2.2– Motivating Example: Nuclear Pump Data

• Multiple failures in a nuclear plant:

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Times	94.32	15.72	62.88	125.76	5.24	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 unknowns consist of $\theta := (\lambda_1, \ldots, \lambda_{10}, \beta)$.

2.1– Summary of Last Lecture

• 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.

• There are two ways to implement incremental strategies.

- Iteratively: Markov chain Monte Carlo.

- Sequentially: Sequential Monte Carlo.

1.1– Outline

• Introduction to Markov chain Monte Carlo

• The Gibbs Sampler

• Examples

Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

2.4– Conditional Distributions

• 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.

2.4– Conditional Distributions

• 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 would not generally converge towards the global maximum of the joint distribution.

2.4– Conditional Distributions

• The conditionals have a familiar form

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

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

• 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 $\lambda_1, \ldots, \lambda_{10}$.

2.3– Bayesian Model for Nuclear Pump Data

• Hierarchical model

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

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

• The posterior distribution is proportional to

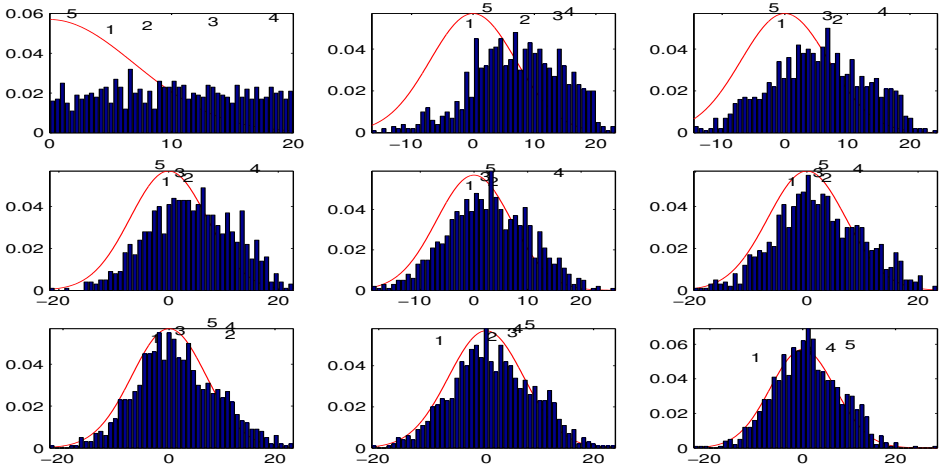
$$\prod_{i=1}^{10} \{(\lambda_i t_i)^{p_i} \exp(-\lambda_i t_i) \lambda_i^{\alpha-1} \exp(-\beta\lambda_i)\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta)$$

$$\propto \prod_{i=1}^{10} \{\lambda_i^{p_i+\alpha-1} \exp(-(t_i+\beta)\lambda_i)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta).$$

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

2.7– Example

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



2.7– Example

• 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

2.6– Example

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

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

• The limiting distribution is

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

• 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.

2.5– Introduction to Markov Chain Monte Carlo

• **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 $A \in \mathcal{B}(\mathbb{X})$

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

and we will write

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

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

$$\frac{1}{N}\sum_{n=1}^{N} \varphi\left(X_n\right) \to \int \varphi\left(x\right)\pi\left(x\right)dx \text{ and/or } X_n \sim \pi.$$

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

2.9– Markov chain Monte Carlo

• 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 $\int_{\mathsf{X}} \varphi(x)\pi(x)dx$.

• In the light of the numerical experiments, one can suggest the estimator

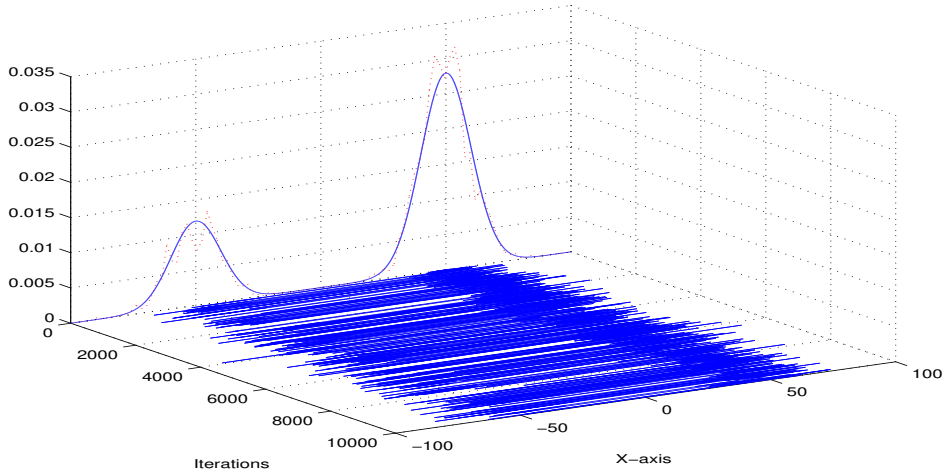
$$\frac{1}{N}\sum_{n=1}^{N} \varphi(X_n).$$

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 $1/\sqrt{N}$.

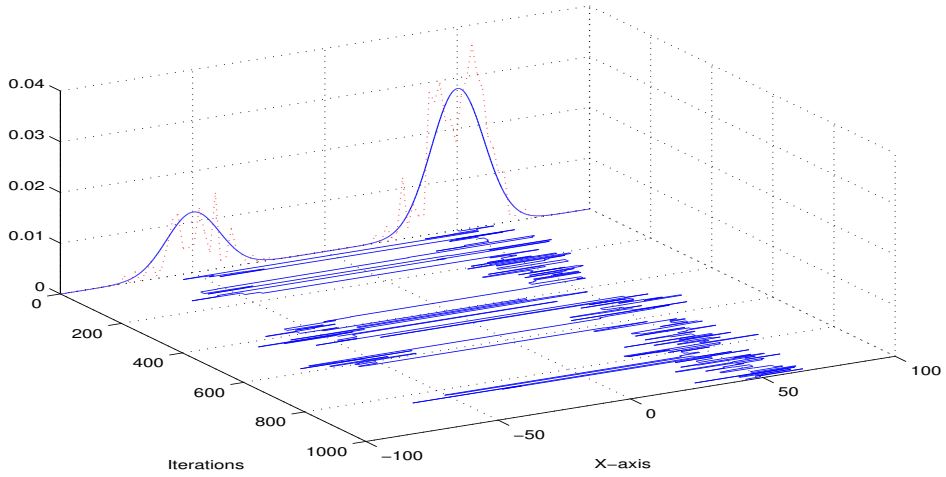
2.8– Example: Mixture of Normals

10000 iterations



2.8– Example: Mixture of Normals

1000 iterations



2.7– Example

• 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.

2.11– Invariant Distribution

• 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

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

2.10– The 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).

2.9– Markov chain Monte Carlo

• 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.

2.9– Markov chain Monte Carlo

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

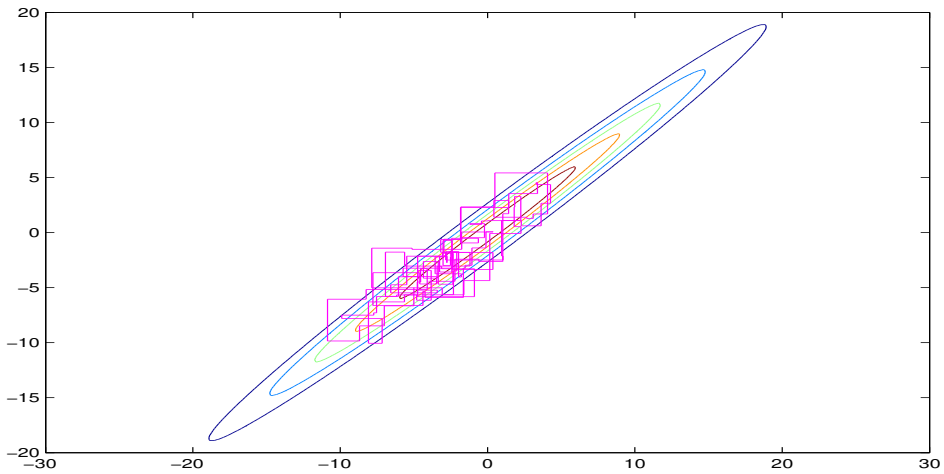
• The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, *i.e.* $\int_{\mathsf{X}} \pi(x) P(x, y) = \pi(y)$.

• The successive distributions of the Markov chains are "attracted" by π, or converge towards π.

• The estimator $\frac{1}{N}\sum_{n=1}^{N} \varphi(X_n)$ converges towards $E_\pi(\varphi(X))$ and asymptotically $X_n \sim \pi$

2.15– About the Gibbs sampler

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



2.14– 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

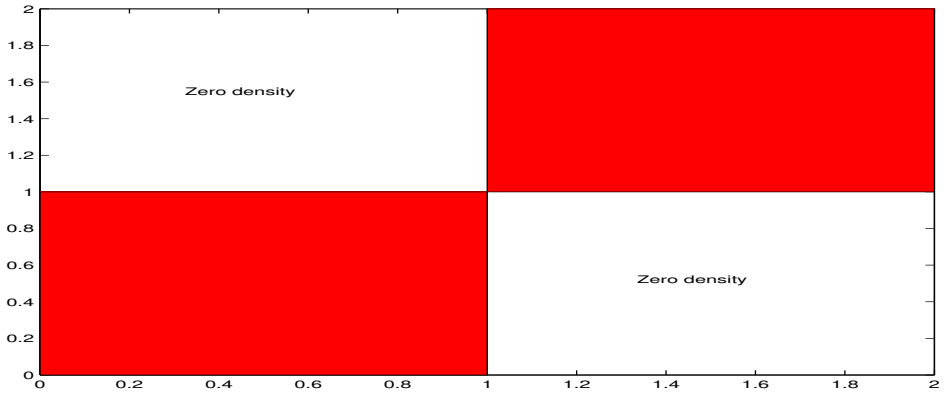
$$\frac{1}{N}\sum_{n=1}^{N}\varphi(X_n) \to \int \varphi(x)\,\pi(x)\,dx$$

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.

2.13– Irreducibility

A distribution that can lead to a reducible Gibbs sampler.



2.12– 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

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

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

2.19– Bayesian Variable Selection Example

• 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

$$\Pr(\gamma_i = 0) = \Pr(\gamma_i = 1) = \tfrac{1}{2},$$

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

2.18– Practical Recommendations

• 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.

2.17– 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 $\theta_i^{-K} = \left(\theta_i^1, ..., \theta_i^{K-1}, \theta_i^{K+1}, ..., \theta_i^p\right)$.

2.16– More about the 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 $\theta_i^{-k} = \left(\theta_i^1, ..., \theta_i^{k-1}, \theta_{i-1}^{k+1}, ..., \theta_{i-1}^p\right)$.

2.23– Bayesian Variable Selection Example

• $\pi(\gamma|D)$ 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 $\left(\beta_{\gamma,i}, \sigma_i^2\right) \sim \pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma_i\right)$.

2.22– A Better Gibbs Sampler

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

• However, we know that

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

where

$$\pi(\gamma|D) \propto \pi(D|\gamma)\,\pi(\gamma)$$

and

$$\begin{aligned}\pi(D|\gamma) &= \int \pi\left(D, \beta_\gamma, \sigma^2 \middle| \gamma\right) d\beta_\gamma d\sigma^2 \\ &\propto \Gamma\left(\frac{\nu_0+n}{2}+1\right)\delta^{-n_\gamma}\left|\Sigma_\gamma\right|^{1/2}\left(\frac{\gamma_0+\sum_{i=1}^n y_i^2-\mu_\gamma^{\mathrm{T}}\Sigma_\gamma^{-1}\mu_\gamma}{2}\right)^{-\left(\frac{\nu_0+n}{2}+1\right)}.\end{aligned}$$

2.21– Bayesian Variable Selection Example

• 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

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

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

2.20– A Bad Gibbs Sampler

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

• 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.

2.23– Bayesian Variable Selection Example

• 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.

2.23– Bayesian Variable Selection Example

• 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.