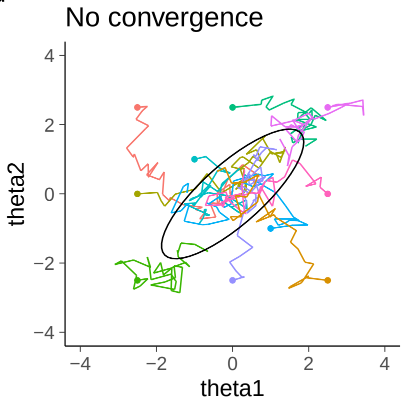
Generic MCMC convergence diagnostics

- Run several chains
- Split-$\widehat{R}$ (Rhat) diagnostic comparing means and variances of chains
- Effective sample size estimate N_{eff} for dependent draws

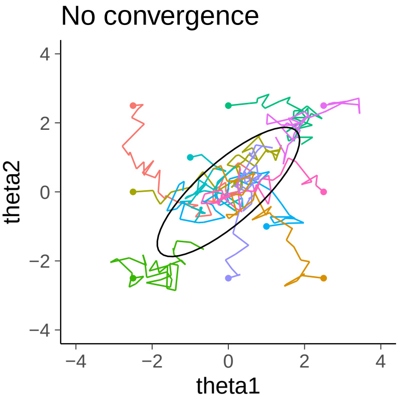
Several chains

- Use of several chains make convergence diagnostics easier
- Start chains from different starting points – preferably overdispersed



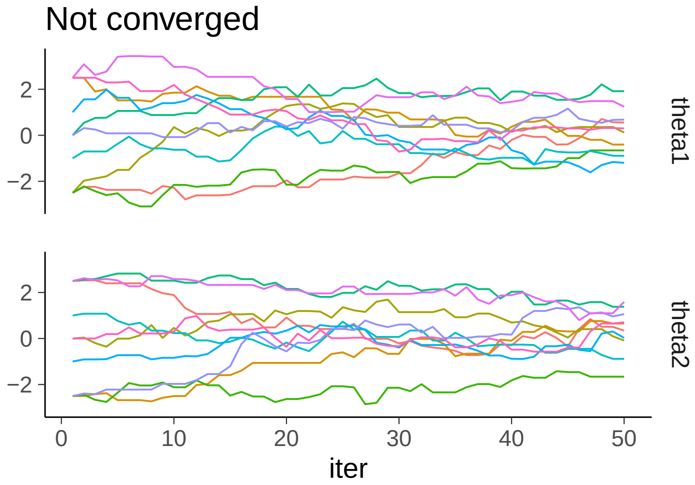
Several chains

- Use of several chains make convergence diagnostics easier
- Start chains from different starting points – preferably overdispersed

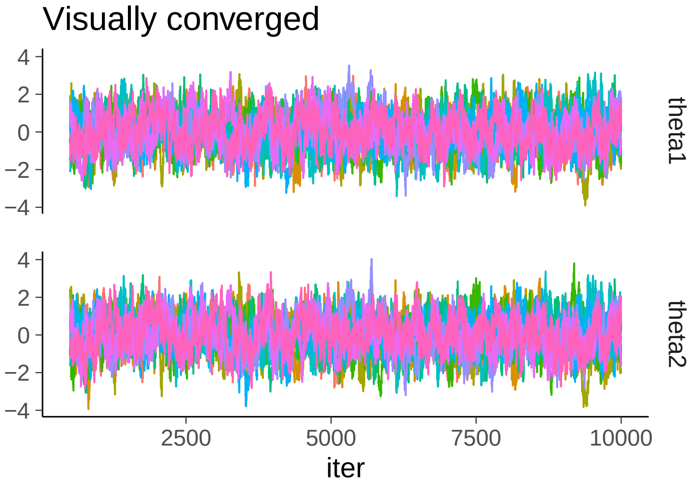


- Remove draws from the beginning of the chains and run chains long enough so that it is not possible to distinguish where each chain started and the chains are well mixed

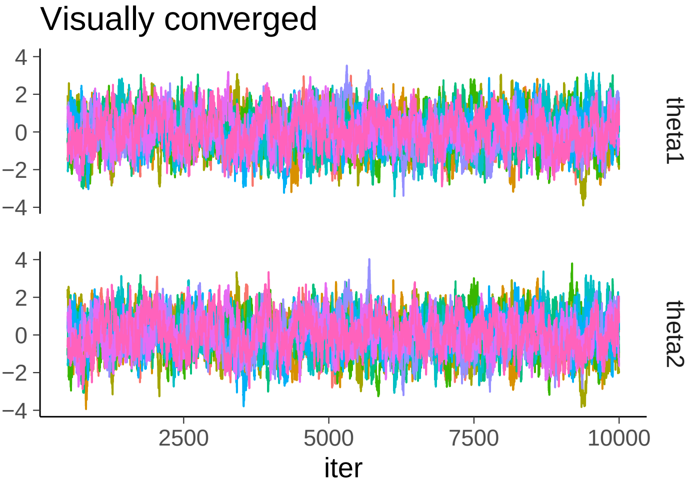
Several chains



Several chains



Several chains



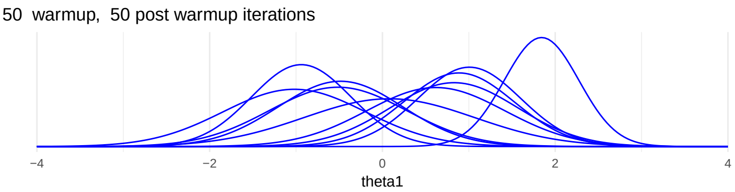
Visual convergence check is not sufficient

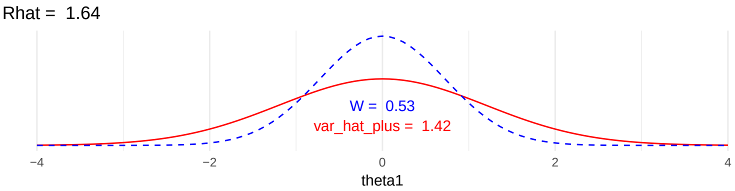
$\widehat{R}$: comparison of within and between variances

- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains

$\widehat{R}$: comparison of within and between variances

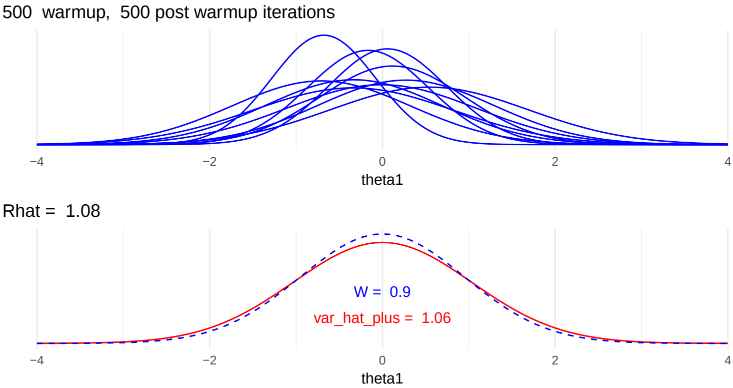
- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate





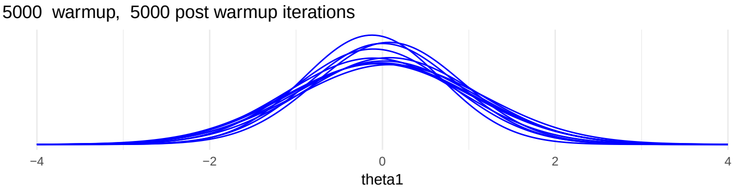
$\widehat{R}$: comparison of within and between variances

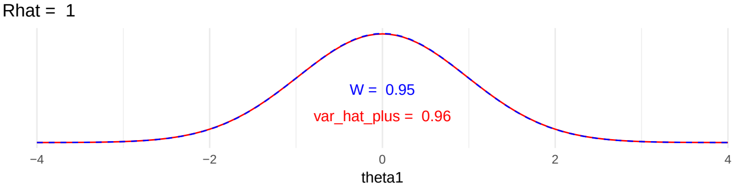
- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate



$\widehat{R}$: comparison of within and between variances

- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate





$\widehat{R}$

- Within chains variance W

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2$$

$\widehat{R}$

- Within chains variance W

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2$$

- Between chains variance B

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2, \text{ where } \bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$$

  - B/n is variance of the means of the chains

$\widehat{R}$

- Within chains variance W

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2$$

- Between chains variance B

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2, \text{ where } \bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$$

  - B/n is variance of the means of the chains

- Estimate total variance $\mathrm{var}(\psi|y)$ as a weighted mean of W and B

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

$\widehat{R}$

- Estimate total variance $\mathrm{var}(\psi|y)$ as a weighted mean of W and B
$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$
  - this overestimates marginal posterior variance if the starting points are overdispersed

$\widehat{R}$

- Estimate total variance $\mathrm{var}(\psi|y)$ as a weighted mean of W and B

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

  - this overestimates marginal posterior variance if the starting points are overdispersed

- Given finite n, W underestimates marginal posterior variance
  - single chains have not yet visited all points in the distribution
  - when $n \to \infty, \quad \mathrm{E}(W) \to \mathrm{var}(\psi|y)$

$\widehat{R}$

- Estimate total variance $\mathrm{var}(\psi|y)$ as a weighted mean of W and B

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

  - this overestimates marginal posterior variance if the starting points are overdispersed

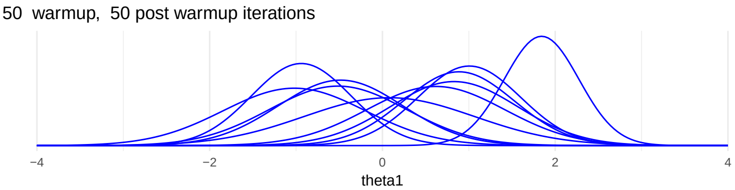
- Given finite n, W underestimates marginal posterior variance
  - single chains have not yet visited all points in the distribution
  - when $n \to \infty$, $\quad \mathrm{E}(W) \to \mathrm{var}(\psi|y)$

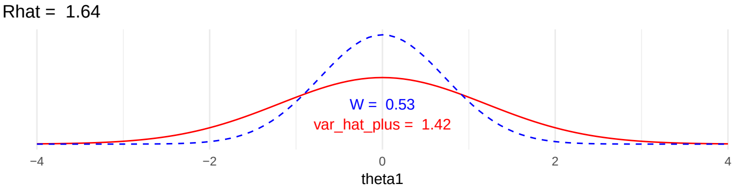
- As $\widehat{\mathrm{var}}^{+}(\psi|y)$ overestimates and W underestimates, compute

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}}$$

$\widehat{R}$

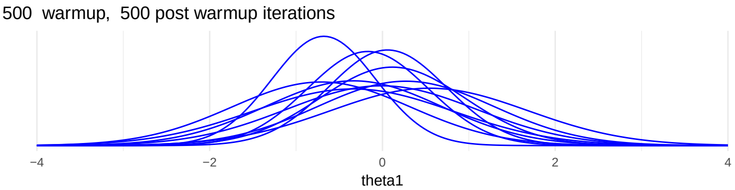
- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate

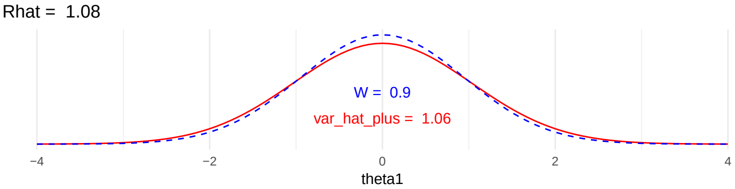




$\widehat{R}$

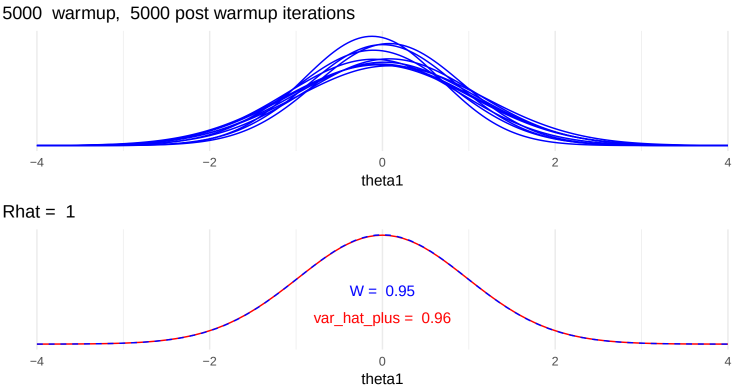
- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate





$\widehat{R}$

- BDA3: $\widehat{R}$ aka *potential scale reduction factor* (PSRF)
- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate



$\widehat{R}$

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}}{W}}$$

- Estimates how much the scale of ψ could reduce if $n \to \infty$
- $R \to 1$, when $n \to \infty$
- if R is big (e.g., $R > 1.01$), keep sampling

$\widehat{R}$

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}}{W}}$$

- Estimates how much the scale of ψ could reduce if $n \to \infty$
- $R \to 1$, when $n \to \infty$
- if R is big (e.g., $R > 1.01$), keep sampling
- If R close to 1, it is still possible that chains have not converged
  - if starting points were not overdispersed
  - distribution far from normal (especially if infinite variance)
  - just by chance when n is finite

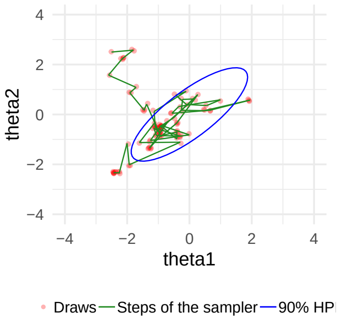
Split-$\widehat{R}$

- BDA3: split-$\widehat{R}$
- Examines *mixing* and *stationarity* of chains
- To examine stationarity chains are splitted to two parts
  - after splitting, we have m chains, each having n draws
  - scalar draws $\psi_{ij} \quad (i = 1, \ldots, n; j = 1, \ldots, m)$
  - compare means and variances of the split chains

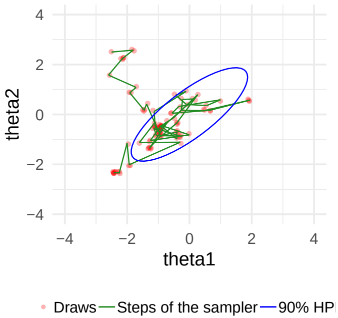
Time series analysis

- Auto correlation function
  - describes the correlation given a certain lag
  - can be used to compare efficiency of MCMC algorithms and parameterizations

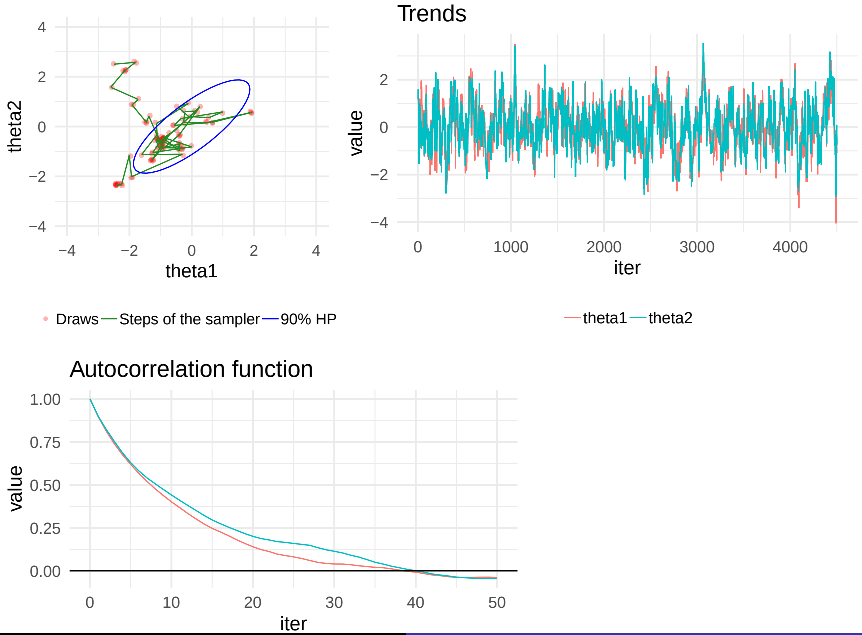
Auto correlation



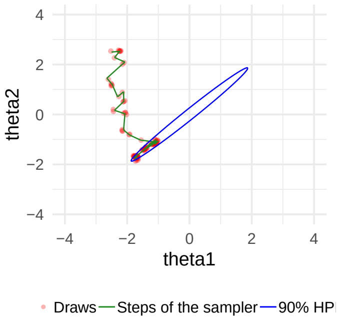
Auto correlation

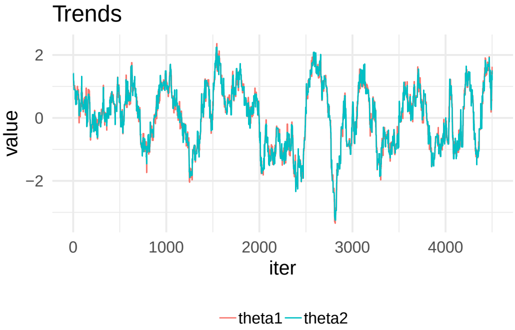


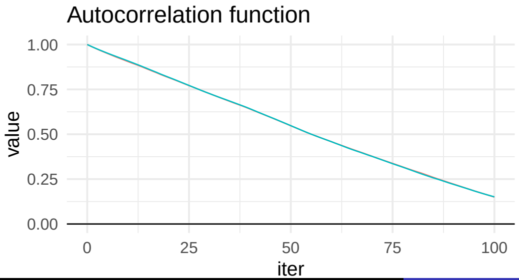
Auto correlation



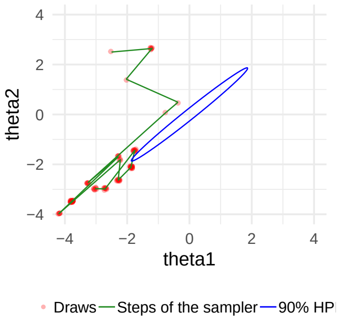
Auto correlation

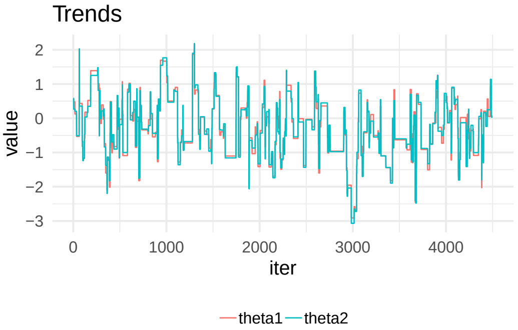


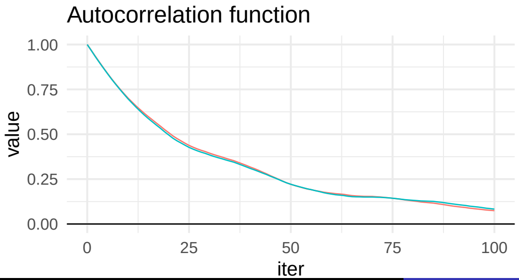




Auto correlation







Time series analysis

- Time series analysis can be used to estimate Monte Carlo error in case of MCMC
- For expectation $\bar{\theta}$

$$\mathrm{Var}[\bar{\theta}] = \frac{\sigma_\theta^2}{N/\tau}$$

where τ is sum of autocorrelations

  - τ describes how many dependent draws correspond to one independent sample
  - in BDA3 $N = nm$
  - $n_{\mathrm{eff}} = nm/\tau$
  - BDA3 focuses on n_{eff} and not the Monte Carlo error directly

Time series analysis

- Estimation of the autocorrelation using several chains

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{j=1}^{m} \hat{\rho}_{t,j}}{2\widehat{\mathrm{var}}^+}$$

where $\hat{\rho}_{t,j}$ is autocorrelation at lag t for chain j

Time series analysis

- Estimation of the autocorrelation using several chains

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{j=1}^{m} \hat{\rho}_{t,j}}{2\widehat{\text{var}}^{+}}$$

where $\hat{\rho}_{t,j}$ is autocorrelation at lag t for chain j

- BDA3 has slightly different less accurate equation. The above equation is used in Stan 2.18+

Time series analysis

- Estimation of the autocorrelation using several chains

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{j=1}^{m} \hat{\rho}_{t,j}}{2\widehat{\text{var}}^{+}}$$

  where $\hat{\rho}_{t,j}$ is autocorrelation at lag t for chain j
- BDA3 has slightly different less accurate equation. The above equation is used in Stan 2.18+
- Compared to usual method which computes the autocorrelation from a single chain, this estimate has smaller variance

Time series analysis

- Estimation of τ

$$\tau = 1 + 2 \sum_{t=1}^{\infty} \hat{\rho}_t$$

where $\hat{\rho}_t$ is empirical autocorrelation

  - empirical autocorrelation function is noisy and thus estimate of τ is noisy
  - noise is larger for longer lags (less observations)
  - less noisy estimate is obtained by truncating

$$\tau \approx 1 + 2 \sum_{t=1}^{T} \hat{\rho}_t$$

Time series analysis

- Estimation of τ

$$\tau = 1 + 2\sum_{t=1}^{\infty} \hat{\rho}_t$$

  where $\hat{\rho}_t$ is empirical autocorrelation

  - empirical autocorrelation function is noisy and thus estimate of τ is noisy
  - noise is larger for longer lags (less observations)
  - less noisy estimate is obtained by truncating

$$\tau \approx 1 + 2\sum_{t=1}^{T} \hat{\rho}_t$$

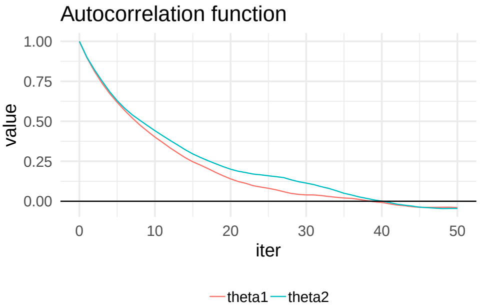
- As τ is estimated from a finite number of draws, it's expectation is overoptimistic
  - if $\tau > mn/20$ then the estimate is unreliable

Geyer's adaptive window estimator

- Truncation can be decided adaptively
  - for stationary, irreducible, recurrent Markov chain
  - let $\Gamma_m = \rho_{2m} + \rho_{2m+1}$, which is sum of two consequent autocorrelations
  - Γ_m is positive, decreasing and convex function of m

Geyer's adaptive window estimator

- Truncation can be decided adaptively
  - for stationary, irreducible, recurrent Markov chain
  - let $\Gamma_m = \rho_{2m} + \rho_{2m+1}$, which is sum of two consequent autocorrelations
  - Γ_m is positive, decreasing and convex function of m
- Initial positive sequence estimator (Geyer's IPSE)
  - Choose the largest m so, that all values of the sequence $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m$ are positive



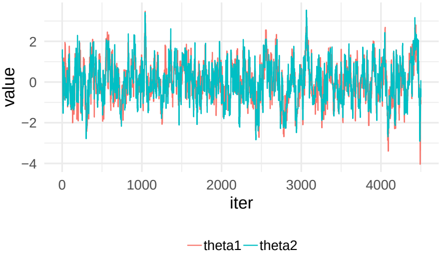
Effective sample size

Effective number of draws $n_{\mathrm{eff}} \approx N/\tau$

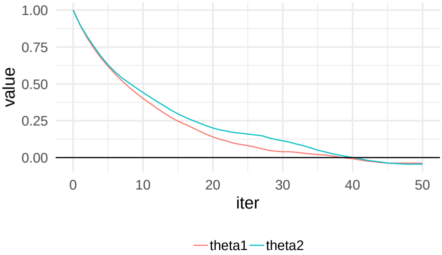
Effective sample size

Effective number of draws $n_{\text{eff}} \approx N/\tau$

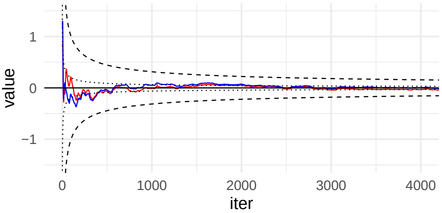
Trends



Autocorrelation function



Cumulative averages

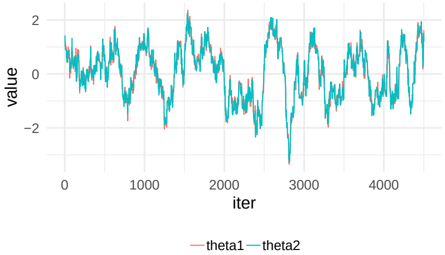


$$\tau \approx 1 + 2\sum_{t=1}^{T} \hat{\rho}_t$$
$$\approx 24$$

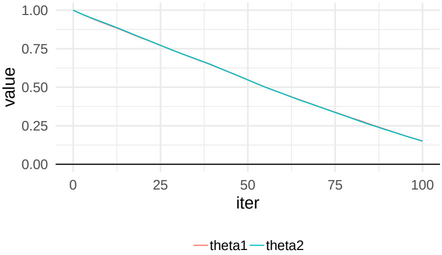
Effective sample size

Effective number of draws $n_{\text{eff}} \approx N/\tau$

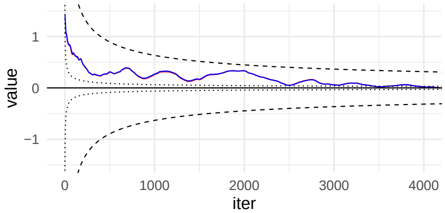
Trends



Autocorrelation function



Cumulative averages



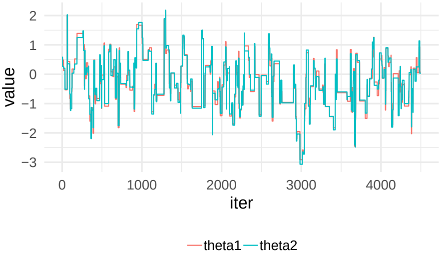
$$\tau \approx 1 + 2 \sum_{t=1}^{T} \hat{\rho}_t$$

$$\approx 104$$

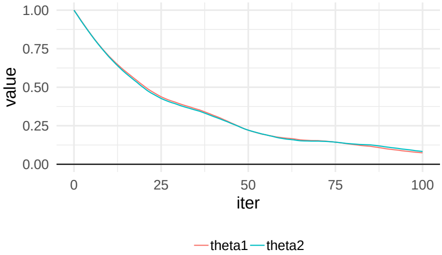
Effective sample size

Effective number of draws $n_{\text{eff}} \approx N/\tau$

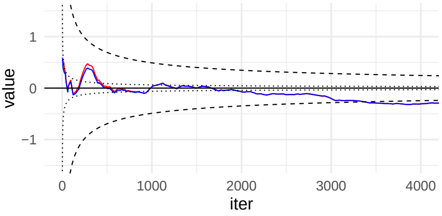
Trends



Autocorrelation function



Cumulative averages

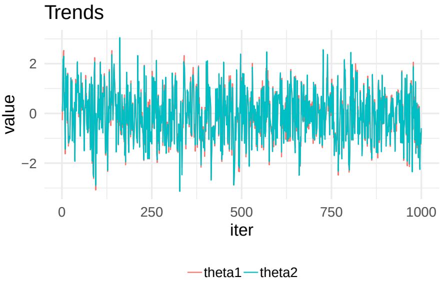


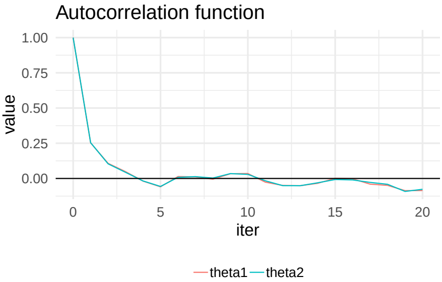
$$\tau \approx 1 + 2\sum_{t=1}^{T} \hat{\rho}_t$$

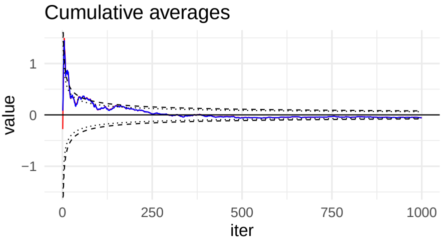
$$\approx 63$$

Dynamic HMC

Effective number of draws $n_{\text{eff}} \approx N/\tau$







$$\tau \approx 1 + 2 \sum_{t=1}^{T} \hat{\rho}_t$$

$$\approx 1.6$$

Problematic distributions

- Nonlinear dependencies
  - optimal proposal depends on location

Problematic distributions

- Nonlinear dependencies
  - optimal proposal depends on location
- Funnels
  - optimal proposal depends on location

Problematic distributions

- Nonlinear dependencies
  - optimal proposal depends on location
- Funnels
  - optimal proposal depends on location
- Multimodal
  - difficult to move from one mode to another

Problematic distributions

- Nonlinear dependencies
  - optimal proposal depends on location
- Funnels
  - optimal proposal depends on location
- Multimodal
  - difficult to move from one mode to another
- Long-tailed with non-finite variance and mean
  - central limit theorem for expectations does not hold

Further diagnostics

- Dynamic HMC/NUTS has additional diagnostics
  - divergences
  - tree depth exceedences
  - fraction of missing information