CPSC 535
Gibbs Sampling

AD

February 2007

- Rejection Sampling and Importance Sampling are two general methods but limited to problems of moderate dimensions.

- Rejection Sampling and Importance Sampling are two general methods but limited to problems of moderate dimensions.
- **Problem**: We try to sample all the components of a potentially high-dimensional parameter simultaneously/sequentially and we can never correct for components already sampled.

- Rejection Sampling and Importance Sampling are two general methods but limited to problems of moderate dimensions.
- **Problem**: We try to sample all the components of a potentially high-dimensional parameter simultaneously/sequentially and we can never correct for components already sampled.
- A powerful class of methods is available to deal with such methods: Markov chain Monte Carlo.

Bayesian Model

- Multiple failures in a nuclear plant

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

Bayesian Model

- Multiple failures in a nuclear plant

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

- Model: Failures of the $i-$th pump follow a Poisson process with parameter λ_i $(1 \leq i \leq 10)$. For an observed time t_i, the number of failures p_i is thus a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable.

Bayesian Model

- Multiple failures in a nuclear plant

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

- Model: Failures of the $i-$th pump follow a Poisson process with parameter λ_i $(1 \leq i \leq 10)$. For an observed time t_i, the number of failures p_i is thus a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable.
- The unknown parameters consist of $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$.

- Hierarchical model

$$\lambda_i | (\alpha, \beta) \overset{\text{iid}}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- Hierarchical model

$$\lambda_i | \, (\alpha, \beta) \overset{\text{iid}}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- The posterior distribution is proportional to

$$\begin{aligned} & p\left(\lambda_{1:10}, \beta | \, p_{1:10}, t_{1:10}\right) \\ \propto & \prod_{i=1}^{10} \{(\lambda_i t_i)^{p_i} \exp(-\lambda_i t_i) \lambda_i^{\alpha-1} \exp(-\beta\lambda_i)\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta) \\ \propto & \prod_{i=1}^{10} \{\lambda_i^{p_i+\alpha-1} \exp(-(t_i+\beta)\lambda_i)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta). \end{aligned}$$

- Hierarchical model

$$\lambda_i | (\alpha, \beta) \overset{\text{iid}}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- The posterior distribution is proportional to

$$\begin{aligned} & p\left(\lambda_{1:10}, \beta | \, p_{1:10}, t_{1:10}\right) \\ \propto & \prod_{i=1}^{10} \{(\lambda_i t_i)^{p_i} \exp(-\lambda_i t_i) \lambda_i^{\alpha-1} \exp(-\beta\lambda_i)\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta) \\ \propto & \prod_{i=1}^{10} \{\lambda_i^{p_i+\alpha-1} \exp(-(t_i+\beta)\lambda_i)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta). \end{aligned}$$

- This multidimensional distribution is rather complex. It is not obvious how the rejection method or importance sampling could be used in this context.

- The conditionals have a familiar form

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

where

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

and

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

- The conditionals have a familiar form

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

where

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

and

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

- Instead of directly sampling the vector $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$ at once, one could suggest sampling it iteratively, starting for example with the λ_i's for a given guess of β, followed by an update of β given the new samples $\lambda_1, \ldots, \lambda_{10}$.

My first Gibbs sampler

- Given a sample, at iteration t, $\theta^t := (\lambda_1^t, \ldots, \lambda_{10}^t, \beta^t)$ one could proceed as follows at iteration $t+1$,

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

My first Gibbs sampler

- Given a sample, at iteration t, $\theta^t := (\lambda_1^t, \ldots, \lambda_{10}^t, \beta^t)$ one could proceed as follows at iteration $t+1$,

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

- Instead of directly sampling in a space with 11 dimensions, one samples in spaces of dimension 1.

My first Gibbs sampler

- Given a sample, at iteration t, $\theta^t := (\lambda_1^t, \ldots, \lambda_{10}^t, \beta^t)$ one could proceed as follows at iteration $t+1$,

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

- Instead of directly sampling in a space with 11 dimensions, one samples in spaces of dimension 1.
- Note that the deterministic version of such an algorithm where sampling is replaced by maximization would not generally converge towards the global maximum of the joint distribution.

- The structure of the algorithm calls for many questions:

- The structure of the algorithm calls for many questions:
  - Are we sampling from the desired joint distribution?

- The structure of the algorithm calls for many questions:
  - Are we sampling from the desired joint distribution?
  - If yes, how many times should the iteration above be repeated?

- The structure of the algorithm calls for many questions:
  - Are we sampling from the desired joint distribution?
  - If yes, how many times should the iteration above be repeated?
- The validity of the approach described here stems from the fact that the sequence $\{\theta^t\}$ defined above is a Markov chain and some Markov chains have very nice properties.

Elements of Markov chains

- **Markov chain**: A sequence of random variables $\{X_n; n \in \mathbb{N}\}$ defined on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ which satisfies the property, for any $A \in \mathcal{B}(\mathbb{X})$

$$\mathbb{P}(X_n \in A | X_0, ..., X_{n-1}) = \mathbb{P}(X_n \in A | X_{n-1}).$$

and we will write

$$\mathbb{P}(X_n \in A | X_{n-1}) = P(x, A) = \int_A P(x, dy).$$

Elements of Markov chains

- **Markov chain**: A sequence of random variables $\{X_n; n \in \mathbb{N}\}$ defined on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ which satisfies the property, for any $A \in \mathcal{B}(\mathbb{X})$

$$\mathbb{P}(X_n \in A | X_0, ..., X_{n-1}) = \mathbb{P}(X_n \in A | X_{n-1}).$$

and we will write

$$\mathbb{P}(X_n \in A | X_{n-1}) = P(x, A) = \int_A P(x, dy).$$

- **Markov chain Monte Carlo**: Given a target π, design a transition kernel P such that asymptotically as $n \longrightarrow \infty$

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \rightarrow \int_{\mathbb{X}} \varphi(x) \pi(x) dx \text{ and/or } X_n \sim \pi.$$

Elements of Markov chains

- **Markov chain**: A sequence of random variables $\{X_n; n \in \mathbb{N}\}$ defined on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ which satisfies the property, for any $A \in \mathcal{B}(\mathbb{X})$

$$\mathbb{P}(X_n \in A | X_0, ..., X_{n-1}) = \mathbb{P}(X_n \in A | X_{n-1}).$$

and we will write

$$\mathbb{P}(X_n \in A | X_{n-1}) = P(x, A) = \int_A P(x, dy).$$

- **Markov chain Monte Carlo**: Given a target π, design a transition kernel P such that asymptotically as $n \longrightarrow \infty$

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \rightarrow \int_{\mathbb{X}} \varphi(x) \pi(x) dx \text{ and/or } X_n \sim \pi.$$

- It should be easy to simulate the Markov chain even if π is complex.

- Consider the autoregression for $|\alpha| < 1$

$$X_n = \alpha X_{n-1} + V_n, \text{ where } V_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

then

$$P\left(x, dy\right) = P\left(x, y\right) dy = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \alpha x\right)^2}{2\sigma^2}\right) dx.$$

- Consider the autoregression for $|\alpha| < 1$

$$X_n = \alpha X_{n-1} + V_n, \text{ where } V_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

then

$$P\left(x, dy\right) = P\left(x, y\right) dy = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \alpha x\right)^2}{2\sigma^2}\right) dx.$$

- The limiting distribution is

$$\pi\left(x\right) = \mathcal{N}\left(x; 0, \frac{\sigma^2}{1 - \alpha^2}\right)$$

and satisfies

$$\int_{\mathbb{X}} \pi\left(x\right) P\left(x, y\right) dx = \pi\left(y\right)$$

- Consider the autoregression for $|\alpha| < 1$

$$X_n = \alpha X_{n-1} + V_n, \text{ where } V_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

then

$$P\left(x, dy\right) = P\left(x, y\right) dy = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \alpha x\right)^2}{2\sigma^2}\right) dx.$$

- The limiting distribution is

$$\pi\left(x\right) = \mathcal{N}\left(x; 0, \frac{\sigma^2}{1 - \alpha^2}\right)$$

and satisfies

$$\int_{\mathbb{X}} \pi\left(x\right) P\left(x, y\right) dx = \pi\left(y\right)$$

- To sample from π, we could just sample the Markov chain and asymptotically we would have $X_n \sim \pi$. [Obviously, in this case this is useless because we can sample from π directly.]

- Graphically, consider 1000 independent Markov chains run in parallel.

- Graphically, consider 1000 independent Markov chains run in parallel.
- We assume that the initial distribution of these Markov chains is $\mathcal{U}_{[0,20]}$. So initially, the Markov chains samples are not distributed according to π

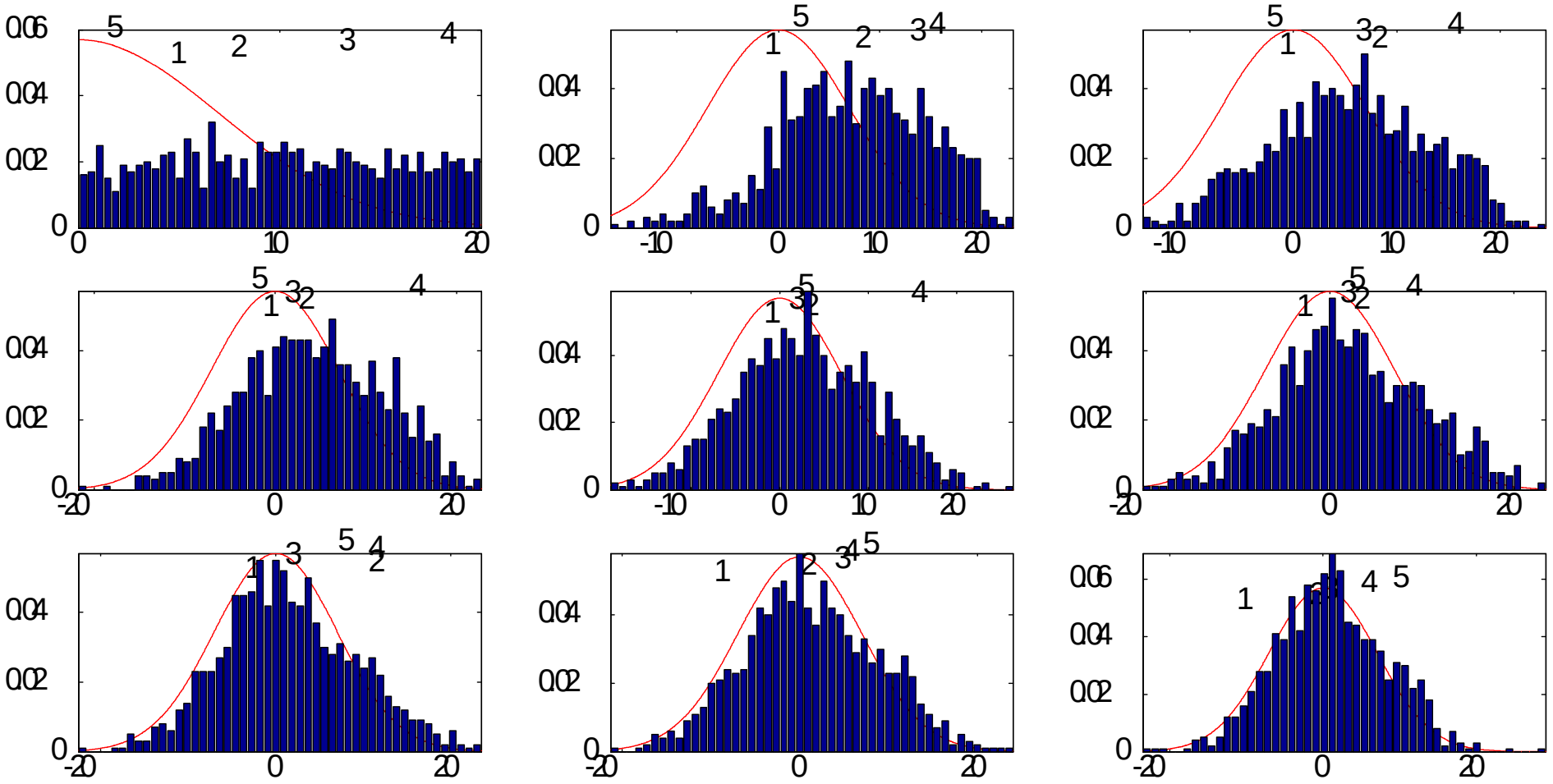

Figure: From top left to bottom right: histograms of 1000 independent Markov chains with a normal distribution as target distribution as n increases.

- The target normal distribution seems to “attract” the distribution of the samples and even to be a fixed point of the algorithm.

- The target normal distribution seems to "attract" the distribution of the samples and even to be a fixed point of the algorithm.
- This is is what we wanted to achieve, *i.e.* it seems that we have produced 1000 independent samples from the normal distribution.

- The target normal distribution seems to "attract" the distribution of the samples and even to be a fixed point of the algorithm.
- This is is what we wanted to achieve, *i.e.* it seems that we have produced 1000 independent samples from the normal distribution.
- In fact one can show that in many (all?) situations of interest it is not necessary to run N Markov chains in parallel in order to obtain 1000 samples, but that one can consider a unique Markov chain, and build the histogram from this single Markov chain by forming histograms from one trajectory.

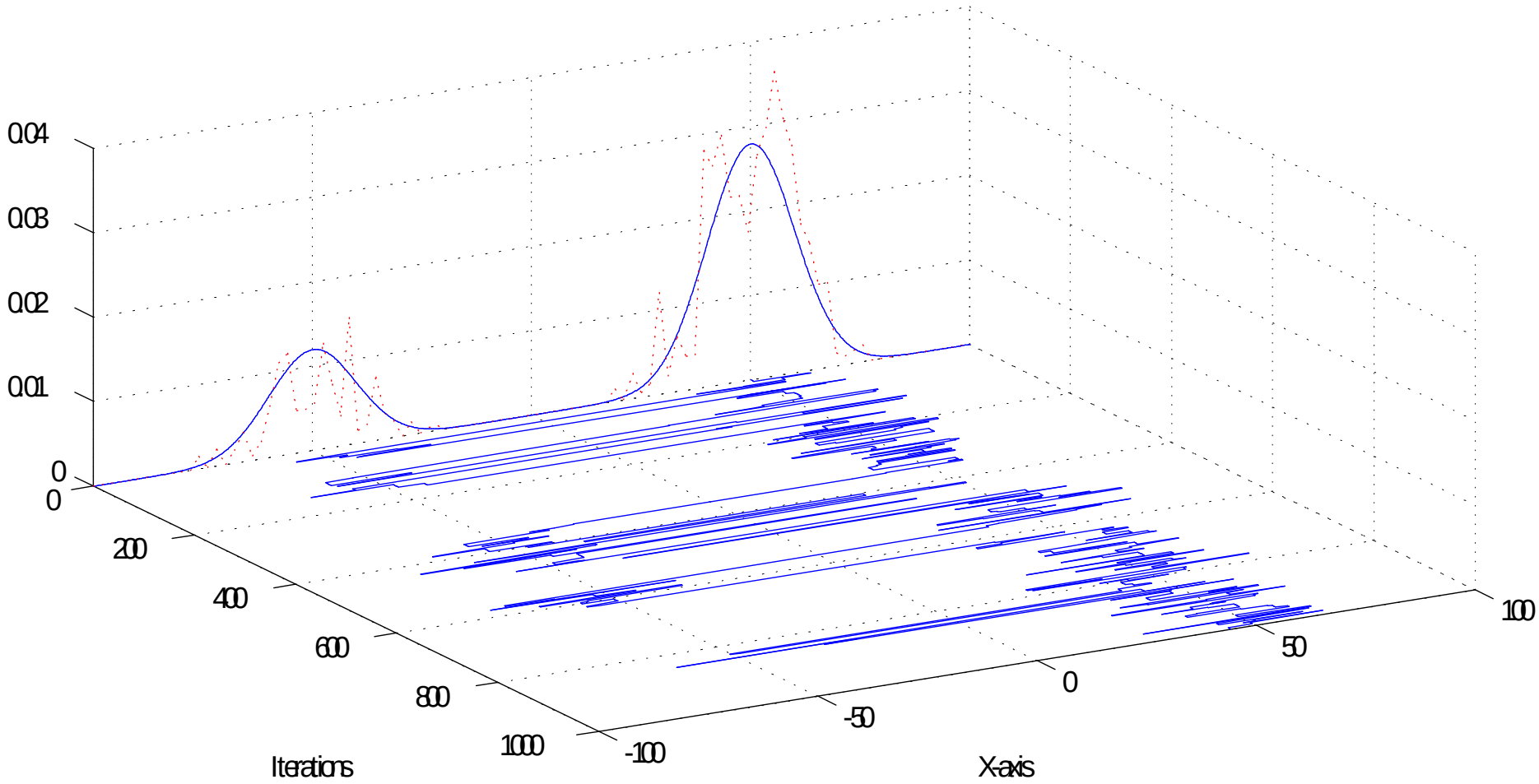

Figure: Bimodal target distributions and simulated Markov chain

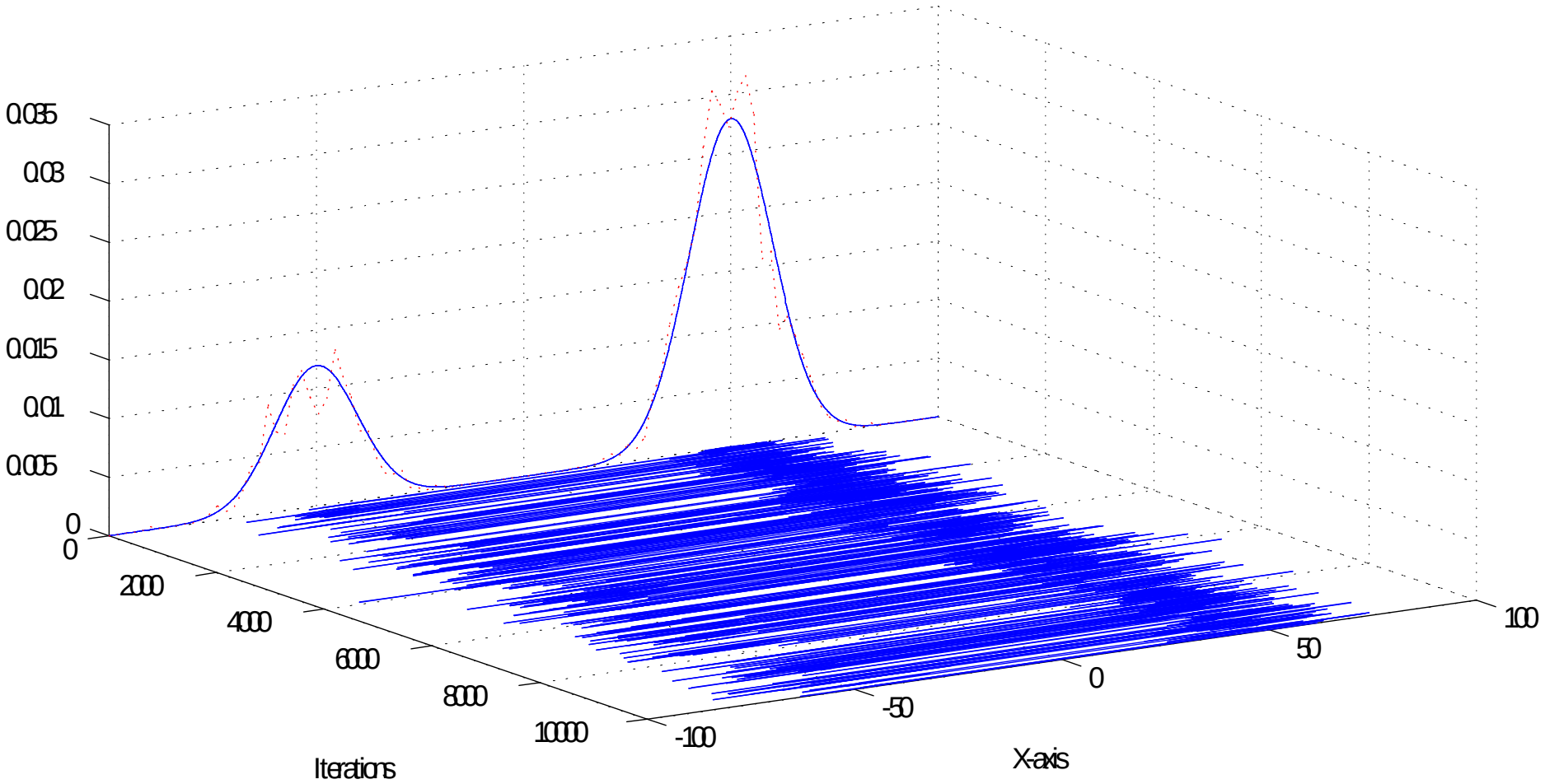

Figure: Bimodal target distributions and simulated Markov chain

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.
- Assume that we have stored $\{X_n, 1 \leq n \leq N\}$ for N large and wish to estimate $\int_{\mathbb{X}} \varphi(x)\pi(x)dx$.

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.
- Assume that we have stored $\{X_n, 1 \leq n \leq N\}$ for N large and wish to estimate $\int_{\mathbb{X}} \varphi(x)\pi(x)dx$.
- In the light of the numerical experiments, one can suggest the estimator

$$\frac{1}{N}\sum_{n=1}^{N} \varphi(X_n).$$

which is exactly the estimator that we would use if $\{X_n, 1 \leq n \leq N\}$ were independent.

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.
- Assume that we have stored $\{X_n, 1 \leq n \leq N\}$ for N large and wish to estimate $\int_{\mathbb{X}} \varphi(x)\pi(x)dx$.
- In the light of the numerical experiments, one can suggest the estimator

$$\frac{1}{N}\sum_{n=1}^{N} \varphi(X_n).$$

  which is exactly the estimator that we would use if $\{X_n, 1 \leq n \leq N\}$ were independent.
- In fact, it can be proved, under relatively mild conditions, that such an estimator is consistent *despite the fact that the samples are NOT independent!* Under additional conditions, a CLT also holds with a rate of CV in $1/\sqrt{N}$.

Markov chains for Monte Carlo

To summarize, we are interested in Markov chains with transition kernel P which have the following three important properties observed above:

- The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, *i.e.* $\int_{\mathbb{X}} \pi(x)P(x,y)dx = \pi(y)$.

Markov chains for Monte Carlo

To summarize, we are interested in Markov chains with transition kernel P which have the following three important properties observed above:

- The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, *i.e.* $\int_{\mathbb{X}} \pi(x)P(x,y)dx = \pi(y)$.
- The successive distributions of the Markov chains converge towards π.

Markov chains for Monte Carlo

To summarize, we are interested in Markov chains with transition kernel P which have the following three important properties observed above:

- The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, *i.e.* $\int_{\mathbb{X}} \pi(x)P(x,y)dx = \pi(y)$.
- The successive distributions of the Markov chains converge towards π.
- The estimator $\frac{1}{N}\sum_{n=1}^{N} \varphi(X_n)$ converges towards $\mathbb{E}_\pi(\varphi(X))$ and asymptotically $X_n \sim \pi$

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which admits $\pi(x)$ as their invariant distribution.

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which admits $\pi(x)$ as their invariant distribution.
- The "art" of MCMC consists of coming up with good ones.

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which admits $\pi(x)$ as their invariant distribution.
- The "art" of MCMC consists of coming up with good ones.
- Convergence is ensured under very weak assumptions; namely irreducibility and aperiodicity.

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which admits $\pi(x)$ as their invariant distribution.
- The "art" of MCMC consists of coming up with good ones.
- Convergence is ensured under very weak assumptions; namely irreducibility and aperiodicity.
- It is usually very easy to establish that an MCMC sampler converges towards π but very difficult to obtain rates of convergence.

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, \theta_0^2\right)$.

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, \theta_0^2\right)$.
- Iteration i; $i \geq 1$

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, \theta_0^2\right)$.
- Iteration i; $i \geq 1$
  - Sample $\theta_i^1 \sim \pi\left(\theta^1 \middle| \theta_{i-1}^2\right)$.

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, \theta_0^2\right)$.
- Iteration i; $i \geq 1$
  - Sample $\theta_i^1 \sim \pi\left(\theta^1 \middle| \theta_{i-1}^2\right)$.
  - Sample $\theta_i^2 \sim \pi\left(\theta^2 \middle| \theta_i^1\right)$.

Two component Gibbs sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = \left(\theta^1, \theta^2\right)$. Then the 2 component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, \theta_0^2\right)$.
- Iteration i; $i \geq 1$
  - Sample $\theta_i^1 \sim \pi\left(\theta^1 \middle| \theta_{i-1}^2\right)$.
  - Sample $\theta_i^2 \sim \pi\left(\theta^2 \middle| \theta_i^1\right)$.
- Sampling from these conditional is often feasible even when sampling from the joint is impossible (e.g. nuclear pump data).

- Clearly $\left\{\left(\theta_i^1, \theta_i^2\right)\right\}$ is a Markov chain and its transition kernel is

$$P\left(\left(\theta^1, \theta^2\right), \left(\widetilde{\theta}^1, \widetilde{\theta}^2\right)\right) = \pi\left(\widetilde{\theta}^1 \middle| \theta^2\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right).$$

- Clearly $\left\{\left(\theta_i^1, \theta_i^2\right)\right\}$ is a Markov chain and its transition kernel is

$$P\left(\left(\theta^1, \theta^2\right), \left(\widetilde{\theta}^1, \widetilde{\theta}^2\right)\right) = \pi\left(\widetilde{\theta}^1 \middle| \theta^2\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right).$$

- Then $\int\int \pi\left(\theta^1, \theta^2\right) P\left(\left(\theta^1, \theta^2\right), \left(\widetilde{\theta}^1, \widetilde{\theta}^2\right)\right) d\theta^1 d\theta^2$ satisfies

$$\begin{aligned}
& \int\int \pi\left(\theta^1, \theta^2\right) \pi\left(\widetilde{\theta}^1 \middle| \theta^2\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right) d\theta^1 d\theta^2 \\
= & \int \pi\left(\theta^2\right) \pi\left(\widetilde{\theta}^1 \middle| \theta^2\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right) d\theta^2 \\
= & \int \pi\left(\widetilde{\theta}^1, \theta^2\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right) d\theta^2 \\
= & \pi\left(\widetilde{\theta}^1\right) \pi\left(\widetilde{\theta}^2 \middle| \widetilde{\theta}^1\right) = \pi\left(\widetilde{\theta}^1, \widetilde{\theta}^2\right)
\end{aligned}$$

Irreducibility

- This does not ensure that the Gibbs sampler does converge towards the invariant distribution!

Irreducibility

- This does not ensure that the Gibbs sampler does converge towards the invariant distribution!
- Additionally it is required to ensure *irreducibility*: loosely speaking the Markov chain can move to any set A such that $\pi(A) > 0$ for (almost) any starting point.

Irreducibility

- This does not ensure that the Gibbs sampler does converge towards the invariant distribution!
- Additionally it is required to ensure *irreducibility*: loosely speaking the Markov chain can move to any set A such that $\pi(A) > 0$ for (almost) any starting point.
- This ensures that

$$\frac{1}{N}\sum_{n=1}^{N}\varphi\left(\theta_n^1,\theta_n^2\right)\rightarrow\int\varphi\left(\theta^1,\theta^2\right)\pi\left(\theta^1,\theta^2\right)d\theta^1 d\theta^2$$

but NOT that asymptotically $\left(\theta_n^1,\theta_n^2\right)\sim\pi$.

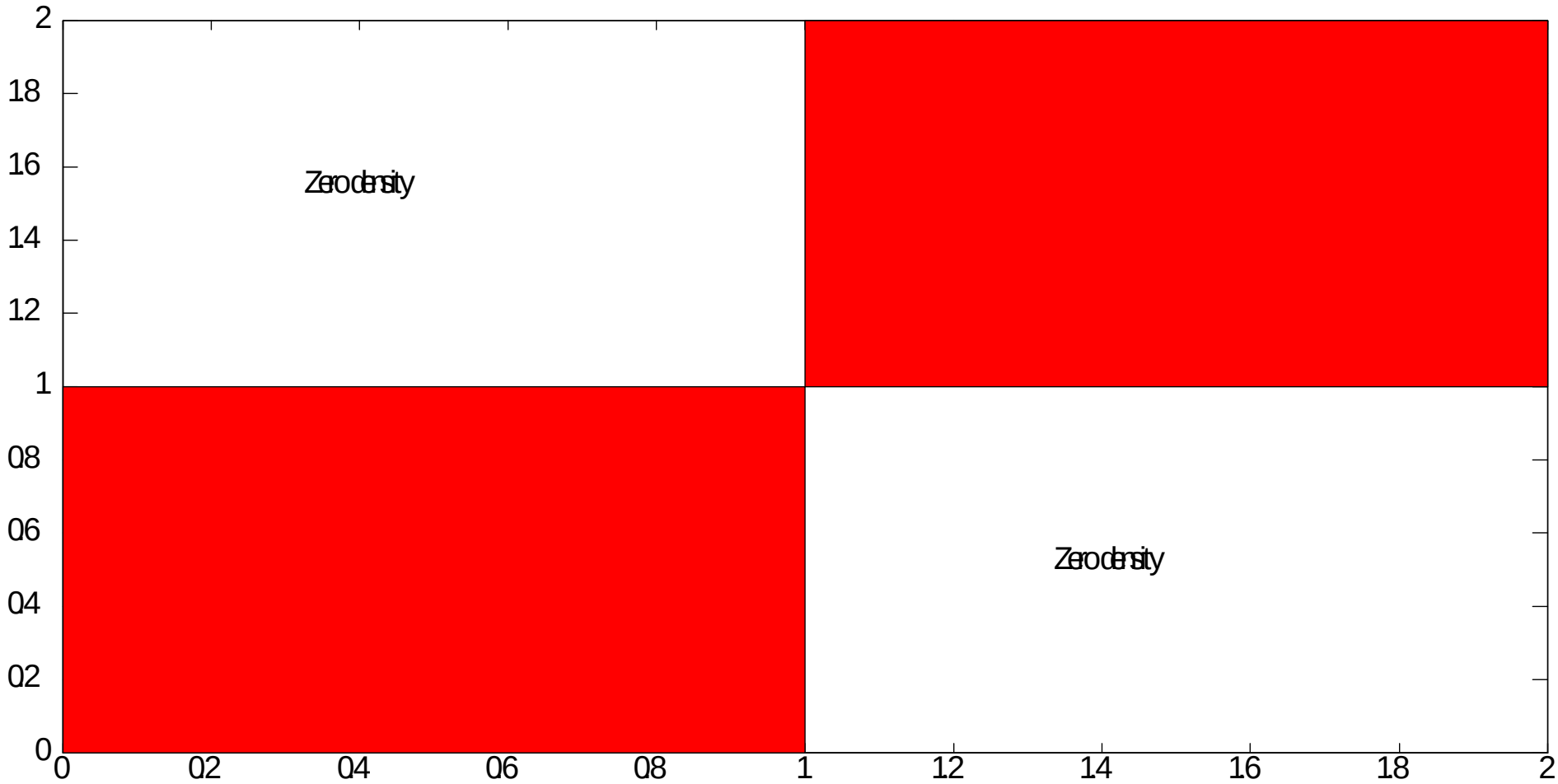

Figure: A distribution that can lead to a reducible Gibbs sampler.

Aperiodicity

- Consider a simple example where $\mathbb{X} = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = \frac{1}{2}$.

Aperiodicity

- Consider a simple example where $\mathbb{X} = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = \frac{1}{2}$.
- However, we know that if the chain starts in $X_0 = 1$, then $X_{2n} = 1$ and $X_{2n+1} = 0$ for any n.

Aperiodicity

- Consider a simple example where $\mathbb{X} = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = \frac{1}{2}$.
- However, we know that if the chain starts in $X_0 = 1$, then $X_{2n} = 1$ and $X_{2n+1} = 0$ for any n.
- We have

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \to \int \varphi(x) \pi(x) dx$$

but clearly X_n is NOT distributed according to π.

Aperiodicity

- Consider a simple example where $\mathbb{X} = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = \frac{1}{2}$.
- However, we know that if the chain starts in $X_0 = 1$, then $X_{2n} = 1$ and $X_{2n+1} = 0$ for any n.
- We have

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \to \int \varphi(x) \pi(x) dx$$

  but clearly X_n is NOT distributed according to π.
- You need to make sure that you do NOT explore the space in a periodic way to ensure that $X_n \sim \pi$ asymptotically.

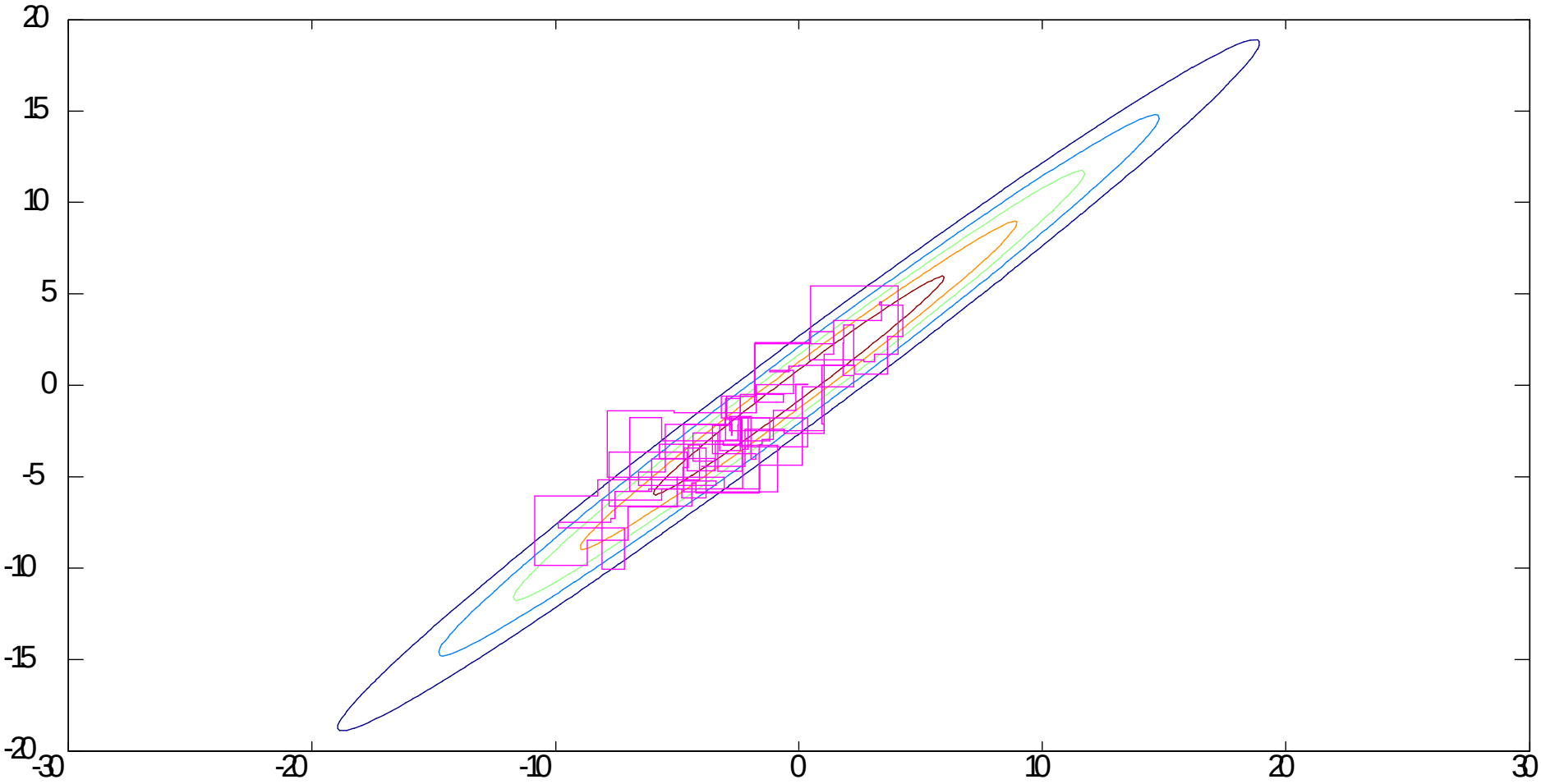

Figure: Even when irreducibility and aperiodicity are ensured, the Gibbs sampler can still converge very slowly.

Deterministic Scan Gibbs Sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.

Deterministic Scan Gibbs Sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.

Deterministic Scan Gibbs Sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:

Deterministic Scan Gibbs Sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$

Deterministic Scan Gibbs Sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.
- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $\theta_i^k \sim \pi\left(\theta^k \middle| \theta_i^{-k}\right)$ where $\theta_i^{-k} = \left(\theta_i^1, ..., \theta_i^{k-1}, \theta_{i-1}^{k+1}, ..., \theta_{i-1}^p\right)$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:
  - Sample $K \sim U_{\{1,...,p\}}$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:
  - Sample $K \sim U_{\{1,...,p\}}$.
  - Set $\theta_i^{-K} = \theta_{i-1}^{-K}$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $\theta_0 = \left(\theta_0^1, ..., \theta_0^p\right)$.
- Iteration i; $i \geq 1$:
  - Sample $K \sim U_{\{1,...,p\}}$.
  - Set $\theta_i^{-K} = \theta_{i-1}^{-K}$.
  - Sample $\theta_i^K \sim \pi\left(\theta^K \middle| \theta_i^{-K}\right)$ where $\theta_i^{-K} = \left(\theta_i^1, ..., \theta_i^{K-1}, \theta_i^{K+1}, ..., \theta_i^p\right)$.

Tricks of the trade

- Try to have as few "blocks" as possible.

Tricks of the trade

- Try to have as few "blocks" as possible.
- Put the most correlated variables in the same block.

Tricks of the trade

- Try to have as few "blocks" as possible.
- Put the most correlated variables in the same block.
- If necessary, reparametrize the model to achieve this.

Tricks of the trade

- Try to have as few "blocks" as possible.
- Put the most correlated variables in the same block.
- If necessary, reparametrize the model to achieve this.
- Integrate analytically as many variables as possible: pretty algorithms can be much more inefficient than ugly algorithms.

Tricks of the trade

- Try to have as few "blocks" as possible.
- Put the most correlated variables in the same block.
- If necessary, reparametrize the model to achieve this.
- Integrate analytically as many variables as possible: pretty algorithms can be much more inefficient than ugly algorithms.
- There is no general result telling strategy A is better than strategy B in all cases: you need experience.

Application to Simulation of Fractal Images

- Consider a 2D black and white 'target' image. We define an distribution ν which assigns $1/P$ mass on each black point and zero on white points where P is the number of black points.

Application to Simulation of Fractal Images

- Consider a 2D black and white 'target' image. We define an distribution ν which assigns $1/P$ mass on each black point and zero on white points where P is the number of black points.
- Now we consider the following simple Markov process on $\mathbb{R}^2$ with

$$P(x, y) = \sum_{i=1}^{k} w_i \delta_{A_i x + b_i}(y)$$

and we select $\{w_i, A_i, b_i\}$ such that $P(x, dy)$ has an invariant distribution π which is an approximation of ν.

Application to Simulation of Fractal Images

- Consider a 2D black and white 'target' image. We define an distribution ν which assigns $1/P$ mass on each black point and zero on white points where P is the number of black points.
- Now we consider the following simple Markov process on $\mathbb{R}^2$ with

$$P(x, y) = \sum_{i=1}^{k} w_i \delta_{A_i x + b_i}(y)$$

and we select $\{w_i, A_i, b_i\}$ such that $P(x, dy)$ has an invariant distribution π which is an approximation of ν.

- To find $\{w_i, A_i, b_i\}$, we write

$$\begin{aligned} \int \pi(x) P(x, y) f(y) dxdy &= \sum_{i=1}^{k} w_i \int f(A_i x + b_i) \pi(x) dx \\ &= \int f(x) \pi(x) dx \approx \int f(x) \nu(x) dx \end{aligned}$$

and solve approximately the equations for some functions f (linear or low order polynoms).

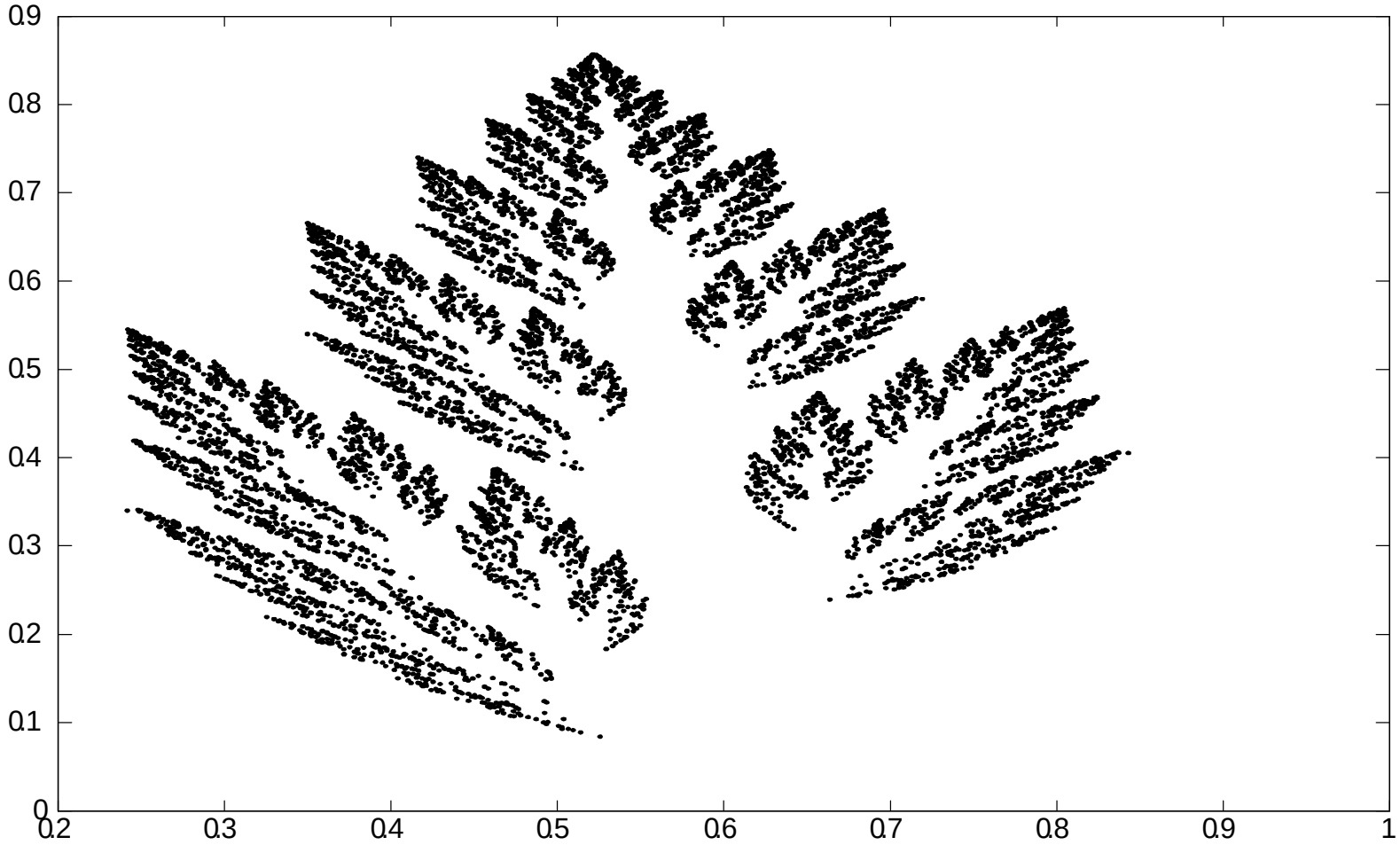

Figure: Fractal image generated using Iterated random functions with $k = 2$ and $N = 10000$ samples

Gibbs Sampler for Bayesian Variable Selection

- We select the following model

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where we assume $\mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$ and for $\alpha^2 << 1$

$$\beta_i \sim \frac{1}{2}\mathcal{N}\left(0, \alpha^2\delta^2\sigma^2\right) + \frac{1}{2}\mathcal{N}\left(0, \delta^2\sigma^2\right)$$

Gibbs Sampler for Bayesian Variable Selection

- We select the following model

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where we assume $\mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$ and for $\alpha^2 << 1$

$$\beta_i \sim \frac{1}{2}\mathcal{N}\left(0, \alpha^2\delta^2\sigma^2\right) + \frac{1}{2}\mathcal{N}\left(0, \delta^2\sigma^2\right)$$

- We introduce a latent variable $\gamma_i \in \{0,1\}$ such that

$$\begin{array}{l} \Pr(\gamma_i = 0) = \Pr(\gamma_i = 1) = \frac{1}{2}, \\ \beta_i | \gamma_i = 0 \sim \mathcal{N}\left(0, \alpha^2\delta^2\sigma^2\right), \quad \beta_i | \gamma_i = 1 \sim \mathcal{N}\left(0, \delta^2\sigma^2\right). \end{array}$$

- We have parameters $\left(\beta_{1:p}, \gamma_{1:p}, \sigma^2\right)$ and observe $D = \{x_i, y_i\}_{i=1}^n$.

- We have parameters $\left(\beta_{1:p}, \gamma_{1:p}, \sigma^2\right)$ and observe $D = \{x_i, y_i\}_{i=1}^n$.
- A potential Gibbs sampler consists of sampling iteratively from $p\left(\beta_{1:p} \middle| D, \gamma_{1:p}, \sigma^2\right)$ (Gaussian), $p\left(\sigma^2 \middle| D, \gamma_{1:p}, \beta_{1:p}\right)$ (inverse-Gamma) and $p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right)$.

- We have parameters $\left(\beta_{1:p}, \gamma_{1:p}, \sigma^2\right)$ and observe $D = \{x_i, y_i\}_{i=1}^n$.
- A potential Gibbs sampler consists of sampling iteratively from $p\left(\beta_{1:p} \middle| D, \gamma_{1:p}, \sigma^2\right)$ (Gaussian), $p\left(\sigma^2 \middle| D, \gamma_{1:p}, \beta_{1:p}\right)$ (inverse-Gamma) and $p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right)$.
- In particular

$$p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right) = \prod_{i=1}^{p} p\left(\gamma_i \middle| \beta_i, \sigma^2\right)$$

and

$$p\left(\gamma_i = 1 \middle| \beta_i, \sigma^2\right) = \frac{\frac{1}{\sqrt{2\pi}\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\delta^2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\delta^2\sigma^2}\right) + \frac{1}{\sqrt{2\pi}\alpha\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\alpha^2\delta^2\sigma^2}\right)}.$$

- We have parameters $\left(\beta_{1:p}, \gamma_{1:p}, \sigma^2\right)$ and observe $D = \{x_i, y_i\}_{i=1}^n$.
- A potential Gibbs sampler consists of sampling iteratively from $p\left(\beta_{1:p} \middle| D, \gamma_{1:p}, \sigma^2\right)$ (Gaussian), $p\left(\sigma^2 \middle| D, \gamma_{1:p}, \beta_{1:p}\right)$ (inverse-Gamma) and $p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right)$.
- In particular

$$p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right) = \prod_{i=1}^{p} p\left(\gamma_i \middle| \beta_i, \sigma^2\right)$$

and

$$p\left(\gamma_i = 1 \middle| \beta_i, \sigma^2\right) = \frac{\frac{1}{\sqrt{2\pi}\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\delta^2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\delta^2\sigma^2}\right) + \frac{1}{\sqrt{2\pi}\alpha\delta\sigma} \exp\left(-\frac{\beta_i^2}{2\alpha^2\delta^2\sigma^2}\right)}.$$

- The Gibbs sampler becomes reducible as α goes to zero.

- This is the result of bad modelling and bad algorithm. You would like to put $\alpha \simeq 0$ and write

$$Y = \sum_{i=1}^{p} \gamma_i \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where $\gamma_i = 1$ if X_i is included or $\gamma_i = 0$ otherwise. However this suggests that β_i is defined even when $\gamma_i = 0$.

- This is the result of bad modelling and bad algorithm. You would like to put $\alpha \simeq 0$ and write

$$Y = \sum_{i=1}^{p} \gamma_i \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where $\gamma_i = 1$ if X_i is included or $\gamma_i = 0$ otherwise. However this suggests that β_i is defined even when $\gamma_i = 0$.

- A neater way to write such models is to write

$$Y = \sum_{\{i:\gamma_i=1\}} \beta_i X_i + \sigma V = \beta_\gamma^{\mathsf{T}} X_\gamma + \sigma V$$

where, for a vector $\gamma = (\gamma_1, ..., \gamma_p)$, $\beta_\gamma = \{\beta_i : \gamma_i = 1\}$, $X_\gamma = \{X_i : \gamma_i = 1\}$ and $n_\gamma = \sum_{i=1}^{p} \gamma_i$.

- This is the result of bad modelling and bad algorithm. You would like to put $\alpha \simeq 0$ and write

$$Y = \sum_{i=1}^{p} \gamma_i \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where $\gamma_i = 1$ if X_i is included or $\gamma_i = 0$ otherwise. However this suggests that β_i is defined even when $\gamma_i = 0$.

- A neater way to write such models is to write

$$Y = \sum_{\{i:\gamma_i=1\}} \beta_i X_i + \sigma V = \beta_\gamma^{\mathsf{T}} X_\gamma + \sigma V$$

where, for a vector $\gamma = (\gamma_1, ..., \gamma_p)$, $\beta_\gamma = \{\beta_i : \gamma_i = 1\}$, $X_\gamma = \{X_i : \gamma_i = 1\}$ and $n_\gamma = \sum_{i=1}^{p} \gamma_i$.

- Prior distributions

$$\pi_\gamma\left(\beta_\gamma, \sigma^2\right) = \mathcal{N}\left(\beta_\gamma; 0, \delta^2\sigma^2 I_{n_\gamma}\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$$

and $\pi(\gamma) = \prod_{i=1}^{p} \pi(\gamma_i) = 2^{-p}$.

- We are interested in sampling from the trans-dimensional distribution $\pi\left(\gamma, \beta_\gamma, \sigma^2 \middle| D\right)$.

- We are interested in sampling from the trans-dimensional distribution $\pi\left(\gamma, \beta_\gamma, \sigma^2 \middle| D\right)$.
- However, we know that

$$\pi\left(\gamma, \beta_\gamma, \sigma^2 \middle| D\right) = \pi\left(\gamma \middle| D\right) \pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma\right)$$

where

$$\pi\left(\gamma \middle| D\right) \propto \pi\left(D \middle| \gamma\right) \pi\left(\gamma\right)$$

and

$$\pi\left(D \middle| \gamma\right) = \int \pi\left(D, \beta_\gamma, \sigma^2 \middle| \gamma\right) d\beta_\gamma d\sigma^2.$$

- $\pi(\gamma|D)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.

- $\pi\left(\gamma \middle| D\right)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.
- Initialization: Select deterministically or randomly $\gamma_0 = \left(\gamma_0^1, ..., \gamma_0^p\right)$.

- $\pi\left(\gamma \middle| D\right)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.
- Initialization: Select deterministically or randomly $\gamma_0 = \left(\gamma_0^1, ..., \gamma_0^p\right)$.
- Iteration i; $i \geq 1$:

- $\pi\left(\gamma \middle| D\right)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.
- Initialization: Select deterministically or randomly $\gamma_0 = \left(\gamma_0^1, ..., \gamma_0^p\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$

- $\pi\left(\gamma \middle| D\right)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.
- Initialization: Select deterministically or randomly $\gamma_0 = \left(\gamma_0^1, ..., \gamma_0^p\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $\gamma_i^k \sim \pi\left(\gamma^k \middle| D, \gamma_i^{-k}\right)$ where $\gamma_i^{-k} = \left(\gamma_i^1, ..., \gamma_i^{k-1}, \gamma_{i-1}^{k+1}, ..., \gamma_{i-1}^p\right)$.

- $\pi\left(\gamma \middle| D\right)$ is a discrete probability distribution with 2^p potential values. We can use the Gibbs sampler to sample from it.
- Initialization: Select deterministically or randomly $\gamma_0 = \left(\gamma_0^1, ..., \gamma_0^p\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $\gamma_i^k \sim \pi\left(\gamma^k \middle| D, \gamma_i^{-k}\right)$ where $\gamma_i^{-k} = \left(\gamma_i^1, ..., \gamma_i^{k-1}, \gamma_{i-1}^{k+1}, ..., \gamma_{i-1}^p\right)$.
  - Optional step: Sample $\left(\beta_{\gamma,i}, \sigma_i^2\right) \sim \pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma\right)$.

- This very simple sampler is much more efficient than the previous one.

- This very simple sampler is much more efficient than the previous one.
- However, it can also mix very slowly because the components are updated one at a time.

- This very simple sampler is much more efficient than the previous one.
- However, it can also mix very slowly because the components are updated one at a time.
- Updating correlated components together would increase significantly the convergence speed of the algorithm at the cost of an increased complexity.

Finite Mixture Models

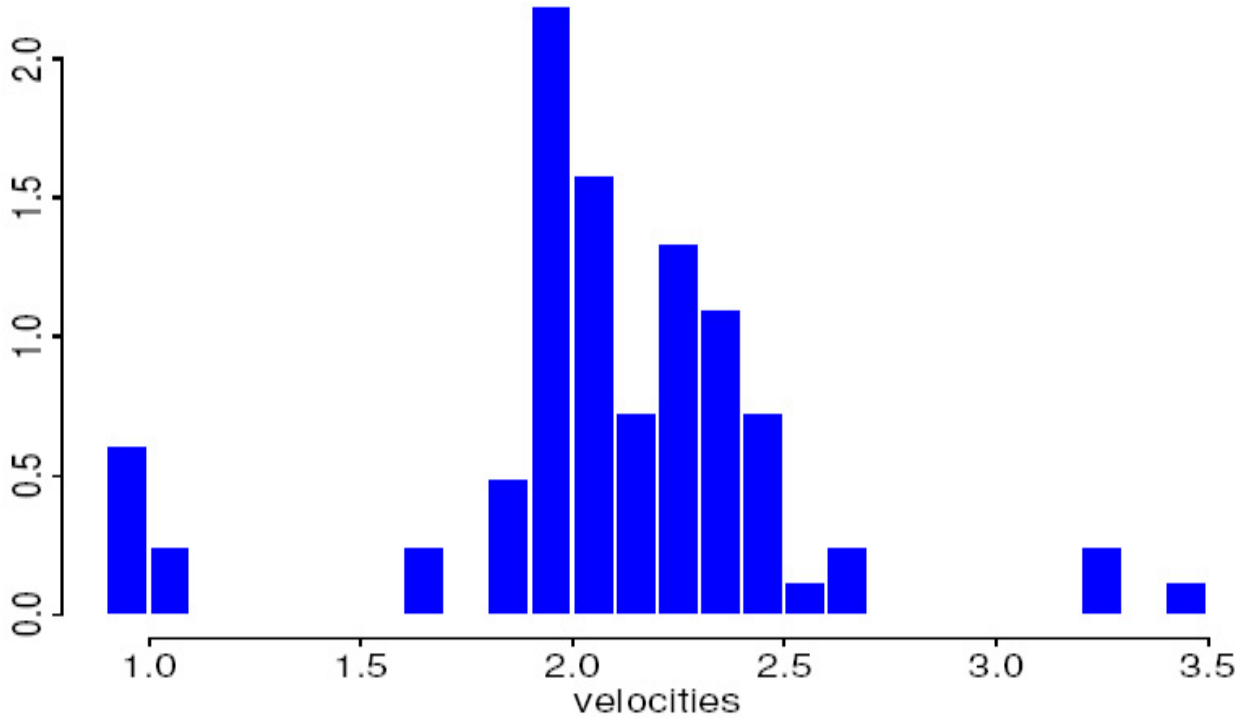

Figure: Velocity (km/sc) of galaxies in the Corona Borealis Region

- Consider the case where one has n data X_i

$$X_i \overset{\text{i.i.d}}{\sim} \sum_{k=1}^{K} p_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where K is fixed and $\theta = \left\{\mu_k, \sigma_k^2, p_k\right\}_{k=1,\ldots,K}$ are unknown.

- Consider the case where one has n data X_i

$$X_i \overset{\text{i.i.d}}{\sim} \sum_{k=1}^{K} p_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where K is fixed and $\theta = \left\{\mu_k, \sigma_k^2, p_k\right\}_{k=1,\ldots,K}$ are unknown.

- A standard approach consists of finding a local maximum of the log-likelihood

$$\log f\left(x_{1:n} \middle| \theta\right) = \sum_{i=1}^{n} \log f\left(x_i \middle| \theta\right)$$

where

$$f\left(x \middle| \theta\right) = \sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(x-\mu_k\right)^2}{2\sigma_k^2}\right).$$

- Consider the case where one has n data X_i

$$X_i \overset{\text{i.i.d}}{\sim} \sum_{k=1}^{K} p_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where K is fixed and $\theta = \left\{\mu_k, \sigma_k^2, p_k\right\}_{k=1,\ldots,K}$ are unknown.

- A standard approach consists of finding a local maximum of the log-likelihood

$$\log f\left(x_{1:n} \middle| \theta\right) = \sum_{i=1}^{n} \log f\left(x_i \middle| \theta\right)$$

where $$f\left(x \middle| \theta\right) = \sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(x-\mu_k\right)^2}{2\sigma_k^2}\right).$$

- *Problem*: The likelihood is unbounded!

Bayesian Model

- We consider the Bayesian framework where we set priors

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

Bayesian Model

- We consider the Bayesian framework where we set priors

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

- We use the following (conditionally conjugate) priors where

$$\begin{aligned} (p_1, ..., p_K) &\sim \mathcal{D}(\gamma_1, ..., \gamma_K). \\ \mu_k | \sigma_k^2 &\sim \mathcal{N}\left(\alpha_k, \frac{\sigma_k^2}{\lambda_k}\right), \ \sigma_k^2 \sim \mathcal{IG}\left(\frac{\lambda_k + 3}{2}, \frac{\beta_k}{2}\right). \end{aligned}$$

Bayesian Model

- We consider the Bayesian framework where we set priors

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

- We use the following (conditionally conjugate) priors where

$$\begin{aligned} (p_1, ..., p_K) &\sim \mathcal{D}(\gamma_1, ..., \gamma_K). \\ \mu_k | \sigma_k^2 &\sim \mathcal{N}\left(\alpha_k, \frac{\sigma_k^2}{\lambda_k}\right), \ \sigma_k^2 \sim \mathcal{IG}\left(\frac{\lambda_k + 3}{2}, \frac{\beta_k}{2}\right). \end{aligned}$$

- It is impossible to use the Gibbs sampler to sample from $\pi(\theta | x_{1:n})$.

Latent Variables

- Like in the EM, we can introduce the missing data $Z_i \in \{1, ..., K\}$ such that

$$X_i | \, Z_i \sim \mathcal{N}\left(\mu_{Z_i}, \sigma^2_{Z_i}\right)$$

and

$$\Pr\left(Z_i = k\right) = p_k .$$

Latent Variables

- Like in the EM, we can introduce the missing data $Z_i \in \{1, ..., K\}$ such that

$$X_i | Z_i \sim \mathcal{N}\left(\mu_{Z_i}, \sigma^2_{Z_i}\right)$$

and

$$\Pr\left(Z_i = k\right) = p_k .$$

- The "complete" likelihood admits a simple form

$$\pi\left(x_{1:n}, z_{1:n} | \theta\right) = \prod_{k=1}^{n} f\left(x_i | \theta, z_i\right) \pi\left(z_i | \theta\right) .$$

Latent Variables

- Like in the EM, we can introduce the missing data $Z_i \in \{1, ..., K\}$ such that

$$X_i | Z_i \sim \mathcal{N}\left(\mu_{Z_i}, \sigma^2_{Z_i}\right)$$

and

$$\Pr\left(Z_i = k\right) = p_k .$$

- The "complete" likelihood admits a simple form

$$\pi\left(x_{1:n}, z_{1:n} | \theta\right) = \prod_{k=1}^{n} f\left(x_i | \theta, z_i\right) \pi\left(z_i | \theta\right) .$$

- Thus we propose to sample the joint posterior $\pi\left(\theta, z_{1:n} | y_{1:n}\right)$ using the Gibbs sampler; that is sampling iteratively from $\pi\left(\theta | y_{1:n}, z_{1:n}\right)$ and $\pi\left(z_{1:n} | y_{1:n}, \theta\right)$.

Gibbs Sampler for Finite Mixture Distributions

- We have

$$\pi\left(z_{1:n}\middle|\,\theta, x_{1:n}\right) = \prod_{i=1}^{n} \pi\left(z_i\middle|\,\theta, x_i\right)$$

where

$$\pi\left(z_i = j\middle|\,\theta, x_i\right) = \frac{f\left(x_i\middle|\,\theta, j\right) p_j}{\sum_{k=1}^{K} f\left(x_i\middle|\,\theta, k\right) p_k}.$$

Gibbs Sampler for Finite Mixture Distributions

- We have

$$\pi\left(z_{1:n}\middle|\,\theta, x_{1:n}\right) = \prod_{i=1}^{n} \pi\left(z_i\middle|\,\theta, x_i\right)$$

where

$$\pi\left(z_i = j\middle|\,\theta, x_i\right) = \frac{f\left(x_i\middle|\,\theta, j\right) p_j}{\sum_{k=1}^{K} f\left(x_i\middle|\,\theta, k\right) p_k}.$$

- We have

$$\pi\left(\theta\middle|\,z_{1:n}, x_{1:n}\right) = \pi\left(p_1, \ldots, p_K\middle|\,z_{1:n}\right) \prod_{k=1}^{K} \pi\left(\mu_k, \sigma_k^2\middle|\,z_{1:n}, x_{1:n}\right)$$

- Introducing

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

- Introducing

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

- We have the full conditionals

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

- Introducing

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

- We have the full conditionals

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

- It is thus trivial to implement the Gibbs sampler.

Simulation Results

- Consider some $n = 100$ simulated data

$$X_i \sim 0.3\mathcal{N}(-2, 1) + 0.7\mathcal{N}(2, 1),$$

i.e. we have well-separated components.

Simulation Results

- Consider some $n = 100$ simulated data

$$X_i \sim 0.3\mathcal{N}(-2, 1) + 0.7\mathcal{N}(2, 1),$$

  i.e. we have well-separated components.
- We set $\gamma_k = 1$, $\alpha_k = 0$, $\lambda_k = 0.01$, $\beta_k = 0.01$ and run the Gibbs sampler for 10000 iterations.

Simulation Results

- Consider some $n = 100$ simulated data

$$X_i \sim 0.3\mathcal{N}(-2, 1) + 0.7\mathcal{N}(2, 1),$$

  i.e. we have well-separated components.
- We set $\gamma_k = 1$, $\alpha_k = 0$, $\lambda_k = 0.01$, $\beta_k = 0.01$ and run the Gibbs sampler for 10000 iterations.
- We obtain $\widehat{\mathbb{E}}(\mu_1 | x_{1:n}) = 2.17$, $\widehat{\mathbb{E}}(\mu_2 | x_{1:n}) = -1.89$, $\widehat{\mathbb{E}}(\sigma_1^2 | x_{1:n}) = 0.92$, $\widehat{\mathbb{E}}(\sigma_2^2 | x_{1:n}) = 1.3$, $\widehat{\mathbb{E}}(p_1 | x_{1:n}) = 0.32$ and $\widehat{\mathbb{E}}(p_2 | x_{1:n}) = 0.68$.

Simulation Results

- Consider some $n = 100$ simulated data

$$X_i \sim 0.3\mathcal{N}(-2, 1) + 0.7\mathcal{N}(2, 1),$$

  i.e. we have well-separated components.
- We set $\gamma_k = 1$, $\alpha_k = 0$, $\lambda_k = 0.01$, $\beta_k = 0.01$ and run the Gibbs sampler for 10000 iterations.
- We obtain $\widehat{\mathbb{E}}(\mu_1 | x_{1:n}) = 2.17$, $\widehat{\mathbb{E}}(\mu_2 | x_{1:n}) = -1.89$, $\widehat{\mathbb{E}}(\sigma_1^2 | x_{1:n}) = 0.92$, $\widehat{\mathbb{E}}(\sigma_2^2 | x_{1:n}) = 1.3$, $\widehat{\mathbb{E}}(p_1 | x_{1:n}) = 0.32$ and $\widehat{\mathbb{E}}(p_2 | x_{1:n}) = 0.68$.
- Increasing the number of iterations to 100000, I obtain similar results. Any good?

- Your algorithm does not work! Indeed we know that

$$\begin{aligned}\mathbb{E}\left(\mu_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(\mu_2 \middle| x_{1:n}\right), \ \mathbb{E}\left(\sigma_1^2 \middle| x_{1:n}\right) = \mathbb{E}\left(\sigma_2^2 \middle| x_{1:n}\right), \\ \mathbb{E}\left(p_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(p_2 \middle| x_{1:n}\right) = 0.5.\end{aligned}$$

- Your algorithm does not work! Indeed we know that

$$\begin{aligned}\mathbb{E}\left(\mu_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(\mu_2 \middle| x_{1:n}\right), \ \mathbb{E}\left(\sigma_1^2 \middle| x_{1:n}\right) = \mathbb{E}\left(\sigma_2^2 \middle| x_{1:n}\right), \\ \mathbb{E}\left(p_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(p_2 \middle| x_{1:n}\right) = 0.5.\end{aligned}$$

- This follows because both the prior and likelihood are exchangeable, that is

$$\begin{aligned}&\pi\left(p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2 \middle| x_{1:n}\right) \\ = \ &\pi\left(p_{\zeta(1)}, ..., p_{\zeta(K)}, \mu_{\zeta(1)}, ..., \mu_{\zeta(K)}, \sigma_{\zeta(1)}^2, ..., \sigma_{\zeta(K)}^2 \middle| x_{1:n}\right)\end{aligned}$$

  for any permutation ζ of the labels.

- Your algorithm does not work! Indeed we know that

$$\begin{aligned}\mathbb{E}\left(\mu_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(\mu_2 \middle| x_{1:n}\right), \ \mathbb{E}\left(\sigma_1^2 \middle| x_{1:n}\right) = \mathbb{E}\left(\sigma_2^2 \middle| x_{1:n}\right), \\ \mathbb{E}\left(p_1 \middle| x_{1:n}\right) &= \mathbb{E}\left(p_2 \middle| x_{1:n}\right) = 0.5.\end{aligned}$$

- This follows because both the prior and likelihood are exchangeable, that is

$$\begin{aligned}&\pi\left(p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2 \middle| x_{1:n}\right) \\ = \ &\pi\left(p_{\zeta(1)}, ..., p_{\zeta(K)}, \mu_{\zeta(1)}, ..., \mu_{\zeta(K)}, \sigma_{\zeta(1)}^2, ..., \sigma_{\zeta(K)}^2 \middle| x_{1:n}\right)\end{aligned}$$

  for any permutation ζ of the labels.

- Clearly, conditional expectations are not useful in this case.
  $\Rightarrow$ This does NOT mean that your Bayesian model is useless.

- One can select another point estimates; e.g. the MAP estimate

$$\theta_{MAP} = \arg\max \pi\left(\theta \middle| x_{1:n}\right).$$

- One can select another point estimates; e.g. the MAP estimate

$$\theta_{MAP} = \arg\max \pi\left(\theta \middle| x_{1:n}\right).$$

- Alternatively, constraints can be set on the priors; e.g. we ensure that

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_P$$

- One can select another point estimates; e.g. the MAP estimate

$$\theta_{MAP} = \arg\max \pi\left(\theta \middle| x_{1:n}\right).$$

- Alternatively, constraints can be set on the priors; e.g. we ensure that

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_P$$

- However, this can lead to "strange" shapes of the posteriors and is not natural in most cases.

- One way to improve the algorithm consists of randomly permuting the labels (Fruwirth-Schnatter, JASA, 2002)

$\Rightarrow$ Realistic only if K is moderate because there are $K!$ permutations.

- One way to improve the algorithm consists of randomly permuting the labels (Fruwirth-Schnatter, JASA, 2002)

$\Rightarrow$ Realistic only if K is moderate because there are $K!$ permutations.

- Alternative ways to improve the algorithm include

- One way to improve the algorithm consists of randomly permuting the labels (Fruwirth-Schnatter, JASA, 2002)

$\Rightarrow$ Realistic only if K is moderate because there are $K!$ permutations.

- Alternative ways to improve the algorithm include
  - Not introducing the latent variables and using sampling strategies different from Gibbs.

- One way to improve the algorithm consists of randomly permuting the labels (Fruwirth-Schnatter, JASA, 2002)

$\Rightarrow$ Realistic only if K is moderate because there are $K!$ permutations.

- Alternative ways to improve the algorithm include
  - Not introducing the latent variables and using sampling strategies different from Gibbs.
  - Integrating out θ as the marginal distribution $\pi\left(z_{1:n} \middle| x_{1:n}\right)$ can be computed analytically (for conjugate priors)

- Initialization: Select deterministically or randomly $z_{1:n}^{(0)}$.

- Initialization: Select deterministically or randomly $z_{1:n}^{(0)}$.
- Iteration i; $i \geq 1$:

- Initialization: Select deterministically or randomly $z_{1:n}^{(0)}$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : n$, sample $Z_k^{(i)} \sim \pi\left(z_k \,|\, x_{1:n}, z_{-k}^{(i)}\right)$ where $z_{-k}^{(i)} = \left(z_1^{(i)}, ..., z_{k-1}^{(i)}, z_{k+1}^{(i-1)}, ..., z_n^{(i-1)}\right)$.

- Initialization: Select deterministically or randomly $z_{1:n}^{(0)}$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : n$, sample $Z_k^{(i)} \sim \pi\left(z_k \middle| x_{1:n}, z_{-k}^{(i)}\right)$ where $z_{-k}^{(i)} = \left(z_1^{(i)}, ..., z_{k-1}^{(i)}, z_{k+1}^{(i-1)}, ..., z_n^{(i-1)}\right)$.
  - Sample $\theta^{(i)} \sim \pi\left(\theta \middle| x_{1:n}, z_{1:n}^{(i)}\right)$.

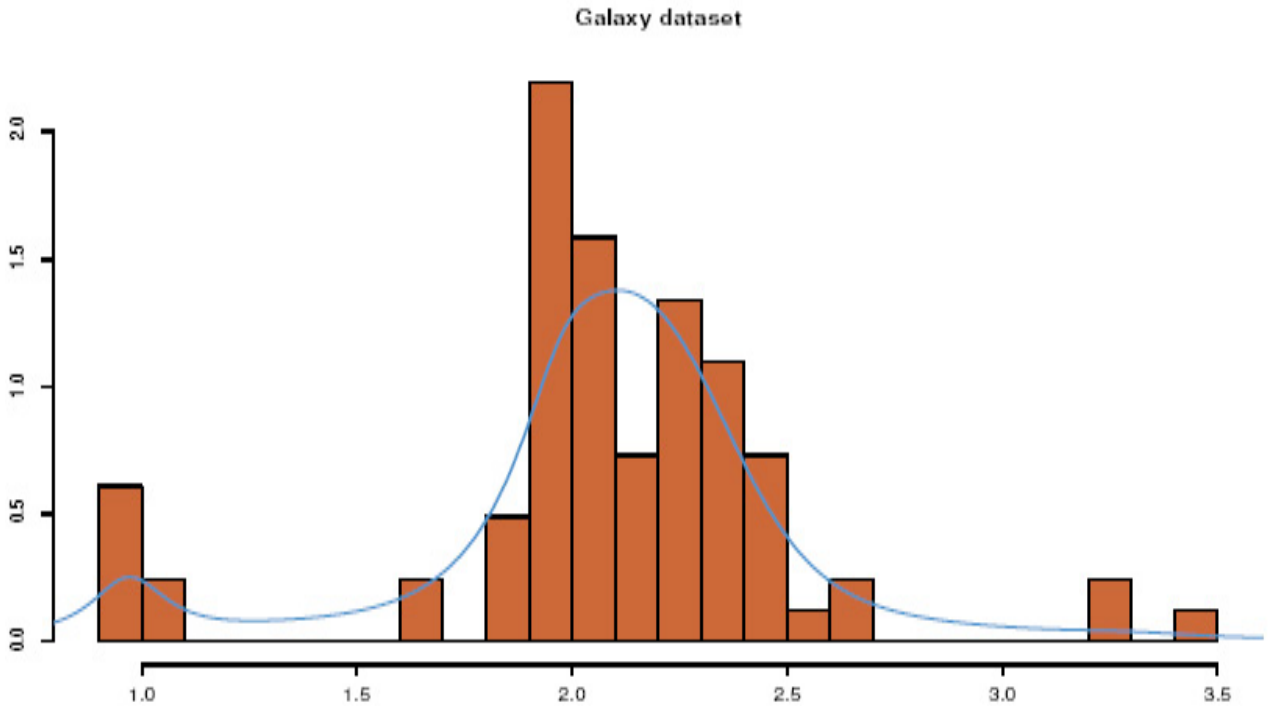

Figure: Predictive distribution for the galaxy dataset.

Discussion

- The Gibbs sampler is a generic tool to sample approximately from high-dimensional distributions.

Discussion

- The Gibbs sampler is a generic tool to sample approximately from high-dimensional distributions.
- Each time you face a problem, you need to think hard about it to design an efficient algorithm.

Discussion

- The Gibbs sampler is a generic tool to sample approximately from high-dimensional distributions.
- Each time you face a problem, you need to think hard about it to design an efficient algorithm.
- Except the choice of the partitions of parameters, the Gibbs sampler is parameter free; this does not mean it is efficient.