VU Minor Applied Econometrics

Bayesian Econometrics for Business & Economics (Bayesian statistics & simulation methods)

Period 2, 2023-2024 (E_MFAE_BEBE)

Lennart Hoogerheide

Vrije Universiteit Amsterdam & Tinbergen Institute

E-mail: l.f.hoogerheide@vu.nl

Lecture 3:

- Exercise 2: Model with geometric distribution.

- Exercise 3: normally distributed data:
  Gibbs sampling method in case of normal prior distribution for μ.

- Simulation method: random walk Metropolis(-Hastings) method:

  - Application to posterior density kernel $P(\theta) = p(\theta)p(y|\theta)$ in Autoregressive Conditional Heteroskedasticity (ARCH) model.

  - Application to uniform target density $P(\theta)$. (Purely for illustration!)

Exercise 2: Bayesian analysis of model with geometric distribution

(a) Suppose we have a set $y = \{y_i | i = 1, \ldots, n\}$ of independent and identically distributed (i.i.d.) random variables y_i, which have a Geometric(θ) distribution. That is, each y_i is the number of Bernoulli trials (with probability of 'success' equal to θ) *before* the first success.
We have probability function:

$$p(y_i|\theta) = \begin{cases} (1-\theta)^{y_i}\theta & \text{if } y_i = 0, 1, 2, \ldots \\ 0 & \text{else.} \end{cases}$$

Suppose we specify a non-informative prior for θ: a uniform distribution on the interval $[0, 1]$:

$$p(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

What is the likelihood?
Use Bayes' rule to derive a *kernel* ($=$ proportionality function) of the posterior density $p(\theta|y)$.

Answer: The likelihood is

$$\begin{aligned} p(y|\theta) &= p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \qquad \text{(due to independence)} \\ &= \prod_{i=1}^{n} (1-\theta)^{y_i}\theta \\ &= \theta^n (1-\theta)^{\sum_{i=1}^{n} y_i} \end{aligned}$$

with $0 \le \theta \le 1$, since θ is a probability.
Note: same as for Bernoulli distribution with $n_1 = n$ *'successes'* and $n_0 = \sum_{i=1}^{n} y_i$ *'failures'*.

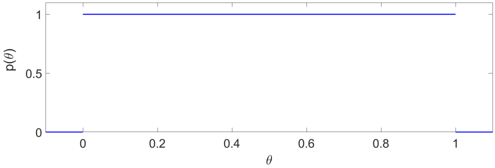
Bayes' rule says that:

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta).$$

So, a kernel of the posterior density $p(\theta|y)$ is given by:

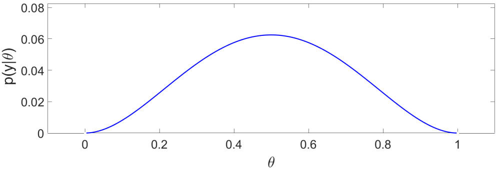
$$p(\theta|y) \propto \begin{cases} \theta^n (1-\theta)^{\sum_{i=1}^{n} y_i} & \text{if } 0 \le \theta \le 1, \\ 0 & \text{else.} \end{cases}$$

Prior $p(\theta)$:



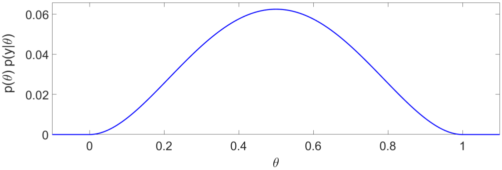
Example of dataset: $n = 2$, $y_1 = y_2 = 1$:
(equivalent with dataset from Bernoulli model with $n = 4$, $y_1 = 0$, $y_2 = 1$, $y_3 = 0$, $y_4 = 1$):

Likelihood: $p(y|\theta) = \theta^2(1-\theta)^2$ for $0 \leq \theta \leq 1$:



Posterior density kernel:

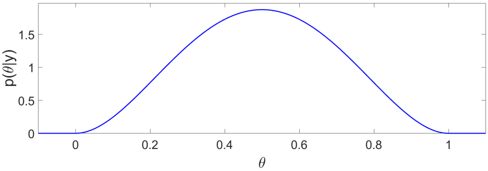
$$p(\theta|y) \propto \begin{cases} \theta^2(1-\theta)^2 & \text{if } 0 \le \theta \le 1, \\ 0 & \text{else.} \end{cases}$$



Note: integral (area under the graph) is **not** equal to 1.

Posterior density:

$$p(\theta|y) = \begin{cases} 30\ \theta^2(1-\theta)^2 & \text{if } 0 \le \theta \le 1, \\ 0 & \text{else.} \end{cases}$$



Note: integral (area under the graph) is equal to 1.
Scaling constant $= \frac{\Gamma(6)}{\Gamma(3)\ \Gamma(3)} = \frac{(6-1)!}{(3-1)!\ (3-1)!} = \frac{120}{2\times 2} = 30.$

(b) What is the exact posterior density $p(\theta|y)$, including the scaling factor? You can make use of Table 1a-1b that provides an overview of some continuous and discrete probability distributions.

Answer: We have kernel of the posterior density $p(\theta|y)$:

$$p(\theta|y) \propto \begin{cases} \theta^n(1-\theta)^{\sum_{i=1}^n y_i} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

Recognize: this is the density of the Beta(a, b) distribution

$$p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \qquad (0 \leq x \leq 1)$$

with $a = n + 1$ and $b = 1 + \sum_{i=1}^n y_i$
(because $a - 1 = n$ and $b - 1 = \sum_{i=1}^n y_i$) and $x = \theta$. So, we have:

$$p(\theta|y) = \begin{cases} \frac{\Gamma(n+\sum_{i=1}^n y_i+2)}{\Gamma(n+1)\Gamma(\sum_{i=1}^n y_i+1)} \theta^n(1-\theta)^{\sum_{i=1}^n y_i} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

Note: here we can replace $\propto$ with $=$.

Table 1a: Continuous distributions

distribution	probability density function	mean
Beta(a, b)	$p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ for $0 \leq x \leq 1$	$\frac{a}{a+b}$
Exponential(b)	$p(x) = \frac{1}{b} \exp\left(-\frac{x}{b}\right)$ for $x \geq 0$	b
Gamma(a, b)	$p(x) = \frac{1}{\Gamma(a) b^a} x^{a-1} \exp(-\frac{x}{b})$ for $x \geq 0$	$a \cdot b$
Normal $N(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ for $-\infty < x < \infty$	μ
Student-t(μ, σ^2, DoF)	$p(x) = \frac{\Gamma(\frac{DoF+1}{2})}{\Gamma(\frac{DoF}{2})\sqrt{DoF\pi}} \frac{1}{\sigma} \left(1 + \frac{(x-\mu)^2}{DoF\,\sigma^2}\right)^{-\frac{DoF+1}{2}}$ for $-\infty < x < \infty$	μ

Table 1b: Discrete distributions

distribution	probability mass function	mean
Bernoulli(a)	$p(x) = a^x(1-a)^{1-x}$ for $x = 0, 1$	a
Binomial(n, a)	$p(x) = \frac{n!}{x!(n-x)!}a^x(1-a)^{n-x}$ for $x = 0, 1, \ldots, n$	$n \cdot a$
Geometric(a)	$p(x) = (1-a)^x a$ for $x = 0, 1, 2, \ldots$	$\frac{1-a}{a}$
Poisson(a)	$p(x) = \frac{a^x \exp(-a)}{x!}$ for $x = 0, 1, 2, \ldots$	a

(c) What is the posterior mean $E(\theta|y)$ of the parameter θ in the model with the Geometric(θ) distribution if we have $n = 2$, $y_1 = y_2 = 1$.

Answer: The posterior mean $E(\theta|y)$ is the mean of a Beta distribution with parameters $a = n + 1$ and $b = 1 + \sum_{i=1}^{n} y_i$, so that

$$E(\theta|y) = \frac{a}{a+b} = \frac{n+1}{n+2+\sum_{i=1}^{n} y_i}.$$

For $n = 2$, $y_1 = y_2 = 1$ we have

$$E(\theta|y) = \frac{n+1}{n+2+\sum_{i=1}^{n} y_i} = \frac{3}{6} = \frac{1}{2}.$$

Note: results are the same as for Bernoulli distribution with $n_1 = n$ 'successes' and $n_0 = \sum_{i=1}^{n} y_i$ 'failures'.
(A dataset of $y_1 = y_2 = 1$ from a geometric distribution is equivalent with a dataset of $y_1 = 0, y_2 = 1, y_3 = 0, y_4 = 1$ from a Bernoulli distribution.)

Exercise 4: Model with Bernoulli distribution with informative, conjugate prior

Bernoulli distribution:

- $y_i = 1$ with probability θ
- $y_i = 0$ with probability $1 - \theta$

Likelihood:

$$
\begin{aligned}
p(y|\theta) &= p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{\sum_{i=1}^{n}(1-y_i)} \\
&= \theta^{n_1}(1-\theta)^{n_0}
\end{aligned}
$$

with $n_1 \equiv \sum_{i=1}^{n} y_i$ the number of ones in the sample, and
with $n_0 \equiv \sum_{i=1}^{n}(1-y_i)$ the number of zeros in the sample.

Suppose that we specify an **informative prior**: e.g., a $Beta(\tilde{n}_1 + 1, \tilde{n}_0 + 1)$ distribution on the interval $[0, 1]$:

$$p(\theta) = \begin{cases} \frac{\Gamma(\tilde{n}_1+\tilde{n}_0+2)}{\Gamma(\tilde{n}_1+1)\Gamma(\tilde{n}_0+1)}\,\theta^{\tilde{n}_1}(1-\theta)^{\tilde{n}_0} & \text{if } 0 \le \theta \le 1, \\ 0 & \text{else.} \end{cases}$$

This is a **conjugate** prior, where the prior has the shape of a posterior that is based on an older dataset.
For example, $\tilde{n}_1 = 2$ and $\tilde{n}_0 = 8$ would have the same effect as adding 10 artificial observations (with two "successes" and eight "failures") to our actual dataset.

(a) Suppose that we specify this $Beta(\tilde{n}_1 + 1, \tilde{n}_0 + 1)$ prior distribution, and that we have a dataset of n observations with n_1 "successes" and n_0 "failures". Use Bayes' rule to derive a kernel of posterior density $p(\theta|y)$.

(b) What is the exact posterior density $p(\theta|y)$, including the scaling factor? What is the posterior mean $E(\theta|y)$? You can make use of Table 1a-1b.

Exercise 3: normally distributed data: Gibbs sampling method in case of normal prior distribution for μ.

Consider the model with i.i.d. normally distributed observations $y_j \sim N\left(\mu, \frac{1}{h}\right)$, $j = 1, \ldots, n$ with prior

$$p(\theta) = p(\mu, h) = p(\mu)p(h)$$

with

$$p(h) \propto \frac{1}{h} \qquad \text{for } h > 0.$$

Now suppose that we specify a normal prior distribution for μ: $\mu \sim N(m_{prior}, v_{prior})$, so that

$$p(\mu) = (2\pi v_{prior})^{-1/2} \ \exp\left(-\frac{(\mu - m_{prior})^2}{2v_{prior}}\right).$$

In this case the steps of the Gibbs sampling method are given on the next slide.

Gibbs sampling method in case of normal prior distribution for μ :

- Choose initial value, for example $\mu_0 = \bar{y}$
- Do for draw $i = 1, \ldots, n_{draws}$:
  - Simulate h_i from Gamma($a = \frac{n}{2}, \;\; b = (\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu_{i-1})^2)^{-1}$) distribution.
  - Simulate μ_i from normal distribution:

$$N\left(\frac{\frac{m_{prior}}{v_{prior}} + h_i n \bar{y}}{\frac{1}{v_{prior}} + h_i n}, \frac{1}{\frac{1}{v_{prior}} + h_i n}\right)$$

- Discard *burn-in* of first draws.

Give a derivation of the abovementioned conditional posterior distributions

$$h \mid \mu, \; y \;\; \sim \;\; \text{Gamma}\left(a = \frac{n}{2}, \;\; b = \left(\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right)^{-1}\right),$$

$$\mu \mid h, \; y \;\; \sim \;\; N\left(\frac{\frac{m_{prior}}{v_{prior}} + h n \bar{y}}{\frac{1}{v_{prior}} + h n}, \frac{1}{\frac{1}{v_{prior}} + h n}\right).$$

Answer:
Model: multiple i.i.d. observations $y = (y_1, \ldots, y_n)'$; $y_j \sim N(\mu, \sigma^2)$ $(j = 1, 2, \ldots, n)$ with **unknown** mean μ and **unknown** precision h $(= 1/\sigma^2)$.

Likelihood:

$$\begin{aligned} p(y|\mu, h) &= p(y_1, \ldots, y_n|\mu, h) \\ &= \prod_{j=1}^{n} \left(\frac{2\pi}{h}\right)^{-1/2} \exp\left(-\frac{h}{2}(y_j - \mu)^2\right) \\ &= \left(\frac{2\pi}{h}\right)^{-n/2} \exp\left(-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right) \end{aligned}$$

The kernel of the joint posterior density becomes:

$$
\begin{aligned}
p(\mu, h|y) &\propto p(\mu, h) \times p(y|\mu, h) \\
&\propto h^{-1} \exp\left(-\frac{1}{2}\frac{(\mu - m_{prior})^2}{v_{prior}}\right) \times \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] \\
&\propto h^{n/2-1} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j - \mu)^2\right] \exp\left(-\frac{1}{2}\frac{(\mu - m_{prior})^2}{v_{prior}}\right).
\end{aligned}
$$

The kernel of the joint posterior density:

$$p(\mu, h|y) \quad \propto \quad h^{n/2-1} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right] \exp\left(-\frac{1}{2}\frac{(\mu-m_{prior})^2}{v_{prior}}\right).$$

If we consider this as a function of h (for fixed μ), then this is proportional to (the same as before):

$$p(h|\mu, y) = \frac{p(\mu, h|y)}{p(\mu|y)} \propto p(\mu, h|y) \propto h^{n/2-1} \exp\left[-\frac{h}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right].$$

$\Rightarrow$ Conditional posterior density of h given μ is the Gamma density:

$$\begin{aligned} p(h|\mu, y) &= \frac{1}{\Gamma(a)b^a} h^{a-1} \exp\left(-\frac{h}{b}\right) \\ &= \frac{\left[\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right]^{n/2}}{\Gamma(n/2)} h^{n/2-1} \exp\left(-\left[\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu)^2\right] h\right) \end{aligned}$$

Conditional posterior distribution of μ *given* h (=precision= $1/\sigma^2$) is the normal posterior distribution for the case with observation $\bar{y}$ with "known" variance $\frac{\sigma^2}{n}$:

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\mu, \frac{1}{hn}\right).$$

So, we have posterior

$$\mu|h, y \sim N(m_{posterior}, v_{posterior})$$

with

$$m_{posterior} = \frac{\frac{m_{prior}}{v_{prior}}}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2/n}} + \frac{\frac{\bar{y}}{\sigma^2/n}}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2/n}} = \frac{\frac{m_{prior}}{v_{prior}} + hn\bar{y}}{\frac{1}{v_{prior}} + hn}$$

$$v_{posterior} = \frac{1}{\frac{1}{v_{prior}} + \frac{1}{\sigma^2/n}} = \frac{1}{\frac{1}{v_{prior}} + hn}$$

Note: choosing $v_{prior} \to \infty$ (so that $\frac{1}{v_{prior}} = 0$ and $\frac{m_{prior}}{v_{prior}} = 0$) corresponds to non-informative prior.

Overview of integration methods:

integration

analytical

numerical

deterministic (possible for dim. $\theta \leq 3$)

simulation (Monte Carlo)

direct simulation

indirect simulation

well-known conditional posteriors:

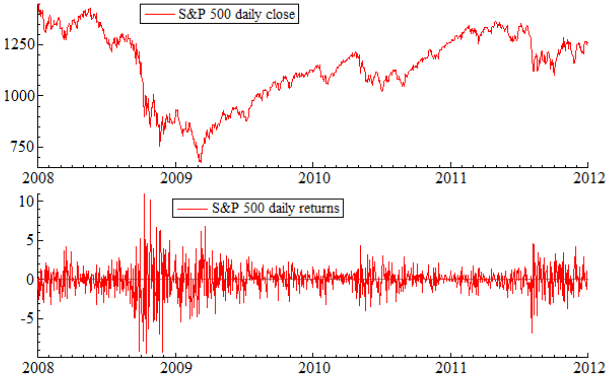
unknown conditional posteriors:

Gibbs sampling

Metropolis-Hastings: random walk Metropolis(-Hastings), independence chain Metropolis-Hastings.

Example: Autoregressive Conditional Heteroskedasticity (ARCH) model

Data: S&P 500 daily close p_t and log-returns $y_t = 100 \times \ln(\frac{p_t}{p_{t-1}})$:



Note: **volatility clustering**: consecutive periods with large variance and consecutive periods with small variance.

Simple case: ARCH(1) model with mean 0 and normal distribution:

$$y_t | I_{t-1} \sim N(0, \sigma_t^2)$$

with conditional variance

$$\sigma_t^2 = \text{var}(y_t | I_{t-1}) = \alpha_0 + \alpha_1 y_{t-1}^2$$

with information set $I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$.

Two parameters:

- α_0 $(\alpha_0 > 0)$: constant term in variance equation
- α_1 $(0 \leq \alpha_1 < 1)$: effect of yesterday's squared return y_{t-1}^2 on today's return's variance $\text{var}(y_t | I_{t-1})$.

Note: the restrictions $\alpha_0 > 0$ and $\alpha_1 \geq 0$ ensure that $\alpha_0 + \alpha_1 y_{t-1}^2 > 0$.

Simple case: ARCH(1) model with mean 0 and normal distribution:

$$y_t | I_{t-1} \sim N(0, \sigma_t^2) \qquad \sigma_t^2 = \mathsf{var}(y_t | I_{t-1}) = \alpha_0 + \alpha_1 y_{t-1}^2.$$

The unconditional variance is:

$$\mathsf{var}(y_t) = \frac{\alpha_0}{1 - \alpha_1}$$

(Derivation:

$$\begin{aligned}
\mathsf{var}(y_t) &= E(y_t^2) \\
&= E(E(y_t^2 | I_{t-1})) \\
&= E(\alpha_0 + \alpha_1 y_{t-1}^2) \\
&= \alpha_0 + \alpha_1 E(y_{t-1}^2) \\
&= \alpha_0 + \alpha_1 \mathsf{var}(y_{t-1}) \\
&= \alpha_0 + \alpha_1 \mathsf{var}(y_t),
\end{aligned}$$

where

$$\mathsf{var}(y_t) = \mathsf{var}(y_{t-1})$$

holds because the ARCH(1) process is stationary for $0 \leq \alpha_1 < 1$.)

Variance targeting: estimate model so that (estimated) unconditional variance is equal to sample variance $s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2$.

Here in ARCH(1) model:

$$\begin{aligned} \frac{\alpha_0}{1-\alpha_1} &= s^2 \\ \alpha_0 &= s^2(1-\alpha_1) \end{aligned}$$

ARCH(1) model becomes:

$$y_t \sim N(0, \sigma_t^2) \qquad \sigma_t^2 = \mathsf{var}(y_t | I_{t-1}) = s^2(1-\alpha_1) + \alpha_1 y_{t-1}^2$$

with only one parameter α_1 (good for *illustrative example* of model with ‘non standard’ posterior distribution!)

Conditional density of y_t given $y_{t-1}, y_{t-2}, \ldots$:

$$
\begin{aligned}
p(y_t|y_{t-1},\alpha_1) &= (2\pi\sigma_t^2)^{-1/2}\exp\left(-\frac{y_t^2}{2\sigma_t^2}\right) \\
&= (2\pi[\alpha_0+\alpha_1 y_{t-1}^2])^{-1/2}\exp\left(-\frac{y_t^2}{2[\alpha_0+\alpha_1 y_{t-1}^2]}\right) \\
&= (2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2])^{-1/2}\exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right)
\end{aligned}
$$

for

- any (G)ARCH model with normal density,
- ARCH(1) model with normal density, and
- ARCH(1) model with normal density with variance targeting,

respectively.

Likelihood (conditional on 'fixed' first observation y_1):

$$
\begin{aligned}
p(y_2,\ldots,y_n|\alpha_1) &= \prod_{t=2}^{n} p(y_t|y_{t-1},y_{t-2},\ldots,\alpha_1) \\
&= \prod_{t=2}^{n} p(y_t|y_{t-1},\alpha_1) \\
&= \prod_{t=2}^{n} \Bigg\{ \left(2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]\right)^{-1/2} \times \\
&\qquad \exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right)\Bigg\}
\end{aligned}
$$

Prior: suppose we specify (non-informative) uniform prior on [0,1) for α_1:

$$
p(\alpha_1) = \begin{cases} 1 & \text{if } 0 \leq \alpha_1 < 1, \\ 0 & \text{else.} \end{cases}
$$

Posterior:

$$
\begin{aligned}
p(\alpha_1|y) &\propto p(y|\alpha_1)p(\alpha_1) \\
&\propto \prod_{t=2}^{n} \left\{ [s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]^{-1/2} \times \right. \\
&\qquad \left. \exp\left(-\frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]} \right) \right\}
\end{aligned}
$$

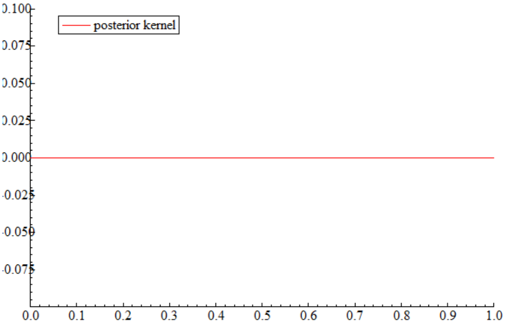
if $0 \leq \alpha_1 < 1$; 0 else.

Note: this is **not** a well-known posterior distribution.
$\Rightarrow$ Use simulation method, for example **Metropolis-Hastings** method.

If we would have two parameters α_0 and α_1, then the *conditional* posterior distributions would also **not** be well-known distributions.
$\Rightarrow$ Gibbs sampling **not** possible for Bayesian analysis of (Generalized) Autoregressive Conditional Heteroskedasticity ((G)ARCH) models.

Numerical problem: posterior density kernel $p(y|\alpha_1)p(\alpha_1)$ often too small (or too large) to be stored on computer.

Plot of $p(y|\alpha_1)p(\alpha_1)$:



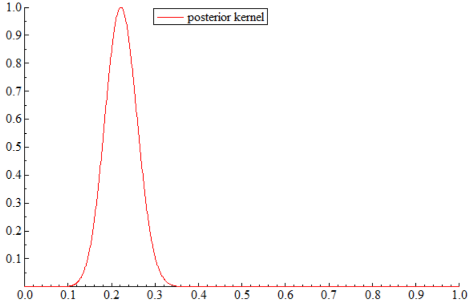
Solution: work with **logarithm** of posterior kernel

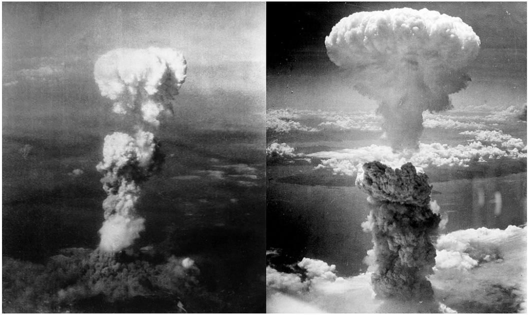
$$f(\alpha_1) = \ln p(y|\alpha_1) + \ln p(\alpha_1).$$

Note: if you want to make a graph of a kernel of the posterior density (which is possible for this 1-dimensional α_1), then we can make a graph of

$$p(\alpha_1|y) \propto \frac{\exp(f(\alpha_1))}{\exp(f(\alpha_{1,mode}))} = \exp(f(\alpha_1) - f(\alpha_{1,mode}))$$

with posterior mode $\alpha_{1,mode}$ $\Rightarrow$ posterior kernel values lie in [0,1] interval:





Possible early applications of Metropolis-Hastings method (August 1945): atomic bombings of Hiroshima (left) and Nagasaki (right).

Publications:

- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953): **random walk Metropolis (-Hastings) method**
- Hastings WK (1970): **independence chain Metropolis-Hastings method**

Random walk Metropolis(-Hastings) method

(random walk: candidate draw from random walk):

- Choose feasible initial value θ_0
- Do for draw $i = 1, \ldots, n_{draws}$:
  - Simulate candidate draw $\tilde{\theta}$ from candidate density $Q(.)$ with mean θ_{i-1} (symmetric candidate density around θ_{i-1})
  - Compute acceptance probability

    $$\alpha = \min\left\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\right\} = \min\left\{\exp[\ln P(\tilde{\theta}) - \ln P(\theta_{i-1})], 1\right\}$$

    with target density kernel $P(\theta)$.
    (In Bayesian estimation, $P(\theta)$ is the posterior density kernel $P(\theta) = p(\theta)p(y|\theta)$, so that $\ln P(\theta) = \ln p(\theta) + \ln p(y|\theta)$.)
  - Simulate U from uniform distribution on $[0, 1]$.
  - If $U \leq \alpha$, then accept: $\theta_i = \tilde{\theta}$ (accept candidate draw).
    If $U > \alpha$, then reject: $\theta_i = \theta_{i-1}$ (repeat previous draw).

Note:

- Acceptance probability α depends on ratio $P(\tilde{\theta})/P(\theta_{i-1})$:
  If $P(\tilde{\theta}) \geq P(\theta_{i-1})$: accept with probability 1.
  If $P(\tilde{\theta}) < P(\theta_{i-1})$: $\tilde{\theta}$ may be rejected.

- We only need ratio $\frac{P(\tilde{\theta})}{P(\theta_{i-1})}$ that does **not** depend on any constant scaling factor in $P(.)$.
  $\Rightarrow$ only need **kernel** $P(\theta)$ of posterior density $p(\theta|y) \propto p(\theta)p(y|\theta)$.

- For numerical reasons we evaluate the **log**-prior and **log**likelihood:

$$\frac{P(\tilde{\theta})}{P(\theta_{i-1})} = \frac{\exp[\ln P(\tilde{\theta})]}{\exp[\ln P(\theta_{i-1})]} = \exp\left[\ln P(\tilde{\theta}) - \ln P(\theta_{i-1})\right]$$

$$= \exp\left[\ln p(\tilde{\theta}) + \ln p(y|\tilde{\theta}) - \ln p(\theta_{i-1}) - \ln p(y|\theta_{i-1})\right]$$

The draws $\theta_0, \theta_1, \theta_2, \ldots$ form a Markov chain that *converges in distribution* to the posterior distribution. $\Rightarrow$

Discard a **burn-in** of the first draws to delete the effect of initial value θ_0 (just like for the Gibbs sampling method).

Purely illustrative example with uniform target density $P(\theta)$:

- Suppose we want to simulate from the uniform distribution on the interval [0,1], so that the target density is

$$P(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1, \\ 0 & \text{else}, \end{cases}$$

  where we use the random walk Metropolis(-Hastings) method.[1]

- Suppose we simulate the candidate draw $\tilde{\theta}$ from the normal distribution $N(\theta_{i-1}, \sigma^2_{candidate})$.

[1]Obviously, this is only for illustrative purposes! In practice, if we want to simulate from the uniform distribution on [0,1], then we **directly** simulate from this.

The acceptance probability is given by

$$\alpha = \min\left\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\right\} =$$

$$= \min\left\{\tfrac{1}{1}, 1\right\} = 1 \qquad \text{if } \tilde{\theta} \in [0,1] \text{ and } \theta_{i-1} \in [0,1]$$

$$= \min\left\{\tfrac{0}{1}, 1\right\} = 0 \qquad \text{if } \tilde{\theta} \notin [0,1] \text{ and } \theta_{i-1} \in [0,1]$$

$$= \min\left\{\tfrac{1}{0}, 1\right\} = \text{undefined} \qquad \text{if } \tilde{\theta} \in [0,1] \text{ and } \underline{\theta_{i-1} \notin [0,1]}$$

$$= \min\left\{\tfrac{0}{0}, 1\right\} = \text{undefined} \qquad \text{if } \tilde{\theta} \notin [0,1] \text{ and } \underline{\theta_{i-1} \notin [0,1]}$$

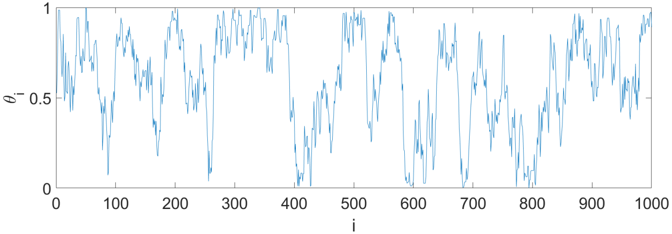
Note: the latter two cases do not occur:

- We choose initial value $\theta_0 \in [0,1]$,
- Each candidate draw $\tilde{\theta}$ outside [0,1] is rejected.
- So, we never have $\theta_{i-1} \notin [0,1]$.

So: we simply accept every $\tilde{\theta} \in [0,1]$ and reject every $\tilde{\theta} \notin [0,1]$.

How to evaluate whether a candidate distribution is 'good' or 'bad'?

- **'trace plot'** of (accepted and repeated) draws $\theta_0, \theta_1, \theta_2, \ldots$:



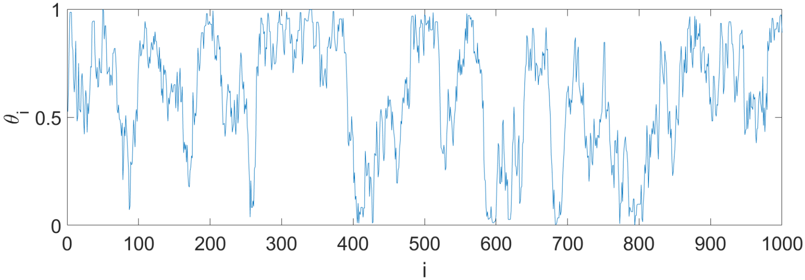
  Do the draws move through the parameter space 'fast enough'?

- **acceptance percentage:** what percentage of the candidate draws is accepted? A percentage close to 0% is bad. But a percentage close to 100% can be bad too!

- **(first order) serial correlation in sequence of (accepted and repeated) draws.**
  The lower the serial correlation, the better. (Close to 1 is bad.)

Case with small candidate steps: $\sigma_{candidate} = 0.1$**:**
$\tilde{\theta} \sim N(\theta_{i-1}, 0.1^2)$ and $n_{draws} = 100000$.

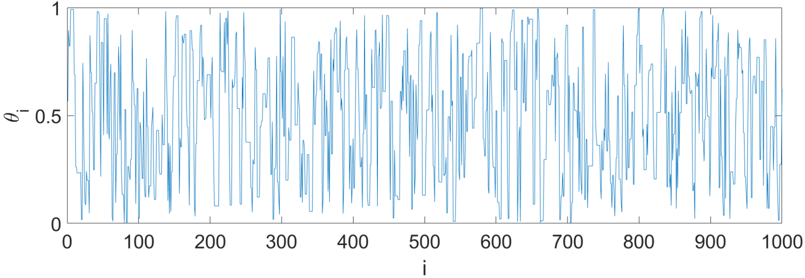
- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



- acceptance percentage = 92%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr(θ_i, θ_{i-1}) = 0.95.

Case with reasonable candidate steps: $\sigma_{candidate} = 0.5$:
$\tilde{\theta} \sim N(\theta_{i-1}, 0.5^2)$ and $n_{draws} = 100000$.

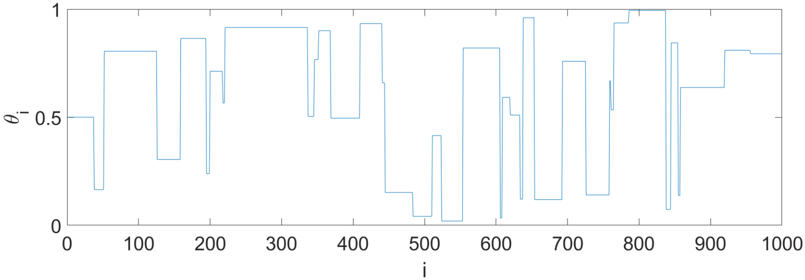
- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



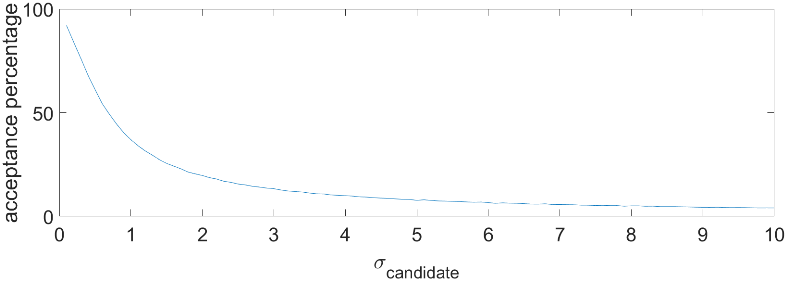
- acceptance percentage $= 61\%$.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr$(\theta_i, \theta_{i-1}) = 0.61$.

Case with large candidate steps: $\sigma_{candidate} = 10$**:**
$\tilde{\theta} \sim N(\theta_{i-1}, 10^2)$ and $n_{draws} = 100000$.

- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



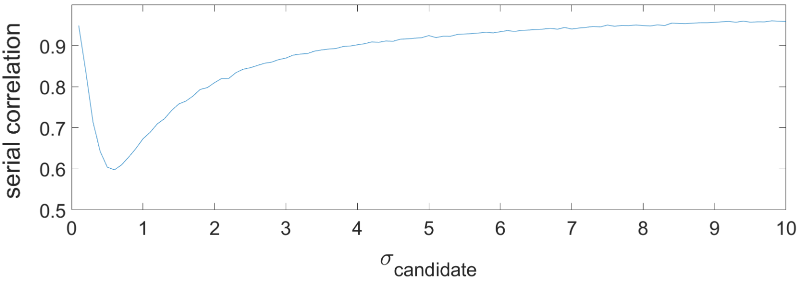
- acceptance percentage = 4%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr(θ_i, θ_{i-1}) = 0.96.



Note: For random walk Metropolis(-Hastings) method we observe that:

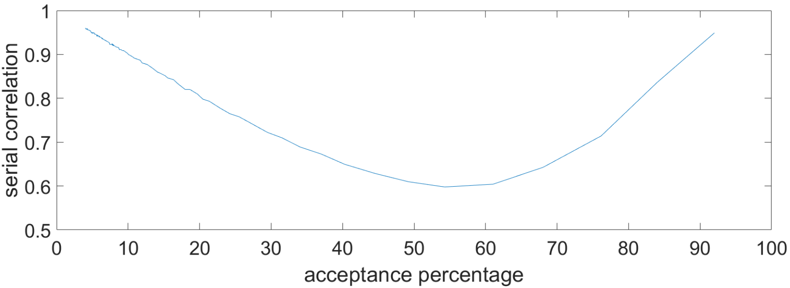
- very small candidate steps are often accepted.
- very large candidate steps are often rejected.

If $\sigma_{candidate} \to 0$, then $\tilde{\theta} \approx \theta_{i-1}$, $P(\tilde{\theta}) \approx P(\theta_{i-1})$, so
$\alpha = \min\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\} \approx 1$.
Then acceptance percentage $\to$ 100%, but also serial correlation $\to$ 1.



Note: for random walk Metropolis(-Hastings) method we have poor performance (i.e., large serial correlation corr(θ_i, θ_{i-1})), if we have

- too small candidate steps (that move too slowly through the parameter space)
- too large candidate steps (that are mostly rejected).



Note: For random walk Metropolis(-Hastings) method **in this example**:

- Best performance (i.e., lowest serial correlation corr(θ_i, θ_{i-1})) if acceptance percentage has 'moderate' value around 50%-60%.

- Reasonable performance (reasonable serial correlation corr(θ_i, θ_{i-1})) if acceptance percentage has value between 20% and 80%.

Literature: For normal target density 23.4% is 'optimal' acceptance rate.

Application to posterior density kernel in ARCH(1) model

Target density kernel $P(\alpha_1) = p(\alpha_1)p(y|\alpha_1)$ with logarithm

$$\ln P(\alpha_1) = \ln p(\alpha_1) + \ln p(y|\alpha_1)$$

with log-prior:

$$\ln p(\alpha_1) \quad = \quad \begin{cases} \ln(1) = 0 & \text{if } 0 \leq \alpha_1 \leq 1, \\ \\ \ln(0) = -\infty & \text{else.} \end{cases}$$

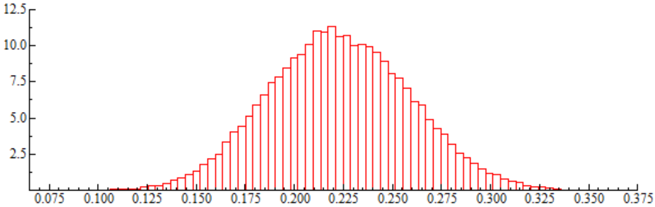
and loglikelihood:

$$\ln p(y|\alpha_1) = \ln\left(\prod_{t=2}^{n} p(y_t|y_{t-1},\alpha_1)\right) = \sum_{t=2}^{n} \ln(p(y_t|y_{t-1},\alpha_1)) =$$

$$= \quad \sum_{t=2}^{n}\left\{-\frac{1}{2}\ln(2\pi[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]) - \frac{y_t^2}{2[s^2(1-\alpha_1)+\alpha_1 y_{t-1}^2]}\right\}$$

In our ARCH(1) model:

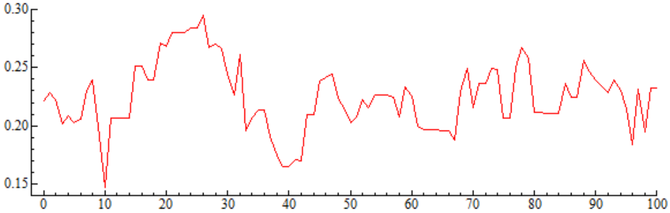
- Initial value $\theta_0 =$ ML estimator.
- Candidate distribution: $\tilde{\theta} \sim N(\theta_{i-1}, 0.03^2)$ (normal distribution with small standard deviation 0.03).
- $n_{draws} = 100100$ draws (with burn-in of 1000 draws).



Posterior mean (stdev):	0.223	(0.036).
Maximum likelihood estimator (standard error):	0.221	(0.037)

Evaluation of quality of candidate distribution $\tilde{\theta} \sim N(\theta_{i-1}, 0.03^2)$:

- **'trace plot'** of (accepted) draws $\theta_0, \theta_1, \theta_2, \ldots$:

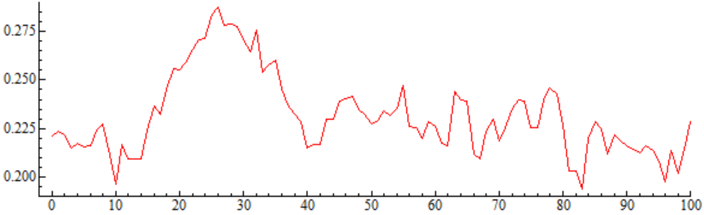


- **acceptance percentage:** what percentage of the candidate draws is accepted?
  Here: acceptance percentage = 75.4% (rather high).

- **(first order) serial correlation in sequence of (accepted) draws.**
  Here: corr$(\alpha_{1,i}, \alpha_{1,i-1}) = 0.816$ (rather high).

Evaluation of quality of candidate distribution $\tilde{\theta} \sim N(\theta_{i-1}, 0.01^2)$ with smaller candidate steps:

- **'trace plot'** of (accepted) draws $\theta_0, \theta_1, \theta_2, \ldots$:



- **acceptance percentage:** what percentage of the candidate draws is accepted?
  Here: acceptance percentage = 91.5% (very high).

- **(first order) serial correlation in sequence of (accepted) draws.**
  Here: corr$(\alpha_{1,i}, \alpha_{1,i-1}) = 0.967$ (very high).

Note again:

- A high acceptance percentage does **not** immediately imply a good quality of (the stdev of) the candidate distribution in the random walk Metropolis(-Hastings) method:

  If variance of candidate distribution $\to 0$, then

  - $\tilde{\theta} \approx \theta_{i-1}$
  - $P(\tilde{\theta}) \approx P(\theta_{i-1})$
  - $\alpha = \min\{\frac{P(\tilde{\theta})}{P(\theta_{i-1})}, 1\} \approx 1$
  - acceptance percentage $\to$ 100%.

  But then also serial correlation $\to 1$.

  Then the random walk Metropolis(-Hastings) method is very **inefficient**: a huge number of draws may be required to 'cover' the whole posterior distribution.