Advanced Monte Carlo methods

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

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

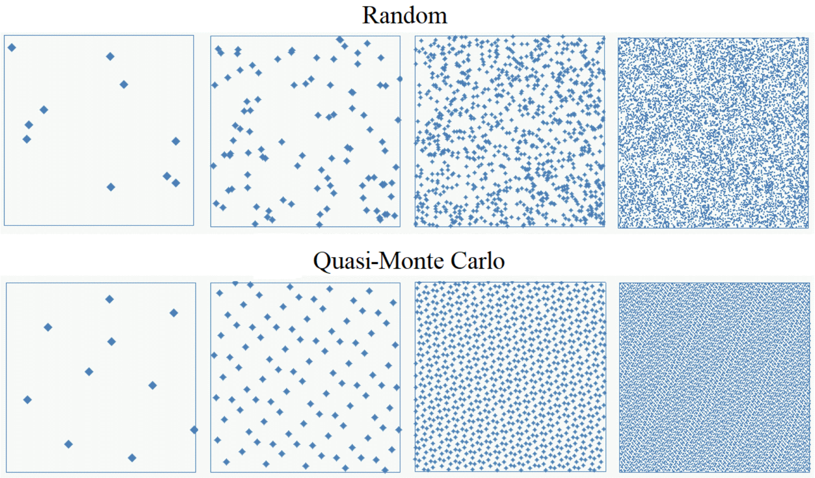
Outline

Monte Carlo techniques we've covered so far

- Simple Monte Carlo
- Importance sampling
- Gibbs sampling
- Metropolis–Hastings
- Combining MCMC moves
- MCMC algorithms for various models

Outline

Quasi-Monte Carlo (QMC)



Quasi-Monte Carlo (QMC)

- Quasi-Monte Carlo uses well-spaced deterministic points instead of random samples to approximate integrals.
- These deterministic sequences of points are also known as "low-discrepancy sequences" (LDSs).
- In simple Monte Carlo, the error is of order $O(1/\sqrt{n})$.
- In quasi-Monte Carlo, the error is of order $O((\log n)^d/n)$ where d is the dimension. (Major improvement!)
- When $d = 1$ or $d = 2$, we could just use a grid, but the number of points in a grid grows exponentially with d.
- Consequently, as d grows, it is far from obvious how to construct good LDSs.

Quasi-Monte Carlo (QMC): Constructions

- Low-discrepancy sequences (LDSs) are usually defined to target the uniform distribution on $[0, 1]^d$.

- Standard choices of LDS:
  - Sobol sequence
  - Halton sequence
  - Hammersley sequence
  - Faure sequence

- By using inverse CDFs, it is easy to transform LDSs to target other probability distributions, e.g., multivariate Gaussians, discrete distributions, copulas.

Quasi-Monte Carlo (QMC): Issues

- Even though QMC is asymptotically superior, in high-dim spaces the benefit does not kick in until n is very large.
- The issue is that when d is large, the $(\log n)^d$ factor often makes QMC slower in practice.
- So far, QMC has had limited success in high-dim applications in statistics.
- Notable exception: In finance applications, QMC is reported to outperform Monte Carlo in many high-dim problems.

Quasi-Monte Carlo (QMC) and MCMC

- Part of the issue is that QMC does not work well when the target distribution is concentrated on a small set, since most of the samples are wasted.
- MCMC does not have this problem since it tends to stay in regions where the target distribution is concentrated.
- It is not clear how to combine QMC and MCMC in a way that obtains the nice properties of both.
- There have been efforts to use QMC in MCMC moves (Owen & Tribble, 2005), but so far there is not a clear benefit of this over standard MCMC.
- This remains an interesting open research problem.

Outline

Outline

Auxiliary variable technique

- Many advanced MCMC methods employ auxiliary variables to construct moves with the desired stationary distribution.

- Suppose $p(x, z)$ is such that $\pi(x) = \int p(x, z)dz$ and T_{aug} is a move on (x, z) that preserves $p(x, z)$.

- Interpretation:
  - $\pi =$ target distribution,
  - $x =$ state of the Markov chain,
  - $z =$ auxiliary variable.

- Define a move T on x as follows:
  1. Suppose the current state is x.
  2. Sample z from $p(z|x)$.
  3. Generate (x', z') by applying the move T_{aug} to (x, z).
  4. Define the new state to be x'.
- It can be shown that $\pi T = \pi$, that is, T preserves π.

Auxiliary variable technique

- An auxiliary variable move T constructed in this way is guaranteed to have π as its stationary distribution.

- However, it is not guaranteed to yield an irreducible Markov chain (and in practice, often it does not).

- Auxiliary moves are often combined with other moves to improve mixing. Irreducibility can be obtained with an appropriate combination of moves.

- Special cases of the auxiliary variable technique:
  - Metropolis–Hastings (and Gibbs)
  - Mixture model sampler with allocation variables
  - Probit regression sampler with data augmentation
  - Slice sampling
  - Reversible jump MCMC
  - Hamiltonian Monte Carlo
  - MCMC for doubly intractable models
  - Swendsen-Wang algorithm for Ising model

Outline

Auxiliary variable technique: MH as a special case

- How can we view Metropolis–Hastings (MH) as an auxiliary variable move?
- First, the Metropolis algorithm is an older and simpler version of MH in which the proposal distribution is symmetric:

$$q(x_{\mathsf{prop}}|x) = q(x|x_{\mathsf{prop}}).$$

- Consequently, the proposal distribution cancels in the MH acceptance probability.
- Thus, in a Metropolis move, the acceptance probability is

$$\alpha = \min\left\{1,\ \frac{\pi(x_{\mathsf{prop}})}{\pi(x)}\right\}.$$

Auxiliary variable technique: MH as a special case

- MH can be viewed as performing a simple Metropolis move in an augmented space.

- Suppose $z = x_{\text{prop}}$ is the proposed new value of x in MH.

- Then $p(x, z) = p(x, x_{\text{prop}}) = \pi(x)q(x_{\text{prop}}|x)$.

- Now, apply a Metropolis move on (x, x_{prop}) as follows: Propose to swap x and x_{prop}, that is, $(x, x_{\text{prop}}) \mapsto (x_{\text{prop}}, x)$.

- The acceptance probability of this Metropolis move on (x, x_{prop}) is

$$\alpha = \min\left\{1,\ \frac{\pi(x_{\text{prop}})q(x|x_{\text{prop}})}{\pi(x)q(x_{\text{prop}}|x)}\right\}.$$

- This coincides with the standard MH acceptance probability!

Outline

Example: Probit regression with data augmentation

- Probit regression for binary outcomes Y_i given covariates x_i:

$$\mathbb{P}(Y_i = 1 \mid \beta, x_i) = \Phi(\beta^{\mathtt{T}} x_i)$$

  where $\Phi(\cdot)$ is the standard normal CDF.

- The posterior on β is not amenable to Gibbs sampling.

- We could use MH, but we would have to tune the proposal distribution.

- Auxiliary variables (a.k.a. data augmentation) allow us to use Gibbs on an augmented space.

Example: Probit regression with data augmentation

- Augmented model:

$$
\begin{aligned}
\beta &\sim \mathcal{N}(\mu, C) \quad \text{(prior on } \beta\text{)} \\
Z_i &\sim \mathcal{N}(\beta^{\mathrm{T}} x_i,\, 1) \quad \text{(auxiliary variables)} \\
Y_i &= \mathrm{I}(Z_i > 0). \quad \text{(outcomes)}
\end{aligned}
$$

- Integrating out the Z_i's, this is equivalent to the probit regression model.

- The nice thing is that given the Z_i's, this is just a linear regression model with the Z_i's playing the role of the data.

- Thus, the full conditional for β is the usual linear regression posterior with Z_i's as data.

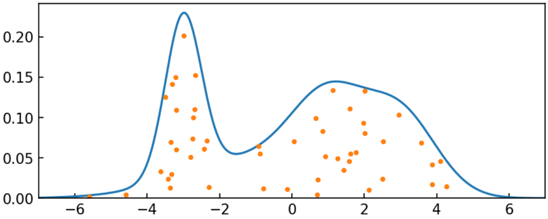
- This approach to probit regression is due to Albert & Chib (1993).

Outline

Slice sampling

- Like MH, slice sampling can be used on complicated target distributions $\pi(x) \propto \tilde{\pi}(x)$.
- Slice sampling is way of constructing MCMC moves that sometimes have better mixing than MH.
- The advantage is that slice sampling makes moves that automatically adapt to the shape of the target distribution.
- Basic idea: Sample uniformly from the region under the target density by alternating between "horizontal" and "vertical" Gibbs updates.
- Slice sampling was introduced by Neal (2003).

Slice sampling: Recall the projection principle



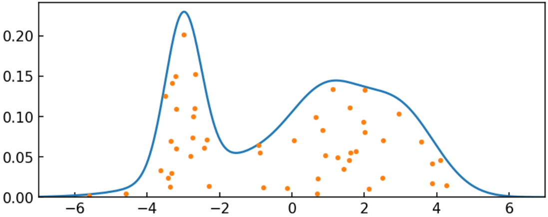
- Suppose we want to sample from a distribution on $\mathbb{R}^d$ with p.d.f. $\pi(x) \propto \tilde{\pi}(x)$.

- Consider the region of $\mathbb{R}^{d+1}$ under $\tilde{\pi}$:

$$A = \left\{ (x, y) \in \mathbb{R}^{d+1} \,:\, 0 < y < \tilde{\pi}(x) \right\}.$$

- It turns out that if $(X, Y) \sim \text{Uniform}(A)$, then $X \sim \pi$.

Slice sampling: Idea of the algorithm



- Slice sampling: Sample uniformly from the region under the target density

$$A = \left\{ (x, y) \in \mathbb{R}^{d+1} \, : \, 0 < y < \tilde{\pi}(x) \right\}$$

by alternating between "horizontal" and "vertical" Gibbs updates to x and y, respectively.

Slice sampling: Algorithm

- The vertical update (sampling from the full conditional of y) is trivial:

$$Y|x \sim \mathrm{Uniform}(0, \tilde{\pi}(x)).$$

- For the horizontal update (sampling from the full conditional of x), ideally we would sample uniformly from the "slice":

$$A_y = \Big\{x \in \mathbb{R}^d \,:\, (x, y) \in A\Big\} = \Big\{x \in \mathbb{R}^d \,:\, y < \tilde{\pi}(x)\Big\}.$$

- Exactly sampling from $\mathrm{Uniform}(A_y)$ is often difficult. Fortunately, however, we only need to make a move that has $\mathrm{Uniform}(A_y)$ as its stationary distribution.

- For instance, we can use an MH move with proposal $\mathrm{Uniform}(B_r(x))$ where $B_r(x) = \{x' \in \mathbb{R}^d \,:\, \|x' - x\| < r\}$ is the ball of radius r, centered at x. The radius r can be chosen adaptively; see Neal (2003) for details.

Outline

Reversible jump MCMC (RJMCMC)

- RJMCMC is a technique for sampling from spaces of varying dimension.

- For instance, suppose models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ have parameter spaces of dimension $d_1, d_2, \ldots$, and we want to sample from the posterior on models, but the marginal likelihood cannot be computed.

- Examples:
  - Variable selection with non-conjugate priors
  - Mixture models with an unknown number of components (with non-conjugate priors on the component parameters)
  - Time-series models of unknown order (with non-conjugate priors)

Reversible jump MCMC (RJMCMC)

- It is not obvious how to construct valid MCMC updates for moving between spaces with different dimension.
- Part of the issue is that the meaning of the density is different in different dimensions.
- RJMCMC (Green, 1995) is a general way of constructing "transdimensional" moves with the correct stationary distribution.
- Basic idea: Introduce auxiliary variables that embed multiple models into spaces of common dimension, and make a move in this augmented space.
- Unfortunately, RJMCMC often mixes very slowly because it is difficult to design good proposals for moving between models.

Outline

Quasi-Monte Carlo

Auxiliary variable methods

- Auxiliary variable technique
- MH as a special case
- Probit regression with data augmentation
- Slice sampling
- Reversible jump MCMC

Handling intractable likelihoods

- Approximate Bayesian computation (ABC)
- MCMC for doubly intractable models

Outline

Approximate Bayesian computation (ABC)

- Sometimes your model is so complicated that you can't even compute the likelihood $p(x|\theta)$ as a function of θ.
- This is sometimes referred to as a "doubly intractable model".
- Suppose you want to sample from the posterior. If you can't compute $p(x|\theta)$, then most MCMC methods cannot be used. E.g., you can't compute the MH acceptance probabilities.
- Suppose, on the other hand, that you can easily generate data sets from $p(x|\theta)$. ABC is designed for such situations.
- This is actually pretty common, for instance, in:
  - evolutionary biology,
  - ecology models,
  - epidemiology,
  - weather and climate models,
  - physics/biophysics models, and
  - dynamical systems models.

ABC: Naive version (Rejection sampling)

- Suppose x is discrete and the observed value is x_{obs}.

- Generate samples of θ according to the following procedure:

  1. Draw $\theta \sim p(\theta)$ and $x|\theta \sim p(x|\theta)$, so that $(x, \theta) \sim p(x, \theta)$.

  2. If $x = x_{\text{obs}}$, then accept θ. Otherwise, reject θ and go back to step 1.

- Then the accepted θ's are samples from the posterior $p(\theta|x_{\text{obs}})$, by the rejection principle.

- (Rejection principle: If we sample Z, and reject unless $Z \in A$, then we get samples from $Z \mid Z \in A$.)

- This naive version of ABC is extremely inefficient unless x can only take a relatively small number of values.

ABC: Standard version (Rejection + summary statistics)

- To make this work more generally, the idea of ABC is accept θ if x is close to x_{obs}, rather than requiring $x = x_{\text{obs}}$.

- Usually, "closeness" is defined by accepting if

$$d(s(x), s(x_{\text{obs}})) \le \varepsilon$$

  where $s(x) \in \mathbb{R}^k$ is a vector of summary statistics and $d(\cdot, \cdot)$ is Euclidean distance.

- This generates samples from the "ABC posterior"

$$p\big(\theta \mid d(s(X), s(x_{\text{obs}})) \le \varepsilon\big),$$

  which is intended to approximate the standard posterior $p(\theta \mid X = x_{\text{obs}})$.

- If $s(x) = x$ and $\varepsilon = 0$, then the ABC posterior equals the standard posterior.

ABC: Pros and cons

- ABC is very easy to use, even for complex models with intractable likelihoods.
- However, the accuracy of the approximation can be poor, unless the summary statistics $s(x)$ are chosen well and the tolerance ε is small.
- Choosing good summary statistics is tricky, and while there is some theory, it remains something of an art.
- Further, when ε is small, the number of rejected samples is large, making ABC very computationally burdensome.
- Adding to the computational burden is the fact that in many models with intractable likelihoods, generating samples from $p(x|\theta)$ can be slow.

ABC: Refined version (Nonparametric regression)

- A more refined version of ABC uses nonparametric regression rather than rejection sampling.
- This allows one to use all of the samples, and tends to improve the accuracy of the approximation.
- The idea is to view the calculation of a posterior expectation $\mathrm{E}(h(\theta)|x)$ as a regression problem, as follows.
- As before, generate a bunch of samples $(x_i, \theta_i) \sim p(x, \theta)$ i.i.d.
- Consider $h(\theta)$ to be the outcome, x to be the predictor, and $(x_1, h(\theta_1)), \ldots, (x_T, h(\theta_T))$ to be the data.
- Use nonparametric regression to fit this "data". Then the estimated regression function $\widehat{\mathrm{E}}(h(\theta)|x_{\text{obs}})$ is an approximation of $\mathrm{E}(h(\theta)|x_{\text{obs}})$, the quantity of interest.

Outline

MCMC for doubly intractable models

- ABC can be used very generally — whenever we can generate data sets from the assumed model.

- In some doubly intractable models, we can compute the likelihood $p(x|\theta)$ up to a normalization constant $c(\theta)$, that is,

  $$p(x|\theta) = f(x, \theta)c(\theta)$$

  where $f(x, \theta)$ can be computed but $c(\theta)$ cannot.

- In such cases, there are very clever MCMC techniques for sampling from the posterior $p(\theta|x)$.

- Basic idea: Introduce an auxiliary data set in such a way that $c(\theta)$ cancels in the MH acceptance probability.

MCMC for doubly intractable models

- The posterior is

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{f(x,\theta)c(\theta)p(\theta)}{p(x)}.$$

- Introduce an auxiliary variable z that takes values in the same space as x and has some conditional density $p(z|x,\theta)$ that can be computed and sampled from.

- Consider using an MH move preserving the augmented posterior,

$$p(z,\theta|x) = p(z|x,\theta)p(\theta|x) = \frac{p(z|x,\theta)f(x,\theta)c(\theta)p(\theta)}{p(x)}.$$

MCMC for doubly intractable models

- If the MH proposal distribution is $q(z', \theta' | z, \theta, x)$, then the MH acceptance probability is

$$\alpha = \min\left\{1,\ \frac{p(z', \theta' | x)}{p(z, \theta | x)} \frac{q(z, \theta | z', \theta', x)}{q(z', \theta' | z, \theta, x)}\right\}$$
$$= \min\left\{1,\ \frac{p(z' | x, \theta')}{p(z | x, \theta)} \frac{f(x, \theta') c(\theta') p(\theta')}{f(x, \theta) c(\theta) p(\theta)} \frac{q(z, \theta | z', \theta', x)}{q(z', \theta' | z, \theta, x)}\right\}.$$

- Now, suppose we choose the proposal distribution such that

$$q(z', \theta' | z, \theta, x) = g(\theta' | x, \theta) f(z', \theta') c(\theta'),$$

in other words, propose θ' from some density $g(\theta' | \theta, x)$ and propose z' by sampling from the model given θ'.

- Plugging this into the MH acceptance probability, the $c(\theta)$ and $c(\theta')$ factors cancel, and we have

$$\alpha = \min\left\{1,\ \frac{p(z' | x, \theta')}{p(z | x, \theta)} \frac{f(x, \theta') p(\theta')}{f(x, \theta) p(\theta)} \frac{g(\theta | x, \theta') f(z, \theta)}{g(\theta' | x, \theta) f(z', \theta')}\right\}.$$

MCMC for doubly intractable models

- Every factor in this MH acceptance probability can be computed.
- Hence, this yields a tractable MCMC sampler for $p(z, \theta|x)$. Discarding the z samples and keeping the θ samples, we obtain samples from $p(\theta|x)$.
- While it works in principle, this sampler can suffer from low acceptance probability.
- Improvements upon this basic idea have been developed by Murray et al. (2006).
- The algorithm above was introduced by Møller et al. (2006).

References and supplements

- Owen, A. B., & Tribble, S. D. (2005). A quasi-Monte Carlo Metropolis algorithm. Proceedings of the National Academy of Sciences, 102(25), 8844-8849.
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88(422), 669-679.
- Neal, R. M. (2003). Slice sampling. The Annals of Statistics, 31(3), 705-741.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4), 711-732.
- Møller, J., Pettitt, A. N., Reeves, R., & Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika, 93(2), 451-458.
- Murray, I., Ghahramani, Z., & MacKay, D. J. (2006). MCMC for doubly-intractable distributions. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (pp. 359-366).