VE414 Lecture 11

Jing Liu

UM-SJTU Joint Institute

June 18, 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

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

- Using a conjugate prior for λ_ℓ,

$$\text{Gamma}\left(\alpha_\ell, \beta_\ell\right)$$

the joint posterior is given by

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

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

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

- 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

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

- 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

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

should we use to reflect the dependency between λ_1 and k specified by

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

- 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,

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

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.

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

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

$$\begin{aligned} f^*_{X_1,X_2} &= f_{X_1|X_2} \cdot f_{X_2} \\ f^{**}_{X_1,X_2} &= f_{X_2|X_1} \cdot f_{X_1} \end{aligned}$$

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

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

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

Theorem (Hammersley-Clifford)

If the joint probability density function being positive

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

guarantees the marginal probability density functions are also positive

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

for all $y_1,\ldots,y_n$ in the support $\mathcal{D}$ of the joint distribution, then we have

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

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

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.

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

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

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

- If we arbitrarily choose k when sample λ_1, and λ_2, then arbitrarily choose λ_1 and λ_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 so-called Markov chain

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

- However, unlike what is covered by elementary courses where a process

$$\left\{X^{(n)}\right\}$$

with a discrete state space is defined as a Markov Chain if the probability

$$\Pr\left(X^{(n)} = j \mid X^{(n-1)} = i_{n-1}, X^{(n-2)} = i_{n-2}, \ldots, X^{(0)} = i_0\right)$$

is equal to the probability

$$\Pr\left(X^{(n)} = j \mid X^{(n-1)} = i_{n-1}\right)$$

- The Markov Chain corresponding to Gibbs is on a continuous state space

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

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

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

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

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

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

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 $\mathbf{y}^{(t)} \in \mathcal{Y}$.

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

Theorem

The joint distribution $f_{\mathbf{Y}}$ is the invariant distribution of the Markov Chain

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

generated by the Gibbs sampling scheme, it is invariant in the sense that

$$\mathbf{Y}^{(t)} \sim f_{\mathbf{Y}} \qquad \text{whenever} \qquad \mathbf{Y}^{(t-1)} \sim f_{\mathbf{Y}}$$

Proof

- Note the above theorem does not guarantee a sample generated by Gibbs follows the joint, it merely states the joint is the invariant distribution.
- To fully understand the situation we need to have a better understanding on

  invariant distribution

- It can be understood as the equilibrium distribution, we still "jump" around as t changes, but the distribution of being in certain states stay the same.

- Consider a small town, in which 30% of the married women get divorced each year and 20% of the single women get married each year.

$$\mathbf{w}_1 = \mathbf{A}\mathbf{w}_0 \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix} \quad \text{and} \quad \mathbf{w}_0 = \begin{bmatrix} 800 \\ 200 \end{bmatrix}$$

Q: Consider the following Julia outputs, what do you notice?

```
julia> A = [0.7 0.2; 0.3 0.8]
```

```
2*2 Array{Float64,2}:
 0.7   0.2
 0.3   0.8
```

```
julia> w0 = [800; 200]
```

```
2-element Array{Int64,1}:
 800
 200
```

```julia
julia> A * w0
```

```
2-element Array{Float64,1}:
 600.0
 400.0
```

```julia
julia> A^2 * w0
```

```
2-element Array{Float64,1}:
 499.99999999999994
 500.0
```

```julia
julia> A^4 * w0
```

```
2-element Array{Float64,1}:
 425.0
 575.0
```

```
julia> A^8 * w0
```

```
2-element Array{Float64,1}:
 401.5625000000006
 598.4375000000002
```

```
julia> A^16 * w0
```

```
2-element Array{Float64,1}:
 400.0061035156251 7
 599.9938964843755
```

```
julia> A^20 * w0
```

```
2-element Array{Float64,1}:
 400.0003814697268
 599.9996185302739
```

```
julia> A^20 * w0
```

```
2-element Array{Float64,1}:
 400.0003814697268
 599.9996185302739
```

```
julia> A^40 * w0
```

```
2-element Array{Float64,1}:
 400.0000000003645
 599.9999999996372
```

- It seems the Markov Chain $\{\mathbf{w}_0, \mathbf{w}_1, \ldots,\}$ converges to $[400, 600]^{\mathrm{T}}$

```
julia> A^80 * w0
```

```
2-element Array{Float64,1}:
 400.00000000000136
 600.000000000002
```

```
julia> w0 = [ 123; 877]; A^80 * w0
```

```
2-element Array{Float64,1}:
 400.000000000014
 600.000000000023
```

```
julia> w0 = [ 877; 123]; A^80 * w0
```

```
2-element Array{Float64,1}:
 400.0000000000136
 600.000000000022
```

```
julia> w0 = [ 159; 841]; A^80 * w0
```

```
2-element Array{Float64,1}:
 400.000000000014
 600.000000000023
```

- And it seems it converges to the same limit independent of the initial $\mathbf{w}_0$.

- Of course, people get married and get divorced change from year to year

$$\mathbf{w}_k \to \begin{bmatrix} 400 \\ 600 \end{bmatrix} \qquad \text{as} \qquad k \to \infty$$

however, it seems the proportion/probability reminds the same, if we set

$$\mathbf{p}_k = \frac{1}{1000} \mathbf{w}_k$$

then $\mathbf{p}_k$ is essentially the pmf of being married or single at the kth year.

- If we denote $x = 1$ as married and $x = 0$ as single, and the limit as

$$\pi_X(x) = \begin{cases} 0.4 & \text{for} \quad x = 1, \\ 0.6 & \text{for} \quad x = 0, \end{cases}$$

then $x_{k-1} \sim \pi_X$ implies $x_k \sim \pi_X$, this is essentially what the last theorem states about samples from Gibbs, but we have yet to see when it converges.

- For this simple model, where $\mathcal{D} = \{0, 1\}$, convergence is easy to show

$$\mathbf{p}_k = \mathbf{A}^k \mathbf{p}_0 = \mathbf{A}^k \left(\alpha_{10}\mathbf{v}_1 + \alpha_{20}\mathbf{v}_2\right)$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors of $\mathbf{A}$ corresponding eigenvalues λ_1 and λ_2.

```
julia> eigvals(A)
```

```
2-element Array{Float64,1}:
 0.5
 1.0
```

which leads to the following convergence result as $k \to \infty$,

$$\mathbf{p}_k = \alpha_{10}\mathbf{A}^k\mathbf{v}_1 + \alpha_{20}\mathbf{A}^k\mathbf{v}_2 = \alpha_{10}\left(\frac{1}{2}\right)^k \mathbf{v}_1 + \alpha_{20}\left(1\right)^k \mathbf{v}_2 \to a_{20}\mathbf{v}_2 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$$

- In this simple case, we can easily identify the conditions lead to convergence, thus the invariant distribution, we need something similar for Gibbs.

Theorem

Suppose the joint probability density function

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = f_{\{Y_1,\ldots,Y_p\}}\left(y_1,\ldots y_p\right) > 0$$

guarantees the marginal probability density functions are also positive

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

for all $y_1,\ldots,y_n$ in the support $\mathcal{D}$ of the joint distribution, then the sequence

$$\left\{f_{\mathbf{Y}^{(1)}}, f_{\mathbf{Y}^{(2)}},\ldots\right\}$$

corresponding to the Gibbs sampling converges to $f_{\mathbf{Y}}$ for every $\mathbf{y}_0 \in \mathcal{D}$, and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} h\left(\mathbf{Y}^{(t)}\right) \to \mathbb{E}\left[h\left(\mathbf{Y}\right)\right]$$

provided the transition kernel $\kappa\left(\mathbf{y}^{(t-1)},\mathbf{y}^{(t)}\right)$ is absolutely continuous.

- Unfortunately, proving the last theorem is beyond the scope of this course.
- However, together with the proceeding theorems, and our understanding on Markov Chains with discrete state space, this theorem gives the conditions under which we can use samples form Gibbs and how to use it properly.

Q: For example, how can we obtain a Monte Carlo estimate of

$$\mathbb{E}\left[h\left(\mathbf{Y}\right)\right]$$

where $h\colon \mathcal{D} \to \mathbb{R}$ is integrable, using samples from Gibbs sampling.

- We could take n samples after many Gibbs iterations, say m, and expect

$$\mathbb{E}\left[h\left(\mathbf{Y}\right)\right] \approx \frac{1}{n}\sum_{t=m}^{m+n} h\left(\mathbf{Y}^{(t)}\right)$$

- Alternatively, we could construct n Markov Chains using Gibbs sampling,

$$\mathbb{E}\left[h\left(\mathbf{Y}\right)\right] \approx \frac{1}{n}\sum_{j=1}^{n} h\left(\mathbf{Y}_j^{(k_j)}\right)$$

where only the last value $\mathbf{Y}_j^{(k_j)}$ of each chain is used.

Q: Can you think of an example that Gibbs sampling will fail sample from $f_{\mathbf{Y}}$?

$$\mathbf{Y} \sim \text{Uniform}\left(\mathcal{C}_1 \cup \mathcal{C}_2\right)$$

where

$$\mathcal{C}_1 = \{(x_1, x_2) \mid (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1\}$$
$$\mathcal{C}_2 = \{(x_1, x_2) \mid (x_1 + 1)^2 + (x_2 + 1)^2 \leq 1\}$$

- Suppose we want to use Gibbs sampling on the Bivariate normal

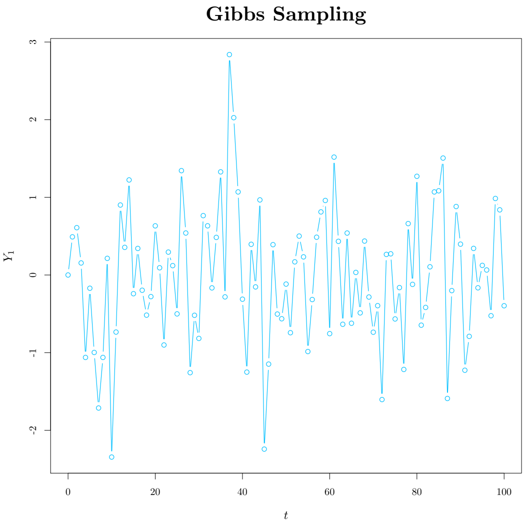
$$\mathbf{Y} \sim \text{Normal}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

where

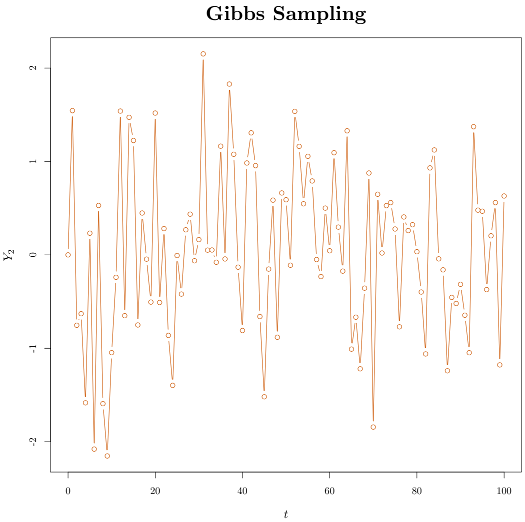
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

Q: What will be our first step?

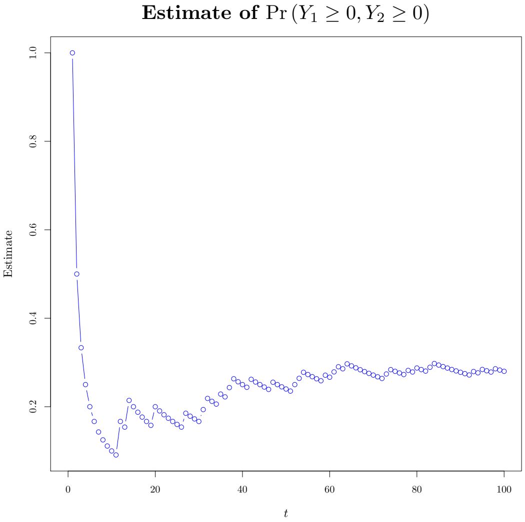
Q: How can we determine whether we have reached the invariant distribution?



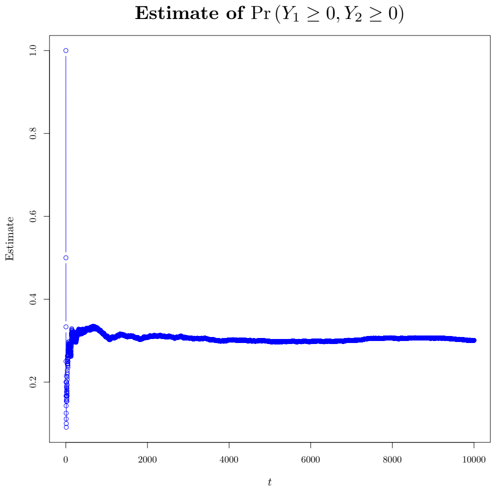
- Various plots based on the sample are usually the way to check.



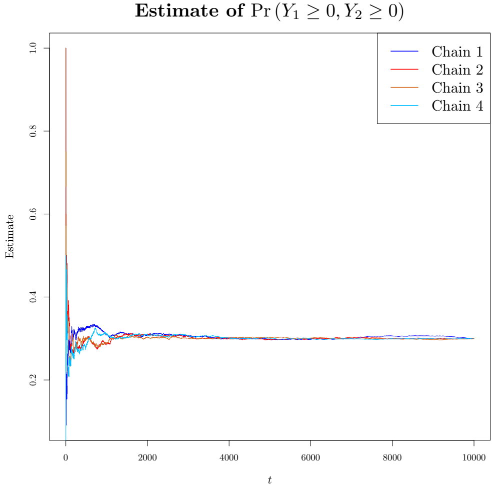
Q: How to estimate the probability $\Pr(Y_1 \geq 0, Y_2 \geq 0)$ base on the sample?



- The last plot suggests the chain is yet to converge, we need a bigger n.



- Of course, we can generate multiple chains using Gibbs sampling.



- In practice, a few chains are run, and each took a certain burn-in period.