# Lecture 9 State Space Models

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP598 Spatiotemporal Analysis
Fall 2021

# Introduction to State Space Models

- It was introduced by Kalman (1960) and Kalman and Bucy (1961) in the space tracking setting.
- It consists of a state equation for the motion of $x_t$ and an observation equation for $y_t$ based on the location $x_t$.
- The most basic special case is linear (state) Gaussian (observation) model, a.k.a. dynamic linear model.
- It has been applied to modeling data from economics, medicine, climate modeling, and engineering.

# Introduction to State Space Models



In this lecture, we focus primarily on linear Gaussian state space models. We will

- present various forms of the model
- introduce the concepts of prediction, filtering and smoothing
- perform maximum likelihood estimation
- present special topics such as hidden Markov models (HMM), stochastic volatility
- discuss a Bayesian approach

# Introduction to State Space Models

- In general, there are two principles assumed in state space models:
  1. The state process $\{x_t\}$ is Markovian, i.e. the future $x_s : s > t$ and the past $x_s : s < t$ are independent conditioned on the present $x_t$;
  2. The observations $\{y_t\}$ are independent given the states $x_t$.

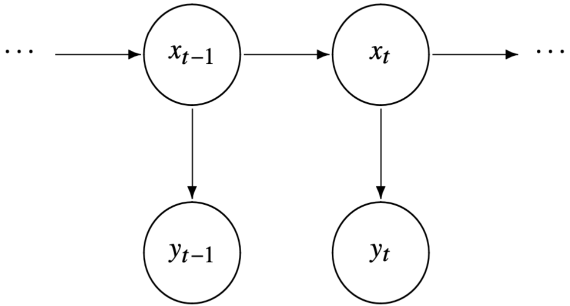



***Fig. 6.1.*** *Diagram of a state space model.*

# Table of Contents

# Linear Gaussian Model

- The liner Gaussian model is the most fundamental state-space model which takes the following form:

$$x_t = \Phi x_{t-1} + w_t, \quad w_t \stackrel{iid}{\sim} N_p(0, Q) \tag{1}$$

$$y_t = A_t x_t + v_t, \quad v_t \stackrel{iid}{\sim} N_q(0, R) \tag{2}$$

where we assume $x_0 \sim N_p(\mu_0, \Sigma_0)$, and $A_t$ is a $q \times p$ measurement/observation matrix.

# Linear Gaussian Model

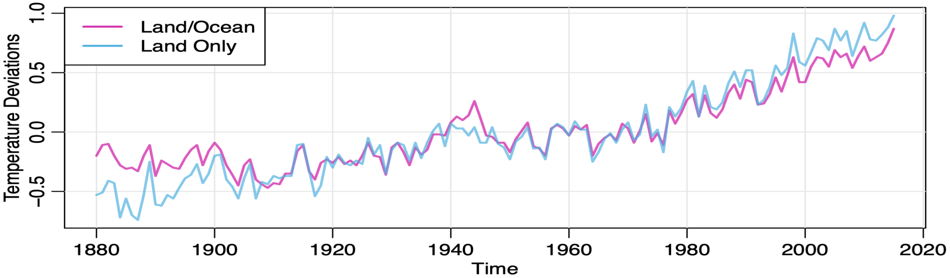



***Fig. 6.3.** Annual global temperature deviation series, measured in degrees centigrade, 1880–2015. The series differ by whether or not ocean data is included.*

- Two time series of temperature $\{y_{t1}, y_{t2}\}$ are modeled as a linear Gaussian state-space model

$$\begin{bmatrix} y_{t_1} \\ y_{t_2} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x_t + \begin{bmatrix} v_{t_1} \\ v_{t_2} \end{bmatrix}, \quad R = \mathrm{Var}\begin{bmatrix} v_{t_1} \\ v_{t_2} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \tag{3}$$
$$x_t = \delta + x_{t-1} + w_t, \quad Q = \mathrm{Var}(w_t)$$

# Linear Gaussian Model

- Let's consider a univariate state-space model

$$\begin{aligned} y_t &= x_t + v_t, \quad v_t \overset{iid}{\sim} N(0, \sigma_v^2) \\ x_t &= \phi x_{t-1} + w_t, \quad w_t \overset{iid}{\sim} N(0, \sigma_w^2) \end{aligned} \tag{4}$$

where $x_0 \sim N(0, \sigma_w^2/(1-\phi^2))$, and $x_0$, $\{v_t\}$, $\{w_t\}$ are independent.

- For AR(1) process, we could compute the autocovariance function of $x_t$

$$\gamma_x(h) = \frac{\sigma_w^2}{1-\phi^2}\phi^h, \quad h = 0, 1, 2, \cdots \tag{5}$$

- Based on the observation equation and independence assumption, we could compute the autocovariance function of $y_t$

$$\gamma_y(h) = \mathrm{Cov}(x_t + v_t, x_{t-h}, +v_{t-h}) = \begin{cases} \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2, & h = 0 \\ \gamma_x(h), & h \geq 1 \end{cases} \tag{6}$$

# Linear Gaussian Model

- Therefore, the ACF of the observations $y_t$ is

$$\rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_v^2}{\sigma_w^2}(1-\phi^2)\right)^{-1} \phi^h \tag{7}$$

- Note, this is NOT the autocorrelation structure of AR(1) unless $\sigma_v^2 = 0$.
- In general, this is ACF of ARMA(1,1), if we identify the following model

$$y_t = \phi y_{t-1} + \theta u_{t-1} + u_t \tag{8}$$

with $u_t \overset{iid}{\sim} N(0, \sigma_u^2)$ for $\sigma_u^2$, $\theta$ to be determined.

# Table of Contents

# Filtering, Smoothing and Forecasting

- The objective of the analysis involving the state-space model is to produce estimators for the underlying unobserved (latent) process $x_t$, given the data $y_{1:s} = \{y_1, \cdots, y_s\}$ up to time $s$.
- Based on the relationship between $t$ and $s$, the problems can be divided into 3 types:
  1. *smoothing*: $t < s$
  2. *filtering*: $t = s$
  3. *forecasting or prediction*: $t > s$
- We fix some notations

$$x_t^s = \mathrm{E}(x_t | y_{1:s}), \quad P_{t_1,t_2}^s = \mathrm{E}\{(x_{t_1} - x_{t_1}^s)(x_{t_2} - x_{t_2}^s)'\} \tag{9}$$

# The Kalman Filter



- First, we present the Kalman filter for the filtering and forecasting problems.
- Consider a linear filter of $x_t^t$ in terms of $y_{1:t}$, $x_t = \sum_{s=1}^t B_s y_s$, for the state-space equation with initial condition $x_0^0 = \mu_0$ and $P_0^0 = \Sigma_0$

$$x_t = \Phi x_{t-1} + \Upsilon u_t + w_t, \quad w_t \overset{iid}{\sim} N_p(0, Q) \tag{10}$$

$$y_t = A_t x_t + \Gamma u_t + v_t, \quad v_t \overset{iid}{\sim} N_q(0, R) \tag{11}$$

- The Kalman filter consists of the following two steps:
  1. *predict*:

$$x_t^{t-1} = \Phi x_{t-1}^{t-1} + \Upsilon u_t, \quad P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q \tag{12}$$

  2. *update*:

$$x_t^t = x_t^{t-1} + K_t(y_t - A_t x_t^{t-1} - \Gamma u_t), \quad P_t^t = [I - K_t A_t] P_t^{t-1} \tag{13}$$

  where $K_t = P_t^{t-1} A_t' [A_t P_t^{t-1} A_t' + R]^{-1}$ is called the *Kalman gain*.

# The Kalman Filter

- The prediction for $t > n$ can be obtained from the *predict* step with initial conditions $x_n^n$ and $P_n^n$.
- The innovations (prediction errors) are

$$\epsilon_t = y_t - \mathrm{E}(y_t | y_{1:t-1}) = y_t - A_t x_t^{t-1} - \Gamma u_t \tag{14}$$

$$\Sigma_t = \mathrm{Var}(\epsilon_t) = \mathrm{Var}[A_t(x_t - x_t^{t-1}) + v_t] = A_t P_t^{t-1} A_t' + R \tag{15}$$

- What are the possible issues?

# The Kalman Filter (density point of view)

- Let's take a density point of view. The two assumptions (Markovianity and conditional independence) translate into

$$p_\theta(x_t|x_{0:t-1}) = p_\theta(x_t|x_{t-1}), \quad p_\theta(y_{1:n}|x_{1:n}) = \prod_{t=1}^{n} p_\theta(y_t|x_t) \tag{16}$$

- For the linear Gaussian model, we have

$$p_\theta(x_t|x_{t-1}) = \phi(x_t; \Phi x_{t-1}, Q), \quad p_\theta(y_t|x_t) = \phi(y_t; A_t x_t, R) \tag{17}$$

# The Kalman Filter (density point of view)



- For the *prediction* step, we could obtain

$$\begin{aligned} p_\theta(x_t|y_{1:t-1}) &= \int_{\mathbb{R}^p} p_\theta(x_t, x_{t-1}|y_{1:t-1}) dx_{t-1} \\ &= \int_{\mathbb{R}^p} \phi(x_t; \Phi x_{t-1}, Q)\phi(x_{t-1}; x_{t-1}^{t-1}, P_{t-1}^{t-1}) dx_{t-1} \\ &= \phi(x_t; x_t^{t-1}, P_t^{t-1}) \end{aligned} \tag{18}$$

- For the *update* step, we have

$$\begin{aligned} p_\theta(x_t|y_{1:t}) &= p_\theta(x_t|y_t, y_{1:t-1}) \propto p_\theta(y_t|x_t)p_\theta(x_t|y_{1:t-1}) \\ &= \phi(y_t; A_t x_t, R)\phi(x_t; x_t^{t-1}, P_t^{t-1}) \\ &\propto \phi(x_t; x_t^t, P_t^t) \end{aligned} \tag{19}$$

# The Kalman Smoother



- For the state-space model (10)(11) with initial conditions $x_n^n$ and $P_n^n$, the Kalman Smoother is

$$x_{t-1}^n = x_{t-1}^{t-1} + J_{t-1}(x_t^n - x_t^{t-1}) \tag{20}$$

$$P_t^n = P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1})J_{t-1}' \tag{21}$$

  for $t = n, n-1, \cdots, 1$, where $J_{t-1} = P_{t-1}^{t-1}\Phi'[P_t^{t-1}]^{-1}$.

- In the following, let's investigate a simple univariate state-space model, the *local level model*.

$$\begin{aligned} y_t &= \mu_t + v_t, \quad v_t \overset{iid}{\sim} N(0, \sigma_v^2) \\ \mu_t &= \mu_{t-1} + w_t, \quad w_t \overset{iid}{\sim} N(0, \sigma_w^2) \end{aligned} \tag{22}$$

# The Local Level Model

- Using the density argument or formulae, we obtain the prediction

$$\mu_t | y_{1:t-1} \sim N(\mu_t^{t-1}, P_t^{t-1}) \tag{23}$$

$$\mu_t^{t-1} = u_{t-1}^{t-1}, \quad P_t^{t-1} = P_{t-1}^{t-1} + \sigma_w^2 \tag{24}$$

- And we can obtain the update (filter)

$$\mu_t | y_{1:t} \sim N(\mu_t^t, P_t^t) \tag{25}$$

$$\mu_t^t = \mu_t^{t-1} + K_t(y_t - \mu_t^{t-1}), \quad K_t = \frac{P_t^{t-1}}{P_t^{t-1} + \sigma_v^2}, \quad P_t^t = (1 - K_t)P_t^{t-1} \tag{26}$$

- We note that $P_t^{t-1} \geq P_t^t \geq P_t^n$ for $t = n, n-1, \cdots, 1$.

# The Local Level Model

- We could use R function `Ksmooth0` (which calls `Kfilter0` for the filtering) to obtain prediction $x_t^{t-1}$ (`xp`), filter $x_t^t$ (`xf`) and smooth $x_t^n$ (`xs`) respectively.
- Their error bounds are $x_t^{t-1} \pm 2\sqrt{P_t^{t-1}}$ (`Pp`), $x_t^t \pm 2\sqrt{P_t^t}$ (`Pf`), and $x_t^n \pm 2\sqrt{P_t^n}$ (`Ps`) respectively.

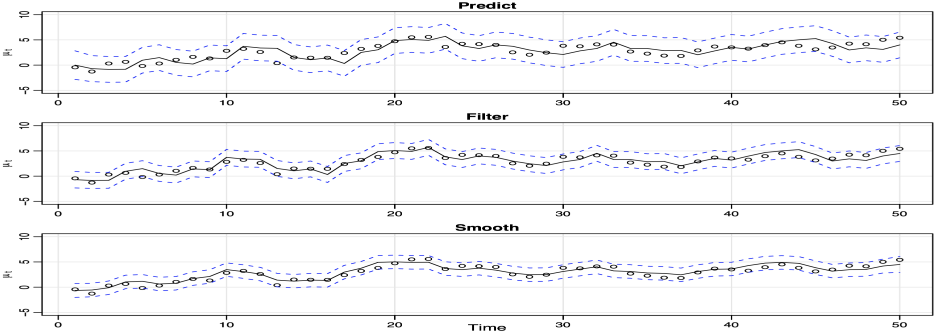



***Fig. 6.4.*** *Displays for Example 6.5. The simulated values of $\mu_t$, for $t = 1, \ldots, 50$, given by (6.51) are shown as points. The top shows the predictions $\mu_t^{t-1}$ as a line with $\pm 2\sqrt{P_t^{t-1}}$ error bounds as dashed lines. The middle is similar, showing $\mu_t^t \pm 2\sqrt{P_t^t}$. The bottom shows $\mu_t^n \pm 2\sqrt{P_t^n}$.*

# Table of Contents

# Maximum Likelihood Estimation



- Recall the state space model with initial condition $x_0^0 = \mu_0$ and $P_0^0 = \Sigma_0$

$$x_t = \Phi x_{t-1} + \Upsilon u_t + w_t, \quad w_t \overset{iid}{\sim} N_p(0, Q)$$
$$y_t = A_t x_t + \Gamma u_t + v_t, \quad v_t \overset{iid}{\sim} N_q(0, R)$$

- The estimation of parameters $\Theta = \{\mu_0, \Sigma_0, \Phi, Q, R, \Upsilon, \Gamma\}$ is quite involved.
- The likelihood is computed using the *innovations* $\{\epsilon_t\}_{t=1}^n$ and their convariances

$$\epsilon_t = y_t - A_t x_t^{t-1} - \Gamma u_t, \quad \Sigma_t = A_t P_t^{t-1} A_t' + R \tag{27}$$

- Then the likelihood, $L_Y(\Theta)$, can be written (up to a constant)

$$-\log L_Y(\Theta) = \frac{1}{2}\sum_{t=1}^n \log|\Sigma_t(\Theta)| + \frac{1}{2}\sum_{t=1}^n \epsilon_t(\Theta)'\Sigma_t(\Theta)^{-1}\epsilon_t(\Theta) \tag{28}$$

# Maximum Likelihood Estimation

We could perform a Newton-Raphson estimation procedure:

1. Select initial values for the parameters, say, $\Theta^{(0)}$.
2. Run the Kalman filter, using the initial parameter values, to obtain a set of innovations and error covariances $\{\epsilon_t^{(0)}\}_{t=1}^n$ and $\{\Sigma_t^{(0)}\}_{t=1}^n$.
3. Run one iteration of a Newton-Raphson optimization on $-\log L_Y(\Theta)$ to obtain a new set of estimates $\Theta^{(1)}$.
4. Iterate between (2) and (3) for step $j$ to obtain $\{\epsilon_t^{(j)}\}_{t=1}^n$ and $\{\Sigma_t^{(j)}\}_{t=1}^n$ until the estimates or the likelihood stabilize, i.e. their consecutive difference fall below some threshold.

# Maximum Likelihood Estimation

- Recall univariate state-space model

$$y_t = x_t + v_t, \quad v_t \overset{iid}{\sim} N(0, \sigma_v^2)$$
$$x_t = \phi x_{t-1} + w_t, \quad w_t \overset{iid}{\sim} N(0, \sigma_w^2)$$

where $x_0 \sim N(0, \sigma_w^2/(1-\phi^2))$, and $x_0$, $\{v_t\}$, $\{w_t\}$ are independent.

- We generate $y_{1:n}$ for $n = 100$ with true values $\phi = 0.8$, $\sigma_w^2 = \sigma_v^2 = 1$.
- Now we use Newton-Raphson procedure (`optim` function) to obtain the MLE of $(\phi, \sigma_w^2, \sigma_v^2)$ given data.
- To start, we initialize parameters using what we calculated before

$$\phi^{(0)} = \frac{\hat{\rho}_y(2)}{\hat{\rho}_y(1)}, \quad (\sigma_w^2)^{(0)} = \frac{1-(\phi^{(0)})^2}{\phi^{(0)}}\hat{\gamma}_y(1), \quad (\sigma_w^2)^{(0)} = \hat{\gamma}_y(0) - \frac{(\sigma_w^2)^{(0)}}{1-(\phi^{(0)})^2} \tag{29}$$

- See more details in R codes.

# Table of Contents

# Hidden Markov Models



- *Hidden Markov models (HMM)* were developed parallel to state-space models in Goldfeld and Quandt (1973) and Lindgren (1978). A good summary is Rabiner and Juang (1986).
- Recall the first principle of state-space model is latent process $x_t$ being Markovian. If the states $x_t$ are a discrete valued Markov chain, then the state-space model is called *hidden Markov model (HMM)*.
- Assume the states, $x_t$, are Markov chain taking values in $1, \cdots, m$, with the following stationary distribution and transition probabilities

$$\pi_j = \Pr(x_t = j), \quad \pi_{ij} = \Pr(x_{t+1} = j | x_t = i) \tag{30}$$

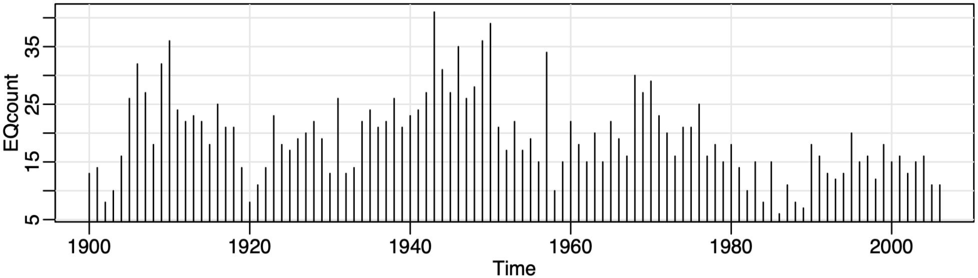
for $t = 0, 1, 2, \cdots$ and $i, j = 1, \cdots, m$.

- We then need to specify the observation model:

$$p_j(y_t) = \Pr(y_t | x_t = j) \tag{31}$$

# Poisson - HMM



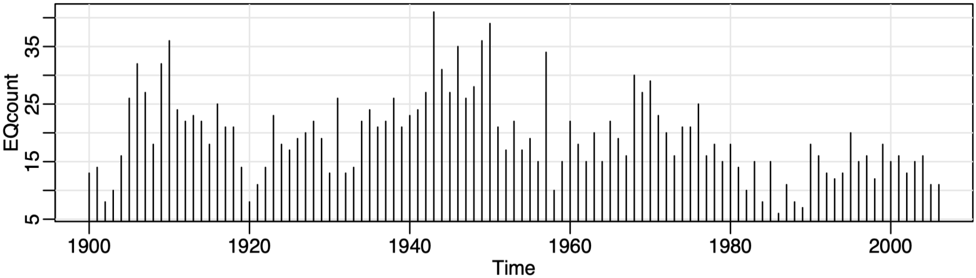



- Consider a Poisson-HMM model for the number of major earthquakes, $y_t$.
- Assume $x_t$ to be a stationary two-state Markov chain taking values in $\{1, 2\}$:

$$\pi_1 = \frac{\pi_{21}}{\pi_{12} + \pi_{21}}, \quad \pi_2 = \frac{\pi_{12}}{\pi_{12} + \pi_{21}} \tag{32}$$

- The number of major earthquakes given $j \in \{1, 2\}$ follows a Poisson distribution: $p_j(y) = \frac{\lambda_j^y e^{-\lambda_j}}{y!}$, $y = 1, 2, \cdots$

# Poisson - HMM




- The marginal distribution of $y_t$ becomes a mixture of Poissons

$$p_\Theta(y_t) = \pi_1 p_1(y_t) + \pi_2 p_2(y_t), \quad \Theta = (\lambda_1, \lambda_2) \tag{33}$$

- The mean is $\mathrm{E}(y_t) = \pi_1\lambda_1 + \pi_2\lambda_2$ and the variance is $\mathrm{Var}(y_t) = \mathrm{E}(y_t) + \pi_1\pi_2(\lambda_2 - \lambda_1)^2 \geq \mathrm{E}(y_t)$.
- The autocovariance is

$$\gamma_y(h) = \sum_{i=1}^{2}\sum_{j=1}^{2} \pi_i(\pi_{ij}^h - \pi_j)\lambda_i\lambda_j = \pi_1\pi_2(\lambda_2 - \lambda_1)^2(1 - \pi_{12} - \pi_{21})^h \tag{34}$$

# HMM Filter

- Denote $\pi_j(t|s) = \Pr(x_t = j|y_{1:s})$.
- *HMM Filter*: For $t = 1, \cdots, n$,

$$\pi_j(t|t-1) = \sum_{i=1}^{m} \pi_i(t-1|t-1)\pi_{ij} \tag{35}$$

$$\pi_j(t|t) = \frac{\pi_j(t)p_j(y_t)}{\sum_{i=1}^{m} \pi_i(t)p_i(y_t)} \tag{36}$$

with initial condition $\pi_j(1|0) = \pi_j$.

# HMM Smoother

- Denote $\varphi_j(t) = p(y_{t+1:n}|x_t = j)$.
- *HMM Smoother*: For $t = n-1, \cdots, 0$,

$$\pi_j(t|n) = \frac{\pi_j(t|t)\varphi_j(t)}{\sum_{i=1}^m \pi_i(t|t)\varphi_j(t)} \tag{37}$$

$$\pi_{ij}(t|n) = \pi_i(t|n)\pi_{ij}p_j(y_{t+1})\frac{\varphi_j(t+1)}{\varphi_i(t)} \tag{38}$$

$$\varphi_i(t) = \sum_{i=1}^m \pi_{ij}p_i(y_{t+1})\varphi_j(t+1) \tag{39}$$

where $\varphi_j(n) = 1$ for $j = 1, \cdots, m$.

# Poisson - HMM

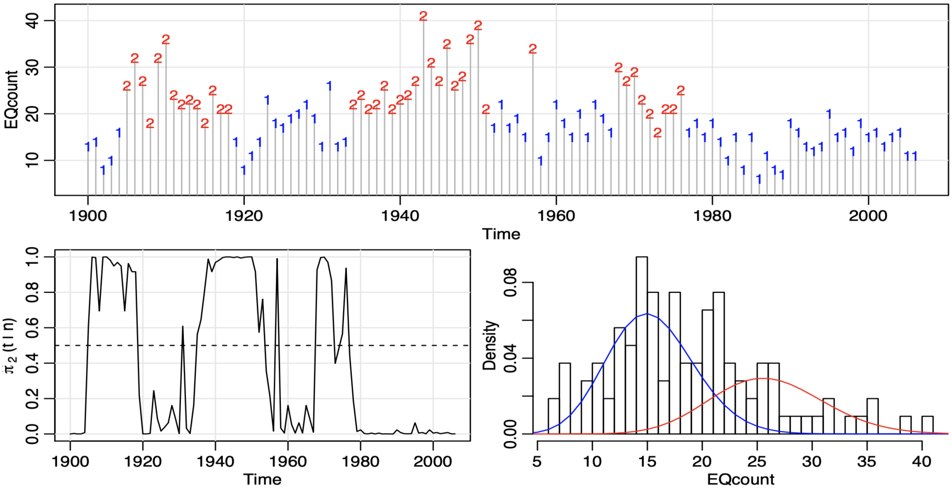



***Fig. 6.13.*** *Top: Earthquake count data and estimated states. Bottom left: Smoothing probabilities. Bottom right: Histogram of the data with the two estimated Poisson densities superimposed (solid lines).*

# Table of Contents

# Bayesian Analysis of State Space Models

- Now we consider some Bayesian approaches for the linear-Gaussian state space model

$$x_t = \Phi x_{t-1} + w_t, \quad w_t \overset{iid}{\sim} N_p(0, Q)$$
$$y_t = A_t x_t + v_t, \quad v_t \overset{iid}{\sim} N_q(0, R)$$

- The main objective is to obtain the posterior density of the parameters $p(\Theta|y_{1:n})$ or $p(x_{0:n}|y_{1:n})$ if the states are meaningful.
- We could apply the following Gibbs sampler for state space models
  1. Draw $\theta' \sim p(\Theta|x_{0:n}, y_{1:n})$
  2. Draw $x'_{0:n} \sim p(x_{0:n}|\Theta', y_{1:n})$

# Bayesian Analysis of State Space Models

- For the first step $\theta' \sim p(\Theta|x_{0:n}, y_{1:n})$, we could use conjugate priors or Metropolis-Hastings algorithm in sampling the posterior

$$p(\Theta|x_{0:n}, y_{1:n}) \propto \pi(\Theta)p(x_0|\Theta)\prod_{t=1}^{n} p(x_t|x_{t-1}, \Theta)p(y_t|x_t, \Theta) \tag{40}$$

  where $\pi(\Theta)$ is the prior on the parameters.
- For the second step $x'_{0:n} \sim p(x_{0:n}|\Theta', y_{1:n})$, it amounts to sampling from the joint smoothing distribution of the latent state sequence:

$$p_\Theta(x_{0:n}|y_{1:n}) = p_\Theta(x_n|y_{1:n})p_\Theta(x_{n-1}|x_n, y_{1:n})\cdots p_\Theta(x_0|x_1) \tag{41}$$

- Note for each component as above, Frühwirth-Schnatter (1994) use the forward-filtering, backward- sampling (FFBS) algorithm

$$p_\Theta(x_t|x_{t+1}, y_{1:t}) \propto p_\Theta(x_t|y_{1:t})p_\Theta(x_{t+1}|x_t) \tag{42}$$

  where in the linear-Gaussian case, $x_t|y_{1:t} \sim N_p^\Theta(x_t^t, P_t^t)$ and $x_{t+1}|x_t \sim N_p^\Theta(\Phi x_t, Q)$.

# Bayesian Analysis of State Space Models

- Let's revisit the local level model

$$\begin{aligned} y_t &= \mu_t + v_t, \quad v_t \overset{iid}{\sim} N(0, \sigma_v^2) \\ \mu_t &= \mu_{t-1} + w_t, \quad w_t \overset{iid}{\sim} N(0, \sigma_w^2) \end{aligned} \tag{43}$$

- We generate $y_{1:n}$ for $n = 100$ with true values $\sigma_w^2 = 0.5$ and $\sigma_v^2 = 1$.
- We adopt inverse gamma priors for variance parameters

$$\sigma_w^2 \sim \Gamma^{-1}(a_0/2, b_0/2), \quad \sigma_v^2 \sim \Gamma^{-1}(c_0/2, d_0/2) \tag{44}$$

where we set $a_0 = b_0 = c_0 = d_0 = 0.02$.

- We run MCMC to obtain 1010 samples and burn in the first 10.

# Bayesian Analysis of State Space Models

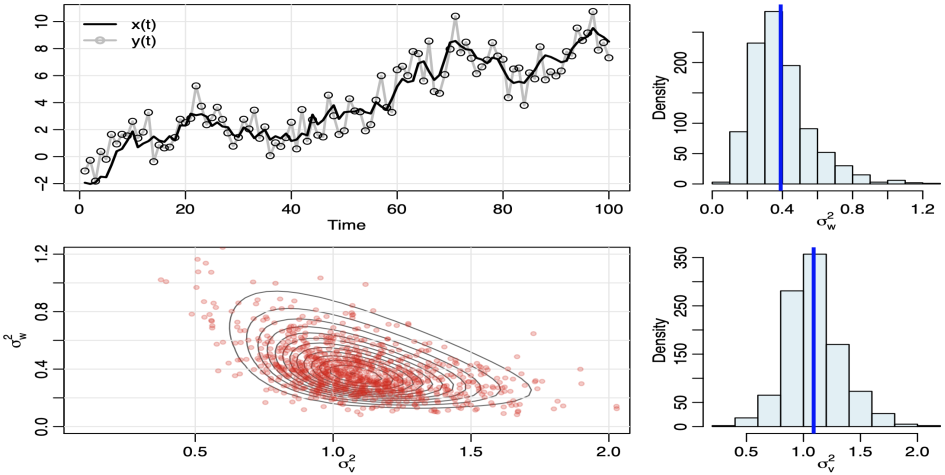



***Fig. 6.20.*** *Display for Example 6.26: Left: Generated states, $x_t$ and data $y_t$. Contours of the likelihood (solid line) of the data and sampled posterior values as points. Right : Marginal sampled posteriors and posterior means (vertical lines) of each variance component. The true values are $\sigma_w^2 = .5$ and $\sigma_v^2 = 1$.*