A Brief Introduction to Diagnosing MCMC Performance

Brian R. Moore
Department of Evolution & Ecology
UC, Davis
2014 Bodega Workshop

Model-Based Phylogenetic Inference

Model-based inference is based on the model

Model specification

model selection

model adequacy

model uncertainty/averaging

➡ Estimating under the model

likelihood optimization

MCMC simulation

MCMC Approximation of the Joint Posterior Probability Density

MCMC in theory and practice

MCMC in theory...

an <u>appropriately constructed</u> and <u>adequately run</u> chain is guaranteed to provide an arbitrarily precise description of the joint stationary density

MCMC in practice...

although a given sampler may work well in most cases, all samplers will fail in some cases, and is not guaranteed to work for any given case

Q. When do we know that the MCMC provides an accurate approximation for a given empirical analysis?

A.

NEVER!

MCMC Approximation of the Joint Posterior Probability Density

MCMC performance and OCD

It is not sufficient to merely be deeply concerned about MCMC performance...you need to be completely obsessed about it!

for any Bayesian inference based on MCMC

particularly for complex models/inference problems

WE ARE HERE

WE SHOULD BE HERE

I'LL BE HERE

careless   **careful**   **paranoid**

MCMC Approximation of the Joint Posterior Probability Density

Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review

Mary Kathryn COWLES and Bradley P. CARLIN

A critical issue for users of Markov chain Monte Carlo (MCMC) methods in applications is how to determine when it is safe to stop sampling and use the samples to estimate characteristics of the distribution of interest. Research into methods of computing theoretical convergence bounds holds promise for the future but to date has yielded relatively little of practical use in applied work. Consequently, most MCMC users address the convergence problem by applying diagnostic tools to the output produced by running their samplers. After giving a brief overview of the area, we provide an expository review of 13 convergence diagnostics, describing the theoretical basis and practical implementation of each. We then compare their performance in two simple models and conclude that all of the methods can fail to detect the sorts of convergence failure that they were designed to identify. We thus recommend a combination of strategies aimed at evaluating and accelerating MCMC sampler convergence, including applying diagnostic procedures to a small number of parallel chains, monitoring autocorrelations and cross-correlations, and modifying parameterizations or sampling algorithms appropriately. We emphasize, however, that it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution.

KEY WORDS: Autocorrelation; Gibbs sampler; Metropolis-Hastings algorithm.

Outline

I. A review of the basics

II. Diagnosing MCMC performance

III. Diagnostics based on single chains

IV. Diagnostics based on the prior

V. Diagnostics based on multiple chains

Approximating the Joint Posterior Probability Density with MCMC

The Metropolis-Hastings algorithm

1. Initialize the chain with some random values for all parameters, including the tree with branch lengths, $\Theta = (\tau, \nu)$

2. Select a parameter to change according to it's proposal probability

3. Propose a change to the selected parameter using the parameter-specific proposal mechanism

4. Calculate the probability of accepting the proposed change

5. Generate a uniform random variable, $U[0,1]$, accept if $R > U$

6. Repeat steps 2–5 an 'adequate' number of times

Metropolis et al. (1953); Hastings (1970)

Outline

Assessing MCMC Performance: Three Main Issues

1. Convergence

Has the chain (robot) successfully targeted the stationary distribution?

2. Mixing

Is the chain (robot) efficiently integrating over the joint posterior probability?

3. Sampling intensity

Have we collected enough samples to adequately describe the posterior probability distribution?

Assessing MCMC Performance: Common Tools

Tracer

Visual inspection of continuous model parameters

AWTY

Visual inspection of discrete (tree) model parameter

Implementation-specific tools

MrBayes

ASDSF, PSRF, comparetrees, etc.

Outline

Assessing MCMC Performance: Diagnostics Based on Single Runs
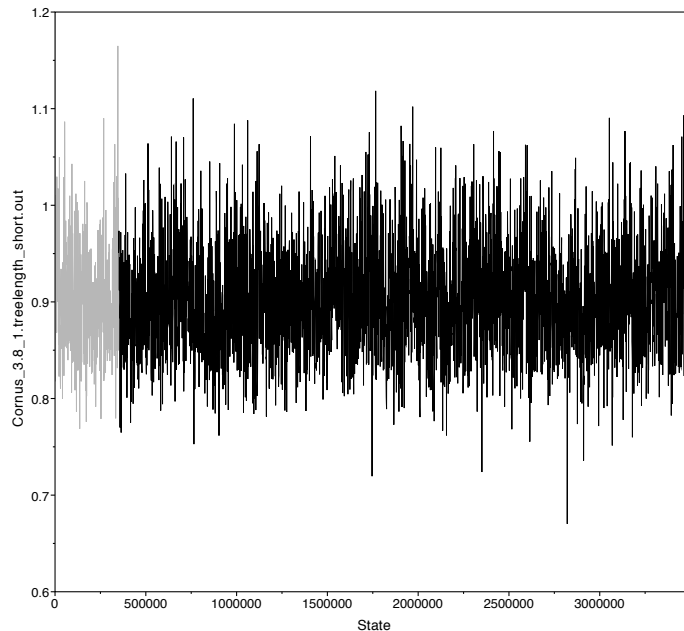
1. Convergence diagnostics

   (i) Time-series plots of parameter estimates

   - continuous parameters (e.g., substitution rates): Tracer
     - some parameters are more reliable than others
     - steps may occur!

Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Tracer plots of tree-length at two stages of a single MrBayes run

all looks good...   until it doesn't

fast*   slow*

lnL   base freq.   sub. rates   ASRV   TL   topology

*somewhat data-set dependent

Assessing MCMC Performance:
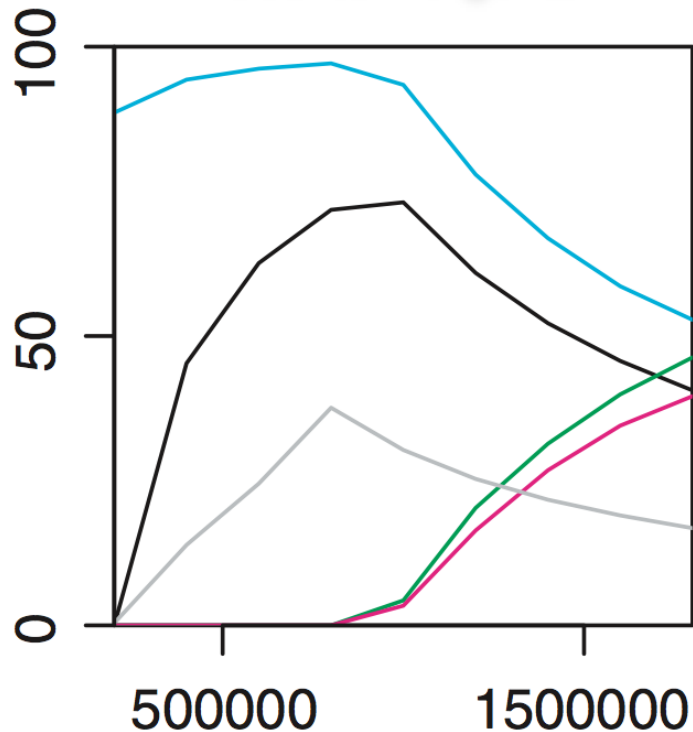# Diagnostics Based on Single Runs

1. Convergence diagnostics

(i) Time-series plots of parameter estimates

- continuous parameters (e.g., substitution rates): Tracer
  - some parameters are more reliable than others
  - steps may occur!
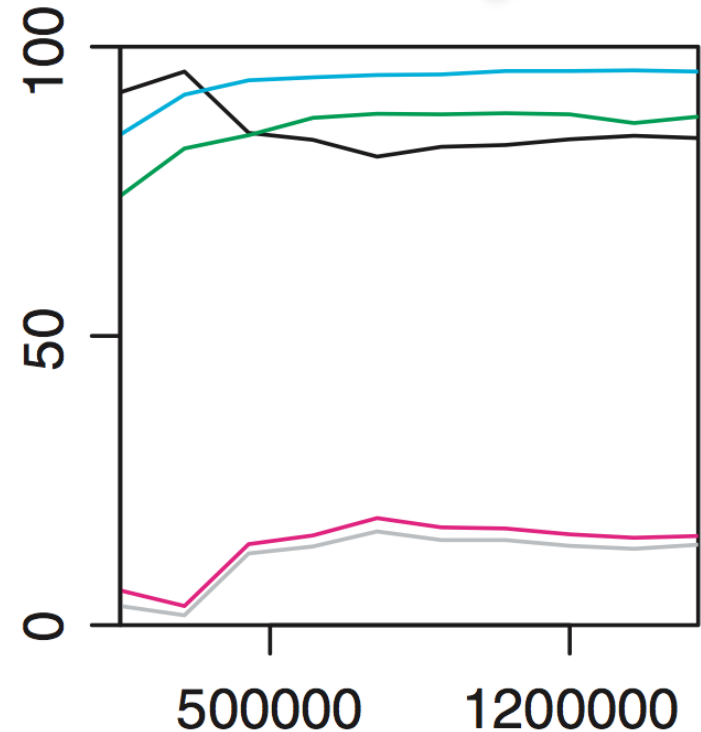- discrete parameters (e.g., cumulative bi-partition frequency): AWTY

Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: AWTY plots of cumulative bi-partition frequency of 5 nodes

Nylander et al. (2008)

Assessing MCMC Performance: Diagnostics Based on Single Runs

1. Convergence diagnostics

(i) Time-series plots of parameter estimates

(ii) Geweke diagnostic: coda, BOA

- A test for equality of the means of the first and last part of a Markov chain (by default the first 10% and the last 50%)

- If the samples are drawn from the stationary distribution, the two means should equal and Geweke's statistic has an asymptotically standard normal distribution

Geweke (1992)

Assessing MCMC Performance: Diagnostics Based on Single Runs

1. Convergence diagnostics

    (i) Time-series plots of parameter estimates

    (ii) Geweke diagnostic: coda, BOA

    (iii) Heidelberg-Welch diagnostic: coda, BOA

- uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution

  - This test is successively applied, first to the whole chain, then after discarding the first 10%, 20%, ... of the samples until either the null hypothesis is accepted, or 50% of the chain has been discarded

  - The latter outcome constitutes "failure" of the test and indicates that a longer run is needed

  - Otherwise, the number of iterations to keep and the number to discard (burn-in) are reported

Heidelberg & Welch (1983)

Assessing MCMC Performance: Diagnostics Based on Single Runs

1. **Convergence diagnostics**

   (i) Time-series plots of parameter estimates

   (ii) Geweke diagnostic: coda, BOA

   (iii) Heidelberg-Welch diagnostic: coda, BOA

   (…) Many others

Heidelberg & Welch (1983)

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

    (i) Form of the time-series plots of parameter estimates

        • continuous parameters (e.g., substitution rates): Tracer warm and fuzzy caterpillars

Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Tracer plots of relative-rate multipliers from two MrBayes runs
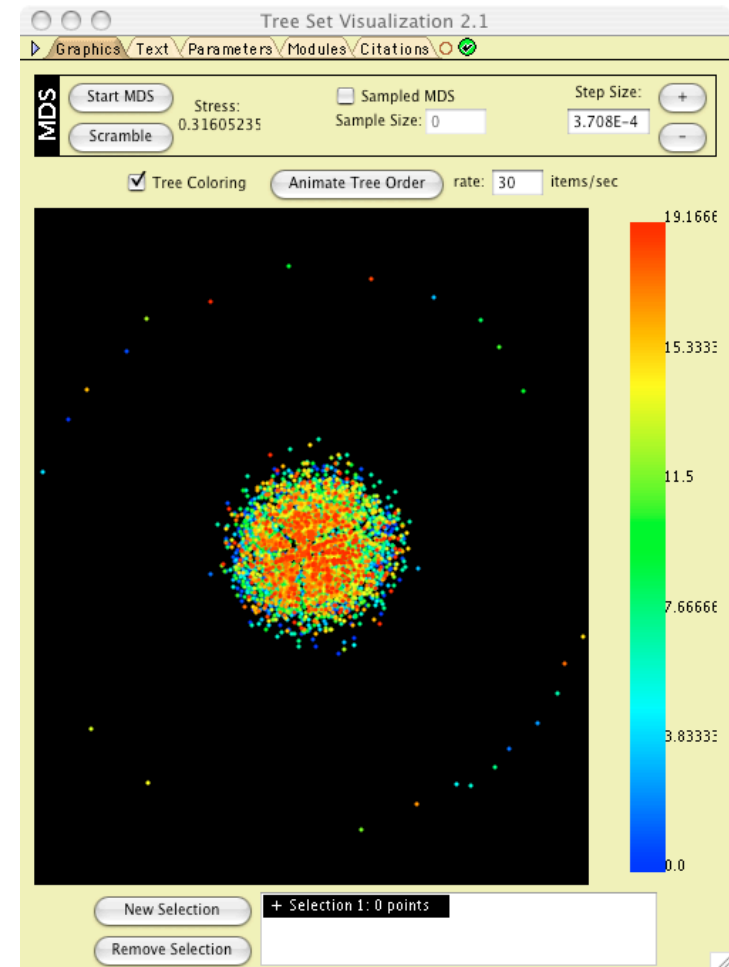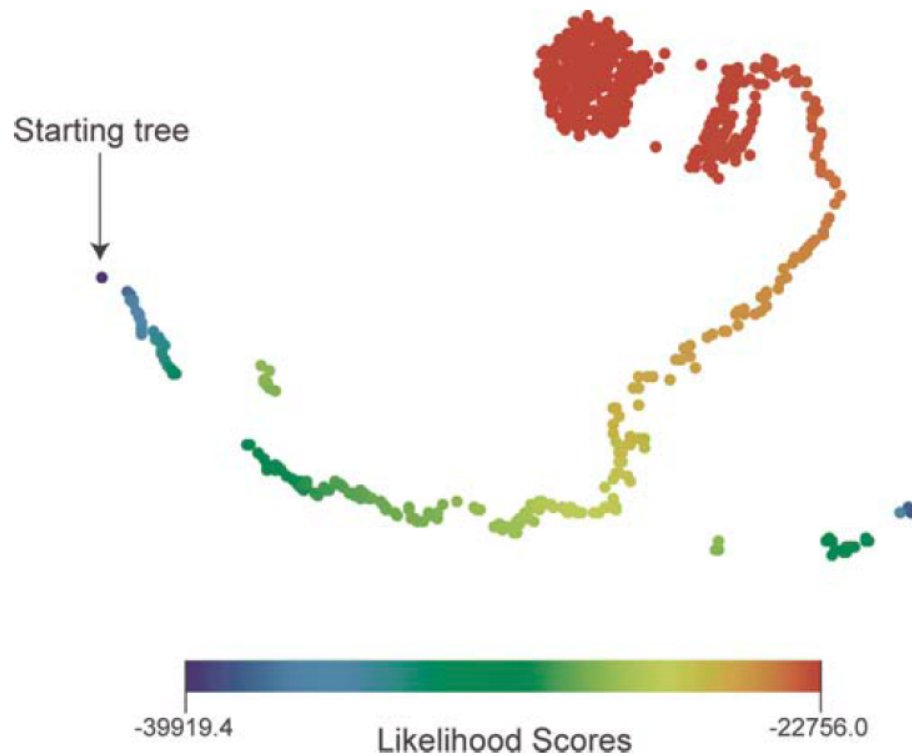
bad mixing

better mixing

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

    (i) Form of the time-series plots of parameter estimates

- continuous parameters (e.g., substitution rates): Tracer warm and fuzzy caterpillars

- discrete parameters:

  - distances among sampled topologies: TreeSetViz

Assessing MCMC Performance:
Diagnostics Based on Single Runs

TreeSetViz

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

(i) Form of the time-series plots of parameter estimates

- continuous parameters (e.g., substitution rates)--Tracer
  warm and fuzzy caterpillars

- discrete parameters:
  - distances among sampled topologies: TreeSetViz

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters: MrBayes, BEAST, etc.
  rates should ideally fall in the ~20–70% range

Assessing MCMC Performance:
Diagnostics Based on Single Runs

Example: Tracer plots of relative-rate multipliers from two MrBayes runs

bad mixing

better mixing


```
Acceptance rates for the moves in the "cold" chain of run 1:
   With prob.   Chain accepted changes to
    13.61 %    param. 1 (revmat) with Dirichlet proposal
         .
         .
         .
     0.04 %    param. 34 (rate multiplier) Dirichlet proposal
     6.59 %    param. 35 (topology and branch lengths) TBR
    14.06 %    param. 35 (topology and branch lengths) LOCAL
```

```
Acceptance rates for the moves in the "cold" chain of run 1:
   With prob.   Chain accepted changes to
    33.30 %    param. 1 (revmat) with Dirichlet proposal
         .
         .
         .
    19.13 %    param. 34 (rate multiplier) Dirichlet proposal
    17.40 %    param. 35 (topology and branch lengths) TBR
    29.76 %    param. 35 (topology and branch lengths) LOCAL
```

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

(i) Form of the time-series plots of parameter estimates

- continuous parameters (e.g., substitution rates)--Tracer
  warm and fuzzy caterpillars

- discrete parameters:

  - distances among sampled topologies: TreeSetViz

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters: MrBayes, BEAST, etc.
  rates should ideally fall in the ~20–70% range

- acceptance rates can be controlled by varying the scale of the
  tuning parameters for the relevant proposal mechanisms
  to increase rates, decrease scale & vice versa

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

(i) Form of the time-series plots of parameter estimates

- continuous parameters (e.g., substitution rates)--Tracer
  warm and fuzzy caterpillars

- discrete parameters:
  - distances among sampled topologies: TreeSetViz

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters--MrBayes, BEAST, etc.
  rates should ideally fall in the ~20-70% range

- acceptance rates can be controlled by varying the scale of the
  tuning parameters for the relevant proposal mechanisms
  to increase rates, decrease scale & vice versa

(iii) Form of the marginal posterior probability densities

- continuous parameters (e.g., substitution rates): Tracer
  beware of porcupine roadkill

Assessing MCMC Performance:
Diagnostics Based on Single Runs

Example: Parameter estimates for relative-rate multipliers from two MrBayes runs

bad mixing

better mixing


```
Acceptance rates for the moves in the "cold" chain of run 1:
   With prob.   Chain accepted changes to
     13.61 %    param. 1 (revmat) with Dirichlet proposal
                 .
                 .
                 .
      0.04 %    param. 34 (rate multiplier) Dirichlet proposal
      6.59 %    param. 35 (topology and branch lengths) TBR
     14.06 %    param. 35 (topology and branch lengths) LOCAL
```
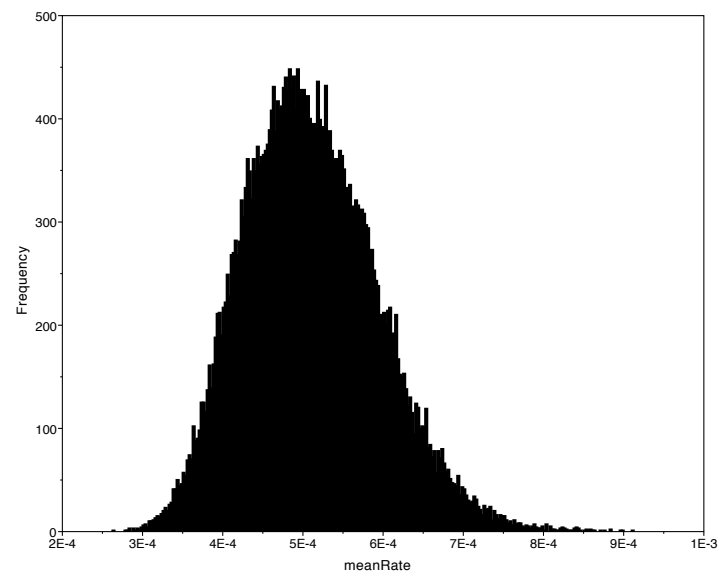
```
Acceptance rates for the moves in the "cold" chain of run 1:
   With prob.   Chain accepted changes to
     33.30 %    param. 1 (revmat) with Dirichlet proposal
                 .
                 .
                 .
     19.13 %    param. 34 (rate multiplier) Dirichlet proposal
     17.40 %    param. 35 (topology and branch lengths) TBR
     29.76 %    param. 35 (topology and branch lengths) LOCAL
```

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

(iv) Autocorrelation time (ACT) of parameter samples

The lag (number of cycles) it takes for autocorrelation in parameter values to break down

The lag k autocorrelation ρ_k is the correlation every draw and its kth lag:

$$\rho_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We would expect the kth lag autocorrelation to be smaller as k increases (our 1st and 100th draws should be less correlated than our 1st and 2nd draws).

If autocorrelation is still relatively high for higher values of k, this indicates high degree of correlation between our draws and slow mixing.

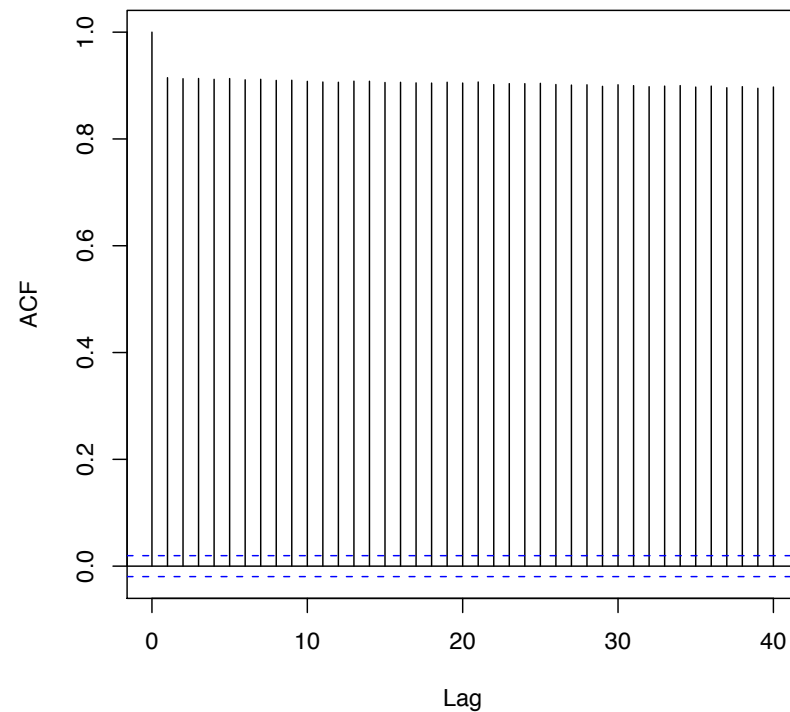# Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

(iv) Autocorrelation time (ACT) of parameter samples
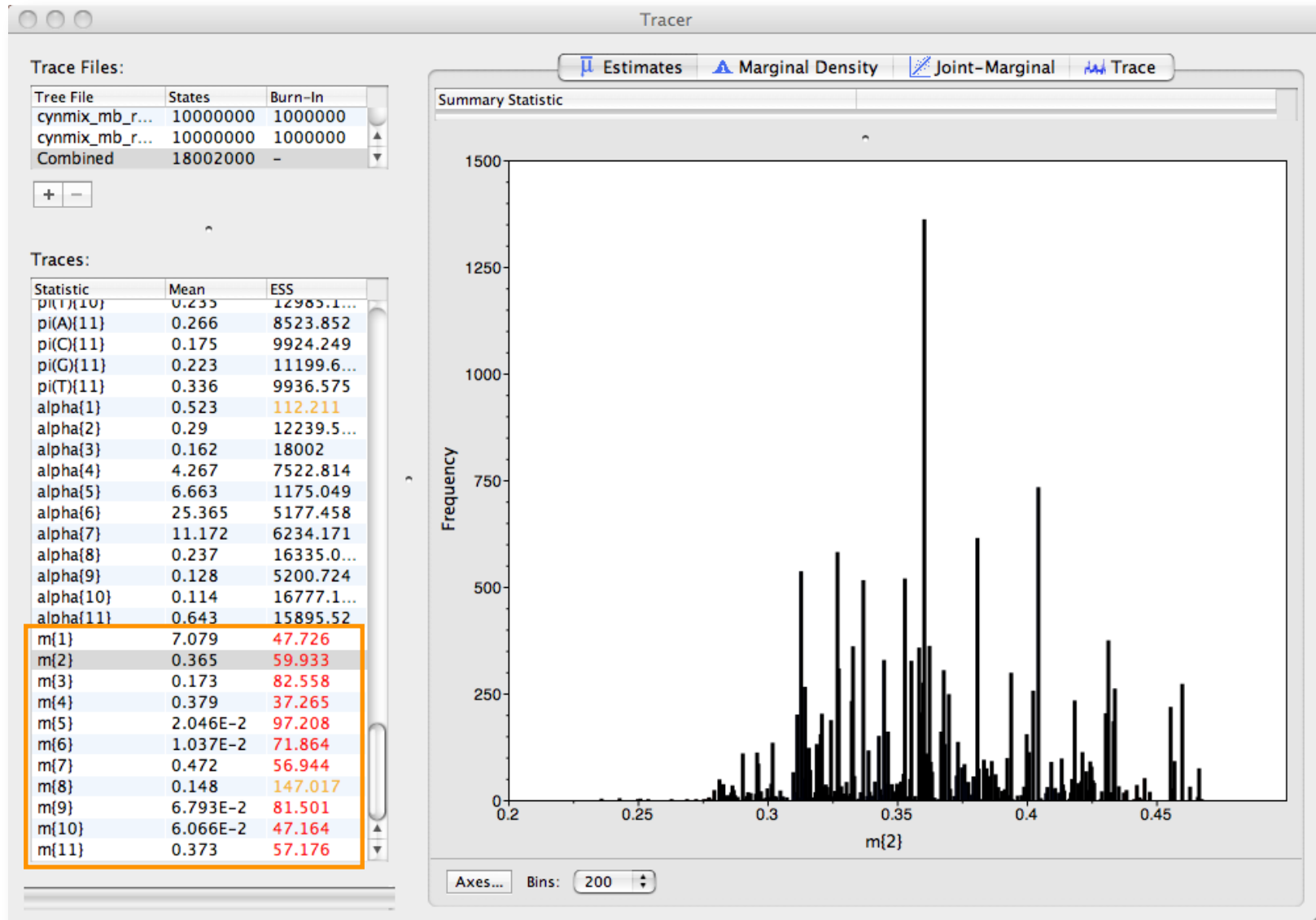

efficient mixing

slow mixing

3. Sample-size diagnostics

    (i) Effective Sample Size (ESS) diagnostic

- number of samples/autocorrelation time (ACT)
- continuous parameters (e.g., substitution rates): Tracer

Assessing MCMC Performance:
Diagnostics Based on Single Runs

Example: ESS values for relative-rate multipliers from two MrBayes runs

low intensity/slow mixing
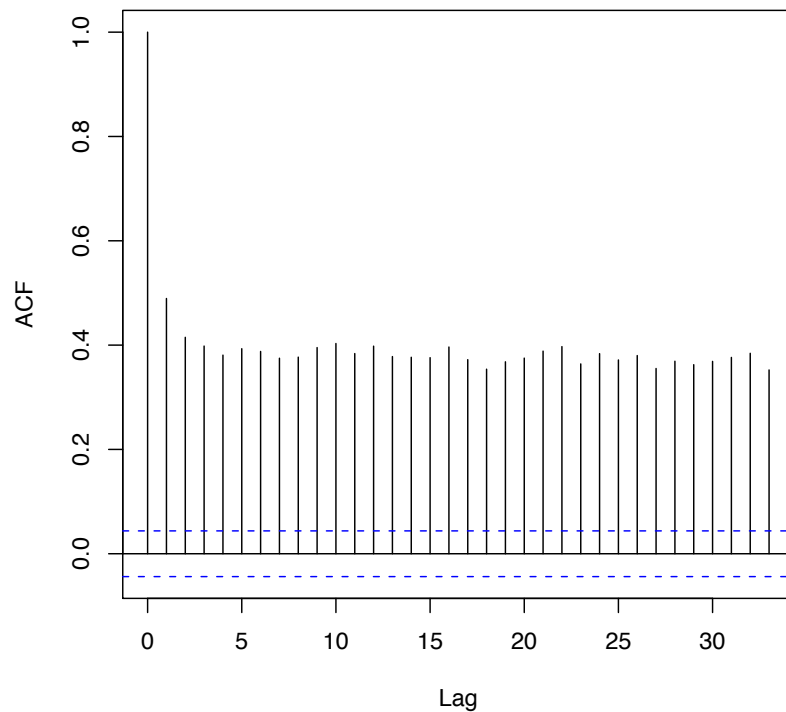
Assessing MCMC Performance: Diagnostics Based on Single Runs

3. Sample-size diagnostics

   (i) Effective Sample Size (ESS) diagnostic
- number of samples/autocorrelation time (ACT)
- continuous parameters (e.g., substitution rates): Tracer

   (ii) Form of the marginal posterior probability densities
- continuous parameters (e.g., substitution rates): Tracer

      brother of porcupine roadkill

      ensure SAE compliance!

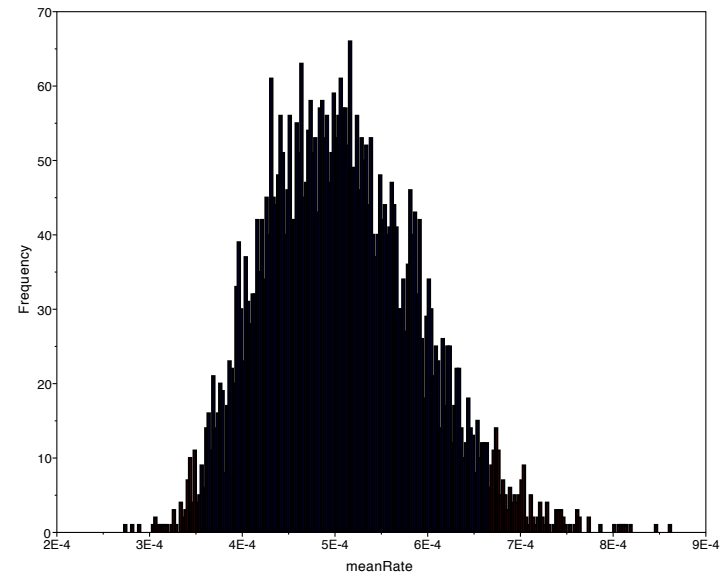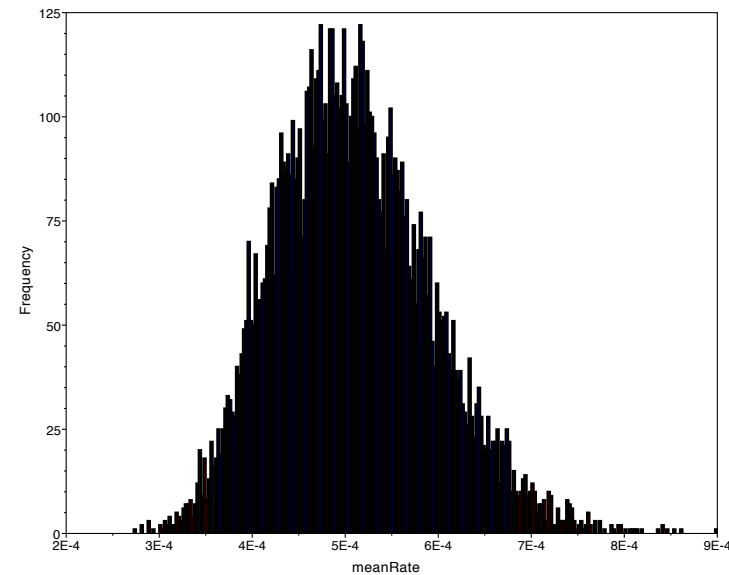# Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Parameter estimates for mean-rate multipliers from BEAST runs

low intensity

better intensity

1M cycles

5M cycles

- inadequate chain length/poor mixing

Assessing MCMC Performance:
Diagnostics Based on Single Runs

Example: Parameter estimates for mean-rate multipliers from BEAST runs

low intensity

better intensity

1M cycles

10M cycles

- inadequate chain length/poor mixing

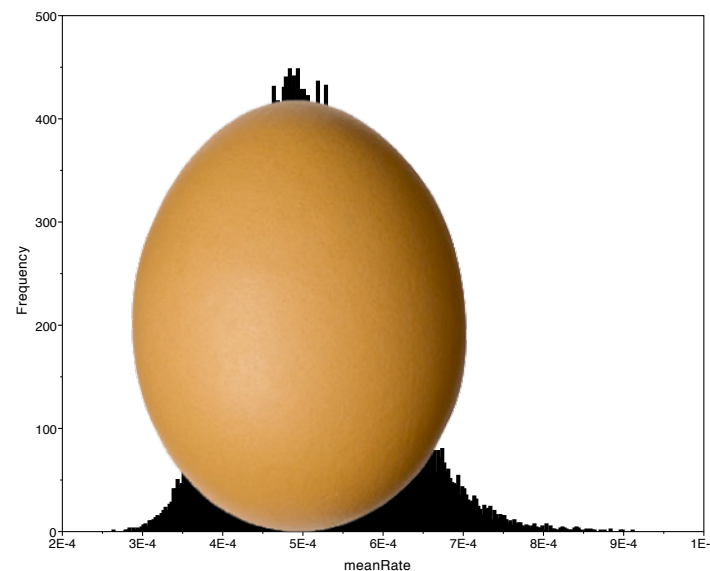# Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Parameter estimates for relative-rate multipliers from two MrBayes runs

low intensity

better intensity

1M cycles

40M cycles

- ESS can be increased by reducing the sampling frequency/increasing burin in
- All continuous parameters should be SAE

Assessing MCMC Performance: Diagnostics Based on Single Runs

MCMC pathologies

Parameter interaction between I+G mixture for among-site rate variation

a relatively large p a

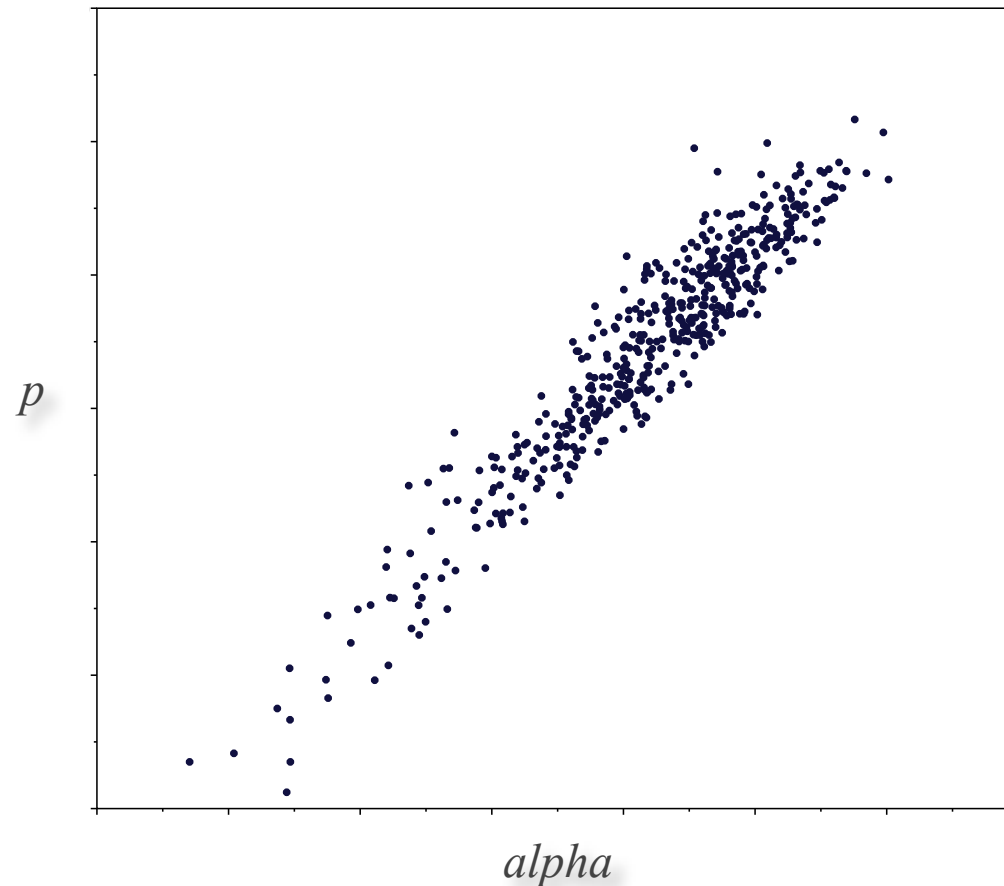a relatively small p a small value of a

alpha

p
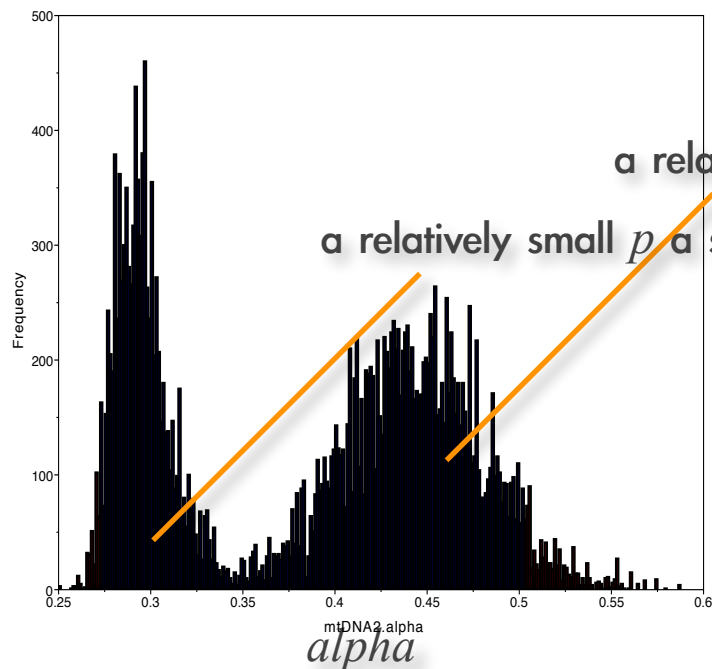
- multi-modal marginal densities indicate parameter interaction/non-identifiability
- use Γ with additional discrete rate categories

Assessing MCMC Performance: Diagnostics Based on Single Runs
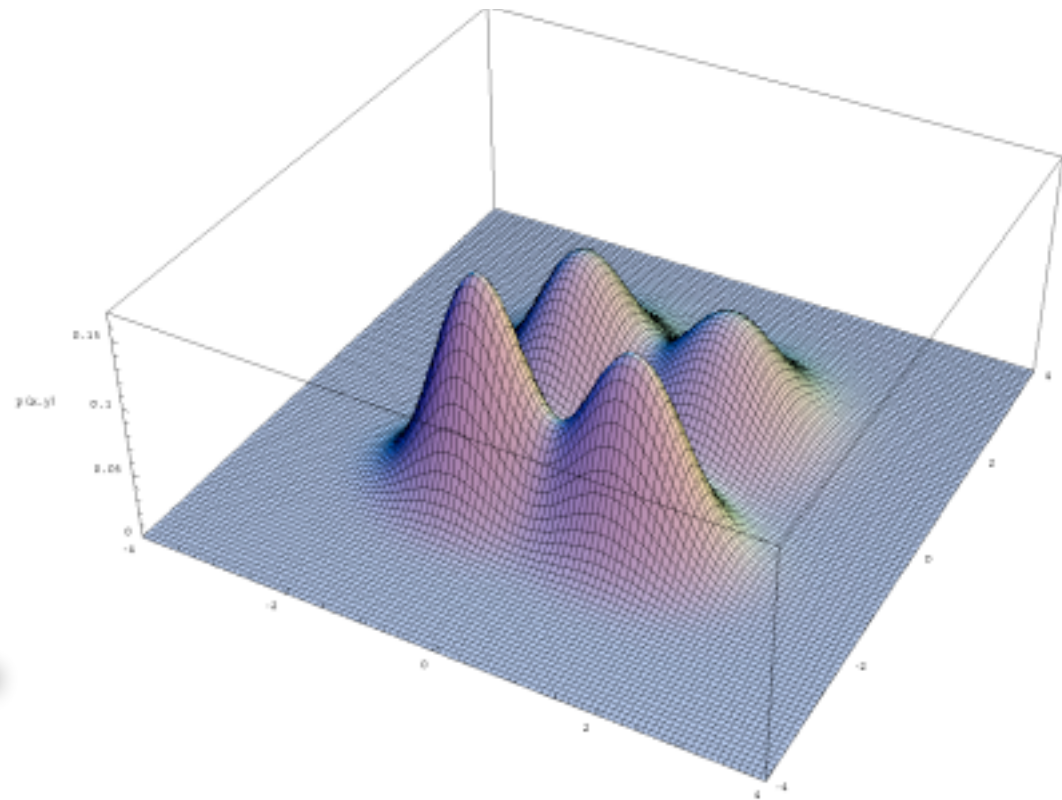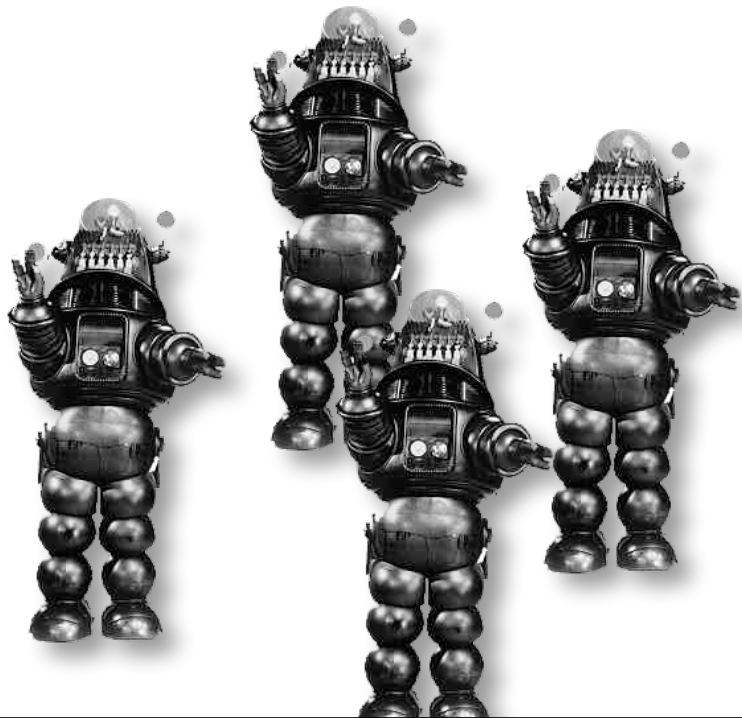
MCMC pathologies

Can identify parameter interaction by plotting joint distribution

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

Robot Squadron!!

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

A slightly more formal description...

To facilitate mixing over the joint posterior probability density, multiple incrementally heated chains may be run

N chains are initiated from random starting point in the joint posterior probability density.

One chain is cold, and $N-1$ are incrementally heated.

Samples are drawn from the cold chain.

The heating distorts the joint posterior probability density, such that chains can more freely traverse regions of the stationary distribution.

Occasionally, a swap is attempted between the cold and one of the randomly chosen heated chains, which ensures that samples are drawn from regions of high posterior probability.

heat of chain $i = 1/(1 + iT)$

chain	0.25	0.20	0.15	0.10
1	1.00	1.00	1.00	1.00
2	0.80	0.83	0.87	0.91
3	0.66	0.71	0.77	0.83
4	0.57	0.62	0.69	0.77

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

Diagnosing MC^3 performance

The primary diagnostic is the acceptance rates for proposed chain swaps:
As a rule of thumb, acceptance rates for proposals should fall in ~20–70% range
- if acceptance rates are too low, decrease the value of the temperature parameter
- if acceptance rates are too high, increase the value of the temperature parameter

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

Example: Tracer plots of relative-rate multipliers from two MrBayes runs

bad mixing

Chain swap information for run 1:

	1	2	3	4
1		0.01	0.00	0.00
2	1666381		0.01	0.00
3	1666964	1664302		0.20
4	1666923	1668351	1667079	

Chain swap information for run 2:

	1	2	3	4
1		0.00	0.00	0.00
2	1664180		0.16	0.00
3	1667247	1669245		0.04
4	1665043	1667632	1666653	

better mixing

Chain swap information for run 1:

	1	2	3	4
1		0.60	0.32	0.17
2	834663		0.65	0.40
3	832631	834125		0.70
4	831509	834020	833052	

Chain swap information for run 2:

	1	2	3	4
1		0.60	0.32	0.17
2	833614		0.65	0.40
3	834623	833715		0.70
4	833536	832594	831918	

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

Diagnosing MC^3 performance

The primary diagnostic is the acceptance rates for proposed chain swaps:
  As a rule of thumb, acceptance rates for proposals should fall in ~20–70% range
  - if acceptance rates are too low, decrease the value of the temperature parameter
  - if acceptance rates are too high, increase the value of the temperature parameter

Other aspects controlling the behavior of the Metropolis coupling can be modified to improve MCMC performance:
  - increase the number of incrementally heated chains (e.g., nchains **parameter**)
  - increase the frequency of attempted chain-swap events (e.g., swapfreq **parameter**)
  - increase the number of swaps attempted per event (e.g., nswaps **parameter**)

Outline

Assessing MCMC Performance: Diagnostics Based on the Prior

Estimating under the prior...

Marginal posterior densities for parameters are updated versions of the corresponding prior probability densities: they are updated by the information in the data via the likelihood function

likelihood function   prior probability

posterior probability

$$\underbrace{\Pr[\tau_i \mid X]}_{} = \frac{\overbrace{\Pr[X \mid \tau_i]}^{} \times \overbrace{\Pr[\tau_i]}^{}}{\underbrace{\sum_{j=1}^{B_{(s)}} \Pr[X \mid \tau_j] \times \Pr[\tau_j]}_{}}$$

marginal likelihood

Assessing MCMC Performance: Diagnostics Based on the Prior

Estimating under the prior...

Marginal posterior densities for parameters are updated versions of the corresponding prior probability densities: they are updated by the information in the data via the likelihood function

We can compare the marginal prior densities to their posterior counterparts to help identify weak parameters
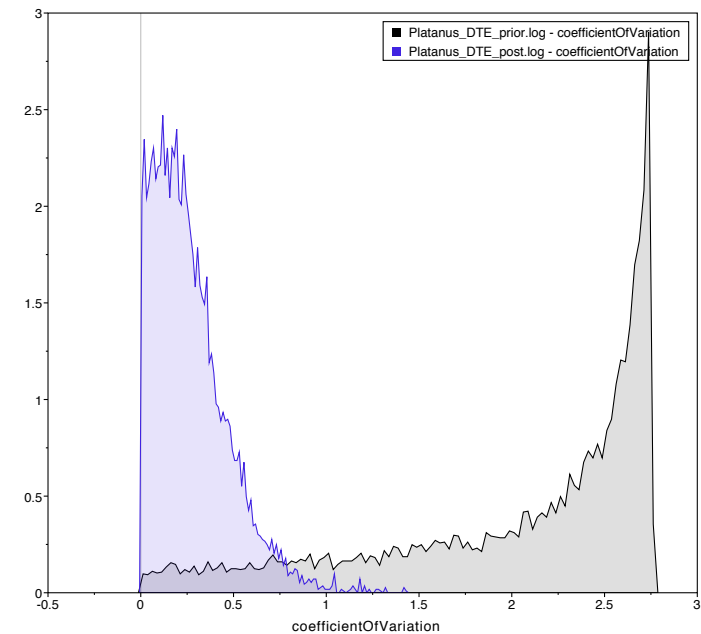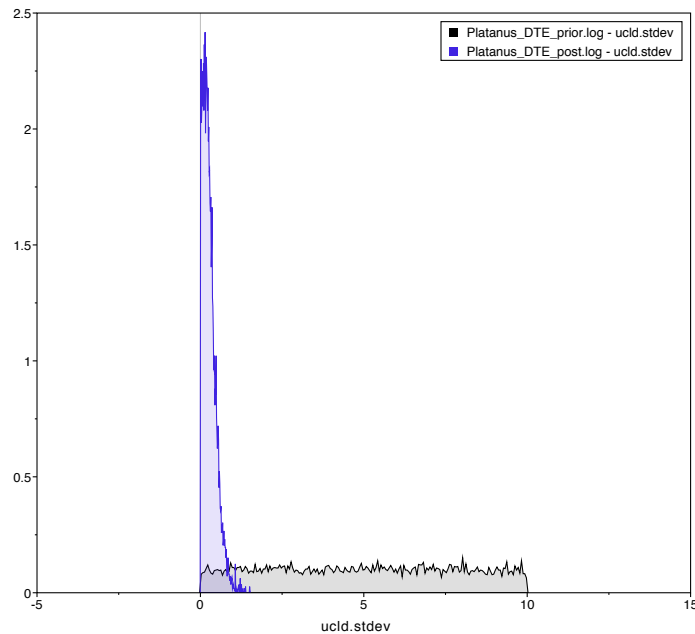
- MCMC can be run to target the joint prior either by estimating with no data or by forcing the likelihood function return 1.

$$R = \min\left[1, \frac{f(X \mid \Theta')}{f(X \mid \Theta)} \cdot \frac{f(\Theta')}{f(\Theta)} \cdot \frac{f(\Theta \mid \Theta')}{f(\Theta' \mid \Theta)}\right]$$

likelihood ratio   prior ratio   proposal ratio

Assessing MCMC Performance: Diagnostics Based on the Prior

Does the marginal prior resemble the marginal posterior?

Strong departure of marginal prior and posterior is always good news

Similarity between the marginal prior and posterior may be good or bad news

Outline

I. A review of where we've been and why

II. Diagnosing MCMC performance

III. Diagnostics based on single chains

IV. Diagnostics based on the prior

➡ V. Diagnostics based on multiple chains

Assessing MCMC Performance:
Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from <u>random</u> parameter values

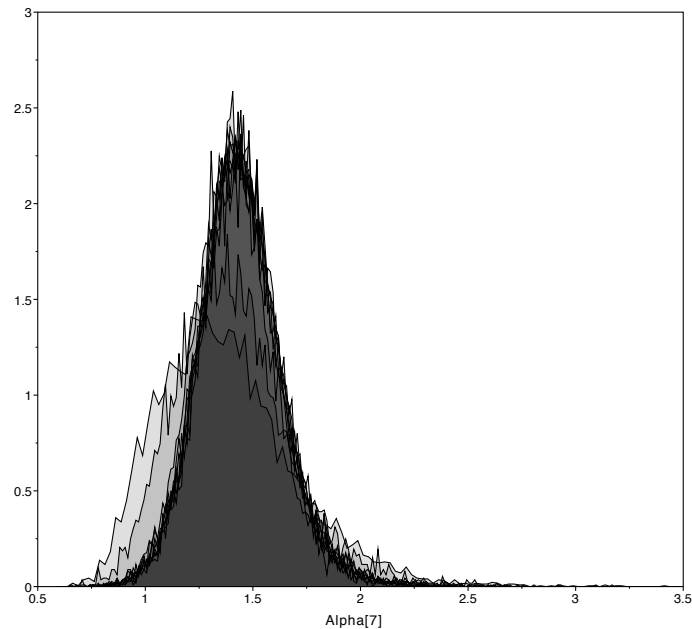Form of the marginal posterior densities for all parameters

- continuous parameters (e.g., substitution rates): Tracer
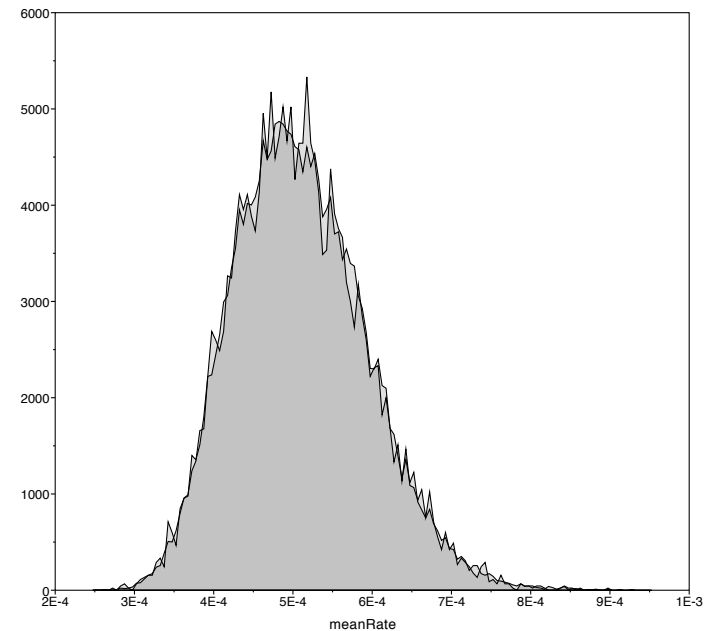
Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Example: Tracer plots of marginal densities from multiple MrBayes runs

bad convergence

better convergence

*Tracer demo

Assessing MCMC Performance:
Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from <u>random</u> parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters:

    - PSRF (Gelman-Rubin) diagnostic: MrBayes

        1. Run $m \geq 2$ chains of length $2n$ from overdispersed starting values.

        2. Discard the first n draws of each chain.

        3. Calculate the within-chain and between-chain variance.

        4. Calculate the estimated variance of the parameter as a weighted sum of the within-chain and between-chain variance.

        5. Calculate the PSRF.

    - Values for all continuous parameters should be 1

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Example: PSRF values for relative-rate multipliers from two MrBayes runs

bad convergence

Parameter	Mean	Variance	95% Cred. Interval Lower	Upper	Median	PSRF *
TL{all}	4.921609	2.998138	2.836000	7.295000	5.056000	9.084
kappa{4,5}	3.095696	0.054125	2.667623	3.587024	3.085271	1.000
alpha{5}	1.006544	0.087721	0.606472	1.738482	0.950093	1.000
pinvar{1}	0.307396	0.009357	0.095913	0.471070	0.316173	1.000
m{1}	0.264226	0.009315	0.146502	0.421870	0.244468	5.507
m{2}	0.040919	0.000227	0.022205	0.065884	0.037425	5.279
m{3}	2.721453	7.157157	0.039001	5.544253	5.030560	69.564
m{4}	2.125810	3.568002	0.199137	4.044249	3.917338	150.012
m{5}	0.188768	0.004373	0.109303	0.295129	0.170624	5.749

better convergence

Parameter	Mean	Variance	95% Cred. Interval Lower	Upper	Median	PSRF *
TL{all}	0.073893	0.000034	0.063000	0.086000	0.074000	1.000
kappa{2,3}	3.236308	0.366904	2.199024	4.587719	3.190195	1.000
m{1}	1.285838	0.028345	0.980634	1.630387	1.278161	1.000
m{2}	1.423906	0.015507	1.182596	1.664627	1.423610	1.000
m{3}	0.589346	0.005341	0.453175	0.736459	0.587617	1.001

Assessing MCMC Performance: Diagnostics Based on Multiple Runs
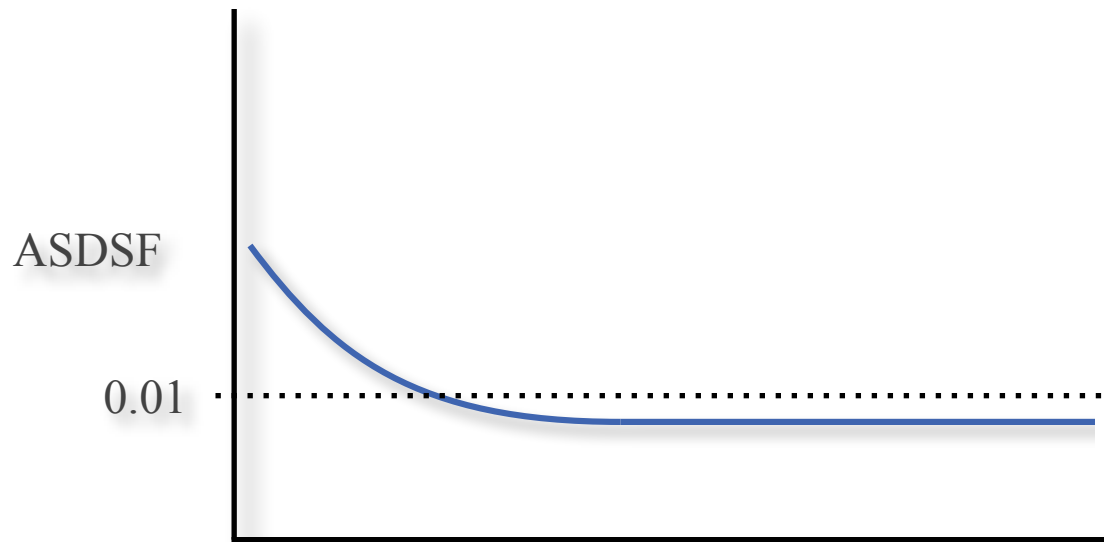
The general idea is to compare estimates from multiple independent chains initiated from <u>random</u> parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters:
  - similarity of marginal densities: Tracer
  - PSRF diagnostic: MrBayes
- discrete parameters:
  - Topology
    - similarity of paired chains (e.g., ASDSF diagnostic in MrBayes)

Assessing MCMC Performance:
Diagnostics Based on Multiple Runs

Example: ASDSF

- stop sampling when **ASDSF** < 0.01

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from <u>random</u> parameter values
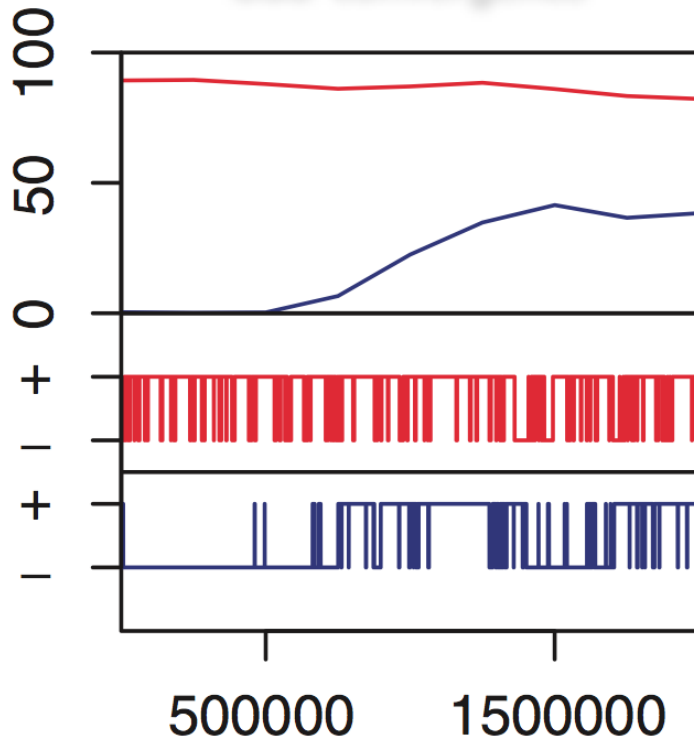
Form of the marginal posterior densities for all parameter

- continuous parameters (e.g., substitution rates): Tracer
- discrete parameters:
    - Topology
        - similarity of paired chains (e.g., ASDSF diagnostic in MrBayes)
        - distances among sampled topologies: TreeSetViz
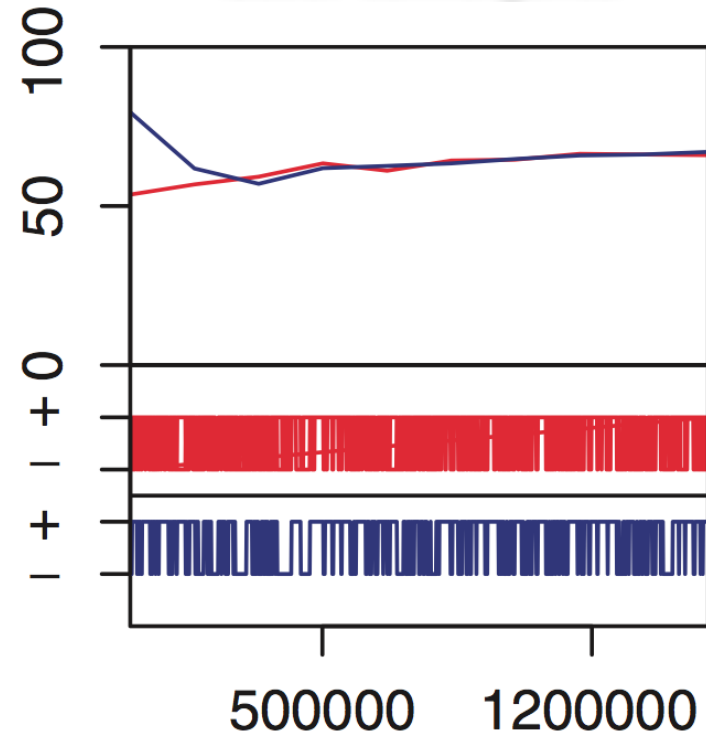   - split frequencies & presence/absence: AWTY

Assessing MCMC Performance:
Diagnostics Based on Multiple Runs

Example: split frequencies & presence/absence in AWTY

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from <u>random</u> parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters (e.g., substitution rates): Tracer
- discrete parameters:
  - Topology
    - similarity of paired chains (e.g., ASDSF diagnostic in MrBayes)
      - distances among sampled topologies: TreeSetViz
    - split frequencies & presence/absence: AWTY
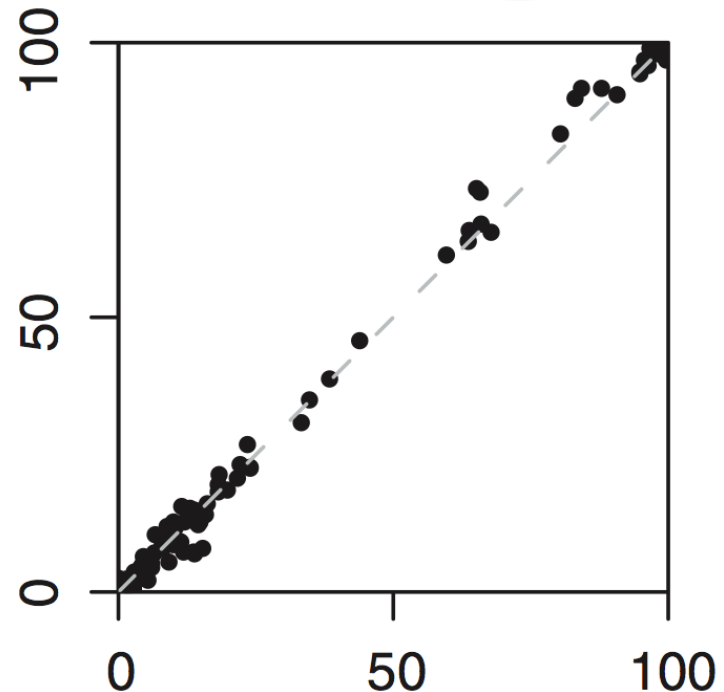   - nodal support--AWTY/MrBayes

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Example: 'comparetrees' plot of trees sampled by two MrBayes runs

bad convergence
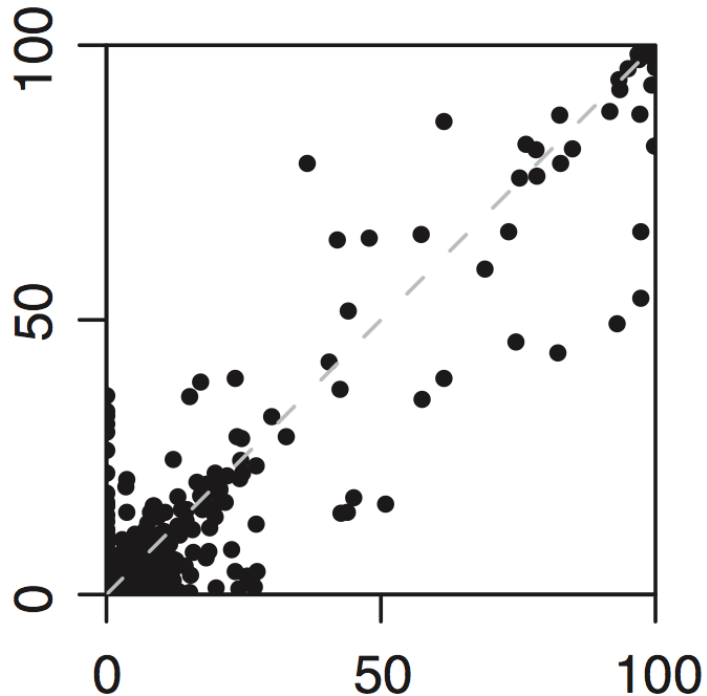
better convergence

Nylander et al. (2008)

Summary: Some General Strategies for Assessing MCMC Performance

You can never be absolutely certain that the MCMC is reliable, you can only identify when something has gone wrong. Gelman

1. When do you need to assess MCMC performance?

   ALWAYS

2. When should you assess the performance of individual runs?

   ALWAYS

3. Which diagnostics should you use to assess individual runs?

   ALL that are relevant for the models/parameters you are estimating under

4. When is a single run sufficient to assess MCMC performance?

   NEVER

5. When should you estimate under the prior?

   WHENEVER POSSIBLE (and be wary of programs where it is not possible)

Summary: Some General Strategies for Assessing MCMC Performance

You can never be absolutely certain that the MCMC is reliable, you can only identify when something has gone wrong. Gelman

6. When should you use Metropolis-Coupling?

Whenever you cannot be certain that standard MCMC is adequate

i.e., ALWAYS (and be wary of programs where it is not possible)

7. When should you perform multiple independent MCMC runs?

ALWAYS (and be wary of pseudo-independence)

8. Which diagnostics should you use to assess individual runs?

ALL that are relevant for the models/parameters you are estimating under

9. How many independent MCMC runs are sufficient?

AS MANY AS POSSIBLE (i.e., as many as you think your data/problem deserve)

10. How long should you run each MCMC analysis?

AS LONG AS POSSIBLE (i.e., as long as you think your data/problem deserve)

Assessing MCMC Performance: Software Tools

Semi-automated analysis using diverse diagnostic tools

Generates an automated report (sup. mat.)

Flags suspicious parameters

R package

BONSAI

Bayesian **O**utput **N**eeds **S**emi-**A**utomated **I**nspection

Mike May