CPSC 535
Metropolis-Hastings: Applications

AD

March 2007

Metropolis-Hastings one-at-a time

- Initialization: Select deterministically or randomly $\theta = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.

Metropolis-Hastings one-at-a time

- Initialization: Select deterministically or randomly $\theta = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:

Metropolis-Hastings one-at-a time

- Initialization: Select deterministically or randomly $\theta = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:
    - For $k = 1 : p$

Metropolis-Hastings one-at-a time

- Initialization: Select deterministically or randomly $\theta = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $\theta_k^{(i)}$ using an MH step of proposal distribution $q_k\left(\left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right), \theta_k'\right)$ and target $\pi\left(\theta_k \middle| \theta_{-k}^{(i)}\right)$ where $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.

- In 1986, Challenger exploded; the explosion being the result of an O-ring failure. It was believed to be a result of a cold weather at the departure time: 31°F.

Logistic Regression

- In 1986, Challenger exploded; the explosion being the result of an O-ring failure. It was believed to be a result of a cold weather at the departure time: 31^{o}F.
- We have access to the data of 23 previous flights which give for flight i: Temperature at flight time x_i and $y_i = 1$ failure and zero otherwise (Robert & Casella, p. 15).

Logistic Regression

- In 1986, Challenger exploded; the explosion being the result of an O-ring failure. It was believed to be a result of a cold weather at the departure time: 31°F.
- We have access to the data of 23 previous flights which give for flight i: Temperature at flight time x_i and $y_i = 1$ failure and zero otherwise (Robert & Casella, p. 15).
- We want to have a model relating Y to x. Obviously this cannot be a linear model $Y = \alpha + x\beta$ as we want $Y \in \{0, 1\}$.

- We select a simple logistic regression model

$$\Pr(Y = 1|x) = 1 - \Pr(Y = 0|x) = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)}.$$

- We select a simple logistic regression model

$$\Pr(Y = 1|x) = 1 - \Pr(Y = 0|x) = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)}.$$

- Equivalently we have

$$\log\text{it} = \log\left(\frac{\Pr(Y = 1|x)}{\Pr(Y = 0|x)}\right) = \alpha + x\beta.$$

- We select a simple logistic regression model

$$\Pr(Y = 1|\, x) = 1 - \Pr(Y = 0|\, x) = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)}.$$

- Equivalently we have

$$\log \mathrm{it} = \log\left(\frac{\Pr(Y = 1|\, x)}{\Pr(Y = 0|\, x)}\right) = \alpha + x\beta.$$

- This ensures that the response is binary.

- We follow a Bayesian approach and select $\pi(\alpha, \beta) = \pi(\alpha | b)\, \pi(\beta) = b^{-1} \exp(\alpha) \exp\left(-b^{-1} \exp(\alpha)\right)$;i.e. exponential prior on $\exp(\alpha)$ and flat prior on β.

- We follow a Bayesian approach and select $\pi(\alpha,\beta) = \pi(\alpha|\, b)\, \pi(\beta) = b^{-1} \exp(\alpha) \exp\left(-b^{-1} \exp(\alpha)\right)$;i.e. exponential prior on $\exp(\alpha)$ and flat prior on β.
- b is selected as the data-dependent prior such that $\mathbb{E}(\alpha) = \widehat{\alpha}$ where $\widehat{\alpha}$ is the MLE of α (Robert & Casella).

- We follow a Bayesian approach and select $\pi(\alpha,\beta)=\pi(\alpha|\,b)\,\pi(\beta)=b^{-1}\exp(\alpha)\exp\left(-b^{-1}\exp(\alpha)\right)$;i.e. exponential prior on $\exp(\alpha)$ and flat prior on β.
- b is selected as the data-dependent prior such that $\mathbb{E}(\alpha)=\widehat{\alpha}$ where $\widehat{\alpha}$ is the MLE of α (Robert & Casella).
- As a simple proposal distribution, we use

$$q\left((\alpha,\beta),(\alpha',\beta')\right)=\pi\left(\alpha'\middle|\,b\right)\mathcal{N}\left(\beta';\beta,\widehat{\sigma}_{\beta}^{2}\right)$$

where $\widehat{\sigma}_{\beta}^{2}$ is the variance associated to the MLE $\widehat{\beta}$.

The algorithm proceeds as follows at iteration i

- Sample $(\alpha^*, \beta^*) \sim \pi(\alpha|\, b)\, \mathcal{N}\left(\beta; \beta^{(i-1)}, \widehat{\sigma}^2_\beta\right)$ and compute

$$\begin{aligned} &\zeta\left(\left(\alpha^{(i-1)}, \beta^{(i-1)}\right), (\alpha^*, \beta^*)\right) \\ = \quad &\min\left(1, \frac{\pi(\alpha^*, \beta^*|\, \text{data})\, \pi\left(\alpha^{(i-1)} \middle|\, b\right)}{\pi\left(\alpha^{(i-1)}, \beta^{(i-1)} \middle|\, \text{data}\right) \pi(\alpha^*|\, b)}\right) \end{aligned}$$

The algorithm proceeds as follows at iteration i

- Sample $(\alpha^*, \beta^*) \sim \pi(\alpha | b) \mathcal{N}\left(\beta; \beta^{(i-1)}, \widehat{\sigma}_\beta^2\right)$ and compute

$$\begin{aligned} & \zeta\left(\left(\alpha^{(i-1)}, \beta^{(i-1)}\right), (\alpha^*, \beta^*)\right) \\ = & \min\left(1, \frac{\pi(\alpha^*, \beta^* | \,\text{data})\, \pi\left(\alpha^{(i-1)} \middle| b\right)}{\pi\left(\alpha^{(i-1)}, \beta^{(i-1)} \middle| \,\text{data}\right) \pi(\alpha^* | b)}\right) \end{aligned}$$

- Set $\left(\alpha^{(i)}, \beta^{(i)}\right) = (\alpha^*, \beta^*)$ with probability $\zeta\left(\left(\alpha^{(i-1)}, \beta^{(i-1)}\right), (\alpha^*, \beta^*)\right)$, otherwise set $\left(\alpha^{(i)}, \beta^{(i)}\right) = \left(\alpha^{(i-1)}, \beta^{(i-1)}\right)$.

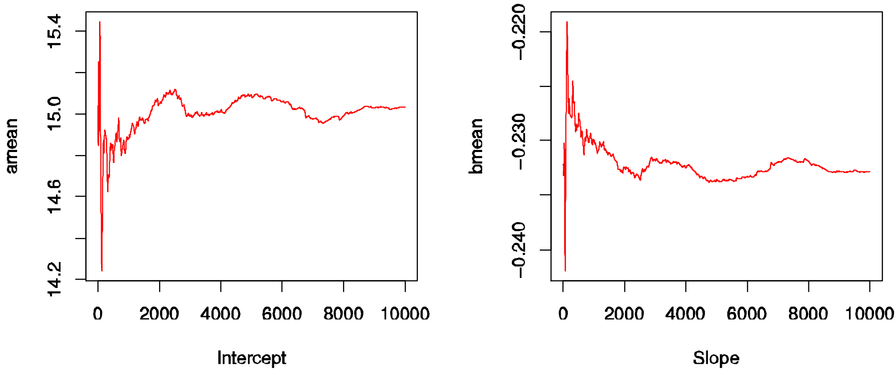

Figure: Plots of $\frac{1}{k}\sum_{i=1}^{k}\alpha^{(i)}$ (left) and $\frac{1}{k}\sum_{i=1}^{k}\beta^{(i)}$ (right).

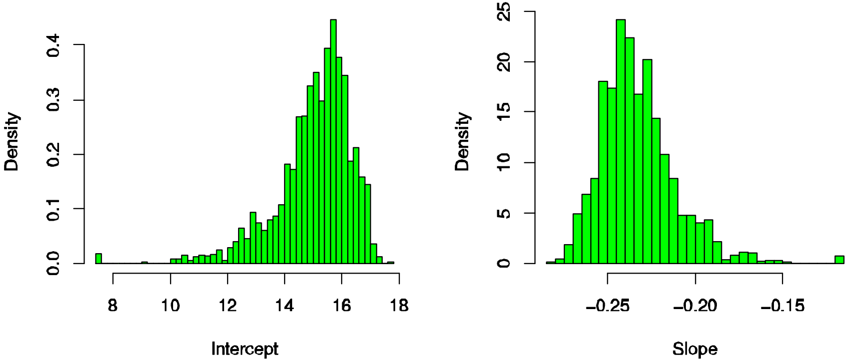

Figure: Histogram estimates of $p(\alpha | data)$ (left) and $p(\beta | data)$ (right).

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.
- The response variable y is 0 for genuine and 1 for counterfeit and the explanatory variables are

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.
- The response variable y is 0 for genuine and 1 for counterfeit and the explanatory variables are
    - x^1 the length,

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.
- The response variable y is 0 for genuine and 1 for counterfeit and the explanatory variables are
  - x^1 the length,
  - x^2: the width of the left edge

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.
- The response variable y is 0 for genuine and 1 for counterfeit and the explanatory variables are
  - x^1 the length,
  - x^2: the width of the left edge
  - x^3: the width of the right edge

Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.
- The response variable y is 0 for genuine and 1 for counterfeit and the explanatory variables are
  - x^1 the length,
  - x^2: the width of the left edge
  - x^3: the width of the right edge
  - x^4: the bottom margin witdth

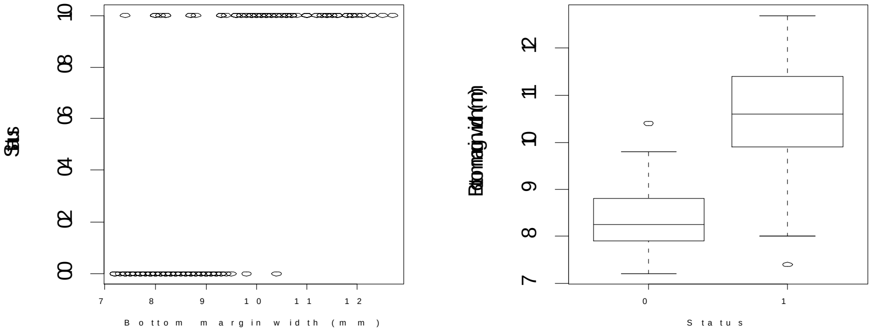

Figure: Left: Plot of the status indicator versus the bottom margin width. Right: Boxplots of the bottom margin width for both counterfeit status.

- Instead of selecting a logistic link, we select a probit one here

$$\Pr(Y = 1 | x) = \Phi\left(x^1\beta_1 + \ldots + x^4\beta_4\right)$$

where

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{v^2}{2}\right) dv$$

- Instead of selecting a logistic link, we select a probit one here

$$\Pr(Y = 1|x) = \Phi\left(x^1\beta_1 + \ldots + x^4\beta_4\right)$$

where

$$\Phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} \exp\left(-\frac{v^2}{2}\right) dv$$

- For n data, the likelihood is then given by

$$f\left(y_{1:n}|\beta, x_{1:n}\right) = \prod_{i=1}^{n} \Phi\left(x_i^T\beta\right)^{y_i}\left(1 - \Phi\left(x_i^T\beta\right)\right)^{1-y_i}.$$

- We assume a vague prior where $\beta \sim \mathcal{N}(0, 100 I_4)$ and we use a simple random walk sampler with $\widehat{\Sigma}$ the covariance matrix associated to the MLE (estimated using simple deterministic method).

- We assume a vague prior where $\beta \sim \mathcal{N}\left(0, 100 I_4\right)$ and we use a simple random walk sampler with $\widehat{\Sigma}$ the covariance matrix associated to the MLE (estimated using simple deterministic method).
- The algorithm is thus simply given at iteration i by

- We assume a vague prior where $\beta \sim \mathcal{N}\left(0, 100 I_4\right)$ and we use a simple random walk sampler with $\widehat{\Sigma}$ the covariance matrix associated to the MLE (estimated using simple deterministic method).
- The algorithm is thus simply given at iteration i by
  - Sample $\beta^* \sim \mathcal{N}\left(\beta^{(i-1)}, \tau^2 \widehat{\Sigma}\right)$ and compute

$$\alpha\left(\beta^{(i-1)}, \beta^*\right) = \min\left(1, \frac{\pi\left(\beta^* \middle| y_{1:n}, x_{1:n}\right)}{\pi\left(\beta^{(i-1)} \middle| y_{1:n}, x_{1:n}\right)}\right).$$

- We assume a vague prior where $\beta \sim \mathcal{N}\left(0, 100 I_4\right)$ and we use a simple random walk sampler with $\widehat{\Sigma}$ the covariance matrix associated to the MLE (estimated using simple deterministic method).
- The algorithm is thus simply given at iteration i by
  - Sample $\beta^* \sim \mathcal{N}\left(\beta^{(i-1)}, \tau^2 \widehat{\Sigma}\right)$ and compute

$$\alpha\left(\beta^{(i-1)}, \beta^*\right) = \min\left(1, \frac{\pi\left(\beta^* \middle| y_{1:n}, x_{1:n}\right)}{\pi\left(\beta^{(i-1)} \middle| y_{1:n}, x_{1:n}\right)}\right).$$

  - Set $\beta^{(i)} = \beta^*$ with probability $\alpha\left(\beta^{(i-1)}, \beta^*\right)$ and $\beta^{(i)} = \beta^{(i-1)}$ otherwise.

- We assume a vague prior where $\beta \sim \mathcal{N}(0, 100 I_4)$ and we use a simple random walk sampler with $\widehat{\Sigma}$ the covariance matrix associated to the MLE (estimated using simple deterministic method).
- The algorithm is thus simply given at iteration i by
  - Sample $\beta^* \sim \mathcal{N}\left(\beta^{(i-1)}, \tau^2 \widehat{\Sigma}\right)$ and compute

$$\alpha\left(\beta^{(i-1)}, \beta^*\right) = \min\left(1, \frac{\pi\left(\beta^* \middle| y_{1:n}, x_{1:n}\right)}{\pi\left(\beta^{(i-1)} \middle| y_{1:n}, x_{1:n}\right)}\right).$$

  - Set $\beta^{(i)} = \beta^*$ with probability $\alpha\left(\beta^{(i-1)}, \beta^*\right)$ and $\beta^{(i)} = \beta^{(i-1)}$ otherwise.
- Best results obtained with $\tau^2 = 1$.

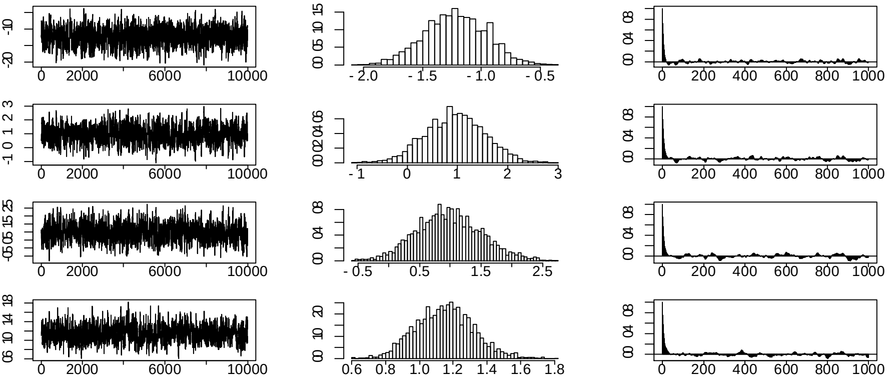

Figure: Traces (left), Histograms (middle) and Autocorrelations (right) for $\left(\beta_1^{(i)}, \ldots, \beta_4^{(i)}\right)$.

Autocorrelation

- One way to monitor the performance of the algorithm of the chain $\left\{X^{(i)}\right\}$ consists of displaying $\rho_k = cov\left[X^{(i)}, X^{(i+k)}\right] / var\left(X^{(i)}\right)$ which can be estimated from the chain, at least for small values of k.

Autocorrelation

- One way to monitor the performance of the algorithm of the chain $\left\{X^{(i)}\right\}$ consists of displaying $\rho_k = cov\left[X^{(i)}, X^{(i+k)}\right] / var\left(X^{(i)}\right)$ which can be estimated from the chain, at least for small values of k.
- Sometimes one uses an effective sample size measure

$$N^{\text{ess}} = N\left(1 + 2\sum_{k=1}^{N_0} \widehat{\rho}_k\right)^{-1/2}.$$

This represents approximately the sample size of an equivalent i.i.d. samples.

Autocorrelation

- One way to monitor the performance of the algorithm of the chain $\left\{X^{(i)}\right\}$ consists of displaying $\rho_k = cov\left[X^{(i)}, X^{(i+k)}\right] / var\left(X^{(i)}\right)$ which can be estimated from the chain, at least for small values of k.
- Sometimes one uses an effective sample size measure

$$N^{\text{ess}} = N\left(1 + 2\sum_{k=1}^{N_0} \widehat{\rho}_k\right)^{-1/2}.$$

  This represents approximately the sample size of an equivalent i.i.d. samples.
- One should be very careful with such measures which can be very misleading.

- We found for $\mathbb{E}\left(\beta \middle| y_{1:n}, x_{1:n}\right) = (-1.22, 0.95, 0.96, 1.15)$ so a simple plug-in estimate of the predictive probability of a counterfeit bill is

$$\widehat{p} = \Phi\left(-1.22x^1 + 0.95x^2 + 0.96x^3 + 1.15x^4\right)$$

For $x = (214.9, 130.1, 129.9, 9.5)$, we obtain $\widehat{p} = 0.59$.

- We found for $\mathbb{E}\left(\beta \mid y_{1:n}, x_{1:n}\right) = (-1.22, 0.95, 0.96, 1.15)$ so a simple plug-in estimate of the predictive probability of a counterfeit bill is

$$\widehat{p} = \Phi\left(-1.22x^1 + 0.95x^2 + 0.96x^3 + 1.15x^4\right)$$

  For $x = (214.9, 130.1, 129.9, 9.5)$, we obtain $\widehat{p} = 0.59$.
- A better estimate is obtained by

$$\int \Phi\left(\beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4\right) \pi\left(\beta \mid y_{1:n}, x_{1:n}\right) d\beta$$

Auxiliary Variables for Probit Regression

- It is impossible to use Gibbs to sample directly from $\pi\left(\beta | y_{1:n}, x_{1:n}\right)$.

Auxiliary Variables for Probit Regression

- It is impossible to use Gibbs to sample directly from $\pi\left(\beta \mid y_{1:n}, x_{1:n}\right)$.
- Introduce the following unobserved latent variables

$$
\begin{aligned}
Z_i &\sim \mathcal{N}\left(x_i^{\mathsf{T}}\beta, 1\right), \\
Y_i &= \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Auxiliary Variables for Probit Regression

- It is impossible to use Gibbs to sample directly from $\pi\left(\beta \mid y_{1:n}, x_{1:n}\right)$.
- Introduce the following unobserved latent variables

$$\begin{aligned} Z_i &\sim \mathcal{N}\left(x_i^{\mathsf{T}}\beta, 1\right), \\ Y_i &= \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- We have now define a joint distribution

$$f\left(y_i, z_i \mid \beta, x_i\right) = f\left(y_i \mid z_i\right) f\left(z_i \mid \beta, x_i\right).$$

- Now we can check that

$$\begin{aligned} f\left(y_i = 1 \middle| x_i, \beta\right) &= \int f\left(y_i, z_i \middle| \beta, x_i\right) dz_i \\ &= \int_0^\infty f\left(z_i \middle| \beta, x_i\right) dz_i = \Phi\left(x_i^{\mathrm{T}} \beta\right). \end{aligned}$$

$\Rightarrow$ We haven't changed the model!

- Now we can check that

$$
\begin{aligned}
f\left(y_i = 1 | x_i, \beta\right) &= \int f\left(y_i, z_i | \beta, x_i\right) dz_i \\
&= \int_0^{\infty} f\left(z_i | \beta, x_i\right) dz_i = \Phi\left(x_i^{\mathrm{T}} \beta\right).
\end{aligned}
$$

$\Rightarrow$ We haven't changed the model!

- We are now going to sample from $\pi\left(\beta, z_{1:n} | x_{1:n}, y_{1:n}\right)$ instead of $\pi\left(\beta | x_{1:n}, y_{1:n}\right)$ because the full conditional distributions are simple

$$
\begin{aligned}
\pi\left(\beta | y_{1:n}, x_{1:n}, z_{1:n}\right) &= \pi\left(\beta | x_{1:n}, z_{1:n}\right) \text{ (standard Gaussian!)}, \\
\pi\left(z_{1:n} | y_{1:n}, x_{1:n}, \beta\right) &= \prod_{i=1}^{n} \pi\left(z_k | y_k, x_k, \beta\right)
\end{aligned}
$$

where

$$
z_k | y_k, x_k, \beta \sim \begin{cases} \mathcal{N}_+\left(x_k^{\mathrm{T}} \beta, 1\right) & \text{if } y_k = 1 \\ \mathcal{N}_-\left(x_k^{\mathrm{T}} \beta, 1\right) & \text{if } y_k = 0. \end{cases}
$$

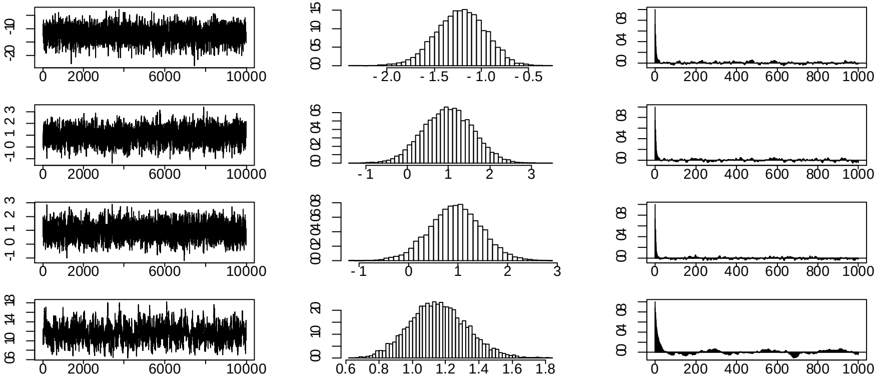

Figure: Traces (left), Histograms (middle) and Autocorrelations (right) for $\left(\beta_1^{(i)}, .., \beta_4^{(i)}\right)$.

- The results obtained through Gibbs are very similar to MH.

- The results obtained through Gibbs are very similar to MH.
- We can also adopt an Zellner's type prior and obtain very similar results.

- The results obtained through Gibbs are very similar to MH.
- We can also adopt an Zellner's type prior and obtain very similar results.
- Very similar were also obtained using a logistic fonction using the MH (Gibbs is feasible but more difficult).

Warning

- Although the introduction of latent variables can be attractive, it can be also very inefficient.

Warning

- Although the introduction of latent variables can be attractive, it can be also very inefficient.
- It is not because you can use the Gibbs sampler that everything works well!

Warning

- Although the introduction of latent variables can be attractive, it can be also very inefficient.
- It is not because you can use the Gibbs sampler that everything works well!
- Consider the following simple generalization of the previous model

$$Z_i \sim \mathcal{N}\left(x_i\beta, \sigma^2\right), \quad Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Warning

- Although the introduction of latent variables can be attractive, it can be also very inefficient.
- It is not because you can use the Gibbs sampler that everything works well!
- Consider the following simple generalization of the previous model

$$Z_i \sim \mathcal{N}\left(x_i\beta, \sigma^2\right), \ \ Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- We complete the model by $\sigma^2 \sim \mathcal{IG}\left(1.5, 1.5\right)$ and $\beta|\,\sigma^2 \sim \mathcal{N}\left(0, 100\right)$.

- Not only the data Z_i and (β, σ^2) are very correlated but we have

$$\Pr\left(Y_i = 1 \middle| x_i, \beta, \sigma^2 \right) = \Phi\left(\frac{x_i \beta}{\sigma} \right)$$

- Not only the data Z_i and $\left(\beta, \sigma^2\right)$ are very correlated but we have

$$\Pr\left(Y_i = 1 \middle| x_i, \beta, \sigma^2\right) = \Phi\left(\frac{x_i \beta}{\sigma}\right)$$

- The likelihood only depends on β/σ and the parameters β and σ are not identifiable.

- Not only the data Z_i and (β, σ^2) are very correlated but we have

$$\Pr\left(Y_i = 1 \middle| x_i, \beta, \sigma^2 \right) = \Phi\left(\frac{x_i \beta}{\sigma} \right)$$

- The likelihood only depends on β/σ and the parameters β and σ are not identifiable.
- One way to improve the mixing consists of using an additional MH step that proposes to randomly rescale the current value.

- Not only the data Z_i and $\left(\beta, \sigma^2\right)$ are very correlated but we have

$$\Pr\left(Y_i = 1 | x_i, \beta, \sigma^2\right) = \Phi\left(\frac{x_i \beta}{\sigma}\right)$$

- The likelihood only depends on β/σ and the parameters β and σ are not identifiable.
- One way to improve the mixing consists of using an additional MH step that proposes to randomly rescale the current value.
- We use a proposal distribution such that

$$\left(\beta', \sigma'\right) = \lambda\left(\beta, \sigma\right) \text{ with } \lambda \sim \mathcal{E}\mathrm{xp}\left(1\right)$$

that proposes to randomly rescale the current value.

Hidden Markov Model

- Consider the following hidden Markov model

$$\begin{aligned} X_k | (X_{k-1} = x_{k-1}) &\sim f_\theta (\cdot | x_{k-1}), \ X_1 \sim \mu \\ Y_n | (X_k = x_k) &\sim g_\theta (\cdot | x_k), \end{aligned}$$

and we set a prior $\pi(\theta)$ on the unknown hyperparameters θ.

Hidden Markov Model

- Consider the following hidden Markov model

$$\begin{aligned} X_k | (X_{k-1} = x_{k-1}) &\sim f_\theta(\cdot | x_{k-1}), \; X_1 \sim \mu \\ Y_n | (X_k = x_k) &\sim g_\theta(\cdot | x_k), \end{aligned}$$

  and we set a prior $\pi(\theta)$ on the unknown hyperparameters θ.
- Given n data, we are interested in the joint posterior

$$\pi(\theta, x_{1:n} | y_{1:n})$$

Hidden Markov Model

- Consider the following hidden Markov model

$$\begin{aligned} X_k \mid (X_{k-1} = x_{k-1}) &\sim f_\theta(\cdot \mid x_{k-1}), \; X_1 \sim \mu \\ Y_n \mid (X_k = x_k) &\sim g_\theta(\cdot \mid x_k), \end{aligned}$$

  and we set a prior $\pi(\theta)$ on the unknown hyperparameters θ.
- Given n data, we are interested in the joint posterior

$$\pi(\theta, x_{1:n} \mid y_{1:n})$$

- There is no closed-form expression for this joint distribution even if the model is linear Gaussian or for finite state-space model.

- In previous lectures, we propose sampling from $\pi\left(\theta, x_{1:n} \middle| y_{1:n}\right)$ using the Gibbs sampler where the variables are updated according to

$$X_k \sim \pi\left(x_k \middle| y_{1:n}, x_{-k}, \theta\right)$$

- In previous lectures, we propose sampling from $\pi\left(\theta, x_{1:n} \mid y_{1:n}\right)$ using the Gibbs sampler where the variables are updated according to

$$X_k \sim \pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right)$$

- For $2 < k < n$, we have

$$\begin{aligned}
\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right) &\propto \pi\left(x_{1:n}, y_{1:n}, \theta\right) \\
&\propto \underbrace{\pi(\theta)\, \mu(x_1) \prod_{i=2}^{n} f_\theta\left(x_i \mid x_{i-1}\right)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} g_\theta\left(y_i \mid x_i\right)}_{\text{likelihood}} \\
&\propto f_\theta\left(x_k \mid x_{k-1}\right) f_\theta\left(x_{k+1} \mid x_k\right) g_\theta\left(y_k \mid x_k\right)
\end{aligned}$$

and $\theta \sim \pi\left(\theta \mid y_{1:n}, x_{1:n}\right)$ (or by subblocks).

- It is often possible to implement the Gibbs sampler even if this can be expensive; e.g. if you use Accept/Reject to sample from $\pi\left(x_k \middle| y_{1:n}, x_{-k}, \theta\right)$ using the proposal $\pi\left(x_k \middle| x_{-k}, \theta\right) \propto f_\theta\left(x_k \middle| x_{k-1}\right) f_\theta\left(x_{k+1} \middle| x_k\right)$.

- It is often possible to implement the Gibbs sampler even if this can be expensive; e.g. if you use Accept/Reject to sample from $\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right)$ using the proposal $\pi\left(x_k \mid x_{-k}, \theta\right) \propto f_\theta\left(x_k \mid x_{k-1}\right) f_\theta\left(x_{k+1} \mid x_k\right)$.
- Even if it is possible to implement the Gibbs sampler, one can expect a very slow convergence of the algorithm is successive variables are highly correlated.

- It is often possible to implement the Gibbs sampler even if this can be expensive; e.g. if you use Accept/Reject to sample from $\pi\left(x_k \middle| y_{1:n}, x_{-k}, \theta\right)$ using the proposal $\pi\left(x_k \middle| x_{-k}, \theta\right) \propto f_\theta\left(x_k \middle| x_{k-1}\right) f_\theta\left(x_{k+1} \middle| x_k\right)$.
- Even if it is possible to implement the Gibbs sampler, one can expect a very slow convergence of the algorithm is successive variables are highly correlated.
- Indeed, as you update x_k with x_{k-1} and x_{k+1} being fixed, then you cannot move much into the space.

- Consider the very simple case where $\theta = \left(\sigma_v^2, \sigma_w^2\right)$

$$\begin{aligned} X_k &= X_{k-1} + V_k \text{ where } V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_v^2\right), \\ Y_k &= X_k + W_k \text{ where } W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_w^2\right) \end{aligned}$$

then we have

$$\begin{aligned} \pi\left(x_k \mid x_{-k}, \theta\right) &\propto f_\theta\left(x_k \mid x_{k-1}\right) f_\theta\left(x_{k+1} \mid x_k\right) \\ &= \mathcal{N}\left(x_k; \frac{x_{k-1} + x_{k+1}}{2}, \frac{\sigma_v^2}{2}\right) \end{aligned}$$

and

$$\begin{aligned} &\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right) \\ \propto\ &\pi\left(x_k \mid x_{-k}, \theta\right) g_\theta\left(y_k \mid x_k\right) \\ =\ &\mathcal{N}\left(x_k; \frac{\sigma_v^2 \sigma_w^2}{\sigma_v^2 + 2\sigma_w^2}\left(\frac{x_{k-1} + x_{k+1}}{\sigma_v^2} + \frac{y_k}{\sigma_w^2}\right), \frac{\sigma_v^2 \sigma_w^2}{\sigma_v^2 + 2\sigma_w^2}\right) \end{aligned}$$

- Assume for the time being that instead of sampling from $\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right)$ directly, we use rejection sampling with $\pi\left(x_k \mid x_{-k}, \theta\right)$ as a proposal distribution.

- Assume for the time being that instead of sampling from $\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right)$ directly, we use rejection sampling with $\pi\left(x_k \mid x_{-k}, \theta\right)$ as a proposal distribution.
- In this case we have to bound

$$g_\theta\left(y_k \mid x_k\right) = \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{\left(y_k - x_k\right)^2}{2\sigma_w^2}\right) \leq \frac{1}{\sqrt{2\pi}\sigma_w}.$$

- Assume for the time being that instead of sampling from $\pi\left(x_k \mid y_{1:n}, x_{-k}, \theta\right)$ directly, we use rejection sampling with $\pi\left(x_k \mid x_{-k}, \theta\right)$ as a proposal distribution.
- In this case we have to bound

$$g_\theta\left(y_k \mid x_k\right) = \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{\left(y_k - x_k\right)^2}{2\sigma_w^2}\right) \leq \frac{1}{\sqrt{2\pi}\sigma_w}.$$

- We accept each proposal $X^* \sim \pi\left(x_k \mid x_{-k}, \theta\right)$ with probability $\exp\left(-\frac{\left(y_k - X^*\right)^2}{2\sigma_w^2}\right)$, so the (unconditional) acceptance probability is given by

$$\begin{aligned} & \int \pi\left(x_k \mid x_{-k}, \theta\right) \exp\left(-\frac{\left(y_k - x_k\right)^2}{2\sigma_w^2}\right) dx_k \\ = & \frac{\sigma_w \exp\left(-\frac{1}{2}\left(y_k^2 / \sigma_w^2 - \left(x_{k-1} + x_{k+1}\right)^2 / \sigma_v^2\right)\right)}{\sqrt{\sigma_v^2 + 2\sigma_w^2}}. \end{aligned}$$

- To improve the algorithm, we would like to be able to sample a whole block of variables simultaneously; i.e. being able to sample for $1 < k < k + L < n$ from

$$\begin{aligned} \pi\left(x_{k:k+L} \middle| y_{1:n}, x_{-(k:k+L)}, \theta \right) &= \pi\left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+L+1}, \theta \right) \\ &\propto \prod_{i=k}^{k+L+1} f_\theta\left(x_i \middle| x_{i-1} \right) \prod_{i=k}^{k+L} g_\theta\left(y_i \middle| x_i \right). \end{aligned}$$

- To improve the algorithm, we would like to be able to sample a whole block of variables simultaneously; i.e. being able to sample for $1 < k < k + L < n$ from

$$\begin{aligned}
\pi\left(x_{k:k+L} \middle| y_{1:n}, x_{-(k:k+L)}, \theta \right) &= \pi\left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+L+1}, \theta \right) \\
&\propto \prod_{i=k}^{k+L+1} f_\theta\left(x_i \middle| x_{i-1} \right) \prod_{i=k}^{k+L} g_\theta\left(y_i \middle| x_i \right).
\end{aligned}$$

- In this case, it is typically impossible to sample from $\pi\left(x_{k:k+L} \middle| y_{1:n}, x_{-(k:k+L)}, \theta \right)$ exactly as L is large, say 5 or 10.

- To improve the algorithm, we would like to be able to sample a whole block of variables simultaneously; i.e. being able to sample for $1 < k < k+L < n$ from

$$\begin{aligned}\pi\left(x_{k:k+L}\middle|\, y_{1:n}, x_{-(k:k+L)}, \theta\right) &= \pi\left(x_{k:k+L}\middle|\, y_{k:k+L}, x_{k-1}, x_{k+L+1}, \theta\right) \\ &\propto \prod_{i=k}^{k+L+1} f_\theta\left(x_i\middle|\, x_{i-1}\right) \prod_{i=k}^{k+L} g_\theta\left(y_i\middle|\, x_i\right).\end{aligned}$$

- In this case, it is typically impossible to sample from $\pi\left(x_{k:k+L}\middle|\, y_{1:n}, x_{-(k:k+L)}, \theta\right)$ exactly as L is large, say 5 or 10.
- We are propose to use a MH step of invariant distribution $\pi\left(x_{k:k+L}\middle|\, y_{1:n}, x_{-(k:k+L)}, \theta\right)$ instead, hence we need to build a proposal distribution $q\left(\left(x_{1:n}, \theta\right), x'_{k:k+L}\right)$.

- We first propose to use the conditional prior

$$\begin{aligned} q\left(\left(x_{1:n},\theta\right),x'_{k:k+L}\right) &= \pi\left(x_{k:k+L}\middle|\,x_{-(k:k+L)},\theta\right) \\ &= \pi\left(x_{k:k+L}\middle|\,x_{k-1},x_{k+L+1},\theta\right) \\ &\propto \prod_{i=k}^{k+L+1} f_{\theta}\left(x_i\middle|\,x_{i-1}\right). \end{aligned}$$

- We first propose to use the conditional prior

$$
\begin{aligned}
q\left(\left(x_{1:n},\theta\right),x'_{k:k+L}\right) &= \pi\left(x_{k:k+L}\middle|\, x_{-(k:k+L)},\theta\right) \\
&= \pi\left(x_{k:k+L}\middle|\, x_{k-1},x_{k+L+1},\theta\right) \\
&\propto \prod_{i=k}^{k+L+1} f_\theta\left(x_i\middle|\, x_{i-1}\right).
\end{aligned}
$$

- In this case, the candidate $X'_{k:k+L} \sim \pi\left(x_{k:k+L}\middle|\, x_{k-1},x_{k+L+1},\theta\right)$ is accepted with probability

$$
\min\left(1,\frac{\pi\left(x'_{k:k+L}\middle|y_{k:k+L},x_{k-1},x_{k+L+1},\theta\right)\pi\left(x_{k:k+L}\middle|x_{k-1},x_{k+L+1},\theta\right)}{\pi\left(x_{k:k+L}\middle|y_{k:k+L},x_{k-1},x_{k+L+1},\theta\right)\pi\left(x'_{k:k+L}\middle|x_{k-1},x_{k+L+1},\theta\right)}\right)
$$
$$
=\min\left(1,\frac{\prod_{i=k}^{k+L} g_\theta\left(y_i|x'_i\right)}{\prod_{i=k}^{k+L} g_\theta\left(y_i|x_i\right)}\right)
$$

- We first propose to use the conditional prior

$$
\begin{aligned}
q\left(\left(x_{1:n},\theta\right),x'_{k:k+L}\right) &= \pi\left(x_{k:k+L}\middle| x_{-(k:k+L)},\theta\right) \\
&= \pi\left(x_{k:k+L}\middle| x_{k-1},x_{k+L+1},\theta\right) \\
&\propto \prod_{i=k}^{k+L+1} f_\theta\left(x_i\middle| x_{i-1}\right).
\end{aligned}
$$

- In this case, the candidate $X'_{k:k+L} \sim \pi\left(x_{k:k+L}\middle| x_{k-1},x_{k+L+1},\theta\right)$ is accepted with probability

$$
\begin{aligned}
&\min\left(1,\frac{\pi\left(x'_{k:k+L}\middle|y_{k:k+L},x_{k-1},x_{k+L+1},\theta\right)\pi\left(x_{k:k+L}\middle|x_{k-1},x_{k+L+1},\theta\right)}{\pi\left(x_{k:k+L}\middle|y_{k:k+L},x_{k-1},x_{k+L+1},\theta\right)\pi\left(x'_{k:k+L}\middle|x_{k-1},x_{k+L+1},\theta\right)}\right)\\
&=\min\left(1,\frac{\prod_{i=k}^{k+L} g_\theta\left(y_i|x'_i\right)}{\prod_{i=k}^{k+L} g_\theta\left(y_i|x_i\right)}\right)
\end{aligned}
$$

- Simple but one cannot expect it to be too efficient when the observations are very informative compared to the prior.

- Consider the case where

$$X_k = AX_{k-1} + BV_k,\ V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I).$$

- Consider the case where

$$X_k = AX_{k-1} + BV_k,\ V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I).$$

- Particular cases include

$$\begin{aligned} X_k &= X_{k-1} + \sigma V_k, \text{ where } V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ X_k &= \begin{pmatrix} \alpha_k \\ \alpha_{k-1} \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} X_{k-1} + \begin{pmatrix} \sigma \\ 0 \end{pmatrix} V_k,\ V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \end{aligned}$$

- In this case, it is simple to see that $\pi\left(x_{k:k+L}\middle|\, x_{k-1}, x_{k+1}, \theta\right)$ is a Gaussian distribution.

- In this case, it is simple to see that $\pi\left(x_{k:k+L}\middle| x_{k-1}, x_{k+1}, \theta\right)$ is a Gaussian distribution.
- In (Knorr-Held, 1999), one samples from this distribution by computing directly the parameters of this joint distribution: complexity $O\left(L^2\right)$.

- In this case, it is simple to see that $\pi\left(x_{k:k+L} \middle| x_{k-1}, x_{k+1}, \theta\right)$ is a Gaussian distribution.
- In (Knorr-Held, 1999), one samples from this distribution by computing directly the parameters of this joint distribution: complexity $O\left(L^2\right)$.
- We can derive a simpler method of complexity $O\left(L\right)$ based on the following decomposition (omitting θ in the notation)

$$\begin{aligned}
\pi\left(x_{k:k+L} \middle| x_{k-1}, x_{k+L+1}\right) &= \prod_{i=k}^{k+L} \pi\left(x_i \middle| x_{k-1}, x_{k+L+1}, x_{i+1}\right). \\
&= \prod_{i=k}^{k+L} \pi\left(x_i \middle| x_{k-1}, x_{i+1}\right)
\end{aligned}$$

- Moreover it is easy to establish the expression for $\pi\left(x_i \mid x_{k-1}, x_{i+1}\right)$

$$\pi\left(x_i \mid x_{k-1}, x_{i+1}\right) \propto \pi\left(x_i \mid x_{k-1}\right) f\left(x_{i+1} \mid x_i\right)$$

as

$$\pi\left(x_i \mid x_{k-1}\right) = \int \pi\left(x_{k:i} \mid x_{k-1}\right) dx_{k:i-1} = \mathcal{N}\left(x_i; \mu_i\left(x_{k-1}\right), \Sigma_i\right)$$

with, for $X_n = AX_{n-1} + BV_n$, $\mu_{k-1}\left(x_{k-1}\right) = x_{k-1}$, $\Sigma_{k-1} = 0$ and for $i \geq k$

$$\begin{aligned} \mu_i\left(x_{k-1}\right) &= A\mu_{i-1}\left(x_{k-1}\right), \\ \Sigma_i &= A\Sigma_{i-1}A^{\mathsf{T}} + \Sigma \text{ with } \Sigma = BB^{\mathsf{T}}. \end{aligned}$$

- Moreover it is easy to establish the expression for $\pi\left(x_i \mid x_{k-1}, x_{i+1}\right)$

$$\pi\left(x_i \mid x_{k-1}, x_{i+1}\right) \propto \pi\left(x_i \mid x_{k-1}\right) f\left(x_{i+1} \mid x_i\right)$$

as

$$\pi\left(x_i \mid x_{k-1}\right) = \int \pi\left(x_{k:i} \mid x_{k-1}\right) dx_{k:i-1} = \mathcal{N}\left(x_i; \mu_i\left(x_{k-1}\right), \Sigma_i\right)$$

with, for $X_n = AX_{n-1} + BV_n$, $\mu_{k-1}\left(x_{k-1}\right) = x_{k-1}$, $\Sigma_{k-1} = 0$ and for $i \geq k$

$$\begin{aligned} \mu_i\left(x_{k-1}\right) &= A\mu_{i-1}\left(x_{k-1}\right), \\ \Sigma_i &= A\Sigma_{i-1}A^{\mathsf{T}} + \Sigma \text{ with } \Sigma = BB^{\mathsf{T}}. \end{aligned}$$

- To obtain $\pi\left(x_i \mid x_{k-1}, x_{i+1}\right)$, we combine the prior $\pi\left(x_i \mid x_{k-1}\right)$ with the "likelihood" $f\left(x_{i+1} \mid x_i\right)$.

- We have $\pi\left(x_i | x_{k-1}\right) = \mathcal{N}\left(x_i; \mu_i\left(x_{k-1}\right), \Sigma_i\right)$ and $f\left(x_{i+1} | x_i\right) = \mathcal{N}\left(x_{i+1}; Ax_i, \Sigma\right)$ then

$$\pi\left(x_i | x_{k-1}, x_{i+1}\right) = \mathcal{N}\left(x_i; \mu_i\left(x_{k-1}, x_{i+1}\right), \widetilde{\Sigma}_i\right)$$

where

$$\begin{aligned} \widetilde{\Sigma}_i &= \left(\Sigma_i^{-1} + A^{\mathsf{T}}\Sigma^{-1}A\right)^{-1}, \\ \mu_i\left(x_{k-1}, x_{i+1}\right) &= \widetilde{\Sigma}_i\left(A^{\mathsf{T}}\Sigma^{-1}x_{i+1} + \Sigma_i^{-1}\mu_i\left(x_{k-1}\right)\right). \end{aligned}$$

- We have $\pi\left(\left.x_i\right| x_{k-1}\right)=\mathcal{N}\left(x_i;\mu_i\left(x_{k-1}\right),\Sigma_i\right)$ and $f\left(\left.x_{i+1}\right| x_i\right)=\mathcal{N}\left(x_{i+1};Ax_i,\Sigma\right)$ then

$$\pi\left(\left.x_i\right| x_{k-1},x_{i+1}\right)=\mathcal{N}\left(x_i;\mu_i\left(x_{k-1},x_{i+1}\right),\widetilde{\Sigma}_i\right)$$

  where

$$\begin{aligned}\widetilde{\Sigma}_i &= \left(\Sigma_i^{-1}+A^{\mathsf{T}}\Sigma^{-1}A\right)^{-1},\\ \mu_i\left(x_{k-1},x_{i+1}\right) &= \widetilde{\Sigma}_i\left(A^{\mathsf{T}}\Sigma^{-1}x_{i+1}+\Sigma_i^{-1}\mu_i\left(x_{k-1}\right)\right).\end{aligned}$$

- To sample a realization of $\pi\left(\left.x_{k:k+L}\right| x_{k-1},x_{k+L+1}\right)$, first compute $\mu_i\left(x_{k-1}\right),\Sigma_i$ for $i=k,...,k+L$ using a forward recursion. Then sample backward $X_{k+L}\sim\pi\left(\left.\cdot\right| x_{k-1},x_{k+L+1}\right)$, $X_{k+L-1}\sim\pi\left(\left.\cdot\right| x_{k-1},X_{k+L}\right),...,X_k\sim\pi\left(\left.\cdot\right| x_{k-1},X_{k+1}\right)$.

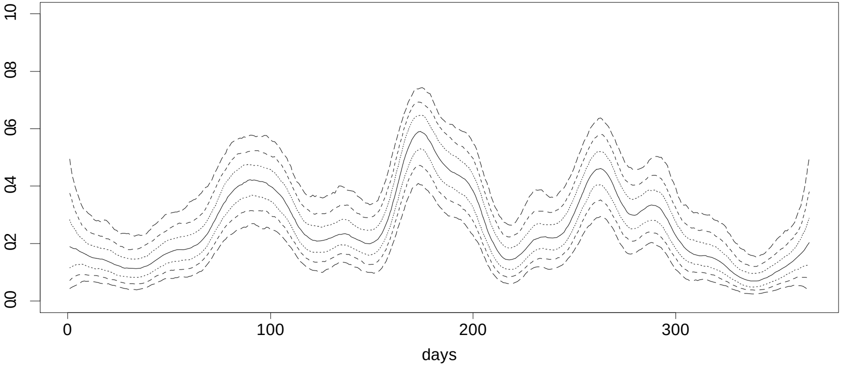

Figure: Number of occurences of rainfall in Tokyo for each day during 1983-1984 reproduced as relative frequencies between 0, 0.5 and 1 $(n = 366)$

- Consider the following model

$$X_k = \begin{pmatrix} \alpha_k \\ \alpha_{k-1} \end{pmatrix} = \begin{pmatrix} 2 & \text{-1} \\ 1 & 0 \end{pmatrix} X_{k-1} + \begin{pmatrix} \sigma \\ 0 \end{pmatrix} V_k, \; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

and

$$Y_k | X_k \sim \begin{cases} B(2, \pi_k) & k \neq 60, \\ B(1, \pi_k) & k = 60 \text{ (February 29)} \end{cases},$$

where

$$\pi_k = \frac{\exp(\alpha_k)}{1 + \exp(\alpha_k)}.$$

- Consider the following model

$$X_k = \left(\begin{array}{c} \alpha_k \\ \alpha_{k-1} \end{array} \right) = \left(\begin{array}{cc} 2 & -1 \\ 1 & 0 \end{array} \right) X_{k-1} + \left(\begin{array}{c} \sigma \\ 0 \end{array} \right) V_k, \; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

and

$$Y_k | X_k \sim \begin{cases} \mathcal{B}(2, \pi_k) & k \neq 60, \\ \mathcal{B}(1, \pi_k) & k = 60 \text{ (February 29)} \end{cases},$$

where

$$\pi_k = \frac{\exp(\alpha_k)}{1 + \exp(\alpha_k)}.$$

- We also use for $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$.

- We use the block sampling strategies discussed before where candidates are sampled according to $\pi\left(x_{k:k+L}\middle|\, x_{k-1}, x_{k+L+1}\right)$ and accepted with proba

$$\min\left(1, \frac{\prod_{i=k}^{k+L} g\left(y_i \middle|\, x_i'\right)}{\prod_{i=k}^{k+L} g\left(y_i \middle|\, x_i\right)}\right).$$

- We use the block sampling strategies discussed before where candidates are sampled according to $\pi\left(x_{k:k+L}\middle| x_{k-1}, x_{k+L+1}\right)$ and accepted with proba

$$\min\left(1, \frac{\prod_{i=k}^{k+L} g\left(y_i\middle| x_i'\right)}{\prod_{i=k}^{k+L} g\left(y_i\middle| x_i\right)}\right).$$

- The parameter σ^2 is updated through a simple Gibbs step

$$\begin{aligned}\sigma^2 &\sim \pi\left(\sigma^2\middle| x_{1:n}, y_{1:n}\right) = \pi\left(\sigma^2\middle| x_{1:n}\right)\\ &= \mathcal{IG}\left(\frac{\nu_0 + n - 1}{2}, \frac{\gamma_0 + \sum_{k=2}^{n}\left(\alpha_k - 2\alpha_{k-1} + \alpha_{k-2}\right)^2}{2}\right)\end{aligned}$$

- We use the block sampling strategies discussed before where candidates are sampled according to $\pi\left(x_{k:k+L}\middle|\,x_{k-1},x_{k+L+1}\right)$ and accepted with proba

$$\min\left(1,\frac{\prod_{i=k}^{k+L}g\left(y_i\middle|\,x_i'\right)}{\prod_{i=k}^{k+L}g\left(y_i\middle|\,x_i\right)}\right).$$

- The parameter σ^2 is updated through a simple Gibbs step

$$\begin{aligned}\sigma^2 &\sim \pi\left(\sigma^2\middle|\,x_{1:n},y_{1:n}\right)=\pi\left(\sigma^2\middle|\,x_{1:n}\right)\\ &= \mathcal{IG}\left(\frac{\nu_0+n-1}{2},\frac{\gamma_0+\sum_{k=2}^{n}\left(\alpha_k-2\alpha_{k-1}+\alpha_{k-2}\right)^2}{2}\right)\end{aligned}$$

- For block size $L=1,5,20$ and 40, we compute the average trajectories of 100 parallel chains after 10, 50, 100 and 500 iterations with initialization $x_k=0$ for all $k,\sigma^2=0.1$.

Figure: Average trajectories over 100 chains for $L = 1, 5, 20$ and 40 from top to bottom.

Figure: Average trajectories over 100 chains for $L = 1, 5, 20$ and 40 from top to bottom.

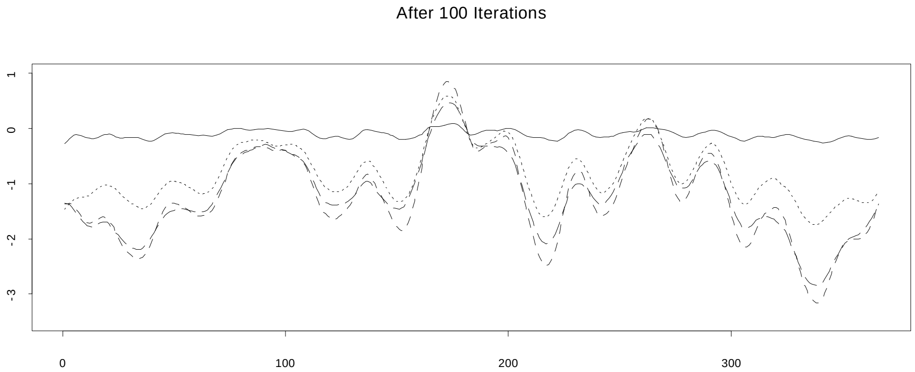

Figure: Average trajectories over 100 chains for $L = 1, 5, 20$ and 40 from top to bottom.

Figure: Average trajectories over 100 chains for $L = 1, 5, 20$ and 40 from top to bottom.

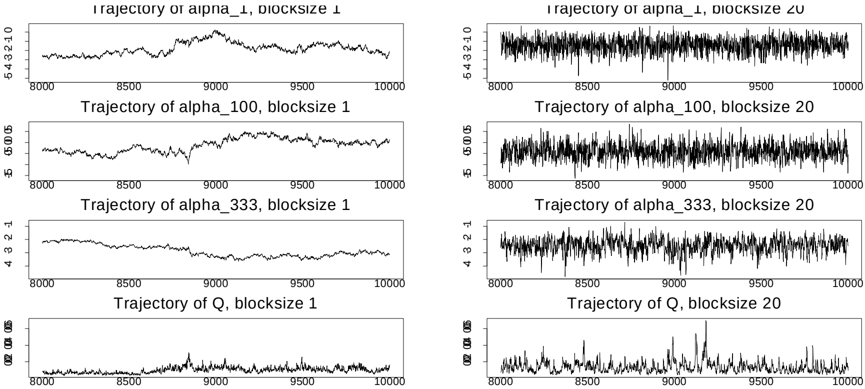

Figure: Traces of $\alpha_1, \alpha_{100}, \alpha_{333}$ and σ^2 for $L = 1$ (left) and $L = 20$ (right).

- This (naive!) block sampling strategy performs well here because the likelihood of the observations is fairly flat.

- This (naive!) block sampling strategy performs well here because the likelihood of the observations is fairly flat.
- For a linear Gaussian observation equation, Knorr-Held compares this strategy to a direct Gibbs sampling implementation. As expected, the conditional proposal strategy is competitive when the observations are not very informative compared to the prior.

- This (naive!) block sampling strategy performs well here because the likelihood of the observations is fairly flat.
- For a linear Gaussian observation equation, Knorr-Held compares this strategy to a direct Gibbs sampling implementation. As expected, the conditional proposal strategy is competitive when the observations are not very informative compared to the prior.
- For more complex problems, such strategies are inefficient and we will need to use the observations to build the proposal.

- (Pitt & Shephard, 1999) propose a more efficient strategy... also more computationally intensive.

- (Pitt & Shephard, 1999) propose a more efficient strategy... also more computationally intensive.
- Consider the log full conditional distribution

$$\begin{aligned}&\log \pi \left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+L+1} \right) \\ &= \sum_{i=k}^{k+L} \log g \left(y_i \middle| x_i \right) + \sum_{i=k}^{k+L+1} \log f \left(x_{i+1} \middle| x_i \right) \\ &\equiv \sum_{i=k}^{k+L} \log g \left(y_i \middle| x_i \right) - \frac{1}{2} \sum_{i=k}^{k+L+1} \left(x_{i+1} - A x_i \right)^{\mathsf{T}} \Sigma^{-1} \left(x_{i+1} - A x_i \right)\end{aligned}$$

which is not quadratic in x_i hence $\pi \left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+1} \right)$ is not Gaussian.

- (Pitt & Shephard, 1999) propose a more efficient strategy... also more computationally intensive.
- Consider the log full conditional distribution

$$
\begin{aligned}
&\log \pi \left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+L+1} \right) \\
&= \sum_{i=k}^{k+L} \log g\left(y_i \middle| x_i \right) + \sum_{i=k}^{k+L+1} \log f\left(x_{i+1} \middle| x_i \right) \\
&\equiv \sum_{i=k}^{k+L} \log g\left(y_i \middle| x_i \right) - \frac{1}{2} \sum_{i=k}^{k+L+1} \left(x_{i+1} - Ax_i \right)^{\mathsf{T}} \Sigma^{-1} \left(x_{i+1} - Ax_i \right)
\end{aligned}
$$

  which is not quadratic in x_i hence $\pi \left(x_{k:k+L} \middle| y_{k:k+L}, x_{k-1}, x_{k+1} \right)$ is not Gaussian.
- The idea is to expand the log-likelihood part around some point estimates

$$
\begin{aligned}
\log g\left(y_i \middle| x_i \right) \simeq\ & \log g\left(y_i \middle| \widehat{x}_i \right) + \nabla \log g\left(y_i \middle| \widehat{x}_i \right) . \left(x_i - \widehat{x}_i \right) \\
& + \frac{1}{2} \left(x_i - \widehat{x}_i \right)^{\mathsf{T}} \nabla^2 \log g\left(y_i \middle| \widehat{x}_i \right) \left(x_i - \widehat{x}_i \right)
\end{aligned}
$$

- By doing this, we have a Gaussian approximation of the log-likelihood and then we obtain a Gaussian proposal $q\left(x_{1:n}, x'_{k:k+L}\right) = q\left(x_{-(k:k+L)}, x'_{k:k+L}\right)$

$$\log q\left(x_{-(k:k+L)}, x'_{k:k+L}\right) \equiv \sum_{i=k}^{k+L} \nabla \log g\left(y_i \middle| \widehat{x}_i\right) . \left(x_i - \widehat{x}_i\right)$$
$$+\frac{1}{2}\left(x_i - \widehat{x}_i\right)^{\mathsf{T}} \nabla^2 \log g\left(y_i \middle| \widehat{x}_i\right)\left(x_i - \widehat{x}_i\right)$$
$$-\frac{1}{2}\sum_{i=k}^{k+L+1}\left(x_{i+1} - Ax_i\right)^{\mathsf{T}} \Sigma^{-1}\left(x_{i+1} - Ax_i\right)$$

- By doing this, we have a Gaussian approximation of the log-likelihood and then we obtain a Gaussian proposal $q\left(x_{1:n}, x'_{k:k+L}\right) = q\left(x_{-(k:k+L)}, x'_{k:k+L}\right)$

$$\log q\left(x_{-(k:k+L)}, x'_{k:k+L}\right) \equiv \sum_{i=k}^{k+L} \nabla \log g\left(y_i \mid \widehat{x}_i\right) . \left(x_i - \widehat{x}_i\right) \\ +\frac{1}{2}\left(x_i - \widehat{x}_i\right)^{\mathrm{T}} \nabla^2 \log g\left(y_i \mid \widehat{x}_i\right)\left(x_i - \widehat{x}_i\right) \\ -\frac{1}{2} \sum_{i=k}^{k+L+1} \left(x_{i+1} - Ax_i\right)^{\mathrm{T}} \Sigma^{-1} \left(x_{i+1} - Ax_i\right)$$

- (Pitt & Shepard, 1999) propose to select

$$\widehat{x}_{k:k+1} = \arg\max \pi\left(x_{k:k+L} \mid y_{k:k+L}, x_{k-1}, x_{k+L+1}\right)$$

and a scheme to sample from $q\left(x_{-(k:k+L)}, x'_{k:k+L}\right)$ which is of complexity $O(L)$.

- This algorithm is applied to the SV model where

$$
\begin{aligned}
X_k &= \phi X_{k-1} + \sigma V_k, \;\; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\
Y_k &= \beta \exp(X_k/2)\, W_k, \;\; W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).
\end{aligned}
$$

- This algorithm is applied to the SV model where

$$\begin{aligned} X_k &= \phi X_{k-1} + \sigma V_k, \;\; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ Y_k &= \beta \exp(X_k/2)\, W_k, \;\; W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

- Prior are set to $\phi \sim \mathcal{U}[-1,1]$, $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\gamma_\sigma}{2}\right)$ and $\beta \sim \mathcal{IG}\left(\frac{\nu_\beta}{2}, \frac{\gamma_\beta}{2}\right)$.

- This algorithm is applied to the SV model where

$$\begin{aligned} X_k &= \phi X_{k-1} + \sigma V_k, \;\; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ Y_k &= \beta \exp(X_k/2)\, W_k, \;\; W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

- Prior are set to $\phi \sim \mathcal{U}[-1,1]$, $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\gamma_\sigma}{2}\right)$ and $\beta \sim \mathcal{IG}\left(\frac{\nu_\beta}{2}, \frac{\gamma_\beta}{2}\right)$.
- Full conditional distributions of the parameters given $x_{1:n}, y_{1:n}$ are standard.

- This algorithm is applied to the SV model where

$$\begin{aligned} X_k &= \phi X_{k-1} + \sigma V_k, \;\; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ Y_k &= \beta \exp(X_k/2) W_k, \;\; W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

- Prior are set to $\phi \sim \mathcal{U}[-1,1]$, $\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\gamma_\sigma}{2}\right)$ and $\beta \sim \mathcal{IG}\left(\frac{\nu_\beta}{2}, \frac{\gamma_\beta}{2}\right)$.
- Full conditional distributions of the parameters given $x_{1:n}, y_{1:n}$ are standard.
- Compared to standard single move strategies, the authors report significant improvement.

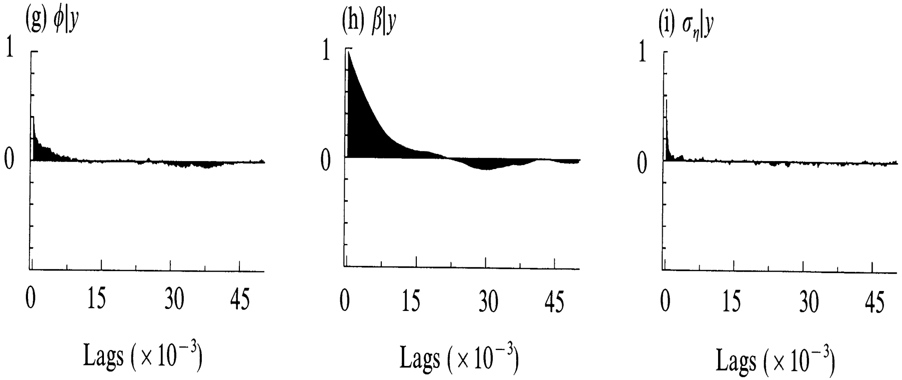

Figure: Autocorrelation plots for (ϕ, σ^2, β) with $L = 1$

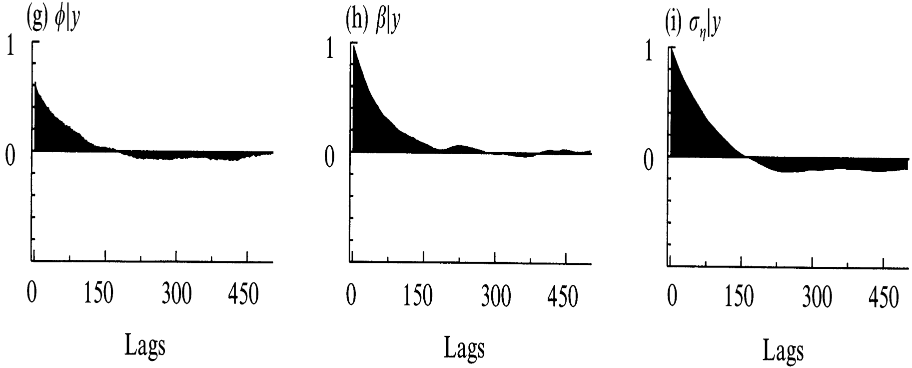

Figure: Autocorrelation plots for (ϕ, σ^2, β) with $L = 50$ on average