MARKOV CHAIN MONTE CARLO AND SEQUENTIAL MONTE CARLO METHODS IN STOCHASTIC VOLATILITY MODELS

Hedibert Freitas Lopes

Associate Professor of Econometrics and Statistics
The University of Chicago Booth School of Business
5807 South Woodlawn Avenue, Chicago, IL 60637
http://faculty.chicagobooth.edu/hedibert.lopes/research

hlopes@chicagobooth.edu

III Summer School
Technical University of Catalonia
Barcelona, Spain

Course Schedule

Day	Weekday	Date	Time	Topic
1	Monday	June 22nd	9:00-11:00	Bayesian inference
	Monday	June 22nd	11:30-13:30	Bayesian model criticism
2	Tuesday	June 23rd	9:00-11:00	Monte Carlo methods
	Tuesday	June 23rd	11:30-13:30	Markov chain Monte Carlo methods
3	Friday	June 26th	9:00-11:00	Dynamic linear models
	Friday	June 26th	11:30-13:30	Nonnormal, nonlinear dynamic models
4	Monday	June 29th	9:00-11:00	Stochastic volatility models
	Monday	June 29th	11:30-13:30	Sequential Monte Carlo (SMC)
5	Wednesday	July 1st	9:00-11:00	SMC with parameter learning
	Wednesday	July 1st	11:30-13:30	SMC in stochastic volatility models

Lecture 1 - Bayesian inference

- Basic concepts such as prior, likelihood, posterior and predictive distributions are introduced.
- The intuitive sequential nature of Bayesian learning is illustrated via conjugate families of distributions.
- The lecture ends with Bayesian inference for normal linear models.

Lecture 2 - Bayesian model criticism

- The lecture starts introducing prior and posterior model probabilities and Bayes factor, key ingredients in assessing model uncertainty.
- Model selection as a decision problem will lead to alternative criteria, such as the posterior predictive criterion (PPC).
- The deviance information criterion (DIC) and cross-validatory measures are also presented.

Lecture 3 - Monte Carlo (MC) methods

- Monte Carlo (MC) integration schemes are introduced to approximate both posterior expectations as well as predictive ordinates.
- Similarly, acceptance-rejection and sampling importance resampling (SIR) algorithms, as well as other iterative resampling schemes, are introduced as tools to approximately sample from posterior distributions.

Lecture 4 - Markov chain Monte Carlo (MCMC) methods

- We start by reviewing basic Markov chain concepts and results that will facilitate the introduction of more general MCMC schemes.
- A few concepts are irreducibility, reversibility, ergodicity, limiting distributions and effective sample size.
- The two most famous MCMC schemes are then introduced: the Gibbs sampler and the Metropolis-Hastings algorithm.

Lecture 5 - Dynamic linear models (DLM)

- The linear model of the first lecture is extended to accommodate time-varying regression coefficients, which is one of many instances in the class of dynamic linear models.
- Sequential learning is provided in closed form by the Kalman filter and smoother.
- Inference for fixed parameters, such as observational and evolutional variances, is performed by integrating out states.

Lecture 6 - Nonnormal, nonlinear dynamic models:

- Forward filtering, backward sampling (FFBS) algorithm and other MCMC schemes.

Lecture 7 - Stochastic volatility models as dynamic models

- FFBS and other MCMC schemes are adapted to stochastic volatility models.
- In particular, we will compare single move and block move MCMC schemes, where block move schemes are based on mixture of normal densities approximation to the distribution of a log chi-square random variable with one degree of freedom.

Lecture 8 - Sequential Monte Carlo (SMC) methods

- The lecture starts with standard particle filters, such as the sequential importance sampling with resampling (SISR) filter and the auxiliary particle filter (APF), to sequentially learn about states in nonnormal and nonlinear dynamic models.
- SMC methods assist inference for fixed parameters, such as observational and evolutional variances. Stochastic volatility models are used to illustrate the filters.

Lecture 9 - SMC with parameter learning

- The APF is coupled with mixture approximation to the posterior distribution of fixed parameters to produced online estimates of both states and parameters in dynamic systems.
- The lecture concentrates on the particle learning (PL) filter and several simulated exercises are performed to compare PL to SISR, APF and MCMC alternatives.

Lecture 10 - SMC in stochastic volatility models

- PL and other filters are compared based on stochastic volatility models.
- The lecture also list current research agenda linking, Markov chain Monte Carlo methods, sequential Monte Carlo methods and general stochastic volatility models.

LECTURE 1

BAYESIAN INFERENCE

Example i. Sequential learning

- John claims some discomfort and goes to the doctor.
- The doctor believes John may have the disease A.
- $\theta = 1$: John has disease A; $\theta = 0$: he does not.
- The doctor claims, based on his expertise (H), that

$$P(\theta = 1|H) = 0.7$$

- Examination X is related to θ as follows

$$\begin{cases} P(X = 1|\theta = 0) = 0.40, & \text{positive test given no disease} \\ P(X = 1|\theta = 1) = 0.95, & \text{positive test given disease} \end{cases}$$

Exam's result: $X = 1$

$$\begin{aligned} P(\theta = 1|X = 1) &\propto l(\theta = 1\,; X = 1)P(\theta = 1) \\ &\propto (0.95)(0.7) = 0.665 \\ P(\theta = 0|X = 1) &\propto l(\theta = 0\,; X = 1)P(\theta = 0) \\ &\propto (0.40)(0.30) = 0.120 \end{aligned}$$

Consequently

$$\begin{aligned} P(\theta = 0|X = 1) &= 0.120/0.785 = 0.1528662 \text{ and} \\ P(\theta = 1|X = 1) &= 0.665/0.785 = 0.8471338 \end{aligned}$$

The information $X = 1$ increases, for the doctor, the probability that John has the disease A from 70% to 84.71%.

2nd exam: Y

John undertakes the test Y, which relates to θ as follows

$$\begin{cases} P(Y = 1|\theta = 1) = 0.99 & P(X = 1|\theta = 1) = 0.95 \\ P(Y = 1|\theta = 0) = 0.04 & P(X = 1|\theta = 0) = 0.40 \end{cases}$$

Predictive:

$P(Y = 0|X = 1) = (0.96)(0.1528662) + (0.01)(0.8471338) = 0.1552229.$

Suppose the observed result was $Y = 0$. This is a reasonably unexpected result as the doctor only gave it roughly 15% chance.

Questions

He should at least consider rethinking the model based on this result. In particular, he might want to ask himself

1. Did 0.7 adequately reflect his $P(\theta = 1|H)$?
2. Is test X really so unreliable?
3. Is the sample distribution of X correct?
4. Is the test Y so powerful?
5. Have the tests been carried out properly?

What is $P(\theta|X = 1, Y = 0)$?

Let $H_2 = \{X = 1, Y = 0\}$ and using the Bayes theorem

$$\begin{aligned} P(\theta = 1|H_2) &\propto l(\theta = 1\,; Y = 0)P(\theta = 1|X = 1) \\ &\propto (0.01)(0.8471338) = 0.008471338 \text{ and} \\ P(\theta = 0|H_2) &\propto l(\theta = 0\,; Y = 0)P(\theta = 0|X = 1) \\ &\propto (0.96)(0.1528662) = 0.1467516 \end{aligned}$$

$$P(\theta = 1|H_i) = \begin{cases} 0.7000 & , H_0\text{: before X and Y} \\ 0.8446 & , H_1\text{: after X=1 and before Y} \\ 0.0546 & , H_2\text{: after X=1 and Y=0} \end{cases}$$

Example ii. Normal model and prior $\Rightarrow$ normal posterior

Suppose X, conditional on θ, is modeled by

$$X|\theta \sim N(\theta, \sigma^2)$$

and the prior distribution of θ is

$$\theta \sim N(\theta_0, \tau_0^2)$$

with σ^2, θ_0 and τ_0^2 known.

Posterior of θ: $(\theta|X = x) \sim N(\theta_1, \tau_1^2)$

$$\begin{aligned} \theta_1 &= w\theta_0 + (1 - w)x \\ \tau_1^{-2} &= \tau_0^{-2} + \sigma^{-2} \end{aligned}$$

where $w = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2})$ measures the relative information contained in the prior distribution with respect to the total information (prior plus likelihood).

Example from Box & Tiao (1973)

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$.

Model: $(X|\theta) \sim N(\theta, (40)^2)$.

Observation: $X = 850$

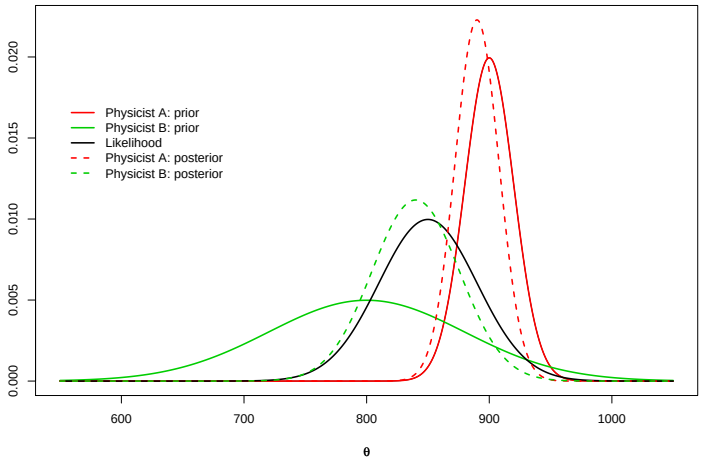
$$\begin{aligned}(\theta|X = 850, H_A) &\sim N(890, (17.9)^2)\\ (\theta|X = 850, H_B) &\sim N(840, (35.7)^2)\end{aligned}$$

Information (precision)

Physicist A: from 0.002500 to 0.003120 (an increase of 25%)

Physicist B: from 0.000156 to 0.000781 (an increase of 400%)

Priors and posteriors



Example iii. Simple linear regression

A simple normal linear regression relates the dependent variable y_i and the the explanatory variable x_i, for $i = 1, \ldots, n$, by

$$\begin{aligned}y_i|\theta, H &\sim N(\theta x_i; \sigma^2)\\ \theta|H &\sim N(\theta_0, \tau_0^2)\end{aligned}$$

Therefore $H = \{\sigma^2, \theta_0, \tau_0^2, x_1, \ldots, x_n\}$.

Example iv. Simple stochastic volatility model

The simplest stochastic volatility model with first-order auregressive log-volatilities, namely SV-AR(1), relates log-return of financial time series y_t to log-volatility θ_t, for $t = 1, \ldots, T$, via

$$\begin{aligned} y_t|\theta, H &\sim N(0; e^{\theta_t}) \\ \theta_t|H &\sim N(\alpha + \beta\theta_{t-1}, \sigma^2) \end{aligned}$$

Therefore $H = \{\alpha, \beta, \sigma^2, \theta_0\}$.

Bayesian ingredients

- Posterior (Bayes' Theorem)

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}, H) &= \frac{p(\boldsymbol{\theta}, \boldsymbol{x} \mid H)}{p(\boldsymbol{x} \mid H)} \\ &= \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}, H) p(\boldsymbol{\theta} \mid H)}{p(\boldsymbol{x} \mid H)} \end{aligned}$$

- Predictive (or marginal) distribution

$$p(\boldsymbol{x}|H) = \int_\Theta p(\boldsymbol{x}, \boldsymbol{\theta}|H)\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}[p(\boldsymbol{x} \mid \boldsymbol{\theta}, H)]$$

$p(\boldsymbol{x}|H)$ is also known as *normalizing constant* and plays an important role in Bayesian model criticism.

Bayesian ingredients (cont.)

- Posterior predictive

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}, H) &= \int_\Theta p(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{x}, H) d\boldsymbol{\theta} \\ &= \int_\Theta p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}, H) p(\boldsymbol{\theta} \mid \boldsymbol{x}, H) d\boldsymbol{\theta} \\ &= \int_\Theta p(\boldsymbol{y} \mid \boldsymbol{\theta}, H) p(\boldsymbol{\theta} \mid \boldsymbol{x}, H) d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}|\boldsymbol{x}}\left[p(\boldsymbol{y} \mid \boldsymbol{\theta}, H)\right] \end{aligned}$$

since, in general, but not always,

$$X, Y \quad \text{are independent given} \quad \boldsymbol{\theta}.$$

It might be more useful to concentrate on prediction rather than on estimation because the former is verifiable, i.e. $\boldsymbol{y}$ is observable while $\boldsymbol{\theta}$ is not.

Sequential Bayes theorem: a rule for updating probabilities

Experimental result: $\boldsymbol{x}_1 \sim p_1(\boldsymbol{x}_1 \mid \boldsymbol{\theta})$

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \propto l_1(\boldsymbol{\theta}; x_1) p(\boldsymbol{\theta})$$

Experimental result: $\boldsymbol{x}_2 \sim p_2(\boldsymbol{x}_2 \mid \boldsymbol{\theta})$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}_2, \boldsymbol{x}_1) &\propto l_2(\boldsymbol{\theta}; \boldsymbol{x}_2) p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \\ &\propto l_2(\boldsymbol{\theta}; \boldsymbol{x}_2) l_1(\boldsymbol{\theta}; \boldsymbol{x}_1) p(\boldsymbol{\theta}) \end{aligned}$$

Experimental results: $\boldsymbol{x}_i \sim p_i(\boldsymbol{x}_i \mid \boldsymbol{\theta})$, for $i = 3, \cdots, n$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}_n, \cdots, \boldsymbol{x}_1) &\propto l_n(\boldsymbol{\theta}; \boldsymbol{x}_n) p(\boldsymbol{\theta} \mid \boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_1) \\ &\propto \left[\prod_{i=1}^{n} l_i(\boldsymbol{\theta}; \boldsymbol{x}_i) \right] p(\boldsymbol{\theta}) \end{aligned}$$

Example iii. Revisited

Combining likelihood and prior

$$\begin{aligned} y_i|\theta, H &\sim N(\theta x_i; \sigma^2) \\ \theta|H &\sim N(\theta_0, \tau_0^2) \end{aligned}$$

leads to posterior

$$\theta|\boldsymbol{y}, H \sim N(\theta_1, \tau_1^2)$$

where

$$\tau_1^{-2} = \tau_0^{-2} + \sigma^{-2} \sum_{i=1}^{n} x_i^2 \text{ and } \theta_1 = \tau_1^2 \left(\tau_0^{-2} \theta_0 + \sigma^{-2} \sum_{i=1}^{n} y_i x_i \right)$$

When $\tau_0^{-2} \to 0$, i.e. with *little* prior knowledge about θ, the above moments converge to ordinary least squares counterparts:

$$\tau_1^{-2} = \sigma^{-2} \sum_{i=1}^{n} x_i^2 \text{ and } \theta_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Example v. Multiple normal linear regression

The standard Bayesian approach to multiple linear regression is

$$(y|X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I_n)$$

where $y = (y_1, \ldots, y_n)$, $X = (x_1, \ldots, x_n)'$ is the $(n \times q)$, design matrix and $q = p + 1$.

The prior distribution of (β, σ^2) is $NIG(b_0, B_0, n_0, S_0)$, i.e.

$$\begin{aligned} \beta|\sigma^2 &\sim N(b_0, \sigma^2 B_0) \\ \sigma^2 &\sim IG(n_0/2, n_0 S_0/2) \end{aligned}$$

for known hyperparameters b_0, B_0, n_0 and S_0.

Example v. Conditionals and marginals

It is easy to show that (β, σ^2) is $NIG(b_1, B_1, n_1, S_1)$, i.e.

$$\begin{aligned} (\beta|\sigma^2, y, X) &\sim N(b_1, \sigma^2 B_1) \\ (\sigma^2|y, X) &\sim IG(n_1/2, n_1 S_1/2) \end{aligned}$$

where

$$\begin{aligned} B_1^{-1} &= B_0^{-1} + X'X \\ B_1^{-1}b_1 &= B_0^{-1}b_0 + X'y \\ n_1 &= n_0 + n \\ n_1 S_1 &= n_0 S_0 + (y - Xb_1)'y + (b_0 - b_1)'B_0^{-1}b_0. \end{aligned}$$

It is also easy to derive the full conditional distributions, i.e.

$$\begin{aligned} (\beta|y, X) &\sim t_{n_1}(b_1, S_1 B_1) \\ (\sigma^2|\beta, y, X) &\sim IG(n_1/2, n_1 S_{11}/2) \end{aligned}$$

where

$$n_1 S_{11} = n_0 S_0 + (y - X\beta)'(y - X\beta).$$

Example v. Ordinary least squares

It is well known that

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\sigma}^2 &= \frac{S_e}{n-q} = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-q} \end{aligned}$$

are the OLS estimates of β and σ^2, respectively.

The conditional and unconditional sampling distributions of $\hat{\beta}$ are

$$\begin{aligned} (\hat{\beta}|\sigma^2, y, X) &\sim N(\beta, \sigma^2(X'X)^{-1}) \\ (\hat{\beta}|y, X) &\sim t_{n-q}(\beta, S_e) \end{aligned}$$

respectively, with

$$(\hat{\sigma}^2|\sigma^2) \sim IG\left((n-q)/2, ((n-q)\sigma^2/2\right).$$

Example v. Sufficient statistics recursions

Recall the multiple linear regression $(y_t|x_t, \beta, \sigma^2) \sim N(x_t'\beta, \sigma^2)$ for $t = 1, \ldots, n$, $\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0)$ and $\sigma^2 \sim IG(n_0/2, n_0 S_0/2)$.

Then, for $y^t = (y_1, \ldots, y_t)$ and $X^t = (x_1, \ldots, x_t)'$, it follows that

$$\begin{aligned} (\beta|\sigma^2, y^t, X^t) &\sim N(b_t, \sigma^2 B_t) \\ (\sigma^2|y^t, X^t) &\sim IG(n_t/2, n_t S_t/2) \end{aligned}$$

where $n_t = n_{t-1} + 1$, $B_t^{-1} = B_{t-1}^{-1} + x_t x_t'$, $B_t^{-1}b_t = B_{t-1}^{-1}b_{t-1} + y_t x_t$ and $n_t S_t = n_{t-1}S_{t-1} + (y_t - b_t'x_t)y_t + (b_{t-1} - b_t)'B_{t-1}^{-1}b_{t-1}$.

The only ingredients needed are: $x_t x_t'$, $y_t x_t$ and y_t^2.

These recursions will play an important role later on when deriving **sequential Monte Carlo** methods for conditionally Gaussian dynamic linear models, like many stochastic volatility models.

Example v. Predictive

The predictive density can be seen as the *marginal likelihood*, i.e.

$$p(y|X) = \int p(y|X, \beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)d\beta d\sigma^2$$

or, by Bayes' theorem, as the *normalizing constant*, i.e.

$$p(y|X) = \frac{p(y|X, \beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)}{p(\beta|\sigma^2, y, X)p(\sigma^2|y, X)}$$

which is valid for all (β, σ^2).

Closed form solution is available for the multiple normal linear regression:

$$(y|X) \sim t_{n_0}(Xb_0, S_0(I_n + XB_0X')).$$

Unfortunately, closed form solutions are rare.

Example iv. Revisited

The posterior distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, H) \prod_{t=1}^{T} p(y_t|\theta_t) \\ &\times \prod_{t=1}^{T} \exp\left\{-\frac{1}{2\sigma^2}(\theta_t - \alpha - \beta\theta_{t-1})^2\right\} \\ &\propto \prod_{t=1}^{T} e^{-\theta_t/2} \exp\left\{-\frac{1}{2} y_t e^{-\theta_t}\right\} \end{aligned}$$

- How to compute $E(\theta_{43}|\boldsymbol{y})$ or $V(\theta_{11}|\boldsymbol{y})$?
- How to compute 95% credible regions for $(\theta_{35}, \theta_{36}|\boldsymbol{y})$?
- How to sample from $p(\boldsymbol{\theta}|\boldsymbol{y})$ or $p(\boldsymbol{\theta}|y_1, \ldots, y_{10})$?
- How to compute $p(\boldsymbol{y})$ or $p(y_t|y_1, \ldots, y_{t-1})$?

LECTURE 2

BAYESIAN MODEL CRITICISM

Outline

- Prior and posterior model probabilities
- Posterior odds
- Bayes factor
- Computing normalizing constants
- Savage-Dickey density ratio
- Bayesian Model Averaging
- Posterior predictive criterion
- Deviance information criterion

Prior and posterior model probabilities

Suppose that the competing models can be enumerated and are represented by the set $M = \{M_1, M_2, \ldots\}$.

Bayesian model comparison is commonly performed by computing posterior model probabilities,

$$Pr(M_j|y) \propto f(y|M_j)Pr(M_j)$$

where $Pr(M_j)$ and

$$f(y|M_j) = \int f(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j$$

are, respectively, the prior model probability and the predictive density of model M_j, for $j = 1, 2, \ldots$

Posterior odds

Posterior odds of model M_j relative to M_k

$$\underbrace{\frac{Pr(M_j|y)}{Pr(M_k|y)}}_{\text{posterior odds}} = \underbrace{\frac{Pr(M_j)}{Pr(M_k)}}_{\text{prior odds}} \times \underbrace{\frac{f(y|M_j)}{f(y|M_k)}}_{\text{Bayes factor}} .$$

The Bayes factor can be viewed as the weighted likelihood ratio of M_j to M_k.

Bayes factor

Jeffreys (1961) recommends the use of the following rule of thumb to decide between models j and k:

$$\begin{aligned} B_{jk} > 100: &\quad \text{decisive evidence against } k \\ 10 < B_{jk} \leq 100: &\quad \text{strong evidence against } k \\ 3 < B_{jk} \leq 10: &\quad \text{substantial evidence against } k \end{aligned}$$

Posterior model probability for model j is

$$Pr(M_j|y) = \left\{ \sum_{k=1}^{\infty} B_{kj} \frac{Pr(M_k)}{Pr(M_j)} \right\}^{-1}$$

for $j = 1, 2, \ldots$.

Computing normalizing constants

A basic ingredient for model assessment is given by the predictive density

$$f(y|M) = \int f(y|\theta, M) p(\theta|M) d\theta \ ,$$

which is the normalizing constant of the posterior distribution.

The predictive density can now be viewed as the likelihood of model M.

It is sometimes referred to as predictive likelihood, because it is obtained after marginalization of model parameters.

For any given model, the predictive density can be written as

$$f(y) = E[f(y|\theta)]$$

where expectation is taken with respect to the prior distribution $p(\theta)$.

Approximate methods (discussed more later)

Several approximations for $f(y)$ based on Monte Carlo and Markov chain Monte Carlo methods are routinely available. Amongst them are:

- Laplace-Metropolis estimator
- Simple Monte Carlo
- Monte Carlo via importance sampling
- Harmonic mean estimator
- Chib's estimator
- Reversible jump MCMC

Key references are DiCiccio, Kass, Raftery and Wasserman (1997) Han and Carlin (2001) and Lopes and West (2004).

Savage-Dickey Density Ratio

$\Rightarrow$ Suppose that M_2 is described by

$$p(y|\omega, \psi, M_2)$$

and M_1 is a restricted verison of M_2, ie.

$$p(y|\psi, M_1) \equiv p(y|\omega = \omega_0, \psi, M_2)$$

$\Rightarrow$ Suppose also that

$$\pi(\psi|\omega = \omega_0, M_2) = \pi(\psi|M_1)$$

$\Rightarrow$ Therefore, it can be proved that the Bayes factor is

$$B_{12} = \frac{\pi(\omega = \omega_0|y, M_2)}{\pi(\omega = \omega_0|M_2)}$$

where $\{\psi^{(1)}, \ldots, \psi^{(N)}\} \sim \pi(\psi|y, M_2)$. See Verdinelli and Wasserman (1995) for further details.

Example i. Normality x Student-t

Suppose we have two competing models

$$\begin{aligned} \mathcal{M}_1 &: x_i \sim N(\mu, \sigma^2) \\ \mathcal{M}_2 &: x_i \sim t_\lambda(\mu, \sigma^2) \end{aligned}$$

Letting $\omega = 1/\lambda$, $\mathcal{M}_1$ is a particular case of $\mathcal{M}_2$ when $\omega = \omega_0 = 0.0$, with $\psi = (\mu, \sigma^2)$.

Let $\omega \sim U(0, 1)$, with $\omega = 1$ corresponding to a Cauchy distribution. Assuming that

$$\pi(\mu, \sigma^2|\mathcal{M}_1) = \pi(\mu, \sigma^2, \omega|\mathcal{M}_2) \propto \sigma^{-2}$$

the Savage-Dickey formula holds and the Bayes factor is

$$B_{12} = \pi(\omega_0|x, \mathcal{M}_2)$$

the marginal posterior of ω evaluated at 0.

Results

Because $\pi(\mu, \sigma^2, \omega|x, \mathcal{M}_2)$ has no closed form solution, they use a Metropolis algorithm.

When $n = 100$ from $N(0, 1)$, then

$$B_{12} = 3.79$$

with standard error of 0.145.

When $n = 100$ from $Cauchy(0, 1)$, then

$$B_{12} = 0.000405$$

with standard error of 0.000240.

Bayesian model averaging

Let $\mathcal{M}$ denote the set that indexes all entertained models.

Assume that Δ is an outcome of interest, such as the future value y_{t+k}, or an elasticity well defined across models, etc.

The posterior distribution for Δ is

$$p(\Delta|y) = \sum_{m \in \mathcal{M}} p(\Delta|m, y) Pr(m|y)$$

for data y and posterior model probability

$$Pr(m|y) = \frac{p(y|m)Pr(m)}{p(y)}$$

where $Pr(m)$ is the prior probability model.

See Hoeting, Madigan, Raftery and Volinsky (1999) for more details.

Posterior predictive criterion

Gelfand and Ghosh (1998) introduced a posterior predictive criterion that, under squared error loss, favors the model M_j which minimizes

$$D_j^G = P_j^G + G_j^G$$

where

$$\begin{aligned} P_j^G &= \sum_{t=1}^{n} V(\tilde{y}_t|y, M_j) \\ G_j^G &= \sum_{t=1}^{n} [y_t - E(\tilde{y}_t|y, M_j)]^2 \end{aligned}$$

and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ are predictions/replicates of y.

The first term, P_j, is a penalty term for model complexity.

The second term, G_j, accounts for goodness of fit.

More general losses

Gelfand and Ghosh (1998) also derived the criteria for more general error loss functions.

Expectations $E(\tilde{y}_t|y, M_j)$ and variances $V(\tilde{y}_t|y, M_j)$ are computed under posterior predictive densities, ie.

$$E[h(\tilde{y}_t)|y, M_j] = \int \int h(\tilde{y}_t) f(\tilde{y}_t|y, \theta_j, M_j) \pi(\theta_j|M_j) d\theta_j d\tilde{y}_t$$

for $h(\tilde{y}_t) = \tilde{y}_t$ and $h(\tilde{y}_t) = \tilde{y}_t^2$.

The above integral can be approximated via Monte Carlo.

Deviance information criterion

Inspired by Dempster's (1997) suggestion to compute the posterior distribution of the log-likelihood, $D(\theta_j) = -2\log f(y|\theta_j, M_j)$, Spiegelhalter et al. (2002) introduced the *deviance information criterion* (DIC)

$$D_j^S = P_j^S + G_j^S$$

where

$$\begin{aligned} P_j^S &= E[D(\theta_j)|y, M_j] - D[E(\theta_j|y, M_j)] \\ G_j^S &= E[D(\theta_j)|y, M_j]. \end{aligned}$$

The DIC is decomposed into two important components:

- One responsible for goodness of fit: (G_j^S)
- One responsible for model complexity: (P_j^S)

P_j^S is also currently referred to as the effective number of parameters of model M_j.

DIC and WinBUGS

The DIC has become very popular in the applied Bayesian community due to its computational simplicity and, consequently, its availability in `WinBUGS`.

Further applications appear, amongst many others, in

- Berg, Meyer and Yu (2002): stochastic volatility models.
- Celeux et al. (2005): mixture models, random effects models and several missing data models.
- Nobre, Schmidt and Lopes (2005): space-time hierarchical models.
- van der Linde (2005): variable selection.
- Lopes and Salazar (2006): nonlinear time series models.
- Silva and Lopes (2008): mixture of copulas models.

LECTURE 3

MONTE CARLO METHODS

Basic Bayesian computation

Main ingredients:

$$\text{Posterior} \quad : \quad \pi(\theta) = \frac{f(x|\theta)p(\theta)}{f(x)}$$

$$\text{Predictive} \quad : \quad f(x) = \int f(x|\theta)p(\theta)d\theta$$

Bayesian Agenda:

- Posterior modes: $\max_\theta \pi(\theta)$;
- Posterior moments: $E_\pi[g(\theta)]$;
- Density estimation: $\hat{\pi}(g(\theta))$;
- Bayes factors: $f(x|M_0)/f(x|M_1)$;
- Decision: $\max_d \int U(d,\theta)\pi(\theta)d\theta$.

Analytic approximations

- Asymptotic approximation (Carlin&Louis, 2000)
- Laplace approximation (Tierney&Kadane, 1986)
- Gaussian quadrature (Naylor and Smith, 1982)

Stochastic approximations/simulations

- Simulated annealing (Metropolis et al, 1953)
- Metropolis-Hastings algorithm (Hastings, 1970)
- Monte Carlo integration (Geweke, 1989)
- Gibbs sampler (Gelfand and Smith, 1990)
- Rejection methods (Gilks and Wild, 1992)
- Importance Sampling (Smith and Gelfand, 1992)

Monte Carlo methods

In what follows we will introduce several Monte Carlo methods for integrating and/or sampling from nontrivial densities.

- Simple Monte Carlo integration
- Monte Carlo integration via importance sampling
- Sampling via the rejection method
- Sampling via importance resampling (SIR)

Monte Carlo integration

The objective here is to compute moments

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

If $\theta_1, \ldots, \theta_n$ is a random sample from $\pi(\cdot)$

$$\Rightarrow \bar{h}_{mc} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i) \rightarrow E_\pi[h(\theta)]$$

If, additionally, $E_\pi[h^2(\theta)] < \infty$, then

$$V_\pi[\bar{h}_{mc}] = \frac{1}{n}\int \{h(\theta) - E_\pi[h(\theta)]\}^2 \pi(\theta)d\theta$$

and

$$v_{mc} = \frac{1}{n^2}\sum_{i=1}^{n}(h(\theta_i) - \bar{h}_{mc})^2 \rightarrow V_\pi[\bar{h}_{mc}]$$

Example i.

The objective here is to estimate

$$p = \int_0^1 [cos(50\theta) + sin(20\theta)]^2 d\theta = 0.965$$

by noticing that the above integral can be rewritten as

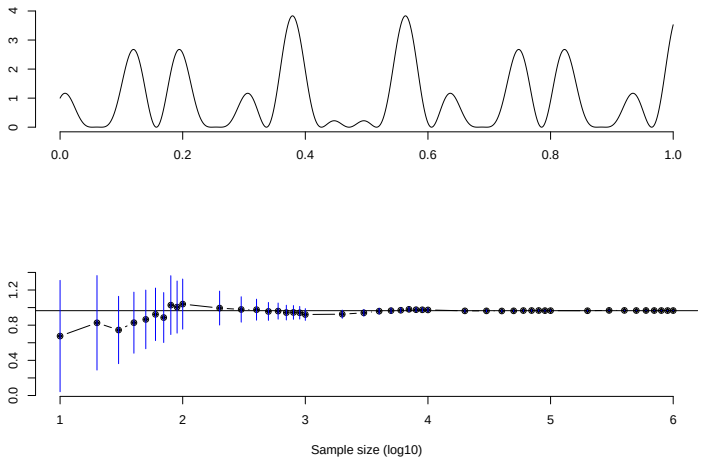
$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

where $h(\theta) = [cos(50\theta) + sin(20\theta)]^2$ and $\pi(\theta) = 1$ is the density of a $U(0,1)$.

Therefore

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i)$$

where $\theta_1, \ldots, \theta_n$ are i.i.d. from $U(0,1)$.



Monte Carlo via importance sampling

The objective is still the same, ie to compute

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

by noticing that

$$E_\pi[h(\theta)] = \int \frac{h(\theta)\pi(\theta)}{q(\theta)} q(\theta)d\theta$$

where $q(\cdot)$ is an *importance function*. Therefore, if $\theta_1, \ldots, \theta_n$ is a random sample from $q(\cdot)$ then

$$\Rightarrow \bar{h}_{is} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i)\pi(\theta_i)/q(\theta_i) \rightarrow E_\pi[h(\theta)]$$

Ideally, $q(\cdot)$ should be (i) as "*close*" as possible to $h(\cdot)\pi(\cdot)$ and (ii) easy to sample from.

Example ii.

The objective here is to estimate

$$p = \int_2^\infty \frac{1}{\pi(1+\theta^2)} d\theta$$

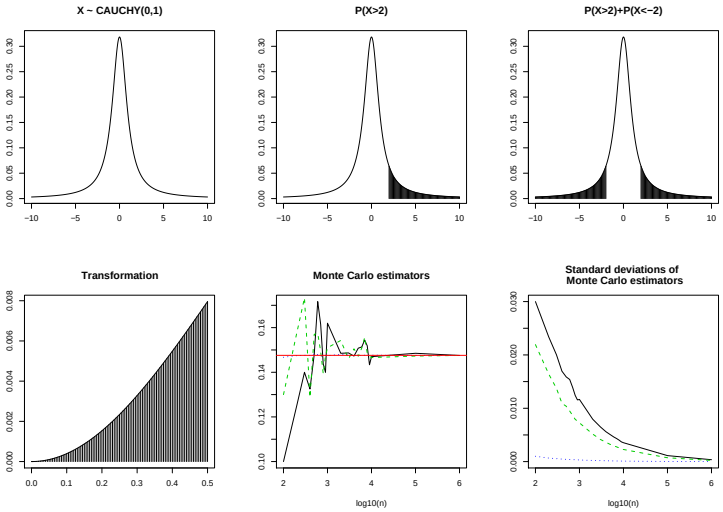
Three Monte Carlo estimators of p are

$$\begin{aligned}
\hat{p}_1 &= \frac{1}{n}\sum_{i=1}^{n} I\{\theta_i \in (2,\infty)\} \\
\hat{p}_2 &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2} I\{\theta_i \in (-\infty,-2) \cup (2,\infty)\} \\
\hat{p}_3 &= \frac{1}{n}\sum_{i=1}^{n} \frac{u_i^{-2}}{2\pi[1+u_i^{-2}]}
\end{aligned}$$

where $\theta_1, \ldots, \theta_n \sim$ Cauchy(0,1) and $u_1, \ldots, u_n \sim U(0, 1/2)$.

n	v_{mc1}	v_{mc2}	v_{mc3}
50	0.051846	0.033941	0.001407
100	0.030000	0.021651	0.000953
700	0.014054	0.008684	0.000359
1000	0.011738	0.007280	0.000308
5000	0.005050	0.003220	0.000138
10000	0.003543	0.002276	0.000097
100000	0.001124	0.000721	0.000031
1000000	0.000355	0.000228	0.000010

If 0.0035 is the desired level of precision in the estimation, then 1 million draws would be necessary for estimator $\hat{p}_1$ while only 700 for estimator $\hat{p}_3$, i.e. roughly 3 orders of magnitude smaller.



Rejection method

The objective is to draw from a target density

$$\pi(\theta) = c_\pi \tilde{\pi}(\theta)$$

when only draws from an auxiliary density

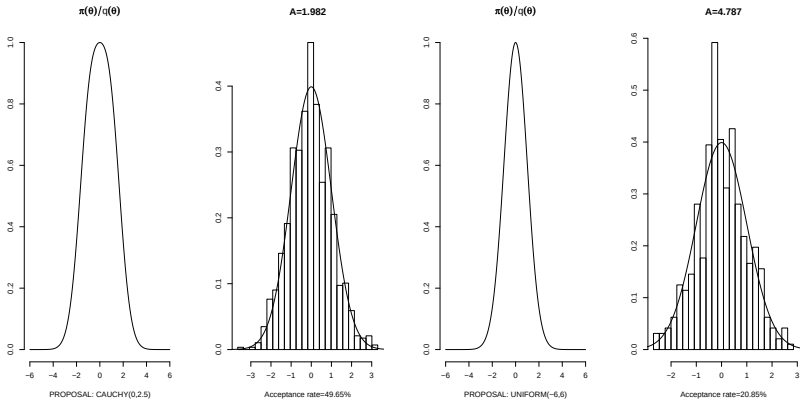
$$q(\theta) = c_q \tilde{q}(\theta)$$

is available. Here c_π and c_q are normalizing constants.

If there exist a constant $A < \infty$ such that

$$\tilde{\pi}(\theta) \leq A\tilde{q}(\theta) \quad \text{for all } \theta$$

then $q(\theta)$ becomes a *blanketing density* or an *envelope* and A the *envelope constant*.



Algorithm

Under the previous conditions the following algorithm can be used to draw from $\pi(\theta)$.

1. Draw θ^* from $q(\cdot)$;
2. Draw u from $U(0,1)$;

3. Accept θ^* if $u \leq \frac{\tilde{\pi}(\theta^*)}{A\tilde{q}(\theta^*)}$;

4. Repeat 1, 2 and 3 until n draws are accepted.

Proof

Applying Bayes' theorem:

$$\begin{aligned} p(\theta \mid Au\tilde{q}(\theta) \leq \tilde{\pi}(\theta)) &= \frac{Pr(Au\tilde{q}(\theta) < \tilde{\pi}(\theta) \mid \theta) q(\theta)}{\int Pr(Au\tilde{q}(\theta) \leq \tilde{\pi}(\theta) \mid \theta) q(\theta) d\theta} \\ &= \frac{Pr\left(u < \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} \mid \theta\right) q(\theta)}{\int Pr\left(u < \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} \mid \theta\right) q(\theta) d\theta} \\ &= \frac{\frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} c_q \tilde{q}(\theta)}{\int \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} c_q \tilde{q}(\theta) d\theta} = \frac{\tilde{\pi}(\theta)}{\int \tilde{\pi}(\theta) d\theta} \\ &= \pi(\theta) \ . \end{aligned}$$

One does not need to known c_π and c_q.

The smaller the A is the larger the acceptance rate.

The theoretical acceptance rate is

$$\begin{aligned} Pr\left(u \leq \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)}\right) &= \int Pr\left(u \leq \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} | \theta\right) q(\theta) d\theta \\ &= \int \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} c_q \tilde{q}(\theta) d\theta \\ &= \frac{1}{A} \frac{\int \tilde{\pi}(\theta) d\theta}{\int \tilde{q}(\theta) d\theta} = \frac{c_q}{A c_\pi}. \end{aligned}$$

Example iii.

Enveloping the N(0,1) density

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\{-0.5\theta^2\}$$

by a multiple of a Cauchy density

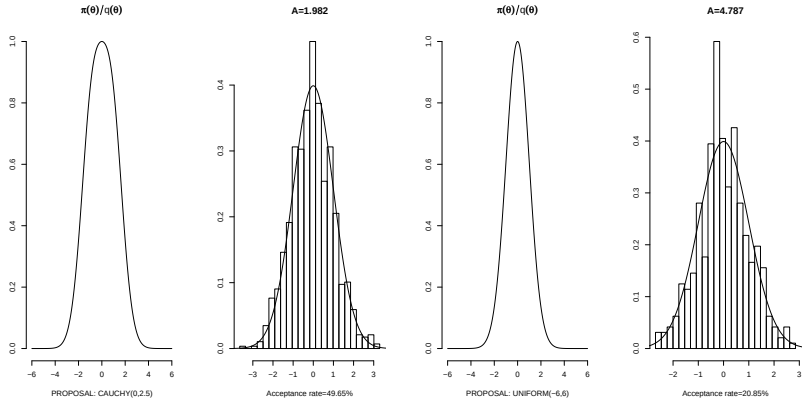
$$q_C(\theta) = \frac{1}{\pi\sqrt{2.5}} \left[1 + \frac{\theta^2}{2.5}\right]^{-1}$$

or a multiple of a Uniform density

$$q_U(\theta) = \frac{1}{12} \qquad \theta \in (-6, 6).$$

$n = 2000$ draws sampled from $q_C(\cdot)$, an observed acceptance rate of 49.65% and true acceptance rate of $1/\sqrt{1.25\pi} = 50\%$.

$n = 2000$ draws sampled from $q_U(\cdot)$, an observed acceptance rate of 20.85% and true acceptance rate of $\sqrt{2\pi}/12 = 21\%$.



Sampling importance resampling

No need to rely on the existance of A!

Algorithm

1. Draw $\theta_1^*, \ldots, \theta_n^*$ from $q(\cdot)$
2. Compute weights
$$w_i = \frac{\pi(\theta_i^*)/q(\theta_i^*)}{\sum_{j=1}^n \pi(\theta_j^*)/q(\theta_j^*)}, \quad i = 1, \ldots, n$$
3. For $j = 1, \ldots, m$, sample
$$\theta_j \text{ from } \{\theta_1^*, \ldots, \theta_n^*\}$$
such that $Pr(\theta_j = \theta_i^*) = \omega_i, \quad i = 1, \ldots, n.$

Rule of thumb: $n/m = 20$.

SIR in the Bayesian context

Let the target distribution is the posterior distribution

$$\pi(\theta) = c_\pi p(\theta) f(x|\theta)$$

A natural (but not necessarily good) choice is

$$q(\theta) = p(\theta)$$

so the weights

$$\omega_i = \frac{f(x|\theta_i)}{\sum_{j=1}^n f(x|\theta_j)}, \quad i = 1, \ldots, n$$

are the normalized likelihoods.

Example iv.

Assume that $\sigma^2/n = 4.5$, $\bar{x} = 7$, $\mu_0 = 0$ and $\tau_0^2 = 1$.

Normal model

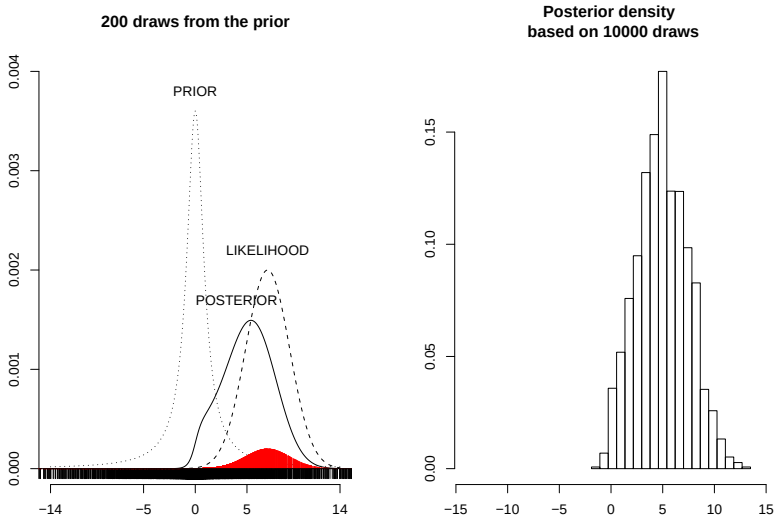
$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\}$$

Cauchy prior

$$p(\mu) \propto \frac{1}{\tau_0^2 + (\mu - \mu_0)^2}$$

Posterior

$$\pi(\mu) \propto \frac{\exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\}}{\tau_0^2 + (\mu - \mu_0)^2}$$



LECTURE 4

MARKOV CHAIN MONTE CARLO

Homogeneous Markov chain

A Markov chain is a stochastic process where given the present state, past and future states are independent, i.e.

$$Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \ldots, \theta^{(0)} \in A_0)$$

equals

$$Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x)$$

for all sets $A_0, \ldots, A_{n-1}, A \subset S$ and $x \in S$.

When the above probability does not depend on n, the chain is said to be *homogeneous* and a transition function, or kernel $P(x, A)$, can be defined as:

1. for all $x \in S$, $P(x, \cdot)$ is a probability distribution over S;
2. for all $A \subset S$, the function $x \mapsto P(x, A)$ can be evaluated.

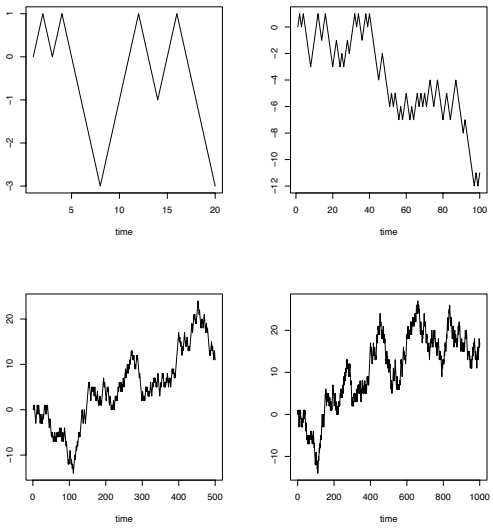
Example i. Random walk

Consider a particle moving independently left and right on the line with successive displacements from its current position governed by a probability function f over the integers and $\theta^{(n)}$ representing its position at instant n, $n \in N$. Initially, $\theta^{(0)}$ is distributed according to some distribution $\pi^{(0)}$. The positions can be related as

$$\theta^{(n)} = \theta^{(n-1)} + w_n = w_1 + w_2 + \ldots + w_n$$

where the w_i are independent random variables with probability function f. So, $\{\theta^{(n)} : n \in N\}$ is a Markov chain in Z.

The position of the chain at instant $t = n$ is described probabilistically by the distribution of $w_1 + \ldots + w_n$.



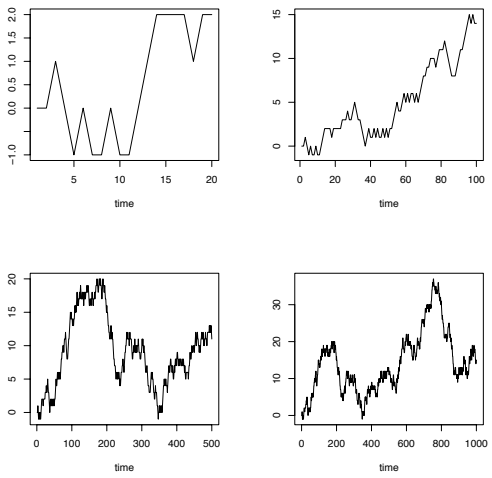
$Pr\{\theta^{(n)} = \theta^{(n-1)} + i\} = 1/2$, for $i = -1, 1$ and $\theta^{(0)} = 0.0$.

Example ii. Birth and death processes

Consider a Markov chain that from the state x can only move in the next step to one of the neighboring states $x-1$, representing a death, x or $x+1$, representing a birth. The transition probabilities are given by

$$P(x,y) = \begin{cases} p_x & \text{, if } y = x+1 \\ q_x & \text{, if } y = x-1 \\ r_x & \text{, if } y = x \\ 0 & \text{, if } |y-x| > 1 \end{cases} .$$

where p_x, q_x and r_x are non-negative with $p_x + q_x + r_x = 1$ and $q_0 = 0$.



$Pr\{\theta^{(n)} = \theta^{(n-1)} + i\} = 1/3$, for $i = -1, 0, 1$ and $\theta^{(0)} = 0.0$.

Discrete state spaces

If S is finite with r elements, $S = \{x_1, x_2, \ldots, x_r\}$, a transition matrix P with (i,j)th element given by $P(x_i, x_j)$ can be defined as

$$P = \begin{pmatrix} P(x_1,x_1) & \ldots & P(x_1,x_r) \\ \vdots & & \vdots \\ P(x_r,x_1) & \ldots & P(x_r,x_r) \end{pmatrix}.$$

Transition probabilities from state x to state y over m steps, denoted by $P^m(x,y)$, is given by the probability of a chain moving from state x to state y in exactly m steps. It can be obtained for $m \geq 2$ as

$$\begin{aligned} P^m(x,y) &= Pr(\theta^{(m)} = y | \theta^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(y, x_{m-1}, \ldots, x_1 | x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(y|x_{m-1}) \ldots Pr(x_1|x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(x_{m-1}, y) \cdots P(x, x_1) \end{aligned}$$

Chapman-Kolmogorov equations

$$\begin{aligned} P^{n+m}(x,y) &= \sum_z Pr(\theta^{(n+m)} = y | \theta^{(n)} = z, \theta^{(0)} = x) \\ &\times \quad Pr(\theta^{(n)} = z | \theta^{(0)} = x) \\ &= \sum_z P^n(x,z) P^m(z,y) \end{aligned}$$

and (more generally)

$$P^{n+m} = P^n P^m.$$

Marginal distributions

Let

$$\pi^{(n)} = (\pi^{(n)}(x_1), \cdots, \pi^{(n)}(x_r))$$

with the initial distribution of the chain when $n = 0$. Then,

$$\pi^{(n)}(y) = \sum_{x \in S} P^n(x,y) \pi^{(0)}(x)$$

or, in matrix notation,

$$\begin{aligned} \pi^{(n)} &= \pi^{(0)} P^n \\ \pi^{(n)} &= \pi^{(n-1)} P \end{aligned}$$

Example iii. 2-state Markov chain

Consider $\{\theta^{(n)} : n \geq 0\}$, a Markov chain in $S = \{0, 1\}$ with $\pi^{(0)}$ given by

$$\pi^{(0)} = (\pi^{(0)}(0), \pi^{(0)}(1))$$

and transition matrix

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

It is easy to see that

$$\begin{aligned} Pr(\theta^{(n)} = 0) &= (1-p) Pr(\theta^{(n-1)} = 0) + q Pr(\theta^{(n-1)} = 1) \\ &= (1-p-q)^n \pi^{(0)}(0) + q \sum_{k=0}^{n-1} (1-p-q)^k \end{aligned}$$

If $p + q > 0$,

$$Pr(\theta^{(n)} = 0) = \frac{q}{p+q} + (1-p-q)^n \left(\pi^{(0)}(0) - \frac{q}{p+q} \right)$$

If $0 < p + q < 2$ then

$$\begin{aligned} \lim_{n \to \infty} Pr(\theta^{(n)} = 0) &= \frac{q}{p+q} \\ \lim_{n \to \infty} Pr(\theta^{(n)} = 1) &= \frac{p}{p+q} \end{aligned}$$

Stationary distributions

A fundamental problem for Markov chains is the study of the asymptotic behavior of the chain as the number of iterations $n \to \infty$.

A key concept is that of a *stationary distribution* π. A distribution π is said to be a stationary distribution of a chain with transition probabilities $P(x, y)$ if

$$\sum_{x \in S} \pi(x) P(x, y) = \pi(y), \quad \forall y \in S$$

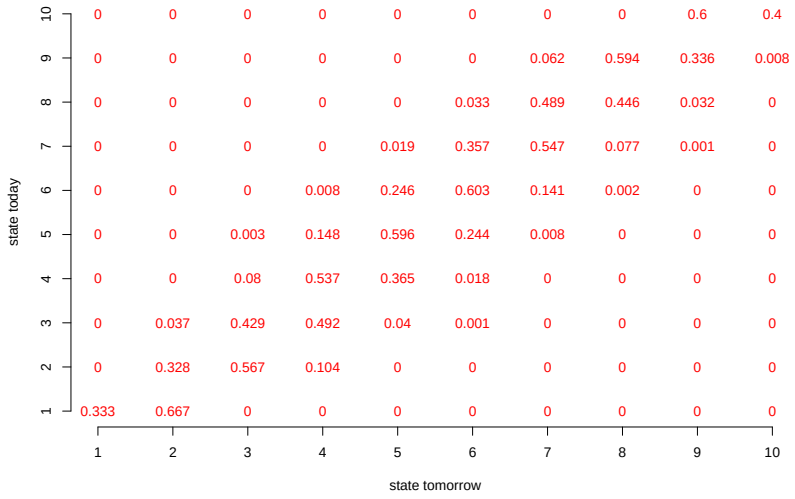
or in matrix notation as $\pi P = \pi$.

If the marginal distribution at any given step n is π then the next step distribution is $\pi P = \pi$.

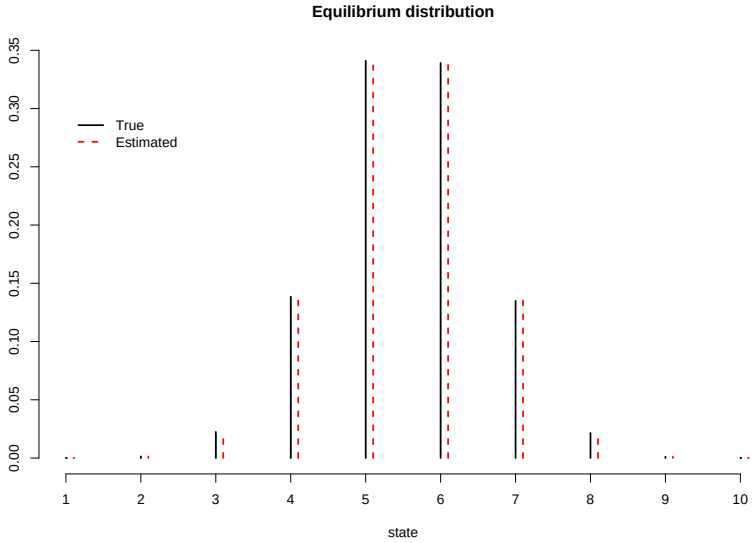
Once the chain reaches a stage where π is its distribution, all subsequent distributions are π.

π is also known as the *equilibrium distribution*.

Example iv. 10-state Markov chain



Equilibrium distribution



Ergodicity

A chain is said to be geometrically ergodic if $\exists \lambda \in [0,1)$, and a real, integrable function $M(x)$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\lambda^n \tag{1}$$

for all $x \in S$.

If $M(x) = M$, then the ergodicity is uniform.

Uniform ergodicity implies geometric ergodicity which implies ergodicity.

The smallest λ satisfying (**??**) is called the rate of convergence.

A very large value of $M(x)$ may slow down convergence considerably.

Ergodic theorem

Once ergodicity of the chain is established, important limiting theorems can be stated. The first and most important one is the ergodic theorem.

The ergodic average of a real-valued function $t(\theta)$ is the average $\bar{t}_n = (1/n)\sum_{i=1}^{n} t(\theta^{(i)})$. If the chain is ergodic and $E_\pi[t(\theta)] < \infty$ for the unique limiting distribution π then

$$\bar{t}_n \overset{a.s.}{\to} E_\pi[t(\theta)] \text{ as } n \to \infty$$

which is a Markov chain equivalent of the law of large numbers.

It states that averages of chain values also provide strongly consistent estimates of parameters of the limiting distribution π despite their dependence.

There are also versions of the central limit theorem for Markov chains.

Inefficiency factor or integrated autocorrelation time

Define the autocovariance of lag k of the chain $t^{(n)} = t(\theta^{(n)})$ as $\gamma_k = Cov_\pi(t^{(n)}, t^{(n+k)})$, the variance of $t^{(n)}$ as $\sigma^2 = \gamma_0$, the autocorrelation of lag k as $\rho_k = \gamma_k/\sigma^2$ and $\tau_n^2/n = Var_\pi(\bar{t}_n)$.

It can be shown that

$$\tau_n^2 = \sigma^2\left(1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\rho_k\right) \tag{2}$$

and that

$$\tau_n^2 \to \tau^2 = \sigma^2 \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right) \tag{3}$$

as $n \to \infty$.

The term between parentheses in Equation (**??**) can be called *inefficiency factor* or *integrated autocorrelation time* because it measures how far $t^{(n)}$s are from being a random sample and how much $Var_\pi(\bar{t}_n)$ increases because of that.

Effective sample size

The inefficiency factor can be used to derive the *effective sample size*

$$n_{\text{eff}} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k} \tag{4}$$

which can be thought of as the size of a random sample with the same variance since

$$Var_\pi(\bar{t}_n) = \sigma^2 / n_{\text{eff}}.$$

It is important to distinguish between

$$\sigma^2 = Var_\pi[t(\theta)] \quad \text{and} \quad \tau^2$$

the variance of $t(\theta)$ under the limiting distribution π and the limiting sampling variance of $\sqrt{n}\,\bar{t}$, respectively.

Note that under independent sampling they are both given by σ^2. They are both variability measures but the first one is a characteristic of the limiting distribution π whereas the second is the uncertainty of the averaging procedure.

Central limit theorem

If a chain is uniformly (geometrically) ergodic and $t^2(\theta)$ ($t^{2+\epsilon}(\theta)$) is integrable with respect to π (for some $\epsilon > 0$) then

$$\frac{\bar{t}_n - E_\pi[t(\theta)]}{\tau/\sqrt{n}} \xrightarrow{d} N(0,1), \tag{5}$$

as $n \to \infty$.

Just as (**??**) provides theoretical support for the use of ergodic averages as estimates, Equation (**??**) provides support for evaluation of approximate confidence intervals.

Reversible chains

Let $(\theta^{(n)})_{n\geq 0}$ be an homogeneous Markov chain with transition probabilities $P(x,y)$ and stationary distribution π.

Assume that one wishes to study the sequence of states $\theta^{(n)}, \theta^{(n-1)}, \ldots$ in reversed order. It can be shown that this sequence is a Markov chain with transition probabilities are

$$\begin{aligned} P_n^*(x,y) &= Pr(\theta^{(n)} = y \mid \theta^{(n+1)} = x) \\ &= \frac{Pr(\theta^{(n+1)} = x \mid \theta^{(n)} = y) Pr(\theta^{(n)} = y)}{Pr(\theta^{(n+1)} = x)} \\ &= \frac{\pi^{(n)}(y) P(y,x)}{\pi^{(n+1)}(x)} \end{aligned}$$

and in general the chain is not homogeneous.

If $n \to \infty$ or alternatively, $\pi^{(0)} = \pi$, then

$$P_n^*(x, y) = P^*(x, y) = \pi(y)P(y, x)/\pi(x)$$

and the chain becomes homogeneous.

If $P^*(x, y) = P(x, y)$ for all x and $y \in S$, the Markov chain is said to be *reversible*. The reversibility condition is usually written as

$$\pi(x)P(x, y) = \pi(y)P(y, x) \tag{6}$$

for all $x, y \in S$.

It can be interpreted as saying that the rate at which the system moves from x to y when in equilibrium, $\pi(x)P(x, y)$, is the same as the rate at which it moves from y to x, $\pi(y)P(y, x)$.

For that reason, (**??**) is sometimes referred to as the *detailed balance equation*; *balance* because it equates the rates of moves through states and *detailed* because it does it for every possible pair of states.

MCMC: a bit of history

Dongarra and Sullivan (2000) [1] list the top 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century (in chronological order):

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform

40s and 50s

Stan Ulam soon realized that computers could be used in this fashion to answer questions of neutron diffusion and mathematical physics;

He contacted John Von Neumann and they developed many Monte Carlo algorithms (importance sampling, rejection sampling, etc);

In the 1940s Nick Metropolis and Klari Von Neumann designed new controls for the state-of-the-art computer (ENIAC);

Metropolis and Ulam (1949) The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335-341;

Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

[1] Guest Editors' Introduction: The Top 10 Algorithms, *Computing in Science and Engineering*, **2**, 22-23.

70s

Hastings and his student Peskun showed that Metropolis and the more general Metropolis-Hastings algorithm are particular instances of a larger family of algorithms.

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

Peskun (1973) Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607-612.

80s and 90s

Geman and Geman (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Pearl (1987) Evidential reasoning using stochastic simulation. *Artificial Intelligence*, **32**, 245-257.

Tanner and Wong (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528-550.

Gelfand and Smith (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Metropolis-Hastings

A sequence $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$ is drawn from a Markov chain whose *limiting equilibrium distribution* is the posterior distribution, $\pi(\theta)$.

Algorithm

1. Initial value: $\theta^{(0)}$
2. Proposed move: $\theta^* \sim q(\theta^*|\theta^{(i-1)})$
3. Acceptance scheme:

$$\theta^{(i)} = \begin{cases} \theta^* & \text{com prob. } \alpha \\ \theta^{(i-1)} & \text{com prob. } 1-\alpha \end{cases}$$

where

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta^{(i-1)})} \frac{q(\theta^{(i-1)}|\theta^*)}{q(\theta^*|\theta^{(i-1)})}\right\}$$

Special cases

1. Symmetric chains: $q(\theta|\theta^*) = q(\theta^*|\theta)$

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta)}\right\}$$

2. Independence chains: $q(\theta|\theta^*) = q(\theta)$

$$\alpha = \min\left\{1, \frac{\omega(\theta^*)}{\omega(\theta)}\right\}$$

where $\omega(\theta^*) = \pi(\theta^*)/q(\theta^*)$.

Random walk Metropolis

The most famous symmetric chain is the random walk Metropolis:

$$q(\theta|\theta^*) = q(|\theta - \theta^*|)$$

Hill climbing: when

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta)}\right\}$$

a value θ^* with higher density $\pi(\theta^*)$ greater than $\pi(\theta)$ is automatically accepted.

Example v. Bivariate mixture of normals

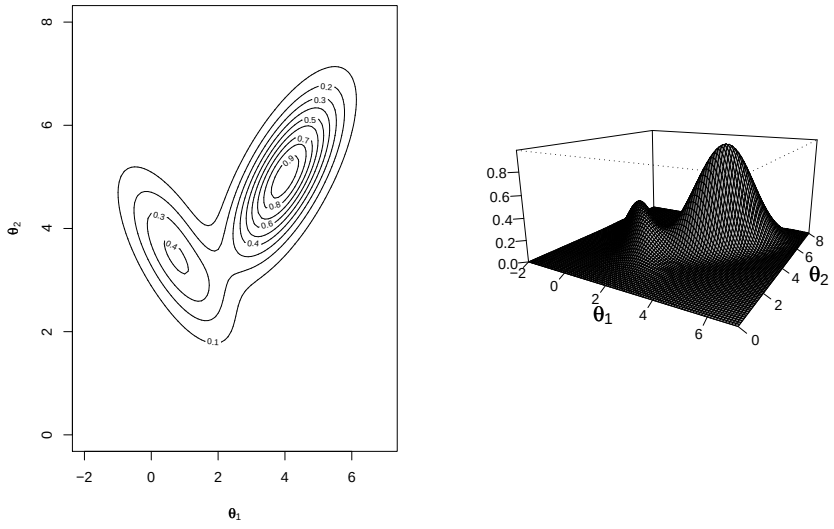
Tthe target distribution is a two-component mixture of bivariate normal densities, ie:

$$\pi(\theta) = 0.7 f_N(\theta; \mu_1, \Sigma_1) + 0.3 f_N(\theta; \mu_2, \Sigma_2).$$

where

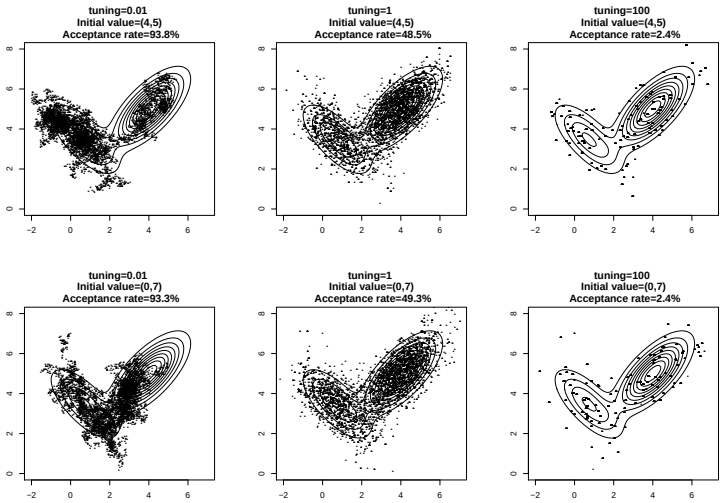
$$\begin{aligned}
\mu_1' &= (4.0, 5.0) \\
\mu_2' &= (0.7, 3.5) \\
\Sigma_1 &= \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix} \\
\Sigma_2 &= \begin{pmatrix} 1.0 & -0.7 \\ -0.7 & 1.0 \end{pmatrix} .
\end{aligned}$$

Targe distribution

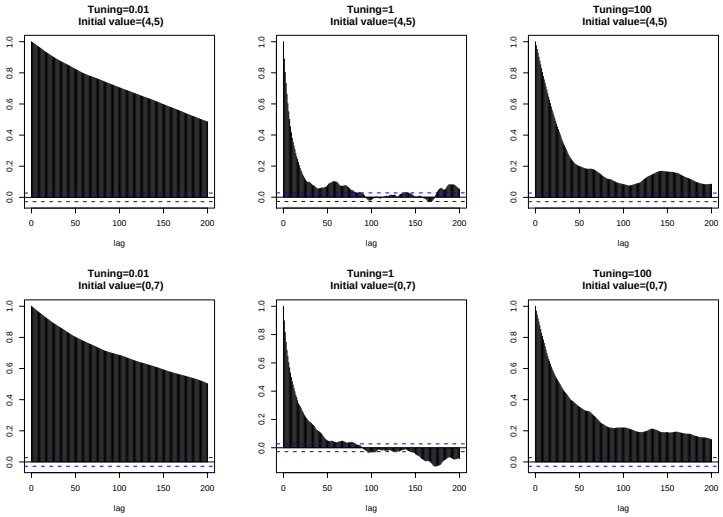


Random walk Metropolis: draws

$q(\theta, \phi) = f_N(\phi; \theta, \nu I_2)$ and ν =tuning.

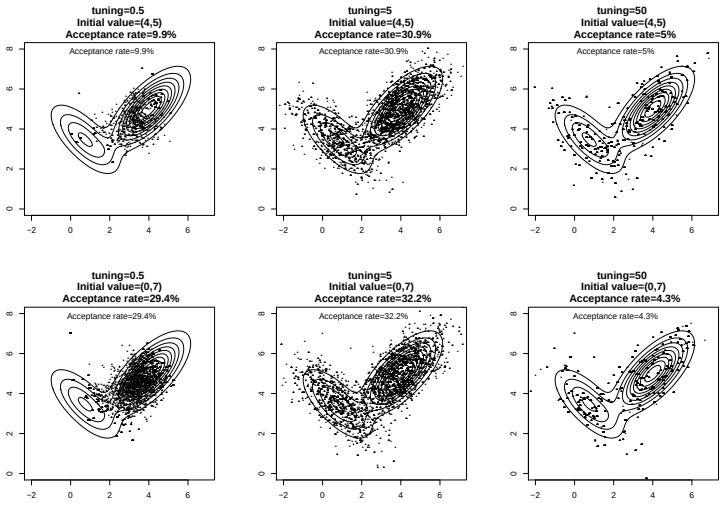


Random walk Metropolis: autocorrelations

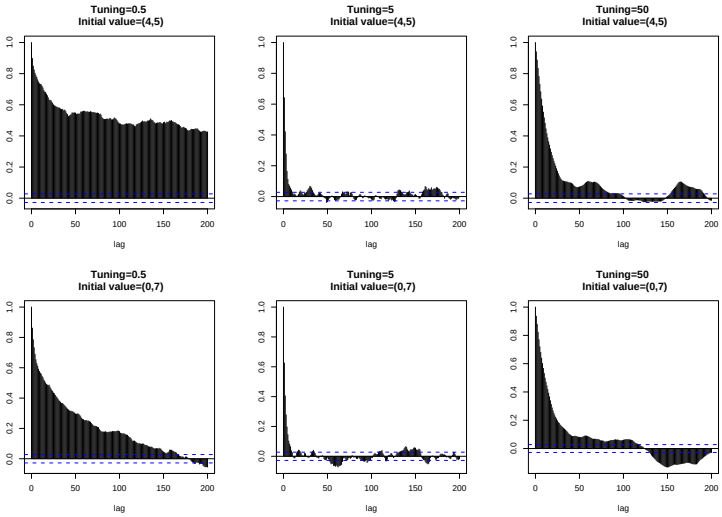


Independent Metropolis

$q(\theta, \phi) = f_N(\phi; \mu_3, \nu I_2)$ and $\mu_3 = (3.01, 4.55)'$.



Independence Metropolis: autocorrelations



Gibbs sampler

Technically, the Gibbs sampler is an MCMC scheme whose transition kernel is the product of the full conditional distributions.

Algorithm

1. Start at $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots)$
2. Sample the components of $\theta^{(j)}$ iteratively:

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots) \\
\theta_2^{(j)} &\sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots) \\
\theta_3^{(j)} &\sim \pi(\theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \ldots) \\
&\vdots
\end{aligned}$$

The Gibbs sampler opened up a new way of approaching statistical modeling by combining simpler structures (the full conditional models) to address the more general structure (the full model).

Example vi: Poisson with a change point

$y_1, \ldots, y_n$ is a sample from a Poisson distribution.

There is a suspicion of a single change point m along the observation process.

Given m, the observation distributions are

$$\begin{aligned} y_i|\lambda &\sim Poi(\lambda), \quad i = 1, \ldots, m \\ y_i|\phi &\sim Poi(\phi), \quad i = m+1, \ldots, n. \end{aligned}$$

Independent prior distributions

$$\begin{aligned} \lambda &\sim G(\alpha, \beta) \\ \phi &\sim G(\gamma, \delta) \\ m &\sim U\{1, \ldots, n\} \end{aligned}$$

with α, β, γ and δ known hyperparameters.

Posterior distribution

Combining the prior and the likelihood

$$\begin{aligned} \pi(\lambda, \phi, m) &\propto f(y_1, \ldots, y_n|\lambda, \phi, m)p(\lambda, \phi, m) \\ &= \prod_{i=1}^{m} f_P(y_i; \lambda) \prod_{i=m+1}^{n} f_P(y_i; \phi) \\ &\times f_G(\lambda; \alpha, \beta) f_G(\phi; \gamma, \delta) \frac{1}{n} \end{aligned}$$

Therefore,

$$\pi(\lambda, \phi, m) \propto \left(\lambda^{\alpha + s_m - 1} e^{-(\beta+m)\lambda}\right) \left(\phi^{\gamma + s_n - s_m - 1} e^{-(\delta+n-m)\phi}\right)$$

where $s_l = \sum_{i=1}^{l} y_i$ for $l = 1, \ldots, n$.

Full conditional distributions

The full conditional distributions for λ, ϕ and m are

$$\pi(\lambda|m) = G(\alpha + s_m, \beta + m)$$

$$\pi(\phi|m) = G(\gamma + s_n - s_m, \delta + n - m)$$

and

$$\pi(m|\lambda, \phi) = \frac{\lambda^{\alpha + s_m - 1} e^{-(\beta+m)\lambda} \phi^{\gamma + s_n - s_m - 1} e^{-(\delta+n-m)\phi}}{\sum_{l=1}^{n} \lambda^{\alpha + s_l - 1} e^{-(\beta+l)\lambda} \phi^{\gamma + s_n - s_l - 1} e^{-(\delta+n-l)\phi}},$$

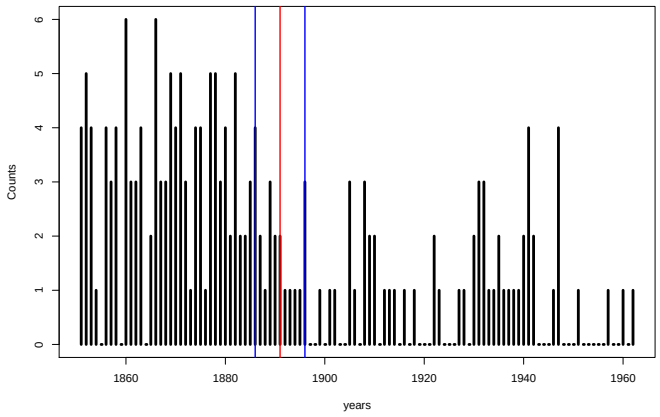
for $m = 1, \ldots, n$, respectively.

Coal mining disasters in Great Britain

This model was applied to the $n = 112$ observed counts of coal mining disasters in Great Britain by year from 1851 to 1962.

Sample mean from 1951 to 1891 = 3.098

Sample mean from 1892 to 1962 = 0.901

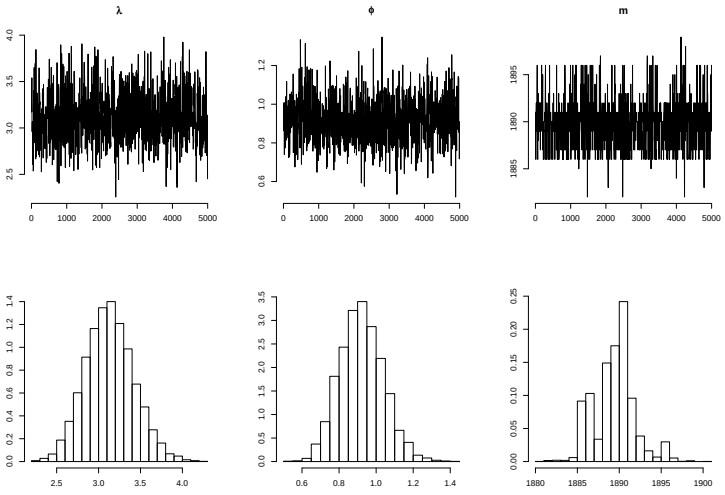


Markov chains

The Gibbs sampler run: 5000 iterations

Starting point: $m^{(0)} = 1891$

Hyperparameters: $\alpha = \beta = \gamma = \delta = 0.001$



Exact and approximate posterior summary

Exact posterior can be obtained by analytically deriving $\pi(m)$ and using it to derive $\pi(\lambda)$ and $\pi(\phi)$.

Par.	Mean	Var	95% C.I.
λ	3.120	0.280	(2.571,3.719)
ϕ	0.923	0.113	(0.684,0.963)
m	1890	2.423	(1886,1895)

Approximate posterior summary based on the Gibbs sampler

Par.	Mean	Var	95% C.I.
λ	3.131	0.290	(2.582,3.733)
ϕ	0.922	0.118	(0.703,1.167)
m	1890	2.447	(1886,1896)

Example vii: AR(1) with normal errors

Let us assume that

$$y_t = \rho y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

for $t = 1, \ldots, n$.

Prior specification

$$\begin{aligned} y_0 &\sim N(m_0, C_0) \\ \rho &\sim N(r_0, V_0) \\ \sigma^2 &\sim IG(n_0/2, n_0 s_0^2/2) \end{aligned}$$

for known hyperparameters m_0, C_0, r_0, V_0, n_0 and s_0^2.

Full conditional distributions

- $(\rho|\sigma^2, y_0, y_{1:n}) \sim N(r_1, V_1)$

$$\begin{aligned} V_1^{-1} &= V_0 + \sigma^{-2} \sum_{t=1}^{n} y_{t-1}^2 \\ V_1^{-1} r_1 &= V_0^{-1} r_0 + \sigma^{-2} \sum_{t=1}^{n} y_{t-1} y_t \end{aligned}$$

- $(\sigma^2|\rho, y_0, y_{1:n}) \sim IG(n_1/2, n_1 s_1^2/2)$

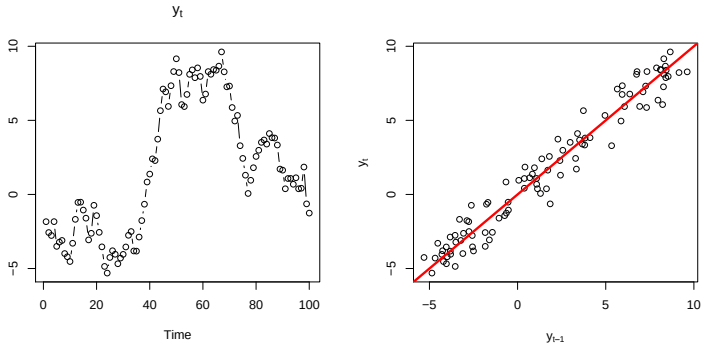
$$\begin{aligned} n_1 &= n_0 + n \\ n_1 s_1^2 &= n_0 s_0^2 + \sum_{t=1}^{n} (y_t - \rho y_{t-1})^2 \end{aligned}$$

- $(y_0|\rho, \sigma^2, y_{1:n}) \sim N(m_1, C_1)$

$$\begin{aligned} C_1^{-1} &= C_0 + \sigma^{-2}\rho^2 \\ C_1^{-1} m_1 &= C_0^{-1} m_0 + \sigma^{-2} \rho y_1 \end{aligned}$$

Simulated data

Set up: $n = 100$, $y_0 = 0.0$, $\rho = 0.95$ and $\sigma^2 = 1.0$.



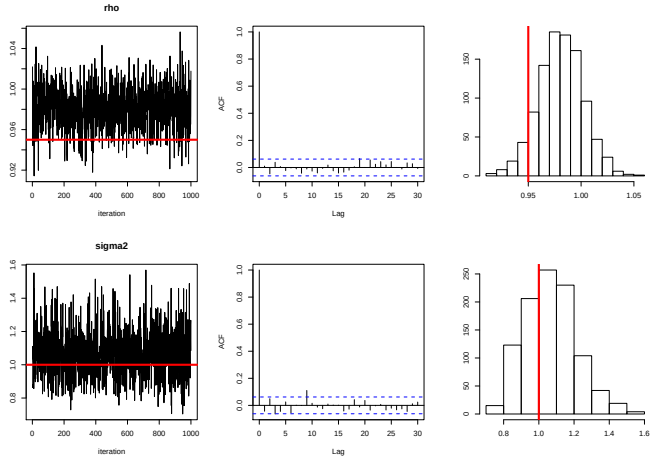
MCMC output

Gibbs sampler run: $(M_0, M, L) = (1000, 1000, 1)$

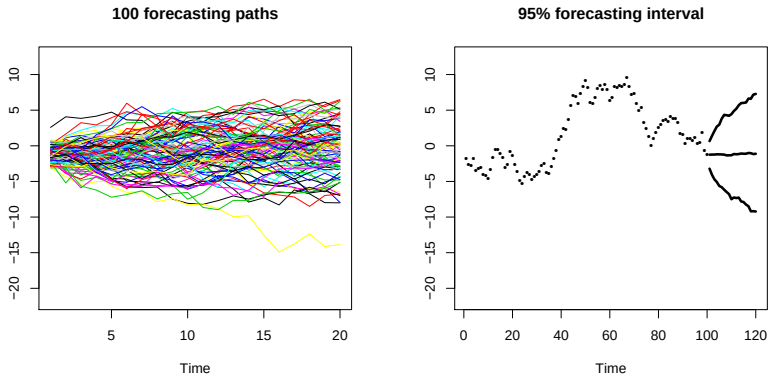
Starting point: true values

Prior of (ρ, σ^2): $r_0 = 0.9$, $V_0 = 10$, $n_0 = 5$ and $s_0^2 = 1$.

Prior of y_0: $m_0 = 0$ and $C_0 = 10$.



Forecasting



Example viii: AR(1) with drift and normal errors

Let us assume that

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

for $t = 1, \ldots, n$.

Prior specification

$$\begin{aligned} y_0 &\sim N(m_0, C_0) \\ \mu &\sim N(\mu_0, W_0) \\ \rho &\sim N(r_0, V_0) \\ \sigma^2 &\sim IG(n_0/2, n_0 s_0^2/2) \end{aligned}$$

for known hyperparameters $m_0, C_0, \mu_0, W_0, r_0, V_0, n_0$ and s_0^2.

Full conditional distributions

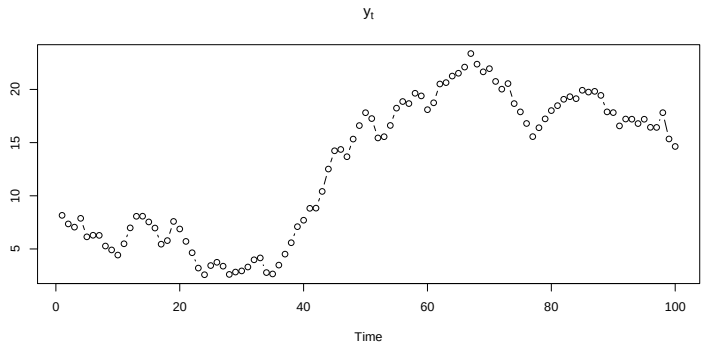
- $(\rho|\mu, \sigma^2, y_0, y_{1:n})$ as before with y_t replaced by $y_t - \mu$.
- $(\sigma^2|\mu, \rho, y_0, y_{1:n})$ as before with y_t replaced by $y_t - \mu$.
- $(y_0|\mu, \rho, \sigma^2, y_{1:n})$ as before with y_1 replaced by $y_1 - \mu$.

- $(\mu|\rho, \sigma^2, y_0, y_{1:n}) \sim N(\mu_1, W_1)$

$$\begin{aligned} W_1^{-1} &= W_0^{-1} + n/\sigma^2 \\ W_1^{-1}\mu_1 &= W_0^{-1}\mu_0 + \sum_{t=1}^{n}(y_t - \rho y_{t-1})/\sigma^2 \end{aligned}$$

Simulated data

Set up: $n = 100$, $y_0 = 0.0$, $\mu = 0.1$, $\rho = 0.99$ and $\sigma^2 = 1.0$.

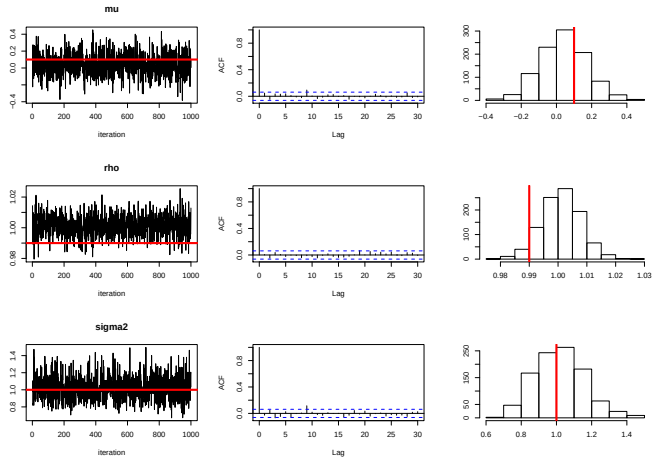


MCMC output

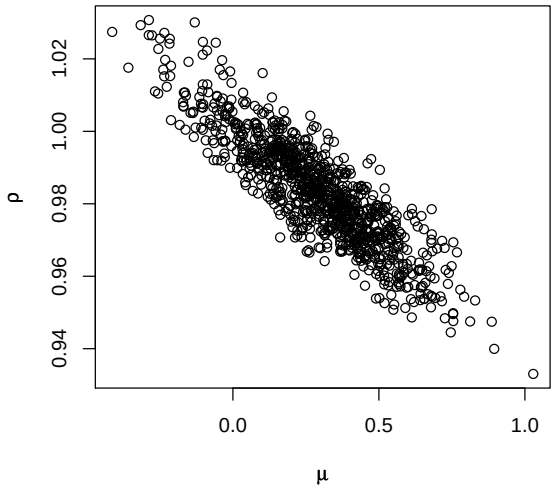
Gibbs sampler run: $(M_0, M, L) = (1000, 1000, 1)$

Starting point: true values

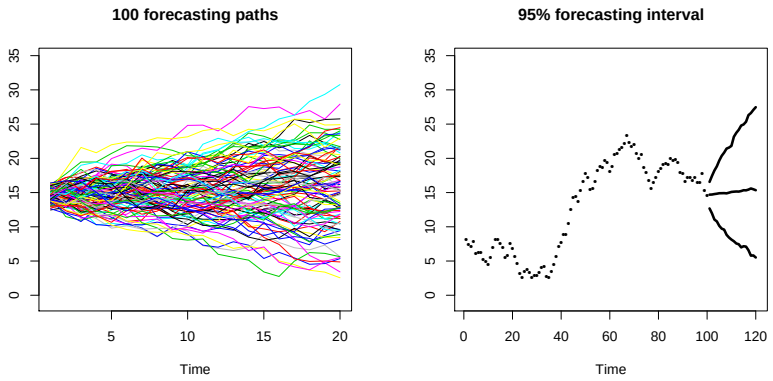
Prior of (μ, ρ, σ^2): $\mu_0 = 0$, $W_0 = 10$, $r_0 = 0.9$, $V_0 = 10$, $n_0 = 5$ and $s_0^2 = 1$. Prior of y_0: $m_0 = 0$ and $C_0 = 10$.



Joint posterior of (μ, ρ)



Forecasting



Example ix: GARCH(1,1) model with normal errors

Let us assume that

$$\begin{aligned} y_t &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= a_1 + a_2 y_{t-1}^2 + a_3 \sigma_{t-1}^2 \end{aligned}$$

for $t = 1, \ldots, n$.

Prior specification

$$\begin{aligned} y_0 &\sim N(m_0, V_0) \\ \sigma_0^2 &\sim IG(n_0/2, n_0 s_0^2/2) \\ a_i &\sim N(a_{0i}, V_{0i}) \qquad i = 1, \ldots, 3 \end{aligned}$$

for known m_0, C_0, n_0, s_0^2 and (α_{0i}, V_{0i}) for $i = 1, 2, 3$.

Metropolis-Hastings algorithm

None of the full conditionals for a_1, a_2 and a_3 are of known form nor easy to sample from.

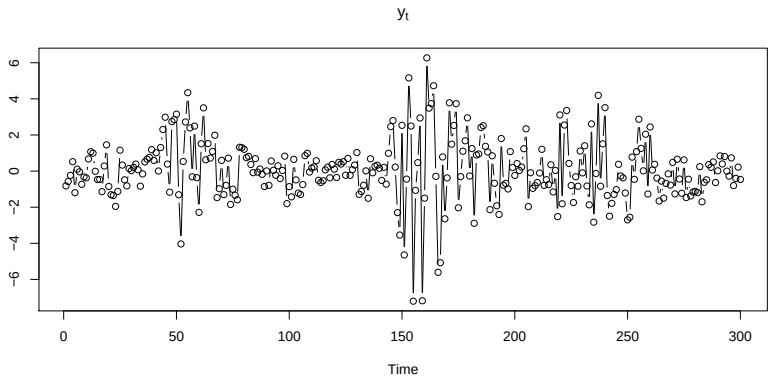
Here we implement a very simple M-H algorithm where a new vector $a^* = (a_1^*, a_2^*, a_3^*)$ is generated from

$$a^* \sim N(a^{(j)}, v^2 I_3)$$

where $a^{(j)}$ is the current state of the chain and v is a tuning standard deviations.

Simulated data

Set up: $n = 300$, $a_1 = 0.1$, $a_2 = 0.4$, $a_3 = 0.59$, $y_0^2 = \sigma_0^2 = 0.1$.



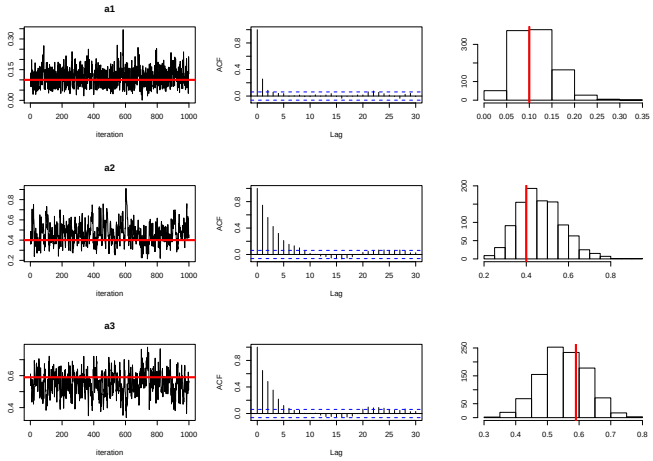
MCMC output

Gibbs sampler run: $(M_0, M, L) = (10000, 1000, 100)$.

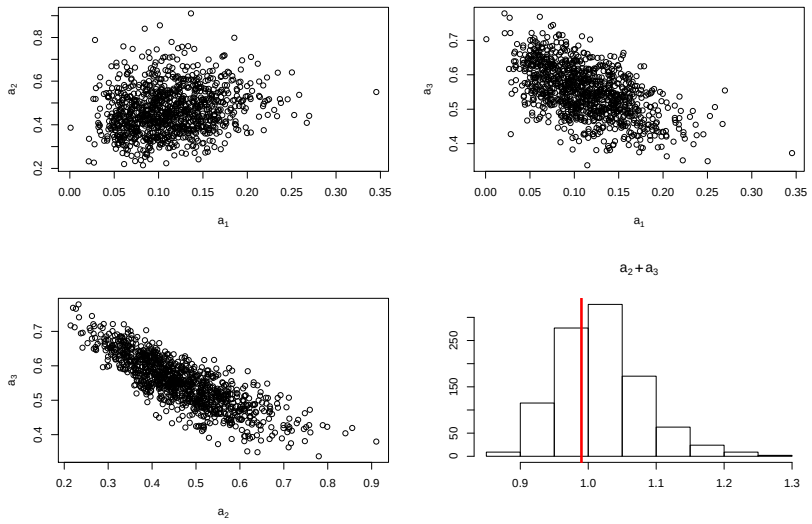
Starting point: true values.

Prior of (μ, ρ, σ^2): $a_{0i} = 0$ and $C_{0i} = 10$ for $i = 1, 2, 3$.

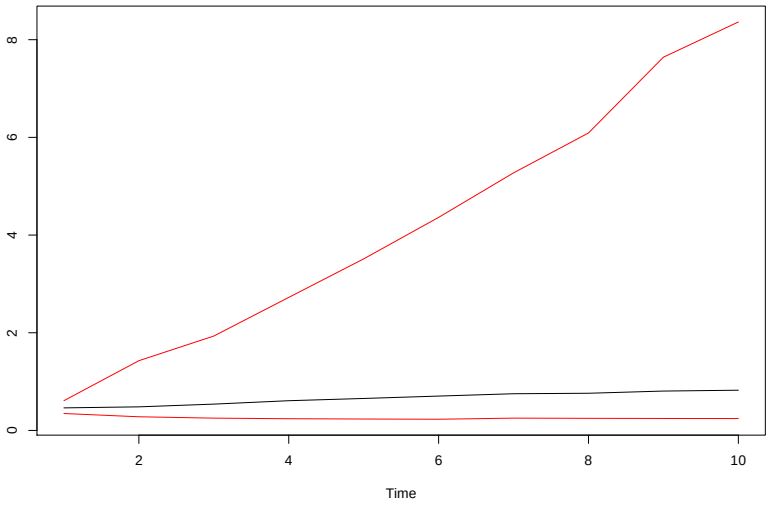
Metropolis-Hastings tuning variance: $v^2 = 0.01^2$.



Pairwise joint posterior distributions



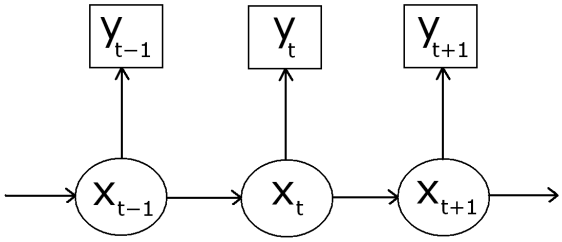
$p(\sigma^2_{n+h}|y_1, \ldots, y_n)$ **for** $h = 1, \ldots, 10$



LECTURE 5

DYNAMIC LINEAR MODELS

Example i. local level model



The local level model (West and Harrison, 1997) is

$$\begin{aligned} y_{t+1}|x_{t+1},\theta &\sim N(x_{t+1},\sigma^2) \\ x_{t+1}|x_t,\theta &\sim N(x_t,\tau^2) \end{aligned}$$

where $x_0 \sim N(m_0, C_0)$ and $\theta = (\sigma^2, \tau^2)$ fixed (for now).

Example i. Evolution, prediction and updating

Let $y^t = (y_1, \ldots, y_t)$.

$$p(x_t|y^t) \Longrightarrow p(x_{t+1}|y^t) \Longrightarrow p(y_{t+1}|x_t) \Longrightarrow p(x_{t+1}|y^{t+1})$$

- Posterior at t: $(x_t|y^t) \sim N(m_t, C_t)$
- Prior at $t+1$: $(x_{t+1}|y^t) \sim N(m_t, R_{t+1})$
- Marginal likelihood: $(y_{t+1}|y^t) \sim N(m_t, Q_{t+1})$
- Posterior at $t+1$: $(x_{t+1}|y^{t+1}) \sim N(m_{t+1}, C_{t+1})$

where $R_{t+1} = C_t + \tau^2$, $Q_{t+1} = R_{t+1} + \sigma^2$, $A_{t+1} = R_{t+1}/Q_{t+1}$, $C_{t+1} = A_{t+1}\sigma^2$, and $m_{t+1} = (1 - A_{t+1})m_t + A_{t+1}y_{t+1}$.

Example i. Backward smoothing

For $t = n$, $x_n|y^n \sim N(m_n^n, C_n^n)$, where

$$\begin{aligned} m_n^n &= m_n \\ C_n^n &= C_n \end{aligned}$$

For $t < n$, $x_t|y^n \sim N(m_t^n, C_t^n)$, where

$$\begin{aligned} m_t^n &= (1-B_t)m_t + B_t m_{t+1}^n \\ C_t^n &= (1-B_t)C_t + B_t^2 C_{t+1}^n \end{aligned}$$

and

$$B_t = \frac{C_t}{C_t + \tau^2}$$

Example i. Backward sampling

For $t = n$, $x_n|y^n \sim N(a_n^n, R_n^n)$, where

$$\begin{aligned} a_n^n &= m_n \\ R_n^n &= C_n \end{aligned}$$

For $t < n$, $x_t|x_{t+1}, y^n \sim N(a_t^n, R_t^n)$, where

$$\begin{aligned} a_t^n &= (1-B_t)m_t + B_t x_{t+1} \\ R_t^n &= B_t \tau^2 \end{aligned}$$

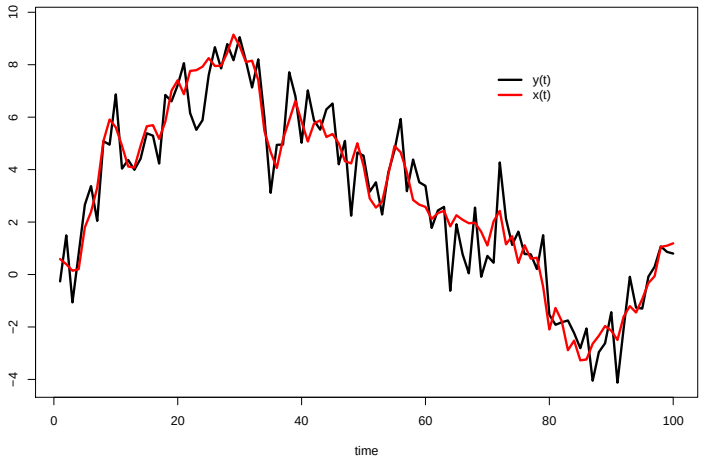
and

$$B_t = \frac{C_t}{C_t + \tau^2}$$

This is basically the Forward filtering, backward sampling algorithm used to sample from $p(x^n|y^n)$ (Carter and Kohn, 1994 and Frühwirth-Schnatter, 1994).

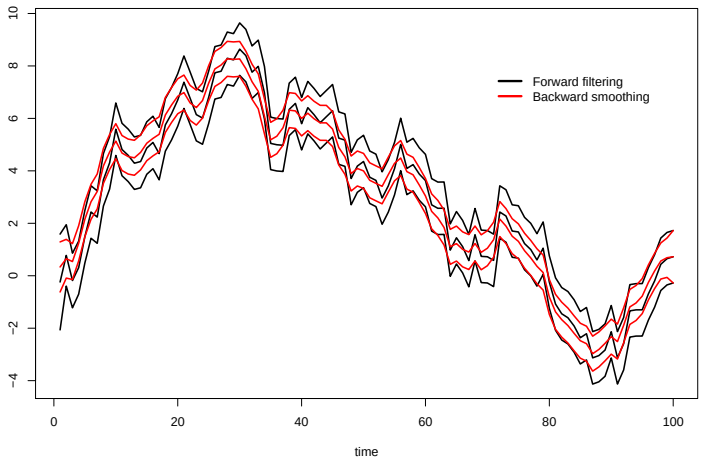
Example i. Simulated data

$n = 100$, $\sigma^2 = 1.0$, $\tau^2 = 0.5$ and $x_0 = 0$.



Example i. $p(x_t|y^t, \theta)$ **versus** $p(x_t|y^n, \theta)$

$m_0 = 0.0$ and $C_0 = 10.0$



Example i. Integrating out states x^n

We showed earlier that

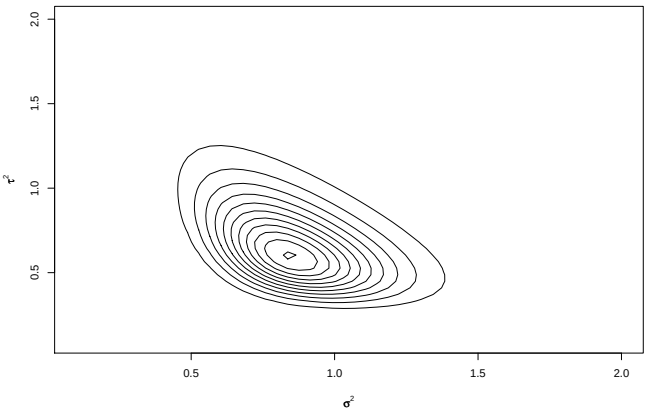
$$(y_t|y^{t-1}) \sim N(m_{t-1}, Q_t)$$

where both m_{t-1} and Q_t were presented before and are functions of $\theta = (\sigma^2, \tau^2)$, y^{t-1}, m_0 and C_0.

Therefore, by Bayes' rule,

$$\begin{aligned} p(\theta|y^n) &\propto p(\theta)p(y^n|\theta) \\ &= p(\theta)\prod_{t=1}^{n} f_N(y_t|m_{t-1}, Q_t). \end{aligned}$$

Example i. $p(y|\sigma^2, \tau^2)$



Example i. MCMC scheme

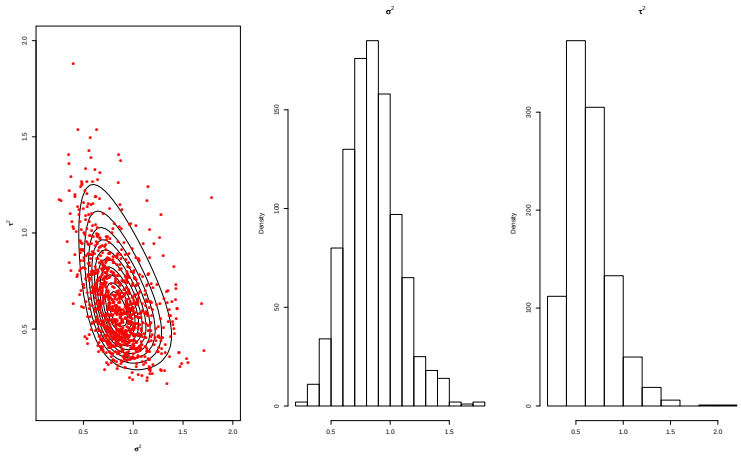
- Sample θ from $p(\theta|y^n, x^n)$

$$p(\theta|y^n, x^n) \propto p(\theta) \prod_{t=1}^{n} p(y_t|x_t, \theta) p(x_t|x_{t-1}, \theta).$$

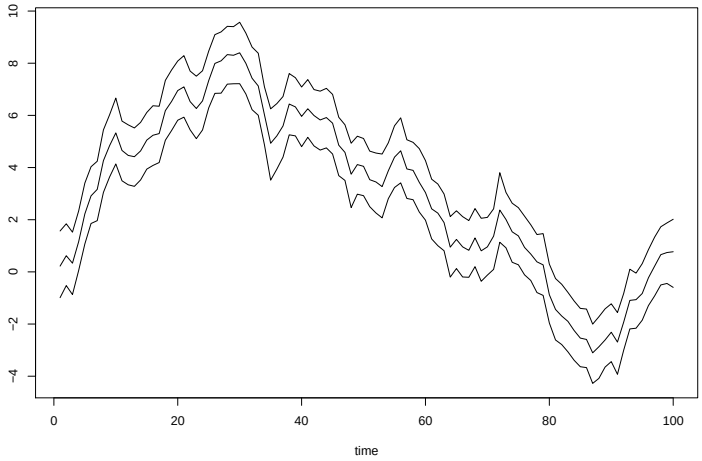
- Sample x^n from $p(x^n|y^n, \theta)$

$$p(x^n|y^n, \theta) = \prod_{t=1}^{n} f_N(x_t|a_t^n, R_t^n)$$

Example i. $p(\sigma^2, \tau^2|y^n)$



Example i. $p(x_t|y^n)$



Lessons from Example i.

Sequential learning in non-normal and nonlinear dynamic models $p(y_{t+1}|x_{t+1})$ and $p(x_{t+1}|x_t)$ in general rather difficult since

$$\begin{aligned} p(x_{t+1}|y^t) &= \int p(x_{t+1}|x_t)p(x_t|y^t)dx_t \\ p(x_{t+1}|y^{t+1}) &\propto p(y_{t+1}|x_{t+1})p(x_{t+1}|y^t) \end{aligned}$$

are usually unavailable in closed form.

Over the last 20 years:

- FFBS for conditionally Gaussian DLMs;
- Gamerman (1998) for generalized DLMs;
- Carlin, Polson and Stoffer (2002) for more general DMs.

Dynamic linear models

Large class of models with time-varying parameters.

Dynamic linear models are defined by a pair of equations, the *observation equation* and the *evolution/system equation*:

$$\begin{aligned} y_t &= F_t'\beta_t + \epsilon_t, \qquad \epsilon_t \sim N(0, V) \\ \beta_t &= G_t\beta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W) \end{aligned}$$

- y_t: sequence of observations;
- F_t: vector of explanatory variables;
- β_t: d-dimensional state vector;
- G_t: $d \times d$ evolution matrix;
- $\beta_1 \sim N(a, R)$.

Example ii. Linear growth model

The linear growth model is slightly more elaborate by incorporation of an extra time-varying parameter β_2 representing the growth of the level of the series:

$$\begin{aligned} y_t &= \beta_{1,t} + \epsilon_t \quad \epsilon_t \sim N(0, V) \\ \beta_{1,t} &= \beta_{1,t-1} + \beta_{2,t} + \omega_{1,t} \\ \beta_{2,t} &= \beta_{2,t-1} + \omega_{2,t} \end{aligned}$$

where $\omega_t = (\omega_{1,t}, \omega_{2,t})' \sim N(0, W)$ and

$$\begin{aligned} F_t &= (1, 0)' \\ G_t &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

Prior, updated and smoothed distributions

Prior distributions

$$p(\beta_t|y^{t-k}) \qquad k > 0$$

Updated/online distributions

$$p(\beta_t|y^t)$$

Smoothed distributions

$$p(\beta_t|y^{t+k}) \qquad k > 0$$

Sequential inference

Let $y^t = \{y_1, \ldots, y_t\}$.

Posterior at time $t-1$:

$$\beta_{t-1}|y^{t-1} \sim N(m_{t-1}, C_{t-1})$$

Prior at time t:

$$\beta_t|y^{t-1} \sim N(a_t, R_t)$$

with $a_t = G_t m_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W$.

predictive at time t:

$$y_t|y^{t-1} \sim N(f_t, Q_t)$$

with $f_t = F_t' a_t$ and $Q_t = F_t' R_t F_t + V$.

Posterior at time t

$$p(\beta_t|y^t) = p(\beta_t|y_t, y^{t-1}) \propto p(y_t|\beta_t)\, p(\beta_t|y^{t-1})$$

The resulting posterior distribution is

$$\beta_t|y^t \sim N(m_t, C_t)$$

with

$$\begin{aligned} m_t &= a_t + A_t e_t \\ C_t &= R_t - A_t A_t' Q_t \\ A_t &= R_t F_t / Q_t \\ e_t &= y_t - f_t \end{aligned}$$

By induction, these distributions are valid for all times.

Smoothing

In dynamic models, the smoothed distribution $\pi(\beta|y^n)$ is more commonly used:

$$\begin{aligned}\pi(\beta|y^n) &= p(\beta_n|y^n)\prod_{t=1}^{n-1} p(\beta_t|\beta_{t+1},\ldots,\beta_n,y^n)\\ &= p(\beta_n|y^n)\prod_{t=1}^{n-1} p(\beta_t|\beta_{t+1},y^t)\end{aligned}$$

Integrating with respect to $(\beta_1,\ldots,\beta_{t-1})$:

$$\begin{aligned}\pi(\beta_t,\ldots,\beta_n|y^n) &= p(\beta_n|y^n)\prod_{k=t}^{n-1} p(\beta_k|\beta_{k+1},y^t)\\ \pi(\beta_t,\beta_{t+1}|y^n) &= p(\beta_{t+1}|y^n)p(\beta_t|\beta_{t+1},y^t)\end{aligned}$$

for $t = 1,\ldots,n-1$.

Marginal smoothed distributions

It can be shown that

$$\beta_t|y^n \sim N(m_t^n, C_t^n)$$

where

$$\begin{aligned}m_t^n &= m_t + C_t G_{t+1}' R_{t+1}^{-1}(m_{t+1}^n - a_{t+1})\\ C_t^n &= C_t - C_t G_{t+1}' R_{t+1}^{-1}(R_{t+1} - C_{t+1}^n)R_{t+1}^{-1}G_{t+1}C_t\end{aligned}$$

Individual sampling from $\pi(\beta_t|\beta_{-t},y^n)$

Let $\beta_{-t} = (\beta_1,\ldots,\beta_{t-1},\beta_{t+1},\ldots,\beta_n)$.

For $t = 2,\ldots,n-1$

$$\begin{aligned}\pi(\beta_t|\beta_{-t},y^n) &\propto p(y_t|\beta_t)\,p(\beta_{t+1}|\beta_t)\,p(\beta_t|\beta_{t-1})\\ &\propto f_N(y_t;F_t'\beta_t,V)f_N(\beta_{t+1};G_{t+1}\beta_t,W)\\ &\times f_N(\beta_t;G_t\beta_{t-1},W)\\ &= f_N(\beta_t;b_t,B_t)\end{aligned}$$

where

$$\begin{aligned}b_t &= B_t(\sigma^{-2}F_t y_t + G_{t+1}'W^{-1}\beta_{t+1} + W^{-1}G_t\beta_{t-1})\\ B_t &= (\sigma^{-2}F_tF_t' + G_{t+1}'W^{-1}G_{t+1} + W^{-1})^{-1}\end{aligned}$$

for $t = 2,\ldots,n-1$.

For $t = 1$ and $t = n$,

$$\pi(\beta_1|\beta_{-1},y^n) = f_N(\beta_1;b_1,B_1)$$

and

$$\pi(\beta_n|\beta_{-n},y^n) = f_N(\beta_t;b_n,B_n)$$

where

$$\begin{aligned} b_1 &= B_1(\sigma_1^{-2}F_1y_1 + G_2'W^{-1}\beta_2 + R^{-1}a) \\ B_1 &= (\sigma_1^{-2}F_1F_1' + G_2'W^{-1}G_2 + R^{-1})^{-1} \\ b_n &= B_n(\sigma_n^{-2}F_ny_n + W^{-1}G_n\beta_{n-1}) \\ B_n &= (\sigma_n^{-2}F_nF_n' + W^{-1})^{-1} \end{aligned}$$

The FFBS algorithm: sampling from $\pi(\beta|y^n)$

For $t = 1, \ldots, n-1$, it can be shown that

$$(\beta_t|\beta_{t+1}, V, W, y^t)$$

is normally distributed with mean

$$(G_t'W^{-1}G_t + C_t^{-1})^{-1}(G_t'W^{-1}\beta_{t+1} + C_t^{-1}m_t)$$

and variance $(G_t'W^{-1}G_t + C_t^{-1})^{-1}$.

Sampling from $\pi(\beta|y^n)$ can be performed by

- Sampling β_n from $N(m_n, C_n)$ and then
- Sampling β_t from $(\beta_t|\beta_{t+1}, V, W, y^t)$, for $t = n-1, \ldots, 1$.

The above scheme is known as the forward filtering, backward sampling (FFBS) algorithm (Carter and Kohn, 1994 and Frühwirth-Schnatter, 1994).

Sampling from $\pi(V, W|y^n, \beta)$

Assume that

$$\begin{aligned} \phi = V^{-1} &\sim Gamma(n_\sigma/2, n_\sigma S_\sigma/2) \\ \Phi = W^{-1} &\sim Wishart(n_W/2, n_W S_W/2) \end{aligned}$$

Full conditionals

$$\begin{aligned} \pi(\phi|\beta, \Phi) &\propto \prod_{t=1}^{n} f_N(y_t; F_t'\beta_t, \phi^{-1})\ f_G(\phi; n_\sigma/2, n_\sigma S_\sigma/2) \\ &\propto f_G(\phi; n_\sigma^*/2, n_\sigma^* S_\sigma^*/2) \\ \pi(\Phi|\beta, \phi) &\propto \prod_{t=2}^{n} f_N(\beta_t; G_t\beta_{t-1}, \Phi^{-1})\ f_W(\Phi; n_W/2, n_W S_W/2) \\ &\propto f_W(\Phi; n_W^*/2, n_W^* S_W^*/2) \end{aligned}$$

where $n_\sigma^* = n_\sigma + n$, $n_W^* = n_W + n - 1$,

$$\begin{aligned} n_\sigma^* S_\sigma^* &= n_\sigma S_\sigma + \sigma(y_t - F_t'\beta_t)^2 \\ n_W^* S_W^* &= n_W S_W + \Sigma_{t=2}^{n}(\beta_t - G_t\beta_{t-1})(\beta_t - G_t\beta_{t-1})' \end{aligned}$$

Gibbs sampler for (β, V, W)

- Sample V^{-1} from its full conditional
$$f_G(\phi; n_\sigma^*/2, n_\sigma^* S_\sigma^*/2)$$
- Sample W^{-1} from its full conditional
$$f_W(\Phi; n_W^*/2, n_W^* S_W^*/2)$$
- Sample β from its full conditional
$$\pi(\beta|y^n, V, W)$$
by the FFBS algorithm.

Likelihood for (V, W)

It is easy to see that

$$p(y^n|V,W) = \prod_{t=1}^{n} f_N(y_t|f_t, Q_t)$$

which is the integrated likelihood of (V, W).

Jointly sampling (β, V, W)

(β, V, W) can be sampled jointly by

- Sampling (V, W) from its marginal posterior
$$\pi(V, W|y^n) \propto l(V, W|y^n)\pi(V, W)$$
by a rejection or Metropolis-Hastings step;
- Sampling β from its full conditional
$$\pi(\beta|y^n, V, W)$$
by the FFBS algorithm.

Jointly sampling (β, V, W) avoids MCMC convergence problems associated with the posterior correlation between model parameters (Gamerman and Moreira, 2002).

Example iii. Comparing sampling schemes

Based on Gamerman, Reis and Salazar (2006) Comparison of sampling schemes for dynamic linear models. *International Statistical Review*, 74, 203-214.

First order DLM with $V = 1$

$$\begin{aligned} y_t &= \beta_t + \epsilon_t\,, \qquad \epsilon_t \sim N(0,1) \\ \beta_t &= \beta_{t-1} + \omega_t\,, \quad \omega_t \sim N(0, W), \end{aligned}$$

with $(n, W) \in \{(100, .01), (100, .5), (1000, .01), (1000, .5)\}$.

400 runs: 100 replications per combination.

Priors: $\beta_1 \sim N(0, 10)$ and V and W have inverse Gammas with means set at true values and coefficients of variation set at 10.

Posterior inference: based on 20,000 MCMC draws.

Schemes

Scheme I: Sampling $\beta_1, \ldots, \beta_n, V$ and W from their conditionals.

Scheme II: Sampling β, V and W from their conditionals.

Scheme III: Jointly sampling (β, V, W).

Scheme	n=100	n=1000
II	1.7	1.9
III	1.9	7.2

Computing times relative to scheme I. For instance, when $n = 100$ it takes almost 2 times as much to run scheme III.

		Scheme		
W	n	I	II	III
0.01	1000	242	8938	2983
0.01	100	3283	13685	12263
0.50	1000	409	3043	963
0.50	100	1694	3404	923

Sample averages (based on the 100 replications) of effective sample size n_{eff} based on V.

Example iv. Spatial dynamic factor model

Based on Lopes, Salazar and Gamerman (2008) Spatial Dynamic Factor Models. *Bayesian Analysis*, 3, 759-792.

Let us consider a simple version of their model

$$y_t|f_t, \theta \sim N(\beta f_t, \Sigma) \tag{7}$$

$$f_t|f_{t-1}, \theta \sim N(\Gamma f_{t-1}, \Lambda) \tag{8}$$

where $\theta = (\beta, \Sigma, \Gamma, \Lambda)$,

$y_t = (y_{1t}, \ldots, y_{Nt})'$ is the N-dimensional vector of observations (locations $s_1, \ldots, s_N$ and times $t = 1, \ldots, T$),

f_t is an m-dimensional vector of common factors, for $m < N$.

$\beta = (\beta_{(1)}, \ldots, \beta_{(m)})$ is the $N \times m$ matrix of factor loadings.

Spatial loadings

The j^{th} column of β, denoted by $\beta_{(j)} = (\beta_{(j)}(s_1), \ldots, \beta_{(j)}(s_N))'$, for $j = 1, \ldots, m$, is modeled as a conditionally independent, distance-based Gaussian random field (GRF), i.e.

$$\beta_{(j)} \sim GRF(\tau_j^2 \rho_{\phi_j}(\cdot)) \equiv N(0, \tau_j^2 R_{\phi_j}), \tag{9}$$

where the (l,k)-element of R_{ϕ_j} is given by $r_{lk} = \rho_{\phi_j}(|s_l - s_k|)$, $l, k = 1, \ldots, N$, for suitably defined correlation functions $\rho_{\phi_j}(\cdot)$, $j = 1, \ldots, m$.

Matérn spatial aucorrelation function

$$\rho_\phi(d) = 2^{1-\phi_2}\Gamma(\phi_2)^{-1}(d/\phi_1)^{\phi_2}\mathcal{K}_{\phi_2}(d/\phi_1)$$

where $\mathcal{K}_{\phi_2}(\cdot)$ is the modified Bessel function of the second kind and of order ϕ_2.

Key references on spatial statistics are Cressie (1993) and Stein (1999).

Forecasting

One is usually interested in learning about the h-steps ahead predictive density $p(y_{T+h}|y)$, i.e.

$$p(y_{T+h}|y) = \int p(y_{T+h}|f_{T+h},\theta)p(f_{T+h}|f_T,\theta)p(f_T,\theta|y)df_{T+h}df_T d\theta \tag{10}$$

where

$$\begin{aligned}(y_{T+h}|f_{T+h},\theta) &\sim N(\beta f_{T+h},\Sigma)\\ (f_{T+h}|f_T,\theta) &\sim N(\mu_h,V_h)\end{aligned}$$

and

$$\mu_h = \Gamma^h f_T \quad \text{and} \quad V_h = \sum_{k=1}^{h}\Gamma^{k-1}\Lambda(\Gamma^{k-1})'$$

for $h > 0$.

$p(y_{T+h}|y)$ can be easily approximated by Monte Carlo integration.

Sampling the common factors

Joint distribution $p(F|y) = \prod_{t=0}^{T-1} p(f_t|f_{t+1},D_t)p(f_T|D_T)$, where $D_t = \{y_1,\ldots,y_t\}$, $t = 1,\ldots,T$ and D_0 represents the initial information.

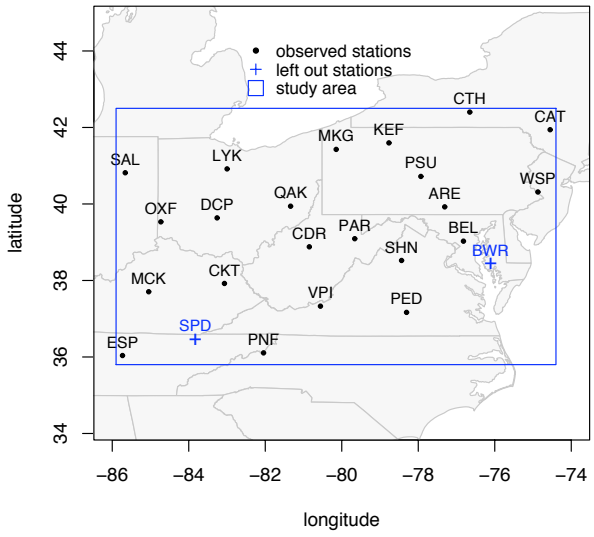
Forward filtering: Starting with $f_{t-1}|D_{t-1} \sim N(m_{t-1},C_{t-1})$, it can be shown that

$$f_t|D_t \sim N(m_t,C_t)$$

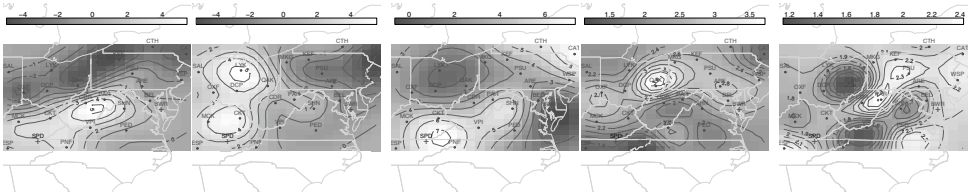
where $m_t = a_t + A_t(y_t - \tilde{y}_t)$, $C_t = R_t - A_t Q_t A_t'$, $a_t = \Gamma m_{t-1}$, $R_l = \Gamma C_{t-1}\Gamma' + \Lambda$, $\tilde{y}_t = \beta a_t$, $Q_t = \beta R_t \beta' + \Sigma$ and $A_t = R_t\beta' Q_t^{-1}$, for $t = 1,\ldots,T$.

Backward sampling: f_T is sampled from $p(f_T|D_T)$. For $t \leq T-1$, f_t is sampled from $p(f_t|f_{t+1},D_t) = f_N(f_t;\tilde{a}_t,\tilde{C}_t)$, where $\tilde{a}_t = m_t + B_t(f_{t+1} - a_{t+1})$, $\tilde{C}_t = C_t - B_t R_{t+1} B_t'$ and $B_t = C_t\Gamma' R_{t+1}^{-1}$.

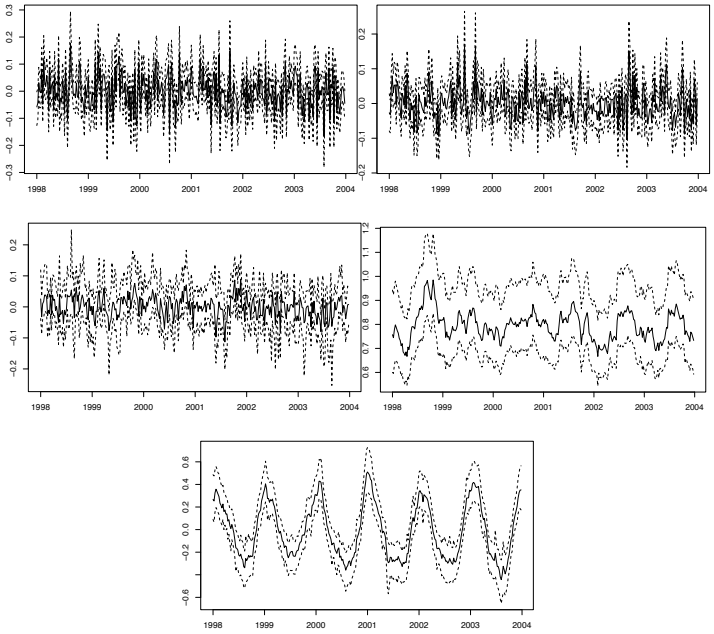
SO_2 in Eastern US



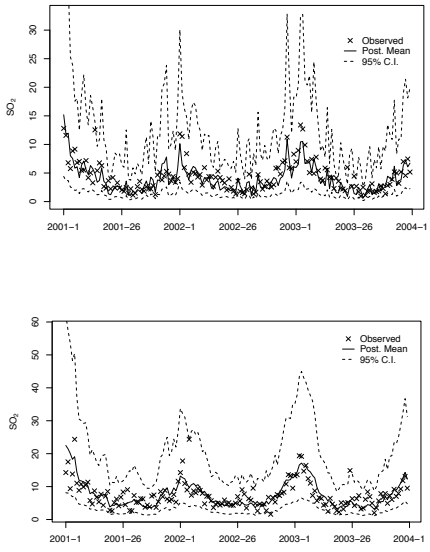
Spatial factor loadings



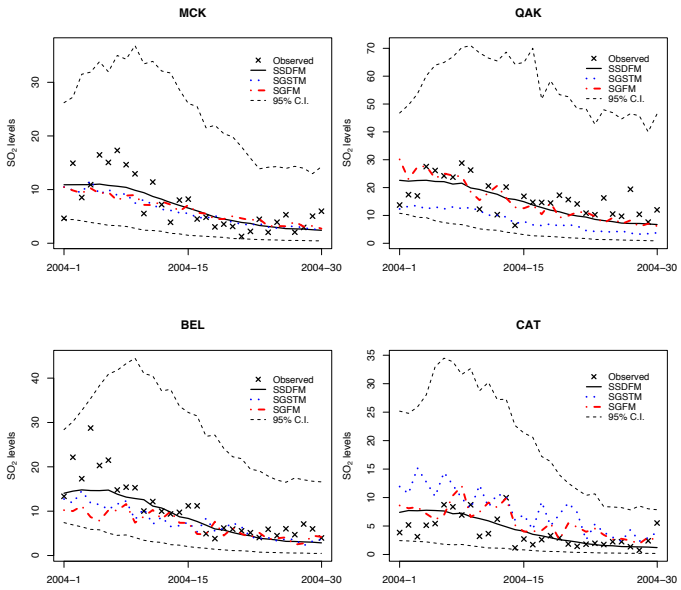
Dynamic factors



Spatial interpolation (stations SPD and BWR)



Forecasting



LECTURE 6

NONNORMAL AND NONLINEAR DYNAMIC MODELS

Generalized linear models

An extension of regression models still preserving linearity and influence of covariates through the mean response is given by generalized linear models.

The observations remain independent but now have distributions in the exponential family.

The model is

$$\begin{aligned} f(y_i|\theta_i) &= a(y_i)\,\exp\{y_i\theta_i + b(\theta_i)\} \\ E(y_i|\theta_i) &= -b'(\theta_i) = \mu_i \\ g(\mu_i) &= \eta_i \\ \eta_i &= x_{i1}\beta_1 + \ldots + x_{id}\beta_d \end{aligned}$$

for $i = 1, \ldots, n$, where the link function g is differentiable.

Binomial regression

Consider $y_i|\pi_i \sim bin(n_i, \pi_i)$, $i = 1, \ldots, n$, and assume that the probabilities π_i are determined by the values of a variable x.

The π_i lie between 0 and 1 and can be associated to a distribution function.

One possibility is the normal distribution and in this case

$$\pi_i = \Phi(\alpha + \beta x_i), \quad i = 1, \ldots, n$$

where Φ is the distribution function of the N(0,1) distribution and α and β are constants.

The binomial distribution belongs to the exponential family and the link function $g_1 = \Phi^{-1}$ is differentiable.

The structure of a generalized linear model is completed with the linear predictor

$$\eta_i = \alpha + \beta x_i, \quad i = 1, \ldots, n.$$

Other possible links include the logistic and complementary log-log transformations

$$\begin{aligned} g_2(\pi_i) &= \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ g_3(\pi_i) &= \log\left\{\log\left(\frac{1}{1-\pi_i}\right)\right\} \end{aligned}$$

associated respectively to the logistic and extreme-value distributions.

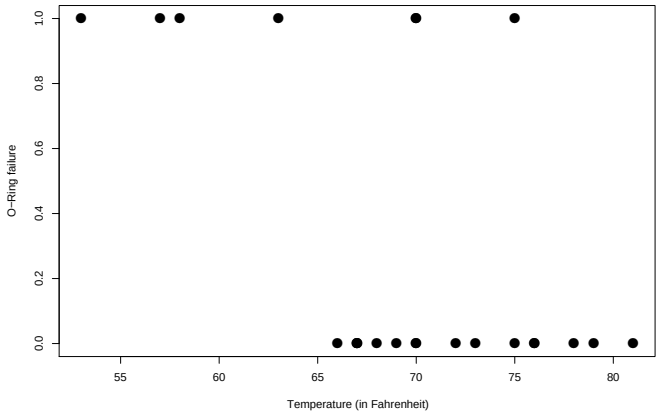
Note that g_1, g_2 and g_3 take numbers from [0,1] to the real line.

Example i. O-ring data

Christensen (1997) analyzed binary observations of O-ring failures y_i (1=failure) in relation to temperature t_i (Fahrenheit).

$y = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0)$

$t = (53, 57, 58, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 81)$



- Bernoulli model: $y_i|\pi_i \sim Bern(\pi_i)$, for $i = 1, \ldots, n = 23$.
- Link function: $g(\pi_i) = \alpha + \beta(t_i - \bar{t})$
- Prior: $\beta \sim N(0, V_\beta)$
- Other constants: $\alpha = -1.26$ and $\bar{t} = 69.6$.
- Links

$$\begin{aligned} \text{Logit} &: g_1(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \\ \text{Probit} &: g_2(\pi) = \Phi(\pi) \\ \text{Log-log} &: g_3(\pi) = \log\left\{\log\left(\frac{1}{1-\pi}\right)\right\} \end{aligned}$$

- Prior variances
  - $V_\beta = 1.0$
  - $V_\beta = 10.0$
  - $V_\beta = 100.0$

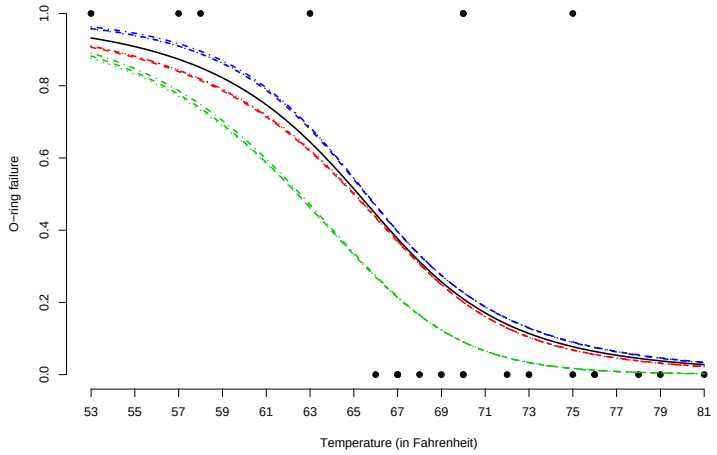
Posterior model probability (PMP)

Assume that $Pr(M_i) = 1/9$ for all i.

V_β	Link	PMP
1	Log-log	0.361
1	Logit	0.311
10	Log-log	0.115
10	Logit	0.101
100	Log-log	0.037
100	Logit	0.033
1	Probit	0.030
10	Probit	0.010
100	Probit	0.003

$Pr(y = 1|t)$

Bayesian model averaging (black), logit models (red), probit models (green) and complementary log-log models (blue).



Dynamic generalized linear model

Dynamic generalized models were introduced by West, Harrison and Migon (1985).

The model is

$$\begin{aligned} f(y_t|\theta_t) &= a(y_t)\,\exp\{y_t\theta_t + b(\theta_t)\} \\ E(y_t|\theta_t) &= \mu_t \\ g(\mu_t) &= F_t'\beta_t \\ \beta_t &= G_t\beta_{t+1} + w_t \end{aligned}$$

with $w_t \sim N(0, W_t)$ and the link function g is again differentiable.

The model is completed with a prior $\beta_1 \sim N(a, R)$.

It combines the prior specification of normal dynamic models with the observational structure of generalized linear models.

Dynamic binomial and Poisson regressions

Dynamic logistic regression with a series of binomial observations y_t with respective success probabilities π_t dynamically related to explanatory variables $x = (x_1, \ldots, x_d)'$ through the logistic link $\text{logit}(\pi_t) = x_t'\beta_t$.

Poisson counts with means λ_t dynamically related through multiplicative perturbations $\lambda_t = \lambda_{t-1} w_t^*$. After a logarithmic transformation, one obtains $\log \lambda_t = \log \lambda_{t-1} + w_t$ with $w_t = \log w_t^*$.

Posterior inference via MCMC

Assuming that the variances of the system disturbances are constant, the model parameters are given by the state parameters $\beta = (\beta_1, \ldots, \beta_n)'$ and the system variance $W = \Phi^{-1}$.

The model is specified with the observation and system equations and completed with the independent prior distributions $\beta_1 \sim N(a, R)$ and $\Phi \sim W(n_W/2, n_W S_W/2)$.

The posterior distribution is given by

$$\pi(\beta, \Phi) \propto \prod_{t=1}^{n} f(y_t|\beta_t) \prod_{i=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\beta_1) p(\Phi) \ .$$

Full conditional for Φ

$$\begin{aligned}
\pi_\Phi(\Phi) &\propto \prod_{t=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\Phi) \\
&\propto \prod_{t=2}^{n} |\Phi|^{1/2} \exp\left\{-\frac{1}{2}\text{tr}[(\beta_t - G_t\beta_{t-1})(\beta_t - G_t\beta_{t-1})'\Phi]\right\} \\
&\times |\Phi|^{[n_W-(p+1)]/2} \exp\left\{-\frac{1}{2}\text{tr}(n_W S_W \Phi)\right\} \\
&\propto |\Phi|^{[n_W^*-(d+1)]/2} \exp\left\{-\frac{1}{2}\text{tr}\left[(n_W^* S_W^*)\,\Phi\right]\right\} \ .
\end{aligned}$$

that is the density of the $W(n_W^*/2, n_W^* S_W^*/2)$ distribution with

$$\begin{aligned}
n_W^* &= n_W + n - 1 \\
n_W^* S_W^* &= n_W S_W + \sum_{t=2}^{n} (\beta_t - G_t\beta_{t-1})(\beta_t - G_t\beta_{t-1})'
\end{aligned}$$

Full conditionals for β

For block β

$$\begin{aligned}
\pi_\beta(\beta) &\propto \prod_{t=1}^{n} f(y_t|\beta_t) \prod_{t=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\beta_1) \\
&\propto \exp\left\{\sum_{t=1}^{n} [y_t\theta_t + b(\theta_t)] - \frac{1}{2}\sum_{t=1}^{n} (\beta_t - G_t\beta_{t-1})'\Phi(\beta_t - G_t\beta_{t-1})\right\} \ .
\end{aligned}$$

For block β_t, $t = 2, \ldots, n-1$

$$\begin{aligned}\pi_t(\beta_t) &\propto f(y_t|\beta_t)\, p(\beta_t|\beta_{t-1}, \Phi) p(\beta_{t+1}|\beta_t, \Phi) \\ &\propto \exp\{y_t\theta_t + b(\theta_t)\} \exp\left\{-\frac{1}{2}\left[(\beta_t - G_t\beta_{t-1})'\Phi(\beta_t - G_t\beta_{t-1})\right.\right. \\ &+ \left.\left.(\beta_{t+1} - G_{t+1}\beta_t)'\Phi(\beta_{t+1} - G_{t+1}\beta_t)\ \right]\right\}\ .\end{aligned}$$

Similar results follow for blocks β_1 and β_n.

Sampling schemes

Knorr-Held (1997) suggested the use of independence chains with prior proposals.

Shephard and Pitt (1997) used independence chains with proposals based on both prior and a normal approximation to the likelihood.

Ravines (2005) used independence normal proposals for the block β with moments given by the approximation of West, Harrison and Migon (1985).

Singh and Roberts (1982) and Fahrmeir and Wagenpfeil (1997) extended to the dynamic setting the method of mode evaluation for static regression.

An alternative previously discussed is the reparametrization in terms of the system disturbances w_t (Gamerman, 1998)

Example ii. Generalized Spatial Dynamic Factor Model

Let $\{s_1, \ldots, s_N\}$ be the N spatial locations in the region of study S, where $S \subset R^2$ and $y_t = (y_{t1}, \ldots, y_{tN})$ be the N-dimensional vector of measurements at time t, for $t = 1, \ldots, T$.

The GSDFM of Lopes, Gamerman and Salazar (2009) is a hierarchical model with first level measurement equation for conditionally independent univariate observations y_{ti} in the one-parameter natural exponential family, i.e.

$$p(y_{ti} \mid \eta_{ti}, \psi) = \exp\{\psi[y_{ti}\eta_{ti} - b(\eta_{ti})] + c(y_{ti}, \psi)\}$$

where η_{ti} is the natural parameter and ψ is a dispersion parameter.

The mean and variance of y_{ti} are, respectively, $b'(\eta_{ti})$ and $b''(\eta_{ti})/\psi$.

The natural parameter η_{ti} is deterministically defined by a linear combination of spatial and temporal components through the link function υ, i.e. $\eta_{ti} = \upsilon(\theta_{ti})$.

The GSDFM is then completed by specifying the spatio-temporal dependence of the θ_{ti}s.

Temporal and spatial variations

The temporal behavior of y_t is modeled by two levels of hierarchy

$$\begin{aligned}\theta_t &= \mu_t + \beta f_t \\ f_t \mid f_{t-1} &\sim N(\Gamma f_{t-1}, \Lambda)\end{aligned}$$

where μ_t is the mean level of the space-time process, f_t contains m common factors, $f_0 \sim N(m_0, C_0)$ and Λ is the evolutional variance.

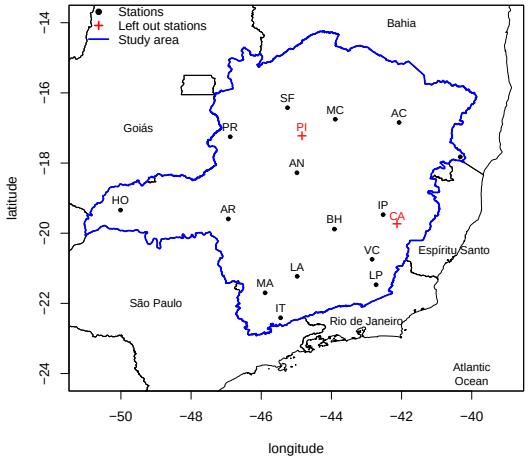
The spatial variation of y_t is modeled via the columns of the factor loadings matrix β, i.e. $\beta_j = (\beta_{1j}, \ldots, \beta_{Nj})'$, for $j = 1, \ldots, m$, is modeled as

$$\beta_j \sim N(\kappa_j, \tau_j^2 R_j)$$

where κ_j is a N-dimensional mean vector. The (l,k)-element of $R_j = R(\phi_j)$ is given by $r_{lk} = \rho_{\phi_j}(|s_l - s_k|)$, $l, k = 1, \ldots, N$, for suitably defined correlation functions $\rho_{\phi_j}(\cdot)$.

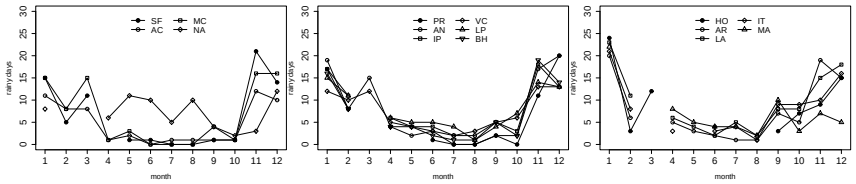
Example ii. Modeling rainfall in Minas Gerais, Brazil

$T = 365$ daily occurrences of rain in 2005 measured at 17 meteorological stations in the state of Minas Gerais, Brazil.



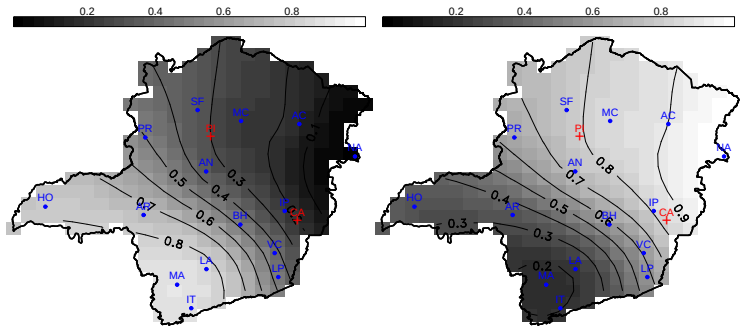
The solid line contours the state. Pirapora (PI) and Caratinga (CA) states (+) were retained for a spatial interpolation exercise.

Counts



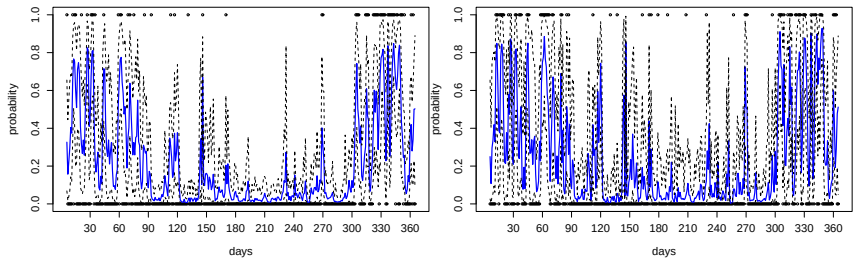
Monthly counts of rainy days for north and northeast stations (top left), center and southeast stations (top right) and south and southwest stations (bottom).

Posterior summary



Mean probability maps for two typical days in 2005 (January 6th and March 2nd) for gauged stations (dot) and ungauged stations (+).

Posterior summary



Daily posterior probability of rain for ungauged Pirapora (left) and Caratinga (right) stations in 2005. Dots are observed rain indicators, solid lines are rain mean probabilities and dashed lines are 95% credibility intervals.

Example iii: Nonlinear model

Let y_t, for $t = 1, \ldots, n$, be generated by the following nonlinear dynamic model

$$\begin{aligned}(y_t|x_t,\psi) &\sim N(x_t^2/20, \sigma^2)\\ (x_t|x_{t-1},\psi) &\sim N(G'_{x_{t-1}}\theta, \tau^2)\\ x_0 &\sim N(m_0, C_0)\end{aligned}$$

where $G'_{x_t} = \big(x_t, x_t/(1+x_t^2), \cos(1.2t)\big)$, $\theta = (\alpha, \beta, \gamma)'$ and $\psi = (\xi', \sigma^2, \tau^2)$.

Prior distribution

$$\begin{aligned}\sigma^2 &\sim IG(n_0/2, n_0\sigma_0^2/2)\\ \theta|\tau^2 &\sim N(\theta_0, \tau^2 V_0)\\ \tau^2 &\sim IG(\nu_0/2, \nu_0\tau_0^2/2)\end{aligned}$$

Sampling $(\psi|x_{0:n}, y^n)$

Let $y^n = (y_1, \ldots, y_n)$ and $x_{0:n} = (x_0, \ldots, x_n)'$.

It follows that

$$\begin{aligned}(\theta, \tau^2|x_{0:n}) &\sim N(\theta_1, \tau^2 V_1) IG(\nu_1/2, \nu_1\tau_1^2/2)\\ (\sigma^2|y^n, x^n) &\sim IG(n_1/2, n_1\sigma_1^2/2)\end{aligned}$$

where $\nu_1 = \nu_0 + n$, $n_1 = n_0 + n$

$$\begin{aligned}Z &= (G_{x_0}, \ldots, G_{x_{n-1}})'\\ V_1^{-1} &= V_0^{-1} + Z'Z\\ V_1^{-1}\theta_1 &= V_0^{-1}\theta_0 + Z'x_{1:n}\\ \nu_1\tau_1^2 &= \nu_0\tau_0^2 + (y - Z\theta_1)'(y - Z\theta_1) + (\theta_1 - \theta_0)'V_0^{-1}(\theta_1 - \theta_0)\\ n_1\sigma_1^2 &= n_0\sigma_0^2 + \sum_{t=1}^{n}(y_t - x_t^2/20)^2\end{aligned}$$

Sampling $x_1, \ldots, x_n$

Let $x_{-t} = (x_0, \ldots, x_{t-1}, x_{t+1}, \ldots, x_n)$, for $t = 1, \ldots, n-1$, $x_{-0} = x^n$, $x_{-n} = x_{0:(n-1)}$ and $y_0 = \emptyset$.

For $t = 0$

$$p(x_0|x_{-0}, y_0, \psi) \propto f_N(x_0; m_0, C_0) f_N(x_1; G'_{x_0}\theta, \tau^2)$$

For $t = 1, \ldots, n-1$

$$p(x_t|x_{-t}, y_t, \psi) \quad \propto \quad f_N(y_t; x_t^2/20, \sigma^2) f_N(x_t; G'_{x_{t-1}}\theta, \tau^2) f_N(x_{t+1}; G'_{x_t}\theta, \tau^2)$$

For $t = n$

$$p(x_n|x_{-n}, y_n, \psi) \propto f_N(y_n; x_n^2/20, \sigma^2) f_N(x_n; G'_{x_{n-1}}\theta, \tau^2)$$

Metropolis-Hastings algorithm

A simple random walk Metropolis algorithm with tuning variance v_x^2 would work as follows. For $t = 0, \ldots, n$

1. Current state: $x_t^{(j)}$
2. Sample x_t^* from $N(x_t^{(j)}, v_x^2)$
3. Compute the acceptance probability

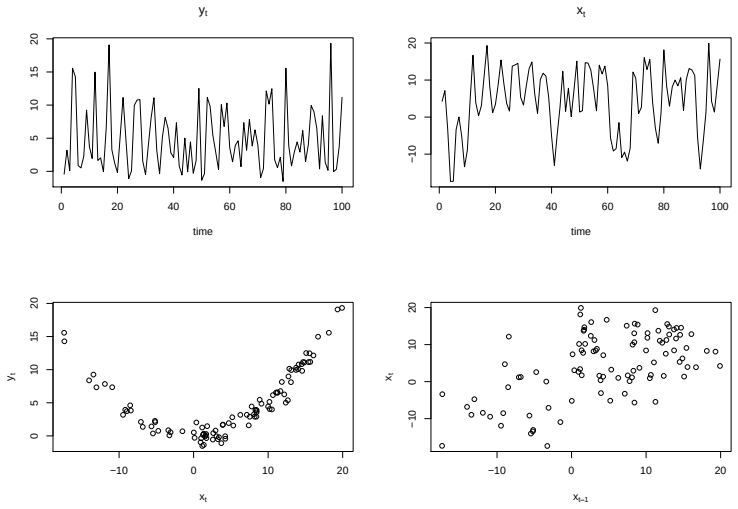
$$\alpha = \min\left\{1, \frac{p(x_t^*|x_{-t}, y_t, \psi)}{p(x_t^{(j)}|x_{-t}, y_t, \psi)}\right\}$$

4. New state:

$$x_t^{(j+1)} = \begin{cases} x_t^* & \text{w. p. } \alpha \\ x_t^{(j)} & \text{w. p. } 1-\alpha \end{cases}$$

Simulation set up

We simulated $n = 100$ observations based on $\theta = (0.5, 25, 8)'$, $\sigma^2 = 1$, $\tau^2 = 10$ and $x_0 = 0.1$.



Prior hyperparameters

- $x_0 \sim N(m_0, C_0)$

$$m_0 = 0.0 \quad \text{and} \quad C_0 = 10$$

- $\theta|\tau^2 \sim N(\theta_0, \tau^2 V_0)$

$$\theta_0 = (0.5, 25, 8)' \quad \text{and} \quad V_0 = \text{diag}(0.0025, 0.1, 0.04)$$

- $\tau^2 \sim IG(\nu_0/2, \nu_0\tau_0^2/2)$

$$\nu_0 = 6 \quad \text{and} \quad \tau_0^2 = 20/3$$

  such that $E(\tau^2) = \sqrt{V(\tau^2)} = 10$.

- $\sigma^2 \sim IG(n_0/2, n_0\sigma_0^2)$

$$n_0 = 6 \quad \text{and} \quad \sigma_0^2 = 2/3$$

  such that $E(\sigma^2) = \sqrt{V(\sigma^2)} = 1$.

MCMC setup

- Metropolis-Hastings tuning parameter

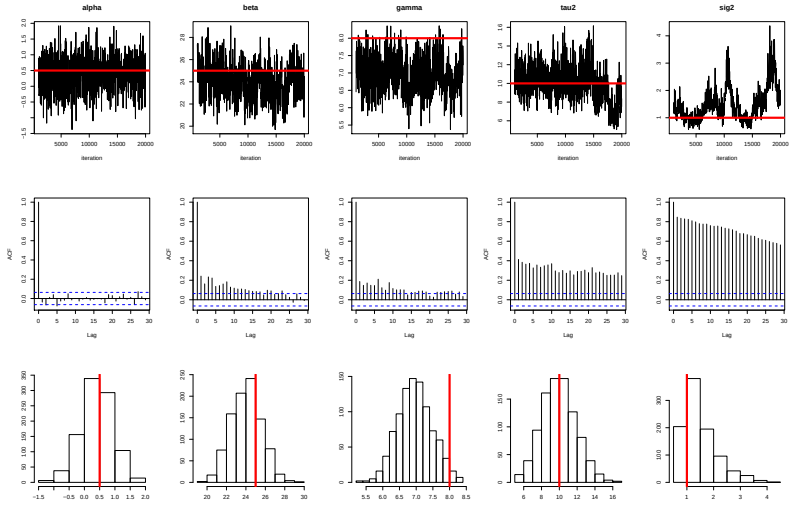
$$v_x^2 = (0.1)^2$$

- Burn-in period, step and MCMC sample size

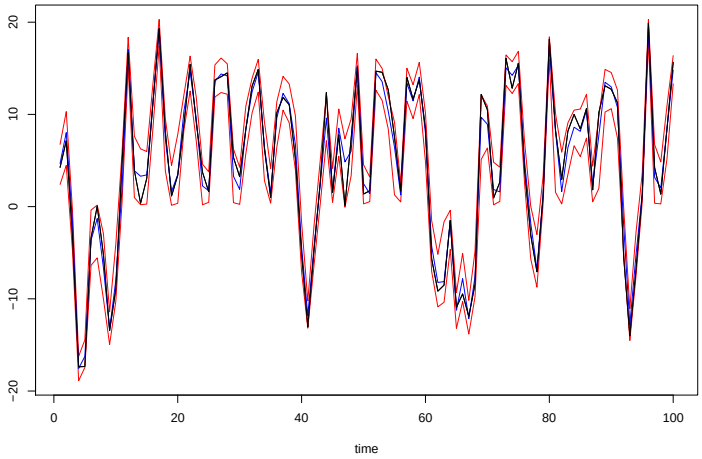
$$M_0 = 1,000 \quad L = 20 \quad M = 950 \;\Rightarrow\; 20,000 \text{ draws}$$

- Initial values
  - $\theta = (0.5, 25, 8)'$
  - $\tau^2 = 10$
  - $\sigma^2 = 1$
  - $x_{0:n} = x_{0:n}^{\text{true}}$

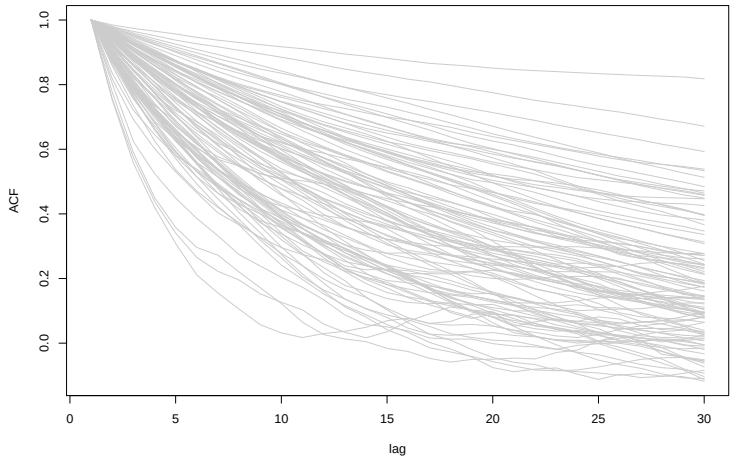
Parameters



States



States



LECTURE 7

STOCHASTIC VOLATILITY MODELS

Stochastic volatility model

The canonical stochastic volatility model (SV-AR(1), hereafter), is

$$\begin{aligned} y_t &= e^{h_t/2}\varepsilon_t \\ h_t &= \mu + \phi h_{t-1} + \tau\eta_t \end{aligned}$$

where ε_t and η_t are $N(0,1)$ shocks with $E(\varepsilon_t\eta_{t+h}) = 0$ for all h and $E(\varepsilon_t\varepsilon_{t+l}) = E(\eta_t\eta_{t+l}) = 0$ for all $l \neq 0$.

τ^2: volatility of the log-volatility.

$|\phi| < 1$ then h_t is a stationary process.

Let $y^n = (y_1, \ldots, y_n)'$, $h^n = (h_1, \ldots, h_n)'$ and $h_{a:b} = (h_a, \ldots, h_b)'$.

Prior information

Uncertainty about the initial log volatility is $h_0 \sim N(m_0, C_0)$.

Let $\theta = (\mu, \phi)'$, then the prior distribution of (θ, τ^2) is normal-inverse gamma, i.e. $(\theta, \tau^2) \sim NIG(\theta_0, V_0, \nu_0, s_0^2)$:

$$\begin{aligned} \theta|\tau^2 &\sim N(\theta_0, \tau^2 V_0) \\ \tau^2 &\sim IG(\nu_0/2, \nu_0 s_0^2/2) \end{aligned}$$

For example, if $\nu_0 = 10$ and $s_0^2 = 0.018$ then

$$\begin{aligned} E(\tau^2) &= \frac{\nu_0 s_0^2/2}{\nu_0/2 - 1} = 0.0225 \\ Var(\tau^2) &= \frac{(\nu_0 s_0^2/2)^2}{(\nu_0/2-1)^2(\nu_0/2-2)} = (0.013)^2 \end{aligned}$$

Hyperparameters: m_0, C_0, θ_0, V_0, ν_0 and s_0^2.

Posterior inference

The SV-AR(1) is a dynamic model and posterior inference via MCMC for the the latent log-volatility states h_t can be performed in at least two ways.

Let $h_{-t} = (h_{0:(t-1)}, h_{(t+1):n})$, for $t = 1, \ldots, n-1$ and $h_{-n} = h_{1:(n-1)}$.

- Individual moves for h_t

  - $(\theta, \tau^2 | h^n, y^n)$
  - $(h_t | h_{-t}, \theta, \tau^2, y^n)$, for $t = 1, \ldots, n$

- Block move for h^n
  - $(\theta, \tau^2 | h^n, y^n)$
  - $(h^n | \theta, \tau^2, y^n)$

Sampling $(\theta, \tau^2 | h^n, y^n)$

Conditional on $h_{0:n}$, the posterior distribution of (θ, τ^2) is also normal-inverse gamma:

$$(\theta, \tau^2 | y^n, h_{0:n}) \sim NIG(\theta_1, V_1, \nu_1, s_1^2)$$

where $X = (1_n, h_{0:(n-1)})$, $\nu_1 = \nu_0 + n$

$$\begin{aligned} V_1^{-1} &= V_0^{-1} + X'X \\ V_1^{-1}\theta_1 &= V_0^{-1}\theta_0 + X'h_{1:n} \\ \nu_1 s_1^2 &= \nu_0 s_0^2 + (y - X\theta_1)'(y - X\theta_1) + (\theta_1 - \theta_0)' V_0^{-1} (\theta_1 - \theta_0) \end{aligned}$$

Sampling $(h_0 | \theta, \tau^2, h_1)$

Combining

$$h_0 \sim N(m_0, C_0)$$

and

$$h_1 | h_0 \sim N(\mu + \phi h_0, \tau^2)$$

leads to (by Bayes' theorem)

$$h_0 | h_1 \sim N(m_1, C_1)$$

where

$$\begin{aligned} C_1^{-1} m_1 &= C_0^{-1} m_0 + \phi \tau^{-2} (h_1 - \mu) \\ C_1^{-1} &= C_0^{-1} + \phi^2 \tau^{-2} \end{aligned}$$

Conditional prior distribution of h_t

Given h_{t-1}, θ and τ^2, it can be shown that, for $t = 1, \ldots, n-1$,

$$\begin{pmatrix} h_t \\ h_{t+1} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mu + \phi h_{t-1} \\ (1+\phi)\mu + \phi^2 h_{t-1} \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & \phi \\ \phi & (1+\phi^2) \end{pmatrix} \right\}$$

so $E(h_t | h_{t-1}, h_{t+1}, \theta, \tau^2)$ and $V(h_t | h_{t-1}, h_{t+1}, \theta, \tau^2)$ are

$$\begin{aligned} \mu_t &= \left(\frac{1-\phi}{1+\phi^2} \right) \mu + \left(\frac{\phi}{1+\phi^2} \right) (h_{t-1} + h_{t+1}) \\ \nu^2 &= \tau^2 (1+\phi^2)^{-1} \end{aligned}$$

respectively. Therefore,

$$\begin{aligned} (h_t | h_{t-1}, h_{t+1}, \theta, \tau^2) &\sim N(\mu_t, \nu^2) \qquad t = 1, \ldots, n-1 \\ (h_n | h_{n-1}, \theta, \tau^2) &\sim N(\mu_n, \tau^2) \end{aligned}$$

where $\mu_n = \mu + \phi h_{n-1}$.

Sampling h_t via random walk Metropolis

Let $\nu_t^2 = \nu^2$ for $t = 1, \ldots, n-1$ and $\nu_n^2 = \tau^2$, then

$$p(h_t|h_{-t}, y^n, \theta, \tau^2) \quad = \quad f_N(h_t; \mu_t, \nu_t^2) f_N(y_t; 0, e^{h_t})$$

for $t = 1, \ldots, n$.

A simple random walk Metropolis algorithm with tuning variance v_h^2 would work as follows:

For $t = 1, \ldots, n$

1. Current state: $h_t^{(j)}$
2. Sample h_t^* from $N(h_t^{(j)}, v_h^2)$
3. Compute the acceptance probability

$$\alpha = \min \left\{ 1, \frac{f_N(h_t^*; \mu_t, \nu_t^2) f_N(y_t; 0, e^{h_t^*})}{f_N(h_t^{(j)}; \mu_t, \nu_t^2) f_N(y_t; 0, e^{h_t^{(j)}})} \right\}$$

4. New state:

$$h_t^{(j+1)} = \begin{cases} h_t^* & \text{w. p. } \alpha \\ h_t^{(j)} & \text{w. p. } 1 - \alpha \end{cases}$$

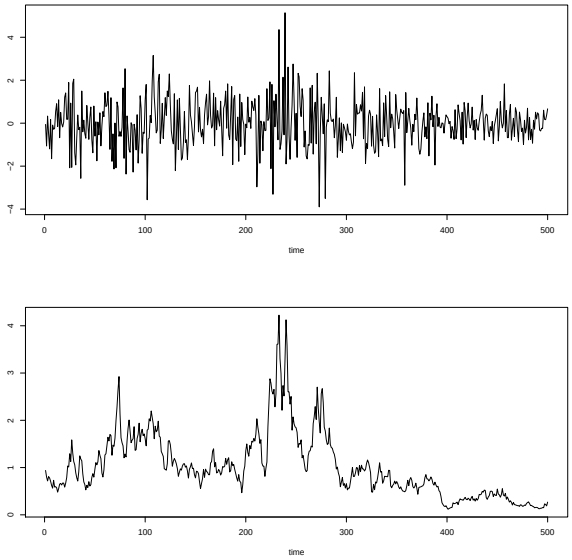
Example i. Simulated data

- Simulation setup
  - $n = 500$
  - $h_0 = 0.0$
  - $\mu = -0.00645$
  - $\phi = 0.99$
  - $\tau^2 = 0.15^2$

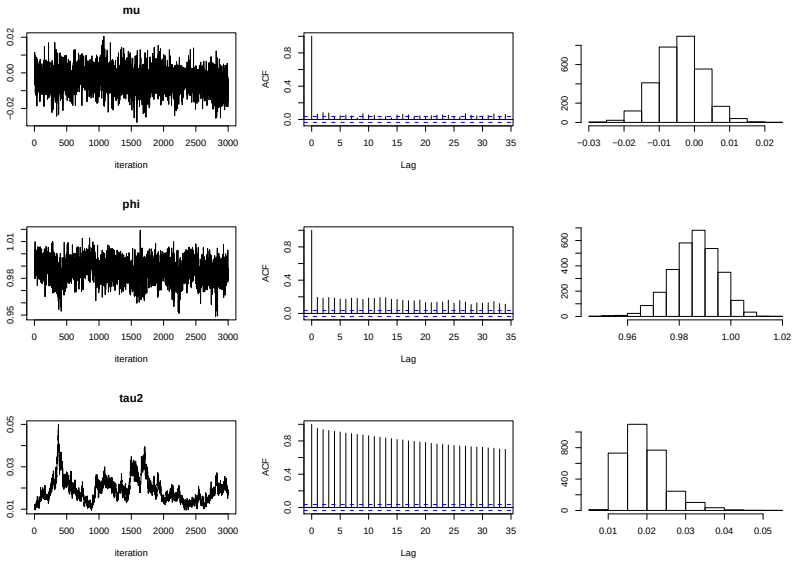
- Prior distribution
  - $\mu \sim N(0, 100)$
  - $\phi \sim N(0, 100)$
  - $\tau^2 \sim IG(10/2, 0.28125/2)$
  - $h_0 \sim N(0, 100)$

- MCMC setup
  - $M_0 = 1,000$
  - $M = 1,000$

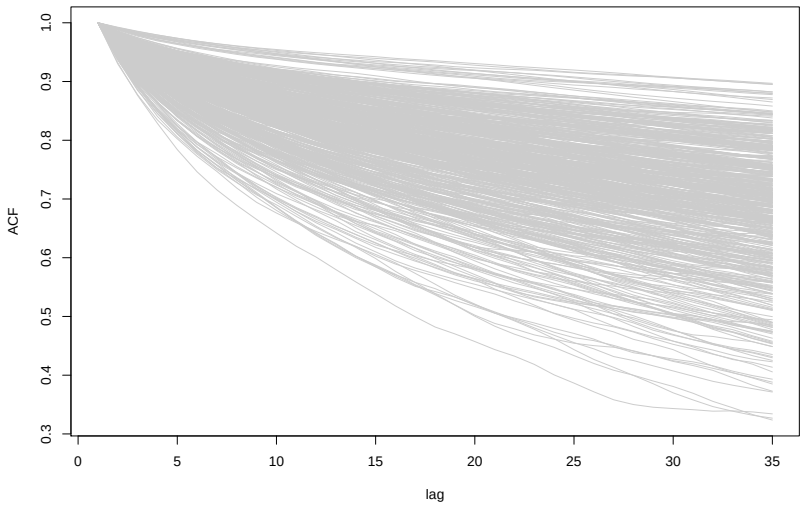
Time series of y_t and $\exp\{h_t\}$



Parameters

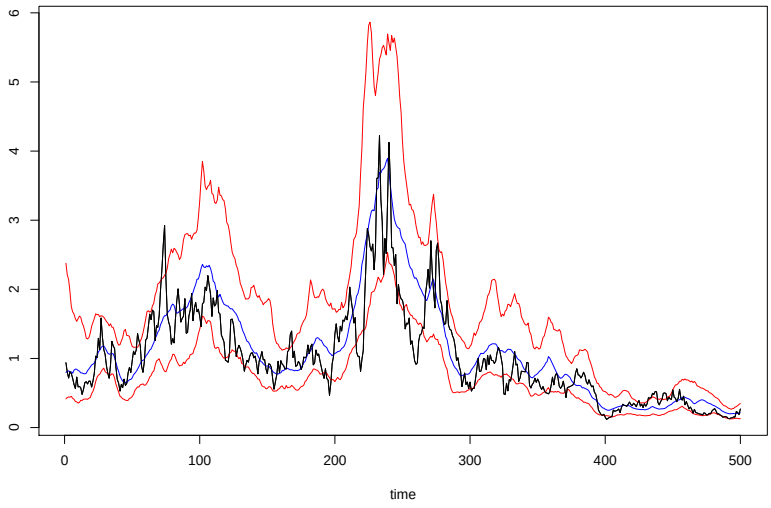


Autocorrelation of h_t



Volatilities

Tuning parameter: $v_h^2 = 0.01$



Sampling h_t via independent Metropolis-Hastings

The full conditional distribution of h_t is given by

$$\begin{aligned} p(h_t|h_{-t}, y^n, \theta, \tau^2) &= p(h_t|h_{t-1}, h_{t+1}, \theta, \tau^2)p(y_t|h_t) \\ &= f_N(h_t; \mu_t, \nu^2) f_N(y_t; 0, e^{h_t}). \end{aligned}$$

Kim, Shephard and Chib (1998) explored the fact that

$$\log p(y_t|h_t) = \text{const} - \frac{1}{2}h_t - \frac{y_t^2}{2}\exp(-h_t)$$

and that a Taylor expansion of $\exp(-h_t)$ around μ_t leads to

$$\begin{aligned} \log p(y_t|h_t) &\approx \text{const} - \frac{1}{2}h_t - \frac{y_t^2}{2}\left(e^{-\mu_t} - (h_t - \mu_t)e^{-\mu_t}\right) \\ g(h_t) &= \exp\left\{-\frac{1}{2}h_t(1 - y_t^2 e^{-\mu_t})\right\} \end{aligned}$$

Proposal distribution

Let $\nu_t^2 = \nu^2$ for $t = 1, \ldots, n-1$ and $\nu_n^2 = \tau^2$.

Then, by combining $f_N(h_t; \mu_t, \nu_t^2)$ and $g(h_t)$, for $t = 1, \ldots, n$, leads to the following proposal distribution:

$$q(h_t|h_{-t}, y^n, \theta, \tau^2) \equiv N\left(h_t; \tilde{\mu}_t, \nu_t^2\right)$$

where $\tilde{\mu}_t = \mu_t + 0.5\nu_t^2(y_t^2 e^{-\mu_t} - 1)$.

Metropolis-Hastings algorithm

For $t = 1, \ldots, n$

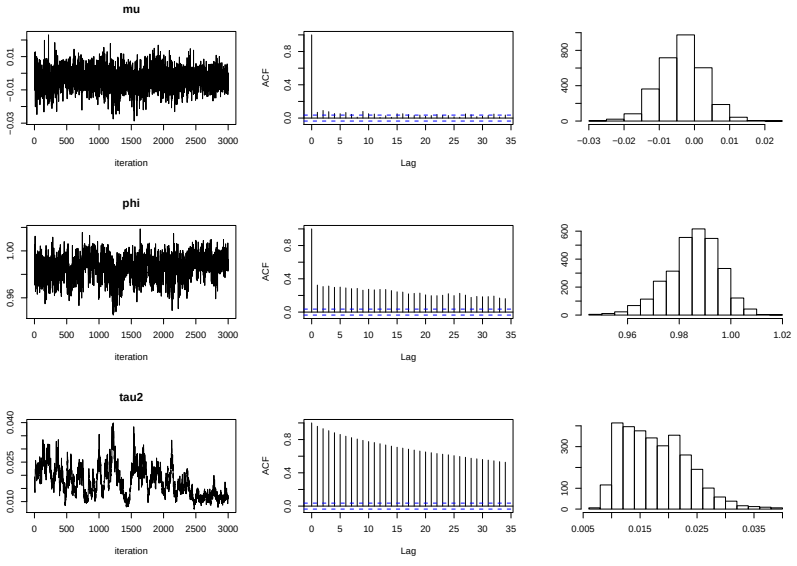
1. Current state: $h_t^{(j)}$
2. Sample h_t^* from $N(\tilde{\mu}_t, \nu_t^2)$
3. Compute the acceptance probability

$$\alpha = \min\left\{1, \frac{f_N(h_t^*; \mu_t, \nu_t^2) f_N(y_t; 0, e^{h_t^*})}{f_N(h_t^{(j)}; \mu_t, \nu_t^2) f_N(y_t; 0, e^{h_t^{(j)}})} \times \frac{f_N(h_t^{(j)}; \tilde{\mu}_t, \nu_t^2)}{f_N(h_t^*; \tilde{\mu}_t, \nu_t^2)}\right\}$$

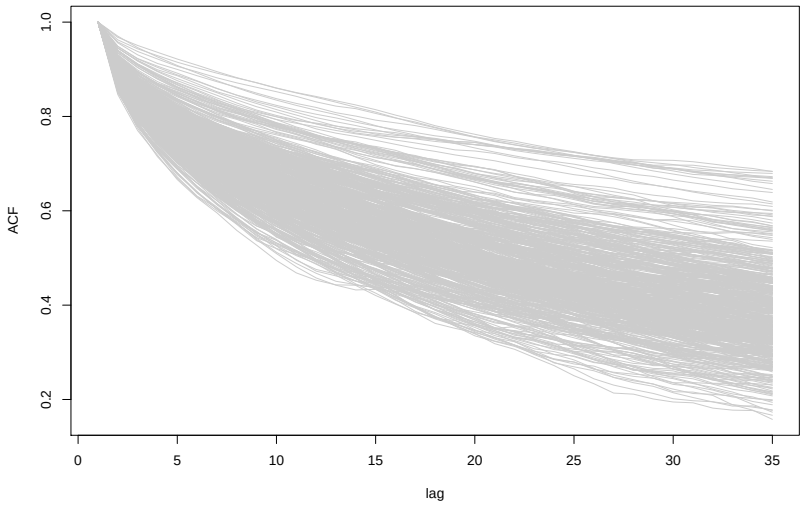
4. New state:

$$h_t^{(j+1)} = \begin{cases} h_t^* & \text{w. p. } \alpha \\ h_t^{(j)} & \text{w. p. } 1 - \alpha \end{cases}$$

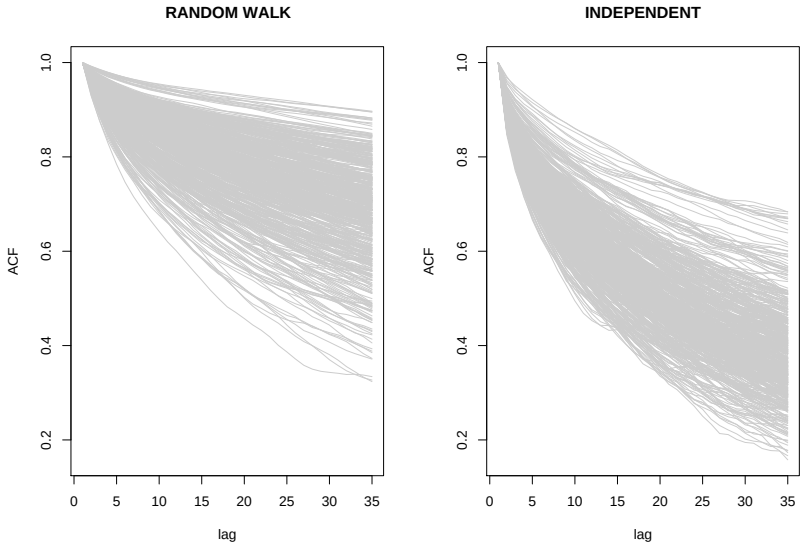
Parameters



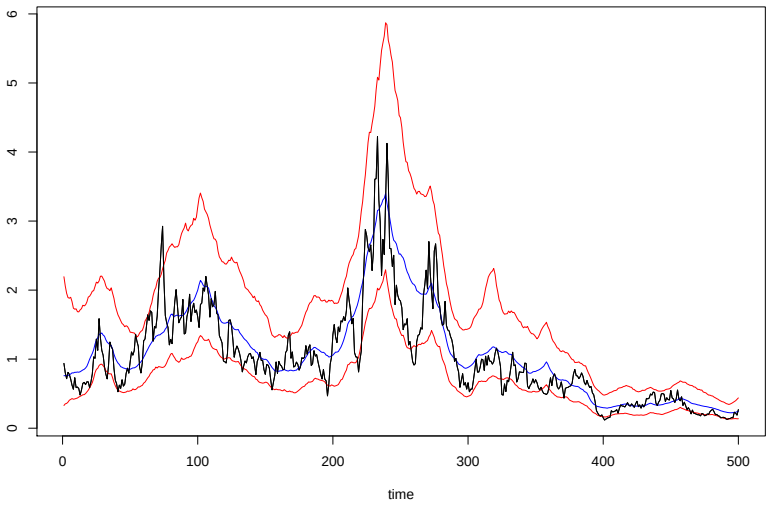
Autocorrelation of h_t



Autocorrelations of h_t for both schemes



Volatilities



Sampling h^n - normal approximation and FFBS

Let $y_t^* = \log y_t^2$ and $\epsilon_t = \log \varepsilon_t^2$.

The SV-AR(1) is a DLM with nonnormal observational errors, i.e.

$$\begin{aligned} y_t^* &= h_t + \epsilon_t \\ h_t &= \mu + \phi h_{t-1} + \tau \eta_t \end{aligned}$$

where $\eta_t \sim N(0,1)$.

The distribution of ϵ_t is $\log \chi_1^2$, where

$$\begin{aligned} E(\epsilon_t) &= -1.27 \\ V(\epsilon_t) &= \frac{\pi^2}{2} = 4.935 \end{aligned}$$

Normal approximation

Let ϵ_t be approximated by $N(\alpha, \sigma^2)$, $z_t = y_t^* - \alpha$, $\alpha = -1.27$ and $\sigma^2 = \pi^2/2$.

Then

$$\begin{aligned} z_t &= h_t + \sigma v_t \\ h_t &= \mu + \phi h_{t-1} + \tau \eta_t \end{aligned}$$

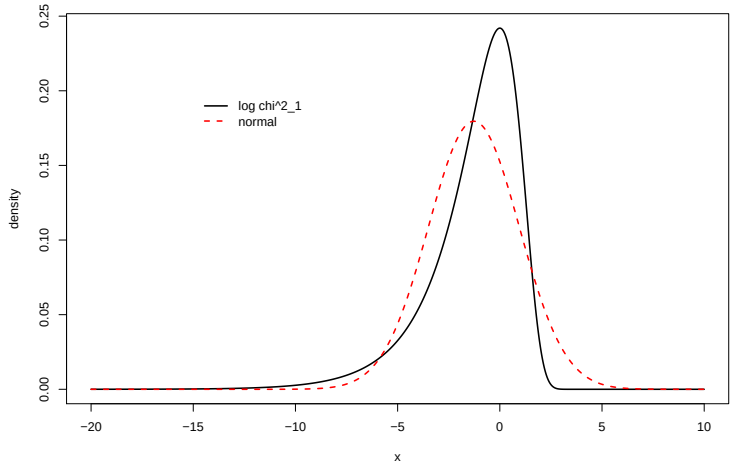
is a simple DLM where v_t and η_t are $N(0,1)$.

Sampling from

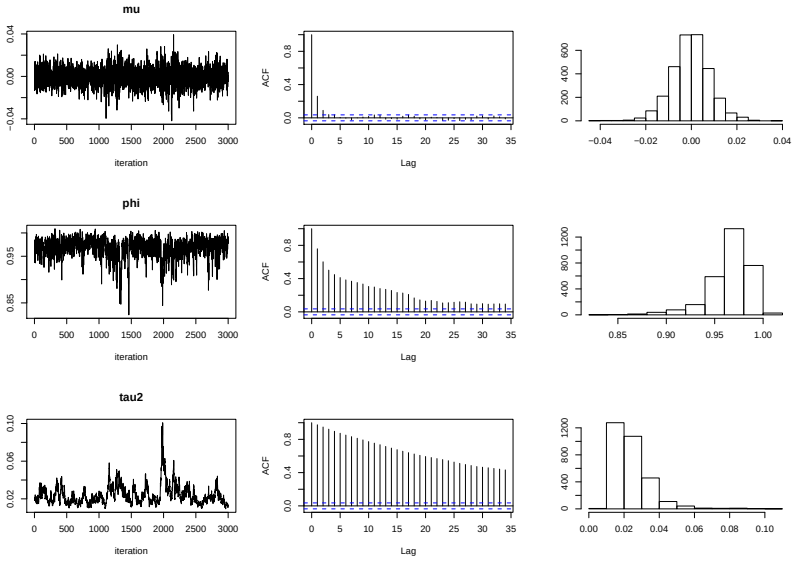
$$p(h^n | \theta, \tau^2, \sigma^2, z^n)$$

can be performed by the FFBS algorithm.

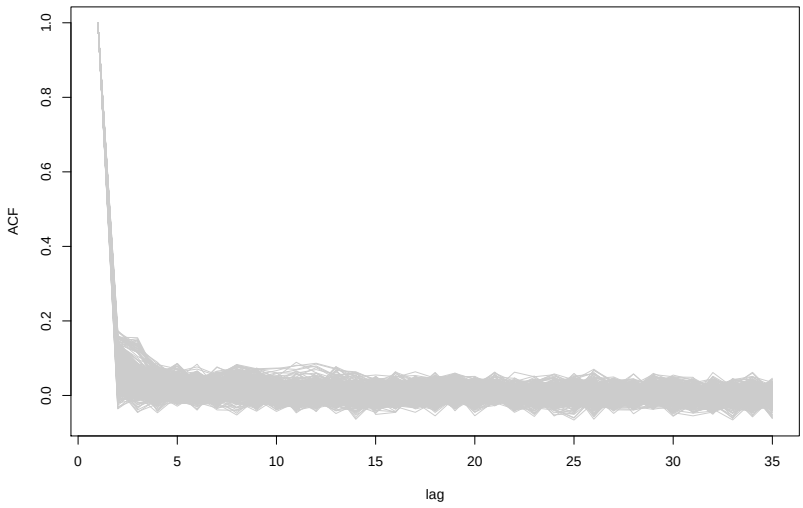
$\log \chi_1^2$ **and** $N(-1.27, \pi^2/2)$



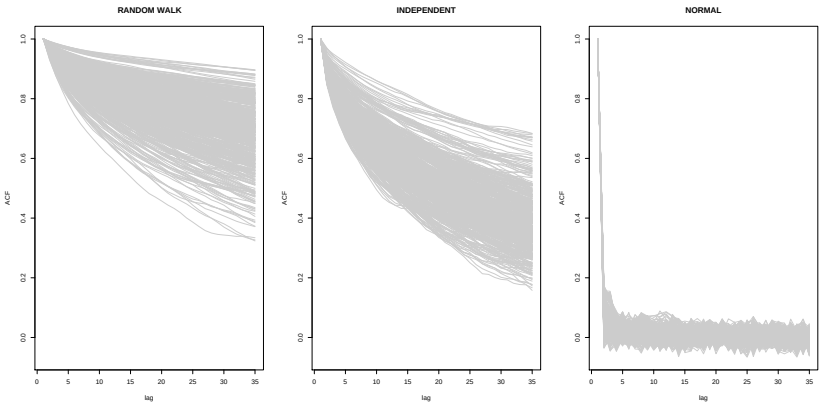
Parameters



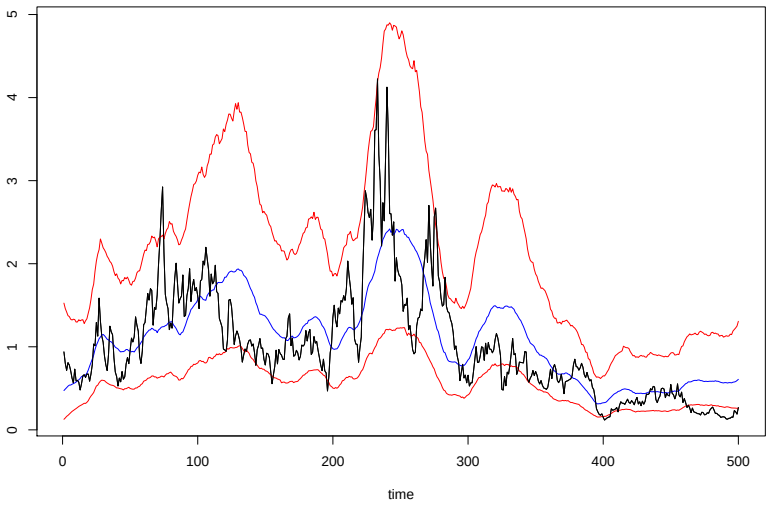
Autocorrelation of h_t



Autocorrelations of h_t for the three schemes



Volatilities



Sampling h^n - mixtures of normals and FFBS

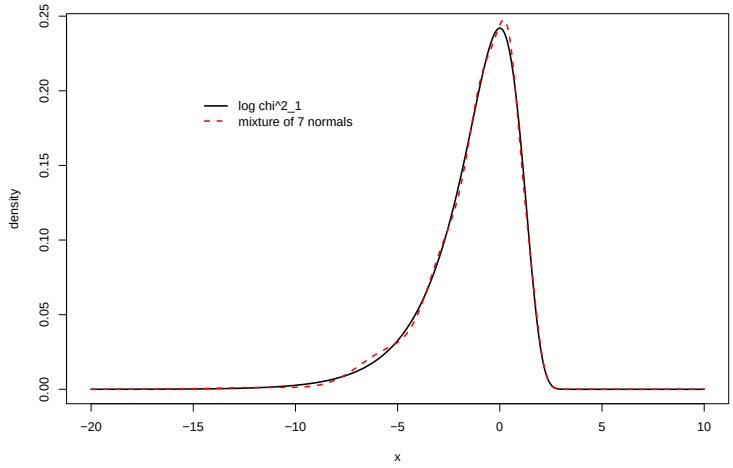
The $\log \chi_1^2$ distribution can be approximated by

$$\sum_{i=1}^{7} \pi_i N(\mu_i, \omega_i^2)$$

where

i	π_i	μ_i	ω_i^2
1	0.00730	-11.40039	5.79596
2	0.10556	-5.24321	2.61369
3	0.00002	-9.83726	5.17950
4	0.04395	1.50746	0.16735
5	0.34001	-0.65098	0.64009
6	0.24566	0.52478	0.34023
7	0.25750	-2.35859	1.26261

$\log \chi_1^2$ **and** $\sum_{i=1}^{7} \pi_i N(\mu_i, \omega_i^2)$



Mixture of normals

Using an argument from the Bayesian analysis of mixture of normal, let $z_1, \ldots, z_n$ be unobservable (latent) indicator variables such that $z_t \in \{1, \ldots, 7\}$ and $Pr(z_t = i) = \pi_i$, for $i = 1, \ldots, 7$.

Therefore, conditional on the z's, y_t is transformed into $\log y_t^2$,

$$\begin{aligned} \log y_t^2 &= h_t + \log \varepsilon_t^2 \\ h_t &= \mu + \phi h_{t-1} + \tau_\eta \eta_t \end{aligned}$$

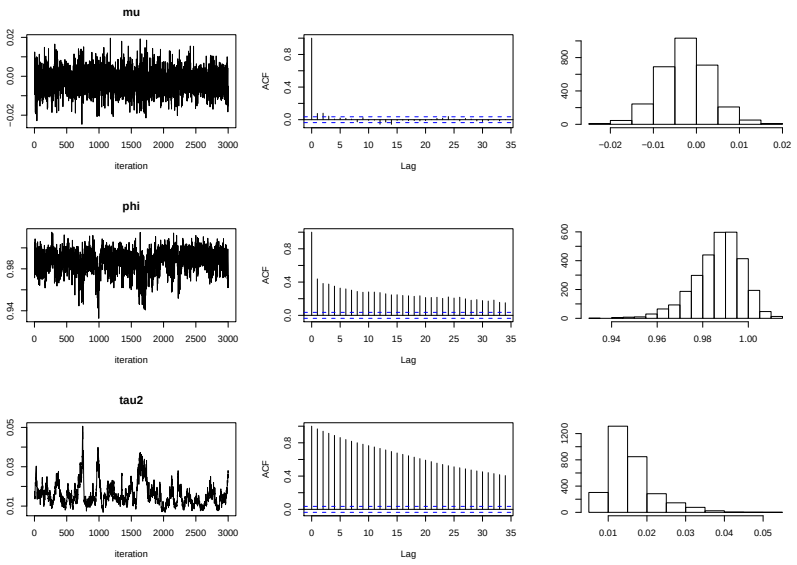
which can be rewritten as a normal DLM:

$$\begin{aligned} \log y_t^2 &= h_t + v_t \qquad\qquad v_t \sim N(\mu_{z_t}, \omega_{z_t}^2) \\ h_t &= \mu + \phi h_{t-1} + w_t \qquad w_t \sim N(0, \tau_\eta^2) \end{aligned}$$

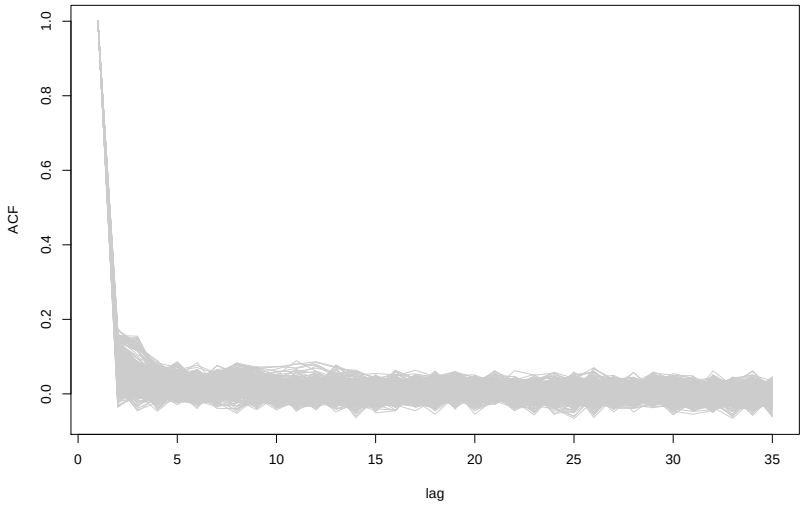
where μ_{z_t} and $\omega_{z_t}^2$ are provided in the previous table.

Then h^n is jointly sampled by using the the FFBS algorithm.

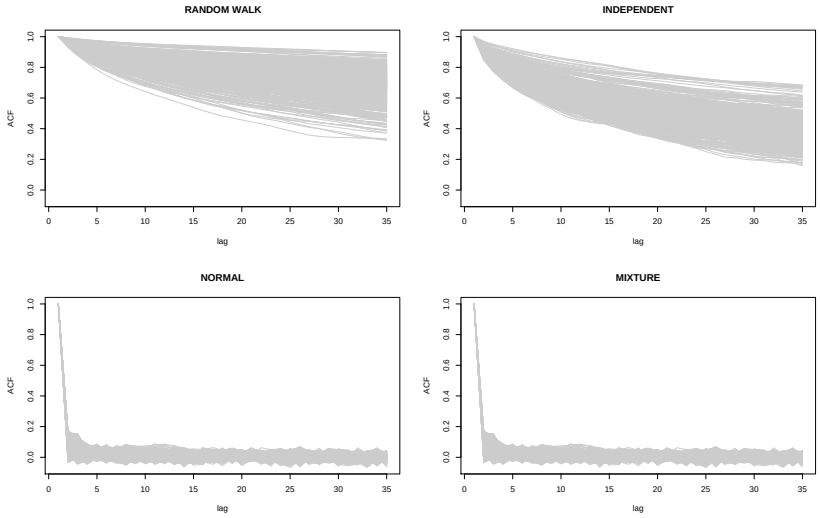
Parameters



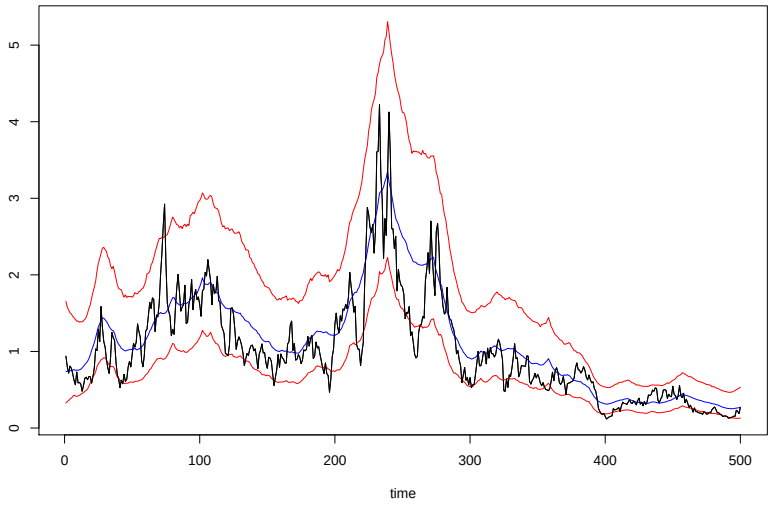
Autocorrelation of h_t



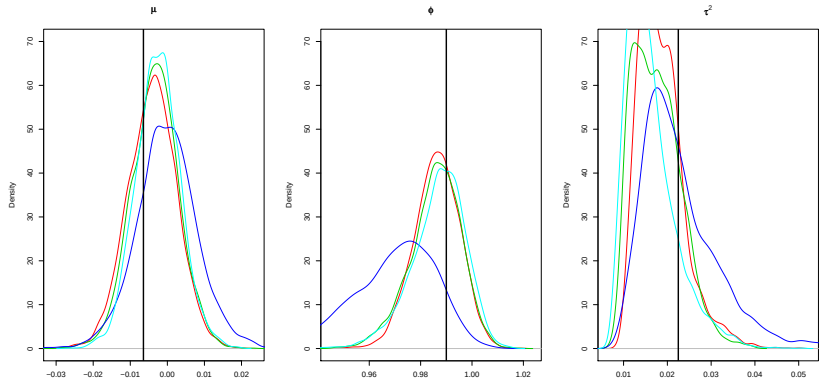
Autocorrelations of h_t for the four schemes



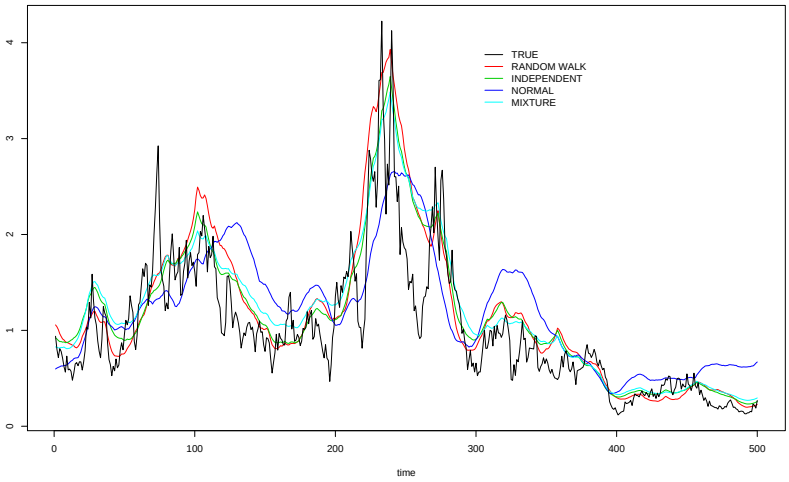
Volatilities



Comparing the four schemes: parameters



Comparing the four schemes: volatilities



Example ii. LSTAR-SV models (Lopes and Salazar, 2006)

Lopes and Salazar (2006) adapt the LSTAR structure to model time-varying variances, SV-LSTAR(k):

$$\begin{aligned} y_t \mid h_t &\sim N(0, e^{h_t}) \\ h_t &\sim N(x_t'\theta_1 + \pi(\gamma, c, h_{t-d})x_t'\theta_2, \sigma^2) \end{aligned}$$

with $x_t' = (1, h_{t-1}, \ldots, h_{t-k})$ and $\theta_i = (\theta_{0i}, \theta_{1i}, \ldots, \theta_{ki})$, for $i = 1, 2$. The logistic transition function is

$$\pi(\gamma, c, h_{t-d}) = 1/(1 + e^{-\gamma(h_{t-d}-c)}).$$

Particular case: $k = d = 1$

$$E(h_t|h_{t-1}, c, \theta) = \left(\theta_{10} + \frac{\theta_{20}}{1 + e^{-\gamma(h_{t-1}-c)}}\right) + \left(\theta_{11} + \frac{\theta_{21}}{1 + e^{-\gamma(h_{t-1}-c)}}\right) h_{t-1}$$

S&P500 stock index

MCMC: Conditional on h^n, sampling $\theta_1, \theta_2, c, \gamma$ and σ^2 are relatively easy. To sample the components of h^n, we use the single parameter move introduced by Jacquier, Polson and Rossi (1994).

North American Standard and Poors 500 index, daily observed from January 7th, 1986 to December 31st, 1997. A total of 3127 observations. We entertained six models for the stochastic volatility:

- $\mathcal{M}_1 : AR(1)$
- $\mathcal{M}_2 : AR(2)$
- $\mathcal{M}_3 : LSTAR(1)$ with $d = 1$
- $\mathcal{M}_4 : LSTAR(1)$ with $d = 2$
- $\mathcal{M}_5 : LSTAR(2)$ with $d = 1$
- $\mathcal{M}_6 : LSTAR(2)$ with $d = 2$

Model comparison

Models	AIC	BIC	DIC
$\mathcal{M}_1 : AR(1)$	12795	31697	7223.1
$\mathcal{M}_2 : AR(2)$	12624	31532	7149.2
$\mathcal{M}_3 : LSTAR(1, d = 1)$	**12240**	**31165**	**7101.1**
$\mathcal{M}_4 : LSTAR(1, d = 2)$	12244	31170	7150.3
$\mathcal{M}_5 : LSTAR(2, d = 1)$	12569	31507	7102.4
$\mathcal{M}_6 : LSTAR(2, d = 2)$	12732	31670	7159.4

AIC: Akaike's information criteria, BIC: Schwarz's information criteria, and DIC: Deviance information criteria.

Posterior inference

Example iii. Factor SV models (Lopes and Carvalho, 2007)

Factor stochastic volatility models appear in Pitt and Shephard (1999), Aguilar and West (2000) and Lopes and Migon (2002) and Lopes and Carvalho (2007), to name just a few. The model is

$$\begin{aligned} (y_t|f_t, \beta_t, \Sigma_t) &\sim N(\beta_t f_t; \Sigma_t) \\ (f_t|H_t) &\sim N(0; H_t) \end{aligned}$$

with $\Sigma_t = \text{diag}(\sigma^2_{1t}, \cdots, \sigma^2_{pt})$ and $H_t = \text{diag}(h_{1t}, \ldots, h_{qt})$. Let $\eta_t = \log(\Sigma_t)$ and $\lambda_t = \log(H_t)$. Then

$$\begin{aligned} (\eta_t|\eta_{t-1}, \mu, \rho, V) &\sim N(\mu + \rho\eta_{t-1}, V) \\ (\lambda_t|\lambda_{t-1}, \alpha, \phi, U) &\sim N(\alpha + \phi\lambda_{t-1}, U) \end{aligned}$$

Pitt and Shephard (1999): diagonal V and U and $\beta_t = \beta$.
Aguilar and West (2000): nondiagonal V and U and $\beta_t = \beta$.
Lopes and Migon (2002): diagonal V and U and β_t.

Lopes and Carvalho (2007): diagonal V and U, β_t and Markov switching for λ_t.

Parameter	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
θ_{01}	-0.060	-0.066	0.292	-0.154	-4.842	-6.081
	(0.184)	(0.241)	(0.579)	(0.126)	(0.802)	(1.282)
θ_{11}	0.904	0.184	0.306	0.572	-0.713	-0.940
	(0.185)	(0.242)	(0.263)	(0.135)	(0.306)	(0.699)
θ_{21}	-	0.715	-	-	-1.018	-1.099
		(0.248)			(0.118)	(0.336)
θ_{02}	-	-	-0.685	0.133	4.783	6.036
			(0.593)	(0.092)	(0.801)	(1.283)
θ_{12}	-	-	0.794	0.237	0.913	1.091
			(0.257)	(0.086)	(0.314)	(0.706)
θ_{22}	-	-	-	-	1.748	1.892
					(0.114)	(0.356)
γ	-	-	118.18	163.54	132.60	189.51
			(16.924)	(23.912)	(10.147)	(0.000)
c	-	-	-1.589	0.022	-2.060	-2.125
			(0.022)	(0.280)	(0.046)	(0.000)
τ^2	0.135	0.234	0.316	0.552	0.214	0.166
	(0.020)	(0.044)	(0.066)	(0.218)	(0.035)	(0.026)

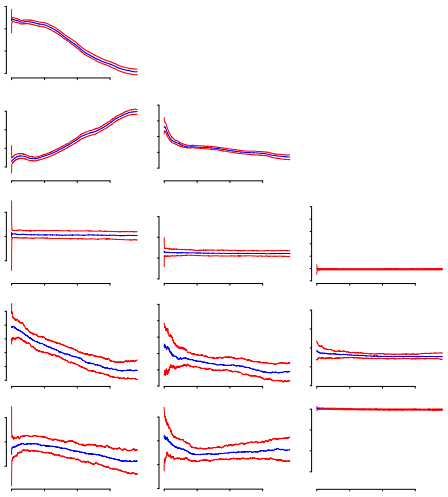
Daily exchange rate returns

To illustrate the time-varying loadings extension we analyze the returns on weekday closing spot prices for six currencies relative to the US dollar (as in Aguilar and West, 2000):

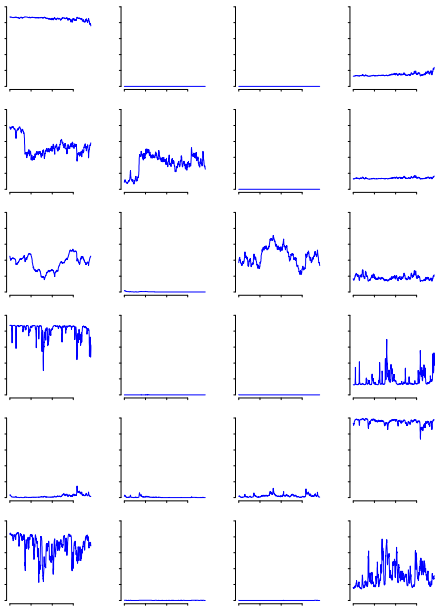
- German Mark (DEM)
- British Pound (GBP)
- Japanese Yen (JPY)
- French Franc (FRF)
- Canadian Dollar (CAD)
- Spanish Peseta (ESP)

To keep the analysis comparable with Aguilar and West (2000) we only use the rst 1000 observations ranging from 1/1/1992 to 10/31/1995.

Time varying factor loadings



Time varying variance decomposition



Importance of time-varying loadings

An interesting observation that highlights the importance of time-varying loadings in the context of this example is the change in the explanatory power of factor 1, the European factor on the British Pound.

The nal months of 1992 marks the withdrawal of Great Britain from the European Union exchange-rate agreement (ERM), fact that is captured in our analysis by changes in the British loading in factor 1

	FSV			FSV+MSSV			
	μ	ρ	v	μ		ρ	v
IBOVESPA	-0.202	0.980	0.040	-0.284	-	0.971	0.047
MEXBOL	-0.440	0.959	0.065	-0.434	-	0.957	0.051
MERVAL	-0.409	0.959	0.083	-0.508	-	0.947	0.068
IPSA	-0.600	0.947	0.058	-0.765	-	0.932	0.071
	α	ϕ	u	α_1	α_2	ϕ	u
Factor	-0.305	**0.971**	0.067	-0.951	-0.588	**0.912**	0.090
$E(\lambda_t)$	-10.517			-10.807	-6.682		

and emphasized by the changes in the percentage of variation of the British Pound explained by factors 1 and 2.

If temporal changes on the factor loadings were not allowed, the only way the model could capture this change in Great Britains monetary policy would be by a shock on the idiosyncratic variation of the Pound, reducing, inturn, thepredictiveabilityof thelatentfactorstructure.

Latin american stock returns

We now illustrate the FSV+MSSV generalization in a extended version of the dataset in Lopes and Migon (2002).

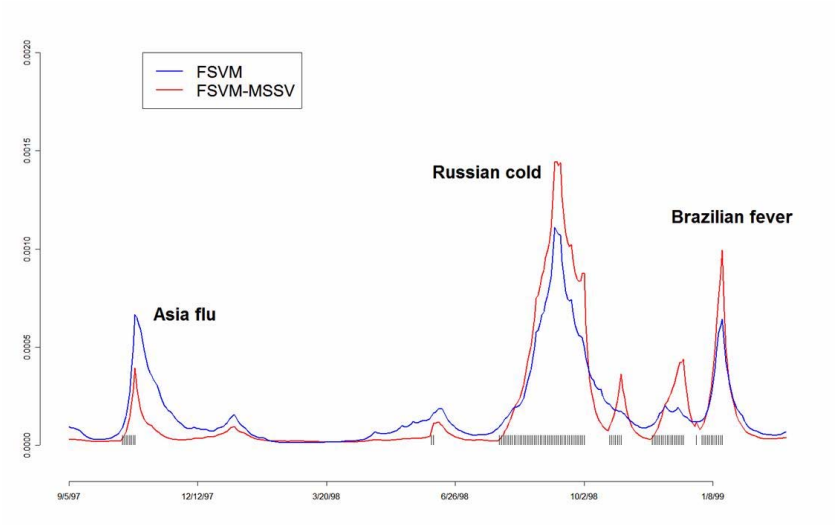
Returns on week-day closing spot prices in Latin American:

- Brazilian Indice Bovespa (IBOVESPA)
- Mexican Indice de Precios y Cotaziones (MEXBOL)
- Argentinean Indice Merval (MERVAL)
- Chilean Indice de Precios Selectivos de Acciones (IPSA).

The series are observed daily from January, 3rd 1994 to May, 26th 2005 (2974 observations), which includes several international currency crises. These crises have directly impacted on Latin American markets, generating higher levels of uncertainty and consequently higher levels of volatility.

Comparing FSV and FSV+MSSV models

Common factor volatilities



Under the model with $k = 2$ the persistency parameter ϕ is likely to be smaller in line with conclusions drawn in Carvalho and Lopes (2007).

The model with $k = 2$ estimates two unconditional means to the log-volatility process that correspond to times of high and low risk in the market. More specifically, the posterior mean of the unconditional standard deviation of the common factor in the FSV model is roughly the same as the one obtained for the low volatility regime in the FSV+MSSV, however the factor is on a high volatility state around 6% of the time, in which the unconditional standard deviation is about eight times higher.

This allows the volatilities to react "faster" once a regime switch is identified, which is highlighted by the previous figure that compares FSV and FSV+MSSV common factor's volatilities.

LECTURE 8

SEQUENTIAL MONTE CARLO METHODS

Nonnormal/nonlinear dynamic models

Most nonnormal and nonlinear dynamic models are defined by

- Observation equation
$$p(y_t|x_t, \psi)$$
- System or evolution equation
$$p(x_t|x_{t-1}, \psi)$$
- Initial distribution
$$p(x_0|\psi)$$

The fixed parameters that drive the state space model, ψ, is kept known and omitted for now.

Evolution and updating

Let the information regarding x_{t-1} at time $t-1$ be summarized by

$$p(x_{t-1}|y^{t-1})$$

Then **Evolution** and **updating** are represented by

$$p(x_t|y^{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y^{t-1})dx_{t-1}$$

and

$$p(x_t|y^t) \propto p(y_t|x_t)p(x_t|y^{t-1})$$

respectively.

These densities are usually **unavailable in closed form**.

The Bayesian boostrap filter

Gordon, Salmond and Smith's (1993) seminal paper uses SIR ideas to obtain draws from $p(x_t|y^t)$ based on draws from $p(x_{t-1}|y^{t-1})$.

SIR: the goal is to draw from $p(x)$ based on draws from $q(x)$.

1. Draw $x_1^*, \ldots, x_N^*$ from q
2. Compute (unnormalized) weights $\omega_i = p(x_i^*)/q(x_i^*)$

3. Draw $x_1, \ldots, x_M$ from $\{x_1^*, \ldots, x_N^*\}$ with weights $\{\omega_1, \ldots, \omega_N\}$

Sampling from the prior: If

$$p(x) \propto \pi(x) l(x)$$

where $\pi(x)$ and $l(x)$ are prior and likelihood, respectively, then a natural (but not necessarily good, actually usually bad!) choice is

$$q(x) = \pi(x).$$

Under this choice, unnormalized weights are likelihoods, i.e.

$$\omega(x) \propto l(x).$$

Example i. Revisiting the 1st order DLM

For illustration, let us reconsider the local level model where closed form solutions are promptly available. The model is

$$\begin{aligned} y_t|x_t &\sim N(x_t, \sigma^2) \\ x_t|x_{t-1} &\sim N(x_{t-1}, \tau^2) \end{aligned}$$

- Posterior at $t = 0$: $(x_0|y_0) \sim N(m_0, C_0)$
- Prior at $t = 1$: $(x_1|y_0) \sim N(m_0, C_0 + \tau^2)$
- Likelihood at time t: $l(x_1; y_1) \propto f_N(x_1; y_1, \sigma^2)$
- Posterior at time t: $(x_1|y_1) \sim N(m_1, C_1)$

where $A_1 = (C_0 + \tau^2)/(C_0 + \tau^2 + \sigma^2)$, $m_1 = (1 - A_1)m_0 + A_1 y_1$ and $C_1 = A_1 \sigma^2$.

Example i. One step update

Let $\{(x_0, \omega_0)^{(i)}\}_{i=1}^N$ summarizes $p(x_0|y_0)$. For example,

$$E(g(x_0)|y_0) \approx \frac{1}{N} \sum_{i=1}^{N} \omega_0^{(i)} g(x_0^{(i)}).$$

Then, $\{(x_1, \omega_0)^{(i)}\}_{i=1}^N$ summarizes $p(x_1|y_0)$, where

$$x_1^{(i)} \sim N(x_0^{(i)}, \tau^2) \qquad i = 1, \ldots, N.$$

are draws from the prior $p(x_1|y_0)$.

Then, $\{(x_1, \omega_1)^{(i)}\}_{i=1}^N$ summarizes $p(x_1|y_1)$, where

$$\omega_1^{(i)} = \omega_0^{(i)} f_N(y_1; x_1^{(i)}, \sigma^2) \qquad i = 1, \ldots, N.$$

Example i. Sequential importance sampling (SIS)

Let $\{(x_{t-1}, \omega_{t-1})^{(i)}\}_{i=1}^N$ summarizes $p(x_{t-1}|y^{t-1})$.

Then, $\{(x_t, \omega_{t-1})^{(i)}\}_{i=1}^N$ summarizes $p(x_t|y^{t-1})$, where

$$\text{Propagation:} \quad x_t^{(i)} \sim N(x_{t-1}^{(i)}, \tau^2) \qquad i = 1, \ldots, N,$$

and $\{(x_t, \omega_t)^{(i)}\}_{i=1}^N$ summarizes $p(x^t|y^t)$, where

$$\text{Reweighting:} \quad \omega_t^{(i)} = \omega_{t-1}^{(i)} f_N(y_t; x_t^{(i)}, \sigma^2) \qquad i = 1, \ldots, N.$$

Effective sample size

Liu (1996) proposed using the following measure of degeneracy of an algorithm:

$$N_{\text{eff},t} = \frac{1}{\sum_{i=1}^N \left(w_t^{(i)}\right)^2}$$

where w_ts are normalized weights, i.e. $w_t^{(i)} = \omega_t^{(i)} / \sum_{j=1}^N \omega_t^{(j)}$.

If $w_t^{(i)} = 1/N$ (equally balanced weights), then

$$N_{\text{eff},t} = N.$$

If $w_t^{(j)} = 1$ for only one j (particle degeneracy) then

$$N_{\text{eff},t} = 1.$$

Example i. SIS with resampling (SISR)

SIS:

- $\{(x_{t-1}, \omega_{t-1})^{(i)}\}_{i=1}^N$ summarizes $p(x_{t-1}|y^{t-1})$.
- $\{(\tilde{x}_t, \omega_{t-1})^{(i)}\}_{i=1}^N$ summarizes $p(x_t|y^{t-1})$, where $\tilde{x}_t^{(i)} \sim N(x_{t-1}^{(i)}, \tau^2)$, for $i = 1, \ldots, N$.
- $\{(\tilde{x}_t, \tilde{\omega}_t)^{(i)}\}_{i=1}^N$ summarizes $p(x^t|y^t)$, where $\tilde{\omega}_t^{(i)} = \omega_{t-1}^{(i)} f_N(y_t; \tilde{x}_t^{(i)}, \sigma^2)$, for $i = 1, \ldots, N$.

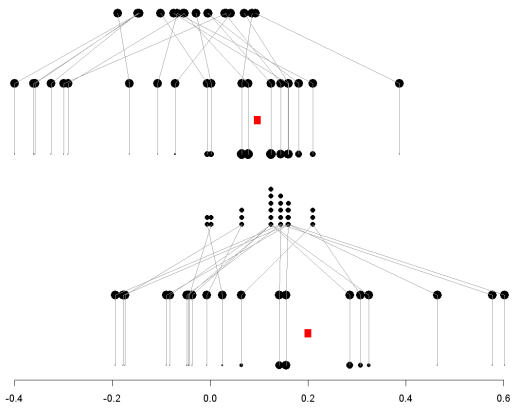
Resampling:

Draw $x_t^{(1)}, \ldots, x_t^{(N)}$ from the set $\left\{\tilde{x}_t^{(1)}, \ldots, \tilde{x}_t^{(N)}\right\}$ with weights $\left\{\tilde{\omega}_t^{(1)}, \ldots, \tilde{\omega}_t^{(N)}\right\}$.

Therefore, $\{(x_t, \omega_t)^{(i)}\}_{i=1}^N$ summarizes $p(x_t|y^t)$, where $\omega_t = 1/N$.

SIS with Resampling (SISR)

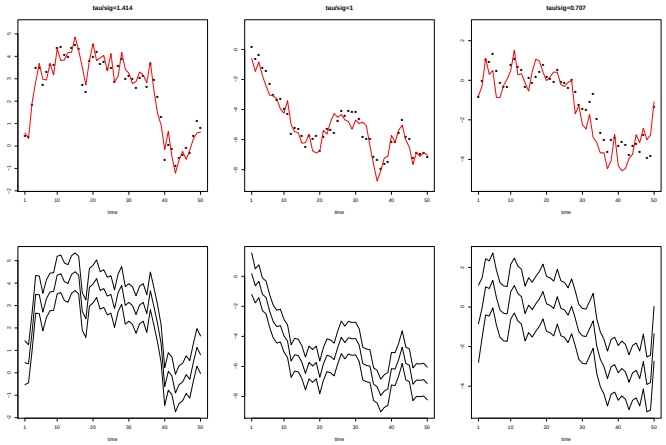
$$
\begin{aligned}
\{x_t^1, \ldots, x_t^N\} &\sim p(x_t|y^t) \\
\tilde{x}_{t+1}^i &\sim p(x_{t+1}|x_t^i) \\
\omega_{t+1}^i &\propto p(y_{t+1}|\tilde{x}_{t+1}^i) \\
\{x_{t+1}^1, \ldots, x_{t+1}^N\} &\sim p(x_{t+1}|y^{t+1}) \\
\tilde{x}_{t+2}^i &\sim p(x_{t+2}|x_{t+1}^i) \\
\omega_{t+2}^i &\propto p(y_{t+2}|\tilde{x}_{t+2}^i)
\end{aligned}
$$



Uniform weights is the goal!

Example i. Simulated data

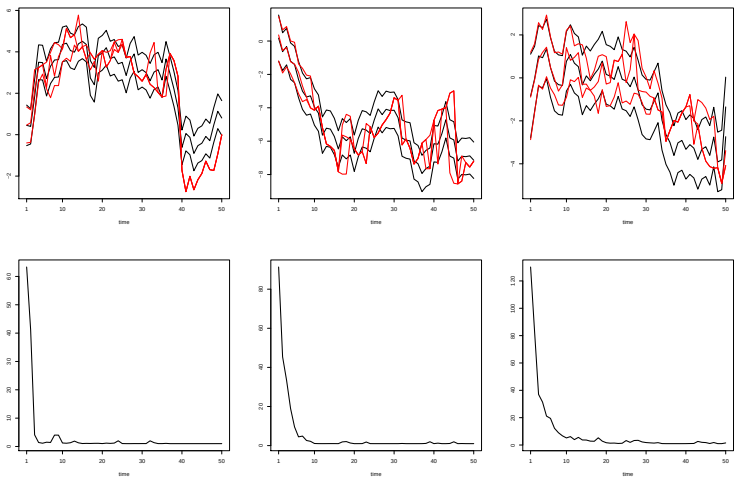
$n = 50$, $x_0 = 0$, $\tau^2 = 0.5$ and $\sigma^2 = (0.25, 0.5, 1.0)$.



Top: y_t and x_t; bottom: m_t and $m_t \pm 2\sqrt{C_t}$.

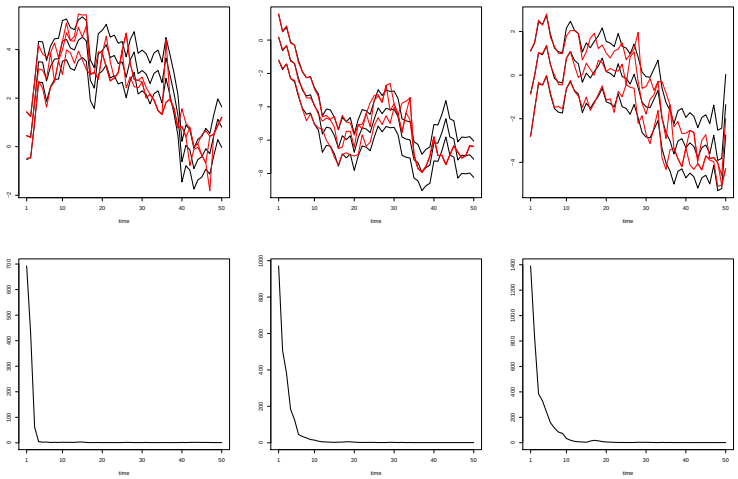
Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Example i. SIS, $N = 1,000$



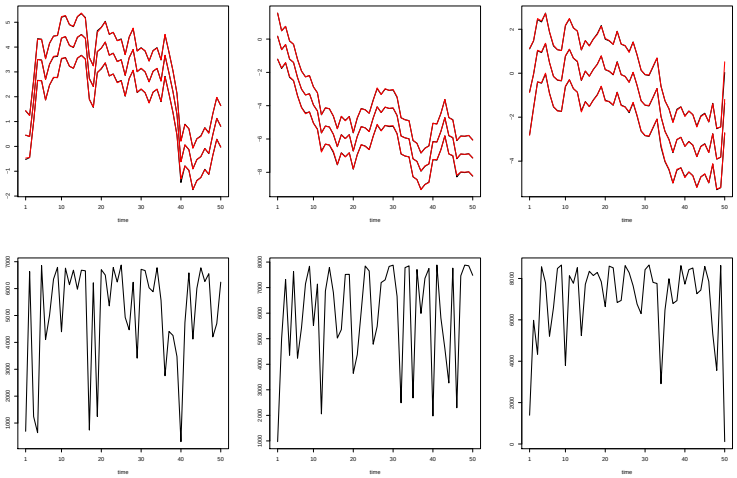
Top: States; Bottom: N_{eff}.
Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Example i. SIS, $N = 10,000$



Top: States; Bottom: N_{eff}.
Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Example i. SISR, $N = 10,000$

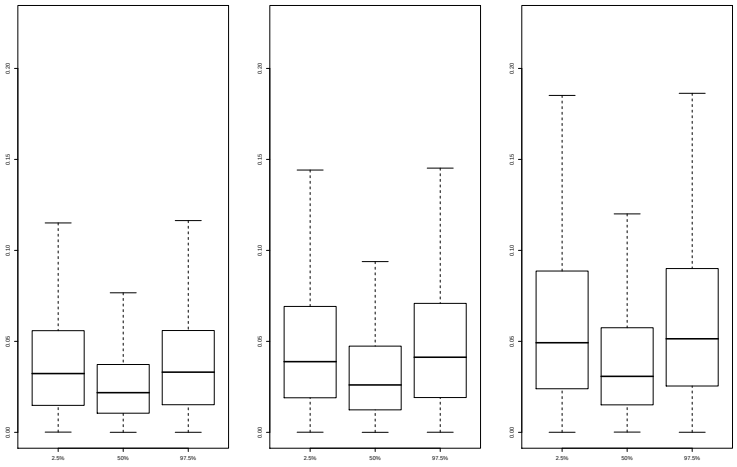


Top: States; Bottom: N_{eff}.

Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Example i. SISR, $n = 1,000$ and $N = 1,000$

$e_{t,\alpha} = |\hat{q}_\alpha(x_t|y^t) - q_\alpha(x_t|y^t)|$, for $\alpha = 0.025, 0.5, 0.975$.



Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Auxiliary particle filter (APF)

Recall the two main steps in any dynamic model:

$$\begin{aligned} p(x_t|y^{t-1}) &= \int p(x_t|x_{t-1})p(x_{t-1}|y^{t-1})dx_{t-1} \\ p(x_t|y^t) &\propto p(y_t|x_t)p(x_t|y^{t-1}) \end{aligned}$$

- $\left\{(x_{t-1}, \omega_{t-1})^{(i)}\right\}_{i=1}^N$ summarizes $p(x_{t-1}|y^{t-1})$.

- Approximating $p(x_t|y^{-1})$ by

$$p_N(x_t|y^{t-1}) = \sum_{i=1}^{N} p(x_t|x_{t-1}^{(i)})\omega_{t-1}^{(i)}$$

- Approximating $p(x_t|y^t)$ by

$$p_N(x_t|y^t) = \sum_{i=1}^{N} p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\omega_{t-1}^{(i)}$$

Pitt and Shephard's (1999) idea

The previous mixture approximation suggests an augmentation scheme where the new target distribution is

$$p_N(x_t, k|y^t) = p(y_t|x_t)p(x_t|x_{t-1}^{(k)})\omega_{t-1}^{(k)}.$$

A natural proposal distribution is

$$q(x_t, k|y^t) = p(y_t|g(x_{t-1}^{(k)}))p(x_t|x_{t-1}^{(k)})\omega_{t-1}^{(k)}$$

where, for instance, $g(x_{t-1}) = E(x_t|x_{t-1})$.

By a simple SIR argument, the weight of the particle x_t is

$$\omega_t \propto \frac{p(y_t|x_t)}{p(y_t|g(x_{t-1}^{(k)}))}$$

APF algorithm

- $\left\{(x_{t-1}, \omega_{t-1})^{(i)}\right\}_{i=1}^{N}$ summarizes $p(x_{t-1}|y^{t-1})$.
- For $j = 1, \ldots, N$
  - Draw k^j from $\{1, \ldots, N\}$ with weights $\{\tilde{\omega}_{t-1}^{(1)}, \ldots, \tilde{\omega}_{t-1}^{(N)}\}$:

$$\tilde{\omega}_{t-1}^{(i)} = \omega_{t-1}^{(i)} p(y_t|g(x_{t-1}^{(i)}))$$

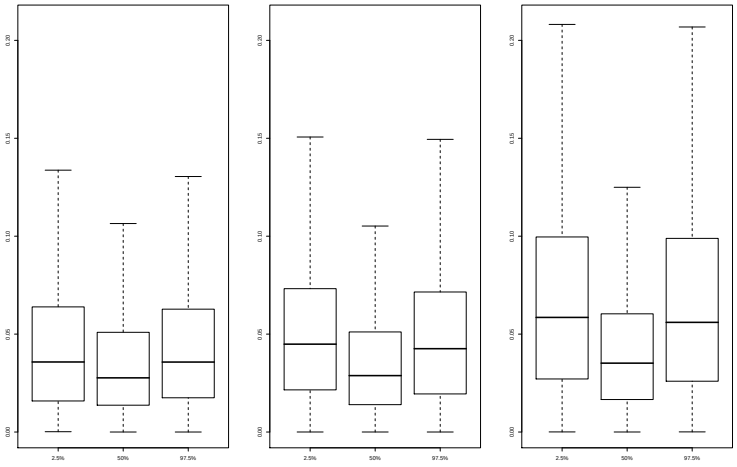
  - Draw $x_t^{(j)}$ from $p(x_t|x_{t-1}^{(k^j)})$.
  - Compute associated weight

$$\omega_t^{(j)} \propto \frac{p(y_t|x_t^{(j)})}{p(y_t|g(x_{t-1}^{(k^j)}))}.$$

- $\left\{(x_t, \omega_t)^{(i)}\right\}_{i=1}^{N}$ summarizes $p(x_t|y^t)$.
- Maybe add a SIR step to replenish x_ts.

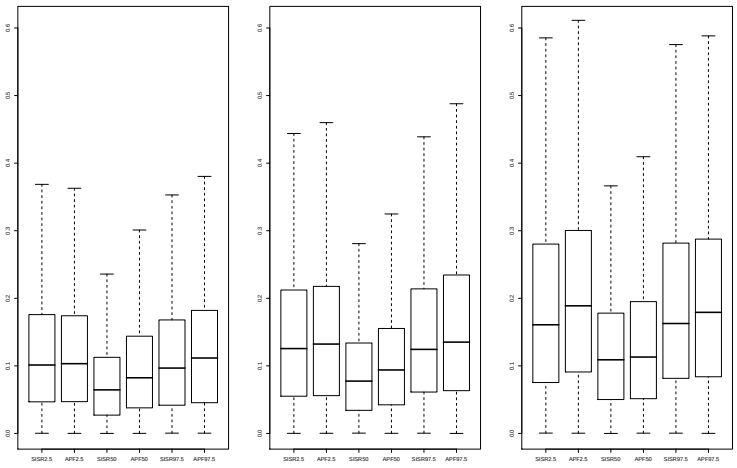
Example i. APF, $n = 1,000$ **and** $N = 1,000$

$e_{t,\alpha} = |\hat{q}_\alpha(x_t|y^t) - q_\alpha(x_t|y^t)|$, for $\alpha = 0.025, 0.5, 0.975$.



Left: $\tau/\sigma = 1.414$; center: $\tau/\sigma = 1.000$; right: $\tau/\sigma = 0.707$.

Example i. SISR & APF, $n = 1,000$ **and** $N = 100$



Example ii. Nonlinear dynamic model

Recall the nonlinear dynamic model previously studied

$$\begin{aligned}(y_t|x_t,\psi) &\sim N(x_t^2/20, \sigma^2)\\ (x_t|x_{t-1},\psi) &\sim N(G'_{x_{t-1}}\theta, \tau^2)\end{aligned}$$

where $x_0 \sim N(m_0, C_0)$, $\theta = (\alpha, \beta, \gamma)'$, $\psi = (\xi', \sigma^2, \tau^2)$ and

$$G'_{x_t} = \big(x_t, x_t/(1+x_t^2), \cos(1.2t)\big)$$

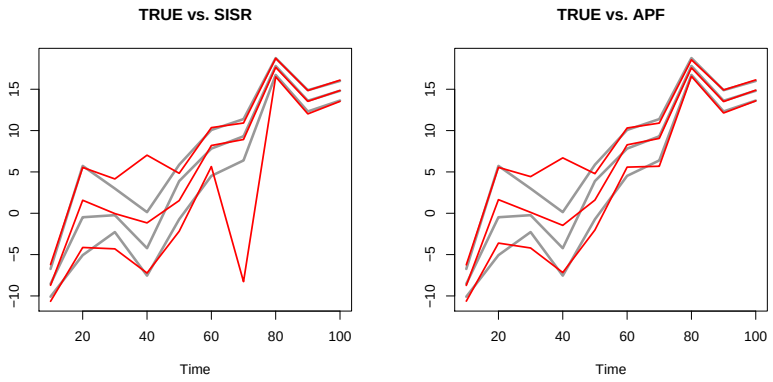
Simulated data: $n = 100$, $\sigma^2 = 1$, $\tau^2 = 10$, $\theta = (0.5, 25, 8)$, $x_0 = 0.1$.

Prior setup: $m_0 = 0$ and $C_0 = 10$.

MCMC setup: $m_0 = 0$, $C_0 = 10$, $v = 0.1$, $M_0 = 1000$ and $M = 5000$.

SMC setup: $N = 5000$.

Comparing SMCMC, SISR and APF



Smoothing

Godsill, Doucet and West (2004) proposed a smoothing scheme based on particle filter draws.

The key results are

$$p(x^n|y^n) = p(x_n|y^n) \prod_{t=1}^{n-1} p(x_t|x_{t+1}, y^t)$$

and (by Bayes rule and conditional independence)

$$p(x_t|x_{t+1}, y^t) \propto p(x_{t+1}|x_t, y^t)p(x_t|y^t).$$

We can now jointly sample from $p(x^n|y^n)$ by sequentially sampling from filtered particles with weights proportional to $p(x_{t+1}|x_t, y^t)$.

Backward sampling algorithm

Repeat the following three steps N times.

- Sample $\tilde{x}_n$ from $\{x_n^{(i)}\}_{i=1}^N$ with weights $\{\omega_n^{(i)}\}_{i=1}^N$.
- For $t = n-1, \ldots, 1$

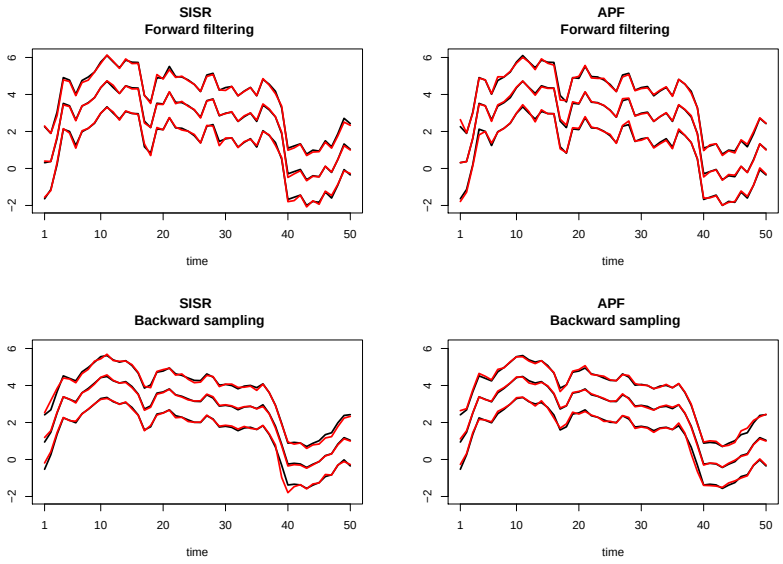
  Sample $\tilde{x}_t$ from $\{x_t^{(i)}\}_{i=1}^N$ with weights $\{\tilde{\omega}_t^{(i)}\}_{i=1}^N$

  $$\tilde{\omega}_t^{(i)} \propto \omega_t^{(i)} p\left(\tilde{x}_{t+1}|x_t^{(i)}\right) \qquad i = 1, \ldots, N$$

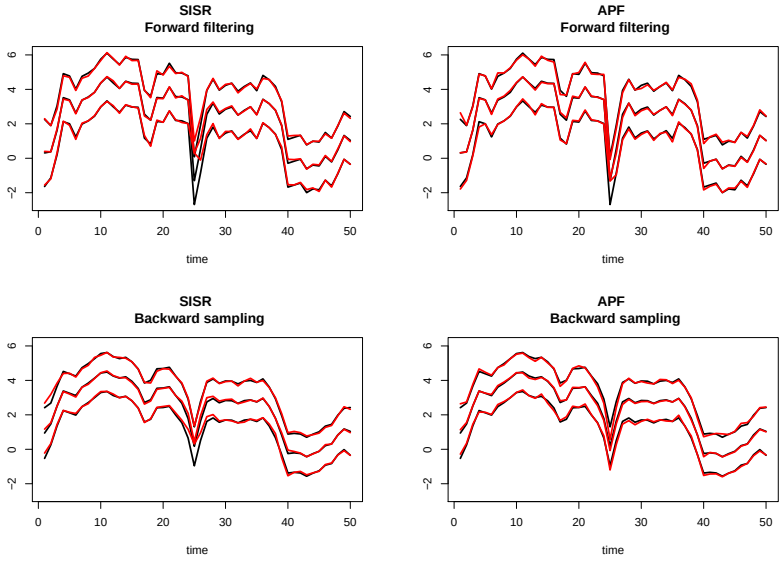
- Then $\{\tilde{x}_1^{(j)}, \ldots, \tilde{x}_n^{(j)}\}$ is a draw from $p(x^n|y^n)$.

Example i. smoothing

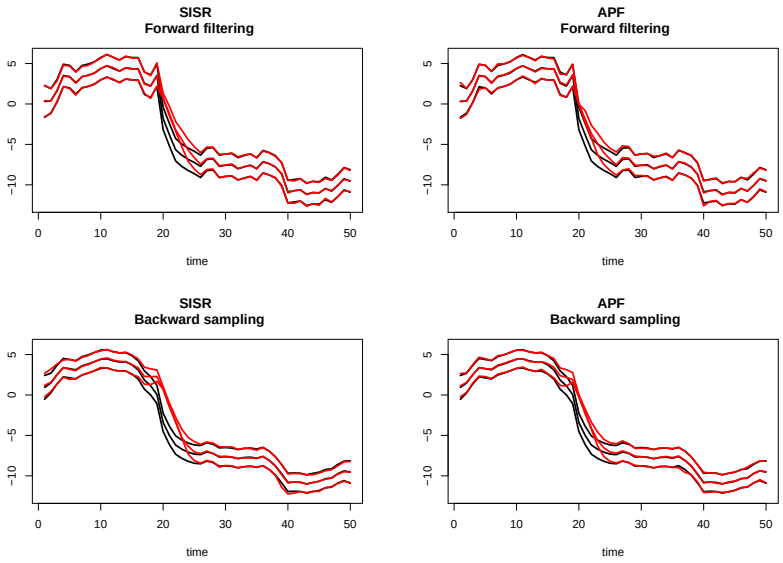
$n = 50$, $\tau^2 = 0.5$, $\sigma^2 = 1$, $x_0 = 0$, $m_0 = 0$, $C_0 = 100$, $N = 1000$.



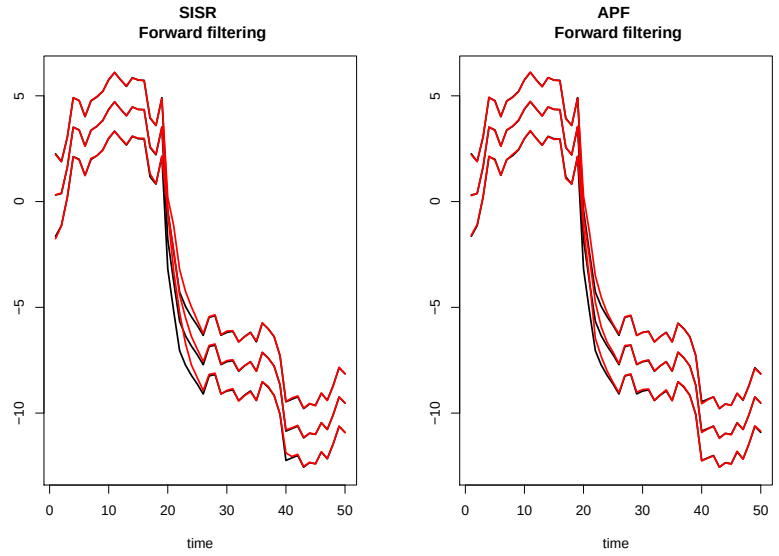
Example i. outlier in y_t



Example i. outlier in x_t



Example i. outlier in x_t (more particles)



LECTURE 9

SEQUENTIAL MONTE CARLO WITH PARAMETER LEARNING

Revisiting the bootstrap and the AP filters

Consider the following general state space model

$$\begin{aligned} \text{Observation equation} &: p(y_{t+1}|x_{t+1}) \\ \text{State equation} &: p(x_{t+1}|x_t) \end{aligned}$$

For a given time t

$$\{(x_t, \omega_t)^{(i)}\}_{i=1}^N$$

is a particle representation of

$$p(x_t|y^t)$$

where $y^t = (y_1, \ldots, y_n)$.

Sample-resample

Goal: $\{(x_{t+1}, \omega_{t+1})^{(i)}\}_{i=1}^N \sim p(x_{t+1}|y^{t+1})$.

Algorithm

- Sample $x_{t+1}^{(i)}$ from $q(x_{t+1}|x_t^{(i)}, y_{t+1})$
- Compute weights

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|x_{t+1}^{(i)}) p(x_{t+1}^{(i)}|x_t^{(i)})}{q(x_{t+1}^{(i)}|x_t^{(i)}, y_{t+1})}$$

Special case: bootstrap filter

In the bootstrap filter

$$q(x_{t+1}|x_t, y_{t+1}) = p(x_{t+1}|x_t),$$

i.e. the transition equation.

This proposal density has no information about y_{t+1}, so we say that the scheme is *blinded*.

The weights are then proportional to the likelihoods

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} p(y_{t+1}|x_{t+1}^{(i)}).$$

Special case: optimal filter

In the optimal filter

$$q(x_{t+1}|x_t, y_{t+1}) = p(x_{t+1}|x_t, y_{t+1}).$$

The weights are then

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} p(y_{t+1}|y^t) \propto \omega_t^{(i)}$$

so, if $\omega_0 \propto 1$, then $\omega_{t+1} \propto 1$ for all t.

This is a **perfectly adapted** filter.

Resample-sample

Goal: $\{(x_{t+1}, \omega_{t+1})^{(i)}\}_{i=1}^N \sim p(x_{t+1}|y^{t+1})$.

Algorithm

- Resample $\tilde{x}_t^{(i)}$ from $\{x_t^{(1)}, \ldots, x_t^{(N)}\}$ with weights

$$q_1(x_t^{(j)}|y_{t+1}) \qquad j = 1, \ldots, N$$

- Sample $x_{t+1}^{(i)}$ from $q_2(x_{t+1}|\tilde{x}_t^{(i)}, y_{t+1})$

- Compute weights

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|x_{t+1}^{(i)}) p(x_{t+1}^{(i)}|\tilde{x}_t^{(i)})}{q_1(\tilde{x}_t^{(i)}|y_{t+1}) q_2(x_{t+1}^{(i)}|\tilde{x}_t^{(i)}, y_{t+1})}$$

Special case: auxiliary particle filter

In the auxiliary particle filter

$$q_1(x_t|y_{t+1}) = p(y_{t+1}|g(x_t))$$

where, for instance, $g(x_t) = E(x_{t+1}|x_t)$.

Also,

$$q_2(x_{t+1}|x_t, y_{t+1}) = p(x_{t+1}|x_t)$$

i.e. the transition equation, so again a *blinded* proposal.

The weights are then equal to

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|x_{t+1}^{(i)})}{p(y_{t+1}|g(\tilde{x}_t^{(i)}))}.$$

Special case: optimal filter

In the optimal filter both proposals q_1 and q_2 depend on y_{t+1}, i.e.

$$q_1(x_t|y_{t+1}) = p(y_{t+1}|x_t).$$

and

$$q_2(x_{t+1}|x_t, y_{t+1}) = p(x_{t+1}|x_t, y_{t+1}).$$

The weights are then equal to

$$\omega_{t+1}^{(i)} = \omega_t^{(i)}$$

so, if $\omega_0 \propto 1$, then $\omega_{t+1} \propto 1$ for all t.

This is a **perfectly adapted** filter.

Resample-sample with learning θ

The objective is to combine $\{(x_t, \theta_t, \omega_t)^{(i)}\}_{i=1}^N \sim p(x_t, \theta|y^t)$ with y_{t+1} to produce $\{(x_{t+1}, \theta_{t+1}, \omega_{t+1})^{(i)}\}_{i=1}^N \sim p(x_{t+1}, \theta|y^{t+1})$.

The index t and $t+1$ in $\theta^{(i)}$ are used to facilitate the identification of the time at which draws are being used.

Algorithm

- Resample $(\tilde{x}_t, \tilde{\theta}_t)^{(i)}$ from $\{(x_t, \theta_t)^{(j)}\}_{j=1}^N$ with weights

$$q_1((x_t, \theta_t)^{(j)}|y_{t+1}) \qquad j = 1, \ldots, N.$$

- Sample $(x_{t+1}, \theta_{t+1})^{(i)}$ from $q_2(x_{t+1}, \theta|(\tilde{x}_t, \tilde{\theta}_t)^{(i)}, y_{t+1})$.

- Compute weights

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|(x_{t+1}, \theta_{t+1})^{(i)})}{q_1((\tilde{x}_t, \tilde{\theta}_t)^{(i)})|y_{t+1})} \frac{p((x_{t+1}, \theta_{t+1})^{(i)}|(\tilde{x}_t, \tilde{\theta}_t)^{(i)})}{q_2((x_{t+1}, \theta_{t+1})^{(i)}|(\tilde{x}_t, \tilde{\theta}_t)^{(i)}, y_{t+1})}$$

Questions:

- How to choose q_1 and q_2?

- What is $p(x_{t+1}, \theta_{t+1}|x_t, \theta_t)$?

- Is it okay to decompose it as

$$p(x_{t+1}, \theta_{t+1}|x_t, \theta_t) = p(x_{t+1}|\theta_t, x_t)p(\theta_{t+1}|x_t, \theta_t)?$$

- If so, then what is $p(\theta_{t+1}|x_t, \theta_t)$?

Liu and West (2001)

They approximate $p(\theta|y^t)$ by a N-component mixture of multivariate normal distributions, i.e.

$$p(\theta|y^t) = \sum_{i=1}^N \omega_t^{(i)} f_N(\theta|a\theta_t^{(i)} + (1-a)\bar{\theta}_t, (1-a^2)V_t)$$

where $\bar{\theta}_t = \sum_{i=1}^N \omega_t^{(i)} \theta_t^{(i)}$ and $V_t = \sum_{t=1}^N \omega_t^{(i)} (\theta_t^{(i)} - \bar{\theta}_t)(\theta_t^{(i)} - \bar{\theta}_t)'$.

This leads to

$$p(\theta_{t+1}|x_t^{(i)}, \theta_t^{(i)}) = f_N(\theta_{t+1}|a\theta_t^{(i)} + (1-a)\bar{\theta}_t, (1-a^2)V_t)$$

They use the same decomposition for q_2. So the weights are

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|(x_{t+1}, \theta_{t+1})^{(i)})}{q_1((\tilde{x}_t, \tilde{\theta}_t)^{(i)})|y_{t+1})}$$

Resampling step

$$q_1(x_t, \theta_t|y_{t+1}) = p(y_{t+1}|g(x_t), m(\theta_t))$$

where

$$g(x_t) = E(x_{t+1}|x_t, m(\theta_t))$$

for instance, and

$$m(\theta_t) = a\theta_t + (1-a)\bar{\theta}_t$$

The weights are then

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|x_{t+1}^{(i)}, \theta_{t+1}^{(i)})}{p(y_{t+1}|g(\tilde{x}_t^{(i)}), m(\tilde{\theta}_t^{(i)}))}$$

Choosing a

Liu and West (2001) use a discount factor argument (see West and Harrison, 1997) to set the parameter a:

$$a = \frac{3\delta - 1}{2\delta}$$

For example,

- $\delta = 0.50$ leads to $a = 0.500$
- $\delta = 0.75$ leads to $a = 0.833$
- $\delta = 0.95$ leads to $a = 0.974$
- $\delta = 1.00$ leads to $a = 1.000$.

In the last case, i.e. $a = 1.0$, the particles of θ will degenerate over time to a single particle.

The LW filter in one page

For particles $\{(x_t, \theta_t, \omega_t)^{(j)}\}_{j=1}^N$ summarizing $p(x_t, \theta|y^t)$, estimates $\bar{\theta}_t = \sum_{i=1}^N \omega_t^{(i)}\theta_t^{(i)}$ and $V_t = \sum_{i=1}^N \omega_t^{(i)}(\theta_t^{(i)} - \bar{\theta}_t)(\theta_t^{(i)} - \bar{\theta}_t)'$, and given shrinkage parameter a, the algorithm runs as follows.

- For $i = 1, \ldots, N$, compute

- $m(\theta_t^{(i)}) = a\theta_t^{(i)} + (1-a)\bar{\theta}_t$.
- $g(x_t^{(i)}) = E(x_{t+1}|x_t^{(i)}, m(\theta_t^{(i)}))$.
- $w_{t+1}^{(i)} = p(y_{t+1}|g(x_t^{(j)}), m(\theta_t^{(j)}))$.

- For $i = 1, \ldots, N$
  - Resample $(\tilde{x}_t, \tilde{\theta}_t)^{(i)}$ from $\{(x_t, \theta_t, w_{t+1})^{(j)}\}_{j=1}^N$.
  - Sample $\theta_{t+1}^{(i)} \sim N(m(\tilde{\theta}_t^{(i)}), h^2 V_t)$.
  - Sample $x_{t+1}^{(i)}$ from $p(x_{t+1}|\tilde{x}_t^{(i)}, \theta_{t+1}^{(i)})$.
  - Compute weight

$$\omega_{t+1}^{(i)} = \omega_t^{(i)} \frac{p(y_{t+1}|x_{t+1}^{(i)}, \theta_{t+1}^{(i)})}{p(y_{t+1}|g(\tilde{x}_t^{(i)}), m(\tilde{\theta}_t^{(i)}))}.$$

Example i. first order dynamic linear model

The revisit the first order dynamic linear model

$$\begin{aligned} y_t &= x_t + \nu_t \qquad \nu_t \sim N(0, \sigma^2) \\ x_t &= x_{t-1} + \omega_t \qquad \omega_t \sim N(0, \tau^2) \end{aligned}$$

where $x_0 = 25$, $\sigma^2 = 0.1$, $\tau^2 = (0.2, 0.1, 0.05)$ and $n = 200$.

Prior setup:

$$\begin{aligned} \sigma^2 &\sim IG(a_0, b_0) \\ x_0 &\sim N(m_0, C_0) \end{aligned}$$

where $a_0 = 5$, $b_0 = 0.4$, $m_0 = 25$ and $C_0 = 100$.

Particle filter setup:

$$\begin{aligned} N &= 2000 \\ \delta &= (0.75, 0.95) \end{aligned}$$

Example i. LW + optimal propagation

Liu and West's (2001) filter with optimal resampling proposal, i.e.

$$p(x_{t+1}|x_t, \sigma^2, y_{t+1}) = f_N(x_{t+1}|m_{t+1}, C_{t+1})$$

where

$$\begin{aligned} C_{t+1}^{-1} &= \tau^{-2} + \sigma^{-2} \\ m_{t+1} &= C_{t+1}(\sigma^{-2} y_{t+1} + \tau^{-2} x_t) \end{aligned}$$

Example i. LW + optimal propagation + kernel for σ^2

Optimal propagation + with mixture approximating σ^2 directly, i.e.

$$q(\sigma^2|x_t, \sigma_t^2, y_{t+1}) \propto f_N(y_{t+1}; x_t, \sigma^2) f_{IG}(\sigma^2|\alpha(\sigma_t^2), \beta(\sigma_t^2))$$

where

$$\begin{aligned} \alpha(\sigma_t^2) &= \frac{\{m(\sigma_t^2)\}^2}{v(\sigma_t^2)} + 2 \\ \beta(\sigma_t^2) &= m(\sigma_t^2)\alpha(\sigma_t^2) \end{aligned}$$

and

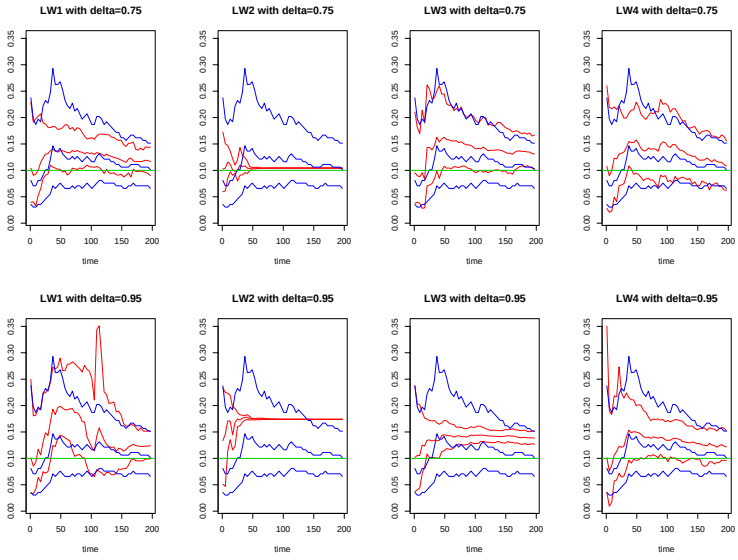
$$\begin{aligned} m(\sigma_t^2) &= a\sigma_t^2 + (1-a)\bar{\sigma}^2 \\ v(\sigma_t^2) &= (1-a^2)S_{\sigma^2}^2 \end{aligned}$$

with $\bar{\sigma}^2$ and $S_{\sigma^2}^2$ the particle approximation to the mean and variance of σ^2 from $p(\sigma^2|y^t)$.

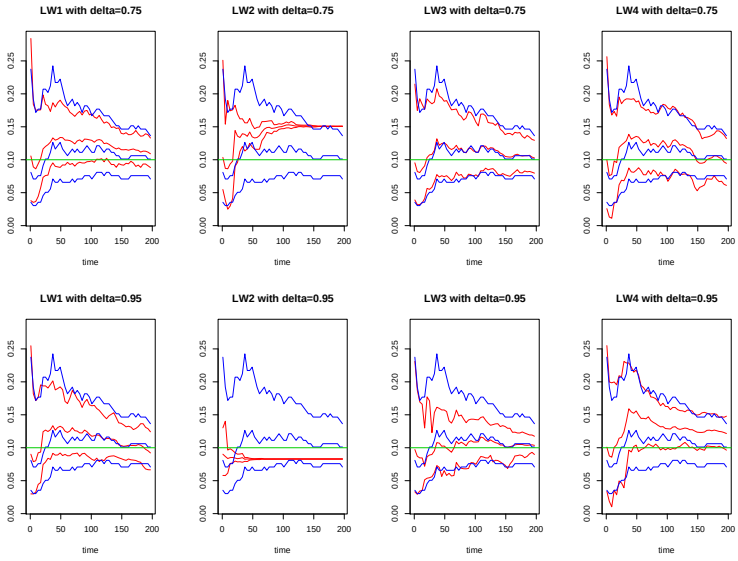
Example i. Comparing various LW filters

- LW1 : LW + $\log \sigma^2$
- LW2 : LW + σ^2
- LW3 : LW + $\log \sigma^2$ + optimal propagation
- LW4 : LW + σ^2 + optimal propagation

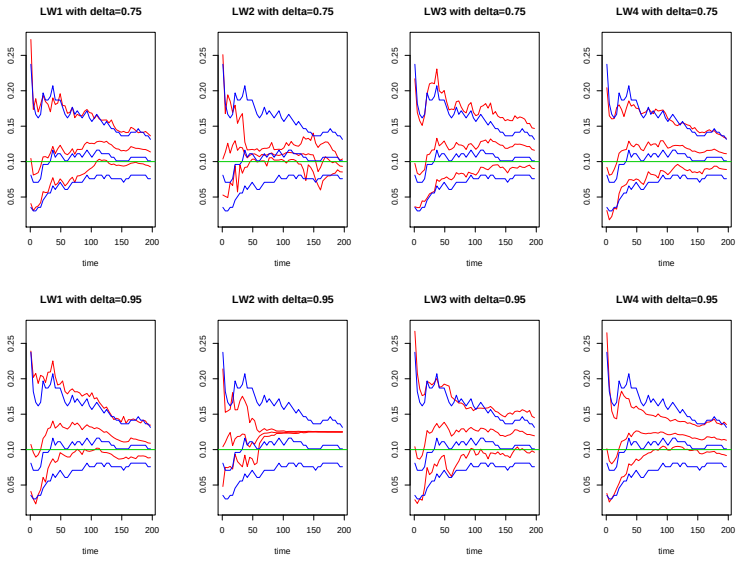
Example i. $\tau/\sigma = 1.4$



Example i. $\tau/\sigma = 1.0$



Example i. $\tau/\sigma = 0.7$



Example ii. nonlinear dynamic model

Let y_t, for $t = 1, \ldots, n$, be modeled as

$$\begin{aligned}(y_t|x_t, \psi) &\sim N(x_t^2/20, \sigma^2)\\ (x_t|x_{t-1}, \psi) &\sim N(G'_{x_{t-1}}\xi, \tau^2)\end{aligned}$$

where $G'_{x_t} = \big(x_t, x_t/(1+x_t^2), \cos(1.2t)\big)$, $\psi = (\xi', \sigma^2, \tau^2)$ and $\xi = (\alpha, \beta, \gamma)'$.

Prior distributions for θ_0, ξ, σ^2 and τ^2 are

$$\begin{aligned} x_0 &\sim N(m_0, V_0) \\ \xi &\sim N(c_0, C_0) \\ \sigma^2 &\sim IG(a_0, A_0) \\ \tau^2 &\sim IG(b_0, B_0) \end{aligned}$$

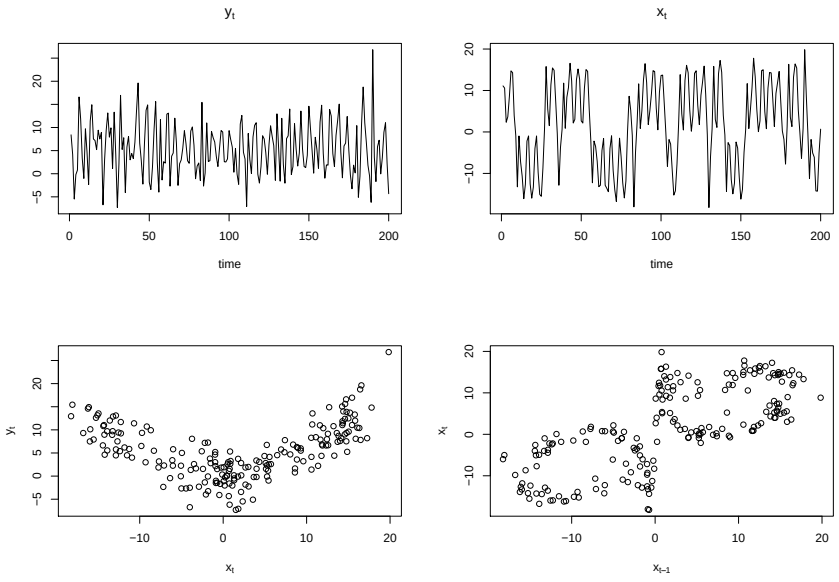
Example ii. Simulation set up

We simulated $n = 200$ observations based on $\xi = (0.5, 25, 8)'$, $\sigma^2 = 10$, $\tau^2 = 1$ and $x_0 = 0.1$.

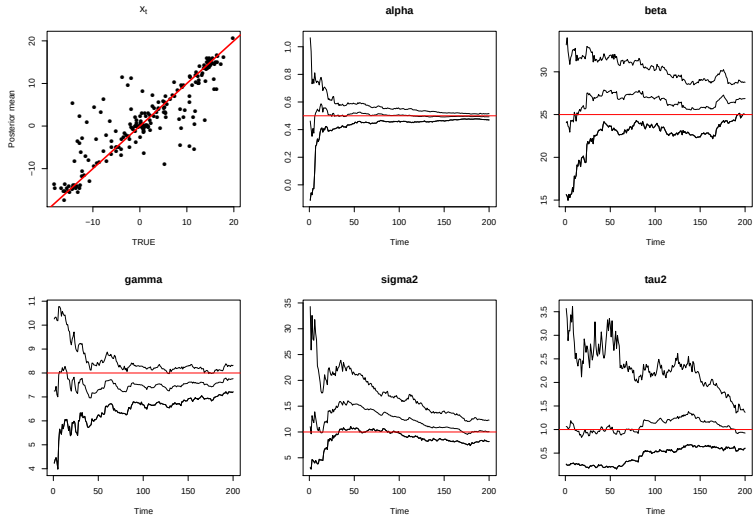
Prior hyperparameters:

$$\begin{aligned} m_0 &= 0.0 \quad \text{and} \quad V_0 = 5 \\ c_0 &= (0.5, 25, 8)' \quad \text{and} \quad C_0 = \text{diag}(0.1, 16, 2) \\ a_0 &= 3 \quad \text{and} \quad A_0 = 20 \\ b_0 &= 3 \quad \text{and} \quad B_0 = 2 \end{aligned}$$

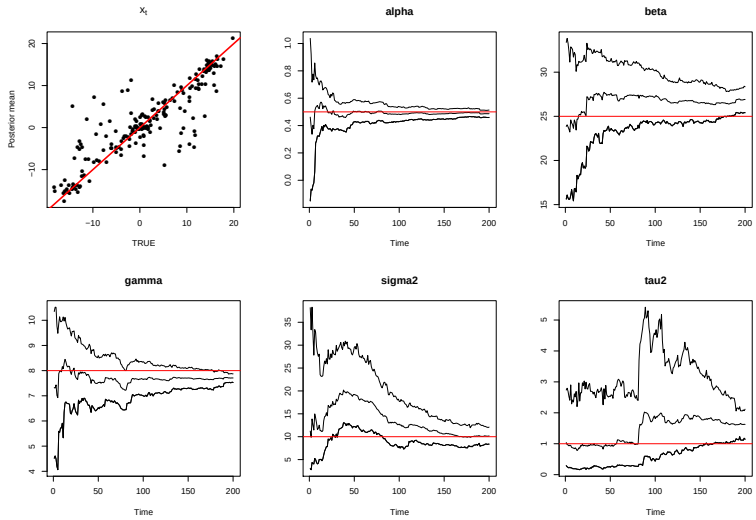
Example ii. Simulated data



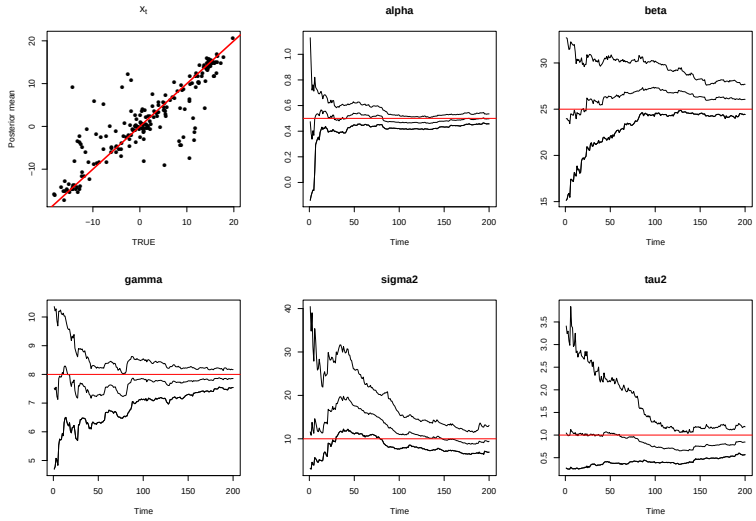
Example ii. $(N, \delta, a) = (2000, 0.75, 0.83)$



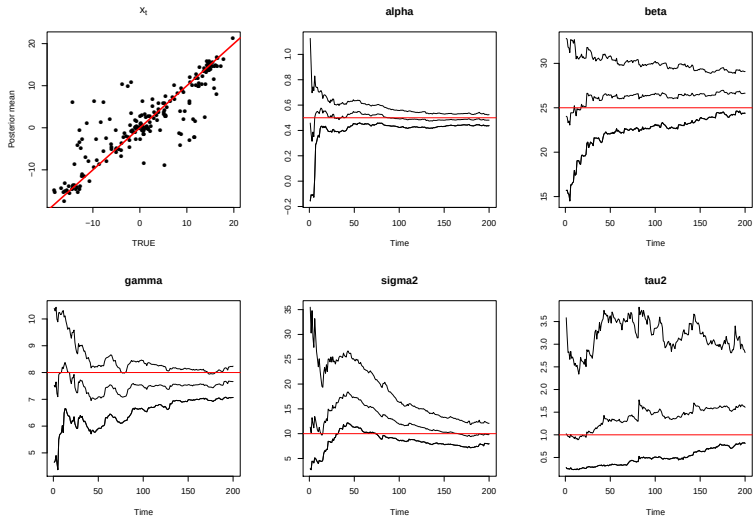
Example ii. $(N, \delta, a) = (2000, 0.90, 0.94)$



Example ii. $(N, \delta, a) = (5000, 0.90, 0.94)$

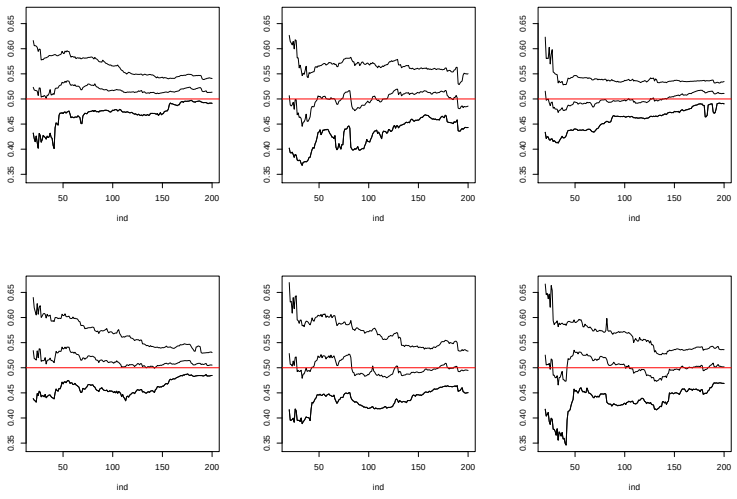


Example ii. $(N, \delta, a) = (10000, 0.90, 0.94)$

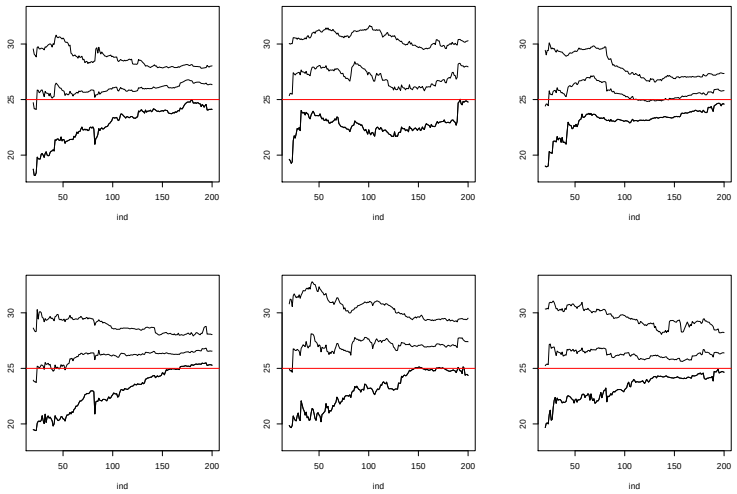


Example ii. Assessing MC error - α

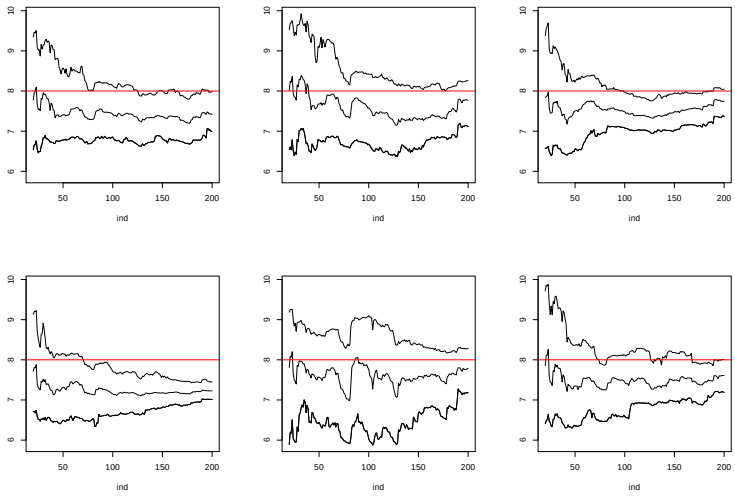
$N = 5000$



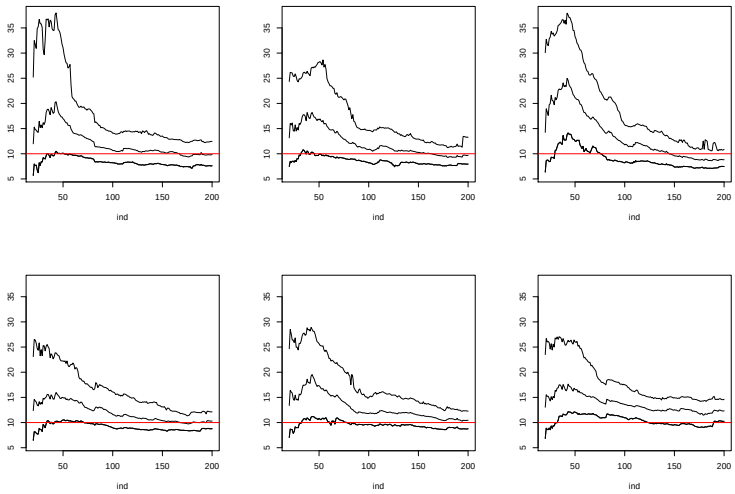
Example ii. Assessing MC error - β



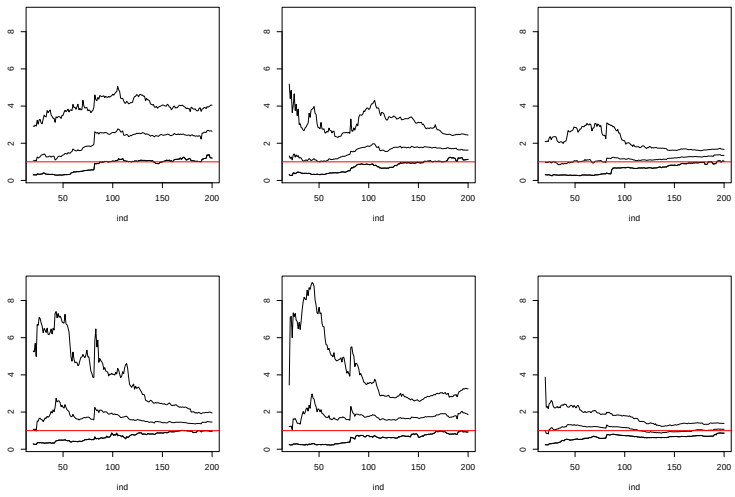
Example ii. Assessing MC error - γ



Example ii. Assessing MC error - σ^2

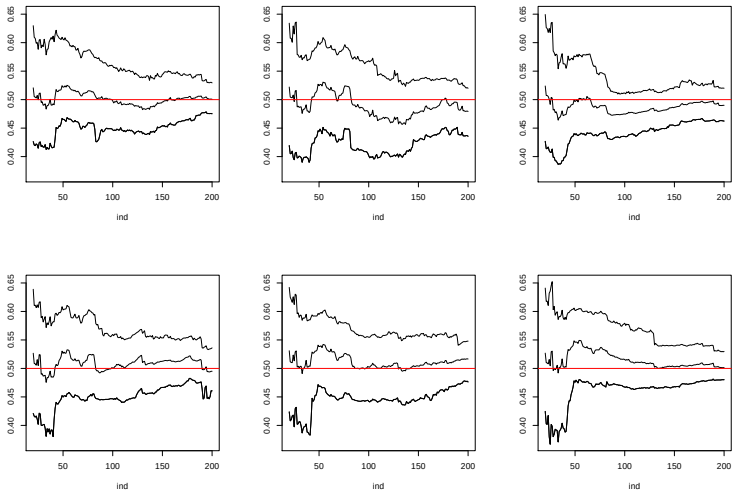


Example ii. Assessing MC error - τ^2

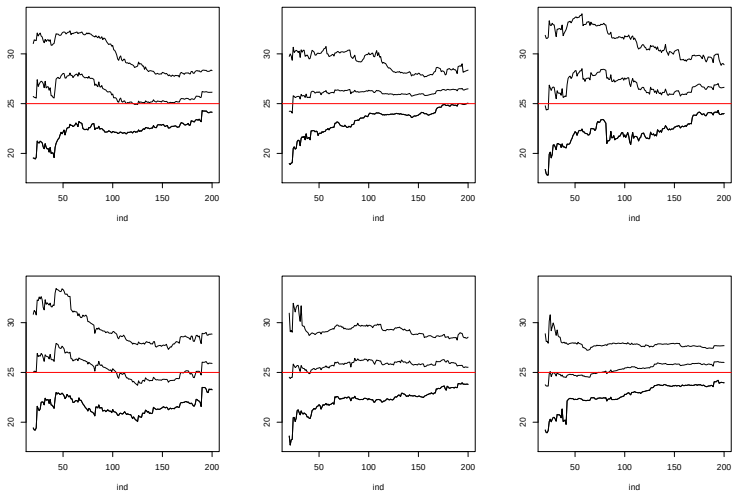


Example ii. Assessing MC error - α

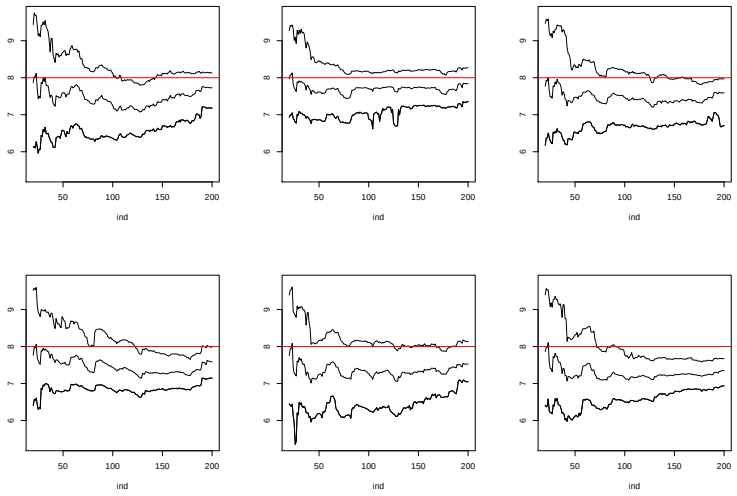
$N = 10000$



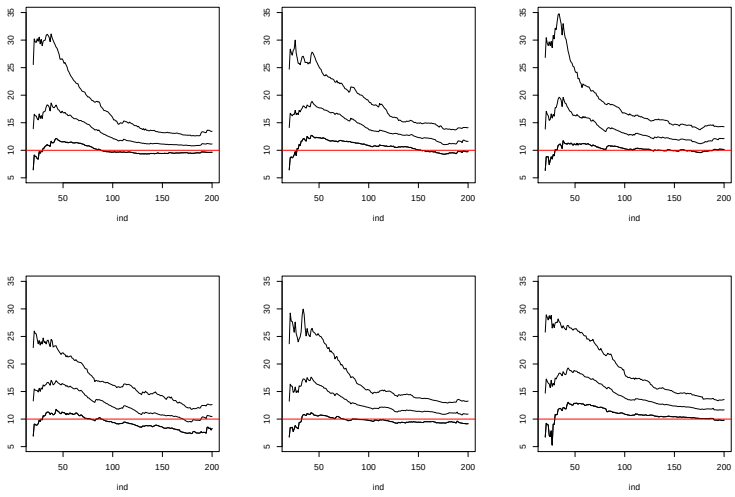
Example ii. Assessing MC error - β



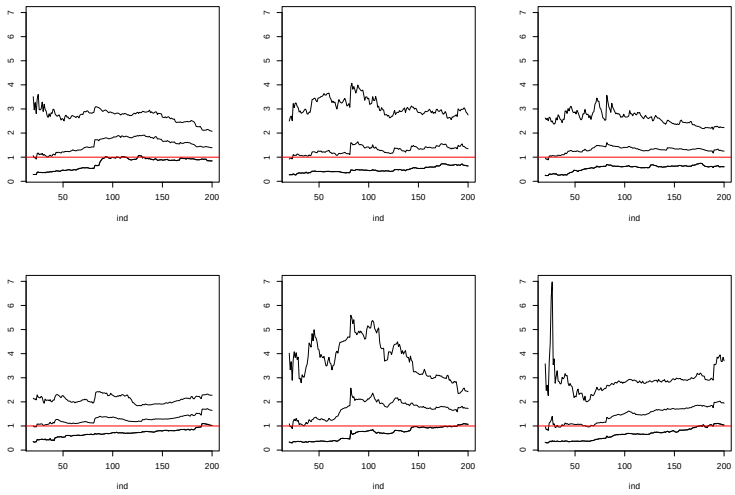
Example ii. Assessing MC error - γ



Example ii. Assessing MC error - σ^2



Example ii. Assessing MC error - τ^2



Particle Learning (PL)

Carvalho, Johannes, Lopes and Polson (2009) introduce Particle Learning (PL) as the following **resample-sample** scheme, where s_t is the vector of sufficient statistics for θ.

- Posterior at t: $\{(x_t, s_t, \theta)^{(i)}\}_{i=1}^N \sim p(x_t, s_t, \theta | y^t)$.
- Resampling weights: $w_{t+1}^{(j)} \propto p(y_{t+1} | x_t^{(j)}, \theta^{(j)})$, $j = 1, \ldots, N$.
- For $i = 1, \ldots, N$
  - Resample: Draw $\{(\tilde{x}_t, \tilde{s}_t, \tilde{\theta})^{(i)}\}_{i=1}^N$ from $\{(x_t, s_t, \theta)^{(j)}, w_{t+1}\}_{j=1}^N$.
  - Sample: Draw $x_{t+1}^{(i)} \sim p(x_{t+1} | (\tilde{x}_t, \tilde{\theta})^{(i)}, y_{t+1})$.

– Recursive sufficient statistics: $s_{t+1}^{(i)} = \mathcal{S}(s_t^{(i)}, x_{t+1}^{(i)}, y_{t+1})$.

– Offline sampling of fixed parameters: $\theta^{(i)} \sim p(\theta|s_{t+1}^{(i)})$.

PL ingredients

Resampling distribution

$$p(y_{t+1}|x_t, \theta)$$

Propagating distribution

$$p(x_{t+1}|x_t, \theta, y_{t+1})$$

Recursive sufficient statistics

$$s_{t+1} = \mathcal{S}(s_t, x_{t+1}, y_{t+1})$$

Example i. 1st order dynamic linear model via PL

- $p(y_{t+1}|x_t, \sigma^2)$ is

$$N(y_{t+1}; x_t, \sigma^2 + \tau^2)$$

- $p(x_{t+1}|x_t, \sigma^2, y_{t+1})$ is

$$N(x_{t+1}; Ay_{t+1} + (1-A)x_t, A\sigma^2)$$

  where $A = \tau^2/(\tau^2 + \sigma^2)$.

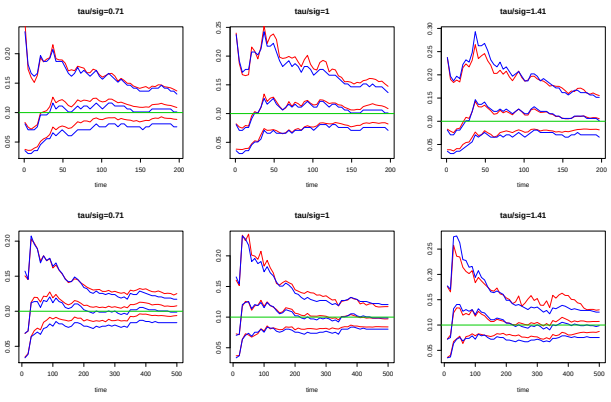
- $p(\sigma^2|s_{t+1})$ is

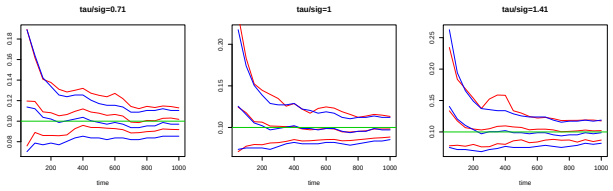
$$IG(a_{t+1}, b_{t+1})$$

  where $s_{t+1} = (a_{t+1}, b_{t+1})$ is recursively updated

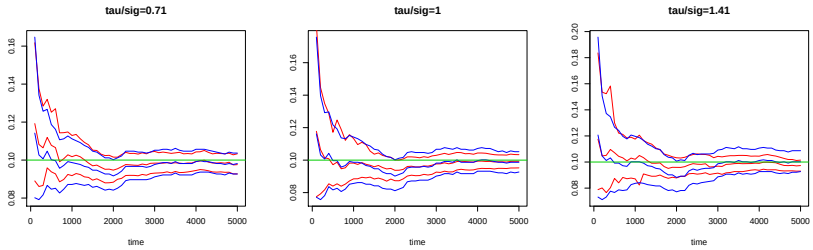
$$\begin{aligned} a_{t+1} &= a_t + \frac{1}{2} \\ b_{t+1} &= b_t + \frac{(y_t - x_t)^2}{2} \end{aligned}$$

Example i. learning τ^2 - $n = 200, 500, 1000$





Example i. learning τ^2 - $n = 5000$ and $N = 2000$



Example iii. MCMC and PL comparison

- **Simulation set up:** For $t = 1, \ldots, n = 300$,

$$p(y_t|x_t, \theta) \equiv f_N(x_t; \sigma^2) \tag{11}$$
$$p(x_t|x_{t-1}, \theta) \equiv f_N(\rho x_{t-1}; \tau^2) \tag{12}$$

  where $\theta = (\rho, \sigma^2, \tau^2) = (1.0, 1.0, 0.25)$ and $x_0 = 0$.

- **Model set up:** Equations (1) and (2) above plus

$$\begin{aligned} p(\theta, x_0) &\equiv f_N(\rho; r_0, W_0) f_{IG}(\sigma^2; a_0, b_0) \\ &\times f_{IG}(\tau^2; c_0, d_0) f_N(x_0; m_0, V_0) \end{aligned}$$

  where $r_0 = 0$, $W_0 = 3$, $a_0 = 3$, $b_0 = 2$, $c_0 = 3$, $d_0 = 0.5$, $m_0 = 0$ and $V_0 = 3$.

- **MCMC set up:** $M_0 = 100K$, $L = 100$ and $M = 20K$. A total of $2100K$ draws.

- **SMC set up:** $M = 20K$ particles.

Example iii. MCMC algorithms

- Gibbs sampler (GIBBS)
  - Sample x from $p(x|y, \theta)$ - FFBS
  - Sample σ^2 from $p(\sigma^2|x, y)$
  - Sample ρ from $p(\rho|x, \tau^2)$
  - Sample τ^2 from $p(\tau^2|x, \rho)$
- Random-walk Metropolis (RW)

– Sample x from $p(x|y,\theta)$ - FFBS

– Sample θ^* from

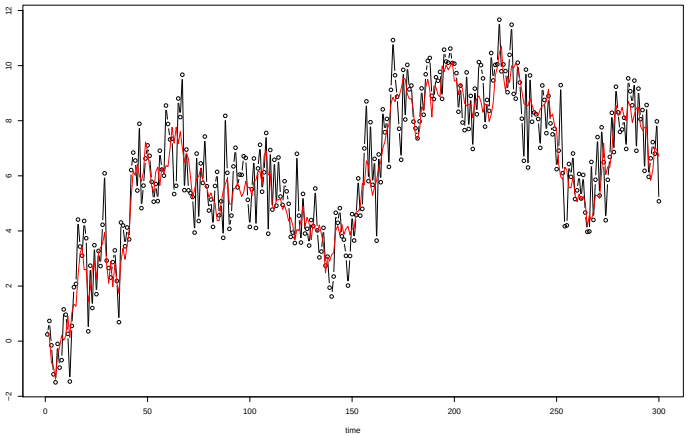
$$q(\theta^*|\theta) = q_\rho(\rho, V_\rho) q_{\sigma^2}(\sigma, V_{\sigma^2}) q_{\tau^2}(\tau^2, V_{\tau^2}),$$

with $V_\rho = 0.01$, $V_{\sigma^2} = 0.01$ and $V_{\tau^2} = 0.01$, and accept with probability

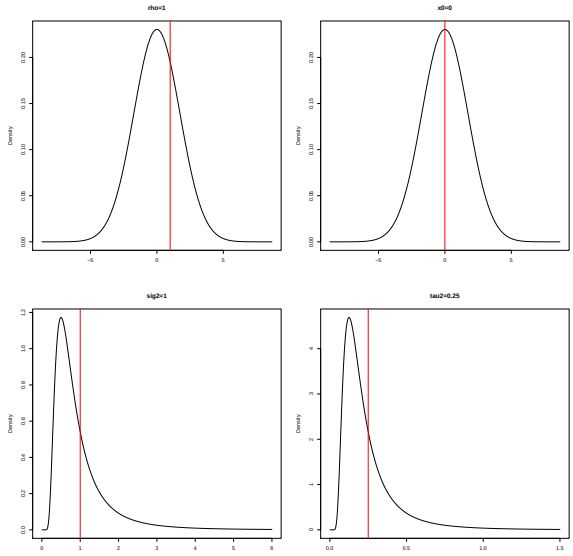
$$\alpha = \min\left\{1, \frac{p(y|\theta^*)p(\theta^*)q(\theta^*|\theta)}{p(y|\theta)p(\theta)q(\theta|\theta^*)}\right\}.$$

Note : Since $p(y|\theta) = \int p(y|x,\theta)p(x|\theta)dx$ can be analytically derived, x and θ are jointly sampled.

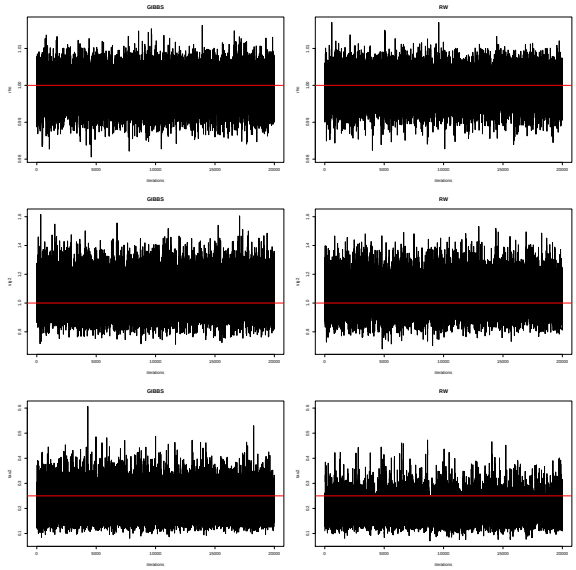
Example iii. Simulated y_t and x_t



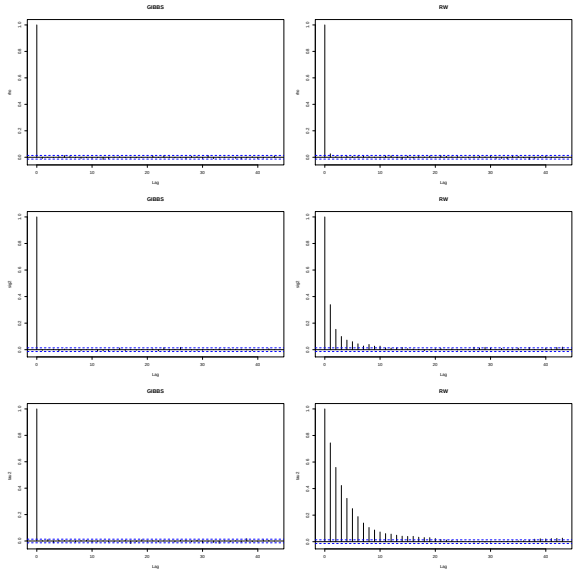
Example iii. Prior of (θ, x_0)



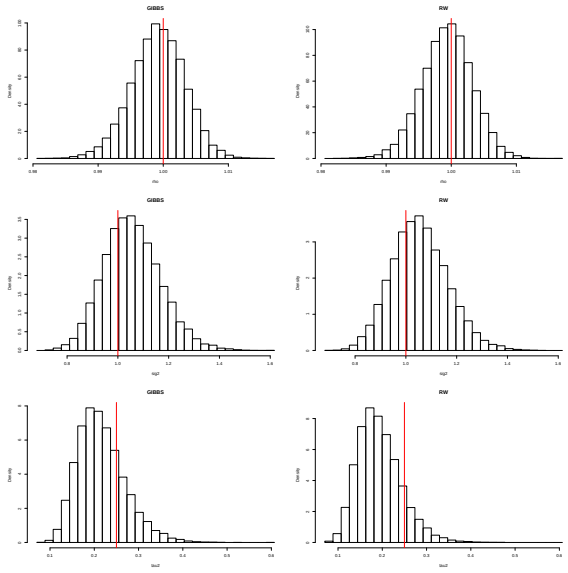
Example iii. MCMC trace plots



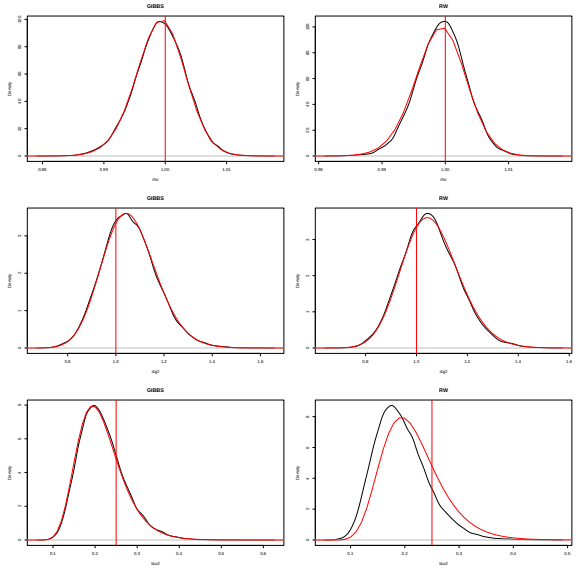
Example iii. MCMC autocorrelation plots



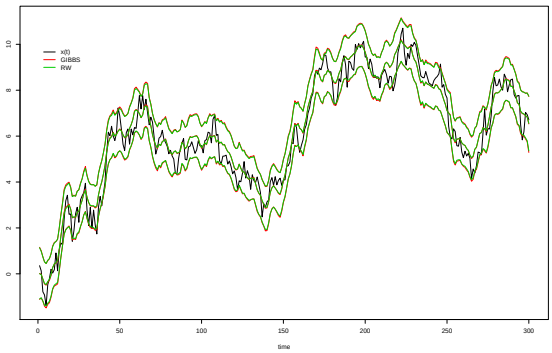
Example iii. MCMC marginal posteriors



Example iii. MCMC and true marginal posteriors

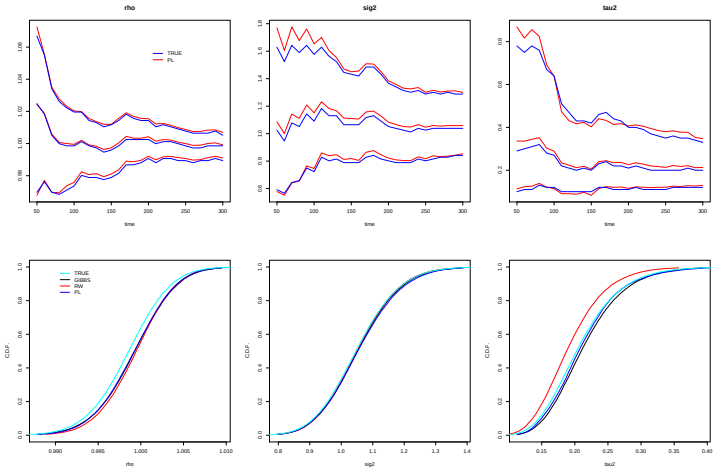


Example iii. $p(x_t|y^n)$ via MCMC



Posterior medians and 95% credibility intervals.

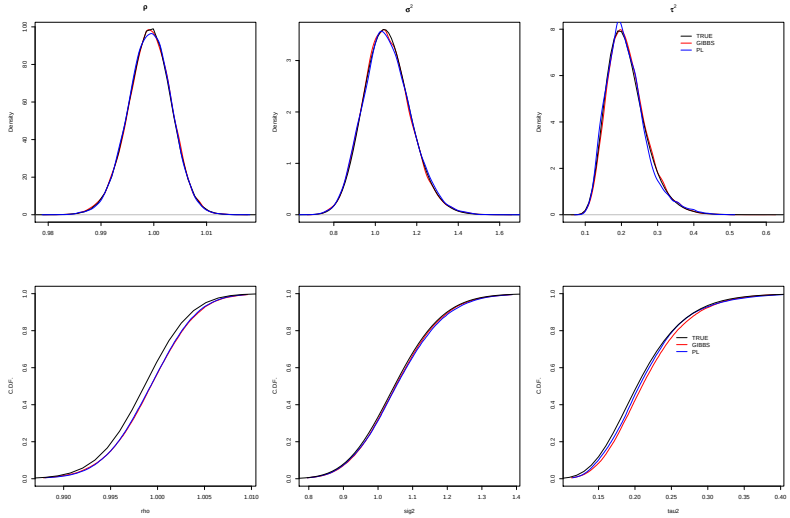
Example iii. PL and MCMC quantiles



TOP: True and PL estimate of percentiles of $p(\theta|y^t)$ for all t. Percentiles are 2.5%,50% and 97.5%.

BOTTOM: True, GIBBS and PL estimates of $F(\theta|y^n)$.

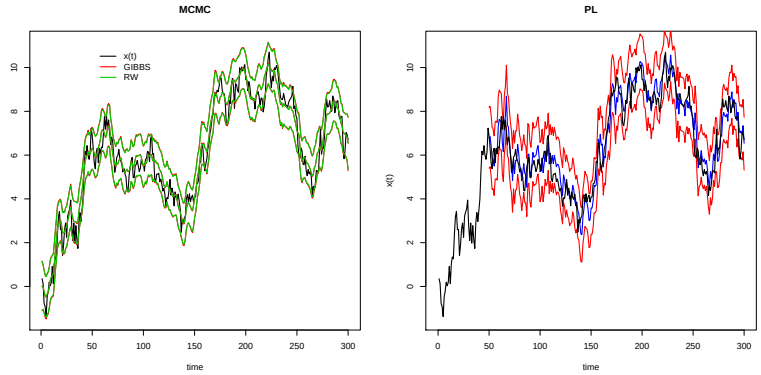
Example iii. PL and GIBBS quantiles



TOP: True, Gibbs and PL estimate of $p(\theta|y^n)$.

BOTTOM: True, Gibbs and PL estimates of $F(\theta|y^n)$.

Example iii. PL and MCMC quantiles

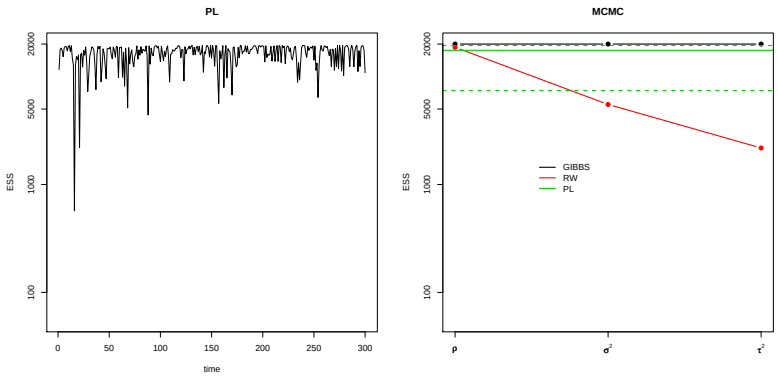


LEFT: Posterior medians and 95% credibility intervals of $p(x_t|y^n)$.

RIGHT: Posterior medians and 95% credibility intervals of $p(x_t|y^t)$.

Example iii. Effective sample sizes

$$ESS_t = M\left(1 + \frac{V(\omega_t)}{E^2(\omega_t)}\right)^{-1} \quad \text{and} \quad ESS = M\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)^{-1}.$$



ω_t are particle weights.

$\rho_k = cov_\pi(t^{(n)}, t^{(n+k)})/var_\pi(t^{(n)})$ and $t^{(n)} = t(\theta^{(n)})$.

Example iv. Dynamic factor model with switching loadings

For $t = 1, \ldots, T$, the model is defined as follows [2]:

- Observation equation

$$y_t|z_t, \theta \sim N(\gamma_t x_t, \sigma^2 I_2)$$

- State equations

$$\begin{aligned} x_t|x_{t-1}, \theta &\sim N(x_{t-1}, \sigma_x^2) \\ \lambda_t|\lambda_{t-1}, \theta &\sim Ber((1-p)^{1-\lambda_{t-1}} q^{\lambda_{t-1}}) \end{aligned}$$

where $z_t = (x_t, \lambda_t)'$, $\gamma_t = (1, \beta_{\lambda_t})'$ is the vector of time-varying loadings and $\theta = (\beta_1, \beta_2, \sigma^2, \sigma_x^2, p, q)'$ is the vector of fixed parameters.

The prior distributions are conditionally conjugate:

$$\begin{aligned} (\beta_i|\sigma^2) &\sim N\left(b_{i0}, \sigma^2 B_{i0}\right) \qquad \text{for } i = 1, 2, \\ \sigma^2 &\sim IG\left(\frac{\nu_{00}}{2}, \frac{d_{00}}{2}\right) \\ \sigma_x^2 &\sim IG\left(\frac{\nu_{10}}{2}, \frac{d_{10}}{2}\right) \\ p &\sim Beta(p_1, p_2) \\ q &\sim Beta(q_1, q_2) \\ x_0 &\sim N(m_0, C_0) \end{aligned}$$

Particle representation

At time t, particles

$$\left\{(x_t, \lambda_t, \theta, s_t^x, s_t)^{(i)}\right\}_{i=1}^N$$

approximating

$$p\left(x_t, \lambda_t, \theta, s_t^x, s_t|y^t\right)$$

where

[2] This example is from Carvalho, Johannes, Lopes and Polson (2009)

- $s_t^x = \mathcal{S}(s_{t-1}^x, \theta)$ are state sufficient statistics
- $s_t = \mathcal{S}(s_{t-1}, x_t, \lambda_t)$ are fixed parameter sufficient statistics

Re-sampling $(x_t, \lambda_t, \theta, s_t^x, s_t)$

Let us redefine $\beta_i = (1, \beta_i)'$ whenever necessary.

Draw an index $k(i) \sim Multi(\omega^{(i)})$ with weights

$$\omega^{(i)} \propto p(y_{t+1}|(s_t^x, \lambda_t, \theta)^{k(i)})$$

with

$$p(y_{t+1}|m_t, C_t, \lambda_t, \theta) = \sum_{j=1}^{2} f_N\left(y_{t+1}; \beta_j m_t, V_j\right) Pr\left(\lambda_{t+1} = j|\lambda_t, \theta\right)$$

where $V_j = (C_t + \sigma_x^2)\beta_j\beta_j' + \sigma^2 I_2$, m_t and C_t are components of s_t^x and f_N denotes the normal density function.

Propagating states

- Draw auxiliary state λ_{t+1}

$$\lambda_{t+1}^{(i)} \sim p(\lambda_{t+1}|(s_t^x, \lambda_t, \theta)^{k(i)}, y_{t+1})$$

where

$$Pr(\lambda_{t+1} = j|s_t^x, \lambda_t, \theta, y_{t+1}) \propto f_N\left(y_{t+1}; \beta_j m_t, V_j\right) p\left(\lambda_{t+1} = j|\lambda_t, \theta\right).$$

- Draw state x_{t+1} conditionally on λ_{t+1}

$$x_{t+1}^{(i)} \sim p(x_{t+1}|\lambda_{t+1}^{(i)}, (s_t^x, \theta)^{k(i)}, y_{t+1})$$

by a simply Kalman filter update.

Updating sufficient statistics for states, s_{t+1}^x

The Kalman filter recursion yield

$$m_{t+1} = m_t + A_{t+1}(y_{t+1} - \beta_{\lambda_{t+1}} m_t)$$
$$C_{t+1} = C_t + \sigma_x^2 - A_{t+1} Q_{t+1}^{-1} A_{t+1}'$$

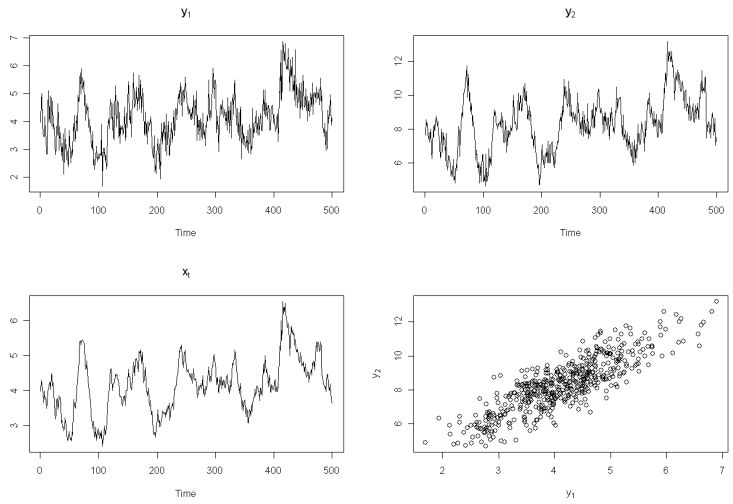
where

$$Q_{t+1} = (C_t + \sigma_x^2)\gamma_{t+1}\gamma_{t+1}' + \sigma^2 I_2$$
$$A_{t+1} = (C_t + \sigma_x^2)\gamma_{t+1}' Q_{t+1}^{-1}$$

Updating sufficient statistics for parameters, s_{t+1}

Recall that $s_{t+1} = \mathcal{S}(s_t, x_{t+1}, \lambda_{t+1})$. Then,

$$\begin{aligned}
(\beta_i|\sigma^2, s_{t+1}) &\sim N\left(b_{i,t+1}, \sigma^2 B_{i,t+1}\right) \qquad \text{for } i = 1, 2, \\
(\sigma^2|s_{t+1}) &\sim IG\left(\frac{\nu_{0t}}{2}, \frac{d_{0,t+1}}{2}\right) \\
(\sigma_x^2|s_{t+1}) &\sim IG\left(\frac{\nu_{1t}}{2}, \frac{d_{1,t+1}}{2}\right) \\
(p|s_{t+1}) &\sim Beta(p_{1,t+1}, p_{2,t+1}) \\
(q|s_{t+1}) &\sim Beta(q_{1,t+1}, q_{2,t+1})
\end{aligned}$$



where $I_{\lambda_{t+1}=i} = I_i$, $I_{\lambda_t=i,\lambda_{t+1}=j} = I_{ij}$, $\nu_{it} = \nu_{i,t-1} + 1$, $B_{i,t+1}^{-1} = B_{it}^{-1} + x_{t+1}^2$, $B_{i,t+1}^{-1} b_{i,t+1} = B_{it}^{-1} b_{it} + x_{t+1} y_{t+1,2} I_i$, $p_{i,t+1} = p_{it} + I_{1i}$ (similarly for $q_{i,t+1}$) for $i = 1, 2$, $d_{0,t+1} = d_{0,t} + (y_{t+1,1} - x_{t+1})^2 + \sum_{j=1}^{2} \left[(y_{t+1,2} - b_{j,t+1} x_{t+1}) \, y_{t+1,2} + B_{j,t+1}^{-1} b_{j,t+1} \right] I_j$, and $d_{1,t+1} = d_{1,t} + (x_{t+1} - x_t)^2$.

CASE I: Randowm walk dynamic factor, static loadings

- Simulation setup:

$$\begin{aligned} n &= 500 \\ \beta &= 2 \\ \sigma^2 &= 0.2 \\ \sigma_x^2 &= 0.05 \end{aligned}$$

- Prior hyperparameters:

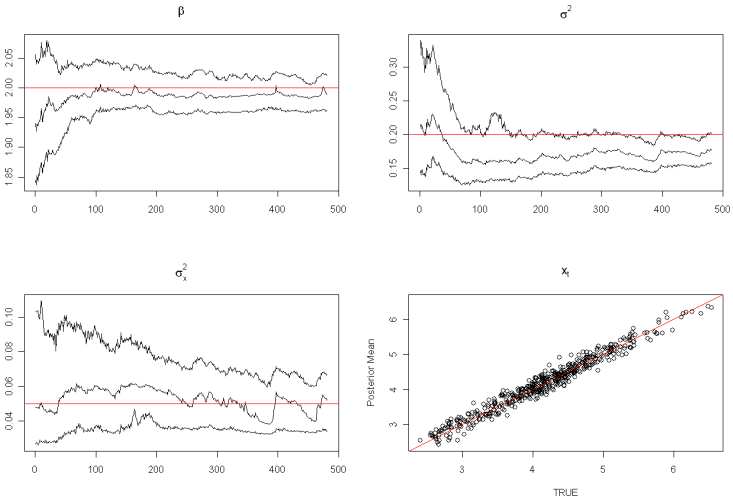
$$\begin{aligned} \beta &: \; b_0 = 0 \;\; B_0 = 10 \\ \sigma^2 &: \; \nu_{00} = 10 \;\; d_{00} = 1.8 \\ \sigma_x^2 &: \; \nu_{10} = 10 \;\; d_{10} = 0.45 \end{aligned}$$

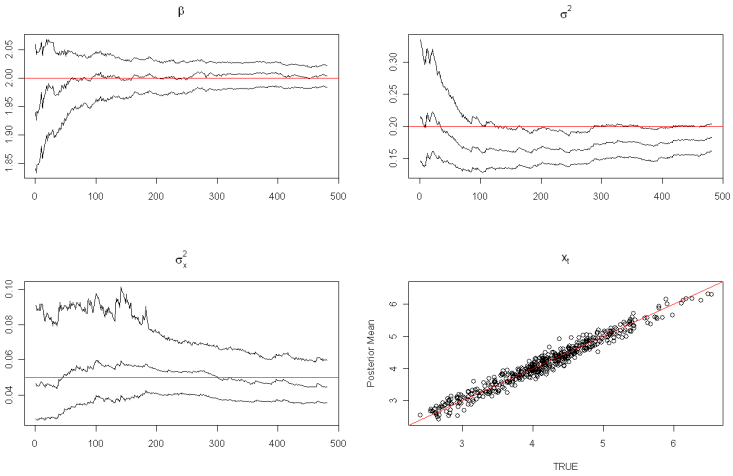
Time series and dynamic factor

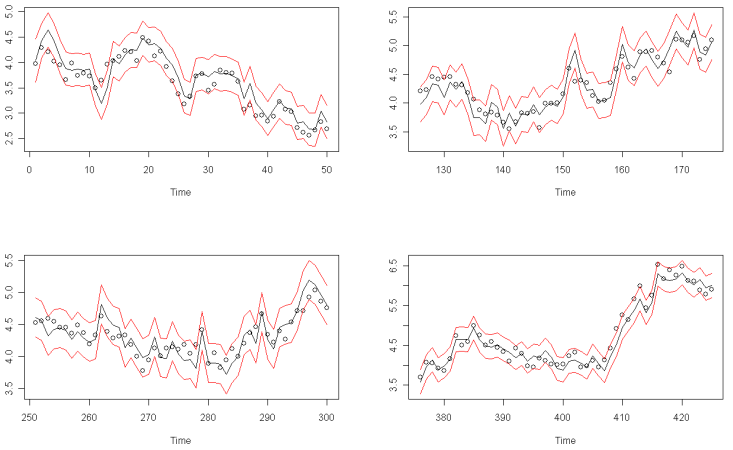
Parameter learning: $M = 1000$

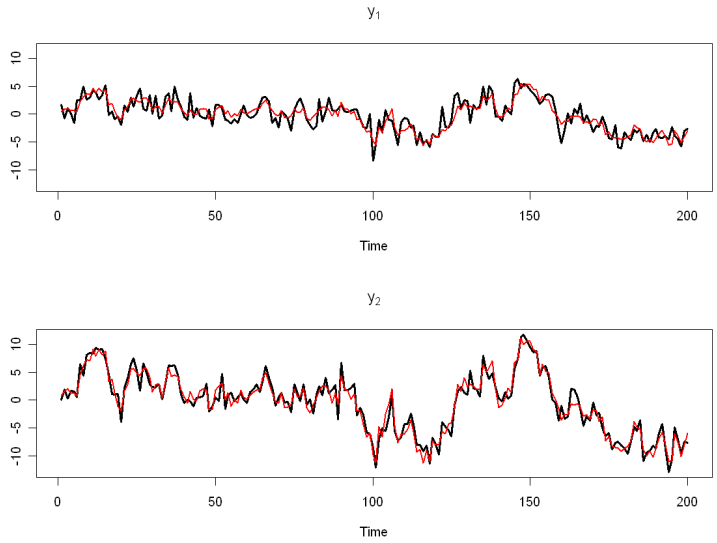
Parameter learning: $M = 5000$

Dynamic factor


β
σ²
σ²ₓ
xₜ
Posterior Mean
TRUE


β
σ²
σ²ₓ
xₜ
Posterior Mean
TRUE


Time
Time
Time
Time



CASE II: AR(1) dynamic factor, static loadings

$$\begin{aligned} x_t|x_{t-1},\theta &\sim N(\rho x_{t-1}, 1.0) \\ \rho &\sim N(\rho_0, V_\rho) \end{aligned}$$

- Simulation setup:

$$\begin{aligned} n &= 200 \\ \beta &= 2 \\ \sigma^2 &= 2.0 \\ \rho &= 9 \end{aligned}$$

- Prior hyperparameters:

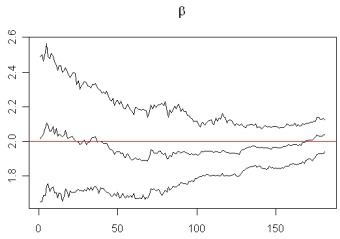
$$\begin{aligned} \beta &: \ b_0 = 0 \ \ B_0 = 100 \\ \sigma^2 &: \ \nu_{00} = 10 \ \ d_{00} = 18 \\ \rho &: \ \rho_0 = 1 \ \ V_\rho = 100 \end{aligned}$$

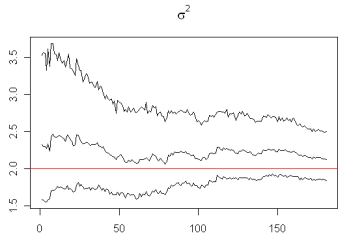
Time series and state variables

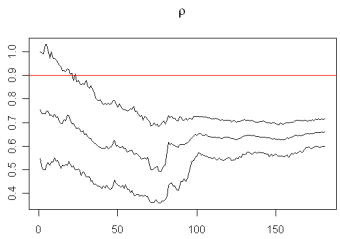
Parameter learning: $M = 1000$ **particles**

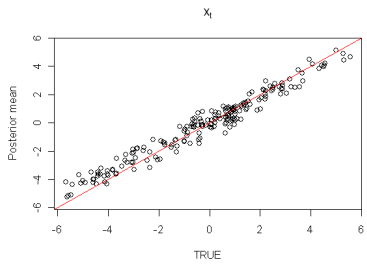
Dynamic factor

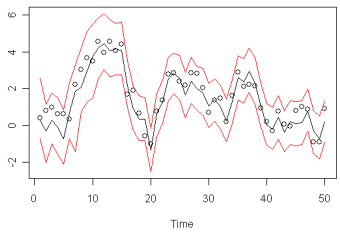
CASE III: Random walk dynamic factor, time-varying loadings

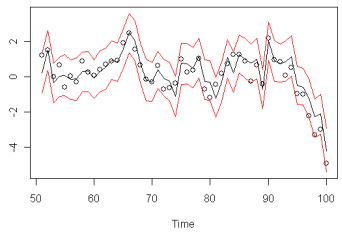

β
1.8
2.0
2.2
2.4
2.6
0
50
100
150

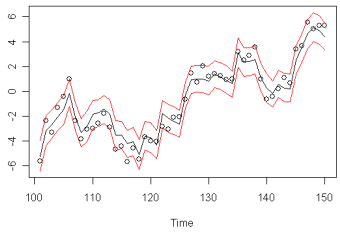

σ2
1.5
2.0
2.5
3.0
3.5
0
50
100
150

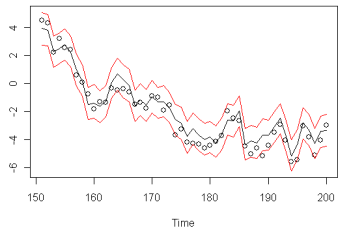

ρ
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
50
100
150

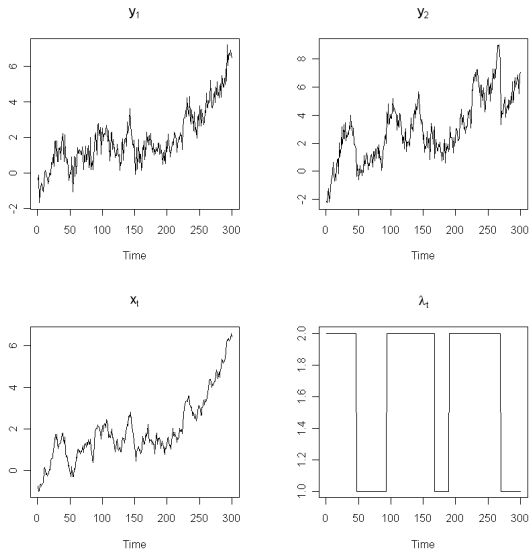

xt
Posterior mean
-6
-4
-2
0
2
4
6
TRUE


-2
0
2
4
6
0
10
20
30
40
50
Time


-4
-2
0
2
50
60
70
80
90
100
Time


-6
-4
-2
0
2
4
6
100
110
120
130
140
150
Time


-6
-4
-2
0
2
4
150
160
170
180
190
200
Time



- Simulation setup:

$$\begin{aligned} n &= 300 \\ (\beta_1, \beta_2) &= (1, 2) \\ \sigma^2 &= 0.2 \\ \sigma_x^2 &= 0.05 \\ p &= q = 0.975 \end{aligned}$$

- Prior hyperparameters:

$$\begin{aligned} \beta_1 &: b_{10} = 0 \;\; B_{10} = 2 \\ \beta_2 &: b_{20} = 3 \;\; B_{20} = 2 \\ \sigma^2 &: \nu_{00} = 5 \;\; d_{00} = 1.0 \\ \sigma_x^2 &: \nu_{10} = 5 \;\; d_{10} = 0.25 \\ p, q &: p_1 = p_2 = q_1 = q_2 = 1 \end{aligned}$$

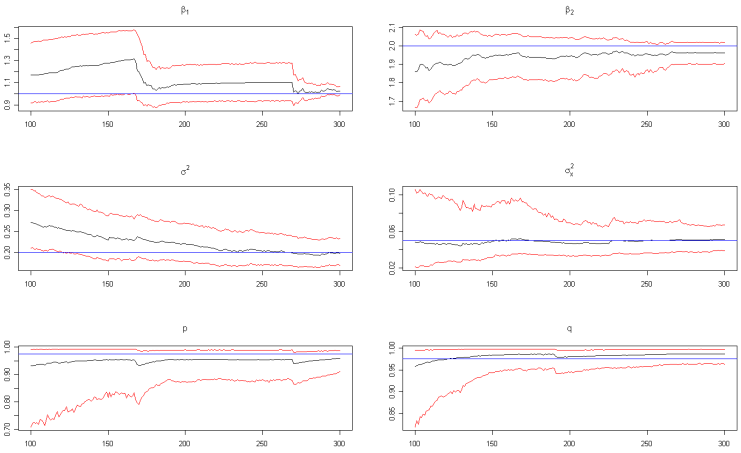
Time series and state variables

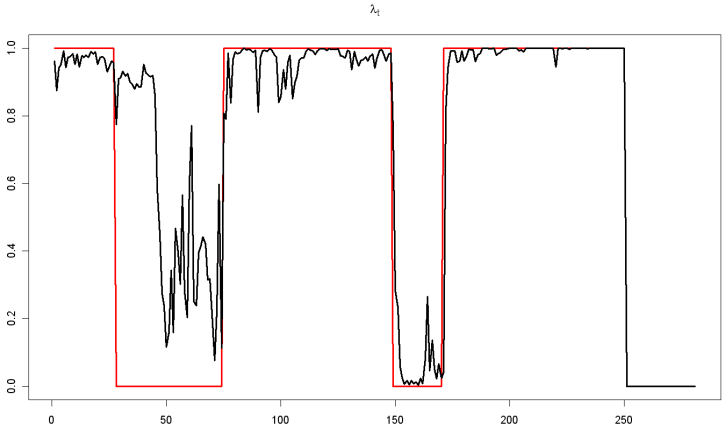
Parameter learning: $M = 5000$ **particles**

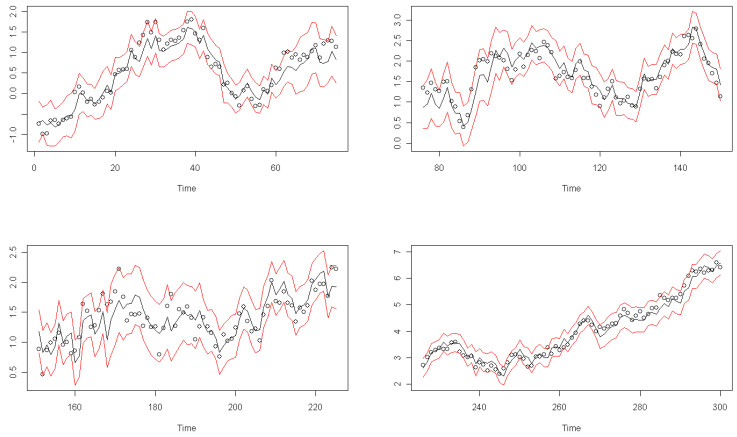
Discrete switching state

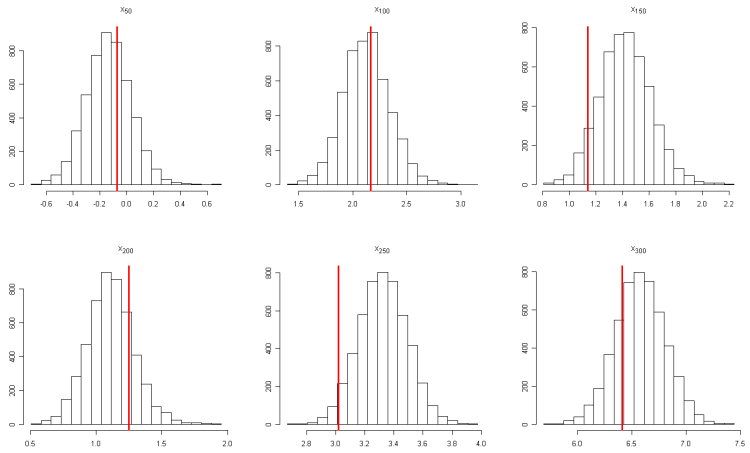
Dynamic factor

Dynamic factor


β_1
β_2
σ^2
σ_x^2
p
q


λ_t


Time
Time
Time
Time



LECTURE 10

STOCHASTIC VOLATILITY via SEQUENTIAL MONTE CARLO METHODS

Example i: Stochastic volatility

Let y_t, for $t = 1, \ldots, n$, be modeled as

$$\begin{aligned} y_t|x_t &\sim N(0, e^{x_t}) \\ (x_t|x_{t-1}, \theta) &\sim N(\alpha + \beta x_{t-1}, \tau^2) \end{aligned}$$

where $\theta = (\alpha, \phi, \tau^2)$.

Simulation setup: $n = 500$, $\alpha = -0.0031$, $\beta = 0.9951$ and $\tau^2 = 0.0074$ and $x_1 = \alpha/(1-\beta) = -0.632653$ (13% of annualized standard deviation).

Prior setup:

$$\begin{aligned} x_0 &\sim N(m_0, C_0) \qquad \alpha \sim N(\alpha_0, V_\alpha) \\ \beta &\sim N(\beta_0, V_\beta) \qquad \tau^2 \sim IG(n_0/2, n_0\tau_0^2/2) \end{aligned}$$

where $m_0 = 0.0$, $C_0 = 0.1$, $\alpha_0 = -0.0031$, $V_\alpha = 0.01$, $\beta_0 = 0.9951$, $V_\beta = 0.01$, $n_0 = 3$ and $\tau_0^2 = 0.0074$.

LW filter with shrinkage factor a

Particles t: $\{(x_t, \theta)^{(j)}, \omega_t^{(j)}\}_{j=1}^M \sim p(x_t, \theta|y^t)$.

Summary of $p(\theta|y^t)$: $\bar{\theta} \approx E(\theta|y^t)$ and $V \approx V(\theta|y^t)$.

Resample quantities: For $j = 1, \ldots, M$

- Compute $m^{(j)} = a\theta^{(j)} + (1-a)\bar{\theta}$
- Compute $g^{(j)} = \alpha^{(j)} + \phi^{(j)} x_t^{(j)}$

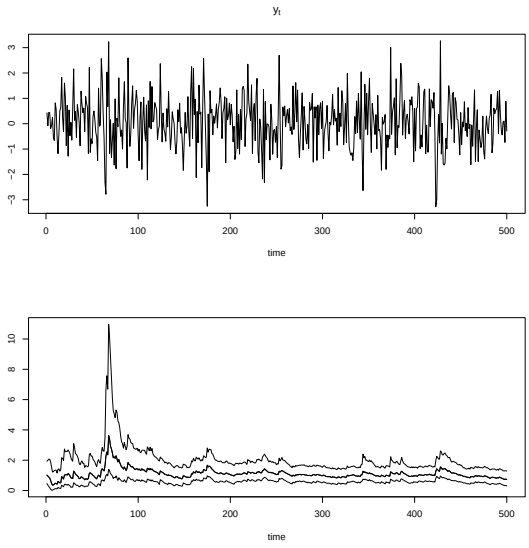
Algorithm: For $l = 1, \ldots, M$

- Draw $k^l \in \{1, \ldots, M\}$, with $P(k^l = j) \propto \omega_t^{(j)} p(y_{t+1}|g^{(j)})$
- Sample $\theta^{(l)}$ from $N(m^{(k^l)}, (1-a^2)V)$
- Sample $x_{t+1}^{(l)}$ from $p(x_{t+1}|x_t^{(k^l)}, \theta^{(l)})$
- Compute weight $\omega_{t+1}^{(l)} \propto p(y_{t+1}|x_{t+1}^{(l)}) / p(y_{t+1}|g^{(k^l)})$

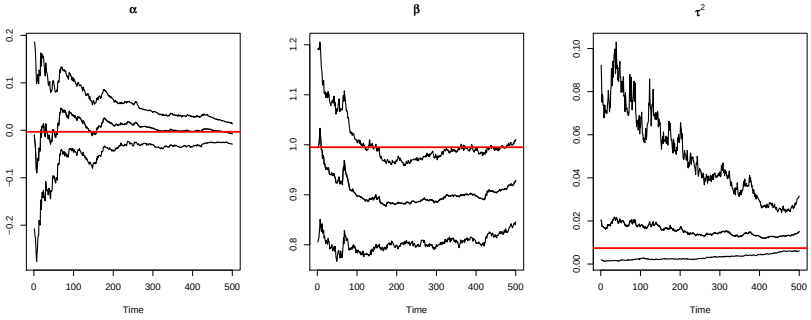
Particles at $t+1$: $\{(x_{t+1}, \theta)^{(j)}, \omega_{t+1}^{(j)}\}_{j=1}^{M} \sim p(x_{t+1}, \theta|y^{t+1})$.

Time series y_t and $p(e^{x_t}|y^t)$

$N = 5000$ and $\theta = (\alpha, \beta, \log(\tau^2))$.



Parameter learning



Example ii: SV-AR(1) via sequential MCMC and LW

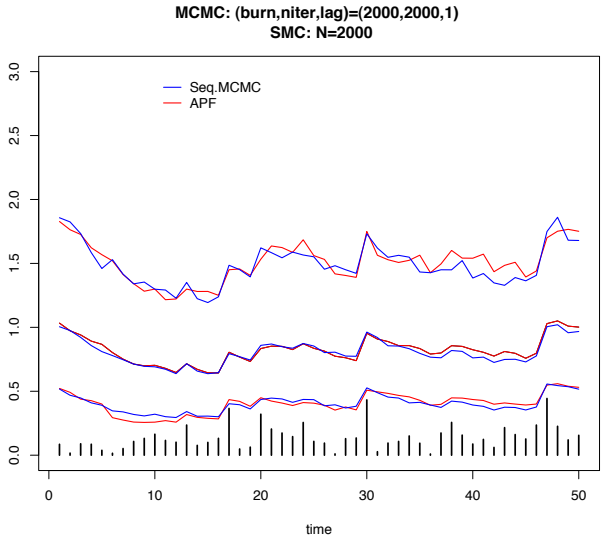
We simulated $n = 50$ observations based on $\alpha = -0.0031$, $\beta = 0.9951$, $\tau^2 = 0.0074$, with $m_0 = 0.0$ and $C_0 = 0.1$.

Also, $x_1 = \alpha/(1-\beta) = -0.632653$, which corresponds to annualized standard deviations around 13%.

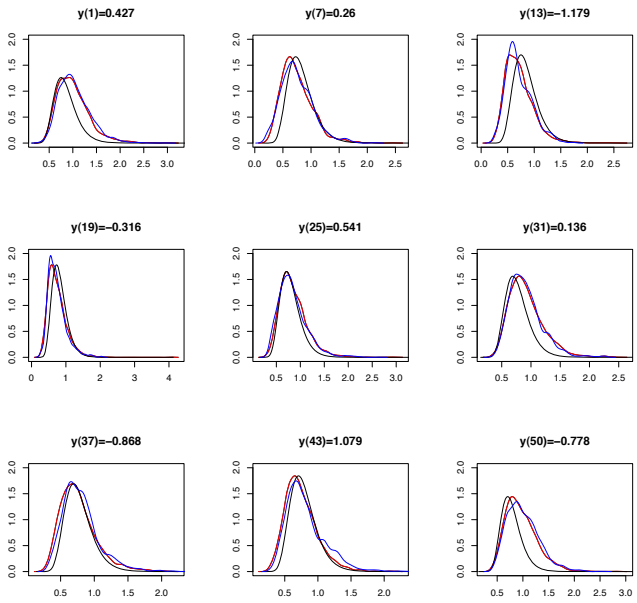
$p(x_t|y^t)$ when θ is known

MCMC: Kim, Shephard and Chib (1994)

SMC: Liu and West (2001) with $\delta = 0.75$ and $a = 0.9521743$.

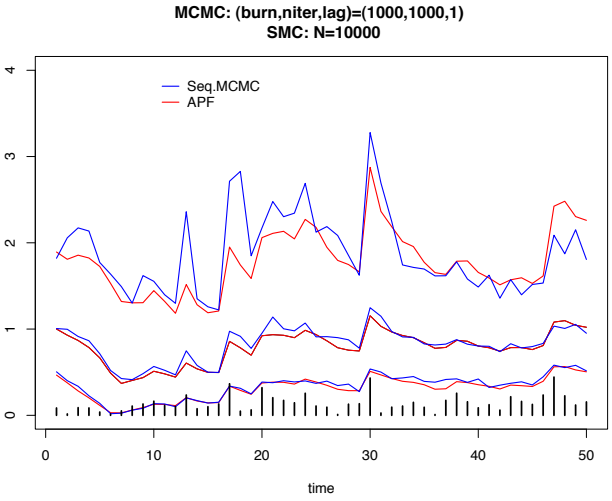


$p(x_t|y^t)$ when θ is known

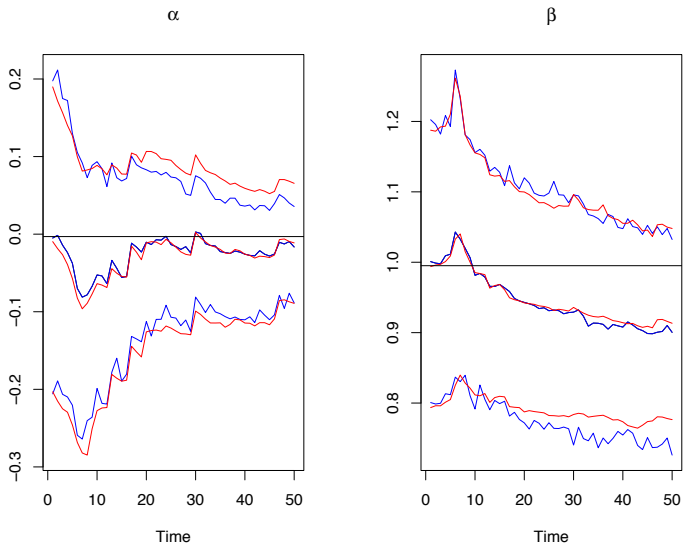


$p(x_t|y^t)$ **when** (α, β) **is unknown**

Prior: $\alpha \sim N(-0.0031, 0.01)$ and $\phi \sim N(0.9951, 0.01)$

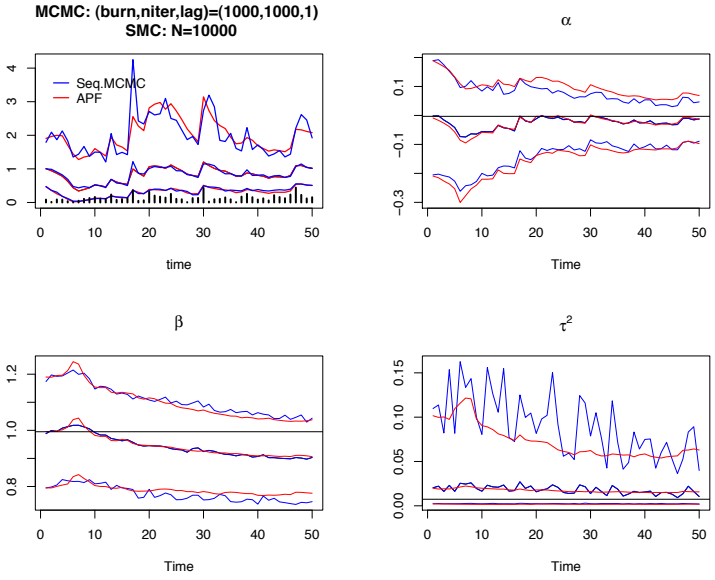


$p(\alpha|y^t)$ **and** $p(\beta|y^t)$

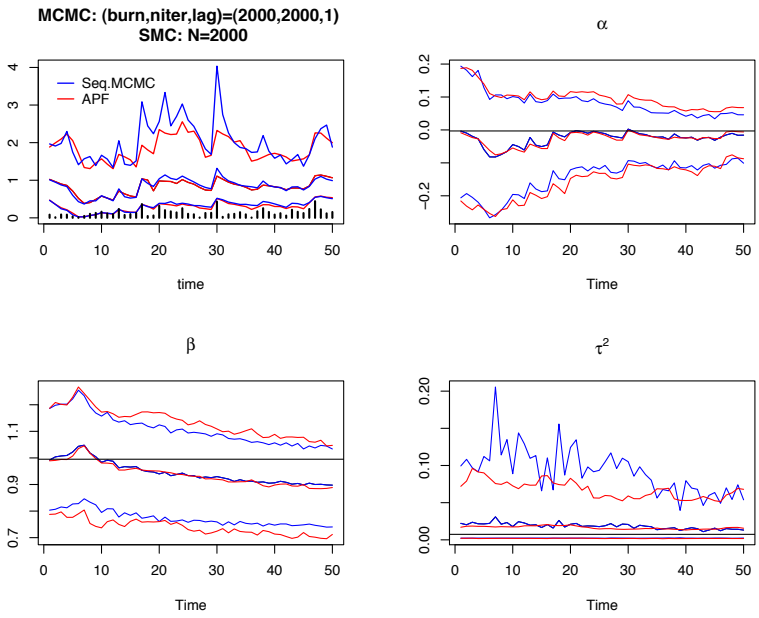


Learning x_t, α, β and τ^2

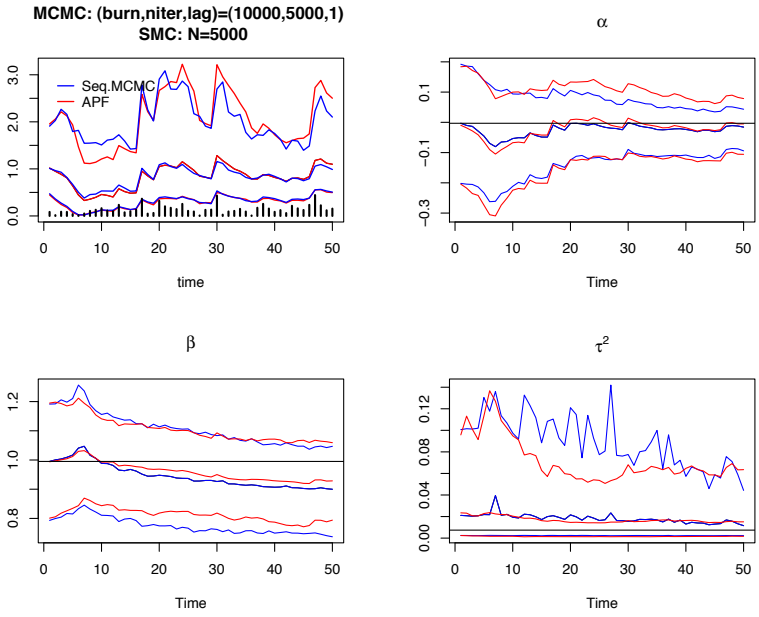
The prior for τ^2 is $IG(1.5, 0.0111)$.



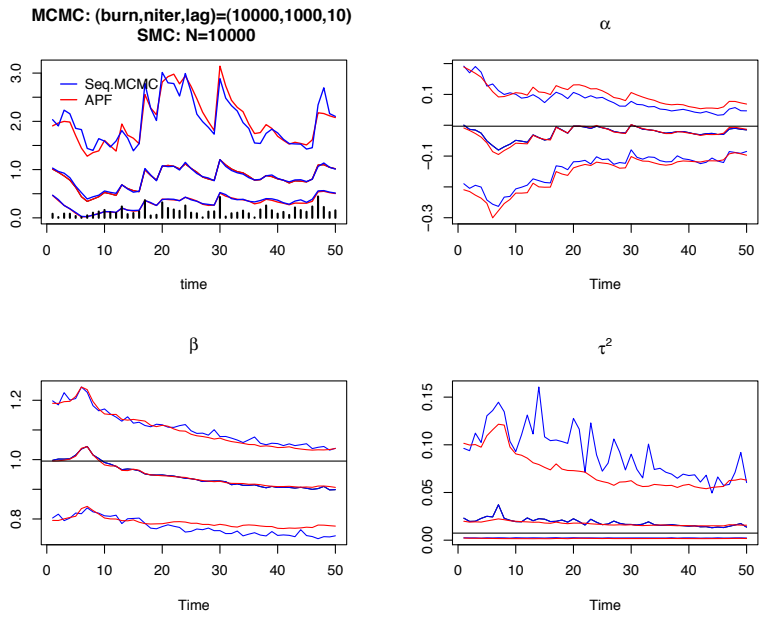
Learning x_t, α, β and τ^2



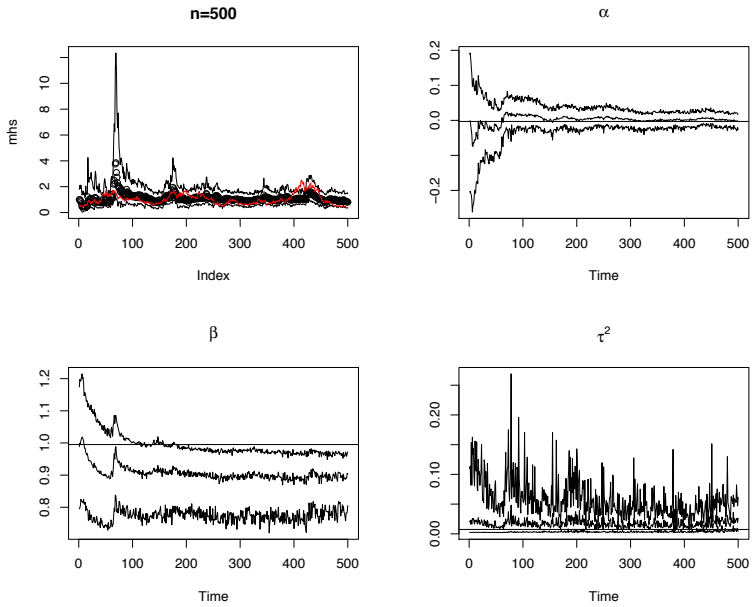
Learning x_t, α, β and τ^2



Learning x_t, α, β and τ^2

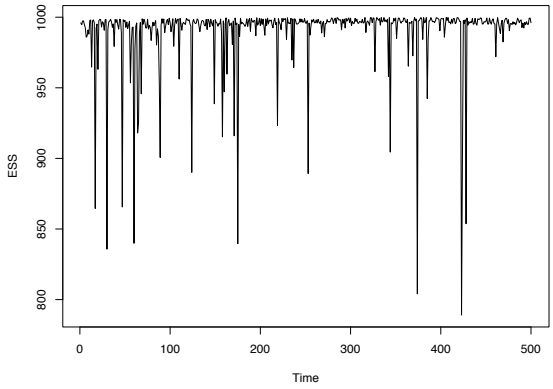


Sequential MCMC when $n = 500$

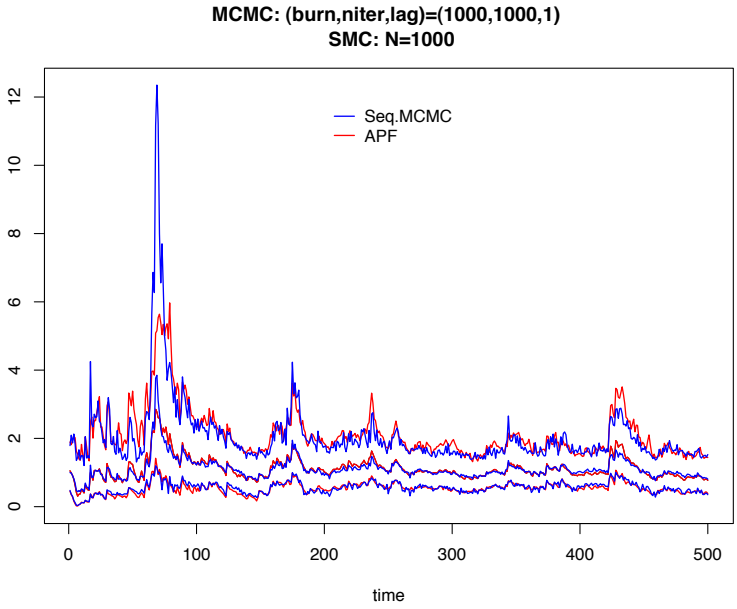


Effective sample size

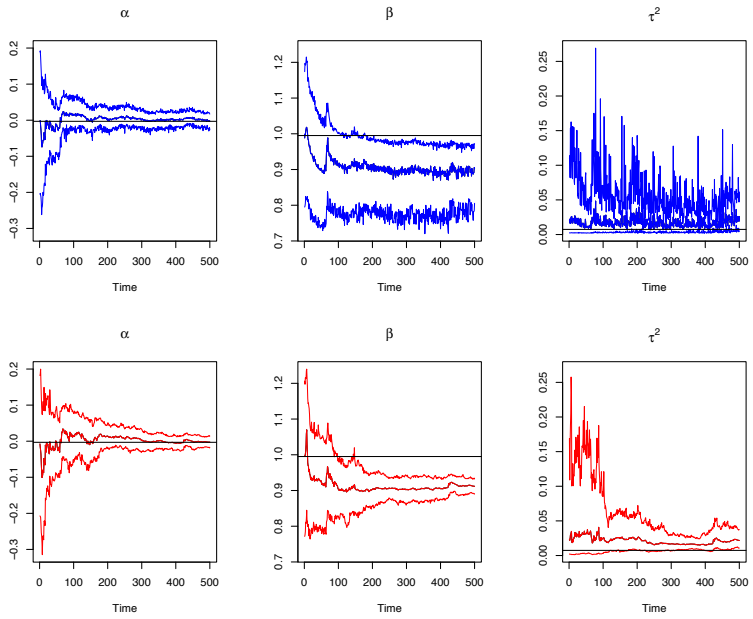
$$ESS_t = \frac{N}{1 + \frac{V(\omega_t)}{E^2(\omega_t)}}$$



Sequential volatilities



Sequential parameter learning



Example iii: Markov switching stochastic volatility

Carvalho and Lopes (2007) adapted APF and LW filters to sequentially estimate states and parameters in Markov switching stochastic volatility (MSSV) models.

Let the daily returns of the IBOVESPA index, y_t, be modeled by a MSSV model, ie.

$$\begin{aligned} y_t|\lambda_t &\sim N(0, \exp(\lambda_t)) \\ (\lambda_t|\lambda_{t-1}, \xi, s_t) &\sim N(\alpha_{s_t} + \phi\lambda_{t-1}, \sigma^2) \end{aligned}$$

where $\xi = (\alpha, \phi, \sigma^2)$, $\alpha = (\alpha_1, \ldots, \alpha_k)$ and regime variables s_t following a k-state first order Markov process,

$$p_{ij} = Pr(s_t = j|s_{t-1} = i) \qquad \text{for } i, j = 1, \ldots, k$$

and $P = (p_{11}, \ldots, p_{1k-1}, \ldots, p_{k1}, \ldots, p_{k,k-1})$.

Particle filter

- *Step 0:* $\left\{\lambda_t^{(j)}, s_t^{(j)}, w_t^{(j)}\right\}_{j=1}^{M} \sim p(\lambda_t, s_t, \theta|D_t)$
- *Step 1:* For $j = 1, \ldots, M$,

$$\begin{aligned} \tilde{s}_{t+1}^{(j)} &= \arg\max_{l \in 1,\ldots,k} Pr(s_{t+1} = l|s_t = s_t^{(j)}) \\ \mu_{t+1}^{(j)} &= \alpha_{\tilde{s}_{t+1}^{(j)}}^{(j)} + \phi_t^{(j)} \lambda_t^{(j)} \end{aligned}$$

- *Step 2:* For $l = 1, \ldots, M$

1. Sample k^l from $\{1, \ldots, k\}$, with $Pr(k^l) \propto p(y_{t+1}|\mu_{t+1}^{(k^l)}) w_t^{(k^l)}$
2. Sample $\theta_{t+1}^{(l)}$ from $N(m_t^{(k^l)}, b^2 V_t)$

Jul 2nd, 97	Thailand devalues the baht by as much as 20%.
Aug 11th, 97	IMF and Thailand set a rescue agreement.
Oct 23rd, 97	Hong Kong's stock index falls 10.4%. South Korea Won weakens.
Dec 2nd, 97	IMF and South Korea set a bailout agreement.
Jun 1st, 98	Russia's stock market crashes.
Jun 20th, 98	IMF gives final approval to a loan package to Russia.
Aug 19th, 98	Russia officially falls into default.
Oct 09th, 98	IMF and World Bank joint meeting + Fed cuts interest rates.
Jan 15th, 99	The real is allowed to float freely by lifting exchange controls.
Feb 2nd, 99	Arminio Fraga is named president of Brazil's Central Bank.

3. Sample $s_{t+1}^{(l)}$ from $1, \ldots, k$ with $Pr(s_{t+1}^{(l)}) = Pr(s_{t+1}^{(l)} | s_t^{(k^l)})$
4. Sample $\lambda_{t+1}^{(l)}$ from $p(\lambda_{t+1} | \lambda_t^{(k^l)}, s_{t+1}^{(l)}, \theta_{t+1}^{(l)})$

- *Step 3:* For $l = 1, \ldots, M$, compute new weights

$$w_{t+1}^{(l)} \propto p(y_{t+1} | \lambda_{t+1}^{(l)}) / p(y_{t+1} | \mu_{t+1}^{(k^l)})$$

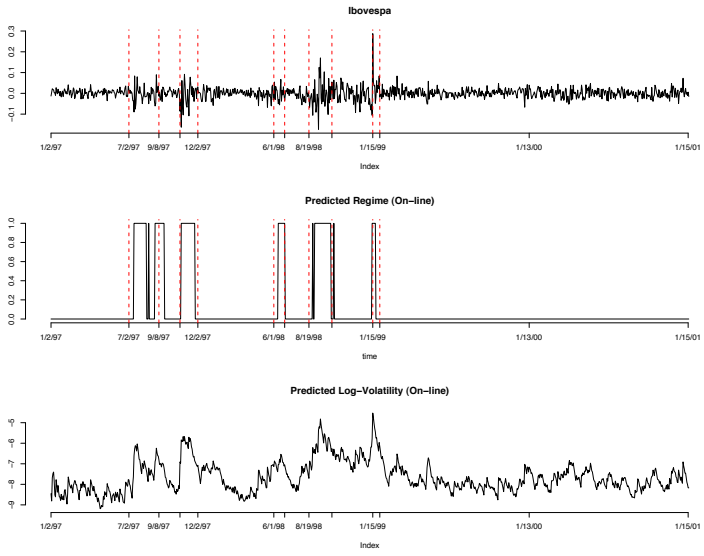
- *Step 4:* $\left\{ \lambda_{t+1}^{(j)}, S_{t+1}^{(j)}, w_{t+1}^{(j)} \right\}_{j=1}^{M} \sim p(\lambda_{t+1}, S_{t+1}, \theta | D_{t+1})$.

Currency crisis

Carvalho and Lopes (2007) used IBOVESPA daily data from January 2nd, 1997 to January, 16th 2001 (1000 observations).

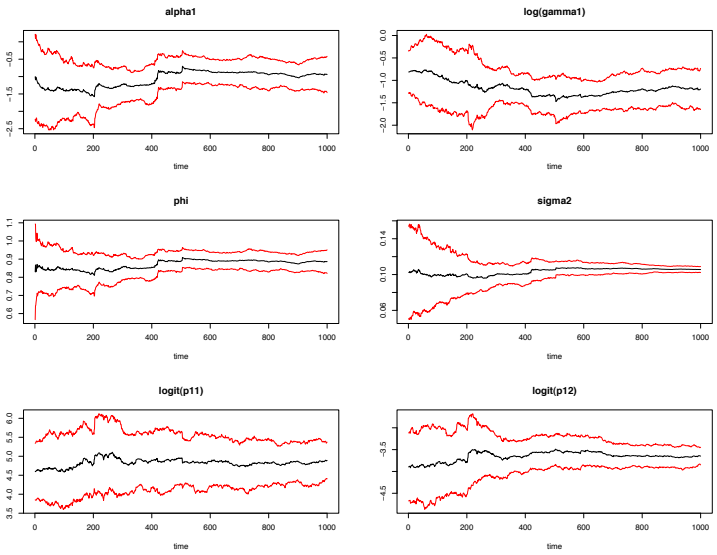
Fitting regime shifts

The vertical lines indicate key market events.

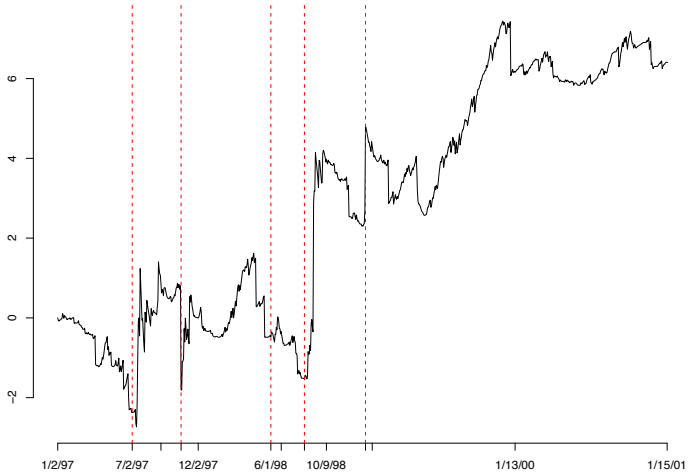


Sequential inference for fixed parameters

Posterior mean, 5% and 95% quantiles of θ.



Sequential Bayes factor: MSSV vs SV



SV-AR(1) via Particle Learning

This example was kindly prepared by my PhD student Samir Warty.

Recall the AR(1) stochastic volatility model:

$$y_{t+1} = \exp\left(\frac{x_{t+1}}{2}\right)\epsilon_{t+1}$$

$$x_{t+1} = \alpha + \beta x_t + \tau\nu_{t+1}$$

where $(\epsilon_t, \nu_t) \sim N(0_2, I_2)$ and $\theta = (\alpha, \beta, \tau)$.

Prior distribution:

$$\tau^2 | s_0 \sim \mathcal{IG}\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right)$$
$$\alpha, \beta | \tau^2, s_0 \sim N(m_0, \tau^2 C_0)$$

where $s_0 = (n_0, S_0, m_0, C_0)$.

Data augmentation argument

Following Kim, Shephard and Chib's (1998) idea:

$$z_{t+1} \equiv \log y_{t+1}^2 = x_{t+1} + \log \epsilon_{t+1}^2 \approx x_{t+1} + u_t$$

where

$$u_t \sim \sum_{i=1}^{7} \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$$

Particle learning (PL) uses augments the state vector to include $\lambda_{t+1} \in \{1, \ldots, 7\}$, the component of the Normal mixture approximation.

Let s_t denote the set of sufficient statistics for (α, β, τ^2) at time t.

Algorithm

- Resample old particles $c_t = (x_t, s_t, \theta)$ with weights

$$w_t \propto p(z_{t+1} | c_t) = \sum_{i=1}^{7} \pi_i f_N(z_{t+1}; \mu_i + \alpha + \beta x_t, \sigma_i^2 + \tau^2)$$

- Propagate new states x_{t+1} from

$$p(x_{t+1} | c_t, z_{t+1}) = \sum_{i=1}^{7} \pi_i f_N(x_{t+1}; \gamma_i, \omega_i)$$

where

$$\begin{aligned} \omega_i &= (\sigma_i^{-2} + \tau^{-2})^{-1} \\ \gamma_i &= \omega_i(\sigma_i^{-2}(z_{t+1} - \mu_i) + \tau^{-2}(\alpha + \beta x_t)) \end{aligned}$$

Algorithm (cont.)

- Update sufficient statistics $s_{t+1} = (n_{t+1}, S_{t+1}, m_{t+1}, C_{t+1})$

$$\begin{aligned} n_{t+1} &= n_t + 1 \\ n_{t+1} S_{t+1} &= n_t S_t + \frac{(x_{t+1} - X_t m_t)^2}{1 + X_t C_t X_t'} \\ C_{t+1}^{-1} &= C_t^{-1} + X_t' X_t \\ C_{t+1}^{-1} m_{t+1} &= C_t^{-1} m_t + X_t' x_{t+1} \end{aligned}$$

where $X_t = (1, x_t)$.

- Sample parameters

$$\tau^2|s_t \sim \mathcal{IG}\left(\frac{n_{t+1}}{2}, \frac{n_{t+1}S_{t+1}}{2}\right)$$
$$\alpha, \beta|\tau^2, s_t \sim \mathcal{N}(m_{t+1}, \tau^2 C_{t+1})$$

Resampling weights

$$\begin{aligned}
w_t &\propto p(z_{t+1}|c_t, \lambda_t, z^t) \\
&\propto \sum_{i=1}^{7} \int_R p(z_{t+1}|x_t, s_t, \theta, \lambda_{t+1} = i, \lambda_t, z^t, x_{t+1}) p(x_{t+1}|x_t, s_t, \theta, \lambda_{t+1} = i, \lambda_t) dx_{t+1} \\
&\qquad \text{(Marginalization over data augmentation)} \\
&\propto \sum_{i=1}^{7} \int_R p(z_{t+1}|\lambda_{t+1} = i, x_{t+1}) p(x_{t+1}|x_t, \theta) dx_{t+1} \qquad \text{(Conditional independence)} \\
&\propto \sum_{i=1}^{7} \int_R f_N(z_{t+1}; \mu_i + x_{t+1}, \sigma_i^2) f_N(x_{t+1}; \alpha + \beta x_t, \tau^2) dx_{t+1} \\
&\propto \sum_{i=1}^{7} \pi_i f_N(z_{t+1}; \mu_i + \alpha + \beta x_t, \sigma_i^2 + \tau^2)
\end{aligned}$$

Posterior distribution for new states

$$\begin{aligned}
p(x_{t+1}|c_t, \lambda_t, z_{t+1}) &= \sum_{i=1}^{7} \pi_i p(x_{t+1}|c_t, \lambda_{t+1} = i, \lambda_t, z_{t+1}) \qquad \text{(Marginalization over data augmentation)} \\
&= \sum_{i=1}^{7} \pi_i \frac{p(z_{t+1}|x_t, s_t, \theta, \lambda_{t+1} = i, \lambda_t, x_{t+1}) p(x_{t+1}|x_t, s_t, \theta, \lambda_{t+1} = i, \lambda_t)}{p(z_{t+1}|x_t, s_t, \theta, \lambda_{t+1} = i, \lambda_t)} \qquad \text{(Bayes theorem)} \\
&= \sum_{i=1}^{7} \pi_i \frac{p(z_{t+1}|\lambda_{t+1} = i, x_{t+1}) p(x_{t+1}|x_t, \theta)}{p(z_{t+1}|x_t, \theta, \lambda_{t+1} = i)} \qquad \text{(Conditional independence)} \\
&= \sum_{i=1}^{7} \pi_i \frac{f_N(z_{t+1}; \mu_i + x_{t+1}, \sigma_i^2) f_N(x_{t+1}; \alpha + \beta x_t, \tau^2)}{f_N(z_{t+1}; \mu_i + \alpha + \beta x_t, \sigma_i^2 + \tau^2)} \\
&= \sum_{i=1}^{7} \pi_i f_N(x_{t+1}; \gamma_i, \omega_i)
\end{aligned}$$

where $\omega_i = (\sigma_i^{-2} + \tau^{-2})^{-1}$ and $\gamma_i = \omega_i(\sigma_i^{-2}(z_{t+1} - \mu_i) + \tau^{-2}(\alpha + \beta x_t))$.

Recursive sufficient statistics

$$
\begin{aligned}
n_{t+1}S_{t+1} &= n_tS_t + (x_{t+1} - X_t(C_t^{-1} + X_t'X_t)^{-1}(C_t^{-1}m_t + X_t'x_{t+1}))'x_{t+1} \\
&\quad + (m_t - (C_t^{-1} + X_t'X_t)^{-1}(C_t^{-1}m_t + X_t'x_{t+1}))'C_t^{-1}m_t \\
&= n_tS_t + (x_{t+1}'x_{t+1} - x_{t+1}'X_t(C_t^{-1} + X_t'X_t)^{-1}(C_t^{-1}m_t + X_t'x_{t+1})) \\
&\quad + (m_t'(C_t^{-1})'m_t - m_t'(C_t^{-1})'(C_t^{-1} + X_t'X_t)^{-1}(C_t^{-1}m_t + X_t'x_{t+1})) \\
&= n_tS_t + (x_{t+1}'x_{t+1} - x_{t+1}'X_t\left(C_t - \frac{C_tX_t'X_tC_t}{1 + X_tC_tX_t'}\right)(C_t^{-1}m_t + X_t'x_{t+1})) \\
&\quad + (m_t'C_t^{-1}m_t - m_t'C_t^{-1}\left(C_t - \frac{C_tX_t'X_tC_t}{1 + X_tC_tX_t'}\right)(C_t^{-1}m_t + X_t'x_{t+1})) \\
&= n_tS_t + x_{t+1}'\left(1 - c + \frac{c^2}{1+c}\right)x_{t+1} - x_{t+1}'\left(1 - \frac{c}{1+c}\right)X_tm_t \\
&\quad - m_t'X_t'\left(1 - \frac{c}{1+c}\right)x_{t+1} + \frac{m_t'X_t'X_tm_t}{1+c} \qquad \text{(where } c \equiv X_tC_tX_t'\text{)} \\
&= n_tS_t + \left(\frac{1}{1+c}\right)(x_{t+1}'x_{t+1} - 2x_{t+1}'X_tm_t + m_t'X_t'X_tm_t) \\
&= n_tS_t + \left(\frac{1}{1+c}\right)(x_{t+1} - X_tm_t)'(x_{t+1} - X_tm_t)
\end{aligned}
$$

where

$$
(C_t^{-1} + X_t'X_t)^{-1} = \left(C_t - \frac{C_tX_t'X_tC_t}{1 + X_tC_tX_t'}\right)
$$

when X_t is a vector.

A few references: MC and MCMC

1. **Casella and George (1992) Explaining the Gibbs sampler. The American Statistician,46,167-74.**

2. **Chib and Greenberg (1995) Understanding the Metropolis-Hastings algorithm. The American Statistician, 49, 327-35.**

3. Gelfand and Smith (1990) Sampling-Based Approaches to Calculating Marginal Densities, JASA, 85, 398-409.

4. Geman and Geman (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-41.

5. Geweke (1989) Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57, 1317-39.

6. Gilks and Wild (1992) Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337-48. Computational Statistics and Data Analysis, 51, 4526-42.

7. Hastings (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57, 97-109.

8. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) Equation of State Calculations by Fast Computing Machines. Journal of Chemical Physics, Number 21, 1087-92.

9. **Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling-resampling perspective. American Statistician, 46, 84-8.**

A few references: Sequential Monte Carlo methods

1. **Carvalho, Johannes, Lopes and Polson (2008) Particle learning and smoothing. University of Chicago Graduate School of Business.**

2. Carvalho and Lopes (2007) Simulation-based sequential analysis of Markov switching stochastic volatility models,Computational Statistics and Data Analysis, 51, 4526-4542.

3. Doucet, de Freitas, and Gordon, 2001, Sequential Monte Carlo Methods in Practice, Springer, New York.

4. Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. Journal of Computational and Graphical Statistics, 11, 848-862.

5. **Gordon, Salmond, and Smith, 1993, Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation, IEE Proceedings, F-140, 107-113.**

6. Liu and West (2001). Combined parameters and state estimation in simulation-based filtering. In Sequential Monte Carlo Methods in Practice. Springer-Verlag New York.

7. **Pitt and Shephard (1999) Filtering via Simulation: Auxiliary Particle Filter, Journal of the American Statistical Association, 94, 590-599.**

8. Polson, Stroud and Muller (2008). Practical filtering with sequential parameter learning. Journal of the Royal Statistical Society, Series B, 70, 413-428.

9. **Storvik (2002). Particle filters in state space models with the presence of unknown static parameters, IEEE. Trans. of Signal Processing, 50, 281-289.**

A few references: Dynamic models

1. **Carlin, Polson and Stoffer (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. Journal of the American Statistical Association, 87, 493-500.**

2. Carter and kohn (1994) On Gibbs sampling for state space models, Biometrika, 81, 541-553.

3. Fruhwirth-Schnatter (1994) Data augmentation and dynamic linear models, Journal of Time Series Analysis, 15, 183-102.

4. **Migon, Gamerman, Lopes and Ferreira (2005) Dynamic models. In Dey and Rao (Eds.) Handbook of Statistics, Volume 25.**

5. West and Harrison (1989/1997) Bayesian Forecasting and Dynamic Models. New York: Springer-Verlag.

A few references: Stochastic volatility models

1. Berg, Meyer and Yu (2004), Deviance Information Criterion for Comparing Stochastic Volatility Models, Journal of Business and Economic Statistics, 22, 107-20.

2. **Eraker, Johannes and Polson (2003) The Impact of Jumps in Volatility and Returns, Journal of Finance, 2003, 58, 1269-300.**

3. Jacquier, Polson and Rossi (1994) Bayesian analysis of stochastic volatility models. Journal of Business and Economic Statistics, 12, 371-415.

4. Jensen and Maheu (2008) Bayesian semiparametric stochastic volatility modeling. Working paper 2008-15, Federal Reserve Bank of Atlanta.

5. **Johannes and Polson (2006) MCMC Methods for Financial Econometrics, Handbook of Financial Econometrics, Yacine Aït-Sahalia and Lars Hansen.**

6. **Kim, Shephard and Chib (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models, Review of economic studies, 65, 36193.**

7. Lopes and Polson (2009) Extracting SP500 and NASDAQ volatility: The credit crisis of 2007-2008, Handbook of Applied Bayesian Analysis.

8. Lopes and Carvalho (2007) Factor stochastic volatility with time varying loadings and Markov switching regimes, Journal of Statistical Planning and Inference, 37, 3082-3091.

9. Lopes and Salazar (2006) Time series mean level and stochastic volatility modeling by smooth transition autoregressions: a Bayesian approach, In Fomby, T.B. (Ed.) Advances in Econometrics: Econometric Analysis of Financial and Economic Time Series/Part B, 2006, Volume 20, 229-242.

10. Polson, Stroud and Muller (2008) Practical Filtering with Sequential Parameter Learning, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 70, pp. 413-28.