MCMC

Péter Antal

Department of Measurement and Information Systems

IDA, 2017

Bayesian inference with Monte Carlo I.

Integration/summation is a central operation in Bayesian statistics (c.f. optimization in the frequentist approach)

$$\bar{f} = E_{\pi(X)}[f(X)] \tag{1}$$

For example

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}, D_N) &= E_{p(G|D_N)}[E_{p(\boldsymbol{\theta}|G,D_N)}[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, G)]] \\ L_{\hat{G}|D_N} &= E_{p(G|D_N)}[L(G,\hat{G})] = \sum_G L(G,\hat{G}) p(G|D_N), \\ p(\alpha(G)|D_N) &= \sum_G 1(\alpha(G) \textit{ is true}) p(G|D_N) \end{aligned}$$

Idea:

1. sampling from $\pi(X)$ to generate i.i.d random samples $\{X_t, t = 1..N\}$;
2. computation of the estimate $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$;
3. providing confidence measure for $|\bar{f} - \hat{f}_N|$, where $\bar{f} = E_{\pi(X)}[f(X)]$.

Consistency and convergence speed I.

The estimate $\hat{f}_N$ is strongly consistent (by the "strong law of large number"), that is

$$P(lim_{N\to\infty}\hat{f}_N = \bar{f}) = 1 \tag{2}$$

The standardized of $\hat{f}_N$ has asymptotically Gaussian distribution (by the "central limit theorem"), that is

$$\frac{\hat{f}_N - \bar{f}}{\sigma_N} \to N(0,1) \; as \; N \to \infty \; \; where \; \sigma_N = Var(f(X))/\sqrt{N}. \tag{3}$$

If $f(X)$ is bounded, then non-asymptotic results about the speed of convergence are also available by the Hoeffding's inequality including the bound and by the Bernstein's inequality. Specifically, if $f(X)$ is within $[0,1]$, then

$$p(|\hat{f}_N - \bar{f}| \geq \epsilon) \leq 2\exp(-2\epsilon^2 N) \leq \delta \tag{4}$$

$$E[\,|\hat{f}_N - \bar{f}|\,] \leq \sqrt{c_0/N}. \tag{5}$$

Markov chains I.

Let $\mathcal{X} = \{X_0, X_1, \ldots\}$ is a sequence of random variables. The values of X_t are frequently interpreted as states from a state space, the index parameter frequently has a temporal or in biological sequence analysis a location interpretation.

Definition

A sequence of random variables $\mathcal{X} = \{X_0, X_1, \ldots\}$ is called a (first-order) Markov chain, if $p(X_t|X_{t-1}, \ldots, X_0) = p(X_t|X_{t-1})$. The Markov chain is (time-)homogeneous, if the so called transition kernels $p(X_t|X_{t-1})$ does not depend on t.

Let us assume that the values of X_t are discrete and finite, denoted by nonnegative integers $S = \{0, 1, \ldots, K\}$. We use the notation $p^{(t)}$ for the distribution of X_t and $p(X_t = i) = p_i^{(t)}$. We always assume homogeneity and these allows a shorthand notation p_{ij} for the transition probabilities as $p_{ij} = p_{ij}^{(1)} = P(X_{t+1} = j|X_t = i)$, which are forming the (one-step) transition probability matrix $P = P^{(1)}[p_{ij}]$ (a stochastic matrix).

Markov chains II.

The "n-step" transition probability matrix $P^{(n)}$ containing $p_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$ is the nth power of P and

$$p'^{(n)} = p'^{(0)} P^{(n)}, \text{ where } P^{(n)} = P^n. \tag{6}$$

A special distribution is the so called invariant distribution p^{inv}.

Definition

The distribution p'^{inv} is called an invariant distribution of a homogeneous Markov chain $\mathcal{X}$ with transition probability matrix P, if $p'^{inv} = p'^{inv} P$ (Consequently, if $p^{(0)} = p'^{inv}$, then $p^{(t)} = p'^{inv}$ for $\forall\, t$.)

For a first-order Markov chain $\mathcal{X}$ the identical marginals $p^{(t)} = p^{inv}$ implies that $\mathcal{X}$ is strongly stationer, that is the distributions of time-shifted finite marginals are identical, so the invariant distribution p^{inv} is frequently called a stationer distribution.

Stability, irreducibility, aperiodicity

Definition

A Markov chain $\mathcal{X}$ is stable, if $\lim_{t\to\infty} p(X_t) = p^{(\infty)}$ exists, independent of the initial distribution $p(X_0)$ and it is a distribution (called limiting distribution or equilibrium distribution).

Definition

The discrete and finite state space Markov chain $\mathcal{X}$ is called

1. irreducible, if there exists $n_{ij} > 0$ for all i, j that $p_{ij}^{(n_{ij})} > 0$,
2. aperiodic, if for some i (and with irreducibility for all), there exists $n_i > 0$ that for all $n \geq n_i$ $p_{ii}^{(n)} > 0$,

Theorem

If a discrete and finite state space Markov chain $\mathcal{X}$ is irreducible and aperiodic, then the chain is stable and there is a unique invariant distribution that is also the limiting distribution (i.e p'^{∞} is a unique, nonnegative solution of $p'^{\infty} = p'^{\infty}P$ and $\sum_i p_i^{(\infty)} = 1$).

(Strong) Law of large numbers for dependent samples

Theorem

If a discrete and finite state space Markov chain $\mathcal{X}$ is stable and $\bar{f} = E_{\pi(X)}[f(X)] < \infty$, then

$$P(lim_{N\to\infty}\hat{f}_N = \bar{f}) = 1, \tag{7}$$

where $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$.

Ergodicity

Definition

The discrete and finite state space Markov chain $\mathcal{X}$ is called geometrically ergodic, if there exists $0 \leq \lambda < 1$ and function $V(.) > 1$ such that

$$\sum_j |p_{ij}^{(t)} - \pi_j| \leq V(i)\lambda^t \text{ for all } i \tag{8}$$

Central limit theorem for dependent samples

Theorem

If a discrete and finite state space Markov chain $\mathcal{X}$ is geometrically ergodic (so stable), started with its invariant distribution $\pi(X)$ and for a real valued function f $\bar{f} = E_\pi[f(X)]$, $\sigma^2 = Var_\pi(f(X))$, $E_\pi[f(X)^{2+\epsilon}] \leq \infty$ with some $\epsilon > 0$, then for $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$

$$\tau^2 = \sigma^2 + 2 \sum_{k=1}^{\infty} E_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})] \tag{9}$$

exists, nonnegative and finite, and

$$\sqrt{N} \frac{\hat{f}_N - \bar{f}}{\tau} \rightarrow N(0,1) \text{ in distribution as } N \rightarrow \infty. \tag{10}$$

Reversibility (detailed balance) and invariance

Definition

The discrete and finite state space Markov chain $\mathcal{X}$ with transition probability matrix P and (invariant) distribution p^{inv} is called reversible, if it satisfies the detailed balance condition

$$\forall\, i,j\; p_i^{inv} P_{ij} = p_j^{inv} P_{ji}. \tag{11}$$

By summation it gives $\boldsymbol{p}^{\boldsymbol{inv}} \boldsymbol{P}_{.j} = p_j^{inv}$, which is the defining equation of an invariant distribution.

Designing invariant and limiting distributions by detailed balance

If for a given P $\boldsymbol{q}$ satisfies detailed balance, then it is an invariant distribution and vice versa, if for a given target distribution $\boldsymbol{q}$ we can construct a P such that it satisfies detailed balance with $\boldsymbol{q}$, then $\boldsymbol{q}$ is its invariant distribution. Furthermore, if the constructed P is such that the corresponding reversible Markov chain is irreducible and aperiodic as well, then $\boldsymbol{q}$ is its unique, invariant, limiting distribution, so we can generate (dependent) samples by sequential simulation and use it to estimate expectations and to provide confidence measures.

The Metropolis-Hastings Algorithm I.

Let $\pi(X)$ denote the unnormalized, strictly positive target distribution over $S = \{0, 1, \ldots, K\}$ ($\pi_i = \pi(X = i) \geq 0$). Let Q be a transition probability matrix ($Q\mathbf{1} = \mathbf{1}$), the so called proposal distribution (for transitions), such that ($q_{ij} \geq 0$) *iff* ($q_{ji} \geq 0$). Define a Markov chain $\mathcal{X}$ with probability transition matrix P such that

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right); \forall i \neq j \tag{12}$$

using $0/0 = 0$ and define $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. Note that the construction needs only the ratios of the target distribution, which fits to the practical case of unnormalized posterior in Bayesian analysis.
Now $\pi(X)$ is the stationary distribution of the defined Markov chain, which can be proved by showing that the detailed balance condition is satisfied. The cases $i = j$ and if $q_{ij} = q_{ji} = 0$ are trivially satisfied. For $i \neq j$ with $q_{ij} \geq 0$, suppose that $\pi_i q_{ij} \geq \pi_j q_{ji}$, then

$$\pi_i p_{ij} = \pi_i \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \pi_j q_{ji} = \pi_j p_{ji} \tag{13}$$

The Metropolis-Hastings Algorithm II.

If Q is irreducible, so will be P and the same is true for aperiodicity. Consequently, if we provide a proposal distribution Q that (its corresponding Markov chain) is irreducible and aperiodic, then for a given target distribution $\pi(X)$ the construction above defines a stable and reversible Markov chain with (invariant) limiting distribution $\pi(X)$.
If Q is symmetric, then we fall back to the original Metropolis Algorithm without ratio of the proposal distributions

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j}{\pi_i}\right); \forall i \neq j. \tag{14}$$

If Q depends on only some distance between the current state x_t and a proposed state x^* ($q(x^*|x_t) = q(|x^* - x_t|)$), then we get the random-walk Metropolis Algorithm (the distance can be semantically defined in discrete spaces). If Q is independent of the current state ($q(x^*|x_t) = q(x^*)$), then we get the independence sampler, which is geometrically convergent determined by $\inf q(x)/\pi(x)$ (by the closeness to the target distribution) [**?**]. If Q is such that changes at most one component of X based on its full conditional distribution, then we get the Gibbs sampler, with an acceptance probability 1.

The Metropolis-Hastings Algorithm III.

0. [Construct an approximate distribution P^S of the posterior using mixture model around modes for checking and initialization of the MCMC.]
1. Construct an irreducible and aperiodic proposal distribution Q specific to the domain.
2. Draw an initial state x_0 from P^S.
3. For $t = 1, 2, \ldots$
   (a) Draw a candidate state x^* from the proposal distribution Q given x_t.
   (b) Calculate the acceptance probability of a step from x_t to x^* $\alpha(x_t, x^*) = min(1, \frac{\pi_{x^*} q_{x_t x^*}}{\pi_{x_t} q_{x_t x^*}})$.
   (c) Set $x_{t+1} = x^*$ with probability $\alpha(x_t, x^*)$, otherwise $x_{t+1} = x_t$.
4. Continue until convergence and specified confidence.
5. [Evaluate speed of convergence and improve efficiency by redesigning Q. Step back to 2.]
6. [Compare against base-line method using importance resampling with P^S. Step back to 1.]

Convergence diagnostic using a single chain

The first method related to burn in is based on a single chain and the mean. It tests the convergence of a single realization from the sequence $\{Y_i; i = 1..N\}$ exploiting that after burn-in (i.e. in case of convergence) the distribution of an ergodic average is asymptotically Gaussian. Formally, define the averages $\hat{Y}_b$ after a (putative) burn-in m and $\hat{Y}_a$ at the end of sequence

$$\hat{Y}_b = \frac{1}{N_b} \sum_{i=m+1}^{m+N_b} Y_i \text{ and } \hat{Y}_a = \frac{1}{N_a} \sum_{i=N-N_a+1}^{N} Y_i \tag{15}$$

with no overlap ($m + N_b + N_a < N$). If N_a/N and N_b/N are fixed, then

$$z_G = \frac{\hat{Y}_b - \hat{Y}_a}{\sqrt{\hat{Var}(Y_b) + \hat{Var}(Y_a)}} \to N(0,1) \text{ in distribution.} \tag{16}$$

Convergence diagnostic using multiple chains

The second method is using independently initialized multiple chains and analyze the variance. Its test is based on the relation of the estimators of the (Monte Carlo) variance of the target quantity using a between-sequence estimation (i.e. the variance of the (independent) estimates for the chains) and a within-sequence estimation (i.e. average of the within-sequence estimates of the variance).

Convergence diagnostic II.

Formally, for M chains with N samples $\{Y_i, j; i = 1..N j = 1 \ldots, M\}$ define

$$B = \frac{N}{M-1} \sum_{j=1}^{M} (\bar{Y}_{+,j} - \bar{(Y)}_{+,+})^2, \text{ where } \bar{Y}_{+,j} = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}, \bar{Y}_{+,+} = \frac{1}{M} \sum_{j=1}^{M} Y_{+,j}$$

$$W = \frac{1}{M} \sum_{j=1}^{M} s_j^2 \text{ where } s_j^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_{ij} - \bar{Y}_{+,j})^2.$$

If the simulations are started independently from an overdispersed starting distribution, then the quantity

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{var}^+(Y)}{W}}, \text{ where var}^+(\mathrm{Y}) = \frac{\mathrm{N}-1}{\mathrm{N}}\mathrm{W} + \frac{1}{\mathrm{NB}} \tag{17}$$

called ”potential scale reduction" can be used to monitor convergence, because $\hat{var}^+(Y)$ overestimates the variance as the chains are still overdispersed and W underestimates the variance as they are still confined to small regions.

Confidence estimation I.

The second task after the determination and elimination of the burn-in period is to determine the stopping time and/or providing confidence measure(s) for the estimate(s). The first method is related to the between-sequence variance of the earlier method, though using a single chain $\{Y_i; i = 1, \ldots, NM\}$. It partitions a sufficiently long chain into M parts with length N such that the ergodic averages are approximately independently Gaussian with mean $E_\pi[f(X)]$ and variance τ^2/N (see Eq. 9). Then approximate τ^2 as follows

$$\hat{\tau}^2 = \frac{N}{(M-1)} \sum_{j=1}^{M} (\bar{Y}_j - \bar{\bar{Y}})^2, \text{ where } \bar{Y}_j = \frac{1}{N} \sum_{i=(j-1)N+1}^{jN} Y_i, \bar{\bar{Y}} = \frac{1}{M} \sum_{j=1}^{M} \bar{Y}_j$$

Confidence estimation II.

Another method is based on the direct estimation of the autocovariance terms $\gamma_k = E_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})]$ in the Eq. 9 of the Monte Carlo variance with

$$\hat{\gamma}_k = \frac{1}{N-k} \sum_{i=1}^{N-k} (Y_i - \bar{f})(Y_{i+k} - \bar{f}) \tag{18}$$

and use a special weighting to eliminate the not reliable autocorrelation terms as follows

$$\hat{\tau}_N^2 = \hat{\gamma}_0 + 2 \sum_{i=1}^{\infty} w_N(i) \hat{\gamma}_0, \text{ where } 0 \leq \mathrm{w_N(i)} \leq 1. \tag{19}$$

Thank you for your attention!

Questions?