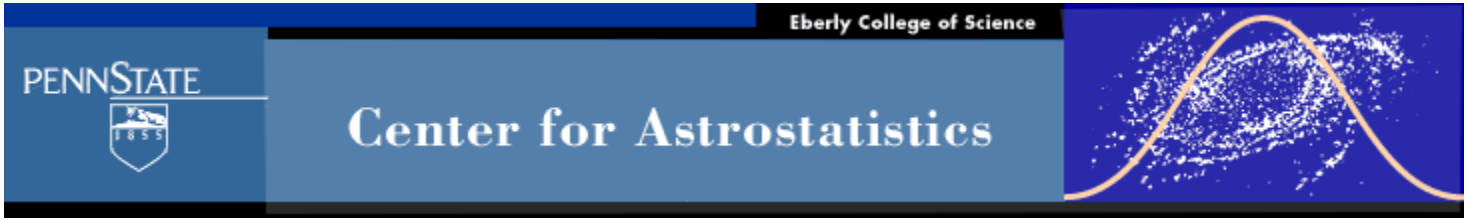
Convergence Diagnostics For Markov chain Monte Carlo

Eric B. Ford (Penn State)

Bayesian Computing for Astronomical Data Analysis

June 5, 2015



MCMC: A Science & an Art

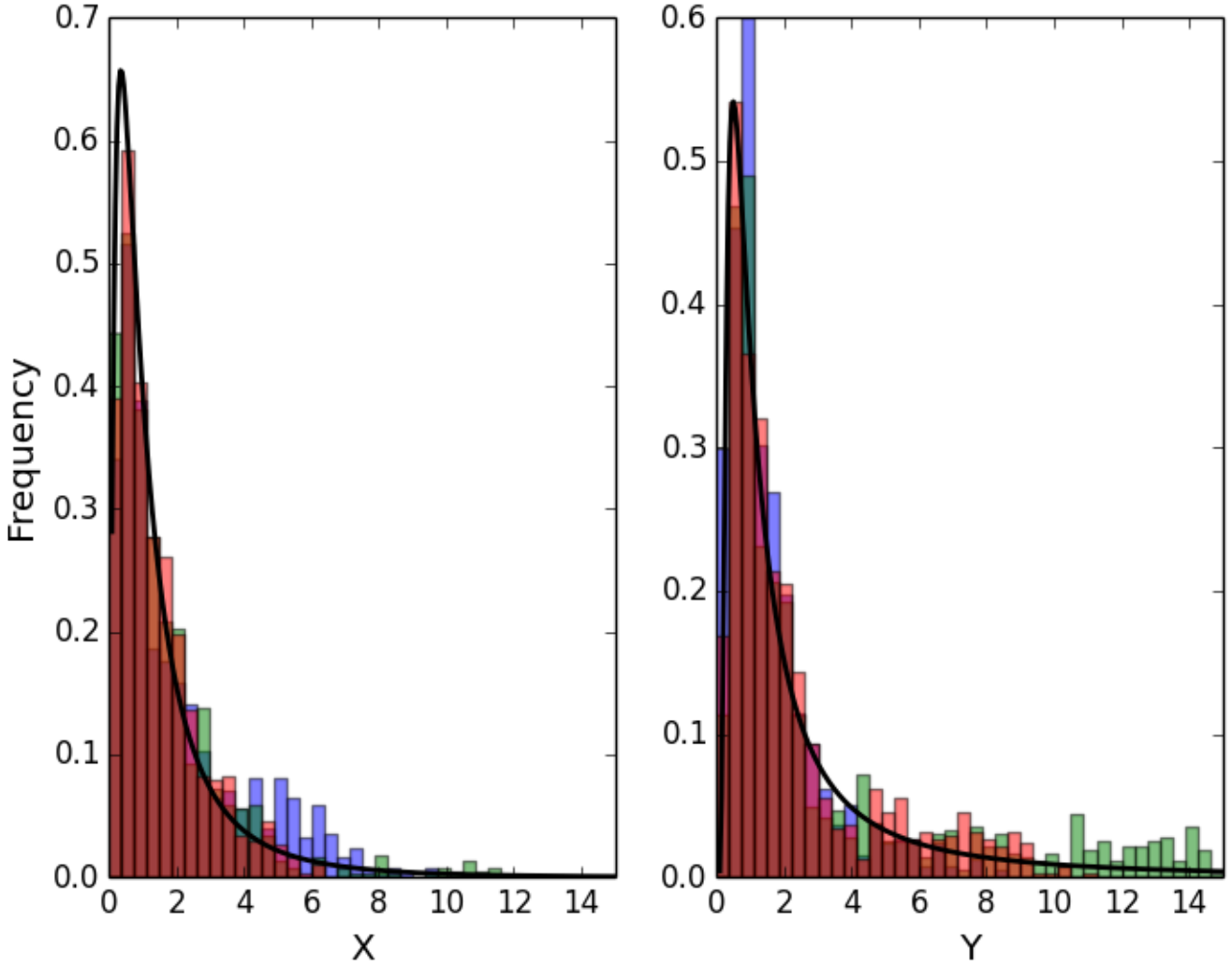
- Science:
  If your algorithm is designed properly, the Markov chain will converge to the target distribution… after infinite iterations
- Art:
  When is it wise to make inferences based on a finite Markov chain

Assessing Convergence is Essential

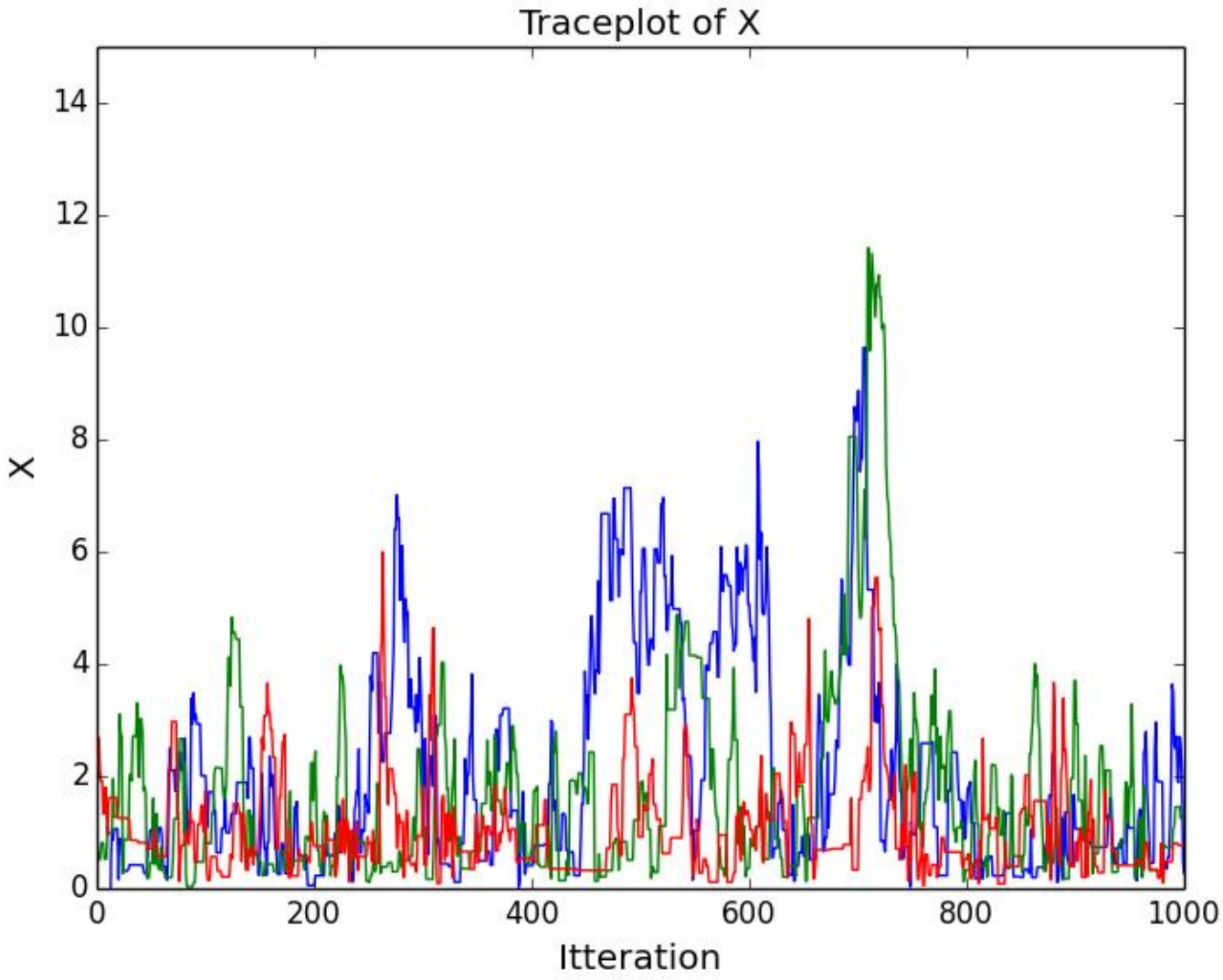
If you want to:

- Base your conclusions on posterior distributions
- Report accurate parameter estimates & uncertainties
- Avoid fooling yourself
- Avoid devoting resources (e.g., your effort, telescope time) to follow-up an “inference” that isn’t supported by data
- Avoid writing an erratum to your paper

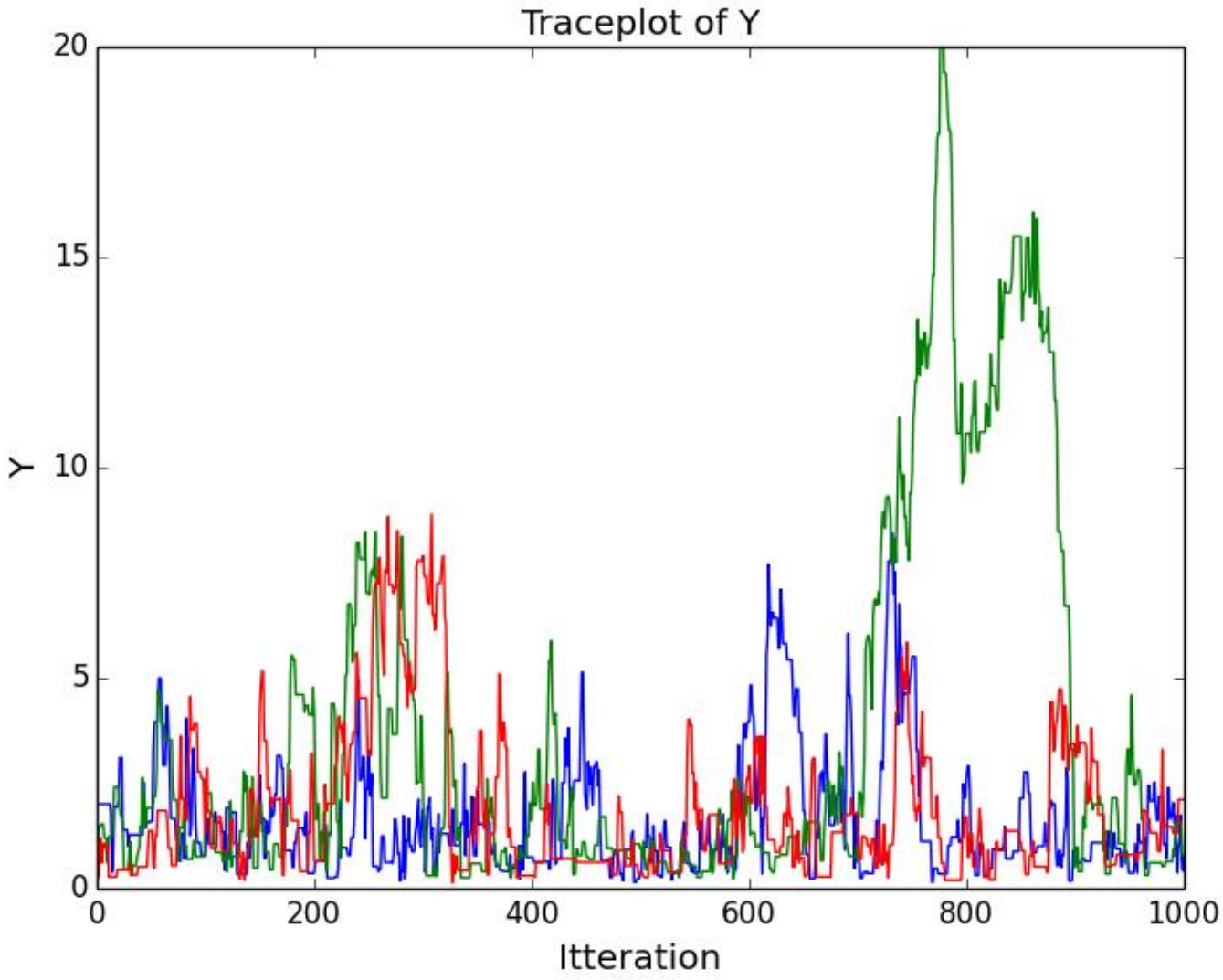
Has this Chain Converged?



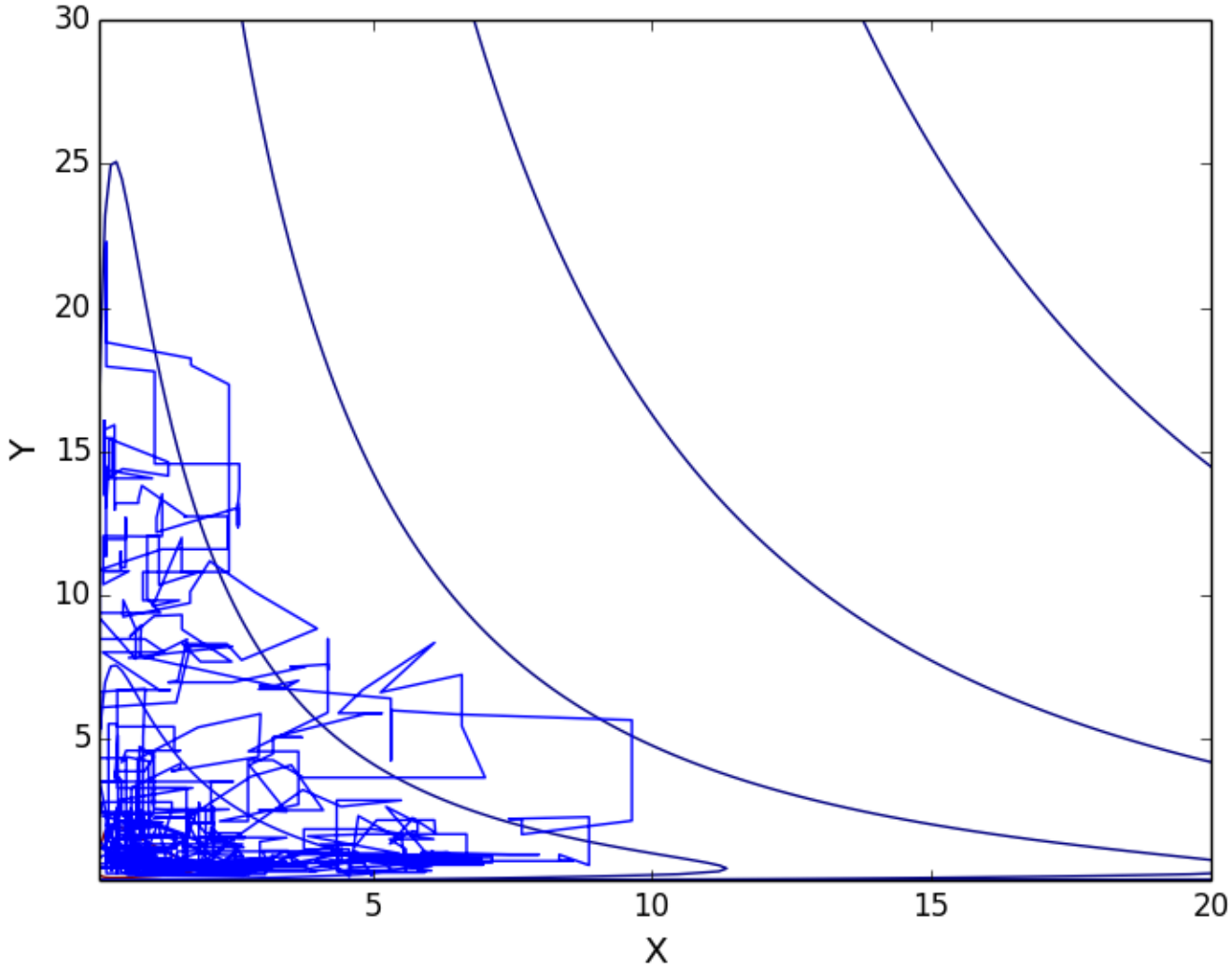
Has this Chain Converged?



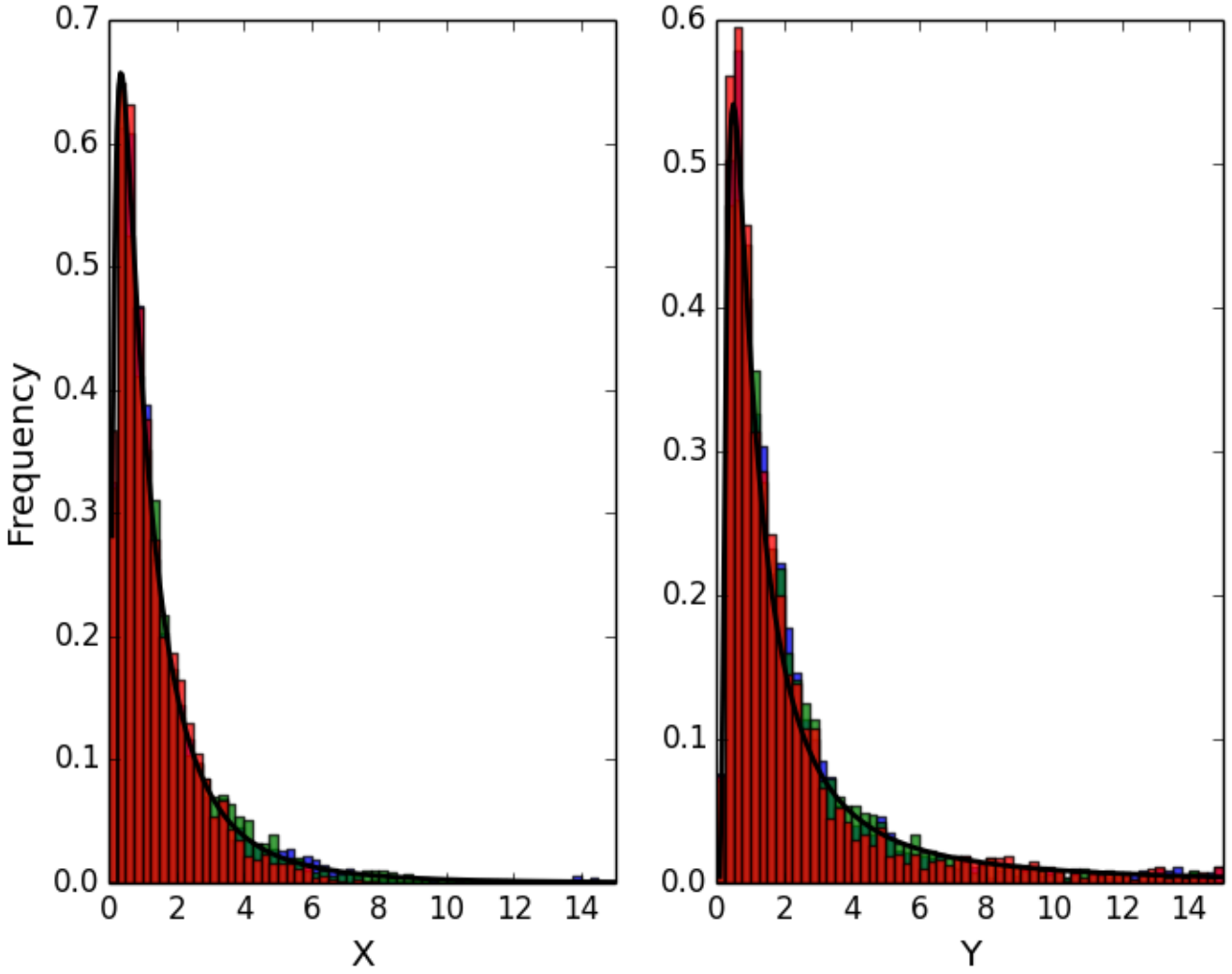
Has this Chain Converged?



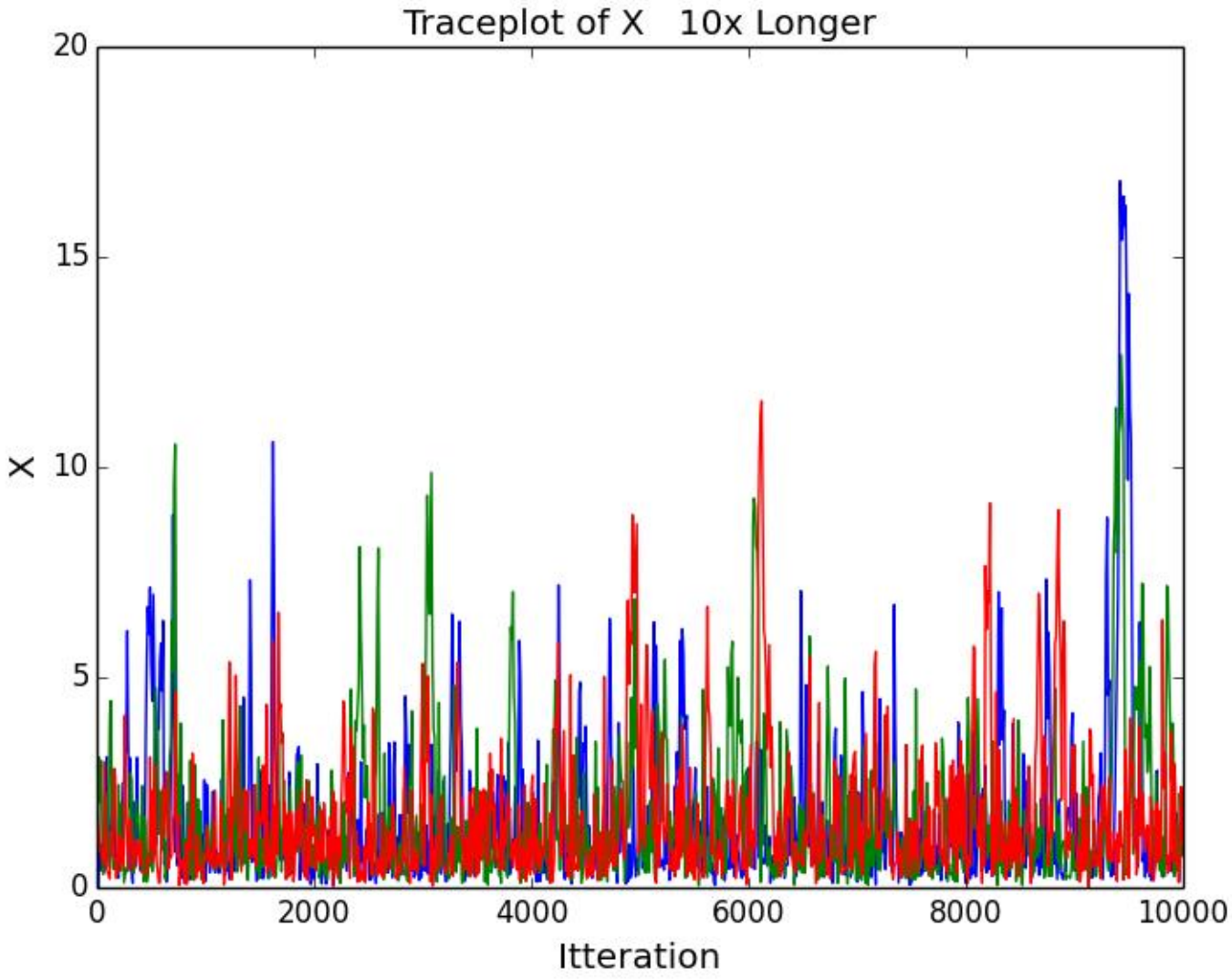
Has this Chain Converged?



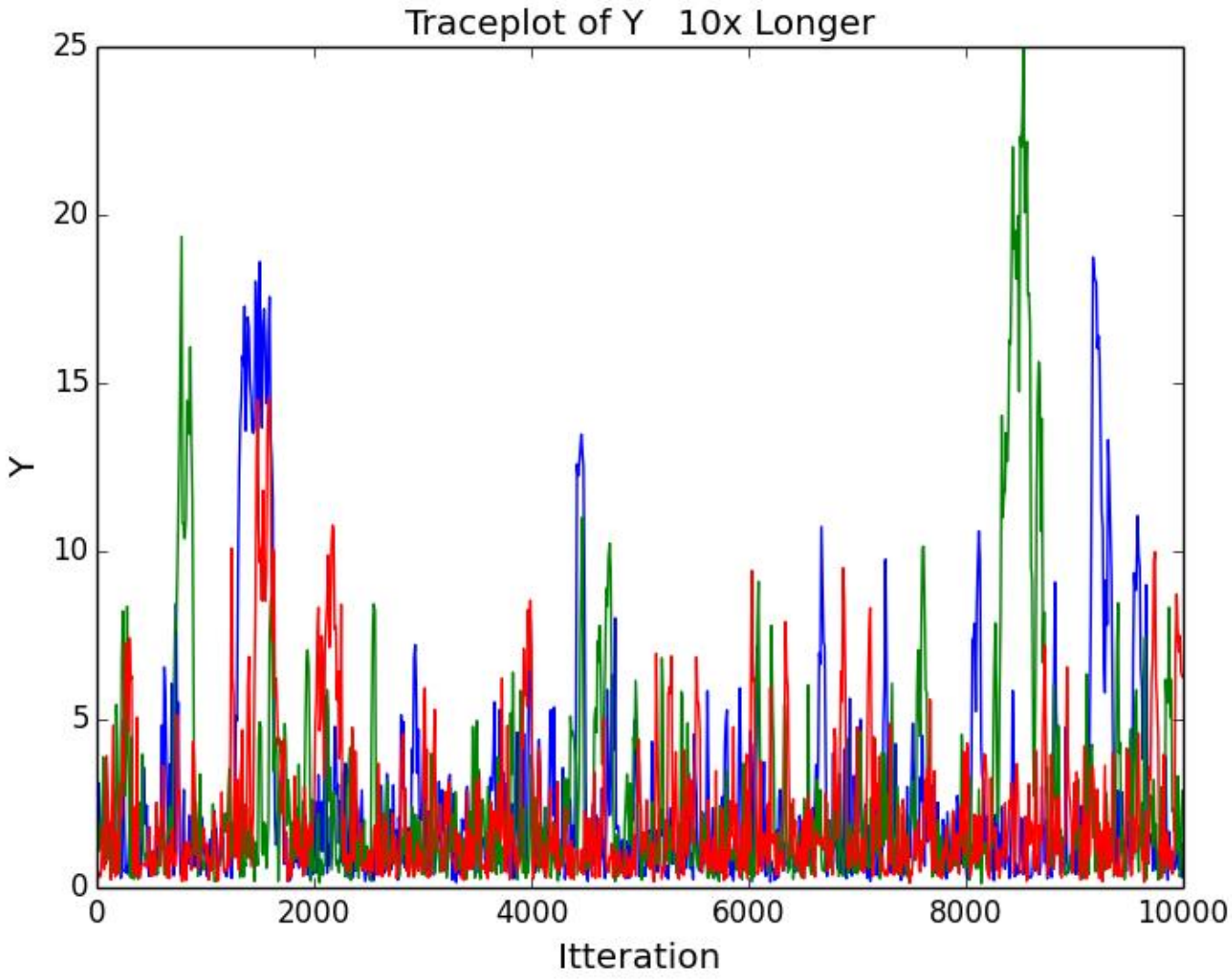
Has this Chain Converged?



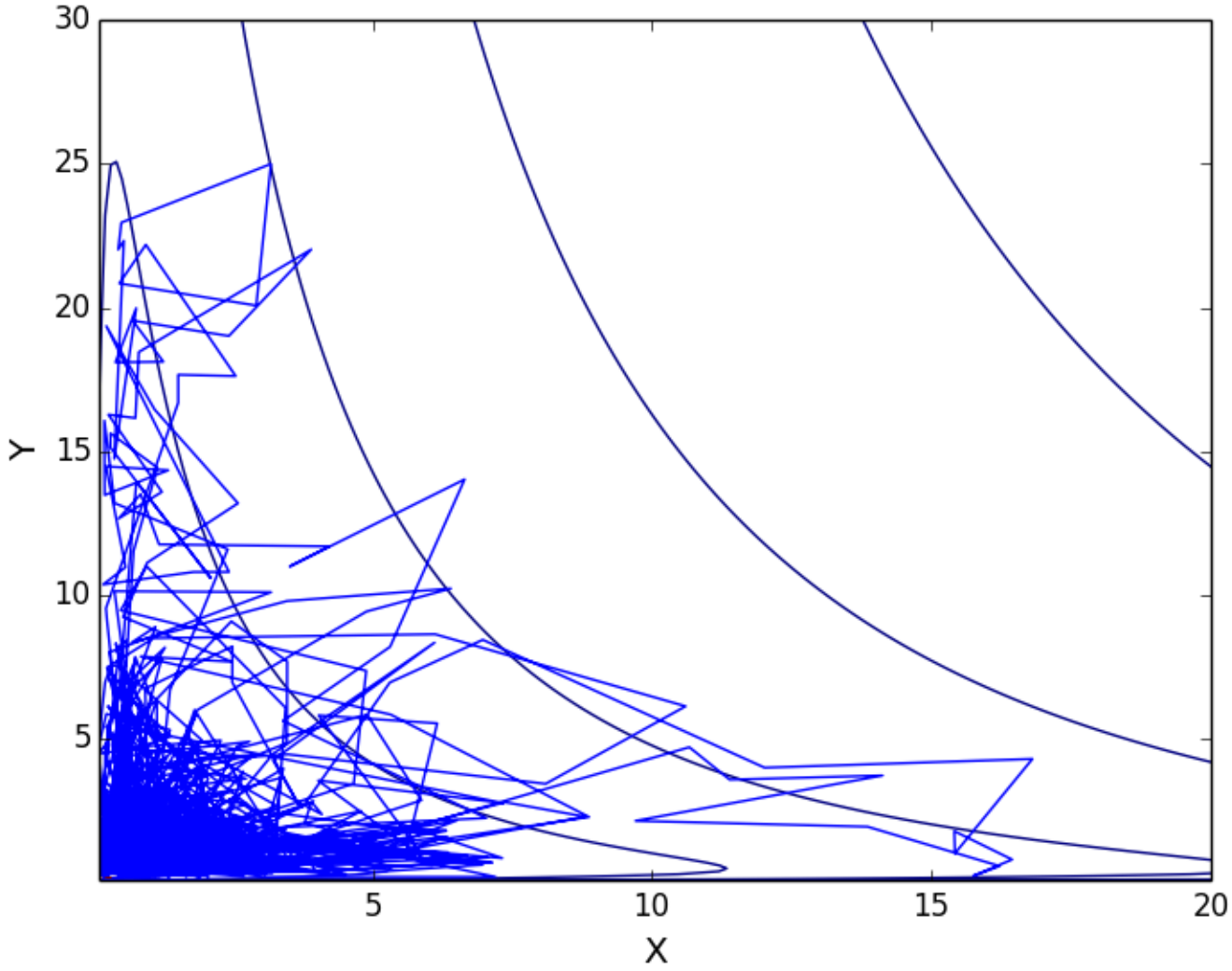
Has this Chain Converged?



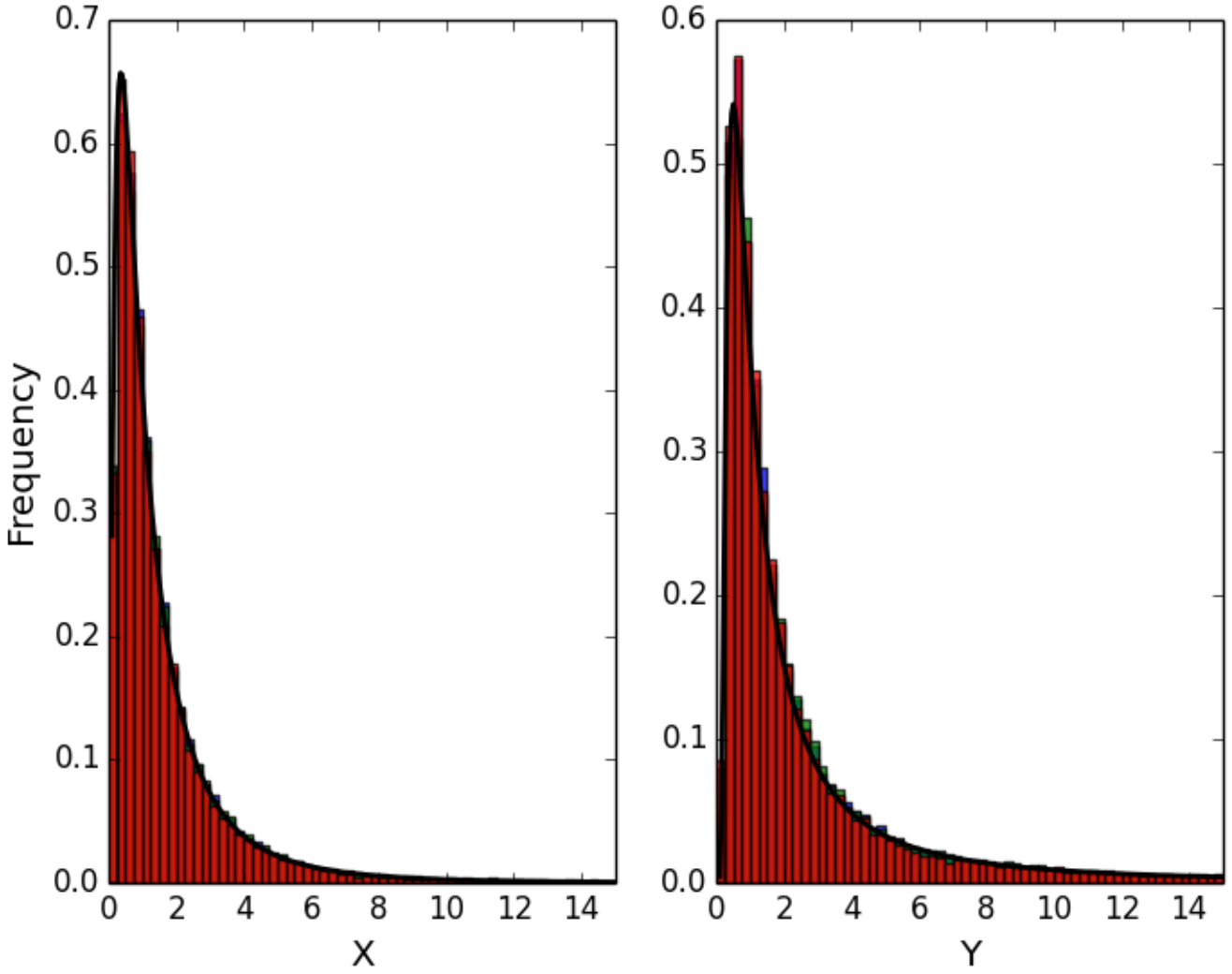
Has this Chain Converged?



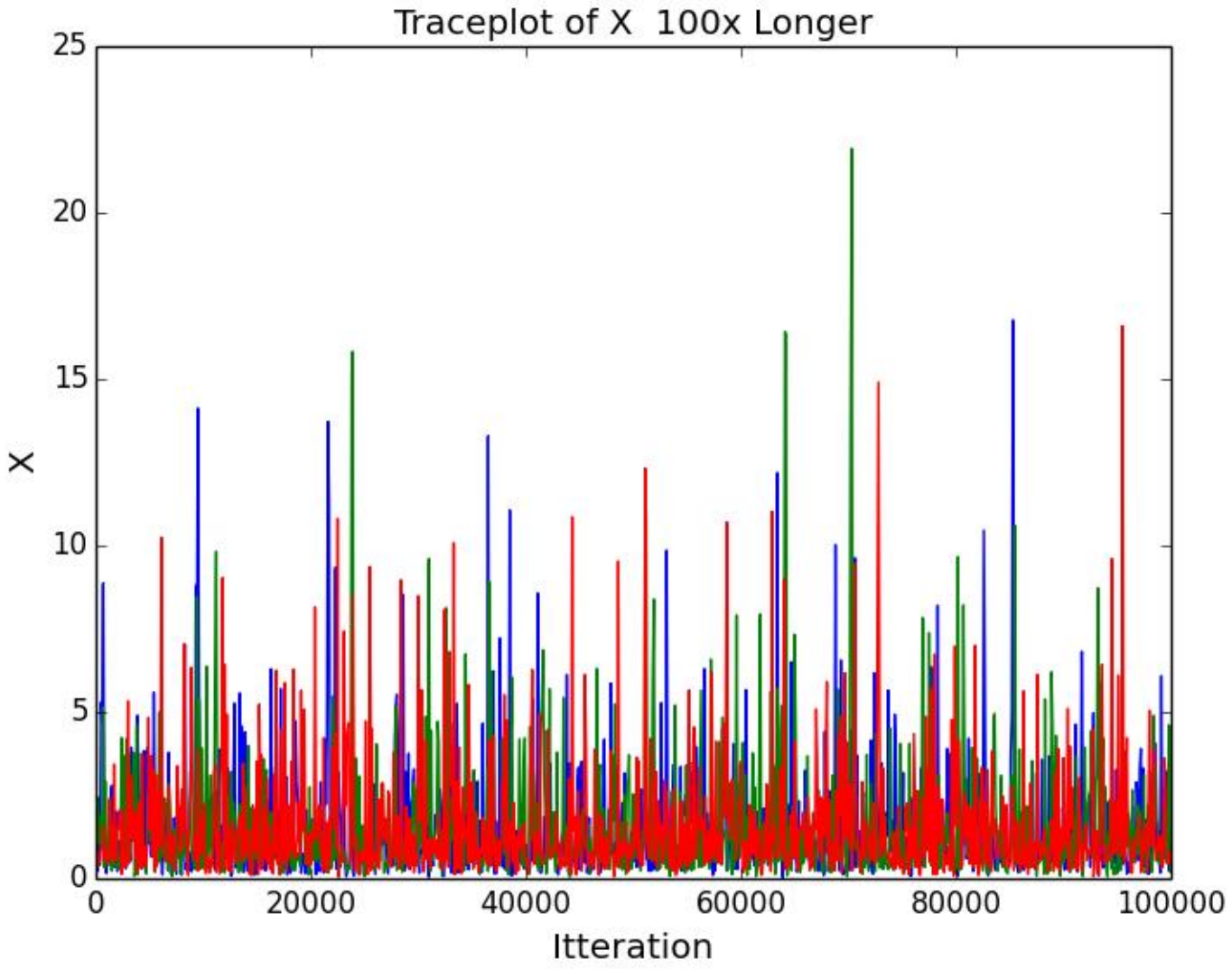
Has this Chain Converged?



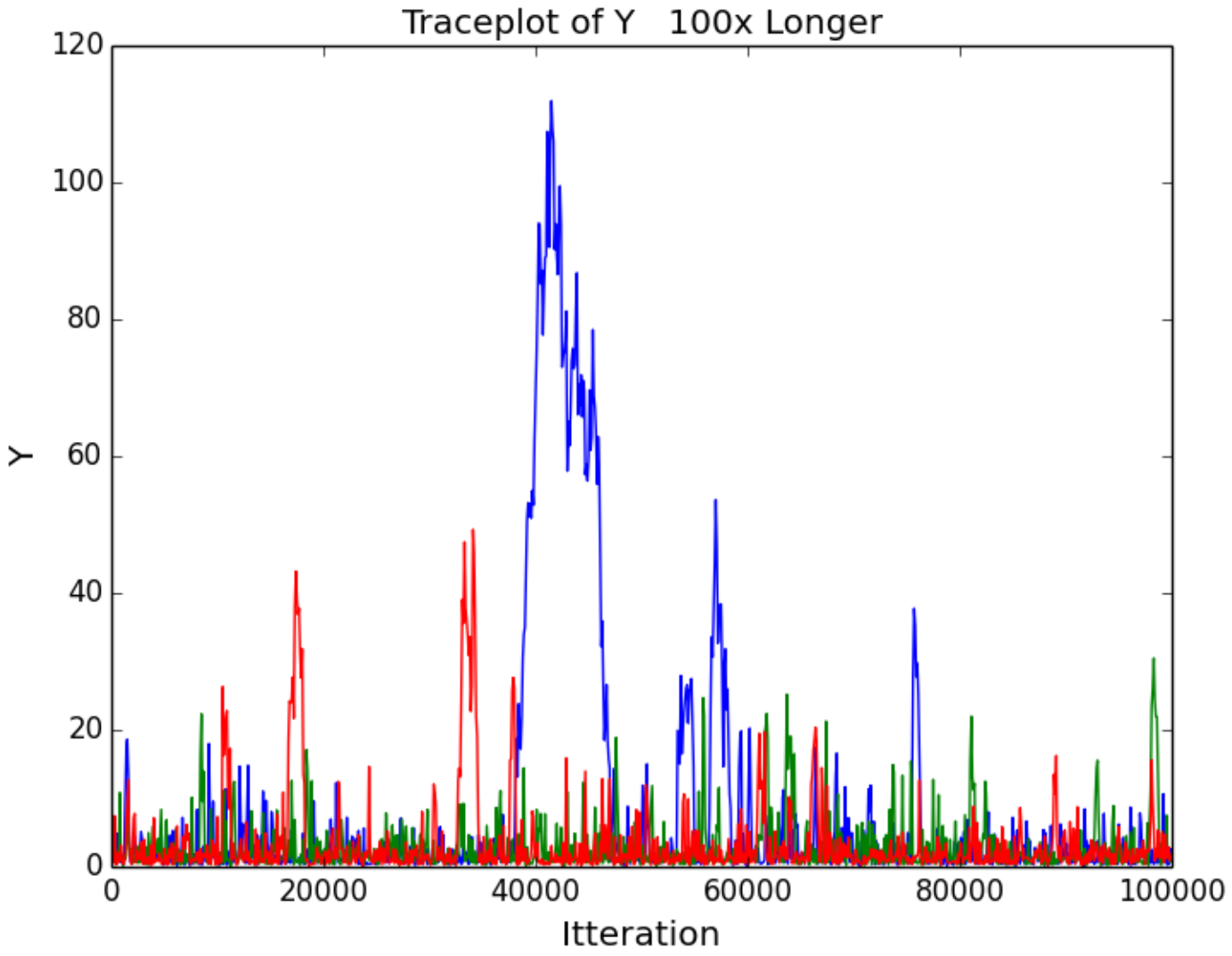
Has this Chain Converged?



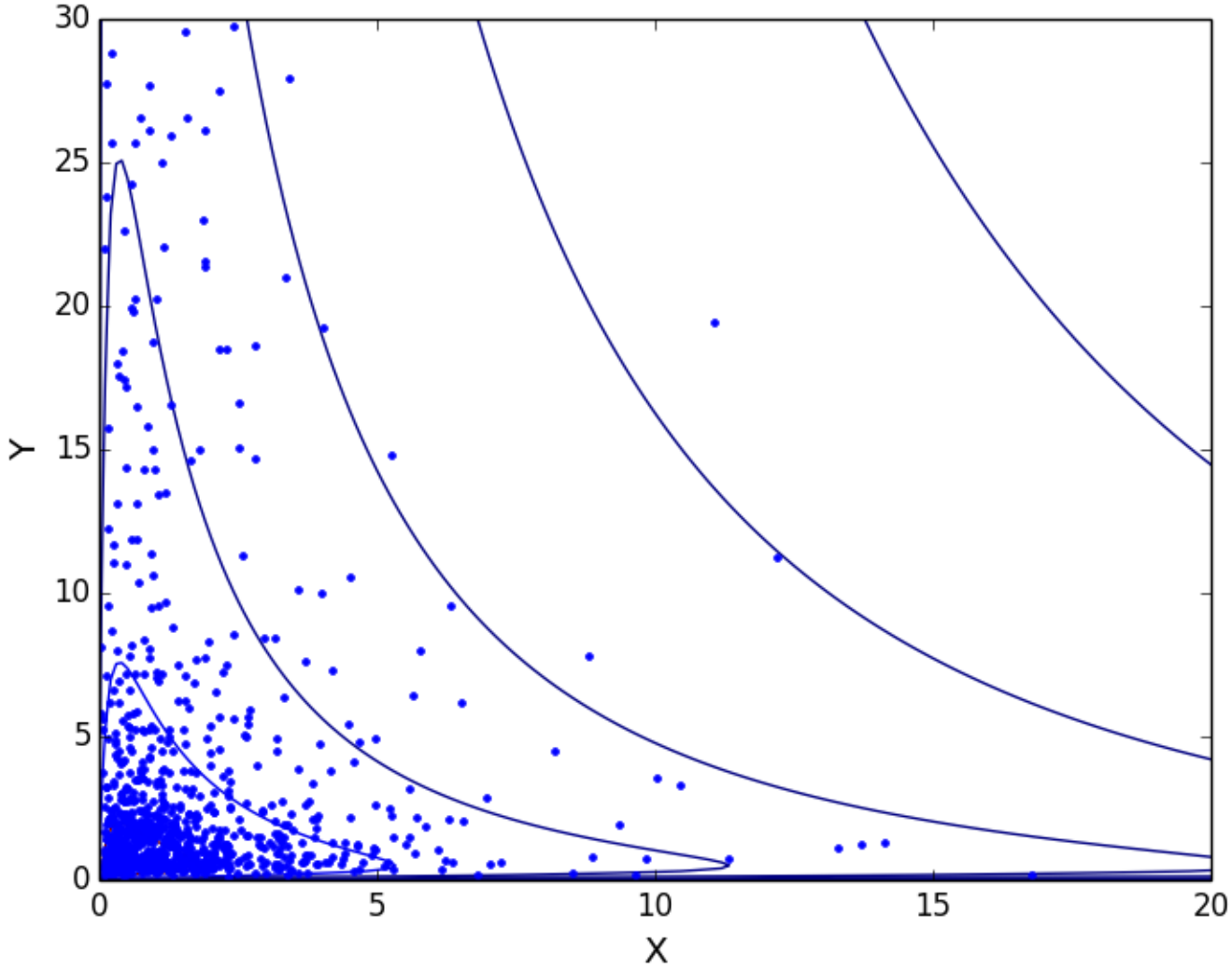
Has this Chain Converged?



Has this Chain Converged?



Has this Chain Converged?



Good Signs

- Any sufficiently large segment of Markov chain would give similar results
- Desirable acceptance rate of proposed steps
- Chain “mixes well”. (i.e., chain has run much longer than any observed timescale for correlation between samples)
- Multiple chains initialized from different initial conditions give similar results
- MCMC analysis of similar problem using simulated data give accurate results, even with significantly fewer iterations

Why Only “Maybe”?

- You can’t prove convergence
  - At best, you fail to prove a failure to converge
- Convergence rate can be exponentially sensitive to barriers between local modes.
- What if there’s a narrow bottleneck between two regions of high probability?
- What if there’s another posterior mode that we’ve completely overlooked?

What should I do?

- Be paranoid
- Run chains longer than you think you need to
- Compute several Markov chains
  - initialized with significantly different initial states
- Look at your Markov chains yourself
  - Trace plots
  - Marginal joint densities

What warning signs should I be looking for?

- Differences within or across Markov chains
- “Poor mixing”
- Low/high acceptance rates
- Autocorrelation between states of Markov chain
- Strongly correlated parameters
- Suspicious tails or posterior shapes

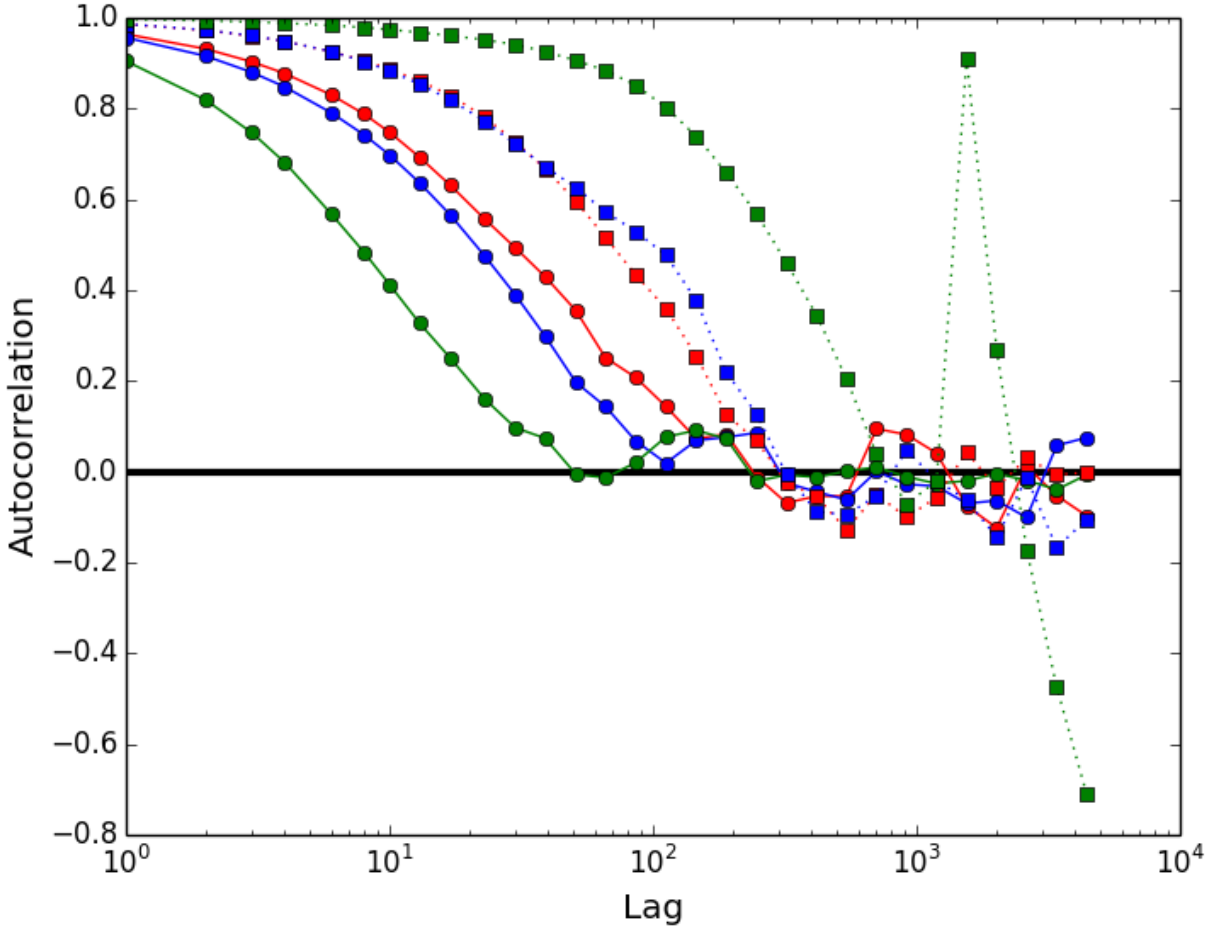
Check Autocorrelation of Markov chain

- Autocorrelation as a function of lag

$$\rho_{lag} = \frac{\sum_i^{N-lag}(\theta_i - \bar{\theta})(\theta_{i+lag} - \bar{\theta})}{\sum_i^{N}(\theta_i - \bar{\theta})^2}$$

- What is smallest lag to give an $\rho_{lag} \approx 0$?
- One of several methods for estimating how many iterations of Markov chain are needed for *effectively* independent samples

Checking Autocorrelation Function



Getting More Quantitative

Calculate convergence diagnostics

- Geweke (1992): Compares means calculated from distinct segments of Markov chain
- Raftery & Lewis (1992): Estimates the *minimum* chain length needed to estimate a percentile to some precision

- Gelman & Rubin (1992): $\hat{R}$ compares variances between chains
- Brooks & Gelman (1998): Several generalizations of $\hat{R}$
  - Account for covariances
  - Can apply to higher moments
  - Scale reduction for arbitrary credible intervals

Estimate Potential Scale Reduction Factor

Gelman-Rubin diagnostic ($\hat{R}$)

- Compute *m* independent Markov chains
- Compares variance of each chain to pooled variance
- If initial states (θ_{1j}) are overdispersed, then $\hat{R}$ approaches unity from above
- Provides estimate of how much variance could be reduced by running chains longer
- It is an *estimate!*

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \qquad \bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m} \bar{\theta}_j$$

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2 \qquad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2$$

$$\hat{\mathrm{Var}}(\theta) = (1 - \frac{1}{n})W + \frac{1}{n}B \qquad \hat{R} = \sqrt{\frac{\hat{\mathrm{Var}}(\theta)}{W}}$$

Estimate Potential Scale Reduction Factor

Bare Minimum:

- Check $\hat{R}$ for each model parameter
- Check $\hat{R}$ for any important functions of model parameters

Better:

- Consider applying a generalization that checks for covariances, moments or intervals of interest

Estimate Potential Scale Reduction Factor

Returning to previous example:

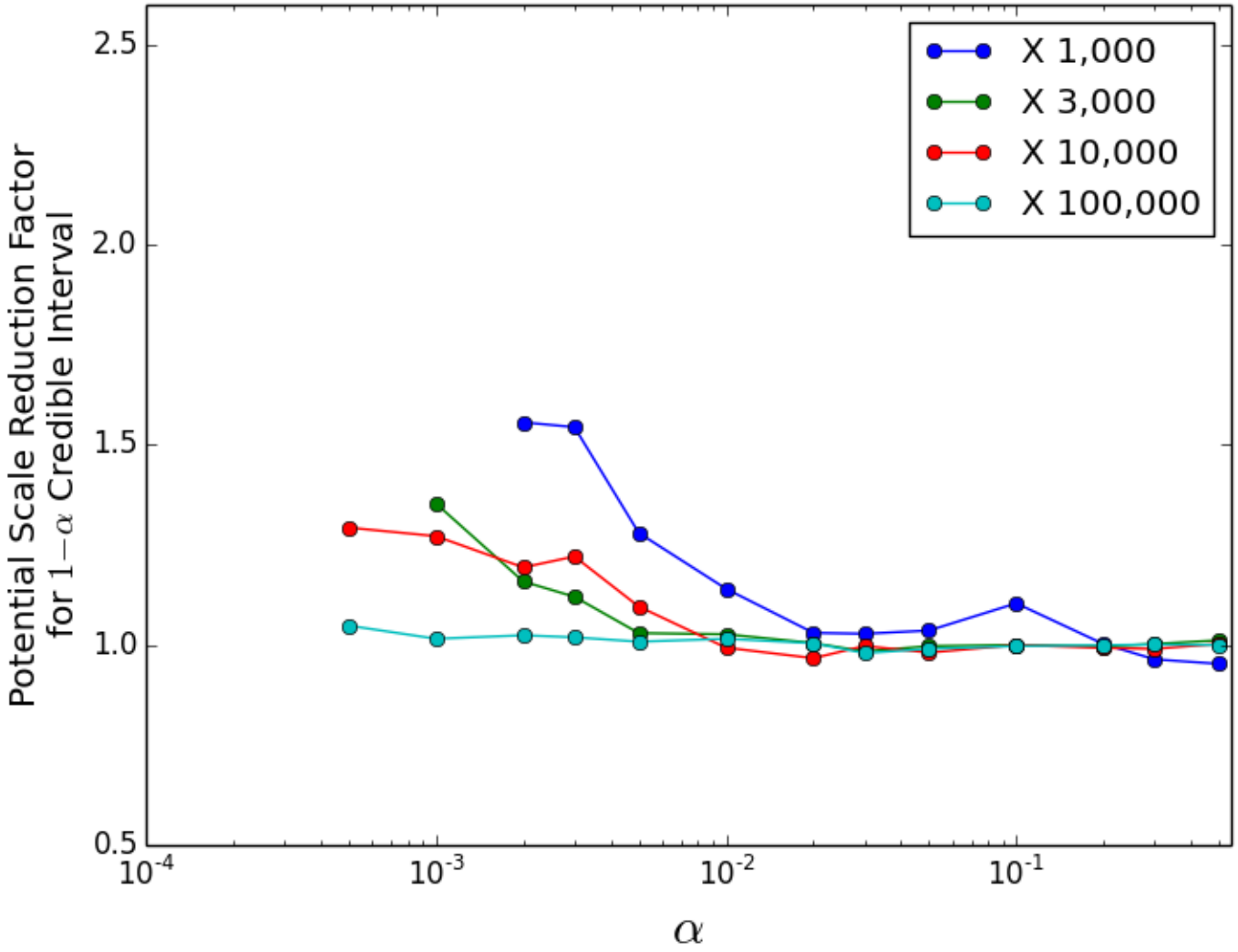
- Gelman-Rubin diagnostic ($\hat{R}$) is <1.001
- Consider generalized statistic

$$\widehat{R}_{\text{interval}} = \frac{\text{length of total-sequence interval}}{\text{mean length of the within-sequence intervals}}$$

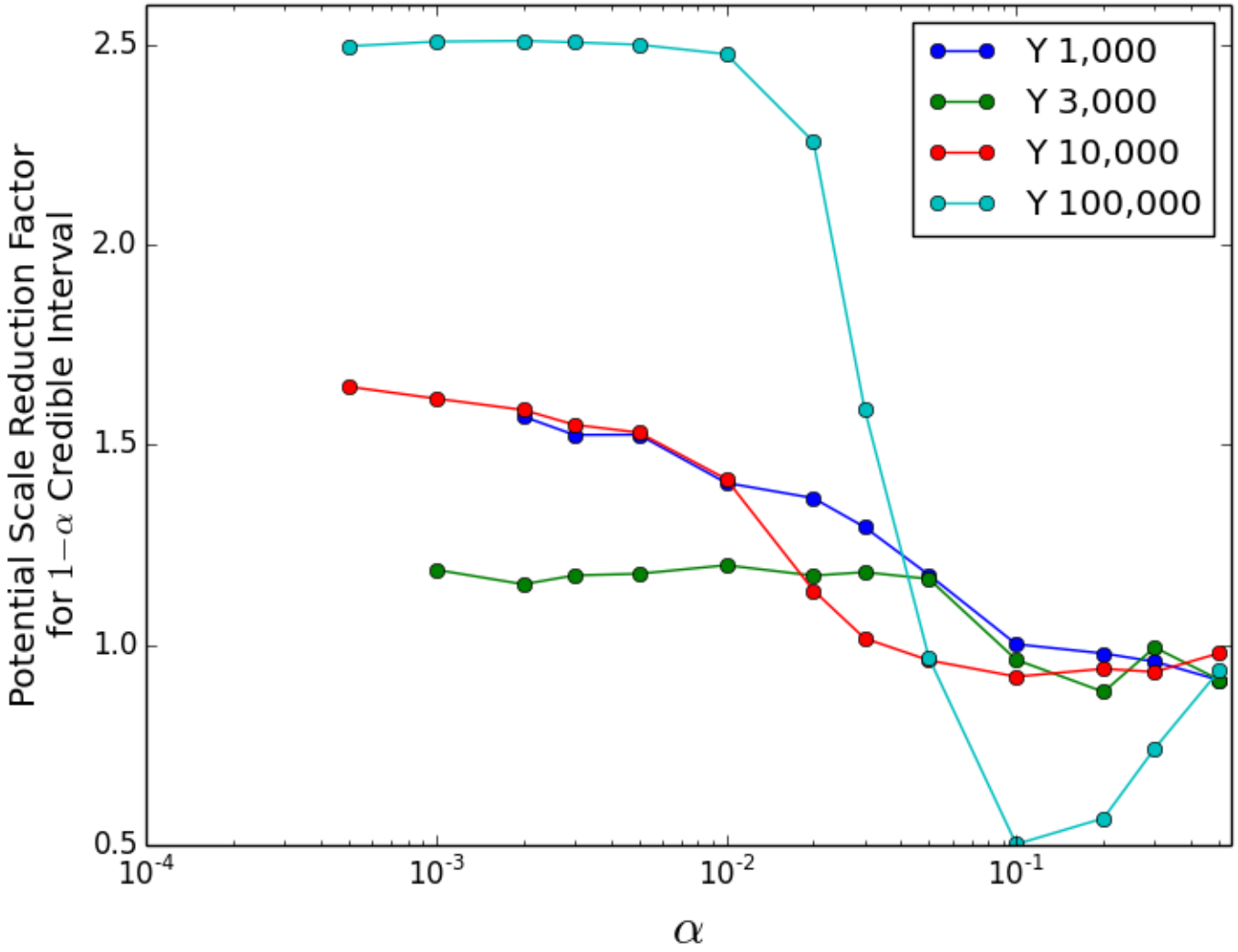
for central (1-α) credible interval

- Plot as function of α

Estimate Potential Scale Reduction Factor



Estimate Potential Scale Reduction Factor



Test using Simplified Problems where You Can Compare to Target Density

This target distribution for the first example was:

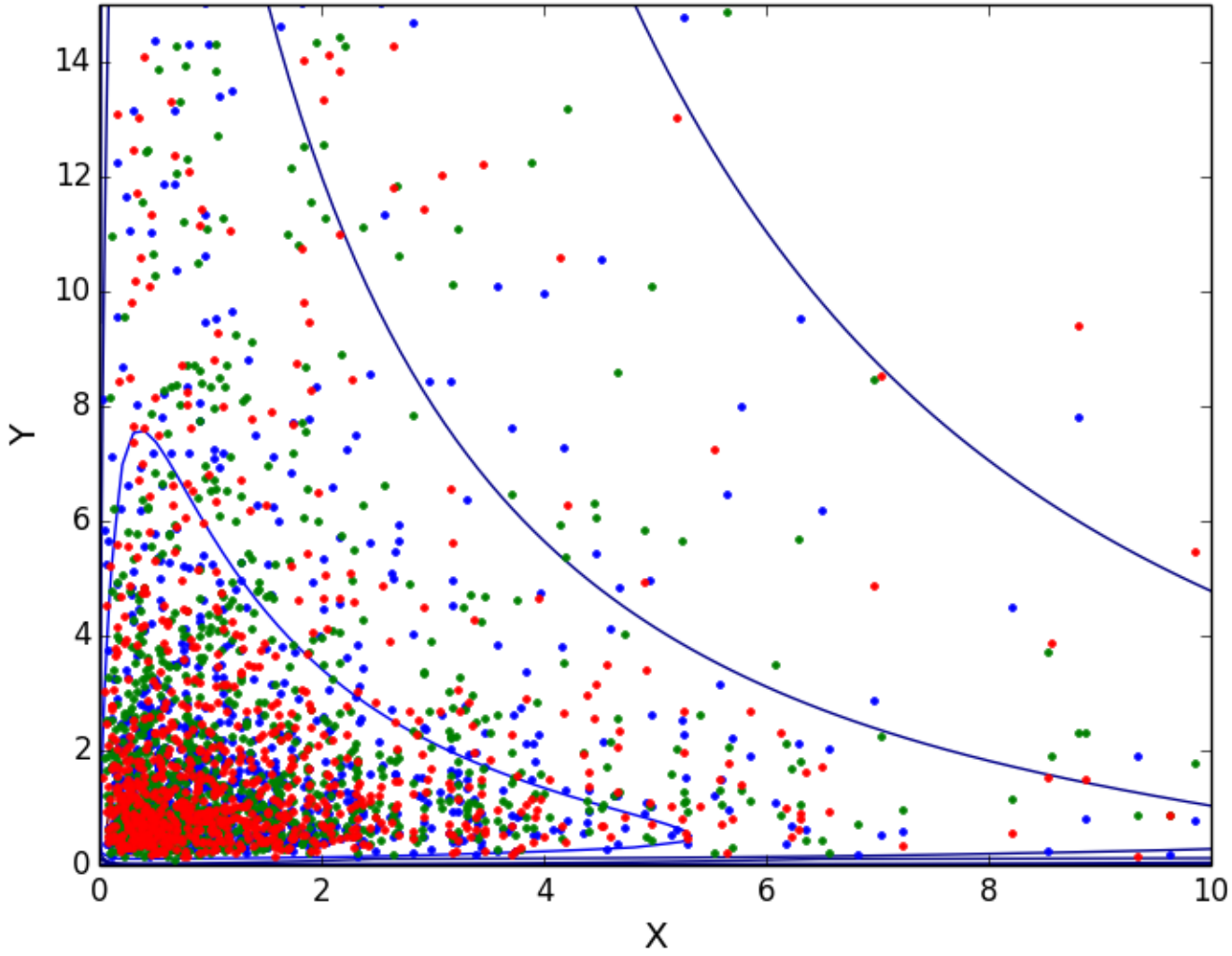
- p(x,y) = p(x) p(y)
- p(x) is LogNormal (zero mean, unit scale)

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0$$

- p(y) is InverseGamma (unit shape, unit scale)

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

Test using Simplified Problems where You Can Compare to Target Density



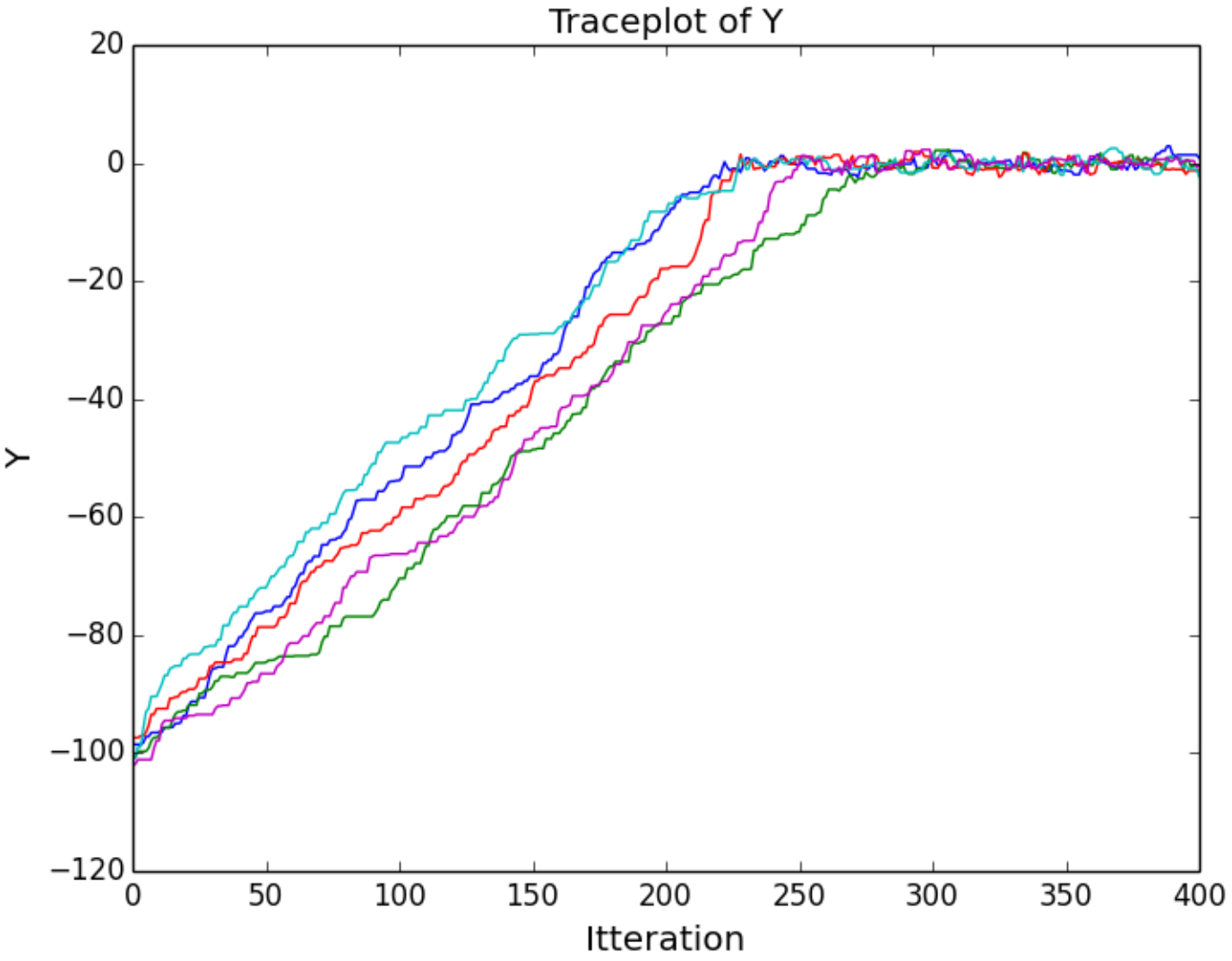
Use Problematic Runs to Improve Your Algorithm

- Why did our Markov chains struggle on a relatively simple target distribution?
- How could we change our algorithm to accelerate convergence?

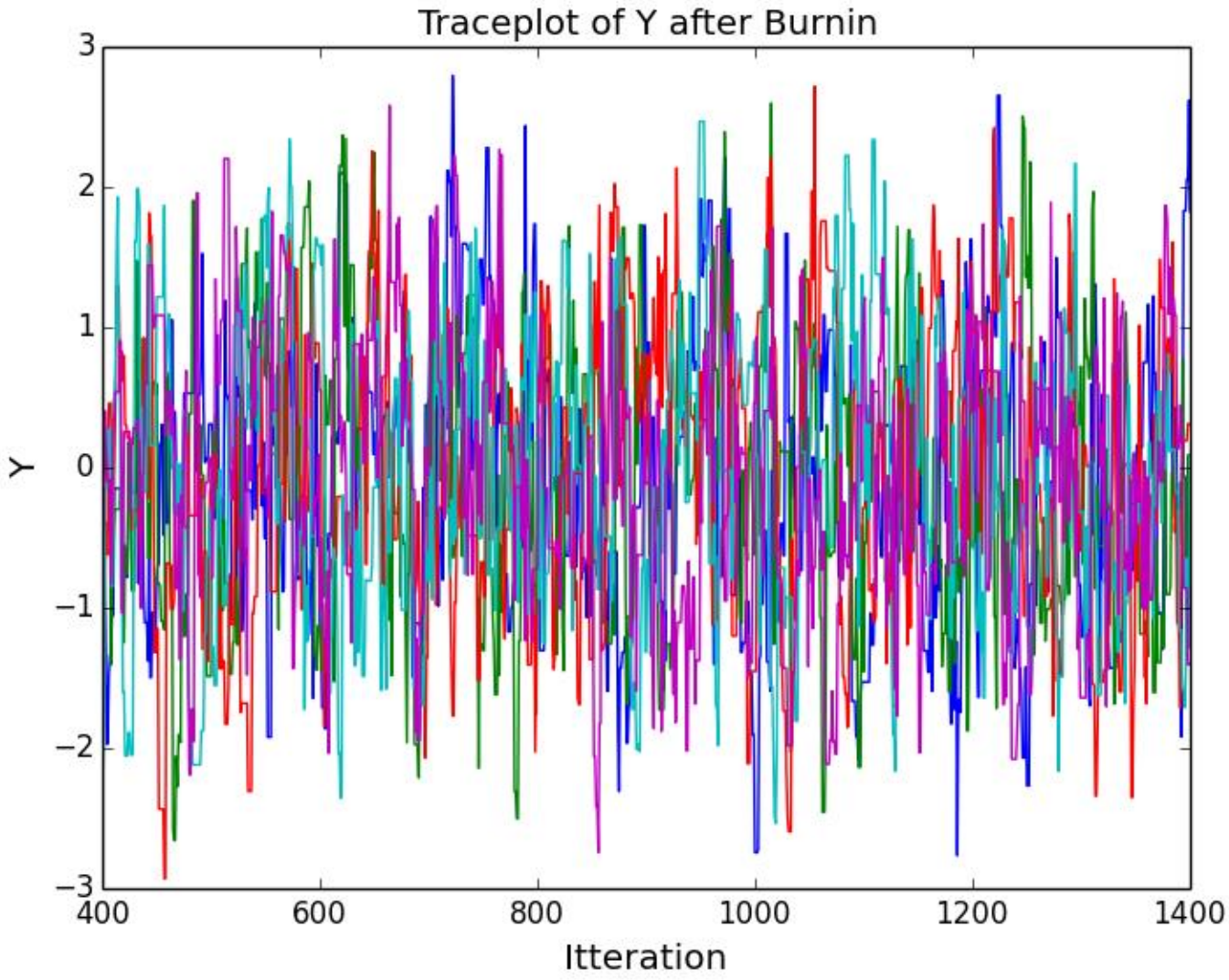
Has this Chain Converged?

- Let’s try this game again…

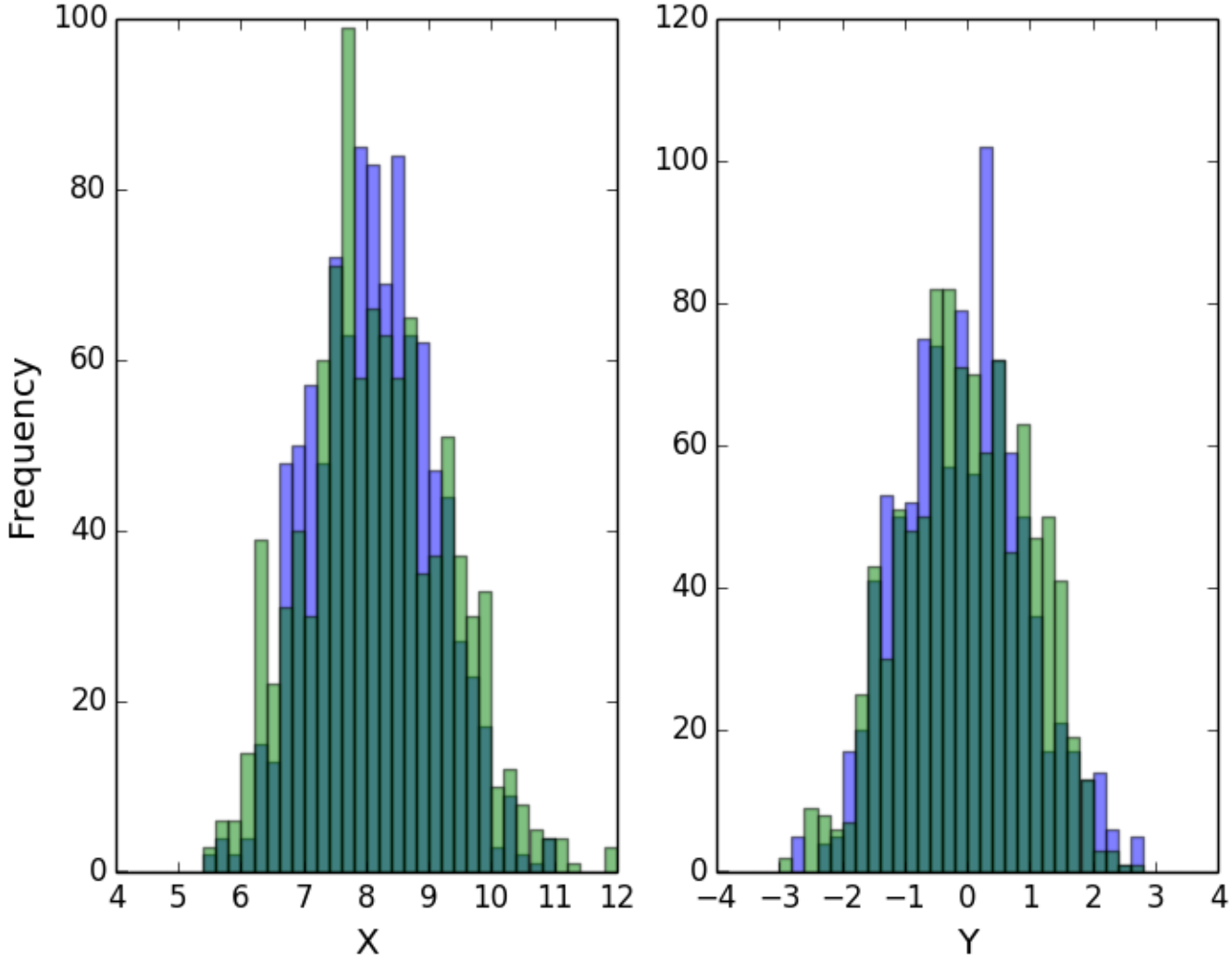
Has this Chain Converged?



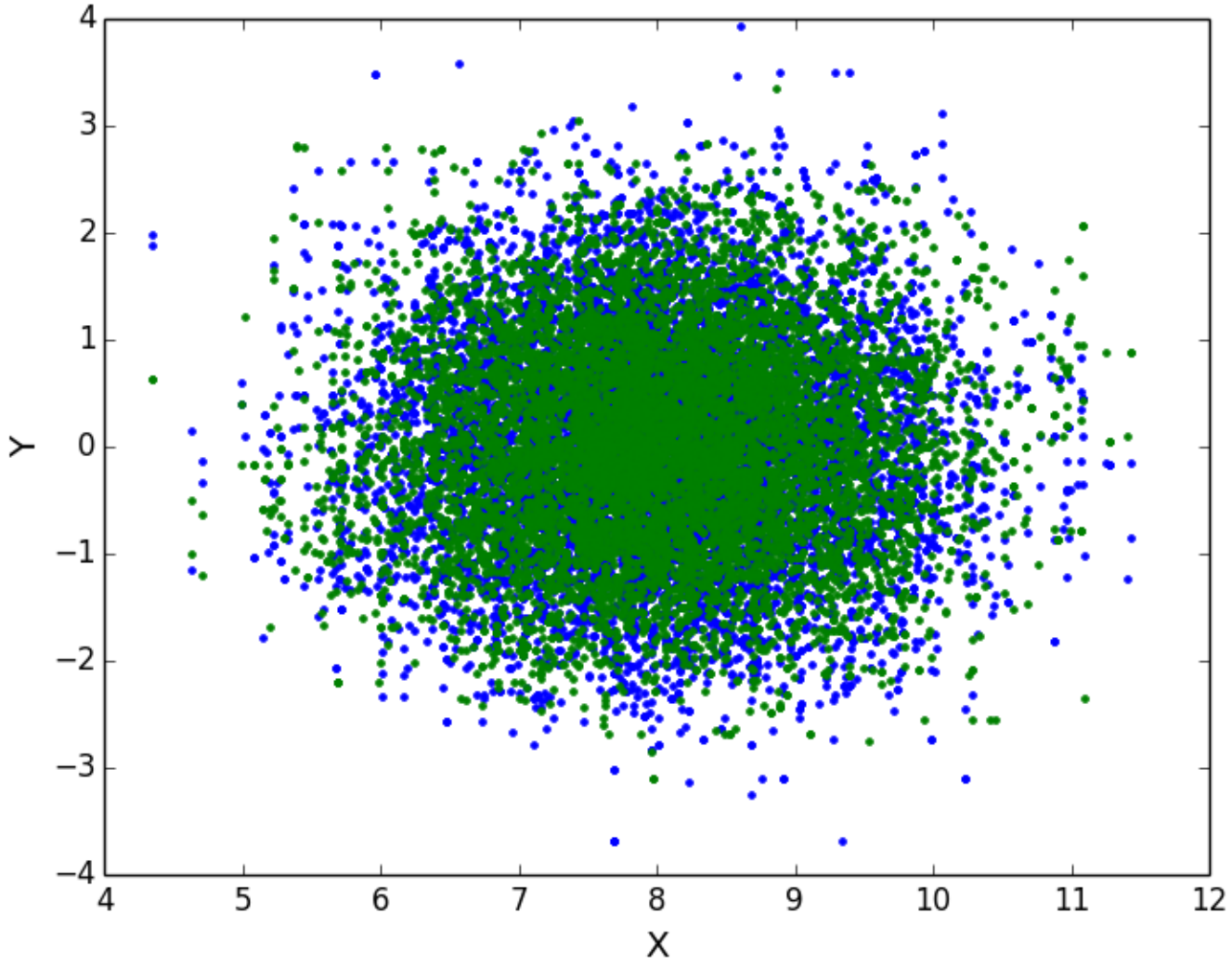
Has this Chain Converged?



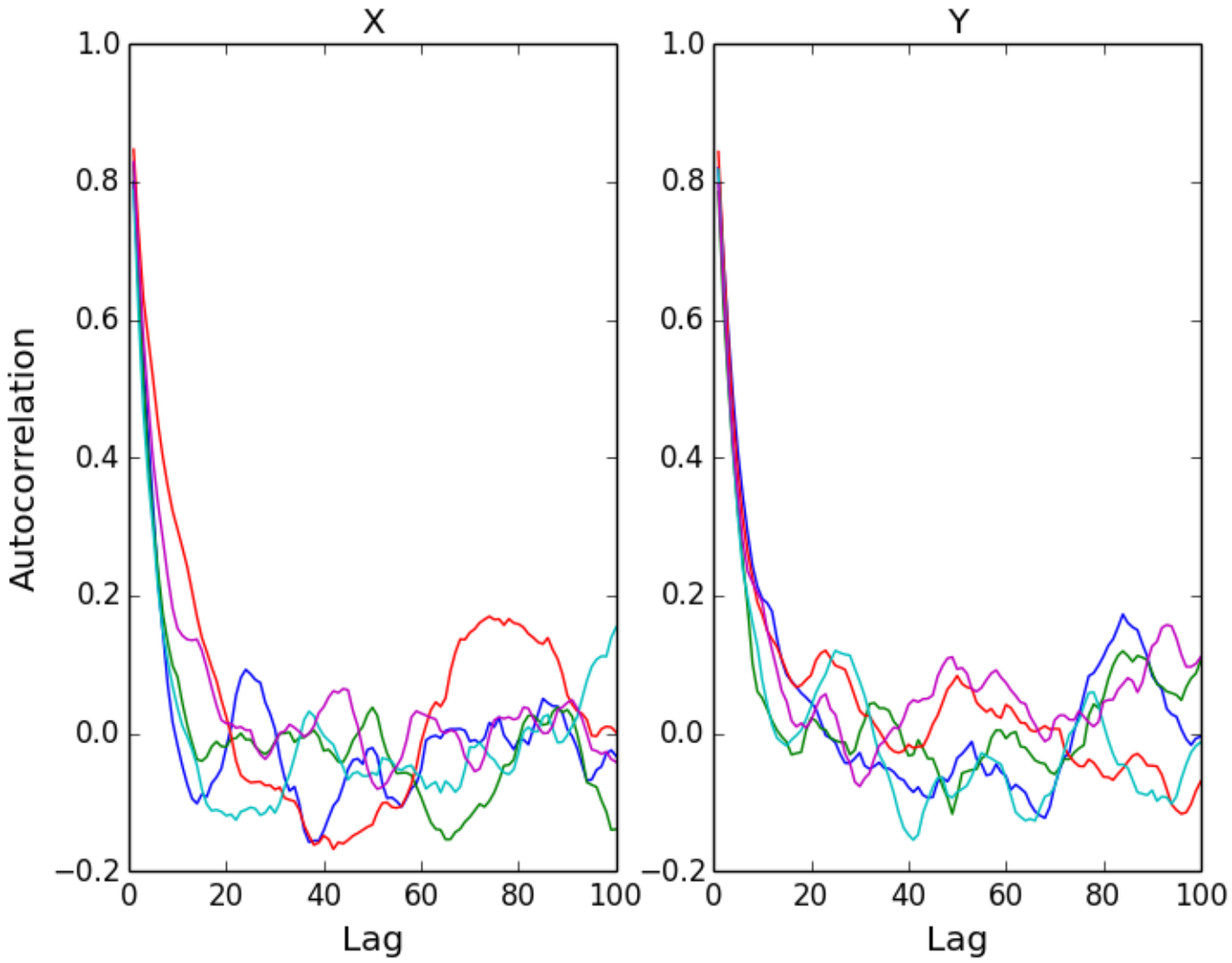
Has this Chain Converged?



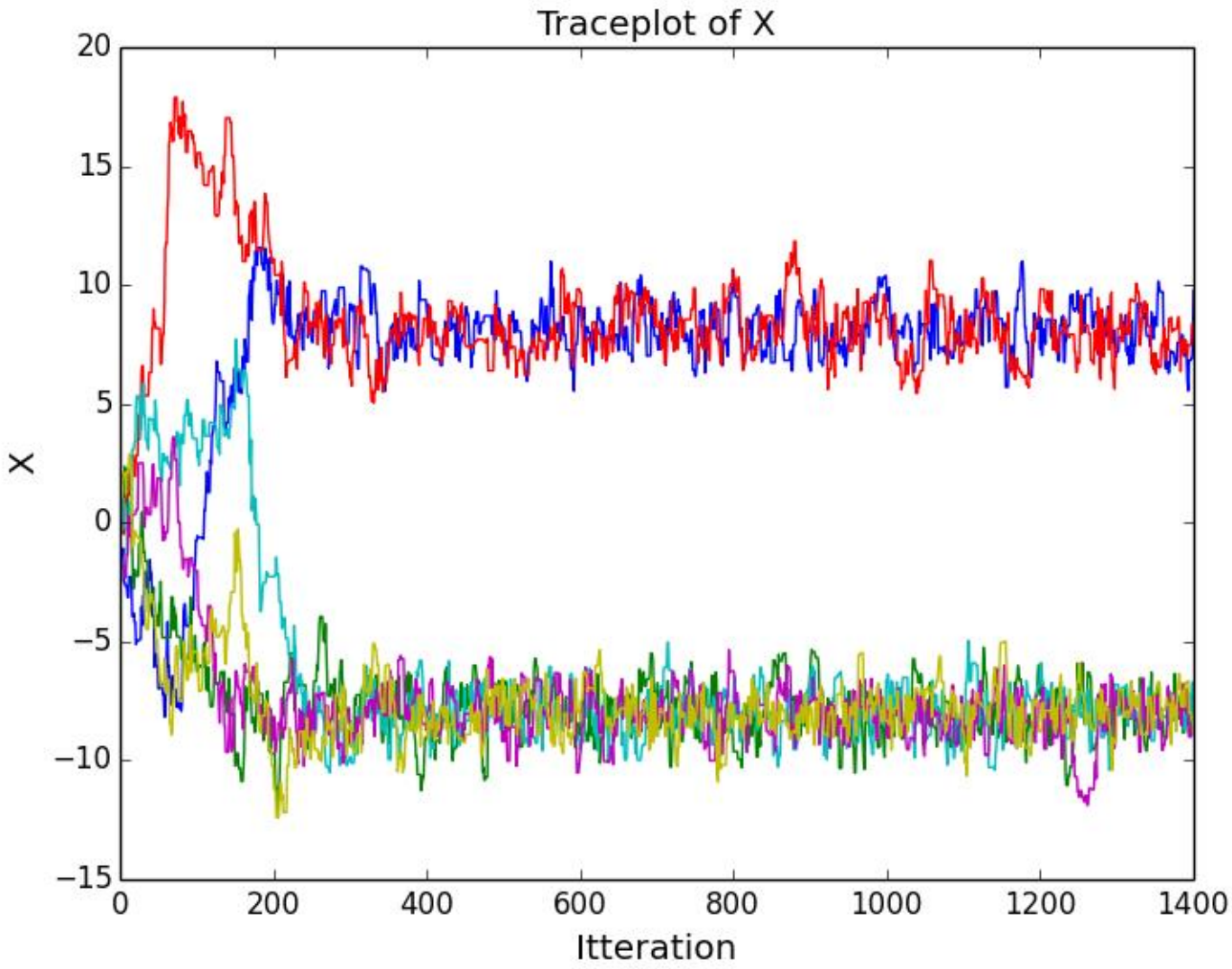
Has this Chain Converged?



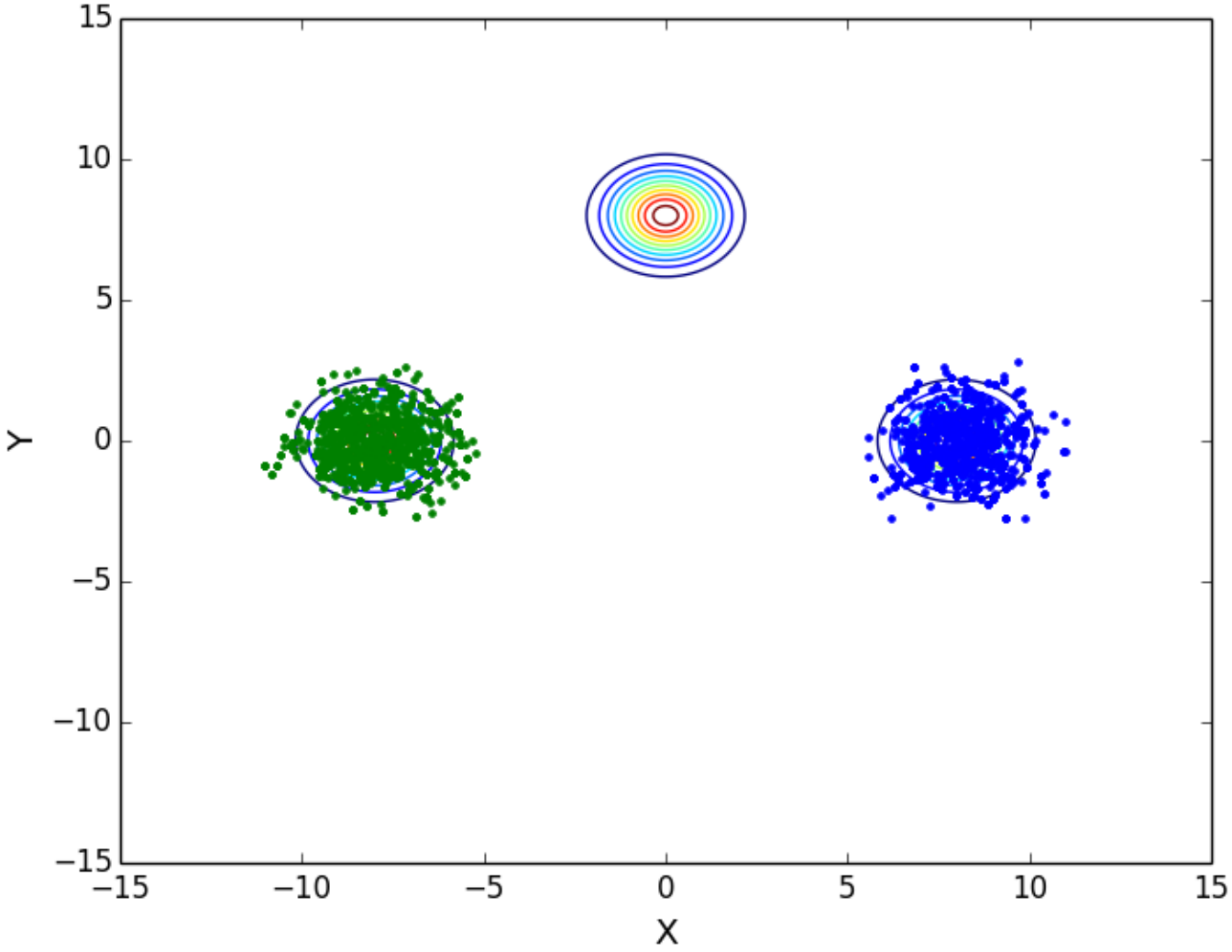
Has this Chain Converged?



Has this Chain Converged?



Has this Chain Converged?



My chain isn't perfect… Now what?

- What is your goal?
  - Ball park estimate of the median value of a parameter
  - Precisely define the boundary of a 99.9% credible region
- What are the consequences?
- Different goals merit different levels of paranoia

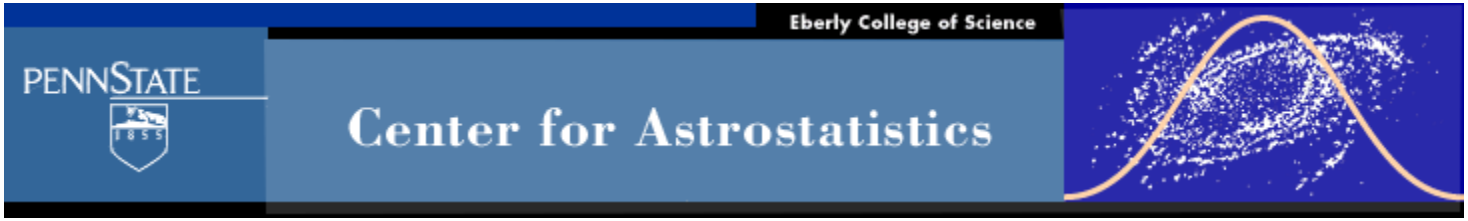
My chain isn’t perfect… Now what?

- Run Markov chains for many more iterations
- Change problem and/or algorithm to accelerate convergence

How a non Non-Converged Markov chain be Useful?

- Even if your chains haven't converged, have they allowed you to learn something about your target density?
- Can you change your algorithm so it will converge more quickly?
  - Change step sizes?
  - Alternative parameterization of problem?
  - Change proposal distribution of MCMC?
  - Solve problem via Importance Sampling?

Pause for Questions

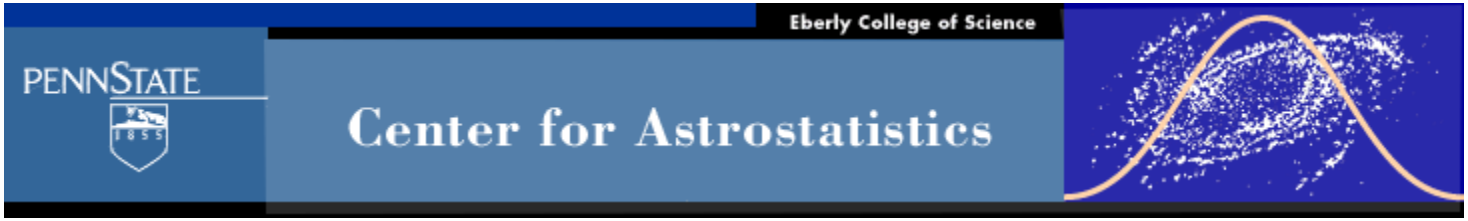


Ensemble MCMC: A Great Tool for Target Densities with Correlations Between Parameters

Eric B. Ford (Penn State)

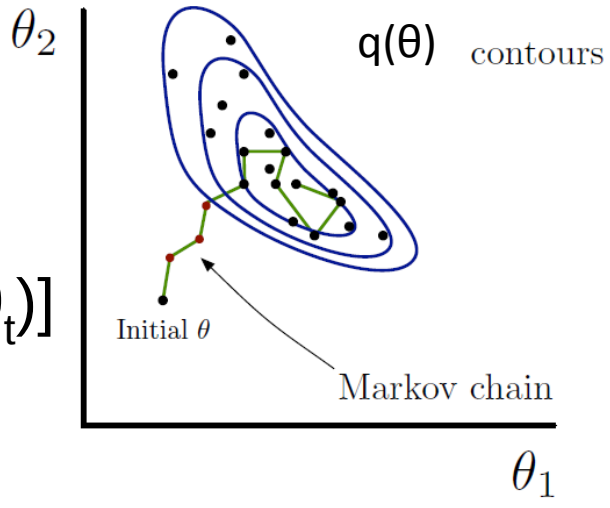
Bayesian Computing for Astronomical Data Analysis

June 5, 2015



Simple Markov Chain Monte Carlo

- Initialise chain with θ_0 (initial guess)
- Loop (iterate over t)
  1. Propose trail state, θ', according to $k(\theta'|\theta_t)$.
  2. Calculate unnormazlized posterior probability for trial state, $q(\theta') \sim p(\theta'|\text{Data,Model})$.
  3. Accept or reject trial state
     - Draw random number, u ~ U[0,1)
     - $\alpha(\theta'|\theta_t) = [q(\theta')\, k(\theta_t|\theta')] / [q(\theta_t)\, k(\theta'|\theta_t)]$
     - If $u \le \alpha(\theta'|\theta_t)$, then set $\theta_{t+1} = \theta'$ (accept)
     - If $u > \alpha(\theta'|\theta_t)$, then set $\theta_{t+1} = \theta_t$ (reject)
- Test for non-convergence



Why Go Beyond Simple MCMC?

- Standard MCMC converges extremely slowly if the proposal distribution is not well chosen
  - It's hard to find a good proposal distribution for complex problems (e.g., many parameters)
  - Want a way to automatically choose good proposal distribution
- Standard MCMC evaluates 1 model at a time
  - Parallelizing standard MCMC requires parallelizing the model evaluation (may be impractical)

What is Ensemble/Population MCMC?

- Instead of updating one set of model parameters at a time, update an ensemble/population of model parameters each “generation”

- Technically, the Markov chain is now over a product space of your model parameters

Advantages of Ensemble MCMC

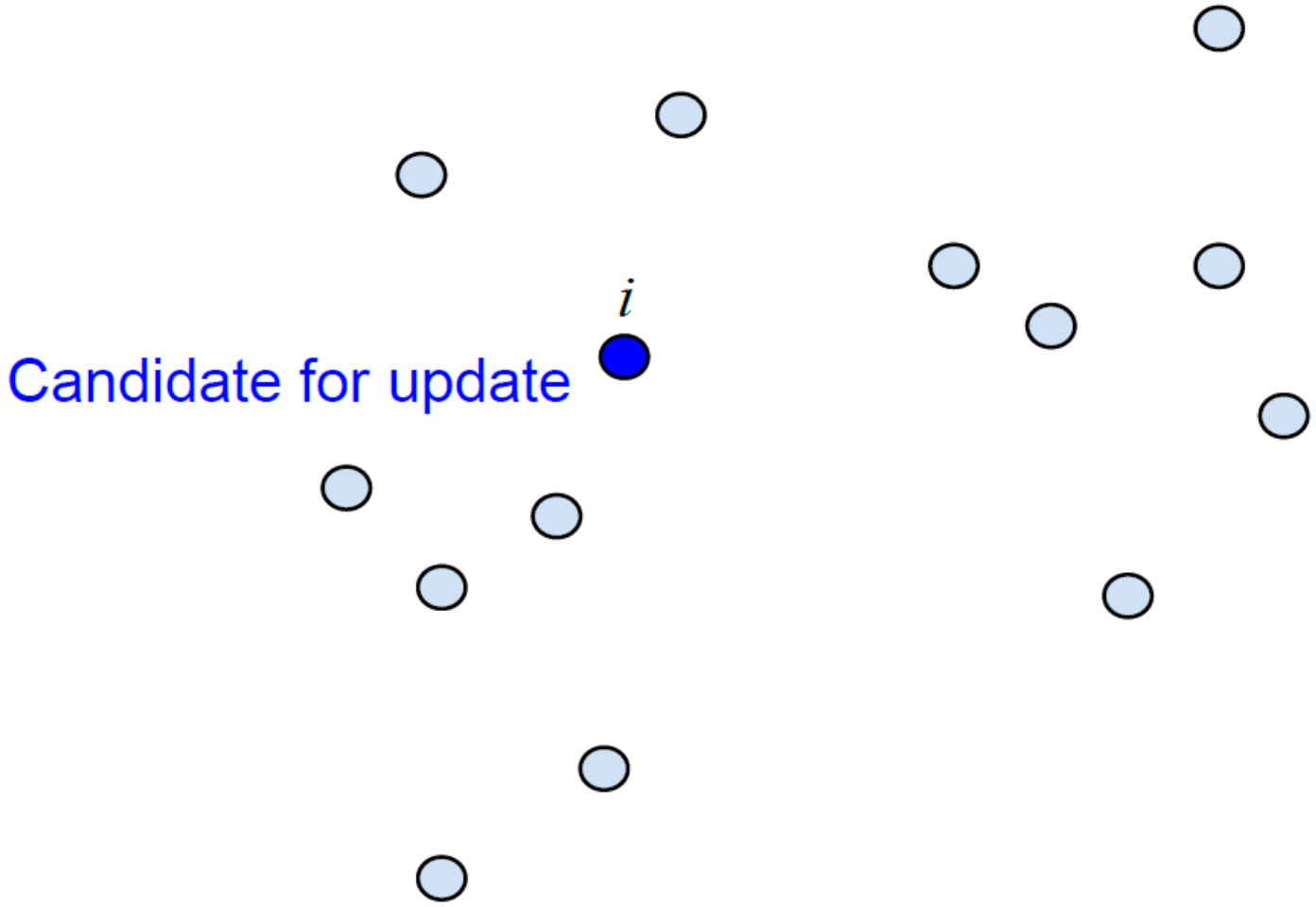
Ensemble MCMC:

- Can take advantage of having a population of model parameters when proposing each trial set of model parameters
- Makes it easy to parallelize over each set of model parameters within a generation

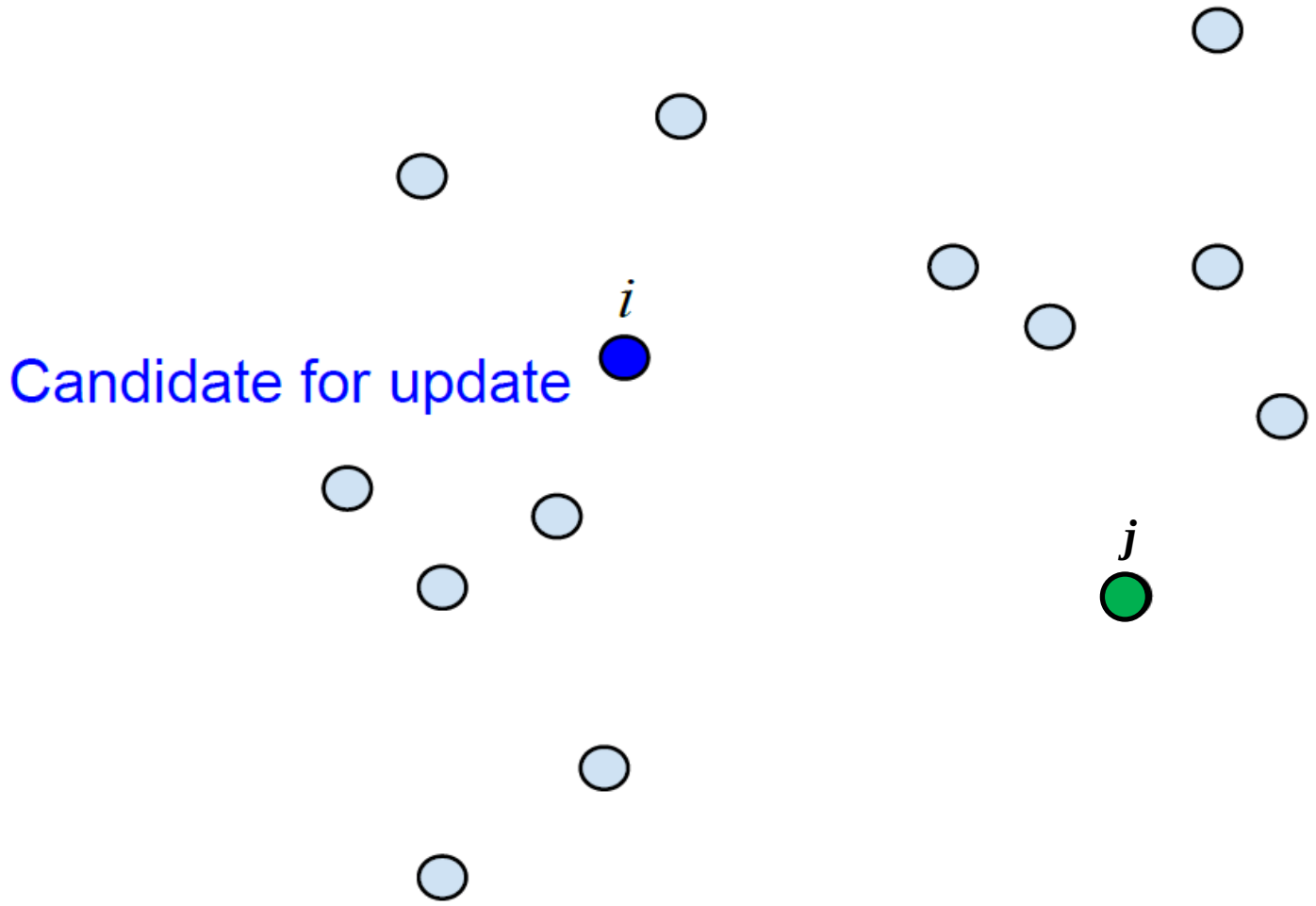
Two Specific Ensemble MCMC Algorithms

- Differential Evolution MCMC
  (ter Braak 2006; ter Braak & Vgurt 2008; Nelson et al. 2014)
  - Combines three states from previous generation for each trial state
- Affine-Invariant Ensemble MCMC
  (Goodman & Weare 2010; Foreman-Mackey et al. 2013)
  - Combines two states from previous generation for each trial state
- Both algorithms
  - Automatically infer shape & scale for proposals
  - Require only a few new parameters (and performance is typically insensitive to their choice)

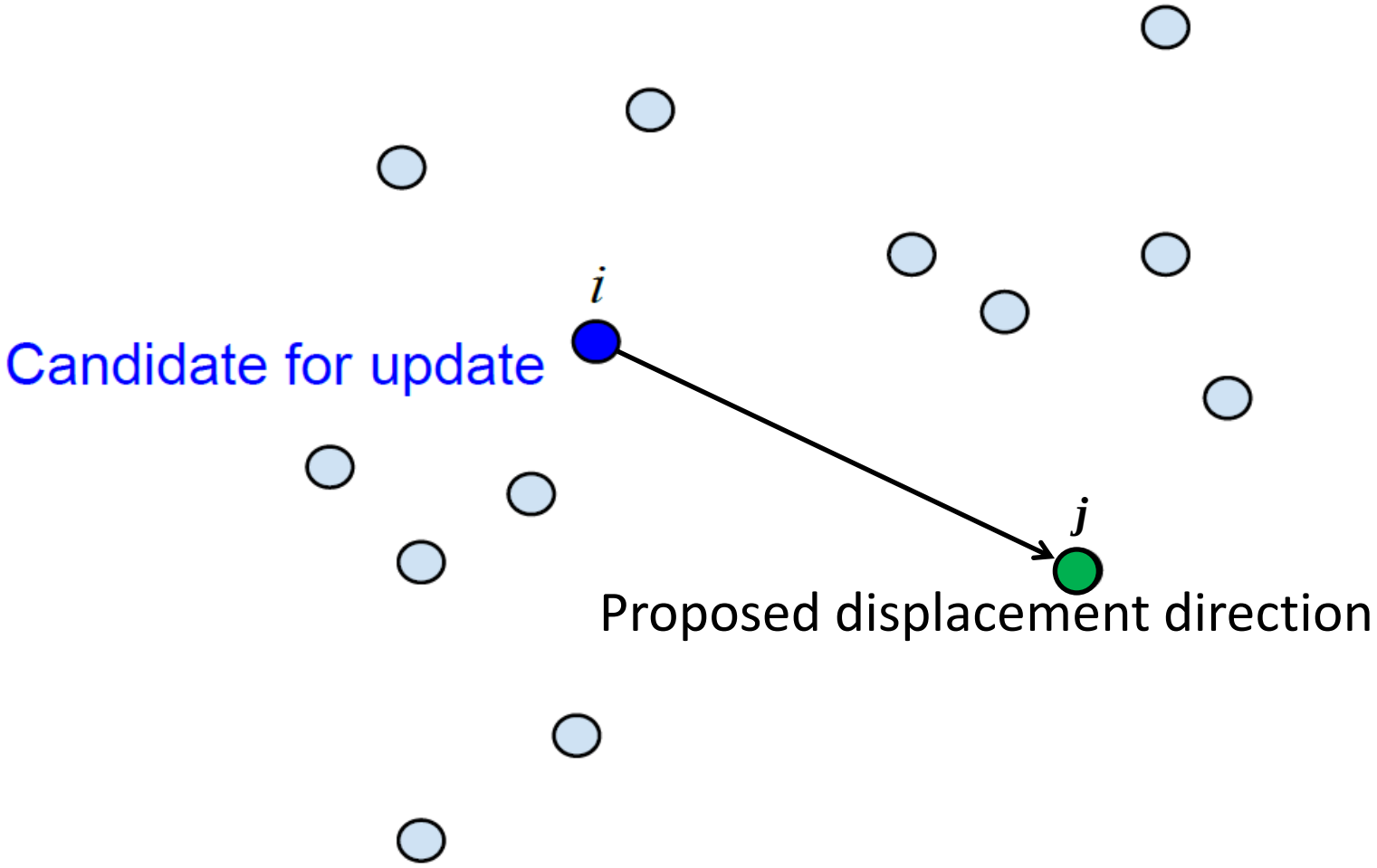
Affine-Invariant Ensemble MCMC



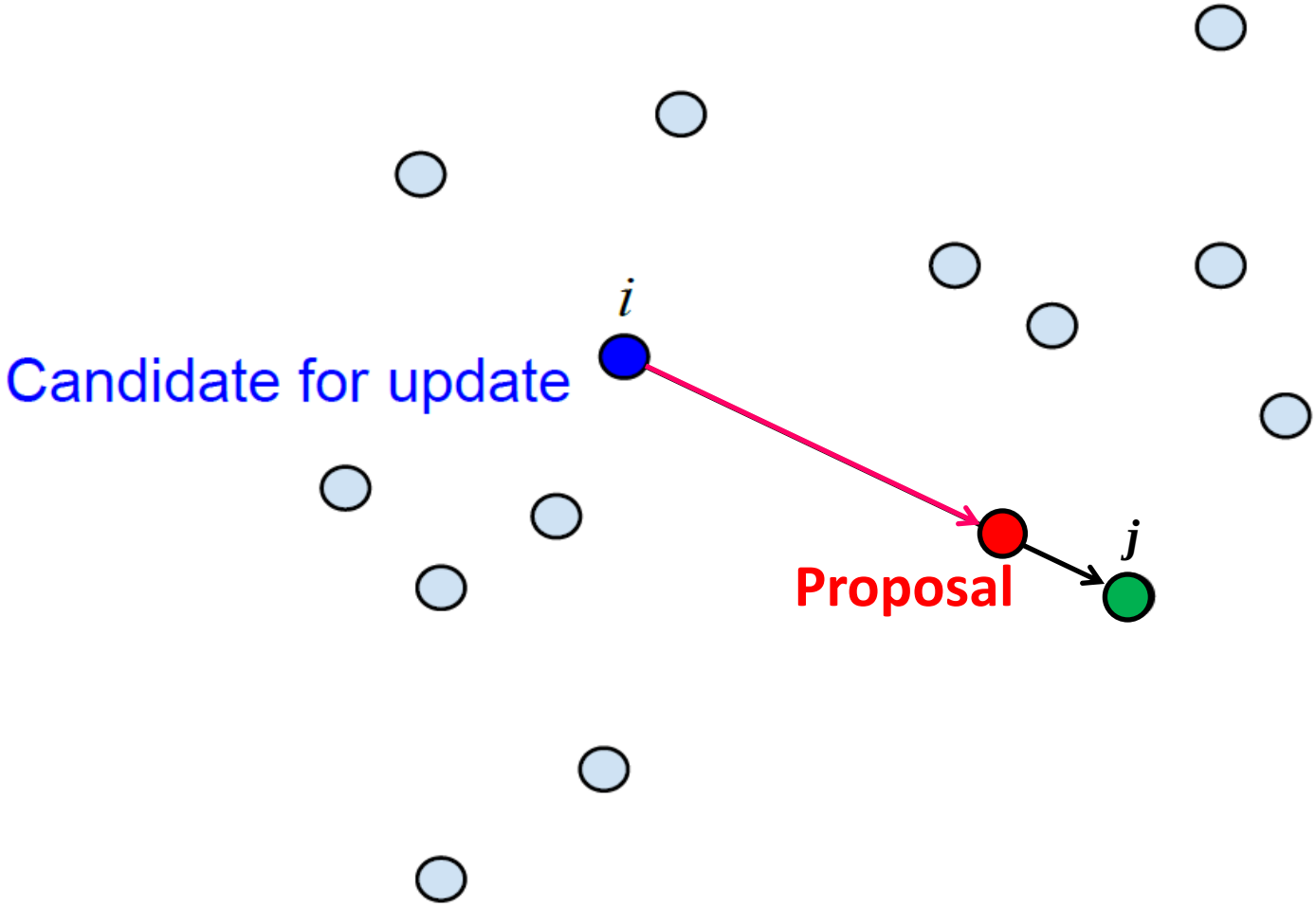
Affine-Invariant Ensemble MCMC



Affine-Invariant Ensemble MCMC



Affine-Invariant Ensemble MCMC



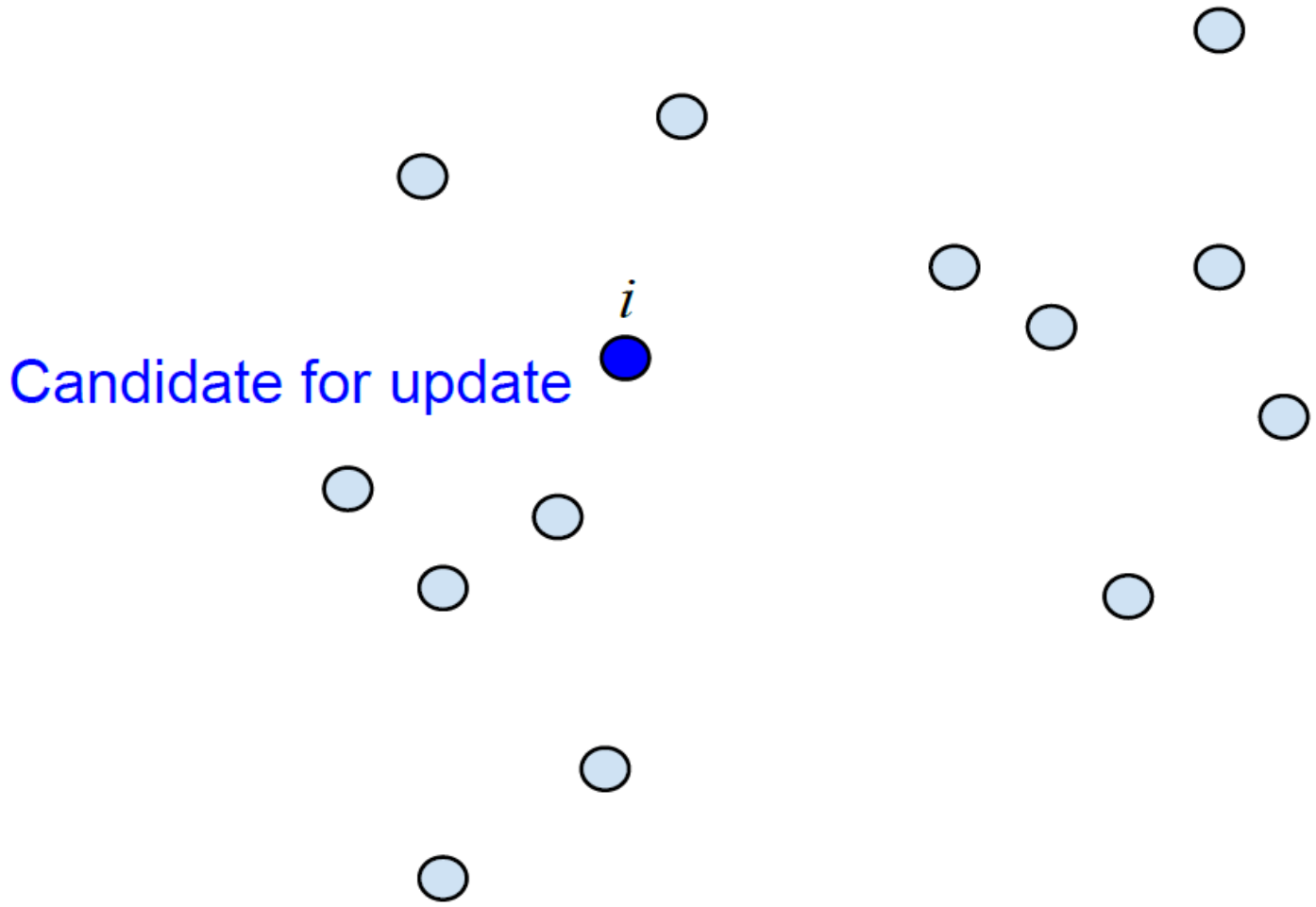
Affine-Invariant Ensemble MCMC

Implementation details

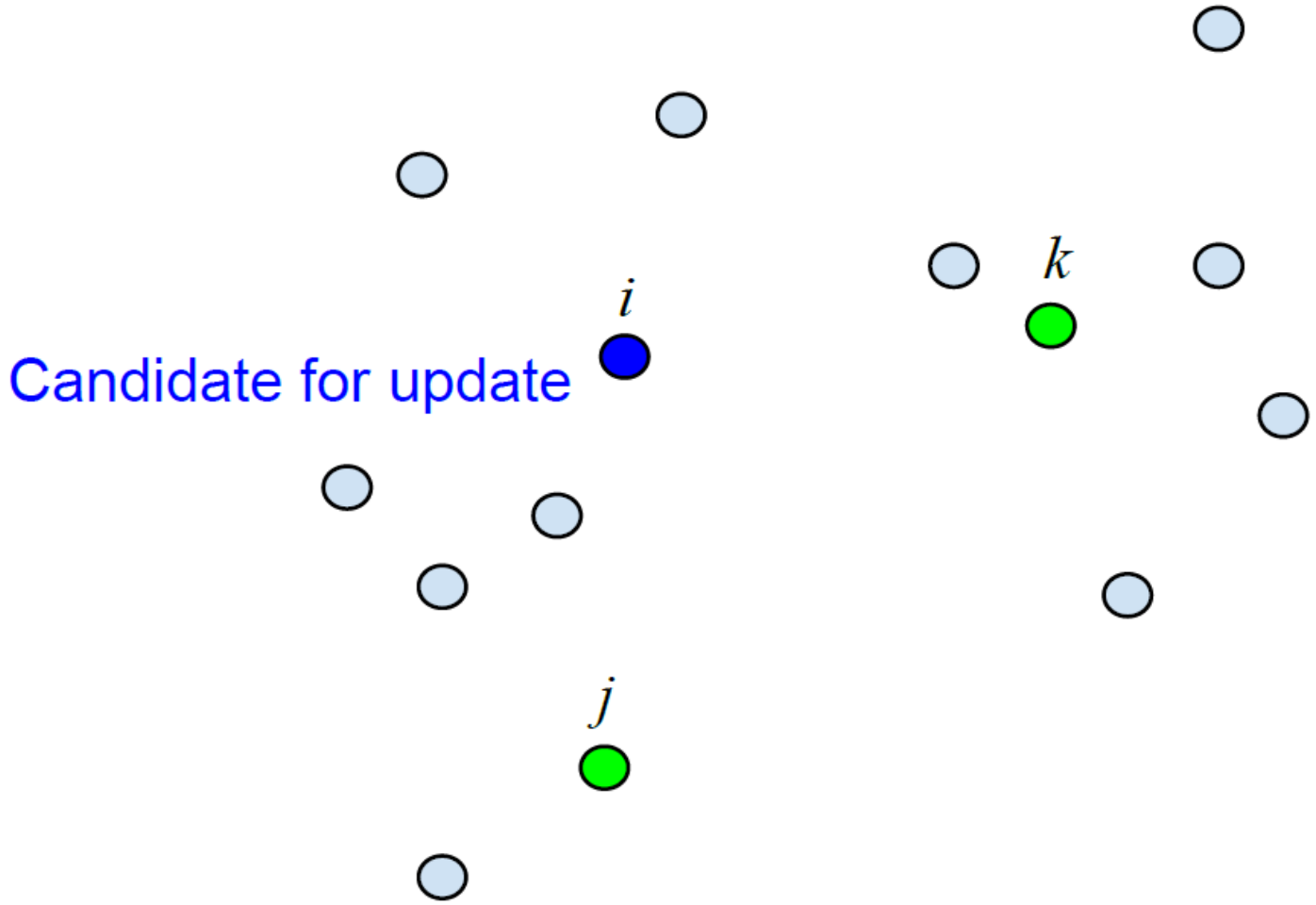
- Proposal step: $\theta_i' = \theta_{t,j} + z\,[\theta_{t,i} - \theta_{t,j}]$
  - z: random variable drawn from distribution $g(z) = z\,g(z)$
  - Update parameters for each "chain" in blocks
- Acceptance probability $\alpha = \min[1, z^{N_d-1}\, q(\theta')/q(\theta_{t,i})\,]$
  - N_d = dimension of parameter space
  - Target distribution: $q(\theta) \sim p(\theta' | \text{Data,Model})$
- Tunable parameters: a, $g(z)$ and N_{chains} (population size)
- Suggestions
  - $g(z) = z^{-1/2}$, $z \in [a^{-1}, a]$,
    0, else
  - $a = 2$
  - $N_{chains} > \text{few} \times N_d$

Goodman & Weare 2010
Foreman-Mackey et al. 2013

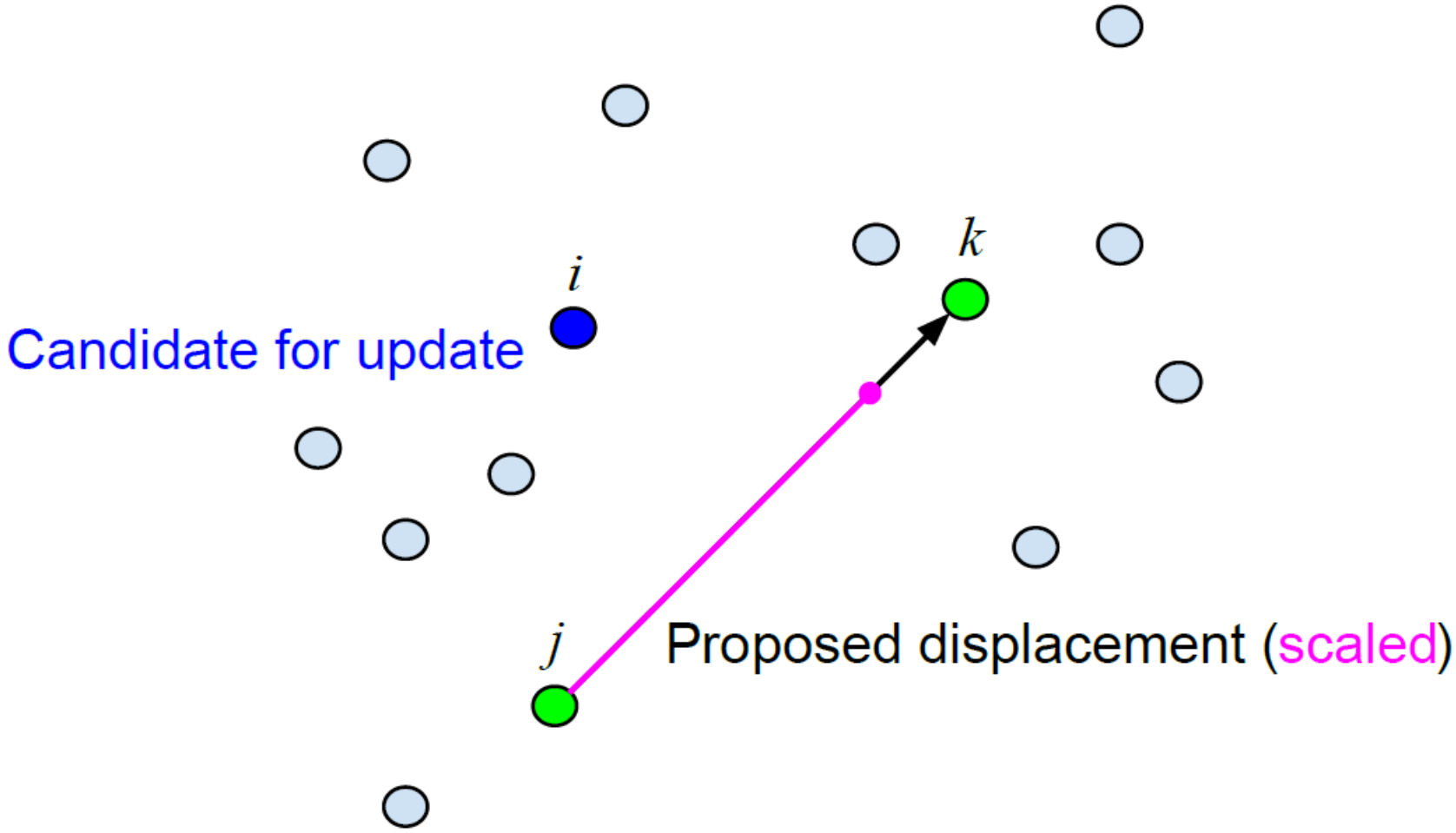
Differential Evolution MCMC



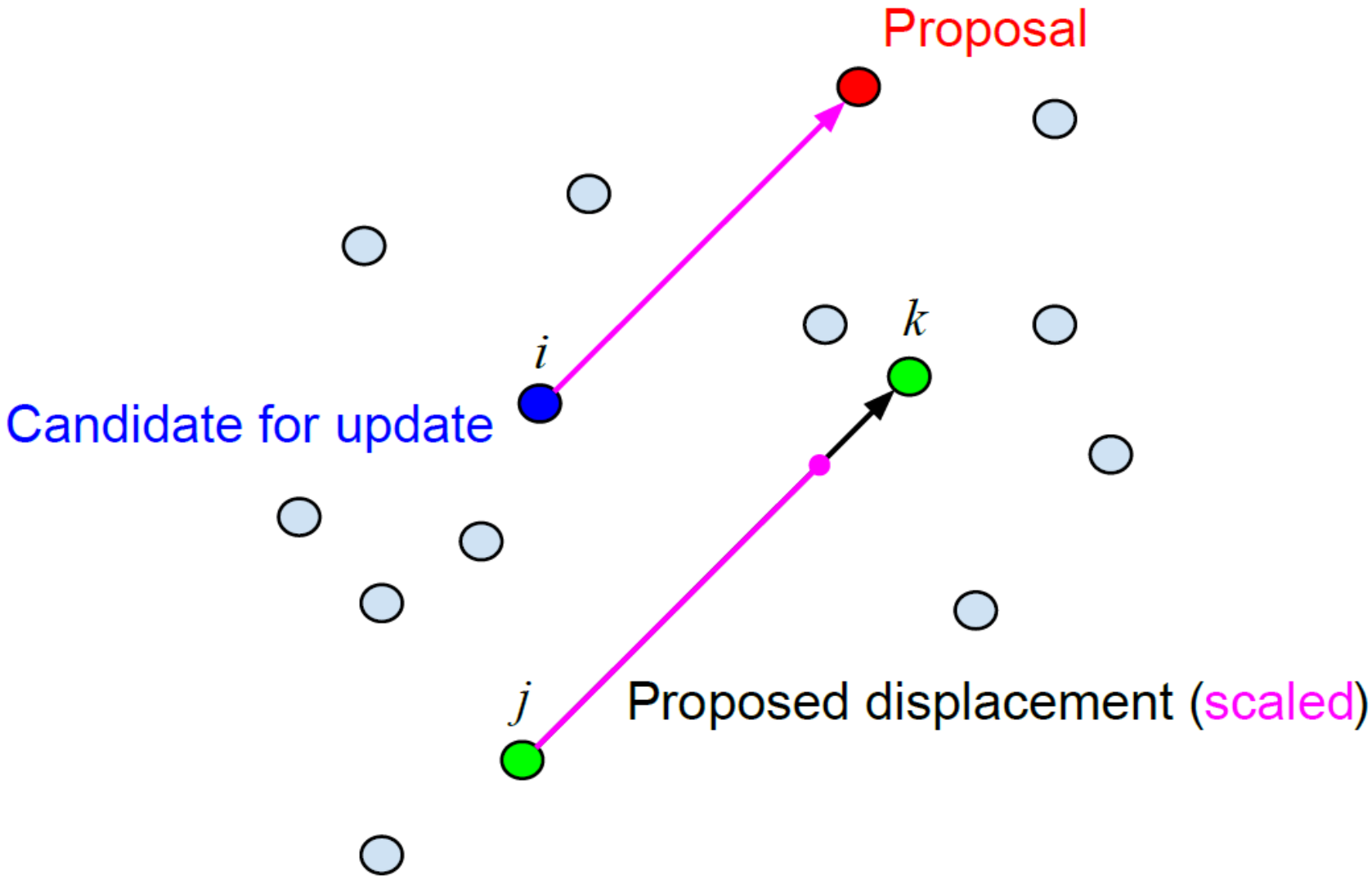
Differential Evolution MCMC



Differential Evolution MCMC



Differential Evolution MCMC



Differential Evolution MCMC

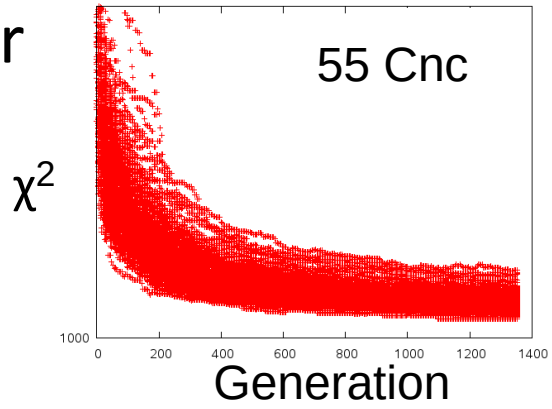
Implementation details

- Proposal step: $\theta' = \theta_{t,i} + \gamma\,[\theta_{t,k} - \theta_{t,j}]$ (most of the time)
  - $\gamma = \gamma_o\,(1 + z)$
  - $z \sim N(0, \sigma_\gamma^2)$
  - $\gamma_o = 2.38 / (2N_{dim})^{1/2}$, (initially, can adjust to improve acceptance rate)
  - Update parameters for each "chain" in blocks
- Optionally, occasionally use large proposal steps
  - $\gamma = (1 + z)$ $z \sim N(0, \sigma_\gamma^2)$
- Acceptance probability: same as standard MCMC
- Tunable parameters: σ_γ, and N_{chains} (population size)
- Suggestions:
  - $N_{chains} >$ few $\times N_d$
  - $0.001 < \sigma_\gamma < 0.1$ (quite insensitive in our tests; Nelson et al 2014)
  - Adapt γ_o to achieve good acceptance rate (~0.25)

ter Braak 2006
Nelson et al. 2014

Choosing Initial Population

- Generate initial population from prior
  - Great… if it works
  - But often get stuck in local maxima, resulting in unreasonable number of generations to complete burn-in



- Generate initial population close to posterior
  - Dramatically reduces burn-in time
  - But what if you missed another important posterior maxima?
- Compromise: Generate initial population to be near posterior, but more dispersed than posterior

How Can Things Still Go Wrong?

- Initial population too far from target density
  - Choose initial population close to target density
  - Test that results insensitive to choice
- Non-linear correlations between parameters
  - Results in long auto-correlation times
  - Increasingly problematic with higher-dimensional parameter spaces
- Multi-modal target density
  - DEMCMC can deal with a few viable modes, but autocorrelation time increases

Example Application of DEMCMC

Measuring planet masses & orbits from Doppler observations of Exoplanet Systems

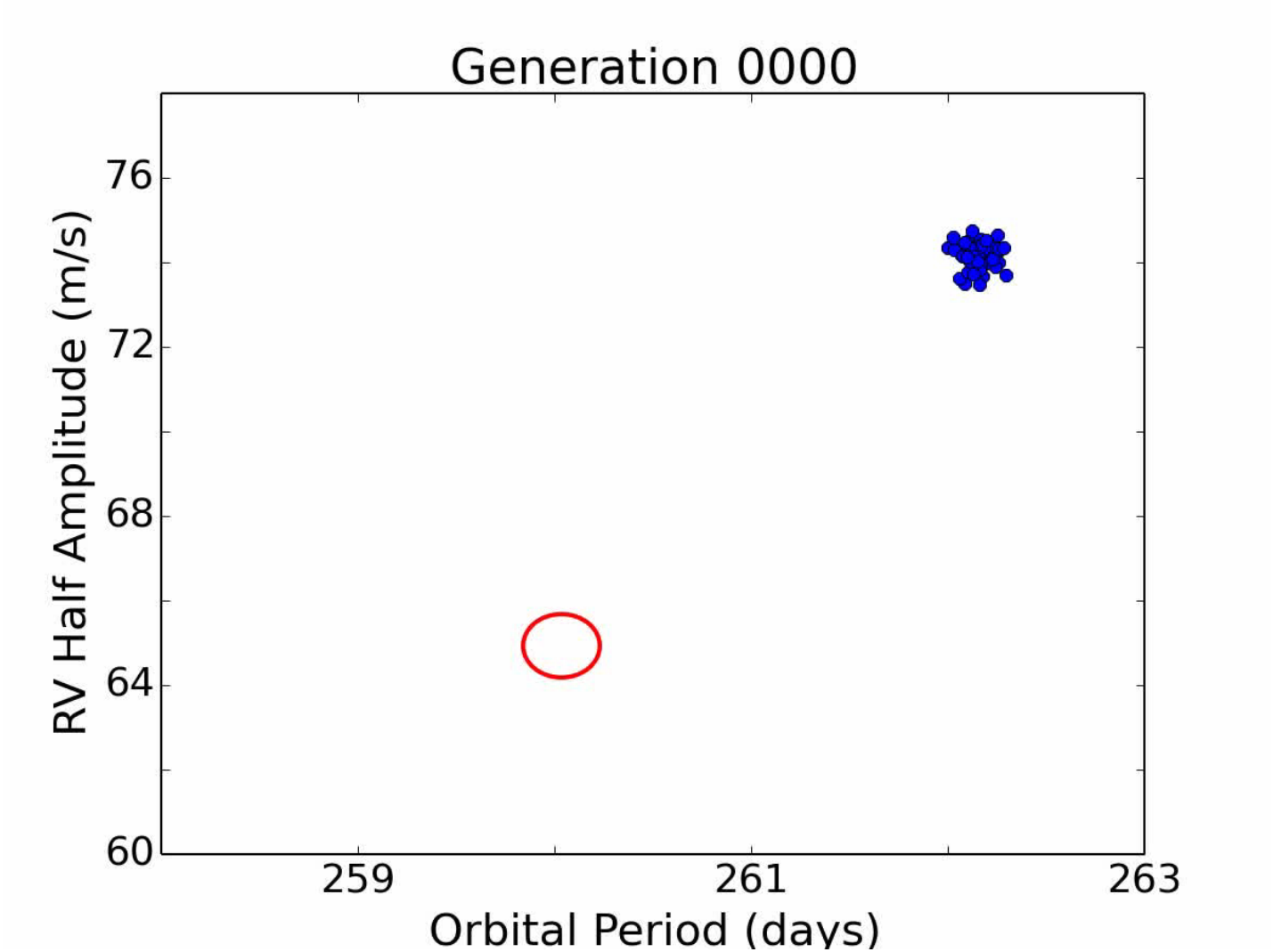
- Physical Model
  - Non-Interacting: Linear superposition of Keplerian orbits $v_{\star,\vec{\theta}}(t,j) = \sum_i K_i \left\{\cos\left[\omega_i + f_i(t)\right] + e_i \cos\omega_i\right\} + C_j$
  - Interacting: Full n-body model $\frac{d^2\vec{r_i}}{dt^2} = -\sum_{j=1}^{N} \frac{Gm_j(\vec{r_i} - \vec{r_j})}{|\vec{r_i} - \vec{r_j}|^3}$
- Likelihood assumes observations with uncorrelated, Gaussian uncertainties

$$\chi^2 = \sum_k \frac{\left[v_{\star,obs}(t_k,j_k) - v_{\star,\vec{\theta}}(t_k,j_k)\right]^2}{(\sigma_{\star,obs}(t_k,j_k)^2 + \sigma_{jit}^2)}$$

How Can Things Still Go Wrong?

- Initial population too far from target density
  - Choose initial population close to target density
  - Test that results insensitive to choice
- Non-linear correlations between parameters
  - Results in long auto-correlation times
  - Increasingly problematic with higher-dimensional parameter spaces
- Multi-modal target density
  - DEMCMC can deal with a few viable modes, but autocorrelation time increases

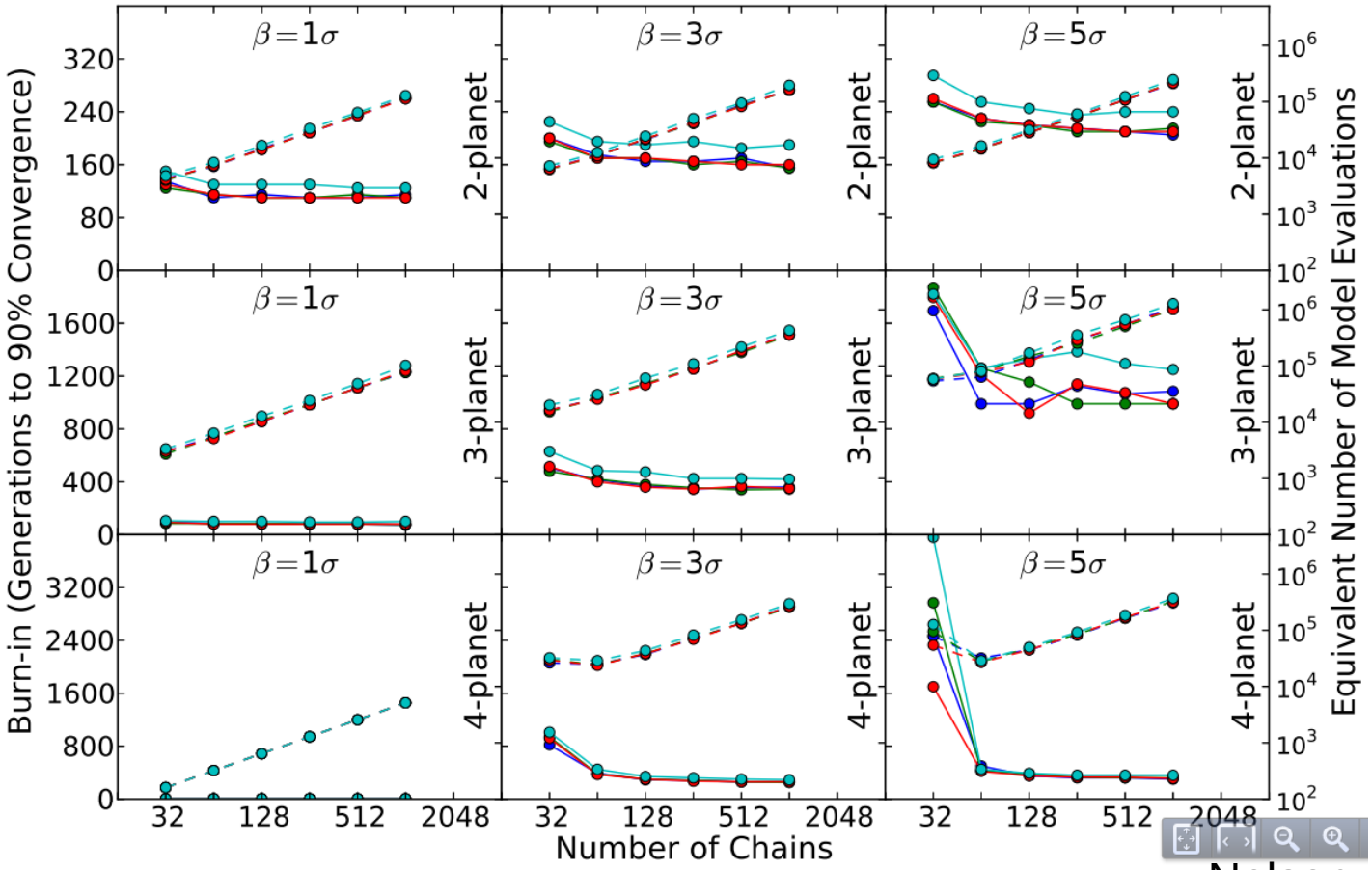
What if Poor Initial Population?



Nelson

Test Burn-In Required using Synthetic Data

- For initial population, take posterior sample and increase dispersion about mean



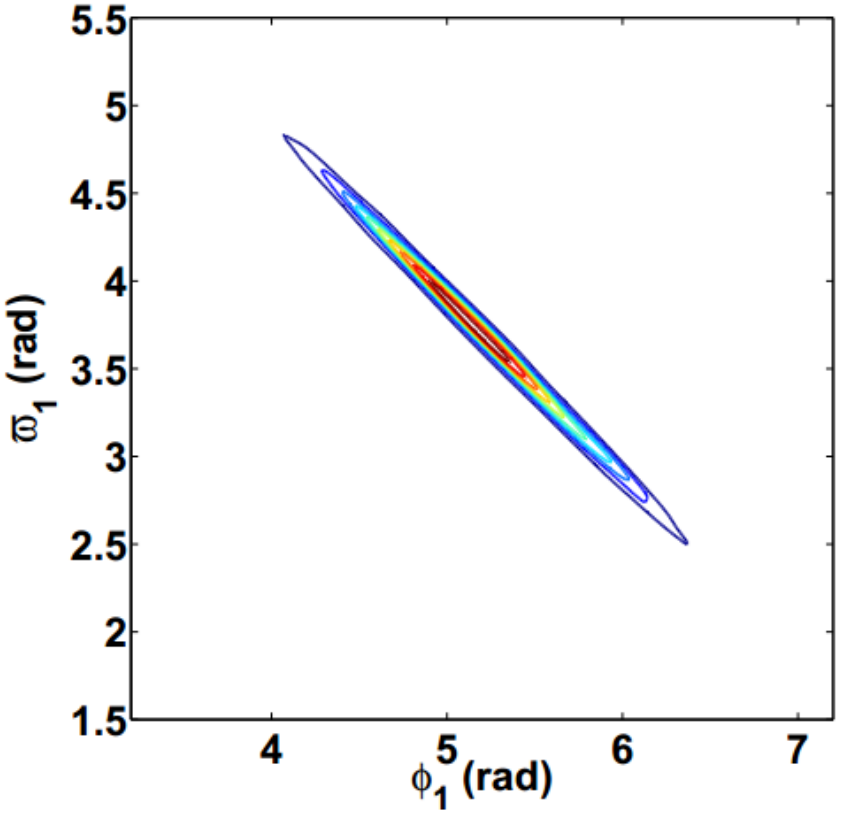
Nelson et al. 2014

How Can Things Still Go Wrong?

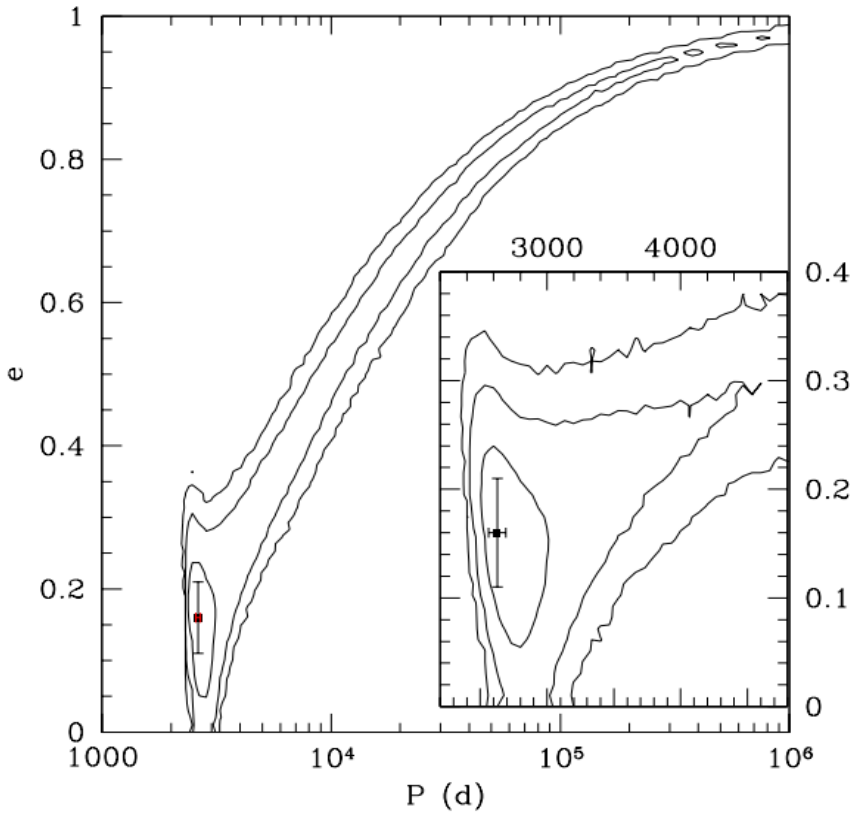
- Initial population too far from target density
  - Choose initial population close to target density
  - Test that results insensitive to choice
- Non-linear correlations between parameters
  - Results in long auto-correlation times
  - Increasingly problematic with higher-dimensional parameter spaces
- Multi-modal target density
  - DEMCMC can deal with a few viable modes, but autocorrelation time increases

Non-Linear Parameter Correlations

Linear Correlations
(still efficient)



Non-Linear Correlations
(reduce efficiency)



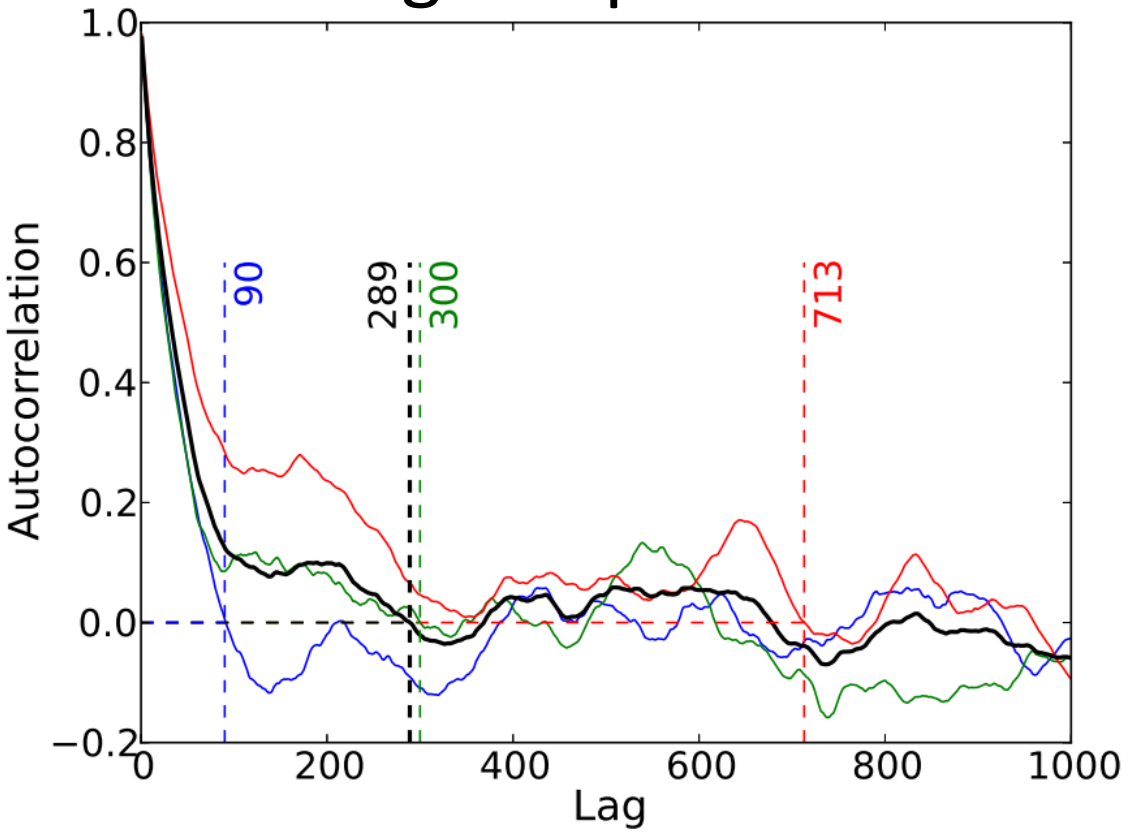
Hou et al. 2012

Ford 2006

Check Sufficient Effective Sample Size

Often, it is practical to run DEMCMC longer to make up for correlations among samples

- Check autocorrelation and other MCMC diagnostics for *all* parameters of interest



Nelson et al. 2014

How Can Things Still Go Wrong?

- Initial population too far from target density
  - Choose initial population close to target density
  - Test that results insensitive to choice
- Non-linear correlations between parameters
  - Results in long auto-correlation times
  - Increasingly problematic with higher-dimensional parameter spaces
- Multi-modal target density
  - DEMCMC can deal with a few viable modes, but autocorrelation time increases

Dealing with Multiple Modes

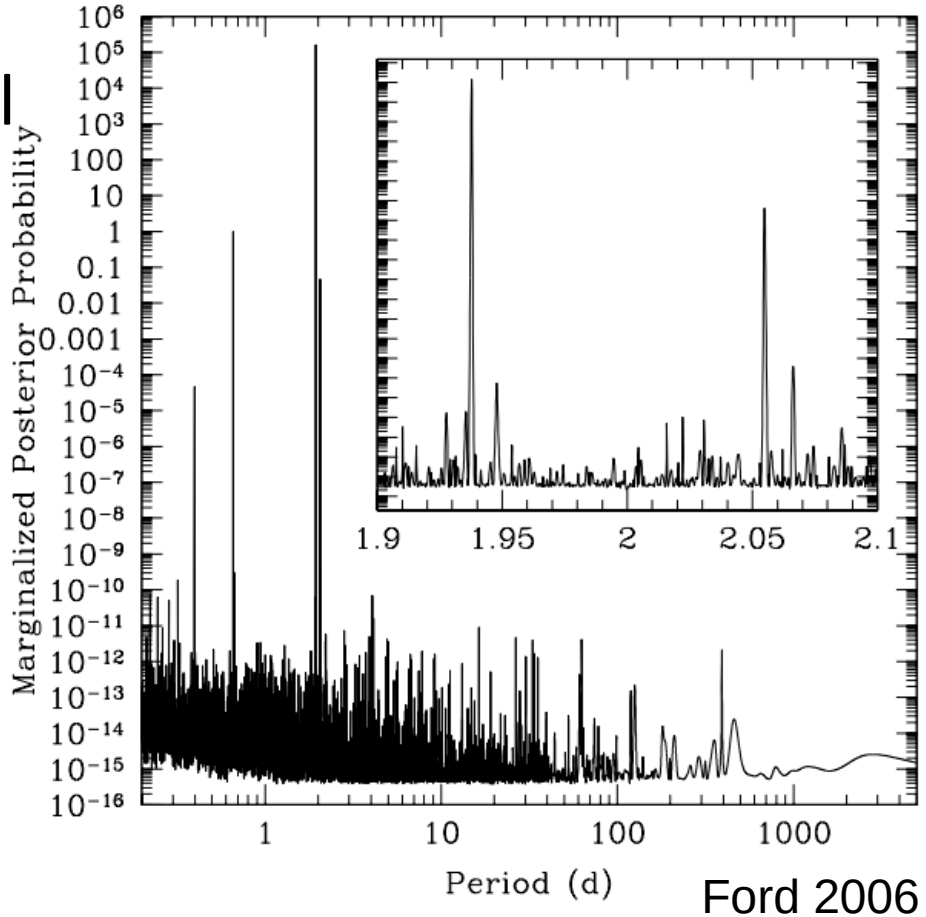
First, Identify Relevant Portion of Parameter Space

- Physical intuition
- Simplified statistical model
- Simplified physical model
- Analyze subset of data

Then, perform MCMC with good initial guesses

- Include samples from each viable mode

(See also Parallel Tempering)

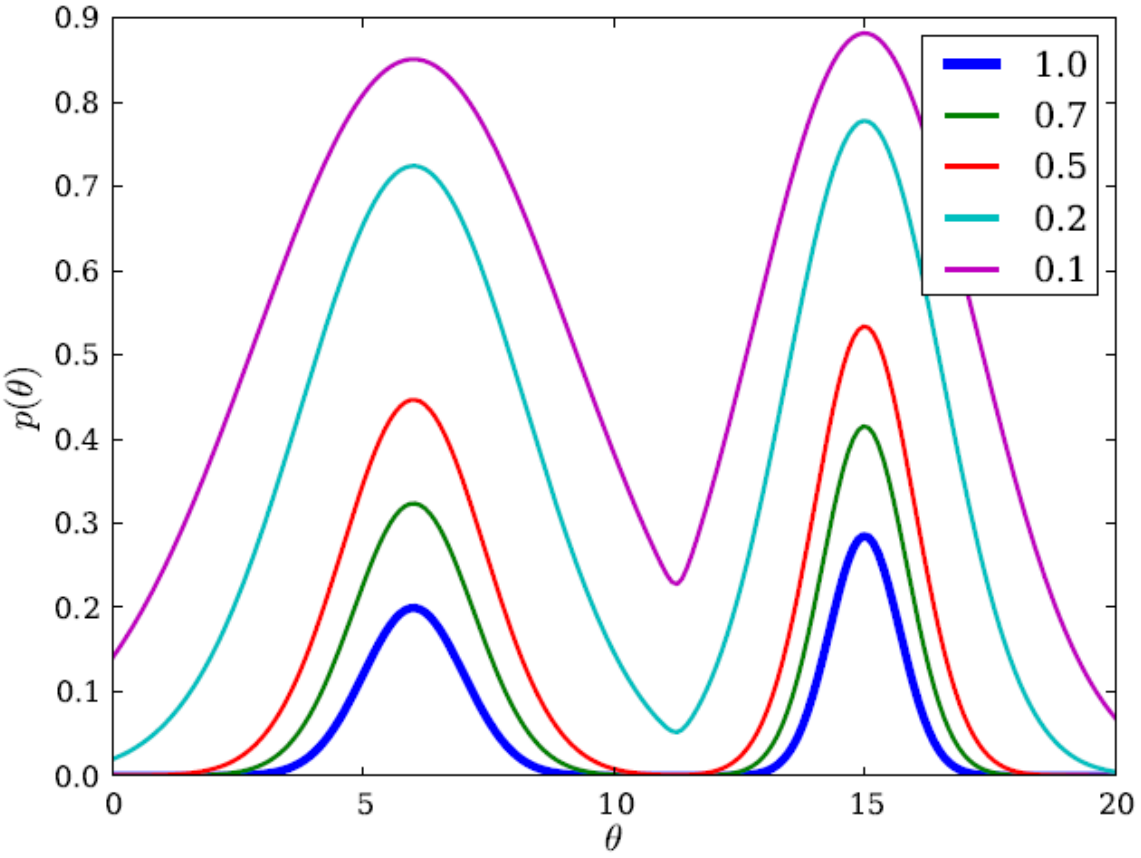


Ford 2006

Parallel Tempering MCMC

To enable large jumps, *anneal* or *temper* the posterior:

$$q_\beta(\theta) = [q(\theta)]^\beta, \qquad \beta \in [0, 1]$$



Loredo

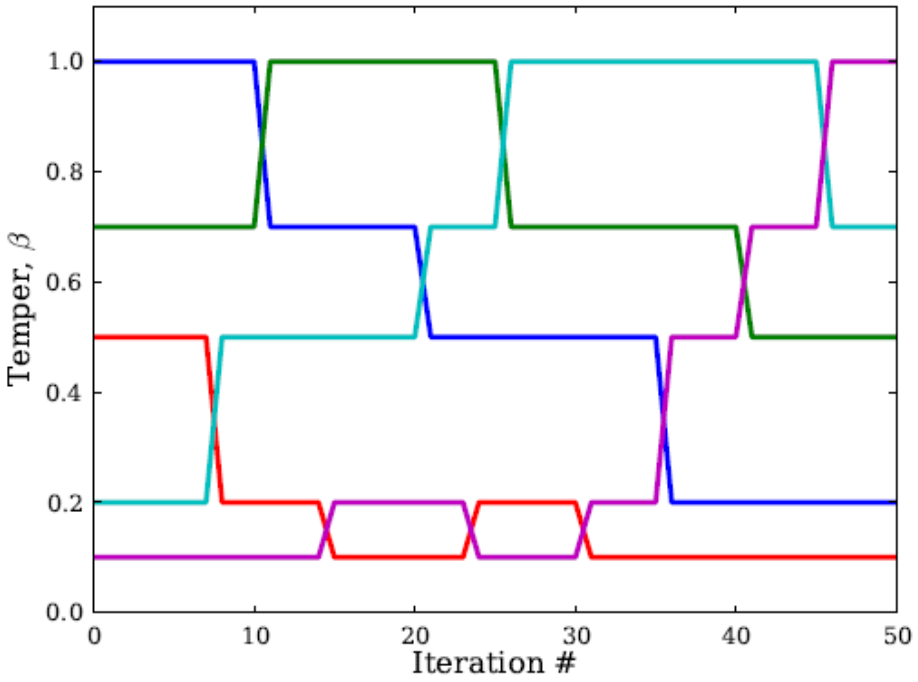
Consider a set of tempers ("inverse temperatures") $\{\beta_i\}$

Think of each $q_i = q_{\beta_i}$ as its own "model" with its own parameters, and construct a sampler for the joint distribution

$$p(\theta_1, \ldots, \theta_m) = \prod_i q_i(\theta_i)$$

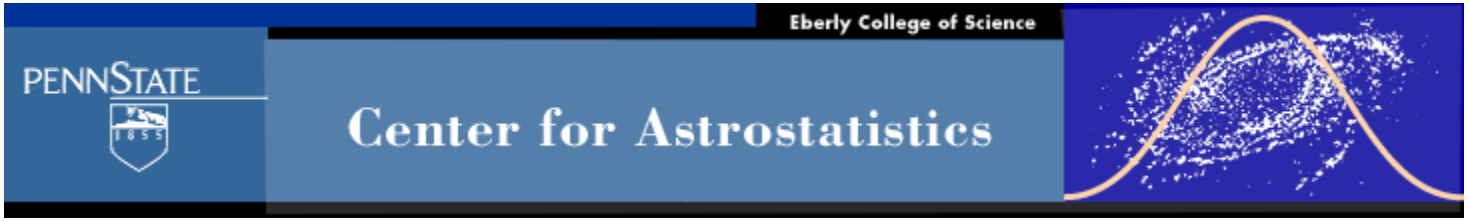
Alternate within-temper proposals and swap proposals between adjacent tempers

Swaps between tempered chains



Loredo

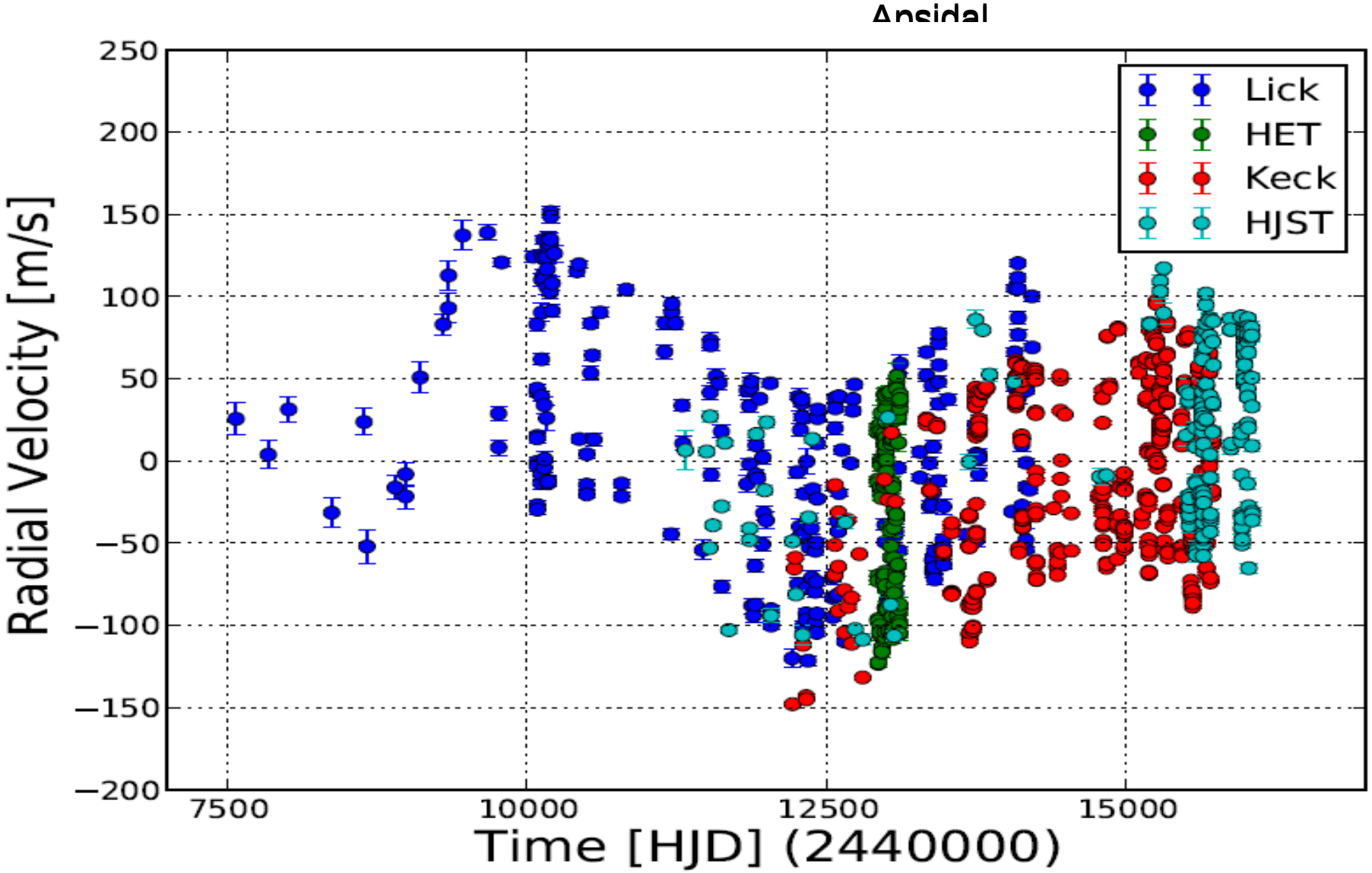
Pause for Questions



Example Application of DEMCMC

- Non-interacting systems & Doppler observations:
  - ~5x$N_{planets}$ physical model parameters
  - Model evaluation is very fast
  - Can require ~10^7 model evaluations
  - Parameter estimation is “solved” problem
  - Use dozens of physically-motivated proposals that deal with non-linear correlations
- Strongly Interacting planetary systems:
  - ~7x$N_{planets}$ physical model parameters
  - Can require ~10^{10} model evaluations
  - Model evaluation is slow, since requires n-body integration
  - Computationally demanding
  - Requires clever algorithms & parallel computation

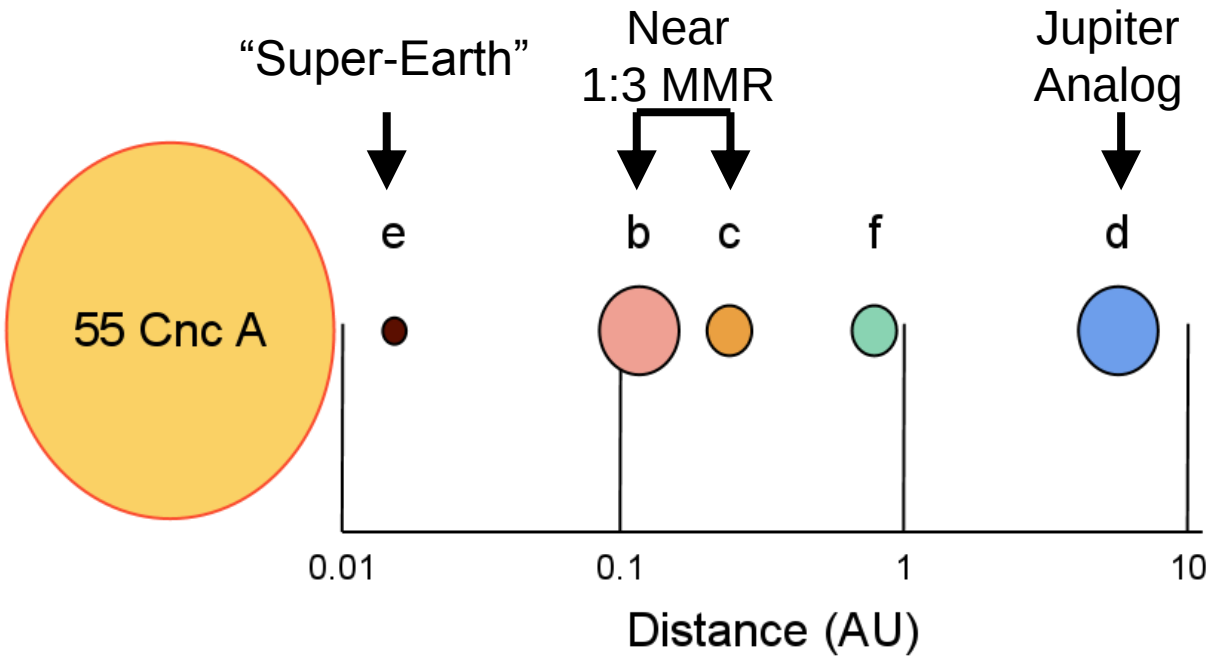
55 Cnc: An RV Classic



Butler+ 1996; Marcy+ 2002; McArthur+ 2004; Endl et al 2012; Nelson et al. 2014

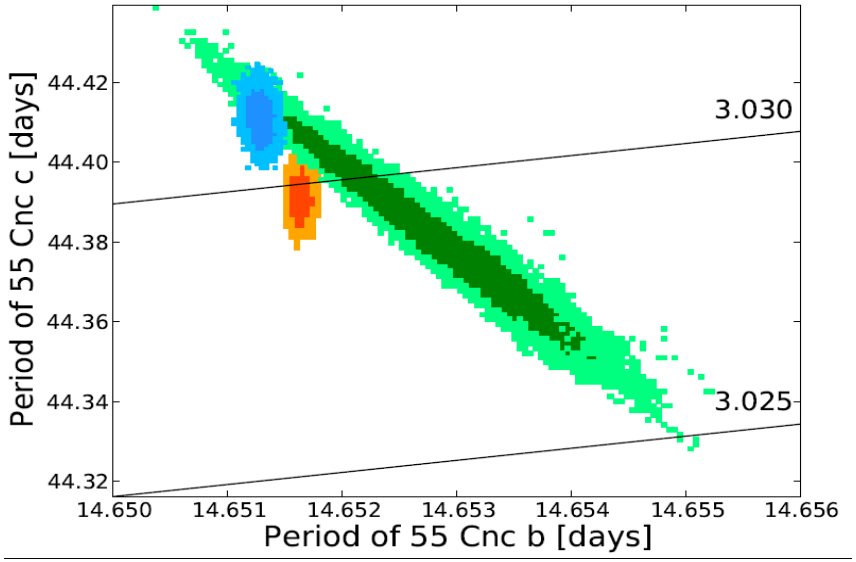
55 Cnc: Astroinformatics in Action

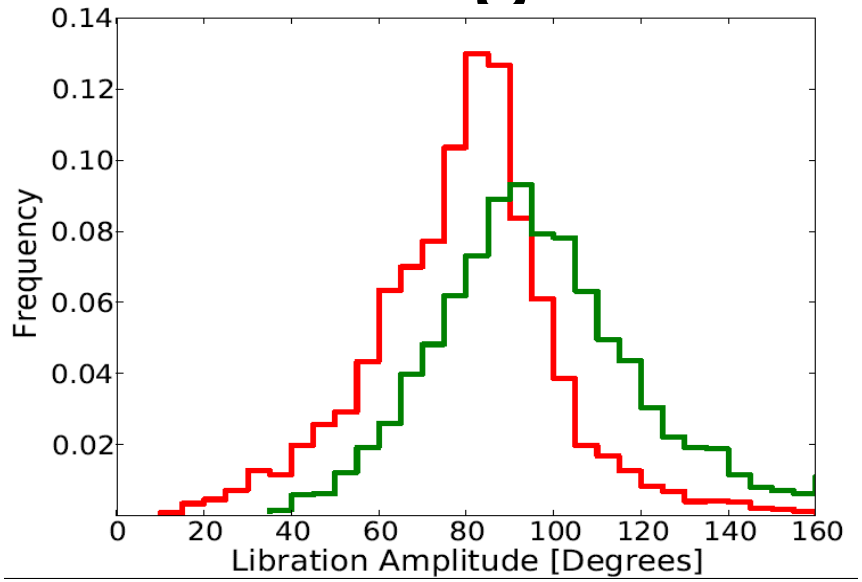
- 1,086 RVs from 4 observatories, spanning over 23 years
- Self-consistent Bayesian Analysis w/ full N-body treatment
- 40 dimensional parameter space
- ~3 weeks of computation w/ GPU (before stability tests)
- N-body integrations using Swarm-NG GPU (Dindar+ 2012)

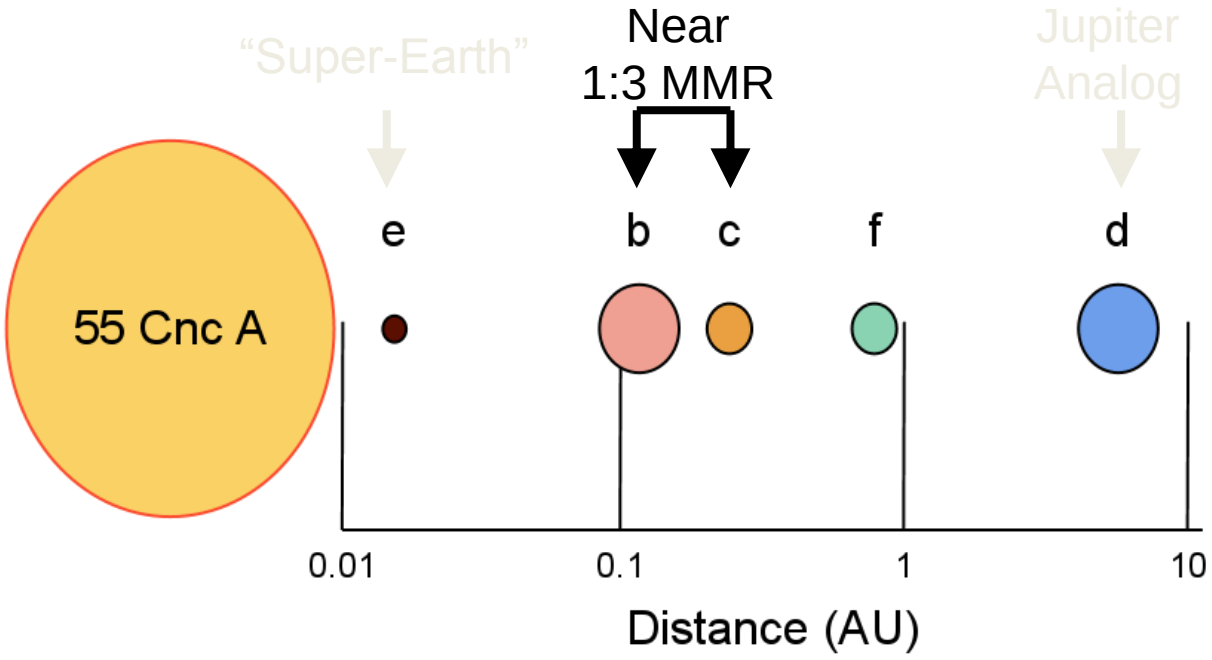


B. Nelson et al. 2014

55 Cnc: Evidence for Disk Migration

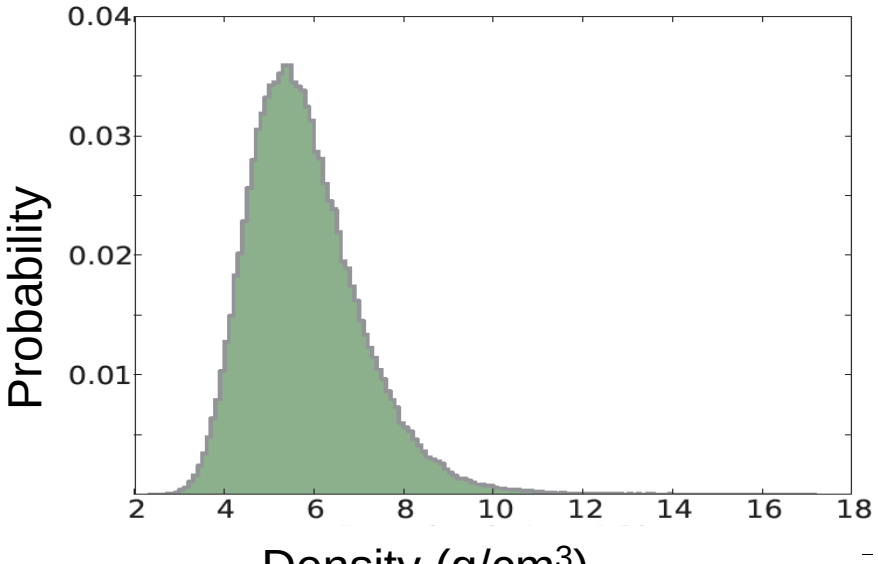


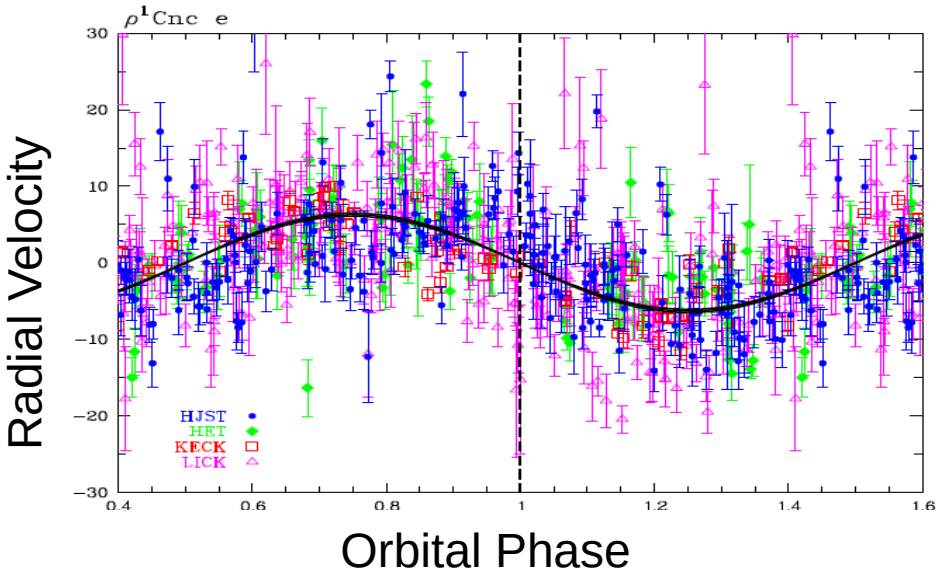




B. Nelson et al. 2014

55 Cnc: Density of a Super-Earth



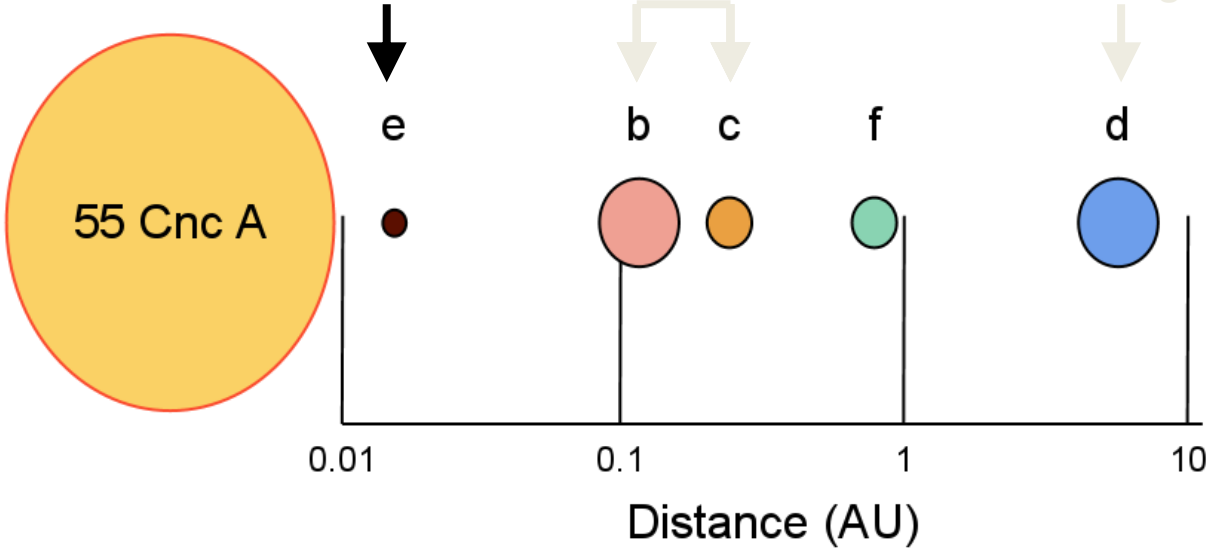


"Super-Earth"

Near 1:3 MMR

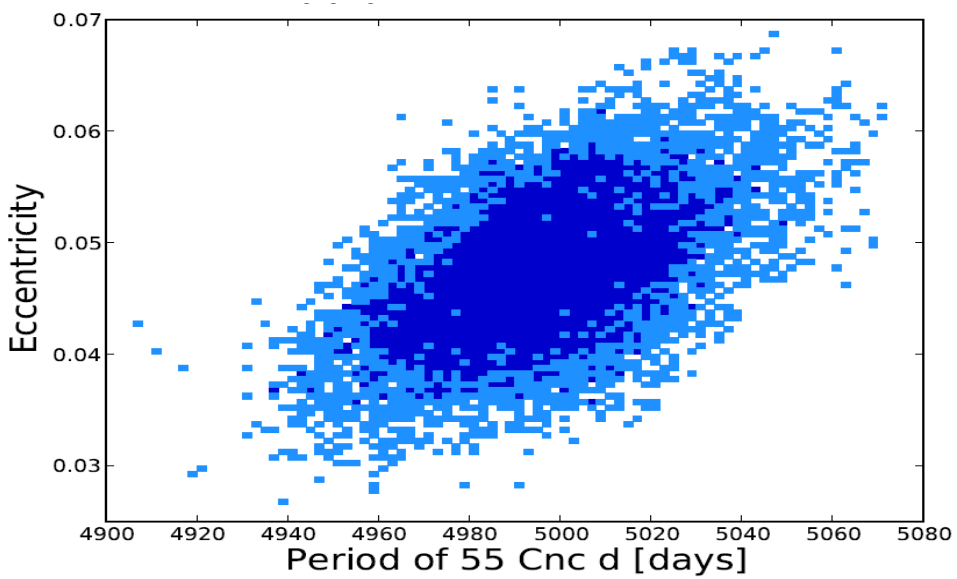
Jupiter Analog

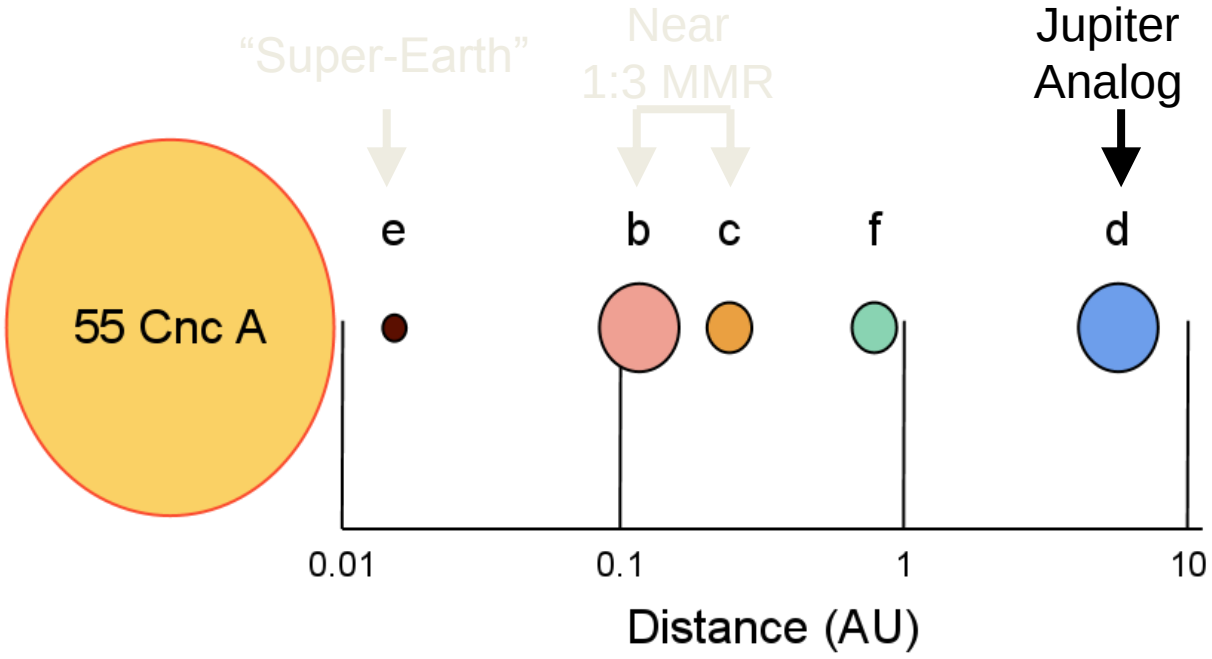
Endl et al. 2012



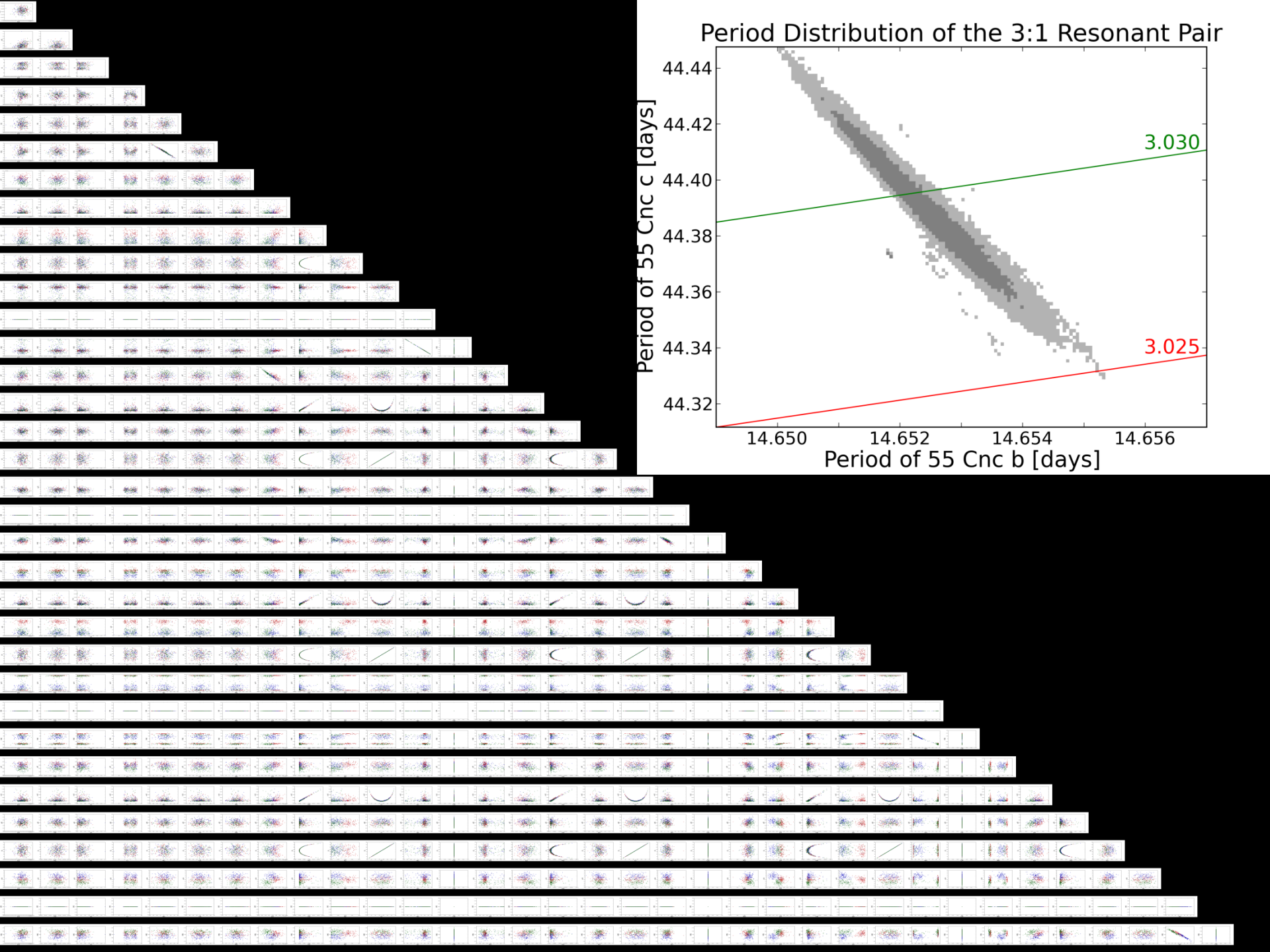
B. Nelson et al. 2014

55 Cnc: A True Jupiter Analog





B. Nelson et al. 2014

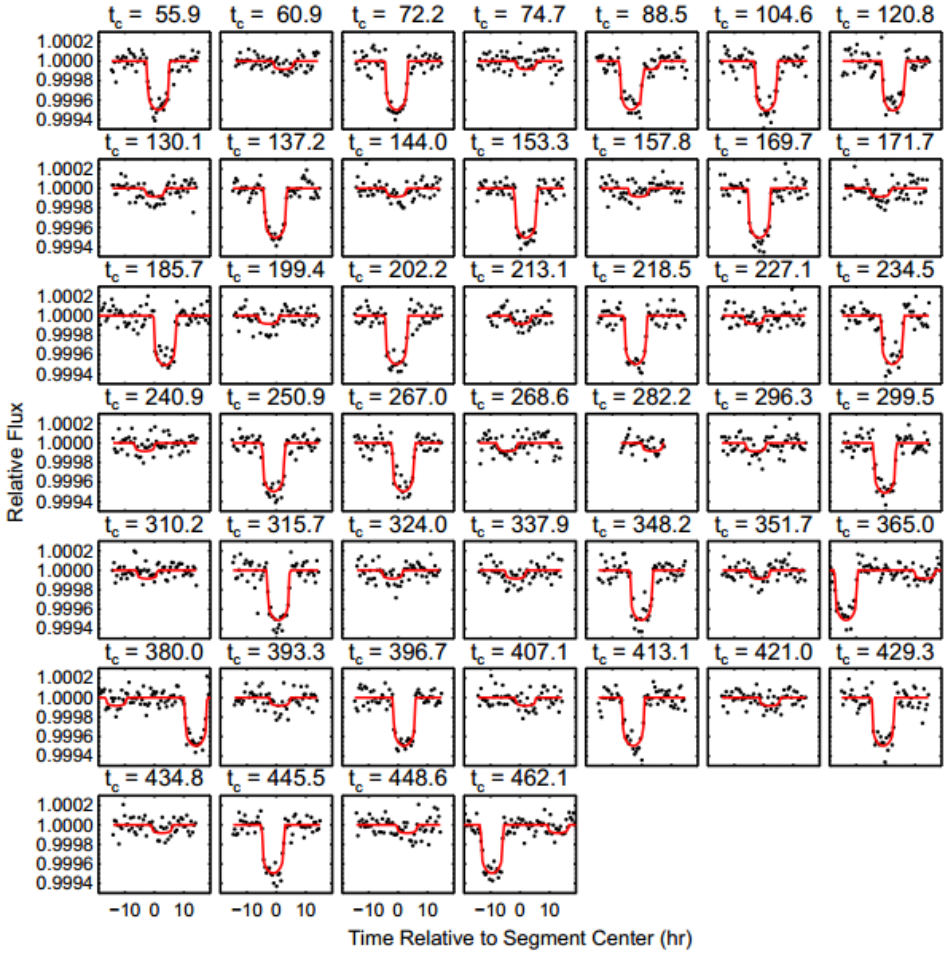

Period Distribution of the 3:1 Resonant Pair
Period of 55 Cnc c [days]
44.44
44.42
44.40
44.38
44.36
44.34
44.32
3.030
3.025
14.650
14.652
14.654
14.656
Period of 55 Cnc b [days]

Example Application of DEMCMC

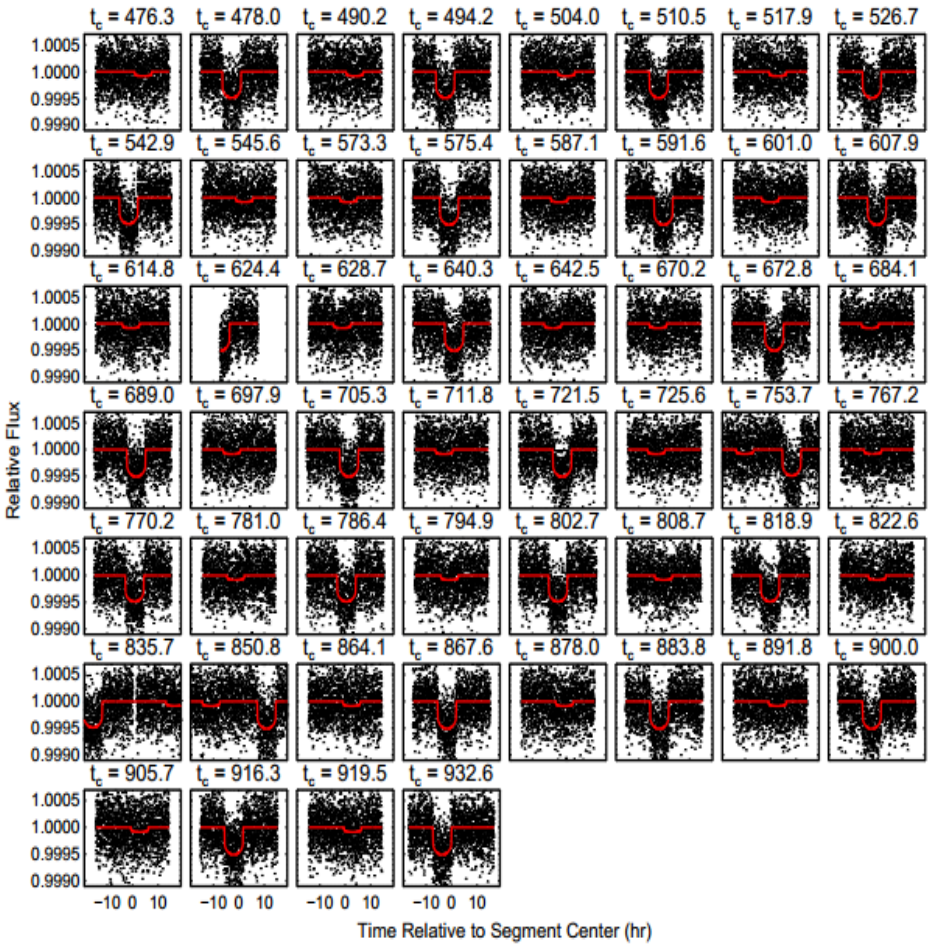
Measuring planet masses & orbits from Kepler light curves of stars with multiple transiting planets

- Physical Model:
  - Orbits: Either non-interacting or full n-body model
  - Light curves: Limb darkening, stellar activity
- Likelihood:
  - Assume each flux measurements has uncorrelated, Gaussian uncertainties, or
  - Could account for correlated noise

Characterizing Kepler's Systems with Interacting Planets: Kepler-36 b & c



30 min exposure time

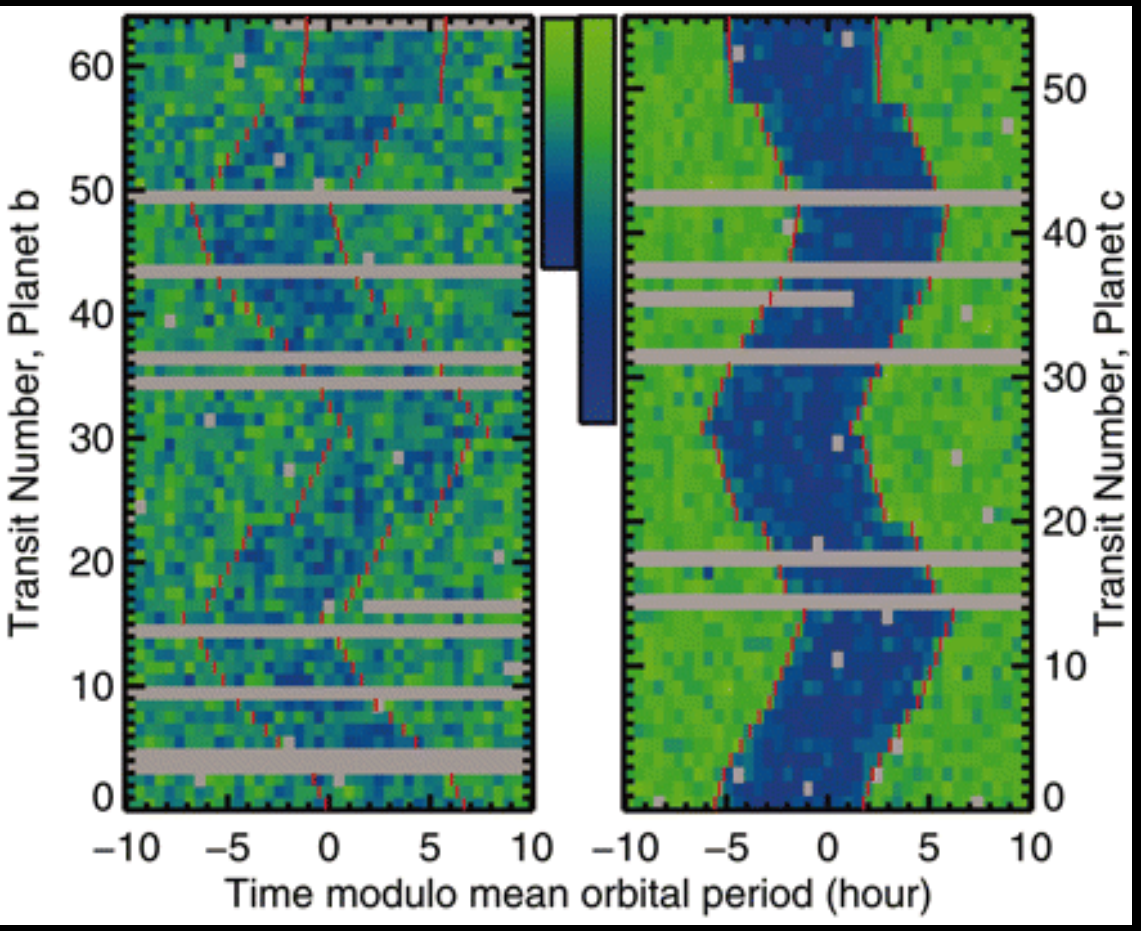


1 min exposure time

Carter et al. 2012

Characterizing Planet Masses for Rapidly Interacting Systems

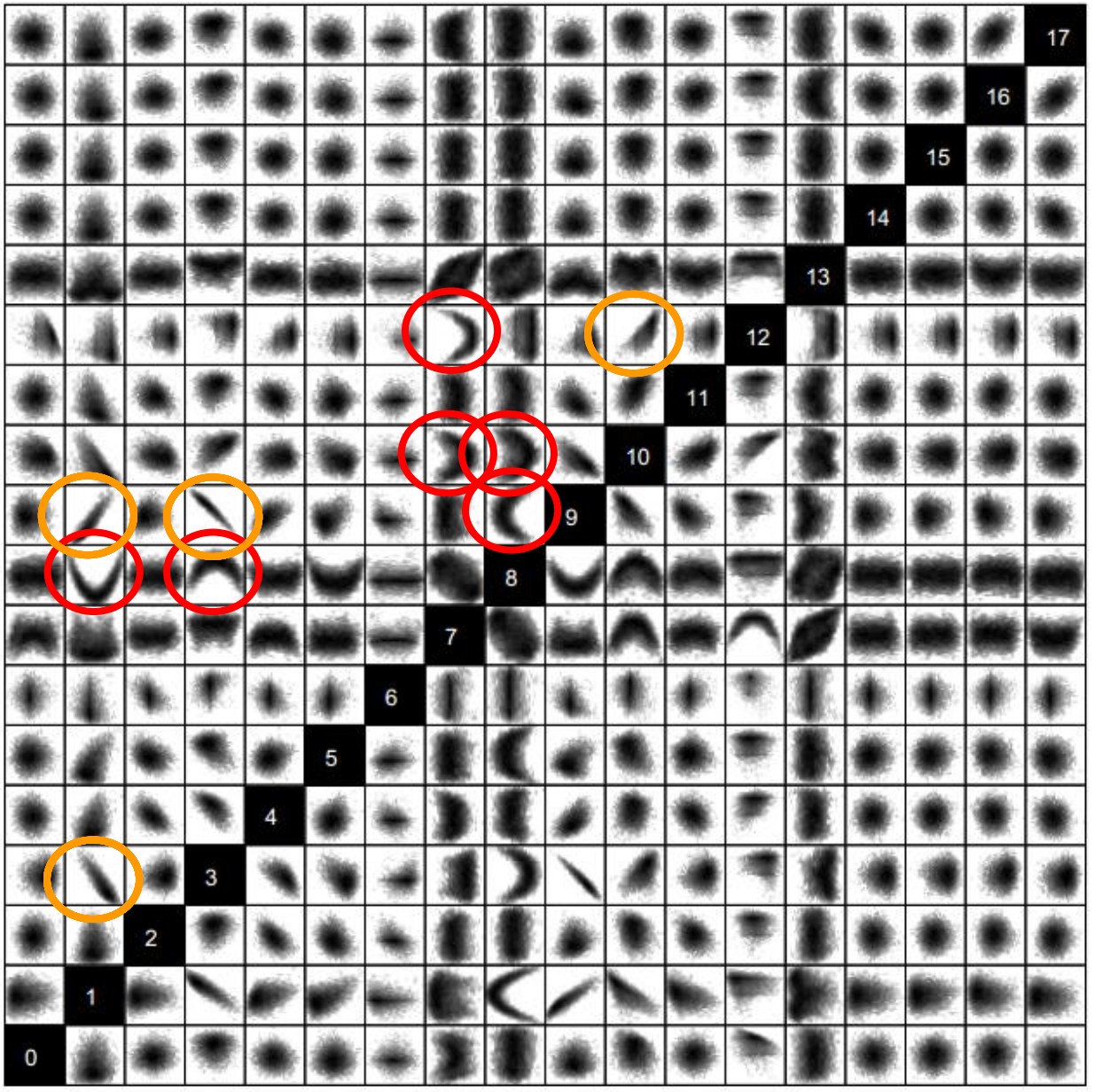
Kepler-36b&c: Chaotic due to 29:34 and 6:7 MMRs!



Carter et al. 2012; Deck et al. 2012

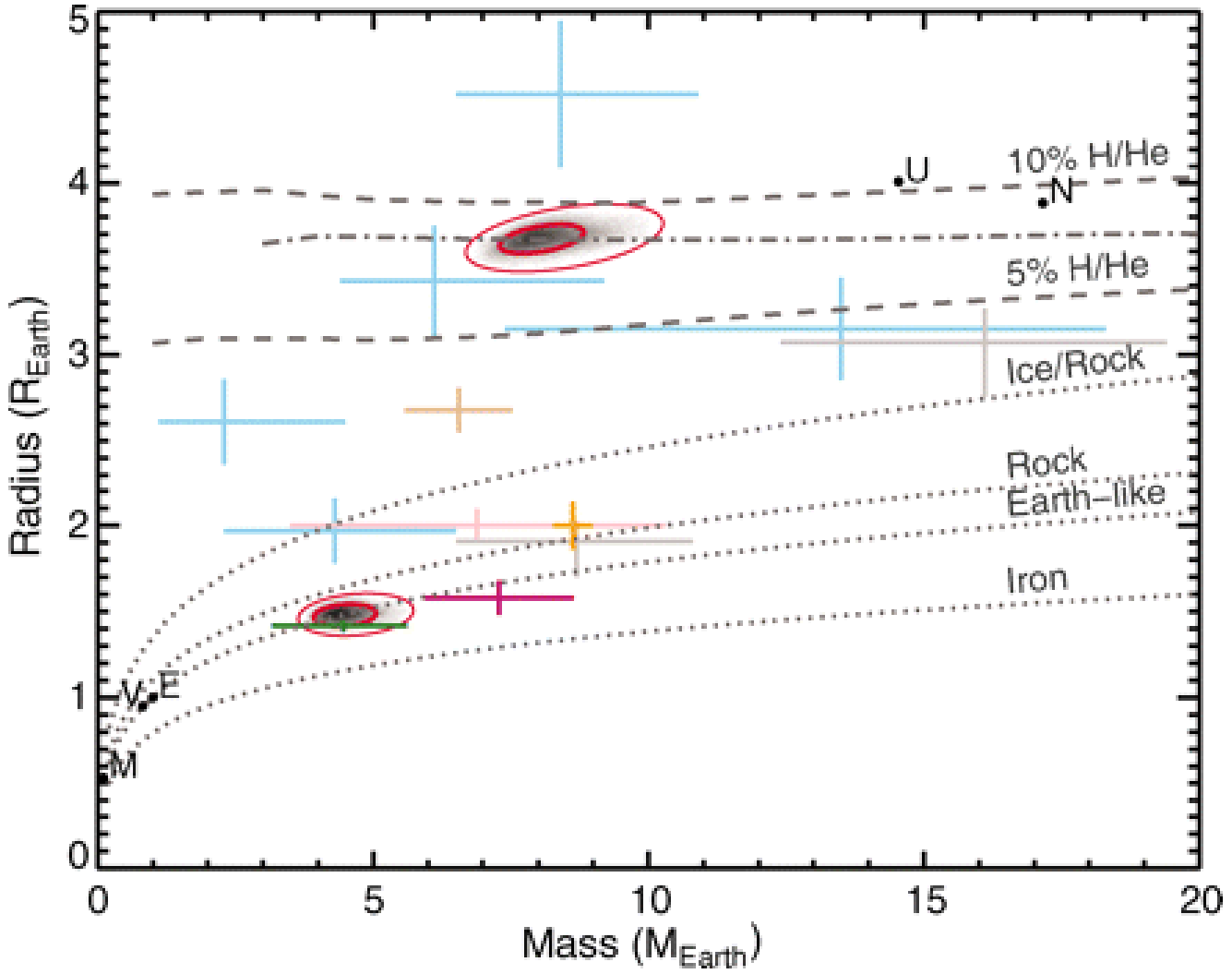
Two-Parameter Marginal Posterior Distributions

- Complex observational constraints
- Impractical to understand correlations a priori
- DEMCMC unphased by correlations
- ~10,000 CPU hours using
  128 cores (MPI) for
  ~1¼ years of observations



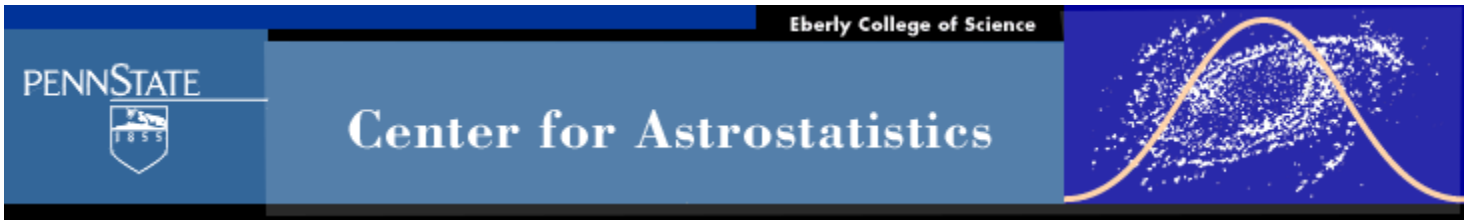
Carter et al. 2012

High-precision masses key for studying planet mass-radius relationship



Carter et al. 2012

Questions


PENNSTATE
1855
Eberly College of Science
Center for Astrostatistics