Outlier Detection in Bayesian Hierarchical Models with Gibbs Sampling

09/14/20, Qiaohong (Joanna) Wang, Dept. of Physics & Astronomy, Vanderbilt University

IPTA Meeting 2020

Collaborator: Professor Stephen Taylor

Outlier Analysis

Motivation: Real-world data is often susceptible to outliers

- Increase uncertainties of the inference model
- Affect the accuracy of statistical inferences, or even dominate the 'fit'

Past & Relevant Work

- Bayesian outlier analysis using Gibbs sampling (Verdinelli & Wasserman, 1991)
- Outliers with Hamiltonian sampling (*Vallisneri & van Haasteren, 2017*)

Challenges in Previous Bayesian Outlier Analysis

- Difficult computations for calculating posterior marginals of an outlier model
- Lack of simplicity and versatility (e.g. Hamiltonian sampling requires derivatives of the density function)

Benefits of Using Gibbs Sampling for Outlier Analysis

- Directly draws samples from the posterior distribution through alternating iterations, thus easy to compute posterior probability of an observation being an outlier
- Requires univariate conditional distributions for each step - simple and easy to implement

Bayesian Hierarchical Model Example 1: Sinusoidal Time Series Analysis with Outliers

$y_i \mid a, \sigma, p, \delta \sim N\left[(1-\delta)\left(a \sin((t_i/p) + \phi)\right),\ (1-\delta)\, rms^2 + \delta\, A^2\right]$

$a \sim N(a_0, \sigma_0^2)$

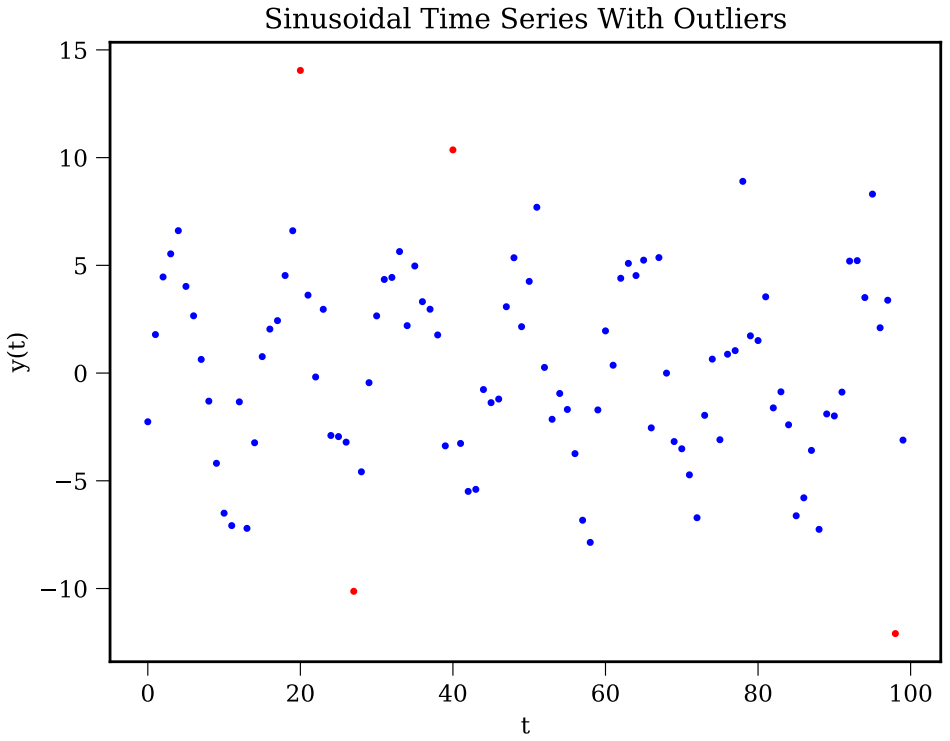
$rms^2 \sim IG(\nu_0, \beta_0)$

$p \sim \text{Uniform}(N)$

$\delta \mid \epsilon \sim \text{Bern}(\epsilon)$

$\epsilon \sim \text{Beta}(\alpha, \beta)$

$y_i = a \sin(t_i/p + \phi) + N(0, rms^2)$
with outliers $\sim N(0, A^2)$



Proposed Solution: Outlier Detection with Gibbs Sampling

Gibbs Sampling: Update multiple parameters by sampling one parameter at a time and cycle through all parameters

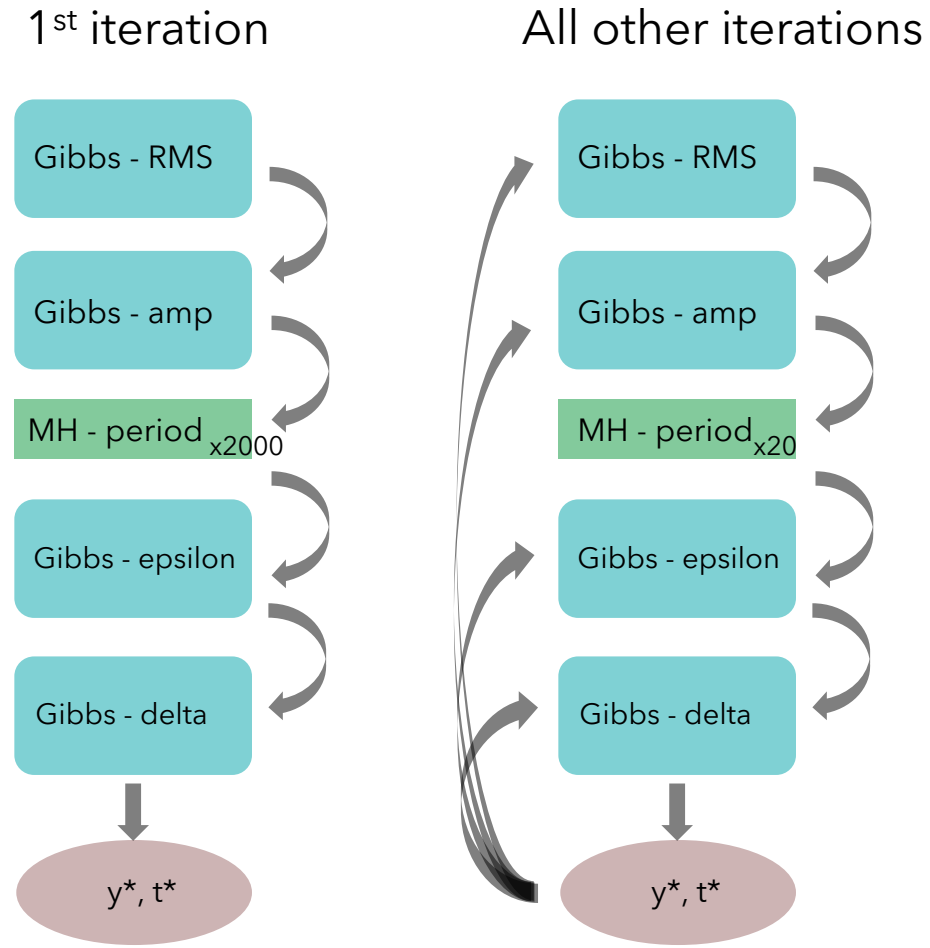
Applicable for:

- Noisy sine wave parameters:
  - Amplitude
  - RMS
- Outlier parameters
  - delta (data labels)
  - epsilon (outlier probability)

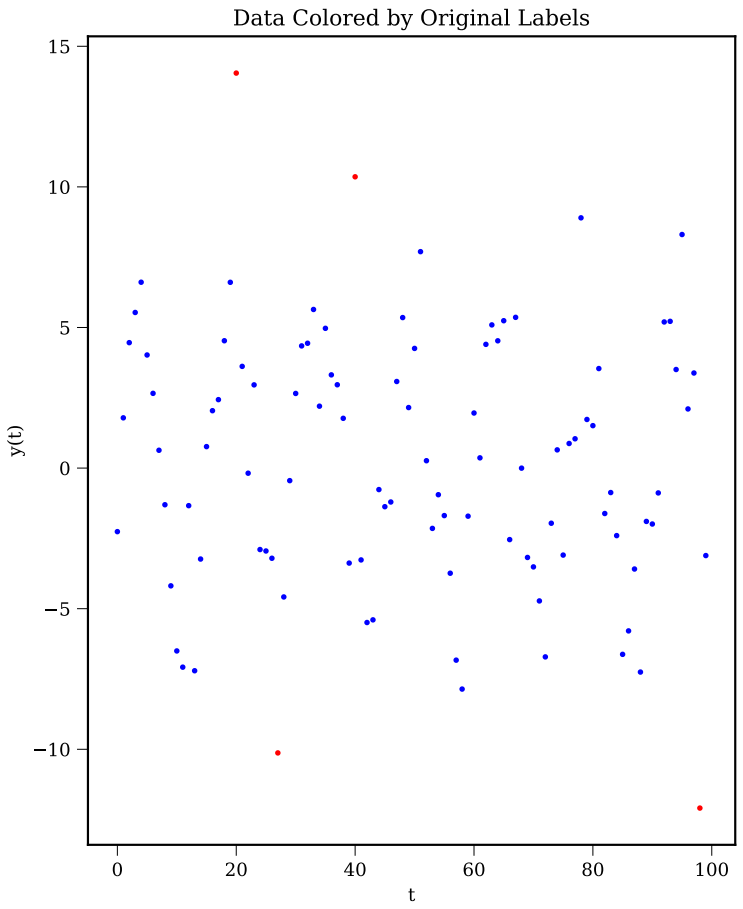
Metropolis-Hasting: Initialize our desired parameter and draw a candidate and decide to move the chain or no

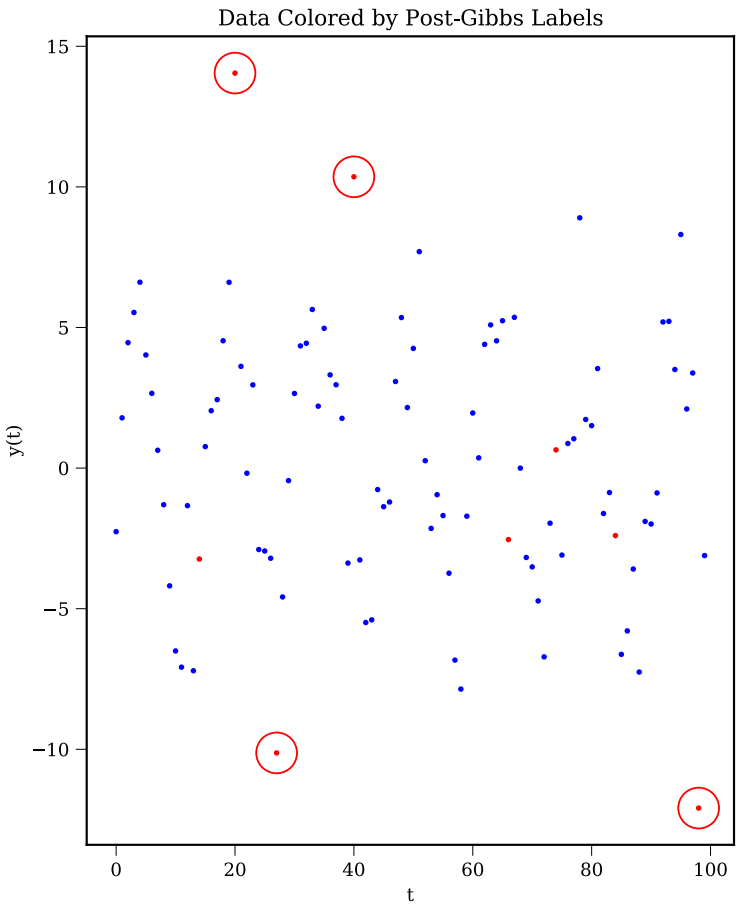
Applicable for:

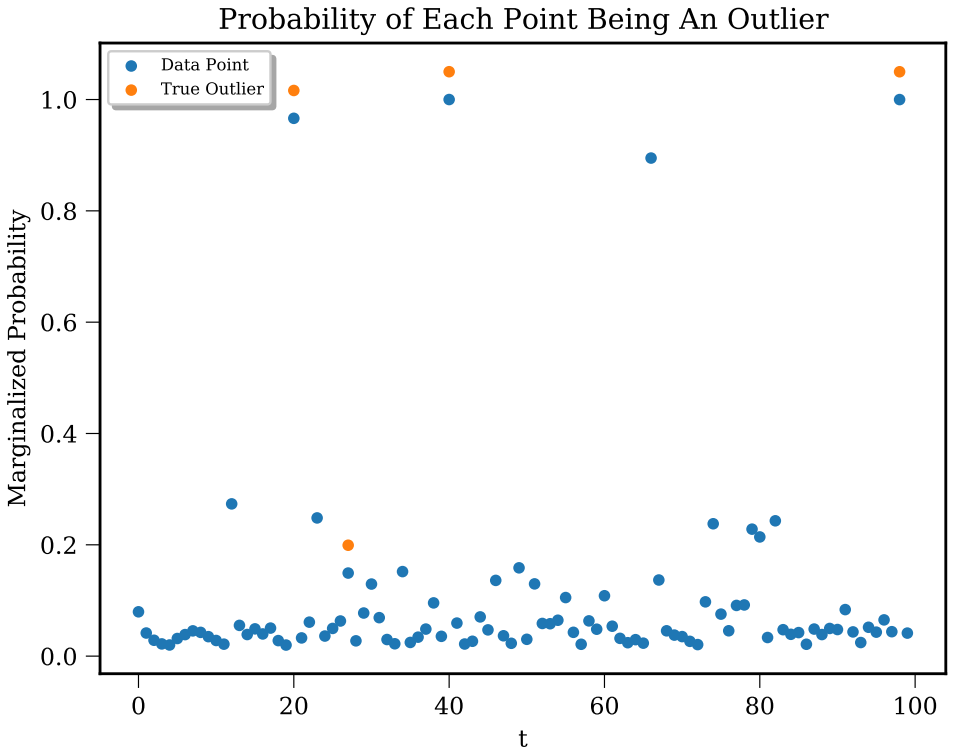
- Noisy sine wave parameters:
  - period



Results







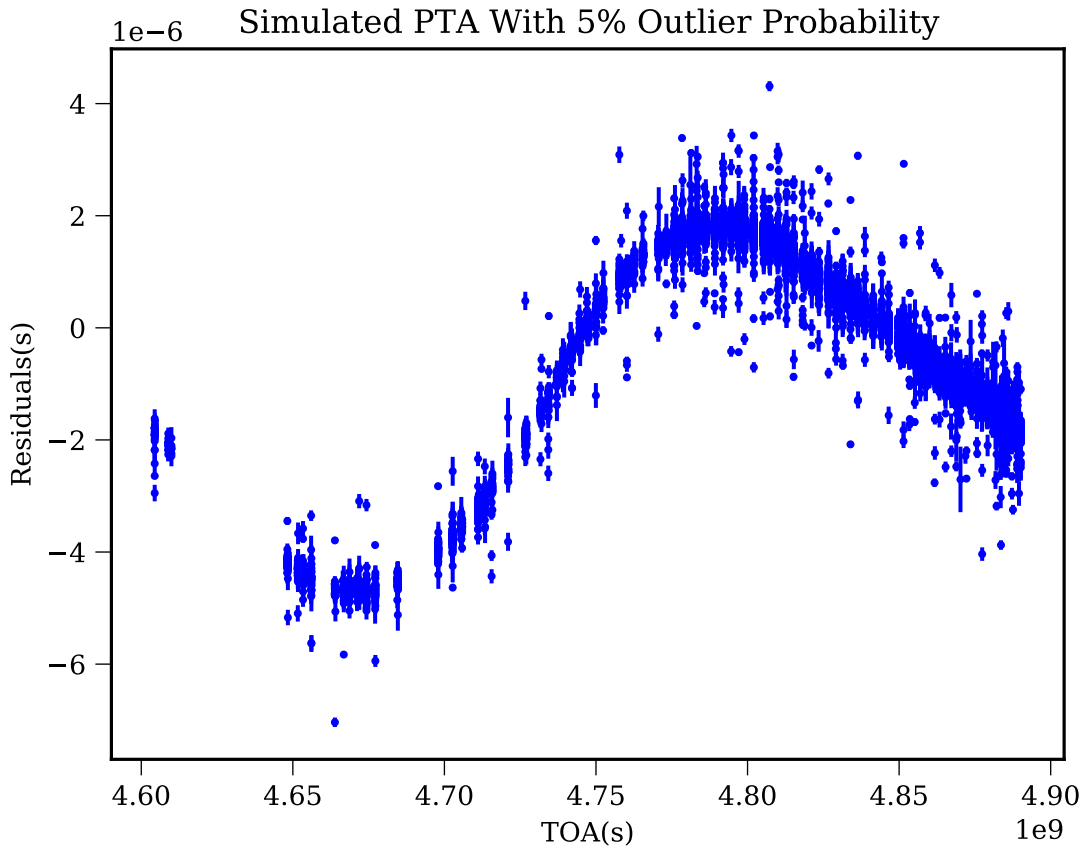
- Blue data points: “true” data
- Red data points: outliers identified by the algorithm
- Red circles: real outliers

- Real outliers are vertically offset for viewing
- With an 80% threshold, the sampling method successfully identified three outliers
- Marginalized outlier probability for the three outliers are:
  - 99.99%, 99.99%, 96.63%

Bayesian Hierarchical Model Example 2: Pulsar Timing Array Analyses

$\delta t = M\varepsilon + Fa + Uj + n$ (Arzoumanian, Brazier, Burke-Spolaor et al., 2016)

$r = \delta t - M\varepsilon - Fa - Uj$



Original Hierarchical Model:

$\delta t \mid b \sim N(Tb, N(\phi))$
$b \mid \phi \sim N(0, B(\phi))$

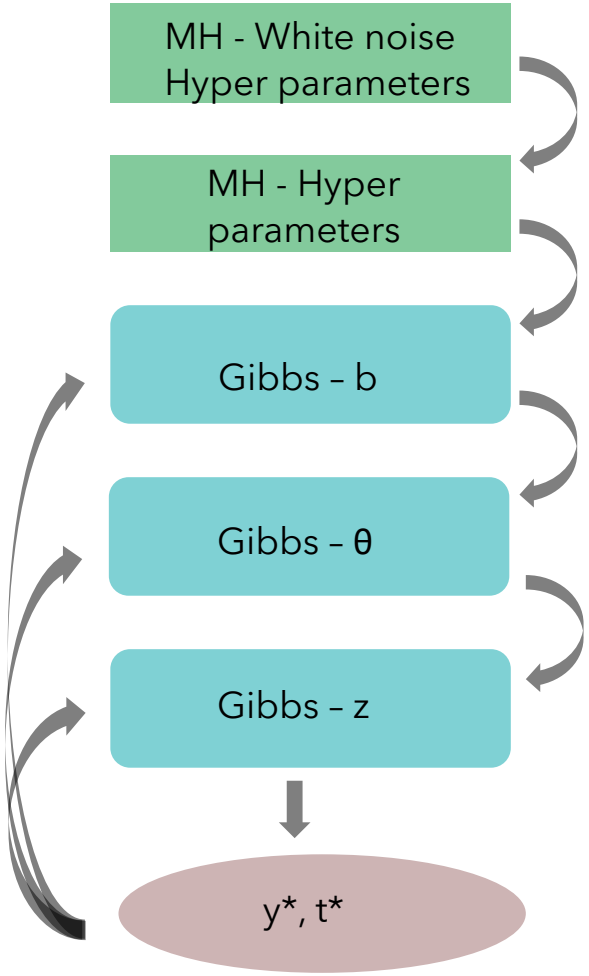
With Outlier Detection:

$\delta t \mid \phi \sim N(Tb, \alpha^z N(\phi))$
$b \mid \phi \sim N(0, B(\phi))$
$z \mid \theta \sim Bern(\theta)$
$\theta \sim$ Uniform/Beta

α: Relative width of outlier distribution to inlier distribution
z: outlier tags
Θ: outlier probability

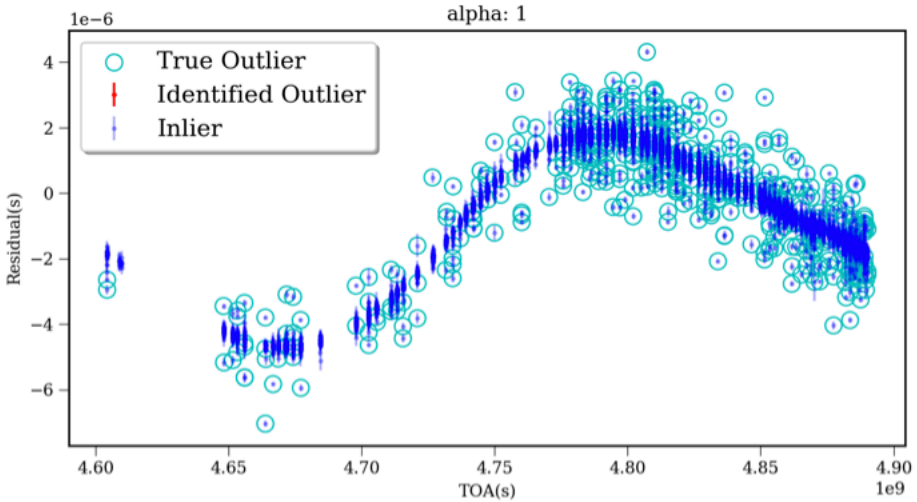
(Tak,Ellis,Ghosh, 2017)

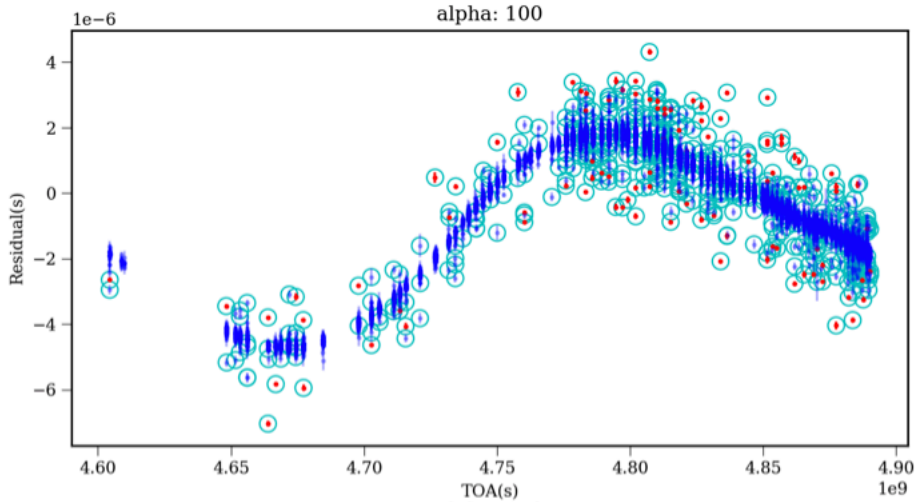
Proposed Solution: Outlier Detection with Gibbs Sampling

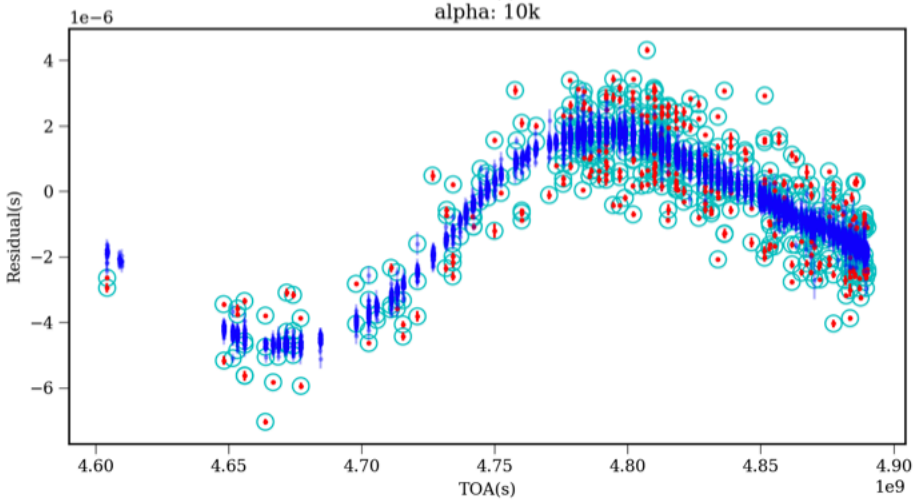


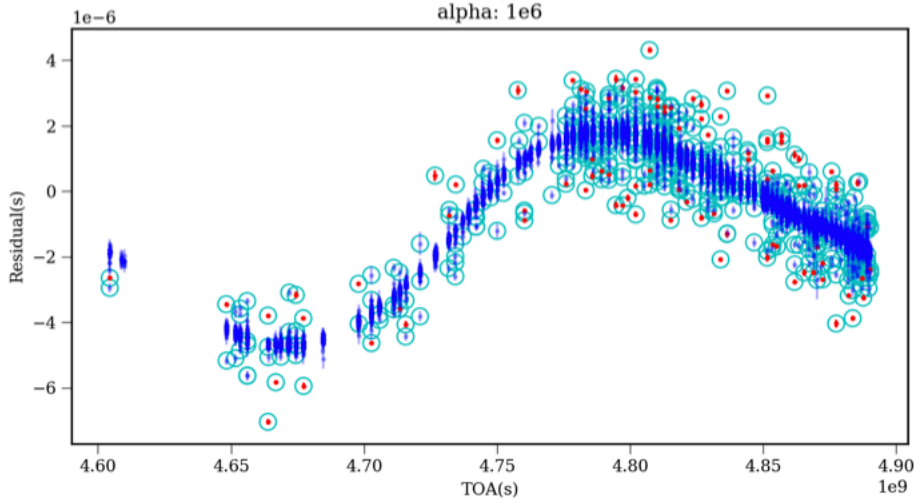
- Update white noise hyper parameters (efac, equad)
- Update RN and ecorr hyper parameters
- Update Fourier coefficients for RN and ecorr
- Update outlier probability
- Update outlier tags
- Outlier tags will "correct" the data and "corrected" data will be further utilized at other iterations

Results on simulated datasets with 5% outlier probability







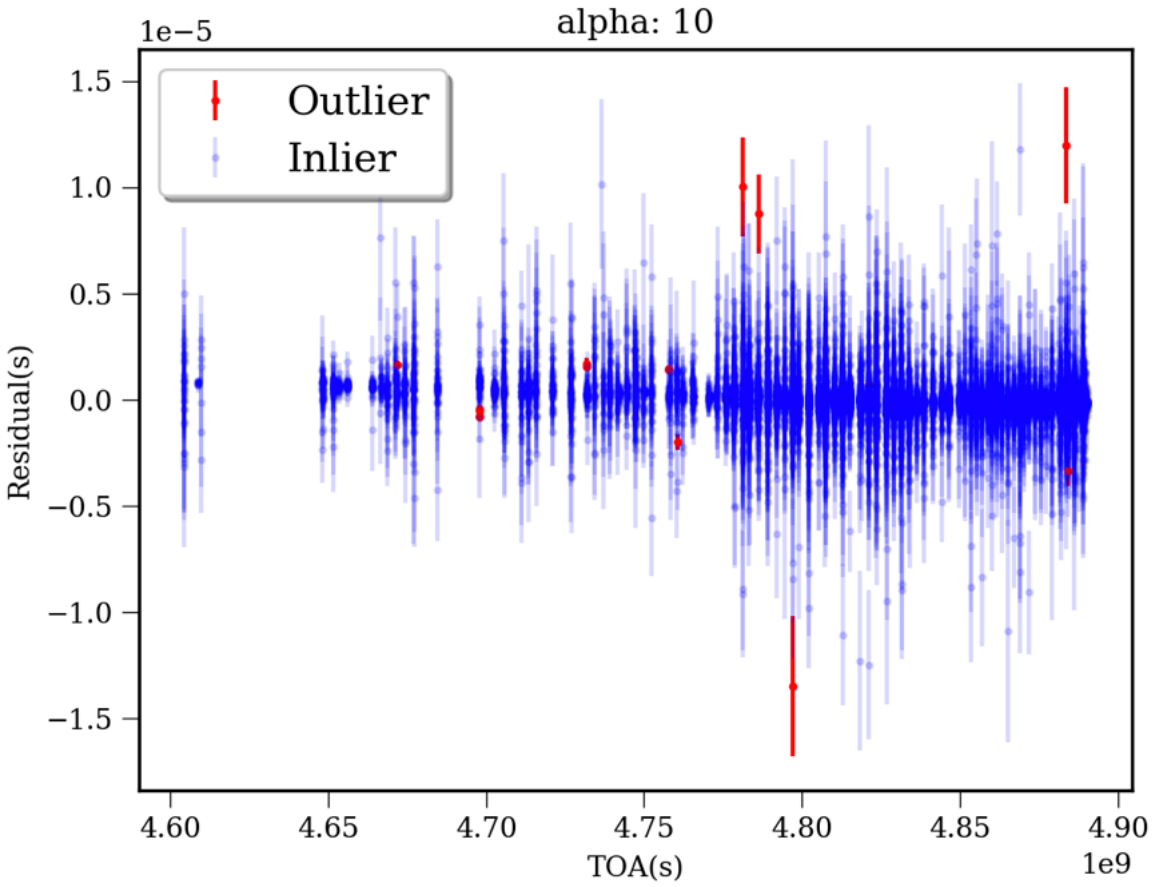


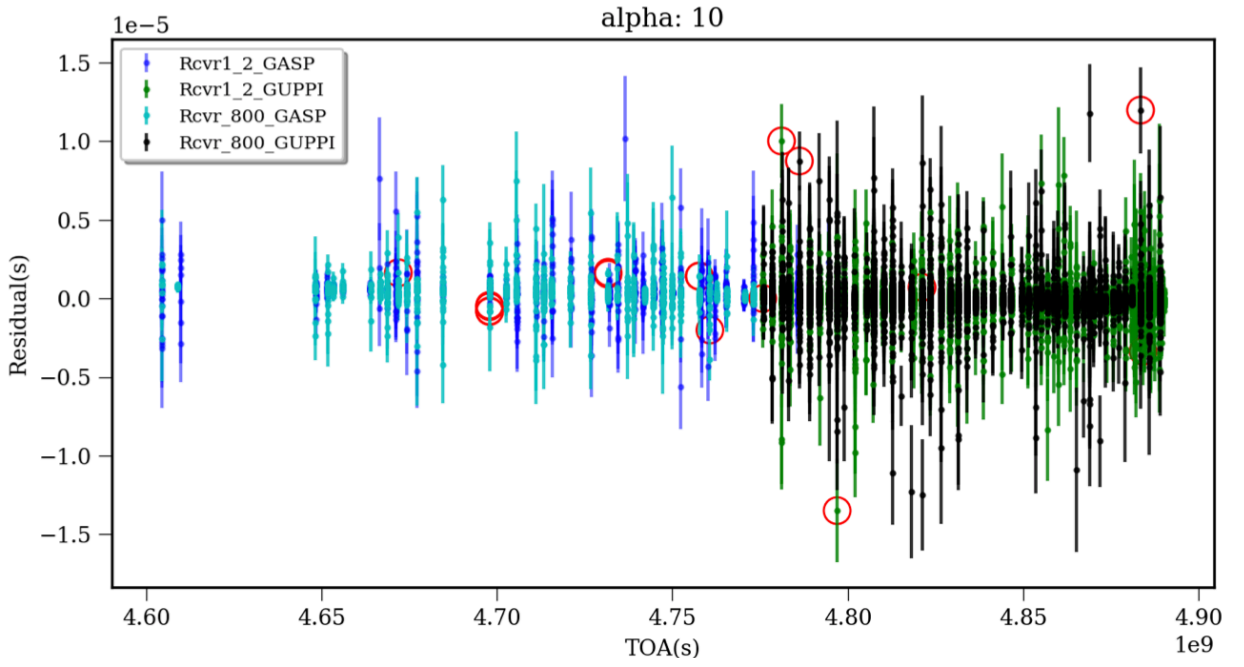
	Threshold				
alpha	50%	80%	90%	95%	99.9%
100	367	343	332	326	282
10k	329	318	310	304	268
1e6	249	240	236	233	97

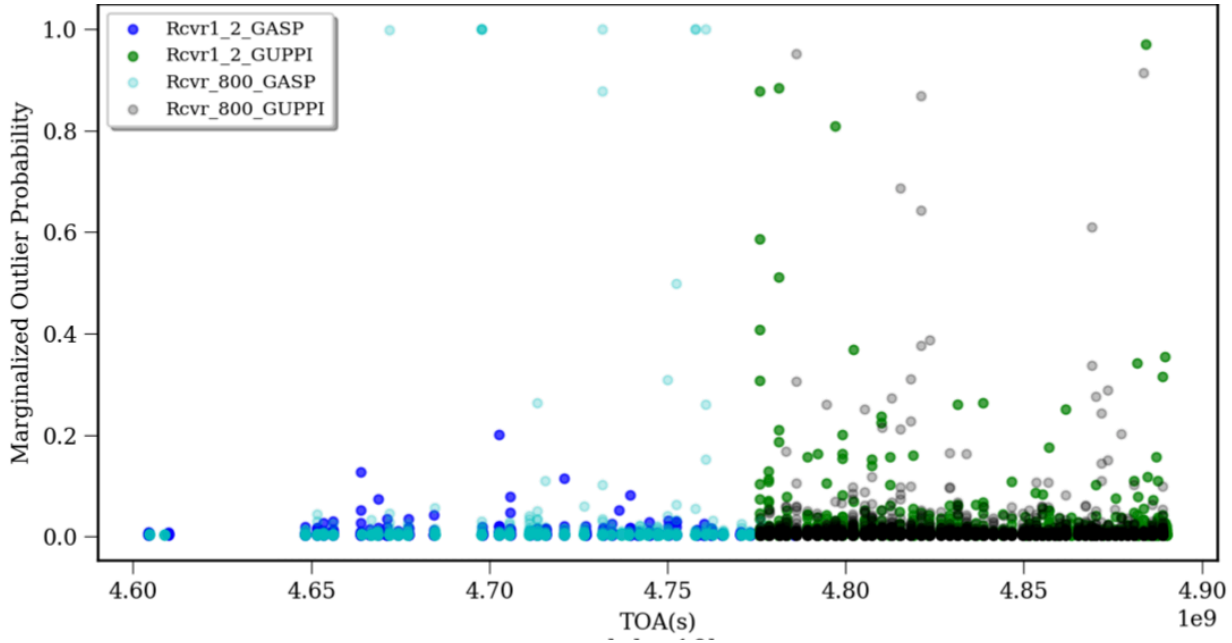
Number of outliers identified based on choices of alpha and threshold

alpha: relative width of outlier distribution to inlier distribution

Results on J1909 – 3744*

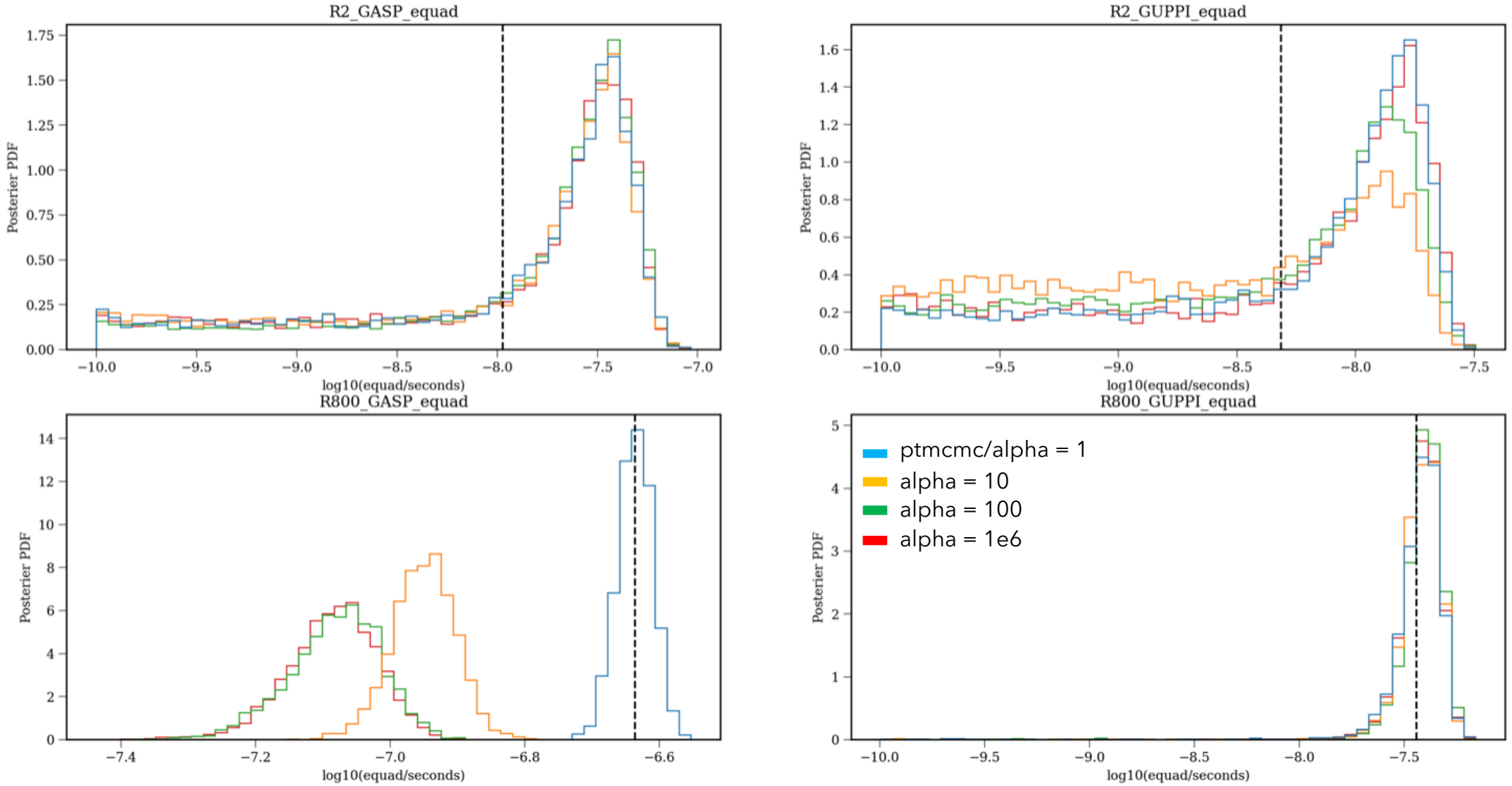






*Data from NANOGrav 9yr Dataset (Arzoumanian. et al, 2015)

equad with alpha 1 v. 10 v. 100 v. 1e6



Thank you!

Acknowledgements

- Professor Stephen Taylor
- Dr. Justin Ellis
- Vanderbilt University Dept. of Physics and Astronomy
- Vanderbilt Data Science Institute