Diagnosing MCMC Performance

Brian R. Moore
Department of Evolution & Ecology
UC, Davis
Bodega Workshop, 2017
Outline

1. 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
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
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
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 \( \sim 20\%-70\% \) range
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

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

bad mixing
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

pi ~ dnDirichlet(pi_prior)

#moves for base frequencies
moves[++mi] = mvSimplexElementScale(pi, alpha=10.0, tune=true, weight=1.0)
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

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
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
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
6.59 % param. 35 (topology and branch lengths) TBR
14.06 % param. 35 (topology and branch lengths) LOCAL

0.04 % param. 34 (rate multiplier) Dirichlet proposal

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
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 density 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
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 $\sim20$–$70\%$ range

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

(iv) Autocorrelation time (ACT) of parameter samples

(iv) Effective sample size (ACT) of parameter samples
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 $\rho_k$ 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
- better sampling

- inadequate chain length/poor mixing
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

- inadequate chain length/poor mixing
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

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

1M cycles

40M cycles
Assessing MCMC Performance: Diagnostics Based on Single Runs

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

- 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)
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
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

Parameter estimates from replicate independent MCMC analyses should be effectively identical.
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
### Asssessing MCMC Performance:
**Diagnostics Based on Multiple Runs**

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

#### bad convergence

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
<th>Lower</th>
<th>Upper</th>
<th>Median</th>
<th>PSRF *</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL{all}</td>
<td>4.921609</td>
<td>2.998138</td>
<td>2.836000</td>
<td>7.295000</td>
<td>5.056000</td>
<td>9.084</td>
</tr>
<tr>
<td>kappa{4,5}</td>
<td>3.095696</td>
<td>0.054125</td>
<td>2.667623</td>
<td>3.587024</td>
<td>3.085271</td>
<td>1.000</td>
</tr>
<tr>
<td>alpha{5}</td>
<td>1.006544</td>
<td>0.087721</td>
<td>0.606472</td>
<td>1.738482</td>
<td>0.950093</td>
<td>1.000</td>
</tr>
<tr>
<td>pinvar{1}</td>
<td>0.307396</td>
<td>0.09357</td>
<td>0.095913</td>
<td>0.471070</td>
<td>0.316173</td>
<td>1.000</td>
</tr>
<tr>
<td>m{1}</td>
<td>0.264226</td>
<td>0.009315</td>
<td>0.146502</td>
<td>0.421870</td>
<td>0.244468</td>
<td>5.507</td>
</tr>
<tr>
<td>m{2}</td>
<td>0.040919</td>
<td>0.000227</td>
<td>0.022205</td>
<td>0.065884</td>
<td>0.037425</td>
<td>5.279</td>
</tr>
<tr>
<td>m{3}</td>
<td>2.721453</td>
<td>7.157157</td>
<td>0.039001</td>
<td>5.544253</td>
<td>5.030560</td>
<td>69.564</td>
</tr>
<tr>
<td>m{4}</td>
<td>2.125810</td>
<td>3.568002</td>
<td>0.199137</td>
<td>4.044249</td>
<td>3.917338</td>
<td>150.012</td>
</tr>
<tr>
<td>m{5}</td>
<td>0.188768</td>
<td>0.004373</td>
<td>0.109303</td>
<td>0.295129</td>
<td>0.170624</td>
<td>5.749</td>
</tr>
</tbody>
</table>

#### better convergence

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
<th>Lower</th>
<th>Upper</th>
<th>Median</th>
<th>PSRF *</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL{all}</td>
<td>0.073893</td>
<td>0.000034</td>
<td>0.063000</td>
<td>0.086000</td>
<td>0.074000</td>
<td>1.000</td>
</tr>
<tr>
<td>kappa{2,3}</td>
<td>3.236308</td>
<td>0.366904</td>
<td>2.667623</td>
<td>3.587024</td>
<td>3.085271</td>
<td>1.000</td>
</tr>
<tr>
<td>m{1}</td>
<td>1.285838</td>
<td>0.028125</td>
<td>0.980634</td>
<td>1.630387</td>
<td>1.278161</td>
<td>1.000</td>
</tr>
<tr>
<td>m{2}</td>
<td>1.423906</td>
<td>0.015507</td>
<td>1.182596</td>
<td>1.664627</td>
<td>1.423610</td>
<td>1.000</td>
</tr>
<tr>
<td>m{3}</td>
<td>0.589346</td>
<td>0.005341</td>
<td>0.453175</td>
<td>0.736459</td>
<td>0.587617</td>
<td>1.001</td>
</tr>
</tbody>
</table>
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 tools are scattered across many programs

<table>
<thead>
<tr>
<th>Software</th>
<th>Manual/visual</th>
<th>Split frequencies</th>
<th>PSRF</th>
<th>ESS</th>
<th>Geweke test</th>
<th>H-W test</th>
<th>S-Stationarity</th>
<th>M-Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWTY</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BOA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CODA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MrBayes</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PhyloBayes</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RevBayes</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Tracer</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Hohna et al. (in prep.)
Assessing MCMC Performance: Software Tools

Software tools are scattered across many programs
Diagnosis is largely manual/by visual inspection

<table>
<thead>
<tr>
<th>Software</th>
<th>Manual/visual</th>
<th>Split frequencies</th>
<th>PSRF</th>
<th>ESS</th>
<th>Geweke test</th>
<th>H-W test</th>
<th>S-Stationarity</th>
<th>M-Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWTY</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BOA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CODA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MrBayes</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PhyloBayes</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RevBayes</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Tracer</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Hohna et al. (in prep.)
Software tools are scattered across many programs
Diagnosis is largely manual/by visual inspection
Use of the methods is time consuming
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
Assessing MCMC Performance: Software Tools

BONSAI

Bayesian Output Needs Semi-Automated Inspection

Semi-automated analysis using diverse diagnostic tools
Generates an automated report (sup. mat.)
Flags suspicious parameters
R package

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

May, Hohna & Moore (in prep.)
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)