STA2311: Advanced Computational Methods for Statistics I

Class 8: MCMC Tuning and Diagnostics

Radu Craiu   Robert Zimmerman

University of Toronto

November 7, 2023

1. Introduction

2. AR(1) Processes

3. Variance Calculations

4. Convergence Analysis

5. Code Analysis

Section 1

Introduction

Gibbs Sampling

- Gibbs sampling is a popular MCMC algorithm for sampling from a complex probability distribution π
- One essential property in reversible MCMC is detailed balance
- The systematic scan Gibbs sampler does *not* satisfy the detailed balance condition
    - Recall that detailed balance condition means that reversibility holds
    - Reversibility essentially means that the distributions of $(X_t, X_{t+1}, \ldots, X_{t+s})$ and $(X_{t+s}, X_{t+s-1}, \ldots, X_t)$ are the same

Detailed Balance Condition

- The detailed balance condition states that

$$\pi(x) \cdot Q(x \to x') = \pi(x') \cdot Q(x' \to x)$$

  where π is the target distribution and Q is the proposal distribution

- Detailed balance simplifies the conditions for a CLT for $\hat{I}$, where $I = \int h(x)\pi(x)\,\mathrm{d}x$
- The CLT says that $\sqrt{M}(I_m - I) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2)$
- This is the same as we would have in the classical iid setup

Gibbs Sampler

- In Gibbs sampling, we update one variable at a time while keeping the others fixed
- At each step, a single variable is sampled from its conditional distribution
- The choice of the next variable to update is deterministic
- The update rule for the i'th component is

$$x_i^{(t+1)} \sim \pi(x_i \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_d^{(t)})$$

Systematic Scan Gibbs Sampling Does Not Have Detailed Balance

- The random scan Gibbs sampler satisfies detailed balance
- However, the systematic scan Gibbs sampler does *not*
- We show this for a 2-component Gibbs sampler
- The deterministic order in which the variables are updated is central to this result

Systematic Scan Gibbs Sampling Does Not Have Detailed Balance (Continued)

- At iteration t, the systematic scan Gibbs sampler samples $X_{t+1} \sim \pi(\cdot \mid Y_t)$ and $Y_{t+1} \sim \pi(\cdot \mid X_{t+1})$
- We have $K(x_{t+1}, y_{t+1} \mid x_t, y_t) = \pi(x_{t+1} \mid y_t) \cdot \pi(y_{t+1} \mid x_t)$ and similarly, $K(x_t, y_t \mid x_{t+1}, y_{t+1}) = \pi(x_t \mid y_{t+1}) \cdot \pi(y_t \mid x_t)$
- But

$$\pi(x_t, y_t) \cdot \pi(x_{t+1} \mid y_t) \cdot \pi(y_{t+1} \mid x_t) \neq \pi(x_{t+1}, y_{t+1}) \cdot \pi(x_t \mid y_{t+1}) \cdot \pi(y_t \mid x_t)$$

- Thus

$$\pi(x_t, y_t) \cdot K(x_{t+1}, y_{t+1} \mid x_t, y_t) \neq \pi(x_{t+1}, y_{t+1}) \cdot K(x_t, y_t \mid x_{t+1}, y_{t+1})$$

- So detailed balance fails

Random Scan Gibbs Sampling Does Have Detailed Balance

- However, suppose we implement a random scan Gibbs sampler, in which at each step we update X with probability $1/2$ and Y with probability $1/2$
- Then

$$
\begin{aligned}
&K(x_{t+1}, y_{t+1} \mid x_t, y_t) \\
&= \frac{1}{2}\pi(x_{t+1} \mid y_t) \cdot \pi(y_{t+1} \mid x_{t+1}) + \frac{1}{2}\pi(y_{t+1} \mid x_t) \cdot \pi(x_{t+1} \mid y_{t+1})
\end{aligned}
$$

and

$$
\begin{aligned}
&K(x_t, y_t \mid x_{t+1}, y_{t+1}) \\
&= \frac{1}{2}\pi(x_t \mid y_{t+1}) \cdot \pi(y_t \mid x_t) + \frac{1}{2}\pi(y_t \mid x_{t+1}) \cdot \pi(x_t \mid y_t)
\end{aligned}
$$

Random Scan Gibbs Sampling Does Have Detailed Balance (Continued)

- With a tedious calculation, one can then check that

$$\pi(x_t, y_t)\left[\frac{1}{2}\pi(x_{t+1} \mid y_t)\cdot\pi(y_{t+1} \mid x_{t+1}) + \frac{1}{2}\pi(y_{t+1} \mid x_t)\cdot\pi(x_{t+1} \mid y_{t+1})\right]$$

  is equal to

$$\pi(x_{t+1}, y_{t+1})\left[\frac{1}{2}\pi(x_t \mid y_{t+1})\cdot\pi(y_t \mid x_t) + \frac{1}{2}\pi(y_t \mid x_{t+1})\cdot\pi(x_t \mid y_t)\right]$$

- So detailed balance is satisfied

Distributions of Subchains

- Let $\boldsymbol{X} \in \mathbb{R}^d$ be a Markov chain with stationary distribution π
- If $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M)}$ are M MCMC samples from π, then $\boldsymbol{X}_j^{(1)}, \ldots, \boldsymbol{X}_j^{(M)}$ are MCMC samples from the j'th marginal distribution $\pi_j(x_j) = \int \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{-j}$
- We demonstrate this for the 2-component Gibbs sampler
- Suppose (x, y) is the current value of the chain and (x', y') is the next value, $K(x', y' \mid x, y) = \pi(y' \mid x) \cdot \pi(x' \mid y')$
- We want to show that $\pi(x)$ is stationary for the x-component; i.e.,

$$\int K(x' \mid x) \cdot \pi(x) \, \mathrm{d}x = \pi(x')$$

Distributions of Subchains (Continued)

- Indeed, since $K(x' \mid x) = \int \pi(y' \mid x) \cdot \pi(x' \mid y') \, \mathrm{d}y'$, we get that

$$\begin{aligned} \int K(x' \mid x) \cdot \pi(x) \, \mathrm{d}x &= \iint \pi(y' \mid x) \cdot \pi(x' \mid y') \, \mathrm{d}y' \cdot \pi(x) \, \mathrm{d}x \\ &= \iint \pi(y', x) \cdot \pi(x' midy') \, \mathrm{d}x \, \mathrm{d}y' \\ &= \int \pi(y') \cdot \pi(x' midy') \, \mathrm{d}y' \\ &= \int \pi(x', y') \, \mathrm{d}y' \\ &= \pi(x') \end{aligned}$$

- The same result holds for the Metropolis-Hastings algorithm (exercise!)

Aside: Functions of Markov Chains are Not Markov Chains

- If (X_n) is a Markov chain and f is some function, does it follow that $(f(X_n))$ is also a Markov chain?
- Not if f is not injective!
- For a counterexample, consider $\mathcal{X} = \{x_1, x_2, x_3\}$ and suppose a Markov chain on $\mathcal{X}$ has initial distribution $\delta(x) = \frac{1}{3}$ and transition kernel satisfying $K(x_1 \mid x_3) = K(x_3 \mid x_1) = 1$ and $K(x \mid x) = 1$ for $x \in \mathcal{X}$
- Now, for any $y \neq z$, let $f(x_1) = f(x_2) = y$ and $f(x_3) = z$, and define the process $Y_n = f(X_n)$ with state-space $\mathcal{Y} = \{y, z\}$
- Then, since $Y_2 = z \Leftrightarrow X_2 = x_3 \Leftrightarrow X_1 = x_1 \Leftrightarrow X_0 = x_3$, Bayes rule gives

$$\mathbb{P}(Y_2 = z \mid Y_1 = y) = \frac{1}{2} \neq 1 = \mathbb{P}(Y_2 = z \mid Y_1 = y, Y_0 = z),$$

so (Y_n) does not satisfy the Markov property

Section 2

AR(1) Processes

AR(1) Processes

- An AR(1) process is defined as

$$X_t = \phi X_{t-1} + \epsilon_t,$$

  where X_t is the value of the process at time t, ϕ is the autoregressive coefficient, and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is white noise

- The process depends on the previous value X_{t-1} and a random disturbance ϵ_t
- Since $X_t \mid X_{t-1} \sim \mathcal{N}(\phi X_{t-1}, \sigma^2)$, the process (X_t) is a Markov chain

AR(1) Stationary Distribution

- For the AR(1) process to have a stationary distribution, $|\phi| < 1$ is a necessary condition
- When $|\phi| < 1$, the process converges to a stationary distribution as $t \to \infty$
- The stationary distribution is Gaussian, and its mean and variance can be determined
- Put $X_0 = \epsilon_0$ (i.e., X_0 is drawn from the noise population) and $X_1 = \phi X_0 + \epsilon_1$ with $\epsilon_1 \sim \mathcal{N}(\mu_0, \sigma^2)$

AR(1) Stationary Distribution (Continued)

- Then

  $$\mathbb{E}[X_1] = \mathbb{E}\left[\mathbb{E}[X_1 \mid X_0]\right] = \mathbb{E}[\phi X_0 + \mu_0] = \phi\mu_0 + \mu_0 = \mu_0(1+\phi)$$

  and

  $$\mathrm{Var}(X_1) = \phi^2 \mathrm{Var}(X_0) + \sigma^2 = \sigma^2(1+\phi^2)$$

- Similarly,

  $$\mathbb{E}[X_2] = \mathbb{E}\left[\mathbb{E}[\phi X_1 + \epsilon_1 \mid X_1]\right] = \mathbb{E}[\phi X_1 + \mu_0] = \mu_0(1+\phi+\phi^2)$$

  and

  $$\mathrm{Var}(X_2) = \phi^2 \mathrm{Var}(X_1) + \mathrm{Var}(\epsilon_1) = \sigma^2(1+\phi^2+\phi^4)$$

- Proceed by inductions and take limits to get $\mathbb{E}[X_n] = \mu_0(1+\phi+\cdots+\phi^n) \to \frac{\mu_0}{1-\phi}$ and $\mathrm{Var}(X_n) = \sigma^2(1+\phi^2+\cdots+\phi^{2n}) \to \frac{\sigma^2}{1-\phi^2}$

Section 3

Variance Calculations

Variance for an MCMC Algorithm

- We want $\mathrm{Var}\left(\frac{1}{M}\sum_{i=1}^{M} h(X_i)\right) = \mathrm{Var}(\hat{I}_M)$
- If $X_i \sim \pi$ (under the stationary regime), then the conditions for an MCMC central limit theorem are satisfied:
- A sufficient condition is *geometric ergodicity*, but for complex samplers, we often do not know that this holds
- The general CLT says

$$\sqrt{M}(\hat{I}_M - I) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2)$$

- If we are under the classical Monte Carlo setup and $X_i \overset{iid}{\sim} \pi$, then $\sigma_h^2 = \mathrm{Var}_\pi(h(X))$

AR(1) Variance

- In the AR(1) model, what is the correlation between X_{t+s} and X_s?

$$\begin{aligned}\mathrm{Cov}(X_{t+s}, X_s) &= \mathrm{Cov}(\phi X_{t+s-1} + \epsilon_{t+s}, \phi X_{s-1} + \epsilon_s)\\ &= \phi^2 \mathrm{Cov}(X_{t+s-1}, X_{s-1})\\ &= \phi^{2s}\mathrm{Cov}(X_t, X_0)\end{aligned}$$

- Using the asymptotic variance for X_{t+s} and X_s, we get

$$\mathrm{Corr}(X_{t+s}, X_s) = \frac{\phi^{2s}\mathrm{Cov}(X_t, X_0)}{\sigma^2/(1-\phi^2)} = \frac{\phi^{2s}}{\sqrt{1-\phi^2}}\mathrm{Corr}(X_t, X_0)$$

- Then

$$\sum_{t>0}\sum_{s\geq 0}\mathrm{Corr}(X_{t+s}, X_s) = \sum_{t>0}\mathrm{Corr}(X_t, X_0)\sum_{s\geq 0}\phi^{2s} = \frac{1}{(1-\phi^2)^{3/2}}\sum_{t>0}\rho_t$$

where $\rho_t = \mathrm{Corr}(X_t, X_0)$

Why Estimate MCMC Variance?

- MCMC estimates often have autocorrelation, which affects the effective sample size
- Accurate variance estimation is crucial for hypothesis testing and interval estimation
- Geyer's estimate provides an efficient way to estimate the variance of MCMC estimates

Autocorrelation in MCMC Chains

- Autocorrelation refers to the correlation between a variable and its lagged values in a time series
- MCMC chains often exhibit high autocorrelation, which reduces the effective sample size and increases uncertainty in parameter estimates

Geyer's Estimate of Variance

- CLT: $\sqrt{M}(\hat{I}_M - I) \to \mathcal{N}(0, \sigma_h^2)$, where $\hat{I}_M = \frac{1}{M}\sum_{i=1}^{M} h(X_i)$
- Geyer's estimation of variance relies on the autocorrelation time τ of the chain
- The formula for Geyer's estimate is given by

$$\sigma_h^2 = \frac{\eta_h^2}{M}\left[1 + \frac{2}{M}\sum_{k=1}^{M}(M-k)\rho_k\right]$$

  where M is the number of iterations, ρ_t is the autocorrelation at lag t, and $\eta_h^2 = \mathrm{Var}_\pi(h(X))$

- If the ρ_t are large and don't decay fast, then trouble!
- In general the sum ends at $\tau << M$, so

$$\sigma_h^2 = \frac{\eta_h^2}{M}\left[1 + 2\sum_{k=1}^{\tau}\rho_k\right]$$

Geyer's Estimate of Variance (Continued)

- Assuming stationarity and time-homogeneity (in the sense that $\mathrm{Corr}(h(X_i), h(X_j)) = \mathrm{Corr}(h(X_0), h(X_{j-i})) =: \rho_{j-i}$), then

$$\begin{aligned}\mathrm{Var}(\hat{I}_M) &= \frac{1}{M}\mathrm{Var}(h(X))\left[1 + \frac{2}{M}\sum_{k=1}^{M}\sum_{i=1}^{M-k}\mathrm{Corr}(h(X_i), h(X_{i+k}))\right] \\ &= \frac{\eta_h^2}{M}\left[1 + 2\sum_{k=1}^{M}\frac{M-k}{M}\mathrm{Corr}(h(X_0), h(X_k))\right] = \\ &= \frac{\eta_h^2}{M}\left(1 + 2\sum_{k=1}^{\tau}\rho_k\right) = \sigma_h^2\end{aligned}$$

- Note that the classical Monte Carlo variance $\frac{\eta_h^2}{M}$ is inflated due to dependence within the samples

Batch Means

- In practice, let L be the maximum t for which $\rho_t > 0.1$, then plug in the estimators for $\rho_1, \ldots, \rho_L$
- But we still need an estimate of η_h^2
- Geyer's estimate is based on the idea of "batch means"
- It divides the MCMC chain into m non-overlapping batches of size b and computes the batch means
- The variance is then computed from the variances of these batch means
- The goal is to assess the variability between batches rather than within each batch

Batch Variance Estimation

- For the ith batch, compute the sample mean, denoted as $\hat{\mu}_i$
- The overall sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ are computed from averages of the batch means:

$$\hat{\mu} = \frac{\sum_{i=1}^{m} \hat{\mu}_i}{m}, \text{ where } \hat{\mu}_i = \frac{1}{b} \sum_{\{j: X_j \in \text{batch i}\}} h(X_j)$$

$$\frac{\hat{\sigma}^2}{b} = \frac{1}{m} \sum_{i=1}^{m} (\hat{\mu}_i - \hat{\mu})^2$$

where k is the number of batches and n_i is the number of samples in each batch

- $\hat{\sigma}^2 = \frac{b}{m} \sum_{i=1}^{m} (\hat{\mu}_i - \hat{\mu})^2$ is Geyer's estimate

Applications and Considerations

- Non-overlapping batch sampling is commonly used for estimating the variance of MCMC estimates, especially in the context of Bayesian analysis
- Choosing an optimal batch size is a trade-off between reducing the variance of the estimator and increasing the bias
- Larger batch sizes tend to yield more precise variance estimates but may introduce bias

Effective Sample Size

- The *effective sample size (ESS)* measures the effective number of independent samples in an MCMC chain, accounting for autocorrelation
- It can be calculated as

$$\text{ESS} = \frac{M}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

  where ρ_k is the autocorrelation at lag k

- A classical Monte Carlo sample of size ESS provides the same variance as the MCMC sample of size M

Section 4

Convergence Analysis

Visual Inspection

- Visual inspection of trace plots is an initial step in convergence assessment
- Trace plots show the sampled values of parameters over time
- A converged chain should exhibit stationary behavior with no significant trends or oscillations

Gelman-Rubin Diagnostic

- The Gelman-Rubin (GR) diagnostic, also known as $\hat{R}$, compares the variance between chains to within chains
- We run m chains in parallel, each for n iterations; write $(X_i^{(k)})$ for the kth chain, $\bar{X}_\cdot^{(k)}$ for the average of the kth chain, and $\bar{X}_\cdot^{(\cdot)}$ for the overall average (averaging over chains and realizations)
- To compute the GR diagnostic, let

$$W = \frac{1}{(n-1)m} \sum_{k=1}^{m} \sum_{i=1}^{n} (X_i^{(k)} - \bar{X}_\cdot^{(k)})^2$$

and

$$B = \frac{n}{m-1} \sum_{k=1}^{m} (\bar{X}_\cdot^{(k)} - \bar{X}_\cdot^{(\cdot)})^2$$

Gelman-Rubin Diagnostic (Continued)

- When the chains behave well, we have $\frac{n-1}{n}W + \frac{1}{n}B \approx W$
- The GR diagnostic is given by

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

- Clearly, as $n \to \infty$ and $B \xrightarrow{d} 0$, we have $R \xrightarrow{d} 1$
- Gelman and Rubin suggest that stationarity has been reached when $R \leq 1.1$

Geweke Diagnostic

- The Geweke convergence diagnostic is based on the idea of comparing means of two segments of the MCMC chain
- It calculates a Z-score, comparing the mean of the early portion of the chain to the mean of the late portion
- The Z-score is then assessed to determine whether the chain has likely reached convergence

Calculating the Z-Score

- The Geweke Z-score is computed as:

$$Z = \frac{\bar{\theta}_a - \bar{\theta}_b}{\sqrt{S_a^2 + S_b^2}}$$

- Here, $\bar{\theta}_a$ and $\bar{\theta}_b$ are the means of the early and late portions of the chain, and S_a and S_b are their corresponding standard errors
- A large magnitude Z-score suggests non-convergence

Section 5

Code Analysis

Bimodal Distribution

- We consider first the bivariate distribution

$$\pi(x_1, x_2) \propto \exp\left(-\frac{1}{2}(Ax_1^2x_2^2 + x_1^2 + x_2^2 - 2Bx_1x2 - 2C_1x_1 - 2C_2x_2)\right),$$

with $A = 8, B = 2, C_1 = 4, C_2 = 4$

- The conditional distributions can be easily determined to be Gaussian:

$$\pi(x_1|x_2) = \mathcal{N}((Bx_2 + C_1)/(Ax_2^2 + 1), 1/(Ax_2^2 + 1))$$

$$\pi(x_2|x_1) = \mathcal{N}((Bx_2 + C_2)/(Ax_2^2 + 1), 1/(Ax_2^2 + 1))$$

- Incidentally, this is one example where the marginals are Gaussian but the joint distribution is not Gaussian

Logistic Regression with Random Intercept

- Consider K clusters, each with N data points.

$$\begin{aligned}
Y_{ij}|u_i, X_{ij}, \beta &\sim \text{Bernoulli}(p_{ij}), \text{ where} \\
logit(p_{ij}) &= \frac{p_{ij}}{1-p_{ij}} = u_i + \beta_0 + \beta_1 X_{ij}, \;\; 1 \leq i \leq K, \; 1 \leq j \leq N \\
u_i &\sim \mathcal{N}(0, \eta^2), \;\; 1 \leq i \leq K \\
\beta &\sim \mathcal{N}_2(0, \sigma_\beta^2 I_2) \\
\eta &\sim \text{Gamma}(a, b)
\end{aligned}$$

Logistic Regression with Random Intercept (Continued)

- Another parametrization:

$$\begin{aligned} Y_{ij}|u_i, X_{ij}, \beta &\sim \text{Bernoulli}(p_{ij}), \text{ where} \\ logit(p_{ij}) &= \frac{p_{ij}}{1-p_{ij}} = R_i + \beta_1 X_{ij}, \;\; 1 \le i \le K, \; 1 \le j \le N \\ R_i|\beta_0 &\sim N(\beta_0, \eta^2), \;\; 1 \le i \le K \\ \beta &\sim \mathcal{N}_2(0, \sigma_\beta^2 I_2) \\ \eta &\sim \text{Gamma}(a, b) \end{aligned}$$

References I