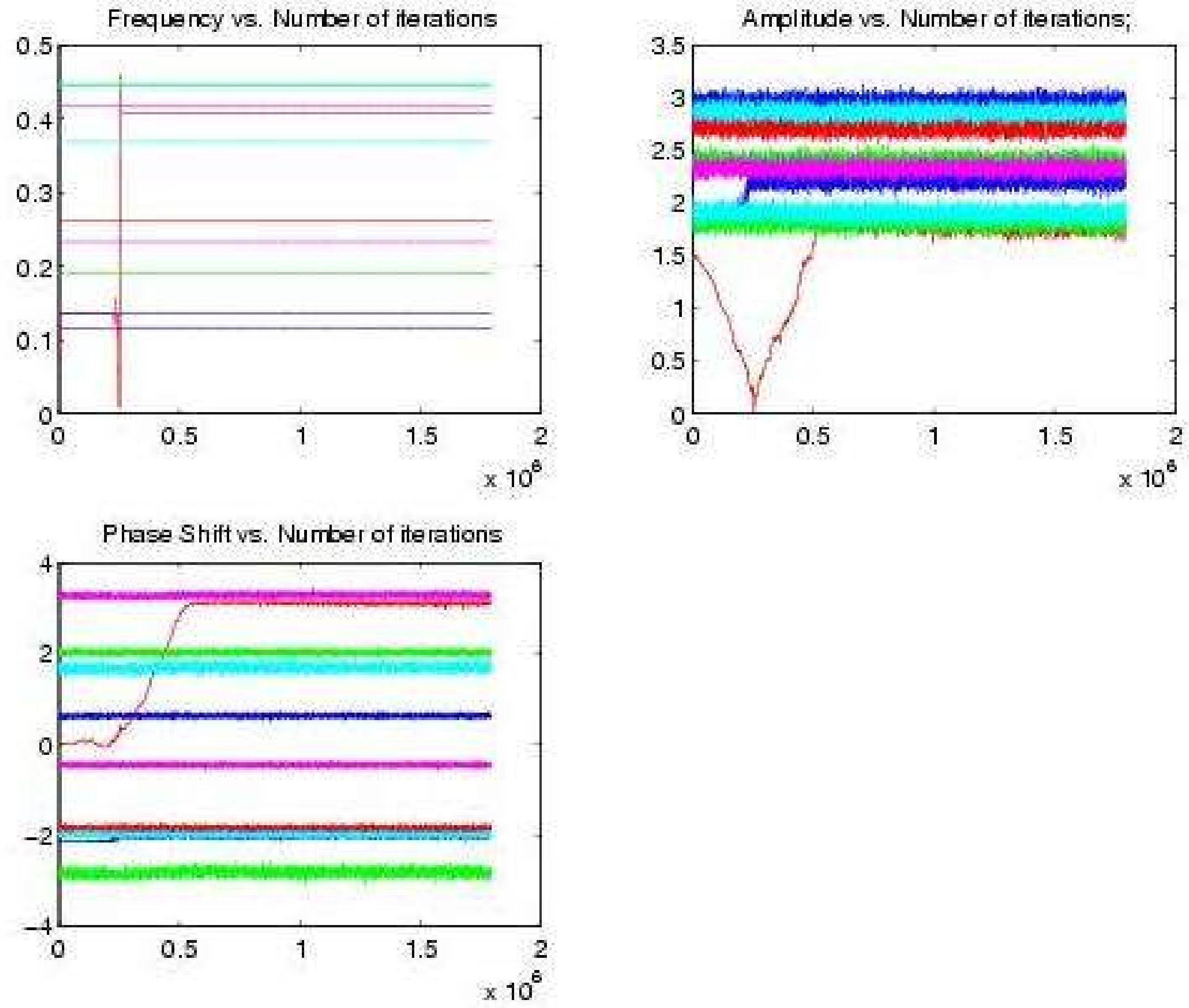
Simple uniform priors are chosen for the amplitude parameters, and frequency. The MCMC provides an estimate of the (unnormalized) posterior density:

$$p(\boldsymbol{a}|\boldsymbol{d}) \propto p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})$$

20 signals (10 shown). Output chains from the MCMC

MCMC takes a quasi-intelligent walk through the parameter space.

Histograms of the post-convergence chains provide estimates of the posterior PDFs



Reversible Jump MCMC

- Total number of signals, N, is unknown
- Need to apply a non-standard MCMC
- Different N, different model
- Create or annihilate signals
- Split and merge signals
- Also estimate noise level

Splitting and Merging Signals

Merge: Two randomly selected signals $i \neq j$ with parameter vectors $S_i=[A_i^{(1)},A_i^{(2)},f_i]$ and $S_j=[A_j^{(1)},A_j^{(2)},f_j]$ are merged into a new signal with parameter vector $S_{i*}=[A_i^{(1)}+ A_j^{(1)}, A_i^{(2)}+ A_j^{(1)}, (f_i+f_j)/2]$

Split: For a randomly chosen signal i with parameter vector $S_i=[A_i^{(1)},A_i^{(2)},f_i]$ a proposal $u=[u_{A(1)}, u_{A(2)}, u_f]$ is sampled from a distribution q_s and split up into signals with parameter vectors $S_{i*}=[A_i^{(1)}/2, A_i^{(2)}/2, f_i] + u$ and $S_{j*}=[A_j^{(1)}/2, A_j^{(2)}/2, f_j] - u$

Creating and Annihilating Signals

Annihilation: A randomly selected chosen signal i with $S_i=[A_i^{(1)},A_i^{(2)},f_i]$ is simply killed.

Creation: A vector $u=[u_{A(1)}, u_{A(2)}, u_f]$ is sampled from a distribution q_c and the new signal with parameter vector $S_{i*} = u$ is added.

Accept or Reject

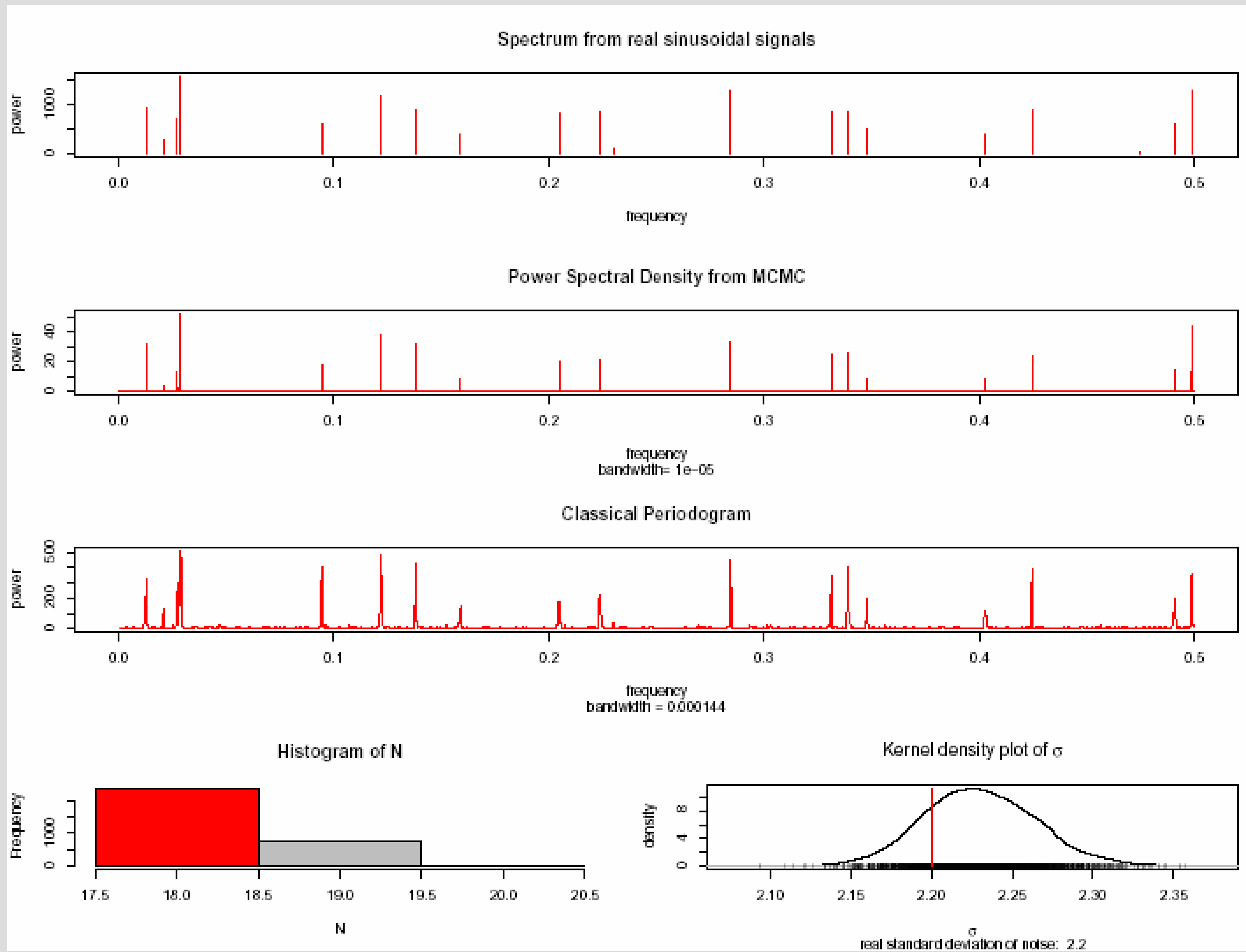
New parameters and new models are accepted or rejected based on Metropolis-Hastings acceptance methods (also using delayed rejection).

Example with N=20 signals,   =2.2, T=2 000s of data sampled at f_s=1 Hz, and max amplitude for a signal of A=1.0

MCMC estimates 18 signals

MCMC generated power spectral density better than classical periodogram

Missed 2 signals f=0.23 Hz and 0.47Hz

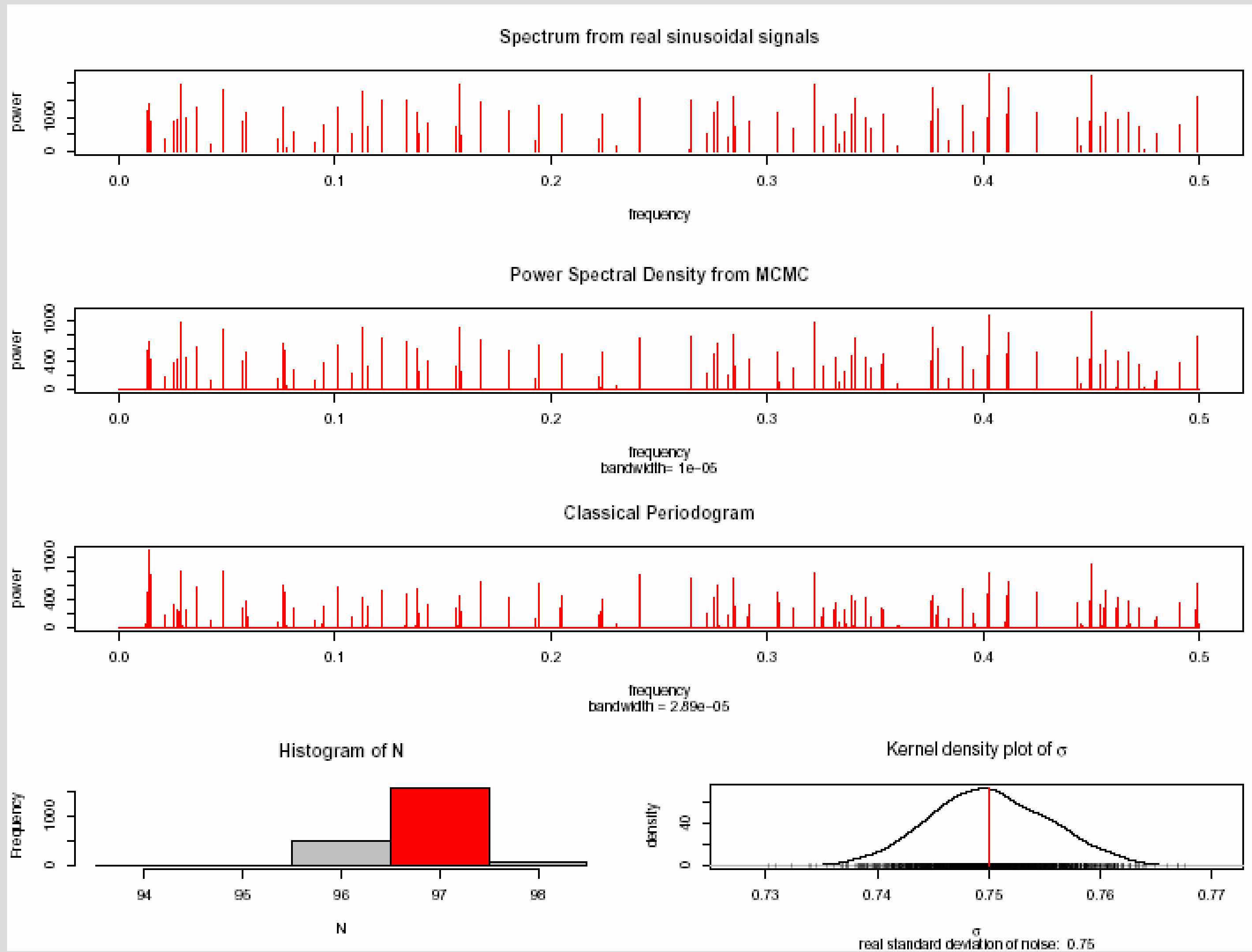


Example with N=100 signals, =0.75, T=10 000s of data sampled at f_s=1 Hz, and max amplitude for a signal of A=0.5

MCMC generated power spectral density is again better than classical periodogram

We can be more sure about the real values of our parameters if there are fewer combinations that plausibly fit the data.

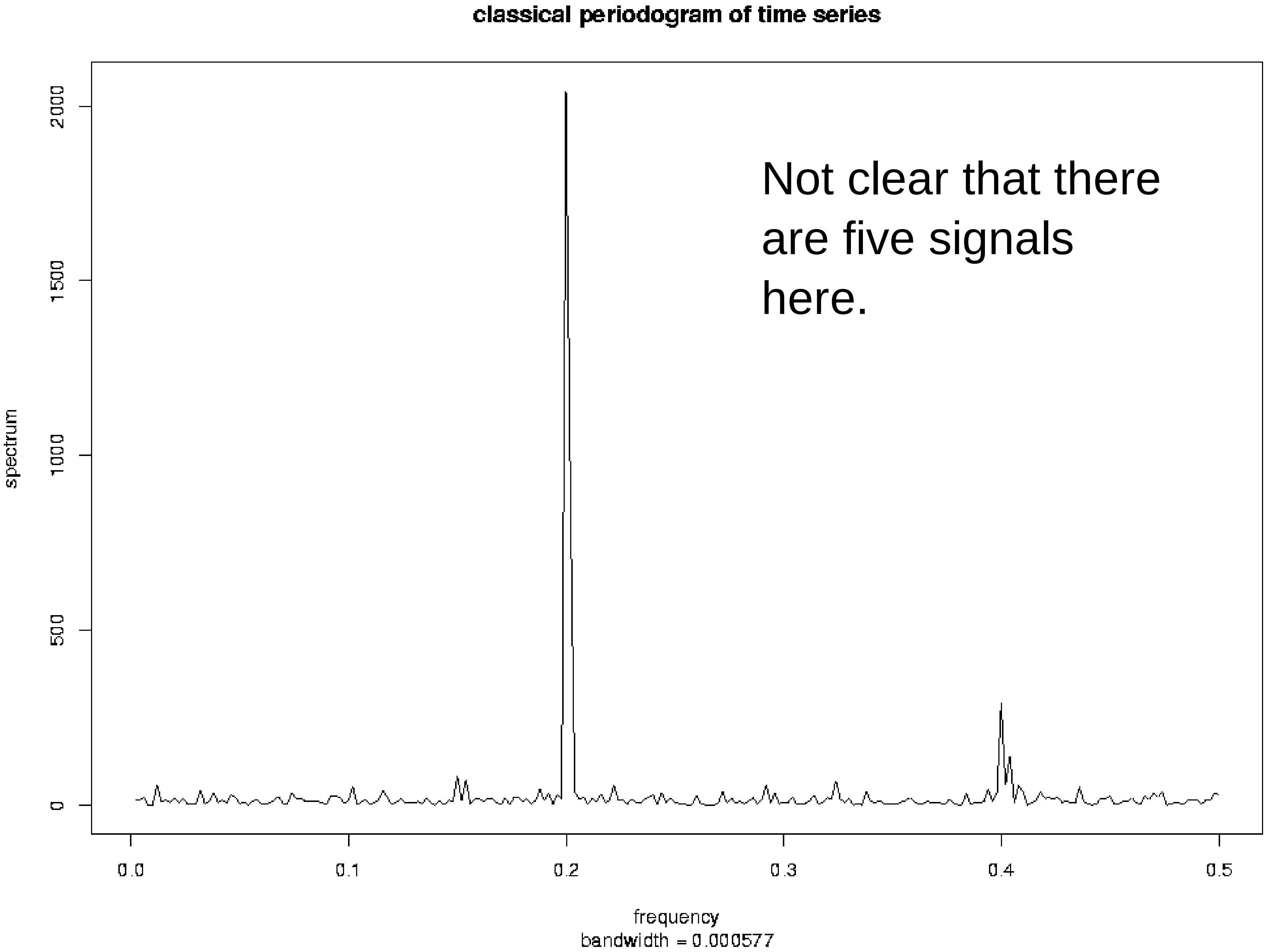
MCMC estimates 97 signals



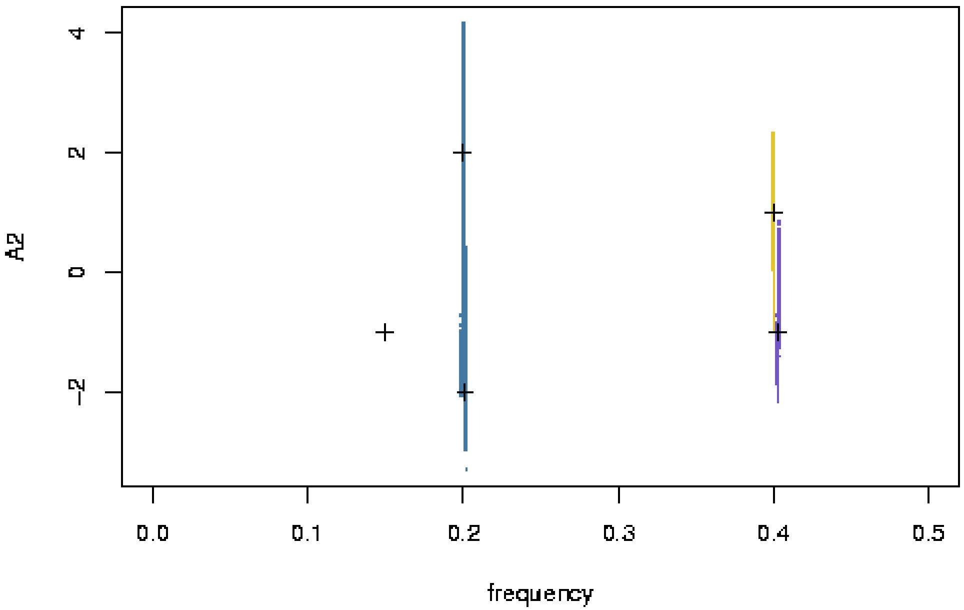
MCMC Good at Resolving Signals with Similar Frequencies

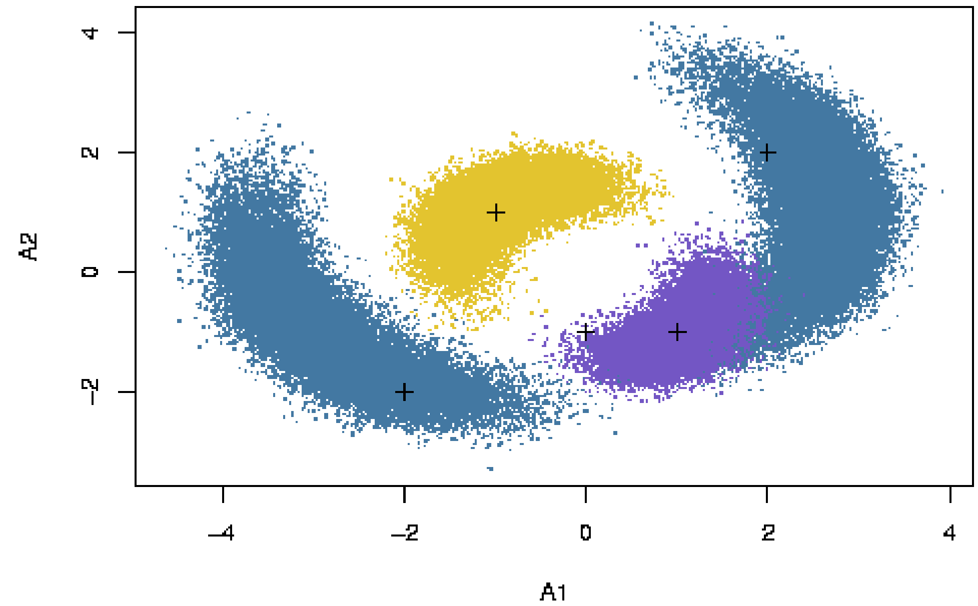
5 injected signals, T = 500s, noise level =3.6, sampled at 1 Hz

n	A1	A2	frequency
1	2	2	0.2
2	-2	-2	0.201
3	-1	1	0.4
4	1	-1	0.403
5	0	-1	0.15



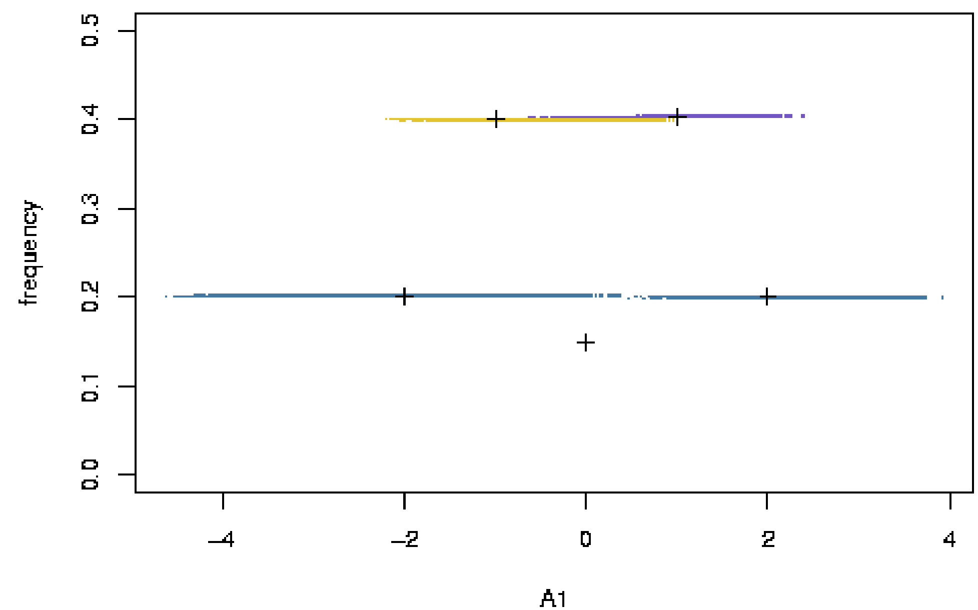
Not clear that there are five signals here.

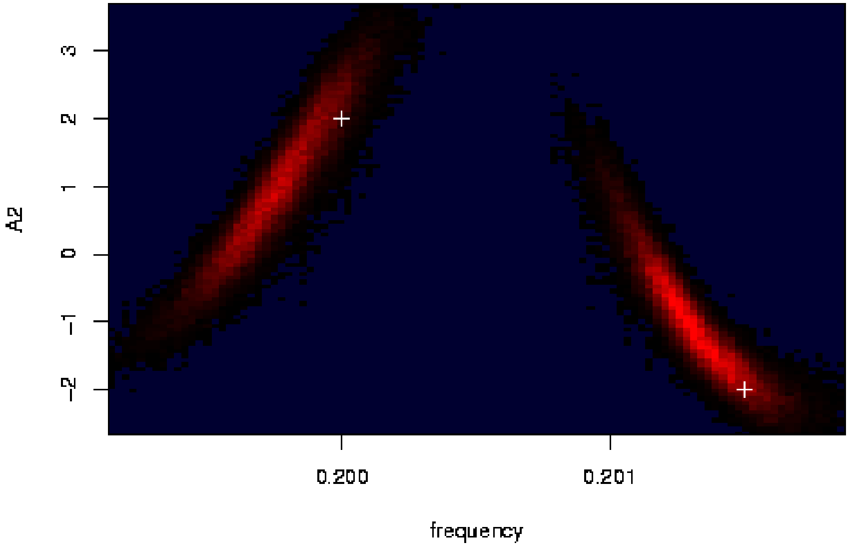


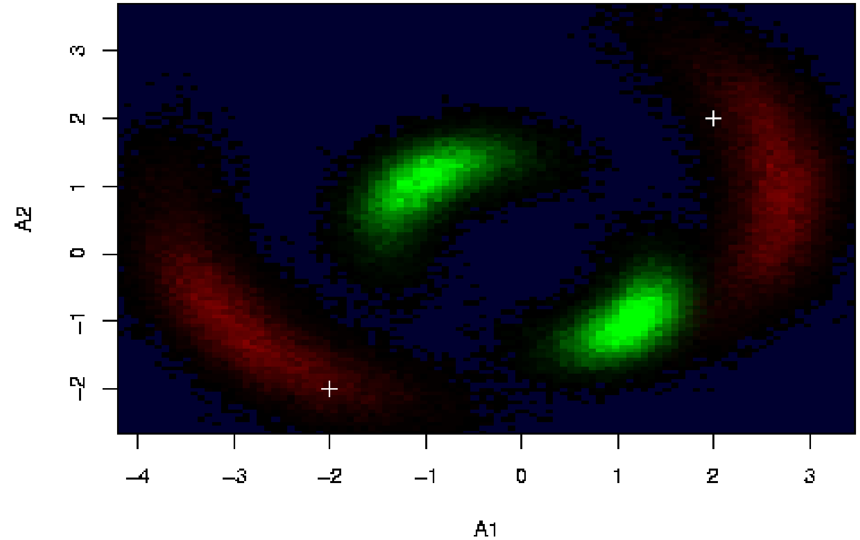


4 of 5 signals
found in output

small signal at
f=0.15 missed



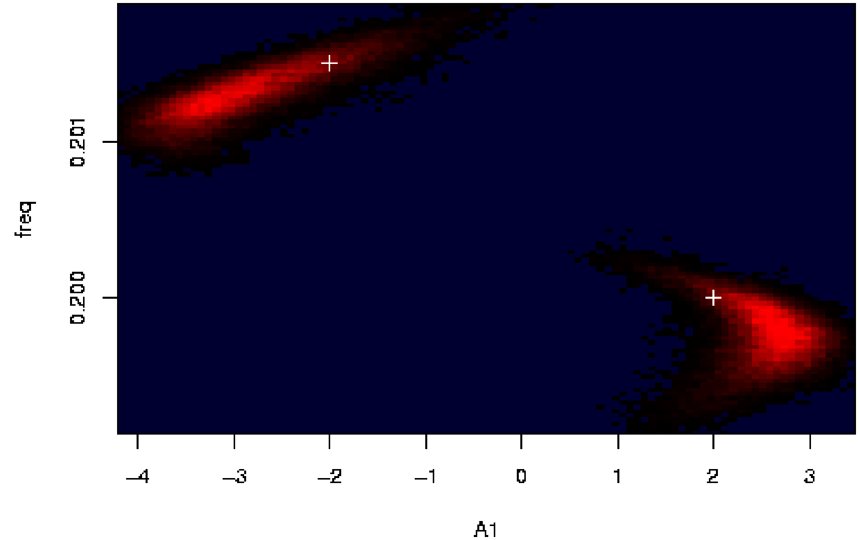


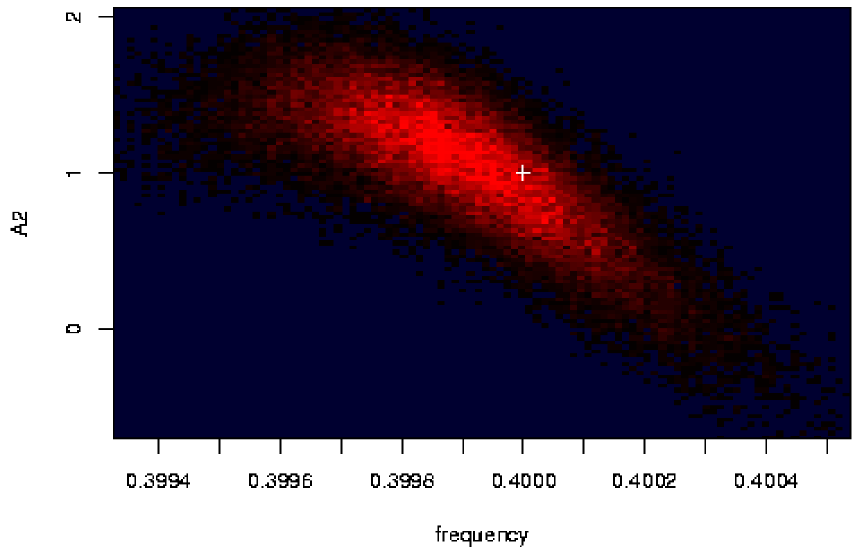


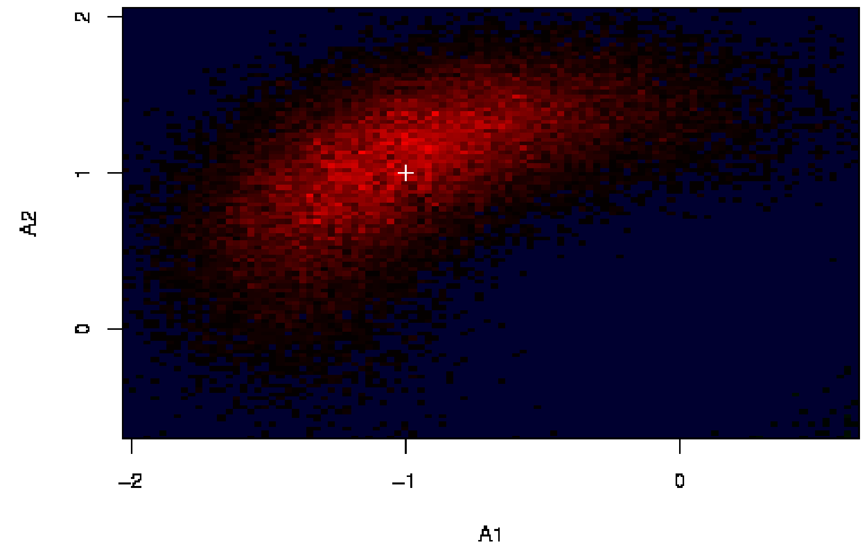
Example of densities

Estimates of posterior PDFs

A1	A2	frequency
2	2	0.2
-2	-2	0.201



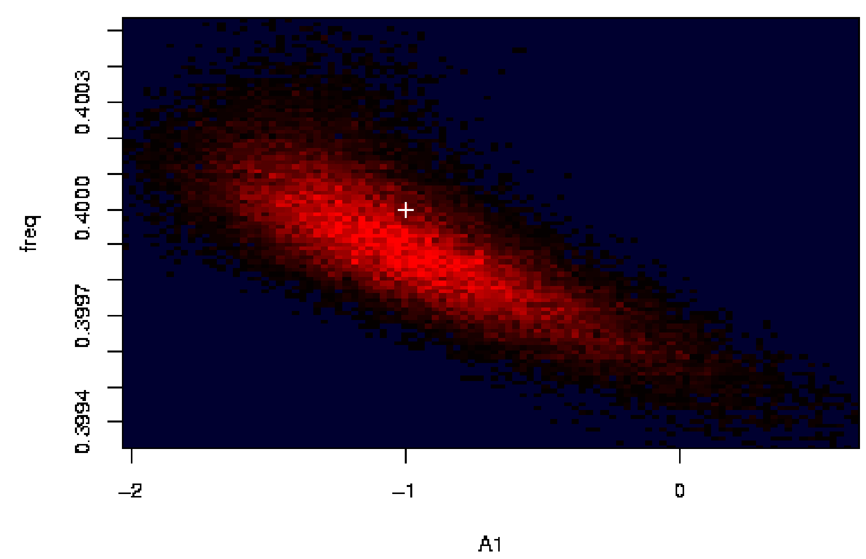


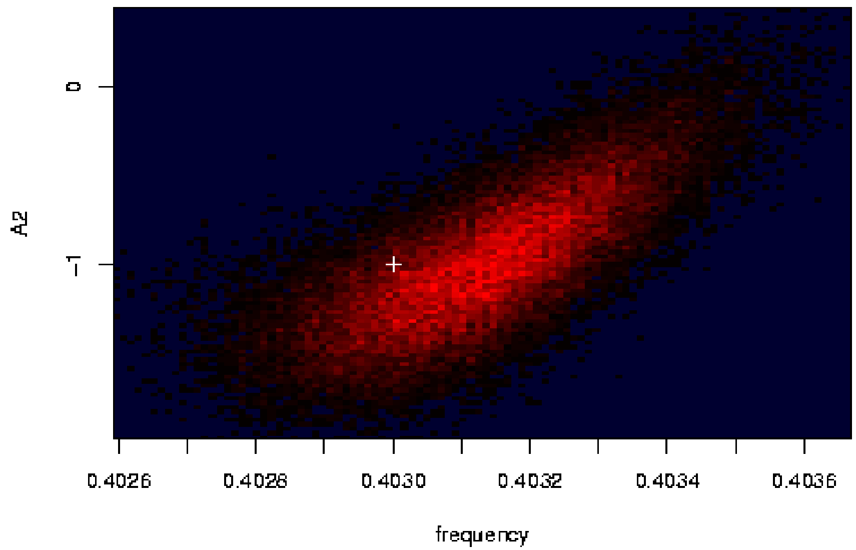


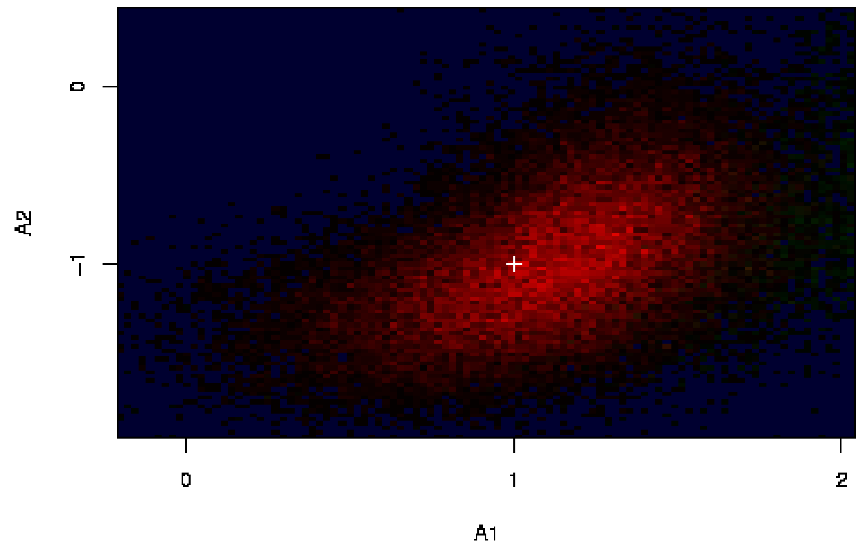
Example of densities

Estimates of posterior PDFs

A1	A2	frequency
-1	1	0.4
1	-1	0.403



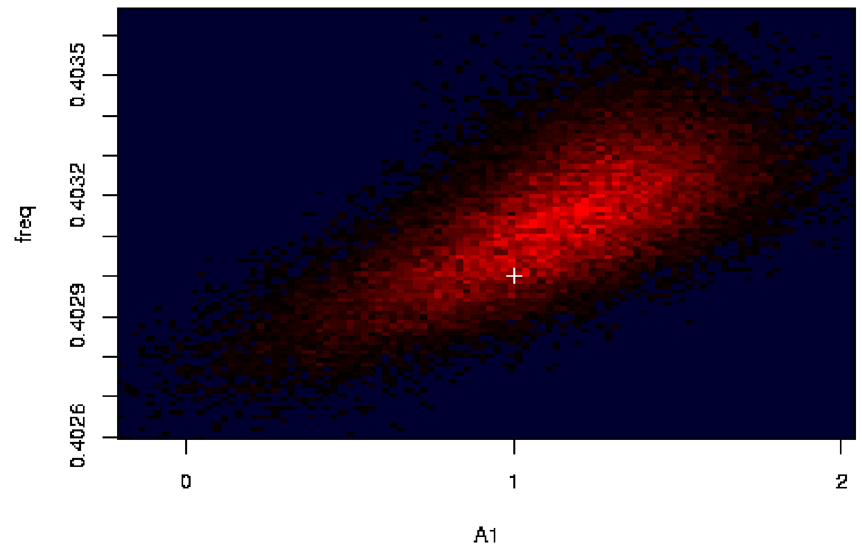


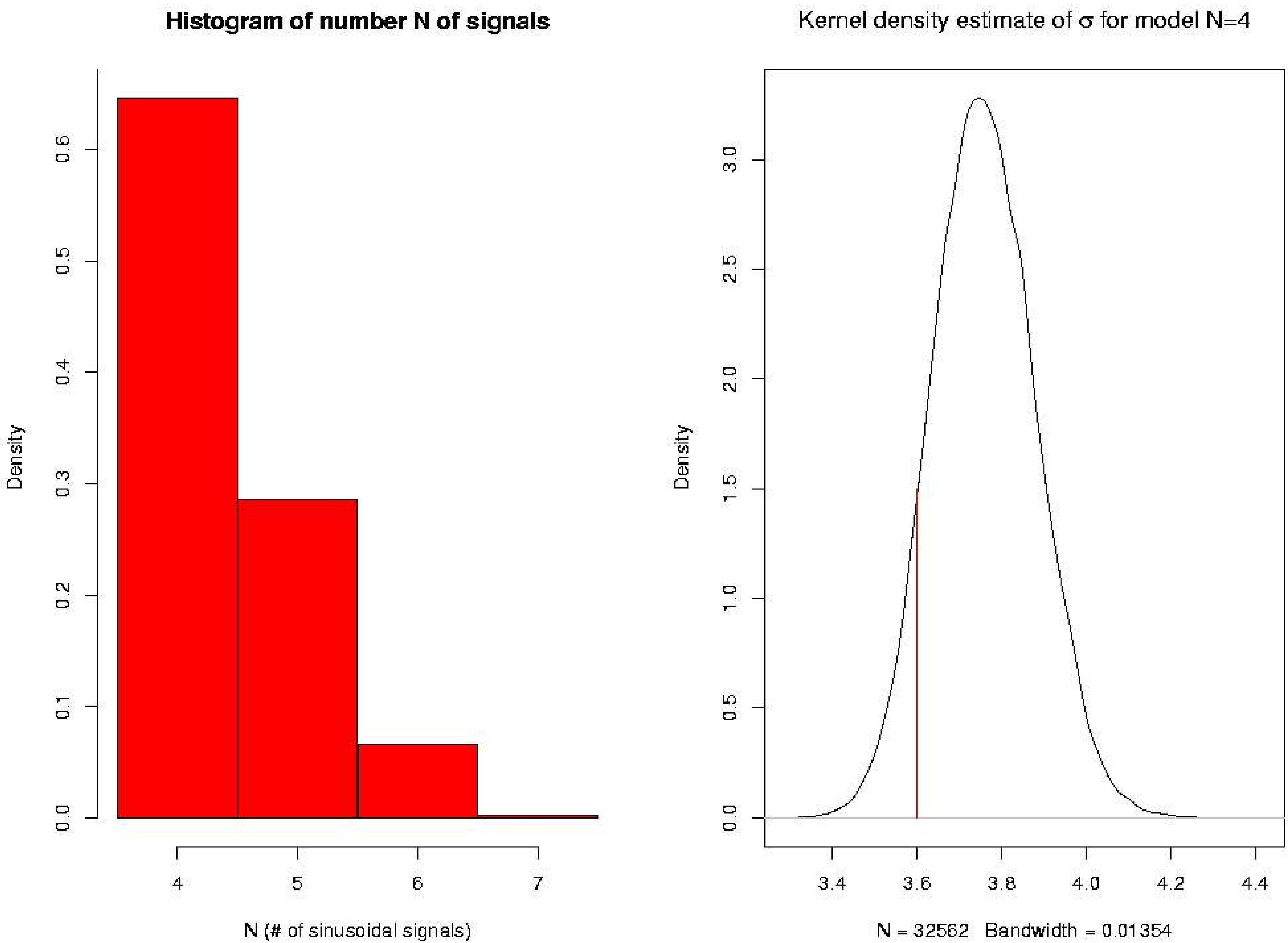


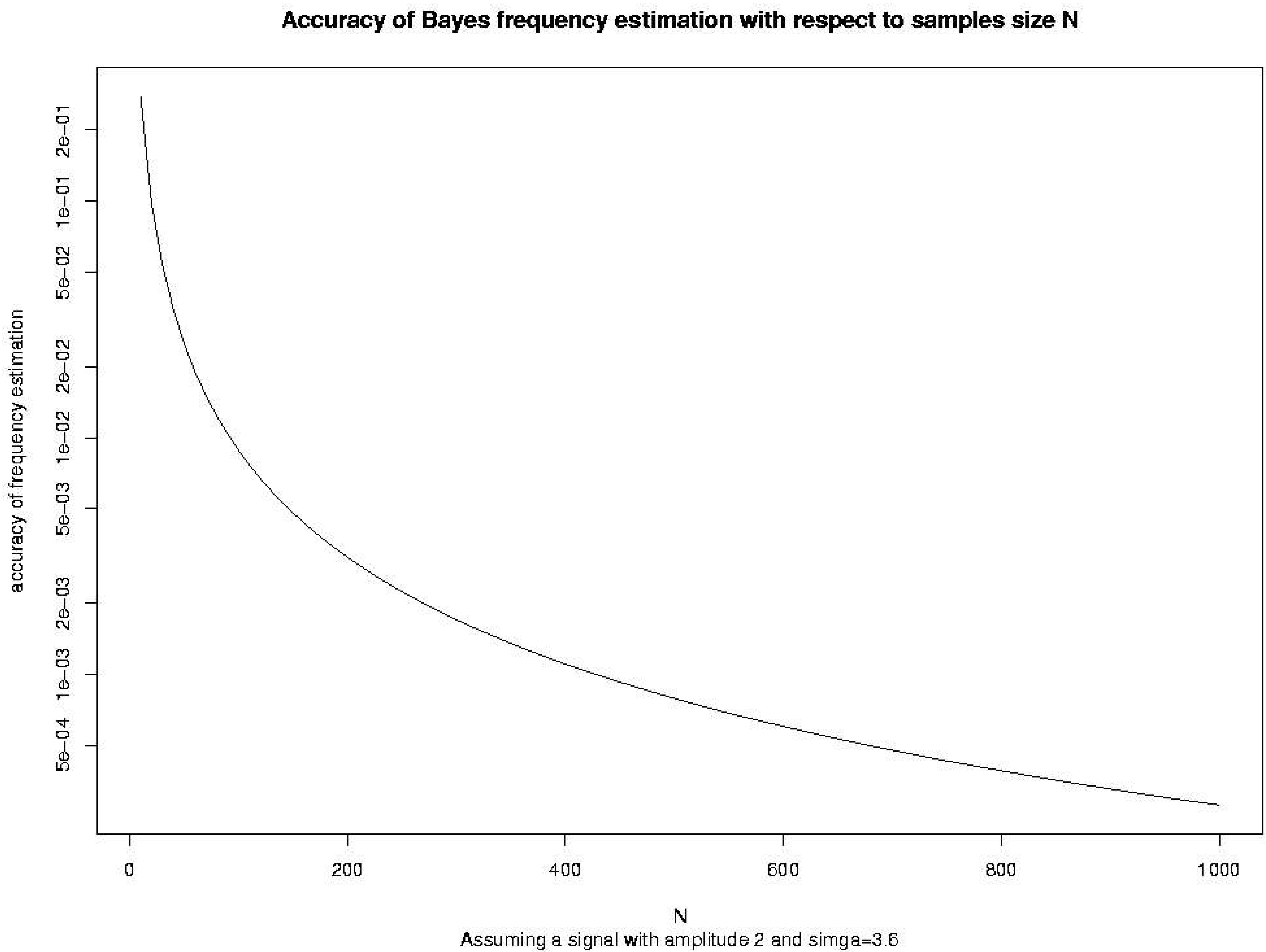
Example of densities

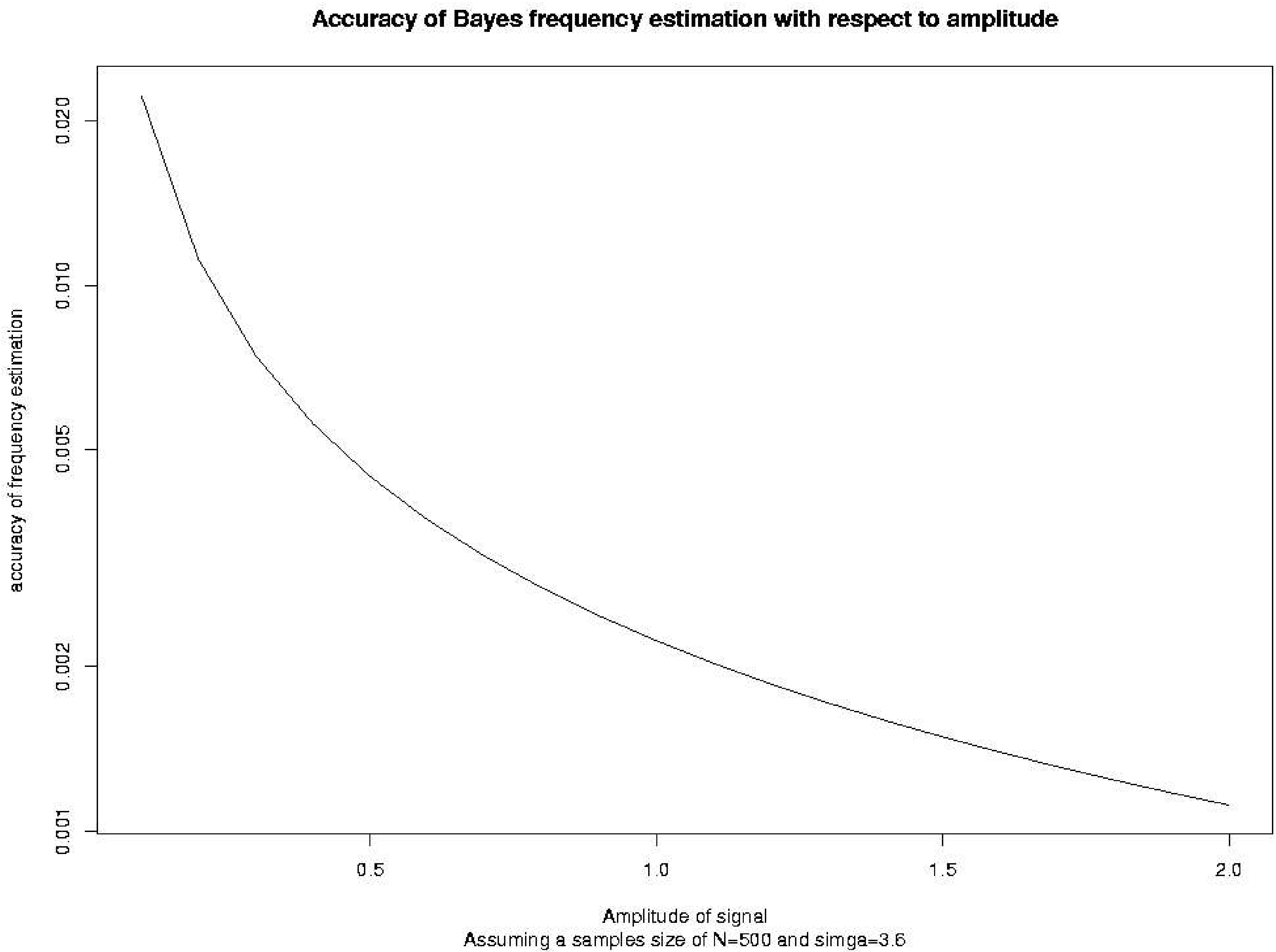
Estimates of posterior PDFs

A1	A2	frequency
-1	1	0.4
1	-1	0.403




Histogram of number N of signals
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4
5
6
7
N (# of sinusoidal signals)
Kernel density estimate of σ for model N=4
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.4
3.6
3.8
4.0
4.2
4.4
N = 32562 Bandwidth = 0.01354


Accuracy of Bayes frequency estimation with respect to samples size N
accuracy of frequency estimation
2e−01
1e−01
5e−02
2e−02
1e−02
5e−03
2e−03
1e−03
5e−04
0
200
400
600
800
1000
N
Assuming a signal with amplitude 2 and simga=3.6


Accuracy of Bayes frequency estimation with respect to amplitude
accuracy of frequency estimation
0.001
0.002
0.005
0.010
0.020
0.5
1.0
1.5
2.0
Amplitude of signal
Assuming a samples size of N=500 and simga=3.6

What Next

- Get good numbers on scalability: computation time as N and T increase. Appears roughly linear so far with N and with T.
-
- Realistic Signals. Start coupling this work with the LISA simulator