Lecture 14
State-Space Models
Readings: Shumway & Stoffer Ch 6.1-6.3; Brockwell & Davis Ch 8.1-8.5

MATH 8090 Time Series Analysis
November 16 & 18 & 23, 2021

Whitney Huang
Clemson University

Agenda

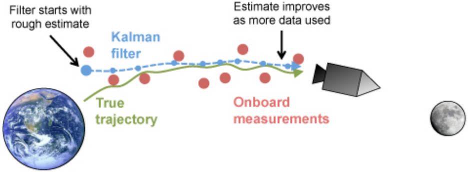
1. **Background**

2. **Forecasting, Filtering, and Smoothing**

3. **Estimating the State-Space Model Parameters**

Historical Background

- The original model arose in the space tracking setting [Kalman, 1960]; [Kalman and Bucy, 1961]
- The "state equation" defines the motion equations for the position of a spacecraft with location x_t



- The data y_t reflect information that can be observed from a tracking device such as velocity and azimuth

The main goal was to retrieve the underling state $\{x_t\}$ based on observed data $\{y_t\}$

State-Space Model

- State equation:

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 1, 2, \cdots,$$

  where

  - $\boldsymbol{X_t} \in \mathbb{R}^p$ is the state vector at time t
  - $\boldsymbol{F}_t$ is the $p \times p$ transition matrix
  - $\boldsymbol{V}_t \overset{i.i.d.}{\sim} \mathrm{WN}(\mathbf{0}, Q_t)$ is the state-transition noise

- Observation equation:

$$\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t, \quad t = 1, 2, \cdots,$$

  where

  - $\boldsymbol{Y_t} \in \mathbb{R}^q$ is the observation vector at time t
  - $\boldsymbol{H}_t$ is the $q \times p$ observation matrix
  - $\boldsymbol{W}_t \overset{i.i.d.}{\sim} \mathrm{WN}(\mathbf{0}, R_t)$ is the observation noise

Applications of State-Space Models

- Through two seemingly simple equations, state-space models define a rich class of processes that have served well as models for time series
- The so-called Kalman recursions for state-space models offer an elegant solution not only for forecasting time series, but also for filtering and smoothing
- State-space models and Kalman recursions can be readily adapted to handle time series with missing values

Additional Assumptions of State-Space Models

State equation:

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 1, 2, \cdots$$

Observation equation:

$$\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t, \quad t = 1, 2, \cdots$$

- $\mathrm{E}(\boldsymbol{W}_s \boldsymbol{V}_t^T) = 0$ for all s and t, that is, every observation noise is uncorrelated with every state-transition noise
- Assuming $\mathrm{E}(\boldsymbol{X}_1) = \boldsymbol{m}_1$, $\mathrm{E}(\boldsymbol{X}_1 \boldsymbol{W}_t^T) = 0$ and $\mathrm{E}(\boldsymbol{X}_1 \boldsymbol{V}_t^T) = 0$ for all t, that is, initial state vector are uncorrelated with both observation and state transition noises

AR(1) Process as a State-Space Model: I

- State-transition equation

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t$$

is reminiscent of a causal AR(1) model:

$$Y_{t+1} = \phi Y_t + Z_{t+1},$$

with $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and $|\phi| < 1$

- AR(1) can be expressed in state-space formulation by setting
  - $\boldsymbol{X}_{t+1} = Y_{t+1}$; $F_t = \phi$
  - $\boldsymbol{V}_t = Z_{t+1}$ along with $Q_t \overset{\text{def}}{=} \mathrm{E}(\boldsymbol{V}_t \boldsymbol{V}_t^T) = \mathrm{E}(Z_{t+1}^2) = \sigma^2$

and by using a degenerate form of the observation equation: $\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t$ in which $H_t = 1$ and $\boldsymbol{W}_t = 0$ so that $\boldsymbol{Y}_t = X_t$

AR(1) Process as a State-Space Model: II

Need to define the initial state X_1 in order to complete the model:

- A natural choice is

$$X_1 = \sum_{j=0}^{\infty} \phi^j Z_{1-j}, \quad \text{for which } \mathrm{Var}(X_1) = \frac{\sigma^2}{1-\phi^2}$$

- With this choice, the required conditions, namely, $\mathrm{E}(X_1 \boldsymbol{W}_t^T) = 0$ and $\mathrm{E}(X_1 \boldsymbol{V}_t^T) = 0$ hold

- Could also set $X_1 = Z_1 \frac{\sigma}{\sqrt{1-\phi^2}}$ to get a AR(1) process, but using $X_1 = Z_1$ would lead to a valid state-space model that is **not** a true AR(1) model

AR(1) Process as a State-Space Model: III

AR(1) process with $0 < \phi < 1$ is known as “red noise”, red noise is related to a 1st order stochastic differential equation, rendering it a model for various geophysical processes:

- Typically only observe red noise process of interest in presence of observational noise (often taken to be white noise)
- Can modify this setup by changing observational noise from $\boldsymbol{W}_t = 0$ to $\boldsymbol{W}_t = W_t \sim \mathrm{WN}(0, \sigma^2)$, where W_t is uncorrelated with Z_t's
- The observation and state-transition equations become

$$Y_t = X_t + W_t \text{ and } X_{t+1} = \phi X_t + Z_{t+1}$$

ARMA(1,1) Process as a State-Space Model: I

Recall ARMA(1,1) process $Y_t - \phi Y_{t-1} = Z_t + \theta Z_{t-1}$

- Expressing ARMA(1,1) as $\phi(B)Y_t = \theta(B)Z_t$, note that one can create Y_t by taking causal AR(1) process $X_t = \phi^{-1}(B)Z_t$ and subjecting it to a $\theta(B)$ filter to obtain output $Y_t = \theta(B)X_t = \theta(B)\phi^{-1}(B)Z_t$
- Can express filtering of AR(1) process by

$$Y_t = [1 \quad \theta]\begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix},$$

which matches up with observation equation

$$\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t$$

if $\boldsymbol{Y}_t = Y_t$, $H_t = [1 \quad \theta]$, $\boldsymbol{X}_t = \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix}$ and $\boldsymbol{W}_t = 0$

ARMA(1,1) Process as a State-Space Model: II

- Given $\boldsymbol{X}_t = \begin{bmatrix} X_t & X_{t-1} \end{bmatrix}^T$, can express $X_{t+1} = \phi X_t + Z_{t+1}$ in the 1st row of matrix equation

$$\begin{bmatrix} X_{t+1} \\ X_t \end{bmatrix} = \begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} Z_{t+1} \\ 0 \end{bmatrix},$$

which matches up with state-transition equation

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t$$

if $F_t = \begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix}$ and $\boldsymbol{V}_t = \begin{bmatrix} Z_{t+1} \\ 0 \end{bmatrix}$ with

$Q_t \stackrel{\text{def}}{=} \mathrm{E}(\boldsymbol{V}_t \boldsymbol{V}_t^T) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}$

- to complete the model, let

$$\boldsymbol{X}_1 = \begin{bmatrix} X_1 \\ X_0 \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{\infty} \phi^j Z_{1-j} \\ \sum_{j=0}^{\infty} \phi^j Z_{-j} \end{bmatrix},$$

noting that $\boldsymbol{X}_1$ and $\boldsymbol{V}_t$ for $t \geq 1$ are uncorrelated, as required

ARMA(1,1) Process as a State-Space Model: III

- Since

$$\mathrm{E}(\boldsymbol{X}_1\boldsymbol{X}_1^T) = \begin{bmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{bmatrix} = \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & \phi \\ \phi & 1 \end{bmatrix},$$

can alternatively stipulate

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & \frac{\phi}{\sqrt{1-\phi^2}} \\ 0 & \frac{\phi}{\sqrt{1-\phi^2}} \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_0 \end{bmatrix},$$

yielding

$$\mathrm{E}(\boldsymbol{X}_1\boldsymbol{X}_1^T) = \begin{bmatrix} 1 & \frac{\phi}{\sqrt{1-\phi^2}} \\ 0 & \frac{\phi}{\sqrt{1-\phi^2}} \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ \frac{\phi}{\sqrt{1-\phi^2}} & \frac{1}{\sqrt{1-\phi^2}} \end{bmatrix}$$

$$= \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & \phi \\ \phi & 1 \end{bmatrix}$$

as required

The Linear Gaussian State-Space Model

- State equation:

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t,$$

  where $\boldsymbol{V}_t \overset{iid}{\sim} \mathrm{N}(0, Q_t)$ with $\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{m}_1, P_1)$

- Observation equation:

$$\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + \boldsymbol{W}_t,$$

  where $\boldsymbol{W}_t \overset{iid}{\sim} \mathrm{N}(\boldsymbol{0}, R_t)$

- Additional assumptions: $\boldsymbol{X}_1$, $\{\boldsymbol{V}_t\}$, and $\{\boldsymbol{W}_t\}$ are uncorrelated

Forecasting, Filtering, and Smoothing

Goal: To estimate the underlying unobserved signal $\boldsymbol{X}_t$, given the data $\boldsymbol{Y}_{1:s} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_s\}$:

- When $s < t$, the problem is called forecasting or prediction
- When $s = t$, the problem is called filtering
- When $s > t$, the problem is called smoothing

In addition to these estimates, we would also want to measure their precision. The solution to these problems is accomplished via the Kalman filter and Kalman smoother

Example: Local Level Model Part I

- **Observation equation**:

$$Y_t = X_t + W_t, \quad \{W_t\} \overset{iid}{\sim} \mathrm{N}(0, \sigma_W^2)$$

- **State equation**:

$$X_{t+1} = X_t + V_t, \quad \{V_t\} \overset{iid}{\sim} \mathrm{N}(0, \sigma_V^2)$$

- Assume $\mathrm{E}(X_1) = m_1$ and $\mathrm{Var}(X_1) = P_1$ and X_1 is uncorrected with W_t's and V_t's

Local Level Model Part II

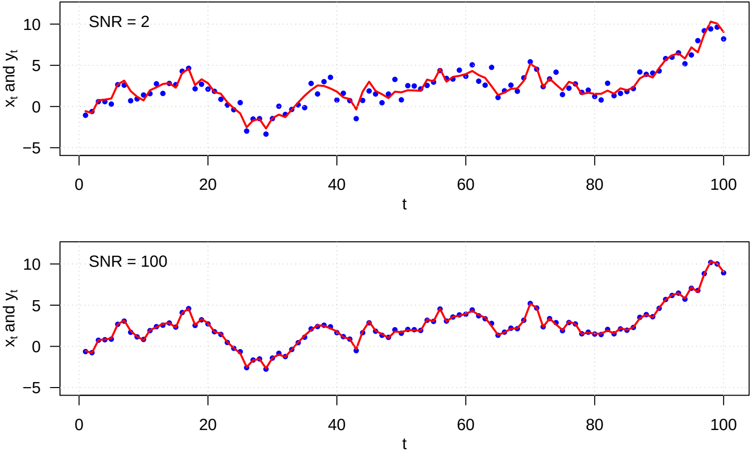
- Since $X_{t+1} = X_t + V_t$, state variable X_t is a random walk starting from m_1 (intended to model a slowly varying trend)
- Since V_t and X_t are uncorrelated,

$$\mathrm{E}(X_{t+1}|X_t) = \mathrm{E}(X_t + V_t|X_t) = X_t + \mathrm{E}(V_t) = X_t;$$

  i.e., if state variable is at a certain 'level' at time t, we can expect no change in its level at time $t+1$

- When $\sigma_W^2 > 0$, trend is corrupted by noise, so ability to pick out trend depends upon "signal to noise" ratio (SNR) $\frac{\sigma_V^2}{\sigma_W^2}$

Local Level Model: Examples



Four Classical Problems in State-Space Models

- Given observations $\{Y_i\}_{i=1}^{t}$ of a local level process,
  1. Filtering: what is best predictor of state X_t?
  2. Forecasting: what is best predictor of state X_{t+1}?
  3. Smoothing: what is best predictor of state X_s for $s < t$?
  4. Estimation: what are best estimates of model parameters $\sigma_W^2, \sigma_V^2, m_1, P_1$?

- Will concentrate first on filtering and forecasting problems with "best" taken to be minimum mean square error (MSE)

- To facilitate discussion, will assume that X_1, V_t's and W_t are multivariate normal (Gaussian) $\Rightarrow Y_t$ and remaining X_t's are also such

Regression Lemma: I

- Suppose random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, to be denoted by

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$$

- Can partition both $\boldsymbol{\mu}$ and Σ:

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{bmatrix}, \begin{bmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} \end{bmatrix} \right),$$

where $\boldsymbol{\mu_X}$ ($\boldsymbol{\mu_Y}$) and $\Sigma_{\boldsymbol{XX}}$ ($\Sigma_{\boldsymbol{YY}}$) are mean and covariance matrix for $\boldsymbol{X}$ ($\boldsymbol{Y}$); $\Sigma_{\boldsymbol{XY}}$ is the cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$

Regression Lemma: II

- Conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is multivariate normal with mean vector

$$\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{y}} = \boldsymbol{\mu}_{\boldsymbol{X}} + \Sigma_{\boldsymbol{X}\boldsymbol{Y}} \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{Y}})$$

and covariance matrix

$$\Sigma_{\boldsymbol{X}|\boldsymbol{y}} = \Sigma_{\boldsymbol{X}\boldsymbol{X}} - \Sigma_{\boldsymbol{X}\boldsymbol{Y}} \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1} \Sigma_{\boldsymbol{X}\boldsymbol{Y}}^{T}$$

- Best (under MSE) predictor of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is

$$\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y}) = \boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Y}} = \boldsymbol{\mu}_{\boldsymbol{X}} + \Sigma_{\boldsymbol{X}\boldsymbol{Y}} \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})$$

Regression Lemma: III

- Recall that, if random vector $\boldsymbol{U}$ has covariance matrix $\Sigma_{\boldsymbol{U}}$, then covariance matrix for $A\boldsymbol{U}$ is $A\Sigma_{\boldsymbol{U}}A^T$

  $\Rightarrow$ covariance matrix of $c + A(\boldsymbol{U} - \mu_{\boldsymbol{U}})$ is also $A\Sigma_{\boldsymbol{U}}A^T$

- Covariance matrix for

$$\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y}) = \boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Y}} = \boldsymbol{\mu}_{\boldsymbol{X}} + \Sigma_{\boldsymbol{X}\boldsymbol{X}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})$$

  is thus

$$\Sigma_{\boldsymbol{X}\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{X}\boldsymbol{Y}}^T = \Sigma_{\boldsymbol{X}\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{X}\boldsymbol{Y}}^T$$

Note: it is not the same as $\Sigma_{\boldsymbol{X}|\boldsymbol{y}} = \Sigma_{\boldsymbol{X}\boldsymbol{X}} - \Sigma_{\boldsymbol{X}\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{X}\boldsymbol{Y}}^T$

Regression Lemma: IV

Consider prediction error $\boldsymbol{U}$ associated with best linear predictor of $\boldsymbol{X}$:

$$\boldsymbol{U} = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X}|\boldsymbol{Y})$$

- Since $\mathrm{E}[\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y})] = \boldsymbol{\mu}_{\boldsymbol{X}} \Rightarrow \mathrm{E}(\boldsymbol{U}) = \mathbf{0}$
- Covariance matrix for $\boldsymbol{U}$ is given by

$$\begin{aligned}
\mathrm{E}(\boldsymbol{U}\boldsymbol{U}^T) =& \mathrm{E}\left([\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}|\boldsymbol{Y})][\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}|\boldsymbol{Y})]^T\right) \\
=& \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) + \mathrm{E}[\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y})]\mathrm{E}[\boldsymbol{X}|\boldsymbol{Y}]^T \\
& - \mathrm{E}[\boldsymbol{X}\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y}]^T] - \mathrm{E}[\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y}]\boldsymbol{X}^T] \\
=& \Sigma_{\boldsymbol{X}\boldsymbol{X}} - \Sigma_{\boldsymbol{X}\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{X}\boldsymbol{Y}}^T,
\end{aligned}$$

which is equal to $\Sigma_{\boldsymbol{X}|\boldsymbol{y}}$, the conditional covariance matrix

Regression Corollary

- Specialize now to case where $\boldsymbol{X}$ has just one element, say, X
- Corollary: conditional distribution of X given $\boldsymbol{Y} = \boldsymbol{y}$ is normal with mean

$$\mu_X + \Sigma_{X\boldsymbol{Y}}^T \Sigma_{\boldsymbol{YY}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu_Y})$$

  and conditional variance

$$\Sigma_{X|\boldsymbol{y}} = \sigma_X^2 - \Sigma_{X\boldsymbol{Y}}^T \Sigma_{\boldsymbol{YY}}^{-1} \Sigma_{X\boldsymbol{Y}},$$

  where $\sigma_X^2 = \mathrm{Var}(X)$ and $\Sigma_{X\boldsymbol{Y}}$ is a column vector containing covariance between X and $\boldsymbol{Y}$
- Since conditional variance is same as MSE for X, will refer to $\Sigma_{X|\boldsymbol{y}}$ as MSE

Aside – Revisiting Time Series Prediction: I

Suppose $\{X_t\}$ is zero mean stationary process with ACF $\gamma(h)$

- Set X to X_{n+1} and put $X_1, \cdots, X_n$ into $\boldsymbol{Y}$
- Corollary says best linear predictor $\hat{X}_{n+1}$ of X_{n+1} given $X_1, \cdots, X_n$ is

$$\hat{X}_{n+1} = \Sigma_{X\boldsymbol{Y}}^T \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1} \boldsymbol{Y} = \gamma_n^T \Gamma_n^{-1} \boldsymbol{Y} \stackrel{\text{def}}{=} \boldsymbol{\phi}_n^T \boldsymbol{Y},$$

  where

  1. $\gamma_n = [\gamma(1), \gamma(2), \cdots, \gamma(n)]^T = \Sigma_{X\boldsymbol{Y}}$
  2. (i, j)th entry of matrix $\Gamma_n = \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}$ is $\gamma(i - j)$
  3. $\boldsymbol{\phi}_n^T \stackrel{\text{def}}{=} \gamma_n^T \Gamma_n^{-1}$ and hence $\boldsymbol{\phi}_n = \Gamma_n^{-1} \gamma_n$

Aside – Revisiting Time Series Prediction: II

- Recall that MSE for $\hat{X}_{n+1}$ is

$$
\begin{aligned}
v_n &= \mathrm{Var}(X_{n+1}) - \boldsymbol{\phi}_n^T \boldsymbol{\gamma}_n \\
&= \sigma_X^2 - \boldsymbol{\gamma}_n^T \Gamma_n^{-1} \boldsymbol{\gamma}_n \\
&= \sigma_X^2 - \Sigma_{X\boldsymbol{Y}}^T \Sigma_{\boldsymbol{Y}\boldsymbol{Y}}^{-1} \Sigma_{X\boldsymbol{Y}} \\
&= \Sigma_{X|\boldsymbol{y}}
\end{aligned}
$$

- This is a special case of regression corollary

Filtering for Local Level Model: I

- Return now to local level model:

$$\begin{aligned} Y_t &= X_t + W_t, \quad \{W_t\} \sim \mathrm{N}(0, \sigma_W^2) \\ X_{t+1} &= X_t + V_t, \quad \{V_t\} \sim \mathrm{N}(0, \sigma_V^2) \end{aligned}$$

  and X_1 is an RV that

  - is uncorrelated with W_t's and V_t's
  - has $\mathrm{E}(X_1) = m_1$ and $\mathrm{Var}(X_1) = P_1$

- Filtering problem is to predict unknown state X_t based on data up to time t, i.e., $Y_1, \cdots, Y_t$
- In what follows, let $Y_{1:t} \stackrel{\text{def}}{=} [Y_1, \cdots, Y_t]^T$

Filtering for Local Level Model: II

- Best linear predictor of X_t given $Y_{1:t}$ is

$$\hat{X}_{t|t} \stackrel{\text{def}}{=} \mathrm{E}(X_t|Y_{1:t}) = m_t + \Sigma_{t,t}^T \Sigma_t^{-1}(Y_{1:t} - \boldsymbol{m}_t),$$

  where
  - $m_t = \mathrm{E}(X_t)$
  - Vector $\Sigma_{t,t}$ contains covarinces between X_t and $Y_{1:t}$
  - (i, j)th element of matrix Σ_t is covariance between Y_i and Y_j
  - $\boldsymbol{m}_t$ is a vector containing, for $j = 1, \cdots, t$,

$$m_j \stackrel{\text{def}}{=} \mathrm{E}(X_j) = \mathrm{E}(X_j + W_j) = \mathrm{E}(Y_j)$$

- Note: $\mathrm{E}(\hat{X}_{t|t}) = \mathrm{E}[\mathrm{E}(X_t|Y_{1:t})] = \mathrm{E}(X_t) = m_t$

- With $P_t \stackrel{\text{def}}{=} \mathrm{Var}(X_t)$, MSE for predictor is

$$\mathrm{E}[(X_t - \hat{X}_{t|t})^2] = P_t - \Sigma_{t,t}^T \Sigma_t^{-1} \Sigma_{t,t} \stackrel{\text{def}}{=} P_{t|t}$$

Forecasting for Local Level Model: I

Forecasting: estimate X_{t+1} given $Y_{1:t}$

- Best linear predictor of X_{t+1} given $Y_{1:t}$ is

$$\hat{X}_{t+1|t} \stackrel{\text{def}}{=} \mathrm{E}(X_{t+1}|Y_{1:t}) = m_{t+1} + \Sigma_{t+1,t}^T \Sigma_t^{-1}(Y_{1:t} - \boldsymbol{m}_t),$$

where vector $\Sigma_{t+1,t}$ has covaraince between X_{t+1} and $Y_{1:t}$

- Note: $\mathrm{E}(\hat{X}_{t+1}|t) = \mathrm{E}[\mathrm{E}(X_{t+1}|Y_{1:t})] = \mathrm{E}(X_{t+1}) = m_{t+1}$

- MSE for predictor is

$$\mathrm{E}[(X_{t+1} - \hat{X}_{t+1|t})^2] = P_{t+1} - \Sigma_{t+1,t}^T \Sigma_t^{-1} \Sigma_{t+1,t} \stackrel{\text{def}}{=} P_{t+1|t}$$

Forecasting for Local Level Model: II

- Let's also consider best linear predictor of Y_{t+1} given $Y_{1:t}$:

$$\hat{Y}_{t+1|t} \stackrel{\text{def}}{=} \mathbb{E}(Y_{t+1}|Y_{1:t}) = m_{t+1} + \tilde{\Sigma}^T_{t+1,t}\Sigma_t^{-1}(Y_{1:t} - \boldsymbol{m}_t),$$

where the vector $\tilde{\Sigma}_{t+1,t}$ has covaraínces between Y_{t+1} and $Y_{1:t}$

- However, note that, for $j = 1, \cdots, t$

$$\mathrm{Cov}(Y_{t+1}, Y_j) = \mathrm{Cov}(X_{t+1} + W_{t+1}, Y_j) = \mathrm{Cov}(X_{t+1}, Y_j)$$

- Thus $\tilde{\Sigma}_{t+1,t} = \Sigma_{t+1,t}$, yielding

$$\hat{Y}_{t+1|t} = m_{t+1} + \Sigma^T_{t+1,t}\Sigma_t^{-1}(Y_{1:t} - \boldsymbol{m}_t) = \hat{X}_{t+1|t}$$

$\Rightarrow$ difference between Y_{t+1} and X_{t+1} is W_{t+1}, therefore they have the same estimator, but their MSEs differ:

$$\mathbb{E}\left[(Y_{t+1} - \hat{Y}_{t+1}|t)^2\right] = P_{t+1|t} + \sigma^2_W$$

Filtering for Local Level Model: III

- To implement filtering (i.e., compute $\hat{X}_{t|t}$), need to determine

  1. $m_j = \mathbb{E}(X_j)$, $j = 1, \cdots, t$
  2. Elements of $\Sigma_{t,t}$, i.e., covarainces between X_t and $Y_1, \cdots, Y_t$
  3. Elements of Σ_t, i.e., covariances between Y_j and Y_k, $1 \le j \le k \le t$

- To compute $P_{t|t}$, i.e., MSE for $\hat{X}_{t|t}$, need $P_t = \mathrm{Var}(X_t)$ in addition to 2 and 3 above

- Since $X_t = X_{t-1} + V_{t-1}$ and $Y_t = X_t + W_t$, telescoping yields $X_j = X_1 + \sum_{l=1}^{j-1} V_l$ and $Y_j = X_1 + \sum_{l=1}^{j-1} V_l + W_j$, $j = 1, \cdots, t$

Filtering for Local Level Model: IV

- Using

$$X_j = X_1 + \sum_{l=1}^{j-1} V_l \text{ and } Y_j = X_1 + \sum_{l=1}^{j-1} + W_j, \quad j = 1, \cdots, t,$$

get $m_j = \mathbb{E}[X_j] = \mathbb{E}[X_1] = m_1$ and (assuming $j \le k \le t$)

$$\begin{aligned}\mathbb{C}\mathrm{ov}(X_t, Y_j) &= \mathbb{C}\mathrm{ov}\left(X_1 + \sum_{l=1}^{t-1} V_l, X_1 + \sum_{l=1}^{j-1} V_l + W_j\right) \\ &= P_1 + (j-1)\sigma_V^2 \\ \mathbb{C}\mathrm{ov}(Y_j, Y_k) &= \mathbb{C}\mathrm{ov}\left(X_1 + \sum_{l=1}^{j-1} V_l + W_j, X_1 + \sum_{l=1}^{k-1} V_l + W_k\right) \\ &= P_1 + (j-1)\sigma_V^2 + \delta_{jk}\sigma_W^2,\end{aligned}$$

where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \ne k$

Filtering for Local Level Model: V

Using

$$X_t = X_1 + \sum_{l=1}^{t-1} V_l,$$

get

$$P_t = \mathbb{V}\mathrm{ar}(X_t) = P_1 + (t-1)\sigma_V^2$$

- Now we have all the pieces needed to form $\hat{X}_{t|t}$ and its MSE $P_{t|t}$
- **Note**: similar argument leads to pieces needed to form forecast $\hat{X}_{t+1|t}$ and its MSE $P_{t+1|t}$

Kalman Recursions for Filtering/Forecasting: I

While straightforward conceptually, forming

$$\hat{X}_{t|t} = m_t + \Sigma_{t,t}^T \Sigma_t^{-1} (Y_{1:t} - \boldsymbol{m}_t)$$

and

$$\hat{X}_{t+1|t} = m_{t+1} + \Sigma_{t+1,t}^T \Sigma_t^{-1} (Y_{1:t} - \boldsymbol{m}_t)$$

via these equations requires inversion of matrix Σ_t whose dimension $t \times t$ becomes problematic as t gets large

- The celebrated Kalman recursions give a recipe that avoids explicit matrix inversion
- **Idea**: at time $t-1$, we have 4 quantities of interest: fitted value $\hat{X}_{t-1|t-1}$, and forecast $\hat{X}_{t|t-1}$ and their associated MSEs $P_{t-1|t-1}$ and $P_{t|t-1}$
- **Note**: $\hat{X}_{t-1|t-1} = \hat{X}_{t|t-1}$ for local level model (but not others)

Kalman Recursions for Filtering/Forecasting: II

- At time t, new observation Y_t becomes available
- Kalman recursion takes $\hat{X}_{t|t-1}$, $P_{t|t-1}$ and Y_t and yields
  - fitted values $\hat{X}_{t|t}$ and forecast $\hat{X}_{t+1|t}$
  - associated MSEs $P_{t|t}$ and $P_{t+1|t}$
- There are six steps in the Kalman recursion:
  1. steps 1 and 2 are preparatory
  2. steps 3 and 4 yield $\hat{X}_{t|t}$ and $P_{t|t}$ (filtering)
  3. steps 5 and 6 yield $\hat{X}_{t+1|t}$ and $P_{t+1|t}$ (forecasting)

Kalman Recursions for Filtering/Forecasting: III

1. Compute innovation:

$$U_t = Y_t - \hat{Y}_{t|t-1} = Y_t - \hat{X}_{t|t-1}$$

2. Compute MSE for $\hat{Y}_{t|t-1}$:

$$P_{t|t-1} + \sigma_W^2 \stackrel{\text{def}}{=} F_t$$

3. Compute new filtered value:

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t U_t,$$

where $K_t \stackrel{\text{def}}{=} P_{t|t-1}/F_t$ is the so-called Kalman gain

4. Compute MSE for new filtered value:

$$P_{t|t} = P_{t|t-1}(1 - K_t)$$

Kalman Recursions for Filtering/Forecasting: IV

5. Compute new forecast:

$$\hat{X}_{t+1|t} = \hat{X}_{t|t-1} + K_t U_t = \hat{X}_{t|t}$$

6. Compute MSE for new forecast:

$$P_{t+1|t} = P_{t|t-1}(1 - K_t) + \sigma_V^2 = P_{t|t} + \sigma_V^2$$

- Recursions are carried out for $t = 1, \cdots, n$ with inputs at $t = 1$ being $\hat{X}_{1|0} \overset{\text{def}}{=} m_1 = \mathbb{E}[X_1]$, $P_{1|0} \overset{\text{def}}{=} P_1 = \mathbb{V}\mathrm{ar}(X_1)$ and Y_1

Kalman Recursions for Filtering/Forecasting: V

To prove validity of steps 3 and 4, need to show that

- $\hat{X}_{t|t-1} + K_t U_t$ is equal to $\hat{X}_{t|t}$
- $P_{t|t-1}(1 - K_t)$ is equal to $P_{t|t}$

$$\begin{aligned}\mathbb{Cov}(X_t, U_t | Y_{1:t-1}) &= \mathbb{Cov}(X_t, Y_t - \hat{Y}_{t|t-1} | Y_{1:t-1}) \\ &= \mathbb{Cov}(X_t, X_t + W_t | Y_{1:t-1}) = \mathbb{Var}(X_t | Y_{1:t-1}) \\ &= P_{t|t-1}\end{aligned}$$

- **Key fact**: X_t conditioned on both $U_t = Y_t - \hat{Y}_{t|t-1}$ and $Y_{1:t-1}$ is the same as X_t conditioned on $Y_{1:t-1}$ We have

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + \frac{P_{t|t-1}}{F_t} U_t$$

and

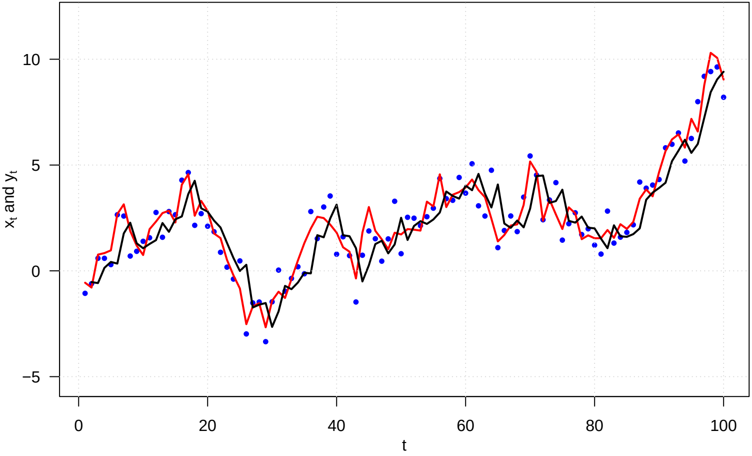
$$P_{t|t} = P_{t|t-1} - \frac{P^2_{t|t-1}}{F_t}$$

since $K_t \stackrel{\text{def}}{=} \frac{P_{t|t-1}}{F_t}$, we get required

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t U_t \text{ and } P_{t|t} = P_{t|t-1}(1 - K_t)$$

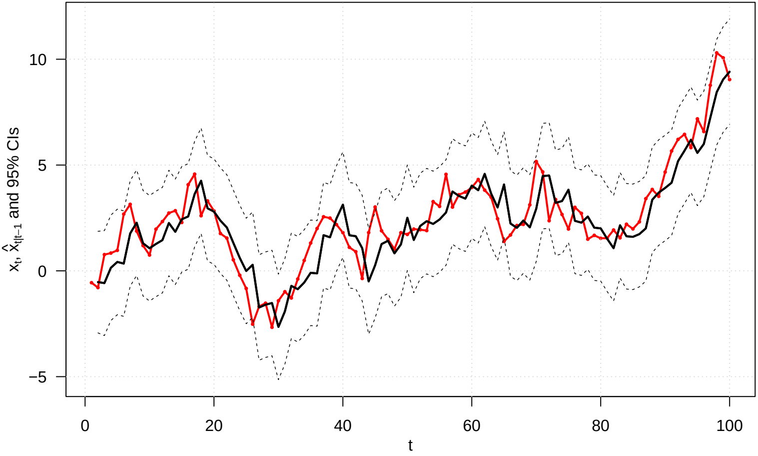
Simulated Example: Local Level Model with SNR $= 2$

Time series Y_t, states X_t, and forecasts $\hat{X}_{t|t-1}$



Simulated Data from Local Level Model with SNR $= 2$

States X_t, forecasts $\hat{X}_{t|t-1}$, and 95% CIs based on $P_{t|t-1}$

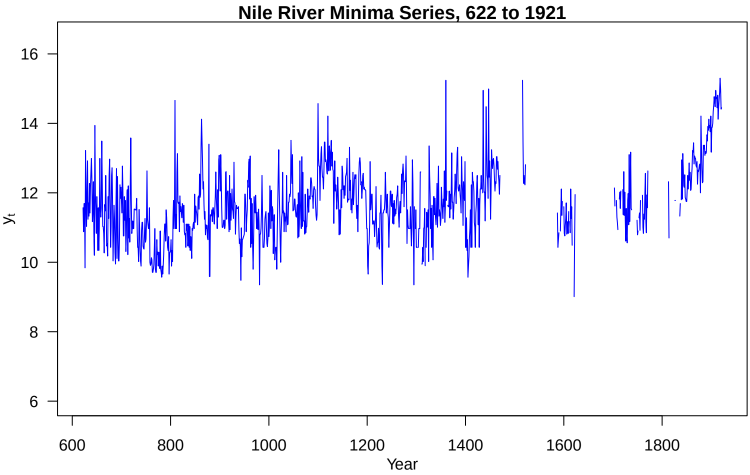


Kalman Recursions for Time Series with Missing Values: I

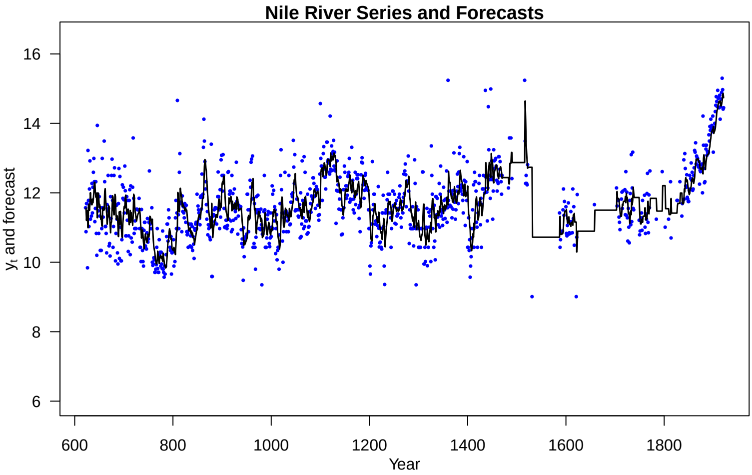
One of the strengths of state-space formulation is the capability to handle time series with missing values. Suppose $Y_1, \cdots, Y_t$ and Y_{t+3} are observed, but not Y_{t+1} and Y_{t+2}:

- use modified recursion (i.e., skip the calculation of the innovation when data is missing)
  - use $\hat{X}_{t+1|t}$ and $P_{t+1|t}$ for $\hat{X}_{t+2|t}$ and $P_{t+2|t}$
  - use $\hat{X}_{t+2|t}$ and $P_{t+2|t}$ for $\hat{X}_{t+3|t}$ and $P_{t+3|t}$
- take $\hat{X}_{t+3|t}$, $P_{t+3|t}$, and Y_{t+3} into usual recursion to obtain $\hat{X}_{t+3|t+3}$ and $P_{t+3|t+3}$ and $\hat{X}_{t+4|t+3}$ and $P_{t+4|t+3}$
- need to interpret "given $t+3$" as conditioning on everything available at time $t+3$, i.e., $Y_1, \cdots, Y_t$ and Y_{t+3}

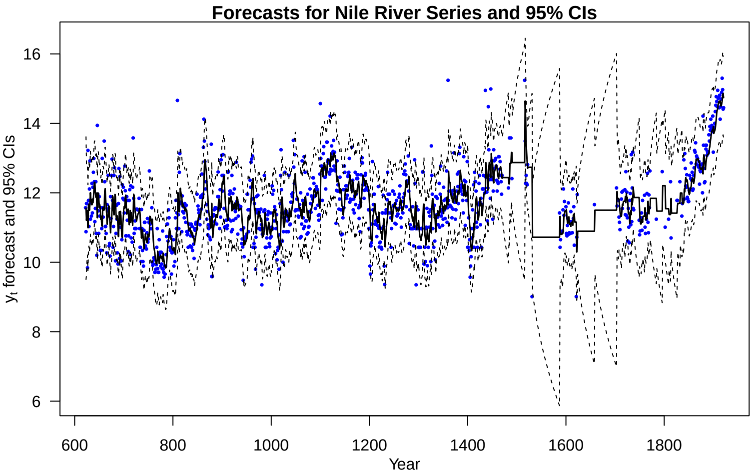
Example: Nile River Annual Minima Series



Nile River Annual Minima Series with Missing Values Imputed



Nile River Annual Minima Series Forecasts with 95 % CI



Kalman Recursions for Smoothing: I

Given time series $Y_1, \cdots, Y_n$, Kalman filter recursions give us $\hat{X}_{t|t}$ for $t = 1, \cdots, n$

- Regression lemma says solution to smoothing problem is

$$\hat{X}_{t|n} \stackrel{\text{def}}{=} \mathbb{E}[X_t | Y_{1:n}] = m_t + \Sigma_{t,n}^T \Sigma_n^{-1} (Y_{1:n} - \boldsymbol{m}_n)$$

- MSE for predictor, i.e., $\mathbb{E}\left[\left(X_t - \hat{X}_{t|n}\right)^2\right]$, is

$$P_t - \Sigma_{t,n}^T \Sigma_n^{-1} \Sigma_{t,n} \stackrel{\text{def}}{=} P_{t|n},$$

where $P_t \stackrel{\text{def}}{=} \mathbb{V}\mathrm{ar}[X_t]$

Kalman Recursions for Smoothing: II

Using innovation U_t, innovation variance F_t, Kalman gains K_t, forecasts $\hat{X}_{t|t-1}$ and associated MSEs $P_{t|t-1}, t = 1, \cdots, n$ computed by Kalman filter recursions, Kalman smoother recursions allow efficient computation of $\hat{X}_{t|n}, t = 1, \cdots, n$

The first two steps yield desired predictor $\hat{X}_{t|n}$

1. Manipulate innovations: starting with $r_n = 0$, recursively form

$$r_{n-1} = \frac{U_t}{F_t} + (1 - K_t) r_t, \quad t = n, \cdots, 1$$

2. Combine manipulated innovations and forecasts:

$$\hat{X}_{t|n} = \hat{X}_{t|t} + P_{t|t-1} r_{t-1}, \quad t = 1, \cdots, n$$

Kalman Recursions for Smoothing: III

Next two steps yield MSE for predictor $\hat{X}_{t|n}$:

3. Manipulate innovation variances: starting with $N_n = 0$, recursively form

$$N_{t-1} = \frac{1}{F_t} + (1 - K_t)^2 N_t, \quad t = n, \cdots, 1$$

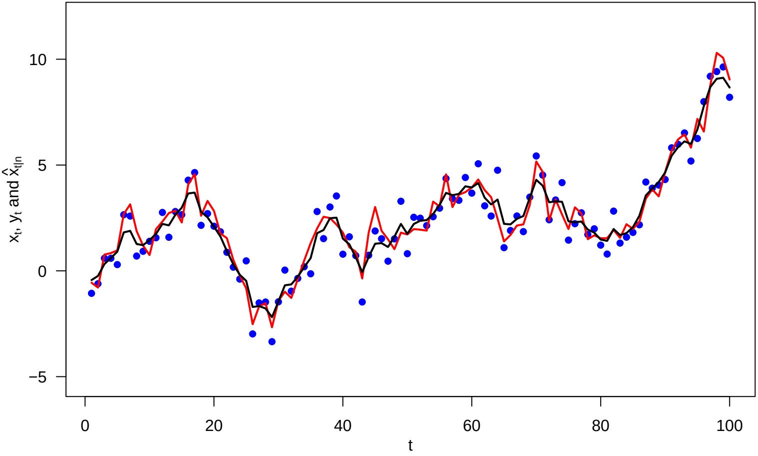
4. Combine manipulated innovation variances and forecast MSEs:

$$V_t = P_{t|t-1} - P_{t|t-1}^2 N_{t-1}, \quad t = 1, \cdots, n,$$

where V_t is the desired MSE

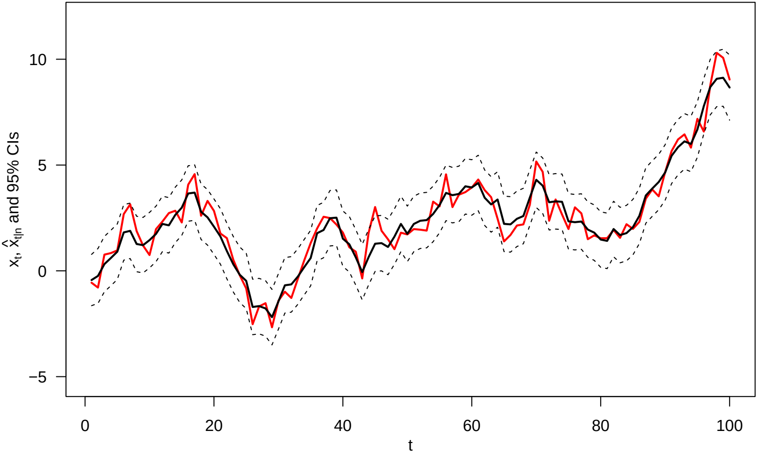
Simulated Example: Local Level Model with SNR $= 2$

Time series Y_t, states X_t, and smooths $\hat{X}_{t|n}$



Simulated Data from Local Level Model with SNR $= 2$

States X_t, smooths $\hat{X}_{t|n}$, and 95% CIs based on $P_{t|n}$



Estimating the State-Space Model Parameters

So far, we've assumed that the parameters $\boldsymbol{\theta} = (\sigma_V^2, \sigma_W^2, m_1, P_1)$ are known. In practice, we need to estimate from the data.

This requires maximizing the marginal likelihood of the data $\boldsymbol{y}$, having integrated the latent time series $\boldsymbol{x}$ out. This is given by:

$$f(\boldsymbol{y}|\sigma_V^2, \sigma_W^2, m_1, P_1) = \int f(\boldsymbol{y}|\boldsymbol{x}, \sigma_W^2) f(\boldsymbol{x}|m_1, P_1, \sigma_V^2)$$

Maximizing over an integral can be difficult

Direct Maximum Marginal Likelihood

Fortunately, our normal distribution facts tell us that the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathrm{N}(\mathbb{E}(\boldsymbol{x}), \mathbb{Var}(\boldsymbol{x}) + \sigma_W^2 I_n).$$

However, the direct evaluation of the marginal likelihood can be challenge due to $n \times n$ matrix inversions

Alternative, we use the innovations $u_t = y_t - \hat{Y}_{t|t-1}$ to compute the likelihood:

$$\ell(\boldsymbol{\theta}) \propto f(u_1) \prod_{t=2}^{n} f(u_t | y_{1:t-1}).$$

We can do the following iteratively:

- Pick an initial guess $\hat{\boldsymbol{\theta}}^0$ and run the Kalman filter to get a set of innovations $\boldsymbol{u}$
- Maximizing $\boldsymbol{\theta}$ (e.g., via Newton–Raphson) with $\boldsymbol{u}$ to obtain new estimate of $\boldsymbol{\theta}$

Expectation-Maximization (EM) Maximum Marginal Likelihood

Another way to compute maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is to use the expectation-maximization (EM) algorithm [Dempster, Laird, and Rubin, 1977]

- Initialize by choosing starting value $\boldsymbol{\theta}^0$, and compute the incomplete likelihood
- Perform the E-step to obtained $\hat{X}_{t|n}$, $P_{t|n}$
- Perform M-step to update the estimate $\boldsymbol{\theta}$ using the complete likelihood
- Recompute the incomplete likelihood
- Repeat until convergence, i.e., $|\hat{\boldsymbol{\theta}}^T - \hat{\boldsymbol{\theta}}^{T-1}| < \epsilon$