WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Bayesian Computation with R

Rainer Hirk
(Laura Vana, Bettina Grün, Paul Hofmarcher, Gregor Kastner)
WS 2021/22

Overview

- Lecture:
  - Bayes approach
  - Bayesian computation
  - Available tools in R
  - Example: stochastic volatility model
- Exercises
- Projects

Deliveries

- Exercises:
  - In groups of 3-4 students;
  - Solutions handed in by e-mail to `rainer.hirk@wu.ac.at` in a `.pdf`-file together with the original `.Rnw`-file;
  - Deadline: **2021-11-2**.
- Projects:
  - In groups of 3–4 students;
  - Data analysis using Bayesian methods in JAGS and frequentist estimation and comparison between the two approaches;
  - Documentation of the analysis consisting of
    (a) Problem description;
    (b) Model specification;
    (c) Model fitting: estimation and convergence diagnostics;
    (d) Interpretation (where available, refer also to cited material).
  - Presentation: **2021-12-06** starting from 09:00.
  - Report deadline: **2021-12-13**.

Material

- Lecture slides
- Further reading:
  - Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer.
  - Albert, J. (2007). *Bayesian Computation with R*. Springer.
  - Marin, J. M. and Robert, C. (2014). *Bayesian Essentials with R*. Springer. (R package **bayess**).

Software tools

- JAGS: Just Another Gibbs Sampler
  - Available from sourceforge: `https://sourceforge.net/projects/mcmc-jags/`
  - Current version: 4.3.0
  - Source code and binaries for Windows and Mac available
- R package **rjags** on CRAN:
  - Bayesian graphical models using MCMC with the JAGS library
  - Compatible version to JAGS: 4.12
  - `install.packages("rjags")`
- R package **coda** on CRAN:
  - Output analysis and diagnostics for MCMC
  - `install.packages("coda")`
- Software documentation: Plummer, M. (2015) JAGS Version 4.0.0 user manual: `https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/jags_user_manual.pdf`.
- Alternatively, R package **rstan** on CRAN: `install.packages("rstan")`)
- Stan software documentation:`http://www.uvm.edu/~bbeckage/Teaching/DataAnalysis/Manuals/stan-reference-2.8.0.pdf`

Frequentist vs. Bayesian

What is the difference between classical frequentist and Bayesian statistics?

- To a frequentist, unknown model parameters are **fixed** and unknown, and only estimable by replications of data from some experiment.
- A Bayesian thinks of parameters as **random**, and thus having distributions for the parameters of interest. So a Bayesian can think about unknown parameters $\boldsymbol{\theta}$ for which no reliable frequentist experiment exists.

Updating beliefs I

Bayes' rule

Event B can be observed directly, while event A cannot be observed directly. Use the information about the observed event B to adjust the probability of event A:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} =$$
$$= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|A^C)\Pr(A^C)},$$

Updating beliefs II

Bayes' theorem

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y},\boldsymbol{\theta})}{\int p(\mathbf{y},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

$$p(\text{latent}|\text{observed}) \propto f(\text{observed}|\text{latent})\pi(\text{latent})$$

$$\text{posterior density} \propto \text{likelihood} \times \text{prior density}$$

Bayesian approach

1. Specify a sampling distribution $f(\mathbf{y}|\boldsymbol{\theta})$ of the data $\mathbf{y}$ in terms of the unknown parameters $\boldsymbol{\theta}$ (**likelihood function**).
2. Specify a **prior distribution** $\pi(\boldsymbol{\theta})$ which is usually chosen to be "non-informative" compared to the likelihood function.
3. Use Bayes' theorem to learn about $\boldsymbol{\theta}$ given the observed data $\Rightarrow$ derive the **posterior distribution** $p(\boldsymbol{\theta}|\mathbf{y})$.
4. Inference is based on summaries of the posterior distribution.

Prior distributions I

- **Elicited priors:** based on expert knowledge.
- **Conjugate priors:** lead to a posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ belonging to the same distributional family as the prior.
  Examples:
  - Beta prior for the success probability parameter of a binomial likelihood.
  - Gamma prior for the rate parameter of a Poisson likelihood.
  - Normal prior for the mean parameter of a normal likelihood with known variance.
  - Gamma prior for the inverse variance (aka precision) of a normal likelihood with known mean.

  See `http://en.wikipedia.org/wiki/Conjugate_prior`.

Prior distributions II

- **Non-informative priors:** do not favor any values of $\boldsymbol{\theta}$ if no a-priori information is available.
  Examples:
  - Uniform distribution (aka **flat** prior):
    - suitable if the parameter space is discrete and finite.
    - leads to **improper** priors for continuous and infinite parameter space.
    - is not (always) invariant under reparameterization.
  - Jeffrey's prior: invariant under reparameterization:

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2} \qquad I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right],$$

    where $I(\boldsymbol{\theta})$ is the Fisher information matrix.
- **Note:** Conjugate priors can be non-informative by choosing the appropriate hyperparameters.

Parameter estimation

- **Point estimation:** given a prior ditribution, what is the best estimator of $\boldsymbol{\theta}$? Each of these estimators may be derived as an optimal estimators with respect to a certain loss function $R(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta})$, which quantifies the loss made when estimating a parameter $\boldsymbol{\theta}$ by an estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$.
  - Posterior mode (aka **generalized ML estimate**) is optimal with respect to the 0/1 loss:

$$R(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) = \begin{cases} 0, & \hat{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{\theta} \\ 1, & \hat{\boldsymbol{\theta}}(\mathbf{y}) \neq \boldsymbol{\theta} \end{cases}$$

  - Posterior mean is optimal with respect to the quadratic loss function $R(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})$.
  - In a single parameter problem, the posterior median is optimal for the absolute loss function $R(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) = |\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}|$.
  - ...

Measuring uncertainty

- **Interval estimation:**

Definition

A $100 \times (1-\alpha)\%$ **credible region** for $\boldsymbol{\theta}$ is a subset $C_{(1-\alpha)}$ of Ω such that

$$1-\alpha = \int_{C_{(1-\alpha)}} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

The probability that $\boldsymbol{\theta}$ lies in $C_{(1-\alpha)}$ given the observed data $\mathbf{y}$ is $(1-\alpha)$.

Examples: quantile based - credible region, **highest posterior density (HPD) region** (region which, for a given α, occupies the smallest possible volume in the parameter space).

Bayesian hypothesis testing I

- **Classical hypothesis testing:**
  - Likelihood ratio test, p-values ...
  - After determining an appropriate test statistic $T(y)$ the p-value is the probability of observing a more extreme value under the null.
  - H_0 must be a simplification of (nested in) H_A.
  - We can only offer evidence against the null hypothesis.
- **Bayesian hypothesis testing:** use **Bayes factors**!
  - It requires some prior knowledge.
  - Based on the data $\mathbf{y}$, one applies Bayes' theorem and computes the posterior probability that the first hypothesis is correct.

Bayesian hypothesis testing II

- **Bayes factors:**

Definition (Bayes factor)

The Bayes factor BF is the ratio of the posterior odds of hypothesis H_1 to the prior odds of H_1:

$$BF = \frac{\Pr(H_1|\mathbf{y})/\Pr(H_2|\mathbf{y})}{\Pr(H_1)/\Pr(H_2)}$$
$$= \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_2)} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta_1}, H_1)\pi(\boldsymbol{\theta_1}|H_1)d\boldsymbol{\theta_1}}{\int f(\mathbf{y}|\boldsymbol{\theta_2}, H_2)\pi(\boldsymbol{\theta_2}|H_2)d\boldsymbol{\theta_2}}$$

i.e., the ratio of the observed marginal densities for the two models.

Bayesian hypothesis testing III

- BF captures the change in the odds in favor of hypothesis H_1 as we move from prior to posterior.
- Jeffrey's scale for interpretation:

BF	Strength of evidence
< 1	Negative (support of H_2)
1–3	Barely worth mentioning
3–10	Substantial
10–30	Strong
30–100	Very strong
> 100	Decisive

A fun reference: Lavine, M (1999). What is Bayesian Statistics and Why Everything Else is Wrong.
The Journal of Undergraduate Mathematics and Its Applications 20, 165–174, `www.math.umass.edu/~lavine/whatisbayes.pdf`

Example: Sleep study

Description: A researcher is interested in the sleeping habits of college students. 27 students are interviewed and in this group 11 record they slept more than 8 hours the previous night.

1. What is the proportion θ of students who sleep more than 8 hours per night?
2. Is the majority of college students getting enough sleep?

Bayesian analysis: we need two components: **likelihood** and **prior**!

Example: Sleep study – likelihood

- We assume that the 27 interviewed students are independent and that the probability θ of sleeping more than 8 hours per night is constant over the students.
- Their answers form a sequence of Bernoulli trials.
- Let Y denote the number of students that recorded sleeping at least 8 hours the previous night.

$$Y|\theta \sim \text{Bin}(n, \theta),$$

which, for $n = 27$ is equivalent to

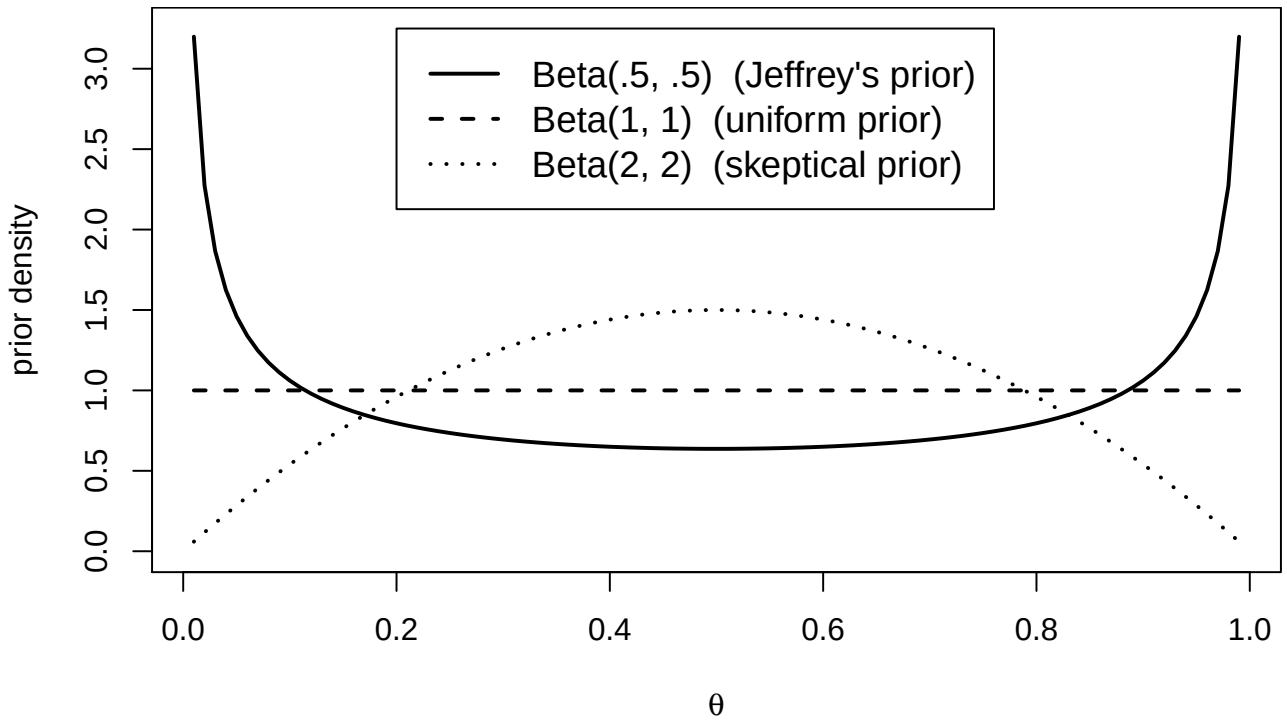
$$f(y|\theta) = \binom{27}{y} \theta^y (1-\theta)^{27-y}.$$

- **Q:** what is the MLE of θ?

Example: Sleep study – prior

Conjugate prior: The Beta distribution is a conjugate family for the binomial distribution.

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$



Example: Sleep study – posterior I

Due to conjugacy, the posterior distribution for θ is

$$\begin{aligned} p(\theta|y) &\propto f(y|\theta)\pi(\theta) \propto \theta^{y+\alpha-1}(1-\theta)^{27-y+\beta-1} \\ &\propto \text{Beta}(y+\alpha, 27-y+\beta). \end{aligned}$$

For Beta(α, β), the expected value is $\alpha/(\alpha+\beta)$. Hence,

$$\mathbb{E}(\theta|y) = \frac{y+\alpha}{y+\alpha+27-y+\beta}.$$

Assume the uniform prior Beta$(1,1)$. The expected value is 0.4.

Example: Sleep study – posterior II

- Compute 95% **credible intervals** for θ:

```
> c(0, round(qbeta(0.95, 12, 17) ,digits = 3))
[1] 0.000 0.565
> round(qbeta(c(0.025, 0.975), 12, 17), digits = 3)
[1] 0.245 0.594
> c(round(qbeta(0.05, 12, 17) ,digits = 3), 1)
[1] 0.269 1.000
```

- **HPD region**? (`coda::HPDinterval`)

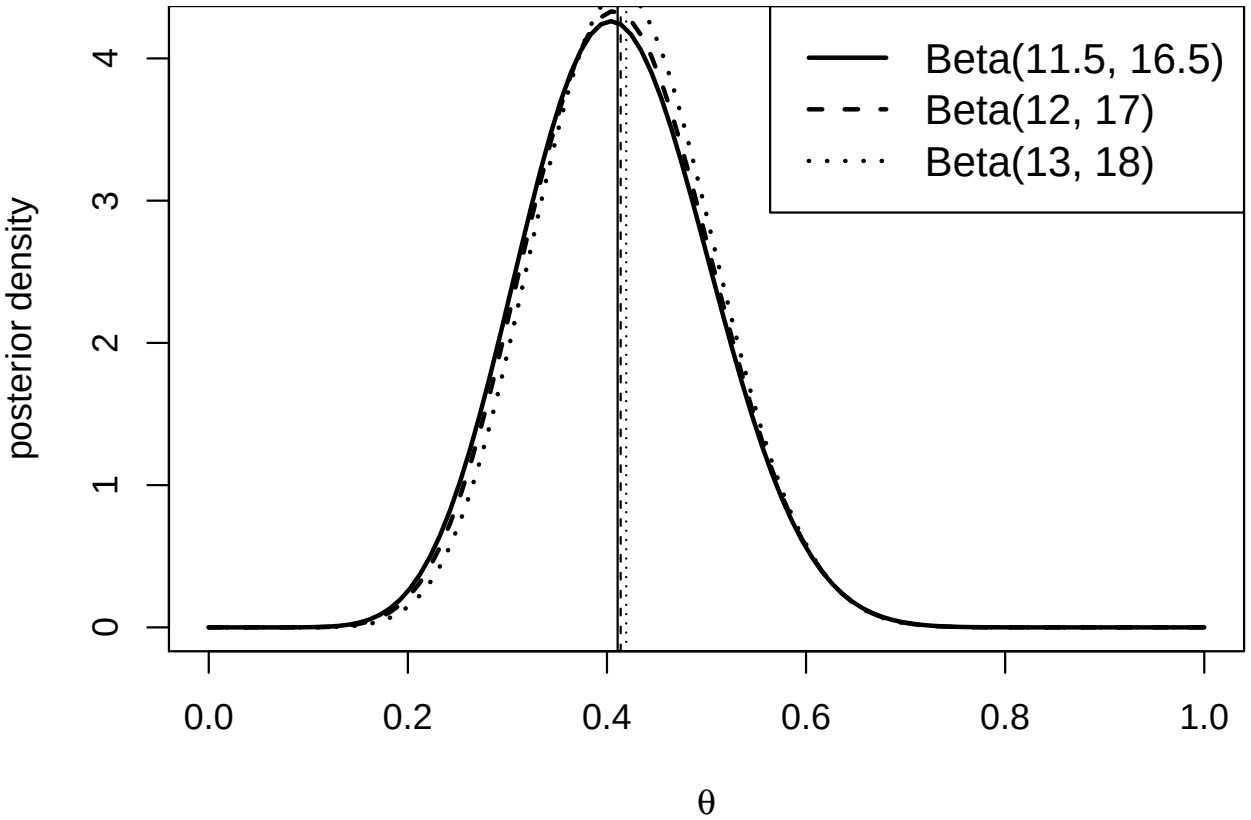
```
      lower  upper
var1 0.2357 0.5839
attr(,"Probability")
[1] 0.95
```

- Frequentist 95% confidence interval:

$$\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

$$0.222 \leq \theta \leq 0.593.$$

Example: Sleep study – posterior III



Example: Sleep study – hypothesis testing I

- We return to the researcher's question whether the majority of college students are getting enough sleep and compare the hypotheses:
  $H_1 : \theta \geq 0.5 \qquad H_2 : \theta < 0.5.$
- Using the uniform prior Beta$(1, 1)$, the prior probability $\Pr(\theta \geq 0.5)$ of H_1 is:

```
> (prior.p1 <- round(pbeta(0.5, 1, 1,
+                          lower.tail = FALSE), digits = 3))
[1] 0.5
```

- From the posterior we compute the posterior probabilty $\Pr(\theta \geq 0.5|y)$ of H_1:

```
> (post.p1 <- round(pbeta(0.5, 12, 17,
+                         lower.tail = FALSE), digits = 3))
[1] 0.172
```

Example: Sleep study – hypothesis testing II

- The Bayes factor is then given by

$$BF = \frac{0.172/(1-0.172)}{0.5/(1-0.5)} = \frac{0.172/0.828}{0.5/0.5} = 0.2.$$

and implies a negative preference for H_1 (support of H_2).

Bayesian computation

- For many advanced problems, the posterior distribution is rather complex and does not belong to a well-known distribution family.
- For such problems computational aspects form a central part of Bayesian statistical modeling.
- Approximate methods:
  - Asymptotic methods
  - Noniterative Monte Carlo methods
  - Markov chain Monte Carlo methods

Normal approximation

Theorem (Bayesian Central Limit Theorem)

Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} f_i(y_i|\boldsymbol{\theta})$ *and that the prior* $\pi(\boldsymbol{\theta})$ *and the likelihood* $f(\mathbf{y}|\boldsymbol{\theta})$ *are positive and twice differentiable near* $\hat{\boldsymbol{\theta}}^{\pi}$, *the posterior mode of* $\boldsymbol{\theta}$.
Then for large n

$$p(\boldsymbol{\theta}|\mathbf{y}) \dot{\sim} N(\hat{\boldsymbol{\theta}}^{\pi}, [I^{\pi}(\mathbf{y})]^{-1}),$$

where $[I^{\pi}(\mathbf{y})]^{-1}$ *is the "generalized" observed Fisher information matrix for* $\boldsymbol{\theta}$ *with*

$$I^{\pi}_{ij}(\mathbf{y}) = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{\pi}}$$

When n is large, $f(\mathbf{y}|\boldsymbol{\theta})$ will be quite peaked relative to $\pi(\boldsymbol{\theta})$, and so $p(\boldsymbol{\theta}|\mathbf{y})$ will be approximately normal.

Example cont.: Sleep study I

Using a flat prior on θ, i.e., $\pi(\theta) \propto 1$, we have

$$\ell(\theta) = \log(f(y|\theta)\pi(\theta)) = y \log \theta + (n - y) \log(1 - \theta) + C.$$

The first derivative is given by

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{n - y}{1 - \theta}.$$

Equating to zero and solving for θ gives the posterior mode by

$$\hat{\theta}^{\pi} = \frac{y}{n}.$$

The second derivative is given by

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{n - y}{(1 - \theta)^2}.$$

Example cont.: Sleep study II

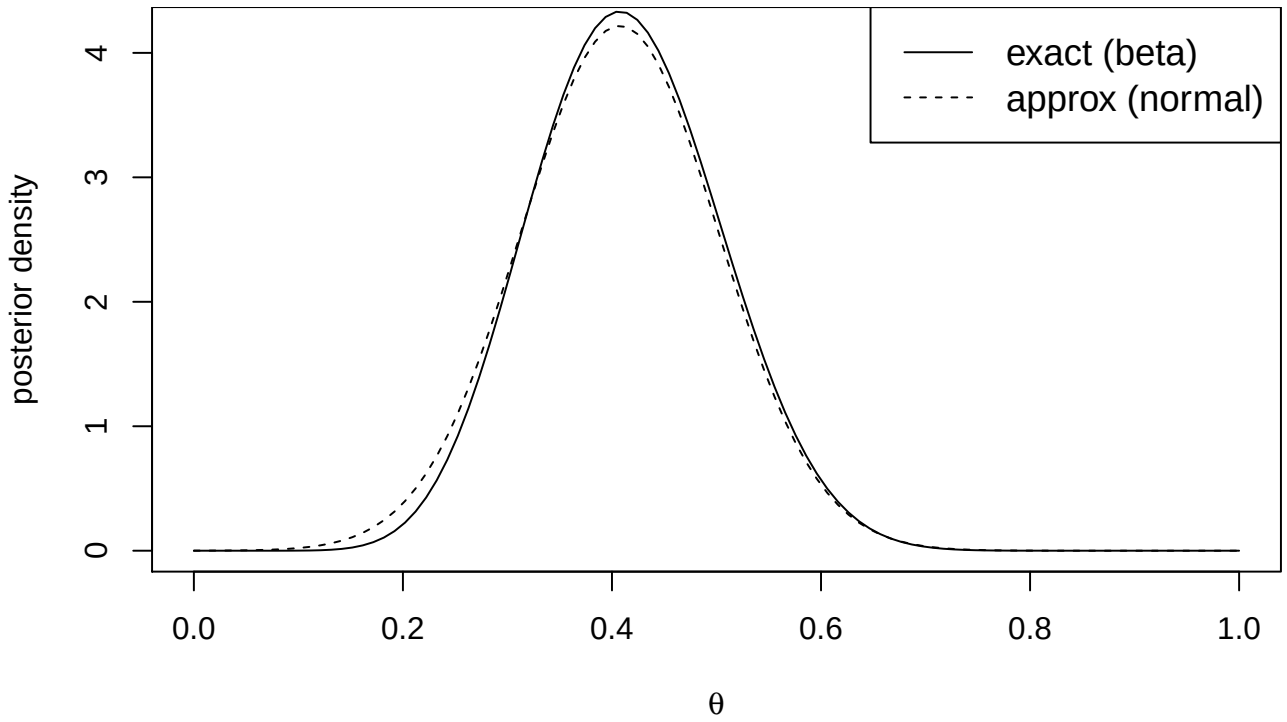
Evaluating at the estimate $\hat{\theta}^{\pi}$ gives

$$\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}^{\pi}} = -\frac{n}{\hat{\theta}^{\pi}(1-\hat{\theta}^{\pi})}.$$

Thus the posterior can be approximated by

$$p(\theta|y) \dot{\sim} N(\hat{\theta}^{\pi}, \frac{\hat{\theta}^{\pi}(1-\hat{\theta}^{\pi})}{n}).$$

Example cont.: Sleep study III



Similar modes, but different tail behavior.

Asymptotic methods

- **Advantages:**
  - Deterministic, noniterative algorithm.
  - Use differentiation instead of integration.
  - Facilitates studies of Bayesian robustness.
- **Disadvantages:**
  - Requires well-parameterized, unimodal posterior.
  - θ must be of at most moderate dimension.
  - n must be large, but is beyond our control.

Noniterative Monte Carlo methods

- Direct sampling
- Indirect methods (e.g., importance sampling, rejection sampling)

Monte Carlo method and direct sampling

Remember the most basic definition of Monte Carlo integration:

- Suppose $\theta \sim f(\theta)$ and we want to compute

$$\gamma := \mathbb{E}[g(\theta)] = \int g(\theta) f(\theta) d\theta.$$

- Then if $\theta_1, \ldots, \theta_n \overset{iid}{\sim} f(\theta)$, we have

$$\hat{\gamma}_n = \frac{1}{n} \sum_{j=1}^{n} g(\theta_j),$$

which converges to $\mathbb{E}[g(\theta)]$ with probability 1 as $n \to \infty$ and

$$\mathbb{V}(\hat{\gamma}_n) = \frac{\mathbb{V}(g(\theta))}{n}$$

Direct sampling

- Using Monte Carlo integration, the computation of posterior expectations requires only a sample size of n from the posterior.
- The joint posterior density for the parameters is analytically converted into a product of conditional and marginal densities from which draws can be made yielding a draw from the joint density.
- Assume we want to estimate a vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ of parameters:

$$p(\theta_1, \theta_2|\mathbf{y}) = p(\theta_1|\mathbf{y})p(\theta_2|\theta_1, \mathbf{y}).$$

  Then θ_1 can be drawn from $p(\theta_1|\mathbf{y})$ and substituted in $p(\theta_2|\theta_1, \mathbf{y})$ and a draw θ_2 is made from $p(\theta_2|\theta_1, \mathbf{y})$.
- Repeating this procedure many times provides a large sample from the joint density from which moments, intervals, etc., can be computed.

Importance sampling

- Often we are interested in the expectation of a function $h(\boldsymbol{\theta})$ with respect to the posterior density, Suppose we wish to approximate

$$\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{y}] = \int h(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int h(\boldsymbol{\theta}) \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

- Suppose we can roughly approximate the normalized likelihood times prior, $cf(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, by some importance density $g(\boldsymbol{\theta})$ from which we can easily sample.
- Then defining the weight function $w(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/g(\boldsymbol{\theta})$,

$$\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int h(\boldsymbol{\theta}) w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}} \approx \frac{\frac{1}{n}\sum_{j=1}^{n} h(\theta_j) w(\theta_j)}{\frac{1}{n}\sum_{j=1}^{n} w(\theta_j)},$$

where $\theta_j \overset{iid}{\sim} g(\boldsymbol{\theta})$.

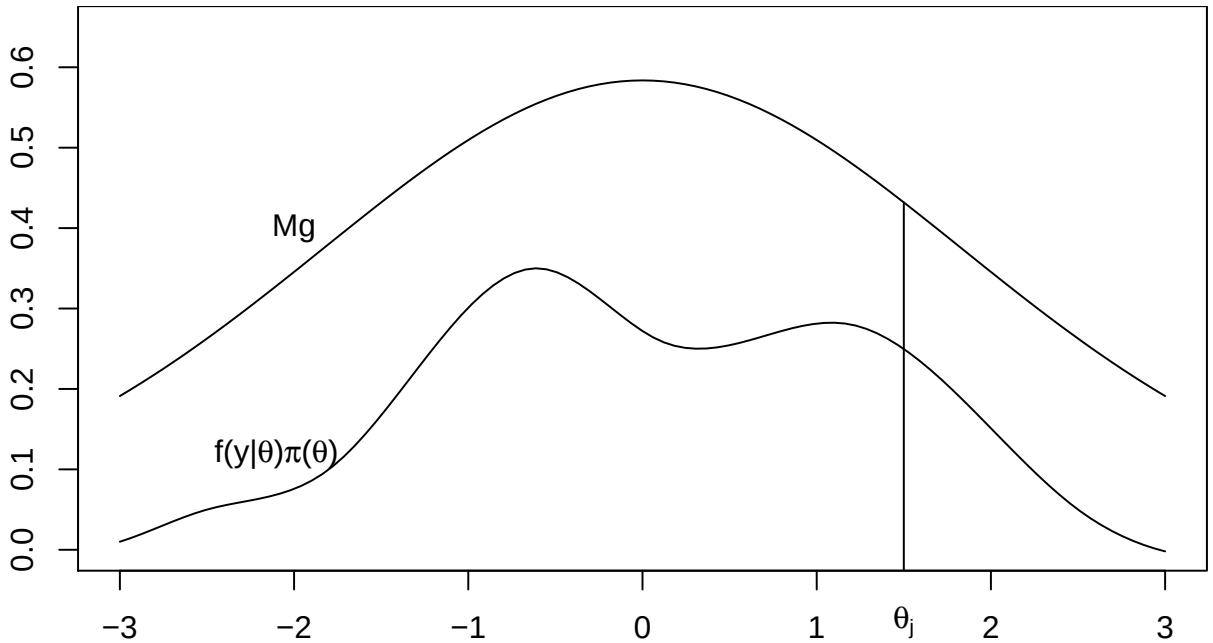
Rejection sampling I

- Instead of trying to approximate the posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

  we try to find a majorizing function.
- Suppose there exists a constant $M > 0$ and a smooth density $g(\boldsymbol{\theta})$, called the envelope function, such that $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) < Mg(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.
- The algorithm proceeds as follows:
  (i) Generate $\theta_j \sim g(\boldsymbol{\theta})$.
  (ii) Generate $U \sim \text{Unif}(0, 1)$.
  (iii) If $MUg(\theta_j) < f(\mathbf{y}|\theta_j)\pi(\theta_j)$, accept θ_j. Otherwise reject θ_j.
  (iv) Return to step (i) and repeat, until the desired sample size is obtained.
- The final sample consists of random draws from $p(\boldsymbol{\theta}|\mathbf{y})$.

Rejection sampling II



- Need to choose M as small as possible (efficiency), and avoid "envelope violations"!

Markov chain Monte Carlo methods I

- Such iterative MC methods are useful when it is difficult or impossible to find a feasible importance or envelope density.
- Complex models have intractable posteriors.
- Combine **Markov chains** and **Monte Carlo integration**.
- **Idea**: to obtain samples from a distribution without this distribution being explicitly available, i.e., a sample from $p(\boldsymbol{\theta}|\mathbf{y})$ is obtained indirectly by generating a realization of a Markov chain $\boldsymbol{\theta}^{(m)}$, $m = 1, 2, \ldots,$ based on some starting value $\boldsymbol{\theta}^{(0)}$.
- **Aim:** constructing an irreducible, aperiodic Markov chain with the posterior as stationary distribution in order to acquire samples from that distribution. Plug sampled values into the Monte Carlo integration.

Markov chains

- A Markov chain $\boldsymbol{\theta}^{(m)}$ is a random variable, with the conditional distribution depending on the past states of the Markov chain.
- The key quantity for characterizing the probabilistic behavior of the Markov chain is the transition kernel $k(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old})$, which is the density of the conditional probability distribution of $\boldsymbol{\theta}^{(m)}$ given $\boldsymbol{\theta}^{(m-1)} = \boldsymbol{\theta}^{old}$:

$$\boldsymbol{\theta}^{(m)}|(\boldsymbol{\theta}^{(m-1)} = \boldsymbol{\theta}^{old}) \sim k(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old}).$$

- Under certain regularity conditions the unconditional distribution converges to an invariant distribution. For the invariant distribution to be the posterior, the transition kernel $k(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old})$ must fulfill the integral equation:

$$p(\boldsymbol{\theta}^{new}|\mathbf{y}) = \int k(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old})p(\boldsymbol{\theta}^{old}|\mathbf{y})d\boldsymbol{\theta}^{(old)}$$

Markov chain Monte Carlo methods

There are many ways of constructing a Markov chain with the stationary distribution being equal to a specific posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. The most widely used are

- Gibbs sampler - most commonly used,
- Metropolis-Hastings algorithm - most universal sampling scheme

Classical Monte Carlo integration uses a sample of independent draws from the density $p(\boldsymbol{\theta}|\mathbf{y})$. In MCMC we have dependent draws, hence **performance evaluation** is needed:

- Convergence monitoring and diagnostics
- Variance estimation

Gibbs sampling I

- Suppose the joint distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is uniquely determined by the full conditional distributions, $\{p_i(\theta_i|\boldsymbol{\theta}_{j\neq i}), i = 1, \ldots, K\}$.
- Given an arbitrary set of starting values $\{\theta_1^{(0)}, \ldots, \theta_K^{(0)}\}$,

$$\begin{aligned}
&\text{Draw } \theta_1^{(1)} \sim p_1(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}),\\
&\text{Draw } \theta_2^{(1)} \sim p_2(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}),\\
&\qquad\vdots\\
&\text{Draw } \theta_K^{(1)} \sim p_K(\theta_K|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{K-1}^{(1)}).
\end{aligned}$$

- Under mild conditions,

$$(\theta_1^{(t)}, \ldots, \theta_K^{(t)}) \xrightarrow{d} (\theta_1, \ldots, \theta_K) \sim p \quad \text{as } t \to \infty.$$

Gibbs sampling II

- For T sufficiently large (say, bigger than t_0), $\{\boldsymbol{\theta}^{(t)}\}_{t=t_0+1}^{T}$ is a (correlated) sample from the true posterior.
- We might use a sample mean to estimate the posterior mean

$$\mathbb{E}(\theta_i|\mathbf{y}) \approx \frac{1}{T - t_o} \sum_{t=t_0+1}^{T} \theta_i^{(t)}.$$

- The time from $t = 0$ to $t = t_0$ is commonly known as the **burn-in** period.
- We may also run m parallel Gibbs sampling chains and obtain

$$\mathbb{E}(\theta_i|\mathbf{y}) \approx \frac{1}{m(T - t_o)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} \theta_i^{(j,t)},$$

where the index j indicates chain number.

Metropolis Hastings algorithm I

- What happens if the full conditional $\{p_i(\theta_i|\boldsymbol{\theta}_{j\neq i})\}$ is not available in closed form?
- Typically, the normalizing constant (denominator in Bayes' theorem) is hard to compute.
- Suppose the true joint posterior for $\boldsymbol{\theta}$ has **unnormalized** density $p(\boldsymbol{\theta})$.
- Choose a proposal density (also called jumping or candidate density) $q(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old})$ that is a valid density function for every possible value of the conditioning variable $\boldsymbol{\theta}^{old}$.

Metropolis Hastings algorithm II

- Given a starting value $\boldsymbol{\theta}^{(0)}$ at iteration $t = 0$, the algorithm proceeds as follows.
  For $t = 1, \ldots, T$ repeat:
  1. Propose $\boldsymbol{\theta}^{new}$ for $\boldsymbol{\theta}^{(t)}$ from $q(\cdot|\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{(t-1)})$.
  2. Compute the ratio

$$r = \frac{p(\boldsymbol{\theta}^{new})q(\boldsymbol{\theta}^{old}|\boldsymbol{\theta}^{new})}{p(\boldsymbol{\theta}^{old})q(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{old})}.$$

  3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{new}$;
     If $r < 1$, set $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^{new} & \text{with probability } r \\ \boldsymbol{\theta}^{old} & \text{with probability } 1 - r \end{cases}$.
- Then a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.

Metropolis Hastings algorithm III

- How to choose the proposal density?
- The random walk proposal density: the usual approach (after $\boldsymbol{\theta}$ has been transformed to have support $\mathbb{R}^K$, if necessary) is to set

$$\boldsymbol{\theta}^{new} \sim N(\boldsymbol{\theta}^{old}, \tilde{\Sigma}).$$

The scale of a random walk proposal density has to be chosen with some care:

- Very small $\tilde{\Sigma}$ will generate small steps $\boldsymbol{\theta}^{new} - \boldsymbol{\theta}^{old}$ with generally high acceptance rates, but also high auto-correlation.
- Large $\tilde{\Sigma}$ will generate large moves $\boldsymbol{\theta}^{new} - \boldsymbol{\theta}^{old}$ and will often propose a value far out in the tails of the distribution, giving generally small acceptance rates.

Convergence assessment

When is it safe to stop and summarize MCMC output?

- We would like to ensure that $\int |\hat{p}_t(\boldsymbol{\theta}) - p(\boldsymbol{\theta})| d\boldsymbol{\theta} < \epsilon$.
  But all we can hope to see is $\int |\hat{p}_t(\boldsymbol{\theta}) - \hat{p}_{t+k}(\boldsymbol{\theta})| d\boldsymbol{\theta} < \epsilon$.
- One can never "prove" convergence of a MCMC algorithm using only a finite realization from the chain.
- A slowly converging sampler may be indistinguishable from one that will never converge (e.g., due to nonidentifiability)!
- Does the eventual mixing of "initially overdispersed" parallel sampling chains provide worthwhile information on convergence?
- YES! Poor mixing of parallel chains can help discover extreme forms of **nonconvergence**.

Convergence diagnostics

Various summaries of MCMC output, such as

- **Sample auto-correlations** in one or more chains:
  - Close to 0 indicates near-independence $\rightarrow$ Chain should quickly traverse the entire parameter space.
  - Close to 1 indicates that the sampler is "stuck".
- Diagnostic tests requiring several chains include for example Gelman & Rubin's shrink factor.
- Other tests for convergence requiring only one chain include among others Heidelberger & Welch's, Raftery & Lewis's and Geweke's diagnostics.

(Possible) Convergence diagnostics strategy

- Run a few (3 to 5) parallel chains, with starting points believed to be overdispersed.
  - E.g., covering ± 3 prior standard deviations from the prior mean.
- Overlay the resulting sample traces for the parameters or a representative subset (if there are many parameters or a hierarchical model is fitted).
- Annotate each plot with **lag 1 sample autocorrelations** and perhaps Gelman & Rubin's diagnostics.
- Look at convergence diagnostic tests output.
- Investigate bivariate plots and **crosscorrelations** among parameters suspected of being confounded, just as one might do regarding collinearity in linear regression.

Variance estimation I

How good is our MCMC estimate once we get it?

- Suppose we have a single long chain of (post-convergence) MCMC samples $\{\theta^{(t)}\}_{t=1}^{T}$. Let

$$\hat{\theta}_T = \hat{\mathbb{E}}[\theta|\mathbf{y}] = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}.$$

- Then by the CLT, under iid sampling we could take

$$\hat{\mathbb{V}}_{\text{iid}}[\hat{\theta}_T] = \frac{s_\theta^2}{T} = \frac{1}{T(T-1)}\sum_{t=1}^{T}(\theta^{(t)} - \hat{\theta}_T)^2.$$

  But this is likely an **underestimate** due to positive autocorrelation in the MCMC samples.

Variance estimation II

- To avoid wasteful parallel sampling or "thinning", compute the effective sample size,

$$\text{ESS} = \frac{T}{\kappa(\theta)},$$

where $\kappa(\theta) = 1 + 2\sum_{k=1}^{\infty} \rho_k(\theta)$ is the **autocorrelation time**, and we cut off the sum when $\rho_k(\theta) < \epsilon$.
Then

$$\hat{\mathbb{V}}_{\text{ESS}}(\hat{\theta}_T) = \frac{s_\theta^2}{\text{ESS}(\theta)}.$$

Note: $\kappa(\theta) \geq 1$, so $\text{ESS}(\theta) \leq T$, and so we have that $\hat{\mathbb{V}}_{\text{ESS}} \geq \hat{\mathbb{V}}_{\text{iid}}$ as expected.

Available tools for estimation

- General purpose estimation tools are provided by the BUGS family:
  1. (WinBUGS)
  2. (OpenBUGS)
  3. JAGS
  - Models are specified via variants of the BUGS language.
  - The software parses the model and determines the samplers automatically to generate draws from the posterior.
- Other major general purpose estimation tool: STAN.

Available tools in R

- **Estimation:**
  - **rjags** provides an interface to the JAGS library.
  - **rstan** provides an interface to the STAN library.
- **Post-processing, convergence diagnostics:**
  - **coda** (Convergence Diagnosis and Output Analysis):
    - contains a suite of functions that can be used to summarize, plot, and and diagnose convergence from MCMC samples.
    - can easily import MCMC output from JAGS or from plain matrices.
    - provides the Gelman & Rubin, Geweke, Heidelberger & Welch, and Raftery & Lewis diagnostics.

For more information see the CRAN Task View: Bayesian Inference.

Data I

- The data consists of a time series of daily USD/EUR exchange rates $\{x_t\}$ from 2000/01/03 to 2012/04/04. We have this data available in package **stochvol** in R.

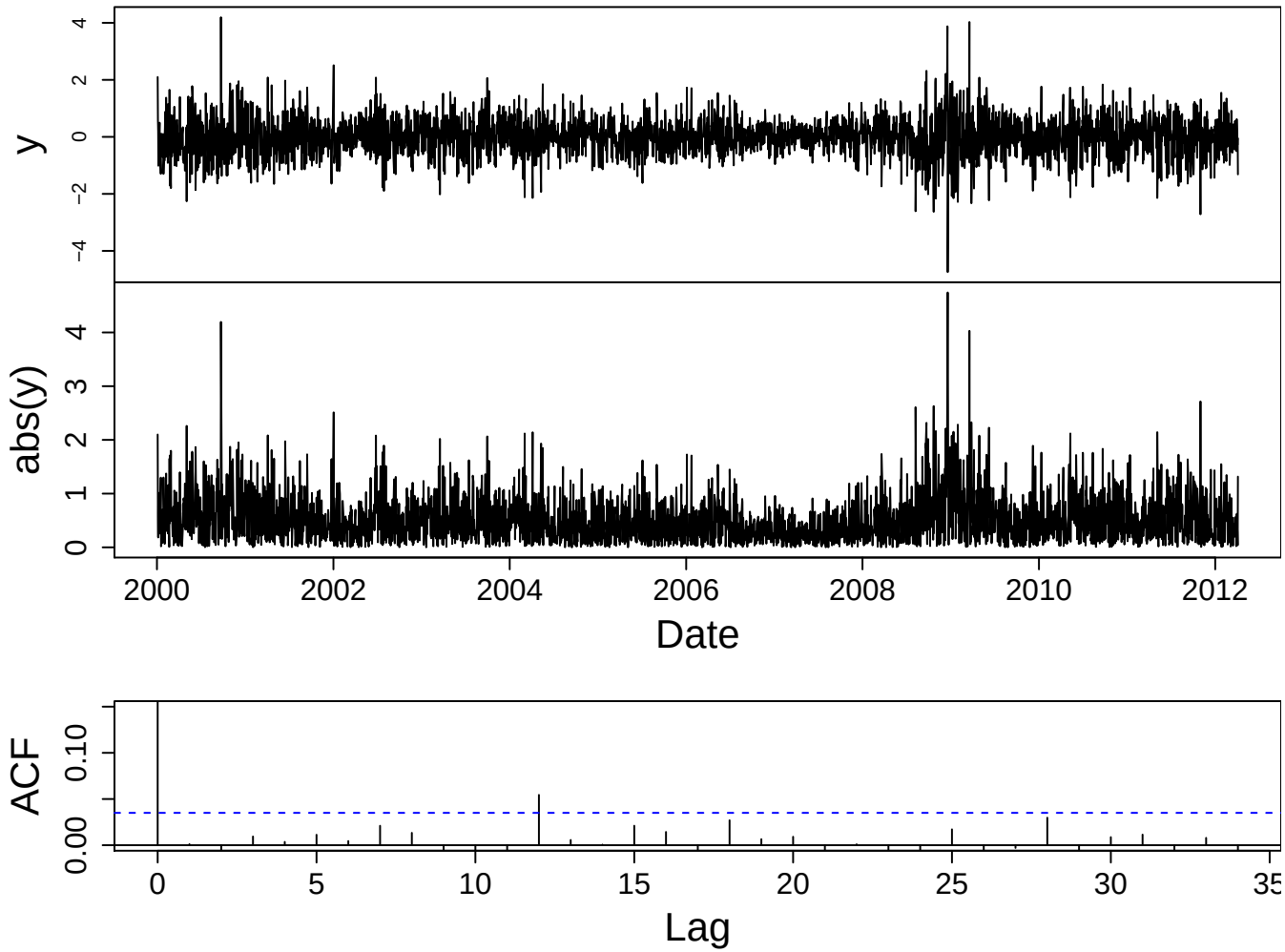
```
> data(exrates, package = "stochvol")
> Garch <- exrates[, c("date", "USD")]
> x <- Garch$USD
```

- The series of interest are the daily mean-corrected returns times hundred, $\{y_t\}$ for $t = 1, \ldots, n$.

$$y_t = 100 \left[\log x_t - \log x_{t-1} - \frac{1}{n} \sum_{i=1}^{n} (\log x_t - \log x_{t-1}) \right],$$

```
> y <- 100 * diff(log(x))
> y <- y - mean(y)
```

Data II



Model I

- Heteroscedasticity can be observed. What can be done?
- GARCH(1,1): $y_t \sim N(0, \sigma_t^2)$ with $\sigma_t^2 = \omega_0 + \omega_1 \epsilon_{t-1}^2 + \lambda_1 \sigma_{t-1}^2$.
- In a stochastic volatility model the variance of a stochastic process is itself randomly distributed and it can be written in the form of a nonlinear state-space model.
- A state-space model specifies the conditional distributions of the observations given unknown states, here the underlying log variances, θ_t, in the observation equations for $t = 1, \ldots, n$

$$y_t | \theta_t \overset{iid}{\sim} N(0, \exp(\theta_t)).$$

Model II

- The unknown states are assumed to follow a Markovian transition over time given by the state equations for $t = 1, \ldots, n$

$$\theta_t | \theta_{t-1}, \mu, \phi, \tau^2 = \mu + \phi(\theta_{t-1} - \mu) + \nu_t, \qquad \nu_t \overset{iid}{\sim} N(0, \tau^2).$$

  with $\theta_0 \sim N(\mu, \tau^2)$.
- The state θ_t determines the amount of log variance on day t.
- ϕ measures the autocorrelation present in the θ_t's and is restricted to be $-1 < \phi < 1$. It can be interpreted as the persistence in the log variance.
- μ can be seen as the level of the log variance.
- τ^2 is the variance of log-variances.

Model III

The full Bayesian model consists of

- a prior for the unobservables
  - 3 parameters: μ, ϕ, τ^2
  - unknown states: $\theta_0, \ldots, \theta_n$

$$p(\mu, \phi, \tau^2, \theta_0, \ldots, \theta_n) = p(\mu, \phi, \tau^2) p(\theta_0 | \mu, \tau^2) \prod_{t=1}^{n} p(\theta_t | \theta_{t-1}, \mu, \phi, \tau^2),$$

- a joint distribution for the observables $y_1, \ldots, y_n$

$$p(y_1, \ldots, y_n | \mu, \phi, \tau^2, \theta_0, \ldots, \theta_n) = \prod_{t=1}^{n} p(y_t | \theta_t).$$

Model specification in BUGS

```
model {
   for (t in 1:length(y)) {
     y[t] ~ dnorm(0, 1/exp(theta[t]));
   }
   theta0 ~ dnorm(mu, itau2);
   theta[1] ~ dnorm(mu + phi * (theta0 - mu), itau2);
   for (t in 2:length(y)) {
     theta[t] ~ dnorm(mu + phi * (theta[t-1] - mu), itau2);
   }
   ## prior
   mu ~ dnorm(0, 0.1);
   phistar ~ dbeta(20, 1.5);
   itau2 ~ dgamma(2.5, 0.025);
   ## transform
   tau <- sqrt(1/itau2);
   phi <- 2 * phistar - 1
 }
```

Estimation with JAGS I

- **Remark:** For Bayesian estimation the parameterization of the normal distribution is in general with respect to mean μ and precision λ, i.e.,

$$y \sim \text{dnorm}(\mu, \lambda),$$

  where $\lambda = \sigma^{-2}$, i.e., the precision is the inverse of the variance. The conjugate prior for the precision is the Gamma distribution (Gamma(0.001, 0.001) is a noninformative conjugate prior for the precision).
- Given the model specification a graphical model is constructed to determine the parents and direct children of each variable/node.
- Based on these relationships, suitable samplers are selected.

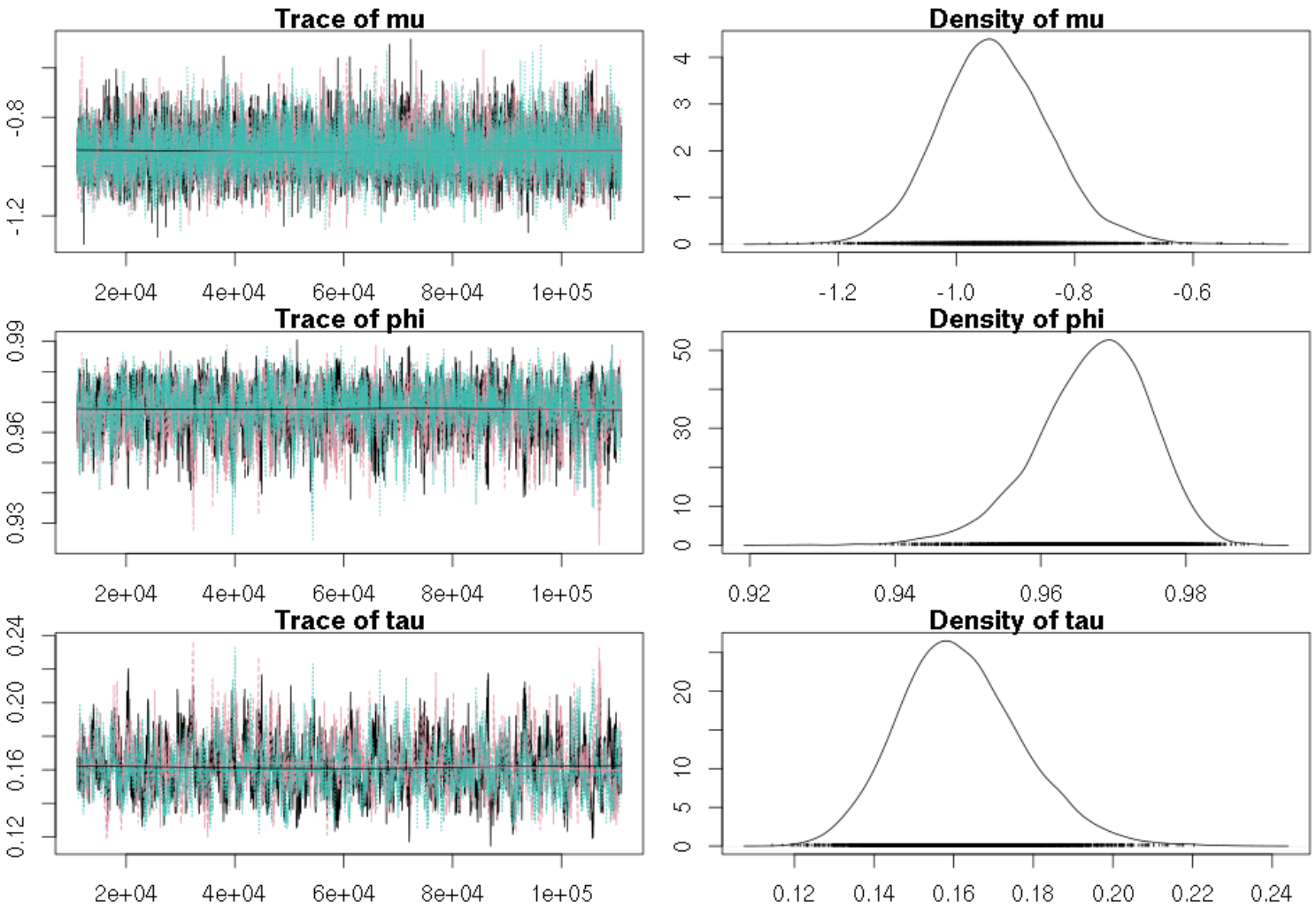
Estimation with JAGS II

```
> library("rjags")
> initials <-
+    list(list(phistar = 0.975, mu = 10, itau2 = 300),
+         list(phistar = 0.5, mu = 0, itau2 = 50),
+         list(phistar = 0.025, mu = -10, itau2 = 1))
> initials <- lapply(initials, "c",
+                    list(.RNG.name = "base::Wichmann-Hill",
+                         .RNG.seed = 2207))
> model <- jags.model("volatility.bug", data = list(y = y),
+                     inits = initials, n.chains = 3)
> update(model, n.iter = 10000)
> draws <- coda.samples(model, c("phi", "tau", "mu", "theta"),
+                       n.iter = 100000, thin = 20)
> effectiveSize(draws[, 1:3])
    mu    phi    tau
6916.7 1283.9  827.3
> summary(draws[, 1:3])
```

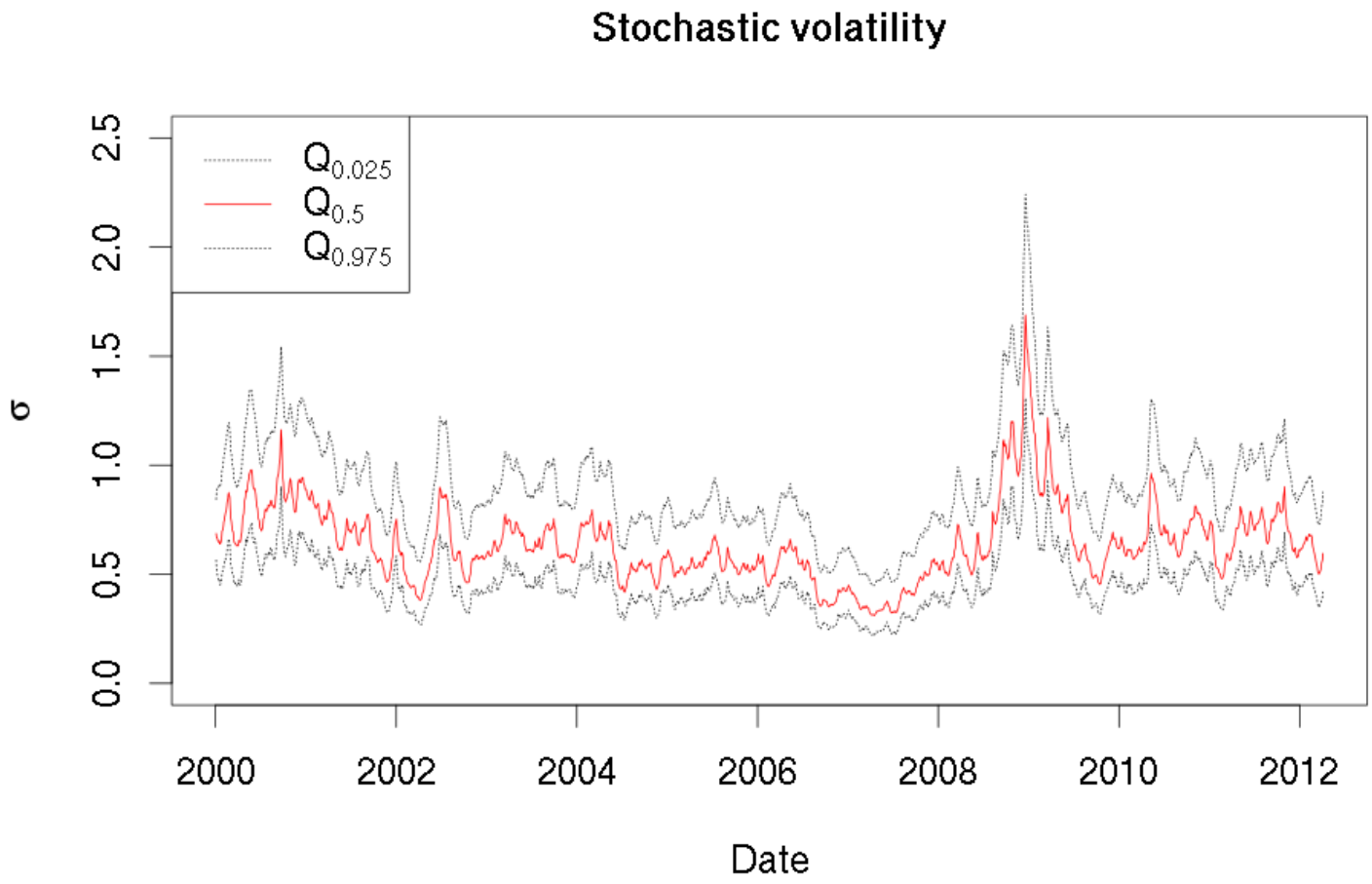
Estimation with JAGS III

```
Iterations = 11020:111000
Thinning interval = 20
Number of chains = 3
Sample size per chain = 5000
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean     SD Naive SE Time-series SE
mu  -0.935 0.0946 7.72e-04       0.001145
phi  0.967 0.0079 6.45e-05       0.000220
tau  0.162 0.0156 1.27e-04       0.000539

2. Quantiles for each variable:

      2.5%    25%    50%    75%  97.5%
mu  -1.116 -0.999 -0.939 -0.875 -0.737
phi  0.950  0.962  0.968  0.973  0.981
tau  0.134  0.151  0.161  0.171  0.195
```

Estimation with JAGS IV



SV model


Stochastic volatility
$Q_{0.025}$
$Q_{0.5}$
$Q_{0.975}$
σ
0.0
0.5
1.0
1.5
2.0
2.5
2000
2002
2004
2006
2008
2010
2012
Date

Diagnostics with **coda**

- Auto- and crosscorrelation: `autocorr.diag`, `autocorr.plot`, `crosscorr`
- Gelman and Rubin diagnostics: `gelman.diag`
- Heidelberger and Welch diagnostics: `heidel.diag`
- Geweke diagnostics: `geweke.diag`, `geweke.plot`
- Raftery and Lewis diagnostics: `raftery.diag`

For more information see the CODA manual at `http://www.stat.ufl.edu/system/man/BUGS/cdaman03/`.