Stat260: Bayesian Modeling and Inference **Lecture Date: April 12, 2010**

Lecture 18: Variable Augmentation & MCMC Foundations

Lecturer: Michael I. Jordan *Scribe: Joshua Paul*

1 Variable Augmentation

Recall that Gibbs sampling is frequently used in the following two (related) settings:

- It is difficult or impossible to sample $x = (x_1, \ldots, x_p)$ directly, but is possible to conditionally sample $x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ for all $i = 1, \ldots, p$.
- It is difficult or impossible to sample x directly, but there exists a "latent variable" y such that it is possible to conditionally sample $x \mid y$ and $y \mid x$.

The latter setting is referred to as *variable augmentation*. Observe that, in this case, the Gibbs sampler returns samples of the form (x, y); marginalizing to obtain samples x may be accomplished by simply ignoring the y component of each (x, y) pair. We consider several concrete examples of variable augmentation.

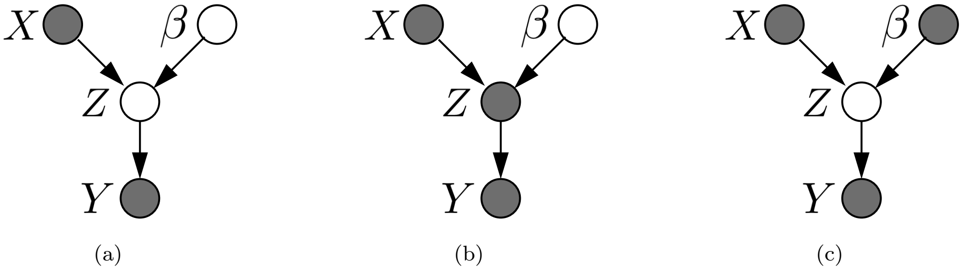

Figure 1: Probit regression model augmented with variable $Z = (Z_1, \ldots, Z_N)$: (a) with the observed data shaded (b) 'observed' data associated with the Gibbs conditional $\beta \mid X, Y, Z$ shaded (c) 'observed' data associated with the Gibbs conditional $Z \mid \beta, X, Y$ shaded.

Example 1 (Probit Regression)**.** A probit model assigns to each $X_i \in \mathbb{R}^p$ a random variable $Y_i \in \{0, 1\}$ using the law $\Pr(Y_i = 1) = \Phi(\beta^{\mathrm{T}} X_i)$, where Φ is the cumulative distribution function for the standard normal. Assuming (for convenience) a flat prior on β, and $X = (X_1, \ldots, X_N)$ with outcomes $Y = (Y_1, \ldots, Y_N)$:

$$p(\beta \mid X, Y) \propto p(Y \mid X, \beta) p(\beta) = \prod_{i=1}^{N} \Phi(\beta^{\mathrm{T}} X_i)^{Y_i} (1 - \Phi(\beta^{\mathrm{T}} X_i))^{1-Y_i}. \tag{1}$$

There is no clear way to sample β from the form given in Eq. (1). Thus, consider the model *augmented* with a random variable $Z = (Z_1, \ldots, Z_N)$, where $Z_i = \beta^{\mathrm{T}} X_i + \varepsilon_i$ with $\varepsilon \sim \mathcal{N}(0, 1)$ and $Y_i = \mathbb{I}(Z_i > 0)$ (see Figure 1(a)). Observe that under the augmented model,

$$\Pr(Y_i = 1 \mid X_i, \beta) = \Pr(Z_i > 0 \mid X_i, \beta) = \Pr(\varepsilon_i < \beta^{\mathrm{T}} X_i \mid X_i, \beta) = \Phi(\beta^{\mathrm{T}} X_i),$$

observed value	0	1	2	3	≥ 4
random count	x_1	x_2	x_3	x_4	x_5
realization	139	128	55	25	13

Table 1: Data for a Poisson model with truncation. 360 samples were drawn from a Poi(λ) distribution, with values as summarized by the table.

and so the models are identical. Moreover, the appropriate conditional sampling formulae are relatively straightforward to obtain. Observe that $\beta \mid Z, X, Y$ is not dependent on Y (see Figure 1(b)), and is a simple linear regression model. In particular, since the prior on β is flat,

$$\beta \mid Z, X \sim \mathcal{N}\left(\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}Z, \left(X^{\mathrm{T}}X\right)^{-1}\right), \tag{2}$$

where Z and X are regarded as matrices. On the other hand, $Z_i \mid \beta, X_i, Y_i$ would be normal, except we are also conditioning on Y_i, which indicates the sign of Z_i (see Figure 1(c)). Thus, $Z_i \mid \beta, X_i, Y_i$ is a *truncated normal*:

$$Z_i \mid \beta, X_i, Y_i \sim \begin{cases} \mathrm{TN}(\beta^{\mathrm{T}}X_i, 1, -\infty, 0), & \text{if } \; Y_i = 0, \\ \mathrm{TN}(\beta^{\mathrm{T}}X_i, 1, 0, \infty), & \text{if } \; Y_i = 1, \end{cases} \tag{3}$$

where the truncated normal distribution is parameterized by the mean and variance of the corresponding normal distribution along with restricted domain $(-\infty, 0)$ or $(0, \infty)$ that Z_i belongs to.

The truncated normal distribution can be efficiently sampled from using rejection sampling; depending on the domain, a reasonable envelope function to use is either a normal or a shifted exponential. Thus, Gibbs sampling, via Eq. (2) and Eq. (3), can be used to generate a posterior sample (β, Z); ignoring Z yields a posterior sample β.

Example 2 (Student's t-distribution)**.** Sampling directly from the Student's t-distribution is difficult. However, variable augmentation may be used to provide an efficient sampler for the t-distribution. The details of this scheme are left as an exercise.

Example 3 (Censored Data)**.** In some cases, due to measurement or experimental constraints, the "full" data is unavailable, having been coarsened, truncated, or otherwise censored. A concrete example is *survival analysis*, which aims to determine the effect of a drug on life-span; if the study proceeds for a fixed amount of time, surviving subjects represent truncated data in the sense that only a minimum life-span is known when the study concludes. In such cases it may be difficult to sample from the posterior for the parameter of interest directly. Augmenting with variables representing the true values of the truncated or coarsened data, coupled with Gibbs sampling, may provide posterior samples.

Consider a Poisson model with truncation, for which a particular outcome is shown in Table 1. In particular, suppose the data is drawn i.i.d.from Poi(λ) with $p(\lambda) \propto 1/\lambda$. Then the posterior density for λ is

$$\begin{aligned} p(\lambda \mid x) \propto p(x \mid \lambda)p(\lambda) &\propto \underbrace{\left(\lambda^{\sum_{i=1}^{4}(i-1)\cdot x_i} e^{-\lambda \sum_{i=1}^{4} x_i}\right)}_{x_1,x_2,x_3,x_4} \cdot \underbrace{\left(1 - \sum_{i=0}^{3} e^{-\lambda}\frac{\lambda^i}{i!}\right)^{x_5}}_{x_5} \cdot \frac{1}{\lambda} \\ &= \left(\lambda^{313}e^{-347\lambda}\right) \cdot \left(1 - \sum_{i=0}^{3} e^{-\lambda}\frac{\lambda^i}{i!}\right)^{13} \cdot \frac{1}{\lambda} \end{aligned}$$

There is no clear way to sample λ from this form. Thus, consider the model *augmented* with "pseudo-observations" $y_1, \ldots, y_{13}$, for the truncated (censored) data. That is, for each of the thirteen censored data points (x_5), introduce y_i which will correspond to the "true" life-span, observing that each $y_i \geq 4$. Since y_i

is unobserved, we must sample it based on the given data. Then

$$p(\lambda \mid x, y) \propto \lambda^{\sum_{i=1}^{4}(i-1)\cdot x_i + \sum_{j=1}^{13} y_i} e^{-\lambda\left(13+\sum_{i=1}^{4} x_i\right)} \frac{1}{\lambda} \quad \Rightarrow \quad \lambda \mid x, y \sim \text{Ga}\left(313 + \sum_{j=1}^{13} y_i, 360\right). \tag{4}$$

As before, $y_i \mid \lambda, x$ would be Poisson, but the model indicates that $y_i \geq 4$. Thus,

$$p(y_i \mid \lambda, x) \propto e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} \cdot \mathbb{I}(y_i \geq 4) \Rightarrow y_i \mid \lambda, x \sim \text{TP}(\lambda, 4, \infty), \tag{5}$$

where the truncated Poisson may be sampled using rejection sampling with, for example, a Poisson or exponential envelope. Together, Eq. (4) and Eq. (5) can be used in a Gibbs sampler, from which posterior samples for $\lambda \mid x$ can ultimately be obtained.

2 MCMC Foundations

General Markov chains may operate in either discrete or continuous time with discrete or continuous state space. In particular, MCMC methods typically construct discrete-time chains with continuous state space, and so this is the relevant portion of Markov chain theory. The remainder of this section introduces several key definitions and theorems. A more thorough treatment of the subject can be found in Meyn and Tweedie (1993).

Definition 4 (transition kernel)**.** Let $\mathcal{X}$ be the (continuous) state space of a Markov process, and $\mathcal{B}(\mathcal{X})$ the Borel subsets of $\mathcal{X}$. Then a transition kernel is a function $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$, such that for all $x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure on $\mathcal{X}$.

The transition kernel governs the evolution of the Markov chain. In particular, for $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, the Markov chain $(X_n)_{n=1,2,\ldots}$ associated with K has the property:

$$\Pr(X_{t+1} \in A \mid X_1 = x_1, \ldots, X_t = x_t) = K(x_t, A) = \int_A K(x_t, dy).$$

The n-th step transition kernel may be calculated as follows:

$$\Pr(X_{t+n} \in A \mid X_1 = x_1, \ldots, X_t = x_t) = K^{(n)}(x_t, A) = \int_{\mathcal{X}} K(x_t, dy) K^{(n-1)}(y, A),$$

where $K^{(1)}(x, A) = K(x, A)$.

Definition 5 (stopping time and number of passages)**.** Consider a Markov chain (X_n) and let $A \in \mathcal{B}(\mathcal{X})$. Define the (random) stopping time $\tau_A = \min\{n : X_n \in A\}$, and the (random) number of passages $\eta_A = \sum_{n=1}^{\infty} \mathbb{I}(X_n \in A)$.

Recall that in discrete-space Markov chain theory, if $\Pr(\tau_{\{y\}} < \infty \mid X_1 = x) > 0$ for all $x, y \in \mathcal{X}$, then the chain (X_n) is said to be *irreducible*. In the continuous-space case, an alternative definition is required.

Definition 6. (ψ-irreducibility) Let (X_n) be a Markov chain with state space $\mathcal{X}$ and $\psi : \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ be a measure. Then if for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ such that $\psi(A) > 0$, there exists an n such that $K^{(n)}(x, A) > 0$, then the (X_n) is said to ψ-irreducible.

Definition 7 (recurrence)**.** Let (X_n) be a Markov chain with state space $\mathcal{X}$. A set $A \in \mathcal{B}(\mathcal{X})$ is said to be recurrent if $\mathbb{E}[\eta_A \mid X_1 = x] = \infty$ for all $x \in \mathcal{X}$. Correspondingly, the chain (X_n) is said to be recurrent if there exists a measure ψ such that (X_n) is ψ-irreducible and for all $A \in \mathcal{B}(\mathcal{X})$ such that $\psi(A) > 0$, A is recurrent.

Remark 8 (Harris recurrence). In addition to the definition of recurrence just given, there is a stronger notion, Harris recurrence. A set $A \in \mathcal{B}(\mathcal{X})$ is Harris recurrent if $\Pr(\eta_A = \infty \mid X_1 = x) = 1$ for all $x \in \mathcal{X}$, and the Markov chain (X_n) is Harris recurrent as in the above definition, appropriately adjusted. It turns out that Harris recurrence is often a sufficient condition when ordinary recurrence is too weak.

Definition 9 (invariance)**.** Let (X_n) be a Markov chain with state space $\mathcal{X}$. A σ-finite measure $\pi : \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ is said to be invariant if, for all $A \in \mathcal{B}(\mathcal{X})$, $\pi(A) = \int_{\mathcal{X}} K(x, A)\pi(dx)$.

Roughly, an invariant measure represents a “distribution” over the state space, which does not change when the Markov chain is iterated. Possessing an invariant measure, particularly an invariant probability measure, is an important property of a Markov chain:

Definition 10 (positivity)**.** A ψ-irreducible Markov chain (X_n) is called positive (with respect to ψ) if it has an invariant *probability* measure.

The following theorem (proved in Meyn and Tweedie (1993)) relates the notions of invariance, positivity, and recurrence:

Theorem 11. *Let (X_n) be a Markov chain. Then:*

1. *If (X_n) is positive, then (X_n) is recurrent.*
2. *If (X_n) is recurrent, then (X_n) possesses a σ-finite invariant measure.*

References

Meyn, S. and Tweedie, R. (1993). *Markov chains and stochastic stability.* Springer-Verlag, London.