Lectures 5 & 6: The Gibbs Sampler

Nick Whiteley

Rejection sampling & Importance sampling

- Objective: approximate an expectation $\mathbb{E}_f(h(X))$ *without* having to sample directly from f.
- Key idea: Sample from an instrumental distribution g and correct for it by the rejection of some values, or by reweighting.
- Yields an *independent* sample $X^{(1)}, X^{(2)}, \ldots$
- Problem: Finding suitable instrumental distributions is hard in high dimensions.

Markov Chain Monte Carlo methods (MCMC)

- Key idea: Create a *dependent* sample, i.e. $X^{(t)}$ depends on the previous value $X^{(t-1)}$.
  $\rightsquigarrow$ allows for "local" updates.
- Only yields an approximate sample from the target distribution.
- More mathematically speaking: yields a Markov chain with the target distribution f as stationary distribution.

4.1 Introduction

4.2 Algorithm

4.3 Hammersley-Clifford theorem

The systematic scan Gibbs sampler

Consider a probability distribution with density $f(x_1, ..., x_p)$, for some $p > 1$.

Algorithm 4.1: (Systematic scan) Gibbs sampler

Starting with $(X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

1. Draw $X_1^{(t)} \sim f_{X_1|X_{-1}}(\cdot|X_2^{(t-1)}, \ldots, X_p^{(t-1)})$.

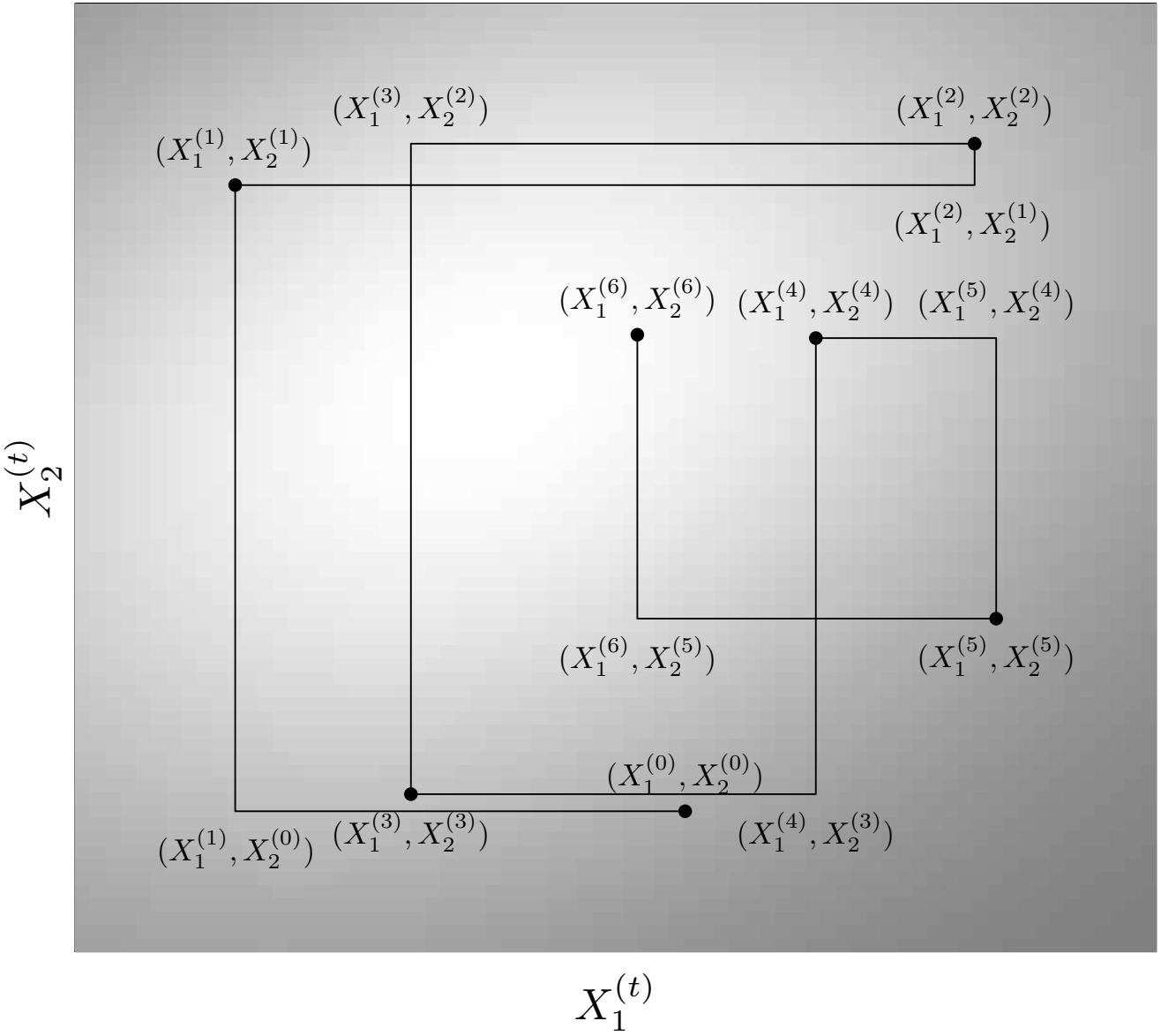
...

j. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$.

...

p. Draw $X_p^{(t)} \sim f_{X_p|X_{-p}}(\cdot|X_1^{(t)}, \ldots, X_{p-1}^{(t)})$.

Illustration of the systematic scan Gibbs sampler



The random scan Gibbs sampler

Algorithm 4.2: Random scan Gibbs sampler

Starting with $(X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

1. Draw an index j from a distribution on $\{1, \ldots, p\}$ (e.g. uniform)
2. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$, and set $X_\iota^{(t)} := X_\iota^{(t-1)}$ for all $\iota \neq j$.

Important questions to ask

- Only the so-called *full-conditional* distributions $X_i|X_{-i}$ are used in the Gibbs sampler.
  - Do the full conditionals fully specify the joint distribution?
- The sequence $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots)$ is a Markov chain.
  - Is the target distribution $f(x_1, \ldots, x_p)$ the invariant distribution of this Markov chain?
  - Will the Markov chain converge to this distribution?
  - If so, what can we use for inference: the whole chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(T)})$ or only the last value $\mathbf{X}^{(T)}$?

The Hammersley-Clifford theorem

Definition 4.1: Positivity condition

A distribution with density $f(x_1, \ldots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the *positivity condition* if $f(x_1, \ldots, x_p) > 0$ for all $x_1, \ldots, x_p$ with $f_{X_i}(x_i) > 0$.

Theorem 4.1: Hammersley-Clifford

Let $(X_1, \ldots, X_p)$ have joint density $f(x_1, \ldots, x_p)$ satisfying the positivity condition. Then for all $(\xi_1, \ldots, \xi_p) \in \text{supp}(f)$

$$f(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)}{f_{X_j|X_{-j}}(\xi_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)}$$

Note the theorem assumes that $f(x_1, \ldots, x_p)$ is a well-defined probability density. In general, it is not guaranteed that every set of full conditionals characterize a well-defined joint distribution.

Example 4.1

- Consider the following "model"

$$\begin{aligned} X_1|X_2 &\sim \text{Expo}(\lambda X_2) \\ X_2|X_1 &\sim \text{Expo}(\lambda X_1), \end{aligned}$$

- It is tempting to write, as in the Hammersley-Clifford theorem:

$$\begin{aligned} f(x_1, x_2) &\propto \frac{f_{X_1|X_2}(x_1|\xi_2) \cdot f_{X_2|X_1}(x_2|x_1)}{f_{X_1|X_2}(\xi_1|\xi_2) \cdot f_{X_2|X_1}(\xi_2|x_1)} \\ &\propto \exp(-\lambda x_1 x_2) \end{aligned}$$

- BUT $\int \int \exp(-\lambda x_1 x_2)\, dx_1\, dx_2 = +\infty$
- There does not exist a well-defined joint density proportional to $\exp(-\lambda x_1 x_2)$.

4.4 Convergence properties

Invariant distribution

Lemma 4.1

The transition kernel of the systematic scan Gibbs sampler is

$$\begin{aligned} K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) &= f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)}, \ldots, x_p^{(t-1)}) \\ &\quad \cdot f_{X_2|X_{-2}}(x_2^{(t)}|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)}) \\ &\quad \cdot \ldots \\ &\quad \cdot f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \ldots, x_{p-1}^{(t)}) \end{aligned}$$

Proposition 4.1

The joint distribution $f(x_1, \ldots, x_p)$ is indeed the invariant distribution of the Markov chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots)$ generated by the Gibbs sampler.

Irreducibility and recurrence

Proposition 4.2

If the joint distribution $f(x_1, \ldots, x_p)$ satisfies the positivity condition, the Gibbs sampler yields an irreducible, recurrent Markov chain.

(less strict conditions exist)

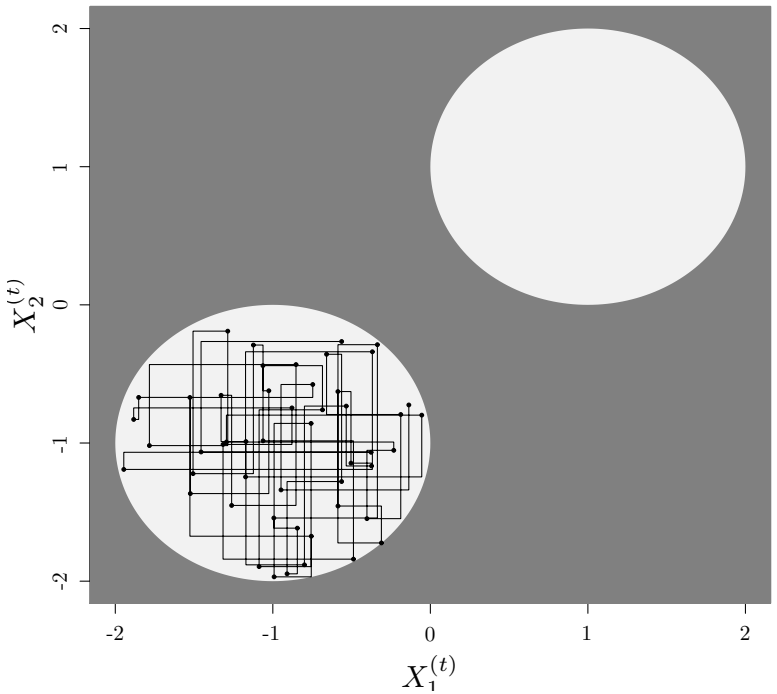
If the transition kernel is absolutely continuous with respect to the dominating measure, then recurrence even implies Harris recurrence.

Example 4.2: Reducible Gibbs sampler

Consider Gibbs sampling from the uniform distribution

$$f(x_1, x_2) = \frac{1}{2\pi} \mathbb{I}_{C_1 \cup C_2}(x_1, x_2),$$

$$\begin{aligned} C_1 &:= \{(x_1, x_2) : \|(x_1, x_2) - (1, 1)\| \leq 1\} \\ C_2 &:= \{(x_1, x_2) : \|(x_1, x_2) + (1, 1)\| \leq 1\} \end{aligned}$$



The resulting Markov chain is *not* irreducible. It stays forever in either C_1 or C_2.

Ergodic theorem

Theorem 4.2

If the Markov chain generated by the Gibbs sampler is irreducible and recurrent (which is e.g. the case when the positivity condition holds), then for any integrable function $h : E \to \mathbb{R}$

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

for almost every starting value $\mathbf{X}^{(0)}$. If the chain is Harris recurrent, then the above result holds for every starting value $\mathbf{X}^{(0)}$.

Thus we can approximate expectations $\mathbb{E}_f\left(h(\mathbf{X})\right)$ by their empirical counterparts using *a single* Markov chain.

Example 4.3 (1)

- Consider

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathsf{N}_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

- Associated marginal distributions

$$\begin{aligned} X_1 &\sim \mathsf{N}(\mu_1, \sigma_1^2) \\ X_2 &\sim \mathsf{N}(\mu_2, \sigma_2^2) \end{aligned}$$

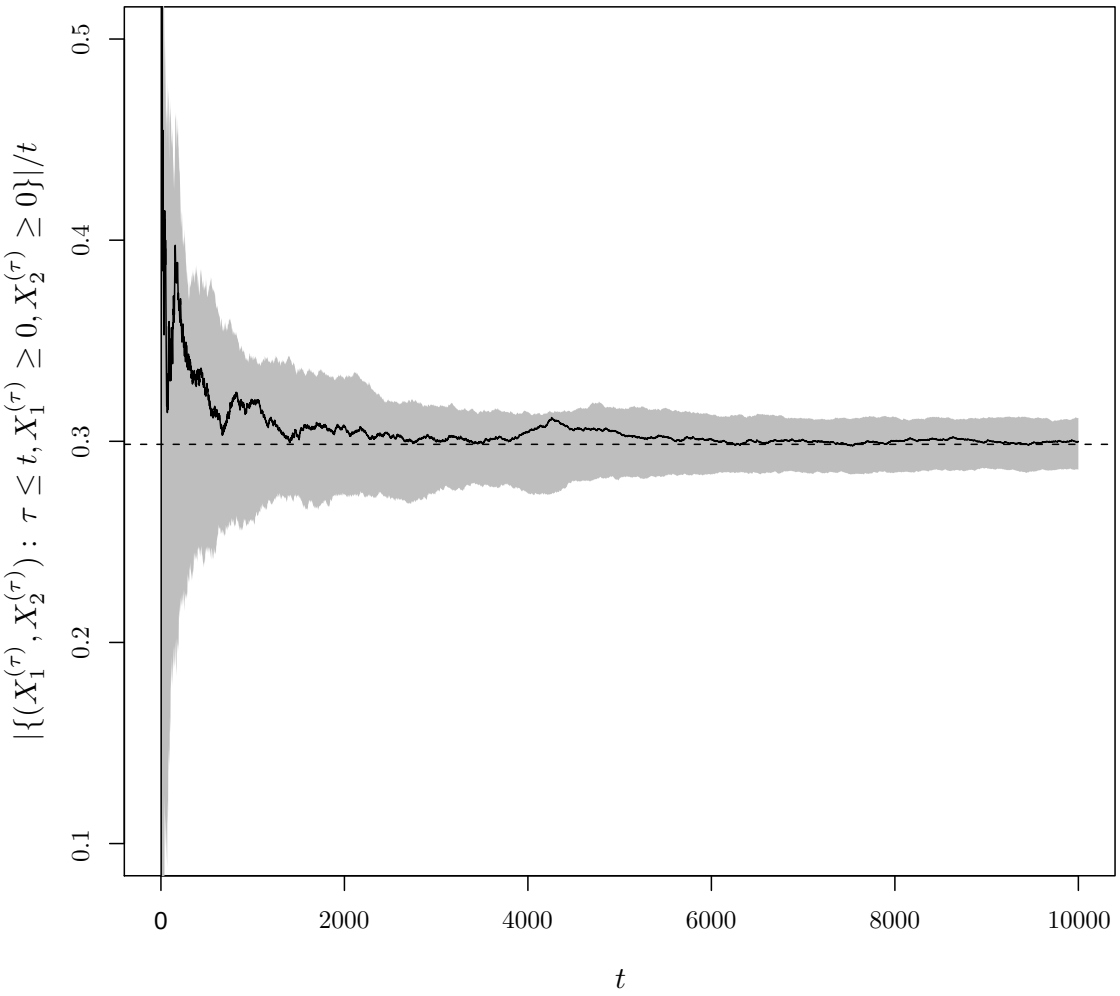
- Associated full conditionals

$$\begin{aligned} X_1 | X_2 = x_2 &\sim \mathsf{N}(\mu_1 + \sigma_{12}/\sigma_2^2(x_2 - \mu_2), \sigma_1^2 - (\sigma_{12})^2\sigma_2^2) \\ X_2 | X_1 = x_1 &\sim \mathsf{N}(\mu_2 + \sigma_{12}/\sigma_1^2(x_1 - \mu_1), \sigma_2^2 - (\sigma_{12})^2\sigma_1^2) \end{aligned}$$

- Gibbs sampler consists of iterating for $t = 1, 2, \ldots$
  1. Draw $X_1^{(t)} \sim \mathsf{N}(\mu_1 + \sigma_{12}/\sigma_2^2(X_2^{(t-1)} - \mu_2), \sigma_1^2 - (\sigma_{12})^2\sigma_2^2)$
  2. Draw $X_2^{(t)} \sim \mathsf{N}(\mu_2 + \sigma_{12}/\sigma_1^2(X_1^{(t)} - \mu_1), \sigma_2^2 - (\sigma_{12})^2\sigma_1^2)$.

Example 4.3 (2)

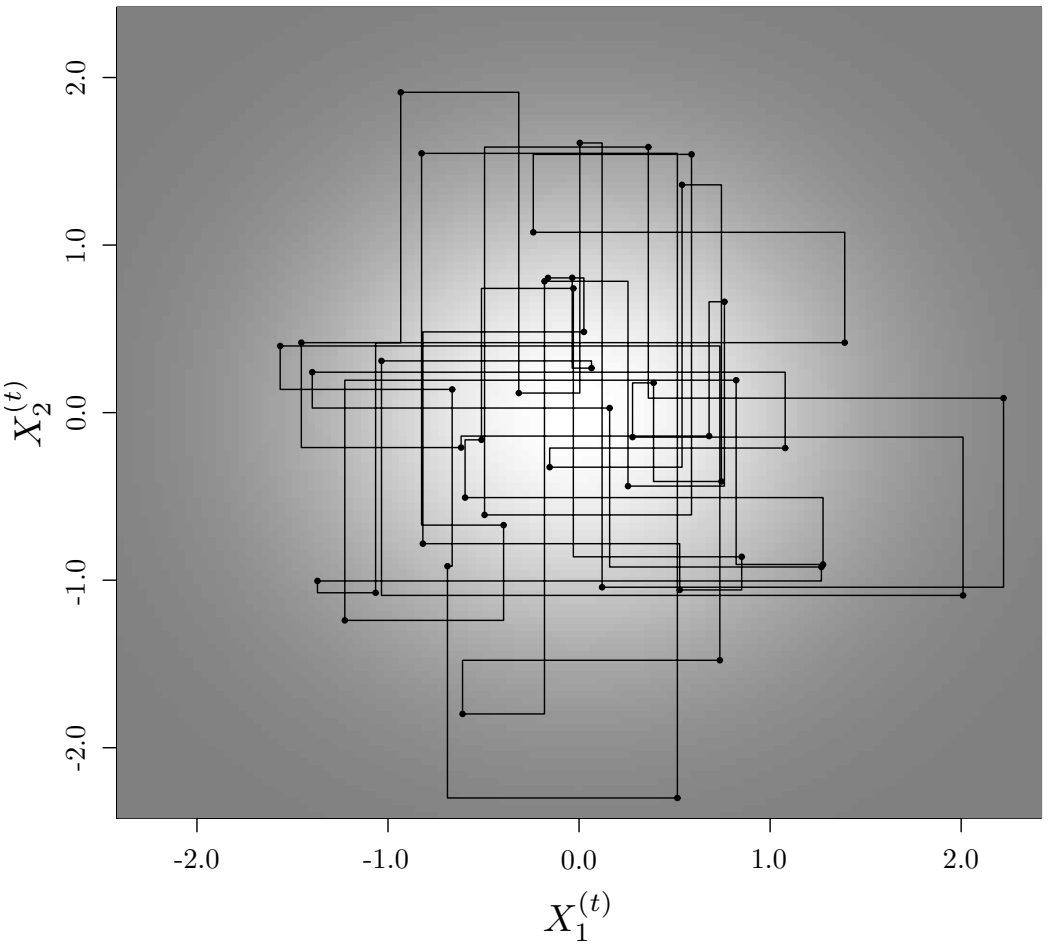
Using the ergodic theorem we can estimate $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$ by the proportion of samples $(X_1^{(t)}, X_2^{(t)})$ with $X_1^{(t)} \geq 0$ and $X_2^{(t)} \geq 0$:

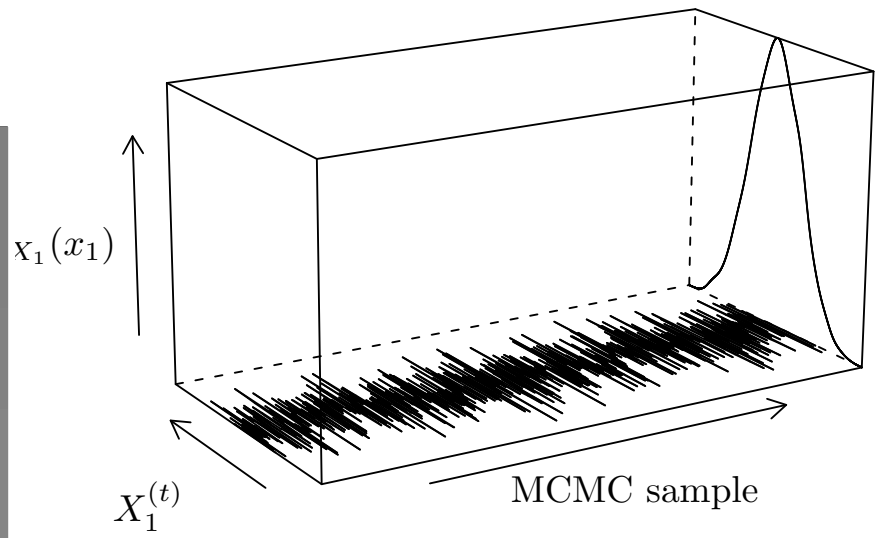


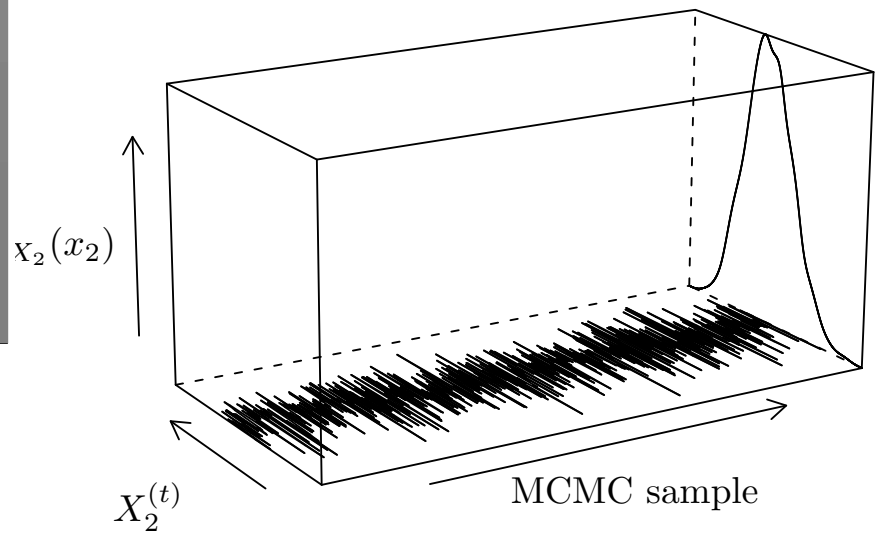
Dependency structure of samples from the Gibbs sampler

- $\mathbf{X}^{(t-1)}$ and $\mathbf{X}^{(t)}$ are dependent and typically positively correlated
  (unless the components $(X_1^{(t)}, \ldots, X_p^{(t)})$ are independent for a fixed t)
- Amount of correlation increases with the dependency (correlation) of the components $(X_1^{(t)}, \ldots, X_p^{(t)})$.
- Consequence: a sample of size n from a Gibbs sampler can (and in most cases will) contain less information than an i.i.d. sample of size n, especially when the correlation between $\mathbf{X}^{(t-1)}$ and $\mathbf{X}^{(t)}$ is large.
  $\rightsquigarrow$ concept of the "effective sample size"

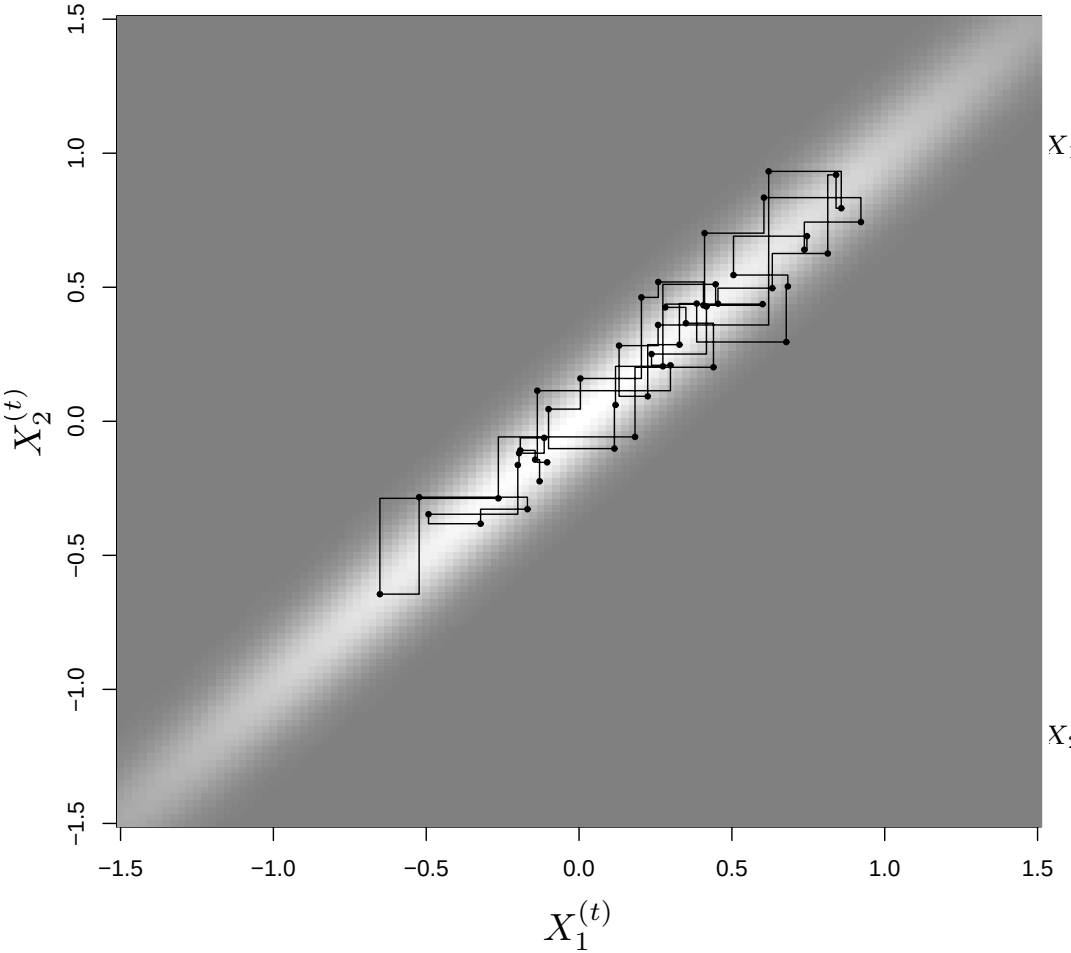
Example 4.4: Bivariate Gaussian $\rho(X_1, X_2) = 0.3$

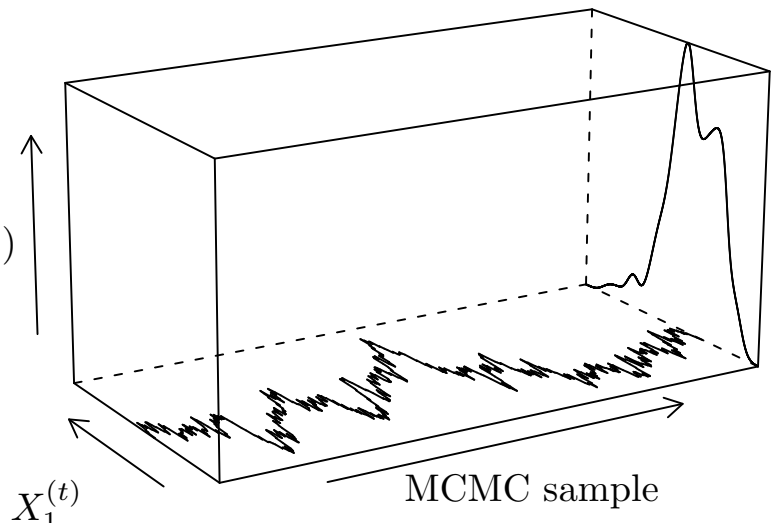


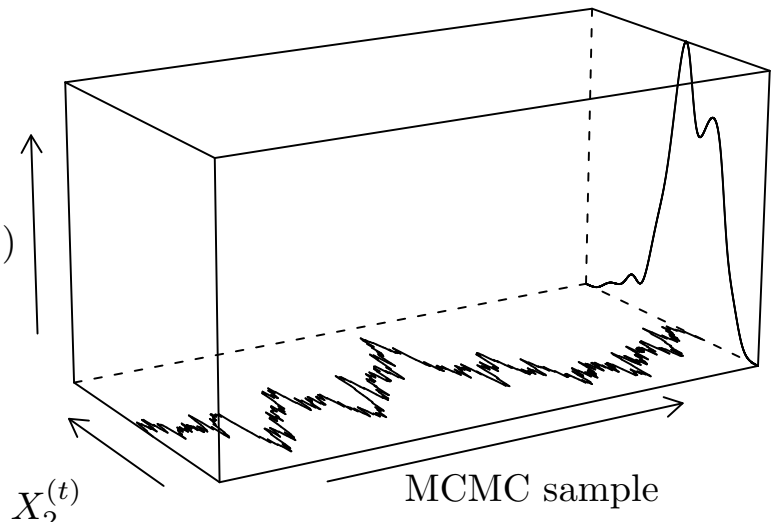




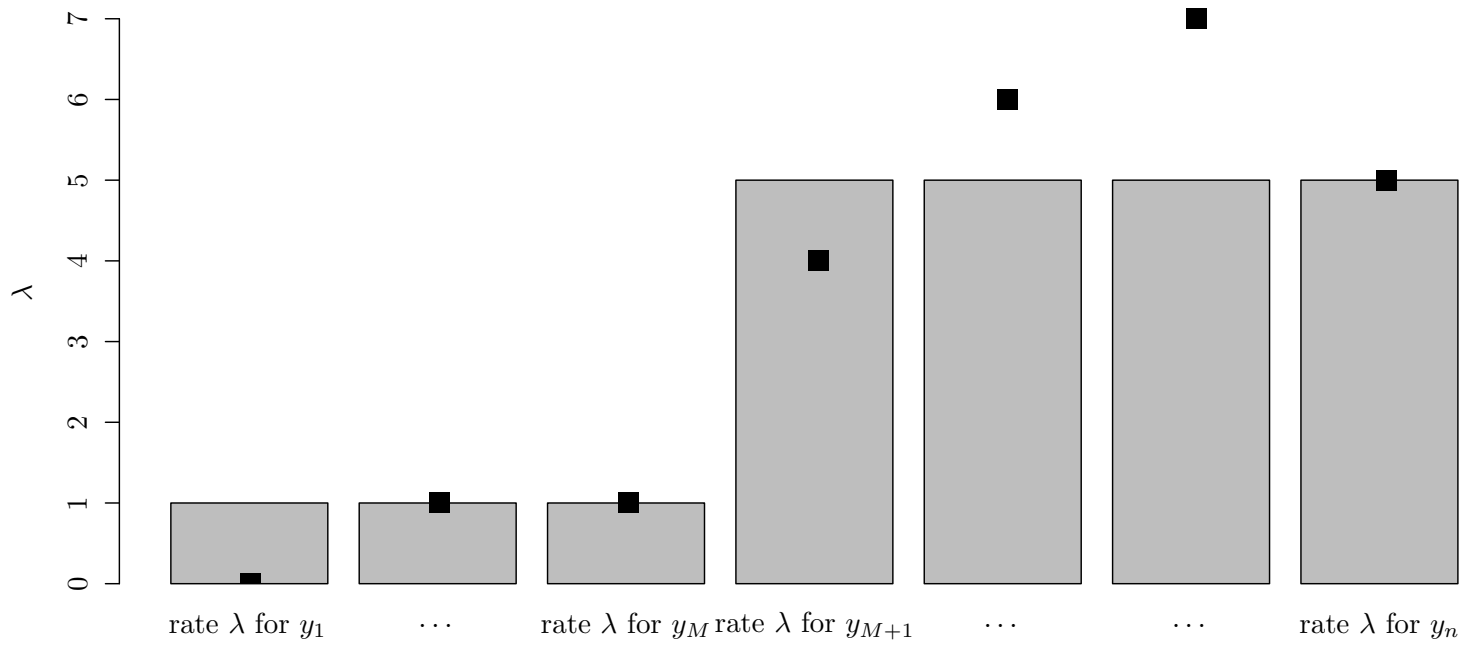
Example 4.4: Bivariate Gaussian $\rho(X_1, X_2) = 0.99$







Example 4.5 Poisson change point model (1)



$$Y_i \sim \text{Poi}(\lambda_1) \quad \text{for} \quad i = 1, \ldots, M$$
$$Y_i \sim \text{Poi}(\lambda_2) \quad \text{for} \quad i = M+1, \ldots, n$$

Objective: (Bayesian) inference about the parameters λ_1, λ_2, and M given observed data $Y_1, \ldots, Y_n$.

Example 4.5 Poisson change point model (2)

- Prior distributions: $\lambda_j \sim \text{Gamma}(\alpha_j, \beta_j)$ $(j = 1, 2)$, i.e.

$$f(\lambda_j) = \frac{1}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} \beta_j^{\alpha_j} \exp(-\beta_j \lambda_j).$$

  (discrete uniform prior on M, i.e. $p(M) \propto 1$).

- Likelihood: $l(y_1, \ldots, y_n | \lambda_1, \lambda_2, M)$

$$= \left(\prod_{i=1}^{M} \frac{\exp(-\lambda_1)\lambda_1^{y_i}}{y_i!} \right) \cdot \left(\prod_{i=M+1}^{n} \frac{\exp(-\lambda_2)\lambda_2^{y_i}}{y_i!} \right)$$

- Joint distribution $f(y_1, \ldots, y_n, \lambda_1, \lambda_2, M)$

$$\begin{aligned} &= l(y_1, \ldots, y_n | \lambda_1, \lambda_2, M) \cdot f(\lambda_1) \cdot f(\lambda_2) \cdot p(M) \\ &\propto \left(\prod_{i=1}^{M} \frac{\exp(-\lambda_1)\lambda_1^{y_i}}{y_i!} \right) \cdot \left(\prod_{i=M+1}^{n} \frac{\exp(-\lambda_2)\lambda_2^{y_i}}{y_i!} \right) \\ &\quad \cdot \frac{1}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1 - 1} \beta_1^{\alpha_1} \exp(-\beta_1 \lambda_1) \cdot \frac{1}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2 - 1} \beta_2^{\alpha_2} \exp(-\beta_2 \lambda_2) \end{aligned}$$

Example 4.5 Poisson change point model (3)

- Joint posterior distribution $f(\lambda_1, \lambda_2, M|y_1, \ldots, y_n)$

$$\begin{aligned} \propto \quad & \lambda_1^{\alpha_1 - 1 + \sum_{i=1}^{M} y_i} \exp(-(\beta_1 + M)\lambda_1) \\ & \cdot \lambda_2^{\alpha_2 - 1 + \sum_{i=M+1}^{n} y_i} \exp(-(\beta_2 + n - M)\lambda_2) \end{aligned}$$

- Conditional on M (i.e. if M was known) we have

$$f(\lambda_1|y_1, \ldots, y_n, M) \propto \lambda_1^{\alpha_1 - 1 + \sum_{i=1}^{M} y_i} \exp(-(\beta_1 + M)\lambda_1),$$

i.e.

$$\lambda_1|Y_1, \ldots Y_n, M \quad \sim \quad \text{Gamma}\left(\alpha_1 + \sum_{i=1}^{M} y_i, \beta_1 + M\right)$$

$$\lambda_2|Y_1, \ldots Y_n, M \quad \sim \quad \text{Gamma}\left(\alpha_2 + \sum_{i=M+1}^{n} y_i, \beta_2 + n - M\right).$$

- $p(M|\ldots) \propto \lambda_1^{\sum_{i=1}^{M} y_i} \cdot \lambda_2^{\sum_{i=M+1}^{n} y_i} \cdot \exp((\lambda_2 - \lambda_1) \cdot M)$

Example 4.5 Poisson change point model (4)

Gibbs sampler (one iteration):

1. Draw λ_1 from the conditional distribution $\lambda_1|Y_1,\ldots,Y_n,M$, i.e. draw

$$\lambda_1 \sim \text{Gamma}\left(\alpha_1 + \sum_{i=1}^{M} y_i, \beta_1 + M\right)$$

2. Draw λ_2 from the conditional distribution $\lambda_2|Y_1,\ldots,Y_n,M$, i.e. draw

$$\lambda_2 \sim \text{Gamma}\left(\alpha_2 + \sum_{i=M+1}^{n} y_i, \beta_2 + n - M\right)$$

3. Draw M from the conditional distribution $M|Y_1,\ldots,Y_n,\lambda_1,\lambda_2$, i.e. draw

$$p(M) \propto \lambda_1^{\sum_{i=1}^{M} y_i} \cdot \lambda_2^{\sum_{i=M+1}^{n} y_i} \cdot \exp((\lambda_2 - \lambda_1)\cdot M)$$

4.5 Data augmentation

Data augmentation

- Gibbs sampling is only feasible when we can sample easily from the full conditionals.
- A technique that can help achieving full conditionals that are easy to sample from is *demarginalisation*:
  Introduce a set of auxiliary random variables $Z_1, \ldots, Z_r$ such that f is the marginal density of $(X_1, \ldots, X_p, Z_1, \ldots, Z_r)$, i.e.

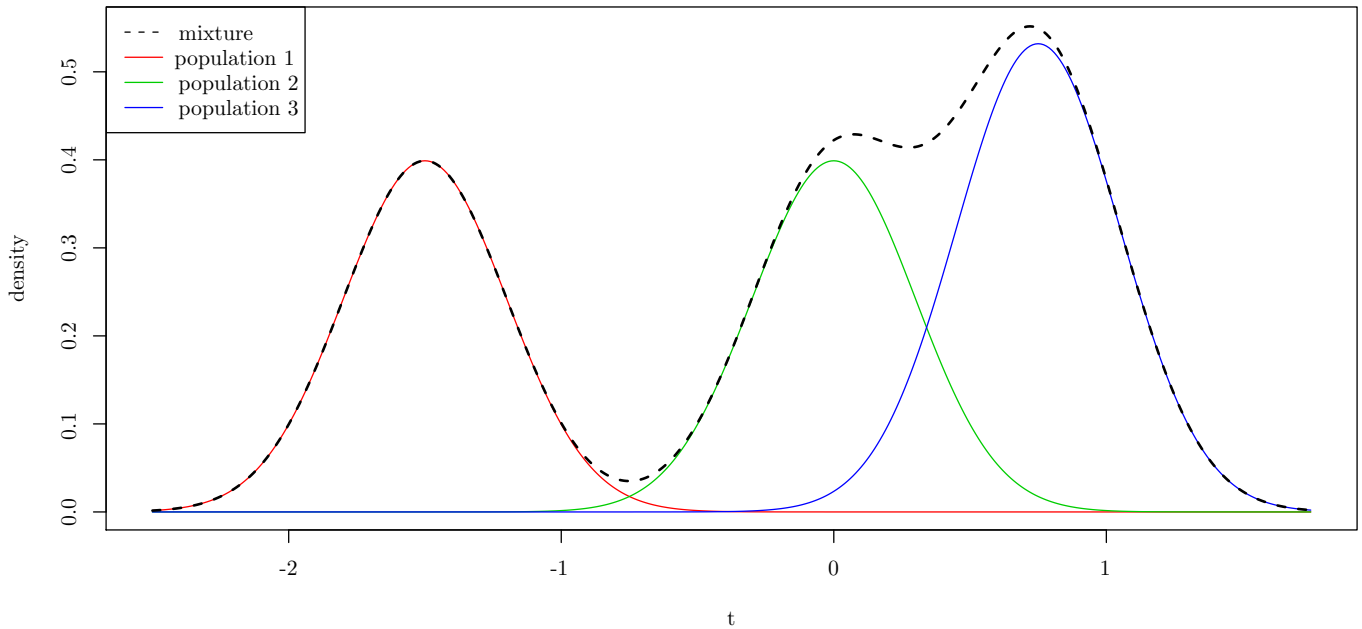
$$f(x_1, \ldots, x_p) = \int f(x_1, \ldots, x_n, z_1, \ldots, z_r)\, d(z_1, \ldots, z_r).$$

- In many cases there is a "natural choice" of the *completion* $(Z_1, \ldots, Z_r)$.

Example 4.6: Mixture of Gaussians: Model

Consider the following K population mixture model for data $Y_1, \ldots, Y_n$:

$$f(y_i) = \sum_{k=1}^{K} \pi_k \phi_{(\mu_k, 1/\tau)}(y_i)$$



Objective: Bayesian inference for the parameters $(\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K)$.

Example 4.6: Mixture of Gaussians: Priors

- The number of components K is assumed to be known.
- The variance parameter τ is assumed to be known.
- $(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, i.e.

$$f_{(\alpha_1,\ldots,\alpha_K)}(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

- $(\mu_1, \ldots, \mu_K) \sim \text{N}(\mu_0, 1/\tau_0)$, i.e.

$$f_{(\mu_0,\tau_0)}(\mu_k) \propto \exp\left(-\tau_0(\mu_k - \mu_0)^2/2\right)$$

Example 4.6: Mixture of Gaussians: Joint distribution

$$f(\mu_1,\ldots,\mu_K,\pi_1,\ldots,\pi_K,y_1,\ldots,y_n) \propto \left(\prod_{k=1}^{K}\pi_k^{\alpha_k-1}\right)\cdot$$
$$\left(\prod_{k=1}^{K}\exp\left(-\tau_0(\mu_k-\mu_0)^2/2\right)\right)\cdot\left(\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\exp\left(-\tau(y_i-\mu_k)^2/2\right)\right)$$

The full conditionals do not seem to come from "nice" distributions.

Use data augmentation: include auxiliary variables $Z_1,\ldots Z_n$ which indicate which population the i-th individual is from, i.e.

$$\mathbb{P}(Z_i=k)=\pi_k \qquad \text{and} \qquad Y_i|Z_i=k\sim N(\mu_k,1/\tau).$$

The marginal distribution of Y is as before, so $Z_1,\ldots Z_n$ are indeed a completion.

Example 4.6: Mixture of Gaussians: Joint distribution (ctd.)

The joint distribution of the augmented system is

$$\begin{aligned} & f(y_1, \ldots, y_n, z_1, \ldots, z_n, \mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K) \\ \propto & \left(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right) \cdot \left(\prod_{k=1}^{K} \exp\left(-\tau_0 (\mu_k - \mu_0)^2 / 2 \right) \right) \\ & \cdot \left(\prod_{i=1}^{n} \pi_{z_i} \exp\left(-\tau (y_i - \mu_{z_i})^2 / 2 \right) \right) \end{aligned}$$

The full conditionals now come from "nice" distributions.

Example 4.6: Mixture of Gaussians: Full conditionals

$$\begin{aligned}
&\mathbb{P}(Z_i = k | Y_1, \ldots, Y_n, \mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K) \\
&\quad = \frac{\pi_k \phi_{(\mu_k, 1/\tau)}(y_i)}{\sum_{\iota=1}^{K} \pi_\iota \phi_{(\mu_\iota, 1/\tau)}(y_i)}
\end{aligned}$$

$$\begin{aligned}
&\mu_k | Y_1, \ldots, Y_n, Z_1, \ldots, Z_n, \pi_1, \ldots, \pi_K \\
&\quad \sim \mathsf{N}\left(\frac{\tau \left(\sum_{i:\ Z_i = k} Y_i \right) + \tau_o \mu_0}{|\{:\ Z_i = k\}| \tau + \tau_0}, \frac{1}{|\{i:\ Z_i = k\}| \tau + \tau_0} \right)
\end{aligned}$$

$$\begin{aligned}
&\pi_1, \ldots, \pi_K | Y_1, \ldots, Y_n, Z_1, \ldots, Z_n, \mu_1, \ldots, \mu_K \\
&\quad \sim \mathsf{Dirichlet}\left(\alpha_1 + |\{i:\ Z_i = 1\}|, \ldots, \alpha_K + |\{I:\ Z_i = K\}| \right).
\end{aligned}$$

Example 4.6: Mixture of Gaussians: Gibbs sampler

Starting with initial values $\mu_1^{(0)}, \ldots, \mu_K^{(0)}, \pi_1^{(0)}, \ldots, \pi_K^{(0)}$ iterate the following steps for $t = 1, 2, \ldots$

1. For $i = 1, \ldots, n$:

   Draw $Z_i^{(t)}$ from the discrete distribution on $\{1, \ldots, K\}$ specified by

$$p(Z_i^{(t)}) = \left(\frac{\pi_k \phi_{(\mu_k^{(t-1)}, 1/\tau)}(y_i)}{\sum_{\iota=1}^{K} \pi_\iota^{(t-1)} \phi_{(\mu_\iota^{(t-1)}, 1/\tau)}(y_i)} \right).$$

2. For $k = 1, \ldots, K$:

   Draw

$$\mu_k^{(t)} \sim \mathrm{N}\left(\frac{\tau \left(\sum_{i:\, Z_i^{(t)}=k} Y_i \right) + \tau_o \mu_0}{|\{i :\ Z_i^{(t)} = k\}|\tau + \tau_0}, \frac{1}{|\{i :\ Z_i^{(t)} = k\}|\tau + \tau_0} \right).$$

3. Draw

$$(\pi_1^{(t)}, \ldots, \pi_K^{(t)}) \sim \text{Dirichlet}\left(\alpha_1 + |\{i :\ Z_i^{(t)} = 1\}|, \ldots, \alpha_K + |\{i :\ Z_i^{(t)} = K\}| \right).$$