# VE414 Lecture 14

Jing Liu

UM-SJTU Joint Institute

October 24, 2019

- Identifying a good proposal distribution in 1-dimensional is fairly simple.
- In *high dimensions*, it is very difficult to find a good proposal for rejection or importance sampling scheme; thus alternatives must be derived.

Q: What is the difference between direct and indirect sampling scheme so far?

- Markov Chain Monte Carlo (MCMC) circumvent a proposal distribution in high dimensions by no sampling from the true target distribution

$$f_{\mathbf{Y}}$$

it aims instead at sampling from a sequence of approximations which have

$$f_{\mathbf{Y}}$$

as their limiting distribution as the number of iterations grows to infinity.

- MCMC generates correlated simulations instead of independent ones.

- Consider the following model of $n$ independent random variables

$$X_i \sim \begin{cases} \mathrm{Poisson}\left(\lambda_1\right) & \text{for} \quad i = 1, \ldots, k \\ \mathrm{Poisson}\left(\lambda_2\right) & \text{for} \quad i = k+1, \ldots, n \end{cases}$$

- Using a conjugate prior for $\lambda_\ell$,

$$\lambda_\ell \sim \mathrm{Gamma}\left(\alpha_\ell, \beta_\ell\right)$$

the joint posterior is given by

$$f_{\{\lambda_1, \lambda_2, K\}|\{X_1, \ldots X_n\}} = \left( \prod_{i=1}^{k} \frac{\exp\left(-\lambda_1\right) \lambda_1^{x_i}}{x_i!} \right) \cdot \left( \prod_{i=k+1}^{n} \frac{\exp\left(-\lambda_2\right) \lambda_2^{x_i}}{x_i!} \right) \cdot \frac{\lambda_1^{\alpha_1 - 1} \beta_1^{\alpha_1}}{\Gamma\left(\alpha_1\right)} \exp\left(-\beta_1 \lambda_1\right) \cdot \frac{\lambda_2^{\alpha_2 - 1} \beta_2^{\alpha_2}}{\Gamma\left(\alpha_2\right)} \exp\left(-\beta_2 \lambda_2\right)$$

where we assume $K$ is unknown and follows a discrete uniform prior.

Q: How to obtain a sample of $\{\lambda_1, \lambda_2, K\}$ according to the joint posterior

$$f_{\{\lambda_1,\lambda_2,K\}|\{X_1,\dots X_n\}} = \left(\prod_{i=1}^{k} \frac{\exp(-\lambda_1)\,\lambda_1^{x_i}}{x_i!}\right) \cdot \left(\prod_{i=k+1}^{n} \frac{\exp(-\lambda_2)\,\lambda_2^{x_i}}{x_i!}\right) \cdot \frac{\lambda_1^{\alpha_1-1}\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \exp(-\beta_1\lambda_1) \cdot \frac{\lambda_2^{\alpha_2-1}\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \exp(-\beta_2\lambda_2)$$

- At the moment, other than sampling direction according to a 3-dimensional grid, we don't have any other way to sample from a multivariate distribution.
- Notice the 1-dimensional conditional posteriors are easy to identify

$$f_{\lambda_1|\{X_1,\dots X_n,\lambda_2,K\}} \sim \mathrm{Gamma}\left(\alpha_1 + \sum_{i=1}^{k} x_i, \beta_1 + k\right)$$

$$f_{\lambda_2|\{X_1,\dots X_n,\lambda_1,K\}} \sim \mathrm{Gamma}\left(\alpha_2 + \sum_{i=k+1}^{n} x_i, \beta_2 + n - k\right)$$

$$f_{K|\{X_1,\dots X_n,\lambda_1,\lambda_2\}} \propto \lambda_1^{\sum_{i=1}^{k} x_i} \lambda_2^{\sum_{i=k+1}^{n} x_i} \exp((\lambda_2 - \lambda_1)\cdot k)$$

- You might be tempted to sample from the conditionals, but the immediate problem follows that idea is what values to conditioning on, e.g. which $k$ in

$$f_{\lambda_1|\{X_1,\ldots X_n,\lambda_2,K\}} \sim \mathrm{Gamma}\left(\alpha_1 + \sum_{i=1}^{k} x_i, \beta_1 + k\right)$$

  should we use to reflect the dependency between $\lambda_1$ and $k$ specified by

$$f_{\{\lambda_1,\lambda_2,K\}|\{X_1,\ldots X_n\}}$$

- Unless all components are independent, having a sample from a joint density

$$f_{\mathbf{Y}}$$

  is not the same as having multiple samples from its conditionals,

$$f_{Y_j|Y_{-j}} = f_{Y_j|\{Y_1,\ldots,Y_{j-1},Y_{j+1},\ldots,Y_p\}} \qquad \text{where} \quad j = 1, 2, \ldots, p$$

  one for each $j$, and arbitrarily putting them together to form a single sample.

- In general, a full set of 1-dimensional conditional density functions, e.g.

$$f_{X_1|X_2} \qquad \text{and} \qquad f_{X_2|X_1}$$

might not even uniquely define a joint density function, i.e.

$$f^{*}_{X_1,X_2} = f_{X_1|X_2} \cdot f_{X_2}$$
$$f^{**}_{X_1,X_2} = f_{X_2|X_1} \cdot f_{X_1}$$

are the same only if the marginals are chosen with respect to the same joint

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_2$$
$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_1$$

Q: Under what condition is the joint defined by the conditionals unique?

Theorem (Hammersley-Clifford)

*If the joint probability density function is positive*

$$f_{\{Y_1,\dots,Y_p\}}\left(y_1,\dots y_p\right) > 0$$

*for all* $y_1,\dots,y_p$ *when the marginal probability density functions are positive*

$$f_{Y_i}\left(y_i\right) > 0$$

*then we have*

$$f_{\{Y_1,\dots,Y_p\}}\left(y_1,\dots y_p\right) \propto \prod_{j=1}^{p} \frac{f_{Y_j|Y_{-j}}\left(y_j \mid y_1,\dots,y_{j-1},\xi_{j+1},\dots,\xi_p\right)}{f_{Y_j|Y_{-j}}\left(\xi_j \mid y_1,\dots,y_{j-1},\xi_{j+1},\dots,\xi_p\right)}$$

*for all* $\xi_1,\dots,\xi_n \in \mathcal{D}$*.*

**Proof**

Q: What is the significance of this theorem?

- Firstly, the last theorem is precisely what we need regarding uniqueness, but it does not guarantee the existence of the joint probability, that we need to be given or determine using some other ways. To see what I mean, consider

$$Y_1 \mid Y_2 \sim \text{Exponential}\left(\lambda y_2\right) \quad \text{and} \quad Y_2 \mid Y_1 \sim \text{Exponential}\left(\lambda y_1\right)$$

- Applying the last theorem, we have

$$\begin{aligned} f_{Y_1,Y_2}\left(y_1,y_2\right) &\propto \frac{f_{Y_1|Y_2}(y_1 \mid \xi_2)}{f_{Y_1|Y_2}(\xi_1 \mid \xi_2)} \cdot \frac{f_{Y_2|Y_1}(y_2 \mid y_1)}{f_{Y_2|Y_1}(\xi_2 \mid y_1)} \\ &= \frac{\lambda\xi_2 \exp\left(-\lambda\xi_2 y_1\right) \cdot \lambda y_1 \exp\left(-\lambda y_1 y_2\right)}{\lambda\xi_2 \exp\left(-\lambda\xi_2\xi_1\right) \cdot \lambda y_1 \exp\left(-\lambda y_1 \xi_2\right)} \propto \exp\left(-\lambda y_1 y_2\right) \end{aligned}$$

- However, the following integral is not finite,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\lambda y_1 y_2\right)\, dy_1\, dy_2$$

thus there is no proper joint distribution behind the two conditionals.

- Secondly, the last theorem provides very little in terms of how to sample from the conditionals so that we can obtain a sample from the joint.

Q: How to obtain ANY sample from ANY one of the conditionals?

- In general, we have unknowns in the conditional densities, e.g.

$$f_{\lambda_1|\{X_1,\ldots X_n,\lambda_2,K\}} \sim \mathrm{Gamma}\left(\alpha_1 + \sum_{i=1}^{k} x_i, \beta_1 + k\right)$$

$$f_{\lambda_2|\{X_1,\ldots X_n,\lambda_1,K\}} \sim \mathrm{Gamma}\left(\alpha_2 + \sum_{i=k+1}^{n} x_i, \beta_2 + n - k\right)$$

$$f_{K|\{X_1,\ldots X_n,\lambda_1,\lambda_2\}} \propto \lambda_1^{\sum_{i=1}^{k} x_i} \lambda_2^{\sum_{i=k+1}^{n} x_i} \exp\left(\left(\lambda_2 - \lambda_1\right) \cdot k\right)$$

- If we arbitrarily choose $k$ when sample $\lambda_1$, and $\lambda_2$, then arbitrarily choose $\lambda_1$ and $\lambda_2$ when sample $k$, we will loose the dependency amongst them.
- It is only sensible to sample from the conditionals alternatingly conditioning on previous sample values to establish some dependency amongst them.

**Algorithm 1:** GIBBS SAMPLING

**Input** : functions $f_{Y_1|Y_{-1}}$, $f_{Y_2|Y_{-2}}$, …, $f_{Y_p|Y_{-p}}$, values $y_1^{(0)}$, …, $y_p^{(0)}$, size $n$

**Output** : sample array $[y_i^{(t)}]_{n\times p}$

**Function** `Gibbs`($f_{Y_1|Y_{-1}}$, $f_{Y_2|Y_{-2}}$, …, $f_{Y_p|Y_{-p}}$, $y_1^{(0)}$, …, $y_p^{(0)}$, $n$):

  **for** $t \leftarrow 1$ **to** $n$ **do**

    **for** $j \leftarrow 1$ **to** $p$ **do**

      $y_j^{(t)} \sim f_{Y_j|Y_{-j}}\left(\cdot \mid y_1^{(t)} \cdots y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, \cdots y_p^{(t-1)}\right)$

      `/* draw from the conditionals */`

    **end for**

  **end for**

  **return** $\left[y_i^{(t)}\right]_{n\times p}$ ; `/* samples */`

**end**

- Gibbs sampling seems very sensible, however, we yet to show the sequence

$$\{\mathbf{Y}^{(0)}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)}, \cdots, \mathbf{Y}^{(n)}\}$$

relates to a distribution, let alone having anything to do with the joint.

- Notice there is a dependency between components within each iteration

$$\mathbf{Y}^{(t)}$$

and there is a dependency between

$$\mathbf{Y}^{(t-1)} \qquad \text{and} \qquad \mathbf{Y}^{(t)}$$

- However, given $\mathbf{Y}^{(t-1)}$, there is no dependency between

$$\mathbf{Y}^{(t-2)} \qquad \text{and} \qquad \mathbf{Y}^{(t)}$$

that is, the following two densities are equivalent,

$$f_{\mathbf{Y}^{(t)}|\{\mathbf{Y}^{(t-1)},\mathbf{Y}^{(t-2)}\}} = f_{\mathbf{Y}^{(t)}|\mathbf{Y}^{(t-1)}}$$

- In fact, Gibbs sampling scheme essentially leads to a Markov chain

$$\{\mathbf{Y}^{(0)}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)}, \cdots, \mathbf{Y}^{(n)}\}$$

- However, unlike what have been covered before where

$$\{X_t\}$$

has a discrete state space, and following the Markov property is satisfied

$$\Pr(X_{t+1} = j \mid X_t, X_{t-1}, \ldots, X_0) = \Pr(X_{t+1} = j \mid X_t)$$

- The Markov Chain corresponding to Gibbs has a continuous state space

$$\mathcal{D} \subset \mathbb{R}^p$$

- In this case, the probability is defined over a set of values

$$\Pr(\mathbf{Y} \in \mathcal{A}) = \int_{\mathcal{A}} f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}$$

where $\mathcal{A}$ is a subset of the continuous state space $\mathcal{D}$.

- A process $\{\mathbf{Y}^{(t)}\}$ on a continuous state space $\mathcal{D}$ is a Markov Chain if

$$\Pr\left(\mathbf{Y}^{(t)} \in \mathcal{Y} \mid \mathcal{B}\right) = \Pr\left(\mathbf{Y}^{(t)} \in \mathcal{Y} \mid \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}\right)$$

  for any $\mathcal{Y} \subset \mathcal{D}$ and $\mathcal{B} = \{\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, \ldots, \mathbf{Y}^{(0)} = \mathbf{y}^{(0)}\}$.

- The transition kernel of the Gibbs sampling scheme is given by

$$\begin{aligned}\kappa\left(\mathbf{y}^{(t-1)}, \mathbf{y}^{(t)}\right) = f_{Y_1|Y_{-1}}&\left(y_1^{(t)} \mid y_2^{(t-1)}, \ldots y_p^{(t-1)}\right) \\ &\cdot f_{Y_2|Y_{-2}}\left(y_2^{(t)} \mid y_1^{(t)}, y_3^{(t-1)} \ldots y_p^{(t-1)}\right) \cdots \\ &\cdot f_{Y_p|Y_{-p}}\left(y_p^{(t)} \mid y_1^{(t)}, \ldots y_{p-1}^{(t)}\right)\end{aligned}$$

  it is the function when integrated with respect to the current state gives the conditional probability of getting from the previous state $\mathbf{y}^{(t-1)}$ to $\mathbf{y}^{(t)} \in \mathcal{Y}$.

$$\Pr\left(\mathbf{Y}^{(t)} \in \mathcal{Y} \mid \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}\right) = \int_{\mathcal{Y}} \kappa\left(\mathbf{y}^{(t-1)}, \mathbf{y}^{(t)}\right) d\mathbf{y}^{(t)}$$