Lecture 9: More MCMC: Adaptive Metropolis, Metropolis-Hastings, and Gibbs

Merlise Clyde

September 23

Example from Last Class

Priors with $\sigma^2 = 1$:

$$p(\mu) \propto 1$$

Example from Last Class

Priors with $\sigma^2 = 1$:

$$p(\mu) \propto 1$$

- Use a Cauchy$(0, 1)$ prior on σ_μ independent of μ and

Example from Last Class

Priors with $\sigma^2 = 1$:

$$p(\mu) \propto 1$$

- Use a Cauchy$(0, 1)$ prior on σ_μ independent of μ and
- Symmetric proposal for μ and σ_τ

Example from Last Class

Priors with $\sigma^2 = 1$:

$$p(\mu) \propto 1$$

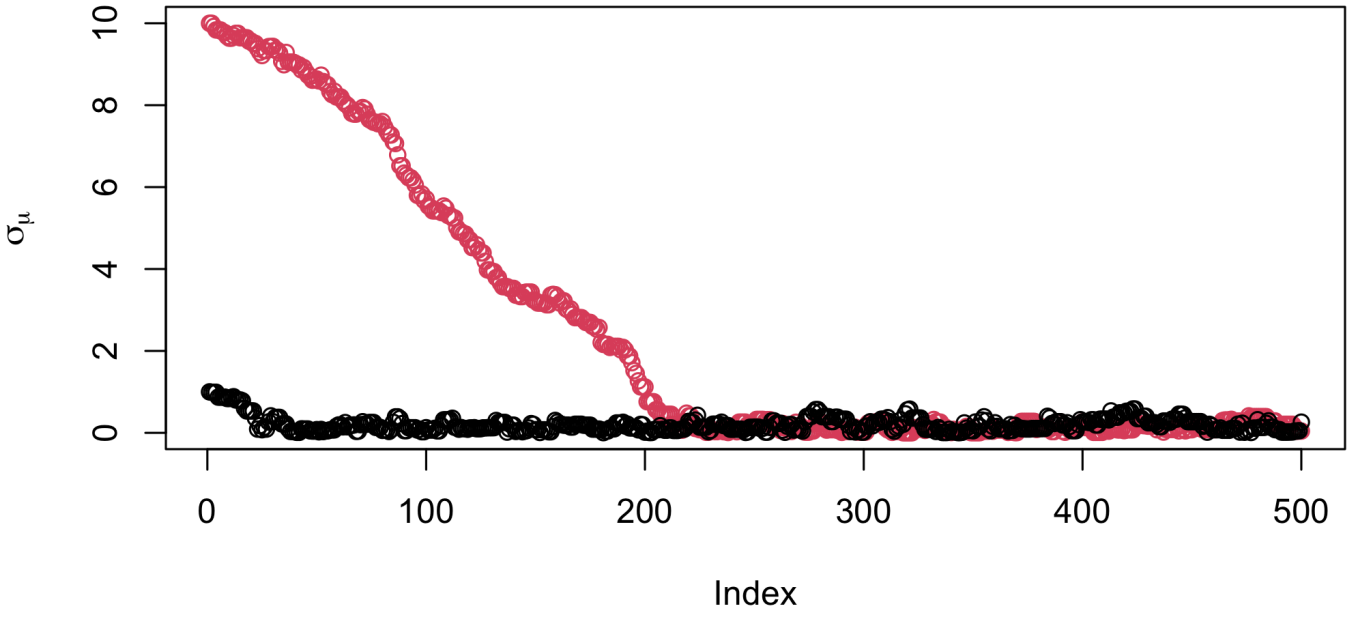
- Use a Cauchy$(0, 1)$ prior on σ_μ independent of μ and
- Symmetric proposal for μ and σ_τ
- Independent normals centered at current values of μ and σ_μ with covariance $\frac{2.4^2}{d}$Cov(θ) where $d = 2$ (the dimension of θ)

Gelman-Rubin

Gelman & Rubin suggested a diagnostic R based on taking separate chains with dispersed initial values to test convergence

Gelman-Rubin

Gelman & Rubin suggested a diagnostic R based on taking separate chains with dispersed initial values to test convergence



Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

- numerator and denominator are both unbiased estimates of the variance if the two chains have converged

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

- numerator and denominator are both unbiased estimates of the variance if the two chains have converged
  - otherwise W is an underestimate (hasn't explored enough)

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

- numerator and denominator are both unbiased estimates of the variance if the two chains have converged
  - otherwise W is an underestimate (hasn't explored enough)
  - numerator will overestimate as B is too large (overdispersed starting points)

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

- numerator and denominator are both unbiased estimates of the variance if the two chains have converged
  - otherwise W is an underestimate (hasn't explored enough)
  - numerator will overestimate as B is too large (overdispersed starting points)
- As $S \to \infty$ and $B \to 0$, $R \to 1$

Gelman-Rubin Diagnostic

- Run m > 2 chains of length 2S from overdispersed starting values.
- Discard the first S draws in each chain.
- Calculate the pooled within-chain variance W and between-chain variance B.

$$R = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}$$

- numerator and denominator are both unbiased estimates of the variance if the two chains have converged
  - otherwise W is an underestimate (hasn't explored enough)
  - numerator will overestimate as B is too large (overdispersed starting points)
- As $S \to \infty$ and $B \to 0$, $R \to 1$
- Note: version in R is slightly different

Gelman-Rubin Diagnostic

```
theta.mcmc = mcmc.list(mcmc(theta1, start=5000), mcmc(theta2, star
gelman.diag(theta.mcmc)
```

```
## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## mu                1          1
## sigma_mu          1          1
##
## Multivariate psrf
##
## 1
```

Gelman-Rubin Diagnostic

```
theta.mcmc = mcmc.list(mcmc(theta1, start=5000), mcmc(theta2, star
gelman.diag(theta.mcmc)
```

```
## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## mu                1          1
## sigma_mu          1          1
##
## Multivariate psrf
##
## 1
```

- Values of $R > 1.1$ suggest lack of convergence

Gelman-Rubin Diagnostic

```
theta.mcmc = mcmc.list(mcmc(theta1, start=5000), mcmc(theta2, star
gelman.diag(theta.mcmc)
```

```
## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## mu                1          1
## sigma_mu          1          1
##
## Multivariate psrf
##
## 1
```

- Values of $R > 1.1$ suggest lack of convergence
- Looks OK

Gelman-Rubin Diagnostic

```
theta.mcmc = mcmc.list(mcmc(theta1, start=5000), mcmc(theta2, star
gelman.diag(theta.mcmc)
```

```
## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## mu                1          1
## sigma_mu          1          1
##
## Multivariate psrf
##
## 1
```

- Values of $R > 1.1$ suggest lack of convergence
- Looks OK

See also `gelman.plot`

Geweke statistic

- Geweke proposed taking two non-overlapping parts of a single Markov chain (usually the first 10% and the last 50%) and comparing the mean of both parts, using a difference of means test

Geweke statistic

- Geweke proposed taking two non-overlapping parts of a single Markov chain (usually the first 10% and the last 50%) and comparing the mean of both parts, using a difference of means test
- The null hypothesis would be that the two parts of the chain are from the same distribution.

Geweke statistic

- Geweke proposed taking two non-overlapping parts of a single Markov chain (usually the first 10% and the last 50%) and comparing the mean of both parts, using a difference of means test
- The null hypothesis would be that the two parts of the chain are from the same distribution.
- The test statistic is a z-score with standard errors adjusted for autocorrelation, and if the p-value is significant for a variable, you need more draws.

Geweke Diagnostic

- The output is the z-score itself (not the p-value).

```
geweke.diag(theta.mcmc)
```

```
## [[1]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##       mu sigma_mu
##  -0.7779   0.7491
##
##
## [[2]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##       mu sigma_mu
##   0.4454   0.6377
```

Practical advice on diagnostics

- There are more tests we can use: Raftery and Lewis diagnostic, Heidelberger and Welch, etc.

Practical advice on diagnostics

- There are more tests we can use: Raftery and Lewis diagnostic, Heidelberger and Welch, etc.
- The Gelman-Rubin approach is quite appealing in using multiple chains

Practical advice on diagnostics

- There are more tests we can use: Raftery and Lewis diagnostic, Heidelberger and Welch, etc.
- The Gelman-Rubin approach is quite appealing in using multiple chains
- Geweke (and Heidelberger and Welch) sometimes reject even when the trace plots look good.

Practical advice on diagnostics

- There are more tests we can use: Raftery and Lewis diagnostic, Heidelberger and Welch, etc.
- The Gelman-Rubin approach is quite appealing in using multiple chains
- Geweke (and Heidelberger and Welch) sometimes reject even when the trace plots look good.
- Overly sensitive to minor departures from stationarity that do not impact inferences.

Practical advice on diagnostics

- There are more tests we can use: Raftery and Lewis diagnostic, Heidelberger and Welch, etc.
- The Gelman-Rubin approach is quite appealing in using multiple chains
- Geweke (and Heidelberger and Welch) sometimes reject even when the trace plots look good.
- Overly sensitive to minor departures from stationarity that do not impact inferences.
- Most common method of assessing convergence is visual examination of trace plots.

Improving

- more iterations and multiple chains

Improving

- more iterations and multiple chains
- thinning to reduce correlations and increase ESS

Improving

- more iterations and multiple chains
- thinning to reduce correlations and increase ESS
- change the proposal distribution q

Proposal Distribution

Common choice

$$\mathsf{N}(\theta^{\star}; \theta^{(s)}, \delta^2 \Sigma)$$

Proposal Distribution

Common choice

$$\mathsf{N}(\theta^\star; \theta^{(s)}, \delta^2 \Sigma)$$

- rough estimate of Σ based on the asymptotic Gaussian approximation $\mathsf{Cov}(\theta \mid y)$ and $\delta = 2.38/\sqrt{\dim(\theta)}$

Proposal Distribution

Common choice

$$\mathsf{N}(\theta^\star; \theta^{(s)}, \delta^2 \Sigma)$$

- rough estimate of Σ based on the asymptotic Gaussian approximation $\mathsf{Cov}(\theta \mid y)$ and $\delta = 2.38/\sqrt{\dim(\theta)}$
  - find the MAP estimate (posterior mode) $\hat{\theta}$

Proposal Distribution

Common choice

$$\mathsf{N}(\theta^\star; \theta^{(s)}, \delta^2 \Sigma)$$

- rough estimate of Σ based on the asymptotic Gaussian approximation $\mathsf{Cov}(\theta \mid y)$ and $\delta = 2.38/\sqrt{\dim(\theta)}$
  - find the MAP estimate (posterior mode) $\hat{\theta}$
  - take

$$\Sigma = \left[-\frac{\partial^2 \log(\mathcal{L}(\theta)) + \log(\pi(\theta))}{\partial \theta \partial \theta^T} \right]^{-1}_{\theta = \hat{\theta}}$$

`

Proposal Distribution

Common choice

$$\mathsf{N}(\theta^{\star}; \theta^{(s)}, \delta^2 \Sigma)$$

- rough estimate of Σ based on the asymptotic Gaussian approximation $\mathsf{Cov}(\theta \mid y)$ and $\delta = 2.38/\sqrt{\dim(\theta)}$
  - find the MAP estimate (posterior mode) $\hat{\theta}$
  - take

$$\Sigma = \left[-\frac{\partial^2 \log(\mathcal{L}(\theta)) + \log(\pi(\theta))}{\partial \theta \partial \theta^T} \right]^{-1}_{\theta = \hat{\theta}}$$

`

- ignore prior and use inverse of Fisher Information (covariance of MLE)

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption
- Workaround? run an initial MCMC for an initial tuning phase (e.g. 1000 samples) and then fix the kernel to depend only on $\theta^{(s-1)}$ and y.

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption
- Workaround? run an initial MCMC for an initial tuning phase (e.g. 1000 samples) and then fix the kernel to depend only on $\theta^{(s-1)}$ and y.
- more elegant approach - formal **adaptive Metropolis**

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption
- Workaround? run an initial MCMC for an initial tuning phase (e.g. 1000 samples) and then fix the kernel to depend only on $\theta^{(s-1)}$ and y.
- more elegant approach - formal **adaptive Metropolis**
  - keep adapting the entire time!

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption
- Workaround? run an initial MCMC for an initial tuning phase (e.g. 1000 samples) and then fix the kernel to depend only on $\theta^{(s-1)}$ and y.
- more elegant approach - formal **adaptive Metropolis**
  - keep adapting the entire time!
  - this may mess up convergence !

Adaptive Metropolis?

- MCMC doesn't allow you to use the full history of the chain $\theta^{(1)}, \ldots, \theta^{(s)}$ in constructing the proposal distributions
- violates the Markov assumption
- Workaround? run an initial MCMC for an initial tuning phase (e.g. 1000 samples) and then fix the kernel to depend only on $\theta^{(s-1)}$ and y.
- more elegant approach - formal **adaptive Metropolis**
  - keep adapting the entire time!
  - this may mess up convergence !
  - need conditions for vanishing adaptation e.g. that the proposal depends less and less on recent states in the chain - Roberts & Rosenthal (2006) and other conditions

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)
- run RWM with a Gaussian proposal for a fixed number of iterations for $s < s_0$

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)
- run RWM with a Gaussian proposal for a fixed number of iterations for $s < s_0$
- estimate of covariance at state s

$$\Sigma^{(s)} = \frac{1}{s}\left(\sum_{i=1}^{s} \theta^{(i)}\theta^{(i)^T} - s\bar{\theta}^{(s)}\bar{\theta}^{(s)^T}\right)$$

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)
- run RWM with a Gaussian proposal for a fixed number of iterations for $s < s_0$
- estimate of covariance at state s

$$\Sigma^{(s)} = \frac{1}{s}\left(\sum_{i=1}^{s} \theta^{(i)}\theta^{(i)^T} - s\bar{\theta}^{(s)}\bar{\theta}^{(s)^T}\right)$$

- proposal for $s > s_0$ with $\delta = 2.38/\sqrt{d}$

$$\theta^* \sim \mathsf{N}(\theta^{(s)}, \delta^2(\Sigma^{(s)} + \epsilon I_d))$$

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)
- run RWM with a Gaussian proposal for a fixed number of iterations for $s < s_0$
- estimate of covariance at state s

$$\Sigma^{(s)} = \frac{1}{s}\left(\sum_{i=1}^{s} \theta^{(i)}\theta^{(i)^T} - s\bar{\theta}^{(s)}\bar{\theta}^{(s)^T}\right)$$

- proposal for $s > s_0$ with $\delta = 2.38/\sqrt{d}$

$$\theta^* \sim \mathsf{N}(\theta^{(s)}, \delta^2(\Sigma^{(s)} + \epsilon I_d))$$

- $\epsilon > 0$ insures covariance is positive definite

Adaptive MCMC

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis (RWM)
- run RWM with a Gaussian proposal for a fixed number of iterations for $s < s_0$
- estimate of covariance at state s

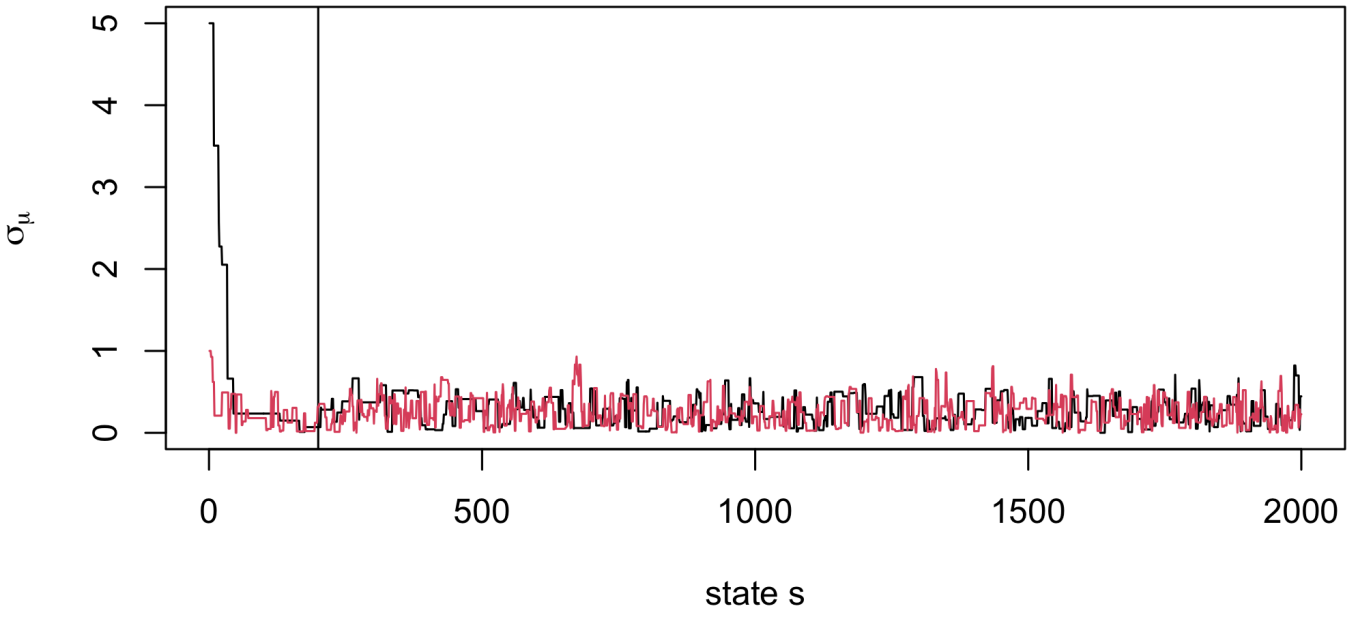
$$\Sigma^{(s)} = \frac{1}{s}\left(\sum_{i=1}^{s} \theta^{(i)}\theta^{(i)^T} - s\bar{\theta}^{(s)}\bar{\theta}^{(s)^T}\right)$$

- proposal for $s > s_0$ with $\delta = 2.38/\sqrt{d}$

$$\theta^* \sim \mathsf{N}(\theta^{(s)}, \delta^2(\Sigma^{(s)} + \epsilon I_d))$$

- $\epsilon > 0$ insures covariance is positive definite
- if s_0 is too large will take longer for adaptation to be seen

Example again



Acceptance rate now around 30-35 % of 10,000 iterations!

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric
- Hastings (1970) generalizes Metropolis algorithms to allow asymmetric proposals - aka Metropolis-Hastings or MH $q(\theta^* \mid \theta^{(s)})$ does not need to be the same as $q(\theta^{(s)} \mid \theta^*)$

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric
- Hastings (1970) generalizes Metropolis algorithms to allow asymmetric proposals - aka Metropolis-Hastings or MH $q(\theta^* \mid \theta^{(s)})$ does not need to be the same as $q(\theta^{(s)} \mid \theta^*)$
- propose $\theta^* \mid \theta^{(s)} \sim q(\theta^* \mid \theta^{(s)})$

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric
- Hastings (1970) generalizes Metropolis algorithms to allow asymmetric proposals - aka Metropolis-Hastings or MH $q(\theta^* \mid \theta^{(s)})$ does not need to be the same as $q(\theta^{(s)} \mid \theta^*)$
- propose $\theta^* \mid \theta^{(s)} \sim q(\theta^* \mid \theta^{(s)})$
- Acceptance probability

$$\min\left\{1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/q(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/q(\theta^{(s)} \mid \theta^*)}\right\}$$

Metropolis-Hastings (MH)

- Metropolis requires that the proposal distribution be symmetric
- Hastings (1970) generalizes Metropolis algorithms to allow asymmetric proposals - aka Metropolis-Hastings or MH $q(\theta^* \mid \theta^{(s)})$ does not need to be the same as $q(\theta^{(s)} \mid \theta^*)$
- propose $\theta^* \mid \theta^{(s)} \sim q(\theta^* \mid \theta^{(s)})$
- Acceptance probability

$$\min\left\{1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/q(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/q(\theta^{(s)} \mid \theta^*)}\right\}$$

- adjustment for asymmetry in acceptance ratio is key to ensuring convergence to stationary distribution!

Special cases

- Metropolis

Special cases

- Metropolis
- Independence chain

Special cases

- Metropolis
- Independence chain
- Gibbs samplers

Special cases

- Metropolis
- Independence chain
- Gibbs samplers
- Metropolis-within-Gibbs

Special cases

- Metropolis
- Independence chain
- Gibbs samplers
- Metropolis-within-Gibbs
- combinations of the above!

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

- what happens if the approximation is really accurate?

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

- what happens if the approximation is really accurate?
- probability of acceptance is ≈ 1

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

- what happens if the approximation is really accurate?
- probability of acceptance is ≈ 1
- Important caveat for convergence: tails of the posterior should be at least as heavy as the tails of the posterior (Tweedie 1994)

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

- what happens if the approximation is really accurate?
- probability of acceptance is ≈ 1
- Important caveat for convergence: tails of the posterior should be at least as heavy as the tails of the posterior (Tweedie 1994)
- Replace Gaussian by a Student-t with low degrees of freedom

Independence Chain

- suppose we have a good approximation $\tilde{\pi}(\theta \mid y)$ to $\pi(\theta \mid y)$
- Draw $\theta^* \sim \tilde{\pi}(\theta \mid y)$ *without* conditioning on $\theta^{(s)}$
- acceptance probability

$$\min \left\{ 1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)/\tilde{\pi}(\theta^* \mid \theta^{(s)})}{\pi(\theta^{(s)})\mathcal{L}(\theta^{(s)})/\tilde{\pi}(\theta^{(s)} \mid \theta^*)} \right\}$$

- what happens if the approximation is really accurate?
- probability of acceptance is ≈ 1
- Important caveat for convergence: tails of the posterior should be at least as heavy as the tails of the posterior (Tweedie 1994)
- Replace Gaussian by a Student-t with low degrees of freedom
- transformations of θ

Block Updates & Gibbs

So far all algorithms update all of the parameters simultaneously

Block Updates & Gibbs

So far all algorithms update all of the parameters simultaneously

- convenient to break problems in to K blocks and update them separately

Block Updates & Gibbs

So far all algorithms update all of the parameters simultaneously

- convenient to break problems in to K blocks and update them separately

- $\theta = (\theta_{[1]}, \ldots, \theta_{[K]}) = (\theta_1, \ldots, \theta_p)$

Block Updates & Gibbs

So far all algorithms update all of the parameters simultaneously

- convenient to break problems in to K blocks and update them separately

- $\theta = (\theta_{[1]}, \ldots, \theta_{[K]}) = (\theta_1, \ldots, \theta_p)$

At iteration s, for $k = 1, \ldots, K$ Cycle thru blocks: (fixed order or random order)

- propose $\theta^*_{[k]} \sim q_k(\theta_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})$

- set $\theta^{(s)}_{[k]} = \theta^*_{[k]}$ with probability

$$\min\left\{1, \frac{\pi(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^*_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^*_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}{\pi(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})\mathcal{L}(\theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[k]}, \theta^{(s-1)}_{[>k]})/q_k(\theta^{(s-1)}_{[k]} \mid \theta^{(s)}_{[<k]}, \theta^{(s-1)}_{[>k]})}\right\}$$

Gibbs Sampler

special case of MH

Gibbs Sampler

special case of MH

- proposal distribution q_k for the kth block is the **full conditional** distribution for $\theta_{[k]}$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) = \frac{\pi(\theta_{[k]}, \theta_{[-k]} \mid y)}{\pi(\theta_{[-k]} \mid y))} \propto \pi(\theta_{[k]}, \theta_{[-k]} \mid y)$$

Gibbs Sampler

special case of MH

- proposal distribution q_k for the kth block is the **full conditional** distribution for $\theta_{[k]}$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) = \frac{\pi(\theta_{[k]}, \theta_{[-k]} \mid y)}{\pi(\theta_{[-k]} \mid y))} \propto \pi(\theta_{[k]}, \theta_{[-k]} \mid y)$$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) \propto \mathcal{L}(\theta_{[k]}, \theta_{[-k]}) \pi(\theta_{[k]}, \theta_{[-k]})$$

Gibbs Sampler

special case of MH

- proposal distribution q_k for the kth block is the **full conditional** distribution for $\theta_{[k]}$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) = \frac{\pi(\theta_{[k]}, \theta_{[-k]} \mid y)}{\pi(\theta_{[-k]} \mid y))} \propto \pi(\theta_{[k]}, \theta_{[-k]} \mid y)$$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) \propto \mathcal{L}(\theta_{[k]}, \theta_{[-k]}) \pi(\theta_{[k]}, \theta_{[-k]})$$

- acceptance probability is always 1!

Gibbs Sampler

special case of MH

- proposal distribution q_k for the kth block is the **full conditional** distribution for $\theta_{[k]}$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) = \frac{\pi(\theta_{[k]}, \theta_{[-k]} \mid y)}{\pi(\theta_{[-k]} \mid y))} \propto \pi(\theta_{[k]}, \theta_{[-k]} \mid y)$$

$$\pi(\theta_{[k]} \mid \theta_{[-k]}, y) \propto \mathcal{L}(\theta_{[k]}, \theta_{[-k]}) \pi(\theta_{[k]}, \theta_{[-k]})$$

- acceptance probability is always 1!
- even though joint distribution is messy, full conditionals may be (conditionally) conjugate and easy to sample from!

Comments

- can use Gibbs steps and Metropolis Hastings steps together

Comments

- can use Gibbs steps and Metropolis Hastings steps together
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)

Comments

- can use Gibbs steps and Metropolis Hastings steps together
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- combine with adaptive Metropolis

Comments

- can use Gibbs steps and Metropolis Hastings steps together
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- combine with adaptive Metropolis
- Adaptive Independence Metropolis Hastings (learn a mixture)