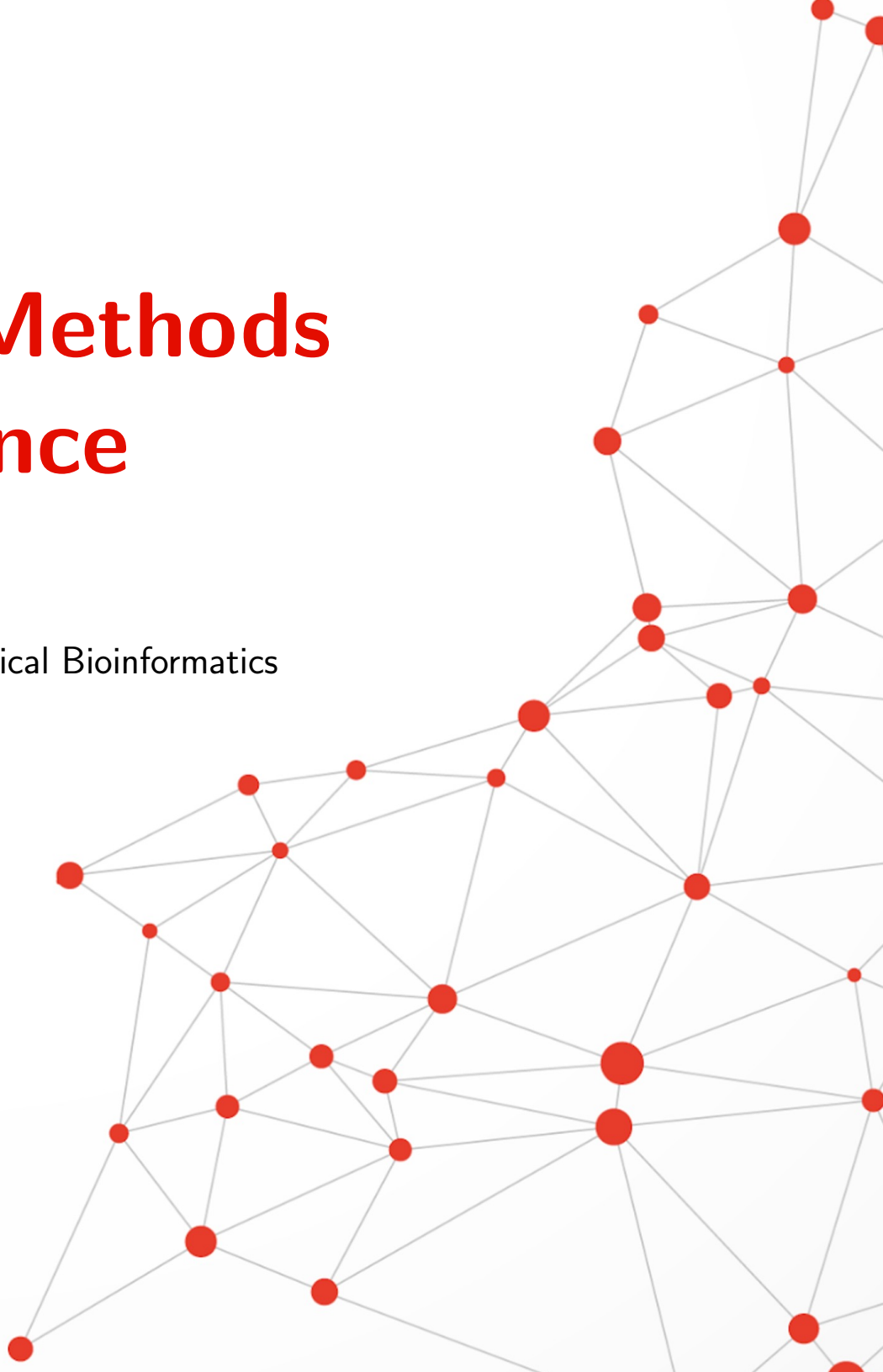
Use of Sampling Methods in Bayesian Inference

Christel Faes
Interuniversity Institute for Biostatistics and statistical Bioinformatics
Hasselt University
christel.faes@uhasselt.be



DSI DATA SCIENCE INSTITUTE
UHASSELT
WWW.UHASSELT.BE/DSI

Contents

Chapter 1
Bayes theorem: computing the posterior distribution

Bayes theorem

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

$$\text{posterior probability} = \frac{\text{likelihood} \times \text{prior probability}}{p(y)}$$

Probability can have two meanings: limiting proportion (objective) or personal belief (subjective)

1.1 Bayes theorem

- Bayes theorem for parameter θ with two possible values, $\theta = 1$ or $\theta = 0$:

$$p(\theta = 1 \mid y) = \frac{p(y \mid \theta = 1)\, p(\theta = 1)}{p(y \mid \theta = 1)\, p(\theta = 1) + p(y \mid \theta = 0)\, p(\theta = 0)}$$

- Bayes theorem for categorical parameter, $\theta_1, \theta_2, \ldots, \theta_K$

$$p(\theta_k \mid y) = \frac{p(y \mid \theta_k)\, p(\theta_k)}{\sum_{k=1}^{K} p(y \mid \theta_k)\, p(\theta_k)}$$

- Bayes theorem for continuous parameter θ:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta) d\theta}$$

- Data y can be binary, categorial, continuous
- i.i.d. sample $\boldsymbol{y} = y_1, \ldots, y_n$
- Joint distribution of sample $= p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) =$ likelihood $L(\theta|\boldsymbol{y})$

$\Rightarrow$ Bayes' Theorem for continuous parameters:

$$p(\theta|\boldsymbol{y}) = \frac{L(\theta|\boldsymbol{y})p(\theta)}{p(\boldsymbol{y})} = \frac{L(\theta|\boldsymbol{y})p(\theta)}{\int L(\theta|\boldsymbol{y})p(\theta)d\theta}$$

Special Cases: use of conjugate priors

- Binomial likelihood + **beta** prior = **beta** posterior
- Normal likelihood + **normal** prior (σ known) = **normal** posterior
- Poisson likelihood + **gamma** prior = **gamma** posterior
- Features:
  - ▷ Posterior compromise between prior and likelihood
  - ▷ Posterior mode/mean/median can be analytically obtained from MLE and prior mode/mean/median
  - ▷ When parameter changes then also prior and posterior must change accordingly using transformation rule

1.2 Numerical techniques to determine the posterior

- In general, posterior distribution cannot be obtained analytically
- Solutions to calculate the normalizing factor are required

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta) d\theta}$$

- Two approaches
  - ▷ Numerical integration
  - ▷ Sampling from the posterior

Example

- Study describing caries experience in Flanders
- Longitudinal oral health study in Flanders from 1996 to 2001
- Caries experience measured by dmft-index in 4351 children
- The sum of the dmft-index in all children was 9758
- Outcome is a count: Poisson likelihood
- Prior information based on literature: Gamma(α_0, β_0)=Gamma(3,1)
- Posterior: Gamma with $\bar{\alpha} = \sum y_i + \alpha_0 = 9758 + 3 = 9761$ and $\bar{\beta} = n + \beta_0 = 4351 + 1 = 4352$

Example

- Replace gamma prior by lognormal prior
- Posterior distribution

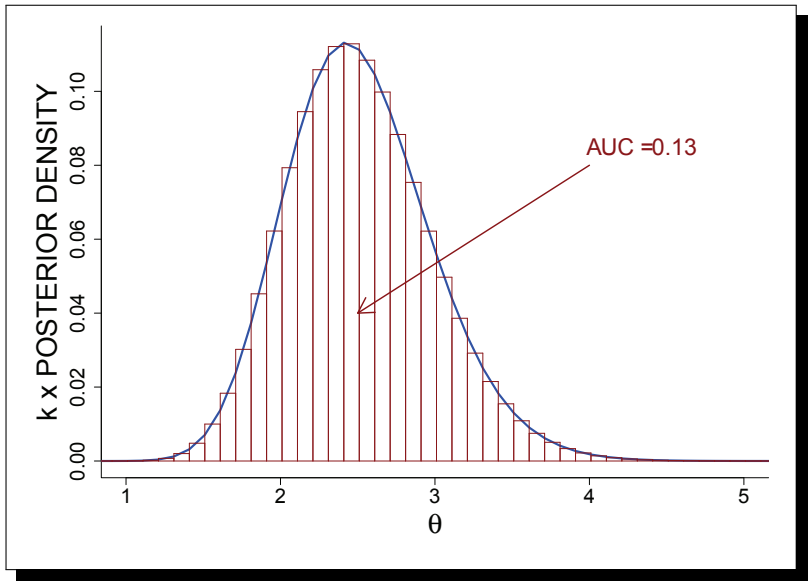
$$\propto \theta^{\sum_{i=1}^{n} y_i - 1} \mathrm{e}^{-n\theta - \left(\frac{\log(\theta)-\mu_0}{2\sigma_0}\right)^2}, \ (\theta > 0)$$

- Posterior moments cannot be evaluated & AUC not known

1.2.1 Numerical integration

- Numerical integration: replacing integral by summation

$$\int_a^b L(\theta|y)p(\theta)d\theta \approx \sum_{m=0}^{M+1} w_m L(\theta_m|y)p(\theta_m)$$



- Simple integration techniques: equidistant grid + approximate sub integrals by polynomial
  - Mid-point rule: constant
  - Trapezoidal rule: linear
  - Simpson's rule: quadratic
- Gaussian quadrature
  - Non-adaptive
  - Adaptive ($M = 1 =$ Laplace approximation)

1.2.2 Sampling from the posterior distribution

- Monte Carlo integration: usefulness of sampling idea
- General purpose sampling algorithms

Monte-Carlo integration

- Monte Carlo integration: replace integral by a Monte Carlo sample $\{\widetilde{\theta}_1, \ldots, \widetilde{\theta}_K\}$
- Approximate $p(\theta|\boldsymbol{y})$ by sample histogram
- Classical Strong Law of Large Numbers:

$$\int t(\theta)\, p(\theta|\boldsymbol{y})\, d\theta \approx \bar{t} = \frac{1}{K} \sum_{k=1}^{K} t(\widetilde{\theta}_k), \quad \text{for } K \text{ large}$$

- Classical Central Limit Theorem: 95% confidence interval

$$[\bar{t} - 1.96\, s_t/\sqrt{K}, \bar{t} + 1.96\, s_t/\sqrt{K}]$$

- 95% equal tail CI: [2.5%, 97.5%] quantile from sample

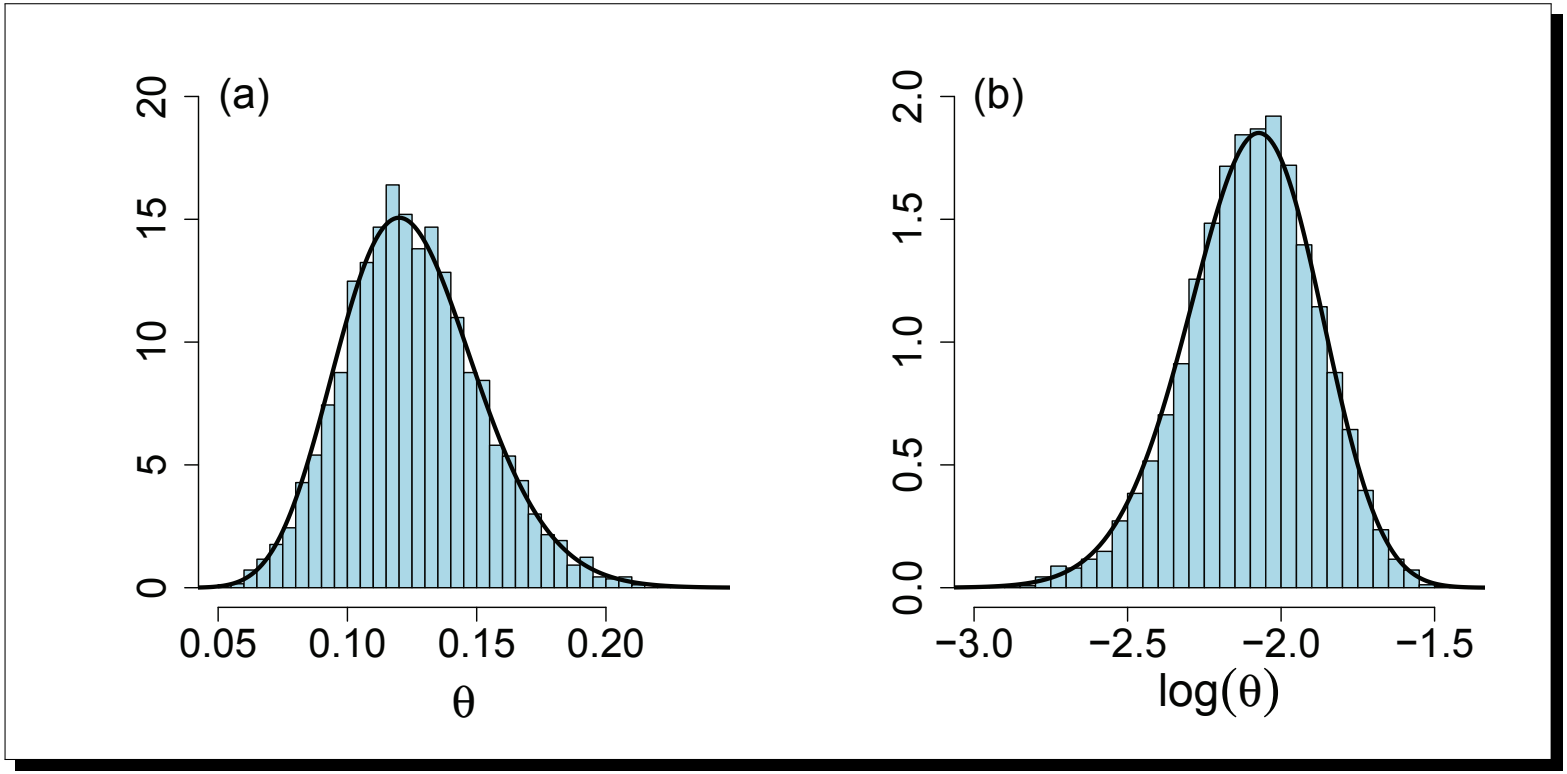
Example

- Stroke study – Monitoring safety
- Rt-PA: thrombolytic for ischemic stroke
- Outcome: complication SICH (Symptomatic intracerebral hemorrhage) in patients with acute ischemic stroke
- Fictive situation:
  - First interim analysis: **50** rt-PA patients with **10** SICHs
  - Historical data: **100** rt-PA patients with **8** ISCHs

Example

- Posterior for θ = probability of SICH with rt-PA = Beta$(19, 133)$
- $5,000$ sampled values of θ from Beta$(19, 133)$-distribution
- Posterior of $\log(\theta)$: one extra line in R-program
- Sample summary measures $\approx$ true summary measures
- 95% equal tail CI for θ: [0.0782, 0.182]
- 95% equal tail CI for $\log(\theta)$: [-2.56, -1.70]
- Approximate 95% HPD interval for θ: [0.0741, 0.179]

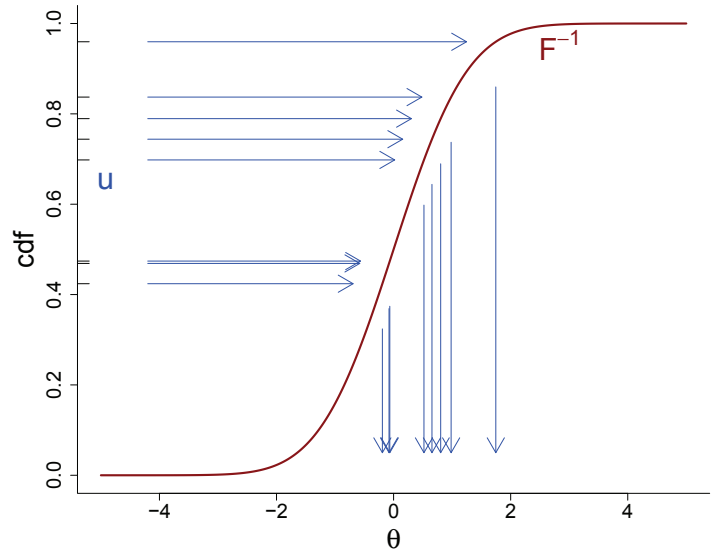
Example



General purpose sampling algorithms

- Many algorithms are available to sample from standard distributions
  - ▷ Inverse cumulative distribution function (ICDF) method:

    sample from $u \sim U(0,1)$ and calculate corresponding sampling point $x = F^{-1}(u)$



- Dedicated procedures/general purpose algorithms for non-standard distributions:

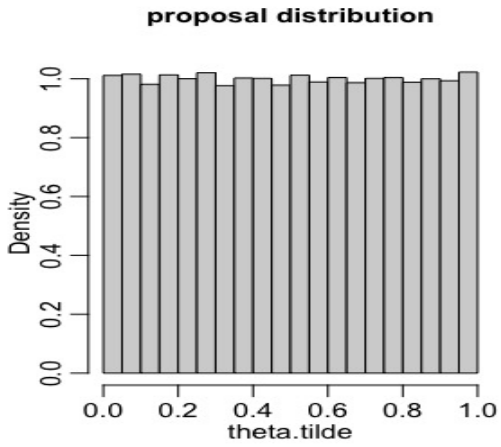
  ▷ Accept-reject (AR) algorithm:

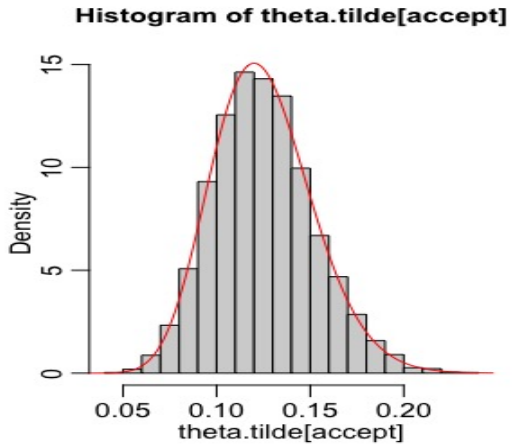
  - sample from a (user-defined) proposal distribution $\widetilde{\theta} \sim q(\theta)$
  - accept this sample from the posterior if $p(\widetilde{\theta} \mid \boldsymbol{y})/A\, q(\widetilde{\theta})$ is large (otherwise reject)

  ▷ Importance sampling if interest in $E\left[t(\theta) \mid \boldsymbol{y}\right]$
  - sample from a (user-defined) proposal distribution $\widetilde{\theta} \sim q(\theta)$
  - estimate weighted average with importance weights $w(\theta^k) = p(\theta^k \mid \boldsymbol{y})/q(\theta^k)$

Example

- ▷ Binomial likelihood + beta prior (for illustration purposes)
- ▷ proposal: uniform distribution
- ▷ $A = \max(L(\widetilde{\theta}|\boldsymbol{y})p(\widetilde{\theta})/q(\widetilde{\theta}))$
- ▷ sample additionaly u from a uniform(0,1) distribution
- ▷ accept when $p(\widetilde{\theta} \mid \boldsymbol{y})/A\, q(\widetilde{\theta}) \geq u$





In practice

- Adaptive rejection sampling algorithm
  - ▷ Builds up envelope distribution in an adaptive manner
  - ▷ Builds up squeezing density in an adaptive manner
  - ▷ Examples: Tangent methods or Derivative-free method
- Weighted sampling-resampling method to compare results from different priors

Chapter 2
More than one parameter

Joint posterior inference

- Let
  - $\boldsymbol{y}$ = sample of n independent observations
  - $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^T$
  - $L(\boldsymbol{\theta} \mid \boldsymbol{y})$
  - Multivariate prior: $p(\boldsymbol{\theta})$
- Multivariate posterior: $$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{L(\boldsymbol{\theta} \mid \boldsymbol{y}) p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta} \mid \boldsymbol{y}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$
  - Posterior mode: $\widehat{\boldsymbol{\theta}}_M$
  - Posterior mean: $\overline{\boldsymbol{\theta}}$
  - HPD region of content (1-α)

2.1 The Method of Composition

A method to yield a random sample from a multivariate distribution

- Stagewise approach
- Based on factorization of joint distribution into a marginal & several conditionals

$$p(\theta_1, \ldots, \theta_d \mid \boldsymbol{y}) = p(\theta_d \mid \boldsymbol{y})\, p(\theta_{d-1} \mid \theta_d, \boldsymbol{y}) \ \ldots \ p(\theta_1 \mid \theta_d, \ldots, \theta_2, \boldsymbol{y})$$

- Sampling approach:
  - ▷ Sample $\widetilde{\theta}_d$ from $p(\theta_d \mid \boldsymbol{y})$
  - ▷ Sample $\widetilde{\theta}_{(d-1)}$ from $p(\theta_{(d-1)} \mid \widetilde{\theta}_d, \boldsymbol{y})$
  - ▷ ...
  - ▷ Sample $\widetilde{\theta}_1$ from $p(\theta_1 \mid \widetilde{\theta}_d, \ldots, \widetilde{\theta}_2, \boldsymbol{y})$

Sampling from N(μ, σ^2), both parameters unknown

- Sampling approach from normal posterior $p(\mu, \sigma^2 \mid \boldsymbol{y}) \equiv \mathsf{N}(\mu, \sigma^2)$
- Three cases:
  - ▷ No prior knowledge
  - ▷ Historical data available
  - ▷ Expert knowledge available

Case 1: No prior knowledge on μ and σ^2

Sample from $p(\mu, \sigma^2 \mid \boldsymbol{y})$: Sample from $p(\sigma^2 \mid \boldsymbol{y})$ & Sample from $p(\mu \mid \sigma^2, \boldsymbol{y})$

1. Sample from $p(\sigma^2 \mid \boldsymbol{y})$:

   - Sample $\widetilde{\nu}^k$ from a $\chi^2(n-1)$-distribution
   - Solve $\widetilde{\sigma}^{2k}$ in $(n-1)s^2/\widetilde{\sigma}^{2k} = \widetilde{\nu}^k$

2. Sample from $p(\mu \mid \sigma^2, \boldsymbol{y})$:

   - Sample $\widetilde{\mu}^k$ from a N($\overline{y}$, $\widetilde{\sigma}^{2k}/n$)-distribution

$\Rightarrow \widetilde{\mu}^1, \ldots, \widetilde{\mu}^K =$ random sample from $p(\mu \mid \boldsymbol{y})$ ($t_{n-1}(\overline{y}, s^2/n)$-distribution)

Case 1 (continued)

To sample from the posterior predictive distribution $p(\widetilde{y} \mid \boldsymbol{y})$, 2 approaches:

1. Sample directly from $t_{n-1}\left[\overline{y}, s^2\left(1+\frac{1}{n}\right)\right]$-distribution

2. Use **Method of Composition**

   ▷ Sample $\widetilde{\sigma}^{2k}$ from Inv-$\chi^2(\sigma^2 \mid n-1, s^2)$

   ▷ Sample $\widetilde{\mu}^k$ from $\mathsf{N}(\mu \mid \overline{y}, \widetilde{\sigma}^{2k}/n)$

   ▷ Sample $\widetilde{y}^k$ from $\mathsf{N}(y \mid \widetilde{\mu}^k, \widetilde{\sigma}^{2k})$

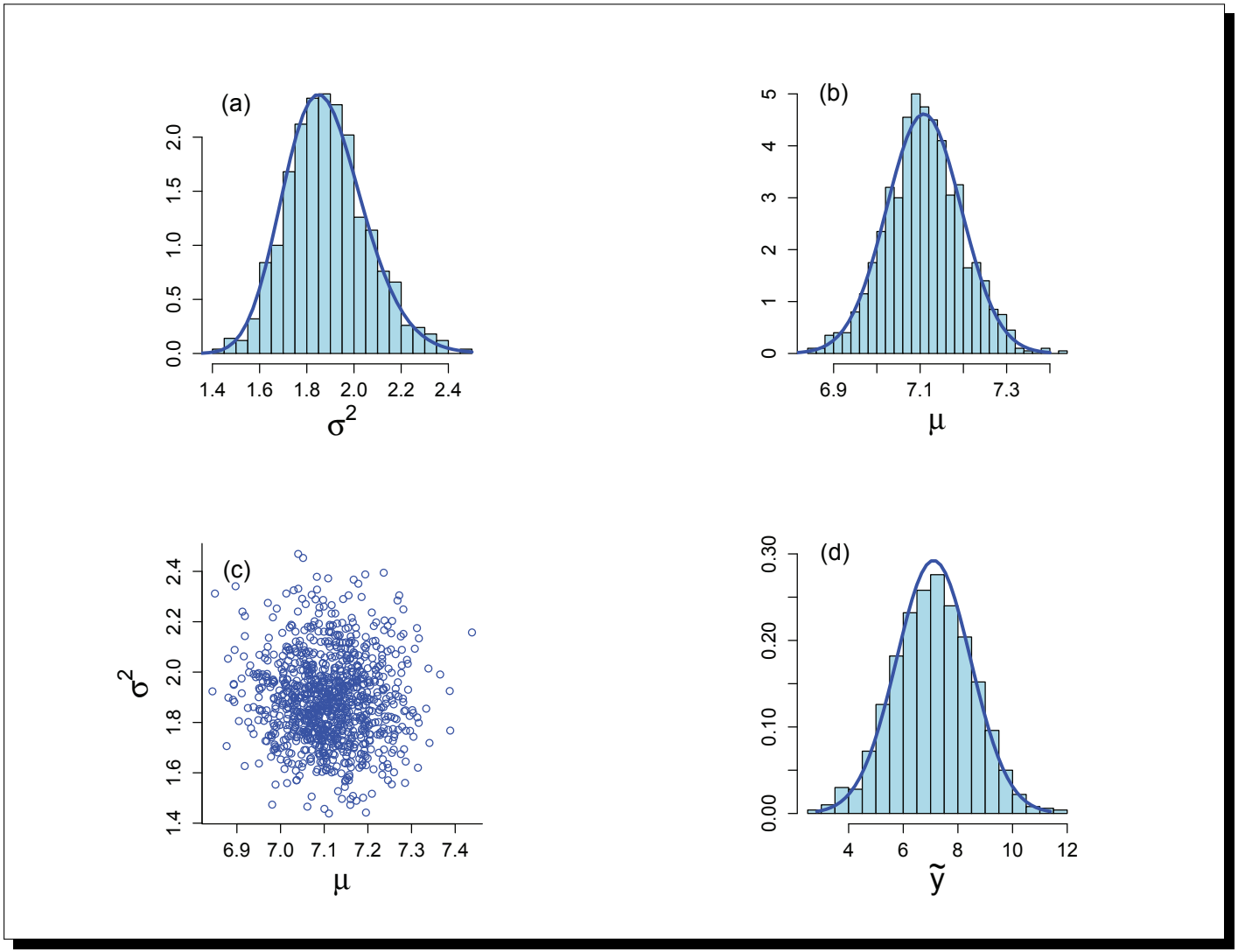
Example

- ▷ retrospective study predicting the incidence of common bile duct (CBD) stones in patients with gallstone disease
- ▷ study on a prospective set of 250 'healthy' patients.
- ▷ outcome: serum alkaline phosphatase (SAP)
- ▷ $y_i = 100/\sqrt{SAP_i}$ has a Gaussian distribution

Example: Sampling the posterior with NI prior

- Sampled posterior distributions on next page ($K = 1000$)
- Posterior mean (95% confidence interval)
  - μ: 7.11 ([7.106, 7.117])
  - σ^2: 1.88 ([1.869, 1.890])
- 95% equal tail CI
  - μ: [6.95, 7.27]
  - σ^2: [1.58, 2.23]

Example (continued)



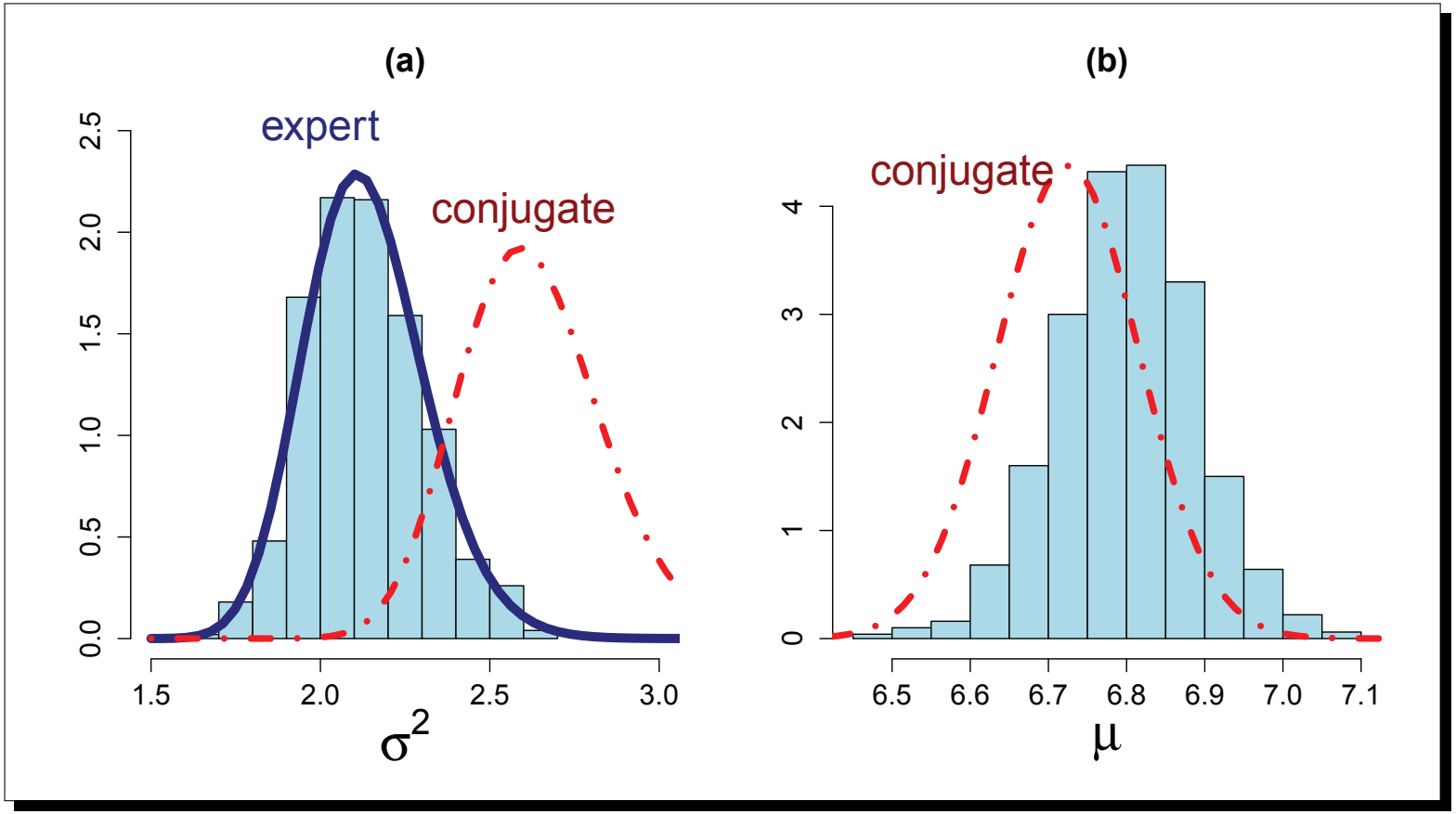
Case 2: Historical data are available

Same procedure as before!

Case 3: Expert knowledge is available

- This is no longer a conjugate setting
- **Problem**: $p(\sigma^2 \mid \boldsymbol{y})$ **does not have a known distribution**
- For a given $\widetilde{\sigma}^2$, sampling $\widetilde{\mu}$ is straightforward
- Use weighted resampling to sample from $p(\sigma^2 \mid \boldsymbol{y})$

Example (continued)



Chapter 3
Markov chain Monte Carlo sampling

Aims:

▷ Introduce the sampling approach(es) that revolutionized Bayesian approach

- ▷ Solving the posterior distribution analytically is often not feasible due to the difficulty in determining the integration constant
- ▷ Computing the integral using numerical integration methods is a practical alternative if only a few parameters are involved

⇒ **New computational approach is needed**

- ▷ **Sampling is the way to go!**
- ▷ With **Markov chain Monte Carlo (MCMC) methods**:
  1. Gibbs sampler
  2. Metropolis-(Hastings) algorithm

MCMC approaches have revolutionized Bayesian methods!

3.1 The Gibbs sampler

- **Gibbs Sampler**: introduced by Geman and Geman (1984) in the context of image-processing for the estimation of the parameters of the Gibbs distribution

- Gelfand and Smith (1990) introduced Gibbs sampling to tackle complex estimation problems in a Bayesian manner

3.1.1 The bivariate Gibbs sampler

Method of Composition:

- $p(\theta_1, \theta_2 \mid \boldsymbol{y})$ is completely determined by:
  - ▷ **marginal** $p(\theta_2 \mid \boldsymbol{y})$
  - ▷ **conditional** $p(\theta_1 \mid \theta_2, \boldsymbol{y})$
- Split-up yields a simple way to sample from joint distribution

Gibbs sampling:

- $p(\theta_1, \theta_2 \mid \boldsymbol{y})$ is completely determined by:
  - ▷ **conditional** $p(\theta_2 \mid \theta_1, \boldsymbol{y})$
  - ▷ **conditional** $p(\theta_1 \mid \theta_2, \boldsymbol{y})$

- Property yields another simple way to sample from joint distribution:
  - ▷ Take starting values θ_1^0 and θ_2^0 (only 1 is needed)
  - ▷ Given θ_1^k and θ_2^k at iteration k, generate the $(k+1)$-th value according to iterative scheme:
    1. Sample $\theta_1^{(k+1)}$ from $p(\theta_1 \mid \theta_2^k, \boldsymbol{y})$
    2. Sample $\theta_2^{(k+1)}$ from $p(\theta_2 \mid \theta_1^{(k+1)}, \boldsymbol{y})$

Result of Gibbs sampling:

- Chain of vectors: $\boldsymbol{\theta}^k = (\theta_1^k, \theta_2^k)^T, k = 1, 2, \ldots$
  - Consists of **dependent** elements
  - **Markov property**: $p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^k, \boldsymbol{\theta}^{(k-1)}, \ldots, \boldsymbol{y}) = p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^k, \boldsymbol{y})$
- Chain depends on starting value + initial portion/burn-in part must **be discarded**
- Under mild conditions: sample from the posterior distribution = target distribution

$\Rightarrow$ From k_0 on: summary measures calculated from the chain consistently estimate the true posterior measures

Gibbs sampler is called a Markov chain Monte Carlo method

Example: SAP study – Gibbs sampling the posterior with NI priors

- Example: sampling from posterior distribution of the normal likelihood based on 250 *alp* measurements of 'healthy' patients with NI prior for both parameters

- Now using Gibbs sampler based on $y = 100/\sqrt{alp}$

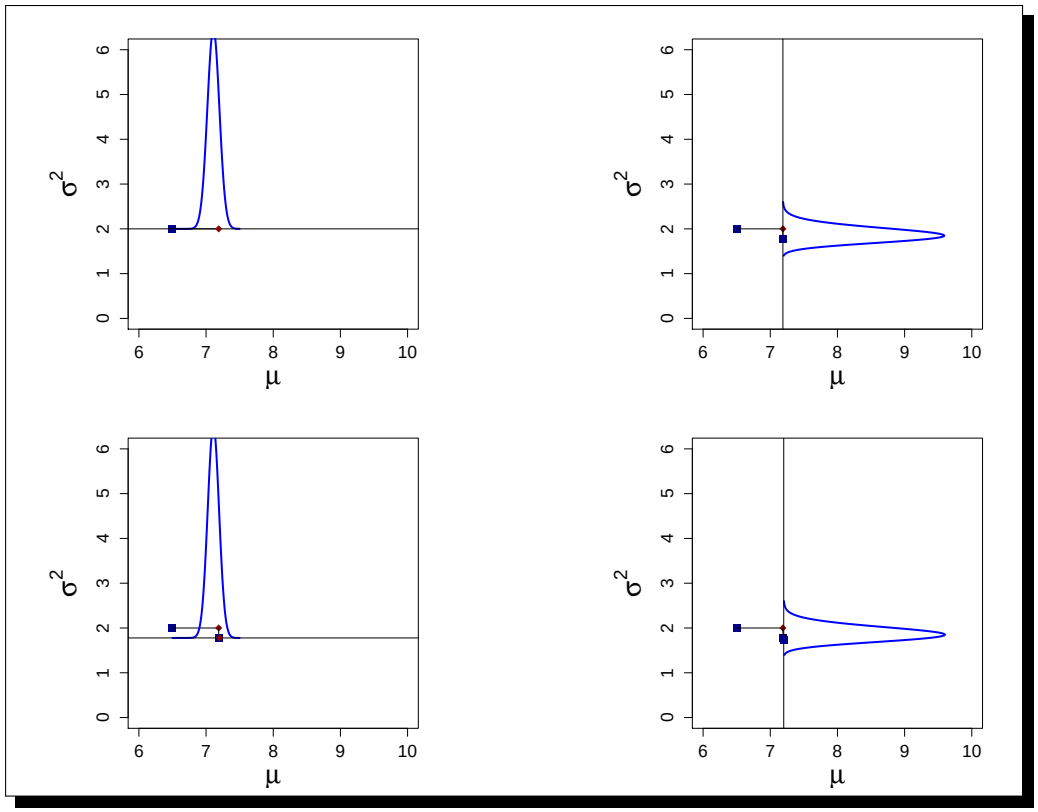
- Determine two conditional distributions:

  1. $p(\mu \mid \sigma^2, \boldsymbol{y})$: $\mathsf{N}(\mu \mid \bar{y}, \sigma^2/n)$
  2. $p(\sigma^2 \mid \mu, \boldsymbol{y})$: $\mathsf{Inv}-\chi^2(\sigma^2 \mid n, s_\mu^2)$ with $s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$

- Iterative procedure: At iteration $(k+1)$

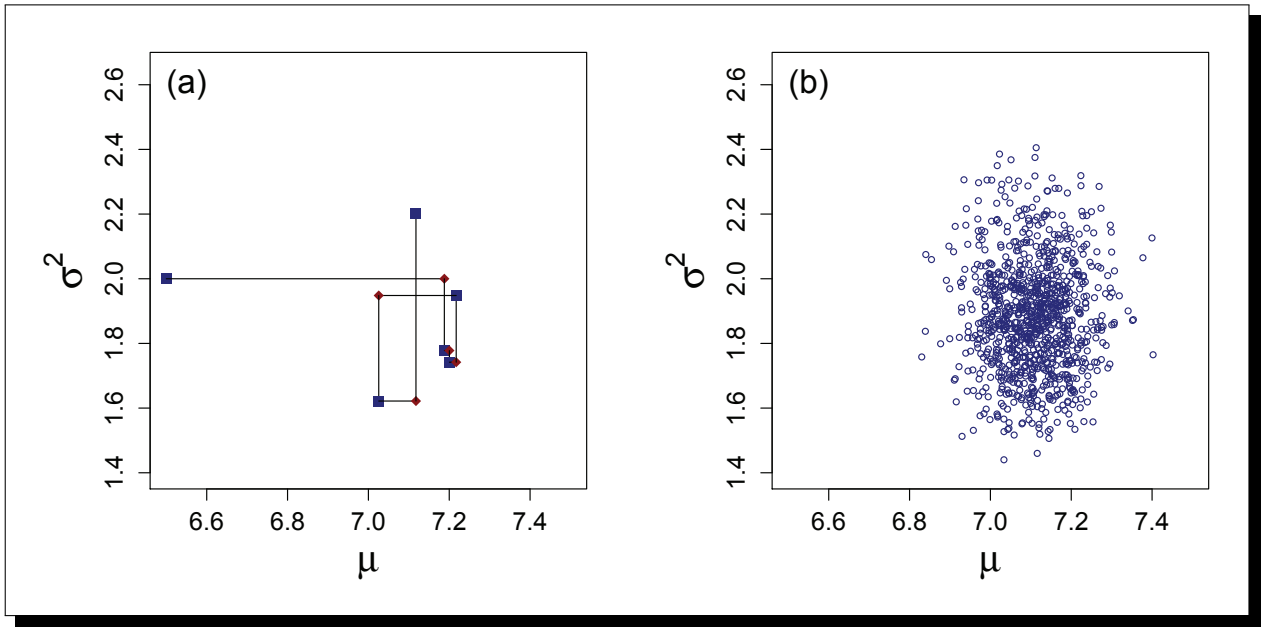
  1. Sample $\mu^{(k+1)}$ from $\mathsf{N}(\bar{y}, (\sigma^2)^k/n)$
  2. Sample $(\sigma^2)^{(k+1)}$ from $\mathsf{Inv}-\chi^2(n, s^2_{\mu^{(k+1)}})$

Gibbs sampling:



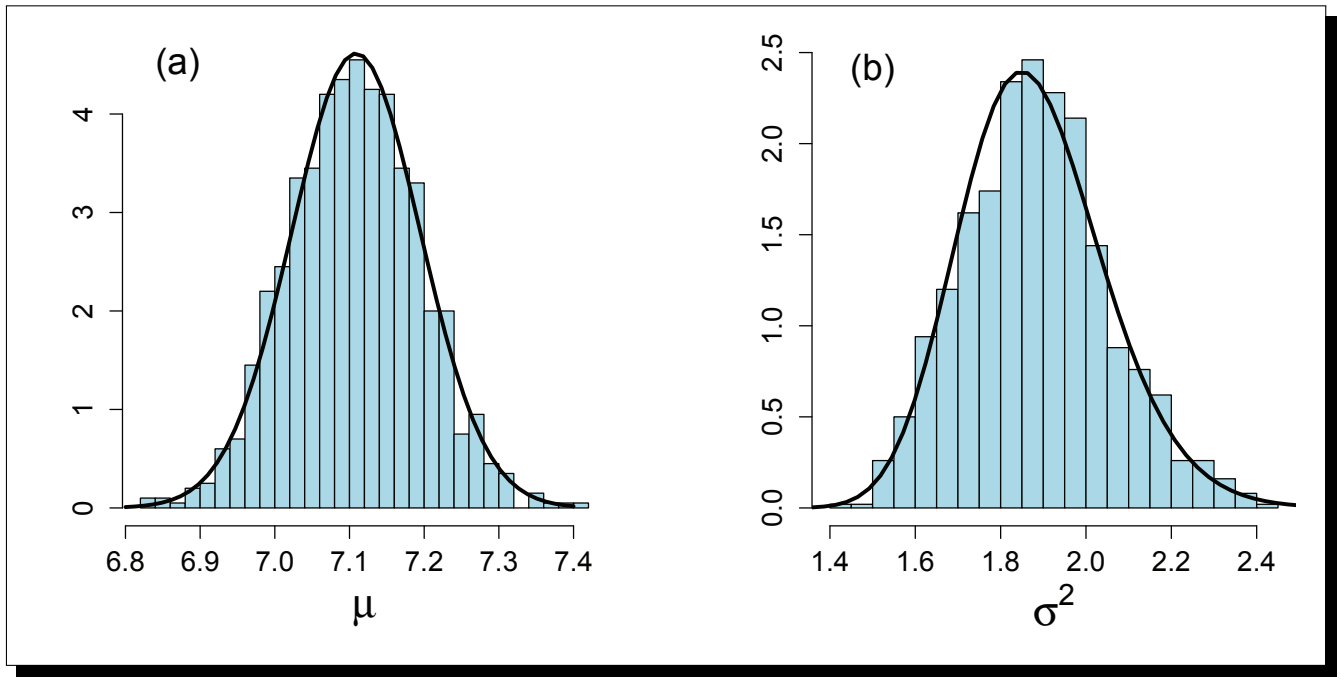
- Sampling from conditional density of μ given σ^2
- Sampling from conditional density of σ^2 given μ

Gibbs sampling path and sample from joint posterior:



- Zigzag pattern in the (μ, σ^2)-plane
- 1 complete step = 2 substeps (blue=genuine element)
- Burn-in = 500, total chain = 1,500

Posterior distributions:



Solid line = true posterior distribution

Example: SAP study – Gibbs sampling the posterior with I priors

- Example: now with independent informative priors (semi-conjugate prior)
  - $\mu \sim \mathsf{N}(\mu_0, \sigma_0^2)$
  - $\sigma^2 \sim \mathsf{Inv}-\chi^2(\nu_0, \tau_0^2)$

- Posterior:

$$
\begin{aligned}
p(\mu, \sigma^2 \mid \boldsymbol{y}) \;&\propto\; \frac{1}{\sigma_0} \mathsf{e}^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \\
&\times\; \left(\sigma^2\right)^{-(\nu_0/2+1)} \mathsf{e}^{-\nu_0\,\tau_0^2/2\sigma^2} \\
&\times\; \frac{1}{\sigma^n} \prod_{i=1}^{n} \mathsf{e}^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\
&\propto\; \prod_{i=1}^{n} \mathsf{e}^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}\; \mathsf{e}^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}\; \left(\sigma^2\right)^{-\left(\frac{n+\nu_0}{2}+1\right)} \mathsf{e}^{-\nu_0\,\tau_0^2/2\sigma^2}
\end{aligned}
$$

Conditional distributions:

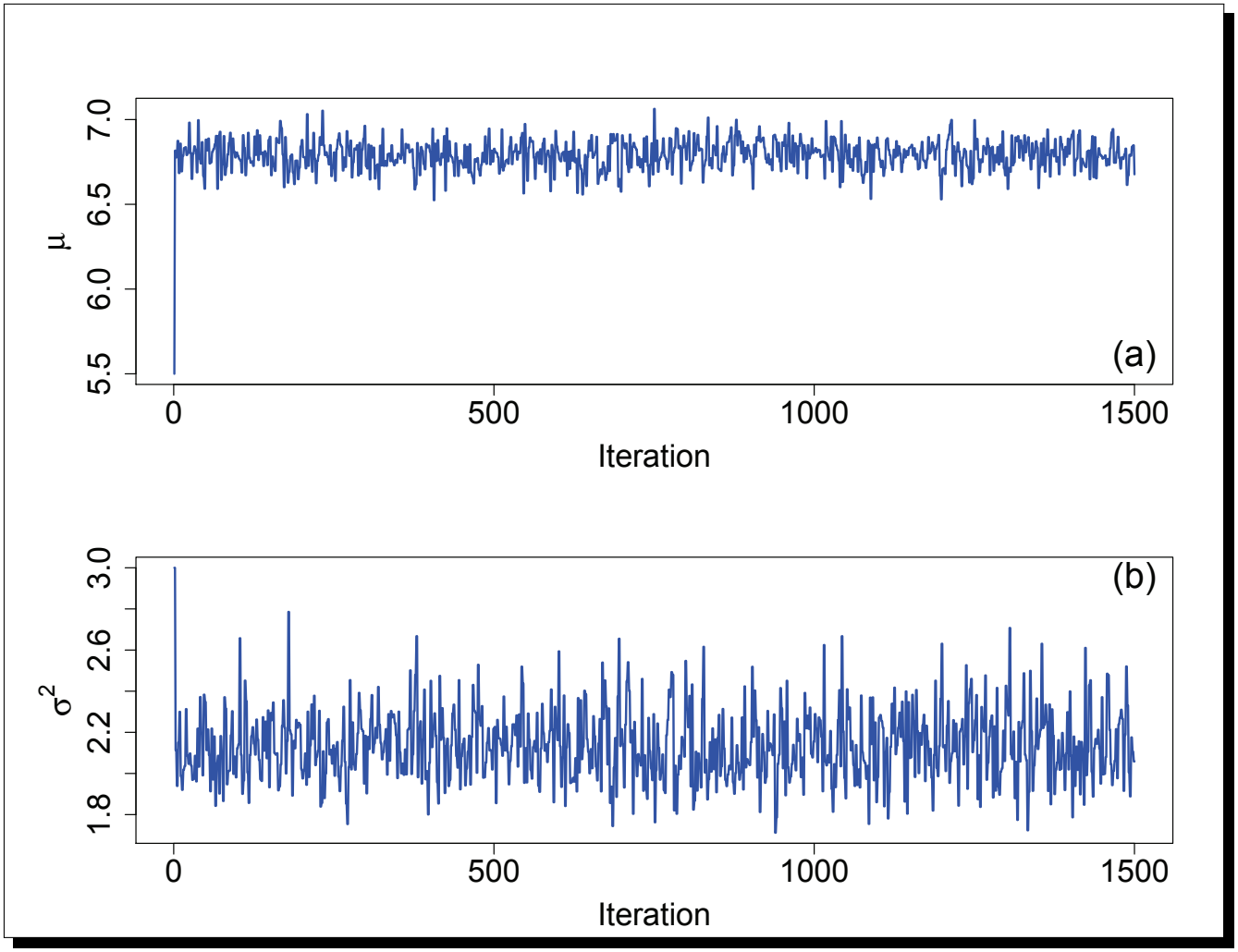
- Determine two conditional distributions:

  1. $p(\mu \mid \sigma^2, \boldsymbol{y})$: $\prod_{i=1}^{n} \mathrm{e}^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \mathrm{e}^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$ $\left(\mathrm{N}\left(\overline{\mu}^k, (\overline{\sigma}^2)^k\right)\right)$
  2. $p(\sigma^2 \mid \mu, \boldsymbol{y})$: $\mathrm{Inv}-\chi^2\left(\nu_0+n, \frac{\sum_{i=1}^{n}(y_i-\mu)^2+\nu_0\tau_0^2}{\nu_0+n}\right)$

- Iterative procedure: At iteration $(k+1)$

  1. Sample $\mu^{(k+1)}$ from $\mathrm{N}\left(\overline{\mu}^k, (\overline{\sigma}^2)^k\right)$
  2. Sample $(\sigma^2)^{(k+1)}$ from $\mathrm{Inv}-\chi^2\left(\nu_0+n, \frac{\sum_{i=1}^{n}(y_i-\mu^{(k+1)})^2+\nu_0\tau_0^2}{\nu_0+n}\right)$

Trace plots:


μ
5.5
6.0
6.5
7.0
(a)
0
500
1000
1500
Iteration
σ2
1.8
2.2
2.6
3.0
(b)
0
500
1000
1500
Iteration

3.1.2 The general Gibbs sampler

Starting position $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_d^0)^T$

Multivariate version of the Gibbs sampler:

Iteration $(k+1)$:

1. Sample $\theta_1^{(k+1)}$ from $p(\theta_1 \mid \theta_2^k, \ldots, \theta_{(d-1)}^k, \theta_d^k, \boldsymbol{y})$

2. Sample $\theta_2^{(k+1)}$ from $p(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^k, \ldots, \theta_d^k, \boldsymbol{y})$

⋮

d. Sample $\theta_d^{(k+1)}$ from $p(\theta_d \mid \theta_1^{(k+1)}, \ldots, \theta_{(d-1)}^{(k+1)}, \boldsymbol{y})$

- Full conditional distributions: $p(\theta_j \mid \theta_1^k, \ldots, \theta_{(j-1)}^k, \theta_{(j+1)}^k, \ldots, \theta_{(d-1)}^k, \theta_d^k, \boldsymbol{y})$
- Also called: full conditionals
- Under mild regularity conditions:
  $\boldsymbol{\theta}^k, \boldsymbol{\theta}^{(k+1)}, \ldots$ ultimately are observations from the posterior distribution

With the help of advanced sampling algorithms (AR, ARS, etc) sampling the full conditionals is done based on the prior $\times$ likelihood

Example: Osteoporosis study – Using the Gibbs sampler

Bayesian linear regression model with NI priors:

- ▷ Regression model: $tbbmc_i = \beta_0 + \beta_1 bmi_i + \varepsilon_i \;\; (i = 1, \ldots, n = 234)$
- ▷ Priors: $p(\beta_0, \beta_1, \sigma^2) \propto \sigma^{-2}$
- ▷ Notation: $\boldsymbol{y} = (tbbmc_1, \ldots, tbbmc_{234})^T$, $\boldsymbol{x} = (bmi_1, \ldots, bmi_{234})^T$

Full conditionals:

$$
\begin{aligned}
p(\sigma^2 \mid \beta_0, \beta_1, \boldsymbol{y}) &= \text{Inv} - \chi^2(n, s^2_{\boldsymbol{\beta}}) \\
p(\beta_0 \mid \sigma^2, \beta_1, \boldsymbol{y}) &= \text{N}(r_{\beta_1}, \sigma^2/n) \\
p(\beta_1 \mid \sigma^2, \beta_0, \boldsymbol{y}) &= \text{N}(r_{\beta_0}, \sigma^2/\boldsymbol{x}^T\boldsymbol{x})
\end{aligned}
$$

with

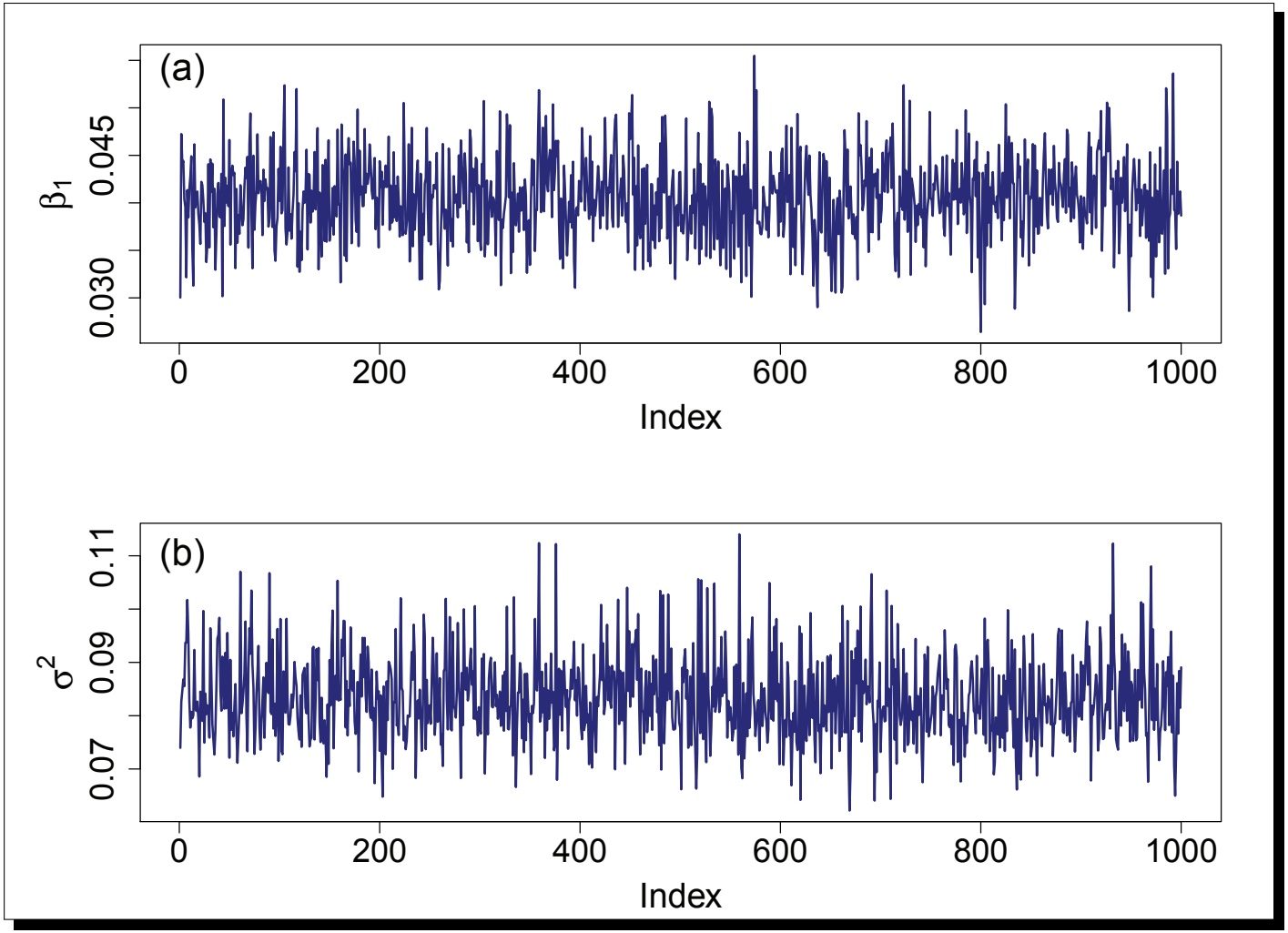
$$
\begin{aligned}
s^2_{\boldsymbol{\beta}} &= \tfrac{1}{n}\textstyle\sum(y_i - \beta_0 - \beta_1 x_i)^2 \\
r_{\beta_1} &= \tfrac{1}{n}\textstyle\sum(y_i - \beta_1 x_i) \\
r_{\beta_0} &= \textstyle\sum(y_i - \beta_0)\, x_i / \boldsymbol{x}^T\boldsymbol{x}
\end{aligned}
$$

Comparison with Method of Composition:

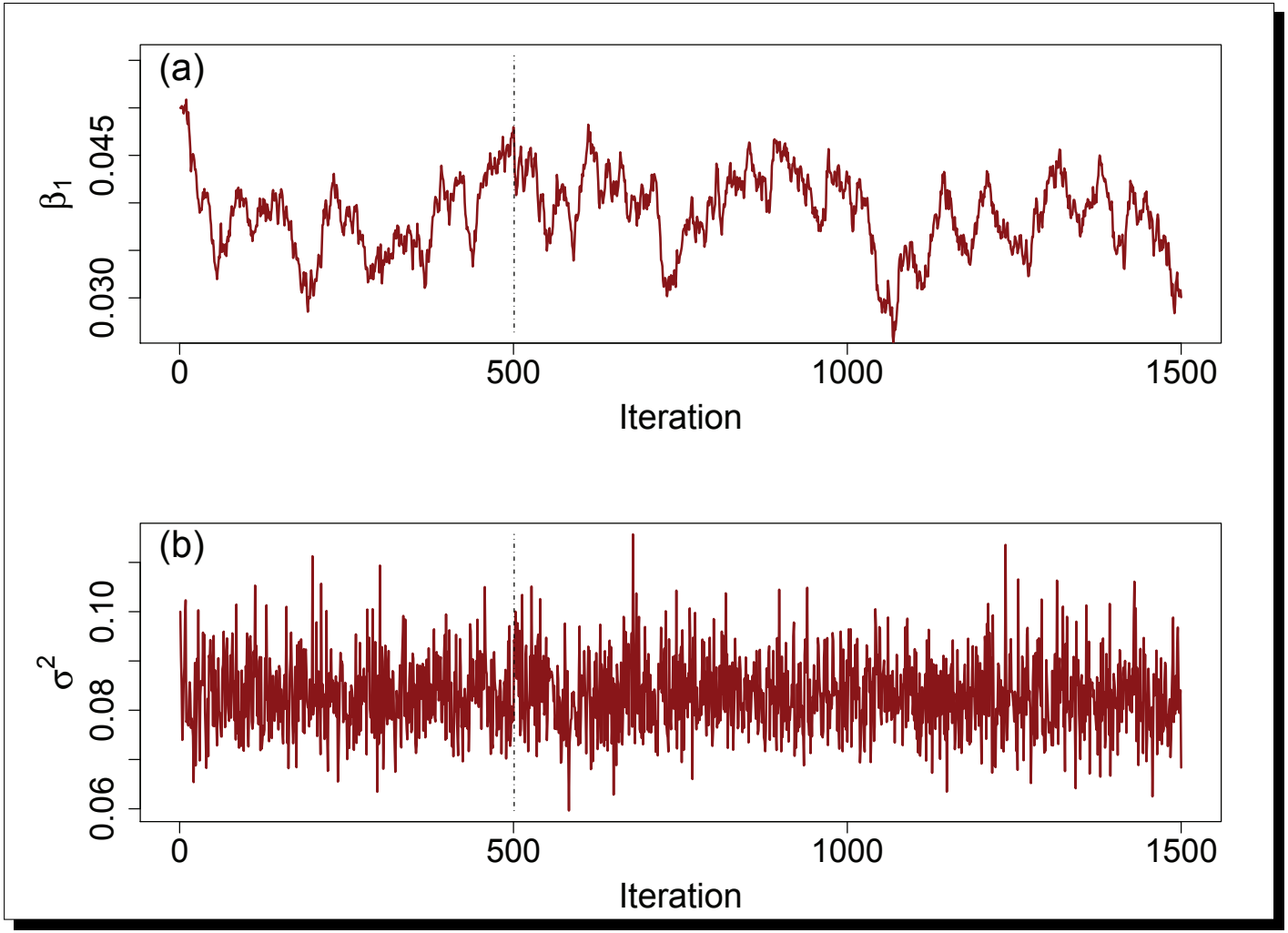
Parameter	Method of Composition						
	2.5%	25%	50%	75%	97.5%	Mean	SD
β_0	0.57	0.74	0.81	0.89	1.05	0.81	0.12
β_1	0.032	0.038	0.040	0.043	0.049	0.040	0.004
σ^2	0.069	0.078	0.083	0.088	0.100	0.083	0.008
	Gibbs sampler						
	2.5%	25%	50%	75%	97.5%	Mean	SD
β_0	0.67	0.77	0.84	0.91	1.10	0.77	0.11
β_1	0.030	0.036	0.040	0.042	0.046	0.039	0.0041
σ^2	0.069	0.077	0.083	0.088	0.099	0.083	0.0077

- Method of Composition = 1,000 independently sampled values
- Gibbs sampler: burn-in = 500, total chain = 1,500

Index plot from Method of Composition:



Trace plot from Gibbs sampler:



Trace versus index plot:

Comparison of index plot with trace plot shows:

- σ^2: index plot and trace plot **similar** $\Rightarrow$ (almost) **independent** sampling
- β_1: trace plot shows **slow mixing** $\Rightarrow$ quite **dependent** sampling

$\Rightarrow$ Method of Composition and Gibbs sampling: **similar** posterior measures of σ^2

$\Rightarrow$ Method of Composition and Gibbs sampling: **less similar** posterior measures of β_1

Autocorrelation:

- ▷ Autocorrelation of lag 1: correlation of β_1^k with $\beta_1^{(k-\mathbf{1})}$ (k=1, ...)
- ▷ Autocorrelation of lag 2: correlation of β_1^k with $\beta_1^{(k-\mathbf{2})}$ (k=1, ...)

...

- ▷ Autocorrelation of lag m: correlation of β_1^k with $\beta_1^{(k-\mathbf{m})}$ (k=1, ...)

High autocorrelation:

$\Rightarrow$ **burn-in part is larger** $\Rightarrow$ takes longer to forget initial positions

$\Rightarrow$ **remaining part** needs to be **longer** to obtain stable posterior measures

3.2 The Metropolis(-Hastings) algorithm

Metropolis-Hastings (MH) algorithm = general Markov chain Monte Carlo technique to sample from the posterior distribution but **does not require full conditionals**

- Special case: Metropolis algorithm proposed by Metropolis in 1953
- General case: Metropolis-Hastings algorithm proposed by Hastings in 1970
- Became popular only after introduction of Gelfand & Smith's paper (1990)
- Further generalization: Reversible Jump MCMC algorithm by Green (1995)

3.2.1 The Metropolis algorithm

Sketch of algorithm:

- New positions are proposed by a **proposal density** q
- Proposed positions will be:
  - ▷ **Accepted:**
    - ○ Proposed location has higher posterior probability: with probability **1**
    - ○ Otherwise: with probability proportional to ratio of posterior probabilities
  - ▷ **Rejected:**
    - ○ Otherwise
- Algorithm satisfies again Markov property $\Rightarrow$ MCMC algorithm
- Similarity with AR algorithm

Metropolis algorithm:

Chain is at $\boldsymbol{\theta}^k \Rightarrow$ Metropolis algorithm samples value $\boldsymbol{\theta}^{(k+1)}$ as follows:

1. Sample a candidate $\widetilde{\boldsymbol{\theta}}$ from the **symmetric** proposal density $q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\theta}^k$
2. The next value $\boldsymbol{\theta}^{(k+1)}$ will be equal to:
   - $\widetilde{\boldsymbol{\theta}}$ with probability $\alpha(\boldsymbol{\theta}^k, \widetilde{\boldsymbol{\theta}})$ (accept proposal),
   - $\boldsymbol{\theta}^k$ otherwise (reject proposal),

   with

$$\alpha(\boldsymbol{\theta}^k, \widetilde{\boldsymbol{\theta}}) = \min\left(r = \frac{p(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^k \mid \boldsymbol{y})}, 1\right)$$

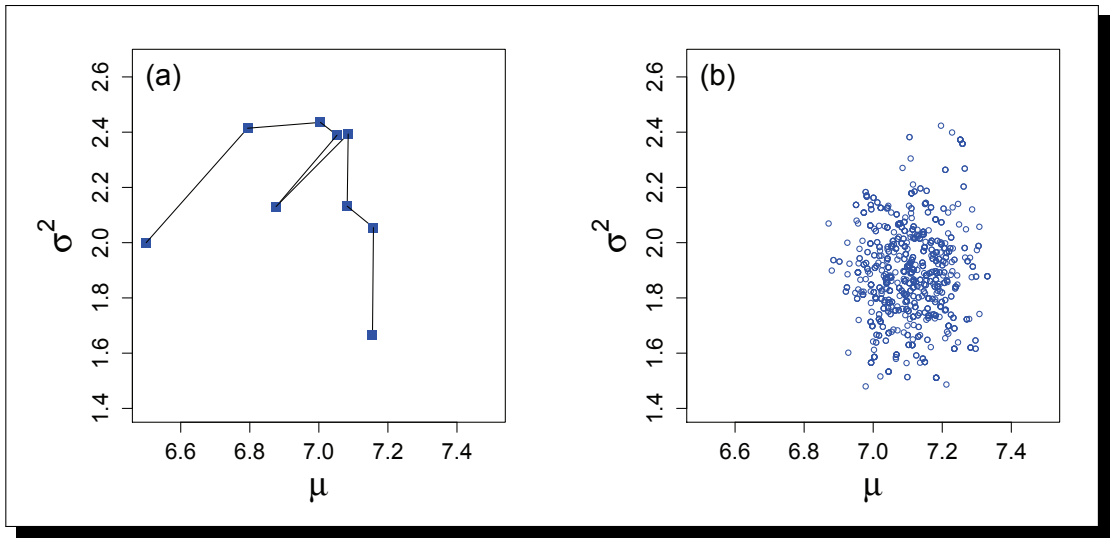
Function $\alpha(\boldsymbol{\theta}^k, \widetilde{\boldsymbol{\theta}}) =$ probability of a move

The MH algorithm only requires the product of the prior and the likelihood to sample from the posterior

Example: SAP study – Metropolis algorithm for NI prior case

Settings as in before, now apply Metropolis algorithm:

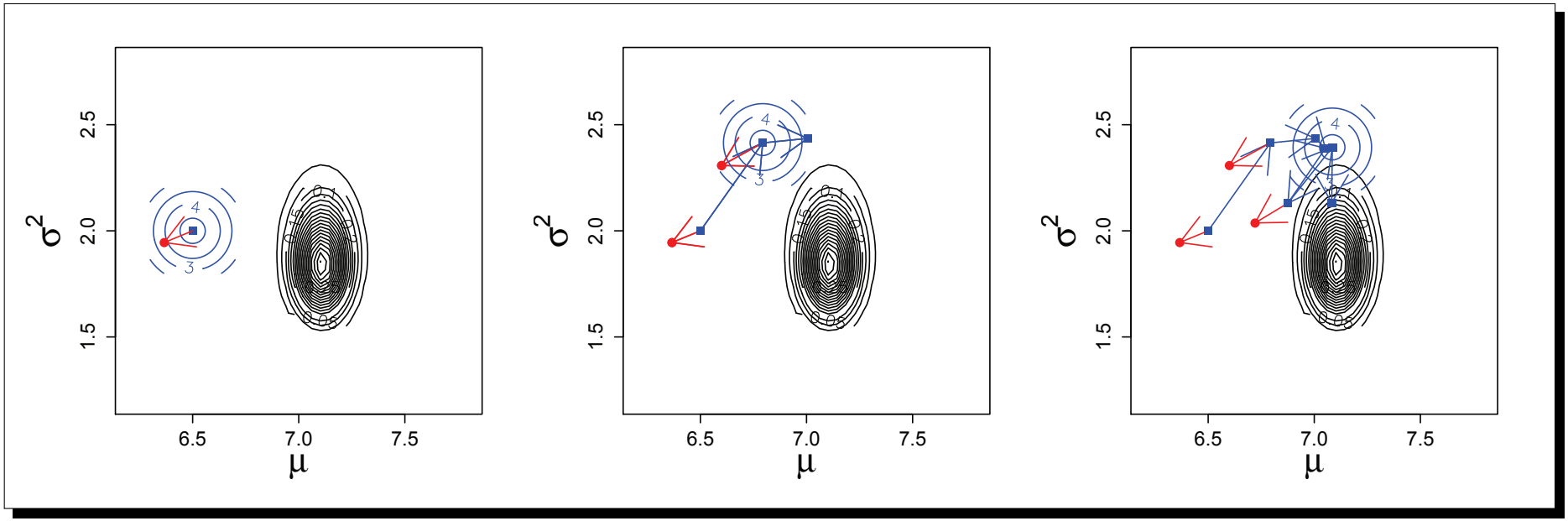
▷ Proposal density: $\mathrm{N}(\boldsymbol{\theta}^k, \Sigma)$ with $\boldsymbol{\theta}^k = (\mu^k, (\sigma^2)^k)^T$ and $\Sigma = \mathrm{diag}(0.03, 0.03)$



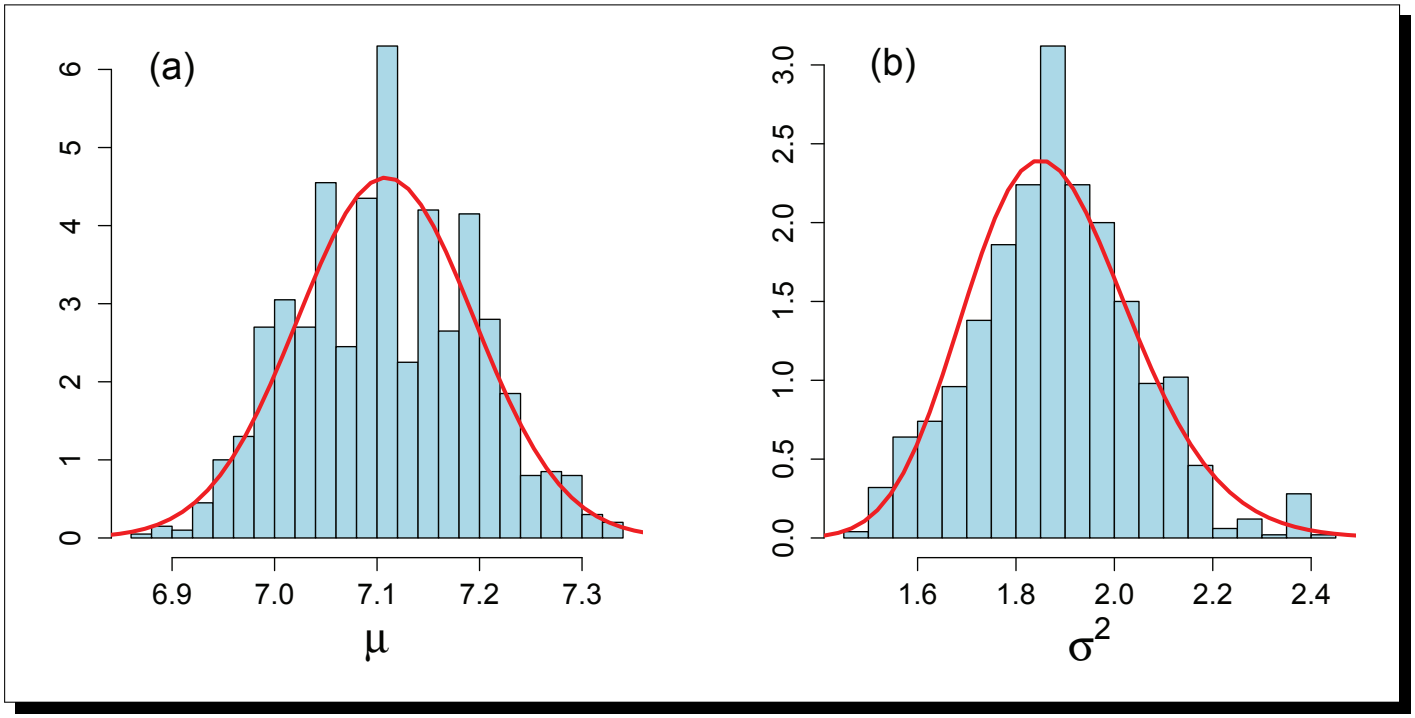
◦ Jumps to any location in the (μ, σ^2)-plane

◦ Burn-in = 500, total chain = 1,500

MH-sampling:

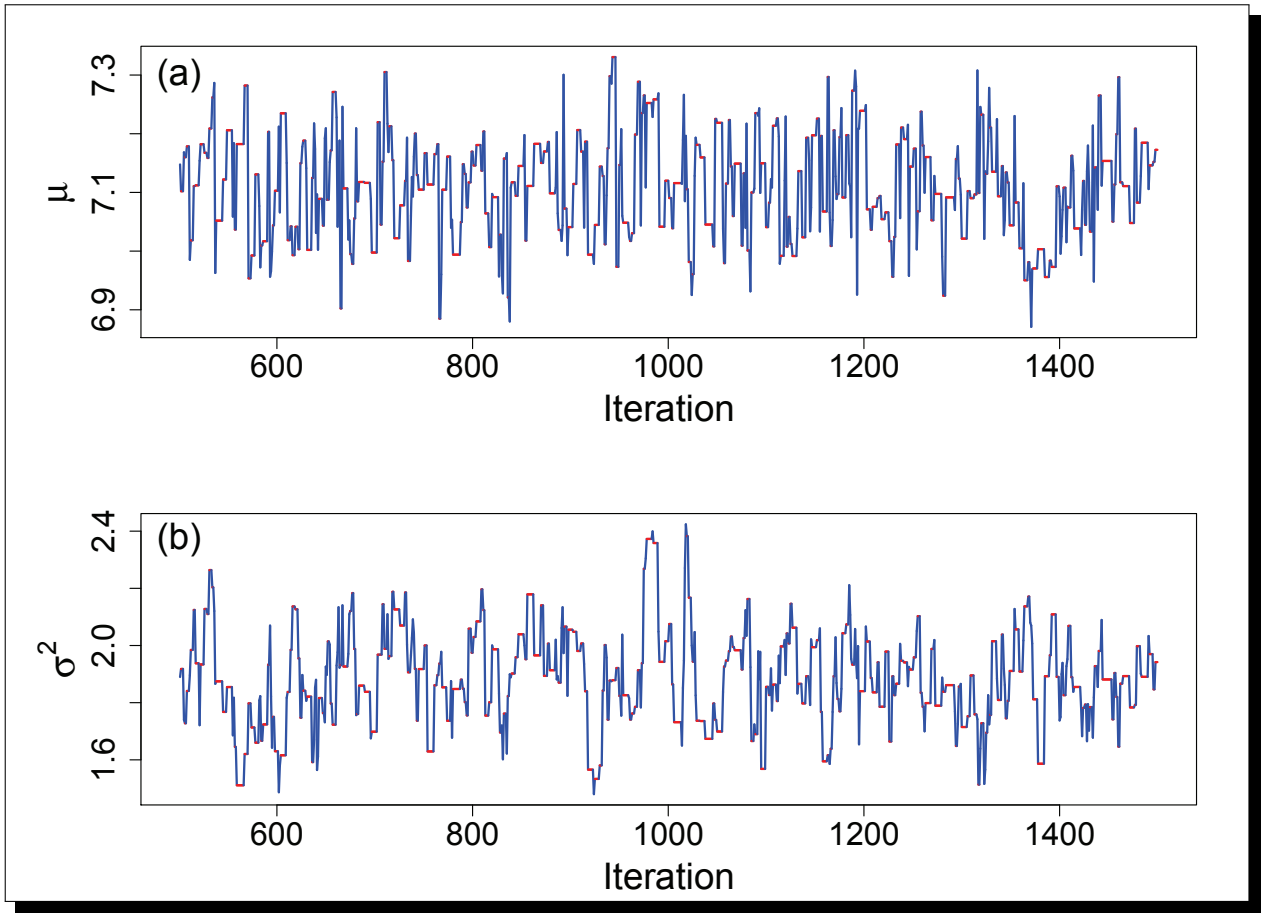


Marginal posterior distributions:



- Acceptance rate = **40%**
- Burn-in = 500, total chain = 1,500

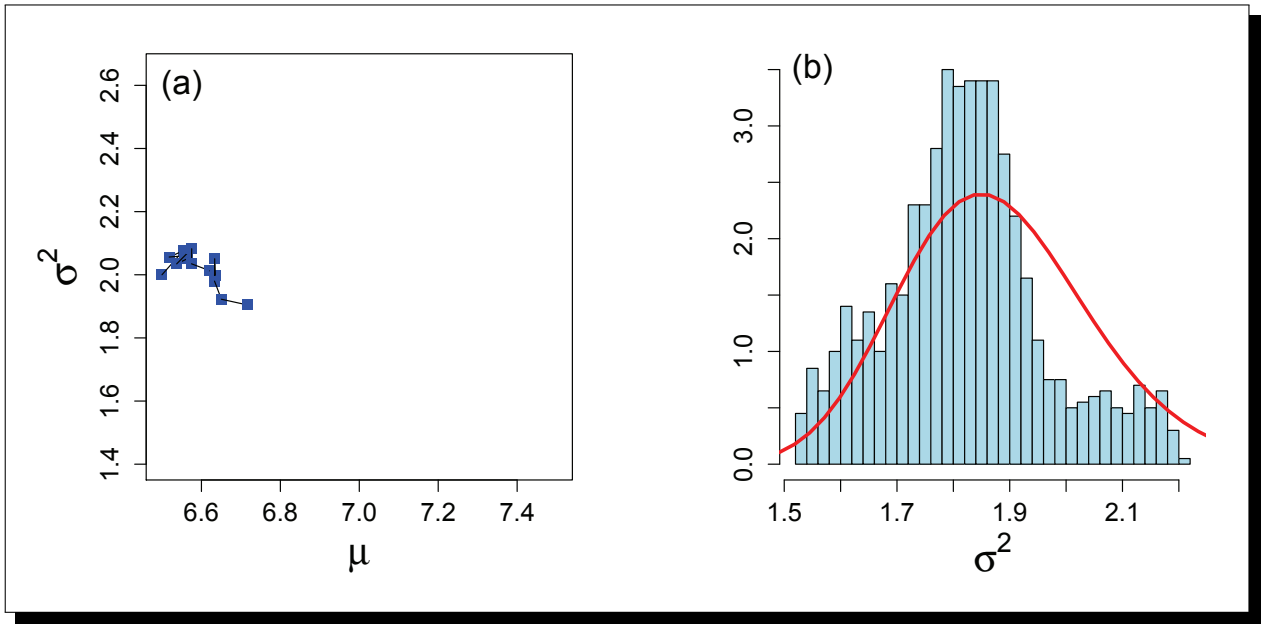
Trace plots:



- Accepted moves = blue color, rejected moves = red color

Second choice of proposal density:

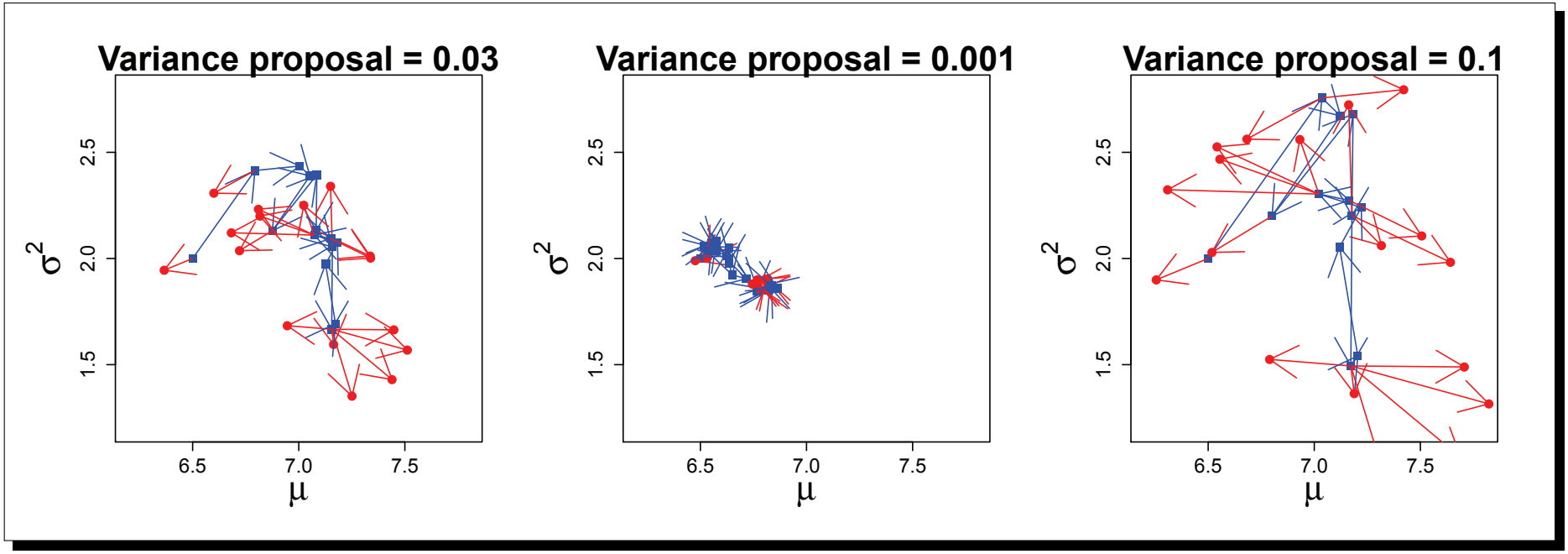
▷ Proposal density: $\mathrm{N}(\boldsymbol{\theta}^k, \Sigma)$ with $\boldsymbol{\theta}^k = (\mu^k, (\sigma^2)^k)^T$ and $\Sigma = \mathrm{diag}(0.001, 0.001)$



◦ Acceptance rate = **84%**

◦ Poor approximation of true distribution

Accepted + rejected positions:



Problem:

> What should be the acceptance rate for a good Metropolis algorithm?

From theoretical work + simulations:

- Acceptance rate: 45% for $d = 1$ and $\approx 24\%$ for $d > 1$

3.2.2 The Metropolis-Hastings algorithm

Metropolis-Hastings algorithm:

Chain is at $\boldsymbol{\theta}^k \Rightarrow$ Metropolis-Hastings algorithm samples value $\boldsymbol{\theta}^{(k+1)}$ as follows:

1. Sample a candidate $\widetilde{\boldsymbol{\theta}}$ from the (**asymmetric**) proposal density $q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\theta}^k$
2. The next value $\boldsymbol{\theta}^{(k+1)}$ will be equal to:
   - $\widetilde{\boldsymbol{\theta}}$ with probability $\alpha(\boldsymbol{\theta}^k, \widetilde{\boldsymbol{\theta}})$ (accept proposal),
   - $\boldsymbol{\theta}^k$ otherwise (reject proposal),

   with

$$\alpha(\boldsymbol{\theta}^k, \widetilde{\boldsymbol{\theta}}) = \min\left(r = \frac{p(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{y})\, q(\boldsymbol{\theta}^k \mid \widetilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}^k \mid \boldsymbol{y})\, q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^k)}, 1\right)$$

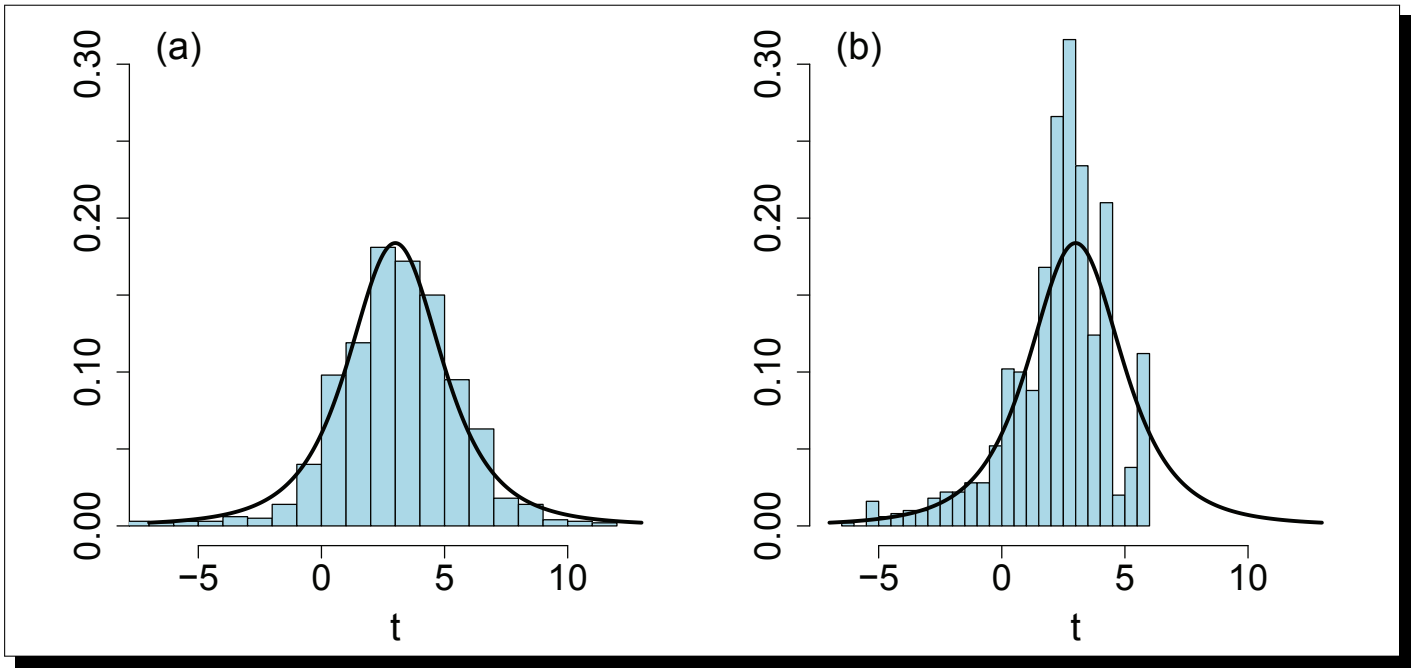
- **Reversibility condition**: Probability of move from $\boldsymbol{\theta}$ to $\widetilde{\boldsymbol{\theta}}$ = probability of move from $\widetilde{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$
- **Reversible chain**: chain satisfying reversibility condition
- Example asymmetric proposal density: $q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^k) \equiv q(\widetilde{\boldsymbol{\theta}})$ (Independent MH algorithm)

Example: Sampling a t-distribution using Independent MH algorithm

Target distribution : $t_3(3, 2^2)$-distribution

(a) Independent MH algorithm with proposal density $\mathrm{N}(3, 4^2)$

(b) Independent MH algorithm with proposal density $\mathrm{N}(3, 2^2)$



3.2.3 Review of Metropolis(-Hastings) approaches

- The Random-Walk Metropolis(-Hastings) algorithm
- The Independent Metropolis-Hastings algorithm
- The Block Metropolis-Hastings algorithm
- The Reversible Jump MCMC (RJMCMC) algorithm

The Random-Walk Metropolis(-Hastings) algorithm

- Proposal density: $q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = q(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$. When $q(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \equiv q(|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|)$ proposal density is symmetric and gives the Metropolis algorithm
  - ▷ Multivariate normal density: WinBUGS & SAS® procedures
  - ▷ Multivariate t-distribution: SAS® PROC MCMC for long tailed posteriors
- Acceptance rate: 45% for $d = 1$ and 23.4% for $d > 1$
- Tuning the proposal density:
  - ▷ WinBUGS (one-dimensional MH algorithm): in first 4000 iterations to produce an acceptance rate between 20% and 40%
  - ▷ SAS® procedure MCMC: in several loops

The Independent Metropolis-Hastings algorithm

- Proposal density: does not depend on the position in the chain, e.g.

$$q(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = \mathsf{N}_d(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\mu}, \Sigma)$$

- One of the possible samplers of the SAS® procedure MCMC
- Similar to AR algorithm but accepts $\widetilde{\boldsymbol{\theta}}$ when $p(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{y})/q(\widetilde{\boldsymbol{\theta}}) > p(\boldsymbol{\theta}^k \mid \boldsymbol{y})/q(\boldsymbol{\theta}^k)$
- High acceptance rate is desirable when proposal density $q(\theta)$ is close to the posterior density
- (Robert and Casella) If $p(\theta \mid \boldsymbol{y}) \leq A\, q(\theta)$ for all θ, then the Markov chain generated by the Independent MH algorithm has excellent convergence properties and is more efficient than the AR algorithm

The Block Metropolis-Hastings algorithm

- MH algorithm within Gibbs sampling: Metropolis-within-Gibbs
- SAS® procedure MCMC: blocks specified by the user
- WinBUGS: regression coefficients in one block (blocking option switched on) and variance parameters in other block

The Reversible Jump MCMC (RJMCMC) algorithm

- Special case of the MH algorithm
- Jumps within space and between spaces
- Important application: Bayesian variable selection

3.3 Choice of the sampler

Choice of the sampler depends on a variety of considerations

Example: Caries study – MCMC approaches for logistic regression

Subset of $n = 500$ children of the Signal-Tandmobiel® study at 1st examination:

- ▷ Research questions:
  - ○ Have girls a different risk for developing caries experience (*CE*) than boys (*gender*) in the first year of primary school?
  - ○ Is there an east-west gradient (*x-coordinate*) in CE?
- ▷ Bayesian model: logistic regression + $\mathrm{N}(0, 100^2)$ priors for regression coefficients
- ▷ No standard full conditionals
- ▷ Three algorithms:
  - ○ Self-written R program: evaluate full conditionals on a grid + ICDF-method
  - ○ WinBUGS program: multivariate MH algorithm (blocking mode on)
  - ○ SAS® procedure MCMC: Random-Walk MH algorithm

Program	Parameter	Mode	Mean	SD	Median	MCSE
MLE	Intercept	-0.5900		0.2800		
	gender	-0.0379		0.1810		
	x-coord	0.0052		0.0017		
R	Intercept		-0.5880	0.2840	-0.5860	0.0104
	gender		-0.0516	0.1850	-0.0578	0.0071
	x-coord		0.0052	0.0017	0.0052	6.621E-5
WinBUGS	Intercept		-0.5800	0.2810	-0.5730	0.0094
	gender		-0.0379	0.1770	-0.0324	0.0060
	x-coord		0.0052	0.0018	0.0053	5.901E-5
SAS®	Intercept		-0.6530	0.2600	-0.6450	0.0317
	gender		-0.0319	0.1950	-0.0443	0.0208
	x-coord		0.0055	0.0016	0.0055	0.00016

Conclusions:

- Posterior means/medians of the three samplers are close (to the MLE)
- Precision with which the posterior mean was determined (high precision = low MCSE) differs considerably
- The clinical conclusion was the same

$\Rightarrow$ Samplers may have quite a different efficiency

3.4 Software

- WinBUGS + MCMC techniques implied a **revolution** in use of Bayesian methods
- WinBUGS has been long the standard software for many Bayesians
- But, WinBUGS is not further updated and is replaced by OpenBUGS
- To avoid the repetitive 'click and point' actions need to start up a Bayesian run with Win/OpenBUGS, a batch version of these programs is available
- R programs such as: R2WinBUGS & R2OpenBUGS, BRugs make use of the script option in Win/OpenBUGS to allow for batch processing starting from R
- But more packages in R have recently been developed: rjags, NIMBLE, STAN, ...

WinBUGS = Windows version of Bayesian inference Using Gibbs Sampling (BUGS)

- Windows-only program for Bayesian estimation with a graphical user interface
- Start: 1989 in MRC Biostatistics at Cambridge with BUGS
- Spiegelhalter et al. 2003
- Final version = 1.4.3
- **Freely available!**
- Difficulties installing WinBUGS under Windows 10
- Available at `http://www.mrc-bsu.cam.ac.uk/software/bugs/`

OpenBUGS = open source version of WinBUGS

- Started in 2004 in Helsinki
- Lunn et al. 2009
- Based on BUGS language
- Larger class of sampling algorithms
- Improved blocking algorithms
- New functions and distributions added to OpenBUGS
- Allows censoring C(lower, upper) and truncation T(lower, upper)
- More details on samplers by default
- Also **freely available**, but only OpenBUGS is kept up-to-date
- OpenBUGS is available as a Windows program at `http://www.openbugs.net`

- WinBUGS and OpenBUGS are the first tools to use for novel users of the Bayesian approach
- But, the repetitive 'clicking and pointing' needed to finalize a Bayesian statistical analysis becomes tiring after some time
- Batch processing of the OpenBUGS analysis then becomes the tool in practice
- There are several programs that make a link between BUGS software and R:
  - R2WinBUGS: makes a link between R and WinBUGS, but also for OpenBUGS
  - R2OpenBUGS: makes a link between R and OpenBUGS

JAGS = "Just Another Gibbs Sampler"

- Plummer 2011
- rjags: makes a link between R and JAGS
- JAGS is written in C++ and is portable to all major operating systems
- A JAGS model is defined in a text file using a dialect of the BUGS language
- It is a free, open-source program
- JAGS manual at `http://sourceforge.net/projects/mcmc-jags/`.

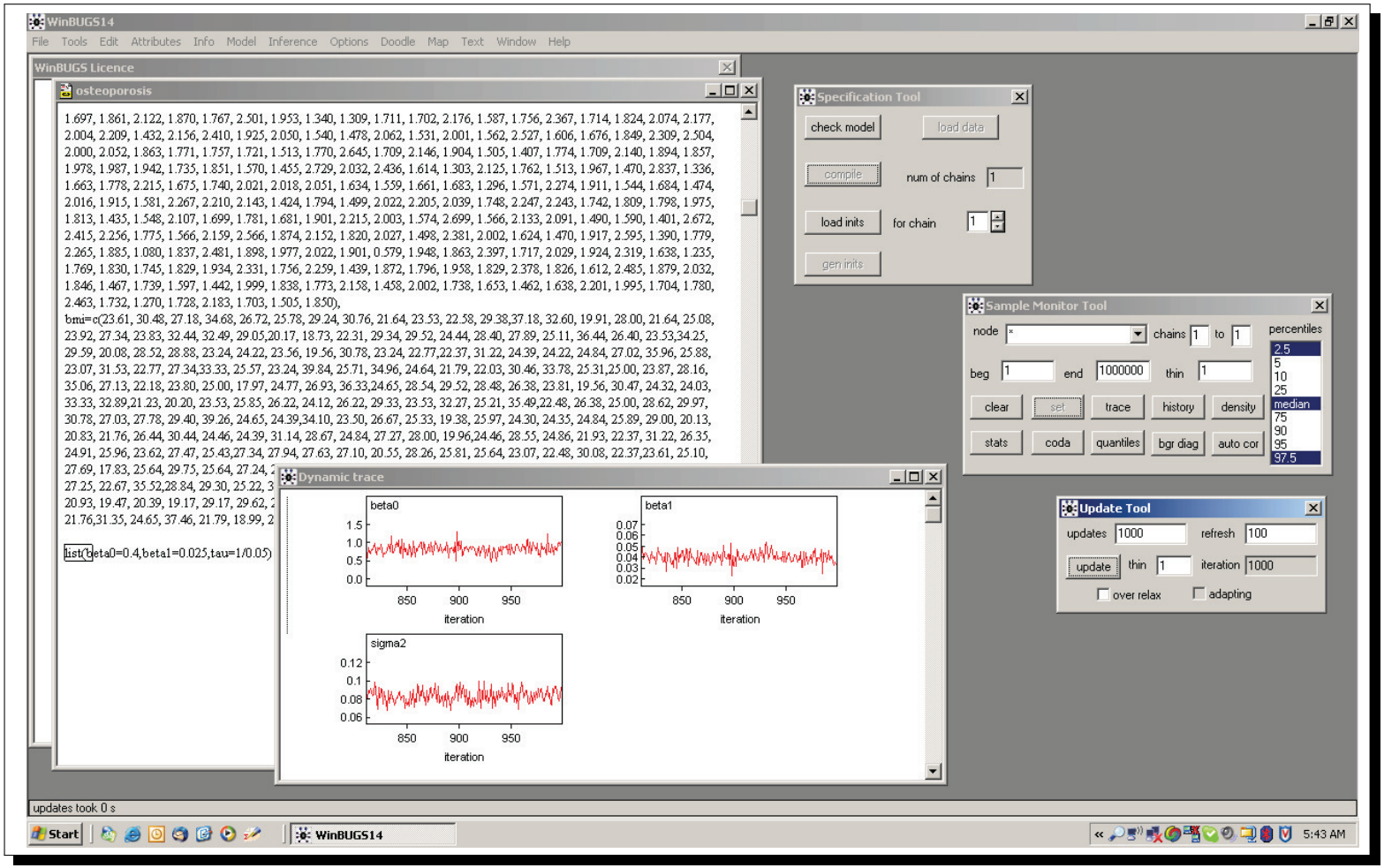
NIMBLE = "Numerical Inference for statistical Models using Bayesian and Likelihood Estimation'

- A framework for statistical models and algorithms.
- Uses almost same model syntax as WinBUGS, OpenBUGS, and JAGS, with C++ in the background for faster computations.
- Extension of BUGS language: Additional syntax, call to existing R functions, and implementation of your own functions/distributions.
- Flexibility in MCMC samplers config: change defaults, write your own algorithms.
- Examples, documentation and download: `https://r-nimble.org/`

Osteoporosis WinBUGS simple linear regression program

```
model
{
    for (i in 1:N)
    {
    tbbmc[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta0+beta1*bmi[i]
    }

    sigma2 <- 1/tau
    sigma <- sqrt(sigma2)

    beta0 ~ dnorm(0,1.0E-6)
    beta1 ~ dnorm(0,1.0E-6)
    tau ~ dgamma(1.0E-3,1.0E-3)
}

# data
list(tbbmc=c(1.798, 2.588, 2.325, 2.236, 1.925, 2.304, .....),
bmi=c(23.61, 30.48, 27.18, 34.68, 26.72, 25.78, 29.24, .....), N=234)

#initial values
list(beta0=0.4,beta1=0.025,tau=1/0.05)
```

Osteoporosis simple linear regression: WinBUGS program



Take home messages

- The two MCMC approaches allow fitting basically any proposed model
- There is no free lunch: computation time can be MUCH longer than with likelihood approaches
- The choice between Gibbs sampling and the Metropolis-Hastings approach depends on computational and practical considerations
- Checking of convergence is necessary when using MCMC!