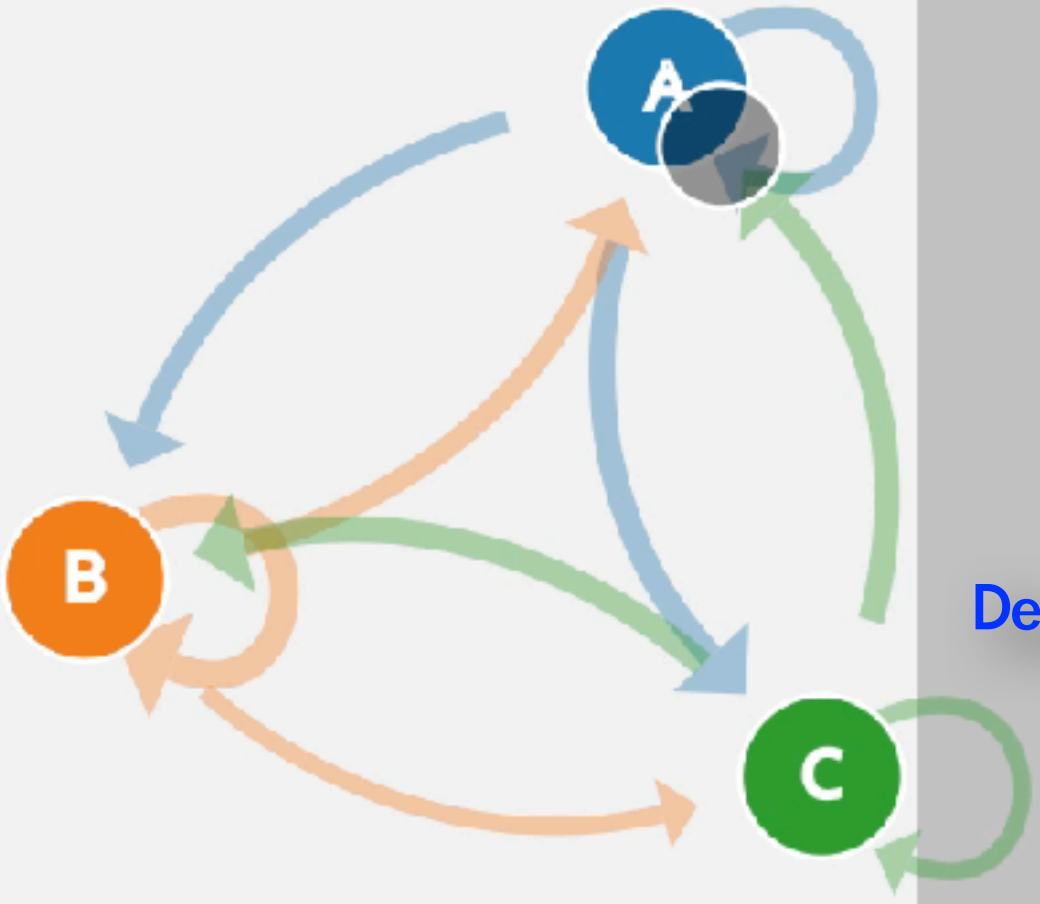
Diagnosing MCMC Performance



Brian R. Moore

Department of Evolution & Ecology
UC, Davis
Bodega Workshop, 2017

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Approximating the Joint Posterior Probability Density using MCMC

MCMC in theory and practice

MCMC in theory...

an appropriately constructed and adequately run 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 particular case

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

A.

NEVER!

Assessing MCMC Performance: Three Main Issues

1. Convergence

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

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?

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?

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Assessing MCMC Performance: Diagnostics Based on Single Runs

1. Convergence diagnostics

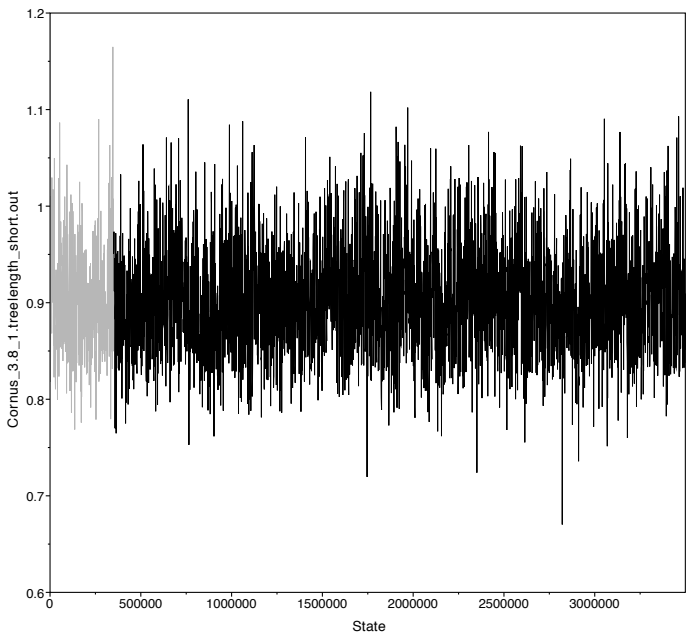
(i) Time-series plots of parameter estimates

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

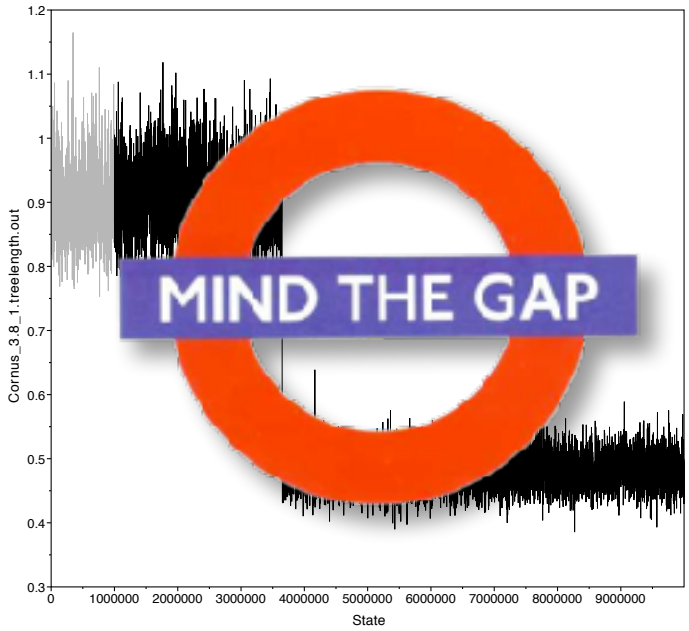
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

(ii) Geweke diagnostic: coda, BOA

(iii) Heidelberg-Welch diagnostic: coda, BOA

(...) Many others

Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

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

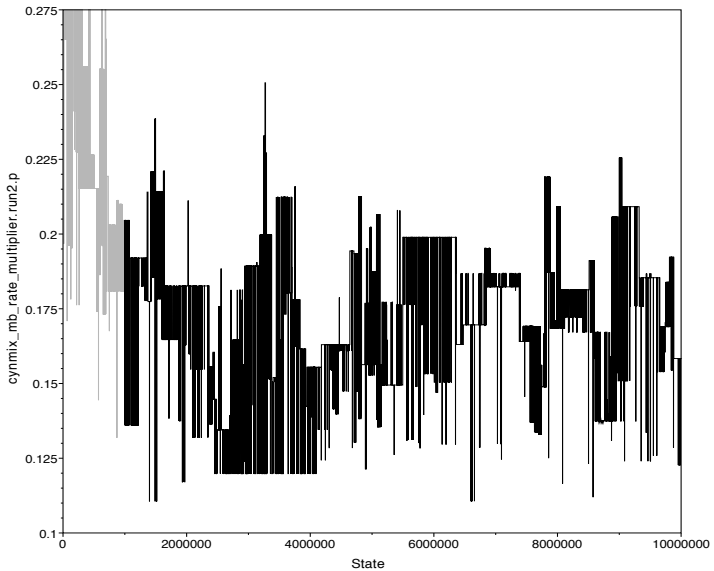
- continuous parameters (e.g., substitution rates): Tracer

warm and fuzzy caterpillars

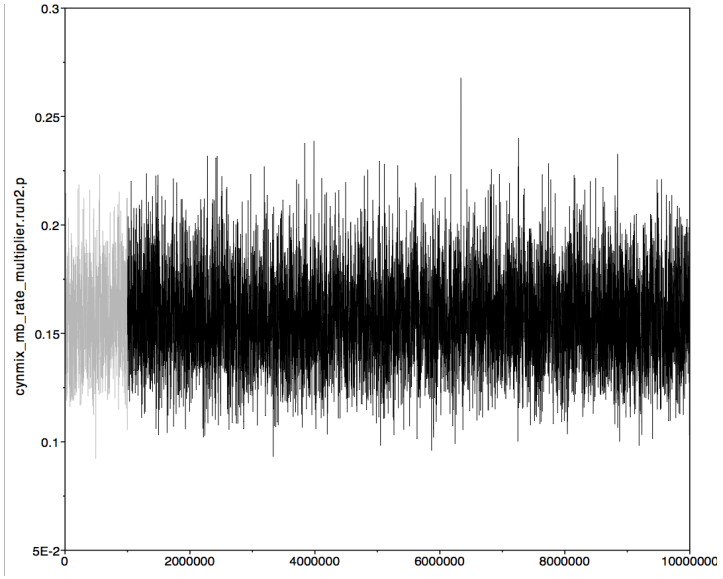
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

bad mixing



better mixing



Assessing MCMC Performance: Diagnostics Based on Single Runs

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

(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

```
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
```

better mixing

```
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

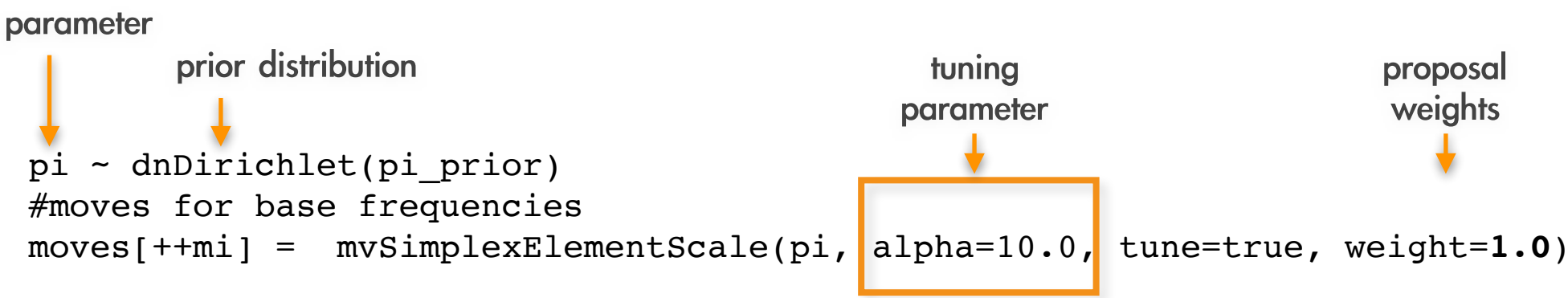
(ii) Acceptance rates of parameter updates

- 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

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters: MrBayes, BEAST, etc.

rates should ideally fall in the ~20–70% range

(iii) Form of the marginal posterior probability densities

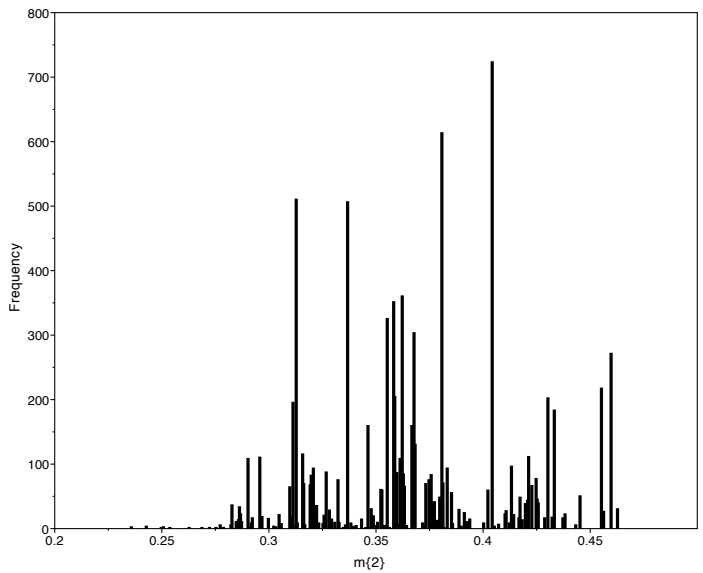
- continuous parameters (e.g., substitution rates): Tracer

beware of porcupine roadkill

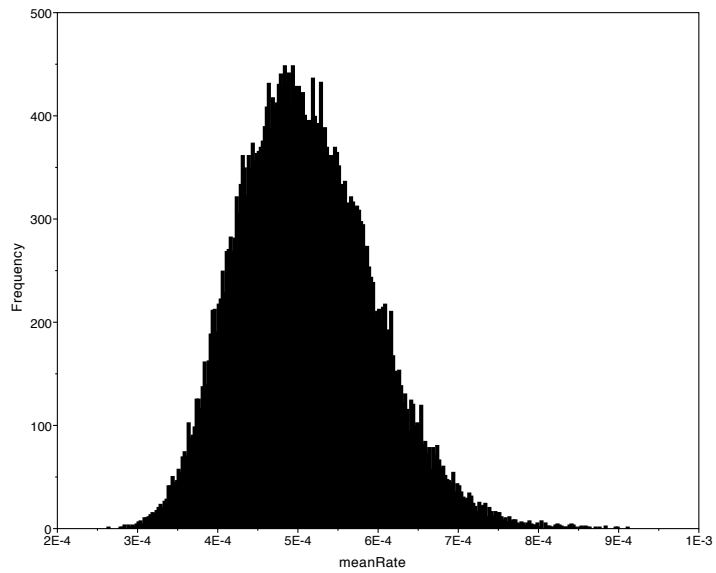
Assessing MCMC Performance: Diagnostics Based on Single Runs

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
```

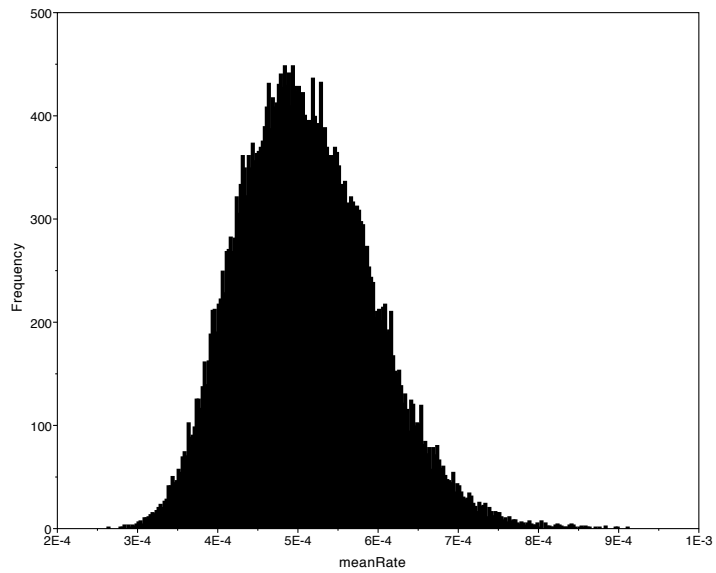
```
Acceptance rates for the moves in the "cold" chain of run 1:
     With prob.  Chain accepted changes to
       33.30 %   param. 1 (revmat) with Dirichlet proposal
          .
          .
          .
       19.13 %   param. 34 (rate multiplier) Dirichlet proposal
       17.40 %   param. 35 (topology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

Assessing MCMC Performance: Diagnostics Based on Single Runs

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

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

- continuous parameters (e.g., substitution rates): Tracer

warm and fuzzy caterpillars

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters: MrBayes, BEAST, etc.

rates should ideally fall in the ~20–70% range

(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

2. Mixing diagnostics

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

- continuous parameters (e.g., substitution rates): Tracer

warm and fuzzy caterpillars

(ii) Acceptance rates of parameter updates

- continuous & discrete parameters: MrBayes, BEAST, etc.

rates should ideally fall in the ~20–70% range

(iii) Form of the marginal posterior probability densities

- continuous parameters (e.g., substitution rates): Tracer

beware of porcupine roadkill

} qualitative diagnostics

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

(ii) Acceptance rates of parameter updates

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

(iii) Form of the marginal posterior probability densities

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

} qualitative diagnostics

(iv) Autocorrelation time (ACT) of parameter samples

(iv) Effective sample size (ACT) of parameter samples

} quantitative diagnostics

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).

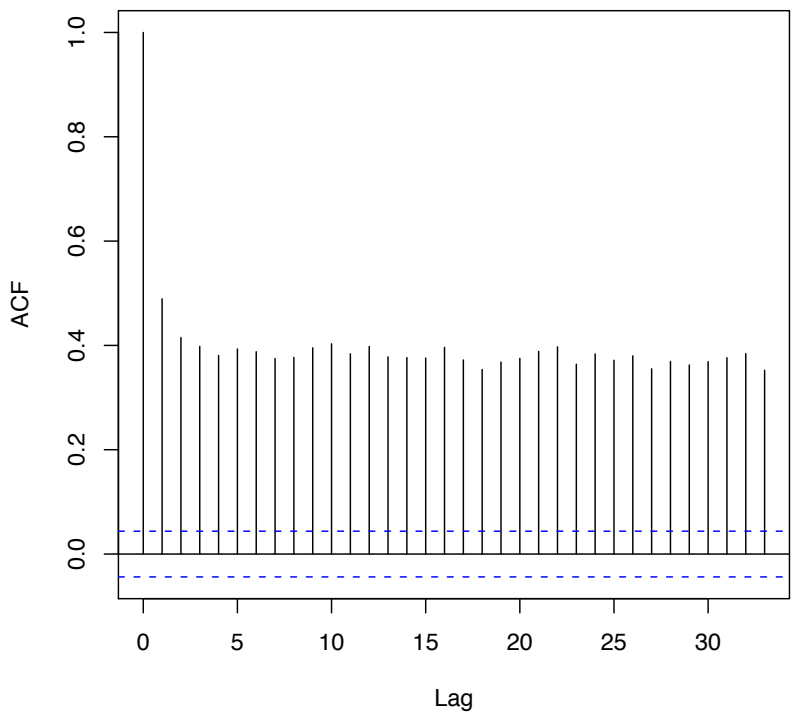
If autocorrelation is still relatively high for higher values of k, this indicates high degree of correlation between our draws and slow mixing.

Assessing MCMC Performance: Diagnostics Based on Single Runs

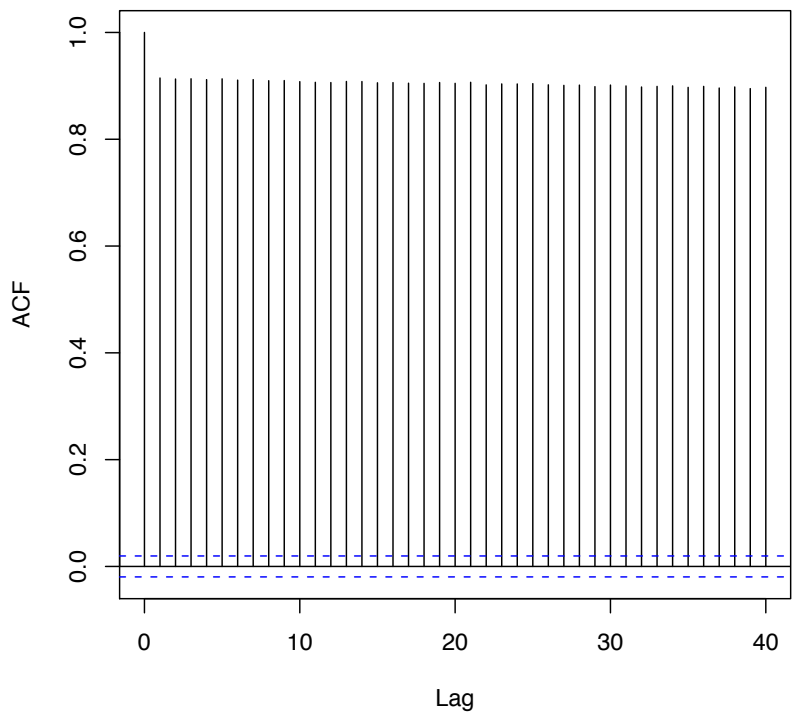
2. Mixing diagnostics

(iv) Autocorrelation time (ACT) of parameter samples

efficient mixing



slow mixing



Assessing MCMC Performance: Diagnostics Based on Single Runs

2. Mixing diagnostics

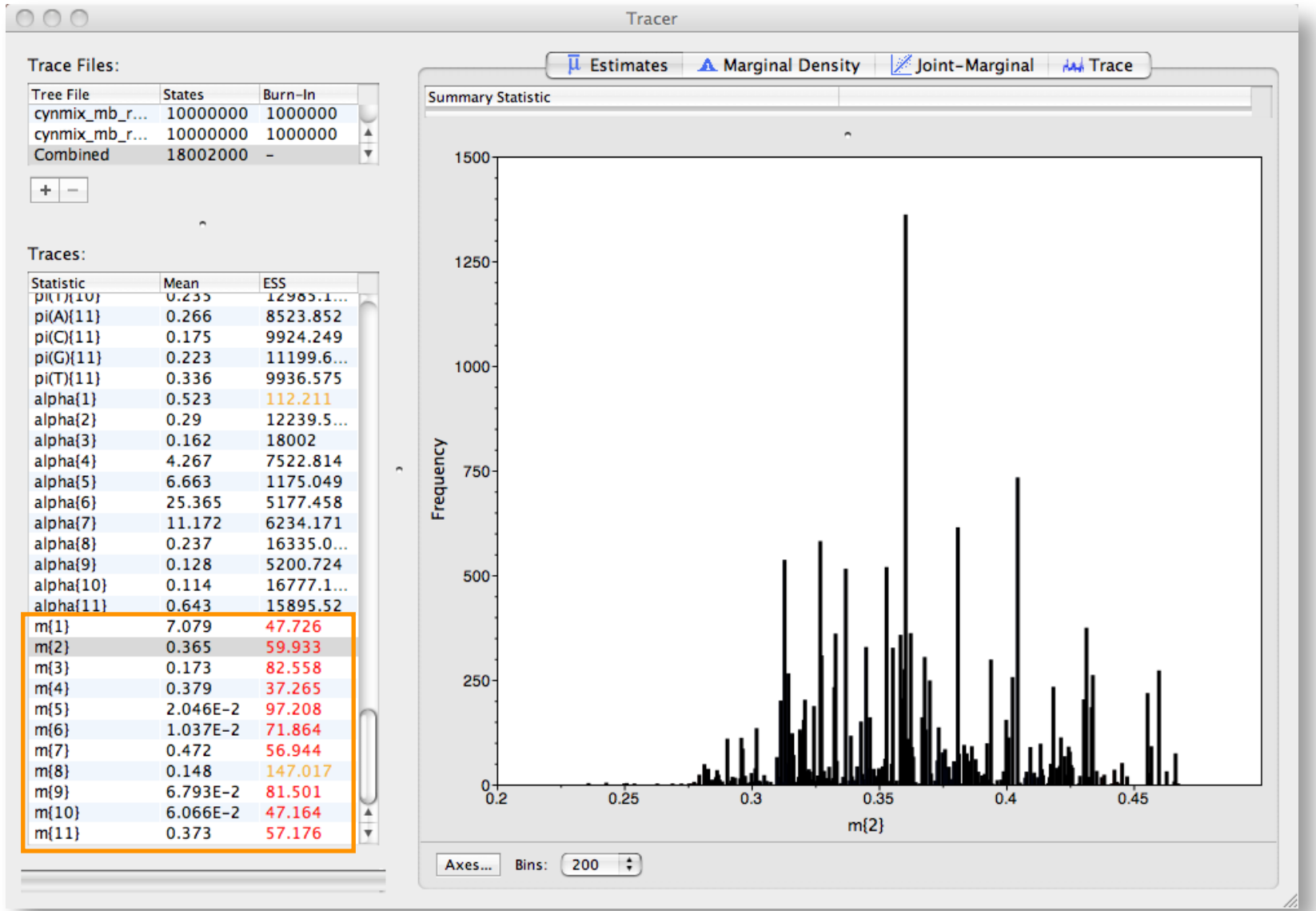
(iv) Effective Sample Size (ESS) diagnostic

- 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 RevBayes runs

poor mixing



Assessing MCMC Performance: Diagnostics Based on Single Runs

3. Sample-size diagnostics

(i) Form of the marginal posterior probability densities

- continuous parameters (e.g., substitution rates): Tracer

  brother of porcupine roadkill

  ensure SAE compliance!

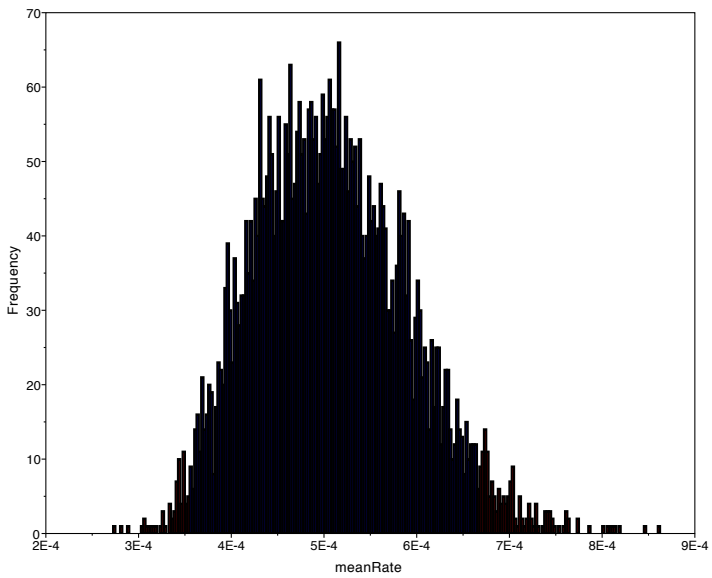
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

poor sampling

1M cycles

better sampling



5M cycles

- inadequate chain length/poor mixing

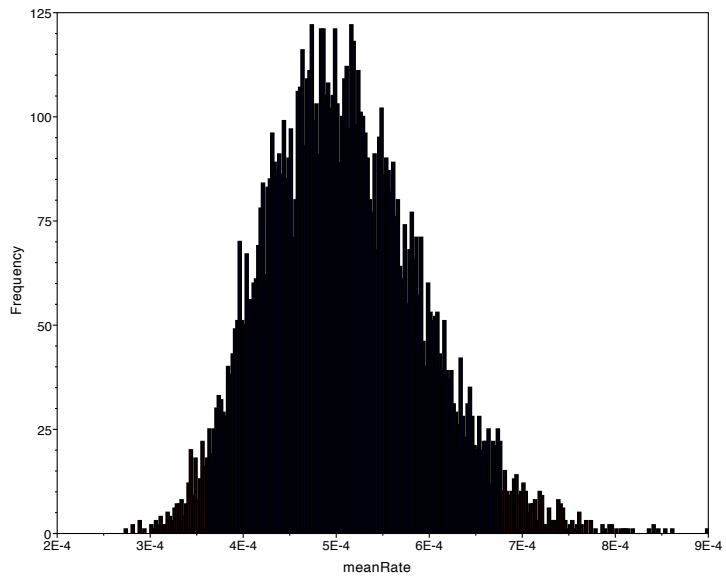
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

poor sampling

1M cycles

better sampling



10M cycles

- inadequate chain length/poor mixing

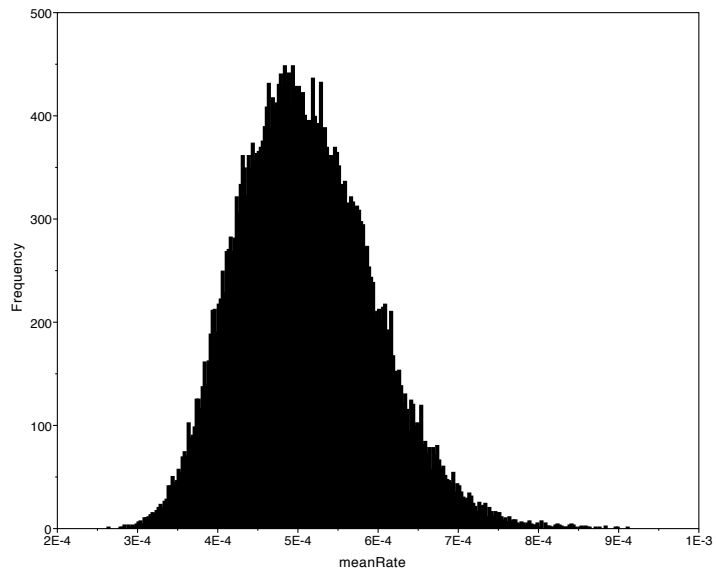
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

poor sampling

1M cycles

better sampling



40M cycles

- ESS can be increased by reducing the sampling frequency/increasing burin in
- All continuous parameters should be SAE
- KDE SAE does not count (use histogram render)

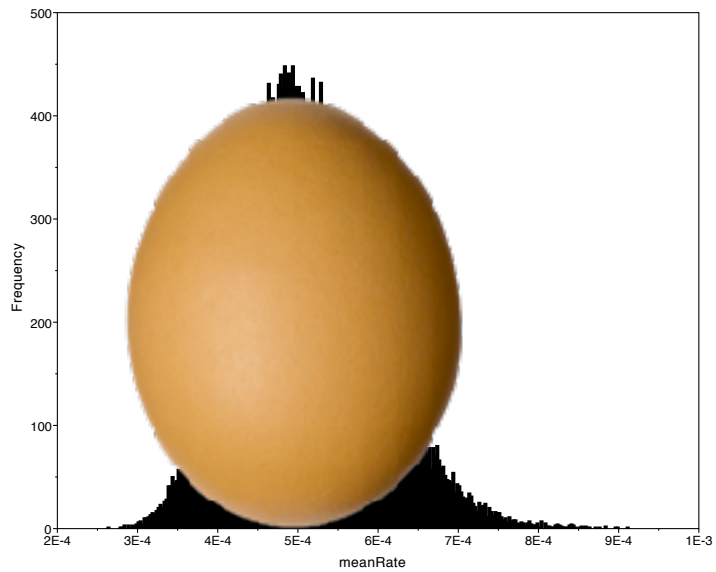
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

poor sampling

1M cycles

better sampling



40M cycles

- ESS can be increased by reducing the sampling frequency/increasing burin in
- All continuous parameters should be SAE
- KDE SAE does not count (use histogram render)

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Outline

I. Diagnosing MCMC performance

motivation and overview of the basics

II. MCMC Diagnostics

diagnostics based on single chains

diagnostics based on the prior

diagnostics based on multiple, replicate chains

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from random parameter values

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from random parameter values

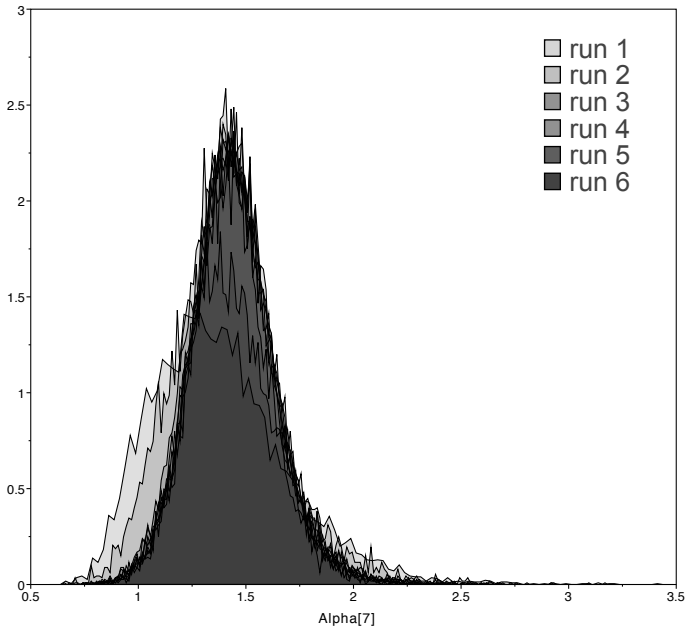
Form of the marginal posterior densities for all parameters

- continuous parameters (e.g., substitution rates): Tracer

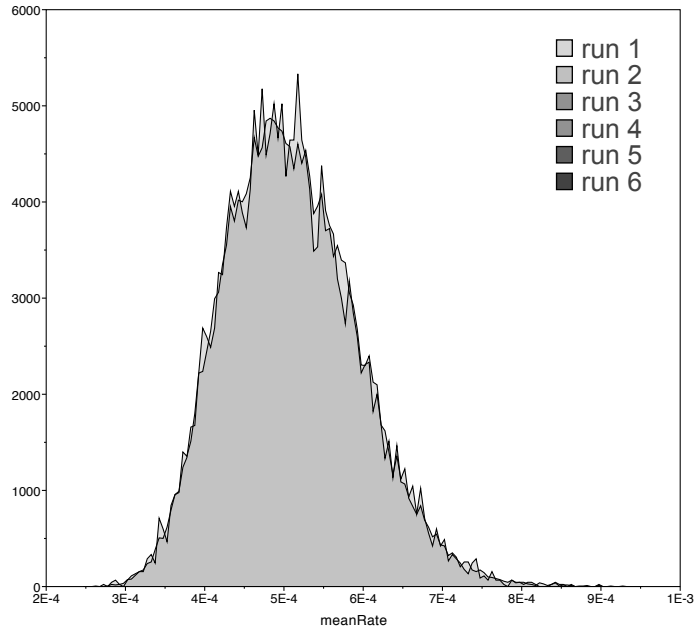
Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Example: Tracer plots of marginal densities from multiple RevBayes runs

bad convergence



better convergence



Parameter estimates from replicate independent MCMC analyses should be effectively identical.

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

The general idea is to compare estimates from multiple independent chains initiated from random parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters:
  - PSRF (Gelman-Rubin) diagnostic: RevBayes

1. Run $m \geq 2$ chains of length $2c$ 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	95% Cred. Interval 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	95% Cred. Interval 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 random parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters:
  - similarity of marginal densities: Tracer
  - PSRF diagnostic: RevBayes
- discrete parameters:
  - Topology
    - similarity of trees sampled by paired, independent chains (e.g., ASDSF)

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

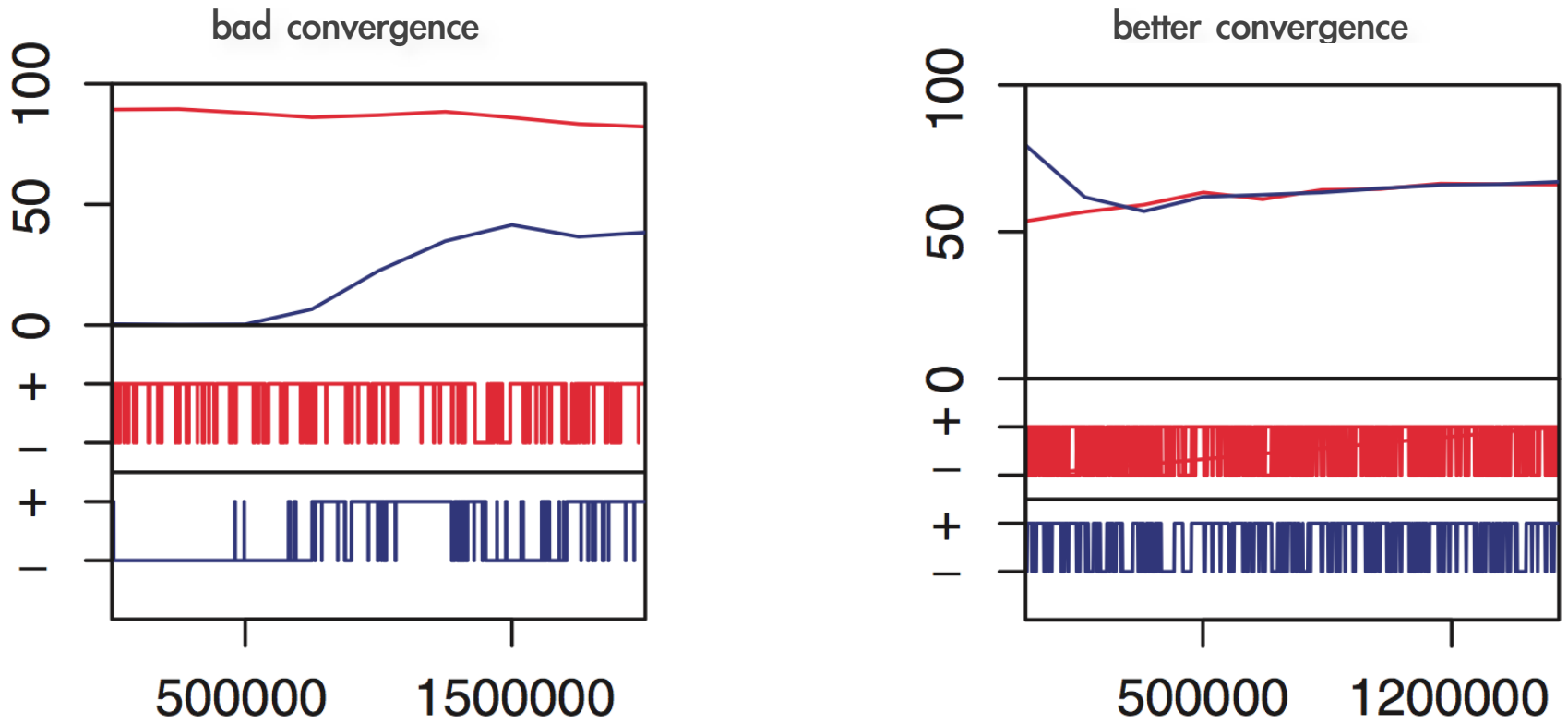
The general idea is to compare estimates from multiple independent chains initiated from random parameter values

Form of the marginal posterior densities for all parameter

- continuous parameters:
  - similarity of marginal densities: Tracer
  - PSRF diagnostic: RevBayes
- discrete parameters:
  - Topology
    - similarity of trees sampled by paired, independent chains (e.g., ASDSF)
    - split frequencies & presence/absence: AWTY

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Example: split frequencies & presence/absence in AWTY



Track the frequency of a single node in trees sampled by two independent chains

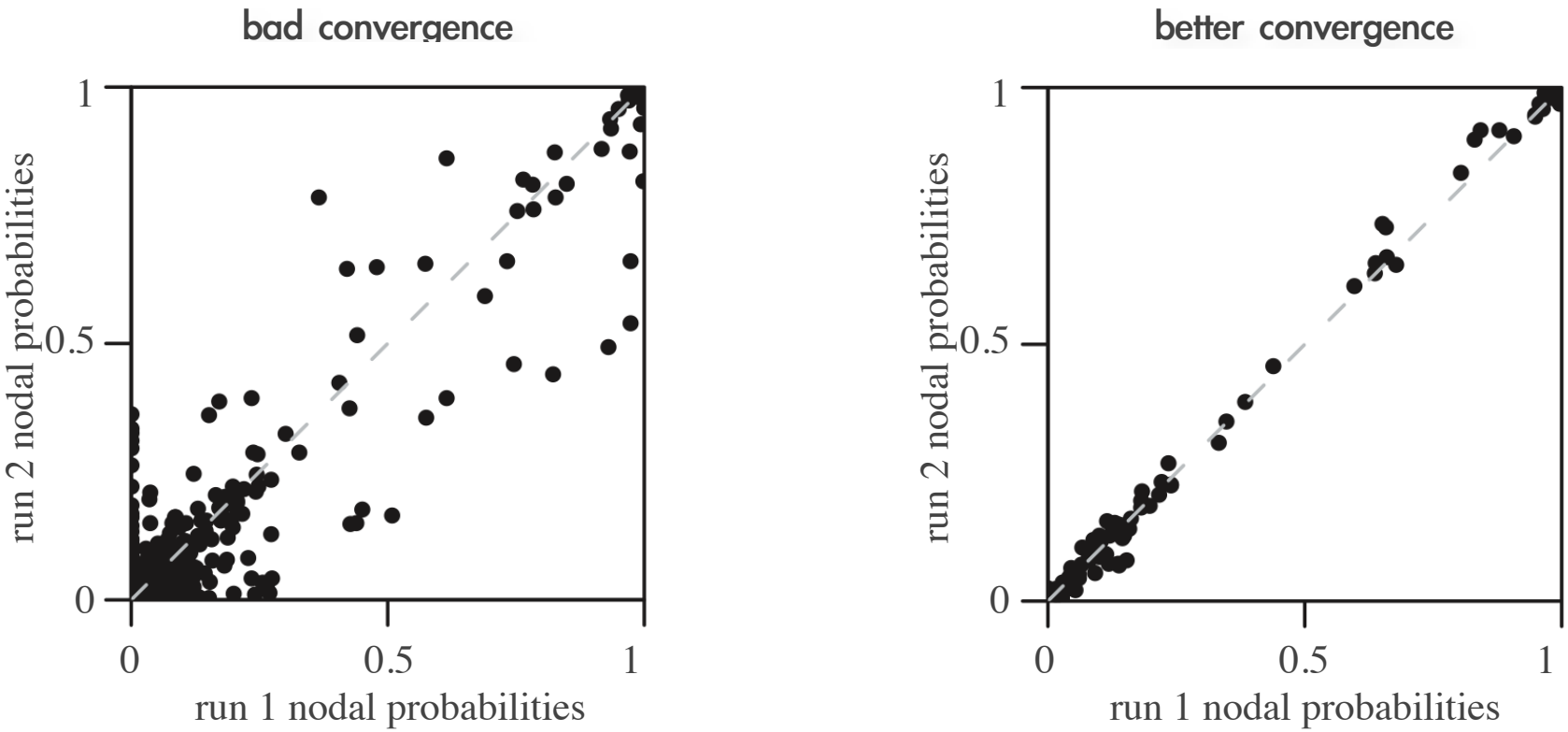
Assessing MCMC Performance: Diagnostics Based on Multiple Runs

Form of the marginal posterior densities for all parameter

- continuous parameters:
  - similarity of marginal densities: Tracer
  - PSRF diagnostic: RevBayes
- discrete parameters:
  - Topology
    - similarity of paired chains (e.g., ASDSF diagnostic in RevBayes)
    - split frequencies & presence/absence: AWTY
    - nodal support (compare-tree plots)

Assessing MCMC Performance: Diagnostics Based on Multiple Runs

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



Compare estimates of node probabilities estimated by two independent chains

Nylander et al. (2008)

Assessing MCMC Performance: Software Tools

Software	Manual/visual	Split frequencies	PSRF	ESS	Geweke test	H-W test	S-Stationarity	M-Stationarity
AWTY	x	x	-	-	-	-	-	-
BOA	x	-	x	x	x	x	-	-
CODA	x	-	x	x	x	x	-	-
MrBayes	-	x	x	x	-	-	-	-
PhyloBayes	-	x	-	-	-	-	-	-
RevBayes	x	x	x	x	x	x	x	x
Tracer	x	-	-	x	-	-	-	-

Software tools are scattered across many programs

Hohna et al. (in prep.)

Assessing MCMC Performance: Software Tools

Software	Manual/visual	Split frequencies	PSRF	ESS	Geweke test	H-W test	S-Stationarity	M-Stationarity
AWTY	x	x	-	-	-	-	-	-
BOA	x	-	x	x	x	x	-	-
CODA	x	-	x	x	x	x	-	-
MrBayes	-	x	x	x	-	-	-	-
PhyloBayes	-	x	-	-	-	-	-	-
RevBayes	x	x	x	x	x	x	x	x
Tracer	x	-	-	x	-	-	-	-

Software tools are scattered across many programs

Diagnosis is largely manual/by visual inspection

Hohna et al. (in prep.)

Assessing MCMC Performance: Software Tools

Software	Manual/visual	Split frequencies	PSRF	ESS	Geweke test	H-W test	S-Stationarity	M-Stationarity
AWTY	x	x	-	-	-	-	-	-
BOA	x	-	x	x	x	x	-	-
CODA	x	-	x	x	x	x	-	-
MrBayes	-	x	x	x	-	-	-	-
PhyloBayes	-	x	-	-	-	-	-	-
RevBayes	x	x	x	x	x	x	x	x
Tracer	x	-	-	x	-	-	-	-

Software tools are scattered across many programs

Diagnosis is largely manual/by visual inspection

Use of the methods is time consuming

Hohna et al. (in prep.)

Assessing MCMC Performance: Software Tools

Software	Manual/visual	Split frequencies	PSRF	ESS	Geweke test	H-W test	S-Stationarity	M-Stationarity
AWTY	x	x	-	-	-	-	-	-
BOA	x	-	x	x	x	x	-	-
CODA	x	-	x	x	x	x	-	-
MrBayes	-	x	x	x	-	-	-	-
PhyloBayes	-	x	-	-	-	-	-	-
RevBayes	x	x	x	x	x	x	x	x
Tracer	x	-	-	x	-	-	-	-

Software tools are scattered across many programs

Diagnosis is largely manual/by visual inspection

Use of the methods is time consuming

Use of the methods is vague and virtual

Hohna et al. (in prep.)

Assessing MCMC Performance: Software Tools

BONSAI

Semi-automated analysis using diverse diagnostic tools

Generates an automated report (sup. mat.)

Flags suspicious parameters

R package

Bayesian Output Needs Semi-Automated Inspection

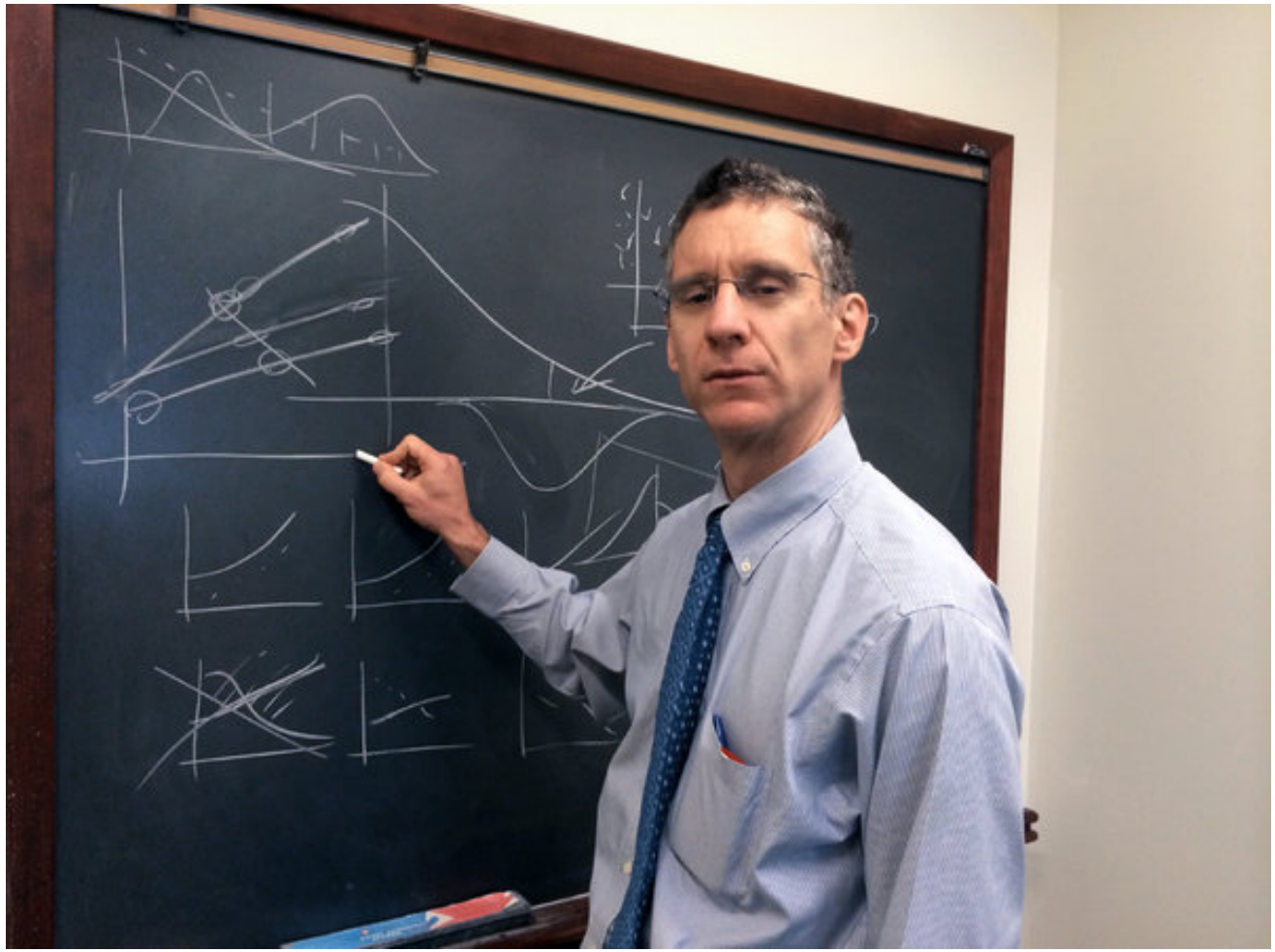
Mike May

https://bitbucket.org/mrmay/bonsai/overview

May, Hohna & Moore (in prep.)

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. Andrew Gelman (hero)



Summary: Some General Strategies for Assessing MCMC Performance:

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:

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 multiple 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)