Hidden Markov Models

Contents

Download PDF here

Authors: Lars Peter Hansen (University of Chicago) and Thomas J. Sargent (NYU)

Fig. 6.1 Two images hidden behind paintings and discovered via x-rays: Vincent Van Gogh's earliest self-portrait (left); portrait of perhaps Joseph Bonaparte by Francisco Goya (right).

> "Truth is like the stars; it does not appear except from behind obscurity of the night." Khalil Gibran

> "If you torture the data long enough, it will confess to anything." Ronald Coase

6.1. Introduction

Markov models provide a natural framework for studying how to learn about states that are hidden from a statistician or decision maker who observes only noisy signals of some or all states. When states evolve over time, the statistician faces the challenge of learning about a moving target.

This chapter presents Hidden Markov Models (HMMs). We begin with a joint probability distribution consisting of a Markov process and a vector of noisy signals about functions of the Markov state. The statistician observes a history of signals, but not the Markov state vector itself. Learning about the Markov state involves constructing a sequence of probability distributions for the state, conditional on the observed signal histories. Recursive representations of these conditional distributions form auxiliary Markov processes that fully summarize the information about the hidden state vector contained in the signal history. The state vector in this auxiliary process serves as a set of sufficient statistics for the probability distribution of the hidden Markov state, given the signal history. We describe how to construct this auxiliary Markov process of sufficient statistics sequentially.

We present four examples of Hidden Markov Models used to learn about:

1. A continuously distributed hidden state vector in a linear state-space system
2. A discrete hidden state vector
3. Unknown parameters treated as hidden invariant states
4. Multiple VAR regimes

6.2. Kalman Filter and Smoother

We assume that a Markov state vector X_t and a vector Z_{t+1} of observations are governed by a linear state space system

$$\begin{aligned} X_{t+1} &= \mathbb{A} X_t + \mathbb{B} W_{t+1} \\ Z_{t+1} &= \mathbb{N} + \mathbb{D} X_t + \mathbb{F} W_{t+1}, \end{aligned} \tag{6.1}$$

where the matrix $\mathbb{F}\mathbb{F}'$ is nonsingular, X_t has dimension n, Z_{t+1} has dimension m and is a signal observed at $t+1$, W_{t+1} has dimension k and is a standard normally distributed random vector that is independent of X_t, of $Z^t = [Z_t, \ldots, Z_1]$, and of X_0. The initial state vector $X_0 \sim Q_0$, where Q_0 is a normal distribution with mean $\overline{X}_0$ and covariance matrix Σ_0.[1] To include the ability to represent an unknown fixed parameter as an invariant state associated with a unit eigenvalue in $\mathbb{A}$, we allow $\mathbb{A}$ not to be a stable matrix.

Remark 6.1

To connect this discussion to the VAR example of Section Vector autoregression in Chapter 4, suppose that X is partitioned as:

$$X_t = \begin{bmatrix} X_t^1 \\ \eta \end{bmatrix}$$

where

$$\begin{aligned} \mathbb{A} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbb{A}_{11} & 0 \\ 0 & 1 \end{bmatrix} \\ \mathbb{B} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbb{B}_1 \\ 0 \end{bmatrix}, \end{aligned}$$

where $\mathbb{A}_{11}$ is a stable matrix. In addition,

$$Z_{t+1} = \begin{bmatrix} Y_{t+1} - Y_t \\ Z_{t+1}^2 \end{bmatrix},$$

so $Y_{t+1} - Y_t$ becomes one of the date $t+1$ observations. Since we may wish to estimate the trend growth rate η, we include this as one of the hidden states. To avoid double counting, we let the first entry of the $\mathbb{N}$ vector be zero. While the previous chapter showed how to construct a martingale component when the state vector X_t is observable at date t, the analysis of the Kalman filter provides us with a different decomposition in which X_t is not known and must be inferred.

Although $\{(X_t, Z_t), t = 0, 1, 2, \ldots\}$ is Markov, $\{Z_t, t = 0, 1, 2, \ldots\}$ is typically not.[2] We want to construct an affiliated Markov process whose date t state is Q_t, defined to be the probability distribution of the time t Markov state X_t conditional on history $Z^t = Z_t, \ldots, Z_1$ and Q_0. The distribution Q_t summarizes information about X_t that is contained in the history Z^t and Q_0. We sometimes use Q_t to indicate conditioning information that is "random" in the sense that it is constructed from a history of observable random vectors. Because the distribution Q_t is multivariate normal, it suffices to keep track only of the mean vector $\overline{X}_t$ and covariance matrix Σ_t of X_t conditioned on Q_0 and Z^t: $\overline{X}_t$ and Σ_t are sufficient statistics for the probability distribution of X_t conditional on the history Z^t and Q_0. Conditioning on Q_t is equivalent to conditioning on these sufficient statistics.

We map sufficient statistics $(\overline{X}_{j-1}, \Sigma_{j-1})$ for Q_{j-1} into sufficient statistics $(\overline{X}_j, \Sigma_j)$ for Q_j by applying formulas for means and covariances of a conditional distribution associated with a multivariate normal distribution. This generates a recursion that maps Q_{j-1} and Z_j into Q_j. It enables us to construct $\{Q_t\}$ sequentially. Thus, consider the following three-step process.

1. Express the joint distribution of X_{t+1}, Z_{t+1} conditional on X_t as

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ \mathbb{N} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} X_t, \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} \begin{bmatrix} \mathbb{B}' & \mathbb{F}' \end{bmatrix} \right).$$

2. Suppose that the distribution Q_t of X_t conditioned on Z^t and Q_0 is normal with mean $\overline{X}_t$ and covariance matrix Σ_t. Use the identity $X_t = \overline{X}_t + (X_t - \overline{X}_t)$ to represent $\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix}$ as

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbb{N} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \overline{X}_t + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} (X_t - \overline{X}_t) + \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} W_{t+1},$$

which is just another way of describing our original state-space system (6.1). It follows that the joint distribution of X_{t+1}, Z_{t+1} conditioned on Z^t and Q_0, or equivalently on $(\overline{X}_t, \Sigma_t)$, is

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ \mathbb{N} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \overline{X}_t, \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \Sigma_t \begin{bmatrix} \mathbb{A}' & \mathbb{D}' \end{bmatrix} + \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} \begin{bmatrix} \mathbb{B}' & \mathbb{F}' \end{bmatrix} \right).$$

Evidently the marginal distribution of Z_{t+1} conditional on $(\overline{X}_t, \Sigma_t)$ is

$$Z_{t+1} \sim \text{Normal}(\mathbb{N} + \mathbb{D}\overline{X}_t, \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}').$$

The resulting density is called the predictive density $\phi(z^*|Q_t)$, i.e., the density of Z_{t+1} conditional on history Z^t and the initial distribution Q_0.

3. Joint normality implies that the distribution for X_{t+1} conditional on Z_{t+1} and $(\overline{X}_t, \Sigma_t)$ is also normal and fully characterized by a conditional mean vector and a conditional covariance matrix. We can compute the conditional mean by running a population regression of $X_{t+1} - \mathbb{A}\overline{X}_t$ on the surprise in Z_{t+1} defined as $Z_{t+1} - \mathbb{N} - \mathbb{D}\overline{X}_t$.[3] Having thus transformed random vectors on both sides of our regression to be independent of past observable information, as ingredients of the pertinent population regression, we have to compute the covariance matrices

$$E\left[\left(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t\right)\left(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t\right)' \mid \overline{X}_t, \Sigma_t\right] = \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \equiv \Omega_t$$
$$E\left[(X_{t+1}-\mathbb{A}\overline{X}_t)\left(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t\right)' \mid \overline{X}_t, \Sigma_t\right] = \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}'.$$

These provide what we need to compute the conditional expectation

$$E[(X_{t+1}-\mathbb{A}\overline{X}_t) \mid Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t, Q_t] = \mathcal{K}(\Sigma_t)(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t),$$

where the matrix of regression coefficients $\mathcal{K}(\Sigma_t)$ is called the **Kalman gain**. To compute the Kalman gain, multiply both sides by $(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t)'$ and take expectations conditioned on $(\overline{X}_t, \Sigma_t)$:

$$\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' = \mathcal{K}(\Sigma_t)\left(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}'\right)$$

Solving this equation for $\mathcal{K}(\Sigma_t)$ gives

$$\mathcal{K}(\Sigma_t) = (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}. \tag{6.2}$$

Recognize formula (6.2) as an application of the population least squares regression formula. This least-squares application is a valid way to compute conditional expectations because of its application in conjunction with the multivariate normal distribution.[4] We compute Σ_{t+1} via the recursion

$$\Sigma_{t+1} = \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' - (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}(\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')'. \tag{6.3}$$

The right side of recursion (6.3) follows directly from substituting the appropriate formulas into the right side of $\Sigma_{t+1} \equiv E(X_{t+1}-\overline{X}_{t+1})(X_{t+1}-\overline{X}_{t+1})'$ and computing conditional expectations. The matrix Σ_{t+1} obeys the formula from standard regression theory for the population covariance matrix of the least squares residual $X_{t+1}-\mathbb{A}\overline{X}_t$. The matrix $\mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}'$ is the covariance matrix of $X_{t+1}-\mathbb{A}\overline{X}_t$ and the remaining term describes the reduction in covariance associated with conditioning on Z_{t+1}. Thus, the probability distribution Q_{t+1} is

$$X_{t+1} \mid Z_{t+1}, \overline{X}_t, \Sigma_t \sim \text{Normal}(\overline{X}_{t+1}, \Sigma_{t+1})$$

where

$$\overline{X}_{t+1} = \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)(Z_{t+1}-\mathbb{N}-\mathbb{D}\overline{X}_t) \tag{6.4}$$

Equations (6.2), (6.3), and (6.4) constitute the Kalman filter. They provide a recursion that describes Q_{t+1} as an exact function of Z_{t+1} and Q_t.

We may view the outcome of the Kalman filter as taking an original process X with unobservable states and replacing it with one that lines up with the relevant conditioning information based on sequentially observed signals. This is reflected in the system:

$$\begin{aligned}\overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)U_{t+1}\\ Z_{t+1} &= \mathbb{N} + \mathbb{D}\overline{X}_t + U_{t+1},\\ \Sigma_{t+1} &= \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' - (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}(\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')'.\end{aligned} \tag{6.5}$$

where $\{(\overline{X}_t, Z_t, \Sigma_t) : t \geq 0\}$ is itself a Markov process. The shock process W used in the original specification is replaced by what is often called an *innovation process*, U, given by

$$U_{t+1} \overset{\text{def}}{=} Z_{t+1} - \mathbb{N} - \mathbb{D}\overline{X}_t.$$

The U process is itself normally distributed and temporally independent and provides new information pertinent for predicting current and future states. Specifically, after we condition on initial inputs $(\overline{X}_0, \Sigma_0)$, $U_t, U_{t-1}, \ldots, U_1$ and $Z_t, Z_{t-1}, \ldots, Z_1$ generate the same information.

Remark 6.2

The covariance matrix Ω_t is presumed to be nonsingular, but it is not necessarily diagonal so that components of the innovation vector U_{t+1} are possibly correlated. We can transform the innovation vector U_{t+1} to produce a new shock process $\overline{W}_{t+1}$ that has the identity as its covariance matrix. To do so, construct a matrix $\overline{\mathbb{F}}(\Sigma_t)$ that satisfies

$$\Omega_t = \overline{\mathbb{F}}(\Sigma_t)[\overline{\mathbb{F}}(\Sigma_t)]'.$$

The factorization on the right side is not unique. For instance, we could find solutions under which $\overline{\mathbb{F}}(\Sigma_t)$ is either lower or upper triangular, but there are other possibilities as well. In what follows we use one such factorization, while the analysis allows any of the possibilities. Construct:

$$\overline{W}_{t+1} \overset{\text{def}}{=} \left[\overline{\mathbb{F}}(\Sigma_t)\right]^{-1} U_{t+1}$$

Then

(6.6)

$$
\begin{aligned}
\overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \overline{\mathbb{B}}(\Sigma_t)\overline{W}_{t+1} \\
Z_{t+1} &= \mathbb{N} + \mathbb{D}\overline{X}_t + \overline{\mathbb{F}}(\Sigma_t)\overline{W}_{t+1}
\end{aligned}
$$

where $\overline{\mathbb{B}}(\Sigma_t) \stackrel{\text{def}}{=} \mathcal{K}(\Sigma_t)\overline{\mathbb{F}}(\Sigma_t)$.

While we may think of $\{(\overline{X}_t, Z_t, \Sigma_t) : t \geq 0\}$ as a Markov process, it will typically not be stationary. Sometimes a stationary counterpart does exist, however. A necessary requirement is that Σ_t be invariant. The recursion (6.3) involves no data, so Σ_t is deterministic and would otherwise vary with t. To construct such a process, find a positive semidefinite fixed point to the recursion (6.3) and initialize $\Sigma_0 = \overline{\Sigma}$. Then $\Sigma_t = \overline{\Sigma}$ for all $t \geq 0$, and

$$
\begin{aligned}
\mathcal{K}(\Sigma_t) &= \mathcal{K}(\overline{\Sigma}) \\
\Omega_t &= \mathbb{D}\overline{\Sigma}\mathbb{D}' + \mathbb{F}\mathbb{F}' \stackrel{\text{def}}{=} \overline{\Omega}
\end{aligned}
$$

for all $t \geq 1$. This simplifies recursive representation (6.6) by making $\overline{\mathbb{B}}$, $\overline{\mathbb{F}}$ and Ω all time-invariant. One way to obtain such a construction of $\overline{\Sigma}$ is to iterate on equation (6.3) to convergence, assuming that such a limit does exist. With this construction, we have a Markov process representation

$$
\begin{aligned}
\overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \overline{\mathbb{B}}\,\overline{W}_{t+1} \\
Z_{t+1} &= \mathbb{N} + \mathbb{D}\overline{X}_t + \overline{\mathbb{F}}\,\overline{W}_{t+1}
\end{aligned}
\tag{6.7}
$$

Compare this to the original state space system (6.1). Key differences are

1. In the original system (6.1), the shock vector W_{t+1} can be of much larger dimension than the time $t+1$ observation vector Z_{t+1}, while in (6.7), the dimension of $\overline{W}_{t+1}$ equals that of the observation vector.
2. The state vector X_t in the original system (6.1) is not observed while in the representation (6.7), the state vector $\overline{X}_t$ is observed.

Remark 6.3

For some applications, the stationary specification may not be the most interesting one. Our decomposition in the previous chapter did not require that we initialize the (now hidden state) Markov process X in its stationary distribution. It was a forward-looking construction conditioning on the current Markov state. We now describe how we might use that apparatus when the Markov state is hidden drawing on the setup in Remark 6.1. To proceed, we partition the constructed hidden state solution (6.6) as

$$
\begin{aligned}
\overline{X}^1_{t+1} &= \mathbb{A}_{11}\overline{X}^1_t + \overline{\mathbb{B}}_1(\Sigma_t)\overline{W}_{t+1} \\
\overline{X}^2_{t+1} &= \overline{X}^2_t + \overline{\mathbb{B}}_2(\Sigma_t)\overline{W}_{t+1} \\
Y_{t+1} - Y_t &= \mathbb{D}_{11}\overline{X}^1_t + \mathbb{D}_{12}\overline{X}^2_t + \overline{\mathbb{F}}_1(\Sigma_t)\overline{W}_{t+1}.
\end{aligned}
$$

Suppose initially that $\mathbb{D}_{12}$ is zero. Since the matrix $\mathbb{A}_{11}$ is a stable matrix, we can imitate the first-order VAR calculation from the previous chapter and construct a martingale component with date $t+1$ increment:

$$
G_{t+1} = \left[\overline{\mathbb{F}}_1(\Sigma_t) + \mathbb{D}_{11}(\mathbb{I} - \mathbb{A}_{11})^{-1}\overline{\mathbb{B}}_1(\Sigma_t)\right]\overline{W}_{t+1}. \tag{6.8}
$$

Next we include the estimation of the second state. By design, this uncertain state is used to introduce an unknown trend coefficient into the analysis. This leads us to modify the computation in a more substantive way. We compute the impact of $\overline{W}_{t+1}$ on $Y_{t+\tau}$ as:

$$
\begin{aligned}
&\left[\overline{\mathbb{F}}_1(\Sigma_t) + \mathbb{D}_{11}\left[\mathbb{I} + \mathbb{A}_{11} + \ldots + (\mathbb{A}_{11})^{\tau-1}\right]\overline{\mathbb{B}}_1(\Sigma_t)\right]\overline{W}_{t+1} \\
&\quad + \tau\mathbb{D}_{12}\overline{\mathbb{B}}_2(\Sigma_t)\overline{W}_{t+1}.
\end{aligned} \tag{6.9}
$$

Notice that the term on the first line of (6.9) converges to the right side of (6.8) as τ gets arbitrarily large. The term on the second line of (6.9), however, diverges for large τ when the trend is to be estimated. This occurs because the second component of the state X_t is invariant over time. A stationary initialization of the constructed state, $\overline{X}_0$ would necessarily reflect full knowledge of the trend coefficient with a zero in the lower-right entry of Σ_0.

Example 6.1

Consider an initial first-order moving-average process $\{Z_t\}$ that obeys:

$$
Z_t = W_t + \lambda W_{t-1},
$$

where $\{W_t\}$ is a univariate i.i.d. process of standardized normal random variables. We represent this process by letting $X_t = W_t$ and write

$$
\begin{aligned}
X_{t+1} &= W_{t+1} \\
Z_{t+1} &= \lambda X_t + W_{t+1} = \lambda\overline{X}_t + \lambda\left(X_t - \overline{X}_t\right) + W_{t+1}
\end{aligned}
$$

Then Z_{t+1} has mean $\lambda\overline{X}_t$ and variance $\Omega_t = \lambda^2\Sigma_t + 1$ conditioned on $\overline{X}_t$. The covariance of Z_{t+1} with X_{t+1} conditioned on $\overline{X}_t$ is one. Thus the Kalman gain is

$$\mathcal{K}(\Sigma_t) = \frac{1}{\lambda^2 \Sigma_t + 1},$$

and

$$\Sigma_{t+1} = 1 - \frac{1}{\lambda^2 \Sigma_t + 1}. \tag{6.10}$$

An invariant solution satisfies:

$$\Sigma = 1 - \frac{1}{\lambda^2 \Sigma + 1},$$

or equivalently the quadratic equation:

$$\lambda^2 \Sigma^2 + (1 - \lambda^2)\Sigma = 0.$$

Thus there are two possible solutions:

$$\Sigma = 0 \text{ or } \Sigma = \frac{\lambda^2 - 1}{\lambda^2}.$$

The second solution is only of interest when $|\lambda| > 1$.

For each of these solutions, form

$$\begin{aligned}
\Omega &= \lambda^2 \Sigma + 1 & &= 1 \text{ or } \lambda^2 \\
\mathcal{K}(\Sigma) &= \frac{1}{\lambda^2 \Sigma + 1} & &= 1 \text{ or } \frac{1}{\lambda^2} \\
\overline{F} &= \sqrt{\lambda^2 \Sigma + 1} & &= 1 \text{ or } |\lambda| \\
\overline{B} &= \frac{1}{\sqrt{\lambda^2 \Sigma + 1}} & &= 1 \text{ or } \frac{1}{|\lambda|}.
\end{aligned}$$

where the time-invariant recursive solution satisfies:

$$\begin{aligned}
\overline{X}_{t+1} &= \overline{B}\,\overline{W}_{t+1} \\
Z_{t+1} &= \lambda \overline{X}_t + \overline{F}\,\overline{W}_{t+1} = \lambda \overline{B}\,\overline{W}_t + \overline{F}\,\overline{W}_{t+1}.
\end{aligned} \tag{6.11}$$

Notice that the solution for the observation process is itself a moving-average representation expressed in terms of the shock process $\overline{W}$.

Return now to the up-dating equation (6.10). One can verify that $\Sigma = 0$ is the locally stable solution when $|\lambda| < 1$ and $\Sigma = (\lambda^2 - 1)/\lambda^2$ is the stable solution when $|\lambda| > 1$. Under the $\Sigma = 0$ solution, $\overline{X}_t = X_t$ and the recursive solution (6.11) just recovers the original moving-average representation. When $|\lambda| > 1$, this is no longer the case; the two moving-average representations differ. To understand the finding when $|\lambda| < 1$, solve the equation

$$W_t + \lambda W_{t-1} = Z_t$$

backwards to obtain:

$$W_t = \sum_{j=0}^{\infty} (-\lambda)^j Z_{t-j},$$

showing that the current shock W_t can be recovered by the current and infinite past of the Z's. The infinite sum is well defined since $|\lambda| < 1$. This same strategy fails when $|\lambda| > 1$ and the Kalman filtering reveals an alternative moving-average representation in which the prediction $\overline{X}_t$ differs from the hidden state X_t. Other informationally equivalent representations can be obtained by replacing the $\overline{W}$ process with its negative counterpart.

6.2.1. Kalman smoother

The Kalman filter provides recursive formulas for computing the distribution of a hidden state vector X_t conditional on a signal history $\{Z_\tau : \tau = 1, 2, \ldots, t\}$ and an initial distribution Q_0 for X_0. This conditional distribution has the form $X_t \sim \mathrm{Normal}(\overline{X}_t, \Sigma_t)$; the Kalman filtering equations provide recursive formulas for the conditional mean $\overline{X}_t$ and the conditional covariance matrix Σ_t.

Knowing outcomes $\{\overline{X}_\tau, \Sigma_\tau\}_{\tau=1}^{T}$ from the Kalman filter provides the foundation for the **Kalman smoother.** The Kalman smoother uses past, present, and **future** values of Z_τ to learn about **current** values of the state X_τ. The Kalman smoother is a recursive algorithm that computes sufficient statistics for the distribution of X_t conditional on the **entire sample** $\{Z_t\}_{t=1}^{T}$, namely, a mean vector and covariance matrix pair $\widehat{X}_t, \widehat{\Sigma}_t$. The Kalman smoother takes outputs $\{\overline{X}_t, \Sigma_t\}_{t=0}^{T}$ from the Kalman filter as inputs and then works **backwards** through the following steps starting from $t = T$.

- Reversed time regression. Write the joint distribution of (X_t, X_{t+1}, Z_{t+1}) conditioned on $\left(\overline{X}_t, \Sigma_t\right)$ as

$$
\begin{bmatrix} X_t \\ X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} \overline{X}_t \\ \mathbb{A}\overline{X}_t \\ \mathbb{N} + \mathbb{D}\overline{X}_t \end{bmatrix}, \begin{bmatrix} \Sigma_t & \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \\ \mathbb{A}\Sigma_t & \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t & \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix} \right)
$$

From this joint distribution, construct the conditional distribution for X_t, given X_{t+1}, Z_{t+1} and $\left(\overline{X}_t, \Sigma_t\right)$. Compute the conditional mean of $X_t - \overline{X}_t$ by using the population least squares formula

$$
\widehat{\mathbb{K}}_1 \left(X_{t+1} - \mathbb{A}\overline{X}_t \right) + \widehat{\mathbb{K}}_2 \left(Z_{t+1} - \mathbb{N} - \mathbb{D}\overline{X}_t \right) \tag{6.12}
$$

where the regression coefficient matrix is

$$
\begin{bmatrix} \widehat{\mathbb{K}}_1 & \widehat{\mathbb{K}}_2 \end{bmatrix} = \widehat{\mathbb{K}} \stackrel{\text{def}}{=} \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1}
$$

and the residual covariance matrix equals

$$
\Sigma_t - \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{A}\Sigma_t \\ \mathbb{D}\Sigma_t \end{bmatrix} \tag{6.13}
$$

- Iterated expectations. Notice that the above reverse regression includes $X_{t+1} - \mathbb{A}\overline{X}_t$ among the regressors. Because X_{t+1} is hidden, that is more information than we have. We can condition down to information that we actually have by instead using $\widehat{X}_{t+1} - \mathbb{A}\overline{X}_t$ as the regressor where $\widehat{X}_{t+1}$ is the conditional expectation of X_{t+1} given the full sample of data $\{Z_t\}_{t=1}^T$ and $\widehat{\Sigma}_{t+1}$ is the corresponding conditional covariance matrix. This gives us a backwards recursion for $\widehat{X}_t$:

$$
\widehat{X}_t - \overline{X}_t = \widehat{\mathbb{K}}_1 \left(\widehat{X}_{t+1} - \mathbb{A}\overline{X}_t \right) + \widehat{\mathbb{K}}_2 \left(Z_{t+1} - \mathbb{N} - \mathbb{D}\overline{X}_t \right)
$$

The law of iterated expectations implies that the regression coefficient matrices $\widehat{\mathbb{K}}_1, \widehat{\mathbb{K}}_2$ equal the ones we have already computed. But since we are using less information, the conditional covariance matrix increases by $\widehat{\mathbb{K}}_1 \widehat{\Sigma}_{t+1} \widehat{\mathbb{K}}_1'$. This implies the backwards recursion:

$$
\widehat{\Sigma}_t = \Sigma_t - \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{A}\Sigma_t \\ \mathbb{D}\Sigma_t \end{bmatrix} + \widehat{\mathbb{K}}_1 \widehat{\Sigma}_{t+1} \widehat{\mathbb{K}}_1'
$$

- Take $\widehat{\Sigma}_T = \Sigma_T$ and $\widehat{X}_T = \overline{X}_T$ as terminal conditions.

6.3. Mixtures

Suppose now that $\{X_t : t \geq 0\}$ evolves as an n-state Markov process with transition probability matrix $\mathbb{P}$.

A date $t+1$ vector of signals Z_{t+1} has density $\psi_i(z^*)$ if hidden state i is realized, meaning that X_t is the ith coordinate vector. We want to compute the probability that X_t is in state i conditional on the signal history. The vector of conditional probabilities equals $Q_t = E[X_t | Z^t, Q_0]$, where Q_0 is a vector of initial probabilities and Z^t is the available signal history up to date t. We construct $\{Q_t : t \geq 1\}$ recursively:

1.

Find the joint distribution of (X_{t+1}, Z_{t+1}) conditional on X_t. Conditional distributions of Z_{t+1} and X_{t+1} are statistically independent by assumption. Write the joint density conditioned on X_t as:

$$\begin{array}{ccc} (\mathbb{P}' X_t) & \times & (X_t)' \text{vec} \{\psi_i(z^*)\} \\ \uparrow & & \uparrow \\ X_{t+1} \ \text{density} & & Z_{t+1} \ \text{density} \end{array} \tag{6.14}$$

where $\text{vec}(r_i)$ is a column vector with r_i in the ith component. We have expressed conditional independence by forming a joint conditional distribution as a product of two conditional densities, one for X_{t+1} and one for Z_{t+1}.

2. Find the joint distribution of X_{t+1}, Z_{t+1} conditioned on Q_t. Since X_t is not observed, we form the appropriate average of (6.14) conditioned on Z^t, Q_0:

$$\mathbb{P}' \text{diag}\{Q_t\} \text{vec} \{\psi_i(z^*)\}, \tag{6.15}$$

   where $\text{diag}(Q_t)$ is a diagonal matrix with the entries of Q_t on the diagonal. Thus, Q_t encodes all pertinent information about X_t that is contained in the history of signals. Conditional on Q_t, X_{t+1} and Z_{t+1} are *not* statistically independent.

3. Find the distribution of Z_{t+1} conditional on Q_t. Summing (6.15) over the hidden states gives

$$(\mathbf{1}_n)' \mathbb{P}' \text{diag}\{Q_t\} \text{vec} \{\psi_i(z^*)\} = Q_t \cdot \text{vec} \{\psi_i(z^*)\}.$$

   Thus, Q_t is a vector of weights used to form a mixture distribution. Suppose, for instance, that ψ_i is a normal distribution with mean μ_i and covariance matrix Σ_i. Then the distribution of Z_{t+1} conditioned on Q_t is a *mixture of normals* with mixing probabilities given by entries of Q_t.

4. Obtain Q_{t+1} by dividing the *joint* density of (Z_{t+1}, X_{t+1}) conditional on Q_t by the *marginal* density for Z_{t+1} conditioned on Q_t and then evaluating this ratio at Z_{t+1}. In this way, we construct the density for X_{t+1} conditioned on (Q_t, Z_{t+1}). It takes the form of a vector Q_{t+1} of conditional probabilities. Thus, we are led to

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\left\{\psi_i(Z_{t+1})\right\}} \right) \mathbb{P}' \text{diag}(Q_t) \text{vec}\left\{\psi_i(Z_{t+1})\right\} \tag{6.16}$$

Together, step 3 and step 4 define a Markov process for Q_{t+1}. As indicated in step 3, Z_{t+1} is drawn from a (history-dependent) mixture of densities ψ_i. As indicated in step 4, the vector Q_{t+1} equals the exact function of Z_{t+1} and Q_t described in (6.16).

6.4. Recursive Regression

A statistician wants to infer unknown parameters of a linear regression model. By treating regression coefficients as hidden states that are constant over time, we can cast this problem in terms of a hidden Markov model. By assigning a prior probability distribution to statistical models that are indexed by parameter values, the statistician can construct a stationary stochastic process as a mixture of statistical models.[5] From increments to a data history, the statistician learns about parameters sequentially. By assuming that the statistician adopts a conjugate prior à la [Raiffa and Schlaifer, 1961], we can construct explicit updating formulas.

Consider the first-order vector autoregressive model

$$\begin{aligned} X_{t+1} &= \mathbb{A} X_t + \mathbb{B} W_{t+1} \\ Z_{t+1} &= \mathbb{N} + \mathbb{D} X_t + \mathbb{F} W_{t+1} \end{aligned} \tag{6.17}$$

where W_{t+1} is an i.i.d. normal random vector with mean vector 0 and covariance matrix I, X_t is an observable state vector, and $\mathbb{A}, \mathbb{B}, \mathbb{D}, \mathbb{F}, \mathbb{N}$ are matrices containing unknown coefficients. When $\mathbb{A}$ is a stable matrix, the vector $\mathbb{N}$ is interpretable as the vector of means of the observation vector Z_{t+1} (conditioned on invariant events).

We now construct a special case of this representation. Suppose that Z_{t+1} and W_{t+1} share the same dimensions, that $\mathbb{F}$ is nonsingular, and that X_t consists of:

$$X_t = \begin{bmatrix} Z_t - \mathbb{N} \\ Z_{t-1} - \mathbb{N} \\ \cdots \\ Z_{t-\ell+1} - \mathbb{N} \end{bmatrix}$$

We obtain a finite-order vector autoregression by writing:

(6.18)

$$
\begin{aligned}
Z_{t+1} &= \mathbb{N} + \mathbb{D}X_t + \mathbb{F}W_{t+1} \\
&= \widetilde{\mathbb{N}} + \mathbb{D}\begin{bmatrix} Z_t \\ Z_{t-1} \\ \cdots \\ Z_{t-\ell+1} \end{bmatrix} + \mathbb{F}W_{t+1}
\end{aligned}
$$

where

$$
\widetilde{\mathbb{N}} \stackrel{\text{def}}{=} \mathbb{N} - \mathbb{D}\begin{bmatrix} \mathbb{N} \\ \mathbb{N} \\ \cdots \\ \mathbb{N} \end{bmatrix}
$$

Our plan is to estimate the coefficients of the matrices $\mathbb{N}, \mathbb{D}$, and $\mathbb{F}$. Notice that $\mathbb{N}$ potentially can be recovered from $\widetilde{\mathbb{N}}$ and $\mathbb{D}$. The matrix $\mathbb{F}$ is not fully identified without further a priori restrictions. What is identified is $\mathbb{F}\mathbb{F}'$. This identification challenge is the topic of so-called "structural vector autoregressions." This literature seeks to motivate sufficient restrictions imposed on the matrix $\mathbb{F}$ to achieve a unique identification. Our previous construction of a permanent shock is one way to contribute to this identification. As with most of the VAR literature, we do not consider the case of over-identification in what follows. With exact identification, we impose a convenient normalization on $\mathbb{F}$. Other observationally equivalent and substantively motivated $\mathbb{F}$'s can be constructed from our estimation.

6.4.1. Conjugate prior updating

In practice, priors are often selected as a matter of convenience. They are constructed with tractability in mind. Conjugate priors are an example of this convenience. Chapter 10 uses this as one motivation for exploring **prior sensitivity.** In this section, we will proceed recursively. In so doing, we will use the insight that "today's posterior is tomorrow's prior." By following suggestions offered by [Zellner, 1962], [Box and Tiao, 1992], [Sims and Zha, 1999], and especially [Zha, 1999], we transform system (6.18) in a way that justifies estimating the unknown coefficients by applying least squares equation-by-equation.

Start by factoring the matrix

$$
\mathbb{F}\mathbb{F}' = \mathbb{J}\Delta\mathbb{J}' \tag{6.19}
$$

where $\mathbb{J}$ is lower triangular with ones on the diagonal and Δ is diagonal.[6] The inverse $\mathbb{J}^{-1}$ is also lower triangular with ones on the diagonal. We make $\mathbb{J}$ lower triangular as a matter of convenience and not necessarily one based on economic interpretation. The literature on

structural VARs explores the identification of the matrix $\mathbb{F}$ from knowledge of the covariance matrix of U_{t+1} given by

$$\mathbb{F}\mathbb{F}'$$

using a variety of different approaches. Constructing a matrix $\mathbb{F}$ amounts to identifying a vector of substantively interesting shocks. The construction of a common permanent shock for a vector time series, described in Section Permanent shocks of Chapter 3, gives a way to identify one of the shocks of interest, but there are many other choices in the literature including different configurations of zeroes in the matrix $\mathbb{F}$. Some researchers stop with only partial identification of the matrix $\mathbb{F}$. For instance, see [Uhlig, 2005] for a comprehensive discussion of how to use sign restrictions to achieve partial identification. Our strategy going forward is to impose a particularly convenient initial identification. Posterior probabilities for the potentially more interesting forms of identification can then be inferred from our computations.

Construct

$$\mathbb{J}^{-1}Z_{t+1} = \mathbb{J}^{-1}\widetilde{\mathbb{N}} + \mathbb{J}^{-1}\mathbb{D}\begin{bmatrix} Z_t \\ Z_{t-1} \\ \cdots \\ Z_{t-\ell+1} \end{bmatrix} + U_{t+1} \tag{6.20}$$

where

$$U_{t+1} = \mathbb{J}^{-1}\mathbb{F}W_{t+1}$$

so that $EU_{t+1}U_{t+1}' = \Delta$. The i^{th} entry of U_{t+1} is uncorrelated with, and consequently statistically independent of, the jth components of Z_{t+1} for $j = 1, 2, \ldots, i-1$. As a consequence, each equation in system (6.20) can be interpreted as a regression equation in which the left-hand side variable in equation i is the i^{th} component of Z_{t+1}. The regressors are a constant, $Z_t, Z_{t-1} \ldots, Z_{t-\ell+1}$, and the j^{th} components of Z_{t+1} for $j = 1, \ldots, i-1$. The ith equation is an unrestricted regression with a disturbance term $U_{t+1,i}$ that is uncorrelated with disturbances $U_{t+1,j}$ to all other equations $j \neq i$. The system of equations (6.20) is thus recursive. The first equation determines the first entry of Z_{t+1}, the second equation determines the second entry of Z_{t+1} given the first entry, and so forth.

We construct estimates of the coefficient matrices $\mathbb{A}, \mathbb{B}, \mathbb{D}, \mathbb{F}, \mathbb{N}$ and the covariance matrix $\Delta = EU_{t+1}U_{t+1}'$ from these regression equations, with the reminder that knowledge of $\mathbb{J}$ and Δ may not directly reveal the matrix $\mathbb{F}$ that is of substantive interest.

Consider, in particular, the ith regression formed in this way and express it as the scalar regression model:

$$Z_{t+1}^{[i]} = R_{t+1}^{[i]}{}' \beta^{[i]} + U_{t+1}^{[i]}$$

where $R_{t+1}^{[i]}$ is the appropriate vector of regressors in the ith equation of system (6.20). To simplify notation, we omit superscripts and understand that we are estimating one equation at a time. To avoid notational confusion, we let the left side variable of the regression be denoted Y_{t+1}. The disturbance U_{t+1} is a normally distributed random variable with mean zero and variance σ^2. Furthermore, U_{t+1} is statistically independent of R_{t+1}. Information observed as of date t consists of X_0 and the current and past history of the right-hand side variables of the regression. Suppose that Z_{t+1} and R_{t+1} are also observed at date $t+1$, but that β and σ^2 remain unknown.

Let the distribution of β conditioned on the history up to time t and on σ^2 be normal with mean b_t and **precision matrix,** $\zeta\Lambda_t$ where $\zeta = \frac{1}{\sigma^2}$. Here the precision matrix equals the inverse of a conditional covariance matrix of the unknown parameters. At date $t+1$, we use the additional information to obtain an updated conditional (inclusive of ζ) distribution of β which we take to be normal with mean b_{t+1} and precision $\zeta\Lambda_{t+1}$. To deduce the recursive updating conditioned on ζ, we observe that both the date t posterior distribution and the date $t+1$ conditional density for Y_{t+1} are normal. After multiplication, the terms inside the exponential involving β are

$$\begin{aligned} &\exp\left(\zeta Y_{t+1} R_{t+1}{}'\beta - \frac{\zeta}{2}\beta' R_{t+1} R_{t+1}{}'\beta + \zeta b_t{}'\Lambda_t\beta - \frac{\zeta}{2}\beta'\Lambda_t\beta\right) \\ &\propto \exp\left[-\frac{\zeta}{2}(\beta - b_{t+1})'\Lambda_{t+1}(\beta - b_{t+1})\right] \end{aligned}$$

where

$$\Lambda_{t+1} = R_{t+1} R_{t+1}{}' + \Lambda_t \tag{6.21}$$

and

$$\Lambda_{t+1} b_{t+1} = [\Lambda_t b_t + R_{t+1} Y_{t+1}]. \tag{6.22}$$

Recursion (6.21) implies that $\Lambda_{t+1} - \Lambda_t$ is a positive semidefinite matrix, which confirms that additional information improves estimation accuracy. Evidently from recursion (6.21), Λ_{t+1} cumulates cross-products of the regressors and adds them to an initial Λ_0. The updated conditional mean b_{t+1} for the normal distribution of unknown coefficients can be deduced from Λ_{t+1} via the updating equation (6.22). Solving difference equation (6.22) backwards shows how $\Lambda_{t+1} b_{t+1}$ cumulates cross-products of R_{t+1} and Y_{t+1} and adds the outcome to an initial condition $\Lambda_0 b_0$.

So far we pretended that we know σ^2 by conditioning on σ^2, which is equivalent to conditioning on its inverse ζ. Assume now that we don't know σ^2 but instead summarize our uncertainty about it with a date t gamma density for ζ conditioned on Z^t, X_0 so that it is proportional to

$$(\zeta)^{\frac{c_t}{2}} \exp(-d_t\zeta/2),$$

where the density is expressed as a function of ζ, so that $d_t\zeta$ has a chi-square density with $c_t + 2$ degrees of freedom. The implied density for ζ conditioned on time $t+1$ information is also a gamma density with updated parameters:

$$\begin{aligned} c_{t+1} &= c_t + 1 \\ d_{t+1} &= (Y_{t+1})^2 - (b_{t+1})'\Lambda_{t+1}b_{t+1} + (b_t)'\Lambda_t b_t + d_t. \end{aligned}$$

The distribution of β conditioned on Z^{t+1}, X_0, and ζ is normal with mean b_{t+1} and precision matrix $\zeta\Lambda_{t+1}$. The distribution of ζ conditioned on Z^{t+1}, X_0 has a gamma density, so that it is proportional to[7]

$$(\zeta)^{\frac{c_{t+1}}{2}} \exp(-d_{t+1}\zeta/2),$$

Standard least squares regression statistics can be rationalized by positing a prior that is not informative. This is commonly done by using an “improper” prior that does not integrate to unity. [8] Setting $\Lambda_0 = 0$ effectively imposes a uniform but improper prior over β. Although Λ_t’s early in the sequence are singular, we can still update $\Lambda_{t+1}b_{t+1}$ via (6.22); b_{t+1} is not uniquely determined until Λ_{t+1} becomes nonsingular. After enough observations have been accumulated to make Λ_{t+1} become nonsingular, the implied normal distributions for the unknown parameters become proper. When $\Lambda_0 = 0$, the specification of b_0 is inconsequential and b_{t+1} becomes a standard least squares estimator. An “improper gamma” prior over σ that is often associated with an improper normal prior over β sets c_0 to minus two and d_0 to zero. This is accomplished by assuming a uniform prior distribution for the logarithm of the precision ζ or for the logarithm of σ^2. With this combination of priors, d_{t+1} becomes a sum of squared regression residuals. [9]

From the posterior of the coefficients of this transformed system we can compute posteriors of nonlinear functions of those coefficients. We accomplish this by using a random number generator repeatedly to take pseudo-random draws from the posterior probability of the coefficients, forming those nonlinear functions, and then using the resulting histograms of those nonlinear functions to approximate the posterior probability distribution of those nonlinear functions. For example, many applied macroeconomic papers report impulse responses as a way to summarize model features. Impulse responses are nonlinear functions of $(\mathbb{A}, \mathbb{B})$.

The conjugate prior approach described above does not generate a posterior for which either the prior or the implied posteriors for the matrix $\mathbb{A}$ have stable eigenvalues with probability one. We therefore modify that approach to impose that $\mathbb{A}$ is a stable matrix. We do this by rescaling the posterior probability so that it integrates to one over the region of the parameter space for which

$\mathbb{A}$ is stable. We in effect condition on $\mathbb{A}$ being stable. This is easy to implement by rejection sampling.[10]

The standard deviation of the martingale increment is a nonlinear function of parameters in $(\mathbb{A}, \mathbb{B})$. We construct a posterior distribution via Monte Carlo simulation. We draw from the posterior of the multivariate regression system and, after conditioning on stability of the $\mathbb{A}$ matrix, compute the nonlinear functions of interest. From the simulation, we construct joint histograms to approximate posterior distributions of functions of interest.[11]

As an illustration, Fig. 6.2 and Fig. 6.3 show posterior histograms for the standard deviation of shocks to short-term consumption growth and for the standard deviation of the martingale increment to consumption, based on the application described in Section An economic rationale of Chapter 4. The standard deviation of the short-term shock contribution is about one-half that of the martingale increment. Fig. 6.2 tells us that short-term risk can be inferred with much more accuracy than long-term risk can be. This evidence says that while there **could** be a long-run risk component to consumption, it is poorly measured. The fat tail in the right side of the distribution of the long-run standard deviation is induced by Monte Carlo draws for which some eigenvalues of $\mathbb{A}$ have absolute values very close to unity. [12]

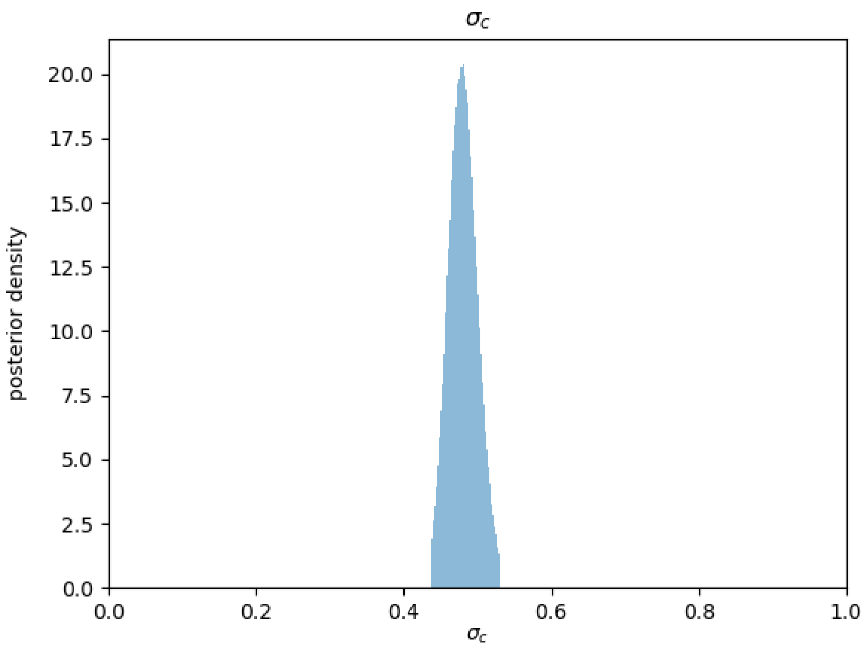

Fig. 6.2 Posterior density for conditional standard deviation of consumption growth.

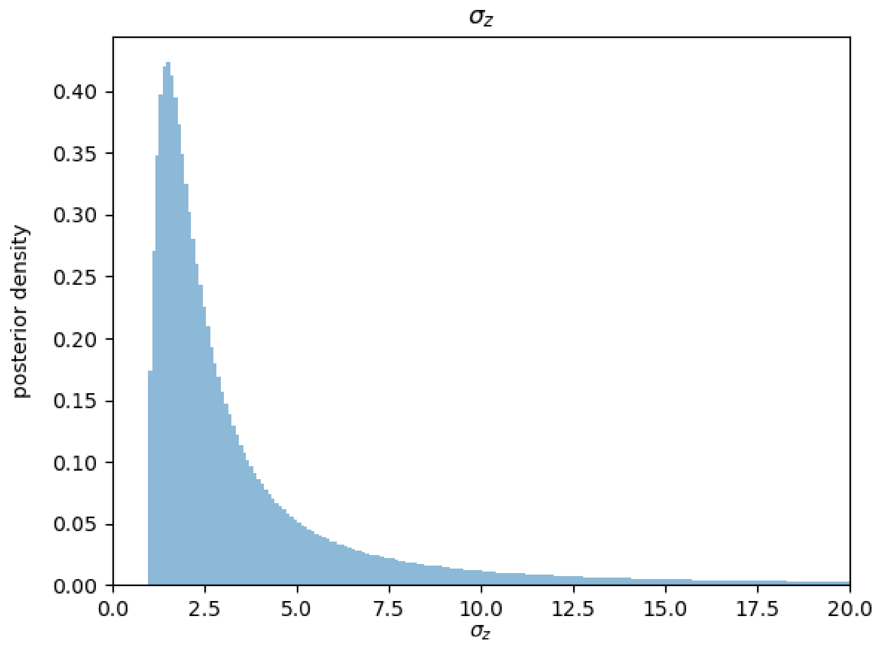

Fig. 6.3 Posterior distribution for the standard deviation of the martingale increment.

Remark 6.4

Following [Carter and Kohn, 1994], we consider a Gibbs sampling approach for making inferences about a linear state-space model with hidden states and unknown parameters. Let θ denote a stand-in for the unknown parameters of the state-space model, and let ϑ denote a stand-in for the entire collection of unknown states. All of the computations that follow condition on a time series of observations, $Z_1, Z_2, \ldots, Z_T$, although we suppress this conditioning in the notation that follows. The Kalman smoother gives a conditional distribution, $P(d\vartheta \mid \theta)$. Given the observations of the composite state vector, ϑ, we extend the conjugate prior method to construct $\widetilde{P}(d\theta \mid \vartheta)$. Finally, let $Q(d\vartheta, d\theta)$ denote the joint probability for (ϑ, θ), which is the target of computation.

To compute the joint distribution, form a Markov process with a one-period transition from (ϑ, θ) to (ϑ^+, θ^+) by following the two-step construction for use in simulation:

- generate ϑ^+ using $P(d\vartheta^+ \mid \theta)$;
- generate θ^+ using $\widetilde{P}(d\theta^+ \mid \vartheta^+)$.

The transition distribution for the Markov process is:

$$\widetilde{P}(d\theta^+ \mid \vartheta^+) P(d\vartheta^+ \mid \theta).$$

We verify that Q is a stationary distribution by first computing

$$\int_{\vartheta,\theta} \widetilde{P}(d\theta^+ \mid \vartheta^+) P(d\vartheta^+ \mid \theta) Q(d\vartheta, d\theta)$$
$$= \widetilde{P}(d\theta^+ \mid \vartheta^+) \int_\theta P(d\vartheta^+ \mid \theta) \int_\vartheta Q(d\vartheta, d\theta).$$

Note that $\int_\vartheta Q(d\vartheta, d\theta)$ is the marginal distribution over θ and thus

$$P(d\vartheta^+ \mid \theta) \int_\vartheta Q(d\vartheta, d\theta) = Q(d\vartheta^+, d\theta).$$

It follows that

$$\int_\theta P(d\vartheta^+ \mid \theta) \int_\vartheta Q(d\vartheta, d\theta) = \int_\theta Q(d\vartheta^+, d\theta)$$

The integral on the right side is the marginal distribution over ϑ^+. Finally, note that

$$\widetilde{P}(d\theta^+ \mid \vartheta^+) \int_\theta Q(d\vartheta^+, d\theta) = Q(d\vartheta^+, d\theta^+),$$

showing that Q is indeed the stationary distribution of this constructed Markov process. This allows us to approximate Q by simulating the Markov process. This approach is an example of a numerical method called Gibbs sampling.

6.5. VAR Regimes

Following [Sclove, 1983] and [Hamilton, 1989], suppose that there are multiple VAR regimes $(\mathbb{A}_i, \mathbb{B}_i, \mathbb{D}_i, \mathbb{F}_i)$ for $i = 1, 2, \ldots, n$, where indices i are governed by a Markov process with transition matrix $\mathbb{P}$. In regime i we have

$$\begin{aligned} X_{t+1} &= \mathbb{A}_i X_t + \mathbb{B}_i W_{t+1} \\ Y_{t+1} - Y_t &= \mathbb{D}_i X_t + \mathbb{F}_i W_{t+1}, \end{aligned} \tag{6.23}$$

where $\{W_{t+1}\}_{t=0}^\infty$ is an i.i.d. sequence of $\text{Normal}(0, I)$ random vectors conditioned on X_0, and $\mathbb{F}_i$ is nonsingular.

We can think of X_t and a regime indicator Z_t jointly as forming a Markov process. In this section Z_t is the hidden regime indicator and Y is observed, reversing the roles that Z and Y played in Section Kalman Filter and Smoother. When regime i is realized, Z_t equals a coordinate vector with one in the i^{th} coordinate and zeros at other coordinates. We study a situation in which regime indicator Z_t is not observed. Let Q_t denote an n-dimensional vector of probabilities over the hidden states Z_t conditioned on Y^t, X_0, and Q_0, where Q_0 is the date zero vector of initial probabilities for Z_0. Equivalently, Q_t is $E(Z_t|Y^t, X_0, Q_0)$.

The vector of conditional probabilities Q_t solves a *filtering problem*. We describe the solution of this problem by representing (X_t, Q_t) as a Markov process via the following four steps.

1. Find the joint distribution for $(Z_{t+1}, Y_{t+1} - Y_t)$ conditioned on (Z_t, X_t). Conditional distributions of Z_{t+1} and Y_{t+1} are statistically independent by assumption. Conditioned on Z_t, X_t conveys no information about Z_{t+1} and thus the conditional density of Z_{t+1} is given by entries of $\mathbb{P}'Z_t$. Conditioned on $Z_t = i$, $Y_{t+1} - Y_t$ is normal with mean $\mathbb{D}_i X_t$ and covariance matrix $\mathbb{F}_i(\mathbb{F}_i)'$. Let $\psi_i(y^*, X_t)$ be the normal density function for $Y_{t+1} - Y_t$ conditioned on X_t when Z_t is in regime i. We can write the joint density conditioned on (Z_t, X_t) as:

$$\underbrace{(\mathbb{P}'Z_t)}_{\substack{\uparrow \\ Z_{t+1}\ \text{density}}} \quad \times \quad \underbrace{(Z_t)'\text{vec}\left\{\psi_i(y^*, X_t)\right\}}_{\substack{\uparrow \\ Y_{t+1}-Y_t\ \text{density}}}$$

where $\text{vec}(r_i)$ is a column vector with r_i in the i^{th} entry. We have imposed conditional independence by forming a joint conditional distribution as a product of two conditional densities, one for Z_{t+1} and one for $Y_{t+1} - Y_t$.

2. Find the joint distribution of $(Z_{t+1}, Y_{t+1} - Y_t)$ conditioned on (X_t, Q_t). Since Z_t is not observed, we form the appropriate average of the above conditioned on the Y^t, X_0, Q_0:

$$\mathbb{P}'\text{diag}\{Q_t\}\text{vec}\left\{\psi_i(y^*, X_t)\right\}$$

where $\text{diag}\{Q_t\}$ is a diagonal matrix with components of Q_t on the diagonal. Thus, Q_t encodes all pertinent information about the time t regime Z_t that is contained in Y^t, X_0 and Q_0. Notice that conditional on (X_t, Q_t), random vectors $Y_{t+1} - Y_t$ and Z_{t+1} are *not* statistically independent.

3. Find the distribution of $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Summing the above over hidden states gives

$$(\mathbf{1}_n)'\mathbb{P}'\text{diag}\{Q_t\}\text{vec}\left\{\psi_i(y^*, X_t)\right\} = Q_t \cdot \text{vec}\left\{\psi_i(y^*, X_t)\right\}.$$

Thus, the distribution for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t) is a *mixture of normals* in which, with probability given by the i^{th} entry of Q_t, $Y_{t+1} - Y_t$ is normal with mean $\mathbb{D}_i X_t$ and covariance matrix $\mathbb{F}_i(\mathbb{F}_i)'$. Similarly, the conditional distribution of X_{t+1} is a mixture of normals.

4. Obtain Q_{t+1} by dividing the *joint* density for $(Y_{t+1} - Y_t, Z_{t+1})$ conditioned on (X_t, Q_t) by the *marginal* density for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Division gives the density for

Z_{t+1} conditioned $(Y_{t+1} - Y_t, X_t, Q_t)$, which in this case is just a vector Q_{t+1} of conditional probabilities. Thus, we are led to the recursion

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \mathrm{vec}\left\{\psi_i(Y_{t+1}, X_t)\right\}} \right) \mathbb{P}' \mathrm{diag}(Q_t) \mathrm{vec}\left\{\psi_i(Y_{t+1}, X_t)\right\}. \tag{6.24}$$

Taken together, steps (3) and (4) provide the one-step-transition equation for Markov state (X_{t+1}, Q_{t+1}). As indicated in step (3), Y_{t+1} is a mixture of normally distributed random variables. As argued in step (4) the vector Q_{t+1} is an exact function of Y_{t+1}, Q_t, and X_t that is given by the above formula (6.24).

Example 6.2

Let $X_t = [\log(B_t/C_t), \log(D_t/C_t)]$ and $Y_t = \log C_t$, where C_t, B_t and D_t are consumption, business income and dividends, respectively.[13][14] We estimate (6.23) using Gibbs sampling with two regimes. Fig. 6.4 plots the two ratios. Fig. 6.5 and Fig. 6.6 report the mean and volatility regimes, Fig. 6.7 the smoothed regime probabilities, and Fig. 6.8 the posterior distributions.

Fig. 6.4 Log business income less consumption $\log(B_t/C_t)$ (purple) and log dividends less consumption $\log(D_t/C_t)$ (green).

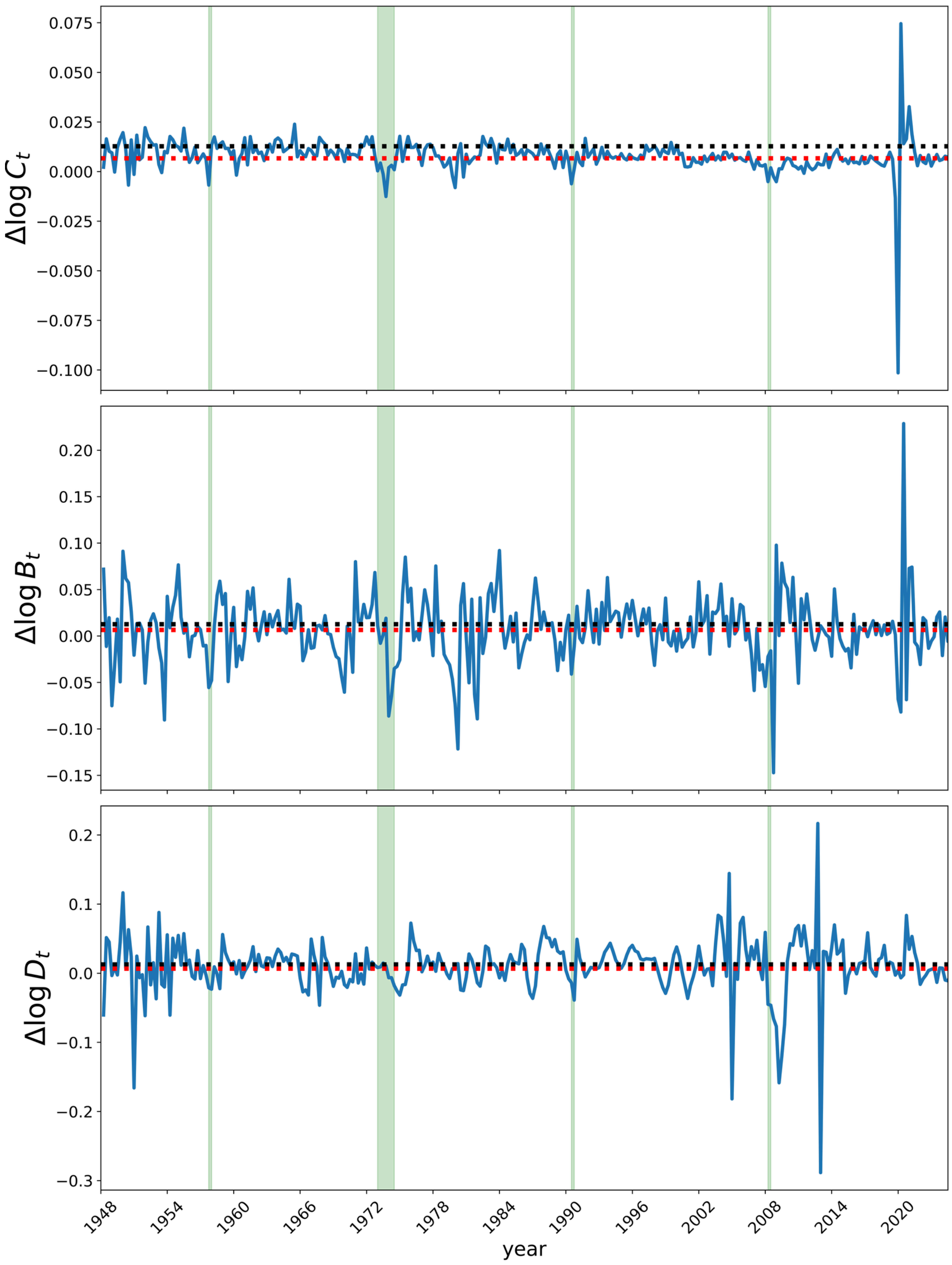

Fig. 6.5 The log growth rates of consumption, business income and dividends are shown as solid lines in each panel. The two regime means (1.276% and 0.667%) are shown as dotted lines in black (high) and red (low). The median smoothed probability of being in the low-mean regime is shown in shaded green. Averaging the two regime means with the stationary regime probabilities gives 1.2480%; the sample means are 0.8185% for consumption, 0.7336% for business income, and 1.0921% for dividends. The half-life for the high mean state is 46 quarters; for the low mean state it is 2 quarters. The correlation matrix for $\Delta \log C_t$, $\Delta \log B_t$ and $\Delta \log D_t$ is

$$\begin{pmatrix} 1.000 & 0.158 & -0.173 \\ & 1.000 & 0.013 \\ & & 1.000 \end{pmatrix}$$

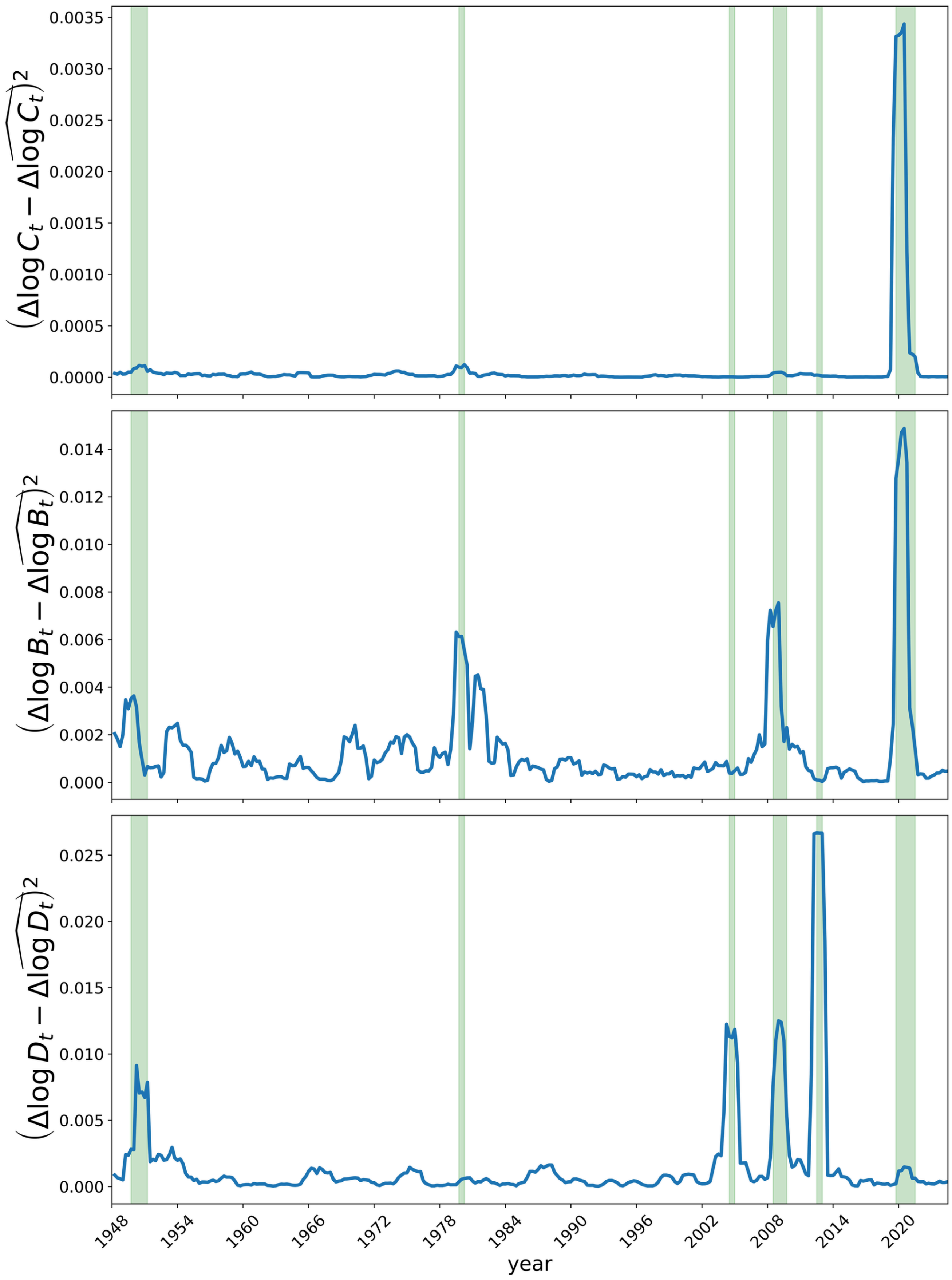

Fig. 6.6 The residual squares of log growth rates of consumption, business income and dividends are shown as solid lines in each panel. A two-quarter window moving average is applied. The median smoothed probability of being in the high-volatility regime is shown in shaded green. The half-life for the low volatility state is 26 quarters; for the high volatility state it is 2 quarters.

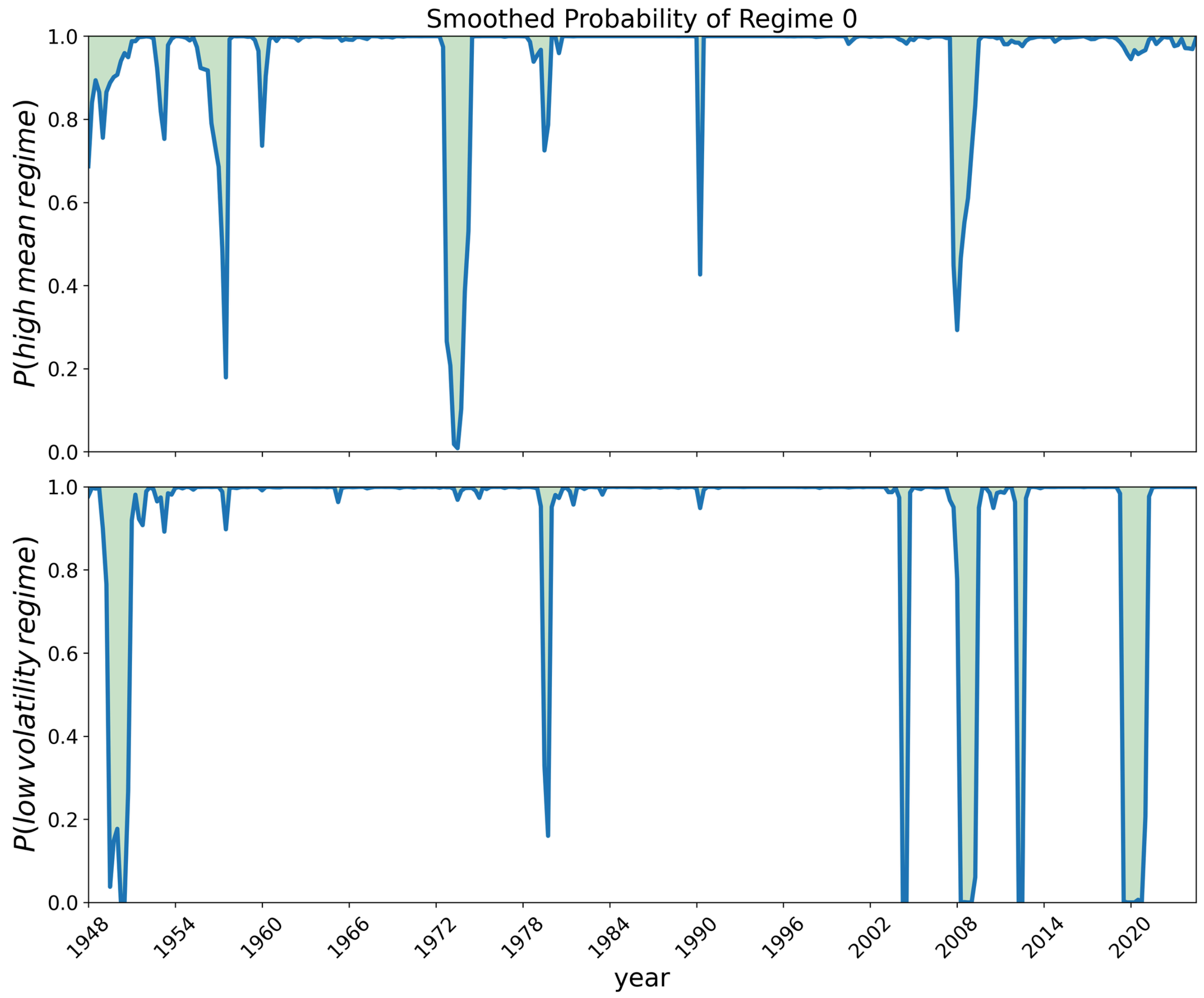

Fig. 6.7 Smoothed probabilities of being in the high-mean and low-volatility regimes.

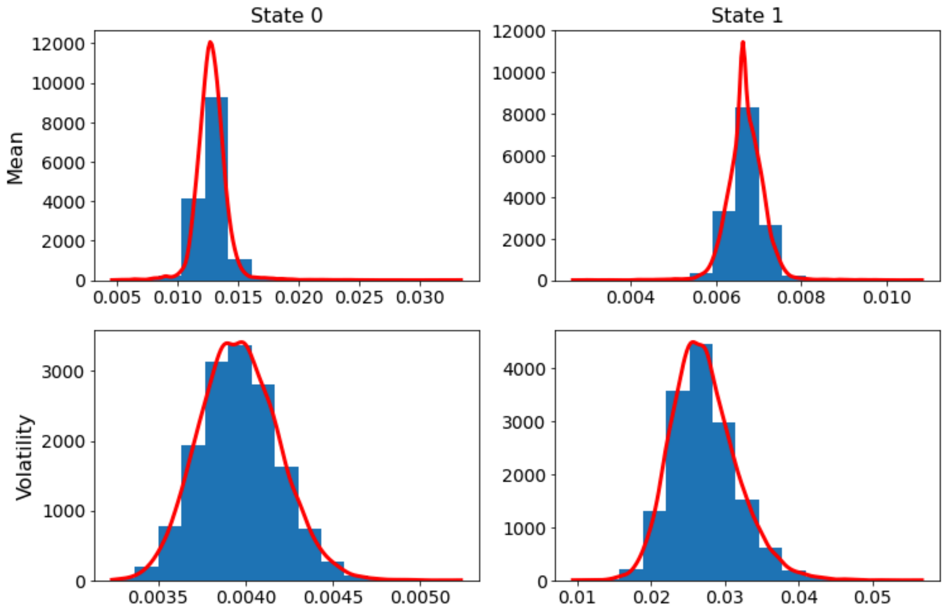

Fig. 6.8 The posterior distributions for the unconditional means and volatilities for $\Delta \log C_t$ across the two states. Sampling 20000 draws from the posterior and burning the first 5000 draws.

6.6. Summary

When a statistician observes signals but not the Markov state, the distribution of the hidden state conditioned on the signal history is itself a Markov process. This chapter constructs that auxiliary process in four settings: the Kalman filter and smoother for a linear-Gaussian system, a discrete hidden state, unknown parameters treated as invariant hidden states, and multiple vector autoregression regimes. In each, a finite list of sufficient statistics summarizes everything the signal history says about the hidden state.

Treating an unknown parameter as an invariant hidden state ties this chapter back to Chapter 1: the parameter indexes a statistical model and never changes, so learning it is learning which invariant event occurred. The likelihood ratios used here to update beliefs are studied as processes in their own right in Chapter 7, and the sensitivity of conclusions to a prior chosen for convenience, raised in the discussion of conjugate priors, is taken up in Chapter 10.

6.7. Exercises

Exercise 6.1 (Muth's inverse optimal prediction problem)

John F. Muth [1960] asked: for what stochastic process $\{Z_t\}$ is Milton Friedman [1957]'s adaptive expectations scheme

$$Z_t^* = \lambda Z_t + (1-\lambda) Z_{t-1}^*, \qquad 0 < \lambda < 1$$

the *optimal* forecast, and over what horizons? This exercise recovers his answer from the Kalman filter.

In the state-space system (6.1) set $\mathbb{N} = 0$, $\mathbb{A} = \mathbb{D} = 1$, $\mathbb{B} = [\mathbb{B}_1 \quad 0]$, and $\mathbb{F} = [0 \quad \mathbb{F}_2]$ with $\mathbb{B}_1 \neq 0$ and $\mathbb{F}_2 \neq 0$, so that

$$X_{t+1} = X_t + \mathbb{B}_1 W_{1,t+1}, \qquad Z_{t+1} = X_t + \mathbb{F}_2 W_{2,t+1}.$$

(a) Show that the best forecast of Z_{t+k} at a date t when the state *is* observed is X_t, for every $k \geq 1$.

(b) Specialize the Riccati recursion (6.3) to this system and solve for its positive invariant solution $\overline{\Sigma}$ in closed form.

(c) Compute the associated Kalman gain $\overline{\mathcal{K}}$ from (6.2) and verify that $0 < \overline{\mathcal{K}} < 1$.

(d) Write the updating equation (6.4) for $\overline{X}_{t+1}$ and compare it with the adaptive expectations scheme. What is Muth's answer to his own question, and over which horizons k does it hold?

Exercise 6.2 (Time averaging and the random walk hypothesis)

Suppose the price of an asset follows a random walk $X_{t+1} = X_t + \mathsf{b} W_{t+1}$, with $\{W_t\}$ an i.i.d. sequence of standard normal random variables and $\mathsf{b} \neq 0$. A researcher wishes to test the random walk hypothesis, but the database reports not X_t but the two-period moving average

$$Z_t = \frac{1}{2}\left(X_t + X_{t-1}\right).$$

(a) Cast this in the form (6.1), taking X_t as the hidden state. Verify that $\mathbb{B}\mathbb{F}^\top \neq 0$, so this is one of the cases the chapter warns is excluded by many textbook treatments.

(b) Solve the Riccati recursion (6.3) for its invariant solution. Explain the answer in one sentence, without reference to the recursion.

(c) Compute the limiting Kalman gain. It lies outside $(0, 1)$; why is that not a contradiction?

(d) Use the innovations representation (6.5) to derive the process followed by $Z_{t+1} - Z_t$. Compute its first- and second-order autocorrelations. What does this say about testing the random walk hypothesis on time-averaged data?

(e) Relate the moving-average coefficient you found in (d) to the boundary case in Example 6.1.

Exercise 6.3 (The likelihood process for a linear state-space model)

Section Kalman Filter and Smoother shows how to compute the distribution of the date $t+1$ hidden state given data through $t+1$. Suppose now that the matrices $\mathbb{A}, \mathbb{B}, \mathbb{N}, \mathbb{D}, \mathbb{F}$ are not known but are parameterized by an underlying vector θ.

(a) Use the recursive solution to represent the *predictive density* of Z_{t+1} conditioned on date t information and on θ. Which outputs of the Kalman filter does it depend on?

(b) Show how to construct the date $t+1$ contribution to the log-likelihood function conditioned on current and past Z's, and write the likelihood process $L_t(\theta)$ for a sample $Z_1, \ldots, Z_t$.

(c) The likelihood process depends on θ through the filtering outputs $\left\{\overline{X}_j, \Omega_j\right\}$, which must be recomputed for each candidate θ. Which of these can be computed *before* seeing any data, and which cannot?

(d) Connect this construction to Chapter 7: when two parameter values θ and θ_0 are compared, which object is the multiplicative martingale increment N_{t+1}?

Exercise 6.4 (Learning an unknown constant)

[Jovanovic, 1979] studies a labor market in which a worker and a firm are uncertain about the quality θ of their match and learn about it from noisy signals. The state-space representation sets

$$X_{t+1} = X_t, \qquad Z_{t+1} = X_t + \mathbb{F} W_{t+1},$$

with $X_t = \theta$ a scalar drawn from $\mathcal{N}\left(\overline{X}_0, \Sigma_0\right)$ and $\mathbb{F} \neq 0$.

(a) Specialize (6.2) and (6.3) to this system. Show that the precision $1/\Sigma_t$ accumulates linearly and solve for Σ_t in closed form.

(b) Deduce the limits of Σ_t and of $\mathcal{K}(\Sigma_t)$, and give the rate at which each converges.

(c) Contrast this with Exercise 6.1, where the hidden state moves. Which feature of the state equation accounts for the difference?

(d) Remark 6.3 observes that the martingale-extraction apparatus of Chapter 4 behaves differently when an unknown *trend* is among the hidden states, because the term on the second line of (6.9) diverges with the horizon. Use your answer to (b) to explain why estimating an invariant state is a qualitatively different exercise from filtering a stationary one, and what that implies for a stationary initialization of $\overline{X}_0$.

(e) In [Jovanovic, 1979]'s application, early in a match Σ_t is large. Explain informally how that sustains a match that a fully informed pair would dissolve, and why quits and layoffs should then decline with job tenure.

6.8. Answers

Solution to Exercise 6.1 (Muth's inverse optimal prediction problem)

Click to hide

(a) X is a random walk, so $\mathbb{E}\left(X_{t+k-1} \mid X_t\right) = X_t$ for every $k \geq 1$. Since $Z_{t+k} = X_{t+k-1} + \mathbb{F}_2 W_{2,t+k}$ and the measurement noise has conditional mean zero, $\mathbb{E}\left(Z_{t+k} \mid X_t\right) = X_t$ for all $k \geq 1$. The forecast does not depend on the horizon: the random walk has no mean reversion to exploit.

(b) With $\mathbb{A} = \mathbb{D} = 1$ we have $\mathbb{B}\mathbb{B}^\top = \mathbb{B}_1^2$, $\mathbb{F}\mathbb{F}^\top = \mathbb{F}_2^2$, and $\mathbb{B}\mathbb{F}^\top = 0$. Recursion (6.3) becomes

$$\Sigma_{t+1} = \Sigma_t + \mathbb{B}_1^2 - \frac{\Sigma_t^2}{\Sigma_t + \mathbb{F}_2^2}.$$

Setting $\Sigma_{t+1} = \Sigma_t = \overline{\Sigma}$ gives $\overline{\Sigma}^2 = \mathbb{B}_1^2 \left(\overline{\Sigma} + \mathbb{F}_2^2 \right)$, a quadratic whose positive root is

$$\overline{\Sigma} = \frac{\mathbb{B}_1^2 + \sqrt{\mathbb{B}_1^4 + 4\mathbb{B}_1^2\mathbb{F}_2^2}}{2}.$$

(c) From (6.2), $\overline{\mathcal{K}} = \overline{\Sigma} / \left(\overline{\Sigma} + \mathbb{F}_2^2 \right)$. Since $\overline{\Sigma} > 0$ and $\mathbb{F}_2^2 > 0$, the ratio lies strictly between zero and one. The two extremes are informative: $\overline{\mathcal{K}} \to 1$ as $\mathbb{F}_2 \to 0$, when the signal is clean and the filter jumps immediately to the newest observation, and $\overline{\mathcal{K}} \to 0$ as $\mathbb{B}_1 \to 0$, when the state barely moves and the filter is reluctant to revise.

(d) Equation (6.4) reads $\overline{X}_{t+1} = \overline{X}_t + \overline{\mathcal{K}} \left(Z_{t+1} - \overline{X}_t \right)$, that is

$$\overline{X}_{t+1} = \left(1 - \overline{\mathcal{K}}\right)\overline{X}_t + \overline{\mathcal{K}} Z_{t+1},$$

which is Friedman's scheme with $Z_t^* = \overline{X}_t$ and $\lambda = \overline{\mathcal{K}}$.

Muth's answer: adaptive expectations are *rational* exactly when the observed series is a random walk observed with independent measurement error, and the smoothing parameter λ is not free but pinned down by the signal-to-noise ratio $\mathbb{B}_1/\mathbb{F}_2$. By (a) the scheme is then optimal at *every* horizon $k \geq 1$ simultaneously, which is the second half of the question and a special feature of the random-walk state.

Solution to Exercise 6.2 (Time averaging and the random walk hypothesis)

Click to hide

(a) Take X_t as the state. Then

$$Z_{t+1} = \frac{1}{2}\left(X_{t+1} + X_t\right) = \frac{1}{2}\left(X_t + \mathsf{b}W_{t+1} + X_t\right) = X_t + \frac{\mathsf{b}}{2}W_{t+1},$$

so in the notation of (6.1), $\mathbb{N} = 0$, $\mathbb{A} = \mathbb{D} = 1$, $\mathbb{B} = \mathsf{b}$, and $\mathbb{F} = \mathsf{b}/2$. The *same* shock drives the state and the measurement, so $\mathbb{B}\mathbb{F}^\top = \mathsf{b}^2/2 \neq 0$. This is the configuration flagged in the footnote to (6.1).

(b) Recursion (6.3) becomes

$$\Sigma_{t+1} = \Sigma_t + \mathsf{b}^2 - \frac{\left(\Sigma_t + \frac{1}{2}\mathsf{b}^2\right)^2}{\Sigma_t + \frac{1}{4}\mathsf{b}^2}.$$

Setting $\Sigma_{t+1} = \Sigma_t = \overline{\Sigma}$ and clearing denominators, the $\mathsf{b}^2\overline{\Sigma}$ and $\frac{1}{4}\mathsf{b}^4$ terms cancel and we are left with $\overline{\Sigma}^2 = 0$, so $\overline{\Sigma} = 0$.

The one-sentence reason: $Z_t = \frac{1}{2}(X_t + X_{t-1})$ can be inverted, since $X_t = 2Z_t - X_{t-1}$, so an observer who knows X_0 and has seen $Z_1, \ldots, Z_t$ knows X_t exactly and there is nothing left to filter.

(c) From (6.2), $\mathcal{K}(\Sigma) = \left(\Sigma + \frac{1}{2}\mathsf{b}^2\right) / \left(\Sigma + \frac{1}{4}\mathsf{b}^2\right)$, which at $\Sigma = 0$ equals 2. There is no contradiction because the Kalman gain is a *regression coefficient*, not a weight in a convex combination. The bound $0 < \mathcal{K} < 1$ in Exercise 6.1 came from $\mathbb{B}\mathbb{F}^\top = 0$, which makes $\mathcal{K}$ the ratio of a variance to a larger variance. Here the numerator carries the extra covariance term $\mathbb{B}\mathbb{F}^\top$ and can exceed the denominator.

(d) With $\Sigma = 0$ the innovations representation (6.5) is

$$\overline{X}_{t+1} = \overline{X}_t + 2U_{t+1}, \qquad Z_{t+1} = \overline{X}_t + U_{t+1},$$

with $\Omega = \mathbb{D}\Sigma\mathbb{D}^\top + \mathbb{F}\mathbb{F}^\top = \frac{1}{4}\mathsf{b}^2$. Differencing the observation equation and substituting $\overline{X}_t - \overline{X}_{t-1} = 2U_t$,

$$Z_{t+1} - Z_t = \left(\overline{X}_t + U_{t+1}\right) - \left(\overline{X}_{t-1} + U_t\right) = U_{t+1} + U_t.$$

So the measured price change is a first-order moving average with coefficient exactly one. Its variance is $2\Omega = \frac{1}{2}\mathsf{b}^2$ and its autocorrelations are

$$\rho(1) = \frac{1}{1+1^2} = \frac{1}{2}, \qquad \rho(j) = 0 \text{ for } j \geq 2.$$

Measured price changes are serially correlated even though the underlying price is a martingale. A researcher who tests the random walk hypothesis by looking for zero autocorrelation in ΔZ will reject it with probability approaching one, and the rejection says nothing whatever about X. The serial correlation is an artifact of time averaging, and its magnitude, one half, is a prediction rather than a free parameter.

(e) Example 6.1 studies $Z_t = W_t + \lambda W_{t-1}$ and finds that the invariant variance is $\Sigma = 0$ when $|\lambda| < 1$, in which case the filter recovers the original representation, and $\Sigma = (\lambda^2 - 1)/\lambda^2$ when $|\lambda| > 1$, in which case it does not. Here $\lambda = 1$ exactly: the boundary between the two regimes, where the two roots coincide. That is why the convergence $\Sigma_t \to 0$ found in (b) is slow rather than geometric, and why the shock U_t can be recovered from current and past Z's only in a limiting sense.

Solution to [Exercise 6.3 (The likelihood process for a linear state-space model)] Click to hide

(a) Step 2 of the filter gives the distribution of Z_{t+1} conditioned on Z^t and Q_0:

$$Z_{t+1} \mid Z^t \sim \text{Normal}\left(\mathbb{N} + \mathbb{D}\overline{X}_t,\ \Omega_t\right), \qquad \Omega_t = \mathbb{D}\Sigma_t\mathbb{D}^\top + \mathbb{F}\mathbb{F}^\top.$$

Writing $\psi\left(z^* \mid \mu, \Sigma\right)$ for a normal density, the predictive density is $\psi\left(z^* \mid \mathbb{N} + \mathbb{D}\overline{X}_t, \Omega_t\right)$. It depends on the data only through the filtered mean $\overline{X}_t$ and on the parameters through $\mathbb{N}, \mathbb{D}$ and the covariance Σ_t.

(b) Evaluating the predictive density at the realized Z_{t+1} gives the date $t+1$ contribution

$$\log \psi\left(Z_{t+1} \mid \mathbb{N} + \mathbb{D}\overline{X}_t, \Omega_t\right) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log\det\Omega_t - \frac{1}{2}U_{t+1}^\top\Omega_t^{-1}U_{t+1},$$

where $U_{t+1} = Z_{t+1} - \mathbb{N} - \mathbb{D}\overline{X}_t$ is the innovation. Factoring the joint density into a product of one-step-ahead conditionals,

$$L_t(\theta) = \prod_{j=1}^{t} \psi\left(Z_j \mid \mathbb{N} + \mathbb{D}\overline{X}_{j-1}, \Omega_{j-1}\right),$$

whose logarithm is the sum of the contributions above. This is the prediction-error decomposition: the Kalman filter turns a joint density over a serially dependent sample into a product of independent normal densities for the innovations.

(c) The covariance sequence $\{\Sigma_j, \Omega_j, \mathcal{K}(\Sigma_j)\}$ solves recursion (6.3), which involves no data at all. For a given θ it can be computed once and stored *before* any Z is observed. The filtered means $\left\{\overline{X}_j\right\}$ cannot: they obey (6.4), which feeds in each realized Z_j. This separation is what makes maximum likelihood for state-space models practical, since the expensive matrix recursion is data independent.

(d) Form the ratio of the likelihood processes at two parameter values:

$$\frac{L_t(\theta)}{L_t(\theta_0)} = \prod_{j=1}^{t} \frac{\psi\left(Z_j \mid \mathbb{N}(\theta) + \mathbb{D}(\theta)\overline{X}_{j-1}(\theta), \Omega_{j-1}(\theta)\right)}{\psi\left(Z_j \mid \mathbb{N}(\theta_0) + \mathbb{D}(\theta_0)\overline{X}_{j-1}(\theta_0), \Omega_{j-1}(\theta_0)\right)}.$$

Each factor is the multiplicative increment N_j of Chapter 7, and the product is the multiplicative martingale M_t of Definition 7.1, a martingale under the θ_0 model. Note that $\overline{X}_{j-1}$ must be recomputed under *each* parameter value: the two models disagree not only about the density of Z_j given the state, but about what the state is inferred to be. Remark 7.2 makes exactly this point when it observes that a hidden-state model requires an "information state vector" constructed by conditioning on the infinite past.

Solution to Exercise 6.4 (Learning an unknown constant) Click to hide

(a) Here $\mathbb{A} = \mathbb{D} = 1$ and $\mathbb{B} = 0$, so $\mathbb{B}\mathbb{B}^\top = 0$ and $\mathbb{B}\mathbb{F}^\top = 0$. Then (6.2) gives $\mathcal{K}(\Sigma_t) = \frac{\Sigma_t}{\Sigma_t + \mathbb{F}^2}$ and (6.3) gives

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t^2}{\Sigma_t + \mathbb{F}^2} = \frac{\Sigma_t \mathbb{F}^2}{\Sigma_t + \mathbb{F}^2}.$$

Inverting, $\frac{1}{\Sigma_{t+1}} = \frac{1}{\Sigma_t} + \frac{1}{\mathbb{F}^2}$: each observation adds a fixed increment $1/\mathbb{F}^2$ to the precision. Hence

$$\frac{1}{\Sigma_t} = \frac{1}{\Sigma_0} + \frac{t}{\mathbb{F}^2}, \qquad \Sigma_t = \frac{\Sigma_0 \mathbb{F}^2}{\mathbb{F}^2 + t\Sigma_0}.$$

(b) $\Sigma_t \downarrow 0$ and $\mathcal{K}(\Sigma_t) \to 0$, both at rate $1/t$: $\Sigma_t \approx \mathbb{F}^2/t$ and $\mathcal{K}(\Sigma_t) \approx 1/t$ for large t. The pair eventually learns θ exactly, but slowly, and the weight placed on each new observation shrinks to zero.

(c) The state equation. In Exercise 6.1 the state is refreshed each period by $\mathbb{B}_1 W_{1,t+1}$, so the filter is chasing a moving target and its uncertainty settles at a strictly positive $\overline{\Sigma}$ rather than vanishing. Here $\mathbb{B} = 0$: the target does not move, information accumulates without ever being made stale, and uncertainty is driven to zero. There is no positive invariant Σ at all.

(d) Filtering a stationary hidden state is a *steady-state* problem: Σ_t converges to a fixed point, the gain becomes time invariant, and the filtered system (6.7) is itself a stationary Markov process. Estimating an invariant state is not: Σ_t never settles, the gain keeps falling, and the model does not admit a stationary representation of the filtered system.

Remark 6.3 identifies this wedge. When one of the hidden states is an unknown trend coefficient, its contribution to the impact of $\overline{W}_{t+1}$ on $Y_{t+\tau}$ enters multiplied by τ in the second line of (6.9), and so diverges as τ grows. A stationary initialization would require the lower-right entry of Σ_0 to be zero, that is, it would require the trend coefficient to be *known*. Assuming stationarity of the filtered system therefore assumes away the very learning problem being posed, which is why Remark 6.3 treats this case separately.

> **(e)** Early in a match, Σ_t is large, so a run of poor signals moves $\overline{X}_t$ only a little: the pair rationally attributes bad draws to noise rather than to low quality, and the match survives. As tenure accumulates, Σ_t falls; posterior beliefs become tight and are no longer easily revised. Matches whose true quality is low are therefore identified and dissolved at low tenure, so the surviving population is increasingly selected toward high θ. Quits and layoffs decline with tenure, and measured wages rise with tenure, without any of the productivity growth that a human-capital account would require. This is [Jovanovic, 1979]'s explanation of two robust features of labor-market data.

---

[1] Many expositions of Kalman filtering assume that $\mathbb{B}\mathbb{F}' = 0$. We shall study some interesting examples in which $\mathbb{B}\mathbb{F}' \neq 0$.

[2] The process $\{X_t, t = 0, 1, 2, \ldots\}$ is also Markov.

[3] This amounts to dividing the joint distribution for (X_{t+1}, Z_{t+1}) conditioned on Q_t by the marginal density for Z_{t+1} conditional on Q_t.

[4] Presentations of multivariate regression theory often report the transpose of this matrix. Those presentations pre-multiply coefficients by regressors whereas Kalman filtering representations post-multiply by regressors.

[5] This stochastic process is not ergodic, being a mixture of statistical models like those described by Proposition 1.2. In the present setting, conditioning on invariant events means knowing parameters, an assumption incompatible with posing a statistical learning problem.

[6] This factorization can be implemented as a Cholesky decomposition.

[7] A decision maker who does not know the underlying parameters in the matrices A, B, D, F, N continues to have a Markov decision problem except that b_t, c_t, d_t must now be included along with the state vector X_t.

[8] Such a procedure can result in estimators that are inadmissible.

[9] [Box and Tiao, 1992] discuss improper priors that include the specification for the regression model here.

[10] Another approach that we don't use here would be to modify how we construct the likelihood function. Currently, the likelihood function conditions on the initial X_0. We could instead impose that X_0 is described by the stationary distribution associated with a stable A matrix.

[11] We could also have used change in variables formulas to deduce posterior distributions of interest, but that would have involved substantial pencil-and-paper work and required additional numerical computation.

[12] Bounding absolute values of these eigenvalues to be less than a pre-specified number strictly less than one would thin the right tail. Doing that amounts indirectly to imposing a particular prior on the size of long-run risk.

[13] Our consumption measure is nondurables plus services consumption per capita. The nominal consumption data come from BEA's NIPA Table 1.1.5 and their deflators from BEA's NIPA Table 1.1.4. The business income data with IVA and CCadj are from BEA's NIPA Table 1.12. Personal dividend income data were obtained from FRED's B703RC1Q027SBEA. Population data comes from FRED's CNP16OV.

[14] By including proprietors' income in addition to corporate profits, we used a broader measure of business income than [Hansen *et al.*, 2008] who used only corporate profits. [Hansen *et al.*, 2008] did not include personal dividends in their VAR analysis.