Chapter Six

State Space Models and Filtering Techniques

Tingguo Zheng (zhengtg@gmail.com)
School of Economics & WISE & Chow, Xiamen University

(For Chow & SOE & WISE Msc and PhD Program, 2026-2027, Autumn Semester)

- **Reference:**
  - "Analysis of Financial Time Series", 2010, Tsay, Chapter 11.
  - Hamilton, 1994 "State-Space Models", Handbook of Econometrics, Volume 4, chapter 50.

- **Outline:**
  - State space representation
  - Kalman filtering for linear Gaussian SS models
  - Approximate filtering for nonlinear SS models
  - Mixed-frequency model and applications
  - Sequential Monte Carlo filtering*
  - Score-driven models

The main applications of state space models in macroeconomics and finance have dealt with the following topics.

- The extraction of signals such as trends and cycles in macroeconomic time series: see Watson (1986), Clark (1987), Harvey and Jaeger (1993), Hodrick and Prescott (1997), Morley, Nelson and Zivot (2003), Proietti (2006), Luati and Proietti (2011).

- Dynamic factor models, for the extraction of a single index of coincident indicators, see Stock and Watson (1989), Frale et al. (2011), and for large dimensional systems (Jungbacker, Koopman and van der Wel, 2011).

- Stochastic volatility models: see Shephard (2005) and Stock and Watson (2007) for applications to US inflation.

- The class of dynamic stochastic general equilibrium (DSGE) models: Sargent (1989), Fernandez-Villaverde and Rubio-Ramirez (2005), Smets and Wouters (2003), Fernandez-Villaverde (2010).

- Term structure of interest rate: Diebold and Li (2006)

- Mixed-frequency data models: mixed-frequency dynamic factor model by Mariano and Murasawa (2003), and mixed-frequency VAR model by Mariano and Murasawa (2010).

1 Linear Gaussian State Space Models

1.1 State space expression

The general representation of linear Gaussian state space models (Harvey, 1991) can be formulated as:

$$\underset{n\times 1}{\boldsymbol{y}_t} = \underset{n\times r}{\boldsymbol{H}_t}\ \underset{r\times 1}{\boldsymbol{\xi}_t} + \underset{n\times 1}{\boldsymbol{d}_t} + \underset{n\times 1}{\boldsymbol{u}_t} \tag{1}$$

$$\underset{r\times 1}{\boldsymbol{\xi}_t} = \underset{r\times r}{\boldsymbol{F}_t}\ \underset{r\times 1}{\boldsymbol{\xi}_{t-1}} + \underset{r\times 1}{\boldsymbol{c}_t} + \underset{r\times 1}{\boldsymbol{v}_t} \tag{2}$$

where the disturbances $\boldsymbol{u}_t$ and $\boldsymbol{v}_t$ are such that:

$$\boldsymbol{u}_t \sim N(\boldsymbol{0}, \underset{n\times n}{\boldsymbol{R}_t}), \quad \boldsymbol{v}_t \sim N(\boldsymbol{0}, \underset{r\times r}{\boldsymbol{Q}_t}), \quad \mathrm{Cov}(\boldsymbol{u}_t\boldsymbol{v}_t) = \boldsymbol{0}. \tag{3}$$

Equation (1) is the *observation equation* or *measurement equation*; and equation (2) is the *state equation* or *transition equation*.

It should be noticed that:

1. The covariance matrices $\boldsymbol{R}_t$ and $\boldsymbol{Q}_t$ may be not full rank and thus singular;

2. It is often assumed that the innovations $\boldsymbol{u}_t$ and $\boldsymbol{v}_t$ are not correlated. If they are correlated, the reader are referred to another state space form advocated by Koopman, Shephard and Doornik (1999, Statistical algorithm for models in state space using SsfPack 2.2) and Durbin and Koopman (2002)'s book.

Time-invariant parametric models

In particular, if all matrices do not depend deterministically on t the state space system is called *time invariant*. This can be rewritten as

$$\underset{n\times 1}{\boldsymbol{y}_t} = \underset{n\times r}{\boldsymbol{H}}\,\underset{r\times 1}{\boldsymbol{\xi}_t} + \underset{n\times 1}{\boldsymbol{d}} + \underset{n\times 1}{\boldsymbol{u}_t} \tag{4}$$

$$\underset{r\times 1}{\boldsymbol{\xi}_t} = \underset{r\times r}{\boldsymbol{F}}\,\underset{r\times 1}{\boldsymbol{\xi}_{t-1}} + \underset{r\times 1}{\boldsymbol{c}} + \underset{r\times 1}{\boldsymbol{v}_t} \tag{5}$$

$$\mathrm{Cov}(\boldsymbol{u}_t\boldsymbol{u}_t) = \underset{n\times n}{\boldsymbol{R}}, \quad \mathrm{Cov}(\boldsymbol{v}_t\boldsymbol{v}_t) = \underset{r\times r}{\boldsymbol{Q}}, \quad \mathrm{Cov}(\boldsymbol{u}_t\boldsymbol{v}_t) = \boldsymbol{0}.$$

- The state space representation is completed by specifying the behavior of the *initial state* $\boldsymbol{\xi}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{P}_0)$.

- Under the weak stationary assumption of $\boldsymbol{y}_t$, the *unconditional mean* of $\boldsymbol{\xi}_t$, $\boldsymbol{a}_0$, may be determined using $\boldsymbol{a}_0 = E[\boldsymbol{\xi}_t] = (\boldsymbol{I}_r - \boldsymbol{F})^{-1}\boldsymbol{c}$.

- Similarly, the *unconditional variance* of $\boldsymbol{\xi}_t$, $\mathrm{Var}(\boldsymbol{\xi}_0)$ may be determined analytically using

$$\mathrm{vec}(\boldsymbol{P}_0) = (\boldsymbol{I}_{r^2} - \boldsymbol{F} \otimes \boldsymbol{F})^{-1}\mathrm{vec}(\boldsymbol{Q}).$$

  Here $\mathrm{vec}(\boldsymbol{P}_0)$ is an $r^2 \times 1$ column vector. It can be easily *reshaped to form* the $r \times r$ matrix $\boldsymbol{P}_0$.

Examples

(1) AR(p) process

Consider the AR(p) process

$$y_t = \mu + x_t, \quad x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

One way to represent AR(p) in state space form is as follows. Define $\boldsymbol{\xi}_t = (x_t, x_{t-1}, \ldots, x_{t-p+1})'$ so that the transition equation for $\boldsymbol{\xi}_t$ becomes

$$\begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \cdot & \cdot & \phi_{p-1} & \phi_p \\ 1 & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \vdots & \ddots & \cdot \\ 0 & \cdot & \cdot & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The transition equation system matrices are

$$\boldsymbol{F} = \begin{bmatrix} \phi_1 & \cdot & \cdot & \phi_{p-1} & \phi_p \\ 1 & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \vdots & \ddots & \cdot \\ 0 & \cdot & \cdot & 1 & 0 \end{bmatrix}, \quad \boldsymbol{c} = \mathbf{0}_{p\times 1}, \quad \boldsymbol{Q} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The measurement equation is $y_t = \mu + (1, 0, \ldots, 0)\boldsymbol{\xi}_t$, which implies that

$$\boldsymbol{H} = (1, 0, \ldots, 0), \quad \boldsymbol{d} = \mu, \quad \boldsymbol{R} = 0.$$

(2) ARMA(p, q) model

The general ARMA(p, q) model

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

may be put in a state space form in the following way. Let $m = \max(p, q+1)$ and rewrite the ARMA(p, q) model as

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_m y_{t-m} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_{m-1} \varepsilon_{t-m+1},$$

where some of the AR or MA coefficients will be zero unless $p = q + 1$.

Harvey (1994, Section 4.4) provides a state-space form with an m-dimensional state vector $\boldsymbol{\xi}_t$, the first element of which is y_t, that is $\xi_{1t} = y_t$. The other elements of $\boldsymbol{\xi}_t$ are obtained recursively.

For the ARMA$(m, m-1)$ model, we have

$$\begin{aligned}\xi_{1,t+1} = y_{t+1} &= \phi_1 y_t + \sum_{j=2}^{m} \phi_j y_{t+1-j} + \sum_{j=1}^{m-1} \theta_j \varepsilon_{t+1-j} + \varepsilon_{t+1} \\ &\equiv \phi_1 \xi_{1t} + \xi_{2t} + \eta_t,\end{aligned}$$

where $\xi_{2t} = \sum_{j=2}^{m} \phi_j y_{t+1-j} + \sum_{j=1}^{m-1} \theta_j \varepsilon_{t+1-j}$, $\eta_t = \varepsilon_{t+1}$, and as defined earlier $\xi_{1t} = y_t$.

Focusing on $\xi_{2,t+1}$, we have

$$\begin{aligned}
\xi_{2,t+1} &= \sum_{j=2}^{m} \phi_j y_{t+2-j} + \sum_{j=1}^{m-1} \theta_j \varepsilon_{t+2-j} \\
&= \phi_2 y_t + \sum_{j=3}^{m} \phi_j y_{t+2-j} + \sum_{j=2}^{m-1} \theta_j \varepsilon_{t+2-j} + \theta_1 \varepsilon_{t+1} \\
&\equiv \phi_2 \xi_{1t} + \xi_{3t} + \theta_1 \eta_t
\end{aligned}$$

where $\xi_{3t} = \sum_{j=3}^{m} \phi_j y_{t+2-j} + \sum_{j=2}^{m-1} \theta_j \varepsilon_{t+2-j}$.

Next, considering $\xi_{3,t+1}$, we have

$$\begin{aligned}
\xi_{3,t+1} &= \sum_{j=3}^{m} \phi_j y_{t+3-j} + \sum_{j=2}^{m-1} \theta_j \varepsilon_{t+3-j} \\
&= \phi_3 y_t + \sum_{j=4}^{m} \phi_j y_{t+3-j} + \sum_{j=3}^{m-1} \theta_j \varepsilon_{t+3-j} + \theta_2 \varepsilon_{t+1} \\
&= \phi_3 \xi_{1t} + \xi_{4t} + \theta_2 \eta_t,
\end{aligned}$$

where $\xi_{4t} = \sum_{j=4}^{m} \phi_j y_{t+2-j} + \sum_{j=3}^{m-1} \theta_j \varepsilon_{t+2-j}$.

Repeating the procedure, we have

$$\xi_{mt} = \sum_{j=m}^{m} \phi_j y_{t+2-j} + \sum_{j=m-1}^{m-1} \theta_j \varepsilon_{t+2-j} = \phi_m y_{t-1} + \theta_{m-1}\varepsilon_t.$$

Finally,

$$\begin{aligned} \xi_{m,t+1} &= \phi_m y_t + \theta_{m-1}\varepsilon_{t+1} \\ &= \phi_m \xi_{1t} + \theta_{m-1}\eta_t. \end{aligned}$$

Define

$$\boldsymbol{\xi}_t = \begin{bmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_m y_{t-m+1} + \theta_1 \varepsilon_t + \cdots + \theta_{m-1} \varepsilon_{t-m+2} \\ \vdots \\ \phi_m y_{t-1} + \theta_{m-1} \varepsilon_t \end{bmatrix}$$

and set

$$y_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{\xi}_t$$

$$\boldsymbol{\xi}_t = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1} & 0 & 0 & \cdots & 1 \\ \phi_m & 0 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{\xi}_{t-1} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-2} \\ \theta_{m-1} \end{bmatrix} \varepsilon_t$$

(3) Stochastic volatility

Define a simple *time-vary variance process* as:

$$y_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1) \tag{6}$$

In a stochastic volatility model, $h_t = \log \sigma_t^2$, the logarithm of the variance, behaves exactly as a stochastic process in the mean, such as *random walks* or *autoregression*.

$$h_t = c + \lambda h_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2) \tag{7}$$

Taking the square and the logarithm, Equation (6) can be reexpressed as:

$$\ln(y_t^2) = h_t + \ln(\varepsilon_t^2). \tag{8}$$

A *quasi-maximum likelihood* (QML) estimator has been proposed by Ruiz (1994) and Harvey et al. (1994) who approximate the measurement equation disturbance term by a *normal* random variable with the same mean

$$\mathrm{E}[\ln(\varepsilon_t^2)] = -1.2704$$

and the variance

$$\mathrm{Var}[\ln(\varepsilon_t^2)] = \pi^2/2.$$

Note: If x is a standard normal random variable, i.e. $x \sim N(0,1)$, then the variable y with log-squared transformation of x is defined by

$$y = \ln(x^2).$$

So the density of y is

$$y = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{x - \exp(x)}{2}\right\}$$

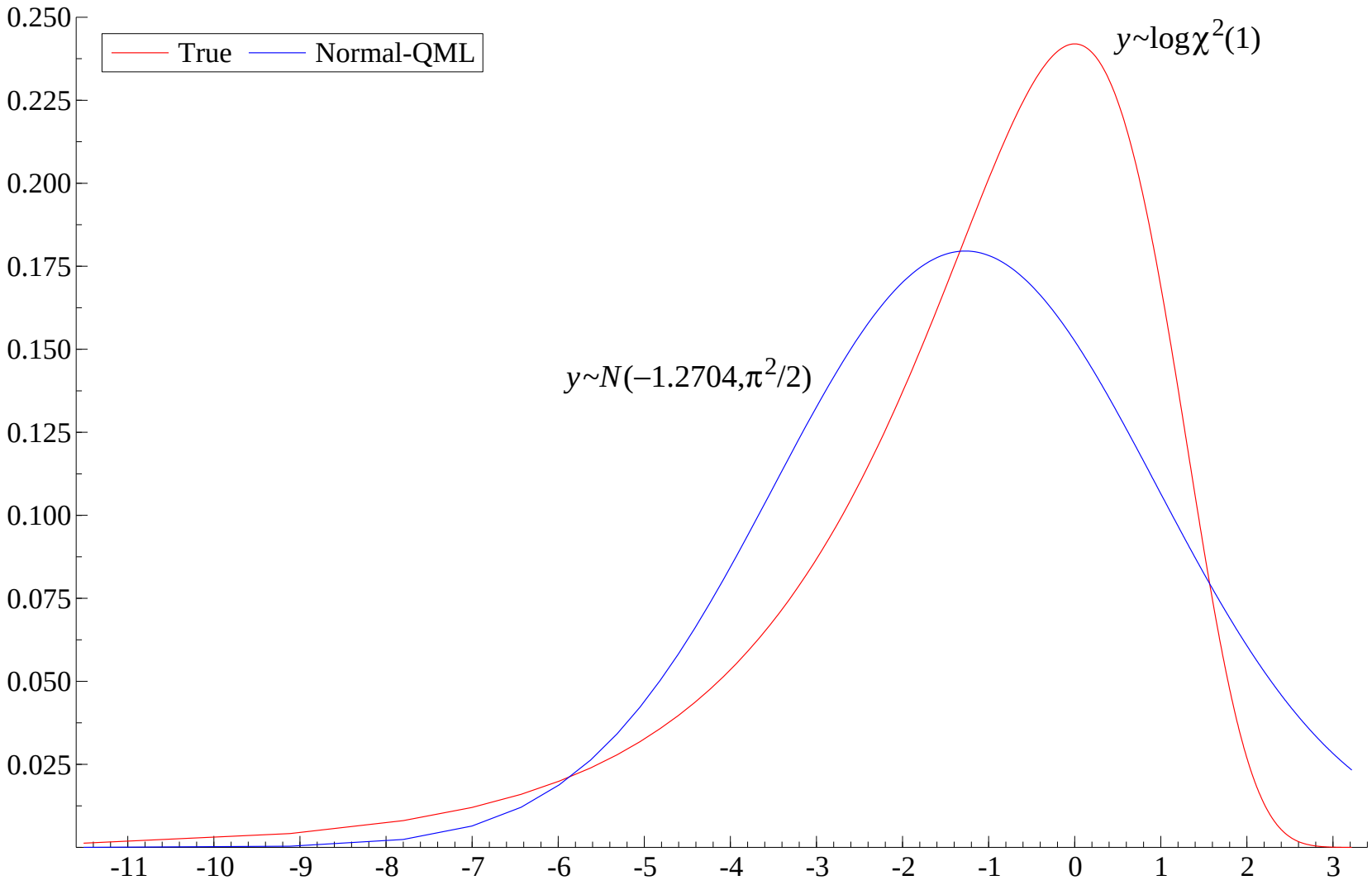


Figure 1 Log-chi-square density and its normal approximation.

(4) Time-varying coefficients

Specify a simple market model modified by using *time-varying coefficients*:

$$
\begin{aligned}
R_t &= \alpha_t + \beta_t R_{mt} + \varepsilon_t \\
\alpha_t &= \alpha_{t-1} + v_{1t} \\
\beta_t &= \beta_{t-1} + v_{2t}
\end{aligned}
\tag{9}
$$

where R_t return on an individual security, R_{mt} is return on the market, and the coefficients follow a *random walk*.

The *observation equation* is:

$$R_t = \begin{bmatrix} 1 & R_{mt} \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} + \varepsilon_t \tag{10}$$

The *state equation* is:

$$\begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} \tag{11}$$

(5) Dynamic factor models

Suppose we have two stationary variables, y_{1t}, and y_{2t} with a *common component* c_t:

$$y_{1t} = \gamma_1 c_t + z_{1t}, \tag{12}$$

$$y_{2t} = \gamma_2 c_t + z_{2t}, \tag{13}$$

$$c_t = \phi_1 c_{t-1} + v_t, \tag{14}$$

$$z_{it} = \alpha_i z_{i,t-1} + e_{it}, \quad i = 1, 2, \tag{15}$$

where e_{1t}, e_{2t}, and v_t are independent of one another.

Stock and Watson (1991) employ a general version of the above dynamic factor model to extract an *coincident index* from four *coincident economic variables*.

Then, a state-space representation of the above model is given by

Measurement Equation

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \gamma_1 & 1 & 0 \\ \gamma_2 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_t \\ z_{1t} \\ z_{2t} \end{bmatrix} \tag{16}$$

Transition Equation:

$$\begin{bmatrix} c_t \\ z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} c_{t-1} \\ z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} v_t \\ e_{1t} \\ e_{2t} \end{bmatrix} \tag{17}$$

(6) A common stochastic trend model

Hai, Mark, and Wu (1996) consider the following *common stochastic trend* model for the *spot* and *forward* exchange

$$y_{1t} = z_t + x_{1t}, \tag{18}$$

$$y_{2t} = z_t + x_{2t}, \tag{19}$$

$$z_t = z_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\mathcal{N}(0, \sigma_\varepsilon^2), \tag{20}$$

$$x_{1t} = \mu_1 + \phi_{11} x_{1,t-1} + \phi_{12} x_{2,t-1} + e_{1t} + \theta_{11} e_{1,t-1} + \theta_{12} e_{2,t-1}, \tag{21}$$

$$x_{2t} = \mu_2 + \phi_{21} x_{1,t-1} + \phi_{22} x_{2,t-1} + e_{2t} + \theta_{21} e_{1,t-1} + \theta_{22} e_{2,t-1}, \tag{22}$$

where

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim i.i.d.\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right),$$

z_t is a common stochastic trend and the stationary component x_{1t} and x_{2t} are assumed to be generated by a vector ARMA(1,1) process.

A state-space representation of the above model is given by

Measurement Equation

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ x_{1t} \\ x_{2t} \\ e_{1t} \\ e_{2t} \end{bmatrix} \tag{23}$$

Transition Equation:

$$\begin{bmatrix} z_t \\ x_{1t} \\ x_{2t} \\ e_{1t} \\ e_{2t} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu_1 \\ \mu_2 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \phi_{11} & \phi_{12} & \theta_{11} & \theta_{12} \\ 0 & \phi_{21} & \phi_{22} & \theta_{21} & \theta_{22} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_{t-1} \\ x_{1,t-1} \\ x_{2,t-1} \\ e_{1,t-1} \\ e_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ e_{1t} \\ e_{2t} \end{bmatrix}. \tag{24}$$

(7) GDP decomposition

A classic trend/cycle decomposition model is given as follows:

$$y_t = \tau_t + c_t \tag{25}$$

$$\tau_t = \tau_{t-1} + g_{t-1} + u_t \tag{26}$$

$$g_t = gc + \lambda g_{t-1} + w_t, \tag{27}$$

$$c_t = \phi_1 c_{t-1} + \phi_2 c_{t-2} + v_t \tag{28}$$

where y_t is log GDP; τ_t is its *trend component* follows a *random walk* with a *stochastic drift* or growth rate which is an autoregressive process; c_t is the *cycle component*.

When $\lambda = \sigma_w = 0$, this becomes the Watson (1986) model.

When $gc = 0$ and $\lambda = 1$, this becomes the Clark (1987) model.

Write equations (25) – (28) in the state space form, the observation equation is:

$$y_t = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tau_t \\ c_t \\ c_{t-1} \\ g_t \end{bmatrix} \tag{29}$$

The state equation is:

$$\begin{bmatrix} \tau_t \\ c_t \\ c_{t-1} \\ g_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & \varphi_1 & \varphi_2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ c_{t-1} \\ c_{t-2} \\ g_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ gc \end{bmatrix} + \begin{bmatrix} v_t \\ u_t \\ 0 \\ w_t \end{bmatrix} \tag{30}$$

(8) BSM model in STAMP

The basic structural model (BSM) is formulated in terms of trend, seasonal, and irregular components. It can be represented by

$$y_t = \tau_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim iidN(0, \sigma_\varepsilon^2), \tag{31}$$

where τ_t is a stochastic trend, γ_t is a stochastic seasonal component, and ε_t is an irregular component; see Harvey (1989).

The trend is specified in the following way:

$$\tau_t = \tau_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim iidN(0, \sigma_\eta^2) \tag{32}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim iidN(0, \sigma_\zeta^2), \tag{33}$$

where β_t is the slope, η_t and ζ_t are assumed to be mutually independent.

The seasonal component γ_t is often specified in two ways such as dummy-variable seasonality and trigonometric seasonality, see Harvey et al. (1998) and Commandeur and Koopman(2007).

A time-varying dummy seasonal specification is given by

$$\sum_{i=0}^{s-1} \gamma_{t-i} = e_t, \tag{34}$$

where s is the seasonal length, e_t is an i.i.d. normal variable with mean zero and variance σ_e^2, i.e., $e_t \sim iidN(0, \sigma_e^2)$.

- In the limiting case $\sigma_e^2 = 0$, the seasonal effects are fixed over time.

- Commonly, $s = 4$ for quarterly data and $s = 12$ for monthly data.

(10) Dynamic Nelson and Siegel model

Diebold and Li (2006) and Diebold, Rudebusch and Arouba (2006) extend the model of Nelson and Siegel (1987) to a state space framework and to estimate the factors of term structure of interest rate. We often call this the dynamic Nelson and Siegel model.

The measurement equation of the state-space form is

$$\begin{bmatrix} y_t(\tau_1) \\ y_t(\tau_2) \\ \vdots \\ y_t(\tau_n) \end{bmatrix} = \begin{bmatrix} 1 & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1} \\ 1 & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} - e^{-\lambda\tau_2} \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-e^{-\lambda\tau_n}}{\lambda\tau_n} & \frac{1-e^{-\lambda\tau_n}}{\lambda\tau_n} - e^{-\lambda\tau_n} \end{bmatrix} \begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_t(\tau_1) \\ \varepsilon_t(\tau_2) \\ \vdots \\ \varepsilon_t(\tau_n) \end{bmatrix}$$

where $\beta_{1,t}$, $\beta_{2,t}$ and $\beta_{3,t}$ are the level, slope and curvature factors.

The transition equation specifies the dynamics of the state variables

$$\begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_2 & 0 \\ 0 & 0 & \phi_3 \end{bmatrix} \begin{bmatrix} \beta_{1,t-1} \\ \beta_{2,t-1} \\ \beta_{3,t-1} \end{bmatrix} + \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \\ \eta_{3,t} \end{bmatrix},$$

$$\begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \\ \eta_{3,t} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}\right),$$

where we have restricted the transition matrix to be diagonal to reflect the theoretical and empirical likelihood that the three factors exhibit little correlation.

2 Kalman Filter and Estimation

Reconsider the model:

$$\begin{aligned} \underset{n\times 1}{\boldsymbol{y}_t} &= \underset{n\times r}{\boldsymbol{H}}\ \underset{r\times 1}{\boldsymbol{x}_t} + \underset{n\times 1}{\boldsymbol{d}} + \underset{n\times 1}{\boldsymbol{u}_t} \\ \underset{r\times 1}{\boldsymbol{x}_t} &= \underset{r\times r}{\boldsymbol{F}}\ \underset{r\times 1}{\boldsymbol{x}_{t-1}} + \underset{r\times 1}{\boldsymbol{c}} + \underset{r\times 1}{\boldsymbol{v}_t} \end{aligned}$$

$$\mathrm{Cov}(\boldsymbol{u}_t\boldsymbol{u}_t) = \underset{n\times n}{\boldsymbol{R}}, \quad \mathrm{Cov}(\boldsymbol{v}_t\boldsymbol{v}_t) = \underset{r\times r}{\boldsymbol{Q}}, \quad \mathrm{Cov}(\boldsymbol{u}_t\boldsymbol{v}_t) = \mathbf{0}.$$

Let the observed data obtained through date $t-1$ be summarized by

$$\mathcal{F}_{t-1} \equiv (\boldsymbol{y}'_{t-1}, \boldsymbol{y}'_{t-2}, \ldots, \boldsymbol{y}'_1)'.$$

It is natural to obtain the one-step-ahead prediction density of $\boldsymbol{y}_t$

conditional on $\mathcal{F}_{t-1}$

$$p(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \int p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) p(\boldsymbol{x}_t \mid \mathcal{F}_{t-1}) d\boldsymbol{x}_t$$

where $p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$ denotes the observation density implied by the observation equation.

The prediction density of $\boldsymbol{x}_t$ given $\mathcal{F}_{t-1}$ is given by

$$p(\boldsymbol{x}_t \mid \mathcal{F}_{t-1}) = \int p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1} \mid \mathcal{F}_{t-1}) d\boldsymbol{x}_{t-1}$$

where $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ denotes the prediction density implied by the transition equation.

The filtered density of $\boldsymbol{x}_t$ given by $\mathcal{F}_t$ is given by

$$p(\boldsymbol{x}_t \mid \mathcal{F}_t) = \frac{p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)p(\boldsymbol{x}_t \mid \mathcal{F}_{t-1})}{p(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})} \propto p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)p(\boldsymbol{x}_t \mid \mathcal{F}_{t-1})$$

by Bayes rule.

Based on the previous prediction and filtered densities, we can implement a recursion from $t-1$ to t. If the input density $p(\boldsymbol{x}_{t-1} \mid \mathcal{F}_{t-1})$ is Gaussian, then the output density $p(\boldsymbol{x}_t \mid \mathcal{F}_t)$ is also Gaussian.

Note that the Gaussian density can be summarized by the corresponding mean and covariance matrix.

This leads to the Kalman filter for the mean and mean-square-error of the state and observation.

The Kalman filter can be better demonstrated in three steps, though at least the first two steps can be easily combined. The three steps are **prediction, updating, and smoothing**.

(1) Prediction

It is noted that, based on information available at $t-1$, the state vector $\boldsymbol{\xi}_t|\mathcal{F}_{t-1} \sim N(\boldsymbol{\xi}_{t|t-1}, \boldsymbol{P}_{t|t-1})$, where:

$$\boldsymbol{\xi}_{t|t-1} := \quad \mathrm{E}(\boldsymbol{\xi}_t|\mathcal{F}_{t-1}) \quad = \boldsymbol{F}\boldsymbol{\xi}_{t-1|t-1} + \boldsymbol{c}, \tag{35}$$

$$\boldsymbol{P}_{t|t-1} := \quad \mathrm{Var}(\boldsymbol{\xi}_t|\mathcal{F}_{t-1}) \quad = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{Q}, \tag{36}$$

where the *input means* $\boldsymbol{\xi}_{t-1|t-1} = \mathrm{E}(\boldsymbol{\xi}_{t-1}|\mathcal{F}_{t-1})$ and *variances* $\boldsymbol{P}_{t-1|t-1} = \mathrm{Var}(\boldsymbol{\xi}_{t-1}|\mathcal{F}_{t-1})$ is assumed to be known at time t. These two equations are known as the prediction equations.

The prediction of $\boldsymbol{y}_t$ conditional on $\mathcal{F}_{t-1}$ can be inferred immediately from (5):

$$\boldsymbol{y}_{t|t-1} = \quad \mathrm{E}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) \quad = \boldsymbol{H}\boldsymbol{\xi}_{t|t-1} + \boldsymbol{d}. \tag{37}$$

Then the *prediction error* can be written as

$$\boldsymbol{\eta}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t|t-1} = \boldsymbol{H}(\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}) + \boldsymbol{u}_t. \tag{38}$$

Since $\boldsymbol{\xi}_{t|t-1}$ is a function of $\mathcal{F}_{t-1}$, the term $\boldsymbol{u}_t$ is *independent* of both $\boldsymbol{\xi}_t$ and $\boldsymbol{\xi}_{t|t-1}$.

Thus the *conditional variance* of (38) is

$$\begin{aligned} \boldsymbol{\Sigma}_t = \quad \mathrm{Var}(\boldsymbol{y}_t - \boldsymbol{y}_{t|t-1}|\mathcal{F}_{t-1}) \quad &= \boldsymbol{H}\mathrm{Var}(\boldsymbol{\xi}_t|\mathcal{F}_{t-1})\boldsymbol{H}' + \boldsymbol{R} \\ &= \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{R}. \end{aligned} \tag{39}$$

(2) Updating

First note that the conditional covariance between (38) and the error in forecasting the state vector is

$$\begin{aligned}\mathrm{E}\{[\boldsymbol{y}_t - \boldsymbol{y}_{t|t-1}][\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}]'|\mathcal{F}_{t-1}\} &= \boldsymbol{H}\cdot\mathrm{E}\{[\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}][\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}]'|\mathcal{F}_{t-1}\} \\ &= \boldsymbol{H}\boldsymbol{P}_{t|t-1}\end{aligned}$$

Thus the distribution of the vector $(\boldsymbol{y}_t', \boldsymbol{\xi}_t')'$ conditional on $\mathcal{F}_{t-1}$ is

$$\begin{bmatrix} \boldsymbol{y}_t|\mathcal{F}_{t-1} \\ \boldsymbol{\xi}_t|\mathcal{F}_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{H}\boldsymbol{\xi}_{t|t-1} + \boldsymbol{d} \\ \boldsymbol{\xi}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{R} & \boldsymbol{H}\boldsymbol{P}_{t|t-1} \\ \boldsymbol{P}_{t|t-1}\boldsymbol{H}' & \boldsymbol{P}_{t|t-1} \end{bmatrix}\right)$$

▶ *Using this joint conditional distribution, our next purpose is to obtain the distribution of $\boldsymbol{\xi}_t|\mathcal{F}_t$, or $\boldsymbol{\xi}_t|\boldsymbol{y}_t, \mathcal{F}_{t-1}$. So we apply the following theorem:*

Theorem 1 (**Gaussian Conditioning Theorem).** *Let* z_1 *and* z_2 *denote* $(n_1 \times 1)$ *and* $(n_2 \times 1)$ *vectors respectively that have a joint normal distribution:*

$$\begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then the distribution of z_2 *conditional on* z_1 *is* $N(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{22|1})$ *where*

$$\boldsymbol{\mu}_{2|1} = \mathrm{E}(\boldsymbol{z}_2|\boldsymbol{z}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1 - \boldsymbol{\mu}_1), \tag{40}$$

$$\boldsymbol{\Sigma}_{22|1} = \mathrm{E}[(\boldsymbol{z}_2 - \boldsymbol{\mu}_{2|1})(\boldsymbol{z}_2 - \boldsymbol{\mu}_{2|1})'|\boldsymbol{z}_1] = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{41}$$

Remark: Thus the *optimal forecast* of z_2 conditional on having observed z_1 is $\boldsymbol{\mu}_{2|1}$, while the *mean squared error* of this forecast is $\boldsymbol{\Sigma}_{22|1}$.

Proof. We have

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11}) \det(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) = \det(\boldsymbol{\Sigma}_{11}) \det(\boldsymbol{\Sigma}_{22|1}).$$

So $f_{2|1}(\boldsymbol{z}_2|\boldsymbol{z}_1) := \frac{f(\boldsymbol{z})}{f_1(\boldsymbol{z}_1)} =$

$$\begin{aligned}
&= \frac{(2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{z} - \boldsymbol{\mu}) \right\}}{(2\pi)^{-n_1/2} \det(\boldsymbol{\Sigma}_{11})^{-1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{z}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_{11}^{-1/2} (\boldsymbol{z}_1 - \boldsymbol{\mu}_1) \right\}} \\
&= (2\pi)^{-n_2/2} \det(\boldsymbol{\Sigma}_{22|1})^{-1/2} \\
&\quad \times \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) - (\boldsymbol{z}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{z}_1 - \boldsymbol{\mu}_1) \right] \right\}.
\end{aligned}$$

where the term $(\boldsymbol{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) - (\boldsymbol{z}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{z}_1 - \boldsymbol{\mu}_1)$

$$= [\boldsymbol{z}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1 - \boldsymbol{\mu}_1)]' \boldsymbol{\Sigma}_{22|1}^{-1} [\boldsymbol{z}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1 - \boldsymbol{\mu}_1)].$$

We have

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{B}_{11} :=& \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \\
=& \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \\
\boldsymbol{B}_{12} :=& -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\
=& -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22|1}^{-1} \\
\boldsymbol{B}_{21} :=& -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \\
=& -\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \\
\boldsymbol{B}_{22} :=& \boldsymbol{\Sigma}_{22|1}^{-1}.
\end{aligned}$$

So

$$
\begin{aligned}
&(\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})-(\boldsymbol{z}_1-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)\\
=&(\boldsymbol{z}_1-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)-2(\boldsymbol{z}_1-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22|1}^{-1}(\boldsymbol{z}_2-\boldsymbol{\mu}_2)\\
&+(\boldsymbol{z}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22|1}^{-1}(\boldsymbol{z}_2-\boldsymbol{\mu}_2)\\
=&(\boldsymbol{z}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22|1}^{-1}(\boldsymbol{z}_2-\boldsymbol{\mu}_2)-2(\boldsymbol{z}_1-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_2-\boldsymbol{\mu}_2)\\
&+\left[\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)\right]'\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)\\
=&\left[(\boldsymbol{z}_2-\boldsymbol{\mu}_2)-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)\right]'\boldsymbol{\Sigma}_{22|1}^{-1}\left[(\boldsymbol{z}_2-\boldsymbol{\mu}_2)-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{z}_1-\boldsymbol{\mu}_1)\right].
\end{aligned}
$$

□

Results from the theorem: It then follows from (40) and (41) that $\boldsymbol{\xi}_t|\mathcal{F}_t = \boldsymbol{\xi}_t|\boldsymbol{y}_t, \mathcal{F}_{t-1}$ is distributed $N(\boldsymbol{\xi}_{t|t}, \boldsymbol{P}_{t|t})$, where

$$\begin{aligned} \boldsymbol{\xi}_{t|t} = &\quad \mathrm{E}(\boldsymbol{\xi}_t|\mathcal{F}_t) &&= \boldsymbol{\xi}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{\eta}_t, && (42) \\ \boldsymbol{P}_{t|t} = &\quad \mathrm{Var}(\boldsymbol{\xi}_t|\mathcal{F}_t) &&= \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{H}\boldsymbol{P}_{t|t-1}, && (43) \end{aligned}$$

where the quantity

$$\boldsymbol{K}_t = \quad \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\boldsymbol{\Sigma}_t^{-1} \quad = \boldsymbol{P}_{t|t-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{R})^{-1},$$

which is also called the *Kalman gain*.

Kalman Filtering

Taken together (35), (35), (42) and (43) make up the Kalman filter.

If desired they can be written as a single set of recursions going directly from $\boldsymbol{\xi}_{t-1|t-1}$ to $\boldsymbol{\xi}_{t|t}$ or, alternatively, from $\boldsymbol{\xi}_{t|t-1}$ to $\boldsymbol{\xi}_{t+1|t}$. We might refer to these as, respectively, the *contemporaneous* and *predictive filter*.

In the latter case

$$\boldsymbol{\xi}_{t+1|t} = \quad \boldsymbol{F}\boldsymbol{\xi}_{t|t} + \boldsymbol{c} = \quad \boldsymbol{F}\boldsymbol{\xi}_{t|t-1} + \boldsymbol{c} + \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{\eta}_t \tag{44}$$

The recursion for the covariance matrix (called as a *Riccati equation*)

$$\boldsymbol{P}_{t+1|t} = \quad \boldsymbol{F}(\boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{H}\boldsymbol{P}_{t|t-1})\boldsymbol{F}' + \boldsymbol{Q}. \tag{45}$$

Starting values

The starting values for the Kalman filter may be specified in terms of $\boldsymbol{\xi}_0$ and $\boldsymbol{P}_0$ or $\boldsymbol{\xi}_{1|0}$ and $\boldsymbol{P}_{1|0}$.

- If $\boldsymbol{\xi}_t$ is stationary, we use the *unconditional mean* and *unconditional covariance matrix* of $\boldsymbol{\xi}_t$

$$\boldsymbol{\xi}_0 = E[\boldsymbol{\xi}_t] = (\boldsymbol{I}_r - \boldsymbol{F})^{-1}\boldsymbol{c}$$

$$\text{vec}(\boldsymbol{P}_0) = (\boldsymbol{I}_{r^2} - \boldsymbol{F} \otimes \boldsymbol{F})^{-1}\text{vec}(\boldsymbol{Q}),$$

  where $\text{vec}(\boldsymbol{P}_0)$ is an $r^2 \times 1$ column vector. It can be easily *reshaped to form* the $r \times r$ matrix $\boldsymbol{P}_0$.

- If $\boldsymbol{\xi}_t$ is non-stationary, a diffuse prior is often used by setting $\boldsymbol{P}_0 = \kappa I$, and letting the scalar κ go to infinity.

(3) Smoothing (for SSM with fixed parameters)

Given initial conditions $\boldsymbol{\xi}_{T|T}$ and $\boldsymbol{P}_{T|T}$ from the Kalman filter.

We can *recursively* obtain $\boldsymbol{\xi}_{t|T}$ and $\boldsymbol{P}_{t|T}$, for $t = T-1, T-2, \ldots, 1$,

$$\boldsymbol{\xi}_{t|T} := \quad \mathrm{E}(\boldsymbol{\xi}_t|\mathcal{F}_T) \quad = \boldsymbol{\xi}_{t|t} + \boldsymbol{\Upsilon}_t(\boldsymbol{\xi}_{t+1|T} - \boldsymbol{\xi}_{t+1|t}) \tag{46}$$

$$\boldsymbol{P}_{t|T} := \quad \mathrm{Var}(\boldsymbol{\xi}_t|\mathcal{F}_T) \quad = \boldsymbol{P}_{t|t} + \boldsymbol{\Upsilon}_t(\boldsymbol{P}_{t+1|T} - \boldsymbol{P}_{t+1|t})\boldsymbol{\Upsilon}_t' \tag{47}$$

where $\boldsymbol{\Upsilon}_t = \boldsymbol{P}_{t|t}\boldsymbol{F}'\boldsymbol{P}_{t+1|t}^{-1}$.

Especially, at time $t = 0$, we can get the estimates of initial state

$$\boldsymbol{\xi}_{0|T} \quad \text{and} \quad \boldsymbol{P}_{0|T},$$

which can be used to set as $\boldsymbol{a}_0$ and $\boldsymbol{\Sigma}_0$.

Derivation of smoothing algorithm

(a) **Computing** $p(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}|\mathcal{F}_t)$, which is Gaussian and it suffices to compute its mean vector and its covariance matrix

- The mean vector is $\begin{bmatrix} \boldsymbol{\xi}_{t|t} \\ \boldsymbol{\xi}_{t+1|t} \end{bmatrix}$, where $\boldsymbol{\xi}_{t+1|t} = \boldsymbol{F}\boldsymbol{\xi}_{t|t} + \boldsymbol{c}$.
- We will use this when calculating the cross covariance:

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}|\mathcal{F}_t) &= E[(\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t})(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t})'|\mathcal{F}_t] \\
&= E[(\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t})(\boldsymbol{F}\boldsymbol{\xi}_t + \boldsymbol{c} + \boldsymbol{v}_{t+1} - \boldsymbol{F}\boldsymbol{\xi}_{t|t} - \boldsymbol{c})'|\mathcal{F}_t] \\
&= \boldsymbol{P}_{t|t}\boldsymbol{F}'.
\end{aligned}$$

- We have $\left.\begin{bmatrix} \boldsymbol{\xi}_t \\ \boldsymbol{\xi}_{t+1} \end{bmatrix}\right| \mathcal{F}_t \sim N\left(\begin{bmatrix} \boldsymbol{\xi}_{t|t} \\ \boldsymbol{\xi}_{t+1|t} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_{t|t} & \boldsymbol{P}_{t|t}\boldsymbol{F}' \\ \boldsymbol{F}\boldsymbol{P}_{t|t} & \boldsymbol{P}_{t+1|t} \end{bmatrix}\right)$.

(b) **Computing** $p(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_T) = p(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_t)$ by Gaussian Conditioning Theorem.

- The corresponding mean and covariance matrix are

$$E[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_T] = E[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_t] = \boldsymbol{\xi}_{t|t} + \boldsymbol{L}_t(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t})$$

$$\mathrm{Cov}(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_T) = \mathrm{Cov}(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_T) = \boldsymbol{P}_{t|t} - \boldsymbol{L}_t\boldsymbol{P}_{t+1|t}\boldsymbol{L}_t',$$

  where $\boldsymbol{L}_t = \boldsymbol{P}_{t|t}\boldsymbol{F}\boldsymbol{P}_{t+1|t}^{-1}$.

- Hence we have

$$\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \mathcal{F}_T \sim N\left(\boldsymbol{\xi}_{t|t} + \boldsymbol{L}_t(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t}), \boldsymbol{P}_{t|t} - \boldsymbol{L}_t\boldsymbol{P}_{t+1|t}\boldsymbol{L}_t'\right).$$

(c) **Computing** $p(\boldsymbol{\xi}_t|\mathcal{F}_T)$

– Two properties are used:

(c.1.) Law of Iterated Expectation (Tower Property):

$$E[Z|X] = E[E[Z|Y,X]|X].$$

(c.2.) Law of Total Conditional Variance:

$$\mathrm{Cov}[Z|X] = \mathrm{Cov}[E[Z|Y,X]|X] + E[\mathrm{Cov}[Z|Y,X]|X].$$

– Let $Z = \boldsymbol{\xi}_t$, $Y = \boldsymbol{\xi}_{t+1}$, $X = \mathcal{F}_T$, and apply the above two properties.

▶ First, we compute the smoothed mean

$$\begin{aligned}
\boldsymbol{\xi}_{t|T} &= E[\boldsymbol{\xi}_t|\mathcal{F}_T] = E[E[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1},\mathcal{F}_T]|\mathcal{F}_T] \\
&= E[\boldsymbol{\xi}_{t|t} + \boldsymbol{L}_t(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t})|\mathcal{F}_T] \\
&= \underbrace{\boldsymbol{\xi}_{t|t}}_{\text{filtered estimate}} + \underbrace{\boldsymbol{L}_t(\boldsymbol{\xi}_{t+1|T} - \boldsymbol{\xi}_{t+1|t})}_{\text{correction term}}.
\end{aligned}$$

- Second, all that remains is to apply the law of total conditional variance to find $\boldsymbol{P}_{t|T}$:

$$
\begin{aligned}
\boldsymbol{P}_{t|T} &= \mathrm{Cov}[\boldsymbol{\xi}_t|\mathcal{F}_T] \\
&= \mathrm{Cov}[E[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1},\mathcal{F}_T]|\mathcal{F}_T] + E[\mathrm{Cov}[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1},\mathcal{F}_T]|\mathcal{F}_T] \\
&= \mathrm{Cov}[\boldsymbol{\xi}_{t|t} + \boldsymbol{L}_t(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t})|\mathcal{F}_T] + E[\boldsymbol{P}_{t|t} - \boldsymbol{L}_t\boldsymbol{P}_{t+1|t}\boldsymbol{L}_t'|\mathcal{F}_T] \\
&= \boldsymbol{L}_t\boldsymbol{P}_{t+1|T}\boldsymbol{L}_t' + (\boldsymbol{P}_{t|t} - \boldsymbol{L}_t\boldsymbol{P}_{t+1|t}\boldsymbol{L}_t') \\
&= \underbrace{\boldsymbol{P}_{t|t}}_{\text{filtered estimate}} + \underbrace{\boldsymbol{L}_t(\boldsymbol{P}_{t+1|T} - \boldsymbol{P}_{t+1|t})\boldsymbol{L}_t'}_{\text{correction term}}.
\end{aligned}
$$

(4) The Lag-One Covariance Smoother *

For the state-space model, with $\boldsymbol{K}_t$, $\boldsymbol{\Upsilon}_t$ ($t = 1, \ldots, T$) obtained from before, and with initial condition

$$\begin{aligned}\boldsymbol{P}_{T,T-1|T} &= \mathrm{Cov}(\boldsymbol{\xi}_T, \boldsymbol{\xi}_{T-1} \mid \mathcal{F}_T) \\ &= (\boldsymbol{I} - \boldsymbol{K}_T\boldsymbol{H})\boldsymbol{F}\boldsymbol{P}_{T-1|T-1},\end{aligned}$$

we have that for $t = T-1, T-2, \ldots, 1$,

$$\begin{aligned}\boldsymbol{P}_{t,t-1|T} &= \mathrm{Cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1} \mid \mathcal{F}_T) \\ &= \boldsymbol{P}_{t|t}\boldsymbol{\Upsilon}'_{t-1} + \boldsymbol{\Upsilon}_t(\boldsymbol{P}_{t+1,t|T} - \boldsymbol{F}\boldsymbol{P}_{t|t})\boldsymbol{\Upsilon}'_{t-1}.\end{aligned}$$

The proof is given in Chapter 6 in Shumway and Stoffer (2006).

1.3 Prediction

As regards multi-step prediction, taking expectations, conditional on the information at time T, of the transition equation at time $T+h$ yields the recursion

$$\boldsymbol{\xi}_{T+h|T} = \boldsymbol{F}\boldsymbol{\xi}_{T+h-1|T} + \boldsymbol{c}, \quad h = 1, 2, 3, \ldots \tag{48}$$

with initial value $\boldsymbol{\xi}_{T|T}$. Similarly,

$$\boldsymbol{P}_{T+h|T} = \boldsymbol{F}\boldsymbol{P}_{T+h-1|T}\boldsymbol{F}' + \boldsymbol{Q}, \quad h = 1, 2, 3, \ldots \tag{49}$$

Thus $\boldsymbol{\xi}_{T+h|T}$ and $\boldsymbol{P}_{T+h|T}$ are evaluated by repeatedly applying the Kalman filter prediction equations.

The MMSE of $\boldsymbol{y}_{T+h}$ can be obtained directly from $\boldsymbol{\xi}_{T+h|T}$. Taking conditional expectations in the measurement equation for $\boldsymbol{y}_{T+h}$ gives

$$\hat{\boldsymbol{y}}_{T+h|T} = E(\boldsymbol{y}_{T+h}|\mathcal{F}_T) = \boldsymbol{H}\boldsymbol{\xi}_{T+h|T} + \boldsymbol{d}, \quad h = 1, 2, 3, \ldots \tag{50}$$

with MSE matrix

$$MSE(\hat{\boldsymbol{y}}_{T+h|T}) = \boldsymbol{H}\boldsymbol{P}_{T+h|T}\boldsymbol{H}' + \boldsymbol{R}, \quad h = 1, 2, 3, \ldots \tag{51}$$

When the normality assumption is relaxed, $\hat{\boldsymbol{y}}_{T+h|T}$ and $MSE(\hat{\boldsymbol{y}}_{T+h|T})$ are still minimum mean square linear estimators.

2.4 Maximum likelihood estimation

The joint density function for the T sets of observations, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, is

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T; \theta) = \prod_{t=1}^{T} p(\boldsymbol{y}_t | \mathcal{F}_{t-1}) \tag{52}$$

where $p(\boldsymbol{y}_t|\mathcal{F}_{t-1})$ denotes the distribution of $\boldsymbol{y}_t$ conditional on the information set at time $t-1$, that is, $\mathcal{F}_{t-1} = \{\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1\}$.

In the Gaussian state space model, the conditional distribution of $\boldsymbol{y}_t$ is normal with mean $\hat{\boldsymbol{y}}_{t|t-1}$ and covariance matrix $\boldsymbol{\Sigma}_t$. Hence the prediction errors or innovations,

$$\boldsymbol{\eta}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1} \sim N(0, \boldsymbol{\Sigma}_t), \quad t = 1, \ldots, T, \tag{53}$$

is serially independent with mean zero and covariance $\boldsymbol{\Sigma}_t$.

Based on the predictive distribution of $\boldsymbol{y}_t$, the conditional density at time t is:

$$p(\boldsymbol{y}_t|\mathcal{F}_{t-1}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}_t|^{-1/2}\exp\left(-\frac{\boldsymbol{\eta}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\eta}_t}{2}\right) \tag{54}$$

where $\boldsymbol{\eta}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t|t-1}$, $\mathcal{F}_{t-1}$ is the information set at time $t-1$.

The parameters can be estimated by maximizing the log likelihood of the density function:

$$\mathcal{L}(\theta) = -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left\{\log|\boldsymbol{\Sigma}_t| + (\boldsymbol{\eta}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\eta}_t)\right\}.$$

Estimated parameters and state variables can be obtained accordingly.

2.5 EM algorithm for dynamic linear models

Shumway and Stoffer (1982) presented a conceptually simpler estimation procedure based on the EM (expectation maximization) algorithm (Dempster et al., 1977).

The basic idea is that if we could observe the states $X_T = \{\xi_0, \xi_1, \ldots, \xi_T\}$ and the observations $Y_T = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, then we would consider $\{X_T, Y_T\}$ as the complete data.

The joint density for the complete data $\{X_T, Y_T\}$ is

$$f_\Theta(X_T, Y_T) = f_{\boldsymbol{a}_0, \boldsymbol{\Sigma}_0}(\boldsymbol{\xi}_0) \prod_{t=1}^{T} f_{\boldsymbol{F}, \boldsymbol{Q}}(\boldsymbol{\xi}_t | \boldsymbol{\xi}_{t-1}) \prod_{t=1}^{T} f_{\boldsymbol{H}, \boldsymbol{R}}(\boldsymbol{y}_t | \boldsymbol{\xi}_t).$$

Under the Gaussian assumption and ignoring constants, the complete data log-likelihood can be written as

$$\begin{aligned}-2\ln L_{X,Y}(\Theta) = &\ln|\boldsymbol{\Sigma}_0| + (\boldsymbol{\xi}_0 - \boldsymbol{a}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\xi}_0 - \boldsymbol{a}_0) \\ &+ T\ln|\boldsymbol{Q}| + \sum_{t=1}^{T}(\boldsymbol{\xi}_t - \boldsymbol{F}\boldsymbol{\xi}_{t-1})'\boldsymbol{Q}^{-1}(\boldsymbol{\xi}_t - \boldsymbol{F}\boldsymbol{\xi}_{t-1}) \\ &+ T\ln|\boldsymbol{R}| + \sum_{t=1}^{T}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)'\boldsymbol{R}^{-1}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)\end{aligned}$$

- If we did have the complete data, we could then use the results from multivariate normal theory to easily obtain the MLEs of Θ.

- However, we do not have the complete data, the EM algorithm gives us an iterative method for finding the MLEs of Θ based on the incomplete data, Y_T, by successively maximizing the conditional expectation of the complete data likelihood.

(1) Expectation step (E-Step):

To implement the EM algorithm, we write, at iteration j, $(j = 1, 2, \ldots)$,

$$Q(\Theta \mid \Theta^{(j-1)}) = E\left\{-2 \ln L_{X,Y}(\Theta) \mid Y_T, \Theta^{(j-1)}\right\}$$

Given $\Theta^{(j-1)}$, we can use the Kalman Smoother to obtain the desired conditional expectations as smoothers. This yields

$$Q(\Theta \mid \Theta^{(j-1)}) = \ln|\boldsymbol{\Sigma}_0| + \operatorname{tr}\{\boldsymbol{\Sigma}_0^{-1} E[(\boldsymbol{\xi}_0 - \boldsymbol{\xi}_{0|T} + \boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0)(\boldsymbol{\xi}_0 - \boldsymbol{\xi}_{0|T} + \boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0)' | \mathcal{F}_T]\}$$

$$+ T \ln|\boldsymbol{Q}| + \operatorname{tr}\left\{\boldsymbol{Q}^{-1} \sum_{t=1}^{T} E[\underbrace{(\boldsymbol{\xi}_t - \boldsymbol{F}\boldsymbol{\xi}_{t-1})(\boldsymbol{\xi}_t - \boldsymbol{F}\boldsymbol{\xi}_{t-1})'}_{=\boldsymbol{\xi}_t\boldsymbol{\xi}_t' - \boldsymbol{\xi}_t\boldsymbol{\xi}_{t-1}'\boldsymbol{F}' - \boldsymbol{F}\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_t' + \boldsymbol{F}\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-1}'\boldsymbol{F}'} | \mathcal{F}_T]\right\}$$

$$+ T \ln|\boldsymbol{R}| + \operatorname{tr}\left\{\boldsymbol{R}^{-1} \sum_{t=1}^{T} E[\underbrace{(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)}_{\text{insert } -\boldsymbol{H}\boldsymbol{\xi}_{t|T} + \boldsymbol{H}\boldsymbol{\xi}_{t|T}} \underbrace{(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)'}_{\text{insert } -\boldsymbol{H}\boldsymbol{\xi}_{t|T} + \boldsymbol{H}\boldsymbol{\xi}_{t|T}} | \mathcal{F}_T]\right\}$$

Using the facts that

$$E[\boldsymbol{\xi}_t\boldsymbol{\xi}_t'|\mathcal{F}_T] = E[\boldsymbol{\xi}_t|\mathcal{F}_T]E[\boldsymbol{\xi}_t'|\mathcal{F}_T] + \mathrm{Var}[\boldsymbol{\xi}_t|\mathcal{F}_T] = \boldsymbol{\xi}_{t|T}\boldsymbol{\xi}_{t|T}' + \boldsymbol{P}_{t|T},$$
$$E[\boldsymbol{\xi}_t\boldsymbol{\xi}_{t-1}'|\mathcal{F}_T] = \boldsymbol{\xi}_{t|T}\boldsymbol{\xi}_{t-1|T}' + \boldsymbol{P}_{t,t-1|T},$$
$$E[\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-1}'|\mathcal{F}_T] = \boldsymbol{\xi}_{t-1|T}\boldsymbol{\xi}_{t-1|T}' + \boldsymbol{P}_{t-1|T},$$

we have

$$\begin{aligned} Q(\Theta \mid \Theta^{(j-1)}) = {} & \ln|\boldsymbol{\Sigma}_0| + \mathrm{tr}\{\boldsymbol{\Sigma}_0^{-1}[\boldsymbol{P}_{0|T} + (\boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0)(\boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0)]'\} \\ & + T\ln|\boldsymbol{Q}| + \mathrm{tr}\{\boldsymbol{Q}^{-1}[\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{F}' - \boldsymbol{F}\boldsymbol{S}_{10}' + \boldsymbol{F}\boldsymbol{S}_{00}\boldsymbol{F}']\} \\ & + T\ln|\boldsymbol{R}| \\ & + \mathrm{tr}\left\{\boldsymbol{R}^{-1}\sum_{t=1}^{T}[(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_{t|T})(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_{t|T})' + \boldsymbol{H}\boldsymbol{P}_{t|T}\boldsymbol{H}']\right\}, \end{aligned} \tag{55}$$

where

$$\boldsymbol{S}_{11} = \sum_{t=1}^{T} (\boldsymbol{\xi}_{t|T} \boldsymbol{\xi}'_{t|T} + \boldsymbol{P}_{t|T}), \quad \boldsymbol{S}_{10} = \sum_{t=1}^{T} (\boldsymbol{\xi}_{t|T} \boldsymbol{\xi}'_{t-1|T} + \boldsymbol{P}_{t,t-1|T})$$

$$\boldsymbol{S}_{00} = \sum_{t=1}^{T} (\boldsymbol{\xi}_{t-1|T} \boldsymbol{\xi}'_{t-1|T} + \boldsymbol{P}_{t-1|T})$$

In the above equations, the smoothers are calculated under the current value of the parameters $\Theta^{(j-1)}$; for simplicity, we have not explicitly displayed this fact.

(2) Maximization step (M-Step):

Minimizing (55) with respect to the unknown parameter set, $\Theta = \{\boldsymbol{H}, \boldsymbol{F}, \boldsymbol{Q}, \boldsymbol{R}\}$, constitutes the maximization step.

At iteration j, for the parameters $\boldsymbol{F}$ and $\boldsymbol{Q}$, this yields the updated estimates,

$$\boldsymbol{F}^{(j)} = \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}$$
$$\boldsymbol{Q}^{(j)} = T^{-1}(\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{10}').$$

To develop estimators for $\boldsymbol{H}$ and $\boldsymbol{R}$ from the last term of (55), it will reduce to minimizing: $\sum_{t=1}^{T} E[(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_t)' \mid Y_T, \Theta^{(j-1)}]$ over $\boldsymbol{H}$ and then evaluating $\boldsymbol{R}$.

Performing the first operation, we obtain

$$\boldsymbol{H}^{(j)} = \sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{\xi}_{t|T}' \boldsymbol{S}_{11}^{-1}$$

Then, we can obtain estimator for $\boldsymbol{R}$ as:

$$\boldsymbol{R}^{(j)} = T^{-1} \sum_{t=1}^{T} [(\boldsymbol{y}_t - \boldsymbol{H}^{(j)} \boldsymbol{\xi}_{t|T})(\boldsymbol{y}_t - \boldsymbol{H}^{(j)} \boldsymbol{\xi}_{t|T})' + \boldsymbol{H}^{(j)} \boldsymbol{P}_{t|T} \boldsymbol{H}^{(j)\prime}].$$

Finally, the updates for the initial mean and covariance matrix are

$$\boldsymbol{a}_0^{(j)} = \boldsymbol{\xi}_{0|T} \quad \text{and} \quad \boldsymbol{\Sigma}_0^{(j)} = (\boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0^{(j)})(\boldsymbol{\xi}_{0|T} - \boldsymbol{a}_0^{(j)})' + \boldsymbol{P}_{0|T}.$$

EM algorithm:

1. Initialize the procedure by selecting starting values for the parameters $\Theta^{(0)} = \{\boldsymbol{a}_0, \boldsymbol{\Sigma}_0, \boldsymbol{F}, \boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{R}\}$.

On iteration j, $(j = 1, 2, \ldots)$:

2. Run the Kalman filter and smoother, and compute the incomplete-data likelihood, $-\ln L_{\mathcal{F}}(\Theta^{(j-1)})$.

3. Perform the E-Step.

4. Perform the M-Step.

5. Repeat Steps 2 - 4 to convergence.

3 Approximate Filters for Nonlinear SS Models

3.1 Extended Kalman Filter

Consider next a multidimensional normal state-space model

$$\boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{\xi}_t) + \boldsymbol{d} + \boldsymbol{u}_t, \tag{56}$$

$$\boldsymbol{\xi}_t = \boldsymbol{\phi}(\boldsymbol{\xi}_{t-1}) + \boldsymbol{c} + \boldsymbol{v}_t, \tag{57}$$

where $\boldsymbol{v}_t \sim i.i.d.N(\boldsymbol{0}, \boldsymbol{Q})$ and $\boldsymbol{u}_t \sim i.i.d.N(\boldsymbol{0}, \boldsymbol{R})$. Suppose $\boldsymbol{\phi}(\cdot)$ in (57) is replaced with a first-order Taylor's approximation around $\boldsymbol{\xi}_{t-1} = \boldsymbol{\xi}_{t-1|t-1}$,

$$\boldsymbol{\xi}_t = \boldsymbol{\phi}_t + \boldsymbol{\Phi}_t \cdot (\boldsymbol{\xi}_{t-1} - \boldsymbol{\xi}_{t-1|t-1}) + \boldsymbol{c} + \boldsymbol{v}_t, \tag{58}$$

where

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}(\boldsymbol{\xi}_{t-1|t-1}), \quad \boldsymbol{\Phi}_t = \left.\frac{\partial \boldsymbol{\phi}(\boldsymbol{\xi}_{t-1})}{\partial \boldsymbol{\xi}'_{t-1}}\right|_{\boldsymbol{\xi}_{t-1}=\boldsymbol{\xi}_{t-1|t-1}}.$$

If the form of ϕ and any parameters it depends on are known, then the inference $\boldsymbol{\xi}_{t|t}$, can be constructed as a function of variables observed at date t through a recursion to be described in a moment.

Similarly the function $\boldsymbol{h}(\cdot)$ in (56) can be linearized around $\boldsymbol{\xi}_{t|t-1}$ to produce

$$\boldsymbol{y}_t = \boldsymbol{h}_t + \boldsymbol{H}_t \cdot (\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}) + \boldsymbol{d} + \boldsymbol{u}_t. \tag{59}$$

where

$$\boldsymbol{h}_t = \boldsymbol{h}(\boldsymbol{\xi}_{t|t-1}), \quad \boldsymbol{H}_t = \left.\frac{\partial \boldsymbol{h}(\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}'_t}\right|_{\boldsymbol{\xi}_t=\boldsymbol{\xi}_{t|t-1}}$$

Treat (58) and (59) as if they were the true model, then the *extended Kalman filter* is implemented as follows:

$$\boldsymbol{\xi}_{t|t-1} = \phi(\boldsymbol{\xi}_{t-1|t-1}) + \boldsymbol{c} \tag{60}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{\Phi}_t \boldsymbol{P}_{t-1|t-1} \boldsymbol{\Phi}_t' + \boldsymbol{Q} \tag{61}$$

$$\boldsymbol{y}_{t|t-1} = \boldsymbol{h}_t + \boldsymbol{d} = \boldsymbol{h}(\boldsymbol{\xi}_{t|t-1}) + \boldsymbol{d} \tag{62}$$

$$\boldsymbol{\eta}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t|t-1} = \boldsymbol{H}_t(\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t|t-1}) + \boldsymbol{u}_t \tag{63}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{H}_t \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t' + \boldsymbol{R} \tag{64}$$

$$\boldsymbol{\xi}_{t|t} = \boldsymbol{\xi}_{t|t-1} + \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t (\boldsymbol{H}_t' \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t + \boldsymbol{R})^{-1} \boldsymbol{\eta}_t \tag{65}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t (\boldsymbol{H}_t' \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t + \boldsymbol{R})^{-1} \boldsymbol{H}_t' \boldsymbol{P}_{t|t-1} \tag{66}$$

3.2 State Space Models with conditional heteroskedasticity

We consider the following state-space model with ARCH disturbances proposed by Harvey, Ruiz and Sentana (1992):

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{\xi}_t + \boldsymbol{d} + \boldsymbol{u}_t + \Lambda\epsilon_t && (67) \\ \boldsymbol{\xi}_t &= \boldsymbol{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{c} + \boldsymbol{v}_t + \lambda\omega_t && (68) \\ \boldsymbol{u}_t &\sim N(0, \boldsymbol{R}), \quad \boldsymbol{v}_t \sim N(0, \boldsymbol{Q}), && (69) \end{aligned}$$

where Λ is $n \times 1$ and λ is $k \times 1$, and ϵ_t and ω_t are 1×1. The ARCH effects are introduced via the following scalar disturbances:

$$\begin{aligned} \epsilon_t | I_{t-1} &\sim N(0, h_{1t}), \quad h_{1t} = \alpha_0 + \alpha_1 \epsilon_{t-1}^2, \\ \omega_t | I_{t-1} &\sim N(0, h_{2t}), \quad h_{2t} = \gamma_0 + \gamma_1 \omega_{t-1}^2. \end{aligned}$$

- If $\Lambda = 0$ and $\lambda = 0$, then there is no ARCH effect and the model reduces to the usual time-invariant model.

- If the first element of Λ is nonzero with the others being zero, only the first shock in the mesurement equation is subject to an ARCH effect.

- If all the elements of Λ are nonzero, then all the shocks in the measurement equation are subject to a common ARCH effect.

- The same argument applies for the transition equation.

- If we define $\boldsymbol{u}_t^* = \boldsymbol{u}_t + \Lambda\epsilon_t$ and $\boldsymbol{v}_t^* = \boldsymbol{v}_t + \lambda\omega_t$, then we have $\boldsymbol{u}_t^*|I_{t-1} \sim N(0, \boldsymbol{R}_t)$ and $\boldsymbol{v}_t^*|I_{t-1} \sim N(0, \boldsymbol{Q}_t)$, where

$$\boldsymbol{R}_t = \boldsymbol{R} + \Lambda h_{1t}\Lambda' \quad \text{and} \quad \boldsymbol{Q}_t = \boldsymbol{Q} + \lambda h_{2t}\lambda'.$$

Difficulty in Kalman filter

$$
\begin{aligned}
\boldsymbol{\xi}_{t|t-1} &= \boldsymbol{F}\boldsymbol{\xi}_{t-1|t-1} + \boldsymbol{c} \\
\boldsymbol{P}_{t|t-1} &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{Q} + \lambda h_{2t}\lambda' \\
\boldsymbol{\eta}_t &= \boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\xi}_{t|t-1} - \boldsymbol{d} \\
\boldsymbol{\Sigma}_t &= \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{R} + \Lambda h_{1t}\Lambda' \\
\boldsymbol{\xi}_{t|t} &= \boldsymbol{\xi}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{H}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\eta}_t \\
\boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{H}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{H}'\boldsymbol{P}_{t|t-1}
\end{aligned}
$$

However, we need to calculate $h_{1t} = \alpha_0 + \alpha_1\epsilon_{t-1}^2$ and $h_{2t} = \gamma_0 + \gamma_1\omega_{t-1}^2$, which are functions of the squares of past unobserved shocks ϵ_{t-1}^2 and ω_{t-1}^2. Thus, the above Kalman filter is not operable.

Approximate Kalman filter

Harvey, Ruiz and Sentana (1992) solve the problem of unobserved ϵ_{t-1}^2 and ω_{t-1}^2 by replacing them with their conditional expectations:

$$h_{1t} = \alpha_0 + \alpha_1 E[\epsilon_{t-1}^2 | I_{t-1}] \quad \text{and} \quad h_{2t} = \gamma_0 + \gamma_1 E[\omega_{t-1}^2 | I_{t-1}].$$

To get the conditional expectations of ϵ_{t-1}^2 and ω_{t-1}^2, Harvey, Ruiz and Sentana augment the heteroskedastic shocks into the original state vector in the transition equation as follows:

$$\begin{bmatrix} \boldsymbol{\xi}_t \\ \epsilon_t \\ \omega_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{F} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{t-1} \\ \epsilon_{t-1} \\ \omega_{t-1} \end{bmatrix} + \begin{bmatrix} I & 0 & \lambda \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_t \\ \epsilon_t \\ \omega_t \end{bmatrix}$$

$$(\boldsymbol{\xi}_t^* = \boldsymbol{c}^* + \boldsymbol{F}^* \boldsymbol{\xi}_{t-1}^* + G^* \boldsymbol{v}_t^*)$$

where

$$E[\boldsymbol{v}_t^* \boldsymbol{v}_t^{*\prime} | I_{t-1}] = \begin{bmatrix} \boldsymbol{Q} & 0 & 0 \\ 0 & h_{1t} & 0 \\ 0 & 0 & h_{2t} \end{bmatrix} = \boldsymbol{Q}_t^*$$

The measurement equation should be replaced by

$$\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{H} & \Lambda & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_t \\ \epsilon_t \\ \omega_t \end{bmatrix} + \boldsymbol{d} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(0, \boldsymbol{R})$$

$$(\boldsymbol{y}_t = \boldsymbol{H}^* \boldsymbol{\xi}_t^* + \boldsymbol{d} + \boldsymbol{u}_t)$$

Applying the Kalman filter recursions to the model, we have

$$\begin{aligned}
\boldsymbol{\xi}^*_{t|t-1} &= \boldsymbol{F}^*\boldsymbol{\xi}^*_{t-1|t-1} + \boldsymbol{c}^* \\
\boldsymbol{P}^*_{t|t-1} &= \boldsymbol{F}^*\boldsymbol{P}^*_{t-1|t-1}\boldsymbol{F}^{*\prime} + G^*\boldsymbol{Q}^*_t G^{*\prime} \\
\boldsymbol{\eta}^*_t &= \boldsymbol{y}_t - \boldsymbol{H}^*\boldsymbol{\xi}^*_{t|t-1} - \boldsymbol{d}^* \\
\boldsymbol{\Sigma}^*_t &= \boldsymbol{H}^*\boldsymbol{P}^*_{t|t-1}\boldsymbol{H}^{*\prime} + \boldsymbol{R}^* \\
\boldsymbol{\xi}^*_{t|t} &= \boldsymbol{\xi}^*_{t|t-1} + \boldsymbol{P}^*_{t|t-1}\boldsymbol{H}^*\boldsymbol{\Sigma}^{*-1}_t\boldsymbol{\eta}^*_t \\
\boldsymbol{P}^*_{t|t} &= \boldsymbol{P}^*_{t|t-1} - \boldsymbol{P}^*_{t|t-1}\boldsymbol{H}^*\boldsymbol{\Sigma}^{*-1}_t\boldsymbol{H}^{*\prime}\boldsymbol{P}^*_{t|t-1}
\end{aligned}$$

In the $\boldsymbol{Q}^*_t$ matrix,

$$h_{1t} = \alpha_0 + \alpha_1 E[\epsilon^2_{t-1}|I_{t-1}] \quad \text{and} \quad h_{2t} = \gamma_0 + \gamma_1 E[\omega^2_{t-1}|I_{t-1}].$$

It is easy to show that

$$E[\epsilon_{t-1}^2|I_{t-1}] = E[\epsilon_{t-1}|I_{t-1}]^2 + E[(\epsilon_{t-1} - E[\epsilon_{t-1}|I_{t-1}])^2],$$
$$E[\omega_{t-1}^2|I_{t-1}] = E[\omega_{t-1}|I_{t-1}]^2 + E[(\omega_{t-1} - E[\omega_{t-1}|I_{t-1}])^2],$$

where

- $E[\epsilon_{t-1}|I_{t-1}]$ and $E[\omega_{t-1}|I_{t-1}]$ are obtained from the last two elements of $\boldsymbol{\xi}_{t-1|t-1}^*$;
- $E[(\epsilon_{t-1} - E[\epsilon_{t-1}|I_{t-1}])^2]$ and $E[(\omega_{t-1} - E[\omega_{t-1}|I_{t-1}])^2]$ are obtained from the last two diagonal elements of $\boldsymbol{P}_{t-1|t-1}^*$.

Example 1 (TVP model with GARCH disturbances)**.**

$$y_t = X_{t-1}\beta_t + \varepsilon_t^*, \quad \varepsilon_t^* | I_{t-1} \sim N(0, h_t)$$

$$\beta_t = \beta_{t-1} + v_t, \quad v_t \sim N(0, Q)$$

$$h_t = \alpha_0 + \alpha_1 {\varepsilon_{t-1}^*}^2 + \alpha_2 h_{t-1},$$

3.3 State Space Models with Markov Switching

Consider the following state-space representation of a dynamic linear model with Markov switching in both measurement and transition equations:

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{H}_{S_t}\boldsymbol{\xi}_t + \boldsymbol{d}_{S_t} + \boldsymbol{u}_t, && (70)\\
\boldsymbol{\xi}_t &= \boldsymbol{F}_{S_t}\boldsymbol{\xi}_{t-1} + \boldsymbol{c}_{S_t} + \boldsymbol{v}_t, && (71)\\
\begin{pmatrix} \boldsymbol{u}_t \\ \boldsymbol{v}_t \end{pmatrix} &\sim N\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{R}_{S_t} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_{S_t} \end{pmatrix} \right), && (72)\\
S_t|S_{t-1} &\sim \text{Markov}(\boldsymbol{P}) && (73)
\end{aligned}$$

where the transition probability matrix $\boldsymbol{P} = (p_{ij})$, $p_{ij} = \Pr(S_t = j|S_{t-1} = i)$ satisfying that $\sum_{j=1}^{M} p_{ij} = 1$ for all i.

(1) Multiprocess Kalman filter

In derivation of the filtering and estimation of the model, we need to solve two filtering problems.

We begin with the recursion between $\boldsymbol{\xi}^i_{t-1|t-1}$ (and $\boldsymbol{P}^i_{t-1|t-1}$) and $\boldsymbol{\xi}^j_{t|t}$ (and $\boldsymbol{P}^j_{t|t}$). Conditional on $S_{t-1} = i$ and $S_t = j$, $i, j = 1, \ldots, M$, we have

$$\begin{aligned}
\boldsymbol{\xi}^{ij}_{t|t-1} &= \boldsymbol{F}_j \boldsymbol{\xi}^i_{t-1|t-1} + \boldsymbol{c}_j, && (74)\\
\boldsymbol{P}^{ij}_{t|t-1} &= \boldsymbol{F}_j \boldsymbol{P}^i_{t-1|t-1} \boldsymbol{F}_j' + \boldsymbol{Q}_j, && (75)\\
\boldsymbol{\eta}^{ij}_{t|t-1} &= \boldsymbol{y}_t - \boldsymbol{H}_j \boldsymbol{\xi}^{ij}_{t|t-1} - \boldsymbol{d}_j, && (76)\\
\boldsymbol{\Sigma}^{ij}_{t|t-1} &= \boldsymbol{H}_j \boldsymbol{P}^{ij}_{t|t-1} \boldsymbol{H}_j' + \boldsymbol{R}_j, && (77)\\
\boldsymbol{\xi}^{ij}_{t|t} &= \boldsymbol{\xi}^{ij}_{t|t-1} + \boldsymbol{P}^{ij}_{t|t-1} \boldsymbol{H}_j' [\boldsymbol{\Sigma}^{ij}_{t|t-1}]^{-1} \boldsymbol{\eta}^{ij}_{t|t-1}, && (78)\\
\boldsymbol{P}^{ij}_{t|t} &= (\boldsymbol{I} - \boldsymbol{P}^{ij}_{t|t-1} \boldsymbol{H}_j' [\boldsymbol{\Sigma}^{ij}_{t|t-1}]^{-1} \boldsymbol{H}_j) \boldsymbol{P}^{ij}_{t|t-1}. && (79)
\end{aligned}$$

To reduce the dimension of the estimation problem, a *re-collapsing* procedure is suggested by Harrison and Stevens (1976) which effectively provides *re-approximation* at each time t.

These approximations are given by

$$\begin{aligned}
\boldsymbol{\xi}_{t|t}^{j} &= E[\boldsymbol{\xi}_t|S_t = j, \mathcal{F}_t] \\
&= \sum_{i=1}^{M} \Pr(S_{t-1} = i|S_t = j, \mathcal{F}_t) E[\boldsymbol{\xi}_t|S_t = j, S_{t-1} = i, \mathcal{F}_t] \\
&= \sum_{i=1}^{M} w_{t|t}^{ij} \boldsymbol{\xi}_{t|t}^{ij}, \qquad (80) \\
\boldsymbol{P}_{t|t}^{j} &= \sum_{i=1}^{M} w_{t|t}^{ij} \{\boldsymbol{P}_{t|t}^{ij} + (\boldsymbol{\xi}_{t|t}^{j} - \boldsymbol{\xi}_{t|t}^{ij})(\boldsymbol{\xi}_{t|t}^{j} - \boldsymbol{\xi}_{t|t}^{ij})'\}, \qquad (81)
\end{aligned}$$

where

$$
\begin{aligned}
w_{t|t}^{ij} &= \Pr(S_{t-1} = i | S_t = j, \mathcal{F}_t) \\
&= \frac{\Pr(S_{t-1} = i, S_t = j | \mathcal{F}_t)}{\Pr(S_t = j | \mathcal{F}_t)},
\end{aligned} \tag{82}
$$

$\boldsymbol{\xi}_{t|t}^{j}$ is an inference about $\boldsymbol{\xi}_t$ based on information up to time t given $S_t = i$, and $\boldsymbol{P}_{t|t}^{j}$ is the corresponding mean squared error.

(2) Hamilton filter

In equation (82), the probability terms $\Pr(S_{t-1} = i, S_t = j|\mathcal{F}_t)$ and $\Pr(S_t = j|\mathcal{F}_t)$ remain unknown. This leads to solving *the second filtering problem* as suggested by Hamilton (1989).

Given the term $\Pr(S_{t-1} = i|\mathcal{F}_{t-1})$ (for $i = 1, 2, \ldots, M$) at the beginning of the t-th iteration, we can compute the joint density of $S_{t-1} = i$ and $S_t = j$ conditional on $\mathcal{F}_{t-1}$, that is

$$\Pr(S_{t-1} = i, S_t = j|\mathcal{F}_{t-1}) = p_{ij} \Pr(S_{t-1} = i|\mathcal{F}_{t-1}) \tag{83}$$

for $i, j = 1, \ldots, M$, where $p_{ij} = \Pr(S_t = j|S_{t-1} = i)$ is the *transition probability*.

By *Bayes' theorem*, we have

$$\Pr(S_{t-1}=i,S_t=j|\mathcal{F}_t)=\frac{f(\boldsymbol{y}_t,S_{t-1}=i,S_t=j|\mathcal{F}_{t-1})}{f(\boldsymbol{y}_t|\mathcal{F}_{t-1})} \tag{84}$$
$$=\frac{f(\boldsymbol{y}_t|S_{t-1}=i,S_t=j,\mathcal{F}_{t-1})\Pr(S_{t-1}=i,S_t=j|\mathcal{F}_{t-1})}{f(\boldsymbol{y}_t|\mathcal{F}_{t-1})}.$$

where

$$f(\boldsymbol{y}_t|S_{t-1}=i,S_t=j,\mathcal{F}_{t-1})=\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_{t|t-1}^{ij}|^{\frac{1}{2}}}\exp\left\{-\frac{\boldsymbol{\eta}_{t|t-1}^{ij}{}'(\boldsymbol{\Sigma}_{t|t-1}^{ij})^{-1}\boldsymbol{\eta}_{t|t-1}^{ij}}{2}\right\},$$

and so the *conditional likelihood function* at time t is given by

$$f(\boldsymbol{y}_t|\mathcal{F}_{t-1})=\sum_{j=1}^{M}\sum_{i=1}^{M}\Pr(S_{t-1}=i,S_t=j|\mathcal{F}_{t-1})f(\boldsymbol{y}_t|S_{t-1}=i,S_t=j,\mathcal{F}_{t-1}). \tag{85}$$

For given parameters of the model, the Kalman filter provides us with the forecast error ($\boldsymbol{\eta}_{t|t-1}^{ij}$) and its variance ($\boldsymbol{\Sigma}_{t|t-1}^{ij}$) as by-products.

Also, the filtered density of S_t based on the information up to time t is given by

$$\Pr(S_t = j|\mathcal{F}_t) = \sum_{i=1}^{M} \Pr(S_{t-1} = i, S_t = j|\mathcal{F}_t). \tag{86}$$

Then we finish the second recursion between $\Pr(S_{t-1} = i|\mathcal{F}_{t-1})$ and $\Pr(S_t = j|\mathcal{F}_t)$.

(3) Kim's (1994) filter

Kim's (1994) filter for the state-space model with Markov switching may be viewed as a *combination* of the *Kalman filter* and the filter provided by Hamilton (1989) along with appropriate approximations.

The Kim's (1994) filter consists of the following steps:

Algorithm 1 (Kim filter)**.**

1. *Run the Kalman filter given in the Eqs.* (74) *to* (79) *for* $i, j = 1, \ldots, M$*.*

2. *Calculate* $\Pr(S_t = j, S_{t-1} = i|\mathcal{F}_t)$ *and* $\Pr(S_t = j|\mathcal{F}_t)$ *for* $i, j = 1, \ldots, M$*.*

3. *Using these probability terms, collapse the* $(M \times M)$ *posteriors in the Eqs.* (78) *and* (79) *into* $(M \times 1)$ *using the Eqs.* (80) *and* (81)*.*

(4) Log-likelihood

As a by-product of running the above Kalman filter, the conditional log-likelihood function can be obtained from step 2 using the equation (85).

The sample conditional log-likelihood is

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{t=1}^{T} \log f(\boldsymbol{y}_t|\mathcal{F}_{t-1}) \\
&= \sum_{t=1}^{T} \log \left\{ \sum_{i,j} \Pr(S_{t-1}=i, S_t=j|\mathcal{F}_{t-1}) f(\boldsymbol{y}_t|S_t=j, S_{t-1}=i, \mathcal{F}_{t-1}) \right\} \\
&= \sum_{t=1}^{T} \log \left\{ \sum_{i,j} \frac{\Pr(S_{t-1}=i, S_t=j|\mathcal{F}_{t-1})}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_{t|t-1}^{ij}|^{\frac{1}{2}}} \exp\left(-\frac{{\boldsymbol{\eta}_{t|t-1}^{ij}}'(\boldsymbol{\Sigma}_{t|t-1}^{ij})^{-1}\boldsymbol{\eta}_{t|t-1}^{ij}}{2} \right) \right\},
\end{aligned}$$

where $i, j \in \{1, \ldots, M\}$.

(5) Smoothing algorithm

- Once parameters are estimated, we can make inferences about S_t and $\boldsymbol{\xi}_t$ by using the entire information in the sample $\mathcal{F}_T$, that is we can determine $\Pr(S_t = j|\mathcal{F}_T)$ and $\boldsymbol{\xi}_{t|T}$ for $t = 1, 2, \ldots, T$.

- For this, consider the *joint probability* of the event that $S_t = j$ and $S_{t+1} = k$ based on all information from the full sample:

$$
\begin{aligned}
&\Pr(S_t = j, S_{t+1} = k|\mathcal{F}_T) = \Pr(S_{t+1} = k|\mathcal{F}_T) \times \Pr(S_t = j|S_{t+1} = k, \mathcal{F}_T) \\
&= \Pr(S_{t+1} = k|\mathcal{F}_T) \times \Pr(S_t = j|S_{t+1} = k, \mathcal{F}_t) \\
&= \Pr(S_{t+1} = k|\mathcal{F}_T) \times \frac{\Pr(S_t = j, S_{t+1} = k|\mathcal{F}_t)}{\Pr(S_{t+1} = k|\mathcal{F}_t)} \\
&= \Pr(S_{t+1} = k|\mathcal{F}_T) \times \frac{p_{jk} \Pr(S_t = j|\mathcal{F}_t)}{\Pr(S_{t+1} = k|\mathcal{F}_t)} \qquad (87)
\end{aligned}
$$

and

$$\Pr(S_t = j|\mathcal{F}_T) \quad = \quad \sum_{k=1}^{M} \Pr(S_t = j, S_{t+1} = k|\mathcal{F}_T). \tag{88}$$

- Keeping Eq.(87) in mind, we now turn to the derivation of the smoothing algorithm for the vector $\boldsymbol{\xi}_t$.

- Given that $S_t = j$ and $S_{t+1} = k$, the smoothing algorithm can be written as follows:

$$\boldsymbol{\xi}_{t|T}^{(j,k)} \quad = \quad \boldsymbol{\xi}_{t|t}^{j} + \boldsymbol{\Upsilon}_t^{(j,k)}(\boldsymbol{\xi}_{t+1|T}^{k} - \boldsymbol{\xi}_{t+1|t}^{(j,k)}), \tag{89}$$

$$\boldsymbol{P}_{t|T}^{(j,k)} \quad = \quad \boldsymbol{P}_{t|t}^{j} + \boldsymbol{\Upsilon}_t^{(j,k)}(\boldsymbol{P}_{t+1|T}^{k} - \boldsymbol{P}_{t+1|t}^{(j,k)})\boldsymbol{\Upsilon}_t^{(j,k)\prime}, \tag{90}$$

where $\boldsymbol{\Upsilon}_t^{(j,k)} = \boldsymbol{P}_{t|t}^{j}\boldsymbol{F}_k'[\boldsymbol{P}_{t+1|t}^{(j,k)}]^{-1}$.

Since $\Pr(S_t = j|\mathcal{F}_T)$ does not depend on $\boldsymbol{\xi}_{t|T}$, we can first calculate smoothed probabilities which we can use then to obtain smoothed values of $\boldsymbol{\xi}_t$, i.e. $\boldsymbol{\xi}_{t|T}$.

So the above smoothing algorithm can be performed by applying the following approximation approach:

Algorithm 2 (Smoothing algorithm)**.**

1. *We run through the basic filter and store the resulting sequences* $\boldsymbol{\xi}_{t|t}^{(i,j)}$, $\boldsymbol{P}_{t|t}^{(i,j)}$, $\boldsymbol{\xi}_{t|t}^{j}$, $\boldsymbol{P}_{t|t}^{j}$, $\Pr(S_t = j|\mathcal{F}_{t-1})$ *and* $\Pr(S_t = j|\mathcal{F}_t)$*.*

2. *For* $t = T-1, T-2, \ldots, 1$*, we obtain the smoothed probabilities* $\Pr(S_t = j, S_{t+1} = k|\mathcal{F}_T)$ *and* $\Pr(S_t = j|\mathcal{F}_T)$ *according to* (87) *and* (88) *and save these. Here,* $\Pr(S_T = j|\mathcal{F}_T)$*, the initial value for smoothing, is given by the final iteration of the basic filter.*

3. At each iteration of (89) *and* (90) *for* $t = T-1, T-2, \ldots, 1$*, we collapse the* $M \times M$ *elements of the matrices* $\boldsymbol{\xi}_{t|T}^{(j,k)}$ *and* $\boldsymbol{P}_{t|T}^{(j,k)}$ *into* M *weighted avarages, by taking weighted avarages over the state variable* S_{t+1}*:*

$$\begin{aligned}
\boldsymbol{\xi}_{t|T}^{j} &= \sum_{k=1}^{M} w_{t|T}^{jk} \boldsymbol{\xi}_{t|T}^{(j,k)}, \\
\boldsymbol{P}_{t|T}^{j} &= \sum_{k=1}^{M} w_{t|T}^{jk} \left\{ \boldsymbol{P}_{t|T}^{(j,k)} + (\boldsymbol{\xi}_{t|T}^{j} - \boldsymbol{\xi}_{t|T}^{(j,k)})(\boldsymbol{\xi}_{t|T}^{j} - \boldsymbol{\xi}_{t|T}^{(j,k)})' \right\},
\end{aligned}$$

where

$$w_{t|T}^{jk} = \frac{\Pr(S_t = j, S_{t+1} = k | \mathcal{F}_T)}{\Pr(S_t = j | \mathcal{F}_T)}.$$

4. By taking a weighted average over the M *Markov-regime at time* t*, the smoothed value is computed as* $\boldsymbol{\xi}_{t|T} = \sum_{j=1}^{M} \Pr(S_t = j | \mathcal{F}_T) \boldsymbol{\xi}_{t|T}^{j}$.

(6) Markov switching SV model

Consider the following Markov switching stochastic volatility model:

$$y_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \tag{91}$$

$$h_t = c_{s_t} + \lambda h_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2). \tag{92}$$

The discrete state variable s_t is assumed to be a first-order Markov process with the *transition probabilities*

$$\Pr(s_t = 0 | s_{t-1} = 0) = p_{00}, \qquad \Pr(s_t = 0 | s_{t-1} = 1) = 1 - p_{11},$$

$$\Pr(s_t = 1 | s_{t-1} = 0) = 1 - p_{00}, \qquad \Pr(s_t = 1 | s_{t-1} = 1) = p_{11}.$$

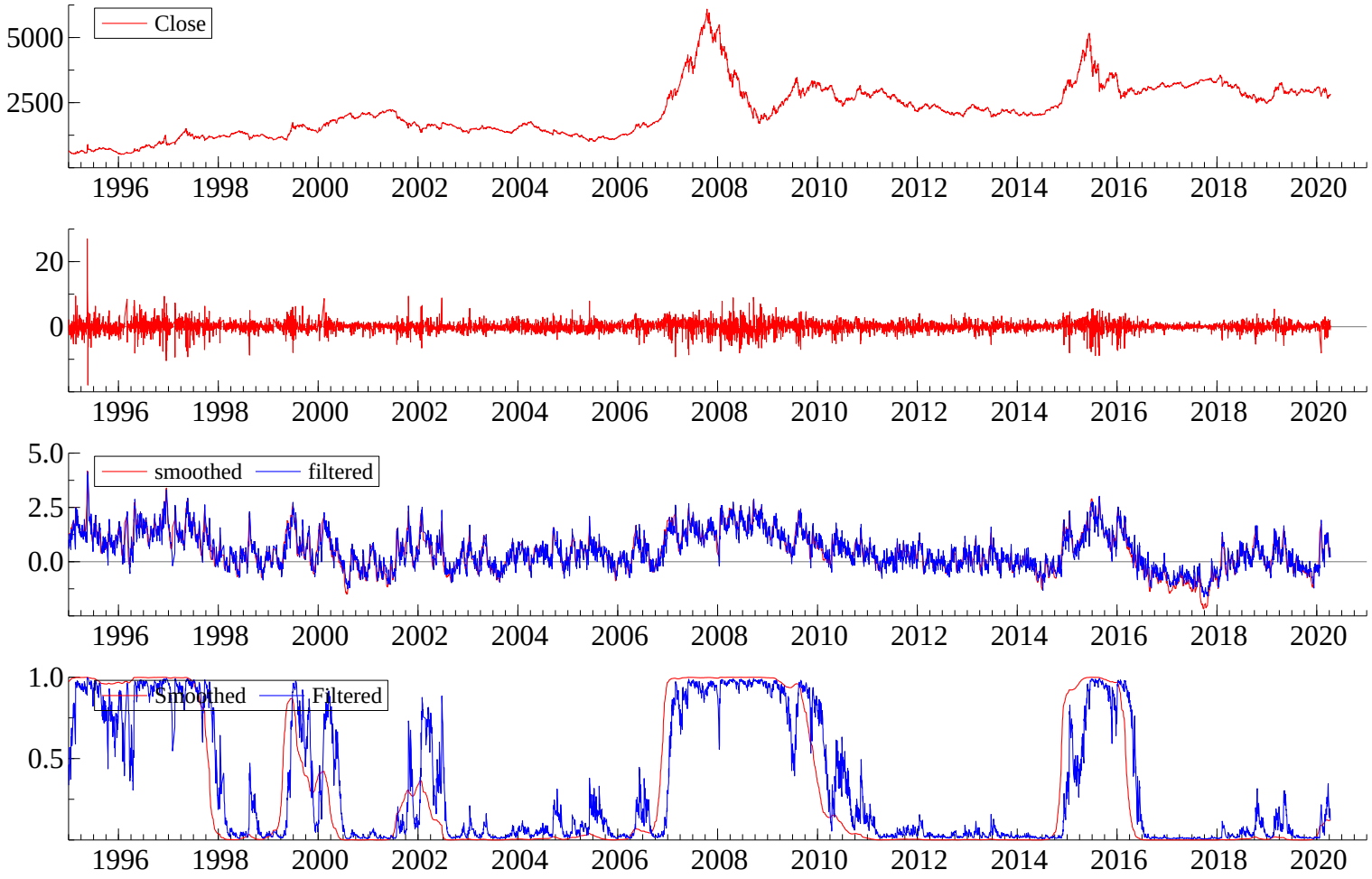
Taking the square and the logarithm, Equation (6) can be rewritten as:

$$\ln(y_t^2) = h_t + \ln(\varepsilon_t^2).$$

We then represent the model as follows:

$$\begin{aligned}
\ln(y_t^2) &= h_t - 1.2704 + u_t, \quad u_t \sim N(0, \pi^2/2), \\
h_t &= c_{s_t} + \lambda h_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2), \\
s_t | s_{t-1}, \mathcal{F}_{t-1} &\sim Markov(P), \\
P &= \begin{bmatrix} p_{00} & 1 - p_{11} \\ 1 - p_{00} & p_{11} \end{bmatrix}.
\end{aligned}$$

Example 2. Fitting Shanghai composite index with the MSSV model. The sample period is from 1995/01/01 to 2020/4/10.



The parameter estimates are reported as follows:

	Coef.Est	Std.Err	t-value	t-prob
p11	0.998794	0.000697	1433.840978	0.000000
p22	0.997625	0.001342	743.540344	0.000000
nu1	0.005172	0.006971	0.741902	0.458175
nu2	0.135570	0.025345	5.348979	0.000000
phi	0.906501	0.016639	54.480587	0.000000
sigma	0.341337	0.030876	11.055149	0.000000

Log likelihood = -13571.1
Akaika Information Criterion = 4.42468
Schwartz Bayesian Criterion = 4.43125

4 Mixed-Frequency Models and Applications

- Mariano and Murasawa (2003): A new coincident index of business cycles based on monthly and quarterly series. Journal of Applied Econometrics, 18, 427 – 443.

- 郑挺国、王霞（2013）：中国经济周期的混频数据测度及实时分析，《经济研究》第6期，58-70.

- Mariano and Murasawa (2010): A Coincident Index, Common Factors, and Monthly Real GDP, Oxford Bulletin of Economics and Statistics, 72(1), 27-46.

4.1 Mixed-frequency dynamic factor model

- $\{Y_{1,t}\}_{t=-\infty}^{\infty}$: an N_1-variate random sequence of quarterly indicators observable every third period; $\{Y_{2,t}\}_{t=-\infty}^{\infty}$: an N_2-variate random sequence of monthly indicators;

- $N = N_1 + N_2$;

- Assume that logs of the indicators are integrated of order 1;

- Let $Y_{1,t}$ be the geometric mean of $Y_{1,t}^*$, $Y_{1,t-1}^*$, and $Y_{1,t-2}^*$, i.e,

$$\ln Y_{1,t} = \frac{1}{3}(\ln Y_{1,t}^* + \ln Y_{1,t-1}^* + \ln Y_{1,t-2}^*)$$

  where $\{Y_{1,t}^*\}_{t=-\infty}^{\infty}$: an N_1-variate latent random sequence such that for all t,

- Taking the three-period differences, for all t,

$$\begin{aligned}
y_{1,t} &= \ln Y_{1,t} - \ln Y_{1,t-3} \\
&= \frac{1}{3}(\ln Y^*_{1,t} - \ln Y^*_{1,t-3}) + \frac{1}{3}(\ln Y^*_{1,t-1} - \ln Y^*_{1,t-4}) \\
&\quad + \frac{1}{3}(\ln Y^*_{1,t-2} - \ln Y^*_{1,t-5}) \\
&= \frac{1}{3}(y^*_{1,t} + y^*_{1,t-1} + y^*_{1,t-2}) + \frac{1}{3}(y^*_{1,t-1} + y^*_{1,t-2} + y^*_{1,t-3}) \\
&\quad + \frac{1}{3}(y^*_{1,t-2} + y^*_{1,t-3} + y^*_{1,t-4}) \\
&= \frac{1}{3}y^*_{1,t} + \frac{2}{3}y^*_{1,t-1} + y^*_{1,t-2} + \frac{2}{3}y^*_{1,t-3} + \frac{1}{3}y^*_{1,t-4}, \qquad (93)
\end{aligned}$$

  where $y_{1,t} = \Delta_3 \ln Y_{1,t}$ and $y^*_{1,t} = \Delta \ln Y^*_{1,t}$.

- $y_{1,t}$ is observed every third period, but $y^*_{1,t}$ is unobserved.

Now build the dynamic factor model for $y_{1,t}^*$ and $y_{2,t}$, which are in the same frequency. The model is given by

$$\begin{bmatrix} y_{1,t}^* \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1^* \\ \mu_2 \end{bmatrix} + \beta f_t + u_t, \tag{94}$$

$$\phi_f(L) f_t = v_{1,t}, \tag{95}$$

$$\Phi_u(L) u_t = v_{2,t}, \tag{96}$$

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \end{bmatrix} = NID\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right). \tag{97}$$

For identification, assume that: (i) the first component of β is 1; (ii) $\Phi_u(\cdot)$ and Σ_{22} is diagonal.

Since we never observe $y_{1,t}^*$, we cannot estimate the above model directly.

Hence we consider the corresponding dynamic one-factor model for $\{y_t\}_{t=-\infty}^{\infty}$ such that for all t,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \beta_1 \left(\frac{1}{3} f_t + \frac{2}{3} f_{t-1} + f_{t-2} + \frac{2}{3} f_{t-3} + \frac{1}{3} f_{t-4} \right) \\ \beta_2 f_t \end{bmatrix} + \begin{bmatrix} \frac{1}{3} u_{1,t} + \frac{2}{3} u_{1,t-1} + u_{1,t-2} + \frac{2}{3} u_{1,t-3} + \frac{1}{3} u_{1,t-4} \\ u_{2,t} \end{bmatrix},$$

where $\mu_1 = 3\mu_1^*$, $(\beta_1', \beta_2')' = \beta$, and $(u_{1,t}', u_{2,t}')' = u_t$.

4.2 Mixed-frequency Markov switching dynamic factor model

By including regime switching dynamics in the intercept, we have

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \beta_1 \left(\frac{1}{3} f_t + \frac{2}{3} f_{t-1} + f_{t-2} + \frac{2}{3} f_{t-3} + \frac{1}{3} f_{t-4} \right) \\ \beta_2 f_t \end{bmatrix} + \begin{bmatrix} \frac{1}{3} u_{1,t} + \frac{2}{3} u_{1,t-1} + u_{1,t-2} + \frac{2}{3} u_{1,t-3} + \frac{1}{3} u_{1,t-4} \\ u_{2,t} \end{bmatrix},$$

$$\phi_f(L) f_t = \nu(S_t) + v_{1,t},$$

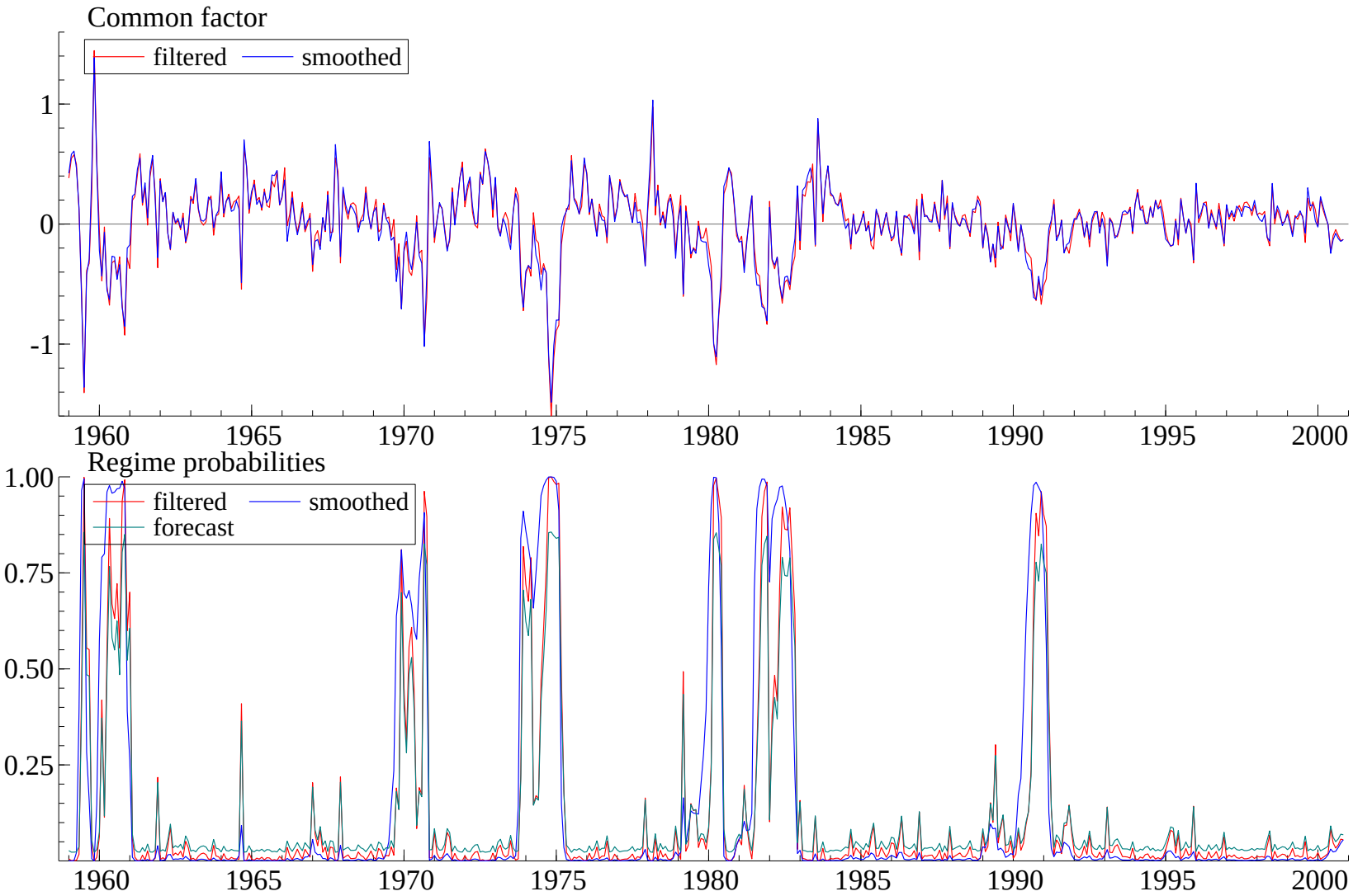
$$\Phi_u(L) u_t = v_{2,t},$$

$$S_t | S_{t-1} \sim Markov(P),$$

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \end{bmatrix} = NID \left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \right).$$

Example 3. Using the data in Mariano and Murasawa (2003).

```
 Strong convergence using analytical derivatives
 Log-likelihood value = -1273.59
============================================================
          Results of Parameter Estimations for Models
------------------------------------------------------------
                 Coef.Est     Std.Err     t-value      t-prob
p11              0.856730    0.068624   12.484343    0.000000
p22              0.977691    0.009061  107.904145    0.000000
beta2            0.508209    0.039172   12.973795    0.000000
beta3            0.817620    0.062672   13.046035    0.000000
beta4            2.154900    0.136030   15.841397    0.000000
beta5            1.755300    0.120708   14.541699    0.000000
nu1             -0.361690    0.066312   -5.454362    0.000000
nu2              0.057276    0.015802    3.624598    0.000320
phi_f1           0.334640    0.060548    5.526812    0.000000
sig_f            0.244907    0.018122   13.514373    0.000000
phi_u_11        -0.821450    0.199619   -4.115092    0.000046
phi_u_12         0.034668    0.253728    0.136634    0.891378
phi_u_21         0.094243    0.045042    2.092351    0.036934
phi_u_22         0.448270    0.050138    8.940794    0.000000
phi_u_31        -0.055209    0.051342   -1.075326    0.282771
phi_u_32         0.032623    0.051215    0.636987    0.524439
phi_u_41        -0.025256    0.064229   -0.393216    0.694335
phi_u_42        -0.058011    0.060438   -0.959844    0.337619
phi_u_51        -0.406940    0.049169   -8.276412    0.000000
phi_u_52        -0.192700    0.047797   -4.031632    0.000064
sig_u_1          0.494360    0.091277    5.416042    0.000000
sig_u_2          0.130292    0.006423   20.285157    0.000000
sig_u_3          0.296111    0.010835   27.328795    0.000000
sig_u_4          0.505410    0.023909   21.138977    0.000000
sig_u_5          0.783600    0.027914   28.071702    0.000000
============================================================
```


Common factor
filtered
smoothed
1
0
-1
1960
1965
1970
1975
1980
1985
1990
1995
2000
Regime probabilities
filtered
smoothed
forecast
1.00
0.75
0.50
0.25
1960
1965
1970
1975
1980
1985
1990
1995
2000

Example 4. Large dynamic factor models, forecasting, and nowcasting

See `https://gist.github.com/ChadFulton/a9172cd6f4194733f47`

5 Nonlinear and non-Gaussian state space models

- **Reference:**
  - Cappé, O., Godsill, S. J., and Moulines, E., (2007), An overview of existing methods and recent advances in sequential Monte Carlo, IEEE Proc., 95, 899 – 924.

- **Outline:**
  - Nonlinear and non-Gaussian state space models
  - Importance sampling and resampling
  - Sequential Monte Carlo filters

5.1 The Nonlinear and Non-Gaussian SS Models

A statistical model generates the observed data, a vector y_t, and the conditional distribution of y_t depends on a latent state variable, x_t. See the Figure.

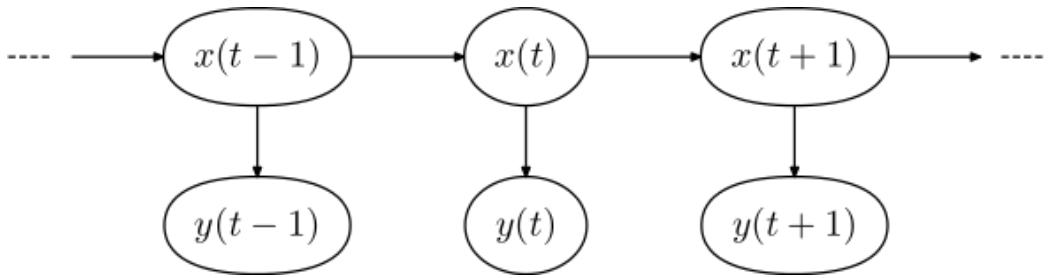


Figure 2 The State-Space (The Hidden states given by the vector x, and the observation states given by the vector y).

Formally, the data is generated by the following state space model

$$\text{Observation equation:} \quad y_t = f(x_t, \varepsilon_t^y) \quad \Rightarrow \quad p(y_t|x_t) \tag{98}$$

$$\text{State equation:} \quad x_t = g(x_{t-1}, \varepsilon_t^x) \quad \Rightarrow \quad p(x_t|x_{t-1}), \tag{99}$$

where ε_t^y is the observation error or "noise", and ε_t^x is the state shock.

The observation equation is often written as a conditional likelihood, $p(y_t|x_t)$, and the state evolution as $p(x_t|x_{t-1})$.

For simplicity, we ignore the static parameters θ in these distributions.

Our interest is mainly in $p(x_t|y_{0:t})$, where $y_{0:t} = (y_0, y_1, \ldots, y_t)$.

Beginning with the Kalman filter, computing $p(x_t|y_{0:t})$ uses a two-step procedure: prediction and Bayesian updating.

- Prediction

$$p(x_t|y_{0:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{0:t-1})dx_{t-1},$$

- Updating by Bayes rule

$$\underbrace{p(x_t|y_{0:t})}_{\text{Posterior}} = \frac{p(x_t, y_t|y_{0:t-1})}{p(y_t|y_{0:t-1})} = \frac{p(y_t|x_t)p(x_t|y_{0:t-1})}{\int p(y_t|x_t)p(x_t|y_{0:t-1})dx_t}$$

$$\propto \underbrace{p(y_t|x_t)}_{\text{Likelihood}} \underbrace{p(x_t|y_{0:t-1})}_{\text{Prior}}.$$

We now need the computation of the two integrals.

- Only for a limited number of settings such as linear Gaussian model, $p(x_t|y_{0:t-1})$ and $p(x_t|y_{0:t})$ are Gaussian. In this case, the Kalman filter recursion is often used to get the mean and variance of $p(x_t|y_{0:t})$.

- However, in nonlinear and non-Gaussian models, it is not possible to analytically compute $p(x_t|y_{0:t})$ since it is a complicated function of $y_{0:t}$. Simulation methods are typically required to characterize $p(x_t|y_{0:t})$.

5.2 Importance Sampling and Resampling

In the Monte Carlo method, we are concerned with estimating the properties of some highly complex probability distribution p, for example computing expectations of the form:

$$\bar{h} = \int h(x)p(x)dx,$$

where $h(\cdot)$ is some useful function for estimation, for example $h(x) = x$.

Sample from p

When the integral cannot be achieved analytically, we can generate random samples (called as *particles*) from p, denote these $\{x^{(i)}\}_{1\leq i\leq N}$, and then approximate the distribution p by point masses so that

$$\bar{h} \approx \frac{1}{N}\sum_{i=1}^{N} h(x^{(i)}).$$

Clearly N needs to be large in order to give a good coverage of all regions of interest.

Sample from q

When we cannot sample directly from the distribution p, we can sample from another distribution q (the importance distribution, or instrumental distribution). This is also called as the importance sampling.

Make N random draws (*particles*) $x^{(i)}$, $i = 1, \ldots, N$ from q instead of p. Consequently, $\bar{h}$ can be estimated using a weighted average:

$$\begin{aligned}\bar{h} &= \int h(x)\frac{p(x)}{q(x)}q(x)dx = \int h(x)r(x)q(x)dx \\ &= \sum_{j=1}^{N} \frac{\tilde{w}^{(j)}}{\sum_{j=1}^{N}\tilde{w}^{(j)}} h(x^{(j)}),\end{aligned}$$

where $\tilde{w}^{(j)} = r(x^{(i)}) = p(x^{(i)})/q(x^{(i)})$ is termed the unnormalised importance weight.

Definition: A sample $(x^{(j)}, \tilde{w}^{(j)})$, $j = 1, \ldots, N$ is said to be properly weighted with respect to distribution p if for all integrable function h, we have

$$\sum_{j=1}^{N} \frac{\tilde{w}^{(j)}}{\sum_{i=1}^{N} \tilde{w}^{(i)}} h(x^{(j)}) \to E_p[h(X)] \qquad \text{as } N \to \infty.$$

Sampling importance resampling (SIR)

In the *first stage of importance sampling*, an i.i.d. sample $(\tilde{x}^{(1)}, \ldots, \tilde{x}^{(M)})$ is drawn from the importance distribution q, and one computes the normalised version of the importance weights,

$$w^{(i)} = \frac{\tilde{w}^{(i)}}{\sum_{i=1}^{M} \tilde{w}^{(i)}}, \quad i = 1, \ldots, M. \tag{100}$$

In the *resampling stage*, a sample of size N denoted by $x^{(1)}, \ldots, x^{(N)}$ is drawn from the intermediate set of points $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(M)}$, taking proper account of the weights computed in (100).

This principle is illustrated in Figure 3.

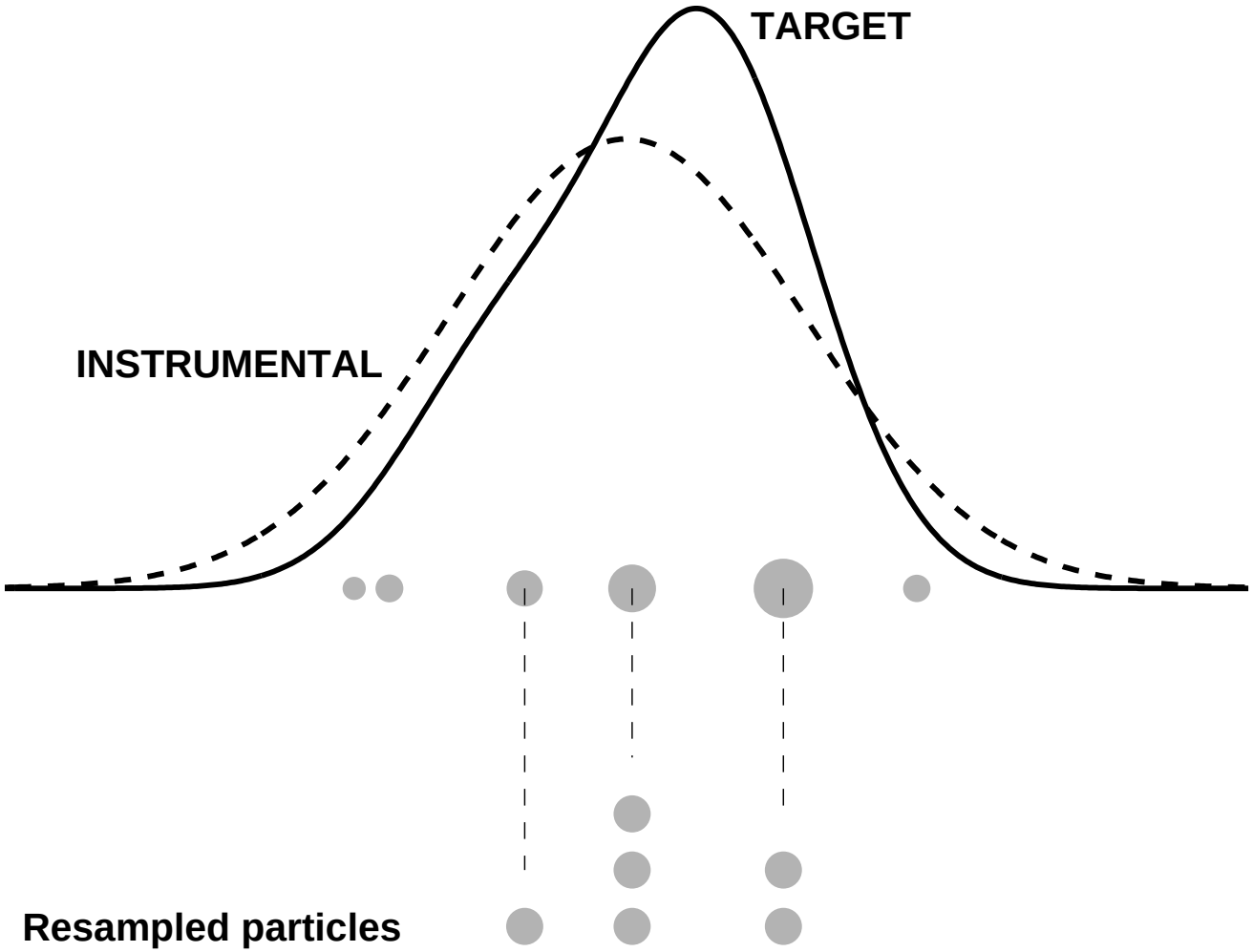


Figure 3 Principle of resampling. $(M = N = 6)$

- The particle approximation can be transformed into an equally weighted random sample from $p(x)$ by resampling from the discrete distribution, $\{\tilde{x}^{(j)}, \tilde{w}^{(j)}\}_{j=1}^{N}$.

- This procedure produces a new sample with uniformly distributed weights, $w_t^{(j)} = 1/N$.

- Resampling can be done in many ways, but the simplest is multinomial sampling. Other methods include stratified sampling, residual resampling, and systematic resampling.

5.3 Sequential Monte Carlo methods

Starting with the initial, or 'prior', distribution $p(x_0)$, *the (on-line) posterior density* $p(x_{0:t}|y_{0:t})$ can be obtained *recursively*:

- Prediction

$$p(x_{0:t}|y_{0:t-1}) = p(x_{0:t-1}|y_{0:t-1})p(x_t|x_{t-1}) \tag{101}$$

- Correction (or updating)

$$\begin{aligned} p(x_{0:t}|y_{0:t}) &= \frac{p(y_t|x_t)p(x_{0:t}|y_{0:t-1})}{\ell(y_t|y_{0:t-1})} \\ &= \frac{p(y_t|x_t)p(x_t|x_{t-1})}{\ell(y_t|y_{0:t-1})} p(x_{0:t-1}|y_{0:t-1}). \end{aligned} \tag{102}$$

We would like to sample from $p(x_{0:t}|y_{0:t})$; since it is generally impossible to sample directly from this distribution, we resort to a sequential version of the importance sampling and resampling procedure.

Conceptually, we $\tilde{x}_{0:t}^{(i)}$, $i = 1, \ldots, N$, from a convenient importance distribution $q(x_{0:t}|y_{0:t})$, and compute the unnormalised importance weights

$$\tilde{w}_t^{(i)} = \frac{p(\tilde{x}_{0:t}^{(i)}|y_{0:t})}{q(\tilde{x}_{0:t}^{(i)}|y_{0:t})}, \quad i = 1, \ldots, N.$$

We can get the normalised weights with $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^N \tilde{w}_t^{(i)}}$. Consequently,

$$\hat{p}(x_{0:t}|y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{\tilde{x}_{0:t}^{(i)}}(x_{0:t}), \tag{103}$$

where $\delta(\cdot)$ is a dirac function.

To achieve the importance distribution we construct the proposal such that it factorises in a form similar to that of the target posterior distribution:

$$q(x_{0:t}|y_{0:t}) = \overbrace{q(x_{0:t-1}|y_{0:t-1})}^{\text{Keep existing path}} \times \overbrace{q(x_t|x_{t-1}, y_t)}^{\text{extend path}}.$$

The unnormalised importance weights then take the following form

$$\tilde{w}_t^{(i)} = \frac{p(\tilde{x}_{0:t}^{(i)}|y_{0:t})}{q(\tilde{x}_{0:t}^{(i)}|y_{0:t})} \propto w_{t-1}^{(i)} \times \frac{p(y_t|\tilde{x}_t^{(i)})p(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)})}{q(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}, y_t)}. \tag{104}$$

In the sequential Monte Carlo literature, the multiplicative update factor on the right-hand side of (104) is often called the *incremental weight*.

(1) SIS/SIR algorithm

The sequential importance sampling (SIS) algorithm is summarized as follows:

Algorithm 3 (*Sequential Importance Sampling (SIS)*)**.**

1. At initial time $t = 0$*,*

(a) Sample $\tilde{x}_0^{(i)} \sim q(x_0)$*, for* $i = 1, \ldots, N$*.*

(b) Compute the importance weights $w_0^{(i)} = \tilde{w}_0^{(i)} / \sum_{i=1}^N \tilde{w}_0^{(i)}$*, where*

$$\tilde{w}_0^{(i)} = \frac{p(y_0|\tilde{x}_0^{(i)})p(\tilde{x}_0^{(i)})}{q(\tilde{x}_0^{(i)})}, \quad \textit{for } i = 1, \ldots, N.$$

2. For $t = 1, \ldots, T$*,*

(a) For $i = 1, \ldots, N$*, propagate particles:* $\tilde{x}_t^{(i)} \sim q(x_t|\tilde{x}_{t-1}^{(i)}, y_t)$*,*

compute weights:

$$\tilde{w}_t^{(i)} = w_{t-1}^{(i)} \frac{p(y_t|\tilde{x}_t^{(i)}) p(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)})}{q(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}, y_t)}.$$

(b) *Normalise weights:* $w_t^{(i)} = \tilde{w}_t^{(i)} / \sum_{i=1}^N \tilde{w}_t^{(i)}$, $i = 1, \ldots, N$, *compute filtering estimate (on-line filtered estimate):* $\bar{h}_t = \sum_{i=1}^N w_t^{(i)} h_t(\tilde{x}_t^{(i)})$.

The problem in SIS is that the algorithm is degenerate. It is known that the variance of the weights increases at every step. This means that we will always converge to a single non-zero weight $w^{(i)} = 1$ for some i and the rest being zero.

Resampling strategy is commonly used to avoid the problem of degeneracy of the algorithm.

The sampling importance resampling (SIR) algorithm can be summarized as follows:

Algorithm 4 (*Sampling Importance Resampling (SIR)*)**.**

1. At initial time $t = 0$*.*

(a) Sample $\tilde{x}_0^{(i)} \sim q(x_0)$*, for* $i = 1, \ldots, N$*.*

(b) Compute importance weights $w_0^{(i)} = \tilde{w}_0^{(i)} / \sum_{i=1}^N \tilde{w}_0^{(i)}$*, where*

$$\tilde{w}_0^{(i)} = \frac{p(y_0|\tilde{x}_0^{(i)})p(\tilde{x}_0^{(i)})}{q(\tilde{x}_0^{(i)})}, \quad \textit{for } i = 1, \ldots, N.$$

(c) Resample N *particles* $\{\tilde{x}_0^{(i)}\}_{i=1}^N$ *with weights* $\{\tilde{w}_0^{(i)}\}_{i=1}^N$ *to obtain the new particles* $\{x_0^{(i)}, 1/N\}_{i=1}^N$*.*

2. For $t = 1, \ldots, T$*,*

(a) For $i = 1, \ldots, N$, propagate particles: $\tilde{x}_t^{(i)} \sim q(x_t | x_{t-1}^{(i)}, y_t)$, compute weights:

$$\tilde{w}_t^{(i)} = w_{t-1}^{(i)} \frac{p(y_t|\tilde{x}_t^{(i)}) p(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})}{q(\tilde{x}_t^{(i)}|x_{t-1}^{(i)}, y_t)}.$$

(b) Normalise weights: $w_t^{(i)} = \tilde{w}_t^{(i)} / \sum_{i=1}^N \tilde{w}_t^{(i)}$, $i = 1, \ldots, N$, compute filtering estimate (on-line filtered estimate): $\bar{h}_t = \sum_{i=1}^N w_t^{(i)} h_t(\tilde{x}_t^{(i)})$.

(c) Resample N particles $\{\tilde{x}_t^{(i)}\}_{i=1}^N$ with weights $\{\tilde{w}_t^{(i)}\}_{i=1}^N$ to obtain the new particles $\{x_t^{(i)}, 1/N\}_{i=1}^N$.

Effective sample size (ESS)

An important technique for resampling and rejuvenating a sequential importance sampler is the *effective sample size* (ESS), which is a useful metric to determine whether resampling is necessary.

Define the coefficients of variation of the importance weights as $cv_t^2 = \frac{\text{Var}(w_t^{(i)})}{E^2(w_t^{(i)})}$, Liu and Chen (1995) and Liu (1996) propose tracking the ESS

$$N_{\text{eff},t} = \frac{N}{1 + cv_t^2},$$

which varies between 1 (N copies of a single particle) and N (equal weights).

If the ESS is large, say, $N_{\text{eff},t} > \alpha_0 N$ ($0 < \alpha_0 < 1$), resampling is undesirable since the pdf over the trajectories is well represented.

Replace Step 2(c) in SIR with the following step

(c) If resampling when $N_{eff,t} \leq \alpha_0 N$, select N particle indices $j_i \in \{1, \ldots, N\}$ according to weights $\{w_t^{(i)}\}_{1 \leq i \leq N}$, and set $x_t^{(i)} = \tilde{x}_t^{(j_i)}$ and $w_t^{(i)} = 1/N$, $i = 1, \ldots, N$; otherwise, set $x_t^{(i)} = \tilde{x}_t^{(i)}$, $i = 1, \ldots, N$.

(2) The bootstrap filter

The bootstrap filter is proposed by Gordon, Salmond, and Smith (1993)

- It uses the state transition density or "prior kernel" as importance distribution, i.e., $q(x_t|x_{t-1}, y_t) = p(x_t|x_{t-1})$.

- The importance weight then simplifies to

$$w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|x_t^{(i)}).$$

- In the original version of the algorithm, resampling is carried out at each and every time step, in which case the term $w_{t-1}^{(i)} = 1/N$ is a constant, which may thus be ignored.

The bootstrap filter can be summarized as follows:

Algorithm 5 (*The bootstrap filter)***.**

1. At initial time $t = 0$*.*

(a) Sample $x_0^{(i)} = \tilde{x}_0^{(i)} \sim p(x_0)$ *with associated weights* $w_0^{(i)} = 1/N$*.*

2. For $t = 1, \ldots, T$*,*

(a) For $i = 1, \ldots, N$*, propagate particles:* $\tilde{x}_t^{(i)} \sim p(x_t | x_{t-1}^{(i)})$*, compute weights:*

$$\tilde{w}_t^{(i)} = p(y_t | \tilde{x}_t^{(i)}).$$

(b) Normalise weights: $w_t^{(i)} = \tilde{w}_t^{(i)} / \sum_{i=1}^N \tilde{w}_t^{(i)}$*,* $i = 1, \ldots, N$*, compute filtering estimate (on-line filtered estimate):* $\bar{h}_t = \sum_{i=1}^N w_t^{(i)} h_t(\tilde{x}_t^{(i)})$*.*

(c) Resample according to weights $\{w_t^{(i)}\}_{1 \leq i \leq N}$*, and set* $x_t^{(i)} = \tilde{x}_t^{(j_i)}$ *and* $w_t^{(i)} = 1/N$*,* $i = 1, \ldots, N$*.*

(3) Auxiliary sampling

The auxiliary particle filter (APF) of Pitt and Shephard (1999, 2001) is a popular algorithm that is simple to implement and works well in many cases. They approximate the incremental target distribution in (102) with the importance distribution

$$\begin{aligned} p(y_t|x_t)p(x_t|x_{t-1}) &\approx g_1(y_t|x_t)g_2(x_t|x_{t-1}) \\ &= g_1(y_t|x_{t-1})g_2(x_t|x_{t-1}, y_t) \end{aligned} \tag{105}$$

The proposal distribution in (105) is decomposed into two parts implying that the sampling of new values $\{x_t^{(i)}\}_{i=1}^N$ from this distribution can potentially be performed in two steps.

The APF is given as follows:

Algorithm 6 (*Auxiliary Particle Filter (APF)*)**.**

1. At initial time $t = 0$*.*

(a) Sample $\tilde{x}_0^{(i)} \sim g_2(x_0)$*, for* $i = 1, \ldots, N$*.*

(b) Compute importance weights $w_0^{(i)} = \tilde{w}_0^{(i)} / \sum_{i=1}^N \tilde{w}_0^{(i)}$*, where*

$$\tilde{w}_0^{(i)} = \frac{p(y_0|\tilde{x}_0^{(i)})p(\tilde{x}_0^{(i)})}{g_2(\tilde{x}_0^{(i)})}, \quad \textit{for } i = 1, \ldots, N.$$

(c) Resample if necessary.

2. For $t = 1, \ldots, T$*,*

(a) For $i = 1, \ldots, N$*, compute* $\tau_{t-1}^{(i)} = g_1(y_t|\tilde{x}_{t-1}^{(i)})$ *and normalize* $v_{t-1}^{(i)} =$

$\frac{\tau_{t-1}^{(i)}}{\sum_{j=1}^{N}\tau_{t-1}^{(j)}}$.

(b) Select N particle indices $j_i \in \{1,\ldots,N\}$ according to weights $\{v_{t-1}^{(i)}\}_{i=1}^{N}$. For $i = 1,\ldots,N$, set $x_{t-1}^{(i)} = \tilde{x}_{t-1}^{(j_i)}$ and set first stage weights

$$u_{t-1}^{(i)} = \frac{w_{t-1}^{(j_i)}}{v_{t-1}^{(j_i)}}.$$

(c) For $i = 1,\ldots,N$, propagate particles: $\tilde{x}_t^{(i)} \sim g_2(x_t|x_{t-1}^{(i)}, y_t)$, compute weights:

$$\tilde{w}_t^{(i)} = u_{t-1}^{(i)} \frac{p(y_t|\tilde{x}_t^{(i)})p(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})}{g_2(\tilde{x}_t^{(i)}|x_{t-1}^{(i)}, y_t)} = \frac{p(y_t|\tilde{x}_t^{(i)})p(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})}{g_1(y_t|\tilde{x}_{t-1}^{(j_i)})g_2(\tilde{x}_t^{(i)}|x_{t-1}^{(i)}, y_t)}.$$

(d) Normalise weights: $w_t^{(i)} = \tilde{w}_t^{(i)} / \sum_{i=1}^{N} \tilde{w}_t^{(i)}$, $i = 1,\ldots,N$, compute filtering estimate (on-line filtered estimate): $\bar{h}_t = \sum_{i=1}^{N} w_t^{(i)} h_t(\tilde{x}_t^{(i)})$.

6 Score-Driven Models

6.1 Basic GAS model

Let $\{y_t\}$ be a time series and $\mathcal{F}_t = \{y_t, y_{t-1}, \ldots\}$ be available information up to t. Creal et al. (2013) proposed the framework of Generalized Autoregressive Score Models (GAS) under a general conditional distribution $p(y_t \mid \mathcal{F}_{t-1})$, which can be parameterized as

$$y_t \mid \mathcal{F}_{t-1} \sim p(y_t \mid f_t, \theta), \tag{106}$$

- f_t is the time-varying parameter of interest which fully characterizes $p(\cdot)$ and only depends on $\mathcal{F}_{t-1}$;

- θ is a collection of static additional parameters.

Furthermore, the time-varying parameter f_t is assumed to follow the familiar autoregressive updating equation

$$f_{t+1} = \omega + \sum_{j=1}^{p} A_j f_{t-j+1} + \sum_{j=1}^{q} B_j s_{t-j+1}, \tag{107}$$

where ω is a vector of constants, A_i and B_j are coefficient matrices. In (107), s_t is a vector which is proportional to the score of (106):

$$s_t = S_t \nabla_t \tag{108}$$

where S_t is a positive definite scaling matrix known at time t and

$$\nabla_t = \frac{\partial \log p(y_t \mid f_t, \theta)}{\partial f_t} \tag{109}$$

is the score of (107) evaluated at f_t.

Creal et al. (2013) suggest to set the scaling matrix S_t to a power $\gamma > 0$ of the inverse of the information matrix of f_t to account for the variance of ∇_t. More precisely:

$$S_t = I_t^{-\gamma} \quad \text{and} \quad I_t = E_{t-1}\left[\nabla_t \nabla_t'\right] \tag{110}$$

where the expectation is taken with respect to the conditional distribution of y_t given $\mathcal{F}_{t-1}$.

- The additional parameter γ is fixed by the user and usually takes values 0, 1/2 or 1.
- When $\gamma = 0$, $S_t = I$ and there is no scaling.
- If $\gamma = 1$ ($\gamma = 1/2$), the conditional score ∇_t is premultiplied by the inverse of (the square root of) its covariance matrix I_t.

6.2 Special cases of GAS models

(1) Regression model:

$$y_t = x_t'\beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2),$$

where x_t is a $k \times 1$ vector x_t of exogenous variables, β_{t-1} is a $k \times 1$ vector of time-varying regression coefficients.

Let $f_t = \beta_t$. The scaled score function based on $S_t = I_t^{-1}$ is given by

$$s_t = (x_t'x_t)^{-1}x_t'(y_t - x_t'f_{t-1}) = (x_t'x_t)^{-1}x_t'\varepsilon_t.$$

The GAS(1,1) specification for the time-varying regression coefficient is

$$f_t = \omega + A_0 s_t + B_1 f_{t-1} \tag{111}$$

In case $x_t \equiv 1$, the above equation for the time-varying intercept reduces to the EWMA recursion by setting $\omega = 0$ and $B_1 = 1$, i.e.

$$f_t = f_{t-1} + A_0(y_t - f_{t-1}).$$

In this case, we obtain the observation driven analogue of the local level (parameter driven) model,

$$y_t = \mu_{t-1} + \varepsilon_t, \qquad \mu_t = \mu_{t-1} + \eta_t,$$

see Durbin and Koopman (2001, Chapter 2).

(2) Unobserved component models

Consider a two component decomposition of a time series $\{y_t\}$. Let $f_t = (f_{1t}, f_{2t})$ be the vector of decomposed components.

For this decomposition we obtain the GAS(1,2) model with observation model

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

The updating equation is specified as

$$f_t = \begin{bmatrix} \omega \\ 0 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} s_t + \begin{bmatrix} 1 & 0 \\ 0 & \phi_1 \end{bmatrix} f_{t-1} + \begin{bmatrix} 0 & 0 \\ 0 & \phi_2 \end{bmatrix} f_{t-2}.$$

The scaled score function is given by

$$s_t = y_t - f_{1,t-1} - f_{2,t-1} = \varepsilon_t$$

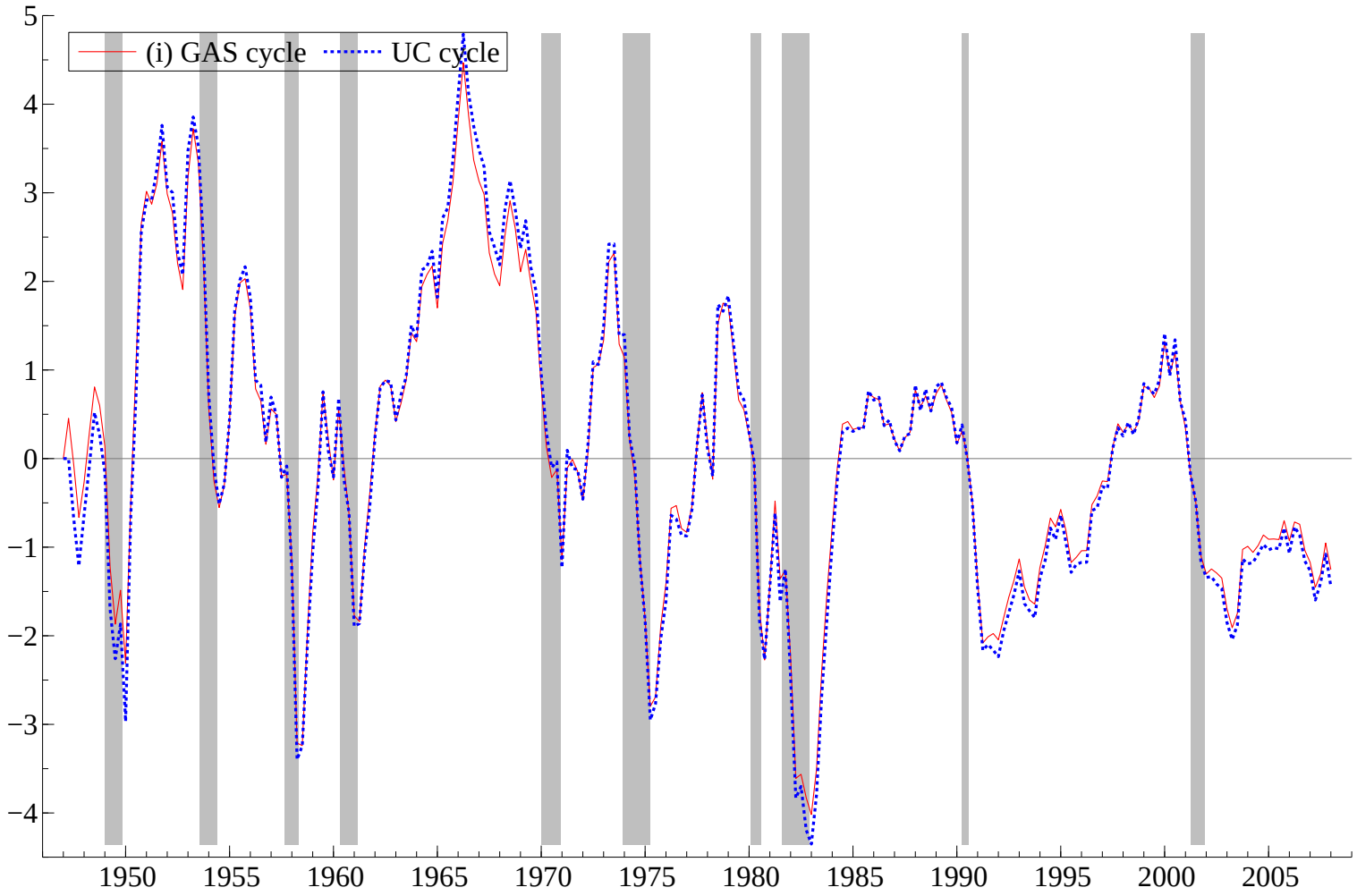
This GAS decomposition can be regarded as the observation driven equivalent of the UC models of Watson (1986) and Clark (1989), who also aim to decompose macroeconomic time series into trend and cycle factors.

The UC trend-cycle decomposition model is then given by

$$
\begin{aligned}
y_t &= f_{1,t} + f_{2,t} \\
f_{1t} &= \omega + f_{1,t-1} + a_1 \xi_{1,t}, \quad \xi_{1,t} \sim N(0,1), \\
f_{2t} &= \phi_1 f_{2,t-1} + \phi_2 f_{2,t-2} + a_2 \xi_{2,t}, \quad \xi_{2,t} \sim N(0,1),
\end{aligned}
$$

where the disturbances $\xi_{1,t}$ and $\xi_{2,t}$ are mutually and serially independent.

Example 5. Trend-cycle illustration: estimated cycles from the GAS and UC trend-cycle models based on quarterly log of U.S. real gdp from 1947(1) through 2008(1).



More GAS models with Python examples are released in

`https://pyflux.readthedocs.io/en/latest/gas.html`.