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 4:

- Exercise 4: Bernoulli distributed data:
  posterior distribution of θ in case of informative, conjugate prior.

- Simulation method: random walk Metropolis(-Hastings) method:
  - Application to posterior density kernel $P(\theta) = p(\theta)p(y|\theta)$ in Poisson regression model (with 2-dimensional $\theta = (\beta_0, \beta_1)'$).
  - Application to uniform target density $P(\theta)$. Purely for illustration! (Slides 21-29 are slides 34-42 from lecture 3.)

- Simulation method: independence chain Metropolis-Hastings method:
  - Application to uniform target density $P(\theta)$. (Purely for illustration!)
  - Application to posterior density kernel $P(\theta) = p(\theta)p(y|\theta)$ in Autoregressive Conditional Heteroskedasticity (ARCH) model.

Exercise 4: Bernoulli model under informative, conjugate prior

Model: 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.

Informative prior: $Beta(\tilde{n}_1 + 1, \tilde{n}_0 + 1)$ distribution:

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

This is a **conjugate** prior: a prior such that prior and posterior are in the same probability distribution family (that is, a prior that has the specification like a posterior 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 $p(\theta|y)$.

Answer: Prior:

$$p(\theta) \propto \begin{cases} \theta^{\tilde{n}_1}(1-\theta)^{\tilde{n}_0} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

Likelihood:

$$p(y|\theta) = \theta^{n_1}(1-\theta)^{n_0}.$$

Posterior density kernel:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \begin{cases} \theta^{\tilde{n}_1+n_1}(1-\theta)^{\tilde{n}_0+n_0} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

Note: same posterior density kernel as in case where we have a uniform prior and where we observe $\tilde{n}_1 + n_1$ "successes" and $\tilde{n}_0 + n_0$ "failures".

(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.

Answer: Posterior density kernel:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \begin{cases} \theta^{\tilde{n}_1+n_1}(1-\theta)^{\tilde{n}_0+n_0} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

This is a Beta$(\tilde{n}_1 + n_1 + 1, \tilde{n}_0 + n_0 + 1)$ distribution with density:

$$p(\theta|y) = \begin{cases} \frac{\Gamma(\tilde{n}_1+n_1+\tilde{n}_0+n_0+2)}{\Gamma(\tilde{n}_1+n_1+1)\Gamma(\tilde{n}_0+n_0+1)}\theta^{\tilde{n}_1+n_1}(1-\theta)^{\tilde{n}_0+n_0} & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{else.} \end{cases}$$

The posterior mean is $E(\theta|y) = \frac{\tilde{n}_1+n_1+1}{\tilde{n}_1+n_1+\tilde{n}_0+n_0+2}$.

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

Overview of integration methods:

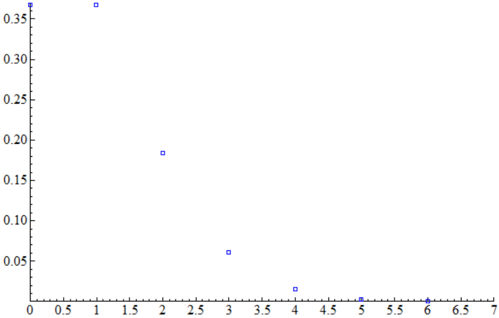
- integration
  - analytical
  - numerical
    - deterministic (possible for dim. $\theta \leq 3$)
    - simulation (Monte Carlo)
      - direct simulation
      - indirect simulation
        - well-known conditional posteriors:
          - Gibbs sampling
        - unknown conditional posteriors:
          - Metropolis-Hastings: random walk Metropolis(-Hastings), independence chain Metropolis-Hastings.

Poisson distribution:

Poisson distribution with mean $E(y_j|\mu) = \mu$: probability mass function:

$$p(y_j|\mu) = \frac{\mu^{y_j} \exp(-\mu)}{y_j!} \qquad y_j = 0, 1, 2, \ldots.$$

Example: Poisson($\mu = 1$) distribution:



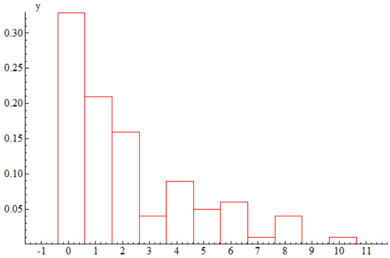
Note: for small mean μ the Poisson(μ) distribution can obviously **not** be approximated well by a normal distribution.

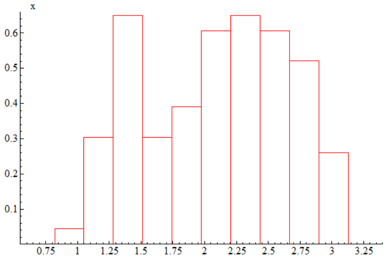
Poisson regression model:

$$y_j \sim \text{Poisson}(\mu_j) \qquad \text{with} \qquad \mu_j = \exp(\beta_0 + \beta_1 x_j).^1$$

Note: exponent to ensure that $\mu_j > 0$.

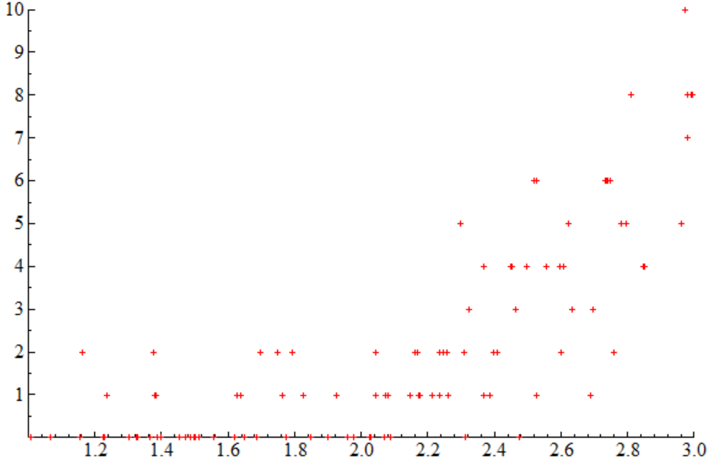
For example, sales y_j (left) and advertisement expenditure x_j (right) $(j = 1, 2, \ldots, n)$ with $n = 100$:



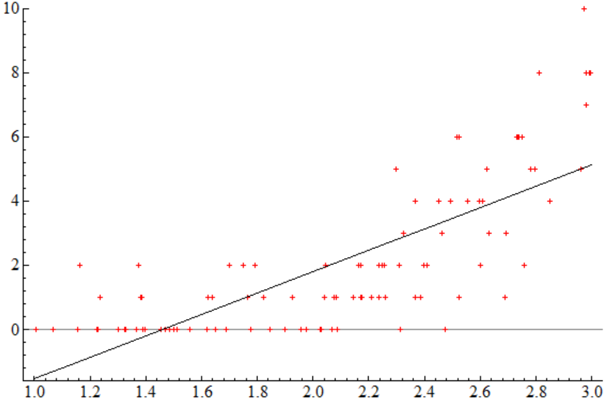


[1] Compare: normal linear regression model has $y_j \sim N(\mu_j, \sigma^2)$ with $\mu_j = \beta_0 + \beta_1 x_j$. Note: everything is conditional on exogenous explanatory variable x_j; $y_j \sim \ldots$ is written instead of $y_j | x_j \sim \ldots$ only to have a shorter/simpler notation here and below.

For example, sales y_j (vertical axis) and advertisement expenditure x_j (horizontal axis) $(j = 1, 2, \ldots, n = 100)$.



Ordinary Least Squares (OLS) may yield negative fitted values $\hat{y}$ for certain values of x:



Note: we can **not** estimate normal regression model for $\ln(y_j)$, because of observations with $y_j = 0$. And the normality assumption for $\ln(y_j)$ or $\ln(1 + y_j)$ would be unrealistic (for example, for prediction interval of y_j).

Poisson probability mass function with mean $\mu_j = \exp(\beta_0 + \beta_1 x_j)$:

$$\begin{aligned} p(y_j|\theta) &= \frac{\mu_j{}^{y_j}\exp(-\mu_j)}{y_j!} \\ &\propto \mu_j{}^{y_j}\exp(-\mu_j) \\ &= \exp(\beta_0 + \beta_1 x_j)^{y_j}\exp(-\exp(\beta_0 + \beta_1 x_j)) \end{aligned}$$

with $\theta = (\beta_0, \beta_1)'$, where everything is conditional on exogenous $x = (x_1, \ldots, x_n)'$.

Likelihood:

$$\begin{aligned} p(y|\theta) &= p(y_1, \ldots, y_n|\theta) = \prod_{j=1}^{n} p(y_j|\theta) \\ &\propto \prod_{j=1}^{n} \exp(\beta_0 + \beta_1 x_j)^{y_j}\exp(-\exp(\beta_0 + \beta_1 x_j)) \end{aligned}$$

Suppose we specify non-informative prior: $p(\theta) \propto 1$ for $\theta = (\beta_0, \beta_1)'$ (with $-\infty < \beta_0 < \infty$, $-\infty < \beta_1 < \infty$).

Posterior:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \prod_{j=1}^{n} \exp(\beta_0 + \beta_1 x_j)^{y_j} \exp(-\exp(\beta_0 + \beta_1 x_j))$$

This is **not** a well-known 'standard' distribution for (β_0, β_1).

For analyzing the posterior we need simulation method like random walk Metropolis(-Hastings) method. Use logarithm of posterior density kernel $P(\theta)$

$$\ln P(\theta) = \sum_{j=1}^{n} \left\{ y_j(\beta_0 + \beta_1 x_j) - \exp(\beta_0 + \beta_1 x_j) \right\}.$$

Random walk Metropolis(-Hastings) method

- Choose feasible initial value θ_0.
  (For example, $(\ln(\bar{y}), 0)$ or the ML estimator.)

- Do for draw $i = 1, \ldots, n_{draws}$:
  - Simulate candidate draw $\tilde{\theta}$ from candidate density $Q(.)$ with mean θ_{i-1}. For example, $\tilde{\theta} \sim N(\theta_{i-1}, \Sigma)$.
  - Compute acceptance probability

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

  - 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).

Common choice of Σ (variance-covariance matrix of candidate distribution) is $-Hessian^{-1}$, where $Hessian$ is the Hessian of the log posterior at the posterior mode, which is here (with flat prior density) the same as $-Hessian^{-1}$ of the loglikelihood at the maximum likelihood estimator:

$$\left(- \frac{\partial^2 \log p(y|\theta)}{\partial\theta\partial\theta'} \bigg|_{\theta=\hat{\theta}_{ML}} \right)^{-1}.$$

Here:

$$\Sigma = \begin{pmatrix} 0.2085 & -0.0796 \\ -0.0796 & 0.0311 \end{pmatrix}.$$

Note: classical/frequentist statistician has for large enough number of observations n (under regularity conditions):

$$\hat{\theta}_{ML} \approx N\left(\theta, \left(- \frac{\partial^2 \log p(y|\theta)}{\partial\theta\partial\theta'} \bigg|_{\theta=\hat{\theta}_{ML}} \right)^{-1}\right).$$

Simulation of $\tilde{\theta}$ from $N(\theta_{i-1}, \Sigma)$, the 2-dimensional normal distribution with mean θ_{i-1} and variance-covariance matrix Σ:

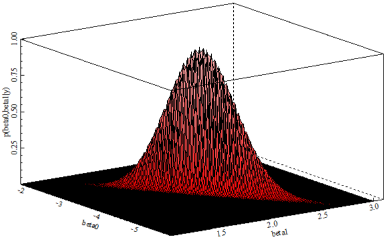
- Use built-in function for simulation from multivariate normal distribution, if this exists.

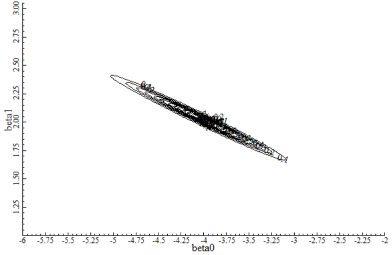
- In Ox/OxMetrics: compute

$$\tilde{\theta} = \theta_{i-1} + \text{choleski}(\Sigma) Z$$

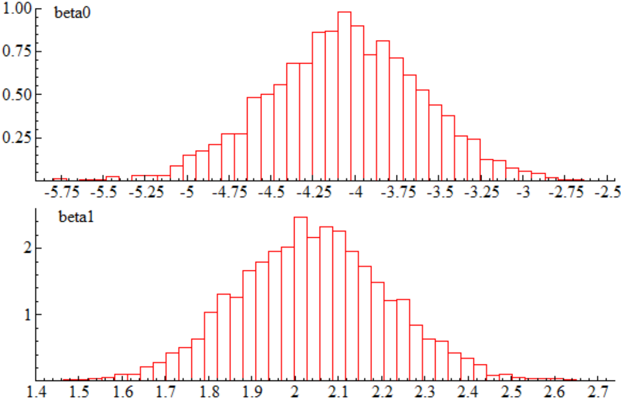
  where $Z = (Z_1, Z_2)'$ is a 2×1 vector of 2 independent standard normal N(0,1) variables.

Kernel of posterior density $p(\beta_0, \beta_1|y)$ (scaled to have maximum 1):





Histograms of 10000 draws (after burn-in of 1000 draws):



Properties of posterior distribution:

	mean	st. dev.	2.5% quantile	97.5% quantile
β_0	-4.0608	0.4577	-4.9744	-3.1738
β_1	2.0415	0.1765	1.7011	2.3919

We have:

- Acceptance percentage: 56.43%.
  $\Rightarrow$ reasonable 'moderate' value
  $\Rightarrow$ this suggests that there are no problems with (too small, too large, or otherwise poorly chosen) variance-covariance matrix Σ of the candidate steps.

- 95% posterior interval of β_1 is [1.7011, 2.3919] which does not include 0.
  $\Rightarrow$ Bayesian statistician using 95% posterior interval rejects $\beta_1 = 0$.
  $\Rightarrow$ Conclusion: There is effect of x (advertisement) on y (sales).

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.[2]

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

[2]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

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

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

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

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

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

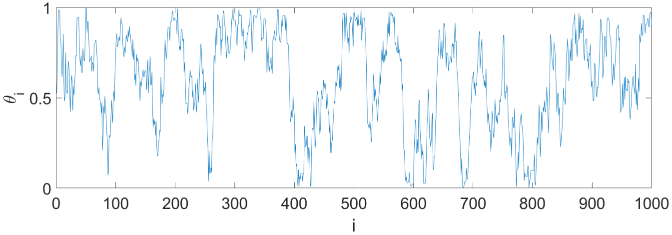
Note: the latter two cases do not occur:

- 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 $\theta_0, \theta_1, \theta_2, \ldots$:



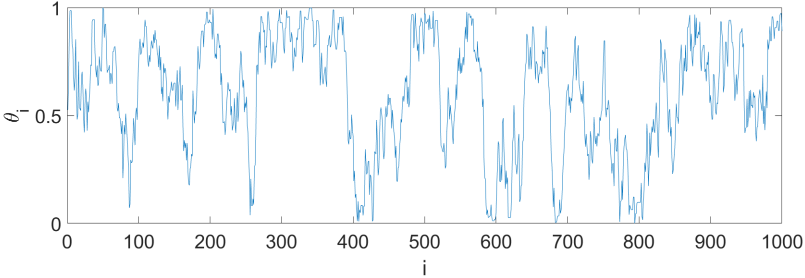
  Do the draws move through the parameter space 'fast enough'?

- **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 $n_{draws} = 100000.$

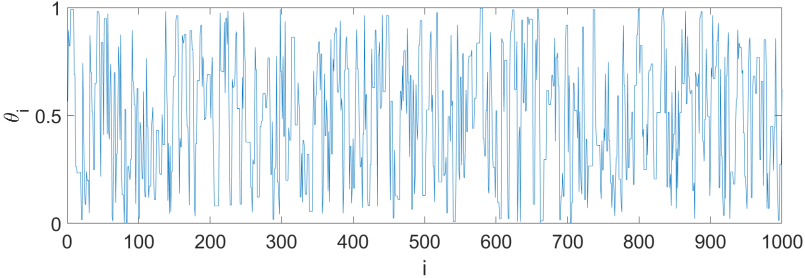
- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



- 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 $n_{draws} = 100000$.

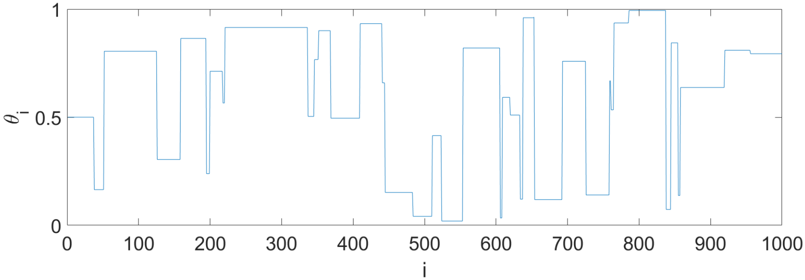
- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



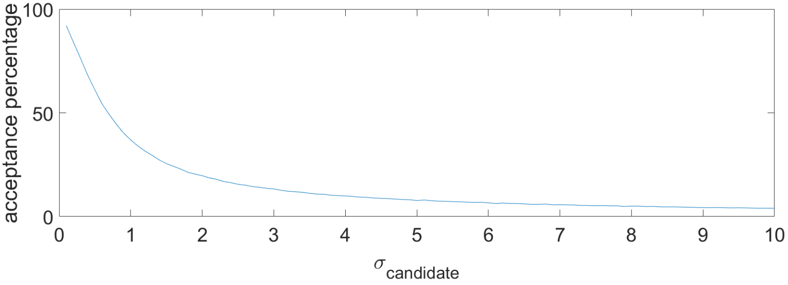
- acceptance percentage = 61%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr(θ_i, θ_{i-1}) = 0.61.

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:



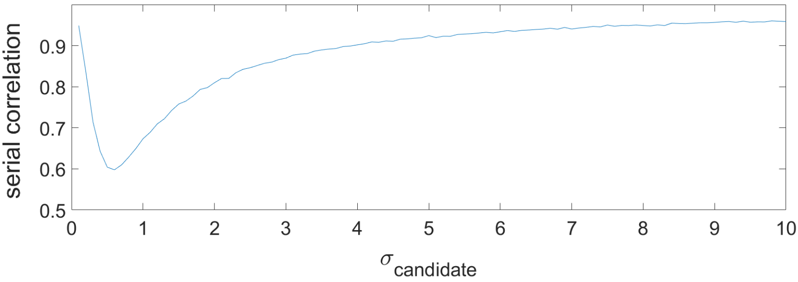
- acceptance percentage = 4%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr(θ_i, θ_{i-1}) = 0.96.



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

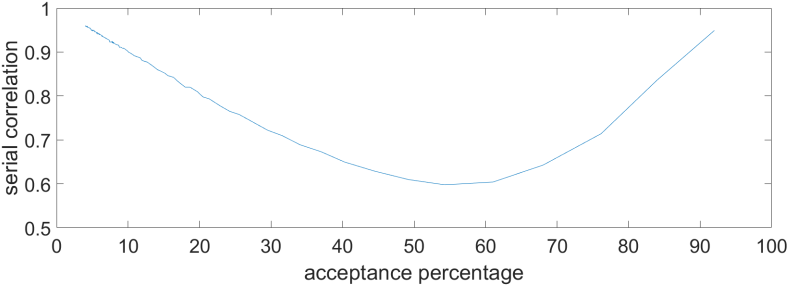
- very small candidate steps are often accepted.
- very large candidate steps are often rejected.

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.

Independence chain Metropolis-Hastings method

(independence chain: candidate draws are independent):

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

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

    with target density kernel $P(\theta)$ and candidate density $Q(\theta)$. (In Bayesian estimation, $P(\theta)$ is the posterior density kernel $P(\theta) = p(\theta)p(y|\theta)$.)

  - 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 $P(\tilde{\theta})/Q(\tilde{\theta})$ that indicates whether posterior density kernel $P(\tilde{\theta}) = p(\tilde{\theta})p(y|\tilde{\theta})$ is low or high in point $\tilde{\theta}$ (compared with candidate $Q(\tilde{\theta})$):
  - If $P(\tilde{\theta})/Q(\tilde{\theta})$ relatively very high $\Rightarrow$ large probability of accepting (and repeating!) $\tilde{\theta}$, because the value $\tilde{\theta}$ (and values close to it) are not enough simulated from candidate density $Q(.)$.
  - If $P(\tilde{\theta})/Q(\tilde{\theta})$ relatively very low $\Rightarrow$ small probability of accepting $\tilde{\theta}$, because the value $\tilde{\theta}$ (and values close to it) are simulated too often from candidate density $Q(.)$.
  - If $P(\tilde{\theta}) = 0 \Rightarrow P(\tilde{\theta})/Q(\tilde{\theta}) = 0 \Rightarrow$ no probability of accepting $\tilde{\theta}$, because this value of $\tilde{\theta}$ is impossible (for example, if $\tilde{\theta}$ lies outside the allowed range of parameter values).

Note (just like for random walk Metropolis(-Hastings) method):

- 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.
  Discard a **burn-in** of the first draws to delete the effect of initial value θ_0.

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 independence chain Metropolis-Hastings method.[3]

- Suppose we simulate the candidate draw $\tilde{\theta}$ from the uniform distribution on the interval $[0, a]$ with density

$$Q(\theta) = \begin{cases} \frac{1}{a} & 0 \leq \theta \leq a, \\ & \\ 0 & \text{else}. \end{cases}$$

[3] 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.

For candidate draw $\tilde{\theta}$ (with $\tilde{\theta} \in [0, a]$) the acceptance probability is

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

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

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

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

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

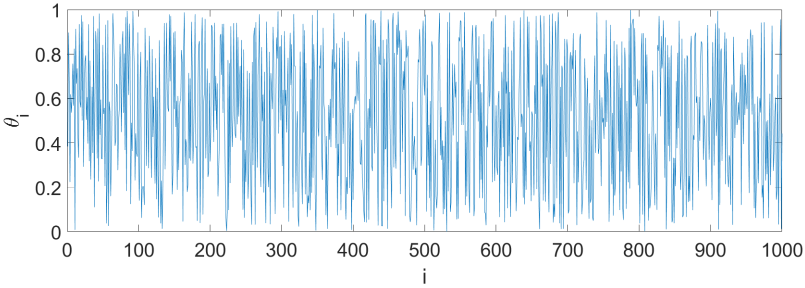
Note: the latter two cases do not occur:

- 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 $\tilde{\theta} \notin [0, 1]$.

Case with perfect candidate density (equal to target density): $a = 1$.
$\tilde{\theta} \sim U(0,1)$ and $n_{draws} = 100000$.

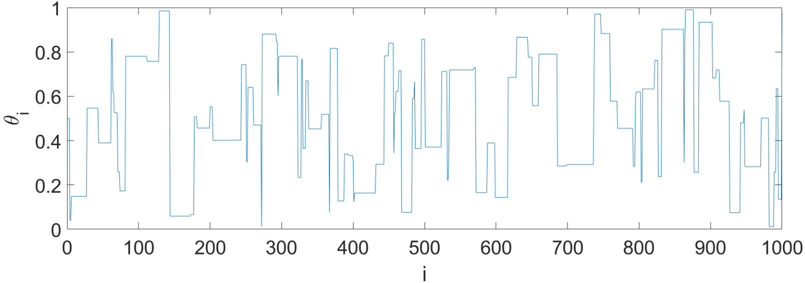
- ‘trace plot’ of first 1000 accepted (and possibly repeated) draws:



- acceptance percentage $= 100\%$.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr$(\theta_i, \theta_{i-1}) = 0.00$.

Case with too wide candidate density: $a = 10$.
$\tilde{\theta} \sim U(0, 10)$ and $n_{draws} = 100000$.

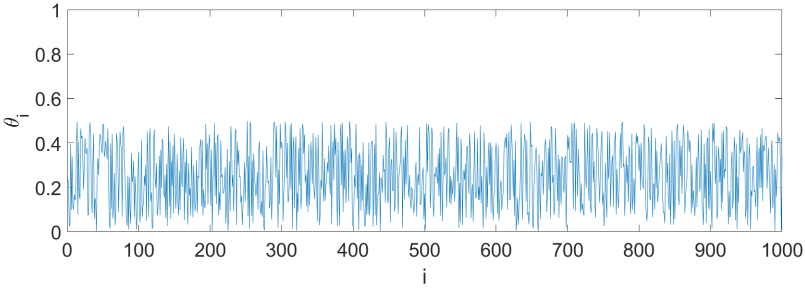
- 'trace plot' of first 1000 accepted (and possibly repeated) draws:



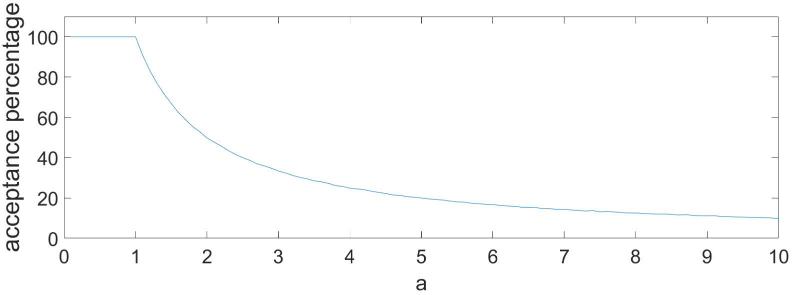
- acceptance percentage = 10%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr$(\theta_i, \theta_{i-1}) = 0.90$.

Case with too narrow candidate density: $a = 0.5$.
$\tilde{\theta} \sim U(0, 0.5)$ and $n_{draws} = 100000$.

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

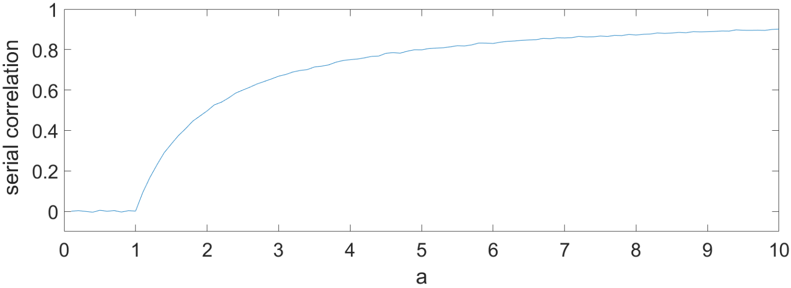


- acceptance percentage = 100%.
- first order serial correlation in sequence of accepted (and possibly repeated) draws: corr$(\theta_i, \theta_{i-1}) = 0.00$.



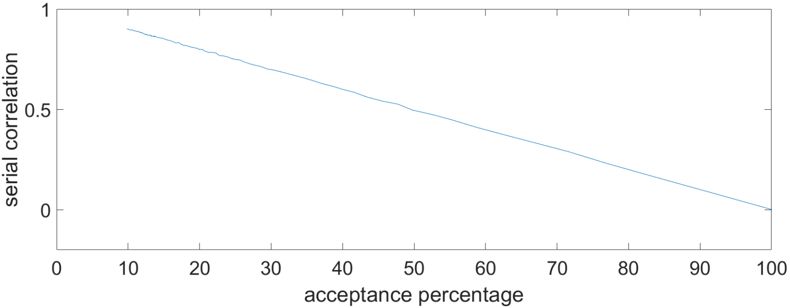
Note: for independence chain Metropolis-Hastings method we observe that

- too wide candidate density leads to poor (low) acceptance percentage.
- here the acceptance percentage does **not** give a warning signal if the candidate density is too narrow.



Note: for independence chain Metropolis-Hastings method we observe that

- too wide candidate density leads to poor (large) serial correlation corr(θ_i, θ_{i-1}) (because of many rejected candidate draws).
- here the serial correlation does **not** give a warning signal if the candidate density is too narrow.



Note: for independence chain Metropolis-Hastings method we have here:

- Best performance (i.e., lowest serial correlation corr(θ_i, θ_{i-1})) if acceptance percentage has highest value 100%.

- However, this is also reached if the candidate density is too narrow.

Check whether candidate density is too narrow **in different way**:
for example, try candidate density with larger variance and check whether estimation results substantially change.

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:

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

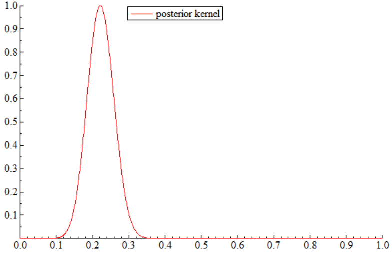
and loglikelihood:

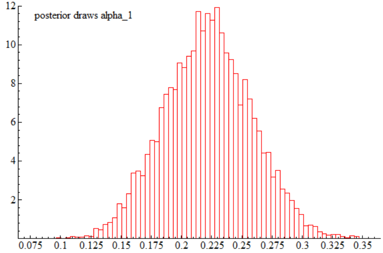
$$\ln p(y|\alpha_1) = \ln\left(\prod_{t=2}^{n} p(y_t|y_{t-1},\alpha_1)\right) = \sum_{t=2}^{n} \ln(p(y_t|y_{t-1},\alpha_1)) =$$

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

In our ARCH(1) model:

- Initial value $\theta_0 =$ ML estimator.
- Candidate distribution: uniform distribution on [0,1].
- $n_{draws} = 100100$ draws (with burn-in of 1000 draws).





Posterior mean (stdev):	0.222	(0.036).
Maximum likelihood estimator (standard error):	0.221	(0.037)

Note:

- In our ARCH(1) model we have acceptance probability:

$$\alpha = \min\left\{\frac{P(\tilde{\alpha}_1)/Q(\tilde{\alpha}_1)}{P(\alpha_{1,i-1})/Q(\alpha_{1,i-1})}, 1\right\}.$$

- We have uniform candidate density on interval [0,1] for α_1:

$$Q(\alpha_1) = \begin{cases} 1 & \text{if } 0 \le \alpha_1 \le 1 \\ 0 & \text{else} \end{cases}$$

  and each (candidate) draw lies in [0,1], so that $Q(\tilde{\alpha}_1) = 1$ and $Q(\alpha_{1,i-1}) = 1$.

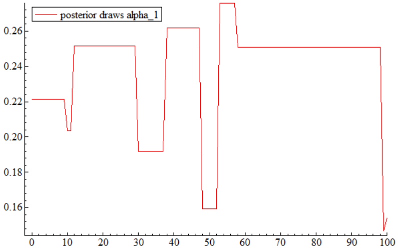
- So **<u>if we use a uniform candidate distribution</u>** we have acceptance probability:

$$\alpha = \min\left\{\frac{P(\tilde{\alpha}_1)}{P(\alpha_{1,i-1})}, 1\right\}.$$

  (The same as for the random walk Metropolis-Hastings method.)

Evaluation of quality of U(0,1) candidate distribution:

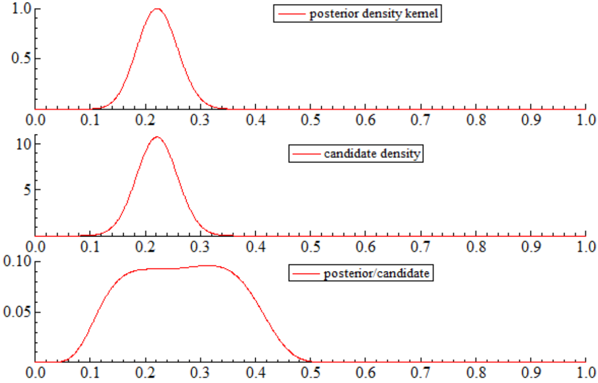
- **'trace plot'** of (accepted) draws $\theta_0, \theta_1, \theta_2, \ldots$:



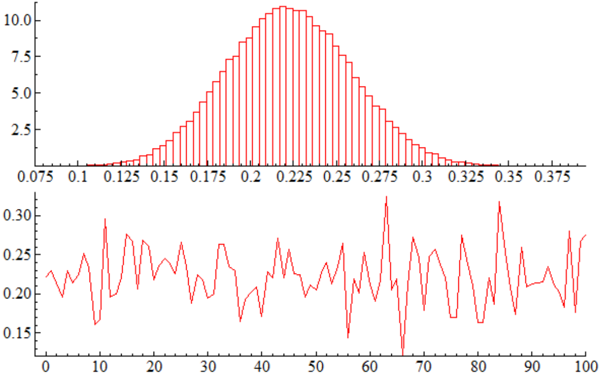
- **acceptance percentage:** what percentage of the candidate draws is accepted?
  Here: acceptance percentage = 11.5% (not so high).

- **(first order) serial correlation in sequence of (accepted) draws.**
  Here: corr$(\alpha_{1,i}, \alpha_{1,i-1}) = 0.850$ (high).

Different candidate distribution: $\tilde{\theta} \sim N(0.221, 0.037^2)$
(normal distribution around maximum likelihood estimator).

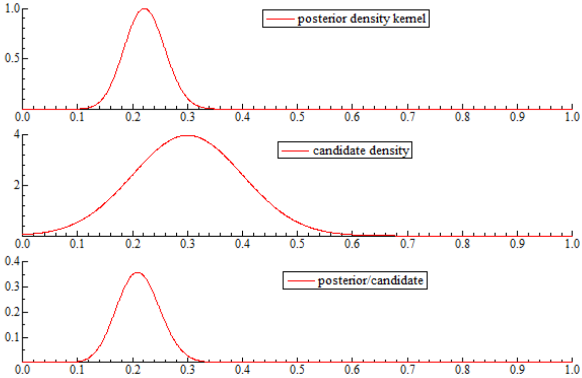


Different candidate distribution: $\tilde{\theta} \sim N(0.221, 0.037^2)$
(normal distribution around maximum likelihood estimator).
Results: histogram (top) and 'trace plot' (bottom) of accepted draws:

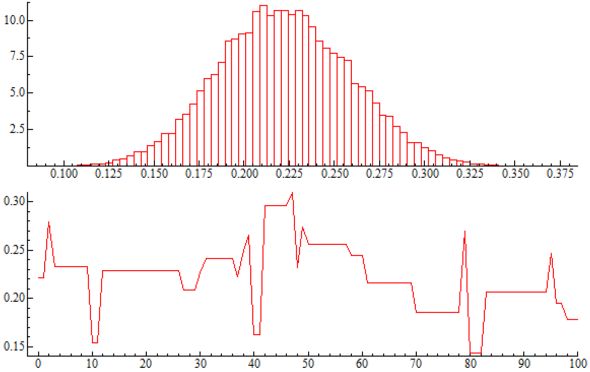


First order serial correlation in (accepted) draws: 0.021 (very low).
Acceptance percentage: 98.5% (very high).
Estimated posterior mean (stdev): 0.223 (0.036)

Different candidate distribution: $\tilde{\theta} \sim N(0.3, 0.1^2)$ (normal distribution with too high mean and large variance).

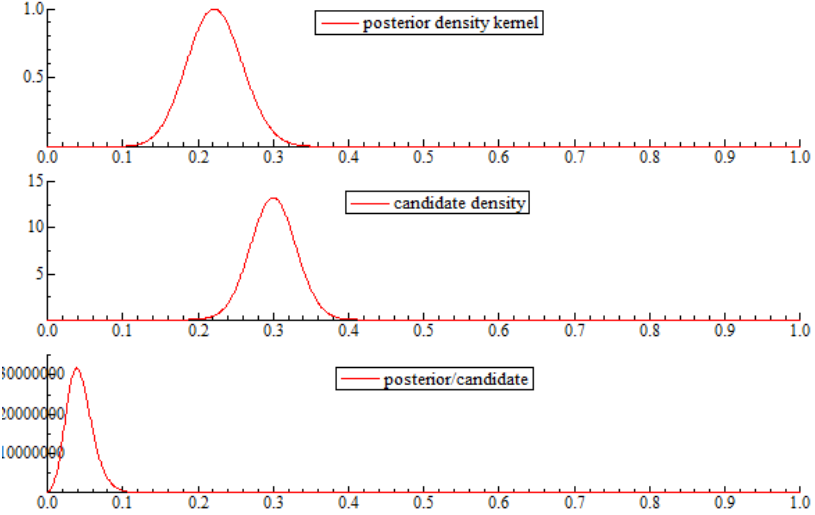


Different candidate distribution: $\tilde{\theta} \sim N(0.3, 0.1^2)$
(normal distribution with too high mean and large variance).
Results: histogram (top) and 'trace plot' (bottom) of accepted draws:

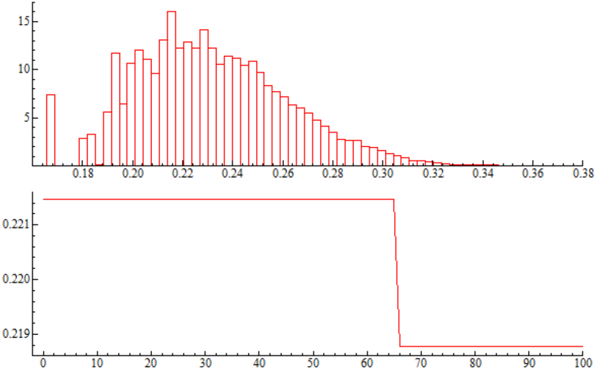


First order serial correlation in (accepted) draws: 0.618.
Acceptance percentage: 32.3%.
Estimated posterior mean (stdev): 0.222 (0.037)

Different candidate distribution: $\tilde{\theta} \sim N(0.3, 0.03^2)$
(normal distribution with too high mean and small variance).



Different candidate distribution: $\tilde{\theta} \sim N(0.3, 0.03^2)$
(normal distribution with too high mean and small variance).
Results: histogram (top) and 'trace plot' (bottom) of accepted draws:



First order serial correlation in (accepted) draws: 0.953.
Acceptance percentage: 11.9%.
Estimated posterior mean (stdev): 0.232 (0.030)

Note:

- Independence chain Metropolis-Hastings method **works fine** if candidate distribution is close to posterior distribution (where candidate distribution 'covers' the whole posterior distribution)

- Independence chain Metropolis-Hastings method **works (but with many rejected candidate draws)** if candidate distribution 'covers' the whole posterior distribution, but is not so close to the posterior distribution.

- Independence chain Metropolis-Hastings method **fails** if candidate distribution does not 'cover' the whole posterior distribution.

How to make it more probable that the candidate 'covers' the posterior distribution?

- Use Student-t candidate distribution (with fatter tails than normal distribution).
- Adapt candidate distribution: run independence chain MH method a second time, now with mean and variance of candidate distribution equal to estimated posterior mean and posterior variance (multiplied with factor ≥ 1) from first run. (This can be repeated several times.)
- Alternatively, use random walk Metropolis(-Hastings) method.

Note:

- Acceptance/rejection/repetition corrects for difference between posterior density with kernel $P(\theta)$ and candidate density $Q(\theta)$.

- Independence chain Metropolis-Hastings method is **not** the same as **acceptance-rejection method**, where we accept with probability

$$\frac{P(\tilde{\theta})/Q(\tilde{\theta})}{M} \qquad M = \max_{\theta} P(\theta)/Q(\theta),$$

  and where rejection means that we try again (without repeating the previous value).

  Advantage of acceptance-rejection method: it yields independent draws.

  Disadvantages of acceptance-rejection method:
  the acceptance percentage ($\approx \frac{1}{M}$) may be very small, especially if the dimension of θ is large, and $\max_{\theta} P(\theta)/Q(\theta)$ needs to be computed.