Lecture 1

State Space Models and the Kalman Filter

Lars A. Lochstoer

UCLA Anderson School of Management

Outline

1. Latent (or Hidden) processes
   - Learning from a noisy signal
2. Sequential Learning
3. The Local Level Model
4. State space models and the Kalman Filter
   - Filtering Probabilities
   - Smoothing Probabilities
   - Maximum Likelihood Estimation

Latent Processes

Economy driven by forces we do not directly observe:

1. Risk aversion / risk capacity
2. The state of the business cycle
3. Beliefs about asset payoffs, macroeconomic conditions, etc.
4. Monetary and fiscal stance (e.g., true policy parameters)
5. +++

These are generally time-varying, dynamic concepts – not constant parameters

We generally do observe outcomes that are related to these latent (hidden) drivers

- We can learn from such (noisy) signals!

Noisy signals: a simple example

Assume inflation π_t follows the process:

$$\pi_t = \mu + \phi\left(\pi_{t-1} - \mu\right) + x_t + \varepsilon_t,$$

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$ and x_t is a latent variable.

Do we observe a (noisy) signal about x_t? Yes!

- Define
$$z_t \equiv \pi_t - \mu - \phi\left(\pi_{t-1} - \mu\right).$$
- Then
$$z_t = x_t + \varepsilon_t, \quad \text{and} \quad z_t | x_t \sim N\left(x_t, \sigma^2\right).$$

Given what we have so far, what is $E_t\left(x_t\right)$?

$$E_t\left(x_t\right) = z_t.$$

Aside: The expectation operator and latent variables

The expectation operator

$$E_t(\cdot) = E(\cdot|I_t)$$

refers to the expectation conditional on the *information set* at time t, I_t.

The latent variable x_t is not in your information set at time t, despite it having a t subscript.

What is in our information set then?

- The history of realized inflation π, and thus signals z, up until time t
- This history is typically written π^t and z^t

Noisy signals and prior beliefs

The above expectation

$$E_t\left(x_t\right) = z_t$$

assumed we had no prior knowledge about what x_t might be.

- Equivalently, we have a "flat" prior about x_t

Priors and posteriors? Let's revisit Bayes rule!

Bayes Rule and Updates in Beliefs

Recall Bayes rule:

$$P(A|B)\,P(B) = P(A,B) \Leftrightarrow P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

So, in our example:

$$P(x_t|z_t) = \frac{P(z_t|x_t)\,P(x_t)}{P(z_t)}$$

The lingo:

- $P(x_t)$: the **prior distribution** of x_t
- $P(x_t|z_t)$: the **posterior distribution** after observing the signal
- $P(z_t|x_t)$: the pdf of the signal conditional on x_t: the **likelihood function**
- $P(z_t)$: the marginal dist. of the observation z_t (thus, $P(z_t)$ is a constant)

Bayes Rule and Updates in Beliefs (cont'd)

Note that in our case:

$$P\left(z_t|x_t\right) = N\left(x_t, \sigma^2\right).$$

Let the prior be Normally distributed:

$$p\left(x_t\right) = N\left(x_{t|t-1}, \sigma^2_{t|t-1}\right)$$

The appendix shows that in this case

$$\frac{P\left(z_t|x_t\right)P\left(x_t\right)}{P\left(z_t\right)} = N\left(x_{t|t}, \sigma^2_{t|t}\right),$$

where

$$x_{t|t} = \frac{\sigma^{-2}}{\sigma^{-2}+\sigma_x^{-2}}z_t + \frac{\sigma_x^{-2}}{\sigma^{-2}+\sigma_x^{-2}}x_{t|t-1},$$

and

$$\sigma^2_{t|t} = \left(\sigma^{-2}_{t|t-1} + \sigma^{-2}\right)^{-1}.$$

Flat vs informative priors

A flat prior means $\sigma_x^2 = \infty$

- We then have:

$$x_{t|t} = \frac{\sigma^{-2}}{\sigma^{-2} + \sigma_{t|t-1}^{-2}} z_t + \frac{\sigma_x^{-2}}{\sigma^{-2} + \sigma_{t|t-1}^{-2}} x_{t|t-1} = z_t.$$

But, if we have an informative prior, $\sigma_x^2 < \infty$

- Such an informative prior could come from prior observations if x_t is not independent across time
- For instance, x_t might follow an AR(1) process

Sequential Learning

We update beliefs every time new information arrives

- This is called **sequential learning**

Let y^t denote available data up until and including time t

- Assume we are trying to infer the value of a latent variable, x_t. If the data only provides noisy signals, we are looking for the distribution of $x_t|y^t$.
- Note that $x_t|y^t = x_t|y_t, y^{t-1}$

Sequential learning and Bayes Rule:

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

- $p\left(x_t|y^{t-1}\right)$ is the *prior distribution* of the state x_t, conditional on data up until time $t-1$
- $p\left(y_t|x_t, y^{t-1}\right)$ is the *likelihood* of the empirical observation y_t, given your prior
- $p\left(x_t|y^t\right)$ is the *posterior distribution* about the latent state, x_t

The Local Level Model

Local Level Model

Consider the **local level model**, for $t = 1, \ldots, T$:

$$y_t = \mu_t + \eta_t, \qquad \eta_t \sim \mathsf{N}(0, \sigma_\eta^2)$$
$$\mu_{t+1} = \mu_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

- We observe the data y_t for $t = 1, \ldots, T$
- But! The trend μ_t is not observable. It is latent.
- The Kalman filter provides a way to estimate the trend μ_t.
- The trend μ_t is also known as a **state variable**.
- In this lecture, I will use the notation: $y_{1:t} = (y_1, \ldots, y_t)$

Local Level Model

$$y_t = \mu_t + \eta_t, \qquad \eta_t \sim \mathsf{N}(0, \sigma_\eta^2)$$
$$\mu_{t+1} = \mu_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

- The Kalman filter, Kalman (1960), is a recursive algorithm that performs:
  1. Filtering
  2. Prediction
  3. Smoothing
  4. Evaluation of the log-likelihood
- Let's introduce each of these ideas one-by-one.
- For now, assume we know the parameters: $\boldsymbol{\theta} = (\sigma_\eta^2, \sigma_\varepsilon^2)$

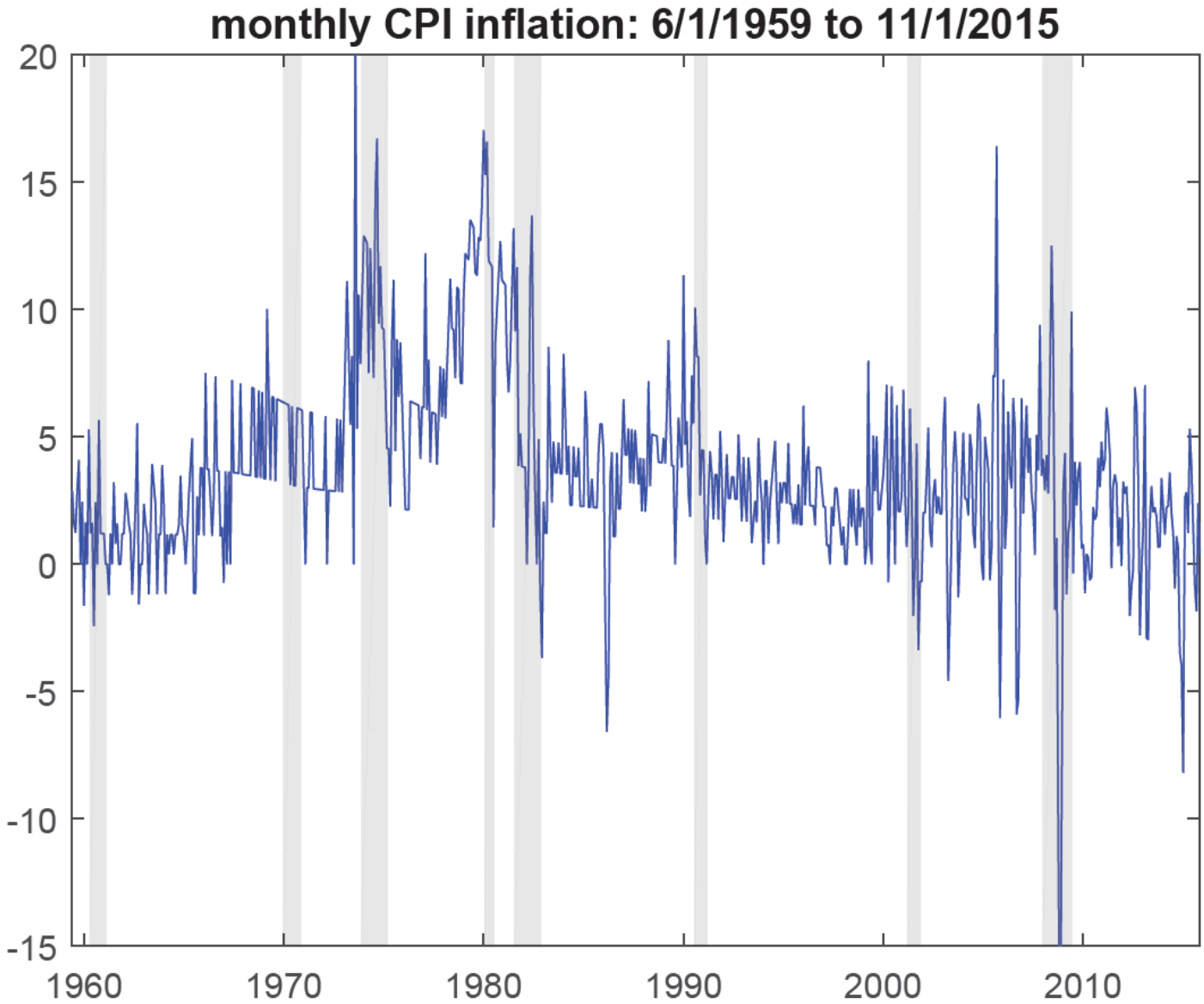
U.S. inflation

- Let $y_t = \pi_t$ denote monthly U.S. CPI inflation
- Stock and Watson (2007) use the local level model to forecast inflation.

$$\pi_t = \bar{\pi}_t + \eta_t, \qquad \eta_t \sim \mathsf{N}(0, \sigma_\eta^2)$$
$$\bar{\pi}_{t+1} = \bar{\pi}_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

- The state variable $\mu_t = \bar{\pi}_t$ represents the trend in inflation and is what many economists would call '*expected inflation*.'

U.S. inflation



U.S. inflation, monthly continuously compounded.

Bayes Theorem

- Let's start at time $t = 0$ before we have seen the first observation y_1.
- Let $p(\mu_1; \boldsymbol{\theta})$ denote a prior distribution that describes our beliefs about μ_1 before we have observed y_1.
- Let's assume that our prior $p(\mu_1; \boldsymbol{\theta})$ is a normal distribution

$$p(\mu_1; \boldsymbol{\theta}) = \mathsf{N}(\mu_{1|0}, \Sigma_{1|0})$$

- The notation $\mu_{1|0}$ means that it is our guess (prediction) of the value of μ_1 at time $t = 1$ but with time $t = 0$ information.
- The variance $\Sigma_{1|0}$ measures our uncertainty about μ_1 with time $t = 0$ information.
- Recall that, for the normal distribution, the mean and covariance matrix are **sufficient statistics**.

Bayes Theorem

- We have the prior: $p(\mu_1; \boldsymbol{\theta}) = \mathsf{N}(\mu_{1|0}, \Sigma_{1|0})$.
- We observe y_1 which is a noisy measure of μ_1.
- Our model says y_1 and μ_1 are related:

$$y_1 = \mu_1 + \eta_1, \qquad \eta_1 \sim \mathsf{N}(0, \sigma_\eta^2)$$

- The (conditional) likelihood is:

$$p(y_1 | \mu_1; \boldsymbol{\theta}) = \mathsf{N}(\mu_1, \sigma_\eta^2)$$

Bayes Theorem

- We have the prior: $p(\mu_1; \boldsymbol{\theta}) = \mathsf{N}(\mu_{1|0}, \Sigma_{1|0})$
- The (conditional) likelihood: $p(y_1|\mu_1; \boldsymbol{\theta}) = \mathsf{N}(\mu_1, \sigma_\eta^2)$
- After we observe the data y_1, how do we revise our beliefs about μ_1?
- We apply Bayes rule:

$$p(\mu_1|y_1; \boldsymbol{\theta}) = \frac{p(y_1|\mu_1; \boldsymbol{\theta})p(\mu_1; \boldsymbol{\theta})}{p(y_1; \boldsymbol{\theta})}$$

- The posterior $p(\mu_1|y_1; \boldsymbol{\theta})$ describes our beliefs about μ_1 after we observe y_1.

Filtering: local level model

- Applying Bayes Rule, we find

$$p(\mu_1|y_1;\boldsymbol{\theta}) \;=\; \frac{p(y_1|\mu_1;\boldsymbol{\theta})p(\mu_1;\boldsymbol{\theta})}{p(y_1;\boldsymbol{\theta})} \;=\; \mathsf{N}(\mu_{1|1},\Sigma_{1|1})$$

  where the mean and variance are

$$\mu_{1|1} \;=\; \mu_{1|0}+\Sigma_{1|0}F_1^{-1}(y_1-\mu_{1|0}) \qquad \Sigma_{1|1} \;=\; \Sigma_{1|0}-\Sigma_{1|0}F_1^{-1}\Sigma_{1|0}$$

  and $F_1 = \Sigma_{1|0}+\sigma_\eta^2$.

- The notation $\mu_{1|1}$ indicates our estimate for μ_1 given $t=1$ information.
- All distributions are normally distributed. Only update the mean and covariance matrix (the sufficient statistics).

Aside: Updating equations versus earlier case

Updating equations look superficially different from those earlier in this note. But:

$$
\begin{aligned}
\mu_{1|1} &= \mu_{1|0} + \Sigma_{1|0} F_1^{-1} (y_1 - \mu_{1|0}) \\
&= \mu_{1|0} \left(1 - \Sigma_{1|0} F_1^{-1}\right) + \Sigma_{1|0} F_1^{-1} y_1 \\
&= \mu_{1|0} \left(1 - \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma_\eta^2}\right) + \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma_\eta^2} y_1 \\
&= \frac{\sigma_\eta^2}{\Sigma_{1|0} + \sigma_\eta^2} \mu_{1|0} + \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma_\eta^2} y_1 \\
&= \frac{\sigma_\eta^2}{\Sigma_{1|0} + \sigma_\eta^2} \frac{\Sigma_{1|0}^{-1} \sigma_\eta^{-2}}{\Sigma_{1|0}^{-1} \sigma_\eta^{-2}} \mu_{1|0} + \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma_\eta^2} \frac{\Sigma_{1|0}^{-1} \sigma_\eta^{-2}}{\Sigma_{1|0}^{-1} \sigma_\eta^{-2}} y_1 \\
&= \frac{\Sigma_{1|0}^{-1}}{\Sigma_{1|0}^{-1} + \sigma_\eta^{-2}} \mu_{1|0} + \frac{\sigma_\eta^{-2}}{\Sigma_{1|0}^{-1} + \sigma_\eta^{-2}} y_1,
\end{aligned}
$$

which is of the same form as the initial updating example.

Aside: Updating equations versus earlier case

Next, the variance:

$$\begin{aligned}
\Sigma_{1|1} &= \Sigma_{1|0} - \Sigma_{1|0} F_1^{-1} \Sigma_{1|0} \\
&= \Sigma_{1|0} - \frac{\Sigma_{1|0}^2}{\Sigma_{1|0} + \sigma_\eta^2} \\
&= \frac{\Sigma_{1|0}^2 + \Sigma_{1|0}\sigma_\eta^2 - \Sigma_{1|0}^2}{\Sigma_{1|0} + \sigma_\eta^2} \\
&= \frac{\Sigma_{1|0}\sigma_\eta^2}{\Sigma_{1|0} + \sigma_\eta^2} \frac{\Sigma_{1|0}^{-1}\sigma_\eta^{-2}}{\Sigma_{1|0}^{-1}\sigma_\eta^{-2}} \\
&= \frac{1}{\Sigma_{1|0}^{-1} + \sigma_\eta^{-2}},
\end{aligned}$$

which is of the same form as the initial updating example.

One-step Ahead Prediction

- Suppose we want to forecast the value of μ_2 given data up to y_1.
- We want the one-step ahead predictive distribution: $p(\mu_2|y_1;\boldsymbol{\theta})$
- The predictive distribution $p(\mu_2|y_1;\boldsymbol{\theta})$ describes our uncertainty about μ_2 given we observe y_1.
- Our model says μ_2 and μ_1 are related:

$$\mu_2 = \mu_1 + \varepsilon_2, \qquad \varepsilon_2 \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

- This defines the (Markov) transition density

$$p(\mu_2|\mu_1;\boldsymbol{\theta}) = \mathsf{N}(\mu_1, \sigma_\varepsilon^2)$$

One-step Ahead Prediction: local level model

- To calculate $p(\mu_2|y_1;\boldsymbol{\theta})$, we integrate out μ_1 by

$$p(\mu_2|y_1;\boldsymbol{\theta}) = \int p(\mu_2|\mu_1;\boldsymbol{\theta})p(\mu_1|y_1;\boldsymbol{\theta})d\mu_1$$

- Since all distributions are Gaussian, integral can be solved analytically

$$p(\mu_2|y_1;\boldsymbol{\theta}) = \mathsf{N}(\mu_{2|1}, \Sigma_{2|1})$$

  where the mean and variance are

$$\mu_{2|1} = \mu_{1|1} \qquad \Sigma_{2|1} = \Sigma_{1|1} + \sigma_\varepsilon^2$$

- We just update the sufficient statistics.
- At time $t = 2$, this is our new prior: $p(\mu_2|y_1;\boldsymbol{\theta}) = \mathsf{N}(\mu_{2|1}, \Sigma_{2|1})$.

Filtering and one-step ahead prediction

- For $t = 2, \ldots, T$, we recursively repeat the two steps:

  Filtering

$$p(\mu_t | y_1, \ldots, y_t; \boldsymbol{\theta}) = \frac{p(y_t | \mu_t; \boldsymbol{\theta}) p(\mu_t | y_{1:t-1}; \boldsymbol{\theta})}{p(y_{1:t}; \boldsymbol{\theta})}$$

  One-step ahead prediction

$$p(\mu_{t+1} | y_{1:t}; \boldsymbol{\theta}) = \int p(\mu_{t+1} | \mu_t; \boldsymbol{\theta}) p(\mu_t | y_{1:t}; \boldsymbol{\theta}) d\mu_t$$

- The predictive distribution $p(\mu_{t+1} | y_{1:t}; \boldsymbol{\theta})$ is the prior at the next iteration, i.e. at time $t + 1$

The Kalman filter for the Local Level Model

- The Kalman filter recursively calculates these two steps
- Start with the initial conditions $\mu_{1|0}$ and $\Sigma_{1|0}$
- For $t = 1, \ldots, T$

$$
\begin{aligned}
v_t &= y_t - \mu_{t|t-1}, \\
F_t &= \Sigma_{t|t-1} + \sigma_\eta^2, \\
K_t &= \Sigma_{t|t-1} / F_t, \\
\mu_{t|t} &= \mu_{t|t-1} + K_t v_t, \qquad \text{Filter step} \\
\Sigma_{t|t} &= \Sigma_{t|t-1} - K_t \Sigma_{t|t-1}
\end{aligned}
$$

$$
\begin{aligned}
\mu_{t+1|t} &= \mu_{t|t}, \qquad \text{Prediction step} \\
\Sigma_{t+1|t} &= \Sigma_{t|t} + \sigma_\varepsilon^2.
\end{aligned}
$$

Filtering and one-step ahead prediction

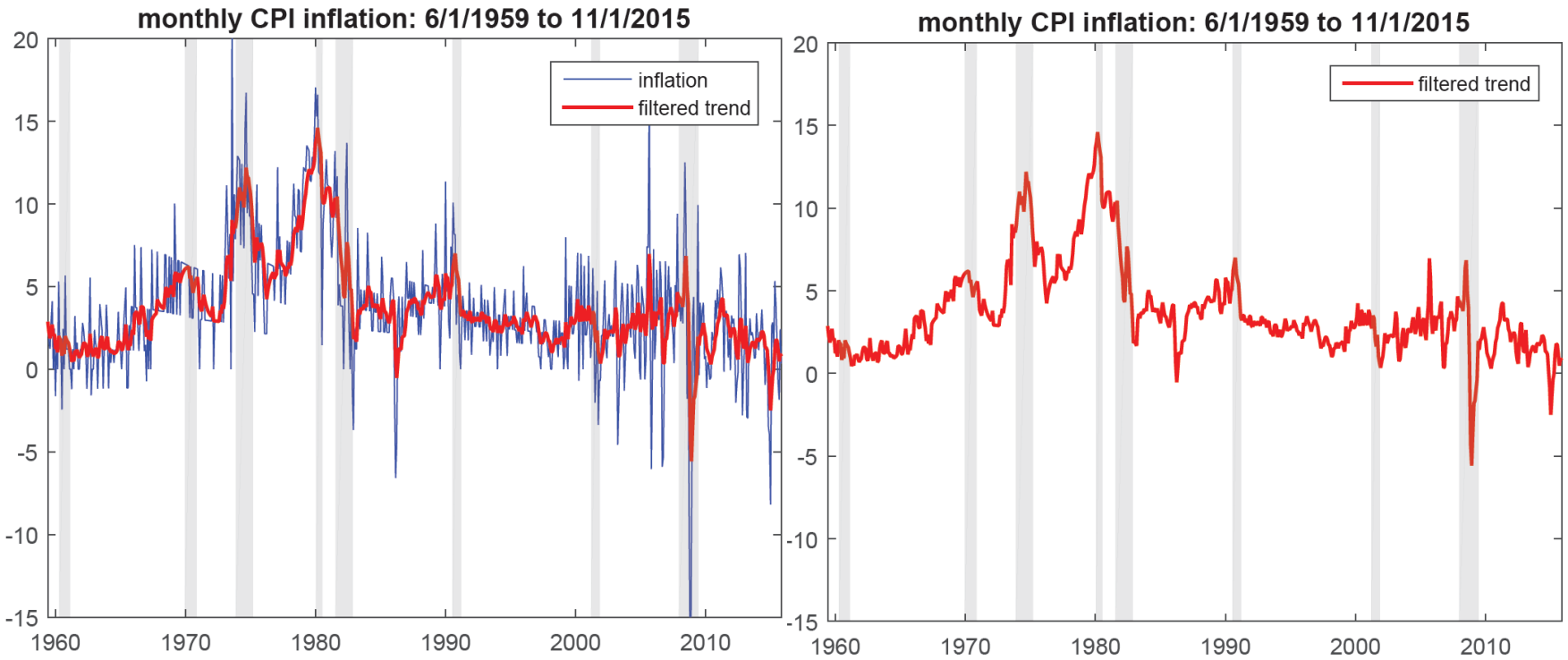
- These distributions describe our uncertainty about μ_t conditional on different information sets.
- All distributions are Gaussian!
- Only calculate means and covariance matrices (sufficient statistics)
- The filtering distribution

$$p(\mu_t|y_1,\ldots,y_t;\boldsymbol{\theta}) = \mathsf{N}(\mu_{t|t},\Sigma_{t|t}) \qquad t=1,\ldots,T$$

- One-step ahead predictive distribution

$$p(\mu_{t+1}|y_1,\ldots,y_t;\boldsymbol{\theta}) = \mathsf{N}(\mu_{t+1|t},\Sigma_{t+1|t}) \qquad t=0,\ldots,T$$

U.S. inflation



U.S. inflation, monthly continuously compounded. Left: inflation & filtered trend. Right: filtered trend.

The State Space Model (SSM)

Linear, Gaussian State Space Models

- Linear, Gaussian state space models are widely applicable
- For example, the following models can be placed in **state space form**
  1. local level model
  2. AR(1) observed in noise
  3. ARMA(p, q) models
  4. VAR(p) models
  5. Linear regression with serially correlated errors
  6. Time-varying parameter models
  7. Structural time series models
  8. Many more!
- First, we need to generalize the model.

Linear, Gaussian state space models

Definition

A **linear, Gaussian state space model** has observation equation

$$y_t = Z_t \alpha_t + d_t + \eta_t \qquad \eta_t \sim \mathrm{N}(0, H_t)$$

state transition equation

$$\alpha_{t+1} = T_t \alpha_t + c_t + R_t \varepsilon_{t+1} \qquad \varepsilon_{t+1} \sim \mathrm{N}(0, Q_t)$$

and initial conditions

$$\alpha_1 \sim \mathrm{N}(a_{1|0}, P_{1|0}).$$

- The **state variable** or **state vector** is α_t.
- The **system matrices** $Z_t, d_t, H_t, T_t, c_t, R_t, Q_t$ are often time invariant: Z, d, H, T, c, R, Q.

Local level model

Consider the local level model

$$y_t = \mu_t + \eta_t, \qquad \eta_t \sim \mathsf{N}(0, \sigma_\eta^2)$$
$$\mu_{t+1} = \mu_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

The model can be written in state space form as:

$$\alpha_t = \mu_t, \qquad Z = 1 \qquad d = 0 \qquad H = \sigma_\eta^2$$

$$T = 1 \qquad c = 0 \qquad R = 1 \qquad Q = \sigma_\varepsilon^2$$

AR(1) observed in noise

$$y_t = \mu_t + \eta_t \qquad \eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$$
$$\mu_{t+1} = \phi_0 + \phi_1 \mu_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

The model can be written in state space form as:

$$\alpha_t = \mu_t \qquad Z = 1 \qquad d = 0 \qquad H = \sigma_\eta^2$$
$$T = \phi_1 \qquad c = \phi_0 \qquad R = 1 \qquad Q = \sigma_\varepsilon^2$$

(Note: the local level model sets $\phi_0 = 0$ and $\phi_1 = 1$.)

ARMA(3,2) (version 1)

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \phi_3(y_{t-3} - \mu) + \varepsilon_t$$
$$+ \vartheta_1\varepsilon_{t-1} + \vartheta_2\varepsilon_{t-2} \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

Let $\phi_0 = (1 - \phi_1 - \phi_2 - \phi_3)\mu$. A state space form is:

$$Z = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad d = 0 \qquad H = 0 \qquad Q = \sigma_\varepsilon^2$$

$$\alpha_t = \begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \varepsilon_t \\ \varepsilon_{t-1} \end{pmatrix} \quad T = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \vartheta_1 & \vartheta_2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad c = \begin{pmatrix} \phi_0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad R = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

ARMA(3,2) (version 2)

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \phi_3(y_{t-3} - \mu) + \varepsilon_t$$
$$+ \vartheta_1\varepsilon_{t-1} + \vartheta_2\varepsilon_{t-2} \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

Let $\phi_0 = (1 - \phi_1 - \phi_2 - \phi_3)\mu$. An alternative state space form is:

$$Z = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \qquad d = 0 \qquad H = 0 \qquad Q = \sigma_\varepsilon^2$$

$$\alpha_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \phi_3 y_{t-2} + \vartheta_1\varepsilon_t + \vartheta_2\varepsilon_{t-1} \\ \phi_3 y_{t-1} + \vartheta_2\varepsilon_t \end{pmatrix}$$

$$T = \begin{pmatrix} \phi_1 & 1 & 0 \\ \phi_2 & 0 & 1 \\ \phi_3 & 0 & 0 \end{pmatrix} \qquad c = \begin{pmatrix} \phi_0 \\ 0 \\ 0 \end{pmatrix} \qquad R = \begin{pmatrix} 1 \\ \vartheta_1 \\ \vartheta_2 \end{pmatrix}$$

Note: the dimension of α_t is smaller than the last slide.

Remarks

- As the example of an ARMA$(3, 2)$ shows us, the state space form of a model is not unique.
- There are multiple ways to place the same model in state space form. The definition of the 'state vector' is not necessarily the same in each case.
- The ARMA$(3, 2)$ example also shows us that the state variable α_t is not always a latent variable.
- For examples on how to write ARMA(p, q) models in state space form; see Hamilton (1994) or Durbin and Koopman (2012).

VAR(p) in state-space form

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim \mathsf{N}(0, \Sigma_\varepsilon)$$

The model can be written in state space form as:

$$Z = \begin{pmatrix} I & 0 & \ldots & 0 \end{pmatrix} \qquad d = 0 \qquad H = 0 \qquad Q = \Sigma_\varepsilon$$

$$\alpha_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix}, \; T = \begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_p \\ I & 0 & \ldots & 0 \\ 0 & I & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix}, \; c = \begin{pmatrix} \Phi_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; R = \begin{pmatrix} I \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Linear regression with serially correlated errors

We can estimate optimally the β in a regression with AR(1) errors (you can use more complicated forms), as opposed to just accounting for it using HAC standard errors:

$$y_t = X_t\beta + \nu_t$$
$$\nu_{t+1} = \phi\nu_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

The model can be written in state space form as:

$$\alpha_t = \nu_t \qquad Z = 1 \qquad d_t = X_t\beta \qquad H = 0$$
$$T = \phi \qquad c = 0 \qquad R = 1 \qquad Q = \sigma_\varepsilon^2$$

Time-varying parameter models

Here, the slopes are changing over time! Note, as in standard OLS, X_t must be observed at time t

$$\begin{aligned} y_t &= X_t\beta_t + \eta_t & \eta_t &\sim \mathrm{N}(0, \Omega) \\ \beta_{t+1} &= \Phi_0 + \Phi_1\beta_t + \varepsilon_{t+1} & \varepsilon_t &\sim \mathrm{N}(0, \Sigma_\varepsilon) \end{aligned}$$

The model can be written in state space form as:

$$\alpha_t = \beta_t \qquad Z_t = X_t \qquad d = 0 \qquad H = \Omega$$

$$T = \Phi_1 \qquad c = \Phi_0 \qquad R = I \qquad Q = \Sigma_\varepsilon$$

Note: a special case of this is a CAPM with time-varying β.

Missing data

Let's say that at time t some elements in

$$y_t = Z_t \alpha_t + d_t + \eta_t \qquad \eta_t \sim \mathrm{N}(0, H_t)$$

are missing.

- Simply set the variance of each missing observed component in H_t to infinity (in practice, a really big number), and zero its covariances. Alternatively, drop those rows of y_t, Z_t, d_t, and H_t in the update so that F_t stays nonsingular.

In Bayesian updating, if a signal has infinite variance of noise, the weight on that signal is zero.

- Next, when we see the updating equations, you will see H_t^{-1} arise which accounts for this effect

Predictive systems

We can use the SSM also on a cross-section for a big-data implementation

1. Let $y_t = R_t^e$ be vector of $N \times 1$ of firms. Can even make this $N_t \times 1$, just make sure Z_t matrix reflects this at each t.
2. Let the i'th row of Z_t be $Z_t^{(i)} = \begin{bmatrix} 1 & X'_{i,t-1} \end{bmatrix}$, where $X_{i,t-1}$ is a $K \times 1$ vector of characteristics

Now, there are $K+1$ latent variables, where the loading on the last K depends on the associated characteristics of the firm at time t.

Be careful of specification of H_t and Q_t, especially for interpretation of estimated standard errors

- But, can view this as a quasi-maximum likelihood approach where we use out-of-sample performance instead of standard errors to evaluate model performance

Important things to remember

- There is more than one way to place a model in **state space form**
- Consequently, the definition of the '**state variable**' may change depending on how you do it.
- No matter how you place a model in **state space form** some things will not change:
  1. forecasts of future data y_{t+h} for $h > 0$
  2. the log-likelihood of the model

The Kalman filter calculates....

- The filtering distribution

$$p(\alpha_t|y_{1:t};\boldsymbol{\theta}) = \frac{p(y_t|\alpha_t;\boldsymbol{\theta})p(\alpha_t|y_{1:t-1}\boldsymbol{\theta})}{p(y_{1:t};\boldsymbol{\theta})}$$

- One-step ahead predictive distribution

$$p(\alpha_{t+1}|y_{1:t};\boldsymbol{\theta}) = \int p(\alpha_{t+1}|\alpha_t;\boldsymbol{\theta})p(\alpha_t|y_{1:t};\boldsymbol{\theta})d\alpha_t$$

- All distributions are Gaussian! We only need their means and covariance matrices (their sufficient statistics).

- Our notation:

$$p(\alpha_t|y_{1:t};\boldsymbol{\theta}) = \mathrm{N}(a_{t|t}, P_{t|t})$$
$$p(\alpha_{t+1}|y_{1:t};\boldsymbol{\theta}) = \mathrm{N}(a_{t+1|t}, P_{t+1|t})$$

The Kalman Filter

- The Kalman filter recursively calculates these two steps.
- Start with the initial conditions $a_{1|0}$ and $P_{1|0}$.
- For $t = 1, \ldots, T$

$$
\begin{aligned}
v_t &= y_t - Z_t a_{t|t-1} - d_t, \\
F_t &= Z_t P_{t|t-1} Z_t' + H_t, \\
K_t &= P_{t|t-1} Z_t' F_t^{-1}, \\
a_{t|t} &= a_{t|t-1} + K_t v_t, \qquad \text{Filter step} \\
P_{t|t} &= P_{t|t-1} - K_t Z_t P_{t|t-1},
\end{aligned}
$$

$$
\begin{aligned}
a_{t+1|t} &= T_t a_{t|t} + c_t, \qquad \text{Prediction step} \\
P_{t+1|t} &= T_t P_{t|t} T_t' + R_t Q_t R_t'
\end{aligned}
$$

Prediction and then filtering

- Some researchers reverse the order of the steps.
- Start with the initial (filtering) conditions $a_{0|0}$ and $P_{0|0}$.
- For $t = 1, \ldots, T$

$$
\begin{aligned}
a_{t|t-1} &= T_{t-1} a_{t-1|t-1} + c_{t-1}, && \text{Prediction step} \\
P_{t|t-1} &= T_{t-1} P_{t-1|t-1} T_{t-1}' + R_{t-1} Q_{t-1} R_{t-1}'
\end{aligned}
$$

$$
\begin{aligned}
v_t &= y_t - Z_t a_{t|t-1} - d_t, \\
F_t &= Z_t P_{t|t-1} Z_t' + H_t, \\
K_t &= P_{t|t-1} Z_t' F_t^{-1}, \\
a_{t|t} &= a_{t|t-1} + K_t v_t, && \text{Filter step} \\
P_{t|t} &= P_{t|t-1} - K_t Z_t P_{t|t-1},
\end{aligned}
$$

The Kalman Predictor

- **Kalman predictor**: The filtered values $a_{t|t}$ and $P_{t|t}$ are never calculated.
- Start with the initial conditions $a_{1|0}$ and $P_{1|0}$.
- For $t = 1, \ldots, T$

$$
\begin{aligned}
v_t &= y_t - Z_t a_{t|t-1} - d_t, \\
F_t &= Z_t P_{t|t-1} Z_t' + H_t, \\
M_t &= T_t P_{t|t-1} Z_t' F_t^{-1}, \\
L_t &= T_t - M_t Z_t, \\
a_{t+1|t} &= T_t a_{t|t-1} + c_t + M_t v_t, \\
P_{t+1|t} &= T_t P_{t|t-1} L_t' + R_t Q_t R_t'
\end{aligned}
$$

- Computationally faster because $a_{t|t}$ and $P_{t|t}$ are not calculated.

Initializing the Kalman filter

- We need values for $a_{1|0}$ and $P_{1|0}$ to start the Kalman filtering recursions.
- There are many suggestions in the literature for how to choose these values. **You should think of them as part of your model!**
- In practice, we encounter two common situations:
  1. α_t is stationary
  2. α_t is non-stationary
- Case 1 is pretty easy. Case 2 is not.
- In the literature, $a_{1|0}$ and $P_{1|0}$ are often called **initial conditions**.

Stationarity of the state equation

- Suppose the state equation α_t is a stationary process.
- This means there exists a stationary (marginal) distribution $\mathrm{N}(\mu_\alpha, V_\alpha)$.
- Stationarity means that all the eigenvalues of the matrix T are inside the unit circle.

Initializing the Kalman filter (stationary case)

- Suppose the state equation α_t is a stationary process.
- Let $\mathrm{N}(\mu_\alpha, V_\alpha)$ denote the stationary distribution of α_t.
- Taking unconditional expectations, we find

$$E[\alpha_{t+1}] = TE[\alpha_t] + c + RE[\varepsilon_t]$$
$$\mu_\alpha = T\mu_\alpha + c$$

$$\Rightarrow \mu_\alpha = (I - T)^{-1} c$$

- If we use the stationary distribution, we set the mean to be

$$a_{1|0} = \mu_\alpha$$

Initializing the Kalman filter (stationary case)

- Let $\mathrm{N}(\mu_\alpha, V_\alpha)$ denote the stationary distribution of α_t.

- Taking unconditional variances, we find

$$\begin{aligned} V[\alpha_{t+1}] &= TV[\alpha_t]T' + RV[\varepsilon_t]R' \\ V_\alpha &= TV_\alpha T' + RQR' \\ \mathrm{vec}(V_\alpha) &= \mathrm{vec}(TV_\alpha T') + \mathrm{vec}(RQR') \\ \mathrm{vec}(V_\alpha) &= (T \otimes T)\mathrm{vec}(V_\alpha) + \mathrm{vec}(RQR') \end{aligned}$$

$$\Rightarrow \mathrm{vec}(V_\alpha) = [I - (T \otimes T)]^{-1}\mathrm{vec}(RQR')$$

- If we use the stationary distribution, we set the covariance matrix as

$$P_{1|0} = V_\alpha$$

Initializing the Kalman filter (non-stationary)

- In some models of interest, elements of the state vector α_t are non-stationary.
- These models have unit roots; e.g. the local level model.
- The stationary distribution of α_t does not exist.
- The simple way to initialize the Kalman filter is set the variance $P_{1|0}$ to a really large number; e.g. 10^4
- Koopman (1997) gives an exact initialization. Effectively, it calculates a conditional log-likelihood function that drops the initial parts of the likelihood. Not easy to implement though. See also Chapter 5 of Durbin and Koopman (2012).

Impact of the initial conditions

- If the model is stationary, the **initial conditions** $a_{1|0}$ and $P_{1|0}$ typically do not have a large influence on the results.
- Filtered estimates $a_{t|t}$ and $P_{t|t}$ will converge to the same thing (they are equal) even if we start the Kalman filter from different initial conditions!!
- This is due to the stationarity of the model.
- The early estimates $a_{t|t}$ and $P_{t|t}$ will be different during the first few iterations: $t < 20$ or so.
- For non-stationary models, the initial conditions can have an impact if the overall (time-series) sample size T is small.

Smoothing distributions

- During the forward pass of the Kalman filter we calculate the filtering and one-step ahead predictive distributions.

$$p(\alpha_t|y_{1:t};\boldsymbol{\theta}) = \mathsf{N}(a_{t|t}, P_{t|t})$$
$$p(\alpha_{t+1}|y_{1:t};\boldsymbol{\theta}) = \mathsf{N}(a_{t+1|t}, P_{t+1|t})$$

  for $t = 1, \ldots, T$

- These distributions describe our uncertainty about the state α_t conditional on different information sets.

- For many time series models, there is still uncertainty about the state vector α_t even after we observe **all** the data.

$$p(\alpha_t|y_1, \ldots, y_T;\boldsymbol{\theta})$$

- This is called the **smoothed distribution**.

Smoothing distributions

- At time $t = T$, the filtered and smoothed distributions are equal!

$$p(\alpha_T | y_{1:T}; \boldsymbol{\theta}) = \mathrm{N}(a_{T|T}, P_{T|T})$$

- At $t = T$, we know the mean $a_{T|T}$ and covariance matrix $P_{T|T}$.
- We can write the sufficient statistics $a_{t|T}, P_{t|T}$ as a recursive function of $a_{t+1|T}, P_{t+1|T}$.
- The **Kalman smoother** recursively calculates the smoothing distributions backwards

$$p(\alpha_t | y_{1:T}; \boldsymbol{\theta}) = \mathrm{N}(a_{t|T}, P_{t|T})$$

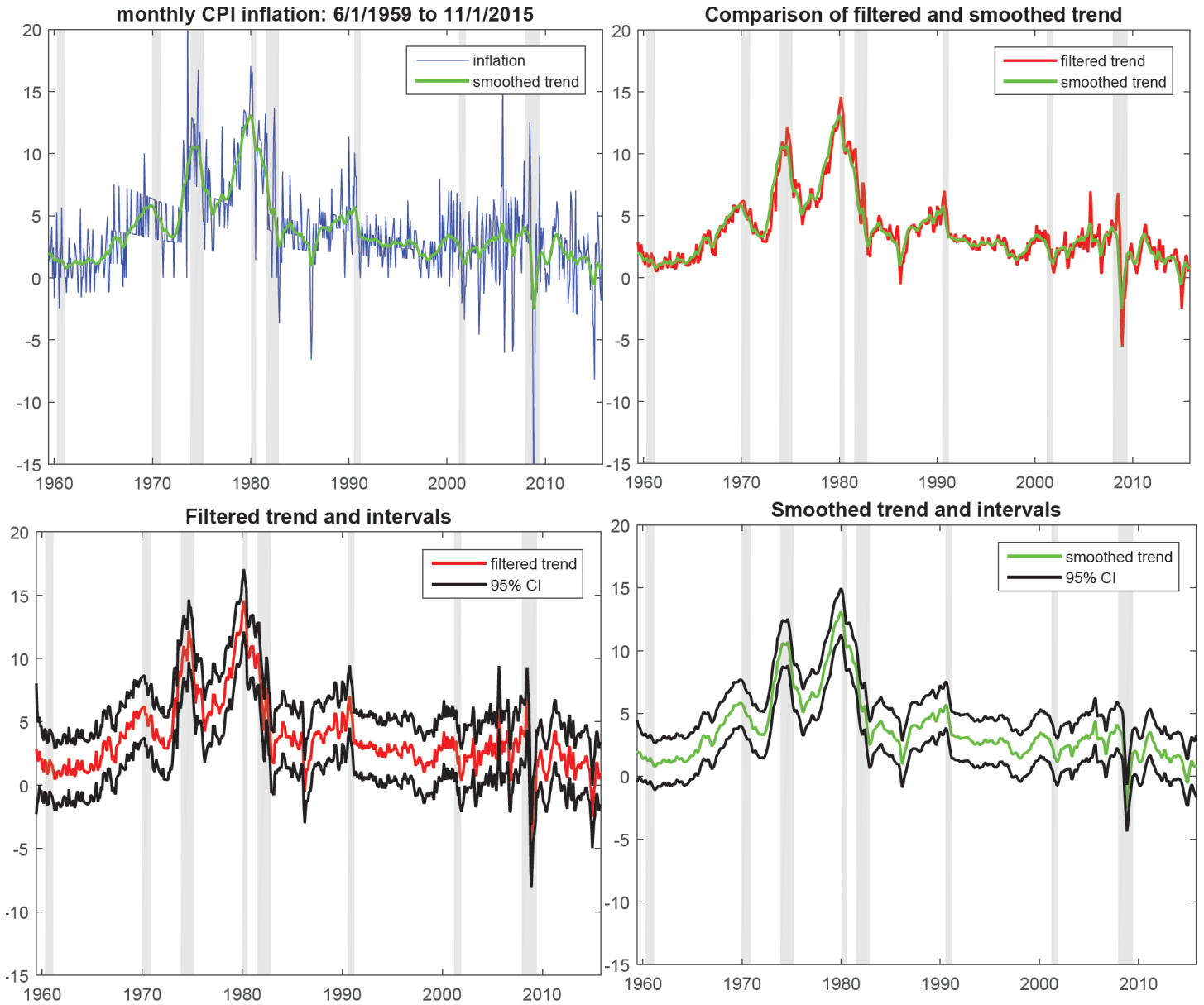
for $t = T - 1, \ldots, 1$

Kalman smoother (Rauch-Tung-Striebel)

- Run the Kalman filter forward in time for $t = 1, \ldots, T$.
- Store the quantities $\left\{ a_{t|t-1}, P_{t|t-1}, v_t, F_t, L_t \right\}_{t=1}^{T}$.
- Set $r_{T+1} = 0$ and $N_{T+1} = 0$
- For $t = T, \ldots, 1$

$$
\begin{aligned}
r_t &= Z_t' F_t^{-1} v_t + L_t' r_{t+1} \\
N_t &= Z_t' F_t^{-1} Z_t + L_t' N_{t+1} L_t \\
a_{t|T} &= a_{t|t-1} + P_{t|t-1} r_t \\
P_{t|T} &= P_{t|t-1} - P_{t|t-1} N_t P_{t|t-1}
\end{aligned}
$$

Example: U. S. inflation



U.S. inflation, monthly continuously compounded.

Comments on the Kalman smoother

- There are different versions of the **Kalman smoother**.
- All of them compute the means $a_{t|T}$ and covariance matrices $P_{t|T}$.
- They differ depending on what you store in computer memory on the forward pass of the Kalman filter.
- In the statistics/econometrics/engineering literature, the terms '**filtering**' and '**smoothing**' are used in different ways.
  - in state space models, filtering means to use data only up to time t.
  - in state space models, smoothing means to use ALL the data T.
  - however, different parts of science use the terms '**filtering**' and '**smoothing**' to mean other things.

Forecasting

- We often want to forecast:
  1. the state variable h steps ahead: α_{t+h}
  2. future data h steps ahead: y_{t+h}
- We also want to characterize our uncertainty of these variables.
- This is easy to do in a state space model.

Forecasting the state variable

- To forecast the state variable α_t at time $t+h$, we need to calculate the predictive distribution.

$$p(\alpha_{t+h}|y_{1:t};\boldsymbol{\theta})$$

- Under the assumption that the errors η_t and ε_t are Gaussian, the predictive distribution is Gaussian.

$$p(\alpha_{t+h}|y_{1:t};\boldsymbol{\theta}) \;=\; \mathrm{N}(a_{t+h|t}, P_{t+h|t})$$

- We can calculate the mean $a_{t+h|t}$ and covariance matrix $P_{t+h|t}$.
- These can be calculated recursively: $t+1$, then $t+2$, then...$t+h$

Forecasting the state variable

- Let's assume that the model is time-invariant. The **system matrices** are constant

$$Z,\ d,\ H,\ T,\ c,\ R,\ Q$$

- At the end of the Kalman filter, we already have

$$p(\alpha_{t+1}|y_{1:t};\boldsymbol{\theta}) \ = \ \mathrm{N}(a_{t+1|t}, P_{t+1|t})$$

- The next predictive distribution is:

$$p(\alpha_{t+2}|y_{1:t};\boldsymbol{\theta}) \ = \ \int p(\alpha_{t+2}|\alpha_{t+1};\boldsymbol{\theta})p(\alpha_{t+1}|y_{1:t};\boldsymbol{\theta})d\alpha_{t+1}$$

- To calculate $a_{t+2|t}, P_{t+2|t}$, we apply the recursion

$$\begin{aligned} a_{t+2|t} &= Ta_{t+1|t} + c, \\ P_{t+2|t} &= TP_{t+1|t}T' + RQR' \end{aligned}$$

Forecasting the state variable

- To calculate, these quantities at longer horizons, we simply iterate

$$a_{t+h|t} = Ta_{t+h-1|t} + c,$$
$$P_{t+h|t} = TP_{t+h-1|t}T' + RQR'$$

- If the system matrices are time-varying, you need to know their future values:

$$H_{t+j}, T_{t+j}, c_{t+j}, R_{t+j}, Q_{t+j} \qquad j = 0, \ldots, h-1$$

Forecasting future data

- To forecast future data y_{t+h}, do the following:
  1. calculate the predictive distribution of α_{t+h} using the earlier recursion

$$a_{t+h|t} = Ta_{t+h-1|t} + c,$$
$$P_{t+h|t} = TP_{t+h-1|t}T' + RQR'$$

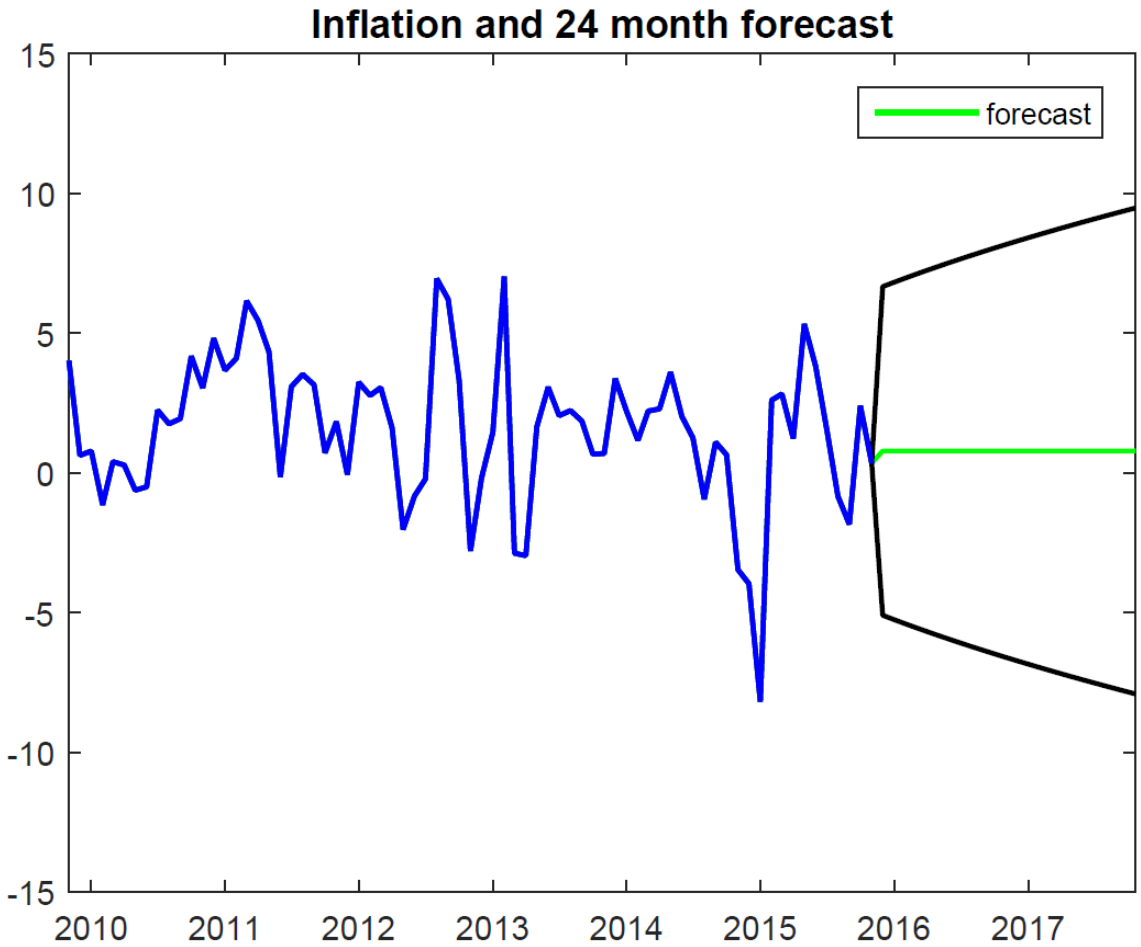
  2. calculate the predictive mean and variance of y_{t+h} via

$$\hat{y}_{t+h|t} = Za_{t+h|t} + d$$
$$F_{t+h|t} = ZP_{t+h|t}Z' + H$$

- If the system matrices are time-varying, you need to know their future values:

$$Z_{t+j}\ d_{t+j}\ H_{t+j}, T_{t+j}, c_{t+j}, R_{t+j}, Q_{t+j} \qquad j = 0, \ldots, h-1$$

Example: U.S. inflation and the local level model



U.S. inflation, monthly continuously compounded.

Example: U.S. inflation: ARMA(1,1)

Suppose we consider an ARMA(1,1) for inflation

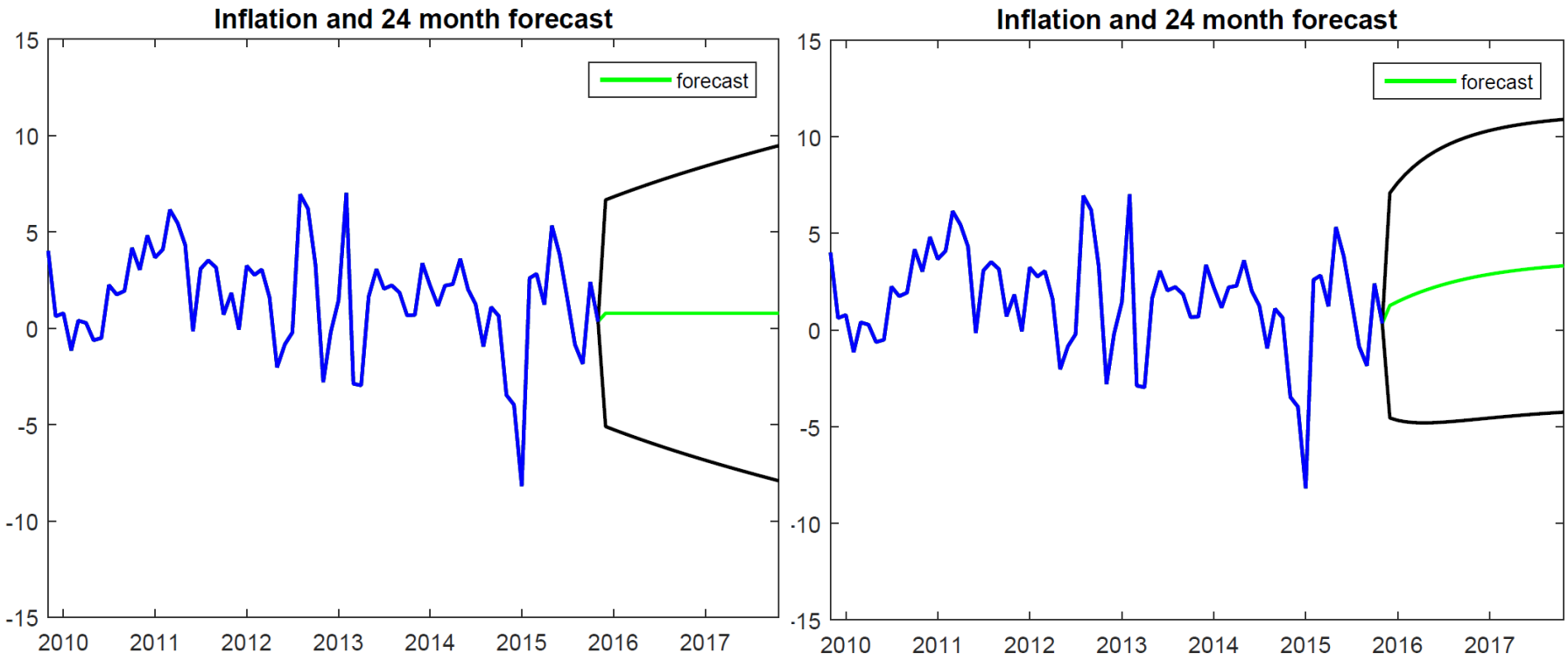
$$\pi_t = \mu + \phi_1(\pi_{t-1} - \mu) + \varepsilon_t + \theta_1 \varepsilon_{t-1} \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_\varepsilon^2)$$

I will use the state space form:

$$Z = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad d = \mu \qquad H = 0 \qquad Q = \sigma_\varepsilon^2$$

$$\alpha_t = \begin{pmatrix} \pi_t - \mu \\ \varepsilon_t \end{pmatrix} \quad T = \begin{pmatrix} \phi_1 & \theta_1 \\ 0 & 0 \end{pmatrix} \quad c = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad R = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Example: U.S. inflation



U.S. inflation, monthly continuously compounded.

Parameter estimation

- Thus far, we have assumed that we know the parameters $\boldsymbol{\theta}$.
- The log-likelihood of the model is the joint distribution of the data:

$$\ln p(y_1, y_2, \ldots, y_T|\boldsymbol{\theta}) = \sum_{t=2}^{T} \ln p(y_t|y_{1:t-1}; \boldsymbol{\theta}) + \ln p(y_1; \boldsymbol{\theta})$$

- The Kalman filter calculates the time t contribution to the log-likelihood

$$\ln p(y_t|y_{1:t-1}; \boldsymbol{\theta})$$

  at each iteration during the forward pass.

- We can maximize the log-likelihood (numerically) to estimate the parameters $\boldsymbol{\theta}$.

What is the likelihood?

- The likelihood at time t is just our predicted value of y_t given information up to time $t-1$

$$p(y_t|y_1,\ldots,y_{t-1};\boldsymbol{\theta})$$

- Under the assumption that the errors are Gaussian, this distribution is Gaussian. Calculate the mean and covariance matrix.

- In the last section, we just showed how to calculate the forecast of y_{t+h} given information up to time t!!!

- We need to calculate the forecast of y_t given information up to time $t-1$.

How to calculate the log-likelihood

- Start with the initial conditions $a_{1|0}$ and $P_{1|0}$
- Initialize the log-likelihood: $\ell_0 = 0$
- For $t = 1, \ldots, T$

$$
\begin{aligned}
v_t &= y_t - Z_t a_{t|t-1} - d_t, \\
F_t &= Z_t P_{t|t-1} Z_t' + H_t, \\
M_t &= T_t P_{t|t-1} Z_t' F_t^{-1}, \\
L_t &= T_t - M_t Z_t, \\
a_{t+1|t} &= T_t a_{t|t-1} + c_t + M_t v_t, \\
P_{t+1|t} &= T_t P_{t|t-1} L_t' + R_t Q_t R_t'
\end{aligned}
$$

$$
\ell_t = \ell_{t-1} - \frac{N}{2} \log(2\pi) - \frac{1}{2} \log |F_t| - \frac{1}{2} v_t' F_t^{-1} v_t
$$

Appendix: Updating from noisy signal – the details

Bayes Rule and Updates in Beliefs

Let's do some math!

- Start with two known distributions:

$$x \sim N\left(\mu_X, \sigma_X^2\right)$$
$$y|x \sim N\left(x, \sigma_{Y|X}^2\right)$$

- Here x corresponds to x_t and y to z_t in our slides at the beginning of the lecture

Bayes Rule and Updates in Beliefs

- We want to get to the distribution of $x|y$, so let's first multiply these two pdf's:

$$\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left\{\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}\frac{1}{\sqrt{2\pi\sigma_{Y|X}^2}}\exp\left\{\frac{(y-x)^2}{2\sigma_{Y|X}^2}\right\}$$

$$=\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}}\exp\left\{\frac{(x-\mu_X)^2}{2\sigma_X^2}+\frac{(y-x)^2}{2\sigma_{Y|X}^2}\right\}$$

$$=\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}}\exp\left\{\begin{array}{l}\frac{(x-\mu_X)^2}{2\sigma_X^2}\frac{\left(\sigma_X^{-2}+\sigma_{Y|X}^{-2}\right)^{-1}/\sigma_X^2}{\left(\sigma_X^{-2}+\sigma_{Y|X}^{-2}\right)^{-1}/\sigma_X^2}\cdots \\ +\frac{(y-x)^2}{2\sigma_{Y|X}^2}\frac{\left(\sigma_X^{-2}+\sigma_{Y|X}^{-2}\right)^{-1}/\sigma_{Y|X}^2}{\left(\sigma_X^{-2}+\sigma_{Y|X}^{-2}\right)^{-1}/\sigma_{Y|X}^2}\end{array}\right\}$$

Oh yeah... Algebra!

Continuing...

Define $k \equiv \left(\sigma_X^{-2} + \sigma_{Y|X}^{-2}\right)^{-1}$:

$$\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}} \exp\left\{ \frac{(x-\mu_X)^2}{2\sigma_X^2} \frac{k/\sigma_X^2}{k/\sigma_X^2} + \frac{(y-x)^2}{2\sigma_{Y|X}^2} \frac{k/\sigma_{Y|X}^2}{k/\sigma_{Y|X}^2} \right\}$$

$$= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}} \exp\left\{ \frac{\left(x^2 - 2x\mu_X + \mu_X^2\right) k/\sigma_X^2 + \left(y^2 - 2yx + x^2\right) k/\sigma_{Y|X}^2}{2k} \right\}$$

$$= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}} \exp\left\{ \frac{x^2 k/\sigma_X^2 - 2x\mu_X k/\sigma_X^2 + \mu_X^2 k/\sigma_X^2 + y^2 k/\sigma_{Y|X}^2 - 2yxk/\sigma_{Y|X}^2 + x^2 k/\sigma_{Y|X}^2}{2k} \right\}$$

$$= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}} \exp\left\{ \frac{x^2 k\left(\sigma_X^{-2} + \sigma_{Y|X}^{-2}\right) - 2x\left(yk/\sigma_{Y|X}^2 + \mu_X k/\sigma_X^2\right) + \mu_X^2 k/\sigma_X^2 + y^2 k/\sigma_{Y|X}^2}{2k} \right\}$$

Finger-lickin' !

Continuing... ...

Note that $k\left(\sigma_X^{-2}+\sigma_{Y|X}^{-2}\right)=1$. So:

$$\frac{1}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}}\exp\left\{\frac{x^2-2x\left(yk/\sigma_{Y|X}^2+\mu_X k/\sigma_X^2\right)+\mu_X^2 k/\sigma_X^2+y^2k/\sigma_{Y|X}^2}{2k}\right\}.$$

Next, complete the square:

$$\frac{1}{\sqrt{2\pi k}}\exp\left\{\frac{\left(x-\left(yk/\sigma_{Y|X}^2+\mu_X k/\sigma_X^2\right)\right)^2}{2k}\right\}$$
$$\times\frac{\sqrt{2\pi k}}{2\pi\sqrt{\sigma_X^2\sigma_{Y|X}^2}}\exp\left\{\frac{-\left(yk/\sigma_{Y|X}^2+\mu_X k/\sigma_X^2\right)^2+\mu_X^2 k/\sigma_X^2+y^2k/\sigma_{Y|X}^2}{2k}\right\}.$$

Note that the first line says $x|y$ is normally distributed with mean $\left(yk/\sigma_{Y|X}^2+\mu_X k/\sigma_X^2\right)$ and variance k.

- The second line is a constant (not a function of x), conditional on y. Since we only were given the distribution up to a proportion (recall the Bayes Rule equation), we can ignore it for our purposes.

Learning with Normal Distributions

In sum, we are looking for the distribution of x conditional on a data point, y.

We found that $x|y$ is normally distributed using Bayes Rule.

- The mean of this distribution is:

$$\begin{aligned} yk/\sigma^2_{Y|X} + \mu_X k/\sigma^2_X &= y\frac{\sigma^{-2}_{Y|X}}{\sigma^{-2}_X + \sigma^{-2}_{Y|X}} + \mu_X \frac{\sigma^{-2}_X}{\sigma^{-2}_X + \sigma^{-2}_{Y|X}} \\ &= y \times (1 \text{ - weight on prior}) + \mu_X \times (\text{weight on prior}) \end{aligned}$$

  Note that the more precise the signal is (the higher $\sigma^{-2}_{Y|X}$ is) and the less precise the prior is (the lower σ^{-2}_X is), the more weight is given to the signal when updating the mean belief about x.

- The variance is $k = \left(\sigma^{-2}_X + \sigma^{-2}_{Y|X}\right)^{-1} < \sigma^2_X$.

References

Burmeister, E., K. D. Wall, and J. D. Hamilton (1986). Estimation of unobserved expected monthly inflation using Kalman Filtering. Journal of Business and Economic Statistics 4(2), 147-160.

Durbin, J. and S. J. Koopman (2012). Time Series Analysis by State Space Methods (2 ed.). Oxford, UK: Oxford University Press.

Hamilton, J. D. (1994). Time Series Analysis. Princeton, NJ: Princeton University Press.

Kalman, R. (1960). A new approach to linear ltering and prediction problems. Journal of Basic Engineering, Transactions of the ASME 82(1), 35-46.

Koopman, S. J. (1997). Exact initial Kalman Filtering and smoothing for non-stationary time series models. Journal of the American Statistical Association 92(440), 1630-1638.

Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? Journal of Money, Credit, and Banking 39(1), 3-34.