Markov chain Monte Carlo

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

Negative-Binomial regression example

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

Negative-Binomial regression example

Markov chains

- Let $(X_t) = (X_0, X_1, X_2, \ldots)$ be a sequence of random vars.

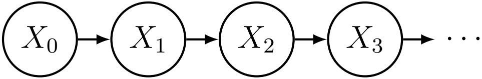
- (X_t) is a *Markov chain* if for all t,

$$X_{t+1} \perp\!\!\!\perp (X_1, \ldots, X_{t-1}) \mid X_t$$

  that is, $p(x_{t+1}|x_{1:t}) = p(x_{t+1}|x_t)$ for all $x_{1:t+1}$.

- In other words, "the future is conditionally independent of the past given the present."

- This is equivalent to saying that the distribution respects the following directed graph:



Ergodic theorem for discrete Markov chains

- For now, we assume (X_t) is a *discrete* Markov chain, that is, X_t is a discrete random variable for all t.
- The same intuitions apply in the continuous case, but the math is considerably more subtle.

- Let's see the theorem first, then define the terminology.

- *Ergodic theorem*: If (X_t) is a time-homogeneous, irreducible, discrete Markov chain with stationary distribution π, then for any bounded function $h(x)$,

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \to \mathrm{E}h(X)$$

as $T \to \infty$, with probability 1, where $X \sim \pi$. If, further, (X_t) is aperiodic, then for all x, x_0,

$$\mathbb{P}(X_t = x \mid X_0 = x_0) \to \pi(x).$$

Definitions (1/2)

- (X_t) is *time-homogeneous* if the distribution of $X_{t+1}|X_t$ is the same for all t, that is, for all a, b, t,

  $$\mathbb{P}(X_{t+1} = b \mid X_t = a) = T_{ab}$$

  for some matrix T that doesn't depend on t.

- T is called the *transition matrix*. Note that the rows of T sum to 1, that is, $\sum_b T_{ab} = 1$ for all a.

- π is a *stationary* (or *invariant*) *distribution* for T if for all b,

  $$\sum_a \pi(a) T_{ab} = \pi(b).$$

  This is often written more succinctly as $\pi T = \pi$, viewing π as a row vector.

Definitions (2/2)

- (X_t) is *irreducible* if for all a, b, there is some t such that

  $$\mathbb{P}(X_t = b \mid X_0 = a) > 0.$$

  (In other words, we can get from point a to point b with positive probability.)

- (X_t) is *aperiodic* if for all a,

  $$\gcd\bigl(\{t : \mathbb{P}(X_t = a \mid X_0 = a) > 0\}\bigr) = 1$$

  where gcd $=$ greatest common divisor. (In other words, the times at which we can return to a are not periodic.)

Some comments on the conditions

- Suppose π is our *target distribution*, that is, we want to generate samples from π.

- Metropolis–Hastings always yields a time-homogeneous Markov chain with stationary distribution π.
    - In fact, it satisfies a stronger condition called detailed balance.

- Thus, irreducibility is the main condition we need to check in practice. Fortunately, irreducibility usually holds in practice.
    - Note: If $T_{ab} > 0$ for all a, b, then the chain is irreducible.

- Aperiodicity is nice to have but is not strictly necessary to justify the use of sample averages. It usually holds anyways.
    - Note: If $T_{aa} > 0$ for some a, and the chain is irreducible, then the chain is aperiodic.

Stationarity

- The term “stationary distribution” comes from this fact:

  If $X_0 \sim \pi$ and (X_t) is a time-homogeneous Markov chain with stationary distribution π, then (X_t) is *stationary*, that is, for all k, the distribution of $(X_t, \ldots, X_{t+k})$ is the same for all t.

- Under the ergodic theorem, X_t converges in distribution to π (this is the 2nd part of the theorem, when aperiodicity holds).

- Informally, when the distribution of X_t is close to π, we say that the chain has “reached stationarity”.

- In MCMC, the burn-in period is the amount of time before the chain is sufficiently close to stationarity.

Detailed balance

- We say that *detailed balance* holds if for all a, b,

$$\pi(a)T_{ab} = \pi(b)T_{ba}.$$

- If detailed balance holds, then π is a stationary distribution for T, since

(Whiteboard exercise)

Detailed balance

- We say that *detailed balance* holds if for all a, b,

$$\pi(a)T_{ab} = \pi(b)T_{ba}.$$

- If detailed balance holds, then π is a stationary distribution for T, since

$$\sum_a \pi(a)T_{ab} = \sum_a \pi(b)T_{ba} = \pi(b)\sum_a T_{ba} = \pi(b).$$

- Interpretation: At stationarity, the probability mass moving from a to b equals the mass moving from b to a.

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

Negative-Binomial regression example

Metropolis–Hastings algorithm

- Nearly all MCMC algorithms are a special case of MH, including Gibbs sampling.

- Suppose the target distribution is $\pi(x)$. For each x', let $q(x|x')$ be a distribution over x (the *proposal distribution*).

- For all x, x', define the *acceptance ratio*

$$\alpha(x', x) = \frac{\pi(x)q(x'|x)}{\pi(x')q(x|x')}.$$

- *MH algorithm*: Initialize x_0, and for $t = 1, \ldots, T$,
  1. Sample $x \sim q(x|x_{t-1})$.
  2. Sample $u \sim \mathrm{Uniform}(0, 1)$.
  3. If $u < \alpha(x_{t-1}, x)$, then set $x_t = x$, otherwise set $x_t = x_{t-1}$.

Metropolis–Hastings algorithm

- Steps 2 and 3 can equivalently be written: With probability $\min\{1,\, \alpha(x_{t-1}, x)\}$, set $x_t = x$, otherwise set $x_t = x_{t-1}$.

- Thus, in short, we propose $x \sim q(x|x_{t-1})$ and accept the proposal with probability $\min\{1,\, \alpha(x_{t-1}, x)\}$.

- The MH algorithm defines a Markov chain with transition matrix T, where

$$T_{ab} = q(b|a) \min\left\{1,\, \frac{\pi(b)q(a|b)}{\pi(a)q(b|a)}\right\}$$

when $a \neq b$, and for all a,

$$T_{aa} = 1 - \sum_{b \neq a} T_{ab}.$$

Metropolis–Hastings: Verifying detailed balance

- Assume $\pi(a) > 0$ and $q(b|a) > 0$ for all a, b.
- We verify that detailed balance holds.
- First, if $a = b$ then it is trivial: (Whiteboard exercise)
- Meanwhile, if $a \neq b$, then

(Whiteboard exercise)

Metropolis–Hastings: Verifying detailed balance

- Assume $\pi(a) > 0$ and $q(b|a) > 0$ for all a, b.

- We verify that detailed balance holds.

- First, if $a = b$ then it is trivial: $\pi(a)T_{aa} = \pi(a)T_{aa}$.

- Meanwhile, if $a \neq b$, then

$$\begin{aligned}\pi(a)T_{ab} &= \pi(a)q(b|a)\min\left\{1,\ \frac{\pi(b)q(a|b)}{\pi(a)q(b|a)}\right\} \\ &= \min\Big\{\pi(a)q(b|a),\ \pi(b)q(a|b)\Big\} \\ &= \pi(b)q(a|b)\min\left\{\frac{\pi(a)q(b|a)}{\pi(b)q(a|b)},\ 1\right\} \\ &= \pi(b)T_{ba}.\end{aligned}$$

Metropolis–Hastings: Intuition

- Students are often mystified by the acceptance probability
$$\min\left\{1, \frac{\pi(b)q(a|b)}{\pi(a)q(b|a)}\right\}.$$
- To understand it, consider an analogy:
  - $\pi(a)$ = amount of money belonging to person a.
  - $q(b|a)$ = fraction of a's money proposed to be transferred to b.
  - $\pi(a)q(b|a)$ = amount proposed to be transferred from a to b.
- We want equal amounts to be transferred between each pair. So, a modification factor is applied to the proposed amounts.
- If $\pi(a)q(b|a) > \pi(b)q(a|b)$ then a would give too much to b.
  - To make it equal, a gives only $\frac{\pi(b)q(a|b)}{\pi(a)q(b|a)}$ times the proposed amount, and keeps the rest.
  - In the reverse direction, b gives her full proposed amount to a.
- This modification factor is precisely the acceptance probability.

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/QZXUx2wx3QxK7P7z8

(Three people per room, randomly assigned. 15 minutes.)

Gibbs sampling is a special case of MH

- Let $\pi(x, y)$ be the target distribution. At iteration $t+1$, a Gibbs update to x can be viewed as sampling from

$$q(x, y|x_t, y_t) := \pi(x|y_t)\,\mathrm{I}(y = y_t).$$

- Now, suppose we do MH with q as the proposal distribution.

- With probability 1, $y = y_t$ when sampling from q, so

$$\begin{aligned}
\alpha((x_t, y_t), (x, y)) &= \frac{\pi(x, y) q(x_t, y_t|x, y)}{\pi(x_t, y_t) q(x, y|x_t, y_t)} \\
&= \frac{\pi(x, y)\pi(x_t|y)\mathrm{I}(y_t = y)}{\pi(x_t, y_t)\pi(x|y_t)\mathrm{I}(y = y_t)} \\
&= \frac{\pi(x, y)\pi(x_t|y)}{\pi(x_t, y)\pi(x|y)} \\
&= 1.
\end{aligned}$$

- Thus, we always accept, so MH reduces to Gibbs in this case.

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

Negative-Binomial regression example

Combining MCMC moves

- One of the many nice things about MCMC is that it is easy to combine various moves when constructing a sampler.

- For instance, we can combine various Gibbs updates or MH moves with different proposal distributions.

- Suppose the target distribution is π. Roughly, a move is a way of updating the variables using an MCMC step targeting π.

- Formally, we define a *move* to be a transition matrix T such that $\pi T = \pi$, that is, π is the stationary distribution of T.

- Two useful ways of combining moves $T_1, \ldots, T_k$ are:
  1. products of moves, and
  2. mixtures of moves.

Products of moves (Deterministic cycle of moves)

- If $T_1, \dots, T_k$ all have stationary distribution π, then the product $T = T_1 \cdots T_k$ has stationary distribution π.

- This is easy to check:

$$\pi T_1 T_2 \cdots T_k = \pi T_2 \cdots T_k = \cdots = \pi T_k = \pi.$$

- This is used in *fixed-scan Gibbs*, where we update the variables by cycling through them in a deterministically chosen order.

- This is also used in *MH-within-Gibbs*, where MH moves on the full conditionals are used in place of some Gibbs updates.

- Note: We do NOT explicitly compute T! All we have to do is apply a sequence of moves.

Mixtures of moves (Random choice of move)

- If $T_1, \ldots, T_k$ all have stationary distribution π, and $w_1, w_2, \ldots, w_k \geq 0$ with $\sum_{i=1}^k w_i = 1$, then the mixture $T = \sum_{i=1}^k w_i T_i$ has stationary distribution π.

- This is also easy to check:

$$\pi T = \sum_{i=1}^k w_i \pi T_i = \sum_{i=1}^k w_i \pi = \pi \sum_{i=1}^k w_i = \pi.$$

- This is used in *random-scan Gibbs*, where we randomly choose which variable to update at each step. Here, w_i is the probability of updating variable i at a given step.

- Note: We do NOT explicitly compute T! All we have to do is randomly choose a move, and apply that move.

Careful! State-dependent moves are typically invalid

- It is important to note that the random choice of move does not depend on the current state.
- In general, the choice of move at each iteration should not depend on the current state of the Markov chain.
- Using a state-dependent move can result in a failure to converge to the correct stationary distribution.
- Note: The fact that the proposal distribution in MH depends on the current state does not violate this principle.
- You need to be very, very careful if you want to try to use state-dependent moves.

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

Negative-Binomial regression example

MCMC rate of convergence

- In basic Monte Carlo, we know $\mathbb{V}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big) = \mathbb{V}(X_t)/T$ since the X_t's are i.i.d. In MCMC, the X_t's are no longer i.i.d., so it is not as easy to assess the rate of convergence.

- However, for any sequence of random variables, we have:

$$\mathbb{V}\Big(\frac{1}{T}\sum_{t=1}^{T} X_t\Big) = \text{(Whiteboard exercise)}$$

MCMC rate of convergence

- In basic Monte Carlo, we know $\mathbb{V}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big) = \mathbb{V}(X_t)/T$ since the X_t's are i.i.d. In MCMC, the X_t's are no longer i.i.d., so it is not as easy to assess the rate of convergence.

- However, for any sequence of random variables, we have:

$$\begin{aligned}\mathbb{V}\Big(\frac{1}{T}\sum_{t=1}^{T} X_t\Big) &= \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathrm{Cov}(X_s, X_t) \\ &= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}(X_t) + \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s}\mathrm{Cov}(X_s, X_t).\end{aligned}$$

- If the X_t's have the same distribution, then the first term equals $\mathbb{V}(X_t)/T$, just like the basic Monte Carlo variance.

- Thus, in MCMC, the approximation error will be small if the $\mathrm{Cov}(X_s, X_t)$ terms are small.

Effective sample size

- Assume (X_t) is stationary. Then $C(\delta) := \mathrm{Cov}(X_t, X_{t+\delta})$ does not depend on t, and $C(\delta) = C(-\delta)$.

- Define $\sigma^2 := \mathbb{V}(X_t) = C(0)$ and $\rho(\delta) := \mathrm{Corr}(X_t, X_{t+\delta})$. Then $C(\delta) = \sigma^2\rho(\delta)$ and

$$\begin{aligned}\mathbb{V}\Big(\frac{1}{T}\sum_{t=1}^{T} X_t\Big) &= \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T} C(s-t) \\ &= \frac{1}{T}C(0) + \frac{2}{T^2}\sum_{\delta=1}^{T-1}(T-\delta)C(\delta) \\ &= \frac{\sigma^2}{T} + \frac{2\sigma^2}{T}\sum_{\delta=1}^{T-1}(1-\delta/T)\rho(\delta) = \frac{\sigma^2}{T_{\text{eff}}}\end{aligned}$$

where

$$T_{\text{eff}} := \frac{T}{1 + 2\sum_{\delta=1}^{T-1}(1-\delta/T)\rho(\delta)}.$$

Effective sample size

- T_{eff} is one way of defining an "effective sample size" (ESS).

- Since $\mathbb{V}\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) = \sigma^2/T_{\text{eff}}$, the idea is that T_{eff} is the number of i.i.d. samples that basic Monte Carlo would require to achieve the same approximation error.

- To compute T_{eff} in practice, one would estimate $\rho(\delta) \approx \widehat{\rho}(\delta)$ using the samples themselves, and use

  $$\widehat{T}_{\text{eff}} := \frac{T}{1 + 2\sum_{\delta=1}^{T-1}(1 - \delta/T)\widehat{\rho}(\delta)},$$

  possibly truncating the sum since $\rho(\delta)$ is harder to estimate for larger δ values.

- If T_{eff} is much smaller than T, then the MCMC sampler is struggling.

Effective sample size

- The formula above is not actually the usual definition of *effective sample size*, which is:

$$T'_{\text{eff}} := \frac{T}{1 + 2\sum_{\delta=1}^{\infty} \rho(\delta)}.$$

- In practice, the infinite sum is approximated by truncating it at an appropriate value.

- The interpretation is the same, but T_{eff} may be more stable than T'_{eff} since it doesn't depend as much on larger δ values, for which $\rho(\delta)$ is harder to estimate in a finite sample.

- It's also worth mentioning that $\tau := 1 + 2\sum_{\delta=1}^{\infty} \rho(\delta)$ is called the *autocorrelation time*.

Outline

Markov chains

Metropolis–Hastings

Combining MCMC moves

MCMC rate of convergence

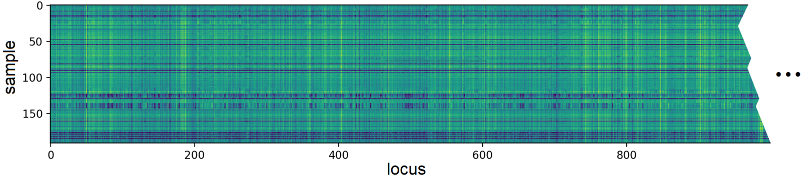
Negative-Binomial regression example

NegBin regression example: Background

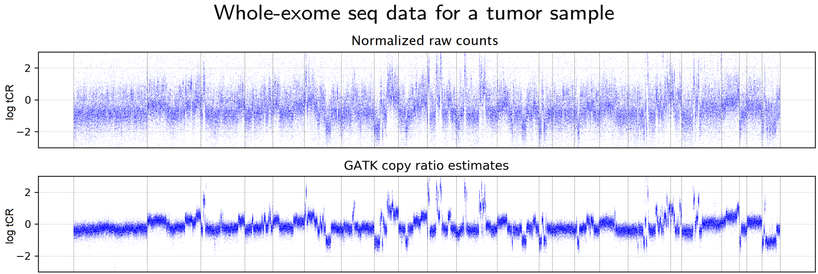
Modern high-throughput sequencing yields large matrices of counts.

- Copy ratio estimation in cancer genomics
  - whole-exome or whole-genome sequencing data
- Copy number variation in genetics
  - whole-exome or whole-genome sequencing data
- Gene expression analysis in biology/medicine
  - RNA-seq data for transcript abundance

log counts for a whole-exome seq data set of 191 samples × 171523 loci



NegBin regression example: Background

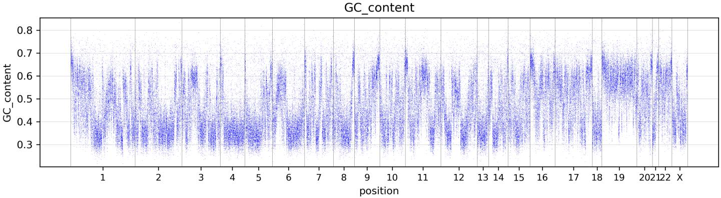


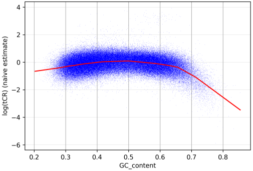
NegBin regression example: Background

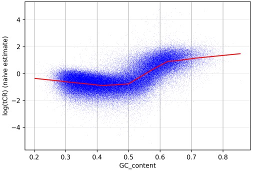
Nontrivial modeling required to deal with many sources of bias:

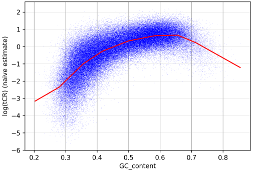
- GC bias
- Mappability bias
  - Repetitive sequences, Tandemly arrayed genes
- Epigenetics
  - Open chromatin, Promoters, Enhancers, etc.
- Fragment length bias
- Batch effects

NegBin regression example: Background









NegBin regression example: Model

- The technical variability in this data is naturally modeled as Poisson, by the “law of small numbers”.
    - Each count is the sum of many Bernoullis with small probability.
- However, there also outliers. A Negative-Binomial model is often used to improve robustness to outliers.
- The Negative-Binomial is an overdispersed Poisson — specifically, a Poisson with a Gamma prior integrated out for each observation.

NegBin regression example: Model

- Suppose the count data are $Y_{ij} \in \{0, 1, 2, \ldots\}$ for loci $i = 1, \ldots, I$ and samples $j = 1, \ldots, J$.
- Let's consider the following model:

$$Y_{ij} \sim \mathrm{NegBin}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i)$$

where

$$\log(\mu_{ij}) = a_i + b_j + c_1 x_i + c_2 x_i^2.$$

- Interpretation:
  - a_i = locus-specific effect
  - b_j = sample-specific effect
  - x_i = locus covariate such as GC content
  - c_1, c_2 = coefficients of linear and quadratic terms
- Assume x_i is standardized to mean zero, unit variance.

NegBin regression example: Identifiability

- This model is not identifiable since an additive constant can be moved between a_i and b_j.
- This non-identifiability can be removed by constraining, say, $\sum_i a_i = 0$.
- However, posterior inference is complicated when constraints are imposed.
- Simple alternative: Run MCMC in the unconstrained (non-identifiable) model, and when MCMC sampling is complete, impose the identifiability constraints on the posterior samples for interpretation purposes.

NegBin regression example: Simulation

- To illustrate, I simulated data from the model using $I = 100$ and $J = 10$, with true parameters generated as $a_i \sim \mathcal{N}(0, 1)$, $b_j \sim \mathcal{N}(5, 1)$, $c_1 = 0$, $c_2 = -1$, and $\alpha_i = 1$ for all i.

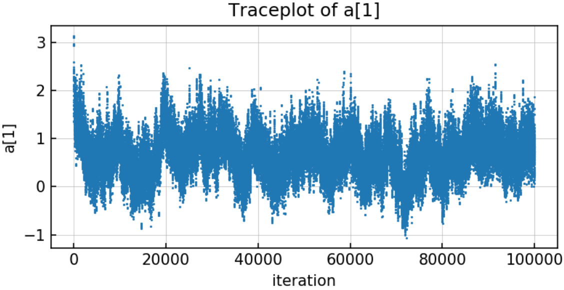
- For simplicity, I assumed $\mathcal{N}(0, 5^2)$ priors on a_i, b_j, and c_k, and fixed $\alpha_i = 1$.

- To perform MCMC, I used an MH-within-Gibbs approach, updating each univariate parameter a_i, b_j, c_k individually.

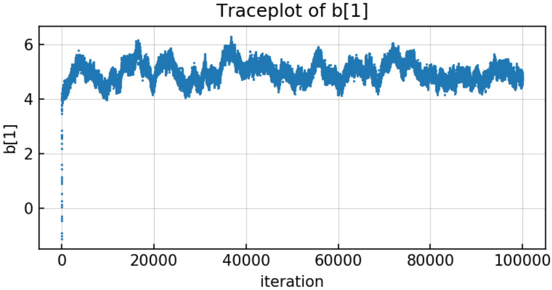
- E.g., MH with proposal $a_i \sim \mathcal{N}(a_{i,t-1}, 0.25^2)$ and target distribution equal to the full conditional for a_i.

- I ran MCMC for 100,000 sweeps, with a burn-in of 20,000. (We'll look at diagnostics below to see if these were good choices.)

NegBin regression example: Traceplots

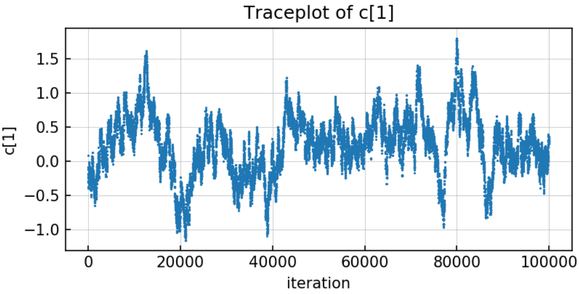


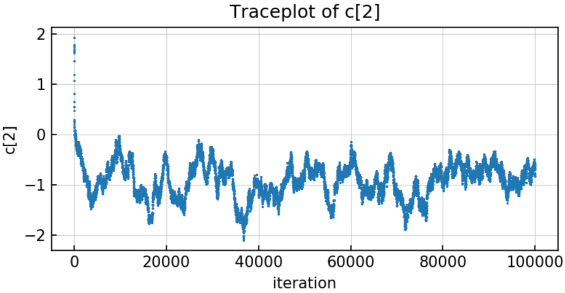
Traceplot of a[1]



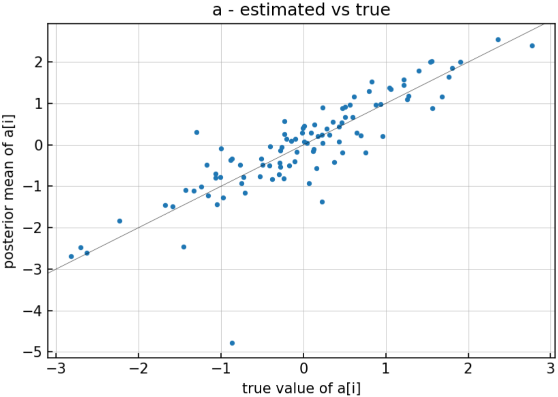
Traceplot of b[1]

NegBin regression example: Traceplots

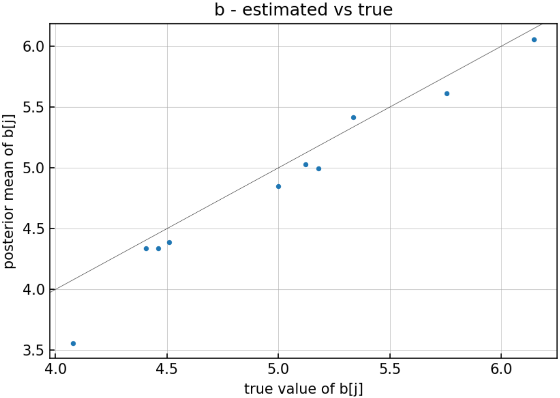

Traceplot of c[1]


Traceplot of c[2]

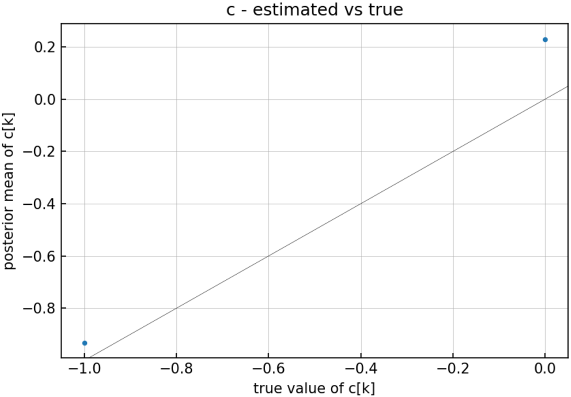
NegBin regression example: Estimated vs true



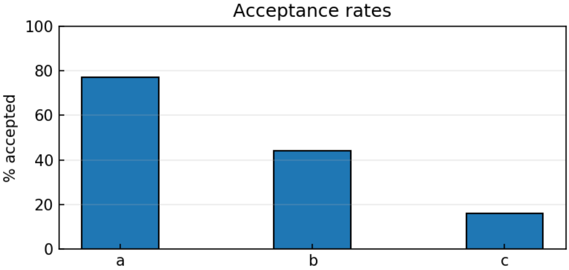
NegBin regression example: Estimated vs true



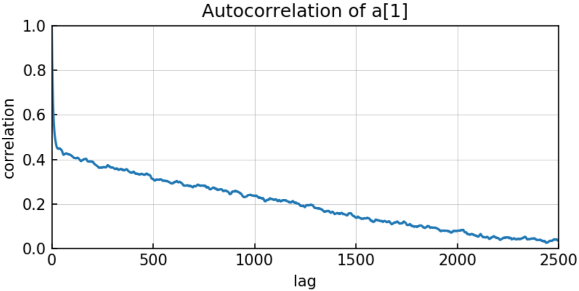
NegBin regression example: Estimated vs true

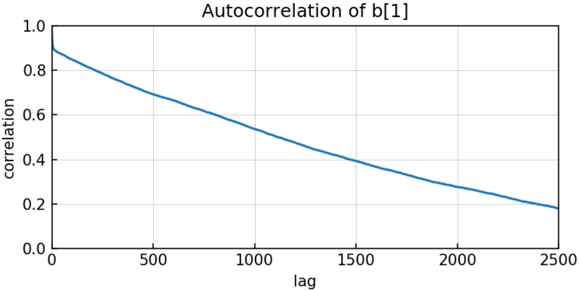


NegBin regression example: Acceptance rate

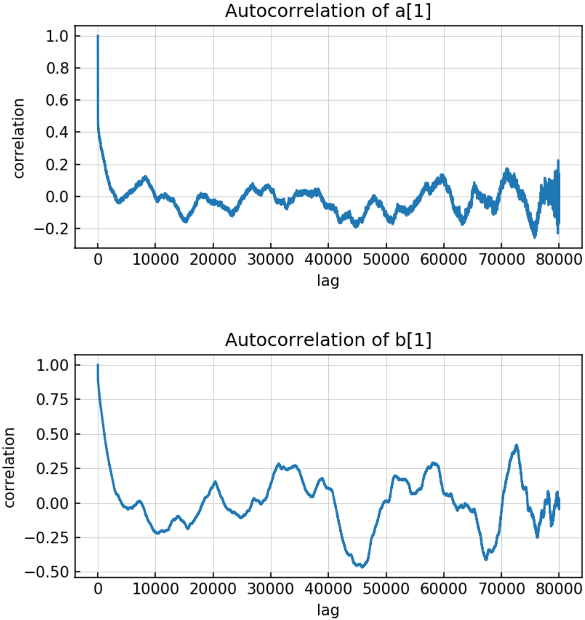


NegBin regression example: Autocorrelation





NegBin regression example: Autocorrelation



References and supplements

- R.E. Kass, B.P. Carlin, A. Gelman, and R.M. Neal (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. The American Statistician, 52(2), pp. 93-100.
- A.E. Gelfand and A.F. Smith (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85(410), 398-409.

Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/r7YzCQFMbP1yQupj7