Advanced Simulation Methods

Chapter 5 - Gibbs Sampling

In this chapter, we will start describing Markov chain Monte Carlo methods. These methods are used to approximate high-dimensional expectations

$$\mathbb{E}_\pi(\phi(X)) = \int_{\mathbb{X}} \phi(x)\,\pi(x)\,dx$$

and do not rely on independent samples from π, or on the use of importance sampling. Instead, the samples are obtained by simulating a Markov chain whose stationary distribution is π. Gibbs sampling and Metropolis-Hastings constitute the two main Markov chain Monte Carlo methods, from which most of the other methods derive. We start with the Gibbs sampler.

1 Motivating Example: A Bayesian Hierarchical Model

The following (real) data give the number of failures (p_i) over time intervals (t_i) of ten nuclear pumps.

Pump i	1	2	3	4	5
# Failures p_i	5	1	5	14	3
Times t_i	94.32	15.72	62.88	125.76	5.24

Pump i	6	7	8	9	10
# Failures p_i	19	1	1	4	22
Times t_i	31.44	1.05	1.05	2.10	10.48

We model the failures of the $i-$th pump as a Poisson process with parameter λ_i, thus, during an observation period of length t_i, the number of failures P_i follows a Poisson distribution of parameters $\lambda_i t_i$. We are interested in inferring the parameters $\lambda_{1:10} = (\lambda_1, \ldots, \lambda_{10})$ from the data. We follow a hierarchical Bayesian approach where we assume that, conditional upon some hyperparameters (α, β), $(\lambda_1, \ldots, \lambda_{10})$ are independent and follow a prior gamma distribution $\mathcal{G}a(\alpha, \beta)$ with density

$$p(\lambda_i|\,\beta) = \frac{\beta^{\alpha-1}}{\Gamma(\alpha)}\lambda_i^{\alpha-1}\exp(-\beta\lambda_i).$$

We assume that β follows itself a prior gamma distribution $\mathcal{G}a(\gamma, \delta)$. The other hyperparameters (α, γ, δ) are fixed to constant values ($\alpha = 1.8$, $\gamma = 0.01$ and $\delta = 1$), and hence are omitted from the conditioning arguments.

In this context, the joint distribution of $\lambda_{1:10}, \beta, P_{1:10}$ is

$$p(\lambda_{1:10}, \beta, p_{1:10}) = p(\beta)\,p(\lambda_i|\,\beta)\prod_{i=1}^{10}\frac{(\lambda_i t_i)^{p_i}}{p_i!}\exp(-\lambda_i t_i)$$

and Bayesian inference relies on

$$p(\lambda_{1:10}, \beta|\,p_{1:10}) = \frac{p(\lambda_{1:10}, \beta, p_{1:10})}{\int p(\lambda_{1:10}, \beta, p_{1:10})\,d\lambda_{1:10}d\beta}.$$

This multidimensional distribution is rather complex. It is not obvious how the rejection method or importance sampling could be efficiently used in this context. However the conditional distributions $p(\lambda_{1:10}|\,p_{1:10}, \beta)$ and $p(\beta|\,p_{1:10}, \lambda_{1:10})$ admit standard parametric forms. Indeed, we have

$$p(\lambda_{1:10}|\,p_{1:10}, t_{1:10}, \beta) = \prod_{i=1}^{10} p(\lambda_i|\,p_i, \beta)$$

where

$$\lambda_i|(\beta, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta) \text{ for } 1 \leq i \leq 10, \tag{1}$$

and $p(\beta|\, p_{1:10}, \lambda_{1:10}) = p(\beta|\, \lambda_{1:10})$ where

$$\beta|(\lambda_1, \ldots, \lambda_{10}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i). \tag{2}$$

In other words, these conditional distributions have a simpler form than the joint distribution on all the parameters. Hence, instead of directly sampling the vector $(\lambda_1, \ldots, \lambda_{10}, \beta)$ at once, one could suggest sampling it alternately, starting for example with the λ_i's for a given guess of β, followed by an update of β given the new samples $\lambda_1, \ldots, \lambda_{10}$.

This sampling strategy raises several important questions.

- Is the joint distribution uniquely specified by the conditional distributions?
- Sampling alternately from these conditional distributions yields a Markov chain: the newly proposed values only depend on the present values and not the past values. Does this provide a Markov chain with the correct stationary distribution? Does the Markov chain converge towards this invariant distribution?

We will see that the answers to both questions is yes under certain conditions.

2 Algorithm

The Gibbs sampler is a very popular technique in Monte Carlo simulation to sample from high-dimensional distributions. Assume you are interested in sampling from the target density

$$\pi(x) = \pi(x_1, x_2, ..., x_d).$$

We use the standard notation $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$.

Algorithm. Systematic scan Gibbs sampler. Let $\left(X_1^{(1)}, ..., X_d^{(1)}\right)$ be the initial state then iterate for $t = 2, 3, ...$

1. Sample $X_1^{(t)} \sim \pi_{X_1|X_{-1}}\left(\cdot|\, X_2^{(t-1)}, ..., X_d^{(t-1)}\right)$.

...

j. Sample $X_j^{(t)} \sim \pi_{X_j|X_{-j}}\left(\cdot|\, X_1^{(t)}, ..., X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)}\right)$.

...

d. Sample $X_d^{(t)} \sim \pi_{X_d|X_{-d}}\left(\cdot|\, X_1^{(t)}, ..., X_{d-1}^{(t)}\right)$.

The conditional distributions used in the Gibbs sampler are often referred to as full conditionals. A popular alternative to the systematic scan Gibbs sampler is the random scan Gibbs sampler.

Algorithm. Random scan Gibbs sampler. Let $\left(X_1^{(1)}, ..., X_d^{(1)}\right)$ be the initial state then iterate for $t = 2, 3, ...$

1. Sample an index J from a distribution on $\{1, ..., d\}$ (typically uniform).
2. Sample $X_J^{(t)} \sim \pi_{X_J|X_{-J}}\left(\cdot|\, X_1^{(t-1)}, ..., X_{J-1}^{(t-1)}, X_{J+1}^{(t-1)}, ..., X_d^{(t-1)}\right)$ and set $X_k^{(t)} := X_k^{(t-1)}$ for $k \neq J$.

Remark. It should be clear that several Gibbs samplers can be defined for a target distribution. Consider for example $\pi(w, y, z)$ where $w, y, z \in \mathbb{R}$ then we can partition (w, y, z) into 3 components $(x_1 = w, x_2 = y, x_3 = z)$ or in 2 components: $x_1 = (w, y)$, $x_2 = z$ or $x_1 = (w, z)$, $x_2 = y$ or $x_1 = (y, z)$, $x_2 = w$. As a rule of thumb, we usually favour Gibbs samplers where the number of components d is the smallest.

3 The Hammersley-Clifford Theorem

An important property of full conditionals is that they fully specify the joint distribution under some weak regularity conditions. This fundamental result was established in an unpublished Oxford technical report by Hammersley and Clifford in 1970.

Definition 3.1. *A distribution with density $\pi(x_1, x_2, ..., x_d)$ and marginal densities $\pi_{X_i}(x_i)$ is said to satisfy the positivity condition if for all $x_1, ..., x_d$ such that $\pi_{X_i}(x_i) > 0$ we have $\pi(x_1, x_2, ..., x_d) > 0$.*

This condition implies that the support of the joint density is the Cartesian product of the support of the marginal densities.

Theorem 3.1. ***(Hammersley-Clifford)*** *Consider a distribution whose density $\pi(x_1, x_2, ..., x_d)$ satisfies the positivity condition. Then for any $(z_1, ..., z_d) \in supp(\pi)$, i.e. $\pi(z_1, ..., z_d) > 0$, we have*

$$\pi(x_1, x_2, ..., x_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(x_j | x_1, ..., x_{j-1}, z_{j+1}, ..., z_d)}{\pi_{X_j|X_{-j}}(z_j | x_1, ..., x_{j-1}, z_{j+1}, ..., z_d)}$$

Proof. We have

$$\pi(x_1, x_2, ..., x_d) = \pi_{X_d|X_{-d}}(x_d | x_1, ..., x_{d-1})\, \pi(x_1, x_2, ..., x_{d-1})$$

and similarly

$$\pi(x_1, x_2, ..., x_{d-1}, z_d) = \pi_{X_d|X_{-d}}(z_d | x_1, ..., x_{d-1})\, \pi(x_1, x_2, ..., x_{d-1}).$$

Hence

$$\begin{aligned}
\pi(x_1, x_2, ..., x_d) &= \pi(x_1, x_2, ..., x_{d-1}, z_d) \frac{\pi_{X_d|X_{-d}}(x_d | x_1, ..., x_{d-1})}{\pi_{X_d|X_{-d}}(z_d | x_1, ..., x_{d-1})} \\
&= \cdots \\
&= \pi(z_1, ..., z_d) \frac{\pi_{X_1|X_{-1}}(x_1 | z_2, ..., z_d)}{\pi_{X_1|X_{-1}}(z_1 | z_2, ..., z_d)} \cdots \frac{\pi_{X_d|X_{-d}}(x_d | x_1, ..., x_{d-1})}{\pi_{X_d|X_{-d}}(z_d | x_1, ..., x_{d-1})}.
\end{aligned}$$

The positivity condition ensures that the conditional densities we introduce are non-zero. ■

It is important to notice that Hammersley-Clifford theorem assumes that $\pi(x_1, x_2, ..., x_d)$ is a well defined probability density. Not every set of full conditionals is compatible; i.e. there is no guarantee that they define a probability density.

Example 3.1. *Consider the following conditionals $\pi_{X_1|X_2}(x_1 | x_2) = x_2 \exp(-x_2 x_1)$ (i.e. an exponential distribution of parameter x_2) and $\pi_{X_1|X_2}(x_1 | x_2) = x_1 \exp(-x_1 x_2)$, defined on $\mathbb{R}^+$. We might expect that these full conditionals define a well defined joint probability density $\pi(x_1, x_2)$. However, Hammersley-Clifford would give*

$$\begin{aligned}
\pi(x_1, x_2, ..., x_d) &\propto \frac{\pi_{X_1|X_2}(x_1 | z_2)}{\pi_{X_1|X_2}(z_1 | z_2)} \frac{\pi_{X_2|X_1}(x_2 | x_1)}{\pi_{X_2|X_1}(z_2 | x_1)} \\
&= \frac{z_2 \exp(-z_2 x_1)\, x_1 \exp(-x_1 x_2)}{z_2 \exp(-z_2 z_1)\, x_1 \exp(-x_1 z_2)} \\
&\propto \exp(-x_1 x_2).
\end{aligned}$$

The problem is that $\int\int \exp(-x_1 x_2)\, dx_1 dx_2$ is not finite, so $\pi_{X_1|X_2}(x_1 | x_2) = x_2 \exp(-x_2 x_1)$ and $\pi_{X_1|X_2}(x_1 | x_2) = x_1 \exp(-x_1 x_2)$ are not compatible.

Since the samples generated by the Gibbs sampler constitute a Markov chain, we can use the tools introduced in the previous lecture notes to study the properties of the chain. A rich literature exists on the theoretical properties of the Gibbs sampler under various conditions. We simply state the main properties, namely that the generated samples allow the estimation of integrals with respect to π. The fact that Gibbs sampling, and in general Markov chain Monte Carlo methods, are more efficient in high dimension than, say, importance sampling, is beyond the scope of this course; but keep in mind that it is essentially the reason why those methods are so popular.

4 Convergence of the Gibbs sampler

We first give the transition kernel of the Gibbs sampler. If $x^{(t)} := \left(x_1^{(t)}, ..., x_d^{(t)}\right)$ then the kernel of the systematic scan Gibbs sampler is simply

$$K\left(x^{(t-1)}, x^{(t)}\right) = \pi_{X_1|X_{-1}}\left(x_1^{(t)} \middle| x_2^{(t-1)}, ..., x_d^{(t-1)}\right) \times \pi_{X_2|X_{-2}}\left(x_2^{(t)} \middle| x_1^{(t)}, x_3^{(t-1)}, ..., x_d^{(t-1)}\right) \times \cdots \\ \times \pi_{X_d|X_{-d}}\left(x_d^{(t)} \middle| x_1^{(t)}, ..., x_{d-1}^{(t)}\right). \tag{3}$$

For the random scan Gibbs sampler, where we pick the index j of the component to be updated uniformly at random, we have

$$K\left(x^{(t-1)}, x^{(t)}\right) = \frac{1}{d}\sum_{j=1}^{d} \pi_{X_j|X_{-j}}\left(x_j^{(t)} \middle| x_{-j}^{(t-1)}\right) \delta_{x_{-j}^{(t-1)}}\left(x_{-j}^{(t)}\right) \tag{4}$$

where $\delta_{x_{-j}^{(t-1)}}$ denotes the Dirac mass located at $x_{-j}^{(t-1)}$. The transition kernel (4) does not admit a density with respect to the Lebesgue measure.

Proposition 4.1. *The systematic scan Gibbs sampler kernel (3) admits $\pi(x_1, x_2, ..., x_d)$ as invariant distribution.*

Proof. Indeed, we can prove that $\int_{\mathbb{X}} \pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) dx^{(t-1)} = \pi\left(x^{(t)}\right)$. To simplify expressions, we limit ourselves to the case $d = 2$. We have

$$\begin{aligned}
&\int \pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) dx^{(t-1)} \\
&= \int \pi\left(x^{(t-1)}\right) \pi_{X_1|X_{-1}}\left(x_1^{(t)} \middle| x_2^{(t-1)}\right) \times \pi_{X_2|X_{-2}}\left(x_2^{(t)} \middle| x_1^{(t)}\right) dx_1^{(t-1)} dx_2^{(t-1)} \\
&= \int \pi_{X_2}\left(x_2^{(t-1)}\right) \pi_{X_1|X_{-1}}\left(x_1^{(t)} \middle| x_2^{(t-1)}\right) \times \pi_{X_2|X_{-2}}\left(x_2^{(t)} \middle| x_1^{(t)}\right) dx_2^{(t-1)} \text{ (integrate } x_1^{(t-1)}\text{)} \\
&= \int \pi\left(x_1^{(t)}, x_2^{(t-1)}\right) \times \pi_{X_2|X_{-2}}\left(x_2^{(t)} \middle| x_1^{(t)}\right) dx_2^{(t-1)} \text{ (as } \pi(x_1, x_2) = \pi_{X_2}(x_2)\, \pi_{X_1|X_{-1}}(x_1|\, x_2)\text{)} \\
&= \pi_{X_1}\left(x_1^{(t)}\right) \times \pi_{X_2|X_{-2}}\left(x_2^{(t)} \middle| x_1^{(t)}\right) \text{ (integrate } x_2^{(t-1)}\text{)} \\
&= \pi\left(x_1^{(t)}, x_2^{(t)}\right).
\end{aligned}$$

The proof is very similar for $d > 2$. ■

Remark: We note that this transition kernel is *not* reversible; i.e. for $d = 2$ we have

$$\pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) \neq \pi\left(x^{(t)}\right) K\left(x^{(t)}, x^{(t-1)}\right).$$

A similar type of proof establishes that the random scan Gibbs sampler kernel (4) also admits $\pi(x_1, x_2, ..., x_d)$ as invariant distribution. It can be additionally proven that this kernel is reversible.

Establishing that the transition kernel admits π as invariant distribution is not sufficient. In particular, we also need the Markov chain to be π-irreducible if we want to have a law of large numbers. It is easy to find example on which the Gibbs sampling chain is not irreducible.

Example 4.1. *Reducible Gibbs sampler. Let $\pi(x_1, x_2)$ be the uniform density on $([-1, 0] \times [-1, 0]) \cup ([0, 1] \times [0, 1])$. For positive values of x_1, the conditional $\pi_{X_2|X_1}(x_2|\, x_1)$ is supported on $[0, 1]$. Similarly for positive values of x_2, $\pi_{X_1|X_2}(x_1|\, x_2)$ is supported on $[0, 1]$. Hence if we start in the positive quadrant, the algorithm will never reach $[-1, 0] \times [-1, 0]$. The chain cannot be π-irreducible. In fact for this problem, although π is a stationary distribution, there are infinitely many different stationary distributions corresponding to arbitrary convex mixtures of the uniform distributions on $([-1, 0] \times [-1, 0])$ and on $([0, 1] \times [0, 1])$. The target and the first steps of the chain generated by a Gibbs sampler are illustrated in Figure 1*

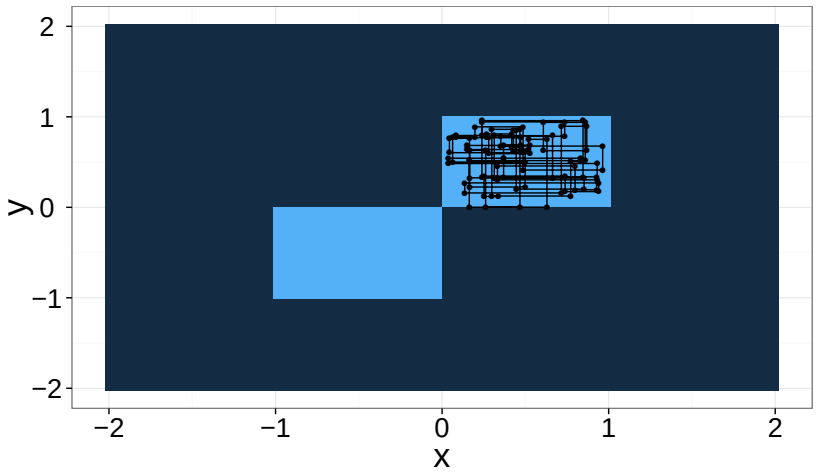

Figure 1: Gibbs sampler on a target π that does not satisfy the positivity condition.

Proposition 4.2. *Assume $\pi(x_1, x_2, ..., x_d)$ satisfies the positivity condition, then the systematic scan Gibbs sampler yields a $\pi-$irreducible and recurrent Markov chain.*

Proof. For any set A such that $\pi(A) := \int_A \pi(x_1, ..., x_d)\, dx_1 \dots dx_d > 0$, we have

$$\begin{aligned}\mathbb{P}\left(X^{(t)} \in A \middle| X^{(t-1)} = x^{(t-1)}\right) &= \int_A K\left(x^{(t-1)}, x^{(t)}\right) dx^{(t)} \\ &= \int_A \pi_{X_1|X_{-1}}\left(x_1^{(t)} \middle| x_2^{(t-1)}, ..., x_d^{(t-1)}\right) \times ... \times \pi_{X_d|X_{-d}}\left(x_d^{(t)} \middle| x_1^{(t)}, ..., x_{d-1}^{(t)}\right) dx^{(t)}\end{aligned}$$

where $\pi_{X_1|X_{-1}}\left(x_1^{(t)} \middle| x_2^{(t-1)}, ..., x_d^{(t-1)}\right), ..., \pi_{X_d|X_{-d}}\left(x_d^{(t)} \middle| x_1^{(t)}, ..., x_{d-1}^{(t)}\right) > 0$ on a set of non-zero measure. Hence we can conclude that

$$\mathbb{P}\left(X^{(t)} \in A \middle| X^{(t-1)} = x^{(t-1)}\right) > 0.$$

It follows that the chain is π-irreducible and actually strongly π-irreducible. We have already established that this kernel admits π as stationary distribution, hence it is also recurrent. ■

It is also the case that if the transition kernel is absolutely continuous with respect to the dominating measure of the target distribution, then π-irreducibility and π-invariance implies Harris recurrence (Theorem 2, Tierney 1994). Hence we have the following theorem.

Theorem 4.1. *Assume the Markov chain generated by the systematic scan Gibbs sampler is $\pi-$irreducible and recurrent (both conditions hold when the positivity condition is satisfied) then we have for any integrable function $\phi : \mathbb{X} \to \mathbb{R}$:*

$$\lim \frac{1}{t} \sum_{i=1}^{t} \phi\left(X^{(i)}\right) = \int_{\mathbb{X}} \phi(x)\, \pi(x)\, dx$$

for $\pi-$almost all starting value $X^{(1)}$.

This result ensures that we can approximate expectations $\mathbb{E}_\pi(\phi(X))$ using a single Markov chain. However this does not guarantee that for a finite number of samples t the approximation will be good.

Example. Assume we are interested in sampling from a simple bivariate normal distribution; i.e. $X := (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2)$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}.$$

It is easy to establish that the Gibbs sampler proceeds as follows in this case

1. Sample $X_1^{(t)} \sim \mathcal{N}\left(\mu_1 + \rho/\sigma_2^2\left(X_2^{(t-1)} - \mu_2\right), \sigma_1^2 - \rho^2/\sigma_2^2\right)$

2. Sample $X_2^{(t)} \sim \mathcal{N}\left(\mu_2 + \rho/\sigma_1^2\left(X_1^{(t)} - \mu_1\right), \sigma_2^2 - \rho^2/\sigma_1^2\right)$.

By proceeding this way, we generate a Markov chain $X^{(t)}$ whose successive samples are correlated. If successive values of $X^{(t)}$ are strongly correlated, then we say that the Markov chain mixes slowly.

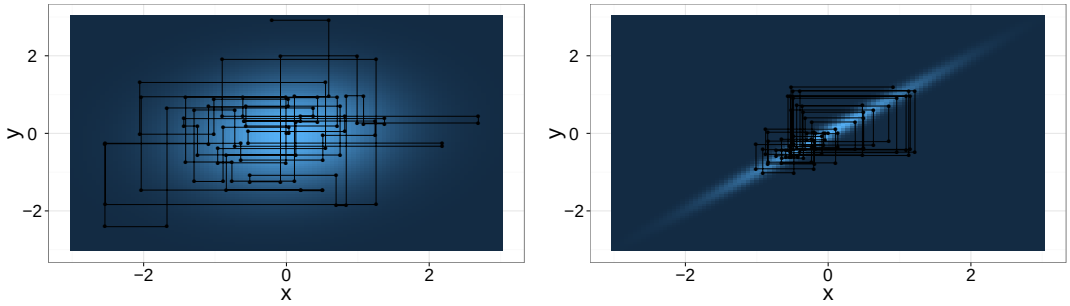

Figure 2: Gibbs sampler on a bivariate normal distribution. Left: $\rho = 0.1$, right $\rho = 0.99$.

We consider the scenario $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$. As $|\rho| \to 1$, we $\sigma_2^2 - \rho^2/\sigma_1^2 \to 0$ and the chain will move very slowly. Figure 2 illustrates this phenomenon, by plotting the first steps of a chain produced by Gibbs sampling, for two values $\rho = 0.1$ and $\rho = 0.99$.

5 Data Augmentation

It is only possible to use Gibbs sampling when we can sample from the full conditionals. For many target distributions of interest, this is not feasible. Thankfully in many scenarios of interest, in particular when dealing with statistical models, it is possible to include a set of auxiliary variables $Z_1, ..., Z_p$ and an associated probability distribution whose joint density $\overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)$ satisfies

$$\int \overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)\, dz_1 ... dz_d = \pi(x_1, ..., x_d)$$

and which is such that its full conditionals are easy to sample. Additionally for many statistical models, these auxiliary variables have a "natural" interpretation.

5.1 Bayesian Inference for Mixture of Gaussians

Mixture of Gaussians are commonly used to model non-normal data. Figure

Assume you have independent data $Y_1, ..., Y_n$ and each observation might come from one of K components/populations. We assume that the distribution within the k-th population is a normal $\mathcal{N}\left(\mu_k, \sigma_k^2\right)$, and the probability of coming from the k-population is p_k. As we do not observe from which populations the observations are coming, we have

$$Y_i | \, \theta \sim \sum_{k=1}^{K} p_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where $\theta = \left(p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2\right)$. We are interested in inferring the parameter θ from the data. In a Bayesian framework, we set a prior

$$p(\theta) = p(p_1, ..., p_K) \prod_{k=1}^{K} p\left(\mu_k, \sigma_k^2\right)$$

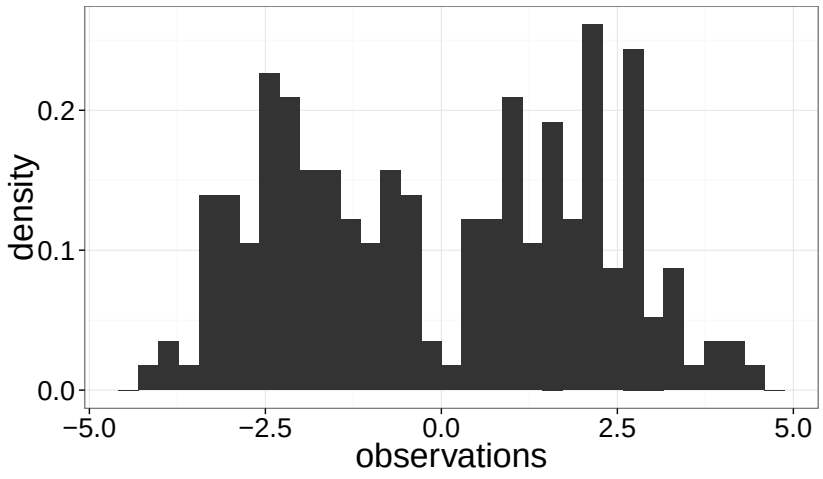

Figure 3: Typical observations that seem to come from a mixture model, here with two normal components.

where $p\left(p_1, ..., p_K\right)$ is a so-called Dirichlet distribution $\mathcal{D}\left(\gamma_1, ..., \gamma_K\right)$ where $\gamma_1, ..., \gamma_K > 0$

$$p\left(p_1, ..., p_K\right) = \frac{\Gamma\left(\sum_{k=1}^K \gamma_k\right)}{\prod_{k=1}^K \Gamma\left(\gamma_k\right)} \prod_{k=1}^K p_k^{\gamma_k - 1}$$

which is defined on the simplex $\left\{(p_1, ..., p_K) : p_i \geq 0 \text{ for any } i,\ \sum_{k=1}^K p_k = 1\right\}$. We also use

$$p\left(\mu_k, \sigma_k^2\right) = p\left(\mu_k \middle| \sigma_k^2\right) p\left(\sigma_k^2\right)$$

where $p\left(\mu_k \middle| \sigma_k^2\right) = \mathcal{N}\left(\mu_k; \alpha_k, \frac{\sigma_k^2}{\lambda_k}\right)$ and $p\left(\sigma_k^2\right) = \mathcal{IG}\left(\sigma_k^2; \frac{\lambda_k+3}{2}, \frac{\beta_k}{2}\right)$.

It appears very difficult to sample directly from the posterior $p\left(\theta \middle| y_1, ..., y_n\right)$, and it is unclear how one could implement a Gibbs sampling algorithm in this context. However, we can introduce some auxiliary variables $Z_1, ..., Z_n$ which tells us from which population data ith is coming from, i.e.

$$\mathbb{P}\left(Z_i = k\right) = p_k \text{ and } \left. Y_i \right| Z_i = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right).$$

Now we consider the extended target distribution $p\left(\theta, z_1, ..., z_n \middle| y_1, ..., y_n\right)$ which is such that

$$\begin{aligned} p\left(\theta, z_1, ..., z_n \middle| y_1, ..., y_n\right) \propto & \left(\prod_{i=1}^n \frac{p_{z_i}}{\sigma_{z_i}} \exp\left(-\frac{\left(y_i - \mu_{z_i}\right)^2}{2\sigma_{z_i}^2}\right)\right) \prod_{k=1}^K p_k^{\gamma_k - 1} \\ & \times \prod_{k=1}^K \exp\left(-\frac{\lambda_k \left(\mu_k - \alpha_k\right)^2}{2\sigma_k^2}\right) \left(\frac{1}{\sigma_k^2}\right)^{\frac{\lambda_k+3}{2}-1} \exp\left(-\frac{\beta_k}{2\sigma_k^2}\right). \end{aligned}$$

We can implement a Gibbs sampler updating $(\theta, z_{1:n})$ by sampling alternately from $\mathbb{P}\left(z_{1:n} \middle| y_{1:n}, \theta\right)$, $p\left(p_{1:n} \middle| y_{1:n}, z_{1:n}, \mu_{1:K}, \sigma_{1:K}^2\right) = p\left(p_{1:n} \middle| z_{1:n}\right)$ and $p\left(\mu_{1:K}, \sigma_{1:K}^2 \middle| y_{1:n}, z_{1:n}, p_{1:n}\right)$. We have

$$\mathbb{P}\left(z_{1:n} \middle| y_{1:n}, \theta\right) = \prod_{i=1}^n \mathbb{P}\left(z_i \middle| y_i, \theta\right)$$

where

$$\mathbb{P}\left(z_i \middle| y_i, \theta\right) = \frac{\frac{p_{z_i}}{\sigma_{z_i}} \exp\left(-\left(y_i - \mu_{z_i}\right)^2 / \left(2\sigma_{z_i}^2\right)\right)}{\sum_{k=1}^K \frac{p_k}{\sigma_k} \exp\left(-\left(y_k - \mu_k\right)^2 / \left(2\sigma_k^2\right)\right)}$$

and introducing

$$n_k = \sum_{i=1}^{n} \mathbf{1}_{\{k\}}(z_i), n_k \overline{y}_k = \sum_{i=1}^{n} x_i \mathbf{1}_{\{k\}}(z_i), s_k^2 = \sum_{i=1}^{n} (y_i - \overline{y}_k)^2 \mathbf{1}_{\{k\}}(z_i)$$

then we have the full conditionals

$$\begin{aligned}
&p_1, ..., p_K | z_{1:n} \sim \mathcal{D}(\gamma_1 + n_1, ..., \gamma_K + n_K), \\
&\sigma_k^2 | z_{1:n}, y_{1:n} \sim \mathcal{IG}\left(\tfrac{\lambda_k + n_k + 3}{2}, \tfrac{\lambda_k s_k^2 + \beta_k + s_k^2 - (\lambda_k + n_k)^{-1}(\lambda_k \alpha_k + n_k \overline{x}_k)^2}{2}\right), \\
&\mu_k | \sigma_k^2, z_{1:n}, y_{1:n} \sim \mathcal{N}\left(\tfrac{\lambda_k \alpha_k + n_k \overline{x}_k}{\lambda_k + n_k}, \tfrac{\sigma_k^2}{\lambda_k + n_k}\right).
\end{aligned}$$

It is thus easy to implement the Gibbs sampler in this scenario.

5.2 Bayesian Probit Regression

Assume that you have access to some data $(x_i, y_i)_{i=1,...,n}$ where $x_i \in \mathbb{R}^d$ is a set of covariates and $y_i \in \{0, 1\}$. A standard regression approach for binary responses consists of using the logistic regression. We present here an alternative known as probit regression. In probit regression, we have

$$\mathbb{P}(Y = 1 | x, \beta) = \Phi\left(x^T \beta\right)$$

where $\beta \in \mathbb{R}^d$ is a set of regression coefficients and $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-v^2/2\right) dv$. Assume we additionally assign a normal prior $p(\beta) = \mathcal{N}(\beta; \mu, \Sigma)$.

Given data $(x_i, y_i)_{i=1,...,n}$, Bayesian inference relies on the posterior

$$p(\beta | y_1, ..., y_n) \propto p(\beta) \prod_{i=1}^{n} \Phi\left(x_i^T \beta\right)^{y_i} \left(1 - \Phi\left(x_i^T \beta\right)\right)^{1 - y_i}.$$

In this scenario, it is unclear how one could perform efficient rejection or importance sampling. Gibbs sampling does not appear to apply either.

Now assume that we associate to each observation (x_i, Y_i) a latent/auxiliary variable Z_i such that

$$\begin{aligned}
&Z_i \sim \mathcal{N}\left(x_i^T \beta, 1\right), \\
&Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

We have now defined a joint distribution

$$p(y_i, z_i | \beta, x_i) = p(y_i | z_i) p(z_i | x_i, \beta).$$

such that marginally, when we integrate out z_i, we have

$$\begin{aligned}
\mathbb{P}(Y_i = 1 | x, \beta) &= \int p(y_i, z_i | \beta, x_i) dz_i \\
&= \int_0^{\infty} \mathcal{N}\left(z_i; x_i^T \beta, 1\right) dz_i = \Phi\left(x_i^T \beta\right).
\end{aligned}$$

We now propose to sample from the extended posterior

$$p(\beta, z_1, \ldots, z_n | y_1, ..., y_n) \propto p(\beta) \prod_{i=1}^{n} p(z_i | x_i, \beta) \prod_{i=1}^{n} p(y_i | z_i).$$

To achieve this, we can use Gibbs sampling as the full conditionals $p(\beta | y_1, ..., y_n, z_1, \ldots, z_n) = p(\beta | z_1, \ldots, z_n)$ and $p(z_1, \ldots, z_n | y_1, ..., y_n, \beta) = \prod_{i=1}^{n} p(z_i | y_i, \beta)$ are standard with

$$p(\beta | z_1, \ldots, z_n) = \mathcal{N}\left(\beta; \widetilde{\mu}, \widetilde{\Sigma}\right)$$

where $\widetilde{\Sigma}^{-1} = \Sigma^{-1} + \sum_{i=1}^{n} x_i x_i^T$, $\widetilde{\mu} = \widetilde{\Sigma}\left(\Sigma^{-1}\mu + \sum_{i=1}^{n} x_i z_i\right)$ and

$$Z_i | y_i, \beta \sim \begin{cases} \mathcal{N}_+\left(x_i^T \beta, 1\right) & \text{if } y_i = 1 \\ \mathcal{N}_-\left(x_i^T \beta, 1\right) & \text{if } y_i = 0. \end{cases}$$

where $\mathcal{N}_+\left(\mu, \sigma^2\right)$, resp. $\mathcal{N}_-\left(\mu, \sigma^2\right)$, is a normal $\mathcal{N}\left(\mu, \sigma^2\right)$ restricted to $(0, \infty)$, resp. $(-\infty, 0)$.