Sampling-Based Approaches to Calculating Marginal Densities

Amber Hu and Sophia Lu

STATS 319

February 13, 2024

The Problem of Bayesian Inference

- Let $\mathbf{x}$ denote observed data and $\boldsymbol{\theta}$ denote model parameters.
- The central problem of Bayesian inference is to compute the posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$, where (by Bayes rule):

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{x})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{x} \mid \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{evidence}}}$$

- For many models, computing the evidence $p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$ is intractable.
- How can we approximate the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$?

Example: Hierarchical Gaussian Model

- Suppose we would like to model differences in test scores from S schools.
- Let $x_{s,n} \in \mathbb{R}$ be the score of the n-th student from the s-th school, for $n = 1, \dots N_s$, $s = 1, \dots S$.
- We could use the following hierarchical model:

$$
\begin{aligned}
x_{s,n} &\sim \mathcal{N}(\theta_s, \sigma_s^2) \quad \text{for } n = 1, \dots, N_s \text{ and } s = 1, \dots, S \\
\theta_s &\sim \mathcal{N}(\mu, \tau^2) \quad \text{for } s = 1, \dots, S \\
\sigma_s^2 &\sim \mathcal{IG}(a_1, b_1) \quad \text{for } s = 1, \dots, S \\
\mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
\tau^2 &\sim \mathcal{IG}(a_2, b_2)
\end{aligned}
$$

- The model parameters are $\boldsymbol{\theta} = \{\{\theta_s, \sigma_s^2\}_{s=1}^S, \mu, \tau^2\}$. Assume all other hyperparameters are known.

MCMC Methods

- Instead of needing to characterize the full posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$, we almost always care about expectations with respect to this distrbution. For example:
  - $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[\boldsymbol{\theta}]$ (posterior mean)
  - $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[\mathbf{1}(\boldsymbol{\theta} \in \mathcal{S})]$ (posterior probability of $\boldsymbol{\theta}$ being in a set $\mathcal{S}$)
  - $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[p(\mathbf{x}' \mid \boldsymbol{\theta})]$ (posterior predictive density of new data $\mathbf{x}'$)
- In general, we care about

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$$

- MCMC methods allow us to draw approximate samples $\boldsymbol{\theta}_n \sim p(\boldsymbol{\theta} \mid \mathbf{x})$. Then, we use the estimate

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[f(\boldsymbol{\theta})] \approx \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{\theta}_n)$$

Prior Work

- Quadrature methods (Reilly (1976), Naylor and Smith (1982))
  - Roughly, these methods approximate the integral numerically via

$$\int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} \approx \sum_{m=1}^{M} f(\boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m \mid \mathbf{x}) \Delta_m$$

    where $\boldsymbol{\theta}_m \in \Theta$ form a grid of points and Δ_m is the partition size.
  - Becomes computationally intensive in high parameter dimensions because we require more points to approximate the integral.
- Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970))
  - MCMC method which allows us to sample from a distribution $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{Z_p}$, where $\tilde{p}(\mathbf{z})$ can be evaluated at any $\mathbf{z}$ and Z_p may be unknown.
  - Involves drawing samples from a proposal distribution and accepting it with some probability.

Overview of Gelfand and Smith (1990)

- "I would like to turn the Bayes thing from being some kind of quasi-religious view of how to do it [Bayesian inference]... to just do it, practically" - Adrian Smith, 2024(?)
- Main contributions of this paper:
  - Argued for using substitution sampling (Tanner and Wong, 1987) and Gibbs sampling (Geman and Geman, 1984) as simple and accessible computational methods for estimating marginal densities
  - Demonstrated a close relationship between the two methods, and showed how substitution sampling can be used to accelerate Gibbs sampling
  - Popularized the usage of MCMC methods in the broader statistics community

Overview of Gelfand and Smith (1990)

- The goal is to estimate **marginal densities** from available **conditional densities**.
  - E.g., if our model has additional parameters or latent variables $\mathbf{z}$, we may want to estimate $p(\boldsymbol{\theta} \mid \mathbf{x})$ from $p(\boldsymbol{\theta} \mid \mathbf{z}, \mathbf{x})$.
  - An "available" density means that we can efficiently sample from it.
- For simplicity, let's ignore the Bayesian framework of conditioning on data $\mathbf{x}$ for now. We'll come back to it later.

Substitution Sampling

- We would like to estimate marginals $p(x)$ and $p(y)$, given that $p(x \mid y)$ and $p(y \mid x)$ are available.
- We can write

$$p(x) = \int p(x \mid y)p(y)dy, \quad p(y) = \int p(y \mid x)p(x)dx$$

- Substituting $p(y)$ into the expression for $p(x)$ yields

$$\begin{aligned} p(x) &= \int p(x \mid y) \left[\int p(y \mid x')p(x')dx' \right] dy \\ &= \int \underbrace{\left[\int p(x \mid y)p(y \mid x')dy \right]}_{:=h(x,x')} p(x')dx' \end{aligned}$$

- This suggests an iterative process to get to $p(x)$, through a series of distributions $p_i(x) \approx p(x)$, such that

$$p_{i+1}(x) = \int h(x, x')p_i(x')dx'$$

Substitution Sampling

Marginal distributions

$$p(x) = \int p(x \mid y)p(y)dy, \quad p(y) = \int p(y \mid x)p(x)dx$$

Algorithm:

- Initialize a distribution $p_0(x)$.
- Sample $x^{(0)} \sim p_0(x)$.
- For $k = 1, \ldots, i$ iterations:
  - Sample $y^{(k)} \sim p(y \mid x = x^{(k-1)})$.
  - Sample $x^{(k)} \sim p(x \mid y = y^{(k)})$.
- At the end of i iterations, collect samples $(x^{(i)}, y^{(i)})$.

Repeat the algorithm m times to generate iid pairs $(x_j^{(i)}, y_j^{(i)})$ for $j = 1, \ldots, m$.

Substitution Sampling

Why does this algorithm work?

- Tanner and Wong (1987) showed that $x^{(i)} \xrightarrow{d} p(x)$ and $y^{(i)} \xrightarrow{d} p(y)$. So for i large enough, we can take $\{x_j^{(i)}, y_j^{(i)}\}_{j=1}^m$ as approximate samples from the marginal distributions.
- We can also obtain an estimate of $p(x)$ via Monte Carlo:

$$p(x) \approx \hat{p}_i(x) = \frac{1}{m} \sum_{j=1}^{m} p(x \mid y = y_j^{(i)})$$

- Can easily extend to 3 or more variables by writing, e.g.,

$$p(x) = \int p(x, z \mid y) p(y) dy dz$$

$$p(y) = \int p(x, y \mid z) p(z) dx dz$$

$$p(z) = \int p(y, z \mid x) p(x) dx dy$$

Gibbs Sampling

- Suppose we instead write the factorization of 3 variables as:

$$p(x) = \int p(x \mid z, y)p(z \mid y)p(y)dydz$$
$$p(y) = \int p(y \mid x, z)p(x \mid z)p(z)dxdz$$
$$p(z) = \int p(z \mid y, x)p(y \mid x)p(x)dxdy$$

  and we only know the full conditionals.
- We can no longer use substitution sampling as is, since it requires knowledge of both full conditionals and reduced conditionals (e.g. $p(z \mid y)$).
- Gibbs sampling is an MCMC method which relies only on sampling from the full conditionals.

Gibbs Sampling

Suppose we would like to sample from K marginal distributions, $p(u_1), \ldots, p(u_K)$. Assume that the full conditionals $p(u_k \mid u_{\neg k})$ for $k = 1, \ldots, K$ are available.

- Initialize $u_1^{(0)}, \ldots, u_K^{(0)}$.
- Sample $u_1^{(1)} \sim p(u_1 \mid u_2^{(0)}, \ldots u_K^{(0)})$.
- Sample $u_2^{(1)} \sim p(u_2 \mid u_1^{(1)}, u_3^{(0)}, \ldots u_K^{(0)})$.
- ... And so on.

After i iterations, collect samples $(u_1^{(i)}, \ldots, u_K^{(i)})$. To obtain multiple iid samples, one could repeat the algorithm m times, or subsample a single sequence of samples (since successive samples will be correlated).

Gibbs Sampling

Why does this algorithm work?

- Geman and Geman (1984) showed that

$$(u_1^{(i)}, \ldots, u_K^{(i)}) \xrightarrow{d} p(u_1, \ldots, u_K)$$

  In fact, this holds under any visiting order, as long as each variable is visited infinitely often.
- They also showed the ergodic theorem for Gibbs sampling. For any measurable f whose expectation exists,

$$\lim_{i \to \infty} \frac{1}{i} \sum_{\ell=1}^{i} f(u_1^{(\ell)}, \ldots, u_K^{(\ell)}) \xrightarrow{\text{a.s.}} \mathbb{E}(f(u_1, \ldots, u_K))$$

- We can obtain an estimate of the density of u_k for $k = 1, \ldots, K$ by Monte Carlo:

$$\hat{p}_i(u_k) = \frac{1}{m} \sum_{j=1}^{m} p(u_k \mid \neg u_k = \neg u_{kj}^{(i)})$$

Relationship Between Substitution and Gibbs Sampling

In the case of two random variables, Gibbs sampling and substitution sampling are identical.

For $K > 2$ variables,

- Gibbs sampling requires K full conditional distributions.
- Substitution sampling requires $K(K-1)$ conditional distributions (including all K full conditional distributions.

Relationship Between Substitution and Gibbs Sampling

Substitution-sampling algorithm may be carried out under availability of just the set of full conditional distributions:

- If $p(y|x)$ is unavailable, we can create a sub-substitution loop to obtain it via

$$p(y|x) = \int p(y|x,z)p(z|x)dz$$

$$p(z|x) = \int p(z|x,y)p(y|x)dy$$

$$\dots$$

- For K variables, this idea can be extended to estimate an arbitrary reduced conditional distribution, given the full conditionals.
- When the set of K full conditionals are available, substitution-sampling algorithm and Gibbs sampler are equivalent.

Relationship Between Substitution and Gibbs Sampling

Accelerated convergence from the substitution-sampling algorithm when some reduced distributions (distinct from the full conditional distributions) are available:

- Write the substitution algorithm with appropriate conditioning to capture these reduced conditionals.
- As we traverse a cycle, we would sample from these distributions as we come to them (otherwise sampling from the full conditional distributions).

Importance Sampling

- Rubin (1987) suggested a nontierative Monte Carlo method for generating marginal distributions using importance-sampling.
- Suppose we want to compute $p(x)$, given $\propto p(x, y)$ and $p(x|y)$, and $p(y)$ is unknown.

Importance Sampling

- Choose an importance-sampling distribution denoted $p_s(y)$ for Y that has positive support wherever Y does, i.e. $\text{supp}(p_s(y)) \supseteq \text{supp}(p(y))$.
- Draw iid pairs (X_l, Y_l) for $l = 1, \ldots, N$ from joint distribution; for example, draw Y_l from $p_s(y)$ and X_l from $p(x|Y_l)$.
- Compute importance weights $w_l := p(X_l, Y_l) / \left(p(X_l, Y_l) p_s(Y_l)\right)$.
- Estimate marginal density $p(x)$ by

$$\hat{p}(x) = \sum_{l=1}^{N} \Big(p(x|Y_l) w_l \Big) / \sum_{j=1}^{N} w_j.$$

- $\hat{p}(x) \overset{a.s.}{\to} p(x)$ as $N \to \infty$ for a.s. x.

Revisiting the Hierarchical Gaussian Model

Recall the hierarchical Gaussian model:

$$\begin{aligned} x_{s,n} &\sim \mathcal{N}(\theta_s, \sigma_s^2) \quad \text{for } n = 1, \ldots, N_s \text{ and } s = 1, \ldots, S \\ \theta_s &\sim \mathcal{N}(\mu, \tau^2) \quad \text{for } s = 1, \ldots, S \\ \sigma_s^2 &\sim \mathcal{IG}(a_1, b_1) \quad \text{for } s = 1, \ldots, S \\ \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \tau^2 &\sim \mathcal{IG}(a_2, b_2) \end{aligned}$$

The full conditional distributions can be found in closed form:

- $p(\theta_s \mid \mu, \tau^2, \sigma_s^2, \{x_{s,n}\})$ and $p(\mu \mid \tau^2, \{\theta_s, \sigma_s^2\}, \{x_{s,n}\})$ are Gaussian distributions
- $p(\sigma_s^2 \mid \mu, \tau^2, \theta_s, \{x_{s,n}\})$ and $p(\tau^2 \mid \mu, \{\theta_s, \sigma_s^2\}, \{x_{s,n}\})$ are inverse gamma distributions.

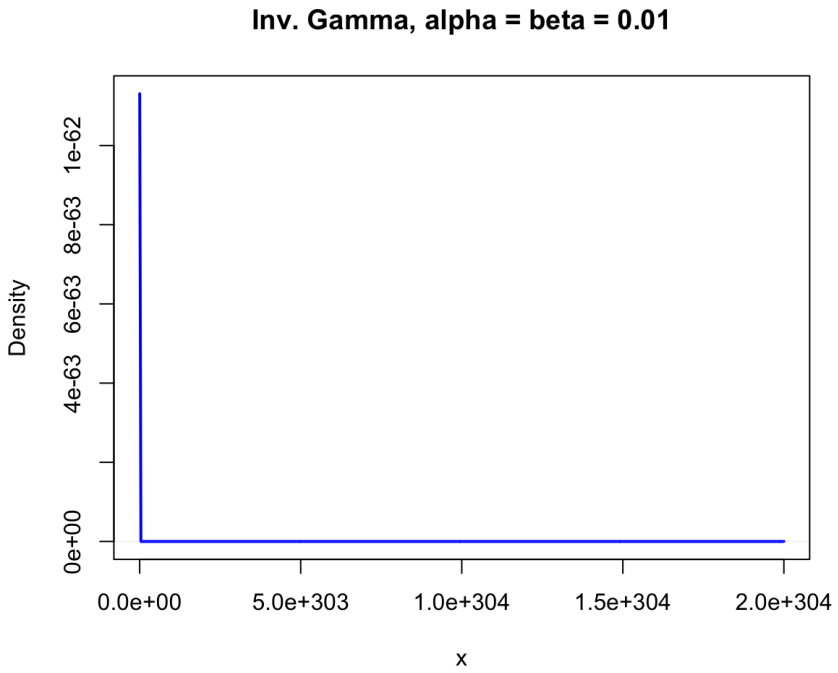


Figure 1: Inverse gamma prior for σ_s^2 and τ^2

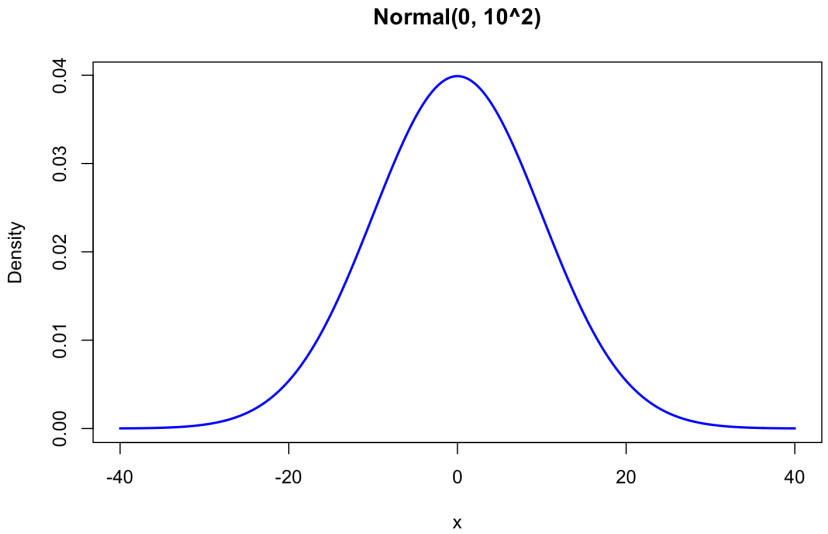
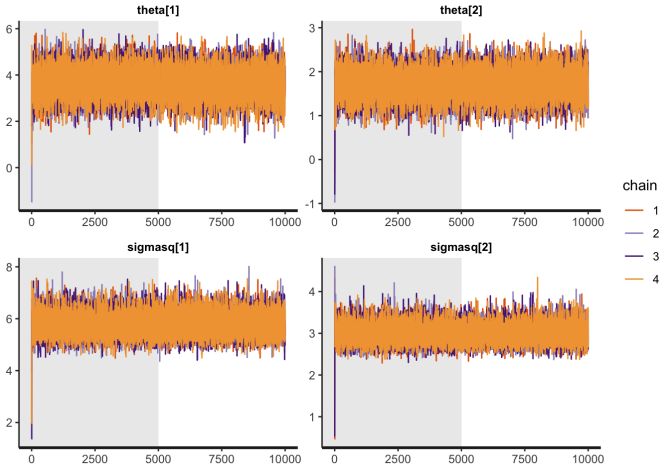
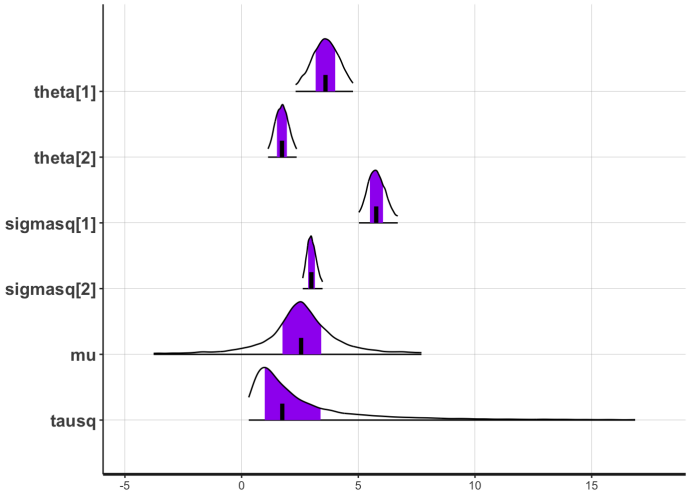


Figure 2: Normal prior for μ

Revisiting the Hierarchical Gaussian Model



(a) Trace plots for model parameters.



(b) Estimated posteriors for model parameters.

Notable developments since then

- Hamiltonian Monte Carlo (HMC) (Neal, 1996):
  - HMC is an instance of the MH algorithm. Proposals are generated via Hamiltonian dynamics evolution simulated through a time-reversible and volume-conserving numerical integrator.
- Stan: A probabilistic programming language for statistical inference written in C++.
  - Allows for easy specification of Bayesian hierarchical model and fast inference based on a variant of the No-U-Turn sampler (NUTS, Hoffman and Gelman, 2014).
- Incorporation of deep neural networks in computing posteriors:
  - A Deep Generative Approach to Conditional Sampling, (Zhou et al., 2023 JASA)
  - Metropolis-Hastings via Classification, (Wang et al., 2022 JMLR)