Metropolis within Gibbs Algorithm: Application to Linear Geostatistical Model

Johnson, Olatunji Olugoke

o.johnson@lanacster.ac.uk

May 23, 2017

Outline

Introduction

What is Markov Chain Monte Carlo (MCMC)?

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

What is the goal of MCMC?

The goal of MCMC is to draw samples from some probability distribution without having to know its exact **height at any point**. The way MCMC achieves this is to "wander around" on that distribution in such a way that the amount of time spent in each location is proportional to the height of the distribution. If the "wandering around" process is set up correctly, you can make sure that this proportionality (between time spent and height of the distribution) is achieved.

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Introduction

Applications of MCMC

1. **Sampling** from any distribution.
2. Heuristic way of **Minimizing/Maximizing any function**.
3. **Crytography**: Breaking the code.
4. Finding out the **best arrangement of a DNA sequence**.
5. ?

Variants of MCMC

1. Metropolis-Hasting Algorithm
2. Gibbs Sampling

Fundamental Mechanisms of MCMC

Gibbs sampler

Given the target multivariate distribution $\pi(x) = \pi(x_1, \ldots, x_p)$, the Gibbs samplers successively and repeatedly generates samples for each of the random variables, X_i , from the full conditional distribution $(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$.

Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm simulates samples from a probability distribution by making use of the full joint density function and (independent) proposal distributions for each of the variables of interest.

Discuss some proposal distribution

Proposal

1. RANDOM WALK PROPOSAL: $q(x^*, x) = q(x^*, -x)$
2. INDEPENDENCE PROPOSAL: $q(x^*, x) = q(y)$
3. TAILORED PROPOSAL: $q(\cdot) = f(\cdot | mode, H^{-1})$
4. ?

Discuss some proposal distribution

Proposal

1. RANDOM WALK PROPOSAL: $q(x^*, x) = q(x^*, -x)$
2. INDEPENDENCE PROPOSAL: $q(x^*, x) = q(y)$
3. TAILORED PROPOSAL: $q(\cdot) = f(\cdot | mode, H^{-1})$
4. ?

Discuss some proposal distribution

Proposal

1. RANDOM WALK PROPOSAL: $q(x^*, x) = q(x^*, -x)$
2. INDEPENDENCE PROPOSAL: $q(x^*, x) = q(y)$
3. TAILORED PROPOSAL: $q(\cdot) = f(\cdot | mode, H^{-1})$
4. ?

Discuss some proposal distribution

Proposal

1. RANDOM WALK PROPOSAL: $q(x^*, x) = q(x^*, -x)$
2. INDEPENDENCE PROPOSAL: $q(x^*, x) = q(y)$
3. TAILORED PROPOSAL: $q(\cdot) = f(\cdot|mode, H^{-1})$
4. ?

Convergence

Challenges in studying convergence

1. **Black Box MCMC:** Not much information about Markov chain transition mechanism, equilibrium distribution and good starting value.
2. **Pseudo-Convergence:** Markov chain can appear to have converged to its equilibrium distribution when it has not.
3. **One Long Run versus Many Short Runs:** Autocorrelation plot on long enough chain can give good information about mixing
4. **Multistart heuristic is worse than useless:** it can give you confidence that all is well when in fact your results are completely erroneous.
5. **Burn-in:** A Markov chain started anywhere near the center of the equilibrium distribution needs no burn-in.

Convergence

Challenges in studying convergence

1. **Black Box MCMC:** Not much information about Markov chain transition mechanism, equilibrium distribution and good starting value.
2. **Pseudo-Convergence:** Markov chain can appear to have converged to its equilibrium distribution when it has not.
3. **One Long Run versus Many Short Runs:** Autocorrelation plot on long enough chain can give good information about mixing
4. **Multistart heuristic is worse than useless:** it can give you confidence that all is well when in fact your results are completely erroneous.
5. **Burn-in:** A Markov chain started anywhere near the center of the equilibrium distribution needs no burn-in.

Convergence

Challenges in studying convergence

1. **Black Box MCMC:** Not much information about Markov chain transition mechanism, equilibrium distribution and good starting value.
2. **Pseudo-Convergence:** Markov chain can appear to have converged to its equilibrium distribution when it has not.
3. **One Long Run versus Many Short Runs:** Autocorrelation plot on long enough chain can give good information about mixing
4. **Multistart heuristic is worse than useless:** it can give you confidence that all is well when in fact your results are completely erroneous.
5. **Burn-in:** A Markov chain started anywhere near the center of the equilibrium distribution needs no burn-in.

Convergence

Challenges in studying convergence

1. **Black Box MCMC:** Not much information about Markov chain transition mechanism, equilibrium distribution and good starting value.
2. **Pseudo-Convergence:** Markov chain can appear to have converged to its equilibrium distribution when it has not.
3. **One Long Run versus Many Short Runs:** Autocorrelation plot on long enough chain can give good information about mixing
4. **Multistart heuristic is worse than useless:** it can give you confidence that all is well when in fact your results are completely erroneous.
5. **Burn-in:** A Markov chain started anywhere near the center of the equilibrium distribution needs no burn-in.

Convergence

Challenges in studying convergence

1. **Black Box MCMC:** Not much information about Markov chain transition mechanism, equilibrium distribution and good starting value.
2. **Pseudo-Convergence:** Markov chain can appear to have converged to its equilibrium distribution when it has not.
3. **One Long Run versus Many Short Runs:** Autocorrelation plot on long enough chain can give good information about mixing
4. **Multistart heuristic is worse than useless:** it can give you confidence that all is well when in fact your results are completely erroneous.
5. **Burn-in:** A Markov chain started anywhere near the center of the equilibrium distribution needs no burn-in.

Convergence

Diagnostics

1. If the chain converges, then the diagnostic will probably say that the chain converged, but they do not say that if the chain pseudo-converges, then the diagnostic will probably say that the chain did not converge.
2. There is only one perfect MCMC diagnostic: perfect sampling (Propp and Wilson, 1996; Kendall and Meller, 2000).
3. Perfect sampling is achieved by sampling a suficiently large sample from the equilibrium distribution of the Markov chain. paradoxically, if it fails to produce an iid from the equilibrium distribution, then the underlying Markov chain is useless for sampling.

Convergence

Diagnostics

1. If the chain converges, then the diagnostic will probably say that the chain converged, but they do not say that if the chain pseudo-converges, then the diagnostic will probably say that the chain did not converge.
2. There is only one perfect MCMC diagnostic: perfect sampling (Propp and Wilson, 1996; Kendall and Meller, 2000).
3. Perfect sampling is achieved by sampling a suficiently large sample from the equilibrium distribution of the Markov chain. paradoxically, if it fails to produce an iid from the equilibrium distribution, then the underlying Markov chain is useless for sampling.

Convergence

Diagnostics

1. If the chain converges, then the diagnostic will probably say that the chain converged, but they do not say that if the chain pseudo-converges, then the diagnostic will probably say that the chain did not converge.
2. There is only one perfect MCMC diagnostic: perfect sampling (Propp and Wilson, 1996; Kendall and Meller, 2000).
3. Perfect sampling is achieved by sampling a suficiently large sample from the equilibrium distribution of the Markov chain. paradoxically, if it fails to produce an iid from the equilibrium distribution, then the underlying Markov chain is useless for sampling.

Convergence

Diagnostics (cont.)

1. Unfortunately, Perfect sampling does not work on black box MCMC because it requires complicated theoretical conditions on the Markov chain dynamics.
2. Unfortunately, no one knows how to do perfect sampling for most MCMC applications.

Convergence

Diagnostics (cont.)

1. Unfortunately, Perfect sampling does not work on black box MCMC because it requires complicated theoretical conditions on the Markov chain dynamics.
2. Unfortunately, no one knows how to do perfect sampling for most MCMC applications.

Gaussian Geostatistical Model

Classical geostatistical Model

$$Y(x_i) = d^\top(x_i)\beta + S(x_i) + Z(x_i)$$

Hierarchical framework

1. Data Model: $Y|\beta, S(x), \tau^2 \sim MVN(D\beta + S(x), \tau^2 I)$
2. Process Model: $S(x)|\theta \sim MVN(0, \Sigma(\theta))$
3. Parameter Model: $[\beta, \tau^2, \theta]$

Combining data and process stages

1. Data Model: $Y|\beta, \theta, \tau^2 \sim MVN(D\beta, \Sigma(\theta) + \tau^2 I)$
2. Parameter Model: $[\beta, \tau^2, \theta]$

Gaussian Geostatistical Model

Classical geostatistical Model

$$Y(x_i) = d^\top(x_i)\beta + S(x_i) + Z(x_i)$$

Hierarchical framework

1. Data Model: $Y|\beta, S(x), \tau^2 \sim MVN(D\beta + S(x), \tau^2 I)$
2. Process Model: $S(x)|\theta \sim MVN(0, \Sigma(\theta))$
3. Parameter Model: $[\beta, \tau^2, \theta]$

Combining data and process stages

1. Data Model: $Y|\beta, \theta, \tau^2 \sim MVN(D\beta, \Sigma(\theta) + \tau^2 I)$
2. Parameter Model: $[\beta, \tau^2, \theta]$

Gaussian Geostatistical Model

Classical geostatistical Model

$$Y(x_i) = d^\top(x_i)\beta + S(x_i) + Z(x_i)$$

Hierarchical framework

1. Data Model: $Y|\beta, S(x), \tau^2 \sim MVN(D\beta + S(x), \tau^2 I)$
2. Process Model: $S(x)|\theta \sim MVN(0, \Sigma(\theta))$
3. Parameter Model: $[\beta, \tau^2, \theta]$

Combining data and process stages

1. Data Model: $Y|\beta, \theta, \tau^2 \sim MVN(D\beta, \Sigma(\theta) + \tau^2 I)$
2. Parameter Model: $[\beta, \tau^2, \theta]$

Gaussian Geostatistical Model

Classical geostatistical Model

$$Y(x_i) = d^\top(x_i)\beta + S(x_i) + Z(x_i)$$

Hierarchical framework

1. Data Model: $Y|\beta, S(x), \tau^2 \sim MVN(D\beta + S(x), \tau^2 I)$
2. Process Model: $S(x)|\theta \sim MVN(0, \Sigma(\theta))$
3. Parameter Model: $[\beta, \tau^2, \theta]$

Combining data and process stages

1. Data Model: $Y|\beta, \theta, \tau^2 \sim MVN(D\beta, \Sigma(\theta) + \tau^2 I)$
2. Parameter Model: $[\beta, \tau^2, \theta]$

Gaussian Geostatistical Model

Classical geostatistical Model

$$Y(x_i) = d^\top(x_i)\beta + S(x_i) + Z(x_i)$$

Hierarchical framework

1. Data Model: $Y|\beta, S(x), \tau^2 \sim MVN(D\beta + S(x), \tau^2 I)$
2. Process Model: $S(x)|\theta \sim MVN(0, \Sigma(\theta))$
3. Parameter Model: $[\beta, \tau^2, \theta]$

Combining data and process stages

1. Data Model: $Y|\beta, \theta, \tau^2 \sim MVN(D\beta, \Sigma(\theta) + \tau^2 I)$
2. Parameter Model: $[\beta, \tau^2, \theta]$

Full conditional distributions

Set of full conditionals of the posterior

- $\beta|y, \tau^2, \theta \sim N(M, V)$
- $\theta|y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \tau^2, \theta)$
- $\tau^2|y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \theta, \tau^2)$

where,

- $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$
- $\pi(\theta) \sim IG(a, b)$
- $\pi(\tau^2) \sim IG(c, d)$
- $V = \left(\Sigma_\beta^{-1} + D'\Sigma_{y|\cdot}^{-1}D\right)^{-1}$
- $M = V\left(\Sigma_\beta^{-1}\mu_\beta + D'\Sigma_{y|\cdot}^{-1}y\right)$

Full conditional distributions

Set of full conditionals of the posterior

- $\beta | y, \tau^2, \theta \sim N(M, V)$
- $\theta | y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \tau^2, \theta)$
- $\tau^2 | y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \theta, \tau^2)$

where,

- $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$
- $\pi(\theta) \sim IG(a, b)$
- $\pi(\tau^2) \sim IG(c, d)$
- $V = \left(\Sigma_\beta^{-1} + D'\Sigma_{y|\cdot}^{-1} D\right)^{-1}$
- $M = V\left(\Sigma_\beta^{-1}\mu_\beta + D'\Sigma_{y|\cdot}^{-1} y\right)$

Full conditional distributions

Set of full conditionals of the posterior

- $\beta|y, \tau^2, \theta \sim N(M, V)$
- $\theta|y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \tau^2, \theta)$
- $\tau^2|y, \beta, \theta \propto f(y|\beta, \theta, \tau^2)\pi(\beta, \theta, \tau^2)$

where,

- $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$
- $\pi(\theta) \sim IG(a, b)$
- $\pi(\tau^2) \sim IG(c, d)$
- $V = \left(\Sigma_\beta^{-1} + D'\Sigma_{y|\cdot}^{-1} D\right)^{-1}$
- $M = V\left(\Sigma_\beta^{-1}\mu_\beta + D'\Sigma_{y|\cdot}^{-1} y\right)$

Limitation

Limitation

1. Hastings-within-Gibbs algorithm requires a lot of computational work.
2. slow mixing.
3. inaccurate.

Limitation

Limitation

1. Hastings-within-Gibbs algorithm requires a lot of computational work.
2. slow mixing.
3. inaccurate.

Limitation

Limitation

1. Hastings-within-Gibbs algorithm requires a lot of computational work.
2. slow mixing.
3. inaccurate.

Conclusions

Conclusions

1. slow mixing and strong autocorrelation: the chains for β.
2. significant dependence: cross-correlations among the covariance parameters θ.
3. It is a good start for a new problem

Conclusions

Conclusions

1. slow mixing and strong autocorrelation: the chains for β.
2. significant dependence: cross-correlations among the covariance parameters θ.
3. It is a good start for a new problem

Conclusions

Conclusions

1. slow mixing and strong autocorrelation: the chains for β.
2. significant dependence: cross-correlations among the covariance parameters θ.
3. It is a good start for a new problem

Recommendations

Recommendations

1. Some properties must be study with care before conclusion.
2. Ensure that the conditional distribution is derived from a valid joint distribution, that is the joint distribution must have a finite integral (though not always easy to validate).

Recommendations

Recommendations

1. Some properties must be study with care before conclusion.
2. Ensure that the conditional distribution is derived from a valid joint distribution, that is the joint distribution must have a finite integral (though not always easy to validate).

THANKS FOR YOUR ATTENTION !!!