Hidden Markov Models

Contents

⤓ Download PDF here

The Markov models provide a setting in which it is natural to study learning about states that are hidden from a statistician or decision maker who observes only possibly noisy signals of some or all of the states. When the state moves over time, the statistician wants to learn about a moving target.

To study such situations, this chapter presents Hidden Markov Models that start from a joint probability distribution consisting of a Markov process and a vector of noise-ridden signals about functions of the Markov state. The statistician observes a history of signals, but not the Markov state vector. Statistical learning about the Markov state proceeds by constructing a sequence of probability distributions of the Markov state conditional on histories of signals. Recursive representations of these conditional distributions form auxiliary Markov processes that summarize all of the information about the hidden state vector that is contained in a history of signals. A state vector in this auxiliary Markov process is a set of sufficient statistics for the probability distribution of the hidden Markov state conditional on the history of signals. We describe how to construct this auxiliary Markov process of sufficient statistics sequentially.

We present four examples of Hidden Markov Models that are 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 cast as hidden invariant states
4. Multiple VAR regimes

5.1. 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

$$\begin{aligned} X_{t+1} &= \mathbb{A}X_t + \mathbb{B}W_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}X_t + \mathbb{F}W_{t+1}, \end{aligned} \tag{5.1}$$

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 A, we allow A not to be a stable matrix.

Although $\{(X_t, Z_t), t = 0, 1, 2, \ldots\}$ is Markov, $\{Z_t, t = 0, 1, 2, \ldots\}$ is 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 can 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

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

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

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

which is just another way of describing our original state-space system (5.1). 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

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

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

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

This is called the predictive conditional density $\phi(z^*|Q_t)$, i.e., the distribution 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 $X_{t+1} - A\overline{X}_t$ on the surprise in Z_{t+1} defined as $Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t$.[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

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

These provide what we need to compute the conditional expectation

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

where the matrix of regression coefficients $\mathcal{K}(\Sigma_t)$ is called the *Kalman gain*. To compute the Kalman gain, multiply both sides by $(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t)'$ 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 for this equation for $\mathcal{K}(\Sigma_t)$ gives

$$\mathcal{K}(\Sigma_t) = (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}. \tag{5.2}$$

We recognize formula (5.2) as an application of the population least squares regression formula associated with the multivariate normal distribution.[4] We compute Σ_{t+1} via the recursion

$$\Sigma_{t+1} = \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' - (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}(\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')'. \tag{5.3}$$

The right side of recursion (5.3) 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 the $X_{t+1} - \mathbb{A}\overline{X}_t$ and the remaining term describes the reduction in covariance associated with conditioning on Z_{t+1}. Thus, the probability distribution Q_{t+1} is

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

where

$$\overline{X}_{t+1} = \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t) \tag{5.4}$$

Equations (5.2), (5.3), and (5.4) constitute the Kalman filter. They provide a recursion that describes Q_{t+1} as an exact function of Z_{t+1} and Q_t.

To summarize, the key idea underlying the Kalman filter is recursively to transform the space spanned by a sequence of signals into a sequence of orthogonal signals. To elaborate, let

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

After we condition on $(\overline{X}_0, \Sigma_0)$, $U_t, U_{t-1}, \ldots U_1$ and $Z_t, Z_{t-1}, \ldots, Z_1$ generate the same information. The Kalman filter synthesizes U_{t+1} from Z^{t+1} via a what is called a Gram-Schmidt process. Conditional on Z^t, $U_{t+1} \sim \mathcal{N}(0, \Omega_t)$, where $\Omega_t = \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}'$, so $U^t = U_t, U_{t-1}, \ldots U_1$ is an orthogonal basis for information contained in Z^t. Step2 computes the innovation U_{t+1} by constructing the predictive density, while step3 computes the Kalman gain $\mathcal{K}(\Sigma_t)$ by regressing $X_{t+1} - \mathbb{A}\overline{X}_t$ on U_{t+1}.

Remark 5.1

Taken together, step2 and step3 present the evolution of $\{Q_{t+1}\}$ as a first-order Markov process. This process is the foundation of an *innovations representation* and its partner the *whitener*. The innovations representation is

$$\begin{aligned} \overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)U_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}\overline{X}_t + U_{t+1}. \end{aligned} \tag{5.5}$$

The *whitener* system is

$$\begin{aligned} U_{t+1} &= Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t \\ \overline{X}_{t+1} &= [\mathbb{A} - \mathbb{D}\mathcal{K}(\Sigma_t)]\overline{X}_t + \mathcal{K}(\Sigma_t)(Z_{t+1} - \mathbb{H}) \end{aligned} \tag{5.6}$$

The innovations representation (5.5) and the whitener system (5.6) both take sequences $\{\Sigma_t, \mathcal{K}(\Sigma_t)\}_{t=0}$ as inputs. These can be precomputed from equations (5.2) and (5.3) before observing any Z_{t+1}'s.

Remark 5.2

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 $\overline{\mathbb{F}}_t$ that satisfies

$$\Omega_t = \overline{\mathbb{F}}_t(\overline{\mathbb{F}}_t)'.$$

The factorization on the right side is not unique. For instance, we could find solutions under which $\overline{\mathbb{F}}_t$ 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;

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

Then

$$\begin{aligned} \overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \overline{\mathbb{B}}_t\overline{W}_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}\overline{X}_t + \overline{\mathbb{F}}_t\overline{W}_{t+1} \end{aligned} \tag{5.7}$$

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

While we may think of $\{(Y_t, \overline{X}_t, \Sigma_t) : t \geq 0\}$, it will typically not be stationary. Often a stationary counterpart does exist, however. A necessary requirement is that Σ_t be invariant, because it does not depend on the underlying shocks. To construct such a process, find a positive semidefinite fixed point to the recursion (5.3) and initialize $\Sigma_0 = \overline{\Sigma}$. Then $\Sigma_t = \overline{\Sigma}$ for all $t \geq 0$ and

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

for all $t \geq 1$. This simplifies recursive representation (5.7) by making $\overline{\mathbb{B}}_t, \overline{\mathbb{F}}_t$ and Ω_t all time-invariant. One way to obtain such a construction of $\overline{\Sigma}$ is to iterate on equation (5.3) to convergence, assuming that such a limit does exits. With this construction, we have a Markov process representation

$$\begin{aligned}\overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \overline{\mathbb{B}}\overline{W}_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}\overline{X}_t + \overline{\mathbb{F}}\overline{W}_{t+1}\end{aligned} \tag{5.8}$$

Compare this to the original state space system (5.1). Key differences are

1. In the original system (5.1), the shock vector W_{t+1} can be of much larger dimension than the time $t+1$ observation vector Z_{t+1}, while in (5.8), the dimension $\overline{W}_{t+1}$ equals that of the observation vector.
2. The state vector X_t in the original system (5.1) is not observed while in representation (5.8), the state vector $\overline{X}_t$ is observed.

5.1.1. Likelihood process

Equations (5.2) and (5.3) together with an initial distribution Q_0 for $X_0 \sim \mathcal{N}(\overline{X}_0, \Sigma_0)$ provide components that allow us to construct a recursive representation for a likelihood process for $\{Z_t : t = 1, 2, \ldots\}$. Let $\psi(z^* \mid \mu, \Sigma)$ denote the density for an m dimensional, normally distributed random vector with mean μ and covariance matrix Σ. With this notation, the density of Z_{t+1} conditional on the hidden state X_t is $\psi(z^* \mid \mathbb{H} + \mathbb{D}X_t, \mathbb{F}\mathbb{F}')$, where z^* is an m dimensional vector of real numbers used to represent potential realizations of Z_{t+1}. The distribution of the hidden state X_t conditioned on history Z^{t-1} and $(\overline{X}_0$ and $\Sigma_0)$ is $Q_t \sim \mathcal{N}(\overline{X}_t, \Sigma_t)$. From these two components, we construct the predictive density $\phi(z^* \mid Z^t)$ for Z_{t+1}:

$$\phi(z^* \mid Z^t, \overline{X}_0, \Sigma_0) = \int \psi(z^* \mid x) Q_t(dx). \tag{5.9}$$

From the Kalman filter, we know that

$$\int \psi(z^* \mid x) Q_t(dx) = \psi(z^* \mid \mathbb{H} + \mathbb{D}\overline{X}_t, \Omega_t)$$

To compute a likelihood process $\{L_t : t = 1, 2, \ldots\}$, factor the joint density for Z^t into a product of conditional density functions in which a time j density function conditions on past information and the initial $(\overline{X}_0, \Omega_0)$. When we evaluate densities at the appropriate random vectors Z_j and the associated histories Z^{j-1} of which $\overline{X}_{j-1}, \Omega_{j-1}$ are functions determined by the Kalman filter, we obtain the likelihood process: [5]

$$L_t = \prod_{j=1}^{t} \psi(Z_j \mid \mathbb{H} + \mathbb{D}\overline{X}_{j-1}, \Omega_{j-1}). \tag{5.10}$$

Via the Kalman filtering formulas for $\{\overline{X}_j, \Omega_j\}_{j=1}^{\infty}$, this construction indicates how the likelihood process depends on the matrices $\mathbb{A}, \mathbb{B}, \mathbb{H}, \mathbb{D}, \mathbb{F}$. Sometimes we regard some entries of these matrices as “free parameters.” Because a likelihood process summarizes information about these parameters, it is the starting point for both frequentist and Bayesian estimation procedures.

1. For fixed values of the parameters that pin down $\mathbb{A}, \mathbb{B}, \mathbb{H}, \mathbb{D}, \mathbb{F}$, $\{L_t\}_{t=1}^{\infty}$ is a stochastic process with some “interesting properties.”
2. For a fixed t and a sample of observations Z^t, L_t becomes a “likelihood function” when viewed as a function of the free parameters.

Example 5.1

John F. Muth [1960] posed and solved the following inverse optimal prediction problem: for what stochastic process $\{Z_t : t \geq 0\}$ is the adaptive expectations scheme of Milton Friedman [1957]

$$Z_t^* = \lambda Z_t + (1-\lambda) Z_{t-1}^* \quad 0 < \lambda < 1 \tag{5.11}$$

optimal for predicting future Z_{t+k}? And over what horizon k, if any, is Z_t^* a good forecast?

Although Muth did not use it to solve his problem, we can convey his answers concisely using the Kalman filter. As described above, initialize the initial covariance matrix for the Kalman filter at $\Sigma_0 = \overline{\Sigma}$ where the latter is the time-invariant solution to the covariance matrix updating equation. Set $\mathbb{A} = \mathbb{D} = 1$, $\mathbb{B} = [\mathbb{B}_1 \quad 0]$, and $\mathbb{F} = [0 \quad \mathbb{F}_2]$ to attain the original state-space system

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

Notice that the best forecast of Z_{t+k} at the time t when the state is observed is X_t for any $k \geq 1$. By the Law of Iterated Expectations, we obtain the mathematical expectation of Z_{t+k} conditional on Z^t by computing $\overline{X}_t$. A time-invariant recursive representation of $\overline{X}_{t+1}$ is

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

where it can be verified that $0 < \overline{\mathcal{K}} < 1$. Notice that

$$\overline{X}_{t+1} = (1 - \overline{\mathcal{K}})\overline{X}_t + \overline{\mathcal{K}} Z_{t+1} \tag{5.12}$$

Comparing (5.11) to (5.12) shows that “adaptive” expectations become “rational” by setting

$$\begin{aligned}\overline{X}_t &= Z_t^* \\ \lambda &= \overline{\mathcal{K}}.\end{aligned}$$

Example 5.2

As state variables for the key Bellman equation in his matching model, Jovanovic [1979] deployed sufficient statistics of conditional distribution Q_t for a univariate hidden Markov state equal to an unknown constant match quality θ drawn from a known initial distribution $\mathcal{N}\left(\overline{X}_0, \Sigma_0\right)$. The state-space representation for Jovanovic [1979]'s model is

$$\begin{aligned}X_{t+1} &= X_t \\ Z_{t+1} &= X_t + \mathbb{F}W_{t+1}\end{aligned}$$

where $\mathbb{F}$ and $X_t = \theta$ are scalars and W_{t+1} is a standardized univariate normal random variable. We fit this model into (5.1) by setting $\mathbb{A} = \mathbb{D} = 1, \mathbb{B} = 0, \mathbb{F} > 0, X_t = \theta$. Evidently, $\overline{X}_{t+1} = (1 - \mathcal{K}(\Sigma_t))\overline{X}_t + \mathcal{K}(\Sigma_t)Z_t$ where $\Sigma_{t+1} = \frac{\Sigma_t \mathbb{F}^2}{\Sigma_t + \mathbb{F}^2}$ and $\mathcal{K}(\Sigma_t) = \frac{\Sigma_t}{\Sigma_t + \mathbb{F}^2}$. Thus, $\frac{1}{\Sigma_{t+1}} = \frac{1}{\Sigma_t} + \frac{1}{\mathbb{F}^2} \downarrow 0$ and $\mathcal{K}(\Sigma_t) \rightarrow 0$. Thus, partners to an ongoing match who observe Z^t eventually learn its true quality θ. In Jovanovic [1979]'s model, especially when $\mathbb{F}$ is large, early on in a match, Σ_t can be large enough to create a situation in which the "he's just been having a few bad days" excuse prevails to sustain the match in hopes of later learning that it is a good one. Jovanovic [1979] put this force to work to help explain why (a) quits and layoffs are negatively correlated with job tenure and (b) wages rise with job tenure.

5.1.2. 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 $X_t \sim \mathcal{N}(\overline{X}_t, \Sigma_t)$; 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, 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** on 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

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

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

$$\widehat{\mathbb{K}}_1\left(X_{t+1} - \mathbb{A}\overline{X}_t\right) + \widehat{\mathbb{K}}_2\left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t\right) \tag{5.13}$$

where the regression coefficient matrix is

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

and the residual covariance matrix equals

$$\Sigma_t - \begin{bmatrix}\Sigma_t\mathbb{A}' & \Sigma_t\mathbb{D}'\end{bmatrix}\begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{D}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1}\begin{bmatrix}\mathbb{A}\Sigma_t \\ \mathbb{D}\Sigma_t\end{bmatrix} \tag{5.14}$$

- 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 $\widehat{X}_t$:

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

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:

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

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

5.2. 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} with 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:

$$\underset{\substack{\uparrow \\ X_{t+1}\ \text{density}}}{(\mathbb{P}'X_t)} \qquad \times \qquad \underset{\substack{\uparrow \\ Z_{t+1}\ \text{density}}}{(X_t)'\text{vec}\left\{\psi_i(y^*)\right\}} \tag{5.15}$$

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 (5.15) conditioned on Y^t, Q_0:

$$\mathbb{P}'\text{diag}\{Q_t\}\text{vec}\left\{\psi_i(z^*)\right\}, \tag{5.16}$$

   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 (5.16) over the hidden states gives

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

   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 $Y_{t+1} - Y_t$ 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 (Q_t, Z_{t+1}). It takes the form of a vector Q_{t+1} of conditional probabilities. Thus, we are led to

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \operatorname{vec}\left\{\psi_i(Z_{t+1})\right\}} \right) \mathbb{P}' \operatorname{diag}(Q_t) \operatorname{vec}\left\{\psi_i(Z_{t+1})\right\} \tag{5.17}$$

Together, step 3 and step4 define a Markov process for Q_{t+1}. As indicated in step3, Z_{t+1} is drawn from a (history-dependent) mixture of densities ψ_i. As indicated in step4, the vector Q_{t+1} equals the exact function of Z_{t+1}, Q_t described in (5.17).

5.3. 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.[6] From increments to a data history, the statistician learns about parameters sequentially. By assuming that the statistician adopts a conjugate prior à la Luce and Raiffa [1957], we can construct explicit updating formulas.

Consider the first-order vector autoregressive model

$$\begin{aligned} X_{t+1} &= \mathbb{A} X_t + \mathbb{B} W_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D} X_t + \mathbb{F} W_{t+1} \end{aligned} \tag{5.18}$$

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 $\mathbb{A}, \mathbb{B}, \mathbb{D}, \mathbb{F}, \mathbb{H}$ are matrices containing unknown coefficients. When $\mathbb{A}$ is a stable matrix, the vector $\mathbb{H}$ is interpretable as the vector of means of the observation vector Z_{t+1} (conditioned on invariant events).

Suppose that Z_{t+1} and W_{t+1} share the same dimensions, that $\mathbb{F}$ is nonsingular, and that X_t consists of $Z_t - H$ and a finite number of lags $Z_{t-j} - \mathbb{H}, j = 0, \ldots, \ell - 1$. After substitution for the state vector, we obtain a finite-order vector autoregression:

(5.19)

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

where

$$
\widetilde{\mathbb{H}} \overset{\text{def}}{=} \mathbb{H} - \mathbb{D} \begin{bmatrix} \mathbb{H} \\ \mathbb{H} \\ \dots \\ \mathbb{H} \end{bmatrix}
$$

Our plan is to estimate the coefficients of the matrices $\mathbb{H}, \mathbb{D}$, and $\mathbb{F}$. Notice that $\mathbb{H}$ potentially can be recovered from $\widetilde{\mathbb{H}}$ 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.” In what follows, we impose a convenient normalization on $\mathbb{F}$. Other observationally equivalent $\mathbb{F}$’s can be constructed from our estimation.

5.3.1. 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. In a later chapter, we will use this 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 can transform system (5.19) in a way that justifies estimating the unknown coefficients by applying least squares equation-by-equation. Factor the matrix $\mathbb{F}\mathbb{F}' = \mathbb{J}\Delta\mathbb{J}'$, where $\mathbb{J}$ is lower triangular with ones on the diagonal and Δ is diagonal.[7] The inverse $\mathbb{J}^{-1}$ is also lower triangular with ones on the diagonal. Construct

$$
\mathbb{J}^{-1} Z_{t+1} = \mathbb{J}^{-1}\widetilde{\mathbb{H}} + \mathbb{J}^{-1}\mathbb{D} \begin{bmatrix} Z_t \\ Z_{t-1} \\ \dots \\ Z_{t-\ell+1} \end{bmatrix} + U_{t+1} \tag{5.20}
$$

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 (5.20) 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 (5.20) 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 can construct estimates of the coefficient matrices $\mathbb{A}, \mathbb{B}, \mathbb{D}, \mathbb{F}, \mathbb{H}$ and the covariance matrix $\Delta = EU_{t+1}U_{t+1}'$ from these regression equations, with the qualification that knowledge of $\mathbb{J}$ and Δ determines $\mathbb{F}\mathbb{F}'$ only up to a factorization. One such factorization is $\mathbb{F} = \mathbb{J}\Delta^{1/2}$, where a diagonal matrix raised to a one-half power can be built by taking the square root of each diagonal entry. Because matrices $\mathbb{F}$ not satisfying this formula also satisfy $\mathbb{F}\mathbb{F}' = \mathbb{J}\Delta\mathbb{J}'$, without additional restrictions $\mathbb{F}$ is not identified.

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 (5.20). To simplify notation, we will omit superscripts and understand that we are estimating one equation at a time. To avoid notational confusion, we will 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 $Z^t = [Z_t', \ldots, Z_1']'$. Suppose that in addition Z_{t+1} and R_{t+1} are also observed at date $t+1$ but that β and σ^2 are unknown.

Let the distribution of β conditioned on Z^t, X_0, and 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 $t+1$, information we add Z_{t+1} to the conditioning set. So we want the distribution of β conditioned on Z^{t+1}, X_0, and σ^2. It is also normal but now has precision $\zeta\Lambda_{t+1}$.

To deduce the recursive updating conditioned on ζ, we observe that both the date t prior distribution and the date $t+1$ conditional density for Y_{t+1} are normal. After multiplication the two terms inside the exponential have two terms involving β are

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

where

$$\Lambda_{t+1} = R_{t+1}{R_{t+1}}' + \Lambda_t \tag{5.21}$$

and

$$\Lambda_{t+1}b_{t+1} = [\Lambda_t b_t + R_{t+1}Y_t]. \tag{5.22}$$

Recursion (5.21) implies that $\Lambda_{t+1} - \Lambda_t$ is a positive semidefinite matrix, which confirms that additional information improves estimation accuracy. Evidently from recursion (5.21), 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 (5.22). Solving difference equation (5.22) 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 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 $c_t + 1$ 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 Y^{t+1}, X_0 has a gamma density, so that it is proportional to[8]

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

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. [9] 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 (5.22); b_{t+1} are 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. [10]

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 the $(\mathbb{A}, \mathbb{B})$.

The conjugate prior approach described above does not generate a posterior for which either the prior or the implied posteriors for the matrix $\mathbb{A}$ has 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.[11]

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.[12]

As an illustration, in Fig. 5.1, we show posterior histograms for the standard deviations of shocks to short-term consumption growth and of the martingale increment to consumption based on the application described in Section An economic rationale of Chapter 4:Processes with Markovian increments. The standard deviation of the short-term shock contribution is about one-half that of the standard deviation of the martingale increment. Fig. 5.1 tells us that short-term risk can be inferred with much more accuracy than is long-term risk. 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 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. [13]

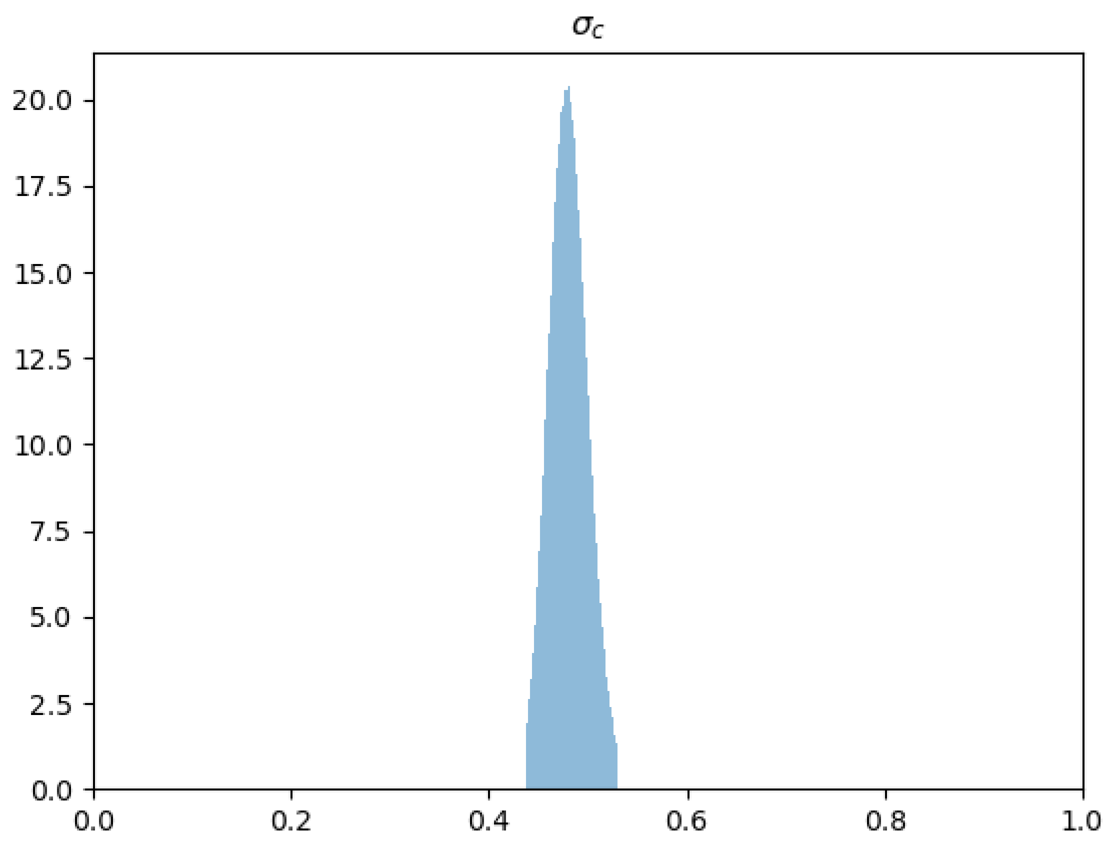


Fig. 5.1 Posterior density for conditional standard deviation of consumption growth.

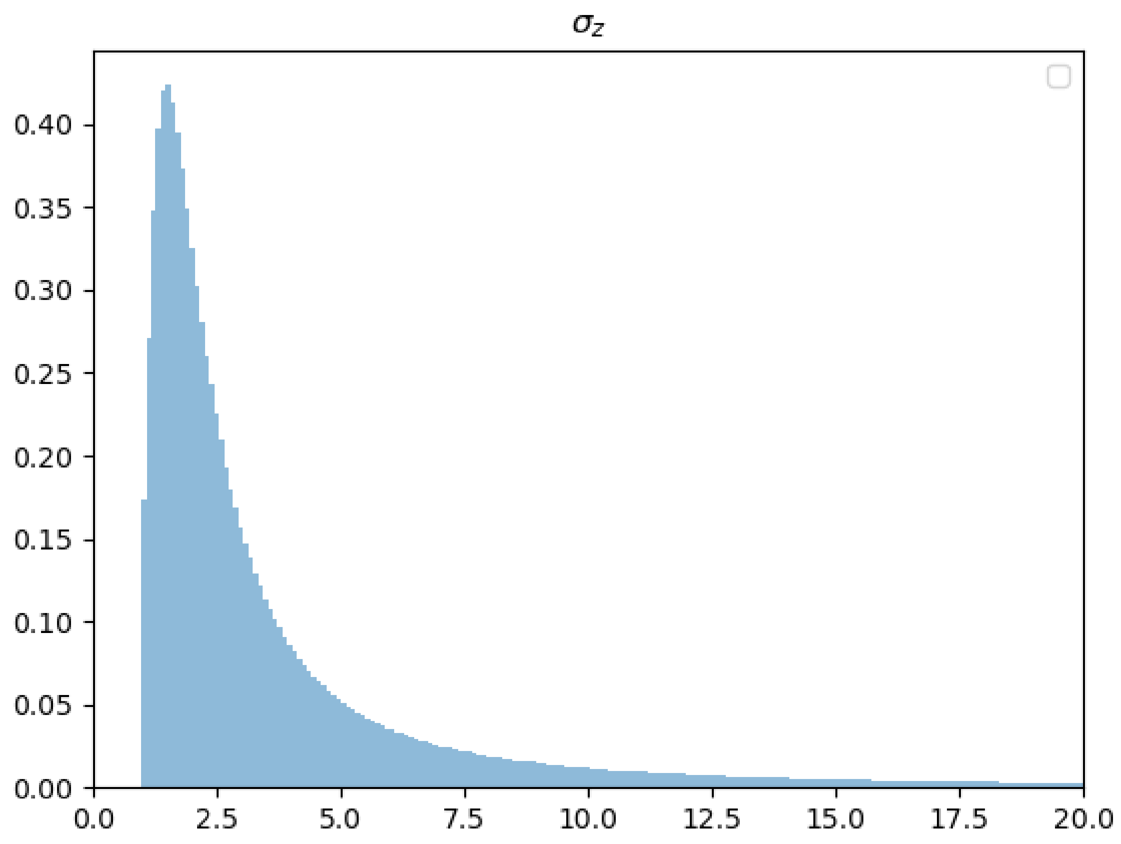


Fig. 5.2 Posterior distribution for the standard deviation of the martingale increment.

Remark 5.3

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

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

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

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

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.

5.4. 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

$$\begin{aligned} X_{t+1} &= \mathbb{A}_i X_t + \mathbb{B}_i W_{t+1} \\ Y_{t+1} - Y_t &= \mathbb{D}_i X_t + \mathbb{F}_i W_{t+1}, \end{aligned} \tag{5.23}$$

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

We can think of X_t and a regime indicator Z_t jointly as forming a Markov process. 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:

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

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:

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

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

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

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

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\,\{\psi_i(Y_{t+1}, X_t)\}}\right)\mathbb{P}'\text{diag}(Q_t)\text{vec}\,\{\psi_i(Y_{t+1}, X_t)\}. \tag{5.24}$$

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 (5.24).

Example 5.3

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. We estimate (5.23) using Gibbs sampling with 2 regimes.

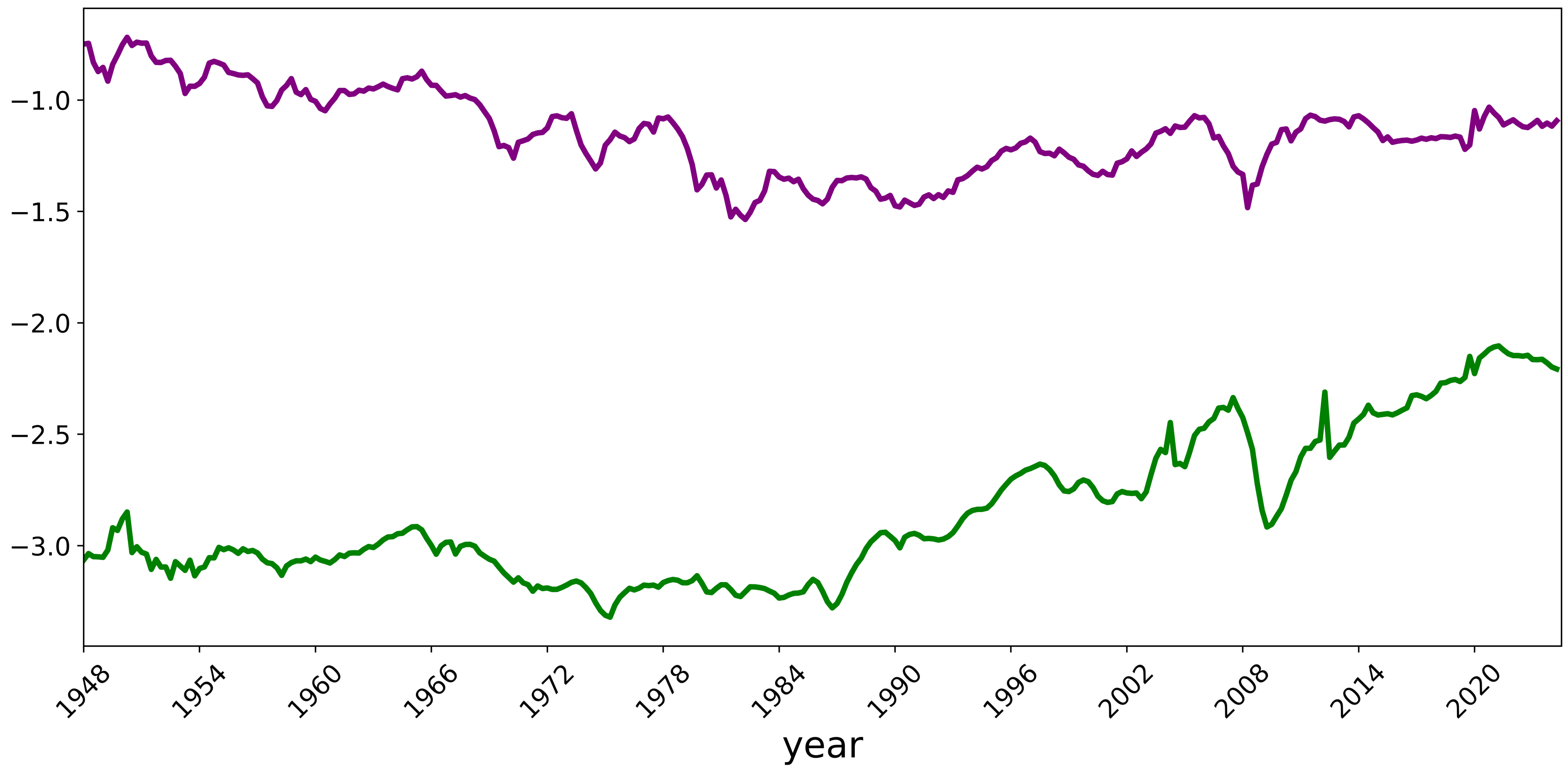


Fig. 5.3 Log business income less consumption $\log\left(B_t/C_t\right)$ (purple) and log dividends less consumption $\log\left(D_t/C_t\right)$ (green).

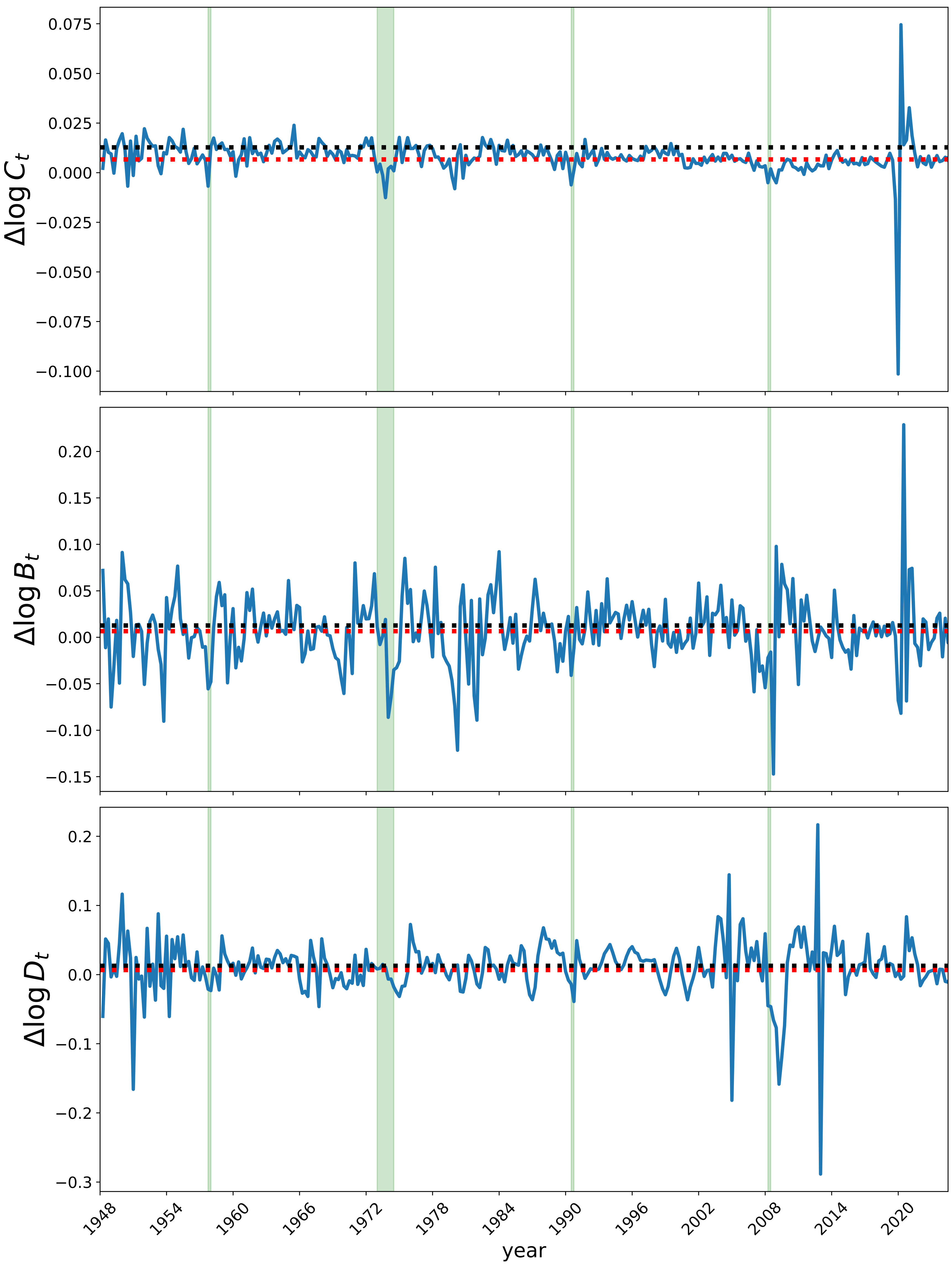

0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
$\Delta \log C_t$
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
$\Delta \log B_t$
0.2
0.1
0.0
−0.1
−0.2
−0.3
$\Delta \log D_t$
1948
1954
1960
1966
1972
1978
1984
1990
1996
2002
2008
2014
2020
year

Fig. 5.4 The log growth rates of consumption, business income and dividends are shown as solid lines in each panel. The stationary 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. The stationary mean is 1.2480% (the observed stationary means are 0.8185% for consumption; 0.7336% for business income and 1.0921% for dividends). The half-life for high mean state is 46 quarters; for the low mean state it is 2 quarters. The correlation matrices for $\Delta \log C_t$, $\Delta \log B_t$ and $\Delta \log D_t$ is

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

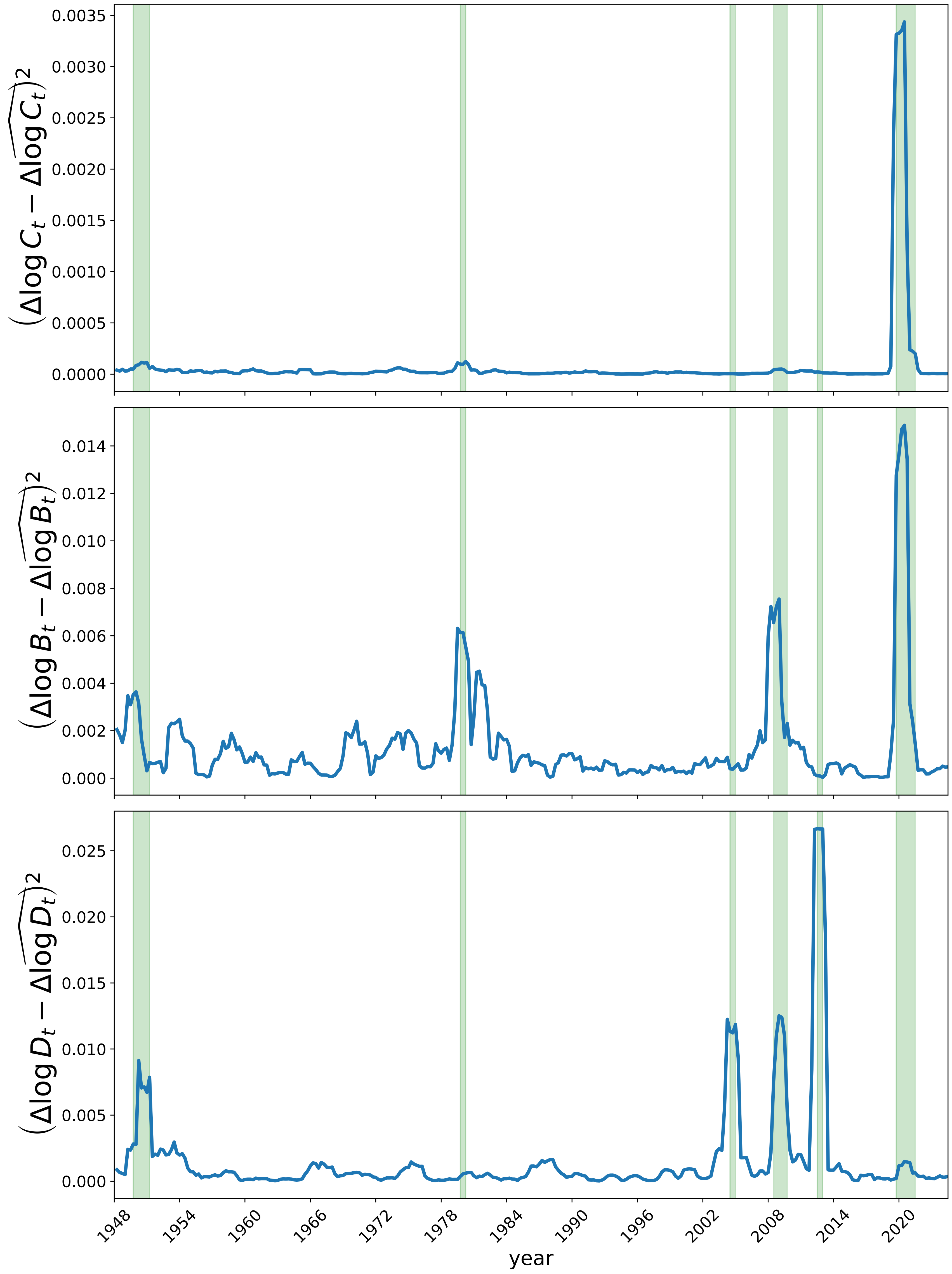


Fig. 5.5 The residual square of log growth rates of consumption, business income and dividends are shown as solid lines in each panel. Five-quarter window moving average is applied. The median smoothed probability of being in the high-volatility regime is shown in shaded green. The half-life for the low volatility state is 26 quarters; for the high volatility state it is 2 quarters.

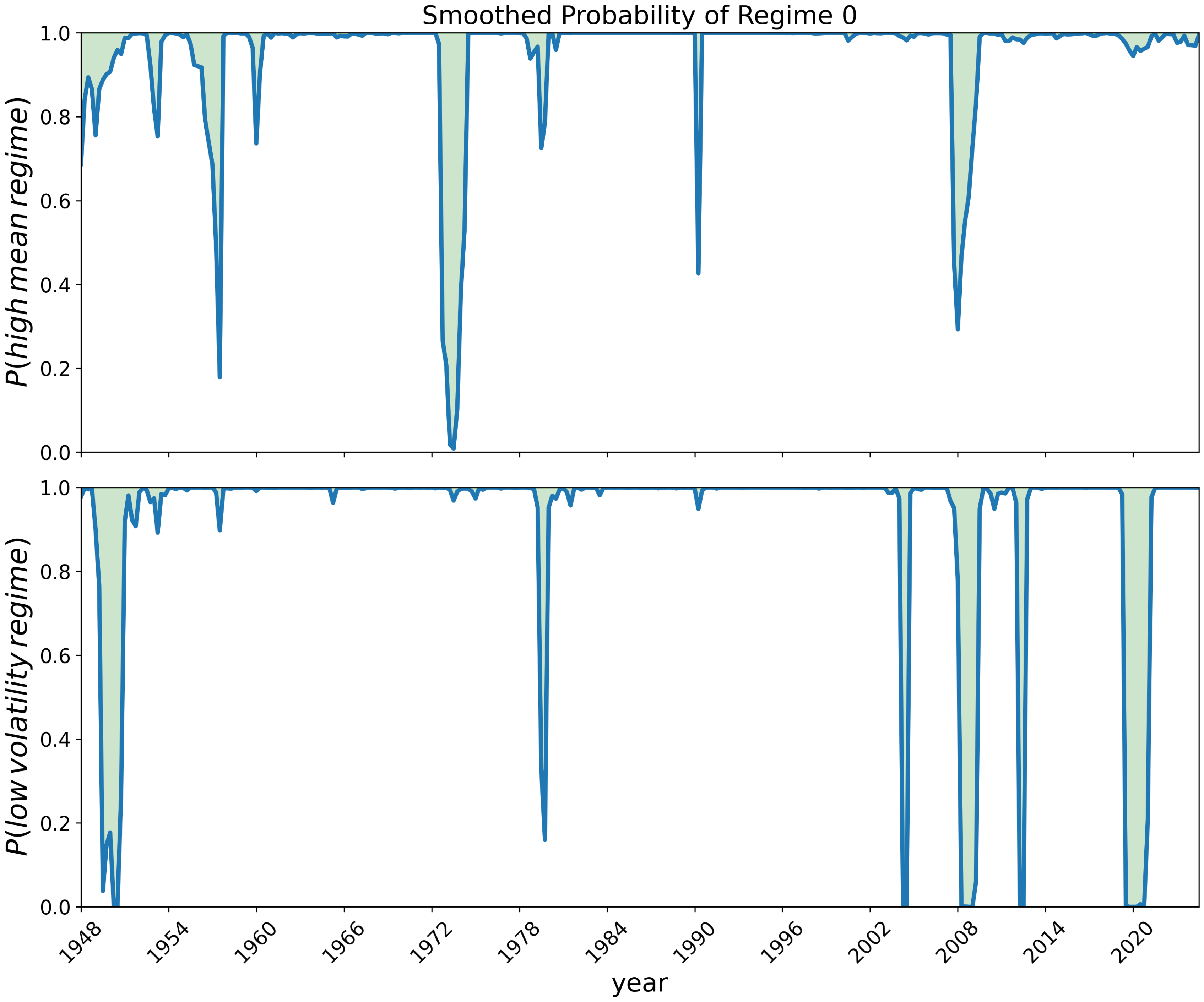
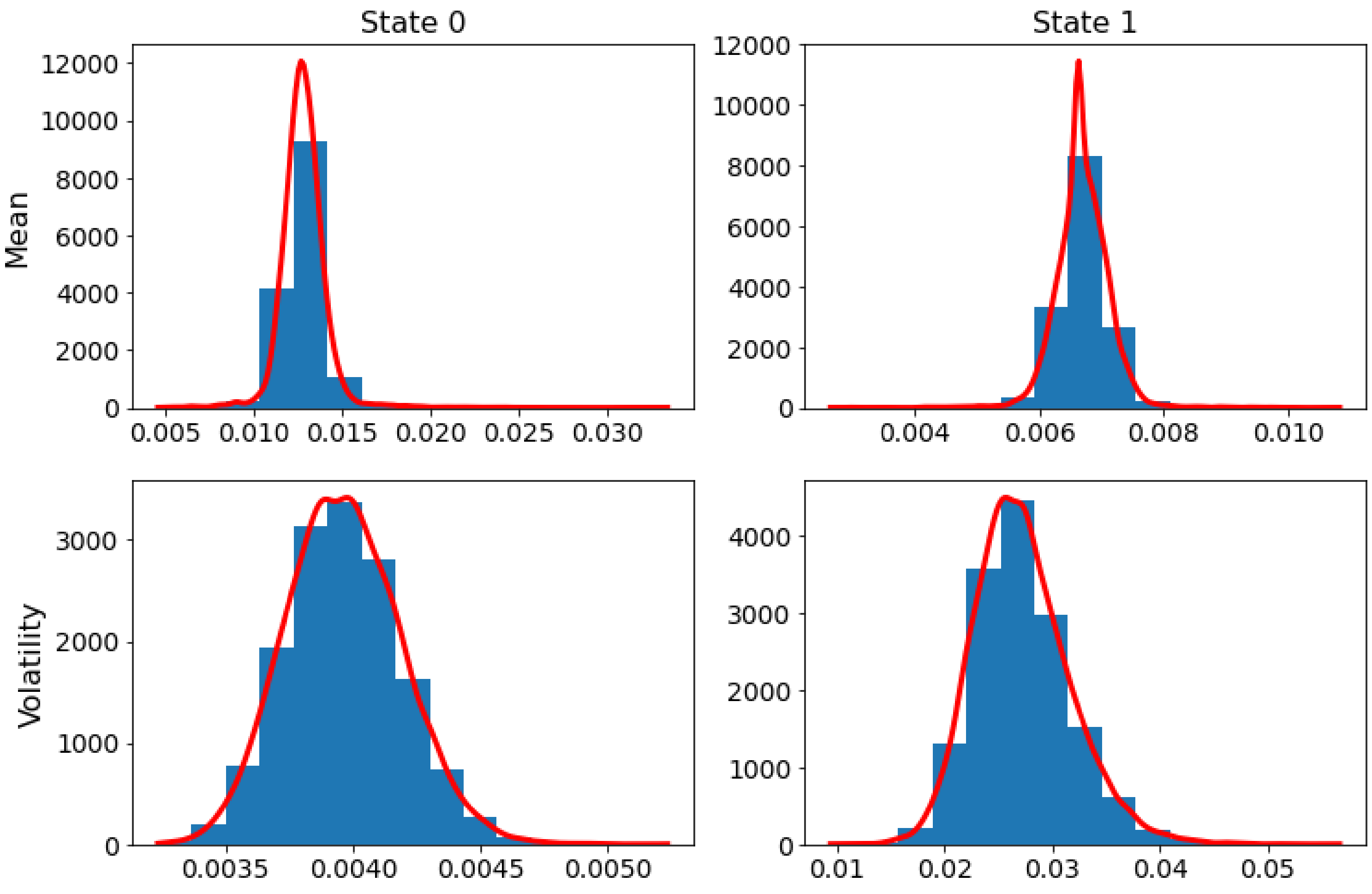


Fig. 5.6 Smoothed probabilities of being in the high-mean and low-volatility regimes.



> *Fig. 5.7* 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.

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

[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 as Kalman filtering representations post-multiply by regressors.

[5] The logarithm of time j component of L_t is evidently
$$\log \psi(Z_j \mid H + D\overline{X}_{j-1}) = -.5m \log(2\pi) - .5 \log \det(\Omega_{j-1})$$
$$-.5(Z_j - H - D\overline{X}_{j-1})'\Omega_{j-1}^{-1}(Z_j - H - D\overline{X}_{j-1})$$

[6] This stochastic process is not ergodic, being a mixture of statistical models like those described by Proposition 1.3. In the present setting, conditioning on invariant events means knowing parameters, an assumption incompatible with posing a statistical learning problem.

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

[8] A decision-maker who does not know the underlying parameters in the matrices A, B, D, F, H 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.

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

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

[11] 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.

[12] 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 require additional numerical computation.

[13] 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.