Markov chain Monte Carlo algorithms: The Metropolis–Hastings algorithm and Gibbs sampling

Anirban Bhattacharya

http://www.stat.tamu.edu/~anirbanb

March 27, 2018

1 Background: The total variation distance and coupling

Let P and Q be two probability measures on a common probability space $(\mathcal{X}, \mathcal{B})$ (here $\mathcal{B}$ denotes the Borel σ-field on $\mathcal{X}$; we shall typically concern ourselves with situations where $\mathcal{X} \subset \mathbb{R}^d$). Let p and q be their respective densities with respect to a dominating measure μ, i.e., $p = dP/d\mu$ and $q = dQ/d\mu$. Recall the total variation distance between P and Q, $d_{\text{TV}}(P, Q)$, is defined as

$$d_{\text{TV}}(P, Q) = \sup_{B \in \mathcal{B}} |P(B) - Q(B)|.$$

The following representation of the TVD is often handy.

Theorem. Let P and Q be probability measures on $(\mathcal{X}, \mathcal{B})$ with densities p and q respectively. Then,

$$d_{\text{TV}}(P, Q) = \int_{p>q} (p - q)d\mu = \int_{q>p} (q - p)d\mu = \frac{1}{2} \int |p - q|d\mu. \tag{1}$$

Remark. From the above theorem, it is easy to see that the total variation distance satisfies the triangle inequality, i.e., $d_{\text{TV}}(P, R) \leq d_{\text{TV}}(P, Q) + d_{\text{TV}}(Q, R)$. This, along with non-negativity and symmetry implies that the total variation distance is a pseudo-metric on the space of probability measures.

From (1), we can also write

$$d_{\text{TV}}(P, Q) = \frac{1}{2}\left[\int_{p>q} (p - q)d\mu + \int_{q>p} (q - p)d\mu \right] = \frac{1}{2} \int \big[(p \vee q) - (p \wedge q)\big] d\mu. \tag{2}$$

Since $(p \vee q) + (p \wedge q) = p + q$, we also have

$$d_{\text{TV}}(P, Q) = 1 - \int (p \wedge q)d\mu.$$

$\int (p \wedge q)d\mu$ can be thought of as the total variation affinity, similar to the Hellinger affinity.

Another useful representation of the TVD is

$$d_{\text{TV}}(P, Q) = \sup_{f \in \mathcal{F}} \int f(p - q)d\mu,$$

where $\mathcal{F} = \{f : \mathcal{X} \to [0, 1]\}$. Note that the supremum is attained at $f^* = \mathbb{1}(p > q)$.

Coupling. A *coupling* of P and Q is a pair of random variables (X, Y) such that $X \sim P$ and

$Y \sim Q$. Equivalently, a coupling can be thought of as a joint distribution on $\mathcal{X} \times \mathcal{X}$ whose marginals are P an Q respectively. A simplest example of a coupling is $X \sim P, Y \sim Q$ and $X \perp Y$. Let us denote $\mathcal{C} \equiv \mathcal{C}(P, Q)$ to be the collection of all possible couplings of P and Q.

The following result connects couplings with the TVD.

Theorem. We have,

$$d_{\mathrm{TV}}(P, Q) = \inf_{(X,Y)\in\mathcal{C}} \mathbb{P}(X \neq Y). \tag{3}$$

Fix any $A \in \mathcal{B}$. We have, for any $(X, Y) \in \mathcal{C}$,

$$\begin{aligned} P(A) - Q(A) &= P(X \in A) - P(Y \in A) \\ &= P(X \in A, Y \notin A) + P(X \in A, Y \in A) - P(X \notin A, Y \in A) - P(X \in A, Y \in A) \\ &\leq P(X \in A, Y \notin A) \\ &\leq P(X \neq Y). \end{aligned}$$

This immediately tells us $d_{\mathrm{TV}}(X, Y) \leq \inf_{(X,Y)\in\mathcal{C}} \mathbb{P}(X \neq Y)$. It can be shown that the infimum is actually attained by producing a coupling such that $d_{\mathrm{TV}}(P, Q) = P(X \neq Y)$.

2 Markov chain Monte Carlo basics

At the beginning of the semester, we studied the Monte Carlo method for approximating an integral

$$I = \int_{\mathcal{X}} g(x)\gamma(x)dx$$

where γ is a probability density function, and g is some integrable function. Provided with *independent samples* $X_1, \ldots, X_N \sim \gamma$, one can form the Monte Carlo estimate

$$\hat{I}_N = N^{-1} \sum_{i=1}^{N} g(X_i).$$

We saw that $\hat{I}_N$ is unbiased for I and $\hat{I}_N \to I$ almost surely as $N \to \infty$ by the SLLN. Moreover, if G has finite second moment σ_g^2 , then, by the central limit theorem,

$$\sqrt{N}(\hat{I}_N - I) \to N(0, \sigma_g^2) \quad \text{as } n \to \infty.$$

This can be used to provide error bounds for a Monte Carlo procedure as follows. Let

$$\hat{\sigma}_g^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[g(X_i) - \hat{I}_N\right]^2.$$

Then, $\hat{I}_N \pm t_{\alpha/2,N-1}\, \hat{\sigma}_g/\sqrt{N}$ is an approximate $100 \times (1-\alpha)\%$ confidence interval for I.

In many situations, direct simulation from γ is not directly possible/computationally prohibitive. The idea behind Markov chain Monte Carlo (MCMC) is to simulate a Markov chain whose "long-run behavior" mimics that of the target distribution γ. Intuitively, this means that irrespective of where the starting point of the chain is, when run long enough, draws from the chain are approximately distributed as γ. A precise characterization is made below in terms of invariance or stationarity; the chain needs to be constructed in a manner such that γ is its invariant/stationary distribution. To make things concrete, let us recall some basic facts.

3 Markov chain basics

A sequence of random variables $X_0, X_1, X_2, \ldots$ taking values in some state-space $\mathcal{X}$ is called a *Markov chain* if for any $t \geq 1$, the conditional distribution $X_t \mid X_{t-1}, \ldots, X_0$ depends only on the previous time point, i.e., $X_t \mid X_{t-1}, \ldots, X_0 \equiv X_t \mid X_{t-1}$.

Discrete state-space. Assume $X_0 = 0$ and for any $t \geq 1$, $X_t = X_{t-1} + \varepsilon_t$, where the ε_ts are independent of the X_ts and $P(\varepsilon_t = \pm 1) = 1/2$. The state-space $\mathcal{X} = \mathbb{Z}$; the integers. This is an example of a discrete random walk.

Continuous state-space Assume $X_0 \sim N(0,1)$ and $X_t = \rho X_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim N(0,1)$. This is an autoregressive model of lag one. The state-space $\mathcal{X} = \mathbb{R}$. We shall also make the assumption $|\rho| < 1$ later on.

A feature of both examples above is that $P[X_t \in A \mid X_{t-1} = x]$ does not depend on t, i.e., it is the same for all t, for any measurable $A \subset \mathcal{X}$. (find these for the examples above!) We shall only concern ourselves with Markov chains which satisfy this property, which is commonly referred to as *time homogeneity*.

Definition. The transition kernel of a time-homogeneous Markov chain is defined as

$$P(x; A) := P[X_1 \in A \mid X_0 = x], \quad x \in \mathcal{X}, A \subset \mathcal{X}.$$

Since the Markov chain is time-homogeneous, $P(x; A) = P[X_t \in A \mid X_{t-1} = x]$ for any t. We shall assume there is a appropriate dominating measure μ on $\mathcal{X}$ w.r.t. which $P(x, \cdot)$ has a (conditional) density $p(\cdot \mid x)$, i.e.,

$$P(x; A) = \int_A p(x' \mid x)\mu(dx').$$

[If you are not familiar with measure theoretic notation, you can always assume the state-space to be discrete and replace integrals by sums.] The Markovian property allows simple convolution formulae for k-step ahead transitions. For example,

$$P^{(2)}(x; A) := P[X_{t+2} \in A \mid X_t = x] = \int_{\mathcal{X}} P(y; A)p(y \mid x)\mu(dy) = \int_A \underbrace{\left[\int_{\mathcal{X}} p(x' \mid y)p(y \mid x)\mu(dy)\right]}_{p^{(2)}(x'|x)} \mu(dx').$$

In general, for any $r \geq 2$, we can recursively define the r-step transition kernel as

$$P^{(r)}(x; A) = \int_{\mathcal{X}} P^{(r-1)}(y; A)p(y \mid x)\mu(dy) = \int_A p^{(r)}(x' \mid x)\mu(dx),$$

where, by convention, we set $P^{(1)} = P$. $p^{(r)}(\cdot \mid \cdot)$ is called the r-step transition density [write down a recursive expression for $p^{(r)}$]

Finite state-spaces: Let us try to understand the above quantities when $\mathcal{X}$ is finite. In that case, the transition kernel is completely determined by the $|\mathcal{X}| \times |\mathcal{X}|$ *transition* matrix $\mathbb{P} = (p(x' \mid x))_{x,x'}$, where $p(x' \mid x) = P[X_1 = x' \mid X_0 = x]$ as defined above. Note that each row of $\mathbb{P}$ (has non-negative entries and) sum to one. Such matrices are called row-stochastic matrices.

The transition matrix plays and important role since the r-step transition kernels can be expressed as powers of the *transition matrix*. For example,

$$P[X_2 = x' \mid X_0 = x] = \sum_y p(x' \mid y)p(y \mid x) = \mathbb{P}^2(x, x'),$$

where $\mathbb{P}^2$ is the square of the transition matrix $\mathbb{P}$. Clearly, $\mathbb{P}^2$ is also a stochastic matrix with each row summing to one. In general, for any r,

$$P[X_r = x' \mid X_0 = x] = \mathbb{P}^r(x, x').$$

3.1 Stationarity

The aspect of Markov chains we are most interested in is stationarity/invariance. A density γ (w.r.t. μ) is stationary w.r.t. the transition kernel $P(x, \cdot)$ if

$$\gamma(x') = \int p(x' \mid x)\gamma(x)\mu(dx), \, \forall \, x' \in \mathcal{X}.$$

γ is called the invariant/stationary distribution. Note that the above equation implies

$$\int_A \gamma(x')\mu(dx') = \int P(x; A)\gamma(x)\mu(dx).$$

Suppose we draw $X_0 \sim \gamma$. Then, from the definition of stationarity, the marginal density of X_1 is also γ! The same is true for $X_2, X_3, \ldots$. In other words, if we initialize the chain by a draw from the stationary distribution, then the marginal distribution of every successive draw is also the stationary distribution.

Example (finite state-space) For finite state-spaces, the invariance/stationary condition reduces to a linear system of equations:

$$\gamma(x') = \sum_{x \in \mathcal{X}} p(x' \mid x)\gamma(x), \, \forall \, x'$$

If we set $\widetilde{\gamma}$ to be the row vector $(\gamma(x) : x \in \mathcal{X})$, then the above set of equations can be concisely represented as the linear system $\widetilde{\gamma} = \widetilde{\gamma}\,\mathbb{P}$.

Example. Consider the AR(1) example above. Let's try to guess what the stationary distribution is. First, since we know that convolutions of normals produce normal distributions, we restrict our search to normal distributions $N(m, b^2)$. Stationarity implies that if $X_{t-1} \sim N(m, b^2)$ and $X_t \mid X_{t-1} \sim N(\rho X_{t-1}, 1)$, then marginally $X_t \sim N(m, b^2)$. Equating means and variances, we get $m = 0$ and $b^2 = (1 - \rho^2)^{-1}$. Note that for b^2 to be positive, we need $|\rho| < 1$.

Example. Finding the stationary distribution may not always be that trivial. For example, consider the Cauchy AR(1) model $X_t = \rho X_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim Ca(0, v)$, a Cauchy distribution with scale $\sqrt{v}$. The transition density is $p(x' \mid x) \propto \{1 + (x' - \rho x)^2/v\}^{-1}$, and the invariant equation is

$$\gamma(x') = \int_{-\infty}^{\infty} p(x' \mid x)\gamma(x)dx.$$

How would you solve this? [Hint: characteristic function!]

Why do we care about stationarity? Roughly speaking, stationary Markov chains, when run long enough, forget where they started from and the r-step transition densities increasingly start behaving the stationary distribution. Mathematically, for any x, $p^{(r)}(x' \mid x)$ gets increasingly closer to $\pi(x')$ as r increases. [Stronger uniform bounds can be improved under suitable conditions]

Example. Let us return to the AR(1) example again. Verify that for any $r \geq 1$, $X_r \mid X_0 = x \sim N(\rho^r x, (1 - \rho^{2r})/(1 - \rho^2))$. If $|\rho| < 1$, then $\rho^r x \to 0$ and $(1 - \rho^{2r})/(1 - \rho^2) \to 1/(1 - \rho^2)$

as $r \to \infty$. Hence, the distribution of $X_r \mid X_0 = x$ approaches $N(0, (1-\rho^2)^{-1})$ as $r \to \infty$. We noted previously that $N(0, (1-\rho^2)^{-1})$ is indeed the stationary distribution. In fact, we can show

$$\left\| p^{(r)}(\cdot \mid x) - \gamma \right\|_{\text{TV}} \leq C(x)\rho^r,$$

where $C(x)$ is some constant depending on x.

This phenomenon of convergence to stationarity is fairly general.

Convergence theorem (finite state-spaces) Suppose $p(x' \mid x) > 0$ for all $x, x' \in \mathcal{X}$ (this condition means that the chain can move from any point to any other point in one step with positive probability. This condition can be replaced by weaker conditions, though in most MCMC applications the above condition itself is satisfied) and let γ be the stationary distribution (since the state-space is finite, γ is a p.m.f. on $\mathcal{X}$). Then, there exists a constant $\alpha \in (0,1)$ and $C > 0$ such that

$$\max_{x \in \mathcal{X}} \left\| p^{(r)}(\cdot \mid x) - \gamma \right\|_{\text{TV}} \leq C\alpha^r.$$

where, recall the total variation distance

$$\left\| p^{(r)}(\cdot \mid x) - \gamma \right\|_{\text{TV}} = \sum_{x' \in \mathcal{X}} \left| p^{(r)}(x' \mid x) - \gamma(x') \right|.$$

This theorem gives a firm mathematical language to the intuition developed above. Irrespective of where we start, the r-step transition probabilities increasingly behave like the stationary distribution. Observe that the quantity on the right $C\alpha^r$ converges to zero geometrically fast as $r \to \infty$. Accordingly, results as above are referred to as *geometric ergodicity*. Let us now see a statement & proof of the convergence result in a more general framework.

Theorem. Suppose there exists a probability measure Q on $\mathcal{X}$ and a constant $\varepsilon > 0$ such that the following *minorization condition* holds,

$$P(x, A) \geq \varepsilon\, Q(A) \quad \text{for all } x \in \mathcal{X}.$$

Then,

$$\sup_{x \in \mathcal{X}} \|P^r(x, \cdot) - \gamma\|_{\text{TV}} \leq (1-\varepsilon)^r.$$

Note that the condition is always satisfied for finite state-spaces when $p(x' \mid x) > 0$ for all x, x'.

Remark. When such a uniform bound on the TV distance holds, the chain is called *uniformly ergodic*. For general state spaces, this is less common. However, geometric ergodicity can be shown under additional conditions, extending the proof below.

Proof. Let us see a really nice proof based on coupling arguments. Given the minorization condition, write

$$P(x, A) = \varepsilon Q(A) + (1-\varepsilon)R(x, A), \quad x \in \mathcal{X},$$

where $R(x, A) = (1-\varepsilon)^{-1}\big[P(x, A) - \varepsilon Q(A)\big]$. Clearly, $R(x, A) \geq 0$ for all x and A. In fact, it can be easily verified that $R(x, \cdot)$ is a transition kernel in itself. The important upshot is that we can write $P(x, \cdot)$ as a mixture of $Q(\cdot)$ and $R(x, \cdot)$, where Q does not depend on x.

We now construct two Markov chains $\{X_j\}$ and $\{Y_j\}$ which eventually couple with probability one as follows.

1. While $X_j \neq Y_j$,

   - Draw $\delta_j \sim \text{Bernoulli}(\varepsilon)$.
   - If $\delta_j = 0$, draw $X_{j+1} \sim R(X_j, \cdot)$ and $Y_{j+1} \sim R(Y_j, \cdot)$ independently.

- Otherwise, if $\delta_j = 1$, draw $x \sim Q(\cdot)$ and set $X_{j+1} = Y_{j+1} = x$.

2. Once $X_j = x = Y_j$, draw $X_{j+1} = Y_{j+1} \sim P(x, \cdot)$.

Observe that both chains X_j and Y_j are Markov chains with transition kernel $P(\cdot, \cdot)$. Let T denote the *coupling time*, i.e., the time j when $\delta_j = 1$:

$$T = \max\{j : \delta_j = 0\} + 1.$$

Here is the key idea. We start the X chain at x and the Y chain with a draw from γ. Then, $P^r(x, A) = P(X_r \in A)$ and $\gamma(A) = P(Y_r \in A)$. We have, using the coupling bound,

$$\|P^r(x, \cdot) - \gamma\|_{\text{TV}} \leq P(X_r \neq Y_r) = P(T > r) = (1 - \varepsilon)^r.$$

The convergence to stationarity is useful to us as it implies an analogue of the SLLN for dependent chains.

Ergodic Theorem: Perhaps most interesting to our purpose is the following ergodic theorem which states that ergodic averages of ergodic stationary Markov chains converge almost surely to their limiting expectations under the stationary distribution, that is

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \to \int h(x)\pi(x)\mu(dx), \text{ a.s.}$$

for any h such that $\int |h(x)|\pi(x)\mu(dx) < \infty$.

Reversibility. A simple and extremely useful sufficient (but not necessary) condition for a density γ to be the stationary distribution is the following reversibility condition: for any $x, x' \in \mathcal{X}$,

$$p(x' \mid x)\gamma(x) = p(x \mid x')\gamma(x').$$

Note that under reversibility, $\int p(x' \mid x)\gamma(x)\mu(dx) = \int p(x \mid x')\gamma(x')\mu(dx) = \gamma(x')$, which implies that γ satisfies the invariance condition and hence is the stationary distribution.

4 The MH algorithm

The theory of Markov chains has been traditionally concerned with finding conditions under which a given Markov chain converges to its stationary distribution (which may or may not be known) and if so, at what rate. The MCMC literature turns the question on its head. It starts with a given distribution γ and tries to find an ergodic Markov chain whose stationary distribution is γ.

The transition kernel of the MH algorithm is

$$p(x' \mid x) = \alpha(x, x')q(x' \mid x) + r(x)\delta_x(x'),$$

where

$$\alpha(x, x') = \min\left\{1, \frac{\gamma(x')q(x \mid x')}{\gamma(x)q(x' \mid x)}\right\}$$

is the usual MH ratio, and $r(x) = 1 - \int \alpha(x, x')q(x' \mid x)\mu(dx)$ is the probability of staying at x. It is easy to verify that the detailed balance condition $p(x' \mid x)\gamma(x) = p(x \mid x')\gamma(x')$ is satisfied. Hence, $\gamma(\cdot)$ is the stationary distribution of the MH chain!

The MH algorithm is very general in its scope of application. On the other hand, the choice of the proposal distribution typically impacts practical performance. A default choice is an

isotropic Gaussian random walk centered at the previous value and with a common variance σ^2_{MH}. If σ_{MH} is too small, the chain accepts too frequently, and there is a lot of autocorrelation. On the other hand, if σ_{MH} is too big, acceptances may get very rare. A default rule of thumb is to tune the proposal variance so that the acceptance rate is between 20 - 30 %.

If some information is available regarding the covariance matrix of the parameters to be sampled, one can use a non-isotropic random walk, with the covariance matrix a constant multiple of the (approximate) covariance matrix of parameters.

There are advanced versions of the MH algorithm, for example, hybrid or Hamiltonian Monte Carlo, that used gradient information of γ. Another way of exploiting additional structure in MH is the Gibbs sampler.

5 The Gibbs sampler

Gibbs sampling refers to a class of Markov chain Monte Carlo algorithms where one samples iteratively from the *full conditional* distributions to create a Markov chain whose stationary distribution is the posterior distribution.

Why full conditionals? Suppose (U, V) have a joint density $\gamma(u, v)$. Then, under mild conditions,

$$\gamma_V(v) \int \frac{\gamma(u \mid v)}{\gamma(v \mid u)} du = 1,$$

which implies we can recover the marginal distributions $\gamma_V(v)$ (and similarly, $\gamma_U(u)$) from the full conditionals. Since we already know the conditionals, combined with the marginals, they determine the joint density.

For example, for $|\rho| < 1$, if $u \mid v \sim N(\rho v, (1 - \rho^2))$ and $v \mid u \sim N(\rho v, (1 - \rho^2))$, then you can verify with a bit of tedious calculation that $u \sim N(0, 1)$ and $v \sim N(0, 1)$.

Note: the existence of the joint density is important. Simply specifying the full conditionals does not mean that they arise from a valid joint distribution. For example, if $u \mid v$ and $v \mid u$ are both exponentials, with $E(u \mid v) = v^{-1}$ and $E(v \mid u) = u^{-1}$, then the integral diverges, i.e., there is no joint distribution on $\mathbb{R}^2$ whose full conditionals are the above distributions.

5.1 Why does the Gibbs sampler work? Normal example

Suppose (u, v) have a bivariate normal distribution with $u \sim N(0, 1)$, $v \sim N(0, 1)$ and $\text{corr}(u, v) = \rho$. Let $r = 1 - \rho^2$. From standard bivariate normal theory, we know $u \mid v \sim N(\rho v, r)$ and $v \mid u \sim N(\rho u, r)$.

A Gibbs sampler proceeds as:

- Initialize $u = u^{(0)}$.
- For $t = 1, \ldots, T$, repeat:
  - Sample $v^{(t)} \sim N(\rho u^{(t-1)}, r)$ by letting $v^{(t)} = \rho u^{(t-1)} + \epsilon^{(t)}$, where $\epsilon^{(t)} \sim N(0, r)$ is independent of everything else.
  - Sample $u^{(t)} \sim N(\rho v^{(t)}, r)$ by letting $u^{(t)} = \rho v^{(t)} + \eta^{(t)}$, where $\eta^{(t)} \sim N(0, r)$ is independent of everything else.

If $u^{(0)} \sim N(0, 1)$, then it follows from a simple calculation that $v^{(1)} \sim N(0, 1)$, $u^{(1)} \sim N(0, 1)$ and in fact $u^{(1)}$ and $v^{(1)}$ have a bivariate normal distribution with correlation ρ. In fact, this is true for every $u^{(t)}$ and $v^{(t)}$. This has to be the case since $N(0, 1)$ is the stationary distribution

of the chain $u^{(t)}$.

In general, $u^{(t)} \mid u^{(t-1)} \sim N(\rho^2 u^{(t-1)}, r(1+\rho^2))$. Similar for $v^{(t)}$. Thus, $\{u^{(t)}\}$ is an AR(1) process. Verify that its stationary distribution is $N(0,1)$. [stationary variance $= r(1-\rho^2)/(1-\rho^4) = 1$]

5.2 Why does the Gibbs sampler work in general

In general, suppose (u, v) have a joint distribution γ. The Gibbs sampler proceeds as

- Initialize $u = u^{(0)}$.
- For $t = 1, \ldots, T$, repeat:
  - Sample $v^{(t)} \sim \gamma(\cdot \mid u^{(t-1)})$.
  - Sample $u^{(t)} \sim \gamma(\cdot \mid v^{(t)})$.

This is clearly a Markov process with transition density $q\big((u', v') \mid (u, v)\big) = \gamma(v' \mid u)\gamma(u' \mid v')$. A couple lines of calculation show us that

$$\gamma(u', v') = \int q\big((u', v') \mid (u, v)\big)\ \gamma(u, v) du dv,$$

that is, γ is indeed the stationary distribution of the Gibbs chain.

5.3 Some basic Gibbs samplers

Binomial with unknown sample size: Suppose $y \mid N, p \sim \text{Binomial}(N, p)$, where N and p are both unknown. Consider independent priors on N and p, with $N \sim \text{Poisson}(\lambda)$, and $p \sim U(0,1)$. The full conditionals are:

- $p \mid N, y \sim \text{Beta}(y+1, N-y+1)$.
- To sample $N \mid p, y$, set $N = y + t$, with $t \sim \text{Poisson}(\lambda(1-p))$. [Verify.]

Linear regression with ridge prior: Suppose $y \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 \mathrm{I}_n)$, where X is a $n \times d$ matrix of covariates and $\beta \in \mathbb{R}^d$ is a vector of covariates. Consider the following prior specification on β, σ^2, with $\beta \mid \sigma^2 \sim N(0, \lambda^{-1}\sigma^2 \mathrm{I}_d)$ and $\sigma^2 \sim \text{IG}(\alpha/2, \gamma/2)$. λ, α, β are hyperparameters which we fix. The full conditionals are:

- $\beta \mid \sigma^2, y \sim N_d((X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1} X^{\mathrm{T}} y, \sigma^2 (X^{\mathrm{T}}X + \lambda \mathrm{I}_d)^{-1})$. [Note: this is of the $N(Q^{-1}b, Q^{-1})$ form.]
- $\sigma^2 \mid \beta, y \sim \text{IG}((n + d + \alpha)/2, \{(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda \beta^{\mathrm{T}}\beta + \gamma\}/2)$.