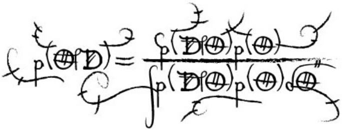


Bayesian Learning 732A46: Lecture 7

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

April 2016

Lecture overview

- Bayesian computations - a recap
- Grid based methods and their curse
- Monte Carlo integration
- First tools to simulate from unknown distributions

- The two **major steps** of any Bayesian analysis
  (1) Obtain the posterior distribution.
  (2) Average some function over the posterior distribution.

(1) The **posterior distribution** $p(\theta|y) = p(\theta|y)$ by Bayes' theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta), \quad p(y) = \int p(y|\theta)p(\theta)d\theta.$$

- For **conjugate priors** $p(\theta|y)$ is a known distribution. Only available for few and simple models.

(2) Examples [$\theta \sim \pi(\cdot)$ continuous. Replace $\int$ by $\sum$ for discrete θ]

Expectation: $E[\theta] = \int \theta p(\theta|y)d\theta$

Prediction : $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$

Probabilities: $\Pr(\theta \in A) = \int_A p(\theta|y)d\theta$. E.g. if $\theta \in [0, \infty)$ then $\Pr(\theta \leq 2) = \int_0^2 p(\theta|y)d\theta$.

- The examples in (2) are special cases of

$$E[h(\theta)] = \int h(\theta)p(\theta|y)d\theta.$$

Expectation: $E[\theta] = \int \theta p(\theta|y)d\theta$. $h(\theta) = \theta$.

Prediction : $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$. $h(\theta) = p(\tilde{y}|\theta)$.

Probabilities: $\Pr(\theta \in A) = \int_A p(\theta|y)d\theta = \int \mathbb{1}_A(\theta)p(\theta|y)d\theta$. $h(\theta) = \mathbb{1}_A(\theta)$,

$$\mathbb{1}_A(\theta) = \begin{cases} 1, & \text{if } \theta \in A, \\ 0, & \text{if } \theta \notin A, \end{cases}$$

- Note: the function of interest is **averaged over the posterior uncertainty** of the parameters.

Grid-based solution to compute $E[h(\theta)]$

- Consider $\theta \in \mathbb{R}$ and form *a grid*

$$\theta^g = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}),$$

  where $\theta^{(1)} < \theta^{(2)} < \cdots < \theta^{(S)}$.

- Important: the grid covers the parameter space where $h(\theta)p(\theta|y) \neq 0$. The expectation is

$$\begin{aligned} E[h(\theta)] &= \int_{\theta^{(1)}}^{\theta^{(S)}} h(\theta)p(\theta|y)d\theta \\ &= \int_{\theta^{(1)}}^{\theta^{(2)}} h(\theta)p(\theta|y)d\theta + \int_{\theta^{(2)}}^{\theta^{(3)}} h(\theta)p(\theta|y)d\theta + \cdots + \int_{\theta^{(S-1)}}^{\theta^{(S)}} h(\theta)p(\theta|y)d\theta \end{aligned}$$

- Let $f(\theta) = h(\theta)p(\theta|y)$,

$$\int_a^b f(\theta)d\theta \approx \text{The area under the curve } f(\theta) \text{ between } a \text{ and } b.$$

Grid-based solution to compute $E[h(\theta)]$, cont.

- Some simple quadrature rules (quadrature = **determining area** *in Latin*)
  - $\int_a^b f(\theta)d\theta \approx (b-a)f(\frac{a+b}{2})$. Midpoint rule. A constant interpolation.
  - $\int_a^b f(\theta)d\theta \approx (b-a)\frac{f(a)+f(b)}{2}$. Trapezoidal rule. A linear interpolation.
- Simpson's rule is obtained with a quadratic interpolation.
- R routines: `gaussquad`, `integrate` (1 dim), `adaptIntegrate` (multi-dimensional).
- Grid-based methods are cursed. Consider $\theta \in \mathbb{R}^p$ and create a grid for each parameter

$$\begin{aligned}
\theta_1^g &= (\theta_1^{(1)}, \theta_1^{(2)}, \ldots, \theta_1^{(S_1)}) \\
&\vdots \\
\theta_p^g &= (\theta_p^{(1)}, \theta_p^{(2)}, \ldots, \theta_p^{(S_p)}).
\end{aligned}$$

- The meshed grid is the tensor product $\theta_1^g \times \theta_2^g \times \ldots \times \theta_p^g$.
- Grows exponentially. Example: If $p = 5$, then 100 grid point in each dimension $\to 100^5$ (10 billion) points on the grid. **The curse of dimensionality.**

Simulation-based solution to compute $E[h(\theta)]$

- Monte Carlo integration to the rescue.
- Suppose we have iid. draws $\{\theta^{(i)}\}_{i=1}^{N}$ from $p(\theta|y)$. By the strong law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \xrightarrow{a.s} E[h(\theta)].$$

- Because of the iid. property I will refer to this as non-Markovian simulation.
- Let I denote the expectation (integral) $E[h(\theta)] = \int h(\theta)p(\theta|y)d\theta$. We estimate it by

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}), \quad \theta^{(i)} \overset{iid.}{\sim} p(\theta|y).$$

- Note that

$$V[\hat{I}] = \frac{\sigma^2}{N}, \quad \text{with } \sigma^2 = V[h(\theta^{(i)})]$$

.

- $V[\hat{I}] \to 0$ (provided σ^2 is bounded). Independent of the dimension of the integral (the number of parameters p).

Our friends revisited with Monte Carlo integration

- Let $\{\theta^{(i)}\}_{i=1}^{N}$ be samples from $p(\theta|y) \propto p(y|\theta)p(\theta)$ (Does not have to be iid.)

- Expectation: $E[\theta] \approx \frac{1}{N}\sum_{i=1}^{N}\theta^{(i)}$.

- Prediction : $p(\tilde{y}|y) \approx \frac{1}{N}\sum_{i=1}^{N} p(\tilde{y}|\theta^{(i)})$.

- Probabilities: $\Pr(\theta \in A) \approx \frac{1}{N}\{\#\theta^{(i)} \text{ draws } \in A\}$.

Simulation of unknown distributions

- **If** we have samples it is (**very**) easy to do **posterior inference**.
- The challenge is to actually **obtain the samples**.
- **Analytic derivations** used so far. **Very cumbersome** - even for simplistic models. Often impossible.
- We start with generating iid. (**non-Markovian**) samples.
  1. The **inverse cdf** for a discrete distribution.
  2. **Rejection sampling**.
- **Note:** Everything I present is **general** for sampling from *any distribution* (not necessarily the **posterior**).
- Since we are **Bayesians** I call the r.v. θ (the parameter) instead of X which you can find in some literature.

The inverse cdf method for a continuous distribution

The inverse cdf for a continuous distribution

Obtain N samples from $F_\theta(\phi) = \Pr(\theta \leq \phi)$. Let F^{-1} denotes the inverse.

- For $i = 1, \ldots, N$, repeat
  1. $u \sim \texttt{uniform}(0, 1)$
  2. $\theta^{(i)} = F_\theta^{-1}(u)$

Proof that we get the correct distribution for θ.

$\Pr(\theta \leq \phi) = \Pr(F_\theta^{-1}(u) \leq \phi) = \Pr(u \leq F_\theta(\phi)) = F_\theta(\phi)$.
This means that

$$\theta \sim F_\theta.$$

$\square$

The inverse cdf for a discrete distribution

- Useful as a **discrete approximation** of a continuous θ.

The inverse cdf for a discretized continuous variable

Obtain N samples from $p(\theta|y)$ known up to a normalizing constant, $\theta \in \mathbb{R}$.

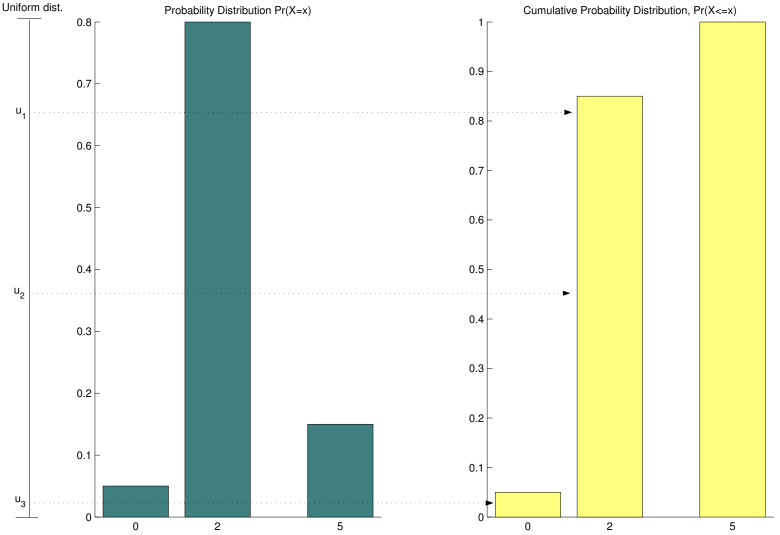
- **Evaluate** $p(y|\theta_j)p(\theta_j)$ for each

$$\theta_j \in (\theta_1, \theta_2, \ldots, \theta_S) \quad \text{(a dense grid)}.$$

- **Normalize** $\hat{f}_j = p(y|\theta_j)p(\theta_j)/\sum_{l=1}^{S} p(y|\theta_l)p(\theta_l)$.
- **Compute** the empirical cdf (cumulative sum) $\hat{F}$ of $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_S)$.
- For $i = 1, \ldots, N$, repeat
  1. $u \sim \texttt{uniform}(0, 1)$
  2. $\theta^{(i)} = \hat{F}^{-1}(u)$

- **Drawback**: a grid, **computationally intractable** for a couple of dimensions.
- Useful for simulating **parts of the posterior** that are one dimensional.

Example of inverse cdf on a discrete sample space $\{0, 2, 5\}$



Recall: Estimating the shrinkage parameter λ

- The normal regression model with **unknown shrinkage**

$$y = X\beta + \varepsilon, \quad \varepsilon \in \mathcal{N}(0, \sigma^2 I)$$

- The **joint posterior** (see priors below) factorizes

$$p(\beta, \sigma^2, \lambda | y) = p(\beta | \sigma^2, \lambda, y) p(\sigma^2 | \lambda, y) p(\lambda | y),$$

$$\begin{array}{rcl}
\textbf{Prior} & \rightarrow & \textbf{Posterior} \\
\beta | \sigma^2, \lambda \sim \mathcal{N}(0, \sigma^2 \Omega_0^{-1}) & \rightarrow & \beta | \sigma^2, \lambda, y \sim \mathcal{N}(\beta_n, \sigma^2 \Omega_n^{-1}) \\
\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2) & \rightarrow & \sigma^2 | \lambda, y \sim \text{Inv-}\chi^2(\nu_n, s_n^2) \\
\lambda \sim p(\lambda) & \rightarrow & \lambda | y \sim \sqrt{\frac{|\Omega_0|}{|\Omega_n|}} \left(\frac{\nu_n s_n^2}{2} \right)^{-\nu_n/2} p(\lambda)
\end{array}$$

and

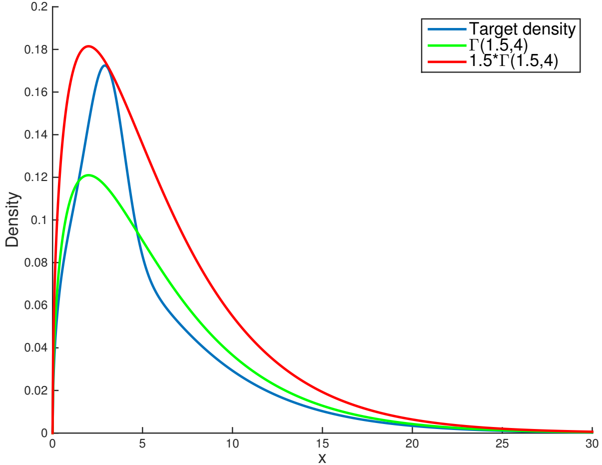
$$\begin{array}{ll}
\beta_n = (X'X + \Omega_0)^{-1} X'y & \Omega_n = X'X + \Omega_0 \\
\nu_n = \nu_0 + n & \nu_n s_n^2 = \nu_0 s_0^2 + y'y - \beta_n' \Omega_n \beta_n
\end{array}$$

- $p(\lambda | y)$ **complex**. Can easily be evaluated on a grid! Inverse cdf to the rescue.

Rejection sampling

- **The setting**:
  - **Not possible to simulate** $p(\theta|y) \propto p(y|\theta)p(\theta)$ directly (not of known form).
  - **We can bound** $p(y|\theta)p(\theta) \leq Mg(\theta), \quad \forall\theta$, where M is **a constant** and $g(\theta)$ is a function with $\int g(\theta)d\theta < \infty$.
  - **We can sample** from a density proportional to $g(\theta)$.

- **The setting**:
  - **Not possible to simulate** $p(\theta|y) \propto p(y|\theta)p(\theta)$ directly (not of known form).
  - **We can bound** $p(y|\theta)p(\theta) \leq Mg(\theta), \quad \forall\theta$, where M is **a constant** and $g(\theta)$ is a function with $\int g(\theta)d\theta < \infty$.
  - **We can sample** from a density proportional to $g(\theta)$.



Rejection sampling

Obtain N samples from $p(\theta|y)$ known up to a normalizing constant.

- Set $\mathtt{i} = 1$.
- While $\mathtt{i} \leq N$ do:
  1. Generate a candidate $\theta' \sim g(\theta)$.
  2. Compute the probality of acceptance

  $$a = \frac{p(y|\theta)p(\theta)}{Mg(\theta)} \quad \text{and draw } u \sim \mathtt{uniform}(0, 1).$$

  3. If $u \leq a \implies \theta^{(i)} = \theta'$, else return to Step 1.
  4. $\mathtt{i} = \mathtt{i} + 1$.

Rejection sampling, cont.

Conditional on acceptance, θ has density $p(\theta|y)$.

For **clarity and simplified** computations consider the ratio $\frac{p(\theta|y)}{Mg(\theta)}$.

$$
\begin{aligned}
\Pr\left(\theta \leq \phi | u \leq \frac{p(\theta|y)}{Mg(\theta)}\right) &= \frac{\Pr\left(\theta \leq \phi, u \leq \frac{p(\theta|y)}{Mg(\theta)}\right)}{\Pr\left(u \leq \frac{p(\theta|y)}{Mg(\theta)}\right)} \\
&= \frac{\int_{-\infty}^{\phi}\int_0^{\frac{p(\theta|y)}{Mg(\theta)}} g(\theta) du d\theta}{\Pr\left(u \leq \frac{p(\theta|y)}{Mg(\theta)}\right)} \\
&= \frac{\int_{-\infty}^{\phi} \frac{p(\theta|y)}{Mg(\theta)} g(\theta) d\theta}{\int_{-\infty}^{\infty} \frac{p(\theta|y)}{Mg(\theta)} g(\theta) d\theta} \\
&= \frac{\frac{1}{M}\int_{-\infty}^{\phi} p(\theta|y) d\theta}{\frac{1}{M}\int_{-\infty}^{\infty} p(\theta|y) d\theta} = \int_{-\infty}^{\phi} p(\theta|y) d\theta.
\end{aligned}
$$

$\square$

Rejection sampling, cont.

- **The density** $g(\theta)$ for **unimodal cases**:
  - **Multivariate t** is a good choice. **Heavy tails** with low degrees of freedom. Let the mean and covariance matrix of $g(\theta)$ match those of the posterior. Use `optim` in R.
  - Choose
    $$M = \sup_{\theta} \frac{p(y|\theta)p(\theta)}{g(\theta)},$$
    gives $a = 1$ at the corresponding θ.
- **Multimodal posterior**: Sample uniformly (at the cost of accepting fewer samples).
- **Drawbacks** of rejection sampling:
  1. If $g(\theta)$ is **"not so proportional"** to $p(y|\theta)p(\theta)$ **few** draws are accepted.
  2. In **high dimensions**: Difficult to find a good M and $g(\theta)$ so that $p(y|\theta)p(\theta) \leq Mg(\theta)$. Making M **too large** gives a **low** the probability of accepting a sample (**acceptance** $\propto \frac{1}{M}$).
- **A precursor** to the **Metropolis-Hastings** algorithm.