WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Bayesian Computation with R

Gregor Kastner, Bettina Grün, Paul Hofmarcher & Kurt Hornik

WS 2013/14

Overview

- Lecture:
  - Bayes approach
  - Bayesian computation
  - A hands-on example: Linear Model
  - Available tools in R
  - Example: Stochastic volatility models
- Exercises
- Projects

Deliveries

- Exercises:
  - Solutions handed in by e-mail to `gregor.kastner@wu.ac.at` in a `.pdf`-file together with the original `.Rnw`-file
  - Deadline: TBA
- Projects:
  - In groups of 2–3 students
  - Data analysis using Bayesian methods
  - Documentation of the analysis consisting of
    (a) Problem description
    (b) Model specification
    (c) Model fitting: estimation and validation
    (d) Interpretation
  - Report via e-mail as a `.pdf`-file (+ `.Rnw`-file)
    Deadline: TBA
  - Presentation: TBA

Material

- Lecture slides
- Further reading:
  - Meyer, R. and Yu J. (2000) BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal* 3, 198–215. DOI: 10.1111/1368-423X.00046
  - Carlin, B. P. and Louis, T. A. (2009) *Bayesian Methods for Data Analysis*. 3rd, CRC Press.

Software tools

- JAGS: Just Another Gibbs Sampler
  - Available from sourceforge:
    `http://sourceforge.net/projects/mcmc-jags/`
  - Current version: 3.4.0
  - 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: 3.11
  - `install.packages("rjags")`
- R package **coda** on CRAN:
  - Output analysis and diagnostics for MCMC
  - `install.packages("coda")`
- Further R packages on CRAN: **Ecdat**, **lme4**

Software documentation

- Plummer, M. (2013) JAGS Version 3.4.0 user manual. Available from sourceforge.
- Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003) WinBUGS user manual. Version 1.4. Available at `www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf`.
- Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003) Examples. Volume 1–3. Also available at `www.mrc-bsu.cam.ac.uk/bugs/winbugs/` as `Vol1.pdf`, `Vol2.pdf` and `Vol3.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 Bayesian can think about unknown parameters $\boldsymbol{\theta}$ for which no reliable frequentist experiment exist.

Bayes approach I

Idea of Bayes approach:

- A Bayesian writes down a prior guess for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, then combines this with the information that the data $\mathbf{y}$ provide. This results in a posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y})$.
- Inference is based on summaries of the posterior.
- posterior information $\geq$ prior information ≥ 0. The second $\geq$ is replaced with $=$ if we have non-informative prior information.

Bayes approach II

The parameter $\boldsymbol{\theta}$ is a **random** quantity with a **prior** distribution

$$\pi(\boldsymbol{\theta}) \equiv \pi(\boldsymbol{\theta}|\boldsymbol{\eta}).$$

$\boldsymbol{\eta}$ are the **hyperparameters** which are assumed **fixed**.
Inference on the parameter $\boldsymbol{\theta}$ is based on its **posterior** distribution given the data

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y},\boldsymbol{\theta})}{\int p(\mathbf{y},\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \end{aligned}$$

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

- **Noninformative priors:** do not favor any values of $\boldsymbol{\theta}$ if no a-priori information is available. E.g.:
  - Uniform distribution:
    - suitable if the parameter space is discrete and finite.
    - leads to **improper** priors (i.e., does not integrate to one) for continuous and infinite parameter space.
    - is not (always) invariant under reparameterization.
  - Jeffrey's prior: invariant under reparameterization.

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2},$$

where $I(\boldsymbol{\theta})$ is the expected Fisher information matrix with

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta})\right].$$

- Exercise: Jeffrey's prior for binomial experiment.

Bayesian inference I

- **Point estimation:**
  - posterior mode (aka **generalized ML estimate**)
  - posterior mean or median
  - ...
- **Interval estimation:**

Definition

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

$$1-\alpha \leq P(C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

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

Bayesian inference II

- The comparison of predictors made by alternative scientific explanations is a mainstay of statistics.

- **Hypothesis testing:** After determining an approporiate test statistics $T(y)$, we get:

$$\text{p-value} = \mathbb{P}[T(y) \text{ more extreme than } T(y_{\text{obs}})|\theta, H_0]$$

- Classical hypothesis testing has some disadvantages:
  - H_0 must be a simplification of H_a, like in nested models.
  - We can only offer evidence against the null hypothesis.
  - The p-value itself offers no direct interpretation as a "weight of evidence".

Bayesian inference III

The Bayesian approach to hypothesis testing is much simpler:

- As in the case for interval estimation, it requires some prior knowledge.
- Based on the data that each of the hypotheses is supported to predict, one applies Bayes' Theorem and computes the posterior probability that the first hypothesis is correct.

Bayesian inference IV

- **Bayes factors:**
  - The Bayes factor BF is the ratio of the posterior odds of model M_1 to the prior odds of M_1:

$$BF = \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)},$$

    i.e., the ratio of the observed marginal densities for the two models.
  - For two **a priori** equally probable models the BF equals the posterior odds of M_1. BF captures the change in the odds in favor of model 1 as we move from prior to posterior.

Bayesian inference V

- Jeffrey's scale for interpretation:

BF	Strength of evidence
< 1	Negative (support of M_2)
1–3	Barely worth mentioning
3–10	Substantial
10–30	Strong
30–100	Very strong
> 100	Decisive

- A fun reference: Wagenmakers, E.-J., Wetzels R., Borsboom D. and van der Maas, H. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi. *Journal of Personailty and Social Psychology* 100(3), 426–432.

Example: Consumer preference I

- 16 consumers have been recruited by a fast food chain to compare the flavour of ground beef patties.
- The patties were kept frozen for eight months in different freezers:
  - a high-quality freezer consistenly maintaining the temperature at $0\,^{\circ}\mathrm{F}$ $(-18\,^{\circ}\mathrm{C})$
  - a freezer where the temperature varies between 0 and $15\,^{\circ}\mathrm{F}$ (-18 to $-9\,^{\circ}\mathrm{C}$).
- In a double-blind study (neither consumers nor waiters know where the patties were stored) each consumer evaluated patties from both fridges.
- The food chain executives are interested in whether the higher-quality freezer leads to a substantial improvement in taste.
- The study result is that 13 out of 16 consumers prefer the more expensive patty.

Example: Consumer preference II

For a Bayesian analysis we need two components:

- **Likelihood:**
  - We assume that consumers are independent and that the probability θ of preferring the more expensive patty is constant over the consumers.
  - Their decisions form a sequence of Bernoulli trials.

  Denoting the number of consumers preferring the more expensive patty by Y gives

$$Y|\theta \sim \text{Bin}(16, \theta),$$

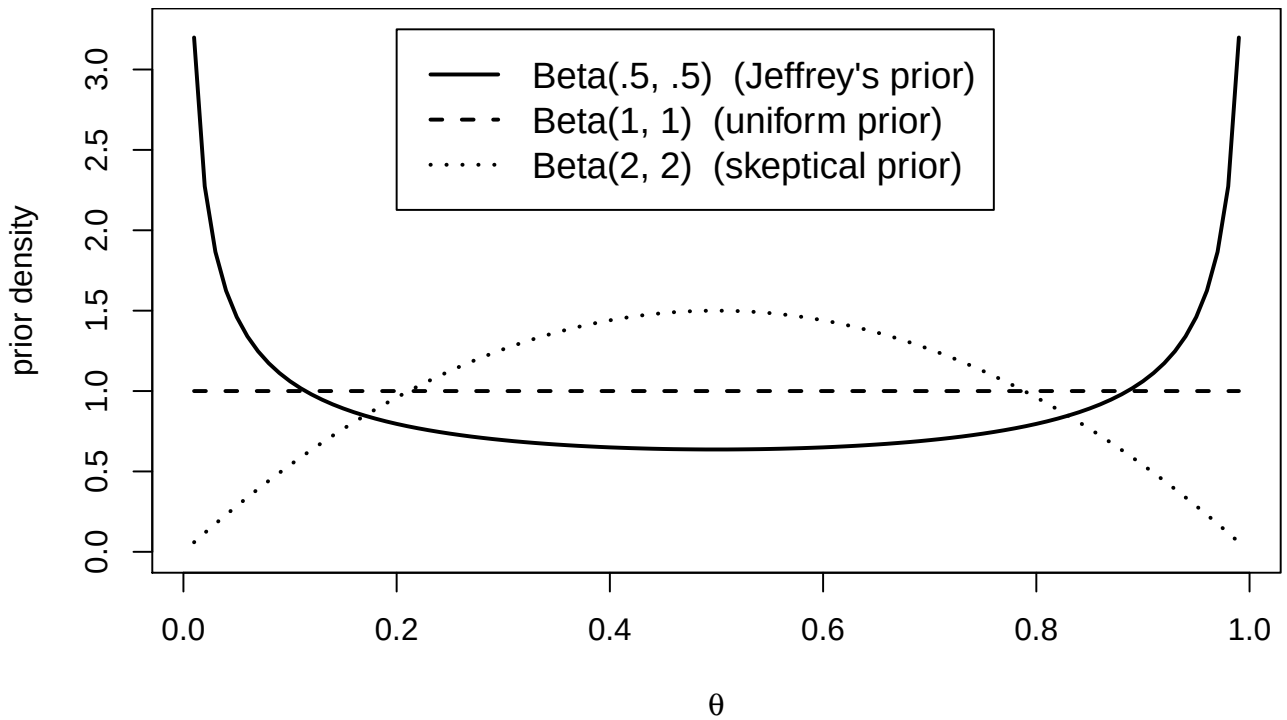
  which is equivalent to

$$f(y|\theta) = \binom{16}{y} \theta^y (1-\theta)^{16-y}.$$

Example: Consumer preference III

- **Prior:** The Beta distribution is a conjugate family for the binomial distribution.

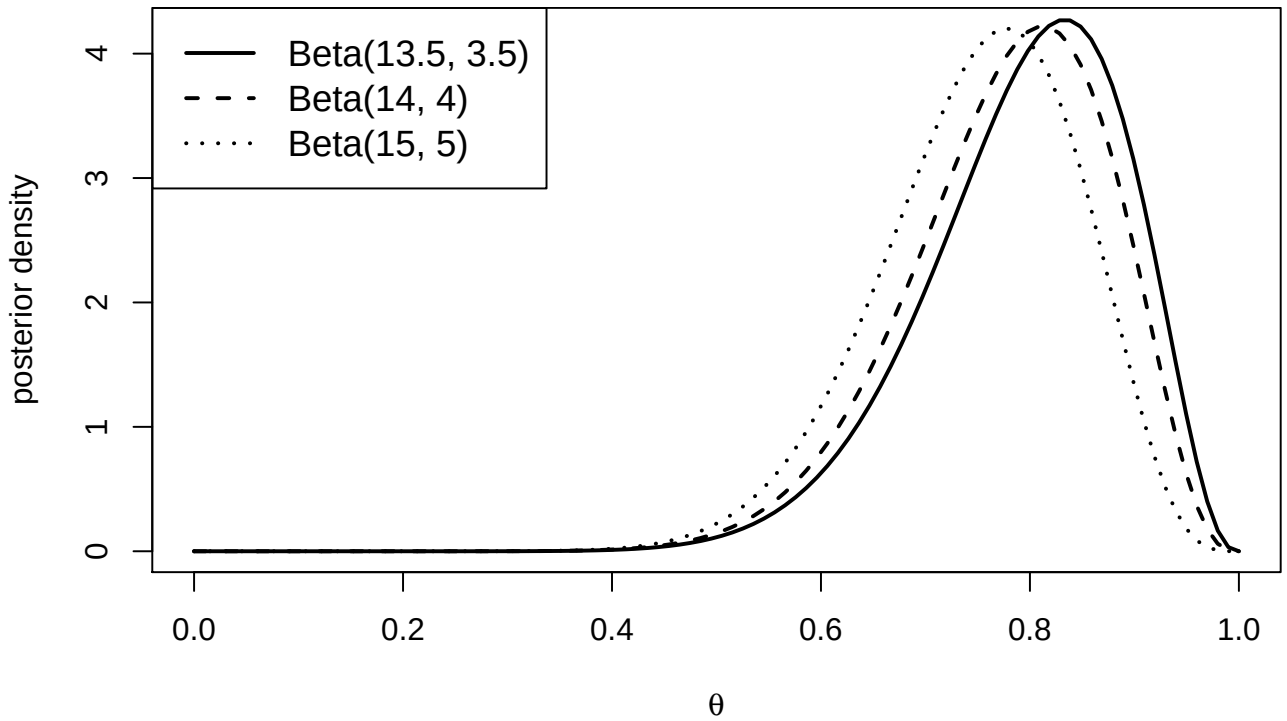
$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$



Example: Consumer preference IV

Thanks to the conjugacy the posterior distribution for θ is

$$p(\theta|y) \propto f(y|\theta)\pi(\theta) \propto \theta^{y+\alpha-1}(1-\theta)^{16-y+\beta-1}$$
$$\propto \text{Beta}(y+\alpha, 16-y+\beta)$$



Example: Consumer preference V

- We return to the executives' question concerning a substantial improvement in taste.
- We select 0.6 as the critical value that θ must exceed in order for the improvement to be regarded as "substantial".
- Given this cutoff value, we compare the hypotheses $M_1 : \theta \geq 0.6$ and $M_2 : \theta < 0.6$.
- Using a uniform prior we get for $\mathbb{P}(\theta > .6|x)$:

```
> (p1 <- round(pbeta(0.6, 14, 4, lower.tail = FALSE),
+              digits = 3))
```

```
[1] 0.954
```

Example: Consumer preference VI

The Bayes factor is then given by

$$BF = \frac{0.954/0.046}{0.4/0.6} = 31.1.$$

This implies a reasonable strong preference for M_1.

Bayesian computation

- Asymptotic methods
- Noniterative Monte Carlo methods
- Markov chain Monte Carlo methods

Development over time

- **prehistory (1763–1960):** Conjugate priors.
- **1960's:** Numerical quadrature (Newton-Cotes methods, Gaussian quadrature, etc.).
- **1970's:** Expectation-Maximization (EM) algorithm (iterative mode finder).
- **1980's:** Asymptotic methods.
- **1980's:** Noniterative Monte Carlo methods (direct posterior sampling and indirect methods, e.g., importance sampling, rejection).
- **1990's:** Markov chain Monte Carlo (MCMC; Gibbs sampling, Metropolis Hastings algorithm, etc.). $\Rightarrow$ broadly applicable, but require care in parametrization and convergence diagnosis!
- **2000's:** Sequential Monte Carlo (SMC)

Normal approximation

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

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 $\boldsymbol{\theta}$, i.e., minus the inverse Hessian of the log posterior evaluated at the mode.

Example cont.: Consumer preference I

Using a flat prior on θ, we have

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

The first derivative is given by

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{n - x}{1 - \theta}.$$

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

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

The second derivative is given by

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

Example cont.: Consumer preference II

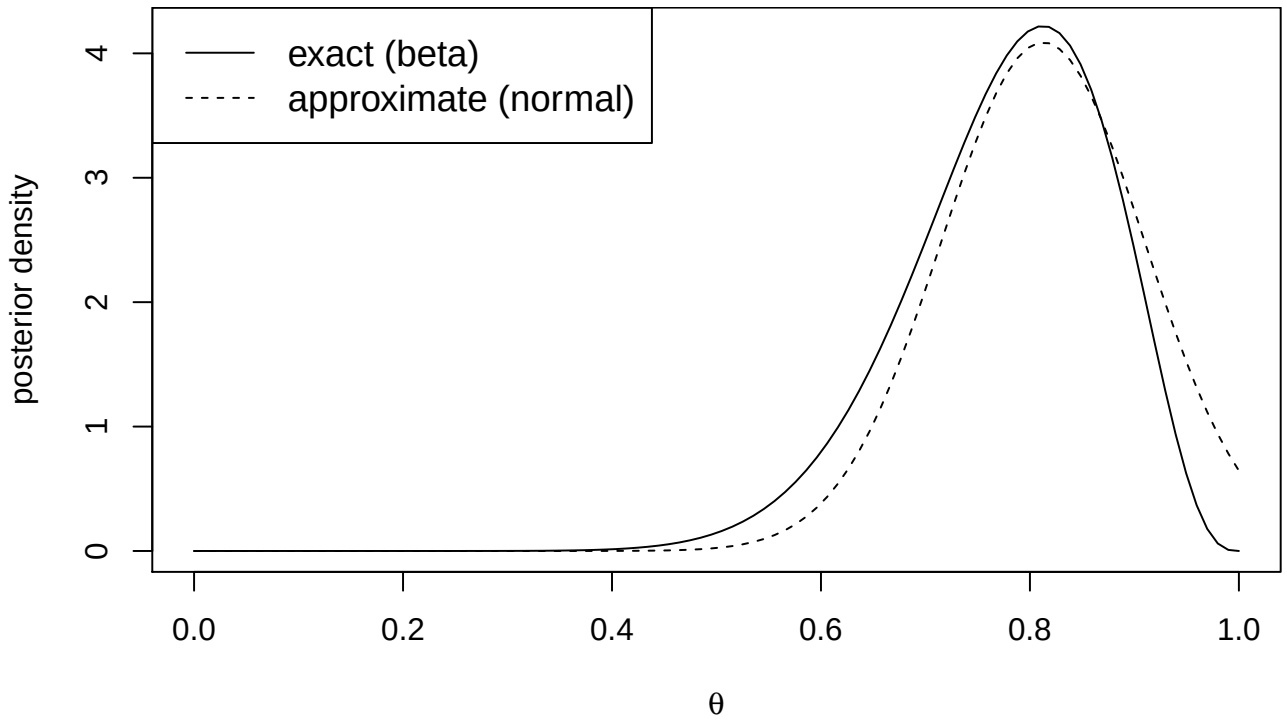
Evaluating at the estimate $\hat{\theta}^{\pi}$ gives

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

Thus the posterior can be approximated by

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

Example cont.: Consumer preference III



Similar modes, but very different tail behavior.

Asymptotic methods

- **Advantages:**
  - Deterministic, noniterative algorithm.
  - Substitutes differentiation for integration.
  - Facilitates studies of Bayesian robustness.
- **Disadvantages:**
  - Requires well-parametrized, 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
  - Importance sampling
  - Rejection sampling

Direct sampling

We begin with the most basic definition of Monte Carlo integration:

- Suppose $\theta \sim p(\theta)$ and we seek $\gamma := \mathbb{E}[c(\theta)] = \int c(\theta)p(\theta)d\theta$.
- Then if $\theta_1, \ldots, \theta_N \overset{iid}{\sim} p(\theta)$, we have

$$\hat{\gamma} = \frac{1}{N}\sum_{j=1}^{N} c(\theta_j),$$

  which converges to $\mathbb{E}[c(\theta)]$ with probability 1 as $N \to \infty$.
- Hence the computation of posterior expectations requires only a sample size of N from the posterior.

Importance sampling I

- Suppose we wish to approximate

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

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

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

  where $\theta_j \overset{iid}{\sim} g(\theta)$.
- Here, $g(\theta)$ is called the importance function; a good match to $cf(\mathbf{y}|\theta)\pi(\theta)$ will produce roughly equal weights.

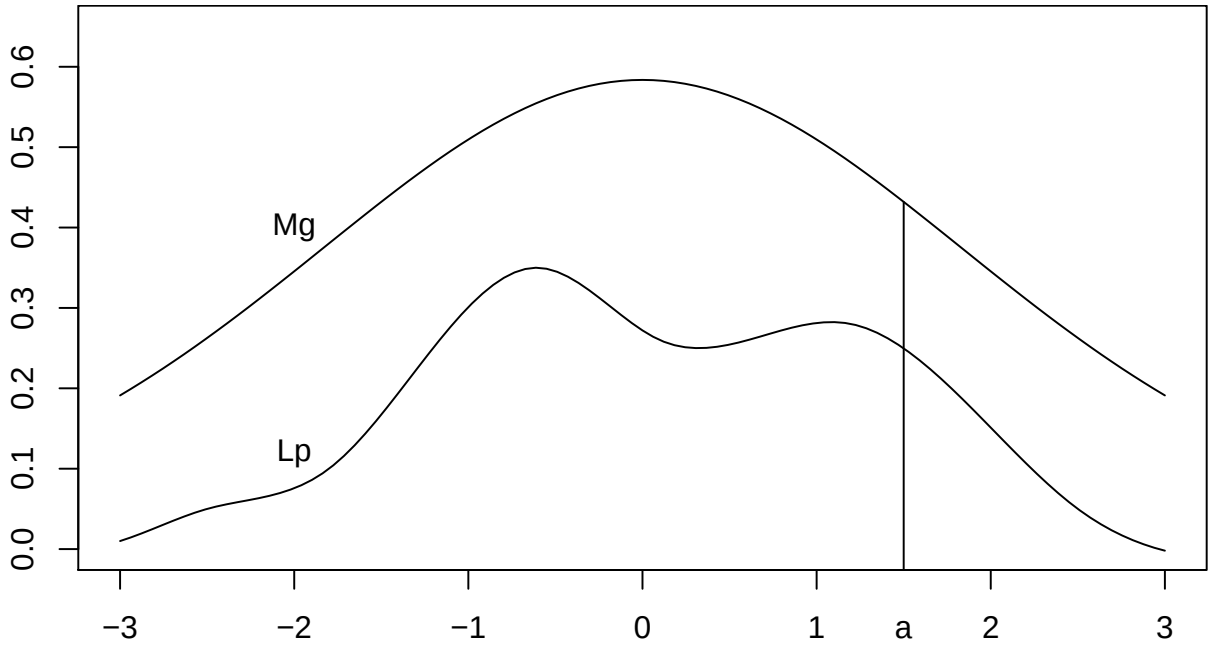
Rejection sampling I

- Instead of trying to approximate the posterior

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

  we try to find a majorizing function.
- Suppose there exists a constant $M > 0$ and a smooth density $g(\theta)$, called the envelope function, such that $f(\mathbf{y}|\theta)\pi(\theta) < Mg(\theta)$ for all θ.
- The algorithm proceeds as follows:
  (i) Generate $\theta_j \sim g(\theta)$.
  (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 $p(\theta|\mathbf{y})$.

Rejection sampling II



- Consider the θ_j samples in the histogram bar centered at a: the rejection step "slices off" the top portion of the bar.
- Repeat for all a: accepted θ_js mimic the lower curve!
- Need to choose M as small as possible (efficiency), and watch for "envelope violations"!

Markov chain Monte Carlo methods I

- Iterative MC methods are useful when it is difficult or impossible to find a feasible importance or envelope density.
- **Algorithms:**
  - Gibbs sampler
  - Metropolis-Hastings algorithm
  - Slice sampler
- **Performance evaluation:**
  - Convergence monitoring and diagnostics
  - Variance estimation

Markov chain Monte Carlo methods II

- Given two unknowns x and y, we can often write

$$p(x) = \int p(x|y)p(y)dy \qquad \text{and } p(y) = \int p(y|x)p(x)dx,$$

where $p(x|y)$ and $p(y|x)$ are known.

- Seeking $p(x)$ the analytical solution via **substitution** is:

$$p(x) = \int p(x|y) \int p(y|x')p(x')dx'dy = \int h(x,x')p(x')dx',$$

where $h(x,x') = \int p(x|y)p(y|x')dy$.

- This determines a **fixed point** system which converges under mild conditions.

$$p_{i+1}(x) = \int h(x,x')p_i(x')dx'$$

Markov chain Monte Carlo methods III

- Tanner and Wong (1987) showed that one can also use a sampling-based approach which they refer to as **data augmentation**.
  1. Draw $X^{(0)} \sim p_0(x)$.
  2. Draw $Y^{(1)} \sim p(y|x^{(0)})$.
  3. Finally, $X^{(1)} \sim p(x|y^{(1)})$.
- Then $X^{(1)}$ has marginal distribution

$$p_1(x) = \int p(x|y)p_1(y)dy = \int h(x, x')p_0(x')dx'.$$

- Repeating this process produces pairs $(X^{(i)}, Y^{(i)})$ such that $X^{(i)} \xrightarrow{d} X \sim p(x)$ and $Y^{(i)} \xrightarrow{d} Y \sim p(y)$.
- The luxury of avoiding the integration above has come at the price of obtaining not the marginal density $p(x)$ itself, but only a **sample** from this density.

Gibbs sampling I

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

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

- 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 algorithm I

- What happens if the full conditional $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$ is not available in closed form?
- Typically, $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$ will be available up to a proportionality constant, since it is proportional to the part of the Bayesian model (likelihood times prior) that involves θ_i.
- Suppose the true joint posterior for $\boldsymbol{\theta}$ has unnormalized density $p(\boldsymbol{\theta})$.
- Choose a candidate density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ that is a valid density function for every possible value of the conditioning variable $\boldsymbol{\theta}^{(t-1)}$, and satisfies

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*),$$

i.e., q is symmetric in its arguments.

Metropolis 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. Draw $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$.
  2. Compute the ratio

$$r = \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)})}.$$

  3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$;
     If $r < 1$, set $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* \text{ with probability } r \\ \boldsymbol{\theta}^{(t-1)} \text{ with probability } 1 - r \end{cases}$.
- Then a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.
- **Note:** When used as a substep in a larger (e.g., Gibbs) algorithm, we often use $T = 1$ (convergence still OK).

Metropolis algorithm III

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

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, \tilde{\Sigma}).$$

  In one dimension $\tilde{\Sigma}$ is often chosen to provide an observed acceptance ratio near 50%:
  - Very small steps $\Rightarrow$ High acceptance rate, but also high auto-correlation.
  - Very large steps $\Rightarrow$ Low acceptance rate and also high auto-correlation.

Metropolis algorithm IV

- **Metropolis-Hastings algorithm:**
  Hastings (1970) showed we can drop the requirement that q be symmetric, provided we use

$$r = \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}.$$

  This is useful for asymmetric target densities.

Auxiliary variables: Simple slice sampler I

- Ease and/or accelerate sampling from $p(\boldsymbol{\theta})$ by adding an **auxiliary** (or **latent**) variable $U \sim p(u|\boldsymbol{\theta})$.
- Suppose we want to sample a univariate θ from $p(\theta|\mathbf{y}) \propto h(\theta)$. We add an auxiliary variable U such that $U|\theta \sim \text{Unif}(0, h(\theta))$. The the joint distribution of θ and U is

$$p(\theta, u) \propto h(\theta)\frac{1}{h(\theta)}I(u < h(\theta)) = I(u < h(\theta)),$$

  where I denotes the indicator function.
- The Gibbs sampler for this joint distribution is given by
  1. $u|\theta \sim \text{Unif}(0, p(\theta))$, and
  2. $\theta|u \sim \text{Unif}(\theta : p(\theta) \geq u)$.
- The second update (over the "slice" defined by u) requires $p(\theta)$ to be invertible, either analytically or numerically.

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?
  - 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.

Variance estimation III

- Another alternative: **Batching**
  Divide the run into m successive batches of length k with batch means $b_1, \ldots, b_m$. Then $\hat{\theta}_T = \bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$, and

$$\hat{\mathbb{V}}_{\text{batch}}(\hat{\theta}_T) = \frac{1}{m(m-1)}\sum_{i=1}^{m}(b_i - \hat{\theta}_T)^2,$$

  provided that k is large enough so that the correlation between batches is negligible.
- For any $\hat{\mathbb{V}}$ used to approximate $\mathbb{V}(\hat{\theta}_N)$, a 95% CI for $\mathbb{E}[\theta|\mathbf{y}]$ is then given by

$$\hat{\theta}_T \pm z_{0.025}\sqrt{\hat{\mathbb{V}}}.$$

Hands-on example: Bayesian Linear Model

- Observation equation: $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$

Hands-on example: Bayesian Linear Model

- Observation equation: $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$
- Prior distributions: $\boldsymbol{\beta}|\sigma^2 \sim N\left(\mathbf{b_0}, \sigma^2\mathbf{B_0}\right)$, $\sigma^2 \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$

Hands-on example: Bayesian Linear Model

- Observation equation: $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$
- Prior distributions: $\boldsymbol{\beta}|\sigma^2 \sim N\left(\mathbf{b_0}, \sigma^2\mathbf{B_0}\right)$, $\sigma^2 \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$

According to Bayes formula, the posterior density is given through

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \quad \propto \quad \underbrace{p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)}_{\text{likelihood}} \underbrace{p(\boldsymbol{\beta}, \sigma^2)}_{\text{prior}}$$

Hands-on example: Bayesian Linear Model

- Observation equation: $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$
- Prior distributions: $\boldsymbol{\beta}|\sigma^2 \sim N\left(\mathbf{b_0}, \sigma^2\mathbf{B_0}\right)$, $\sigma^2 \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$

According to Bayes formula, the posterior density is given through

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \quad \propto \quad \underbrace{p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)}_{\text{likelihood}} \underbrace{p(\boldsymbol{\beta}, \sigma^2)}_{\text{prior}} = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$$

Hands-on example: Bayesian Linear Model

- Observation equation: $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$
- Prior distributions: $\boldsymbol{\beta}|\sigma^2 \sim N\left(\mathbf{b_0}, \sigma^2\mathbf{B_0}\right)$, $\sigma^2 \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$

According to Bayes formula, the posterior density is given through

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto \underbrace{p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)}_{\text{likelihood}} \underbrace{p(\boldsymbol{\beta}, \sigma^2)}_{\text{prior}} = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\
&\propto \left(\frac{1}{\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \times \\
&\quad \left(\frac{1}{\sigma}\right)^p \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{b_0})'\mathbf{B_0}^{-1}(\boldsymbol{\beta}-\mathbf{b_0})\right\} \times \\
&\quad \left(\frac{1}{\sigma^2}\right)^{c_0+1} \exp\left\{-\frac{C_0}{\sigma^2}\right\}
\end{aligned}
$$

Bayesian Linear Model II: A Solution

The posterior $\beta, \sigma^2|\mathbf{y}$ follows a so-called "normal-inverse-gamma" distribution, for which it can be shown that

$$\hat{\beta}_{\text{Bayes}} := \mathbb{E}[\beta|\mathbf{y}] = (\mathbf{X}'\mathbf{X} + \mathbf{B_0}^{-1})^{-1}(\mathbf{B_0}^{-1}\mathbf{b_0} + \mathbf{X}'\mathbf{y})$$

Bayesian Linear Model II: A Solution

The posterior $\beta, \sigma^2|\mathbf{y}$ follows a so-called "normal-inverse-gamma" distribution, for which it can be shown that

$$\hat{\beta}_{\text{Bayes}} := \mathbb{E}[\beta|\mathbf{y}] = (\mathbf{X}'\mathbf{X} + \mathbf{B_0}^{-1})^{-1}(\mathbf{B_0}^{-1}\mathbf{b_0} + \mathbf{X}'\mathbf{y})$$

Setting $\mathbf{A} = (\mathbf{X}'\mathbf{X} + \mathbf{B_0}^{-1})^{-1}\mathbf{X}'\mathbf{X}$, we can interpret $\hat{\beta}_{\text{Bayes}}$ as the weighted mean of prior expectation $\mathbf{b_0}$ and OLS estimate $\hat{\beta}_{\text{OLS}}$:

$$\hat{\beta}_{\text{Bayes}} = (\mathbf{I} - \mathbf{A})\mathbf{b_0} + \mathbf{A}\hat{\beta}_{\text{OLS}}$$

Note that when the diagonal elements of $\mathbf{B_0}$ are large, $\mathbf{A}$ approaches $\mathbf{I}$ and thus $\hat{\beta}_{\text{Bayes}}$ approaches $\hat{\beta}_{\text{OLS}}$. Vice versa, when $\mathbf{B_0}$ has small diagonal elements, $\mathbf{A}$ approaches $\mathbf{0}$, thus $\hat{\beta}_{\text{Bayes}}$ approaches $\mathbf{b_0}$.

Bayesian Linear Model III: Another Solution

What to do if you don't speak normal-inverse-gamma-ish, or you want to find a more flexible way of learning about the posterior distribution?

Bayesian Linear Model III: Another Solution

What to do if you don't speak normal-inverse-gamma-ish, or you want to find a more flexible way of learning about the posterior distribution?

$$\Downarrow$$

Use the Gibbs-sampler to "surf the posterior" by alternately simulating values from the (full) conditional parameter densities $\boldsymbol{\beta}|\mathbf{y}, \sigma^2$ and $\sigma^2|\mathbf{y}, \boldsymbol{\beta}$.

Bayesian Linear Model IV: Hands On!

- $\beta|\mathbf{y}, \sigma^2 \sim N(\mathbf{b_n}, \mathbf{B_n})$ with

$$\begin{aligned} \mathbf{B_n} &= \left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{1}{\sigma^2}\mathbf{B_0}^{-1}\right)^{-1} \\ \mathbf{b_n} &= \mathbf{B_n}\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{y} + \frac{1}{\sigma^2}\mathbf{B_0}^{-1}\mathbf{b_0}\right) \end{aligned}$$

Bayesian Linear Model IV: Hands On!

- $\beta|\mathbf{y}, \sigma^2 \sim N(\mathbf{b_n}, \mathbf{B_n})$ with

$$\begin{aligned} \mathbf{B_n} &= \left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{1}{\sigma^2}\mathbf{B_0}^{-1}\right)^{-1} \\ \mathbf{b_n} &= \mathbf{B_n}\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{y} + \frac{1}{\sigma^2}\mathbf{B_0}^{-1}\mathbf{b_0}\right) \end{aligned}$$

- $\sigma^2|\mathbf{y}, \beta \sim \mathcal{G}^{-1}(c_n, C_n)$ with

$$\begin{aligned} c_n &= c_0 + \frac{n}{2} + \frac{p}{2} \\ C_n &= C_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \frac{1}{2}(\beta - \mathbf{b_0})'\mathbf{B_0}^{-1}(\beta - \mathbf{b_0}) \end{aligned}$$

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.

Available tools in R

- **Estimation:**
  - **R2WinBUGS** allows to run WinBUGS & OpenBUGS from R.
  - **rjags** provides an interface to the JAGS 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 WinBUGS, OpenBUGS, and 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 Pound/Dollar exchange rates $\{x_t\}$ from 01/10/81 to 28/06/85. We have this data available in package **Ecdat** in R.

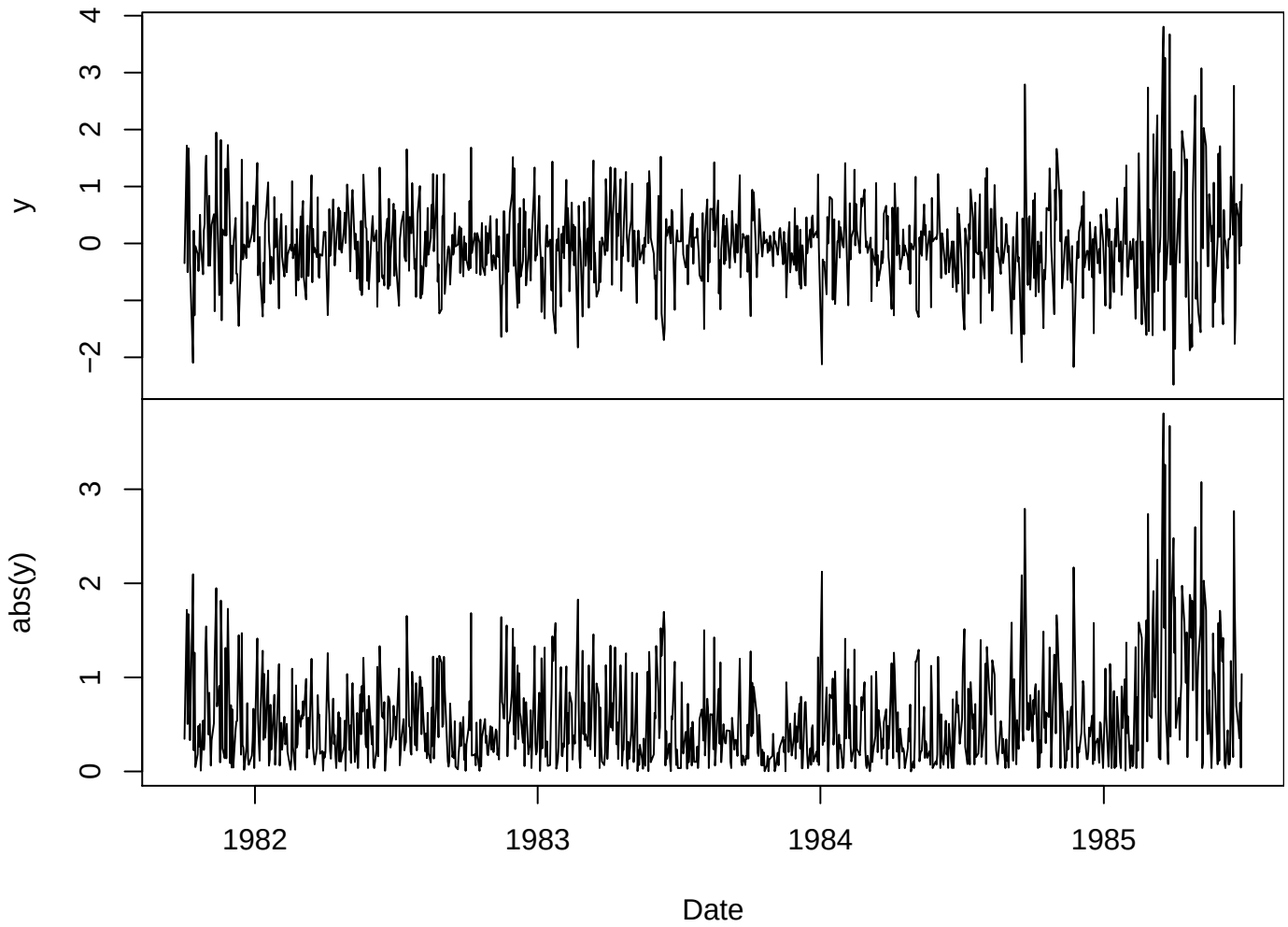
```
> data("Garch", package = "Ecdat")
> Garch <- subset(Garch,
+                 date >= 811001 & date <= 850628,
+                 c(date, bp))
> x <- Garch$bp
```

- 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

- The stochastic volatility model can be written in the form of a nonlinear state-space model.
- A state-space model specifices the conditional distributions of the observations given unkown states, here the underlying latent volatilities, θ_t, in the observation equations for $t = 1, \ldots, n$

$$y_t|\theta_t = \exp\left(\frac{1}{2}\theta_t\right) u_t, \qquad u_t \overset{iid}{\sim} N(0,1).$$

- The unkown 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 $\theta_0 \sim N(\mu, \tau^2)$.

Model II

- The state θ_t determines the amount of volatility on day t.
- ϕ measures the autocorrelation present in the logged squared data and is restricted to be $-1 < \phi < 1$. It can be interpreted as the persistence in the volatility.
- The constant scaling factor $\beta = \exp(\mu/2)$ can be seen as the modal volatility.
- τ^2 is the volatility of log-volatilities.

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

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

where $\tau = \sigma^{-2}$, i.e., the precision is the inverse of the variance. The conjugate prior for the precision is the Gamma distribution.

Model III

The full Bayesian model consists of

- a prior for the unobservables
  - 3 parameters: μ, ϕ, τ^2
  - unkown 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 I

```
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
   beta <- exp(mu/2);
   tau <- sqrt(1/itau2);
   phi <- 2 * phistar - 1
 }
```

Estimation with JAGS I

- 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 (from the **base** and **bugs** module):
  - **Conjugate sampler:** for Gibbs sampling.
  - **Finite sampler:** discrete valued node with fixed support of less than 20 possible values, not bounded using truncation.
  - **Discrete slice sampler:** for any scalar discrete-valued stochastic node.
  - **Real slice sampler:** for any scalar real-valued stochastic node.

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", "beta"),
+                       n.iter = 100000, thin = 20)

> summary(draws)
```

Estimation with JAGS III

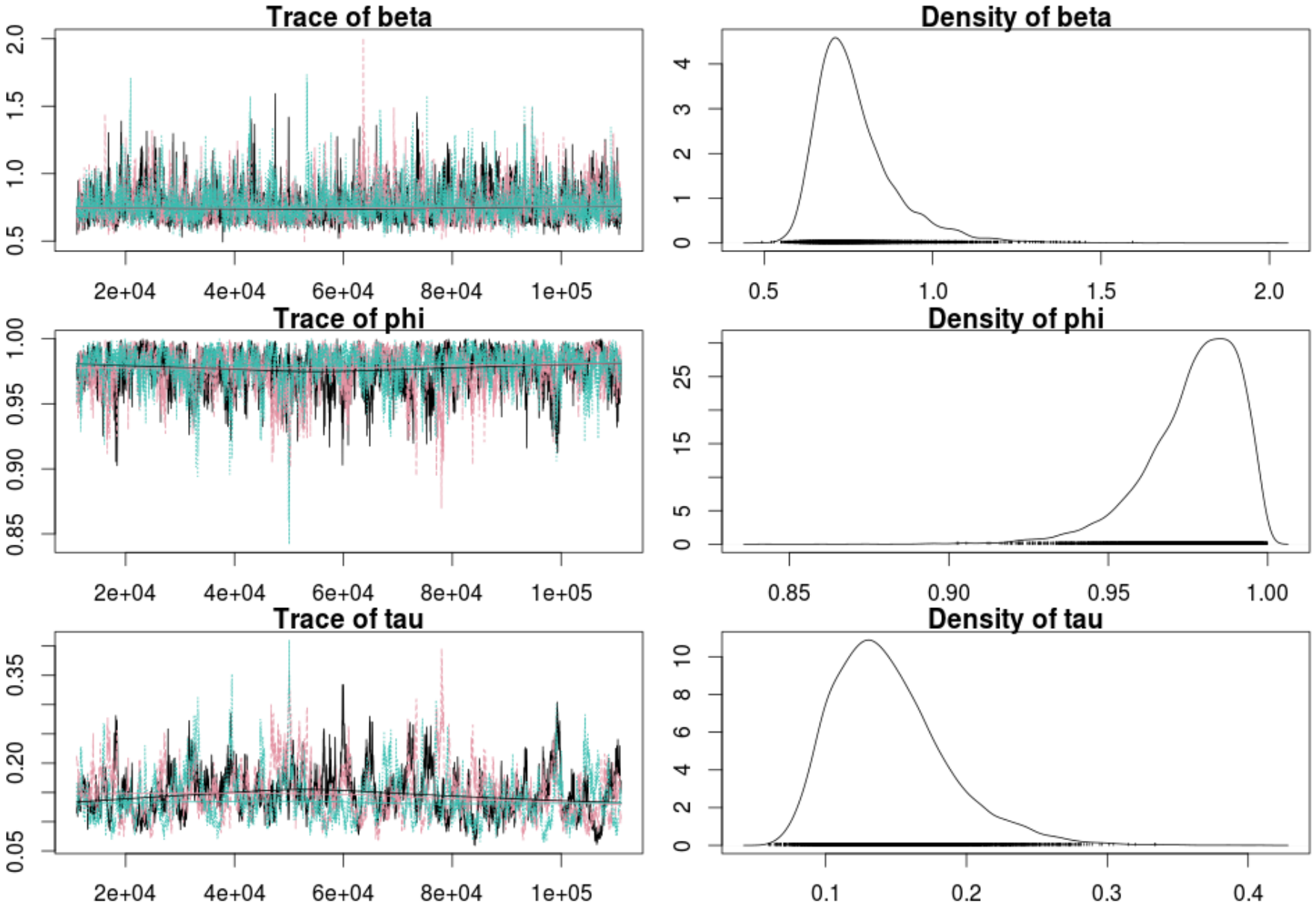
```
Iterations = 11020:111000
Thinning interval = 20
Number of chains = 3
Sample size per chain = 5000
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean     SD Naive SE Time-series SE
beta 0.770 0.1231 0.001005       0.003436
phi  0.977 0.0150 0.000122       0.000676
tau  0.145 0.0408 0.000333       0.002469

2. Quantiles for each variable:

      2.5%   25%   50%   75% 97.5%
beta 0.611 0.689 0.744 0.821 1.084
phi  0.940 0.969 0.980 0.988 0.997
tau  0.085 0.116 0.139 0.167 0.243
```

Estimation with JAGS IV



Diagnostics with coda I

- 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.mrc-bsu.cam.ac.uk/bugs/documentation/Download/cdaman03.pdf` and the addendum to the manual at `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/Download/cdaman04.pdf`

Literature I

Albert, J. (2007) *Bayesian Computation with R*. Springer.

Carlin, B. P. (2010) *Introduction to Bayesian Analysis.* Course material available at `http://www.biostat.umn.edu/~brad`.

Carlin, B. P. and Louis, T. A. (2009) *Bayesian Methods for Data Analysis*. 3rd, CRC Press.

Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 91(434), 883–904.

Chib, S., Griffiths W. and Koop G. (2008) *Bayesian Econometrics*. Emeral Group Publishing Ltd.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.

Literature II

Neal, R. M. (2003) Slice sampling. *Annals of Statistics* 31(3), 705–741.

Meyer, R. and Yu J. (2000) BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal* 3, 198–215.

Tanner, M. A. and Wong, W. H. (1987) The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398): 528–540.

Watsham, T. J. and Parramore, K. (1997) *Quantitative Methods in Finance.* Cengage Learning EMEA.