Inclusion Frequencies in Null Setting and Preliminary TF Discovery Results

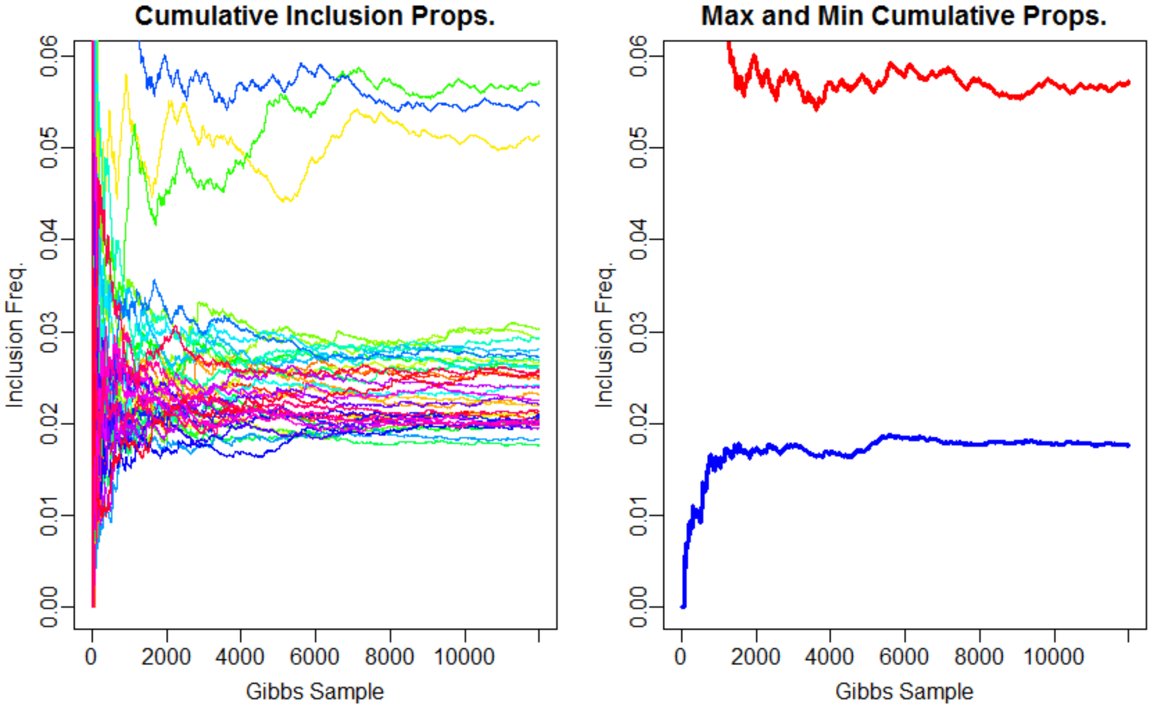
November 15, 2012

For the following simulations, $y_i \overset{iid}{\sim} \mathcal{N}(0,\, \sigma^2)$ and $\boldsymbol{X}$ consists of entries that are all $x_{ij} \overset{iid}{\sim} \mathcal{N}(0,\, 1)$. Shown below are the cumulative inclusion probabilities. The burn-in is purposely excluded to see the behavior of an entire Gibbs chain. 12000 Gibbs samples were generated with no thinning.

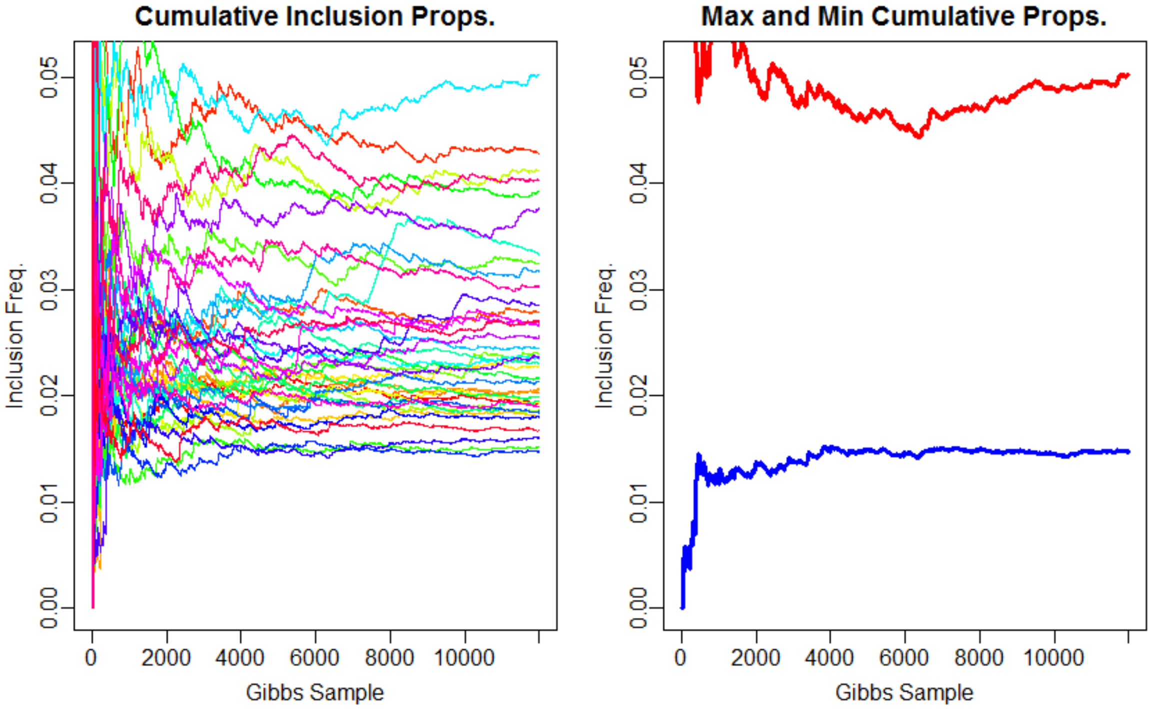
A number of different settings for σ and n will be included. The number of covariates is 39, which is the same as the number of TFs in the real data. Hence, the expectation might be that inclusion frequencies would converge to .0256.

1 σ=.1

N=250

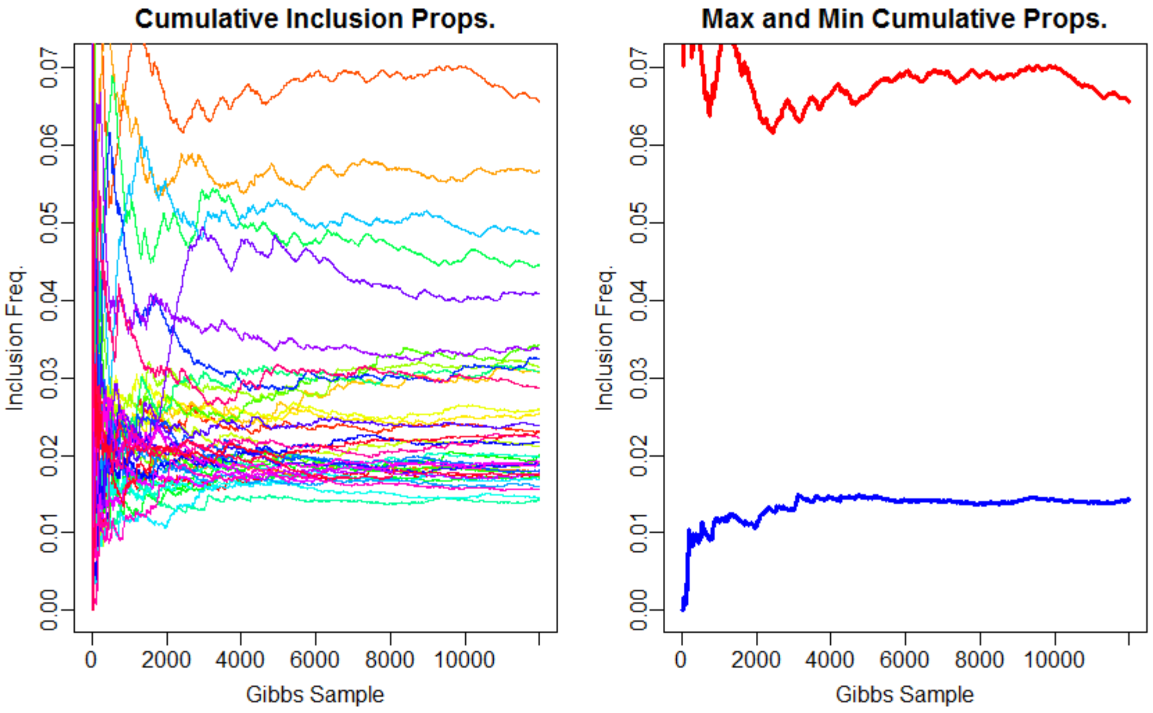


N=1000

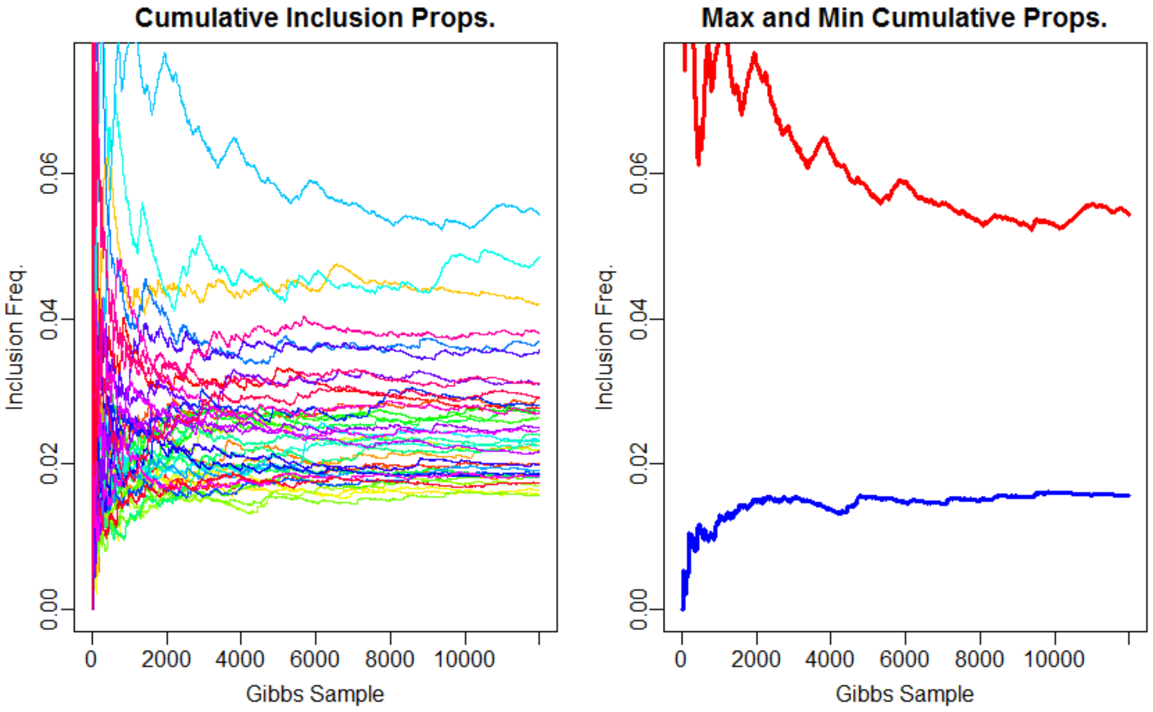


2 σ=1

N=250

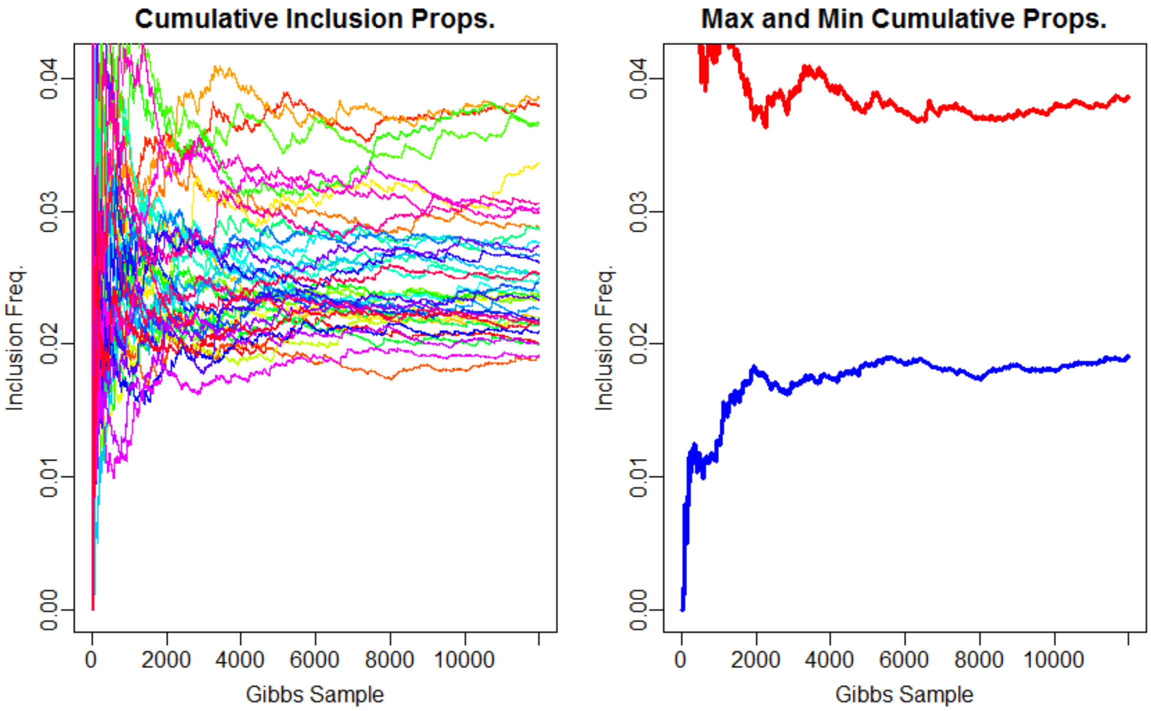


N=1000

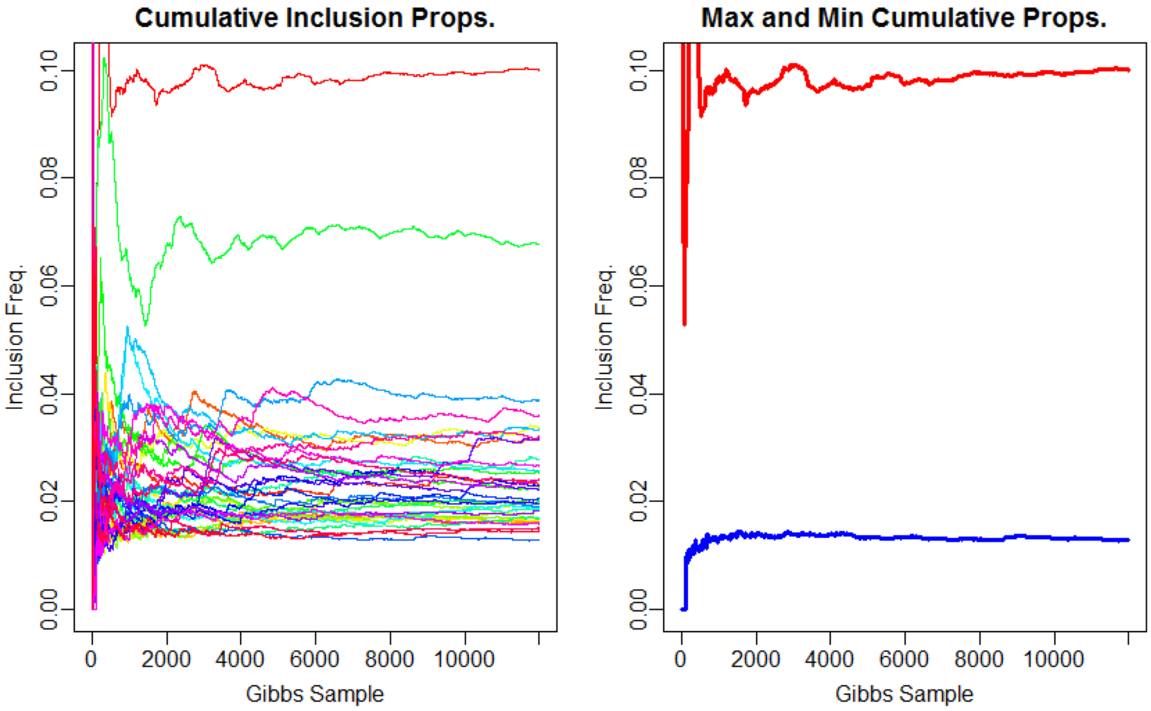


3 σ=2

N=250

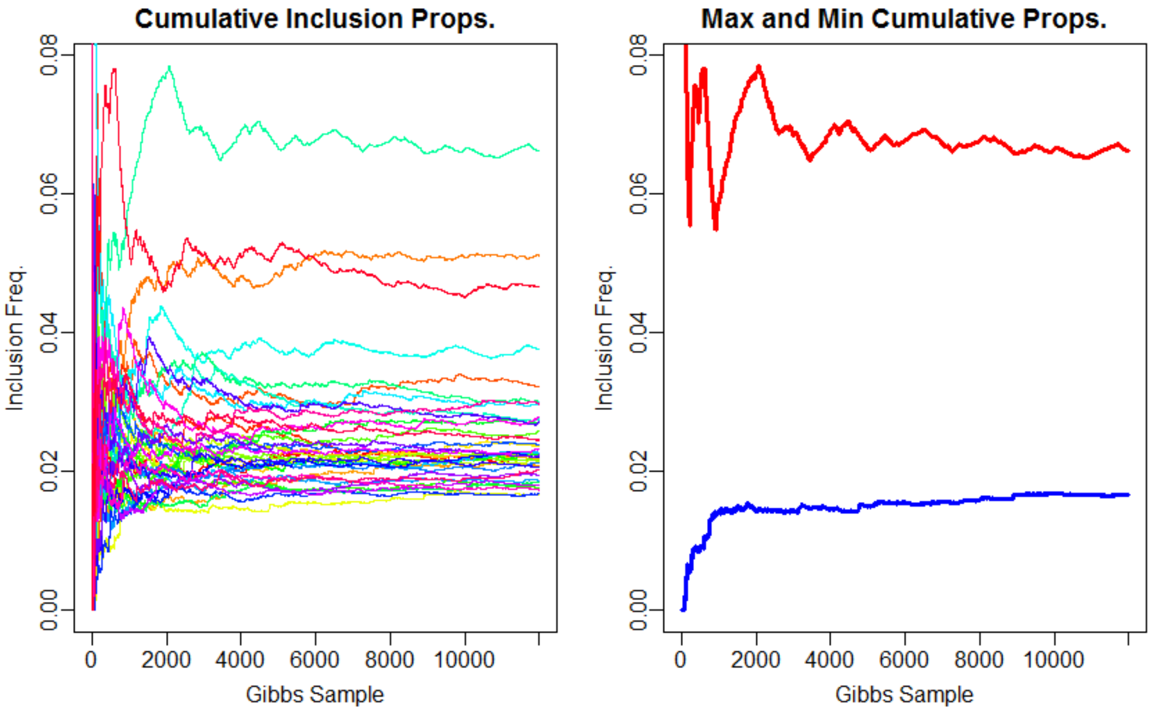


N=1000

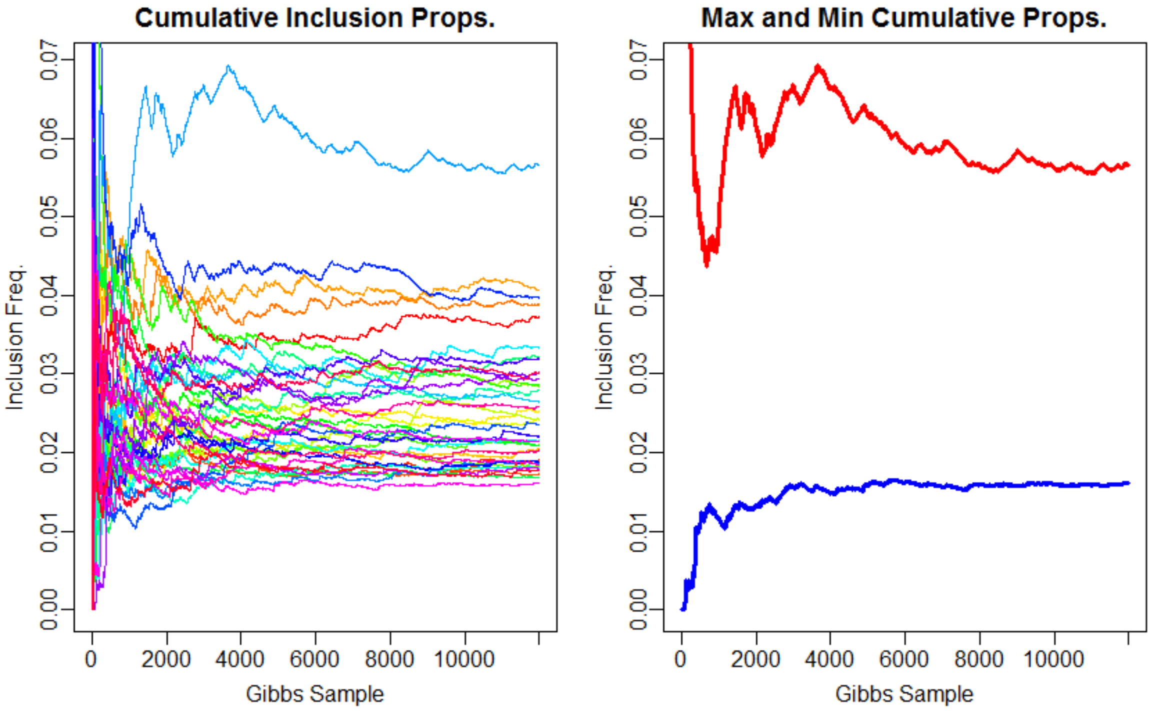


4 σ=8

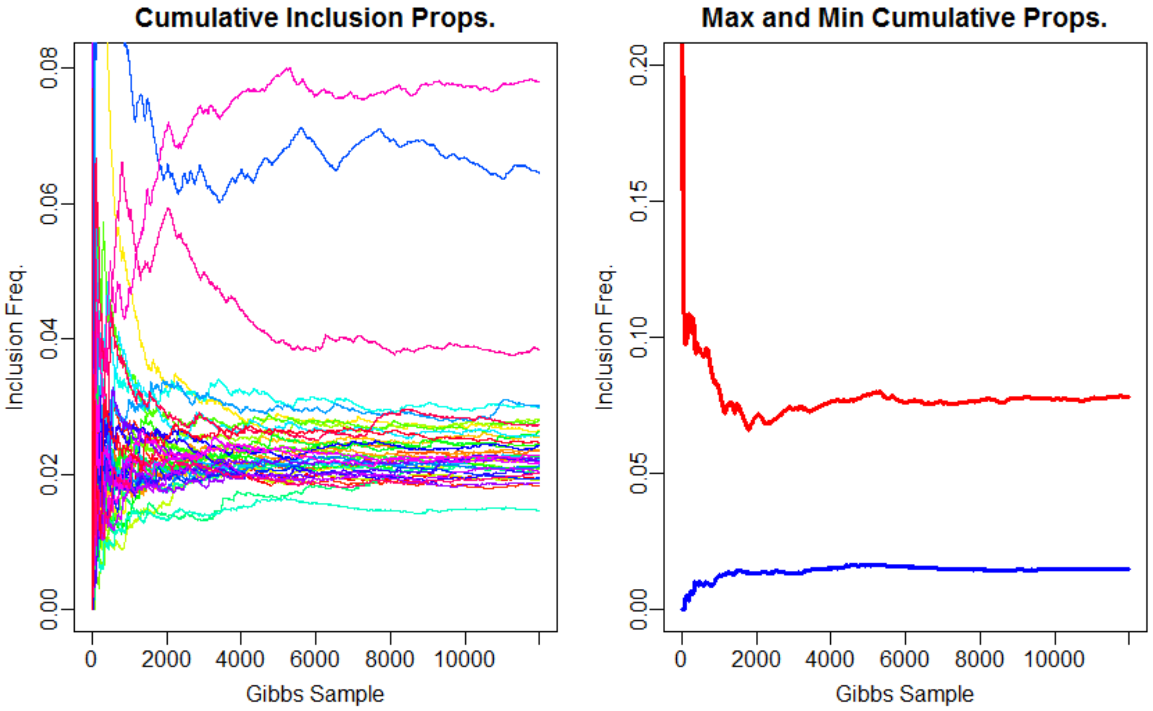
N=250



N=1000



5 4 Aggregated Chains-Multiple Starting Points

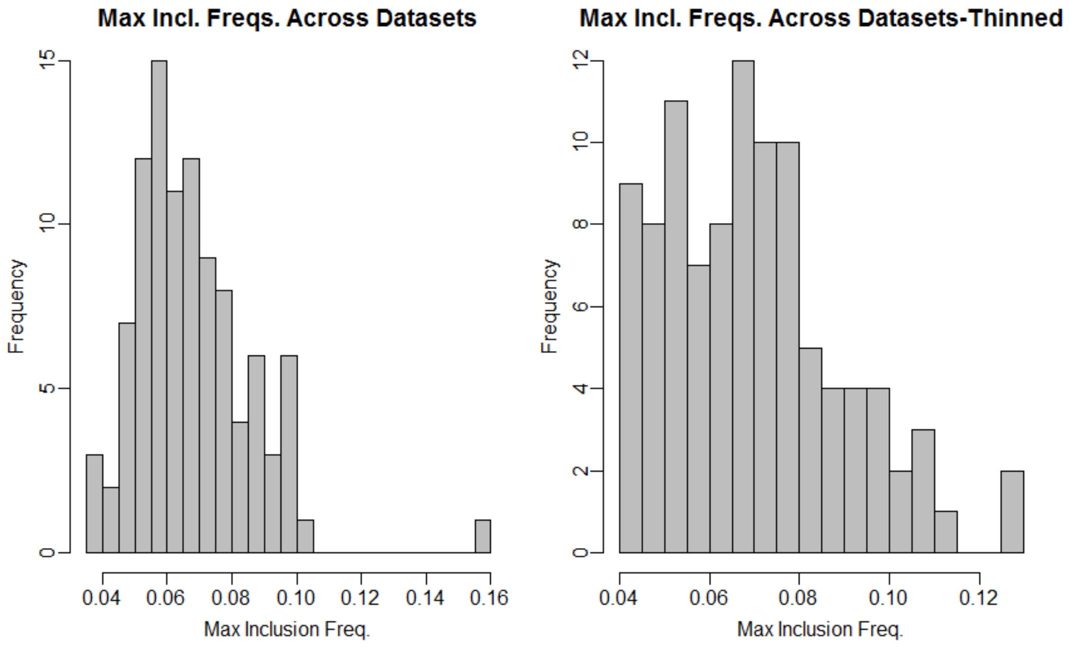


Hypothesis: From the above plots as well as the attempt to use multiple chains, it seems that autocorrelation is not necessarily the root of the problem. The issue seems to be more linked to the existence of posterior modes that BART is discovering. These modes tend to move around from dataset to dataset, but nonetheless, the posterior probability surface doesn't seem to be flat enough to allow the algorithm to wander randomly and create a uniform distribution on the splitting rules.

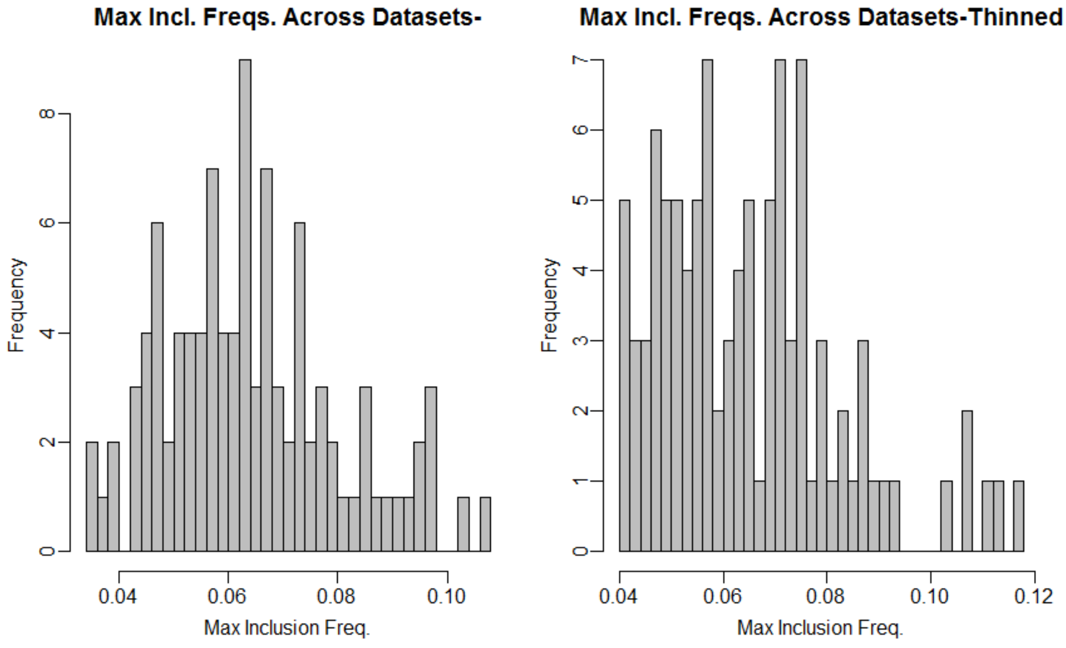
6 Maximum Inclusion Frequencies Across Datasets (N=100)

2000 Burn-in and 5000 posterior samples. 10 Trees. Different $\boldsymbol{y}$ vector for each of 100 iterations. Thinning is by 25.

$\sigma = 1$



$\sigma = 8$



7 TF Discoveries-500 Genes

1000 Burn-in and 2000 posterior samples at 10 trees with no thinning. 100 bootstrap iterations for null setting. Algorithm ran in 2 hours on 20 cores.

Shown below are the various proportions of times a TF was “discovered” in the 500 genes examined.

No Simultaneous Coverage

```
ABF1  ACE2  BAS1  CAD1  CBF1  FKH1  FKH2  GAL4  GCN4  GCR1  GCR2
0.148 0.072 0.250 0.260 0.176 0.150 0.180 0.102 0.316 0.122 0.158
 HAP2  HAP3  HAP4  HSF1  INO2  LEU3  MBP1  MCM1 MET31  MSN4  NDD1
0.074 0.222 0.118 0.308 0.176 0.152 0.184 0.136 0.158 0.298 0.098
 PDR1  PHO4  PUT3  RAP1  RCS1  REB1  RLM1  RME1  ROX1  SKN7  SMP1
0.146 0.104 0.076 0.284 0.056 0.174 0.074 0.178 0.070 0.262 0.108
 STB1 STE12  SWI4  SWI5  SWI6  YAP1
0.148 0.166 0.130 0.162 0.170 0.286
```

Simultaneous Coverage Bands

```
 ABF1  ACE2  BAS1  CAD1  CBF1  FKH1  FKH2  GAL4  GCN4  GCR1  GCR2
0.038 0.006 0.072 0.056 0.026 0.024 0.014 0.008 0.116 0.008 0.022
 HAP2  HAP3  HAP4  HSF1  INO2  LEU3  MBP1  MCM1 MET31  MSN4  NDD1
0.008 0.054 0.014 0.070 0.022 0.016 0.046 0.008 0.008 0.060 0.006
 PDR1  PHO4  PUT3  RAP1  RCS1  REB1  RLM1  RME1  ROX1  SKN7  SMP1
0.010 0.010 0.006 0.066 0.004 0.022 0.002 0.028 0.006 0.058 0.004
 STB1 STE12  SWI4  SWI5  SWI6  YAP1
0.024 0.024 0.020 0.044 0.030 0.064
```

The maximum cut-off method is omitted, but similar to the simultaneous coverage scenario, which is clearly too restrictive. There is some inherent Type I error control build into this model as their is a constraint on the total budget on inclusion frequencies.