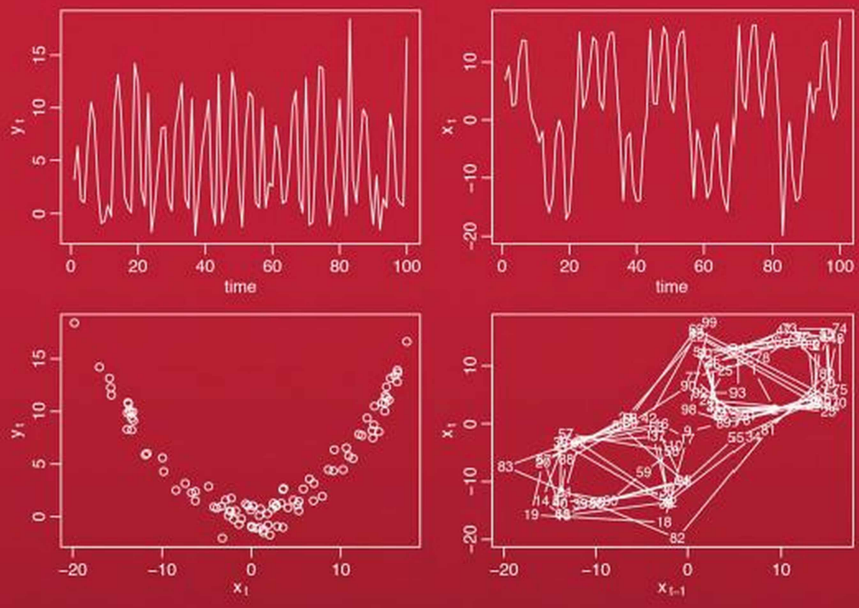
Preview Chapters 3, 9, and 10

Texts in Statistical Science

Nonlinear Time Series

Theory, Methods, and Applications with R Examples



Randal Douc
Eric Moulines
David S. Stoffer

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Texts in Statistical Science

Nonlinear Time Series

Theory, Methods, and Applications with R Examples

Randal Douc
Telecom SudParis
Evry, France

Eric Moulines
Telecom ParisTech
Paris, France

David Stoffer
University of Pittsburgh
Pennsylvania, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

Chapter 3

Beyond Linear Models

The goal of this chapter is to to provide a cursory introduction to nonlinear processes and models that may be used for data analysis. We motivate the need for nonlinear and non-Gaussian models through real data examples, discussing why there is a need for such models. We give some examples that may help explain some of the similarities seen in nonlinear or non-Gaussian process from many different disciplines. Then we exhibit some of the models used to analyze such processes and briefly discuss their properties. Our intention is not to be exhaustive in covering these topics, but rather to give a sampling of various situations and approaches to modeling nonlinear and non-Gaussian processes.

As discussed in Chapter 1, the main goal of classical time series is the analysis of the second order structure of stationary processes. This structure is fully determined by the autocovariance or autocorrelation functions, or alternatively by the associated spectral measure. The second order structure of the process fully determines the structure of stationary Gaussian processes. It is known from the Wold decomposition (see Theorem 1.25) that a regular second order stationary process $\{X_t,\, t \in \mathbb{Z}\}$ may be represented as

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}\,, \tag{3.1}$$

where $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ and $\{Z_t,\, t \in \mathbb{Z}\} \sim \mathrm{WN}(0, \sigma_z^2)$ is white noise. In this case, the spectral measure of the process $\{X_t,\, t \in \mathbb{Z}\}$ has a density $f_x(\omega) = \frac{\sigma_z^2}{2\pi} \left|\psi(\mathrm{e}^{-\mathrm{i}\omega})\right|^2$, where $\psi(\mathrm{e}^{-\mathrm{i}\omega})$ is the transfer function associated to the impulse response $\{\psi_j,\, j \in \mathbb{N}\}$; see Example 1.33. If we are only interested in the second order structure, $\{X_t,\, t \in \mathbb{Z}\}$ is equivalent to the (strong sense) causal linear process $\{\tilde{X}_t,\, t \in \mathbb{Z}\}$ given by

$$\tilde{X}_t = \sum_{j=0}^{\infty} \psi_j \tilde{Z}_{t-j}\,, \tag{3.2}$$

where $\{\tilde{Z}_t,\, t \in \mathbb{Z}\}$ is strong (i.i.d.) white noise with variance σ_z^2. The structure of a linear process is therefore intimately related to the properties of causal linear systems.

For simplicity, assume that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. First, if the input of a linear system is a sine wave of pulsation ω_0, i.e., $Z_t = A\cos(\omega_0 t + \varphi)$, then the output, $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, is a sine wave of same frequency ω_0 but with the amplitude scaled

by $|\psi(e^{-i\omega_0})|$ and phase shifted by $\arg[\psi(e^{-i\omega_0})]$; see Exercise 3.1. This property is typically lost in nonlinear systems. If we input a sine wave into a nonlinear system, then the output (provided it is well defined) contains not only a component at the fundamental frequency ω_0 but also components at the harmonics, i.e., multiples of the fundamental frequencies $2\omega_0, 3\omega_0$, and so on. Second, a linear process satisfies a superposition principle, i.e., if we input a sum $\{Z_t^{(1)},\ t \in \mathbb{Z}\}$ and $\{Z_t^{(2)},\ t \in \mathbb{Z}\}$ into a linear system, then the output will be the sum $X_t = X_t^{(1)} + X_t^{(2)}$, where $X_t^{(i)} = \sum_{j=0}^{\infty} \psi_j Z_{t-j}^{(i)}$, $i = 1, 2$. This property clearly extends to an arbitrary number of components and explains why the process $\{X_t,\ t \in \mathbb{Z}\}$ may be represented as

$$X_t = \int_{-\pi}^{\pi} e^{i\omega t}\, \psi(e^{-i\omega})\, dZ(\omega)\ ,$$

where $Z(\omega)$ is the spectral field associated with $\{Z_t,\ t \in \mathbb{Z}\}$, i.e., $Z_t = \int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega)$. In the linear world, there is a kind of natural duality between the time-domain and the frequency-domain, since linear transformation preserves the frequencies and obeys a superposition principle. In the nonlinear world, there is no such thing as *impulse response* or *transfer function* and there is no longer a nice correspondence between time and frequency domains.

When a linear process is invertible, the innovations $\{Z_t,\ t \in \mathbb{Z}\}$ can be expressed in terms of the process $\{X_t,\ t \in \mathbb{Z}\}$. If the process is causally invertible, then there exists a sequence $\{\pi_j,\ j \in \mathbb{N}\}$ such that

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}\,;$$

see Definition 1.30. If $Z_t \sim \text{iid}\,(0, \sigma_z^2)$, then, according to (3.1), for any $t \in \mathbb{Z}$, X_t is a (linear function) of Z_s, for $s \le t$. Hence, X_t is $\mathcal{F}_t^Z$ measurable, where $\mathcal{F}_t^Z = \sigma(Z_s, s \le t)$ is the past history at time t of the process $\{Z_t,\ t \in \mathbb{Z}\}$; this implies that $\mathcal{F}_t^X \subset \mathcal{F}_t^Z$, where $\mathcal{F}_t^X = \sigma(X_s, s \le t)$. Thus, for any $t \in \mathbb{Z}$, Z_t is independent of $\mathcal{F}_{t-1}^X \subset \mathcal{F}_{t-1}^Z$. Therefore $\mathbb{E}\left[Z_t \,\middle|\, \mathcal{F}_{t-1}^X\right] = 0$ showing that the conditional expectation of X_t of the process given the past $\mathcal{F}_{t-1}^X$ can be linearly expressed as a function of its past values, the optimal predictor is linear.

3.1 Nonlinear non-Gaussian data

In Chapter 1, we indicated that linear Gaussian models can handle a broad range of problems, but that it is often necessary to go beyond these models. One might say that in the linear Gaussian world, "$\infty = 2$". In Section 1.5, however, we argued that the annual sunspot numbers were not a linear Gaussian process because the data are not time reversible, i.e., the data plotted in time order as $X_{1:n} = \{X_1, X_2, \ldots, X_n\}$ does not look the same as the data plotted in reverse time order $X_{n:1} = \{X_n, X_{n-1}, \ldots, X_1\}$. In that section we pointed out that such an occurrence will not happen for a linear Gaussian process because, in that case, $X_{1:n}$ and $X_{n:1}$ have the same distribution.

Trying to model something that is *not* linear or *not* Gaussian might seem like a

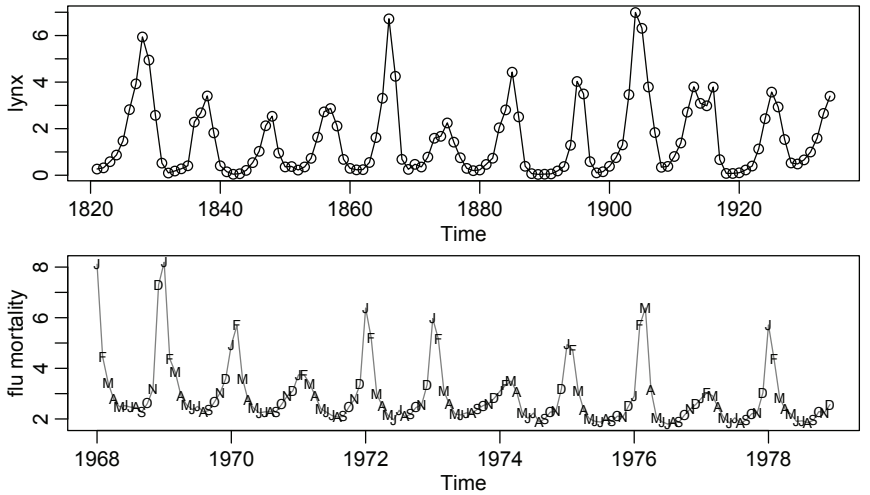


Figure 3.1 *Top: Annual numbers (÷1000) of lynx trappings for 1821–1934 in Canada. Bottom: Monthly rates of pneumonia and influenza deaths in the United States for 1968–1978.*

daunting task at first, because how does one model something in the negative? That is, how do you model a process that is *not* linear or *not* Gaussian; there are so many different ways to be not something. Fortunately, there are patterns of nonlinearity and non-Gaussianity that are common to many processes. In recognizing these similarities, we are able to develop general strategies and models to cover a wide range of processes observed in diverse disciplines. The following examples will help in explaining some of the commonalities of processes that are in the complement of linear Gaussian processes.

Example 3.1 (Sunspots, felines, and flu). The irreversiblity of the sunspot data (Figure 1.1) is a trait that is observed in a variety of processes. For example, the data shown in Figure 3.1 are typical of predator-prey relationships; the data are the annual numbers of lynx trappings for 1821–1934 in Canada; see Campbell and Waker (1977). Such relationships are often modeled by the Lotka-Volterra equations, which are a pair of simple nonlinear differential equations used to describe the interaction between the size of predator and prey populations; e.g., see Edelstein-Keshet (2005, Ch. 6). Note that, as opposed to the sunspot data set, the lynx data tend to increase slowly to a peak and then decline quickly to a trough ($\nearrow\!\!\downarrow$). Another process that has a similar pattern is the influenza data also shown in Figure 3.1. These data are taken from Shumway and Stoffer (2011) and are monthly pneumonia and influenza deaths per 1,000 people in the United States for 11 years. ◇

Example 3.2 (EEG, S&P500, and explosions). The data shown in the top of Figure 3.2 are a single channel EEG signal taken from the epileptogenic zone of a subject with epilepsy, but during a seizure free interval of 23.6 seconds, and is series (d) shown in Andrzejak et al. (2001, Figure 3). The bottom of Figure 3.2 shows the

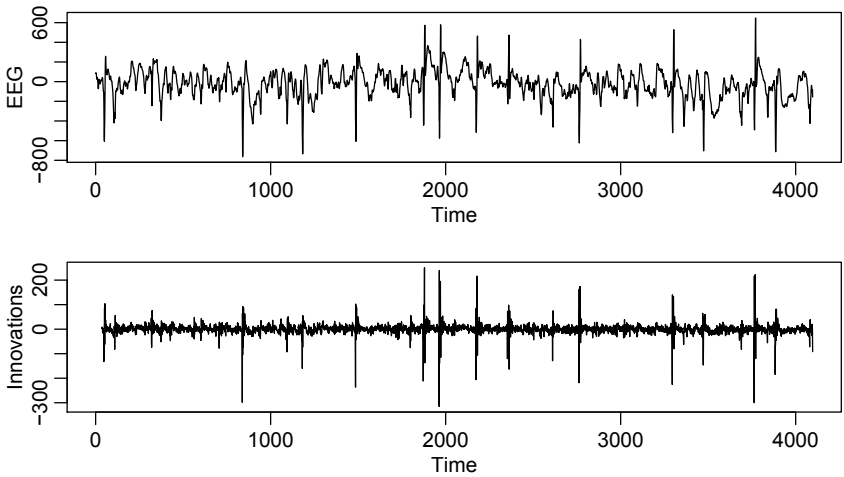


Figure 3.2 *Top: A single channel EEG signal taken from the epileptogenic zone of a subject with epilepsy during a seizure free interval of 23.6 seconds; see Andrzejak et al. (2001). Bottom: The innovations after removal of the signal using an autoregression based on AIC.*

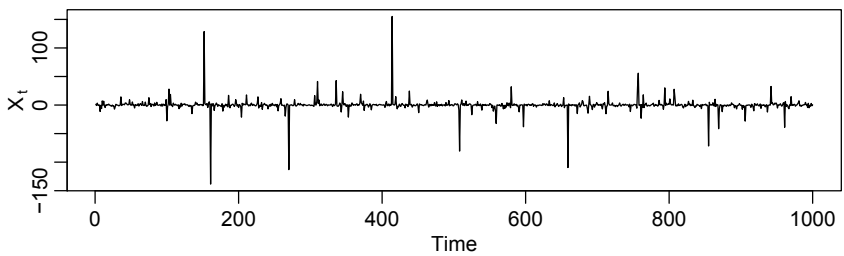


Figure 3.3: *Simulated infinite variance series generated as i.i.d. standard Cauchy errors.*

innovations (residuals) after the signal has been removed based on fitting an AR(p) using AIC to determine the order.

Due to the large spikes in the EEG trace, it is apparent that the data are not normal. In fact, the innovations in Figure 3.2 look more like the simulated infinite variance noise series shown in Figure 3.3, which were generated from i.i.d. standard Cauchy errors.

Moreover, the left side of Figure 3.4 shows the sample ACF of the EEG innovations. The fact that the values are small indicates that the innovations are white noise. However, the right side of Figure 3.4 shows the sample ACF of the squared EEG innovations, where we clearly see significant autocorrelation. Thus, while the innovations appear to be white, the are clearly not independent, and hence not Gaussian.

The behavior seen in the EEG trace is not particular to EEGs, and in fact is quite common in financial data. For example, the top of Figure 3.5 shows the daily

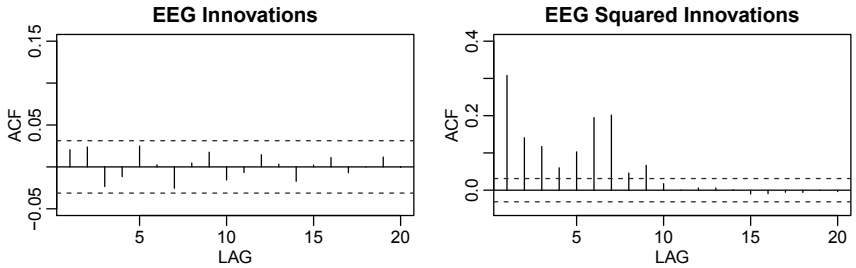


Figure 3.4 *The sample ACF of the EEG innovations (left) and the squared innovations (right); the EEG innovations series is shown in Figure 3.2.*

returns[1] of the S&P 500 from 2001 to the end of 2011. There, the data exhibit what is called volatility clusters, that is, regions of highly volatile periods tend to be clustered together. As with the EEG series, the data have very little autocorrelation, but the squares of the returns have significant autocorrelation; this is demonstrated in the bottom part of Figure 3.5.

Figure 3.6 shows the two phases or arrivals (the P-wave and then the S-wave) along the surface of an explosion at a seismic recording station. The recording instruments are observing earthquakes and mining explosions with the general problem of interest being to distinguish between waveforms generated by earthquakes and those generated by explosions. This distinction is key in enforcing a comprehensive nuclear test ban treaty. The general problem of interest, which is discussed in more detail in Shumway and Stoffer (2011, Chapter 7), is in discriminating between waveforms generated by earthquakes and those generated by explosions. The data behave in a similar fashion to the EEG trace and the S&P500 returns; see Exercise 3.2. ◇

Example 3.3 (Great discoveries and polio). Another situation in which normality is an unreasonable assumption is when the data are discrete-valued and small. Two such process are the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959, shown in Figure 3.7, and the number of poliomyelitis cases reported to the U.S. Centers for Disease Control for the years 1970 to 1983, displayed in Figure 3.8.

These two unrelated processes have striking similarities in that the marginal distributions appear to be Possion, or more specifically generalized Poisson or negative binomial (which are a mixture of Poissons; this is often used to account for over- or under-dispersion, where the mean and the standard deviation are not equal; e.g., see Joe and Zhu (2005). Moreover, we see that the ACFs of each process seems to imply a simple autocorrelation structure, which might be modeled as a simple non-Gaussian AR(1) type of model.

[1] If X_t is the price of an asset at time t, the *return* or *growth rate* of that asset, at time t, is $R_t = (X_t - X_{t-1})/X_{t-1}$. Alternately, we may write $X_t = (1+R_t)X_{t-1}$, or $\nabla \ln X_t = \ln(1+R_t)$. But $\ln(1+R_t) = R_t - R_t^2/2 + R_t^3/3 - \cdots$ for $-1 < R_t \leq 1$. If R_t is a small percentage, then the higher order terms are negligible, and $\ln(1+R_t) \approx R_t$. It is easier to program $\nabla \ln X_t$, so this is often used instead of calculating R_t directly. Although it is a misnomer, $\nabla \ln X_t$ is often called the *log-return*.

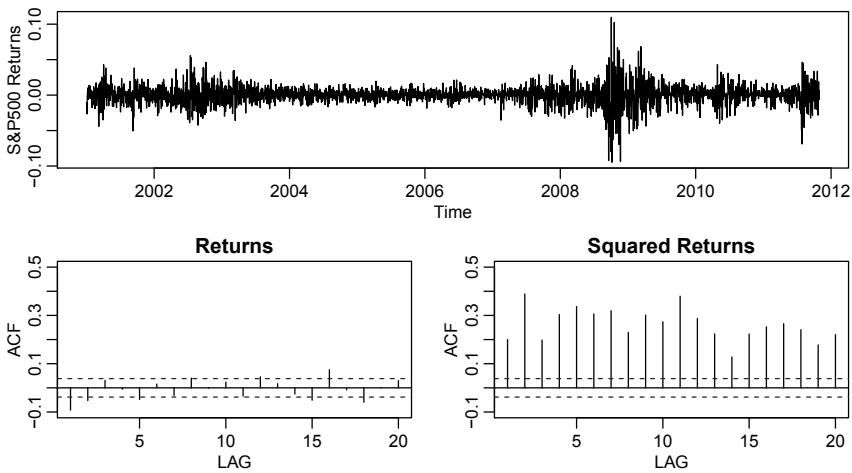


Figure 3.5 *Top: Daily returns of the S&P 500 from 2001 to the end of 2011. Bottom: The sample ACF of the returns and of the squared returns.*

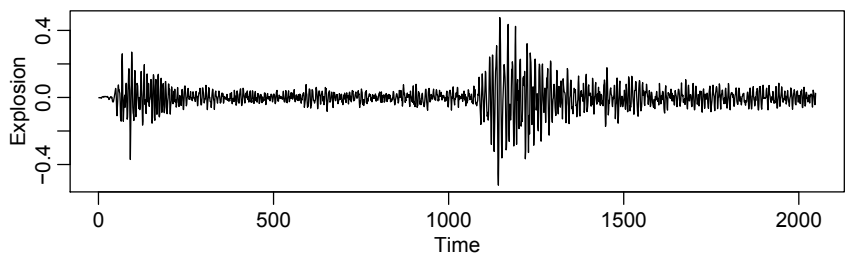


Figure 3.6 *Two phases or arrivals along the surface of an explosion at a seismic recording station. Compressional waves, also known as primary or P-waves, travel fastest, at speeds between 1.5 and 8 kilometers per second in the Earth's crust. Shear waves, also known as secondary or S-waves, travel more slowly, usually at 60% to 70% of the speed of P-waves.*

These data sets make it clear that, in addition to the problems discussed in the previous examples, there is a need to have non-Gaussian time series models that can take into account processes that produce discrete-valued observations that may have an autocorrelation structure similar to what is seen in ARMA models. The data set `discoveries` is an R data set that was taken from McNeil (1977). The polio data set is taken from Zeger (1988) and can be found in the R package `gamlss.data`. We test if the marginal number of reported polio cases is Poisson or negative binomial using `goodfit` from the R package `vcd`.

```
>  summary(goodfit(as.integer(polio)))              # Poisson
 Goodness-of-fit test for poisson distribution
                      X^2 df      P(> X^2)
  Likelihood Ratio 78.04415  9  3.949539e-13
```

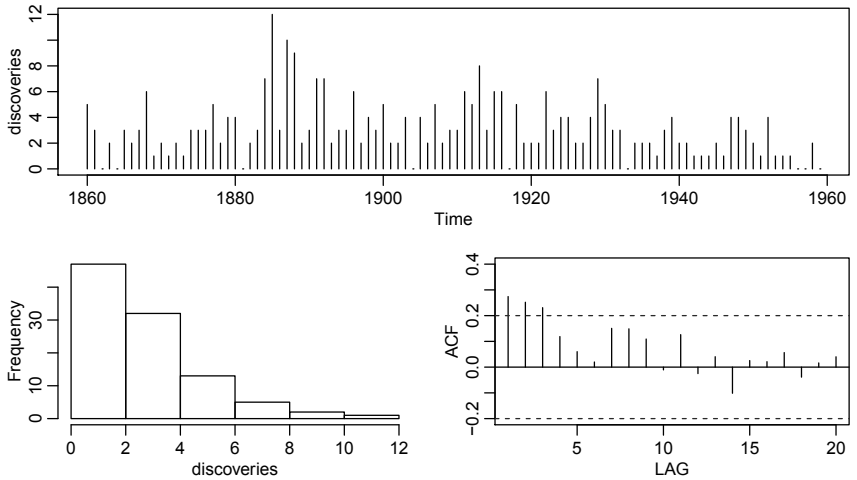


Figure 3.7 *The numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959. Source: The World Almanac and Book of Facts, 1975 Edition.*

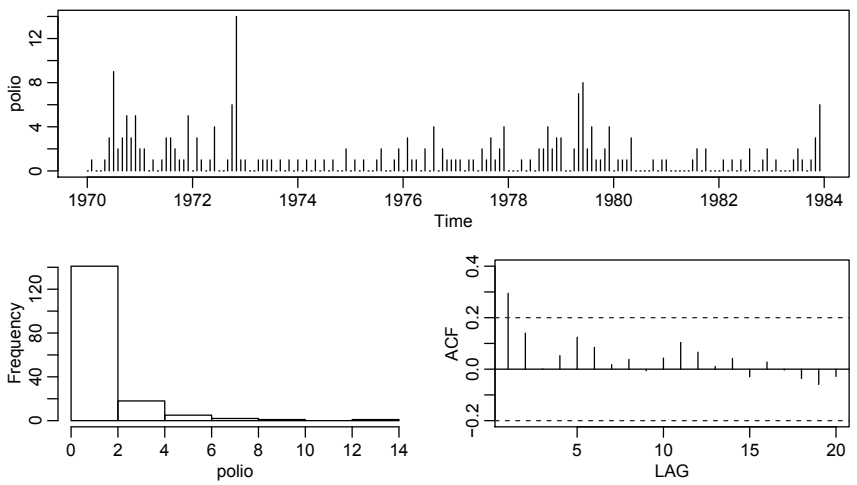


Figure 3.8 *Poliomyelitis cases reported to the U.S. Centers for Disease Control for the years 1970 to 1983.*

```
> summary(goodfit(as.integer(polio), "nbinomial")) # Neg Binomial
  Goodness-of-fit test for nbinomial distribution
                       X^2  df  P(> X^2)
  Likelihood Ratio 12.8793   8  0.116070
```

Clearly the Poisson distribution does not fit, while the negative binomial appears to be satisfactory. ◇

The essential points of Examples 3.1, 3.2 and 3.3 are that (i) linear or Gaussian

time series models are limited, and not all situations can be handled by such models even after transformation; (ii) similar types of departures from the linear or Gaussian process are observed in data from many diverse fields, and in varied and unrelated situations. Hence, the development of general nonlinear models has played a prominent role in the field of time series for decades.

3.2 Volterra series expansion

A natural idea for going beyond the linear structure of (3.1) is to consider the following Volterra series. An M-th order Volterra process is given by

$$X_t = \sum_{i=1}^{M} \sum_{m_1=0}^{\infty} \cdots \sum_{m_i=0}^{\infty} \psi^{(i)}_{m_1,\ldots,m_i} \prod_{j=1}^{i} Z_{t-m_j} \,, \tag{3.3}$$

where $\{Z_t,\, t \in \mathbb{Z}\}$ is an i.i.d. sequence. The coefficients $\{\psi^{(i)}_{m_1,\ldots,m_i},\, (m_1,\ldots,m_i) \in \mathbb{N}^i\}$ are the coefficients determining the i-th order Volterra kernel. For simplicity, it is assumed that $\sum |\psi^{(i)}_{m_1,\ldots,m_i}| < \infty$, but of course, this assumption may be weakened. These types of expansions were first considered by Wiener (1958), where the concern was with the case when both the input $\{Z_t,\, t \in \mathbb{Z}\}$ and the output $\{X_t,\, t \in \mathbb{Z}\}$ were observable. In the context of time series, only $\{X_t,\, t \in \mathbb{Z}\}$ is observable. The first term $H_1[Z_s,\, s \le t] = \sum_{m=0}^{\infty} \psi^{(1)}_m Z_{t-m}$ is a linear model. The second term $H_2[Z_s,\, s \le t] = \sum_{m_1,m_2=0}^{\infty} \psi^{(2)}_{m_1,m_2} Z_{t-m_1} Z_{t-m_2}$ is a linear combination of quadratic terms. The higher order terms may be called the cubic component, the quartic component, and so on. This expansion might be seen as the M-th order principal part of the multidimensional Taylor expansion of the generic nonlinear model $X_t = g(Z_s,\, s \le t)$ (assuming that the operator g is analytic). By the Stone-Weierstrass Theorem, any continuous function $g\colon (z_1,\ldots,z_m) \mapsto g(z_1,\ldots,z_m)$ on a compact set of $\mathbb{R}^m$ can be approximated with an arbitrary precision in the topology of uniform convergence by a polynomial $p(z_1,\ldots,z_m)$. Hence, it is not difficult to guess that, under rather weak assumptions, an arbitrary finite memory nonlinear system $X_t = g(Z_t, Z_{t-1},\ldots,Z_{t-m+1})$ can be approximated arbitrarily well by a Volterra series expansion. Infinite memory processes $X_t = g(Z_s, s \le t)$ can also be approximated arbitrarily well by a finite order Volterra series provided that the infinite memory possesses some forgetting property (roughly speaking, the influence of the infinite past should fade away in some appropriate sense).

Even simple Volterra series expansions display properties that are markedly different from linear processes. For example, consider the process $\{X_t,\, t \in \mathbb{Z}\}$ defined by,

$$X_t = Z_t + \beta Z_{t-1} Z_{t-2} \tag{3.4}$$

where $\{Z_t,\, t \in \mathbb{Z}\}$ is a strong (i.i.d.) white noise sequence with zero mean and constant variance. It follows immediately that $\{X_t,\, t \in \mathbb{Z}\}$ has zero mean, constant vari-

ance, and autocovariance function given by,

$$\mathbb{E}[X_t X_{t+h}] = \\ \mathbb{E}\left[Z_t Z_{t+h} + \beta Z_{t-1} Z_{t-2} Z_{t+h} + \beta Z_t Z_{t+h-1} Z_{t+h-2} + \beta^2 Z_{t-1} Z_{t-2} Z_{t+h-1} Z_{t+h-2}\right] .$$

For all $h \neq 0$, each of the terms on the right-hand side is zero because Z_t is a strong white noise. Thus, as far as its second order properties are concerned, $\{X_t,\, t \in \mathbb{Z}\}$ behaves just like a white noise process. However, given observations up to time t, one can clearly construct a non-trivial prediction of X_{t+h}. Specifically, if we adopt the mean square error criterion, the optimal forecast of a future observation, X_{t+h}, is its conditional expectation, $X_{t+h|t} = \mathbb{E}\left[X_{t+h} \,\middle|\, \mathcal{F}_t^X\right]$, where $\mathcal{F}_t^X = \sigma(X_s, s \leq t)$. Computing this conditional expectation is not entirely trivial as we shall see. Assume that the system is invertible, i.e., that there exists a measurable non-anticipative function such that $Z_t = g(X_s, s \leq t)$. In this case, for any $t \in \mathbb{Z}$, Z_t belongs to $\mathcal{F}_t^X$ and therefore

$$\mathbb{E}\left[X_{t+1} \,\middle|\, \mathcal{F}_t^X\right] = \beta g(X_s, s \leq t) \times g(X_s, s \leq t-1) \quad \mathbb{P}\text{-a.s.}$$

Note that such inverse does not always exist; in this case, more delicate arguments are used to compute forecasts.

There is a substantial literature on the theoretical properties of these models, which plays an important role in nonlinear system theory. Estimating the coefficients of the Volterra kernels individually is difficult for two reasons. First, the kernels of the Volterra series are strongly dependent. A direct approach leads to the problem of simultaneously solving a strongly coupled set of nonlinear equations for the kernel coefficients. Second, the canonical representation (3.3) contains, in general, far too many parameters to estimate efficiently from a finite set of observations. To alleviate this problem, following the original suggestion by Wiener, the estimation of the Volterra kernels is generally performed by developing the coefficients on appropriately chosen orthogonal basis function, such as the Laguerre and Kautz functions or generalized orthonormal basis functions (GOBFs); see Campello et al. (2004) and the references therein. This technique requires us to assume that the coefficients in the expansion may be expressed as known functions of some relatively small number of other parameters.

3.3 Cumulants and higher-order spectra

We have seen that in the linear Gaussian world, it is sufficient to work with second-order statistics. Now, reconsider Example 1.35 where X, Y, Z are i.i.d. N(0,1) random variables with $Y = X^2 + Z$. This could be a model (appropriately parameterized) for automobile fuel consumption Y versus speed X; i.e., fuel consumption is lowest at moderate speeds, but is highest at very low and very high speeds. In that example, we saw that the BLP ($\widehat{Y} = 1$) was considerably worse than the minimum mean square predictor ($\widehat{Y} = X^2$). If, however, we consider linear prediction on $\mathcal{M} = \overline{\text{sp}}\{1, X, X^2\}$, then from Proposition 1.34, the prediction equations are

$$(i)\ \mathbb{E}[Y - P_{\mathcal{M}} Y] = 0;\ \ (ii)\ \mathbb{E}[Y - P_{\mathcal{M}} Y]X = 0;\ \ (iii)\ \mathbb{E}[Y - P_{\mathcal{M}} Y]X^2 = 0 \qquad (3.5)$$

where $P_{\mathcal{M}}Y = a + bX + cX^2$. Solving these equations will yield $a = b = 0$, and $c = 1$ (see Exercise 3.3) so that $P_{\mathcal{M}}Y = X^2$, which was also the optimal predictor $\mathbb{E}\left[Y \,\middle|\, X\right]$. The problem with the BLP in Example 1.35 was that it only considered moments up to order 2, e.g., $\mathbb{E}[YX]$ and $\mathbb{E}\left[X^2\right]$. But here, we have improved the predictor by considering slightly higher-order moments such as $\mathbb{E}\left[YX^2\right]$ and $\mathbb{E}\left[X^4\right]$.

For a collection of random variables $\{X_1, \ldots, X_k\}$, let $\varphi(\xi_1, \ldots, \xi_k) = \varphi(\xi)$ be the corresponding joint characteristic function,

$$\varphi(\xi) = \mathbb{E}\left[\exp\left\{\mathrm{i}\sum_{j=1}^{k} \xi_j X_j\right\}\right] . \tag{3.6}$$

For $r = (r_1, \ldots r_k)$, if the moments $\mu_r = \mathbb{E}\left[X_1^{r_1} \cdots X_k^{r_k}\right]$ exist up to a certain order $|r| := \sum_{j=1}^{k} r_j \leq n$, then they are the coefficients in the expansion of $\varphi(\xi)$ around zero,

$$\varphi(\xi) = \sum_{|r| \leq n} (\mathrm{i}\xi)^r \mu_r / r! + o(|\xi|^n), \tag{3.7}$$

where $r! = \prod_{j=1}^{k} r_j!$ and $\xi^r = \xi_1^{r_1} \ldots \xi_k^{r_k}$. Similarly, the joint cumulants $\boldsymbol{\kappa_r} \equiv \mathrm{cum}[X_1^{r_1} \cdots X_k^{r_k}]$ are the coefficients in the expansion of the *cumulant generating function*, defined as the logarithm of the characteristic function:

$$\ln \varphi(\xi) = \sum_{|r| \leq n} (\mathrm{i}\xi)^r \kappa_r / r! + o(|\xi|^n). \tag{3.8}$$

A special case of (3.8) is $X_j = X$ for $j = 1, \ldots, k$, in which one obtains the r-th cumulant of X. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\ln \varphi(\xi) = \mathrm{i}\mu\xi - \frac{1}{2}\sigma^2\xi^2$, so that $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, and $\kappa_r = 0$, for $r > 2$. In fact, the normal distribution is the only distribution for which this is true (i.e., there are a finite number of non-zero cumulants, Marcinkiewicz, 1939). Another interesting case is the Poisson(λ) distribution wherein $\ln \varphi(\xi) = \lambda(\mathrm{e}^{\xi} - 1)$ and consequently $\kappa_r = \lambda$ for all r.

Some special properties of cumulants are:

- The cumulant is invariant with respect to the permutations: $\mathrm{cum}(X_1, \ldots, X_k) = \mathrm{cum}(X_{\sigma(1)}, \ldots, X_{\sigma(k)})$ where σ is any permutation on $\{1, \ldots, k\}$.
- For every $(a_1, \ldots, a_k) \in \mathbb{R}^k$, $\mathrm{cum}(a_1 X_1, \ldots, a_k X_k) = a_1 \cdots a_k \, \mathrm{cum}(X_1, \ldots, X_k)$.
- The cumulant is multilinear:

$$\mathrm{cum}(X_1 + Y_1, X_2, \ldots, X_k) = \mathrm{cum}(X_1, X_2, \ldots, X_k) + \mathrm{cum}(Y_1, X_2, \ldots, X_k) .$$

- If $\{X_1, \ldots, X_k\}$ can be partitioned into two disjoint sets that are independent of each other, then $\mathrm{cum}(X_1, \ldots, X_k) = 0$.
- If $\{X_1, \ldots, X_k\}$ and $\{Y_1, \ldots, Y_k\}$ are independent, then $\mathrm{cum}(X_1 + Y_1, \ldots X_k + Y_k) = \mathrm{cum}(X_1, \ldots, X_k) + \mathrm{cum}(Y_1, \ldots, Y_k)$.
- $\mathrm{cum}(X) = \mathbb{E}[X]$ and $\mathrm{cum}(X, Y) = \mathrm{Cov}(X, Y)$.
- $\mathrm{cum}(X, Y, Z) = \mathbb{E}[XYZ]$ if the means of the random variables are zero.

Further information on the properties of cumulants may be found in Brillinger (2001), Leonov and Shiryaev (1959), and Rosenblatt (1983).

Using Theorem 1.23 for a zero-mean stationary time series $\{X_t, t \in \mathbb{Z}\}$, and defining $\kappa_x(r) = \text{cum}(X_{t+r}, X_t) = \gamma_x(r)$ we may write

$$\begin{aligned} \kappa_x(r) = \mathbb{E}[X_{t+r}X_t] &= \int_{-\pi}^{\pi}\int e^{i(t+r)\omega} e^{it\lambda}\, \mathbb{E}[dZ(\omega)\, dZ(\lambda)] \\ &= \int_{-\pi}^{\pi}\int e^{it(\omega+\lambda)} e^{ir\omega}\, \mathbb{E}[dZ(\omega)\, dZ(\lambda)]. \end{aligned} \tag{3.9}$$

Because the left-hand side of (3.9) does not depend on t, the right-hand side cannot depend on t. Thus, it must be the case that $\mathbb{E}[dZ(\omega)\, dZ(\lambda)] = 0$ unless $\lambda = -\omega$, and consequently, as pointed out in Theorem 1.23, $\mathbb{E}[dZ(-\omega)\, dZ(\omega)] = \mathbb{E}\left[|dZ(\omega)|^2\right] = dF(\omega)$. If $\kappa_x(r) = \gamma_x(r)$ is absolutely summable, then by Proposition 1.19, $dF(\omega) = f(\omega)d\omega$, where $f(\omega)$ is the spectral density of the process.

This concept may be applied to higher order moments. For example, suppose the cumulant $\kappa_x(r_1, r_2) = \text{cum}(X_{t+r_1}, X_{t+r_2}, X_t) = \mathbb{E}[X_{t+r_1} X_{t+r_2} X_t]$ exists and does not depend on t. Then,

$$\begin{aligned} \kappa_x(r_1, r_2) &= \int_{-\pi}^{\pi}\int\int e^{i(t+r_1)\omega_1} e^{i(t+r_2)\omega_2} e^{it\lambda}\, \mathbb{E}[dZ(\omega_1)\, dZ(\omega_2)\, dZ(\lambda)] \\ &= \int_{-\pi}^{\pi}\int\int e^{it(\omega_1+\omega_2+\lambda)} e^{ir_1\omega_1} e^{ir_2\omega_2}\, \mathbb{E}[dZ(\omega_1)\, dZ(\omega_2)\, dZ(\lambda)]\ . \end{aligned} \tag{3.10}$$

Because $\kappa_x(r_1, r_2)$ does not depend on t, it must be that $\mathbb{E}[dZ(\omega_1)\, dZ(\omega_2)\, dZ(\lambda)] = 0$ unless $\omega_1 + \omega_2 + \lambda = 0$. Consequently, we may write

$$\kappa_x(r_1, r_2) = \int_{-\pi}^{\pi}\int e^{ir_1\omega_1} e^{ir_2\omega_2} \mathbb{E}[dZ(\omega_1)\, dZ(\omega_2)\, dZ(-[\omega_1 + \omega_2])]\,. \tag{3.11}$$

Hence, the *bispectral distribution* may be defined as

$$dF(\omega_1, \omega_2) = \mathbb{E}[dZ(\omega_1)\, dZ(\omega_2)\, dZ(-[\omega_1 + \omega_2])]\,. \tag{3.12}$$

Following Proposition 1.19, under absolute summability conditions, we may define the bispectral density or *bispectrum* as $f(\omega_1, \omega_2)$ where

$$\kappa_x(r_1, r_2) = \int_{-\pi}^{\pi}\int e^{ir_1\omega_1} e^{ir_2\omega_2} f(\omega_1, \omega_2)\, d\omega_1 d\omega_2 \tag{3.13}$$

and

$$f(\omega_1, \omega_2) = (2\pi)^{-2} \sum_{-\infty < r_1, r_2 < \infty} \sum \kappa_x(r_1, r_2)\, e^{-ir_1\omega_1} e^{-ir_2\omega_2}. \tag{3.14}$$

If $\{X_t\}$ is Gaussian, then $\kappa_x(r_1, r_2) = 0$ for all $(r_1, r_2) \in \mathbb{Z}^2$, and thus the bispectrum $f(\omega_1, \omega_2) \equiv 0$ for $(\omega_1, \omega_2) \in [-\pi, \pi]^2$. Consequently, tests of linearity and Gaussianity may rely on the bispectrum; see Exercise 3.4.

Finally, higher order cumulant spectra may be defined analogously to the bispectrum. That is, let $\kappa_x(r) = \kappa_x(r_1, \ldots, r_k) = \mathrm{cum}(X_{t+r_1}, \ldots, X_{t+r_k}, X_t)$ and assume that

$$\sum_{-\infty<r<\infty} \cdots \sum |\kappa_x(r)| < \infty.$$

Then, the $k+1$-*st order cumulant spectrum* is defined by

$$f_x(\omega_1, \ldots, \omega_k) = \sum_{-\infty<r<\infty} \cdots \sum \kappa_x(r) \exp\Big\{-\mathrm{i} \sum_{j=1}^{k} r_j \omega_j\Big\}. \tag{3.15}$$

We note that higher-order spectra are generally complex-valued. The inverse relationship is,

$$\kappa_x(r_1, \ldots, r_k) = \int_{-\pi}^{\pi} \cdots \int f_x(\omega_1, \ldots, \omega_k) \exp\Big\{\mathrm{i} \sum_{j=1}^{k} r_j \omega_j\Big\} \mathrm{d}\omega_1 \ldots \mathrm{d}\omega_k. \tag{3.16}$$

For further details, the reader is referred to Brillinger (1965, 2001) and Rosenblatt (1983).

3.4 Bilinear models

In Example 3.2, we saw that time series data may exhibit simple or no autocorrelation structure, but still be highly dependent; this dependence was seen in the squares of the observations. For example, the EEG innovations shown in Figure 3.2 and the autocorrelation structure shown in Figure 3.4 suggest that the innovations are white (i.e., uncorrelated noise), but the squared innovations indicate that there is still a dependence structure. This is a common occurrence, especially in financial time series. For example, in Figure 3.5, the daily returns of the S&P 500 exhibit obvious dependence, whereas the sample ACF indicates only a small correlation structure. The squares of the process, however, indicate a strong correlation structure is present. Exercise 3.2 explores the fact that the innovations of the explosion series shown in Figure 3.6 also have this property.

One early exploration of models for this type of behavior was the bilinear model developed for the statistical analysis of time series by Granger and Andersen (1978) and explored further by Subba Rao (1981) and others. The basic idea is that of using higher order terms of a Volterra expansion of the noise. For example, think of an ARMA model, $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-i}$, as a first order (linear) approximation of the Volterra expansion (3.3). In the nonlinear case, it seems reasonable to, at least, use a second order approximation, and that is the idea behind the bilinear model,

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{P} \sum_{j=1}^{Q} b_{ij} X_{t-i} Z_{t-j} + Z_t, \tag{3.17}$$

which is denoted as BL(p,q,P,Q). In this case, the process $\{X_t,\, t\in\mathbb{Z}\}$ is said to be causal (or nonanticipative) if for all $t\in\mathbb{Z}$, X_t is measurable with respect to $\mathcal{F}_t^Z = \sigma(Z_s, s\le t)$, i.e., X_t can be expressed as a measurable (but nonlinear) function of Z_s, for $s\le t$.

While the model appears to be a simple extension of the ARMA model to include nonlinearity, the existence, stationarity and invertibility of bilinear processes is a delicate topic. In fact, the model is too complicated to be examined in full generality. Consequently, investigations such as Granger and Andersen (1978), Pham and Tran (1981), Subba Rao (1981) and Subba Rao and Gabr (1984) focus on restricted models. Although the model has the desired property of exhibiting ARMA-type correlation structure for X_t with dependent innovations, the problem of examining the model in its generality seems to be the reason the model lost favor as a procedure to analyze nonlinear time series. Priestley (1988) has a nice discussion of the model and its history.

As a simple illustration of the properties of the model, consider the following bilinear model, BL(0,0,2,1),

$$X_t = b\,Z_{t-1}\,X_{t-2} + Z_t, \tag{3.18}$$

where $Z_t\sim \text{iid}\,(0,\sigma_z^2)$ with Z_t independent of X_s for $s<t$. If we assume that $\mathbb{E}\left[Z_t^4\right] < \infty$ and $b^2\sigma_z^2<1$, we can show that X_t is stationary using the techniques of Chapter 4; see Exercise 4.11. Let $\mathcal{F}_t=\sigma(X_t,X_{t-1},\dots)$, then direct calculation (see Exercise 3.6) establishes that $\mathbb{E}[X_t]=0$ and

$$\mathbb{E}\left[X_t\,X_{t-h}\,\middle|\,\mathcal{F}_{t-2}\right] = \begin{cases} \sigma_z^2 + b^2\,\sigma_z^2\,X_{t-2}^2, & h=0,\\ b\,\sigma_z^2\,X_{t-2}, & h=1,\\ 0, & h\ge 2, \end{cases} \tag{3.19}$$

with probability one. From these facts we can establish that X_t is white noise, but X_t^2 is predictable from its history. Such a model could be used to describe the innovations of the EEG data set shown in Figure 3.2. Recall Figure 3.4 where the innovations are white, but the squares of the innovations are correlated.

3.5 Conditionally heteroscedastic models

An autoregressive conditional heteroscedastic model of order p, ARCH(p), is defined as

$$X_t = \sigma_t\,\varepsilon_t\,, \tag{3.20}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2\,, \tag{3.21}$$

where the coefficients $\alpha_j\ge 0$, $j\in\{0,\dots,p\}$ are non-negative and $\varepsilon_t\sim\text{iid}\,(0,1)$ is the driving noise. If the driving noise is assumed to be Gaussian, the model implies that the conditional distribution of X_t given $X_{t-1},\dots,X_{t-p}$ is Gaussian,

$$X_t\,\big|\,X_{t-1},\dots,X_{t-p} \sim \text{N}(0,\alpha_0+\alpha_1X_{t-1}^2+\cdots+\alpha_pX_{t-p}^2). \tag{3.22}$$

Often, the driving noise is not normal and other distributions, such as the t-distribution, are used to model the noise.

To explore the properties of the model, first define $\mathcal{F}_t^X = \sigma(X_s, s \le t)$. In general, we are mainly interested in causal (or nonanticipative) solutions in which X_t is measurable with respect to $\mathcal{F}_t^\varepsilon = \sigma(\varepsilon_s, s \le t)$. This implies that $\mathcal{F}_t^X \subseteq \mathcal{F}_t^\varepsilon$ for all $t \in \mathbb{Z}$. Assume that the parameters $\{\alpha_i; i = 0, \ldots, p\}$ are chosen in such a way that there exists a nonanticipative second-order stationary solution to (3.20)-(3.21). It then follows from (3.20) that

$$\begin{aligned}\mathbb{E}\left[X_t \,\middle|\, \mathcal{F}_{t-1}^X\right] &= \mathbb{E}\left[\sigma_t \varepsilon_t \,\middle|\, \mathcal{F}_{t-1}^X\right] \overset{(1)}{=} \sigma_t \mathbb{E}\left[\varepsilon_t \,\middle|\, \mathcal{F}_{t-1}^X\right] \\ &\overset{(2)}{=} \sigma_t \mathbb{E}\left[\mathbb{E}\left[\varepsilon_t \,\middle|\, \mathcal{F}_{t-1}^\varepsilon\right] \,\middle|\, \mathcal{F}_{t-1}^X\right] \overset{(3)}{=} 0 \quad \mathbb{P}\text{-a.s.}, \end{aligned} \tag{3.23}$$

where we have used that for all $t \in \mathbb{Z}$: (1) σ_t is $\mathcal{F}_{t-1}^X$-measurable (the process $\{\sigma_t^2, t \in \mathbb{Z}\}$ is previsible); (2) that $\mathcal{F}_{t-1}^X \subset \mathcal{F}_{t-1}^\varepsilon$ (the process $\{X_t, t \in \mathbb{Z}\}$ is nonanticipative); (3) $\mathbb{E}\left[\varepsilon_t \,\middle|\, \mathcal{F}_{t-1}^\varepsilon\right] = \mathbb{E}(\varepsilon_t) = 0$. Note that, as a consequence of (3.23), we have $\mathbb{E}[X_t] = \mathbb{E}\left[\mathbb{E}\left[X_t \,\middle|\, \mathcal{F}_{t-1}^X\right]\right] = 0$.

Because $\mathbb{E}\left[X_t \,\middle|\, \mathcal{F}_{t-1}^X\right] = 0$ $\mathbb{P}$-a.s., for all $t \in \mathbb{Z}$, the process $\{X_t, t \in \mathbb{Z}\}$ is said to be a *martingale difference* or *increment process* (see Appendix B). Assume that $\mathbb{E}\left[X_t^2\right] < \infty$. The fact that $\{X_t, t \in \mathbb{Z}\}$ is a martingale difference implies that it is also an uncorrelated sequence. To see this, let $h > 0$, then

$$\begin{aligned}\operatorname{Cov}(X_{t+h}, X_t) &= \mathbb{E}[X_t X_{t+h}] = \mathbb{E}\left[\mathbb{E}\left[X_t X_{t+h} \,\middle|\, \mathcal{F}_{t+h-1}^X\right]\right] \\ &= \mathbb{E}\left[X_t \mathbb{E}\left[X_{t+h} \,\middle|\, \mathcal{F}_{t+h-1}^X\right]\right] = 0. \end{aligned} \tag{3.24}$$

The last line of (3.24) follows because X_t is $\mathcal{F}_{t+h-1}^X$-measurable for $h > 0$, and $\mathbb{E}\left[X_{t+h} \,\middle|\, \mathcal{F}_{t+h-1}\right] = 0$ $\mathbb{P}$-a.s., as determined in (3.23).

While (3.24) implies that the ARCH process is white noise, it is still a dependent sequence. In fact, it is possible to write the ARCH(p) model as a non-Gaussian AR(p) model in the squares, X_t^2. First, square (3.20), $X_t^2 = \sigma_t^2 \varepsilon_t^2$, and then subtract (3.21), to obtain

$$X_t^2 - (\alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2) = Z_t\,, \tag{3.25}$$

where $Z_t = \sigma_t^2(\varepsilon_t^2 - 1)$. If the driving noise is Gaussian, then ε_t^2 is the square of a standard normal random variable, and $\varepsilon_t^2 - 1$ is a shifted (to have mean-zero) χ_1^2 random variable. The fact that $\{Z_t, t \in \mathbb{Z}\}$ is white noise follows from the fact that it is a martingale difference, $\mathbb{E}\left[Z_t \,\middle|\, \mathcal{F}_{t-1}^X\right] = \sigma_t^2 \mathbb{E}\left[\varepsilon_t^2 - 1\right] = 0$, $\mathbb{P}$-a.s., noting that $\sigma_t^2 \in \mathcal{F}_{t-1}^X$.

ARCH models were introduced by Engle (1982) to model the varying (conditional) variance or volatility of time series. It is often found in economics and finance that the larger values of time series (shocks) also cause instability at later times (i.e., larger variances); this phenomenom is referred to as *conditional heteroscedasticity*. For example, as illustrated in Figure 3.5 the returns of the S&P 500 exhibit largest variance after shocks. Allowing the conditional variance of X_t to depend on $X_{t-1}^2, \ldots, X_{t-p}^2$ is a first step in this direction.

The limitation of the ARCH model is that the squared process admits an AR correlation structure, which is not always the case. Bollerslev (1986) generalized the ARCH model by allowing the conditional variance $\mathbb{E}\left[X_t^2 \,\middle|\, \mathcal{F}_{t-1}^X\right]$ to depend not only on the lagged squared returns $(X_{t-1}^2, \ldots, X_{t-p}^2)$ but also on the lagged conditional variances, leading to the *generalized autoregressive conditional heteroscedastic* model, GARCH(p,q), where (3.20) still holds, but now

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \,, \tag{3.26}$$

where the coefficients α_j, $j \in \{0, \ldots, p\}$ and β_j, $j \in \{1, \ldots, q\}$ are nonnegative (although this assumption can be relaxed). The extension of the ARCH process to the GARCH process bears much similarity to the extension of standard AR models to ARMA models described in Section 1.3.2. To see this, consider a GARCH$(1,1)$ model (for ease of notation). As with the ARCH model, square (3.20), $X_t^2 = \sigma_t^2 \varepsilon_t^2$, and then subtract σ_t^2 to obtain

$$X_t^2 - \sigma_t^2 = \sigma_t^2(\varepsilon_t^2 - 1) := Z_t \,. \tag{3.27}$$

Consequently,

$$\beta_1(X_{t-1}^2 - \sigma_{t-1}^2) = \beta_1 Z_{t-1} \,, \tag{3.28}$$

and thus subtracting (3.28) from (3.27), we obtain

$$(X_t^2 - \sigma_t^2) - \beta_1(X_{t-1}^2 - \sigma_{t-1}^2) = Z_t - \beta_1 Z_{t-1} \,,$$

or

$$X_t^2 - \beta_1 X_{t-1}^2 - (\sigma_t^2 - \beta_1 \sigma_{t-1}^2) = Z_t - \beta_1 Z_{t-1} .$$

But $\sigma_t^2 - \beta_1 \sigma_{t-1}^2 = \alpha_0 + \alpha_1 X_{t-1}^2$, so finally

$$X_t^2 - (\alpha_1 + \beta_1) X_{t-1}^2 = Z_t - \beta_1 Z_{t-1} \,, \tag{3.29}$$

implying $\{X_t^2,\ t \in \mathbb{Z}\}$ is a non-Gaussian ARMA$(1,1)$. We note that this technique generalizes to any GARCH(p,q) by writing it as a GARCH(m,m) model where $m = \max(p,q)$ and setting any additional coefficients to zero, i.e., $\alpha_{p+1} = \cdots = \alpha_q = 0$ if $p < q$ or $\beta_{q+1} = \cdots = \beta_p = 0$ if $p > q$; see Exercise 3.7.

Summarizing, in general, if $\{X_t,\ t \in \mathbb{Z}\}$ is GARCH(p,q), then it is a martingale difference, $\mathbb{E}\left[X_t \,\middle|\, \mathcal{F}_{t-1}^X\right] = 0$ $\mathbb{P}$-a.s., and consequently is white noise. In addition, $\{X_t^2,\ t \in \mathbb{Z}\}$ is a non-Gaussian ARMA(p,q) process. This type of result was the goal of the bilinear model presented in Section 3.4, but as opposed to the bilinear model, the correlation structure of the GARCH model easily generalizes.

Another reason for the popularity of GARCH(p,q) models is that parameter estimation is straight-forward by conditioning on initial values. That is, the conditional likelihood of the data $X_{p+1}, \ldots, X_n$ given $X_1, \ldots, X_p$, and $\sigma_p^2 = \cdots = \sigma_{p+1-q}^2 = 0$ (if $q > 0$) is

$$\mathrm{L}(\theta; X_1, \ldots, X_p, \sigma_p^2 = \cdots = \sigma_{p+1-q}^2 = 0) = \prod_{t=p+1}^{n} \mathrm{p}^{\theta}(X_t \,|\, X_{t-1}, \ldots, X_1), \tag{3.30}$$

where

$$\theta = \{\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\}.$$

If $\varepsilon_t \sim$ iid $\mathrm{N}(0,1)$, then the conditional densities $\mathrm{p}^{\theta}(\cdot|\cdot)$ in (3.30) are Gaussian, i.e.,

$$X_t \,\big|\, X_{t-1}, \ldots, X_1 \sim \mathrm{N}(0, \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2),$$

for $t = p+1, \ldots, n$, with $\sigma_p^2 = \cdots = \sigma_{p+1-q}^2 = 0$. The sample size is typically large in financial applications so that conditioning on a few initial values is not problematic. Because the conditional likelihood is easily evaluated for a specified θ, a numerical routine such as Newton–Raphson (Section 1.3.4) is typically employed. In addition, the gradient of the likelihood is easily evaluated; see Exercise 3.8.

Some drawbacks of the GARCH model are that the likelihood tends to be flat unless n is very large, and the model tends to overpredict volatility because it responds slowly to large isolated returns. Returns are rarely conditionally normal or symmetric, so various extensions to the basic model have been developed to handle the various situations noticed empirically. Interested readers might find the general discussions in Bollerslev et al. (1994) and Shephard (1996) worthwhile reading. Also, Gouriéroux (1997) gives a detailed presentation of ARCH and related models with financial applications and contains an extensive bibliography. Excellent texts on financial time series analysis are Chan (2002), Teräsvirta et al. (2011), and Tsay (2005).

Finally, we briefly mention *stochastic volatility models*; a detailed treatment of these models is given in Chapter 9. The volatility component, σ_t^2, in the GARCH model is conditionally nonstochastic. For example, in the ARCH(1) model, any time the previous return is zero, i.e., $X_{t-1} = 0$, it must be the case that $\sigma_t^2 = \alpha_0$, and so on. This assumption seems a bit unrealistic in that one would expect some variability in this outcome. The stochastic volatility model adds a stochastic component to the volatility in the following way. The GARCH model assumes $X_t = \sigma_t \varepsilon_t$, or equivalently,

$$\ln X_t^2 = \ln \sigma_t^2 + \ln \varepsilon_t^2. \tag{3.31}$$

Thus, the observations, $\ln X_t^2$, are generated by two components, the unobserved volatility $\ln \sigma_t^2$ and the unobserved non-Gaussian noise $\ln \varepsilon_t^2$. While, for example, the GARCH$(1,1)$ models volatility without error, $\sigma_{t+1}^2 = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \sigma_t^2$, the basic stochastic volatility model assumes the latent variable is an autoregressive process,

$$\ln \sigma_{t+1}^2 = \phi_0 + \phi_1 \ln \sigma_t^2 + Z_t \tag{3.32}$$

where $Z_t \sim$ iid $\mathrm{N}(0, \sigma_z^2)$. The introduction of the noise term Z_t makes the latent volatility process stochastic.

Together (3.31) and (3.32) comprise the stochastic volatility model. In fact, the model is a non-Gaussian state space model. Let $h_t = \ln \sigma_t^2$, $Y_t = \ln X_t^2$, and $V_t = \ln \varepsilon_t^2$, then the basic stochastic volatility model may be written as

$$\begin{aligned} h_{t+1} &= \phi_0 + \phi_1 h_t + Z_t \quad \text{(state)} \\ Y_t &= h_t + V_t \quad\quad\quad\quad\ \text{(observation)} \end{aligned}$$

where Z_t is a Gaussian process, but V_t is not a Gaussian process. Given n observations, the goals are to estimate the parameters ϕ_0, ϕ_1 and σ_z^2, and then predict future volatility. Further details and extensions are discussed in Chapter 9.

3.6 Threshold ARMA models

Self-exciting threshold ARMA (SETARMA or TARMA) models, introduced by Tong (1978, 1983, 1990), have been widely employed as a model for nonlinear time series. Threshold models are piecewise linear ARMA models for which the linear relationship varies according to delayed values of the process (hence the term *self-exciting*). In this class of models, it is hypothesized that different autoregressive processes may operate and that the change between the various ARMA is governed by threshold values and a time lag. A k-regimes TARMA model has the form

$$X_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} X_{t-i} + Z_t^{(1)} + \sum_{j=1}^{q_1} \theta_j^{(1)} Z_{t-j}^{(1)} & \text{if } X_{t-d} \le r_1\,, \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} X_{t-i} + Z_t^{(2)} + \sum_{j=1}^{q_2} \theta_j^{(2)} Z_{t-j}^{(2)} & \text{if } r_1 < X_{t-d} \le r_2\,, \\ \quad\vdots & \quad\vdots \\ \phi_0^{(k)} + \sum_{i=1}^{p_k} \phi_i^{(k)} X_{t-i} + Z_t^{(k)} + \sum_{j=1}^{q_k} \theta_j^{(k)} Z_{t-j}^{(k)} & \text{if } r_{k-1} < X_{t-d}\,, \end{cases} \tag{3.33}$$

where $Z_t^{(j)} \sim \text{iid } \mathrm{N}(0, \sigma_j^2)$, for $j = 1, \ldots, k$, the positive integer d is a specified delay, and $-\infty < r_1 < \cdots < r_{k-1} < \infty$ is a partition of $\mathsf{X} = \mathbb{R}$. These models allow for changes in the ARMA coefficients over time, and those changes are determined by comparing previous values (back-shifted by a time lag equal to d) to fixed threshold values. Each different ARMA model is referred to as a *regime*. In the definition above, the values (p_j, q_j) of the order of ARMA models can differ in each regime, although in many applications, they are equal. Stationarity and invertibility are obvious concerns when fitting time series models. For the threshold time series models, such as TAR, TMA and TARMA models, however, the stationary and invertible conditions in the literature are less well-known in general. If known, often they are restricted to TAR or TMA processes with order one, and/or only sufficient conditions for higher orders; see e.g., Petruccelli and Woolford (1984), Brockwell et al. (1992), Ling (1999), and Ling et al. (2007).

The model can be generalized to include the possibility that the regimes depend on a collection of the past values of the process, or that the regimes depend on an exogenous variable (in which case the model is not self-exciting). For example, in the case such as that of the lynx, its prey varies from small rodents to deer, with the Snowshoe Hare being its overwhelmingly favored prey. In fact, in certain areas the lynx is so closely tied to the Snowshoe that its population rises and falls with that of the hare, even though other food sources may be abundant. In this case, it seems reasonable to replace X_{t-d} in (3.33) with say Y_{t-d}, where Y_t is the size of the Snowshoe Hare population.

The popularity of TAR models is due to their being relatively simple to specify, estimate, and interpret as compared to many other nonlinear time series models. In addition, despite its apparent simplicity, the class of TAR models could reproduce

many nonlinear phenomena such as stable and unstable limit cycles, jump resonance, harmonic distortion, modulation effects, chaos and so on.

As a simple example, Tong (1990, p. 377) fit the following TAR model with two regimes with delay variable $d = 2$ to the logarithm (base 10) of the lynx data,

$$X_t = \begin{cases} 0.62 + 1.25X_{t-1} - 0.43X_{t-2} + Z_t^{(1)}, & X_{t-2} \le 3.25\,, \\ 2.25 + 1.52X_{t-1} - 1.24X_{t-2} + Z_t^{(2)}, & X_{t-2} > 3.25\,, \end{cases} \tag{3.34}$$

although more complicated models were also fit to these data. Tong and Lim (1980) fit a two-regime TAR(11) to the sunspot data, the square root of which is shown in Figure 1.1. Also, Shumway and Stoffer (2011, Section 5.5) fit a threshold model to the differenced influenza and pneumonia mortality data set shown at the bottom of Figure 3.1.

3.7 Functional autoregressive models

In its basic form, a functional AR(p) model is written as

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + Z_t \tag{3.35}$$

where $\{Z_t,\, t \in \mathbb{N}\}$ is a strong white noise and is independent of X_s for $s < t$. The function $f(\cdot)$ is understood to be the conditional expectation, $f(X_{t-1}, \ldots, X_{t-p}) = \mathbb{E}\left[X_t \,\middle|\, X_{t-1}, \ldots, X_{t-p}\right]$, and can be left unspecified but with various smoothness conditions on $f(\cdot)$ and often under weak-dependence conditions on the process $\{X_t\}$. Sometimes the noise process is written as

$$Z_t = h(X_{t-1}, \ldots, X_{t-p})\,\varepsilon_t, \tag{3.36}$$

where $\varepsilon_t \sim \text{iid}\,(0,1)$. The function $h(\cdot)$ represents the possibility of conditionally heteroscedastic variance, with $h(\cdot) \equiv \sigma_z$ representing the homoscedastic case.

The basic goal is to estimate $f(\cdot)$, often via nonparametric methods, and then use the estimated relationship for prediction. We note, however, that many parametric models fit into this general model. For example, the TAR model in Section 3.6 would be considered a parametric form of the model with $f(\cdot)$ being an AR(q_j) with parameters depending on X_{t-d} as specified in (3.33), and with $\sigma(\cdot) = \sigma_j$ also depending on the observed value of X_{t-d}, with d specified as in (3.33). The model with (3.36) added clearly includes various forms of the ARCH model. As another example, we mention the amplitude-dependent exponential autoregressive (EXPAR) model introduced in Haggan and Ozaki (1981), which assumes that

$$f(x_1, \ldots, x_p) = (\phi_1 + \pi_1 \mathrm{e}^{-\gamma x_1^2})x_1 + \cdots + (\phi_p + \pi_p \mathrm{e}^{-\gamma x_1^2})x_p\,. \tag{3.37}$$

In this case, the autoregressive part retains an additive form, but the coefficients entering the regression are made to change instantaneously with x_1^2.

In more recent works, estimation of f or h is performed using some of the same tools used in non- or semi-parametric estimation of regression functions. Note, however, that some care should be exercised in controlling the functions f and h in such

a way that there exists a (strict-sense or second-order) stationary solution for (3.35)–(3.36). This of course involves some rather non-trivial conditions on the behavior of f and h; we will have to wait until Chapter 4 to develop the tools required to show the existence of such solutions. Various versions of the non- or semi-parametric approach have been explored. For example, Hastie and Tibshirani (1990) examined the additive model,

$$X_t = f_1(X_{t-1}) + \cdots + f_p(X_{t-p}) + Z_t \tag{3.38}$$

and Chen and Tsay (1993) explored the functional coefficient AR model, which, in its simplest form, is written as

$$X_t = f_1(X_{t-d})X_{t-1} + \cdots + f_p(X_{t-d})X_{t-p} + Z_t \tag{3.39}$$

where $d > 0$ is some specified delay. Another interesting model is the *partially linear model*, where for example, we might have

$$f(X_{t-1}, \ldots, X_{t-p}) = \mu(t) + \sum_{j=1}^{p} \phi_j X_{t-j} \tag{3.40}$$

where $\mu(t)$ is a local trend function of time t that we do not wish to model parametrically. For example, the rates of pneumonia and influenza mortality series shown at the bottom of Figure 3.1 exhibits some negative, but not necessarily linear trend over the decade (e.g., it appears that the decline in the average annual mortality is more pronounced over the first part of the series than at the end of the series). In this case, we may wish to fit $f(\cdot)$ via semiparametric methods.

Semiparametric and nonparametric estimation for time series models in various forms runs the gamut of the methods used for independent data. These mainly involve some type of local smoothing such as running means or medians, kernel smoothing, local polynomial regression, smoothing splines, and backfitting algorithms such as the ACE algorithm. There are a number of excellent modern expositions on this topic and we refer the reader to texts by Fan and Yao (2003) and by Gao (2007). In addition, the comprehensive review by Härdle et al. (1997) provides an accessible overview of the field.

3.8 Linear processes with infinite variance

In Example 3.2, we argued that the EEG data shown at the top of Figure 3.2 may be best described as having infinite variance, and we compared the innovations after an AR(p) fit to the data to Cauchy noise, a realization of which is shown in Figure 3.3. Such models have been used in a variety of situations, for example Fama (1965) used them to examine stock market prices.

An important property of Gaussian random variables is that the sum of two of them is itself a normal random variable. One consequence of this is that if Z is normal, then for Z_1 and Z_2 independent copies of Z and any positive constants a and b, $aZ_1 + bZ_2 =_d cZ + d$, for some positive c and some $d \in \mathbb{R}$. (The symbol $=_d$ means equality in distribution). In other words, the shape of Z is preserved (up to scale and

shift) under addition. One typically defines an infinite variance linear process via symmetric (about zero) stable innovations.

Definition 3.4 (Stable law). *A random variable Z is said to be* stable*, or have a stable distribution, if Z_1 and Z_2 are two independent copies of Z, and for any positive constants a and b, the linear combination $aZ_1 + bZ_2$ has the same distribution as $cZ + d$, for some positive c and $d \in \mathbb{R}$. A random variable is said to be strictly stable if $d = 0$ for all positive a and b. A random variable is symmetric stable if X and $-X$ have the same stable distribution.*

Equivalently, the random variable Z is stable if for every $n \in \mathbb{N}^*$, there exists constants $a_n > 0$ and b_n such that the sum of i.i.d. copies, $Z_1 + \cdots + Z_n$, has the same distribution as $a_n Z + b_n$. We say that Z is strictly stable if $b_n = 0$.

Remark 3.5. It is possible to show that the only possible choice for the scaling constant a_n is $a_n = n^{1/\alpha}$ for some $\alpha \in (0, 2]$.

Remark 3.6. The addition rule for independent random variables says that the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances. Suppose $Z \sim \mathrm{N}(\mu,\ \sigma^2)$. Let Z_1 and Z_2 be two independent copies of Z. Then, $aZ_1 \sim \mathrm{N}(a\mu,\ (a\sigma)^2)$, $bZ_2 \sim \mathrm{N}(b\mu,\ (b\sigma)^2)$, and $cZ + d \sim \mathrm{N}(c\mu + d,\ (c\sigma)^2)$. The addition rule implies that $c^2 = a^2 + b^2$ and $d = (a + b - c)\mu$.

The most effective way to define the set of stable distributions is through their characteristic functions; see Billingsley (1995, Chapters 5, 26).

Theorem 3.7. *A random variable X is* stable *if and only if $X =_d aZ + b$, where $a > 0$, $b \in \mathbb{R}$ and Z is a random variable with characteristic function*

$$\varphi(\xi) = \mathbb{E}\exp(\mathrm{i}\xi Z) = \begin{cases} \exp\left(-|\xi|^\alpha [1 - \mathrm{i}\beta \tan\frac{\pi\alpha}{2}\ (\operatorname{sign}\xi)]\right) & \alpha \neq 1 \\ \exp\left(-|\xi|[1 + \mathrm{i}\beta\frac{2}{\pi}\ (\operatorname{sign}\xi)\ln|\xi|]\right) & \alpha = 1 \end{cases},$$

where $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, and sign *is the sign function given by* $\operatorname{sign}(\xi) = -1$ *if $\xi < 0$,* $\operatorname{sign}(\xi) = 0$ *if $\xi = 0$ and* $\operatorname{sign}(\xi) = 1$ *if $\xi > 0$.*

When $\beta = 0$ and $b = 0$, these distributions are symmetric around zero, in which case the characteristic function of aZ has the simpler form

$$\varphi(\xi) = \mathrm{e}^{-a^\alpha |\xi|^\alpha}$$

Remark 3.8. The Gaussian distribution is stable with parameters $\alpha = 2,\ \beta = 0$. The Cauchy distribution is stable with parameters $\alpha = 1,\ \beta = 0$. A random variable Z is said to be Lévy$(\gamma,\ \delta)$ if it has density

$$f(x) = \sqrt{\frac{\gamma}{2\pi}}\frac{1}{(x-\delta)^{3/2}}\exp\left(-\frac{\gamma}{2(x-\delta)}\right), \quad \delta < x < \infty.$$

The Lévy distribution is stable with parameters $\alpha = 1/2,\ \beta = 1$.

Remark 3.9. Both the Gaussian and Cauchy distributions are symmetric and bell-shaped, but the Cauchy distribution has much heavier tails. If Z is standard normal, $\mathbb{P}(Z \geq 3)$ is 1.310^{-3}, whereas if Z is standard Cauchy (equivalently, a t-distribution

with 1 degree of freedom), $\mathbb{P}(Z \geq 3) = 10^{-1}$. In a sample from these two distributions, there will be (on average) more than 100 times as many values above 3 in the Cauchy case than in the normal case. This is the reason stable distributions are called heavy tailed. In contrast to the normal and Cauchy distributions, the Lévy distribution is highly skewed. The distribution is concentrated on $x > 0$, and it has even heavier tails than the Cauchy.

Remark 3.10. For non-normal stable random variables Z (i.e., $\alpha < 2$), it can be shown that $\mathbb{E}\left[|Z|^{\delta}\right] < \infty$ only for $0 < \delta < \alpha$. Consequently, $\mathrm{Var}\, Z = \infty$ for $0 < \alpha < 2$ and $\mathbb{E}\left[|Z|\right] = \infty$ for $0 < \alpha \leq 1$.

It is possible to define an ARMA-type model with stable innovations. That is, we may define a process $\{X_t,\, t \in \mathbb{Z}\}$ such that $\mathbb{P}$-a.s.,

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

where $\{Z_t,\, t \in \mathbb{Z}\}$ is a sequence of i.i.d. stable random variables, and $\sum_j |\psi_j|^{\delta} < \infty$ for some $\delta \in (0, \alpha) \cap [0, 1]$. Moreover, it is possible to write the process as

$$\phi(B) X_t = \theta(B) Z_t,$$

where $\{X_t,\, t \in \mathbb{Z}\}$ is strictly stationary (but, of course, not covariance stationary unless $\alpha = 2$) and $\phi(B)$ and $\theta(B)$ are as in Section 1.3.2. These models are described in a fair amount of detail in Brockwell and Davis (1991, §12.5), who also discuss fitting these models to data.

3.9 Models for counts

In Example 3.3, we presented two time series that are discrete-valued and take on small values. These series should be contrasted with the series discussed in Section 1.5 and Example 3.1, which are also counts (the number of sunspots in Figure 1.1; the number of lynx trappings and the number of flu deaths in Figure 3.1), but are quite different in that one could use, for example, a TAR model with Gaussian noise as a reasonable approximation in the latter cases, but any use of normality is out of the question for the great discoveries series displayed in Figure 3.7 and for the cases of polio time series displayed in Figure 3.8.

There are two basic approaches to the problem. One approach is to develop models that produce integer-valued outcomes, and the other is to develop generalized linear models for dependent data. We briefly describe each approach in the following sections. Our presentation is very brief, the texts by MacDonald and Zucchini (1997) and by Kedem and Fokianos (2002) present rather extensive discussions of these models. In addition, the second part of Durbin and Koopman (2012) details the generalized linear model approach to the problem.

3.9.1 Integer valued models

In the late 1970s and through the 1980s, there were a number of researchers who worked on models with specific non-Gaussian marginals. The driving force behind

these models is that discrete-valued time series can have ARMA-type autocorrelation structures along with marginal distributions that follow standard distributions such as Poisson, negative binomial, and so on. For a linear model to have marginals that match the innovations, the distributions must be stable; see Section 3.8. However, a number of researchers focused on random mixing or random summation as a method to obtain models that admit desired marginals and correlation structures.

For example, Jacobs and Lewis (1978a,b, 1983) developed DARMA, or discrete ARMA, models via mixing. For example, a DAR(1) model is of the form

$$X_t = V_t X_{t-1} + (1 - V_t) Z_t \tag{3.41}$$

where V_t is i.i.d. Bernoulli with $\Pr\{V_t = 1\} = 1 - \Pr\{V_t = 0\} = \rho$, and $\{Z_t\}$ is i.i.d. according to some specified discrete-valued distribution. Clearly, the support of X_t is the support of the noise Z_t. It is easy to show that (3.41) has the ACF structure of an AR(1), i.e., $\rho(h) = \rho^h$ for $h \in \mathbb{N}$; see Exercise 3.9. The authors also developed a 'new' DARMA, or NDARMA model with similar properties. These types of models are discussed further in Example 4.22 of Chapter 4.

Langberg and Stoffer (1987), and Block et al. (1988, 1990) developed (bivariate) exponential and geometric time series with ARMA correlation structure. We briefly discuss the univariate aspects of the geometric model; the bivariate model was called the BGARMA model. The authors developed the model using both random mixing and random summation, and then showed that the two methods are equivalent. The basic idea is as follows. If $X \sim \mathrm{G}(p)$ and $Z \sim \mathrm{G}(p/\pi)$ are independent geometric random variables [e.g., $\Pr(X = k) = p(1-p)^{k-1}$ for $k \in \mathbb{N}^*$], independent of I, which is Bernoulli$(1 - \pi)$, then $X' = IX + Z$, has the same distribution as X. For random summation, suppose $N \sim G(\pi)$ independent of $Z_j \sim$ iid $\mathrm{G}(p/\pi)$, then $X' = \sum_{j=1}^{N} Z_j$ has the representation $X' = IX + Z$. This basic idea can be extended and used to formulate various non-Gaussian multivariate processes, and we refer the reader to Block et al. (1988) for a thorough presentation. As a simple example, let $X_0 \sim \mathrm{G}(p)$ and define, for $t \in \mathbb{N}$,

$$X_t = I_t X_{t-1} + Z_t \tag{3.42}$$

where $I_t \sim$ iid Bernoulli$(1 - \pi)$, and $Z_t \sim$ iid $\mathrm{G}(p/\pi)$ is the noise process. Then, $\{X_t,\ t \in \mathbb{N}\}$ is a process with the autocorrelation structure of an AR(1), and where $X_t \sim \mathrm{G}(p)$; see Exercise 3.9.

Finally we mention some models that are based on the notion of *thinning* that was discussed in Steutel and Van Harn (1979) as an integer-valued analog to stability for continuous-valued random variables. The idea is closely related to the random summation concept in Block et al. (1988), but in this case, the decomposition is given by $X' = \alpha \circ X + X_\alpha$ where X and X_α are independent, and $\alpha \circ X = \sum_{j=1}^{X} N_j$ where $N_j \sim$ iid Bernoulli(α). Under this decomposition, X' and X have the same distribution, and such processes are called discrete stable; Steutel and Van Harn (1979) show, for example, that the Poisson distribution is discrete stable. McKenzie (1986), Al-Osh and Alzaid (1987) and others used the idea of thinning to obtain the INAR, or integer-valued AR, model. For example, an INAR(1) has the form

$$X_t = \alpha \circ X_{t-1} + Z_t \tag{3.43}$$

for $\alpha \in [0,1)$, where Z_t is an i.i.d. sequence of integer-valued random variables such as Poisson(λ) wherein the marginal of X_t is also Poisson with rate $\lambda/1-\alpha$. Moreover, the ACF of X_t is like an AR(1) and is given by $\rho(h) = \alpha^h$ for $h \in \mathbb{N}$; see Exercise 3.9.

3.9.2 Generalized linear models

The basic idea here is to extend the theory of generalized linear models to dependent data. This approach is apparently more successful for data analysis than the integer-valued models discussed in the previous subsection because there seem to be fewer pathologies in this setup. This topic is best discussed in more generality than is done in this section, and should be presented after the material in Part III on nonlinear state space models. Our brief discussion here can be supplemented with the texts mentioned in the introduction to this section.

We restrict attention to the univariate case. Let U_t be a vector of deterministic exogenous inputs or covariates, let $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots; U_t)$, and denote the conditional mean and variance by $\mu_t = \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right]$ and $\sigma_t^2 = \mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$. Assume that conditionally, the observations are from an exponential family,

$$f(x_t \mid \theta_t, \mathcal{F}_{t-1}) = \exp\left[\frac{x_t\theta_t - b(\theta_t)}{\phi} + c(x_t;\phi)\right]. \tag{3.44}$$

The parameter ϕ is called the dispersion or scale parameter. It is assumed that $b(\theta_t)$ is twice differentiable, $c(x_t;\phi)$ does not involve θ_t, and θ_t is the (monotone) canonical link function. For this family, it can be shown that $\mu_t = b'(\theta_t)$ and $\sigma_t^2 = \phi\, b''(\theta_t)$. As an example, consider the Poisson distribution with mean function μ_t, in which case $\phi = 1$, $\theta_t = \ln\mu_t$ is the canonical link, $b(\theta_t) = \exp(\theta_t)$, and $c(x_t;\phi) = -\ln(x_t!)$; see Exercise 3.9. In the basic overdispersed or quasi-Poisson model, the scale parameter ϕ is left unspecified and estimated from the data rather than fixing it at 1. Typically, an estimating function is used and a quasi-Poisson model does not correspond to models with a fully specified likelihood.

Various approaches to modeling θ_t exist and include non- and semi-parametric methods, observation-driven models (i.e., μ_t is driven by the past data) and parameter driven models (i.e., μ_t is driven by the past parameter values) and various combinations of these models. In these settings, we have a link function,

$$\theta_t := \theta_t(\mu_t) = h(U_t, X^{t-1}, \mu^{t-1}, \varepsilon_t) \tag{3.45}$$

where $X^{t-1} = \{X_{t-1}, X_{t-2}, \ldots\}$ represents the data history, $\mu^{t-1} = \{\mu_{t-1}, \mu_{t-2}, \ldots\}$, and ε_t represents a vector of latent processes.

For time series of counts, the Poisson distribution is used most often. In the case of time series, it is typically necessary to account for over-dispersion and autocorrelation found in the data. For example, in Example 3.3 we saw overdispersion in that the data seem to have negative binomial marginals, and in Figure 3.7 and Figure 3.8, where autocorrelation is evident. Static models have $h(\cdot) = \beta' U_t$ where β is a vector of regression parameters and U_t is a vector of fixed inputs as previously explained. It

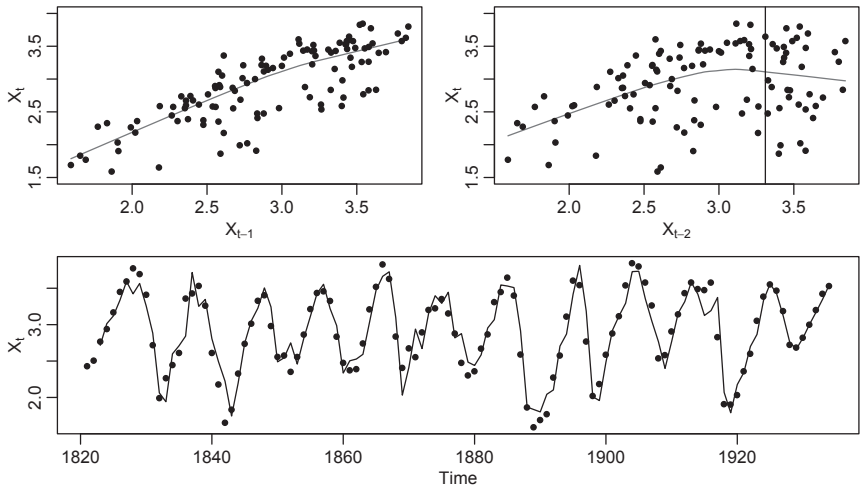


Figure 3.9 *Display for Example 3.11. Top: Lag plots of the Lynx series with a lowess fit emphasizing nonlinearity. The vertical line shows the threshold value of 3.31. Bottom: The logged Lynx series as points, and the one-step-ahead predictions as a solid line.*

is also possible to have $h(\cdot)$ be non- or semi-parametric, particularly to evaluate nonlinear trend; see Example 3.13. An extension that is treated in Kedem and Fokianos (2002) and reviewed in Fokianos (2009) is the case where $h(\cdot) = \beta' U_t + \sum_{j=1}^{p} \phi_j X_{t-j}$. Zeger (1988) introduced a stochastic element via a stationary latent process; i.e., $h(\cdot) = \beta' U_t + \varepsilon_t$. Here, overdispersion is introduced via the latent variable. Davis et al. (2003) and Shephard (1995) extended this idea to the *generalized linear ARMA*, or GLARMA, model by writing $\varepsilon_t = \sum_{j=1}^{p} \alpha_j(\varepsilon_{t-j} + e_{t-j}) + \sum_{j=1}^{q} \beta_j e_{t-j}$, where $e_t = (X_t - \mu_t)/\sqrt{\mu_t}$, and $\varepsilon_t = e_t = 0$ for $t \leq 0$. Finally, we mention GARCH-type Poisson models wherein $\mu_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{k=1}^{q} \beta_j \mu_{t-j}$; see Engle and Russell (1998). There is a considerable amount of literature on this topic, and we have only presented a few approaches. For an extensive and up-to-date review, see Jung and Tremayne (2011), which also presents an empirical comparison of various methods for analyzing discrete-valued time series.

3.10 Numerical examples

In this section, we use some of the models presented in this chapter to analyze a few of the data sets presented in Section 3.1. In particular, we will fit a SETAR model to the lynx data set using the R package `tsDyn`, an asymmetric GARCH-type model to the S&P 500 data set using the package `fGarch`, and overdispersed Poisson models to the polio data set using the packages `dyn`, `mgcv`, and `glm`. We also present some tests for detecting nonlinearity.

Example 3.11 (SETAR model). We used the `tsDyn` package to fit the SETAR model specified in (3.34) to the logarithm (base 10) of the lynx data. However, we allow the program to choose the optimal value of the threshold, rather than the one

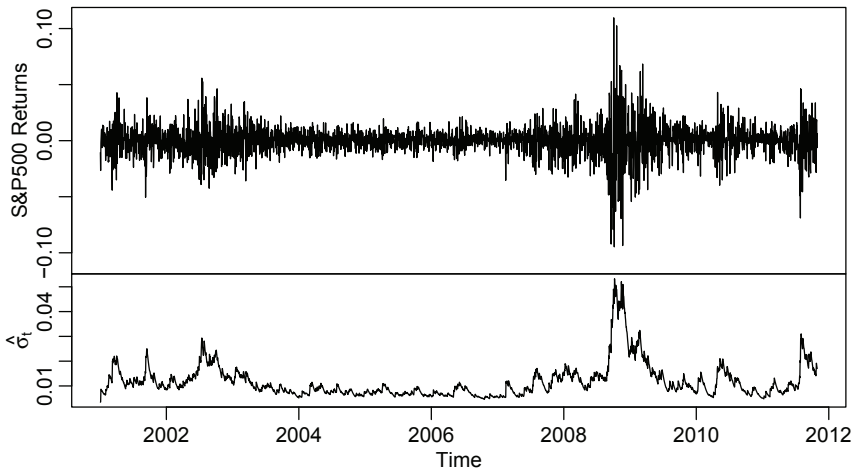


Figure 3.10 *Display for Example 3.12. Top: The S&P 500 returns. Bottom: The predicted one-step-ahead volatility from an Asymmetric Power ARCH fit.*

specified in (3.34), namely, 3.25. The top of Figure 3.9 shows two lag plots with lowess fits superimposed; these plots clearly indicate nonlinear behavior. The vertical line in the lag 2 plot indicates the optimal threshold value for the model, which is 3.31. The fitted model is similar to the fitted model displayed in (3.34). The results of the fit are as follows.

```
SETAR model ( 2 regimes)
Coefficients:
 Low regime:                          High regime:
    phiL.1    phiL.2   const L         phiH.1    phiH.2   const H
  1.264279 -0.428429  0.588437       1.599254 -1.011575  1.165692
Threshold:
-Variable: Z(t) = + (0) X(t)+ (1)X(t-1)
-Value: 3.31
Proportion of points in low regime: 69.64%   High regime: 30.36%
```

Finally, we note that it is not necessary to use a special package to fit a SETAR model. The model can be fit using piecewise linear regressions, `lm` in R, once a threshold value has been determined. ◇

Example 3.12 (Asymmetric power ARCH). The R package `fGarch` was used to fit a model to the S&P 500 returns discussed in Example 3.2. The data are displayed in Figure 3.5, where a small amount of autocorrelation is noticed. Hence, we include an AR(1) in the model to account for the conditional mean. For the conditional variance, we fit an Asymmetric Power ARCH (APARCH) model to the data; see Exercise 8.17 for details. In this case, the model is $X_t = \mu_t + \varepsilon_t$ where μ_t is an AR(1), and ε_t is GARCH-type noise where the conditional variance is modeled as

$$\sigma_t^\delta = \alpha_0 + \sum_{j=1}^{p} \alpha_j f_j(\varepsilon_{t-j}) + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^\delta \,, \tag{3.46}$$

where

$$f_j(\varepsilon) = (|\varepsilon| - \gamma_j \varepsilon)^{\delta} . \tag{3.47}$$

Note that the model is GARCH when $\delta = 2$ and $\gamma_j = 0$, $j \in \{1, \ldots, p\}$. The parameters γ_j ($|\gamma_j| \leq 1$) are the *leverage* parameters, which are a measure of asymmetry, and $\delta > 0$ is the parameter for the power term. A positive (resp. negative) value of γ_j's means that past negative (resp. positive) shocks have a deeper impact on current conditional volatility than past positive shocks (Black, 1976). This model couples the flexibility of a varying exponent with the asymmetry coefficient to take the *leverage effect* into account. Further, to guarantee that $\sigma_t > 0$, we assume that $\alpha_0 > 0$, $\alpha_j \geq 0$ with at least one $\alpha_j > 0$, and $\beta_j \geq 0$.

We fit an AR(1)-APARCH(1,1) model to the data. The (partial) results of the fit are outlined below, and Figure 3.10 displays the returns as well as the estimated one-step-ahead predicted volatilty, $\widehat{\sigma}_t$.

```
          Estimate  Std. Error  t value   Pr(>t)
mu       5.456e-05   1.685e-04    0.324  0.74605
ar1     -6.409e-02   1.989e-02   -3.221  0.00128
alpha0   1.596e-05   3.419e-06    4.668 3.04e-06
alpha1   4.676e-02   8.193e-03    5.708 1.15e-08
gamma1   1.000e+00   4.319e-02   23.156  < 2e-16
beta1    9.291e-01   7.082e-03  131.207  < 2e-16
delta    1.504e+00   2.054e-01    7.323 2.42e-13

Standardised Residuals Tests:
                                Statistic p-Value
 Shapiro-Wilk Test  R    W      0.9810958 0
 Ljung-Box Test     R    Q(20)  19.58712  0.4840092
 Ljung-Box Test     R^2  Q(20)  25.55894  0.1808778
```

Finally, we mention that ACF of the squared returns shown in Figure 3.5 indicates persistent volatility, and it seems reasonable that some type of integrated, or IGARCH(1,1) model could be fit to the data. In this case, (3.26) would be fit but with $\alpha_1 + \beta_1 \equiv 1$; recall the discussion following (3.29). ◇

Example 3.13 (Overdispersed Poisson model). At this point, we do not have all the tools necessary to fit complex models to dependent count data, so we use some existing R packages for independent data to fit simple time series models. In particular, we fit two overdispersed Poisson models to the polio data set displayed in Figure 3.8. In both cases, we followed Zeger (1988) by adding sinusoidal terms to account for seasonal behavior, namely, $C_{kt} = \cos(2\pi tk/12)$ and $S_{kt} = \sin(2\pi tk/12)$, for $k = 1, 2$. The first model is fully parametric, while the second model is semiparametric. The link functions, (3.45), for the two models are:

$$\textit{Model } 1: \ \ln(\mu_t) = \alpha_0 + \alpha_1 t + \beta_1 C_{1t} + \beta_2 S_{1t} + \beta_3 C_{2t} + \beta_4 S_{2t} + \varphi X_{t-1} \tag{3.48}$$

$$\textit{Model } 2: \ \ln(\mu_t) = \alpha_0 + \mathrm{sm}(t) + \beta_1 C_{1t} + \beta_2 S_{1t} + \beta_3 C_{2t} + \beta_4 S_{2t} \tag{3.49}$$

A lagged value of the series was included in Model 2 in a first run, but it was not needed when the semi-parametrically fit smooth trend term, $\mathrm{sm}(t)$, was included in the model. In each case, a scale parameter ϕ is estimated. The results are displayed in Figure 3.11. ◇

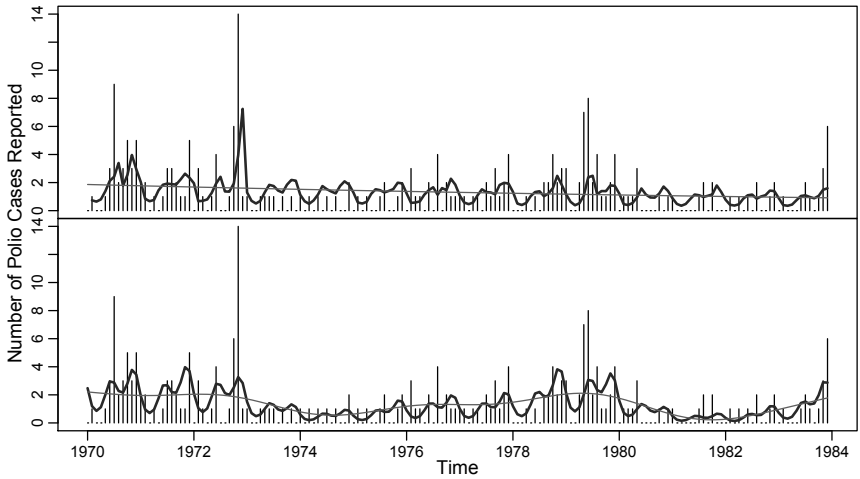


Figure 3.11 *Display for Example 3.13. In both cases, an overdispersed Poisson model is used. Top: The result of the Model 1, (3.48), fit superimposed on the polio data set. Displayed are the estimated mean function and linear trend lines. Bottom: The result of the Model 2, (3.49), fit superimposed on the polio data set. Displayed are the estimated mean function and trend lines; the assumed smooth trend is estimated via semi-parametric techniques.*

Example 3.14 (BiSpectrum). A number of researchers have suggested tests of nonlinearity based on the bispectrum given in 3.14. Given the results displayed in Exercise 3.4, it is clear that an estimate of

$$B(\omega_1, \omega_2) = \frac{|f(\omega_1, \omega_2)|^2}{f(\omega_1) f(\omega_2) f(\omega_1 + \omega_2)}$$

could be used to determine whether or not a process is linear. Because the value of $B(\omega_1, \omega_2)$ is unbounded, some researchers have proposed normalizing it so that, like squared coherence given in (1.82), it lives on the unit interval, with larger values indicating nonlinear (specifically, quadratic) dynamics.

The method proposed in Hinich and Wolinsky (2005) uses "frame averaging" wherein one first partitions time into blocks. Then, DFTs are calculated in each block and their averages are used to estimate the spectrum and bispectrum and to form an estimate of $B(\omega_1, \omega_2)$. This estimate is then transformed using a normalization based on a noncentral chi-squared distribution under the null hypothesis that the process is linear. For details, we refer the reader to Hinich and Wolinsky (2005). We provide an R script, `bi.coh`, that can be used to estimate and plot the normalized bispectrum.

Figure 3.12 shows the graphic produced by the script for the S&P 500 returns discussed in Example 3.12. Note that values over .95 are dark (or pink if color is used); numerous dark values indicate nonlinearity. ◇

Example 3.15 (Time domain tests for nonlinearity). There are a number of time domain tests for nonlinearity in the conditional mean, and many of them are discussed and compared in Lee et al. (1993). An obvious approach is to assess whether

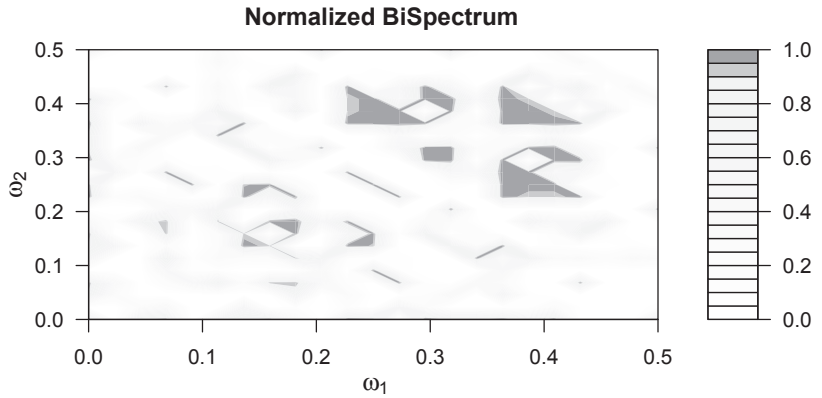


Figure 3.12 *Display for Example 3.14. Estimated normalized bispectrum of the S&P 500 returns displayed in Figure 3.5, with highlighted regions indicating departure from the linearity assumption.*

the coefficients of the higher-order ($M \geq 2$) terms in the Volterra series (3.3) are zero. For practical purposes, given a finite set of data, tests typically focus on whether or not there is the existence of second-order terms.

Keenan (1985) developed a one-degree-of-freedom test by first fitting a linear AR(p) model, where p is chosen arbitrarily or by some model choice criterion such as those described in (1.65). Given data, $\{X_1, \ldots, X_n\}$, a model is fit, and the one-step-ahead predictions,

$$\widehat{X}_{t|t-1} = \widehat{\phi}_0 + \widehat{\phi}_1 X_{t-1} + \cdots + \widehat{\phi}_p X_{t-p},$$

for $t = p+1, \ldots, n$ are calculated. Then, the AR(p) model is fit again, but now with the squared predictions included in the model. That is, the model

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta \widehat{X}^2_{t|t-1} + Z_t$$

is fit for $t = p+1, \ldots, n$, and the null hypothesis that $\theta = 0$ is tested against the alternative hypothesis that $\theta \neq 0$, in the usual fashion. In a sense, one is testing if the squared forecasts have additional predictive ability.

Tsay (1986) extended this idea by testing whether any of the second-order terms are additionally predictive. That is, the model,

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \sum_{1 \leq i \leq j \leq p} \sum \theta_{i,j} X_{t-i} X_{t-j} + Z_t,$$

is fit to the data and the null hypothesis $\theta_{i,j} = 0$ for all $1 \leq i \leq j \leq p$ is tested against the alternative hypothesis that at least one $\theta_{i,j} \neq 0$.

Both of these tests are available in the R package TSA, and we perform both tests on the logged Lynx data set analyzed in Example 3.11. Note that the series is fairly short.

```
> Keenan.test(log10(lynx))
   $test.stat  $p.value   $order
   [1] 11.669  [1] 0.001  [1] 11
> Tsay.test(log10(lynx))
   $test.stat  $p.value   $order
   [1]  1.316  [1] 0.226  [1] 11
```

We see that AIC chooses $p = 11$ and that the Keenan test identifies nonlinearity in the conditional mean whereas the Tsay test does not. ◇

Exercises

3.1. Suppose that $Z_t = A\cos(\omega_0 t + \varphi)$, where A, ω_0 and φ are fixed.

(a) If Z_t is the input process in (3.1), show that the output, X_t is a sinusoid of frequency ω_0 with squared amplitude $|\psi(\mathrm{e}^{-\mathrm{i}\omega_0})|^2$ and phase shifted by $\arg(\psi(\mathrm{e}^{-\mathrm{i}\omega_0}))$.

(b) Let $X_t = Z_t^2$. What are the frequency, amplitude and phase of X_t? Comment.

3.2. Using Example 3.2 as a guide, remove the signal from the explosion series shown in Figure 3.6 using the R command `ar`. Then calculate the sample ACF of the innovations of explosion series and compare it to the sample ACF of the squared innovations. How do these results compare to the results of Example 3.2 for the EEG and the S&P500 series?

3.3. Show that the solution to (3.5) is $a = b = 0$ and $c = 1$, as claimed.

3.4. Suppose $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, where $\sum_j |\psi_j| < \infty$ and $\{Z_t,\, t \in \mathbb{Z}\}$ are i.i.d. with mean-zero, variance σ_z^2 and finite third moment $\mathbb{E}\left[Z_t^3\right] = \mu_3$.

(a) Let $\kappa_x(r_1, r_2) = \mathrm{cum}(X_{t+r_1}, X_{t+r_2}, X_t)$. Show that $\kappa_x(r_1, r_2) = \mu_3 \sum_j \psi_j \psi_{j+r_1} \psi_{j+r_2}$.

(b) Use part (a) and Proposition 1.22 (see also Example 1.33) to show that the bispectrum of $\{X_t,\, t \in \mathbb{Z}\}$ is $f(\omega_1, \omega_2) = \frac{\mu_3}{(2\pi)^2} \psi(e^{-\mathrm{i}\omega_1}) \psi(e^{-\mathrm{i}\omega_2}) \psi(e^{\mathrm{i}(\omega_1+\omega_2)})$.

(c) Finally, show that

$$\frac{|f(\omega_1, \omega_2)|^2}{f(\omega_1) f(\omega_2) f(\omega_1 + \omega_2)}$$

is equal to $\mu_3^2 / 2\pi\sigma_z^6$, independent of frequency.

(d) How can the facts of this exercise be used to determine if a process $\{X_t,\, t \in \mathbb{Z}\}$ is (i) linear, and (ii) Gaussian?

3.5. Consider the process given by $X_t = Z_t + \theta Z_{t-1} Z_{t-2}$, where $\{Z_t,\, t \in \mathbb{Z}\}$ is a sequence of i.i.d. Gaussian variables with zero mean and variance σ^2.

(a) Show that $\{X_t,\, t \in \mathbb{Z}\}$ is strict-sense and weak-sense stationary.

(b) Show that $\{X_t,\, t \in \mathbb{Z}\}$ is a (weak) white-noise.

(c) Compute the bispectrum of $\{X_t,\, t \in \mathbb{Z}\}$.

3.6. For the bilinear model $\mathrm{BL}(0, 0, 2, 1)$ shown in (3.18),

(a) Show that $\mathbb{E}[X_t] = 0$ and then verify (3.19).

(b) Verify the statement that $\{X_t,\, t \in \mathbb{Z}\}$ is white noise but that $\{X_t^2,\, t \in \mathbb{Z}\}$ is predictable from its history.

3.7. If $\{X_t,\, t \in \mathbb{Z}\}$ is GARCH(p,q), show that $\{X_t^2,\, t \in \mathbb{Z}\}$ is non-Gaussian ARMA.

3.8. If $\{X_t,\, t \in \mathbb{Z}\}$ is ARCH(1), show that the gradient of the conditional log-likelihood $l(\alpha_0, \alpha_1 \,|\, X_1)$ is given by the 2×1 gradient vector,

$$\begin{pmatrix} \partial l/\partial \alpha_0 \\ \partial l/\partial \alpha_1 \end{pmatrix} = \sum_{t=2}^{n} \begin{pmatrix} 1 \\ X_{t-1}^2 \end{pmatrix} \times \frac{\alpha_0 + \alpha_1 X_{t-1}^2 - X_t^2}{2\left(\alpha_0 + \alpha_1 X_{t-1}^2\right)^2}.$$

3.9. The following problems are based on the material in Section 3.9.

(a) For the DAR(1) model in (3.41), show that the support of X_t and Z_t are the same, and then derive the ACF.

(b) Show that the marginal distribution of X_t defined by (3.42) is Geometric with parameter p, and then show the ACF is that of an AR(1) model.

(c) Show that the marginal of X_t in (3.43) is Poisson if Z_t is Poisson, and then derive the ACF of X_t.

(d) Suppose X_t is Poisson with conditional mean given by μ_t. Verify that the marginal of X_t is in the exponential family given by (3.44), identify the components, θ_t, ϕ, $b(\cdot)$, and $c(\cdot)$, and verify that $\mu_t = b'(\theta_t)$ and $\sigma_t^2 = \phi\, b''(\theta_t)$.

3.10. Using Section 3.10 as a guide, perform the following analyses.

(a) Fit a threshold model to the sunspots series displayed in Figure 1.1.

(b) For the mortality series, say M_t, shown at the bottom of Figure 3.1, calculate the normalized bispectrum of the series itself and then of $X_t = \nabla \ln M_t$. What is the interpretation of X_t? Comment on the difference between the results. Then, fit a threshold model to X_t. Explain why it is better to fit such a model to X_t rather than M_t.

(c) Fit a GARCH (or GARCH-type) model to the explosion series in Figure 3.6 and comment.

(d) Analyze the great discoveries and innovations series displayed in Figure 3.7.

3.11. Using Example 3.12 as guide, use the `fGarch` package to fit an AR-GARCH-type model to the returns of either (a) the CAC40, or (b) the NASDAQ. Include a complete set of residual diagnostics.

3.12. Using Example 3.14 and Example 3.15 as guides:

(a) Generate an AR(2) with $\phi_1 = 1$, $\phi_2 = -.9$, and $n = 2048$ and calculate the normalized bispectrum. Comment on the results.

(b) Calculate the normalized bispectrum of the R data set `sunspots` and comment.

(c) For the data generated in (a) and used in (b), perform the Keenan and Tsay tests for nonlinearity and comment.

Chapter 9

Non-Gaussian and Nonlinear State Space Models

The state space model has become a powerful tool for time series modeling and forecasting. Such models, in conjunction with the Kalman filter, have been used in a wide range of applications (see Chapter 3). A *nonlinear state space model* (NLSS) or equivalently a *Hidden Markov Model* (HMM), keeps the hierarchical structure of the Gaussian linear state space model, but removes the limitations of linearity and Gaussianity. An HMM is a discrete time process $\{(X_t, Y_t),\ t \in \mathbb{N}\}$, where $\{X_t,\ t \in \mathbb{N}\}$ is a Markov chain and, conditional on $\{X_t,\ t \in \mathbb{N}\}$, $\{Y_t,\ t \in \mathbb{N}\}$ is a sequence of independent random variables such that the conditional distribution of Y_t only depends on X_t. We denote by $(\mathsf{X}, \mathcal{X})$ the state space of the hidden Markov chain $\{X_t,\ t \in \mathbb{N}\}$ and by $(\mathsf{Y}, \mathcal{Y})$ the state space of the observations.

Of the two processes $\{X_t,\ t \in \mathbb{N}\}$ and $\{Y_t,\ t \in \mathbb{N}\}$, only $\{Y_t,\ t \in \mathbb{N}\}$ is actually observed, so that inference on the parameters of the model must be achieved using $\{Y_t,\ t \in \mathbb{N}\}$. Inference on the latent or state process, $\{X_t,\ t \in \mathbb{N}\}$, is often also of interest. As we shall see, these two statistical objectives are strongly intertwined.

In this chapter, we consider a number of prototype HMMs (used in some of these applications) in order to illustrate the variety of situations; e.g., finite-valued state spaces, nonlinear Gaussian state-space models, conditionally Gaussian state-space models, and so on.

9.1 Definitions and basic properties

9.1.1 Discrete-valued state space HMM

If both X and Y are discrete-valued, the hidden Markov model is said to be *discrete*, which is the case originally considered by Baum and Petrie (1966). Let M be a Markov transition matrix on X, so that for any $x \in \mathsf{X}$, $x' \mapsto M(x,x')$ is a probability on X. Thus, for any $x' \in \mathsf{X}$, $M(x,x') \geq 0$ and $\sum_{x' \in \mathsf{X}} M(x,\ x') = 1$. In the discrete state-space setting, we identify any function $f : \mathsf{X} \to \mathbb{R}$, i.e., $f : x \mapsto f(x)$, with a column vector $f = (f(x))_{x \in \mathsf{X}}$ and any finite measure ξ on X, with a row vector $\xi = (\xi(x))_{x \in \mathsf{X}}$; ξ is a probability if $\sum_{x \in \mathsf{X}} \xi(x) = 1$. Let $\{X_t,\ t \in \mathbb{N}\}$ be a Markov chain with initial distribution ξ and Markov transition matrix M. For $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$ and any $x \in \mathsf{X}$, we

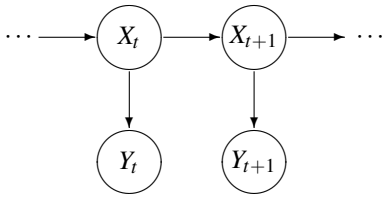


Figure 9.1 *Representation of the dependence structure of a hidden Markov model, where* $\{Y_t,\, t \in \mathbb{N}\}$ *are the observations and* $\{X_t,\, t \in \mathbb{N}\}$ *is the state sequence.*

get

$$\mathbb{E}\left[f(X_t) \,\middle|\, X_{t-1} = x\right] = Mf(x) = \sum_{x' \in \mathsf{X}} M(x,x') f(x') \,.$$

For any integer t and any $f_{t+1} \in \mathbb{F}_b(\mathsf{X}^{t+1}, \mathcal{X}^{\otimes(t+1)})$, the joint distribution of the chain is given by

$$\mathbb{E}_\xi[f_{t+1}(X_0, X_1, \ldots, X_t)] = \sum_{x_0 \in \mathsf{X}} \ldots \sum_{x_t \in \mathsf{X}} \xi(x_0) M(x_0,x_1) \ldots M(x_{t-1},x_t) f_{t+1}(x_0,x_1,\ldots,x_t) \,.$$

The previous identity implies that $\xi_t(f) := \mathbb{E}_\xi[f(X_t)] = \xi M^t f$, where ξ_t denotes the marginal distribution of X_t.

Let G be a Markov transition matrix from X to Y, i.e., for any $x \in \mathsf{X}$, $G(x,\cdot)$ is a probability on Y, so that for any $y \in \mathsf{Y}$, $G(x,y) \geq 0$ and $\sum_{y \in \mathsf{Y}} G(x,y) = 1$. Consider K, the Markov kernel on $\mathsf{X} \times (\mathcal{X} \otimes \mathcal{Y})$ given by

$$K(x; x', y') = M(x,x') G(x',y') \,, \quad (x,x',y') \in \mathsf{X}^2 \times \mathsf{Y} \,.$$

For all $x \in \mathsf{X}$ and $(x',y') \in \mathsf{X} \times \mathsf{Y}$, $K(x; x', y') \geq 0$ and for any $x \in \mathsf{X}$,

$$\sum_{(x',y') \in \mathsf{X} \times \mathsf{Y}} K(x; x', y') = \sum_{x' \in \mathsf{X}} M(x,x') \sum_{y' \in \mathsf{Y}} G(x',y') = \sum_{x' \in \mathsf{X}} M(x,x') = 1 \,.$$

Let ξ be a probability on X. Consider the stochastic process $\{(X_t, Y_t),\, t \in \mathbb{N}\}$ with joint distribution given, for any $t \in \mathbb{N}$ and function $h_{t+1} \in \mathbb{F}_b((\mathsf{X} \times \mathsf{Y})^{t+1}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes(t+1)})$ by

$$\mathbb{E}_\xi[h_{t+1}((X_0,Y_0), \ldots, (X_t,Y_t))] \tag{9.1}$$

$$= \sum_{(x_0,y_0)} \ldots \sum_{(x_t,y_t)} h_{t+1}((x_0,y_0),\ldots,(x_t,y_t)) \xi(x_0,y_0) G(x_0,y_0) \prod_{s=1}^{t} K(x_{s-1}; x_s, y_s) \,.$$

The process $\{(X_t,Y_t),\, t \in \mathbb{N}\}$ is a Markov chain on $\mathsf{X} \times \mathsf{Y}$ with initial distribution ξ_G where $\xi_G(x,y) = \xi(x)G(x,y)$, $(x,y) \in \mathsf{X} \times \mathsf{Y}$ and transition kernel K. The marginal distribution of $\{X_t,\, t \in \mathbb{N}\}$ is obtained by marginalizing with respect to the observations:

$$\mathbb{E}_\xi[f_{t+1}(X_0, X_1, \ldots, X_t)] = \sum_{x_0 \in \mathsf{X}} \ldots \sum_{x_t \in \mathsf{X}} f_{t+1}(x_0,\ldots,x_t) \xi(x_0) \prod_{i=1}^{t} M(x_{i-1},x_i) \,, \tag{9.2}$$

where $f_{t+1} \in \mathbb{F}_b(\mathsf{X}^{t+1}, \mathcal{X}^{\otimes(t+1)})$, showing that $\{X_t,\, t \in \mathbb{N}\}$ is a Markov chain on X with initial distribution ξ and transition kernel M.

On the other hand, let $\{h_0, \ldots, h_t\}$ be a set of functions, $h_i \in \mathbb{F}_b(\mathsf{Y}, \mathcal{Y})$ and $f_{t+1} \in \mathbb{F}_b(\mathsf{X}^{t+1}, \mathcal{X}^{\otimes(t+1)})$. We get

$$\begin{aligned}\mathbb{E}_\xi\Big[\prod_{s=0}^{t} h_s(Y_s) f_{t+1}(X_0,\ldots,X_t)\Big] &= \mathbb{E}_\xi\Big[f_{t+1}(X_0,\ldots,X_t)\mathbb{E}\Big\{\prod_{s=0}^{t} h_s(Y_s) \mid X_0,\ldots,X_t\Big\}\Big] \\ &= \sum_{x_0\in\mathsf{X}} \cdots \sum_{x_t\in\mathsf{X}} \xi(x_0) \prod_{s=1}^{t} M(x_{s-1},x_s) \prod_{s=0}^{t} \sum_{y_s\in\mathsf{Y}} G(x_s,y_s) h_s(y_s)\,,\end{aligned}$$

showing that the components of the vector of observations $(Y_0, \ldots, Y_t)$ are conditionally independent given the state sequence $X_0, \ldots, X_t$ and that the conditional distribution of Y_s given X_s is $G(X_s, \cdot)$:

$$\mathbb{E}\left[\prod_{s=0}^{t} h_s(Y_s) \,\middle|\, X_0,\ldots,X_t\right] = \prod_{i=0}^{t} \mathbb{E}\left[h_s(Y_s) \,\middle|\, X_s\right] = \prod_{s=0}^{t} G h_s(X_s)\,. \tag{9.3}$$

The joint distribution of the sequence of observations $Y_0, \ldots, Y_t$ may be deduced from (9.1) by marginalizing with respect to the state sequence:

$$\mathrm{p}_{\xi,t}(Y_{0:t}) = \sum_{x_0\in\mathsf{X}} \cdots \sum_{x_t\in\mathsf{X}} \xi(x_0) \prod_{s=1}^{t} M(x_{s-1},x_s) G(x_s,Y_s)\,. \tag{9.4}$$

This expression of the joint distribution of the observations might look a little daunting at first sight, because it involves evaluating the joint distribution of the state-sequence and the observations, and then marginalizing the state sequence. If the number of states is m, then the number of state sequences is m^{t+1}, so that the numerical complexity seems to grow exponentially with t. We will see later that the likelihood can be computed with an algorithm whose complexity grows quadratically in the number of states and linearly in the number of time steps t.

The marginal distribution of the t-th observation Y_t is obtained by marginalizing (9.4) with respect to $(Y_0, \ldots, Y_{t-1})$

$$\begin{aligned}\mathrm{p}_{\xi,t}(Y_t) &= \sum_{x_0\in\mathsf{X}} \cdots \sum_{x_t\in\mathsf{X}} \xi(x_0) \prod_{s=1}^{t} M(x_{s-1},x_s) G(x_t,Y_t) \\ &= \sum_{x\in\mathsf{X}} \mathbb{P}_\xi[X_t = x] G(x,Y_t)\,.\end{aligned}$$

The marginal distribution is a mixture of the distributions $\{G(x,Y_t), x \in \mathsf{X}\}$ with weights given by $\{\mathbb{P}_\xi[X_t = x], x \in \mathsf{X}\}$. If the Markov kernel M admits a stationary distribution π, then $\mathbb{P}_\pi[X_t = x] = \pi(x)$, and the weights of the mixture remain constants $\{\pi(x), x \in \mathsf{X}\}$. Such behavior is a key property of HMMs; their marginal distribution is a mixture of state-dependent distributions.

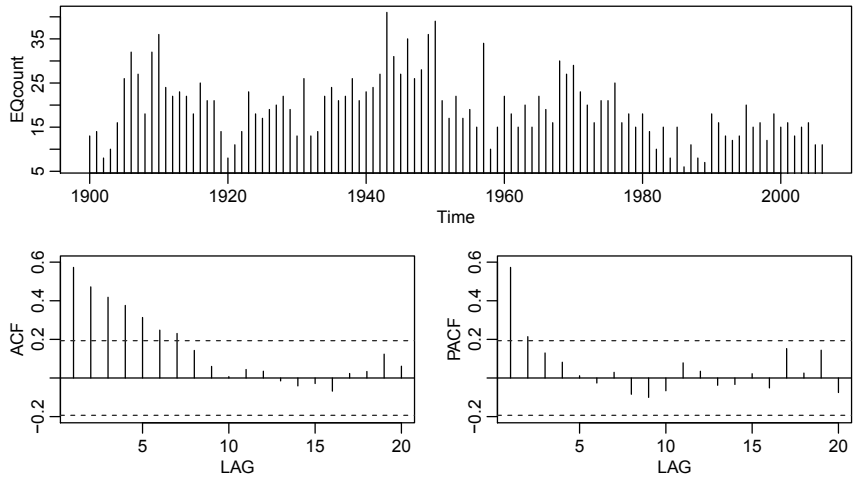


Figure 9.2 *Top: Series of annual counts of major earthquakes (magnitude 7 and above) in the world between 1900-2006. Bottom: Sample ACF and PACF of the square root of the counts.*

Example 9.1 (Number of major earthquakes). Consider the time series of annual counts of major earthquakes displayed in Figure 9.2; see MacDonald and Zucchini (2009, Chapter 1). As discussed in Example 3.3 and Example 3.13 an overdispersed Poisson or Negative Binomial distribution may be a satisfactory model for the marginal distribution; however, given the serial correlation, it is perhaps more important to model joint distributions. As suggested by MacDonald and Zucchini (2009), a simple and convenient way to capture both the marginal distribution and the serial dependence is to consider HMM model with a Poisson distribution. We denote the number of major earthquakes in year t as Y_t, whereas the state, or latent variable, is denoted by X_t. For simplicity, we consider the process $\{X_t,\, t \in \mathbb{N}\}$ to be a two-state Markov chain, $\mathsf{X} = \{1,2\}$, where M is a 2×2 matrix given by,

$$M = \begin{bmatrix} M(1,1) & M(1,2) \\ M(2,1) & M(2,2) \end{bmatrix}$$

with $M(1,1), M(2,2) \in (0,1)$, $M(1,2) = 1 - M(1,1)$, $M(2,1) = 1 - M(2,2)$. The stationary distribution of this Markov chain is given by

$$\pi(1) = \frac{M(2,1)}{2 - M(1,1) - M(2,2)}\,, \quad \text{and} \quad \pi(2) = \frac{M(1,2)}{2 - M(1,1) - M(2,2)}\,.$$

For $x \in \mathsf{X}$, denote λ_x as the parameter of the Poisson distribution:

$$G(x,y) = \frac{(\lambda_x)^y}{y!} \mathrm{e}^{-\lambda_x}\,, \quad y \in \mathbb{N}\,.$$

Assuming that the Markov chain is stationary ($\xi = \pi$), the marginal distribution is a

mixture of Poisson distribution

$$\mathrm{p}_{\pi,t}(Y_t) = \pi(1)G(1,Y_t) + \pi(2)G(2,Y_t) = \pi(1)\frac{(\lambda_1)^{Y_t}}{Y_t!}\mathrm{e}^{-\lambda_1} + \pi(2)\frac{(\lambda_2)^{Y}_t}{Y_t!}\mathrm{e}^{-\lambda_2} .$$

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function. The mean and the variance of $f(Y_t)$ are given by (see Exercise 9.1)

$$\mathbb{E}_\pi[f(Y_t)] = \pi(1)\mu(1) + \pi(2)\mu(2) , \tag{9.5}$$

$$\mathrm{Var}_\pi[f(Y_t)] = \mathbb{E}_\pi[Y_0] + \pi(1)\pi(2)(\mu(2)-\mu(1))^2 \geq \mathbb{E}_\pi[f(Y_0)] , \tag{9.6}$$

where

$$\mu(x) = \sum_{y\in\mathbb{N}} G(x,y)f(y) = \sum_{y\in\mathbb{N}} \frac{\lambda_x^y}{y!} f(y) .$$

The marginal distribution of a two-state Poisson-HMM model is therefore overdispersed compared to the Poisson distribution. Using the conditional independence of the observations and the states (9.3),

$$\mathbb{E}_\pi[f(Y_t)f(Y_{t+k})] = \mathbb{E}_\pi\{\mathbb{E}_\pi\left[f(Y_t)f(Y_{t+k})\,|\,X_t,X_{t+k}\right]\} ,$$

$$\mathbb{E}_\pi[\mu(X_t)\mu(X_{t+k})] = \mathbb{E}_\pi[\mu(X_0)\mu(X_k)] = \sum_{x_0=0}^{1}\sum_{x_k=0}^{1} \pi(x_0)M^k(x_0,x_k)\mu(x_0)\mu(x_k) ,$$

where $M^k(x,x')$ denotes the (x,x')-element of the k-th iterate of the transition matrix M. Therefore, the process $\{f(Y_t), t \in \mathbb{N}\}$ is a covariance stationary process with autocovariance function

$$\begin{aligned}\mathrm{Cov}_\pi(f(Y_0),f(Y_k)) &= \sum_{x_0=0}^{1}\sum_{x_k=0}^{1} \pi(x_0)\{M^k(x_0,x_k)-\pi(x_k)\}\mu(x_0)\mu(x_k) ,\\ &= \pi(1)\pi(2)(\mu(2)-\mu(1))^2(1-M(1,2)-M(2,1))^k .\end{aligned}$$

Therefore, for any function $f : \mathbb{R}_+ \to \mathbb{R}$, a two-state Poisson-HMM has an exponentially decaying autocorrelation function (see Exercise 9.3). It is worthwhile to note that the rate of decay of the autocorrelation does not depend upon the choice of f. If we increase the number of states, more complex dependence structure may be obtained; see Exercise 9.3. ◇

A slightly more general example is when the state is discrete, but the observations take values in a general state space. Let $(\mathsf{Y},\mathcal{Y})$ be a measurable space, and let G be a kernel on $\mathsf{X}\times\mathcal{Y}$ (see Definition 5.2). Denote by K, the Markov kernel on $\mathsf{X}\times(\mathcal{X}\otimes\mathcal{Y})$ given for all $x,x' \in \mathsf{X}$, and $A \in \mathcal{Y}$ by

$$K(x;\{x'\}\times A) = M(x,x')G(x',A) . \tag{9.7}$$

Let ξ be a probability on X. Consider the stochastic process $\{(X_t,Y_t),\ t \in \mathbb{N}\}$ with

joint distribution given, for any $t \in \mathbb{N}$ and function $h_{t+1} \in \mathbb{F}_b((\mathsf{X}\times\mathsf{Y})^{t+1}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes(t+1)})$ by

$$\mathbb{E}_\xi[h_{t+1}(\{(X_s,Y_s)\}_{s=0}^t)] = \sum_{x_0}\cdots\sum_{x_t}\int\cdots\int h_{t+1}(\{(x_s,y_s)\}_{s=0}^t) \times \xi(x_0,\mathrm{d}y_0)G(x_0,\mathrm{d}y_0)\prod_{s=1}^t K(x_{s-1};x_s,\mathrm{d}y_s)\,. \quad (9.8)$$

The process $\{(X_t,Y_t),\, t\in\mathbb{N}\}$ is a Markov chain on $\mathsf{X}\times\mathsf{Y}$ with initial distribution ξ_G where $\xi_G(\{x\}\times A) = \xi(x)G(x,A)$, $x\in\mathsf{X}$ and $A\in\mathcal{Y}$ and transition kernel K. By marginalizing with respect to the observations, (9.8) implies that $\{X_t,\, t\in\mathbb{N}\}$ is a Markov chain with transition kernel M and initial distribution ξ; see (9.2). Similarly, proceeding as in (9.3), the sequence of observations $Y_0,\ldots,Y_t$ are independent conditional to the states.

If, for all $x\in\mathsf{X}$, $G(x,\cdot)$ is absolutely continuous with respect to μ, $G(x,\cdot)\ll\mu(\cdot)$, with transition density function $g(x,\cdot)$. Then, for $A\in\mathcal{Y}$, $G(x,A) = \int_A g(x,y)\,\mu(\mathrm{d}y)$ and the joint transition kernel K can be written as

$$K(x,C) = \iint_C M(x,dx')g(x',y')\,\mu(dy')\,,\quad C\in\mathcal{X}\otimes\mathcal{Y}\,. \quad (9.9)$$

In this case, the joint distribution of the sequence of observations $Y_0,\ldots,Y_t$ has a density with respect to the product measure $\mu^{\otimes(t+1)}$ given by

$$\mathrm{p}_{\xi,t}(Y_{0:t}) = \sum_{x_0\in\mathsf{X}}\cdots\sum_{x_t\in\mathsf{X}}\xi(x_0)\prod_{s=1}^t M(x_{s-1},x_s)g(x_s,Y_s)\,. \quad (9.10)$$

The marginal distribution of the t-th observation Y_t is obtained by marginalizing (9.10) with respect to the observations $(Y_0,\ldots,Y_{t-1})$ and is therefore a mixture of the densities $\{g(x,Y_t), x\in\mathsf{X}\}$.

$$\mathrm{p}_{\xi,t}(Y_t) = \sum_{x\in\mathsf{X}}\mathbb{P}_\xi[X_t=x]g(x,Y_t)\,.$$

If f is a function and ξ is chosen to be the stationary distribution of the Markov chain P (assuming that it exists) then $\mathbb{E}_\pi[f(Y_t)] = \sum_{x\in\mathsf{X}}\pi(x)\mu(f;x)$, where $\mu(f;x)$ is the conditional expectation of $f(Y_t)$ given state x, $\mu(f;x) = \mathbb{E}\left[f(Y_0)\,\middle|\, X_0=x\right]$, $x\in\mathsf{X}$. For instance, if $f(y)=y^2$, then $\mu(f;x)$ equals the conditional second moment. Assume that the number of states, d, is finite. Defining $\Gamma(f) = \mathrm{diag}\{\mu(f;x), x\in\mathsf{X}\}$, the unconditional mean can be written more compactly as $\mathbb{E}_\pi[f(Y_t)] = \pi\Gamma(f)\mathbf{1}$. Furthermore, for $h>0$,

$$\begin{aligned}\mathbb{E}_\pi[f(Y_t)f(Y_{t+h})] &= \sum_{x,x'}\mathbb{E}\left[f(Y_t)f(Y_{t+h})\,\middle|\,(X_t,X_{t+h})=(x,x')\right]\mathbb{P}_\pi[(X_t,X_{t+h})=(x,x')]\\ &= \sum_{x,x'}\mu(f;x)\mu(f;x')\pi(x)P^h(x,x')\,,\end{aligned}$$

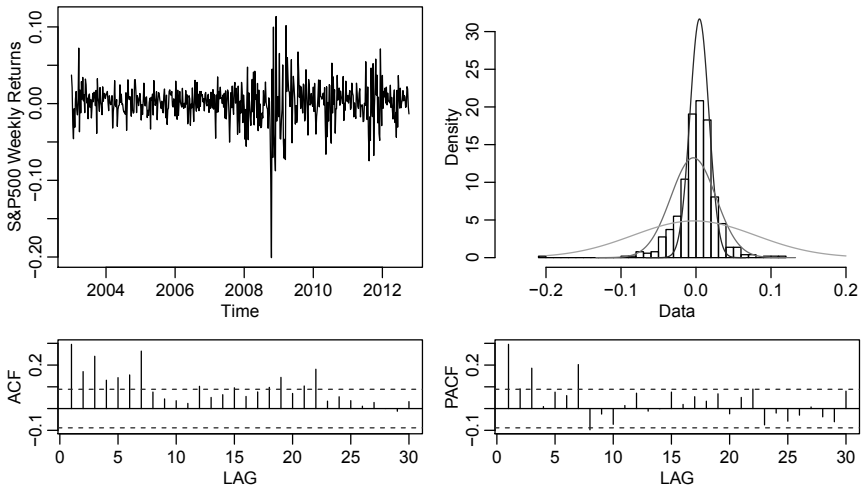


Figure 9.3 *Top: Weekly log-returns of S&P500 from January 2, 2003 to September 28, 2012. Histogram superimposed with a mixture of three Gaussian distributions. Bottom: Sample ACF and PACF of the squares of the log-returns.*

where we have used

$$\begin{aligned}&\mathbb{E}\left[f(Y_t)f(Y_{t+h}) \,\middle|\, (X_t,X_{t+h}) = (x,x')\right] \\ &\qquad = \mathbb{E}\left[f(Y_t) \,\middle|\, X_t = x\right]\mathbb{E}\left[f(Y_{t+h}) \,\middle|\, X_{t+h} = x'\right] = \mu(f;x)\mu(f;x')\,,\end{aligned}$$

which follows from the conditional independence of the observations given the state and $\mathbb{P}_\pi[(X_t,X_{t+h}) = (x,x')] = \mathbb{P}(X_{t+h} = x' \,|\, X_t = x)\,\mathbb{P}_\pi[X_t = x] = \pi(x)P^h(x,x')$. In matrix form, this can be written as

$$\mathbb{E}_\pi[f(Y_t)f(Y_{t+h})] = \pi\Gamma(f)P^h\Gamma(f)\mathbf{1}\,.$$

Finally, the covariance of $f(Y_t)$ and $f(Y_{t+h})$ is

$$\mathrm{Cov}_\pi(f(Y_t),f(Y_{t+h})) = \pi\Gamma(f)P^h\Gamma(f)\mathbf{1} - (\pi\Gamma(f)\mathbf{1})^2\,.$$

Example 9.2 (S&P500 weekly returns). In this example, we consider the weekly S&P500 log-returns from January 3, 2003 until September 28, 2012. The time series is displayed in the upper left of Figure 9.3; the bottom row of the figure shows the sample ACF and PACF of the squared returns indicating that there is dependence among the returns.

Financial markets are usually characterized as *bullish* (most investors expect upward price movement), *neutral* or *bearish* (most investors expect downward price movement). It has also been reported that the equity market returns and volatility tend to move in opposite directions. To assess this assumption, we modeled the marginal distribution of the log-return by a mixture of three Gaussian distributions. We fitted the marginal distributions using the R package `mixtools`, which can be used

to implement MLE to fit the normal mixtures via the EM algorithm. The procedure, however, assumes the data are independent, which is obviously not the case. We will consider a more appropriate analysis later in Example 9.12. Under the assumption of independence, we fitted the observations with three components yielding means $(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (.005, -.003, -.002)$, standard deviations $(\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3) = (.013, .030, .082)$, and mixing probabilities $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (.55, .42, .03)$. A histogram of the data along with the three fitted normals are displayed in the upper right of Figure 9.3. Note that state 1 may be interpreted as the bullish state, with positive mean and a lower volatility. State 2 may be seen as the bearish state with negative mean and comparatively higher volatility. The meaning of state 3 should be interpreted more carefully, since it captures mostly the outliers occurring during the 2008 and 2011 crises.

As previously indicated, a Gaussian mixture model is not appropriate here because the log-returns are serially correlated. The market may stay in a bullish regime for some time before moving to another regime at a later date. As suggested by Rydén et al. (1998), a Hidden Markov model with Gaussian emission probability is a good candidate to capture these stylized facts. ◇

The previous example leads us to examine a more general model. Suppose that $\{X_t,\, t \in \mathbb{N}\}$ is a Markov chain with state space $\mathsf{X} := \{1, \ldots, m\}$ and that the observations $\{Y_t,\, t \in \mathbb{N}\}$, conditional on $\{X_t,\, t \in \mathbb{N}\}$, are independent Gaussian with means $\{\mu_{X_t},\, t \in \mathbb{N}\}$ and variance $\{\sigma^2_{X_t},\, t \in \mathbb{N}\}$. The distribution of the Markov chain is specified by a Markov transition matrix $M = \{M(x,x')\}_{(x,x') \in \mathsf{X}^2}$, which is assumed to have a unique invariant distribution denoted π. Assume for simplicity that the Markov chain is stationary. Suppose the marginal distribution of $\{Y_t,\, t \in \mathbb{N}\}$ is a mixture of m Gaussian distributions with mixing weights $(\pi(1), \ldots, \pi(m))$. The observations may be expressed as $Y_t = \mu_{X_t} + \sigma_{X_t} V_t$, where $\{V_t,\, t \in \mathbb{N}\}$ are i.i.d. $\mathrm{N}(0,1)$. The autocorrelation function of $\{Y_t,\, t \in \mathbb{N}\}$ is given by, for $h > 0$,

$$\frac{\mathrm{Cov}(Y_t, Y_{t+h})}{\mathrm{Var}(Y_t)} = \frac{\pi \Gamma_1 P^h \Gamma_1 \mathbf{1} - (\pi \Gamma_1 \mathbf{1})^2}{\pi \Gamma_2 \mathbf{1} - (\pi \Gamma_1 \mathbf{1})^2} \,,$$

where $\Gamma_p = \mathrm{diag}\{\int y^p \mathfrak{g}(y; \mu_x, \sigma_x^2) \mathrm{d}y,\, x \in \mathsf{X}\}$. For a two-state model, the autocorrelation is given by

$$\frac{\mathrm{Cov}(Y_t, Y_{t+h})}{\mathrm{Var}(Y_t)} = \frac{\pi(1)\pi(2)(\mu_1 - \mu_2)^2}{\pi(1)\sigma_1^2 + \pi(2)\sigma_2^2} \lambda^h \,,$$

where $\lambda := 1 - M(1,2) - M(2,1)$. The process is not autocorrelated if $\mu_1 = \mu_2$. For the squared process, we have

$$\frac{\mathrm{Cov}(Y_t^2, Y_{t+h}^2)}{\mathrm{Var}(Y_t^2)} = \frac{\pi \Gamma_2 P^h \Gamma_2 \mathbf{1} - (\pi \Gamma_2 \mathbf{1})^2}{\pi \Gamma_4 \mathbf{1} - (\pi \Gamma_2 \mathbf{1})^2} \,.$$

For a two-state model, the autocovariance of the squared process is given by

$$\mathrm{Cov}(Y_t^2, Y_{t+h}^2) = \pi(1)\pi(2)(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2)\lambda^h, \quad h > 0 \,.$$

Note that if $\mu_1 = \mu_2$ and $\lambda \neq 0$, the process $\{Y_t, t \in \mathbb{N}\}$ is white noise, but $\{Y_t^2, t \in \mathbb{N}\}$

is autocorrelated. State dependent variances are neither necessary nor sufficient for autocorrelation in the squared process. Even if $\sigma_1 = \sigma_2$, the marginal process shows conditional heteroscedasticity provided that $M(1,1) \neq M(2,1)$. On the other hand, if $M(1,1) = M(2,1)$, the squared process is not autocorrelated even if $\sigma_1 \neq \sigma_2$. These results extend directly to more general families of distributions.

9.1.2 Continuous-valued state-space models

It is not necessary restrict the definition of HMM to discrete state-spaces.

Definition 9.3 (Hidden Markov model). *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces and let M and G denote, respectively, a Markov kernel on $(\mathsf{X}, \mathcal{X})$ and a Markov kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Denote by K the Markov kernel on $\mathsf{X} \times (\mathcal{X} \otimes \mathcal{Y})$ by*

$$K(x; C) = \iint_C M(x, \mathrm{d}x')\, G(x', \mathrm{d}y'), \quad x \in \mathsf{X},\ C \in \mathcal{X} \otimes \mathcal{Y}. \tag{9.11}$$

The Markov chain $\{(X_t, Y_t),\ t \in \mathbb{N}\}$ with Markov transition kernel K and initial distribution $\xi \otimes G$, where ξ is a probability measure on $(\mathsf{X}, \mathcal{X})$, is called a hidden Markov Model *(HMM).*

The definition above specifies the distribution of $\{(X_t, Y_t),\ t \in \mathbb{N}\}$; the term *hidden* is justified because $\{X_t,\ t \in \mathbb{N}\}$ is not observable. As before, we shall denote by $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ the probability measure and corresponding expectation associated with the process $\{(X_t, Y_t),\ t \in \mathbb{N}\}$, respectively.

An HMM is said to be *partially dominated* if there exists a probability measure μ on $(\mathsf{Y}, \mathcal{Y})$ such that for all $x \in \mathsf{X}$, $G(x, \cdot)$ is absolutely continuous with respect to μ, $G(x, \cdot) \ll \mu(\cdot)$, with transition density function $g(x, \cdot)$. Then, for $A \in \mathcal{Y}$, $G(x, A) = \int_A g(x, y)\, \mu(\mathrm{d}y)$ and the joint transition kernel K can be written as

$$K(x; C) = \iint_C M(x, dx')\, g(x', y')\, \mu(dy'), \quad C \in \mathcal{X} \otimes \mathcal{Y}. \tag{9.12}$$

A partially dominated HMM is *fully dominated* if there exists a probability measure λ on $(\mathsf{X}, \mathcal{X})$ such that $\xi \ll \lambda$ and, for all $x \in \mathsf{X}$, $M(x, \cdot) \ll \lambda(\cdot)$ with transition density function $m(x, \cdot)$. Then, for $A \in \mathcal{X}$, $M(x, A) = \int_A m(x, x')\, \lambda(\mathrm{d}x')$ and the joint Markov transition kernel K has a density k with respect to the product measure $\lambda \otimes \mu$

$$k\left(x; x', y'\right) := m(x, x')\, g(x', y'), \quad (x, x', y') \in \mathsf{X}^2 \times \mathsf{Y}. \tag{9.13}$$

Note that for the fully dominated model, we will generally use the notation ξ to denote the *probability density function* of the initial state X_0 (with respect to λ) rather than the distribution itself.

Proposition 9.4. *Let $\{(X_t, Y_t),\ t \in \mathbb{N}\}$ be a Markov chain over the product space $\mathsf{X} \times \mathsf{Y}$ with transition kernel K given by* (9.11)*. Then, for any integer p and any ordered set $\{t_1 < \cdots < t_p\}$ of indices the random variables $Y_{t_1}, \ldots, Y_{t_p}$ are $\mathbb{P}_\xi$-conditionally independent given $(X_{t_1}, X_{t_2}, \ldots, X_{t_p})$, i.e., and all functions $f_1, \ldots, f_p \in \mathbb{F}_b(\mathsf{Y}, \mathcal{Y})$,*

$$\mathbb{E}_\xi \left[\prod_{i=1}^{p} f_i(Y_{t_i}) \;\middle|\; X_{t_1}, \ldots, X_{t_p} \right] = \prod_{i=1}^{p} G f_i(X_{t_i}), \tag{9.14}$$

where $Gf(x) = \int_{\mathsf{Y}} G(x, \mathrm{d}y) f(y)$.

Proof. See Exercise 9.10. ∎

Assume that the HMM is partially dominated (see (9.12)). The joint probability of the unobservable states and observations up to index t is such that for any function $h_{t+1} \in \mathbb{F}_b((\mathsf{X} \times \mathsf{Y})^{t+1}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes(t+1)})$,

$$\mathbb{E}_\xi[h_{t+1}(X_0, Y_0, \ldots, X_t, Y_t)] = \int \cdots \int h_{t+1}(x_0, y_0, \ldots, x_t, y_t) \times \xi(\mathrm{d}x_0) g(x_0, y_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s) g(x_s, y_s) \prod_{s=0}^{t} \mu(\mathrm{d}y_s) \,, \quad (9.15)$$

Marginalizing with respect to the unobservable variables $X_0, \ldots, X_t$, one obtains the joint distribution of the observations $Y_{0:t}$

$$\mathrm{p}_{\xi,t}(Y_{0:t}) = \int \cdots \int \xi(\mathrm{d}x_0) g(x_0, Y_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s) g(x_s, Y_s) \,. \quad (9.16)$$

Example 9.5 (Stochastic volatility). Denote by Y_t the daily *log-returns* of some financial asset. Most models for return data that are used in practice are of a multiplicative form,

$$Y_t = \sigma_t V_t \,, \quad (9.17)$$

where $\{V_t,\, t \in \mathbb{N}\}$ is an i.i.d. sequence and the *volatility process* $\{\sigma_t,\, t \in \mathbb{N}\}$ is a non-negative stochastic process such that V_t is independent of σ_s for all $s \leq t$. It is often assumed that V_t has zero mean and unit variance.

We have already discussed the ARCH/GARCH models in Section 3.5. An alternative to the ARCH/GARCH models is stochastic volatility (SV) models, in which the volatility is a non-linear transform of a hidden linear autoregressive process. The canonical model in SV for discrete-time data has been introduced by Taylor (1982) and worked out since then by many authors; see Hull and White (1987) and Jacquier et al. (1994) for early references and Shephard and Andersen (2009) for an up-to-date survey. In this model, the hidden volatility process, $\{X_t,\, t \in \mathbb{N}\}$, follows a first order autoregression,

$$X_{t+1} = \phi X_t + \sigma W_t \,, \quad (9.18a)$$

$$Y_t = \beta \exp(X_t/2) V_t \,. \quad (9.18b)$$

where $\{W_t,\, t \in \mathbb{N}\}$ is a white Gaussian noise with mean zero and unit variance and $\{V_t,\, t \in \mathbb{N}\}$ is a strong white noise. The error processes $\{W_t,\, t \in \mathbb{N}\}$ and $\{V_t,\, t \in \mathbb{N}\}$ are assumed to be mutually independent and $|\phi| < 1$. As W_t is normally distributed, X_t is also normally distributed. All moments of V_t exist, so that all moments of Y_t in (9.18) exist as well. Assuming that $X_0 \sim \mathrm{N}(0, \sigma^2/(1-\phi^2))$ (the stationary distribution of the Markov chain) the kurtosis[1] of Y_t is given by (see Exercise 9.12)

$$\kappa_4(Y) = \kappa_4(V) \exp(\sigma_X^2), \quad (9.19)$$

[1] For an integer m and a random variable U, $\kappa_m(U) := \mathbb{E}[|U|^m] / (\mathbb{E}[|U|^2])^{m/2}$. Typically, κ_3 is called *skewness* and κ_4 is called *kurtosis*.

where $\sigma_X^2 = \sigma^2/(1-\phi^2)$ is the (stationary) variance of X_t. Thus $\kappa_4(Y_t) > \kappa_4(V_t)$, so that if $V_t \sim \mathrm{N}(0,1)$, the distribution of Y_t is leptokurtic. The autocorrelation function of $\{Y_t^{2m}, t \in \mathbb{N}\}$ for any integer m is given by

$$\mathrm{Cor}(Y_t, Y_{t+h}) = \frac{\exp(m^2\sigma_X^2\phi^h) - 1}{\kappa_{4m}(V)\exp(m^2\sigma_X^2) - 1}\,, \quad h \in \mathbb{N}\,. \tag{9.20}$$

The decay rate of the autocorrelation function is faster than exponential at small time lags and then stabilizes to ϕ for large lags. ◇

Example 9.6 (NGM model). We consider the univariate model introduced in Netto, Gimeo, and Mendes (1978)—hereafter referred to as the *NGM model*—discussed by Kitagawa (1987) and Carlin et al. (1992), given, in state-space form, by

$$X_t = F_t^\theta(X_{t-1}) + W_t \quad \text{and} \quad Y_t = H_t(X_t) + V_t\,, \tag{9.21}$$

with

$$F_t^\theta(X_{t-1}) = \alpha X_{t-1} + \beta X_{t-1}/(1+X_{t-1}^2) + \gamma\cos[1.2(t-1)]\,, \tag{9.22a}$$

$$H_t(X_t) = X_t^2/20\,, \tag{9.22b}$$

where $X_0 \sim \mathrm{N}(\mu_0, \sigma_0^2)$, with $W_t \sim \text{iid}\,\mathrm{N}(0, \sigma_w^2)$ independent of $V_t \sim \text{iid}\,\mathrm{N}(0, \sigma_v^2)$ and each sequence independent of X_0. Figure 9.4 shows a typical data sequence Y_t and the corresponding state process X_t with all the variances equal to unity and, as in Kitagawa (1987) and Carlin et al. (1992), $\theta = (\alpha = .5, \beta = 25, \gamma = 8)$. Additionally, Figure 9.4 demonstrates the nonlinearity by exhibiting a scatterplot of the observations versus the states, and a phase space trajectory of the states that demonstrates that the states are bifurcating near ± 10.

Note that, in this case, there is no closed form for the covariance of the observations. However, because of the nonlinearity of the processes, the autocovariance function contains little information about the dynamics of the Y_t. In addition, the marginal distribution of the observations is highly complex and no longer known. This model has become a standard model for testing numerical procedures and is used throughout Chapter 12. ◇

9.1.3 Conditionally Gaussian linear state-space models

Conditionally Gaussian linear state-space models belong to a class of models that we will refer to as *hierarchical hidden Markov models*, whose dependence structure is depicted in Figure 9.5. In such models the variable I_t, which is the highest in the hierarchy, influences both the transition from W_{t-1} to W_t as well as the observation Y_t.

Conditionally Gaussian models related to the previous example are also commonly used to approximate non-Gaussian state-space models. Imagine that we are interested in the linear model given by (2.1)–(2.2) with both noise sequences still being i.i.d. but at least one of them with a non-Gaussian distribution. Assuming a very

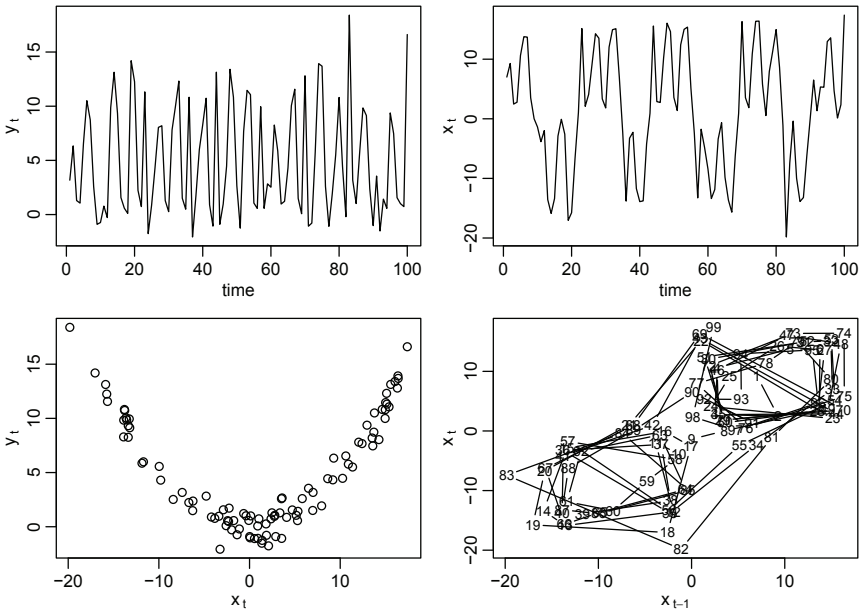


Figure 9.4 *Typical realization of the observations (Y_t) and state process (X_t), for $t = 1,\ldots,100$, generated from the model* (9.21)*. The bottom row shows the quadratic relationship between the observations and states, and a phase space trajectory of the states indicating the bifurcating dynamics of the process.*

general form of the noise distribution would directly lead us into the world of (general) continuous state-space HMMs. As a middle ground, however, we can assume that the distribution of the noise is a mixture of Gaussians.

Let $\{I_t,\, t \in \mathbb{N}\}$ be a sequence of random variables taking values in a set $\mathbb{I}$, which can be finite or infinite. We often refer to these variables as the *indicator variables* when $\mathbb{I}$ is finite. To model non-Gaussian system dynamics, we will typically model the dynamic of the partial state sequence $\{W_t,\, t \in \mathbb{N}\}$ as follows

$$W_{t+1} = \mu_W(I_{t+1}) + A(I_{t+1})W_t + R(I_{t+1})W_t\ , \quad W_t \sim \mathrm{N}(0, I)\ ,$$

where, μ_W, A and R are respectively vector-valued and matrix-valued functions of suitable dimensions on $\mathbb{I}$. When $\mathbb{I} = \{1,\ldots,r\}$ is finite, the distribution of the noise, $\mu_W(I_{t+1}) + R(I_{t+1})W_t$, driving the state equation is a finite mixture of multivariate Gaussian distributions. Similarly, the observation equation is modeled by

$$Y_t = \mu_Y(I_t) + B(I_t)W_t + S(I_t)V_t\ , \quad V_t \sim \mathrm{N}(0, I)\ ,$$

where μ_Y, B and S are respectively vector-valued and matrix-valued functions. Here again, when $\mathbb{I} = \{1,\ldots,r\}$ is finite, then the distribution of the observation noise $\mu_Y(I_t) + S(I_t)V_t$ is a finite mixture of multivariate distribution, allowing us to model outliers, for example. Since B is also a function of I, this model may accommodate changes in the way the state is observed.

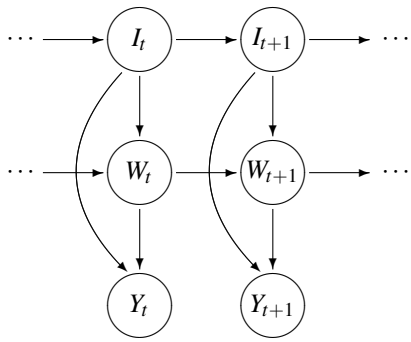


Figure 9.5: *Graphical representation of the dependence structure of a hierarchical HMM.*

Example 9.7 (Level shifts and outliers). Gerlach et al. (2000) have considered the following model to analyze data with level shifts and outliers in both the observations and innovations:

$$Y_t = W_t + \sigma_V I_{t,1} V_t \,, \tag{9.23}$$

$$W_t - \mu_t = \sum_{i=1}^{p} \phi_i (W_{t-i} - \mu_{t-i}) + \sigma_Z I_{t,2} Z_t \,, \tag{9.24}$$

$$\mu_t = \mu_{t-1} + \sigma_W I_{t,3} W_t \,, \tag{9.25}$$

where $\{(V_t, Z_t, W_t), t \in \mathbb{Z}\}$ is an i.i.d. sequence of Gaussian vectors with zero mean and identity covariance, and $\{(I_{t,1}, I_{t,2}, I_{t,3}), t \in \mathrm{Z}\}$ is i.i.d. taking values in $\mathbb{I}$, which is typically discrete. The time series $\{W_t,\ t \in \mathbb{Z}\}$ has mean level μ_t and is generated by an autoregressive model with coefficients $\phi = (\phi_1,\ \ldots,\ \phi_p)$.

If $I_{t,1}$ and $I_{t,2}$ are equal to 1, then the observations are a noisy version of an AR(p) process; recall Example 2.2. Observational outliers are modeled by assuming that $I_{t,1}$ takes some large values (like 10, 20). Similarly, innovation outliers are modeled by large values of $I_{t,2}$. If $I_{t,3} = 0$, then $\mu_t = \mu_{t-1}$. Level shifts occur at time points t for which $I_{t,3} \neq 0$. ◇

Example 9.8 (Stochastic volatility cont.). Another example of the use of mixtures is in the observational noise of the SVM, (9.18),

$$X_t = \phi X_{t-1} + \sigma W_t \,, \tag{9.26a}$$

$$\ln Y_t^2 = \beta + X_t + \ln V_t^2 \,, \tag{9.26b}$$

where $W_t \sim$ iid $\mathrm{N}(0,1)$, but where now, the observational noise, V_t, is not assumed to be normal. The assumption that the V_t are normal comes from the original ARCH model, which is an assumption that is typically violated empirically. Under the normal assumption, $\ln V_t^2$ is the log of a χ_1^2 random variable with density given by

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(e^x - x\right) \right\} \,, \quad -\infty < x < \infty \,. \tag{9.27}$$

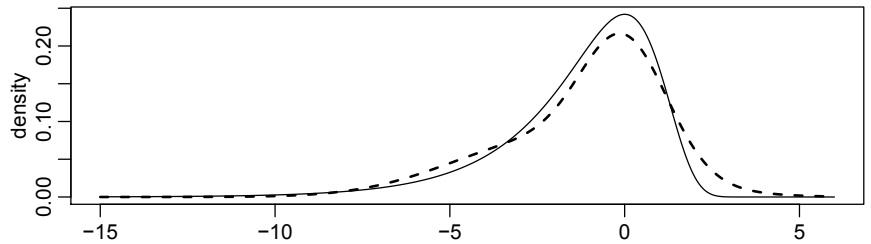


Figure 9.6 *Density of the log of a χ^2_1 as given by (9.27) (solid line) and a fitted normal mixture (dashed line) from Shumway and Stoffer (2011, Example 6.18).*

The mean of the distribution is $-(\gamma + \ln 2)$, where $\gamma \approx 0.5772$ is Euler's constant, and the variance of the distribution is $\pi^2/2$. It is a highly skewed density but it is, of course, not flexible because the distribution is fixed; i.e., there are no parameters to be estimated.

To avoid having a fixed observational noise distribution, Kim and Stoffer (2008) and Shumway and Stoffer (2011, Chapter 6) assumed that the observational noise in (9.26b) is a mixture of two normals with parameters to be estimated. That is,

$$\ln V_t^2 = I_t Z_{t,0} + (1 - I_t) Z_{t,1} \,, \tag{9.28}$$

where $I_t \sim$ iid Ber(π), with $\pi \in [0,1]$, $Z_{t,0} \sim$ iid N$(0, \sigma_0^2)$, and $Z_{t,1} \sim$ iid N(μ_1, σ_1^2). The advantage to this model is that it is easy to fit because it uses conditional normality and there are three additional parameters to provide flexibility in the analysis. Figure 9.6 compares the $\ln \chi^2_1$ density to a fitted mixture distribution taken from Shumway and Stoffer (2011, Example 6.18). Note that the mixture distribution is able to accommodate kurtosis when the volatility is large. ◇

9.1.4 Switching processes with Markov regimes

Markov-switching models perhaps constitute the most significant generalization of HMMs. In such models, the conditional distribution of Y_{t+1}, given all the past variables, depends not only on X_{t+1} but also on Y_t (and possibly more lagged Y-variables). Thus, conditional on the state sequence $\{X_t,\, t \in \mathbb{N}\}$, $\{Y_t,\, t \in \mathbb{N}\}$ forms a (non-homogeneous) Markov chain. Graphically, this is represented as in Figure 9.7. In state-space form, a Markov-switching model may be written as

$$X_{t+1} = a_t(X_t, W_t) \,, \tag{9.29}$$

$$Y_{t+1} = b_t(X_{t+1}, Y_t, V_{t+1}) \,. \tag{9.30}$$

We can even go a step further and assume that $\{(X_t, Y_t),\, t \in \mathbb{N}\}$ jointly forms a Markov chain, but that only $\{Y_t,\, t \in \mathbb{N}\}$ is actually observed.

A switching linear autoregression is a model of the form

$$Y_t = \mu(I_t) + \sum_{i=1}^{p} a_i(I_t)(Y_{t-i} - \mu(I_{t-i})) + \sigma(I_t) V_t \,, \qquad p \geq 1 \,, \tag{9.31}$$

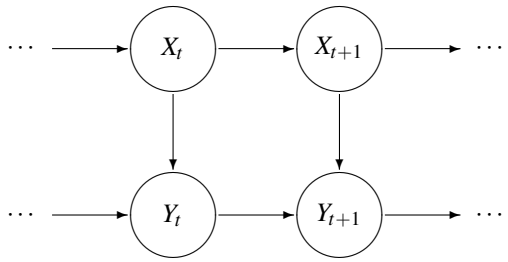


Figure 9.7 *Graphical representation of the dependence structure of a Markov-switching model, where* $\{Y_t,\, t \in \mathbb{N}\}$ *is the observable process and* $\{X_t,\, t \in \mathbb{N}\}$ *is the hidden chain.*

where $\{I_t,\, t \in \mathbb{N}\}$, called the *regime*, is a Markov chain on a finite state space $\mathbb{I} = \{1,2,\ldots,r\}$, and $\{V_t,\, t \in \mathbb{N}\}$ is white noise independent of the regime; the functions $\mu : \mathbb{I} \to \mathbb{R}$, $a_i : \mathbb{I} \to \mathbb{R}$, $i = 1,\ldots,r$, and $\sigma : \mathbb{I} \to \mathbb{R}$ describe the dependence of the parameters on the realized regime.

This model can be rewritten in state-space form as follows. Let

$$\begin{aligned} \boldsymbol{Y}_t &= [Y_t, Y_{t-1}, \ldots, Y_{t-p+1}]' \,, \\ \boldsymbol{I}_t &= [I_t, I_{t-1}, \ldots, I_{t-d+1}]' \,, \\ \mu(\boldsymbol{I}_t) &= [\mu(I_t), \ldots, \mu(I_{t-p+1})]' \,, \\ \boldsymbol{V}_t &= [V_t, 0, \ldots, 0]' \,, \end{aligned}$$

and denote by $A(i)$ the $p \times p$ companion matrix associated with the autoregressive coefficients of the state i,

$$A(i) = \begin{bmatrix} a_1(i) & a_2(i) & \ldots & \ldots & a_p(i) \\ 1 & 0 & & & 0 \\ 0 & 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{bmatrix} . \tag{9.32}$$

The stacked observation vector $\boldsymbol{Y}_t$ then satisfies

$$\boldsymbol{Y}_t = \mu(I_t) + A(I_t)\,(\boldsymbol{Y}_{t-1} - \mu(\boldsymbol{I}_{t-1})) + \sigma(I_t)\boldsymbol{V}_t \,. \tag{9.33}$$

Note that the model is a random coefficient vector autoregression as discussed in Chapter 4.

Example 9.9 (Influenza mortality). In Example 3.1, we discussed the monthly pneumonia and influenza mortality series shown in Figure 3.1. We pointed out the non-reversibility of the series, which rules out the possibility that the data are generated by a linear Gaussian process. In addition, note that the series is irregular, and while mortality is highest during the winter, the peak does not occur in the same month each year. Moreover, some seasons have very large peaks, indicating flu epidemics, whereas other seasons are mild. In addition, it can be seen from Figure 3.1

that there is a slight negative trend in the data set (this is best seen by focusing on the troughs), indicating that flu prevention is getting better over the eleven year period.

Although it is not necessary, to ease the discussion, we focus on the differenced data, which will remove the trend. In this case, we denote $Y_t = \nabla \text{flu}_t$, where flu_t represents the data discussed in Example 3.1. Shumway and Stoffer (2011, Example 5.6) fit a threshold model to Y_t, but we might also consider a switching autoregessive model given in (9.31) or (9.33) where there are two hidden regimes, one for epidemic periods and one for more mild periods. In this case, the model is given by

$$Y_t = \begin{cases} \phi_0^{(1)} + \sum_{j=1}^{p} \phi_j^{(1)} Y_{t-j} + \sigma^{(1)} Z_t \,, & \text{for } I_t = 1 \,, \\ \phi_0^{(2)} + \sum_{j=1}^{p} \phi_j^{(2)} Y_{t-j} + \sigma^{(2)} Z_t \,, & \text{for } I_t = 2 \,, \end{cases} \tag{9.34}$$

where $Z_t \sim$ iid $\mathrm{N}(0,1)$, and I_t is a hidden, two-state Markov chain. ◇

9.2 Filtering and smoothing

Statistical inference for nonlinear state-space models involves computing the *posterior distribution* of a collection of state variables $X_{s:s'} := (X_s, \ldots, X_{s'})$, with $s < s'$ conditioned on a batch of observations, $Y_{0:t} = (Y_0, \ldots, Y_t)$, which we denote by $\phi_{\xi,s:s'|t}$ (where ξ is the initial distribution), the dependence on the observations being implicit for ease of notation. Specific problems include *filtering*, which corresponds to $s = s' = t$, *fixed lag smoothing*, when $s = s' = t - L$ and fixed interval smoothing, if $s = 0$ and $s' = t$ (see Section 2.2).

Definition 9.10 (Smoothing, filtering, prediction). *For non-negative indices s, t, and n with $t \geq s$, and any initial distribution ξ on $(\mathsf{X}, \mathcal{X})$, denote by $\phi_{\xi,s:t|n}$ (the dependence in the observations is implicit to avoid overloading the notation) the conditional distribution of $X_{s:t}$ given $Y_{0:n}$. Specific choices of s,t and n give rise to several particular cases of interest:*

Joint Smoothing: $\phi_{\xi,0:n|n}$, *for* $n \geq 0$;

(Marginal) Smoothing: $\phi_{\xi,t|n}$ *for* $0 \leq t \leq n$;

Prediction: $\phi_{\xi,t+1|t}$ *for* $t \geq 0$;

***p*-step Prediction:** $\phi_{\xi,t+p|t}$ *for* $t, p \geq 0$.

Filtering: $\phi_{\xi,t|t}$ *for* $t \geq 0$; *Because the use of filtering will be preeminent in the following, we shall most often abbreviate* $\phi_{\xi,t|t}$ *to* $\phi_{\xi,t}$.

Despite the apparent simplicity of the above problems, the smoothing distribution can be computed in closed form only in very specific cases, principally, the linear Gaussian model (see Section 2.2) and the discrete-valued Hidden Markov model (where the state $\{X_t,\ t \in \mathbb{N}\}$ takes its values in a finite alphabet).

9.2.1 Discrete-valued state-space HMM

We denote by $\phi_{\xi,t}$ the filtering distribution, i.e., the distribution of X_t given the observations up to time t, $Y_{0:t}$. To simplify the notations, the dependence of the filtering distribution with respect to the observations is implicit.

Denote by $\gamma_{\xi,t}$ the joint distribution of the state X_t and the observations $Y_{0:t}$:

$$\gamma_{\xi,t}(x_t) = \sum_{x_0} \cdots \sum_{x_{t-1}} \xi(x_0) g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, x_s) \,,$$

where we have set, for $s \in \mathbb{N}$

$$g_s(x_s) = g(x_s, Y_s) \quad \text{and} \quad Q_s(x_{s-1}, x_s) = M(x_{s-1}, x_s) g(x_s, Y_s) \,. \tag{9.35}$$

This equation may be rewritten in matrix form as follows

$$\gamma_{\xi,t} = \xi Q_0 Q_1 \cdots Q_t \,,$$

where $Q_0 = \mathrm{diag}\{g(x, Y_0), x \in \mathsf{X}\}$ and

$$Q_s = M \,\mathrm{diag}\{g(x, Y_s), x \in \mathsf{X}\}, \quad \text{for } s \geq 1 \,. \tag{9.36}$$

This distribution may be computed recursively as follows

$$\gamma_{\xi,0}(x_0) = \xi(x_0) g_0(x_0) \quad \text{and} \quad \gamma_{\xi,t}(x_t) = \sum_{x_{t-1} \in \mathsf{X}} \gamma_{\xi,t-1}(x_{t-1}) Q_t(x_{t-1}, x_t) \tag{9.37}$$

or equivalently in matrix form $\gamma_{\xi,0} = \xi Q_0$ and for $t \geq 1$, $\gamma_{\xi,t} = \gamma_{\xi,t-1} Q_t$. The computational complexity grows like the square of the number of states. The joint distribution of $(Y_0, \ldots, Y_t)$ may be obtained by marginalizing the joint distribution $\gamma_{\xi,t}$ of $(Y_0, \ldots, Y_t, X_t)$ with respect to the state X_t, i.e., $\mathrm{p}_{\xi,t}(Y_{0:t}) = \sum_{x_t \in \mathsf{X}} \gamma_{\xi,t}(x_t)$ or, in matrix form, $\mathrm{p}_{\xi,t}(Y_{0:t}) = \gamma_{\xi,t} \mathbf{1}$. The filtering distribution is the conditional distribution of the state X_t given $(Y_0, \ldots, Y_t)$. It is obtained by dividing the joint distribution of $(X_t, Y_0, \ldots, Y_t)$ by $\mathrm{p}_{\xi,t}(Y_{0:t})$,

$$\phi_{\xi,t}(x_t) = \frac{\gamma_{\xi,t}(x_t)}{\sum_{x_t \in \mathsf{X}} \gamma_{\xi,t}(x_t)} \,, \tag{9.38}$$

or in matrix form $\phi_{\xi,t} = \gamma_{\xi,t} / \gamma_{\xi,t} \mathbf{1}$. By plugging the recursion (9.37), the filtering distribution can thus be updated recursively as follows

$$\phi_{\xi,t}(x_t) = \frac{\sum_{x_{t-1} \in \mathsf{X}} \phi_{\xi,t-1}(x_{t-1}) Q_t(x_{t-1}, x_t)}{\sum_{(x_{t-1}, x'_{t-1}) \in \mathsf{X}^2} \phi_{\xi,t-1}(x_{t-1}) Q_t(x_{t-1}, x'_t)} \,. \tag{9.39}$$

In matrix form, this recursion reads

$$\phi_{\xi,t} = \frac{\phi_{\xi,t-1} Q_t}{\phi_{\xi,t-1} Q_t \mathbf{1}} \,.$$

Algorithm 9.1 (Forward Filtering)

Initialization: For $x \in \mathsf{X}$,

$$\phi_{\xi,0|-1}(x) = \xi(x) .$$

Forward Recursion: For $t = 0, \ldots, n$,

$$c_{\xi,t} = \sum_{x \in \mathsf{X}} \phi_{\xi,t|t-1}(x) g_t(x) , \tag{9.40}$$

$$\phi_{\xi,t}(x) = \phi_{\xi,t|t-1}(x) g_t(x) / c_{\xi,t} , \tag{9.41}$$

$$\phi_{\xi,t+1|t}(x) = \sum_{x' \in \mathsf{X}} \phi_{\xi,t}(x') M(x,x') , \tag{9.42}$$

for each $x \in \mathsf{X}$.

The algorithm is summarized in Algorithm 9.1, which in the Rabiner (1989) terminology, corresponds to the normalized forward recursion. The computational cost of filtering is thus proportional to n, the number of observations, and scales like $|\mathsf{X}|^2$ (squared cardinality of the state space X) because of the $|\mathsf{X}|$ vector matrix products corresponding to (9.42).

The predictive distribution of the observation Y_t given $Y_{0:t-1}$ is equal to the ratio

$$\frac{\mathrm{p}_{\xi,t}(Y_{0:t})}{\mathrm{p}_{\xi,t-1}(Y_{0:t-1})} = \frac{\sum_{x_t \in \mathsf{X}} \gamma_{\xi,t}(x_t)}{\sum_{x_{t-1} \in \mathsf{X}} \gamma_{\xi,t-1}(x_{t-1})} = \sum_{(x_{t-1},x_t) \in \mathsf{X}^2} \phi_{\xi,t-1}(x_{t-1}) Q_t(x_{t-1},x_t) . \tag{9.43}$$

or in matrix form

$$\frac{\mathrm{p}_{\xi,t}(Y_{0:t})}{\mathrm{p}_{\xi,t-1}(Y_{0:t-1})} = \phi_{\xi,t-1} Q_t \mathbf{1} .$$

The likelihood of $n+1$ the observations may therefore be written as

$$\begin{aligned} \mathrm{p}_{\xi,n}(Y_{0:n}) &= \mathrm{p}_{\xi,0}(Y_0) \prod_{t=1}^{n} \frac{\mathrm{p}_{\xi,t}(Y_{0:t})}{\mathrm{p}_{\xi,t-1}(Y_{0:t-1})} \\ &= \mathrm{p}_{\xi,0}(Y_0) \prod_{t=1}^{n} \sum_{(x_{t-1},x_t) \in \mathsf{X}^2} \phi_{\xi,t-1}(x_{t-1}) Q_t(x_{t-1},x_t) . \end{aligned} \tag{9.44}$$

In matrix form, the likelihood may be expressed as

$$\mathrm{p}_{\xi,n}(Y_{0:n}) = \mathrm{p}_{\xi,0}(Y_0) \prod_{t=1}^{n} \phi_{\xi,t-1} Q_t \mathbf{1} .$$

The complexity to evaluate this joint distribution grows linearly with the number of observations n and quadratically with the number of states m, whereas the complexity of the direct evaluation of the likelihood (summing up on all the possible sequences of states) grows exponentially fast $O(m^n)$. The direct evaluation of the likelihood is

therefore manageable even when the number of observations is large, which enables likelihood inference. We will discuss this issue in depth in Chapter 12.

The filtering recursion yields the probability distribution of the state X_t given the observations up to time t. When analyzing a time series by batch, the inference of the state X_t that incorporates all the observations $(Y_0, \ldots, Y_n)$ is in general preferable. Such probability statements are given by the fixed interval smoothing probabilities. To simplify the derivations, we denote by $p_\xi(x_{s:t}, y_{s':t'})$ the density with respect to the counting measure of the vector $(X_{s:t}, Y_{s':t'})$. Note first that, for any $s \in \{0, \ldots, n-1\}$,

$$\begin{aligned} p_\xi(x_s|x_{s+1:n}, y_{0:n}) &= \frac{p_\xi(y_{0:s}, x_s, x_{s+1}, y_{s+1:n}, x_{s+2:n})}{\sum_{x'_s \in \mathsf{X}} p_\xi(y_{0:s}, x'_s, x_{s+1}, y_{s+1:n}, x_{s+2:n})} \\ &\overset{(1)}{=} \frac{p_\xi(y_{0:s}, x_s, x_{s+1})}{\sum_{x'_s \in \mathsf{X}} p_\xi(y_{0:s}, x'_s, x_{s+1})} \\ &\overset{(2)}{=} \frac{\phi_{\xi,s}(x_s) M(x_s, x_{s+1})}{\sum_{x'_s \in \mathsf{X}} \phi_{\xi,s}(x'_s) M(x'_s, x_{s+1})} = p_\xi(x_s|x_{s+1}, y_{0:s}) \,, \end{aligned}$$

where (1) follows from

$$p_\xi(y_{s+1:t}, x_{s+2:n}|y_{0:s+1}, x_{0:s+1}) = p_\xi(y_{s+1:t}, x_{s+2:n}|x_{s+1})$$

which cancels in the numerator and the denominator, (2) from

$$p_\xi(x_{s+1}|x_s, y_{0:s}) = p(x_{s+1}|x_s) = M(x_s, x_{s+1}) \,,$$

where we have used (9.7) and the fact that $\{(X_t, Y_t),\, t \in \mathbb{N}\}$ is a Markov chain. This shows that $\{X_{n-s}, s \in \{0, 1, \ldots, n\}\}$ conditioned on the observations $Y_{0:n}$ is a Markov chain, with initial distribution $\phi_{\xi,n}$ and transition kernel $B_{\phi_{\xi,s}}$ where for any measure η on X, B_η is the Markov matrix given by

$$B_\eta(x, x') := \frac{\eta(x')\, m(x', x)}{\sum_{x'' \in \mathsf{X}} \eta(x'')\, m(x'', x)} \,. \tag{9.45}$$

In matrix form, the backward kernel may be written as

$$B_\eta = \mathrm{diag}(M' D_\eta \mathbf{1})^{-1} M' D_\eta \,, \quad D_\eta := \mathrm{diag}(\eta(x), x \in \mathsf{X}) \,.$$

Here, $\mathbf{1}$ denotes the matrix with all entries equal to one. For any integers $n > 0$, $s \in \{0, \ldots, n-1\}$, the posterior distribution $\phi_{\xi,s:n|n}$ may be expressed as

$$\phi_{\xi,s:n|n}(x_{s:n}) = \phi_{\xi,n}(x_n) B_{\phi_{\xi,n-1}}(x_n, x_{n-1}) \ldots B_{\phi_{\xi,s}}(x_{s+1}, x_s) \,. \tag{9.46}$$

In particular, the marginal smoothing distribution $\phi_{\xi,s|n}$ may be expressed in matrix form as

$$\phi_{\xi,s|n} = \phi_{\xi,n} B_{\phi_{\xi,n-1}} B_{\phi_{\xi,n-2}} \ldots B_{\phi_{\xi,s}} \,.$$

Algorithm 9.2 (Backward marginal smoothing)

Given stored values of $\phi_{\xi,0},\ldots,\phi_{\xi,n}$ and starting from n, backwards in time.

Initialization: For $x \in \mathsf{X}$,

$$\phi_{\xi,n|n}(x) = \phi_{\xi,n}(x) .$$

Backward Recursion: For $t = n-1,\ldots,0$,

- Compute the backward transition kernel according to

$$B_{\phi_{\xi,t}}(x,x') = \frac{\phi_{\xi,t}(x')M(x',x)}{\sum_{x''\in\mathsf{X}} \phi_{\xi,t}(x'')M(x'',x)}$$

  for $(x,x') \in \mathsf{X}\times\mathsf{X}$.
- Compute

$$\phi_{\xi,t|n}(x) = \sum_{x'\in\mathsf{X}} \phi_{\xi,t+1|n}(x')B_{\phi_{\xi,t}}(x',x) .$$

  for $(x,x') \in \mathsf{X}\times\mathsf{X}$.

The marginal smoothing distribution can be generated recursively, backwards in time as follows

$$\phi_{\xi,s|n} = \phi_{\xi,s+1|n}B_{\phi_{\xi,s}} . \tag{9.47}$$

This recursion, summarized in Algorithm 9.2, is the *forward-backward* algorithm or the *Baum-Welch* algorithm for discrete Hidden Markov Models. In the forward pass, the filtering distributions $\{\phi_{\xi,t}, t \in \{0,\ldots,n\}\}$ are computed and stored. In the backward pass, these filtering distributions are corrected by recursively applying the backward kernels.

When X is finite, it turns out that it is also possible to determine the path $\hat{X}_{0:n}$ which maximizes the joint smoothing probability

$$\hat{X}_{0:n} := \underset{x_{0:n}\in\mathsf{X}^{n+1}}{\arg\max}\, \mathbb{P}_\xi(X_{0:n} = x_{0:n} \mid Y_{0:n}) = \underset{x_{0:n}\in\mathsf{X}^{n+1}}{\arg\max}\, \phi_{\xi,0:n|n}(x_{0:n}) . \tag{9.48}$$

Solving the maximization problem (9.48) over all possible state sequences $x_{0:m}$ by brute force would involve m^{n+1} function evaluations, which is clearly not feasible except for small n. The algorithm that makes it possible to efficiently compute the *a posteriori most likely sequence of states* is known as the *Viterbi algorithm*, which is based on the well-known *dynamic programming* principle. The logarithm of the joint smoothing distribution may be written as

$$\begin{aligned}\ln \phi_{\xi,0:t|t}(x_{0:t}) = (\ell_{\xi,t-1} - \ell_{\xi,t}) \\ + \ln \phi_{\xi,0:t-1|t-1}(x_{0:t-1}) + \ln m(x_{t-1},x_t) + \ln g_t(x_t) ,\end{aligned} \tag{9.49}$$

where $\ell_{\xi,t}$ denotes the log-likelihood of the observations up to index t. The salient feature of (9.49) is that, except for a constant term that does not depend on the state

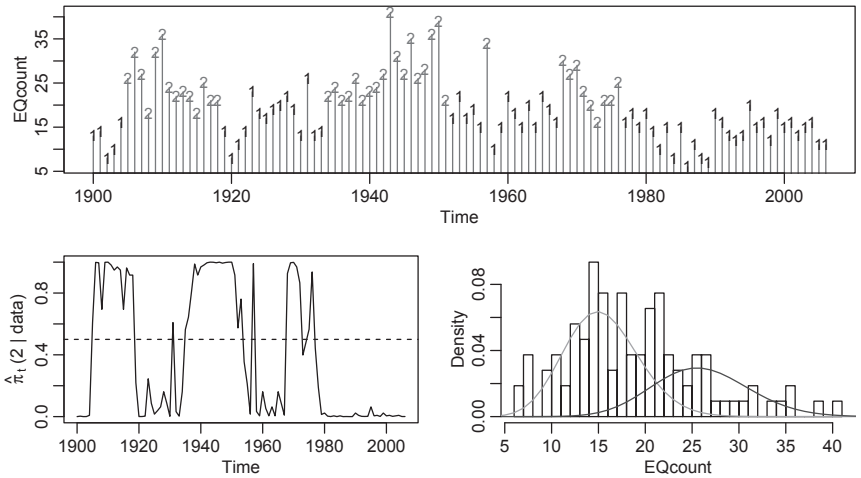


Figure 9.8 *Top: Earthquake count data and estimated states. Bottom left: Smoothing probabilities. Bottom right: Histogram of the data with the two estimated Poisson densities superimposed (solid lines).*

sequence (on the right-hand side of the first line), the *a posteriori* log-probability of the subsequence $x_{0:t}$ is equal to that of $x_{0:t-1}$ up to terms that only involve the pair (x_{t-1}, x_t). Define

$$\mu_t(x) = \max_{x_{0:t-1} \in \mathsf{X}^t} \ln \phi_{\xi,0:t|t}(x_{0:t-1}, x) + \ell_{\xi,t} \,, \tag{9.50}$$

that is, up to a number independent of the state sequence, the maximal conditional probability (on the log scale) of a sequence up to time t and ending with state $x \in \mathsf{X}$. Also define $b_t(x)$ to be that value in X of x_{t-1} for which the optimum is achieved in (9.50); in other words, $b_t(x)$ is the second final state in an optimal state sequence of length $t+1$ and ending with state x. Using (9.49), we then have the simple recursive relation

$$\mu_t(x') = \max_{x \in \mathsf{X}} \left[\mu_{t-1}(x) + \ln m(x, x')\right] + \ln g_t(x') \,, \tag{9.51}$$

and $b_t(x')$ equals the state for which the maximum is achieved. The backward recursion first identifies the final state of the optimal state sequence. Then, once the final state is known, the next to final one can be determined as the state that gives the optimal probability for sequences ending with the now known final state. After that, the second next to final state can be determined in the same manner, and so on.

Example 9.11 (Number of major earthquakes; Example 9.1, cont.). For a model with two states, we assume that the parameters of the Poisson distribution $(\lambda_1, \lambda_2) \in \mathbb{R}^+$ associated with each state and of the transition matrix $[M(x,x')]_{(x,x') \in \mathsf{X}^2}$ are unknown, where $\mathsf{X} = \{1,2\}$. Denote by θ these parameters, which are assumed to be-

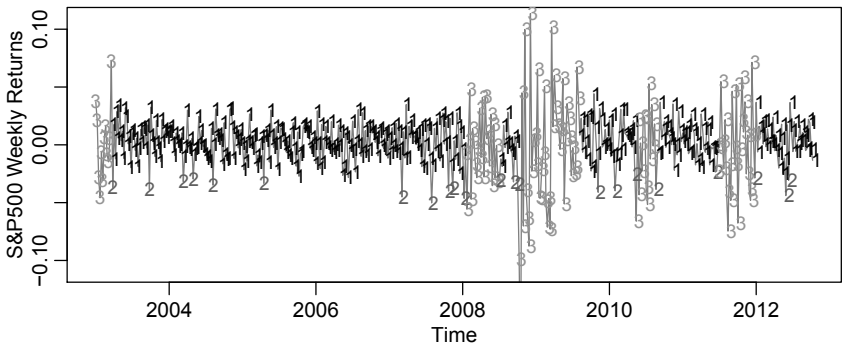


Figure 9.9 *S&P 500 weekly return from January 3, 2003 to September 30, 2012 and the estimated state based on the smoothing distributions. The states are indicated by points labeled 1,2 or 3. For display purposes, the vertical axis has been truncated; cf. Figure 9.3.*

long to a compact subset of

$$\Theta = \Big\{ \{\lambda_x\}_{x\in\mathsf{X}}, [M(x,x')]_{(x,x')\in\mathsf{X}^2}, M(x,x') \geq 0 \text{ and } \sum_{x'\in\mathsf{X}} M(x,x') = 1 \Big\}. \tag{9.52}$$

Given observations $Y_0,\ldots,Y_n$, we may use (9.44) to write the log-likelihood as

$$\ln \mathrm{p}^{\theta}_{\xi,n}(Y_{0:n}) \;=\; \ln\Big(\mathrm{p}^{\theta}_{\xi,0}(Y_0)\Big) + \sum_{t=1}^{n} \ln\left(\sum_{(x_{t-1},x_t)\in\mathsf{X}^2} \phi^{\theta}_{\xi,t-1}(x_{t-1}) Q^{\theta}_t(x_{t-1},x_t) \right).$$

Consequently, MLE can be performed via numerical maximization.

We fit the model to the time series of earthquake counts using the R package `depmixS4`. The package does not provide standard errors, so we obtained them by a parametric bootstrap procedure; see Remillard (2011) for justification. We note, however, that the standard errors may be obtained as a by-product of the estimation procedure; see Chapter 12. The MLEs of the intensities, along with their standard errors, were $(\hat\lambda_1,\hat\lambda_2) = (15.4_{(.7)}, 26.0_{(1.1)})$. The MLE of the transition matrix was $[\hat M(1,1),\hat M(1,2),\hat M(2,1),\hat M(2,2)] = [.93_{(.04)},.07_{(.04)},.12_{(.09)},.88_{(.09)}]$. Figure 9.8 displays the counts, the estimated state (displayed as points) and the smoothing distribution for the earthquakes data, modeled as a 2-state Poisson HMM model with parameters fitted using the MLEs. Finally, a histogram of the data is displayed along with the two estimated Poisson densities superimposed as solid lines. ◇

Example 9.12 (S&P500; Example 9.2, cont.). In Example 9.2, we fitted a mixture of three Gaussian distributions to the weekly S&P500 log-returns from January 3, 2003 until September 28, 2012. The results, which are displayed in Figure 9.3, are obtained under the unlikely assumption that the data are independent.

Here, we fit an HMM using the R package `depmixS4`, which takes into account that the data are dependent. As in Example 9.2, we chose a three-state model and we

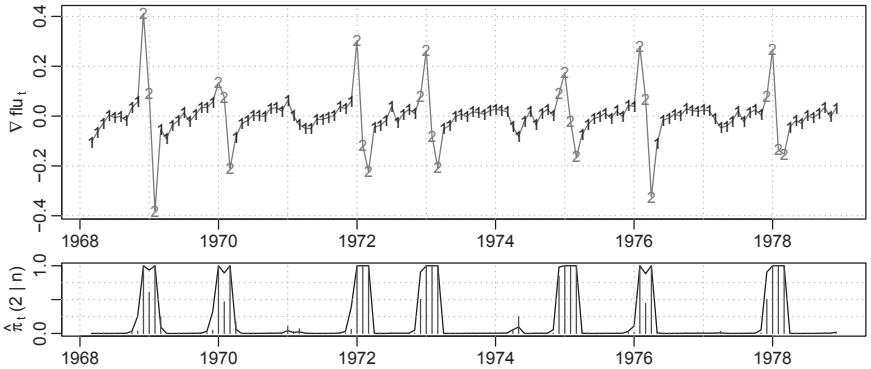


Figure 9.10 *The differenced flu mortality data of Figure 3.1 along with the estimated states (displayed as points). The smoothed state 2 probabilities are displayed in the bottom of the figure as a straight line. The filtered state 2 probabilities are displayed as vertical lines.*

leave it to the reader to investigate a two-state model (see Exercise 9.11). Standard errors (shown in parentheses below) were obtained via a parametric bootstrap based on a simulation script provided with the package.

The fitted transition matrix was

$$\widehat{M} = \begin{bmatrix} .945_{(.107)} & .055_{(.107)} & .000_{(.005)} \\ .739_{(.362)} & .000_{(.069)} & .261_{(.351)} \\ .031_{(.029)} & .027_{(.069)} & .942_{(.062)} \end{bmatrix},$$

and the three fitted normals were $\mathrm{N}(\hat{\mu}_1 = .004_{(.018)}, \hat{\sigma}_1 = .014_{(.020)})$, $\mathrm{N}(\hat{\mu}_2 = -.034_{(.020)}, \hat{\sigma}_2 = .009_{(.006)})$, and $\mathrm{N}(\hat{\mu}_3 = -.003_{(.006)}, \hat{\sigma}_3 = .044_{(.012)})$. The data, along with the predicted state (based on the smoothing distribution), are plotted in Figure 9.9.

The major differences between these results and the results from Example 9.2 are that regime 2 appears to represent a somewhat large-in-magnitude negative return, and may be a lone dip, or the start or end of a highly volatile period. States 1 and 3 represent clusters of regular or high volatility, respectively. Note that there is a large amount of uncertainty in the fitted normals, and in the transition matrix involving transitions from state 2 to states 1 or 3. ◇

Example 9.13 (Influenza mortality; Example 9.9, cont.). In Example 9.9, we considered fitting a two-state switching AR model given by (9.34). In particular, the idea was that data exhibit two different dynamics, one during an epidemic period, and another during a non-epidemic period.

We used the R package `MSwM` to fit the model specified in (9.34), with $p = 2$. The results were

$$\hat{Y}_t = \begin{cases} .006_{(.003)} + .293_{(.039)} Y_{t-1} + .097_{(.031)} Y_{t-2} + .024 Z_t, & \text{for } I_t = 1, \\ .199_{(.063)} - .313_{(.281)} Y_{t-1} - 1.604_{(.276)} Y_{t-2} + .112 Z_t, & \text{for } I_t = 2, \end{cases}$$

with estimated transition matrix

$$\widehat{M} = \begin{bmatrix} .927 & .073 \\ .300 & .700 \end{bmatrix} .$$

Figure 9.10 displays the data $Y_t = \nabla \text{flu}_t$ along with the estimated states (displayed as points labeled 1 or 2). The smoothed state 2 probabilities are displayed in the bottom of the figure as a straight line. The filtered state 2 probabilities are displayed in the same graph as vertical lines. ◇

9.2.2 *Continuous-valued state-space HMM*

The recursion developed for the discrete-valued state space extends directly to the general state-space setting. At time $t-1$, the filtering distribution $\phi_{\xi,t-1}$ summarizes all information the observations $Y_0, \ldots, Y_{t-1}$ contain about the state X_{t-1}.

Denote by $\gamma_{\xi,t}$

$$\gamma_{\xi,t}(f) = \int \cdots \int \xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s) f(x_t) \tag{9.53}$$

where $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$ and

$$g_t(x_t) := g(x_t, Y_t) \text{ and } Q_t(x_{t-1}, A) := \int_A M(x_{t-1}, \mathrm{d}x_t) g_t(x_t) \text{ for all } A \in \mathcal{X}. \tag{9.54}$$

This distribution may be computed recursively as follows

$$\begin{aligned} \gamma_{\xi,0}(f) &= \int \xi(\mathrm{d}x_0) g_0(x_0) f(x_0) = \xi(g_0 f) \\ \gamma_{\xi,t}(f) &= \iint \gamma_{\xi,t-1}(\mathrm{d}x_{t-1}) Q_t(x_{t-1}, \mathrm{d}x_t) f(x_t) = \gamma_{\xi,t-1} Q_t(f) \, . \end{aligned}$$

The joint distribution of the observations $(Y_0, \ldots, Y_t)$ is obtained by marginalizing the joint distribution $\gamma_{\xi,t}$ with respect to the state X_t, i.e.,

$$\mathrm{p}_{\xi,t}(Y_{0:t}) = \int \gamma_{\xi,t}(\mathrm{d}x_t) = \gamma_{\xi,t}(\mathbf{1}) \, . \tag{9.55}$$

The filtering distribution is the conditional distribution of the state X_t given $(Y_0, \ldots, Y_t)$. It is obtained by dividing the joint distribution of $(X_t, Y_0, \ldots, Y_t)$ by $\mathrm{p}_{\xi,t}(Y_{0:t})$,

$$\begin{aligned} \phi_{\xi,t}(f) = \frac{\gamma_{\xi,t}(f)}{\gamma_{\xi,t}(\mathbf{1})} &= \frac{\int \cdots \int f(x_t) \xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s) g_s(x_s)}{\int \cdots \int \xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s) g_s(x_s)} \\ &= \frac{\int \cdots \int f(x_t) \phi_{\xi,t-1}(\mathrm{d}x_{t-1}) M(x_{t-1}, \mathrm{d}x_t) g_t(x_t)}{\int \cdots \int \phi_{\xi,t-1}(\mathrm{d}x_{t-1}) M(x_{t-1}, \mathrm{d}x_t) g_t(x_t)} \\ &= \frac{\phi_{\xi,t-1} Q_t f}{\phi_{\xi,t-1} Q_t \mathbf{1}} \, . \end{aligned} \tag{9.56}$$

The forward recursion in (9.56) may be rewritten to highlight a two-step procedure involving both the predictive and filtering distributions. For $t \in \{0,1,\dots,n\}$ and $f \in \mathbb{F}_b(\mathsf{X},\mathcal{X})$, with the convention that $\phi_{\xi,0|-1} = \xi$, (9.56) may be decomposed as

$$\phi_{\xi,t|t-1} = \phi_{\xi,t-1} M \,, \tag{9.57a}$$

$$\phi_{\xi,t}(f) = \frac{\phi_{\xi,t|t-1}(f g_t)}{\phi_{\xi,t|t-1}(g_t)} \,. \tag{9.57b}$$

***Filter to Predictor*:** The first equation in (9.57) means that the updated predictive distribution $\phi_{\xi,t|t-1}$ is obtained by applying the transition kernel M to the current filtering distribution $\phi_{\xi,t-1}$. The predictive distribution is the one-step distribution of the Markov chain with kernel M given its initial distribution.

***Predictor to Filter*:** The second equation in (9.57) is recognized as Bayes' rule to correct the predictive distribution through the information contained in the actual observation Y_t.

- X_t is distributed *a priori* according to the predictive distribution $\phi_{\xi,t|t-1}$,
- g_t is the conditional probability density function of Y_t given X_t.

Although the recursions (9.57) appear relatively simple, they in fact must be approximated by numerical methods. We discuss particle approximations in Chapter 10.

The joint smoothing distribution $\phi_{\xi,0:t|t}$ then satisfies, for $f_{t+1} \in \mathbb{F}_b(\mathcal{X}^{\otimes(t+1)})$,

$$\phi_{\xi,0:t|t}(f_{t+1}) = \Big(\mathrm{p}_{\xi,t}(Y_{0:t})\Big)^{-1} \int \cdots \int f_{t+1}(x_{0:t}) \xi(\mathrm{d}x_0) g(x_0,y_0) \prod_{s=1}^{t} Q_s(x_{s-1},\mathrm{d}x_s) \tag{9.58}$$

assuming that $\mathrm{p}_{\xi,t}(Y_{0:t}) > 0$. Likewise, for indices $p \geq 0$,

$$\phi_{\xi,0:t+p|t}(f_{t+p+1}) = \int \cdots \int f_{t+p+1}(x_{0:t+p}) \times \phi_{\xi,0:t|t}(\mathrm{d}x_{0:t}) \prod_{s=t+1}^{t+p} M(x_{s-1},\mathrm{d}x_s) \tag{9.59}$$

for all functions $f_{t+p+1} \in \mathbb{F}_b(\mathcal{X}^{\otimes(t+p+1)})$. Eq. (9.58) implicitly defines the filtering, the predictive and the smoothing distributions as these are obtained by marginalization of the joint smoothing distribution; see Exercise 9.13.

The expression of the joint smoothing distribution (9.58) implicitly defines all other particular cases of smoothing kernels as these are obtained by marginalization. For instance, the marginal smoothing kernel $\phi_{\xi,t|n}$ for $0 \leq t \leq n$ is such that for $f \in \mathbb{F}_+(\mathsf{X},\mathcal{X})$,

$$\phi_{\xi,t|n}(f) := \int \cdots \int f(x_t)\, \phi_{\xi,0:n|n}(\mathrm{d}x_{0:n}) \,, \tag{9.60}$$

where $\phi_{\xi,0:n|n}$ is defined by (9.58).

Similarly, we note that the p-step predictive distribution $\phi_{\xi,n+p|n}$ may be obtained by marginalization of the joint distribution $\phi_{\xi,0:n+p|n}$ with respect to all variables x_t

except the last one (the one with index $t = n+p$). A closer examination of (9.59) directly shows that $\phi_{\xi,n+p|n} = \phi_{\xi,n} M^p$.

We now derive recursion for the smoothing distribution. For $\eta \in \mathbb{M}_1(\mathcal{X})$, assume that there exists a kernel B_η on $(\mathsf{X}, \mathcal{X})$ that satisfies, for all $h \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

$$\iint h(x,x')\eta(\mathrm{d}x)M(x,\mathrm{d}x') = \iint h(x,x')\eta M(\mathrm{d}x')B_\eta(x',\mathrm{d}x) \; . \tag{9.61}$$

This kernel is referred to as the *backward kernel* . When the HMM is fully dominated (see Definition 9.3), then the backward kernel may be explicitly written as

$$B_\eta(x,A) := \frac{\int \eta(\mathrm{d}x')\, m(x',x)\mathbb{1}_A(x')}{\int \eta(\mathrm{d}x')\, m(x',x)} \; , \quad A \in \mathcal{X} \; . \tag{9.62}$$

In other words, denote by η the distribution of X_0, and assume that the conditional distribution of X_1 given X_0 is $M(X_0,\cdot)$. Then the joint distribution of (X_0,X_1) is given by $\eta \otimes M$, i.e., for any $h \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$, $\mathbb{E}_\eta[h(X_0,X_1)] = \iint \eta(\mathrm{d}x_0)M(x_0,\mathrm{d}x_1)h(x_0,x_1)$. The marginal distribution of X_1 is ηM, and the conditional distribution of X_0 given X_1 is specified by the kernel $B_\eta(X_1,\cdot)$.

Proposition 9.14. *Given a strictly positive index t, initial distribution ξ, and index $t \in \{0,\ldots,n-1\}$,*

$$\mathbb{E}\left[f(X_t) \,\middle|\, X_{t+1:n}, Y_{0:n}\right] = \mathbb{E}\left[f(X_t) \,\middle|\, X_{t+1}, Y_{0:t}\right] = B_{\phi_{\xi,t}} f(X_{t+1})$$

for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$. In addition,

$$\mathbb{E}_\xi\left[f(X_{0:n}) \,\middle|\, Y_{0:n}\right] = \int \cdots \int f(x_{0:n})\, \phi_{\xi,n}(\mathrm{d}x_n) \prod_{s=0}^{n-1} B_{\phi_{\xi,s}}(x_{s+1},\mathrm{d}x_s) \tag{9.63}$$

for any $f \in \mathbb{F}_b(\mathsf{X}^{n+1}, \mathcal{X}^{\otimes(n+1)})$.

Proof. See Exercise 9.14. ∎

It follows from Proposition 9.14 that, conditionally on $Y_{0:n}$, the joint distribution of the index-reversed sequence $\{X_n, X_{n-1}, \ldots, X_0\}$ is that of a non-homogeneous Markov chain with initial distribution $\phi_{\xi,n}$ and transition kernels $\{B_{\phi_{\xi,t}}\}_{n-1 \geq t \geq 0}$. The backward smoothing kernel depends neither on the future observations nor on the index n. Therefore, the sequence of backward transition kernels $\{B_{\phi_{\xi,t}}\}_{0 \leq t \leq n-1}$ may be computed by forward recurrence on t. This decomposition suggests Algorithm 9.3 to recursively compute the marginal smoothing decomposition. Although the algorithm is apparently simple, the smoother must be approximated numerically. We discuss particle methods in Chapter 11.

Example 9.15 (The Rauch-Tung-Striebel smoother). For a linear Gaussian state-space model, all the conditional distributions are Gaussian distributions. Therefore, in that case, only the mean vectors and the covariance matrices need to be evaluated, and correspondingly the filtering or smoothing equations become equivalent to the

Algorithm 9.3 (Forward Filtering/Backward Smoothing)

Forward Filtering: Compute, forward in time, the filtering distributions $\phi_{\xi,0}$ to $\phi_{\xi,n}$ using the recursion (9.56). At each index t, the backward transition kernel $B_{\phi_{\xi,t}}$ may be computed according to (9.61).

Backward Smoothing: From $\phi_{\xi,n}$, compute, for $t = n-1, n-2, \ldots, 0$,

$$\phi_{\xi,t|n} = \phi_{\xi,t+1|n} B_{\phi_{\xi,t}} .$$

ordinary Kalman filter / smoother. The smoothing algorithm introduced above leads to an alternative derivation of Proposition 2.7, which is referred to as the Rauch-Tung-Striebel smoother; see Rauch et al. (1965). Let $X_{t+1} = \Phi X_t + W_t$ and $Y_t = AX_t + V_t$, where $\{W_t,\ t \in \mathbb{N}\}$ is i.i.d. zero-mean Gaussian with covariance Q and $\{V_t,\ t \in \mathbb{N}\}$ is i.i.d. zero mean-mean Gaussian with covariance R, $\{V_t,\ t \in \mathbb{N}\}$ and $\{W_t,\ t \in \mathbb{N}\}$ are independent. The initial state X_0 has a Gaussian distribution and is independent of $\{V_t,\ t \in \mathbb{N}\}$ and $\{W_t,\ t \in \mathbb{N}\}$.

We first determine the backward kernel. Let η be a Gaussian distribution with mean μ_0 and covariance Γ_0, i.e., $\eta = \mathrm{N}(\mu_0, \Gamma_0)$. Assume that $X_0 \sim \eta$ and let $X_1 = \Phi X_0 + W_0$. Note that, under this model,

$$\begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Gamma_0 & \Gamma_0\Phi' \\ \Phi\Gamma_0 & Q + \Phi\Gamma_0\Phi' \end{bmatrix} \right), \tag{9.64}$$

where $\mu_1 = \Phi\mu_0$. For any $x_1 \in \mathsf{X}$, the backward kernel, (9.61), is the conditional distribution of X_0 given $X_1 = x_1$, in the model (9.64). This conditional distribution is Gaussian, with mean and covariance

$$\mu_{0|1} = \mu_0 + J_0(x_1 - \mu_1) \tag{9.65}$$

$$\Gamma_{0|1} = \Gamma_0 - \Gamma_0\Phi'(\Phi\Gamma_0\Phi' + Q)^{-1}\Phi\Gamma_0 = \Gamma_0 - J_0(\Phi\Gamma_0\Phi' + Q)J_0' \tag{9.66}$$

where J_0 is the Kalman gain

$$J_0 = \Gamma_0\Phi'(\Phi\Gamma_0\Phi' + Q)^{-1} . \tag{9.67}$$

The action of this kernel is best understood by considering the Gaussian random vector

$$\tilde{X}_0 = \mu_0 + J_0(X_1 - \mu_1) + Z_0 \tag{9.68}$$

where Z_0 is a Gaussian random vector with zero-mean and covariance $\Gamma_{0|1}$ independent of X_1. Conditional to $X_1 = x_1$, $\tilde{X}_0$ is distributed according to $B_\eta(x_1, \cdot)$, provided that $\eta = \mathrm{N}(\mu_0, \Gamma_0)$. Assume now that $X_1 \sim \mathrm{N}(\mu_1, \Gamma_1)$. Then, the (unconditional) distribution of $\tilde{X}_0$ is Gaussian with mean and covariance given by

$$\tilde{\mu}_0 = \mu_0 + J_0(\mu_1 - \mu_0) , \tag{9.69}$$

$$\begin{aligned} \tilde{\Gamma}_0 &= J_0\Gamma_1 J_0' + \Gamma_0 - J_0(\Phi\Gamma_0\Phi' + Q)J_0' \\ &= \Gamma_0 + J_0\left(\Gamma_1 - (\Phi\Gamma_0\Phi' + Q)\right)J_0' . \end{aligned} \tag{9.70}$$

To obtain the recursion for the smoother covariance, it suffices to replace (μ_0, Γ_0) in (9.65)-(9.66) by the filtering mean and covariance $(X_{t|t}, P_{t|t})$, defined in (2.11)-(2.12) and (μ_1, Γ_1) by $(X_{t+1|n}, P_{t+1|n})$, to obtain the *forward-filtering, backward smoothing* recursion

$$X_{t|n} = X_{t|t} + J_t(X_{t+1|n} - X_{t|t}) \tag{9.71}$$

$$P_{t|n} = P_{t|t} + J_t\left(P_{t+1|n} - P_{t+1|t}\right)J_t' \tag{9.72}$$

where $J_t = P_{t|t}\Phi'(\Phi P_{t|t}\Phi' + Q)^{-1}$ is the Kalman Gain. The filtering mean and covariance $(X_{t|t}, P_{t|t})$ are computed using the Kalman filter. The smoothing mean and covariance are obtained by running (9.71)-(9.72) backwards in time, starting from $(X_{n|n}, P_{n|n})$. ◇

9.3 Endnotes

Nonlinear state space models and their generalizations are nowadays used in many different areas. Several specialized books are available that largely cover applications of HMMs to some specific areas such as speech recognition (Rabiner and Juang, 1993, Jelinek, 1997), econometrics (Hamilton, 1989, Kim and Nelson, 1999), computational biology (Durbin et al., 1998, Koski, 2001), or computer vision (Bunke and Caelli, 2001). The elementary theory of HMM is covered in MacDonald and Zucchini (2009) and Fraser (2008), which discuss a lot of interesting examples of applications.

Most of the early references on filtering and smoothing, which date back to the 1960s, focused on the specific case of Gaussian linear state-space models, following the pioneering work by Kalman and Bucy (1961). The classic book by Anderson and Moore (1979) on *optimal filtering*, for instance, is fully devoted to linear state-space models; see also Kailath et al. (2000, Chapter 10) for a more exhaustive set of early references on the smoothing problem. Although some authors, for example, Ho and Lee (1964) considered more general state-space models, it is fair to say that the Gaussian linear state-space model was the dominant paradigm. Until the early 1980s, the works that *did not* focus on the linear state-space model were usually advertised by the use of the words "Bayes" or "Bayesian" in their title; see, e.g., Ho and Lee (1964) or Askar and Derin (1981).

Almost independently, the work by Baum and his colleagues on hidden Markov models (Baum et al., 1970) dealt with the case where the state space X of the hidden state is finite. These two streams of research (on Gaussian linear models and finite state space models) remained largely separated. The forward-backward algorithm is known to many, especially in the field of speech processing, as the *Baum-Welch algorithm*, although the first published description of the approach is due to Baum et al. (1970, p. 168).

The forward-backward algorithm was discovered several times in the early 1970s; see Fraser (2008) and MacDonald and Zucchini (2009). A salient example is the paper by Bahl et al. (1974) on the computation of posterior probabilities for a finite-state Markov channel encoder for transmission over a discrete memoryless

channel. The algorithm described by Bahl et al. (1974) is fully equivalent to the forward-backward and is known in digital communication as the BCJR (for Bahl, Cocke, Jelinek, and Raviv) algorithm. Chang and Hancock (1966) is another less well-known reference, contemporary with the work of Baum and his colleagues, which also describes the forward-backward decomposition and its use for decoding in communication applications.

Approximately at the same time, in applied probability, the seminal work by Stratonovich (1960) stimulated a number of contributions that were to compose a body of work generally referred to as *filtering theory*. The object of filtering theory is to study inference about partially observable Markovian processes in *continuous time*. A number of early references in this domain indeed consider some specific form of discrete state space continuous-time equivalent of the HMM (Shiryaev 1966, Wonham 1965; see also Lipster and Shiryaev 2001, Chapter 9). Working in continuous time, however, implies the use of mathematical tools that are definitely more complex than those needed to tackle the discrete-time model of Baum et al. (1970). As a matter of fact, filtering theory and hidden Markov models evolved as two mostly independent fields of research. A poorly acknowledged fact is that the pioneering paper by Stratonovich (1960) (translated from an earlier Russian publication) describes, in its first section, an equivalent to the forward-backward smoothing approach of Baum et al. (1970). It turns out, however, that the formalism of Baum et al. (1970) generalizes well to models where the state space is *not* discrete anymore, in contrast to that of Stratonovich (1960).

Exercises

9.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and Y a random variable such that $\mathbb{E}\left[Y^2\right] < \infty$. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Show that

$$\mathbb{E}[(Y - \mathbb{E}[Y \,|\, \mathcal{G}])^2] = \mathbb{E}[(Y - \mathbb{E}[Y \,|\, \mathcal{G}])^2] + \mathbb{E}\left[(\mathbb{E}[Y \,|\, \mathcal{G}] - \mathbb{E}[Y])^2\right] ,$$

and check (9.6).

9.2. Consider a discrete state-space HMM. Denote by $\mathsf{X} = \{1, \ldots, m\}$ the state-space of the Markov chain, M, the $m \times m$ transition matrix and, for $y \in \mathsf{Y}$, by $\Gamma = \mathrm{diag}(G(1,y), \ldots, G(m,Y))$. Assume that M admits a unique stationary distribution denoted by $\pi = [\pi(1), \ldots, \pi(m)]$.

(a) Show that the likelihood of the observations (9.4) may be expressed as

$$\mathrm{p}_{\xi,t}(Y_{0:t}) = \xi \Gamma(Y_0) M \Gamma(Y_1) M \ldots M \Gamma(Y_t) \mathbb{1}$$

where $\mathbb{1} = [1, 1, \ldots, 1]'$.

(b) Show that for any $h \in \mathbb{N}$,

$$\mathrm{p}_{\xi,t}(Y_{h:t+h}) = \xi M^h \Gamma(Y_0) M \Gamma(Y_1) M \ldots M \Gamma(Y_t) \mathbb{1}$$

and check that $\mathrm{p}_{\xi,t}(Y_{h:t+h}) = \mathrm{p}_{\xi,t}(Y_{0:t})$.

9.3. We use the notations of Exercise 9.2. Let $(\mu_x, x \in \mathsf{X})$ and $(\sigma_x^2, x \in \mathsf{X})$ denote the mean and variance of the distributions $(G(x,\cdot), x \in \mathsf{X})$.

(a) $\mathbb{E}_\pi[Y_t] = \sum_{x\in\mathsf{X}} \pi(x)\mu_x$.

(b) $\mathbb{E}_\pi[Y_t^2] = \sum_{x\in\mathsf{X}} \pi(x)(\sigma_x^2 + \mu_x^2)$.

(c) $\mathrm{Var}_\pi(Y_t) = \sum_{x\in\mathsf{X}} \pi(x)(\sigma_x^2 + \mu_x^2) - (\sum_{x\in\mathsf{X}} \pi(x)\mu_x)^2$.

(d) If $m = 2$, $\mathrm{Var}_\pi(Y_t) = \pi(1)\sigma(1)^2 + \pi(2)\sigma_2^2 + \pi(1)\pi(2)(\mu_1 - \mu_2)^2$.

(e) For $k \in \mathbb{N}$,

$$\mathbb{E}_\pi[Y_t Y_{t+k}] = \sum_{x_0\in\mathsf{X}} \sum_{x_k\in\mathsf{X}} \pi(x_0)\mu_{x_0} M^k(x_0,x_k)\mu_{x_k} = \pi \mathrm{diag}(\mu) M^k \mu \ ,$$

where $\mu = [\mu_1, \mu_2, \ldots, \mu_m]'$.

(f) Show that, if the eigenvalues of M are distinct, then $\mathrm{Cov}_\pi(X_0, X_k)$ may be expressed as a linear combination of the k-th powers of those eigenvalues.

9.4. Consider the state-space model with non-linear state evolution equation

$$X_t = A(X_{t-1}) + R(X_{t-1})W_t \ , \qquad W_t \sim \mathrm{N}(0, I) \ , \tag{9.73}$$

$$Y_t = BX_t + SV_t \ , \qquad V_t \sim \mathrm{N}(0, I) \ , \tag{9.74}$$

where A and R are matrix-valued functions of appropriate dimensions. Show that the conditional distribution of X_t given $X_{t-1} = x$ and Y_t is multivariate Gaussian with mean $m_t(x)$ and covariance matrix $\Sigma_t(x)$, given by

$$\begin{aligned} K_t(x) &= R(x)R'(x)B' \left[BR(x)R'(x)B' + SS^t\right]^{-1} \ , \\ m_t(x) &= A(x) + K_t(x)\left[Y_{t+1} - BA(x)\right] \ , \\ \Sigma_t(x) &= \left[I - K_t(x)B\right] R(x)R'(x) \ . \end{aligned}$$

9.5. Assume that $Y_t = \mu_{X_t} + \sigma_{X_t} Z_t$, where $\{Z_t,\, t \in \mathbb{N}\} \sim$ iid $\mathrm{N}(0,1)$ and $\{X_t,\, t \in \mathbb{N}\}$ is a two-state stationary Markov chain, independent of $\{Z_t,\, t \in \mathbb{N}\}$.

(a) Show that the unconditional distribution of Y_t is given by a mixture of two normal distributions: $p(y_t) = \pi_1 \mathfrak{g}(y_t; \mu_1,\ \sigma_1^2) + \pi_2 \mathfrak{g}(y_t; \mu_2,\ \sigma_2^2)$ where $\pi_x = \mathbb{P}(X_t = x)$, $x \in \{1,2\}$.

(b) Show that the skewness is given by

$$\frac{\mathbb{E}\left[(Y_t - \mu)^3\right]}{(\mathbb{E}\left[(Y_t - \mu)^2\right])^{3/2}} = \pi_1\pi_2(\mu_1 - \mu_2)\frac{3(\sigma_2^2 - \sigma_1^2)^2 + (\pi_2 - \pi_1)(\mu_2 - \mu_1)^2}{\sigma^3} \ ,$$

with $\mu = \mathbb{E}[Y_t]$ and $\sigma^2 = \mathrm{Var}(Y_t)$ being the mean and variance of the mixture distribution: $\mu = \pi_1\mu_1 + \pi_2\mu_2$ and $\sigma^2 = \pi_1\sigma_1^2 + \pi_2\sigma_2^2 + \pi_1\pi_2(\mu_2 - \mu_1)^2$.

(c) Show that the excess kurtosis is given by

$$\frac{\mathbb{E}\left[(Y_t - \mu)^4\right]}{(\mathbb{E}\left[(Y_t - \mu)^2\right])^2} - 3 = \pi_1\pi_2\frac{3(\sigma_2^2 - \sigma_1^2)^2 + c(\mu_1,\mu_2)}{\sigma^4} \ ,$$

where

$$c(\mu_1,\mu_2) = 6(\pi_1 - \pi_2)(\sigma_2^2 - \sigma_1^2)(\mu_2 - \mu_1)^2 + (\mu_2 - \mu_1)^4(1 - 6\pi_1\pi_2) \ .$$

Remark 9.16. Note that skewness in the marginal distribution will be present whenever both the means and the variances are different. For a model where the means are the same, no skewness is present. If the variances are the same and the means are different, skewness is possible only if $\pi_1 \neq \pi_2$. Thus, for a Markov mixture model with different means but equal variances, asymmetry is introduced into the marginal distribution only through asymmetry in the persistence probabilities, namely $M(1,1) \neq M(2,2)$. If $\mu_1 = \mu_2$, the marginal distribution has fatter tails than a normal distribution as long as $\sigma_1^2 \neq \sigma_2^2$; see Frühwirth-Schnatter (2006, p. 309).

9.6 (Filtering distribution for a 2-state HMM). Let $\{X_t,\, t \in \mathbb{N}\}$ be a two-state stationary Markov chain with transition kernel M and stationary distribution π. Let $\{(X_t, Y_t),\, t \in \mathbb{N}\}$ be a partially dominated HMM.

(a) Show that the predictive distribution of $X_t = 1$,

$$\phi_{t|t-1}(1) = \pi_1 + \lambda \pi_2 \phi_{t-1}(1) - \lambda \pi_1 \phi_{t-1}(2),$$

where $\lambda = M(1,1) - M(1,2)$ is equal to the second eigenvalue of M.

(b) Show that

$$\phi_{t|t-1}(1) = (1-\lambda)\pi_1 + \lambda \phi_{t-1}(1)\,.$$

Remark 9.17. When λ is close to 0 (the Markov chain is not very persistent), the predictive distribution for X_t is dominated by the stationary distribution of the chain. When λ is close to 1 (highly persistent Markov chains), the predictive distribution for X_t will be dominated by the filtered state probability ϕ_{t-1}.

9.7. Consider a two-state Gaussian HMM, $Y_t = \mu_{X_t} + \sigma_{X_t} V_t$, where $\{V_t,\, t \in \mathbb{N}\}$ is a strong white Gaussian noise and $\{X_t,\, t \in \mathbb{N}\}$ is a stationary two-state Markov chain with transition kernel M such that $M(1,2) \in (0,1)$ and $M(2,1) \in (0,1)$. Show that $\{Y_t,\, t \in \mathbb{N}\}$ is an ARMA(1,1) process.

9.8 (Spectral density of a discrete-valued Markov chain). Let $\mathsf{X} = \{x_1, \ldots, x_n\}$ be a finite set and M be a transition kernel on X. Assume that M admits a unique stationary distribution π. Let $\{X_t,\, t \in \mathbb{Z}\}$ be a stationary Markov chain on X with transition kernel M. Define by $M_\infty = \lim_{k\to\infty} M^k = \pi \mathbf{1}$.

(a) Show that $M_\infty = M M_\infty = M_\infty M$ and M_∞ is idempotent.

(b) Let $F = M - M_\infty$. Show that $(M^k - M_\infty) = F^k,\ k = 1, 2, \cdots$.

(c) Set $S = \mathrm{diag}(x_1, \ldots, x_n)$ $R = \mathrm{diag}(\pi_1, \cdots, \pi_n)$. Show that $\mu_X = \pi S \mathbf{1}$, $\gamma_X(0) = \pi S(\mathrm{I} - M_\infty)\mathbf{1}$ and $\gamma_X(k) = \pi S F^{|k|} S \mathbf{1}$, $k = \pm 1, \pm 2, \cdots$.

(d) Show that the eigenvalues of F are zero and $\lambda_i,\ |\lambda_i| < 1,\ i = 2, \cdots, n$, the subdominant eigenvalues of P counted with their algebraic multiplicity.

(e) Denote by $f_X(\omega)$ denote the spectral density of $\{X_t,\, t \in \mathbb{Z}\}$. Show that

$$2\pi f_X(\omega) = \pi S[(\mathrm{I} - M_\infty) + 2\,\mathrm{Re}(\mathrm{e}^{-\mathrm{i}\omega} F(\mathrm{I} - \mathrm{e}^{-\mathrm{i}\omega} F)^{-1})] S \mathbf{1}\,. \tag{9.75}$$

9.9 (Exercise 9.8, cont.). (a) Show that $z^{-1}F(\mathrm{I} - z^{-1}F)^{-1} = F(Iz - F)^{-1} = QJ(Iz - J)^{-1}Q^{-1}$ where $J = Q^{-1}FQ$ is the Jordan canonical form of F, where

the Jordan matrix $J = \text{diag}(J_1 : \cdots : J_u)$ where each diagonal block J_i is a $\nu_i \times \nu_i$ matrix of the form

$$\begin{pmatrix} \varphi_i & 1 & 0 & \cdots & 0 \\ 0 & \varphi_i & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & 1 \\ 0 & 0 & \cdots & 0 & \varphi_i \end{pmatrix},$$

$i = 1, \cdots, u$, the φ_j, $i = 1, \cdots, u$, being the unique eigenvalues of F and $\nu_i \geqq 1$, $\nu_1 + \cdots + \nu_u = n$, their respective algebraic multiplicities. Set $Q = [Q_1 : \cdots : Q_u]$ and $Q^{-1} = [Q^1 : \cdots : Q^u]'$ where Q_i and Q^i are $n \times \nu_i$ matrices, $i = 1, \cdots u$.

(b) Show that $J(\text{I}z - J)^{-1} = \text{diag}(J_1(\text{I}z - J_1)^{-1} : \cdots : J_n(\text{I}z - J_u)^{-1})$ and hence that

$$QJ(\text{I}z - J)^{-1}Q^{-1} = \sum_{i=1}^{u} Q_i J_i (\text{I}z - J_i)^{-1}(Q^i)' . \tag{9.76}$$

(c) Let q_{ij}, $\cdots q_{i\nu_i}$, and $q^{i1}, \cdots, q^{i\nu_i}$ denote the columns of Q_i and Q^j, respectively. Show that $J_j(\text{I}z - J_i)^{-1}$ equals an upper triangular Toeplitz matrix with first row

$$[\varphi_i(z - \varphi_i)^{-1},\ (z - 2\varphi_i)(z - \varphi_i)^{-2},\ \cdots,\ (z - 2\varphi_i)(z - \varphi_i)^{-\nu_i}(-1)^{\nu_i}] .$$

(d) Deduce that each of the summands in (9.76) gives rise to an expansion of the form

$$R_{i1}\frac{\varphi_i}{(z - \varphi_i)} + \sum_{j=2}^{\nu_i} R_{ij}\frac{(z - 2\varphi_i)(-1)^j}{(z - \varphi_i)^j} , \tag{9.77}$$

where the second and subsequent terms only appear if $\nu_i \geqq 2$ and

$$R_{ij} = \sum_{l=1}^{\nu_i - j + 1} q_{il}(q^{il+j-1})' .$$

(e) Show that $\pi S Q_i J_i (\text{I}z - J_i)^{-1}(Q^i)' S\mathbf{1} = b_i(z)/\beta_j(z)$ where

$$b_i(z) = \pi S[R_{i1}\varphi_i(z - \varphi_i)^{\nu_i - 1} + \sum_{j=2}^{\nu_i} R_{ij}(z - 2\varphi_i)(z - \varphi_i)^{\nu_i - j}(-1)^j]S\mathbf{1}$$

and $\beta_i(z) = (z - \varphi_i)^{\nu_i}$, $\quad i = 1, \cdots, u$.

(f) Show that

$$2\pi f_X(\omega) = \gamma_X(0) + \sum_{i=2}^{u} \frac{b_i(\text{e}^{\text{i}\omega})}{\beta_i(\text{e}^{\text{i}\omega})} + \frac{b_i(\text{e}^{-\text{i}\omega})}{\beta_i(\text{e}^{-\text{i}\omega})} \quad = \gamma_X(0) + \frac{a(\text{e}^{\text{i}\omega})}{\alpha(\text{e}^{\text{i}\omega})} + \frac{a(\text{e}^{-\text{i}\omega})}{\alpha(\text{e}^{-\text{i}\omega})}$$

where $\alpha(z) = z^{1-n}\prod_{i=2}^{u}\beta_i(z) = \prod_{j=2}^{n}(1 - \lambda_j z^{-1})$ and $a(z) = z^{1-n}\sum_{i=2}^{u}\{b_i(z)\prod_{j=2, j\neq i}^{u}\beta_j(z)\}$.

(g) By combining the results above, show that the special density of a discrete-valued Markov chain can be expressed in the rational form

$$f_X(\omega) = \frac{1}{2\pi} \frac{|m(\mathrm{e}^{\mathrm{i}\omega})|^2}{|\alpha(\mathrm{e}^{\mathrm{i}\omega})|^2}, \; -\pi < \omega < \pi,$$

where $m(z) = m_0 + m_1 z^{-1} + \cdots + m_{n-1} z^{-n+1}$ and $\alpha(z) = 1 + \alpha_1 z^{-1} + \cdots + \alpha_{n-1} z^{-n+1}$ are relatively prime and n is the state dimension. Furthermore, if $\lambda_j,\ j = 2,\ \cdots,\ n$, are the sub-dominant eigenvalues of P, then $\alpha(z) = \prod_{j=2}^{n}(1 - \lambda_j z^{-1})$.

9.10. (a) Show that

$$\mathbb{E}_\xi \left[\prod_{i=1}^{p} f_i(Y_{t_i}) h(X_{t_1}, \ldots, X_{t_p}) \right] = \mathbb{E}_\xi \left[h(X_{t_1}, \ldots, X_{t_p}) \prod_{i=1}^{p} G f_i(X_{t_i}) \right]$$

and check (9.14).

(b) Show that, for any integers t and p and any ordered t-tuple $\{t_1 < \cdots < t_p\}$ of indices such that $t \notin \{t_1, \ldots, t_p\}$, the random variables Y_t and $(X_{t_1}, \ldots, X_{t_p})$ are $\mathbb{P}_\xi$-conditionally independent given X_t.

9.11. Fit a two-state model to the S&P 500 weekly returns discussed in Example 9.12. Compare the AIC and BIC of the two-state model with the three-state model and state your conclusions. Note: For the 3-state model, `depmix` reports:

```
'log Lik.' 1236.996 (df=14),  AIC: -2445.992,  BIC: -2386.738.
```

9.12. We consider the autoregressive stochastic volatility model Example 9.5.

(a) Show that for any integer m,

$$\mathbb{E}\left[Y_t^{2m}\right] = \beta^{2m} \mathbb{E}\left[V_t^{2m}\right] \exp(m^2 \sigma_X^2/2)\,,$$

where $\sigma_X^2 = \sigma^2/(1-\phi^2)$.

(b) Show (9.19).

(c) Show that for any positive integer h, $\mathrm{Var}(X_t + X_{t+h}) = 2\sigma_X^2(1+\phi^h)$.

(d) Show that

$$\mathrm{Cov}(Y_t^{2m}, Y_{t+h}^{2m}) = \beta^{4m} \left(\mathbb{E}\left[V_t^{2m}\right]\right)^2 \left(\exp(m^2\sigma_X^2(1+\phi^h)) - \exp(m^2\sigma_X^2)\right)\,.$$

(e) Establish (9.20).

9.13. The purpose of this exercise is to prove (9.59). Consider two functions $f \in \mathbb{F}_b(\mathsf{X}^{n+p+1}, \mathcal{X}^{\otimes(n+p+1)})$ and $h \in \mathbb{F}_b(\mathsf{Y}^{n+1}, \mathcal{Y}^{\otimes(n+1)})$.

(a) Show that

$$\begin{aligned}\mathbb{E}_\xi[h(Y_{0:n}) f(X_{0:n+p})] = &\int \cdots \int f(x_{0:n+p}) \xi(\mathrm{d}x_0) g(x_0, y_0) \\ &\times \left[\prod_{s=1}^{n} Q_s(x_{s-1}, \mathrm{d}x_s)\right] h(y_{0:n}) \left[\prod_{s=n+1}^{n+p} Q_s(x_{s-1}, \mathrm{d}x_s)\right] \mu_{n+p}(\mathrm{d}y_{0:n+p})\,,\end{aligned}$$

where $Q_s(x_{s-1}, \mathrm{d}x_s) = M(x_{s-1}, \mathrm{d}x_s) g - x_s, y_s)$.

(b) Show that

$$\mathbb{E}_\xi[h(Y_{0:n})f(X_{0:n+p})] = \int\cdots\int h(y_{0:n})f(x_{0:n+p}) \\ \phi_{\xi,0:n|n}(y_{0:n},\mathrm{d}x_{0:n})\left[\prod_{s=n+1}^{n+p} M(x_{s-1},\mathrm{d}x_s)\right] \mathrm{p}_{\xi,n}(y_{0:n})\mu_n(\mathrm{d}y_{0:n}) .$$

9.14. Let $t \in \{0,\dots,n-1\}$ and $h \in \mathbb{F}_b(\mathsf{X}^{n-t},\mathcal{X}^{\otimes(n-t)})$.

(a) Show that

$$\mathbb{E}_\xi\left[f(X_t)h(X_{t+1:n}) \mid Y_{0:n}\right] = \int\cdots\int f(x_t)h(x_{t+1:n})\,\phi_{\xi,t:n|n}(\mathrm{d}x_{t:n}) .$$

(b) Show that

$$\mathbb{E}_\xi\left[f(X_t)h(X_{t+1:n}) \mid Y_{0:n}\right] = \frac{\mathrm{p}_{\xi,t}(Y_{0:t})}{\mathrm{p}_{\xi,n}(Y_{0:n})}\iint \phi_{\xi,t}(\mathrm{d}x_t)M(x_t,\mathrm{d}x_{t+1})f(x_t)\tilde{h}(x_{t+1}) ,$$

where

$$\tilde{h}(x_{t+1}) := g(x_{t+1},Y_{t+1})\int\cdots\int\left[\prod_{s=t+2}^{n} M(x_{s-1},\mathrm{d}x_s)g(x_s,Y_s)\right]h(x_{t+1:n}) .$$

(c) Show that

$$\iint \phi_{\xi,t}(\mathrm{d}x_t)M(x_t,\mathrm{d}x_{t+1})f(x_t)\tilde{h}(x_{t+1}) \\ = \int \phi_{\xi,t}M(\mathrm{d}x_{t+1})B_{\phi_{\xi,t}}f(x_{t+1})\tilde{h}(x_{t+1}) .$$

(d) Show that for any $\hat{h} \in \mathbb{F}_b(\mathsf{X}^{n-t},\mathcal{X}^{\otimes(n-t)})$,

$$\phi_{\xi,t+1:n|n}(\hat{h}) = \frac{\int\cdots\int \phi_{\xi,t}(\mathrm{d}x_t)\prod_{s=t+1}^{n} M(x_{s-1},\mathrm{d}x_s)g(x_s,Y_s)\hat{h}(x_{t+1:n})}{\int\cdots\int \phi_{\xi,t}(\mathrm{d}x_t)\prod_{s=t+1}^{n} M(x_{s-1},\mathrm{d}x_s)g(x_s,Y_s)} .$$

(e) Deduce that

$$\mathbb{E}_\xi\left[f(X_t)h(X_{t+1:n}) \mid Y_{0:n}\right] = \int\cdots\int B_{\phi_{\xi,t}}f(x_{t+1})h(x_{t+1:n})\phi_{\xi,t+1|n}(\mathrm{d}x_{t+1:n}) .$$

(f) Show that

$$\mathbb{E}_\xi\left[f(X_t)h(X_{t+1:n}) \mid Y_{0:n}\right] = \mathbb{E}_\xi\left[h(X_{t+1:n})B_{\phi_{\xi,t}}f(X_{t+1}) \mid Y_{0:n}\right] ,$$

and conclude.

9.15. Using Proposition 9.14, show the validity of Algorithm 9.3.

Chapter 10

Particle Filtering

Prior to the mid-1980s, a number of methods were developed to approximate the filtering/smoothing distribution for non-normal or nonlinear state-space models in an attempt to circumvent the computational complexity of inference for such models. With the advent of cheap and fast computing, a number of authors developed computer-intensive methods based on numerical integration. For example, Kitagawa (1987) proposed a numerical method based on piecewise linear approximations to the density functions for prediction, filtering, and smoothing for non-Gaussian and nonstationary state-space models. Pole and West (1989) used Gaussian quadrature techniques; see West and Harrison (1997, Chapter 13) and the references therein.

Sequential Monte Carlo (SMC) refers to a class of methods designed to approximate a *sequence of probability distributions* by a set of *particles* such that each have an assigned non-negative weight and are updated recursively. SMC methods are a combination of the sequential importance sampling method introduced in Handschin and Mayne (1969) and the sampling importance resampling algorithm proposed in Rubin (1987).

10.1 Importance sampling

Throughout this section, μ denotes a probability measure on a measurable space $(\mathsf{X}, \mathcal{X})$, which is referred to as the *target distribution*. The aim of importance sampling is to approximate integrals of the form $\mu(f) = \int_{\mathsf{X}} f(x)\,\mu(\mathrm{d}x)$ for $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$. The plain Monte Carlo approach consists in drawing an i.i.d. sample $\{X^i\}_{i=1}^N$, from the target distribution μ and then evaluating the sample mean $N^{-1}\sum_{i=1}^N f(X^i)$.

Importance sampling is based on the idea that in certain situations it is more appropriate to sample from a *proposal distribution* ν, and then to apply a change-of-measure formula. Assume that the target distribution μ is absolutely continuous with respect to ν and denote by $w = \mathrm{d}\mu/\mathrm{d}\nu$ the Radon-Nikodym derivative of μ with respect to ν, referred to in the sequel as the *weight function*. Then, for $f \in \mathrm{L}^1(\mu)$, the change of measure formula implies

$$\mu(f) = \int f(x)\,\mu(\mathrm{d}x) = \int f(x)\,w(x)\,\nu(\mathrm{d}x)\,. \tag{10.1}$$

If $\{X^i\}_{i=1}^N$ is an i.i.d. sample from ν, (10.1) suggests the following estimator of $\mu(f)$:

$$N^{-1}\sum_{i=1}^{N} f(X^i)w(X^i)\,. \tag{10.2}$$

Because $\{X^i\}_{i=1}^N$ is an i.i.d. sample from ν, the Strong Law of Large Numbers (SLLN) implies that $N^{-1}\sum_{i=1}^N f(X^i)w(X^i)$ converges to $\nu(fw)=\mu(f)$ almost surely as N tends to infinity; see Exercise 10.1. In addition, moments bounds, deviations inequalities, and central limit theorem for i.i.d. variables may be used to assess the fluctuations of this estimator around its mean.

In many situations, the target probability measure μ is known only up to a normalizing factor. This happens in particular when applying importance sampling ideas to solve filtering and smoothing problems in NLSS. The weight function w is then known up to a (constant) scaling factor only. It is however still possible to use the importance sampling paradigm, by adopting the self-normalized form of the importance sampling estimator,

$$\sum_{i=1}^{N} \frac{w(X^i)}{\sum_{j=1}^N w(X^j)} f(X^i)\,. \tag{10.3}$$

The self-normalized importance sampling estimator is defined as a ratio of the sample means $N^{-1}\sum_{i=1}^N f(X^i)w(X^i)$ and $N^{-1}\sum_{i=1}^N w(X^i)$. The SLLN implies that these two sample means converge, almost surely, to $\nu(fw)=\nu(w)\mu(f)$ and $\nu(w)$, respectively, showing that the self-normalized importance sampling estimator is a consistent estimator of $\mu(f)$; see Exercise 10.5. Importance sampling is of considerable generality and interest since it introduces very little restrictions on the choice of the proposal distribution. This choice is typically guided by two requirements: the proposal distribution should be easy to simulate and should lead to an efficient estimator. This is discussed in the very simple example below.

Example 10.1. Assume that the target distribution is a Gaussian mixture, with density $p(x)=\alpha\mathfrak{g}(x;m_1,\sigma_1^2)+(1-\alpha)\mathfrak{g}(x;m_2,\sigma_2^2)$. A natural choice for the proposal distribution is the Student t_κ-distribution,

$$q_\kappa(x)=\frac{\Gamma((\kappa+1)/2)}{\sqrt{\kappa\pi}\Gamma(\kappa/2)}\left(1+\frac{x^2}{\kappa}\right)^{-\frac{\kappa+1}{2}}\,,$$

where κ is the *number of degrees of freedom* and Γ is the Gamma function. The t_κ-distribution is symmetrical about $x=0$ and has a single mode at $x=0$. It is easy to show that $\lim_{\kappa\to\infty} q_\kappa(x)=(\sqrt{2\pi})^{-1}\mathrm{e}^{-x^2/2}$: As $\kappa\to\infty$, the t_κ distribution tends to the unit normal distribution. For small to moderate number of degrees of freedom, the fat-tailed behavior of the distribution is characterized by the kurtosis relative to that of a normal distribution that is equal to $6/(\kappa-4)$. When κ is an integer, a draw of a student t_κ may be obtained by sampling $Z_1,\ldots,Z_\kappa$ independent standard normal random variables and computing

$$T_\kappa=Z_0\left(\kappa^{-1}\sum_{i=1}^{\kappa} Z_i^2\right)^{-1/2}\,.$$

◇

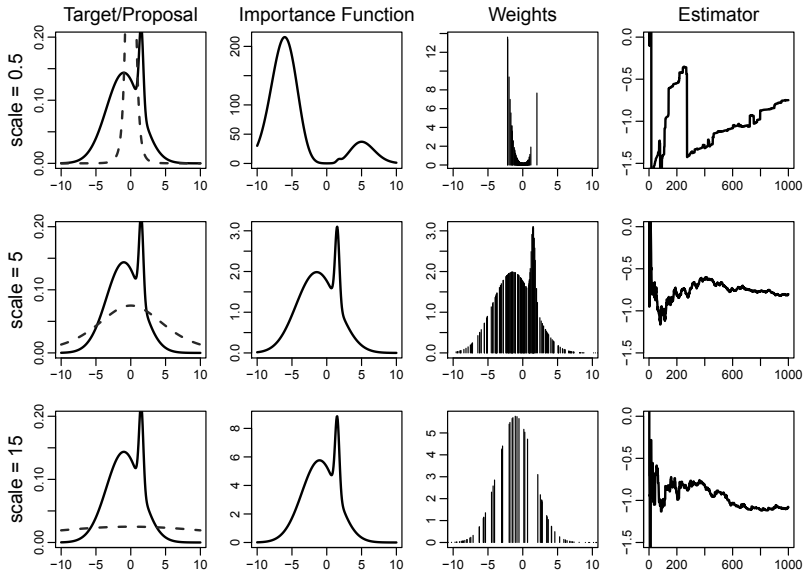


Figure 10.1 *The importance sampling estimator for different choices of proposal distributions. The proposal distribution is a Student t with 4 degrees of freedom. The scales are 0.5 (top row), 5 (middle row), and 15 (bottom row). The number of samples in each case is 1000. In the first column of each row, the target pdf is displayed as a solid line and the proposal is displayed as a dashed line.*

Another possibility consists in resorting to the inversion method. We use the importance sampling estimator to estimate the mean of the target distribution (here equal to $\alpha m_1 + (1-\alpha)m_2$) using t_κ distribution with $\kappa = 4$ and different scales (denoting s the scale, the proposal distribution is $x \mapsto s^{-1} q_\kappa(s^{-1}x)$). As illustrated in Figure 10.1, the choice of the scale plays a crucial role. When the scale is either too small or too large, then the importance sampling estimator becomes very poor.

We can take the idea further. Assume that instead of drawing an independent sample from ν, the distribution ν is already approximated by a set *particles*, each associated to a non-negative *weight*.

Definition 10.2. *A weighted sample* $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ *is said to be* adapted *to* $\mathcal{F}^N \subset \mathcal{F}$ *if* $\sigma(\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N) \subset \mathcal{F}^N$.

A weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ *is* consistent *for the probability measure* $\mu \in \mathbb{M}_1(\mathcal{X})$ *if, as* $N \to \infty$,

$$\sum_{i=1}^N \frac{\omega^{N,i}}{\Omega^N} f\left(X^{N,i}\right) \xrightarrow{\mathbb{P}} \mu(f)\,, \quad \text{for any } f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})\,, \tag{10.4}$$

$$\max_{1 \le i \le N} \frac{\omega^{N,i}}{\Omega^N} \xrightarrow{\mathbb{P}} 0\,, \tag{10.5}$$

where Ω^N is the sum of the importance weights

$$\Omega^N := \sum_{i=1}^{N} \omega^{N,i} . \tag{10.6}$$

A weighted sample is a triangular array of random variables: For different values of N, say $N \neq M$, the $(\omega^{N,i}, X^{N,i})$ and $(\omega^{M,i}, X^{M,i})$ are not necessarily equal for any given $i \leq M \wedge N$. To simplify the notation, this dependence is not mentioned explicitly when it is obvious from the context.

Remark 10.3. It is not necessary for the weighted sample size to be equal to N. The weighted sample size may be equal to M_N, where M_N is a deterministic or even random sequence of integers satisfying $M_N \to \infty$ as $N \to \infty$. For simplicity, we assume in this chapter that $M_N = N$, for the weighted sample size. More general statements can be found in Douc and Moulines (2008).

We may transform a weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ consistent for ν into a weighted sample $\{(X^{N,i}, \tilde{\omega}^{N,i})\}_{i=1}^N$ consistent for μ, simply by modifying the weights. Setting $\tilde{\omega}^{N,i} = w(X^{N,i})\omega^{N,i}$, then $\{X^{N,i}, \tilde{\omega}^{N,i}\}_{i=1}^N$ is a weighted sample consistent for μ; see Exercise 10.6. We can even consider a more complex transformation. Let Q be a finite kernel on $\mathsf{X} \times \mathcal{X}$ (not necessarily Markov) and assume that we are willing to construct a weighted sample consistent for μ, where μ is given by

$$\mu = \frac{\nu Q}{\nu Q(\mathbf{1})} . \tag{10.7}$$

This type of recursive update is ubiquitous in NLSS; see Section 10.2. Assume that we have already constructed a weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ consistent for ν. We wish to apply a transformation to $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ to obtain a new weighted sample $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})\}_{i=1}^N$ consistent for μ. We will describe below one possible method to achieve this goal. Consider a Markov kernel denoted R on $\mathsf{X} \times \mathcal{X}$. Assume that there exists a function $w : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$, such that, for each $x \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$Q(x, A) = \int_{\mathsf{X}} w(x, x') R(x, \mathrm{d}x') \mathbb{1}_A(x') . \tag{10.8}$$

If the kernels Q and R have densities denoted by q and r with respect to the same dominating measure, then we simply have to set

$$w(x, x') = \begin{cases} q(x, x')/r(x, x') & r(x, x') \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10.9}$$

The new weighted sample $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})\}_{i=1}^N$ is constructed as follows. For $i = 1, \ldots, N$, we draw $\tilde{X}^{N,i}$ from the proposal kernel $R(X^{N,i}, \cdot)$ conditionally independently given $\mathcal{F}^N$, where $\sigma(\{(X^{N,j}, \omega^{N,j})\}_{j=1}^N) \subset \mathcal{F}^N$. By construction, for any $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$,

$$\mathbb{E}\left[f(\tilde{X}^{N,i}) \,\middle|\, \mathcal{F}^N \right] = R f(X^{N,i}) . \tag{10.10}$$

Note that we can take $\mathcal{F}^N = \sigma(\{(X^{N,j}, \omega^{N,j})\}_{j=1}^N)$ if we are only performing a single

step analysis, but the σ-algebra $\mathcal{F}^N$ can be chosen larger than that (we will see examples of this later, when we will apply these results sequentially). We then associate to each new particle positions the importance weight:

$$\tilde{\omega}^{N,i} = \omega^{N,i} w(X^{N,i}, \tilde{X}^{N,i}) \,, \quad \text{for } i = 1, \ldots, N. \tag{10.11}$$

We may now state the main consistency result for the importance sampling estimator.

Theorem 10.4. *Assume that the weighted sample* $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ *is adapted to* $\mathcal{F}^N$ *and consistent for* ν. *Then, the weighted sample* $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})\}_{i=1}^N$ *defined by* (10.10) *and* (10.11) *is consistent for* μ.

Proof. We show first that for any $f \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

$$\frac{1}{\Omega^N} \sum_{j=1}^N \tilde{\omega}^{N,j} f(X^{N,j}, \tilde{X}^{N,j}) \xrightarrow{\mathbb{P}} \nu \otimes Q(f) \,, \tag{10.12}$$

where $\tilde{X}^{N,j}$ and $\tilde{\omega}^{N,j}$ are defined in (10.10) and (10.11), respectively. Here, $\nu \otimes Q$ is the tensor product of the probability ν and the kernel Q, which is the measure defined for any $C \in \mathcal{X}^{\otimes 2}$,

$$\nu \otimes Q(C) = \iint \nu(\mathrm{d}x) Q(x, \mathrm{d}x') \mathbb{1}_C(x, x') \,. \tag{10.13}$$

The definition (10.8) implies that,

$$\begin{aligned} \mathbb{E}\left[\tilde{\omega}^{N,i} f(X^{N,i}, \tilde{X}^{N,i}) \,\middle|\, \mathcal{F}^N\right] &= \omega^{N,i} \int w(X^{N,i}, x') R(X^{N,i}, \mathrm{d}x') f(X^{N,i}, x') \\ &= \omega^{N,i} \int Q(X^{N,i}, \mathrm{d}x') f(X^{N,i}, x') = \omega^{N,i} \delta_{X^{N,i}} \otimes Q(f) \,. \end{aligned} \tag{10.14}$$

Therefore, we get

$$\sum_{i=1}^N \mathbb{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) \,\middle|\, \mathcal{F}^N \right] = \sum_{i=1}^N \frac{\omega^{N,i}}{\Omega^N} \delta_{X^{N,i}} \otimes Q(f) \,.$$

Noting that the weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is consistent, and that the function $x \mapsto \delta_x \otimes Q(f)$ is bounded, using that $\int \nu(\mathrm{d}x) \delta_x \otimes Q(f) = \nu \otimes Q(f)$, $\sum_{i=1}^N (\omega^{N,i}/\Omega^N) \delta_{X^{N,i}} \otimes Q(f) \to_{\mathbb{P}} \nu \otimes Q(f)$ as N goes to infinity. We will now show that

$$\sum_{j=1}^N \left\{ \frac{\tilde{\omega}^{N,j}}{\Omega^N} f(X^{N,j}, \tilde{X}^{N,j}) - \mathbb{E}\left[\frac{\tilde{\omega}^{N,j}}{\Omega^N} f(X^{N,j}, \tilde{X}^{N,j}) \,\middle|\, \mathcal{F}^N \right] \right\} \xrightarrow{\mathbb{P}} 0 \,. \tag{10.15}$$

Put $U_{N,j} = (\tilde{\omega}^{N,j}/\Omega^N) f(X^{N,j}, \tilde{X}^{N,j})$ for $j = 1, \ldots, N$ and appeal to Theorem B.18 on the convergence of triangular array of random variables. There are two key conditions

to check, the *tightness* (B.4) and the asymptotic negligibility (B.5). We first check (B.4). Note that

$$\sum_{j=1}^{N} \mathbb{E}\left[\,|U_{N,j}|\,\middle|\,\mathcal{F}^N\right] = \sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N}\,\delta_{X^{N,i}} \otimes Q(|f|) \xrightarrow{\mathbb{P}} \nu \otimes Q(|f|)\,,$$

showing that the sequence $\left\{\sum_{j=1}^{N} \mathbb{E}\left[\,|U_{N,j}|\,\middle|\,\mathcal{F}^N\right]\right\}_{N\geq 0}$ is tight (Theorem B.18-Eq.(B.4)). We now check the *negligibility* condition (B.5), i.e., for any $\varepsilon > 0$, put $A_N := \sum_{j=1}^{N} \mathbb{E}\left[\,|U_{N,j}|\mathbb{1}\{|U_{N,j}| \geq \varepsilon\}\,\middle|\,\mathcal{F}^N\right] \to_{\mathbb{P}} 0$. For all $C,\ \varepsilon > 0$,

$$A_N \mathbb{1}\left\{\max_{1\leq i\leq N} \omega^{N,i}/\Omega^N \leq \varepsilon/C\right\} \leq \sum_{j=1}^{N} (\omega^{N,j}/\Omega^N)\delta_{X^{N,j}} \otimes R\left([w|f|]_C\right) \xrightarrow{\mathbb{P}} \nu \otimes R\left([w|f|]_C\right)\,, \quad (10.16)$$

where for $u \in \mathbb{R}^+$, $[u]_C = u\mathbb{1}_{\{u\geq C\}}$. By dominated convergence, the right-hand side can be made arbitrarily small by letting $C \to \infty$. Since $\max_{1\leq i\leq N} \omega^{N,i}/\Omega^N \to_{\mathbb{P}} 0$, A_N tends to zero in probability, showing (B.5). Thus Theorem B.18 applies and (10.12) holds; in addition, $\sum_{j=1}^{N} \tilde{\omega}^{N,j}/\Omega^N \to_{\mathbb{P}} \nu Q(1)$. Combined with (10.12) this shows that, for $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$,

$$\sum_{j=1}^{N} \frac{\tilde{\omega}^{N,j}}{\widetilde{\Omega}^N} f(\tilde{X}^{N,j}) \xrightarrow{\mathbb{P}} \mu(f)\,.$$

It remains to prove that $\max_{1\leq j\leq N} \tilde{\omega}^{N,j}/\widetilde{\Omega}^N \to_{\mathbb{P}} 0$. Because $\widetilde{\Omega}^N/\Omega^N \to_{\mathbb{P}} \nu Q(1)$, it suffices to show that $\max_{1\leq j\leq N} \tilde{\omega}^{N,j}/\Omega^N \to_{\mathbb{P}} 0$. For any $C > 0$, by applying (10.12), we get

$$\max_{1\leq j\leq N} \frac{\tilde{\omega}^{N,j}}{\Omega^N} \mathbb{1}_{\{w(X^{N,j},\tilde{X}^{N,j})\leq C\}} \leq C \max_{1\leq i\leq N} \frac{\omega^{N,i}}{\Omega^N} \xrightarrow{\mathbb{P}} 0\,,$$

$$\max_{1\leq j\leq N} \frac{\tilde{\omega}^{N,j}}{\Omega^N} \mathbb{1}_{\{w(\tilde{X}^{N,j})>C\}} \leq \sum_{j=1}^{N} \frac{\tilde{\omega}^{N,j}}{\Omega^N} \mathbb{1}_{\{w(X^{N,j},\tilde{X}^{N,j})>C\}} \xrightarrow{\mathbb{P}} \nu \otimes Q\left(\{w > C\}\right)\,.$$

The term in the RHS of the last equation goes to zero as $C \to \infty$, which concludes the proof. ∎

Next, we discuss the asymptotic normality of the estimator. Asymptotic normality is crucial to assess the dispersion of the estimators and compute, in particular, confidence intervals. We first need to extend our definition of consistent weighted samples.

Definition 10.5 (Asymptotically normal weighted samples). *Let $\mu \in \mathbb{M}_1(\mathcal{X})$ and $\zeta \in \mathbb{M}_+(\mathcal{X})$. A weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^{N}$ on X is said to be* asymptotically

normal *for* (μ,σ,ζ) *if, for any* $f\in\mathbb{F}_b(\mathsf{X},\mathcal{X})$,

$$N^{1/2}\sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\{f(X^{N,i})-\mu(f)\}\stackrel{\mathbb{P}}{\Longrightarrow}\mathrm{N}\{0,\sigma^2(f)\}\,, \tag{10.17}$$

$$N\sum_{i=1}^{N}\left(\frac{\omega^{N,i}}{\Omega^N}\right)^2 f(X^{N,i})\stackrel{\mathbb{P}}{\longrightarrow}\zeta(f)\,, \tag{10.18}$$

$$N^{1/2}\max_{1\le i\le N}\frac{\omega^{N,i}}{\Omega^N}\stackrel{\mathbb{P}}{\longrightarrow}0\,, \tag{10.19}$$

where Ω^N *is defined in* (10.6).

We establish the asymptotic normality of the importance sampling estimator defined by (10.10) and (10.11) in the following theorem.

Theorem 10.6. *Suppose that the assumptions of Theorem 10.4 hold. Assume in addition that the weighted sample* $\{(X^{N,i},\omega^{N,i})\}_{i=1}^N$ *is asymptotically normal for* (ν,σ,ζ)*. Then, the weighted sample* $\{(\tilde{X}^{N,i},\tilde{\omega}^{N,i})\}_{i=1}^N$ *is asymptotically normal for* $(\mu,\tilde{\sigma},\tilde{\zeta})$ *with*

$$\tilde{\zeta}(f):=\{\nu Q(1)\}^{-2}\iint\zeta(\mathrm{d}x)R(x,\mathrm{d}x')w^2(x,x')f(x')\,,$$

$$\tilde{\sigma}^2(f):=\frac{\sigma^2\{Q[f-\mu(f)]\}}{\{\nu Q(1)\}^2}+\tilde{\zeta}([f-\mu(f)]^2)-\frac{\zeta(\{Q[f-\mu(f)]\}^2)}{\{\nu Q(1)\}^2}\,.$$

Proof. Pick $f\in\mathbb{F}_b(\mathsf{X},\mathcal{X})$ and assume, without loss of generality, that $\mu(f)=0$. Write $\sum_{i=1}^N\tilde{\omega}^{N,i}/\widetilde{\Omega}^N f(\tilde{X}^{N,i})=(\Omega^N/\widetilde{\Omega}^N)\,(A_N+B_N)$, with

$$A_N=\sum_{j=1}^{N}\mathbb{E}\left[\frac{\tilde{\omega}^{N,j}}{\Omega^N}f(\tilde{X}^{N,j})\,\middle|\,\mathcal{F}^N\right]=\sum_{j=1}^{N}\frac{\omega^{N,j}}{\Omega^N}Qf(X^{N,j})\,,$$

$$B_N=\sum_{j=1}^{N}\left\{\frac{\tilde{\omega}^{N,j}}{\Omega^N}f(\tilde{X}^{N,j})-\mathbb{E}\left[\frac{\tilde{\omega}^{N,j}}{\Omega^N}f(\tilde{X}^{N,j})\,\middle|\,\mathcal{F}^N\right]\right\}\,.$$

Because $\widetilde{\Omega}^N/\Omega^N\to_{\mathbb{P}}\nu Q1$ (see Theorem 10.4), the conclusion of the theorem follows from Slutsky's theorem if we prove that $N^{1/2}(A_N+B_N)\Rightarrow_{\mathbb{P}}\mathrm{N}(0,\sigma^2(Qf)+\eta^2(f))$ where

$$\eta^2(f):=\zeta\otimes R(w^2f^2)-\zeta(\{Qf\}^2)\,, \tag{10.20}$$

with w given in (10.8). Because $\{(X^{N,i},\omega^{N,i})\}_{i=1}^N$ is asymptotically normal for (ν,σ,ζ), $N^{1/2}A_N\Rightarrow_{\mathbb{P}}\mathrm{N}(0,\sigma^2(Qf))$. Next we prove that for any real u,

$$\mathbb{E}\left[\exp(\mathrm{i}uN^{1/2}B_N)\,\middle|\,\mathcal{F}^N\right]\stackrel{\mathbb{P}}{\longrightarrow}\exp\left(-(u^2/2)\eta^2(f)\right)\,,$$

where $\eta^2(f)$ is defined in (10.20). For that purpose we use Theorem B.20, and we

thus need to check (B.10)-(B.11) with $U_{N,j} := N^{1/2}(\tilde{\omega}^{N,j}/\Omega^N) f(\tilde{X}^{N,j})$, $j = 1, \ldots, N$ Because $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is asymptotically normal for (ν, σ, ζ), (10.18) implies

$$\sum_{j=1}^N \mathbb{E}\left[U_{N,j}^2 \,\middle|\, \mathcal{F}^N\right] \xrightarrow{\mathbb{P}} \zeta \otimes R(w^2 f^2)\,, \quad \sum_{j=1}^N (\mathbb{E}\left[U_{N,j} \,\middle|\, \mathcal{F}^N\right])^2 \xrightarrow{\mathbb{P}} \zeta\{Qf\}^2\,,$$

showing (B.10). It then remains for us to check (B.11). For $\varepsilon > 0$, denote $C_N := \sum_{j=1}^N \mathbb{E}\left[U_{N,j}^2 \mathbb{1}_{\{|U_{N,j}| \geq \varepsilon\}} \,\middle|\, \mathcal{F}^N\right]$. Proceeding like in (10.16), for all $C > 0$, it is easily shown that

$$C_N \leq N \sum_{i=1}^N \left(\frac{\omega^{N,i}}{\Omega^N}\right)^2 \delta_{X^{N,i}} \otimes R\left([wf]_C\right) + \mathbb{1}\left\{\frac{N^{1/2} \max_{1 \leq i \leq N} \omega^{N,i}}{\Omega^N} \geq \frac{\varepsilon}{C}\right\} \sum_{j=1}^N \mathbb{E}\left[U_{N,j}^2 \,\middle|\, \mathcal{F}^N\right]\,,$$

where for $u \in \mathbb{R}^+$, $[u]_C = u^2 \mathbb{1}_{\{u \geq C\}}$. The RHS of the previous display converges in probability to $\zeta \otimes R\left([wf]_C\right)$, which can be made arbitrarily small by taking C sufficiently large. Therefore, condition (B.11) is satisfied and Theorem B.20 applies, showing that $N^{1/2}(A_N + B_N) \Rightarrow_{\mathbb{P}} \mathrm{N}\left(0, \sigma^2(Qf) + \eta^2(f)\right)$.

Consider now (10.18). Recalling that $\widetilde{\Omega}^N/\Omega^N \to_{\mathbb{P}} \nu Q(1)$, it is sufficient to show that for $f \in \mathbb{F}_b(\mathcal{X}^{\otimes 2}, \mathcal{X}^{\otimes \in})$,

$$\left(\frac{N^{1/2}}{\Omega^N}\right)^2 \sum_{j=1}^N (\tilde{\omega}^{N,j})^2 f(X^{N,j}, \tilde{X}^{N,j}) \xrightarrow{\mathbb{P}} \zeta \otimes R(w^2 f)\,, \tag{10.21}$$

Define $U_{N,j} = N(\tilde{\omega}^{N,j}/\Omega^N)^2 f(\tilde{X}^{N,j})$. Because $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is asymptotically normal for (ν, σ, ζ),

$$\sum_{j=1}^N \mathbb{E}\left[U_{N,j} \,\middle|\, \mathcal{F}^N\right] = N \sum_{i=1}^N \left(\frac{\omega^{N,i}}{\Omega^N}\right)^2 \delta_{X^{N,i}} \otimes R(w^2 f) \xrightarrow{\mathbb{P}} \zeta \otimes R(w^2 f)\,.$$

The proof of (10.21) follows from Theorem B.18 along the same lines as above. Details are omitted. Thus Theorem B.18 applies and condition (10.18) is proved.

Consider finally (10.19). Combining with $\widetilde{\Omega}^N/\Omega^N \to_{\mathbb{P}} \nu Q(1)$ (see proof of Theorem 10.4) it is sufficient to show that $C_N := (N^{1/2}/\Omega^N)^2 \max_{1 \leq j \leq N} (\tilde{\omega}^{N,j})^2 \to_{\mathbb{P}} 0$. For any $C > 0$,

$$C_N \leq C^2 N \max_{1 \leq i \leq N} \left(\frac{\omega^{N,i}}{\Omega^N}\right)^2 + N \sum_{j=1}^N \left(\frac{\tilde{\omega}^{N,j}}{\Omega^N}\right)^2 \mathbb{1}_{\{w(X^{N,j}, \tilde{X}^{N,j}) \geq C\}}\,.$$

Applying (10.21) with $f \equiv \mathbb{1}_{\{w > C\}}$, the RHS of the previous display converges in probability to $\zeta \otimes R\left(w^2 \mathbb{1}\{w \geq C\}\right)$. The proof follows since this quantity can be made as small as we wish by taking C large enough. ∎

10.2 Sequential importance sampling

We now specialize the importance sampling to NLSS models. We use the notations introduced in Definition 9.3 where M denotes the Markov transition kernel of the hidden chain, ξ is the distribution of the initial state X_0, and g denotes the transition density function of the observation given the state with respect to the measure μ on $(\mathsf{Y},\mathcal{Y})$. We denote the filtering distribution by ϕ_t, omitting the dependence with respect to the initial distribution ξ and the observations for notational simplicity, and by Q_t the kernel on $\mathsf{X}\times\mathcal{X}$ defined, for all $x\in\mathsf{X}$ and $f\in\mathbb{F}_+(\mathsf{X},\mathcal{X})$, by

$$Q_t f(x) = \int_{\mathsf{X}} M(x,\mathrm{d}x') g(x',Y_t) f(x') \,. \tag{10.22}$$

According to (9.56), the filtering distribution ϕ_t is given by

$$\phi_t(f) = \frac{\gamma_t(f)}{\gamma_t(1)} \,, \quad \text{for all } f\in\mathbb{F}_+(\mathsf{X},\mathcal{X}) \tag{10.23}$$

where $\{\gamma_t, t\in\mathbb{N}\}$ are computed recursively as follows

$$\gamma_0(f) := \xi[g_0 f] \,, \quad \gamma_t(f) := \gamma_{t-1} Q_t(f) \,, \quad t\geq 1 \,,\, f\in\mathbb{F}_+(\mathsf{X},\mathcal{X}) \,. \tag{10.24}$$

Let $\{R_t \,, t\geq 1\}$ be a family of Markov kernels on $(\mathsf{X},\mathcal{X})$ and $r_0\in\mathbb{M}_1(\mathcal{X})$. The kernels R_t will be referred to as the *proposal kernels*. We assume that there exist weight functions $w_0:\mathsf{X}\to\mathbb{R}_+$ and $w_t:\mathsf{X}\times\mathsf{X}\to\mathbb{R}_+$ such that, for any $(x,x')\in\mathsf{X}$ and $f\in\mathbb{F}_+(\mathsf{X},\mathcal{X})$,

$$\xi[g_0 f] = r_0[w_0 f] \,, \tag{10.25}$$

$$Q_t f(x) = \int w_t(x,x') R_t(x,\mathrm{d}x') f(x') \,. \tag{10.26}$$

When the kernels Q_t and R_t have densities with respect to a common dominating measure, then

$$w_t(x,x') = \frac{q_t(x,x')}{r_t(x,x')} \,, \quad (x,x')\in\mathsf{X}\times\mathsf{X} \,. \tag{10.27}$$

Assume that the weighted sample $\{(X_{t-1}^{N,i},\omega_{t-1}^{N,i})\}_{i=1}^N$ is consistent for ϕ_{t-1}. We construct a weighted sample $\{(X_t^{N,i},\omega_t^{N,i})\}_{i=1}^N$ consistent for ϕ_t as follows. In the proposal step, each particle $X_{t-1}^{N,i}$ gives birth to a single offspring, $X_t^{N,i}$, $i\in\{1,\ldots,N\}$ which is sampled conditionally independently from the past of the particles and weights, i.e., $\mathcal{F}_{t-1}^N = \sigma\{\{(X_s^{N,i},\omega_s^{N,i})\}_{i=1}^N \,,\, s\leq t-1\}$. The distribution of this offspring is specified by the proposal kernel $R_t(X_{t-1}^{N,i},\cdot)$. Next we assign to the new particle $X_t^{N,i}$, $i=1,\ldots,N$, the importance weight

$$\omega_t^{N,i} = \omega_{t-1}^{N,i} w_t(X_{t-1}^{N,i},X_t^{N,i}) \,. \tag{10.28}$$

This construction yields to the *Sequential Importance Sampling (SIS)* algorithm (Algorithm 10.1) . The first obvious choice is that of setting $R_t = M$. The weight function

Algorithm 10.1 (SIS: Sequential Importance Sampling)

Initial State: Draw an i.i.d. sample $X_0^1,\dots,X_0^N$ from r_0 and set

$$\omega_0^i = g_0(X_0^i)w_0(X_0^i) \qquad \text{for } i = 1,\dots,N\,.$$

Recursion: For $t = 1,2,\dots,$

- Draw $(X_t^1,\dots,X_t^N)$ conditionally independently given $\{X_s^j\,, j = 1,\dots,N, s = 0,\dots,t-1\}$ from the distribution $X_t^i \sim R_t(X_{t-1}^i,\cdot)$.
- Compute the updated importance weights

$$\omega_t^i = \omega_{t-1}^i w_t(X_{t-1}^i, X_t^i), \qquad i = 1,\dots,N\,.$$

then simplifies to

$$w_t(x,x') = g_t(x') = g(x',Y_t) \qquad \text{for all } (x,x') \in \mathsf{X}^2\,, \tag{10.29}$$

which *does not depend on x*. The prior kernel is often convenient: sampling from M is often straightforward, and computing the incremental weight amounts to evaluating the conditional likelihood of the new observation given the current particle. The *optimal kernel* is defined as

$$P_t^\star(x,A) = \frac{Q_t(x,A)}{Q_t(x,\mathsf{X})}\,, \tag{10.30}$$

where Q_t is given by (10.22). The kernel $P_t^\star$ may be interpreted as the conditional distribution of the hidden state X_t given X_{t-1} and the current observation Y_t. The optimal kernel was introduced in Zaritskii et al. (1975) and Akashi and Kumamoto (1977) and has been used since by many authors Liu and Chen (1995), Chen and Liu (2000), Doucet et al. (2000, 2001), Tanizaki (2003). The associated weight function

$$w_t(x,x') = Q_t(x,\mathsf{X}) \qquad \text{for } (x,x') \in \mathsf{X}^2, \tag{10.31}$$

is the conditional likelihood of the observation Y_t given the previous state $X_{t-1} = x$. Note that this weight does not depend on x'. The optimal kernel (10.30) incorporates information both on the state dynamics and on the current observation. There are however two problems with using $P_t^\star$. First, sampling from this kernel is most often computationally costly. Second, calculation of the incremental importance weight $Q_t(x,\mathsf{X})$ may be analytically intractable. However, when the observation equation is linear (Zaritskii et al., 1975), these difficulties can be overcome as illustrated in the following example:

Example 10.7 (Noisy ARCH(1)). We consider an ARCH(1) model observed in additive noise:

$$\begin{aligned} X_t &= \sigma_w(X_{t-1})W_t\,, & W_t &\sim \text{iid } \mathrm{N}(0,1)\,,\\ Y_t &= X_t + \sigma_v V_t\,, & V_t &\sim \text{iid } \mathrm{N}(0,1)\,, \end{aligned}$$

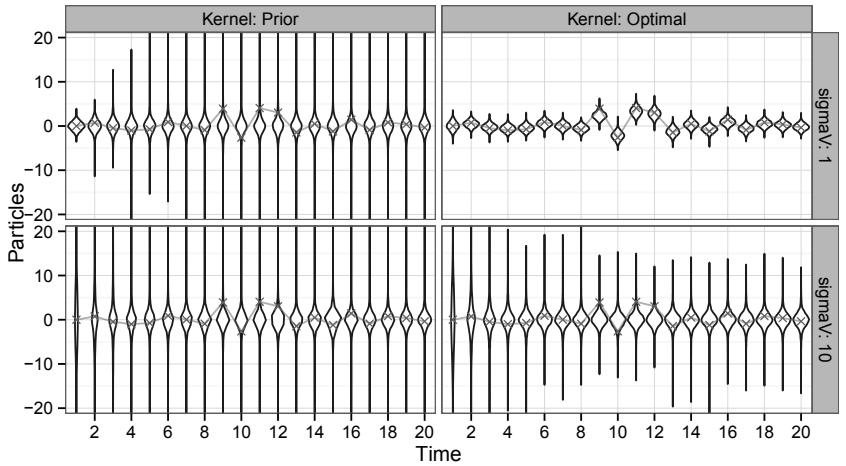


Figure 10.2 *Violin plots of the particle distributions, ignoring the importance weights, for* 5000 *particles, with prior kernel and with optimal kernel, for the ARCH(1) model of Example 10.7. The continuous line represents the actual hidden state trajectory, which is also the observation.*

with $\sigma_w^2(x) = \alpha_0 + \alpha_1 x^2$, where α_0 and α_1 are positive. The optimal kernel has density

$$p_t^\star(x, x') = \mathfrak{g}\left(x'; \tilde{m}_t(x, Y_t), \tilde{\sigma}_t^2(x)\right) , \tag{10.32}$$

where

$$\tilde{m}_t(x, Y_t) = \frac{\sigma_w^2(x) Y_t}{\sigma_w^2(x) + \sigma_v^2} , \quad \tilde{\sigma}_t^2(x) = \frac{\sigma_w^2(x)\sigma_v^2}{\sigma_w^2(x) + \sigma_v^2} . \tag{10.33}$$

The associated weight function is given by

$$Q_t(x, \mathsf{X}) = \mathfrak{g}\left(Y_t; 0, \sigma_w^2(x) + \sigma_v^2\right) . \tag{10.34}$$

When $\sigma_v^2 \gg \sigma_w^2(x)$ (the observation is non-informative), then $\tilde{m}_t(x, Y_t) \approx 0$ and $\tilde{\sigma}_t^2(x) \approx \sigma_w^2(x)$: the prior and the optimal kernel are almost identical. On the other hand, when $\sigma_v^2 \ll \sigma_w^2(x)$, then $\tilde{m}_t(x, Y_t) \approx Y_t$ and $\tilde{\sigma}_t^2(x) = \sigma_v^2$ (i.e., the observation is informative), the optimal kernel is markedly different from the prior kernel and it is expected that the optimal kernel would display better performance. This is illustrated in Figure 10.2, which compares the distribution of the posterior mean estimates achieved with both kernels, on one single simulated dataset of 10 timesteps, generated with parameters $(\alpha_0 = 1, \alpha_1 = 0.99)$. The observations Y_t were set to the simulated hidden state X_t, i.e., the mode of the local likelihood: this setting favors the prior kernel over the optimal kernel, by avoiding "extreme" observations, but makes the comparison meaningful across different values of σ_V.

SIS was then run in turn with 5000 particles for each of the two proposal kernels, for two distinct values of σ_V. The procedure was repeated independently 100 times, to obtain a sample of posterior mean estimates in the four cases. The results are

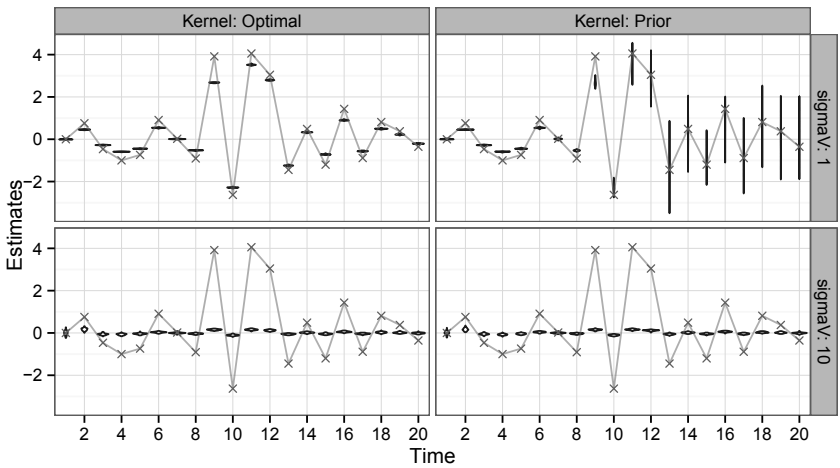


Figure 10.3 *Boxplots of the posterior mean estimates of* 100 *independent SIS runs, each using* 5000 *particles, with prior kernel and with optimal kernel, for the ARCH(1) model of Example 10.7. The continuous line represents the actual hidden state trajectory, which is also the observation.*

exhibited in Figure 10.3. With poorly informative observations, i.e., $\sigma_V = 10$, the distribution of the posterior mean estimates for both kernels are similar: the filtering distribution is mostly influenced by the dynamic equation. However, for informative observations, i.e., $\sigma_V = 1$, while the SIS for both kernels is centered around a same value, much closer to the actual hidden state, the variance of the point estimate is markedly smaller when using the optimal kernel. ◇

The weights ω_t^i measure the adequacy of the particle X_t^i to the target distribution ϕ_t. A particle such that the associated weight ω_t^i is orders of magnitude smaller than the sum Ω_t^N does not contribute to the estimator. If there are too many ineffective particles, the particle approximation is inefficient.

Unfortunately, this situation is the rule rather than the exception, as the importance weights will degenerate as the time index t increases, with most of the normalized importance weights ω_t^i/Ω_t^N close to 0 except for a few.

Example 10.8 (Example 10.7, cont.). Figure 10.4 displays the Lorenz curve of the normalized importance weights after 5, 10, 25, and 50 time steps for the noisy ARCH(1) with $\sigma_V = 1$ (see Example 10.7) for the prior kernel and the optimal kernel. The number of particles is set to 5000. The Lorenz curve is a graphical representation of the cumulative distribution function of the empirical probability distribution. Assume that X is a random variable with cumulative distribution function F and quantile function $F^{-1}(t) = \inf\{x : F(x) \geq t\}$. The Lorenz curve corresponding to any random variable X with cumulative distribution function F and finite mean $\mu = \int x \mathrm{d}F(x)$ is defined to be

$$L(p) = \mu^{-1} \int_0^p F^{-1}(t)\mathrm{d}t \,, \quad 0 \leq p \leq 1 \,. \tag{10.35}$$

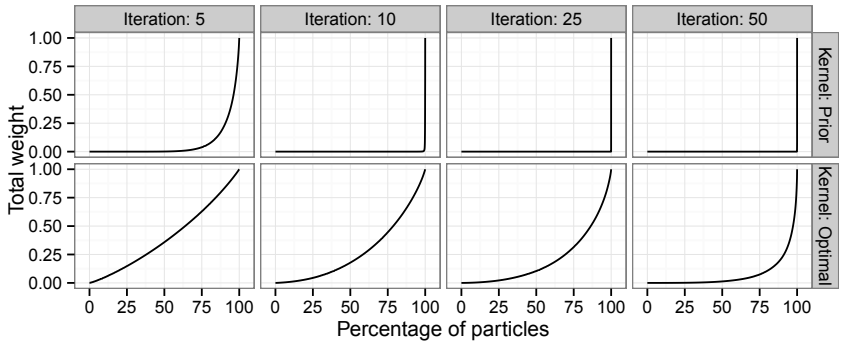


Figure 10.4 *Lorenz curves of the importance weights for the noisy ARCH model after 5, 10, 25, and 50 iterations with 5000 particles. Top panel: prior kernel; Bottom panel: optimal kernel.*

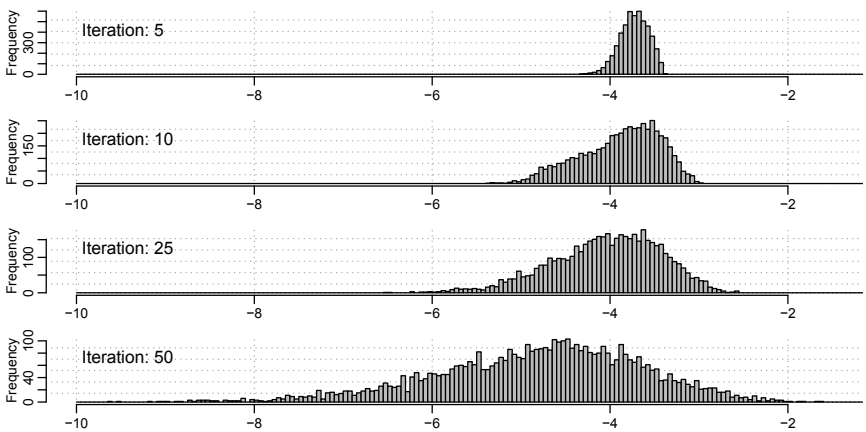


Figure 10.5 *Histograms of the base 10 logarithm of the normalized importance weights after (from top to bottom) 5, 10, 25, and 50 iterations for the noisy ARCH model with N=5000 particles.*

Applied in our setting, Figure 10.4 shows the fraction of the total sum of the importance weights that the particles of the lowest $x\%$ fraction possess. The Lorenz curves show that the normalized weights for the prior kernel quickly degenerate: the total mass is concentrated on a small fraction of particles. For the optimal kernel, the degeneracy is much slower, but after 50 steps, the bottom 75% of the importance weights accounts for less than 10% of the total weights.

Figure 10.5 displays the histogram of the base 10 logarithm of the normalized importance weights after 5, 10, 25, and 50 time steps for the same model using the optimal kernel. Figure 10.5 shows that the normalized importance weights degenerate as the number of iterations of the SIS algorithm increases. ◇

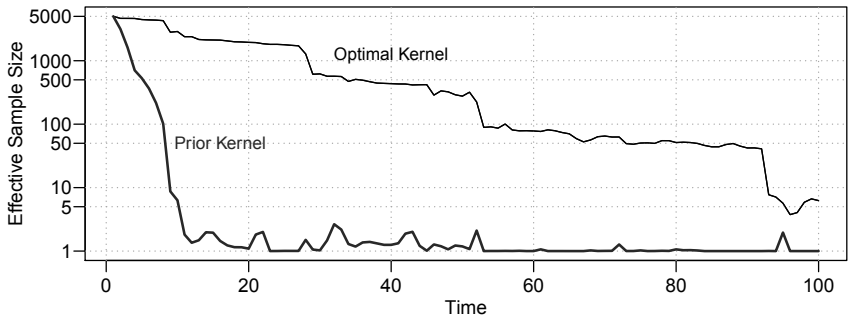


Figure 10.6 *Effective Sample Size curves of the importance weights for the noisy ARCH model, 5000 particles, 100 time points. Note how the optimal kernel improves over the prior kernel but eventually degenerates.*

A simple criterion to quantify the degeneracy of a set of importance weights $\{\omega^j\}_{i=1}^N$ is the *coefficient of variation* used by Kong et al. (1994), which is defined by

$$\mathrm{CV}\left(\{\omega^i\}_{i=1}^N\right) := \left[\frac{1}{N}\sum_{i=1}^N\left(N\frac{\omega^i}{\Omega^N}-1\right)^2\right]^{1/2}. \tag{10.36}$$

The coefficient of variation is minimal when the normalized weights are all equal to $1/N$, and then $\mathrm{CV}(\{\omega^i\}_{i=1}^N)=0$. The maximal value of $\mathrm{CV}(\{\omega^i\}_{i=1}^N)$ is $\sqrt{N-1}$, which corresponds to one of the normalized weights being one and all others being null. A related criterion is the *effective sample size* ESS (Liu, 1996), defined as

$$\mathrm{ESS}(\{\omega^i\}_{i=1}^N) = \left[\sum_{i=1}^N\left(\frac{\omega^i}{\Omega^N}\right)^2\right]^{-1} = \frac{N}{1+\left[\mathrm{CV}(\{\omega^i\}_{i=1}^N)\right]^2}, \tag{10.37}$$

which varies between 1 and N (equal weights). The effective sample size may be understood as a proxy for the equivalent number of i.i.d. samples, but this interpretation can sometimes be rather misleading.

As an example, Figure 10.6 shows the ESS curves of the importance weights for a simulated noisy ARCH model as given in Example 10.7, with 100 time points. Using 5000 particles, note how the optimal kernel performs better than the prior kernel, but eventually degenerates.

10.3 Sampling importance resampling

The solution proposed by Gordon et al. (1993) to avoid the degeneracy of the importance weights is based on *resampling* using the normalized weights as probabilities of selection. Thus, particles with small importance weights are eliminated, whereas those with large importance weights are replicated. After resampling, all importance weights are reset to one.

Algorithm 10.2 (SIR: Sampling Importance Resampling)

Sampling: Draw an i.i.d. sample $X^1,\dots,X^N$ from the instrumental distribution ν.

Weighting: Compute the (normalized) importance weights

$$\omega^i = \frac{w(X^i)}{\sum_{j=1}^N w(X^j)} \quad \text{for } i = 1,\dots,N\,.$$

Resampling:

- Draw, conditionally independently given $(X^1,\dots,X^N)$, N discrete random variables $(I^1,\dots,I^N)$ taking values in the set $\{1,\dots,N\}$ with probabilities $(\omega^1,\dots,\omega^N)$, i.e.,

$$\mathbb{P}(I^1 = j) = \omega^j\,, \quad j = 1,\dots,N\,. \tag{10.39}$$

- Set, for $i = 1,\dots,N$, $\tilde{X}^i = X^{I^i}$.

10.3.1 Algorithm description

The resampling method is rooted in the *sampling importance resampling* (or SIR) method to sample a distribution μ, introduced by Rubin (1987, 1988). We discuss this procedure first and we will then explain how this procedure can be used in combination with the SIS procedure. We first consider the SIR in the simple setting of a single step importance estimator. In this setting, the SIR proceeds in two stages. In the *sampling stage*, an i.i.d. sample $\{X^i\}_{i=1}^N$ is drawn from the proposal distribution ν. The importance weights are then evaluated at particle positions,

$$\omega^i = w(X^i)\,, \tag{10.38}$$

where w is the importance weight function defined in (10.1). In the *resampling stage*, a sample of size N denoted by $\{\tilde{X}^i\}_{i=1}^N$ is drawn from the set of points $\{X^i\}_{i=1}^N$, with probability proportional to the weights (10.38). The rationale is that particles X^i associated to large importance weights ω^i are more likely under the target distribution μ and should thus be selected with higher probability during the resampling than particles with low (normalized) importance weights.

This idea is easy to extend. Assume that $\{(X^i,\omega^i)\}_{i=1}^N$ is a weighted sample consistent for ν. We may apply the resampling stage to $\{(X^i,\omega^i)\}_{i=1}^N$, i.e., draw a sample $\{\tilde{X}^i\}_{i=1}^N$ from the set of points $\{X^i\}_{i=1}^N$ with probability proportional to the weights $\{\omega^i\}_{i=1}^N$. Doing so we obtain an equally weighted sample $\{(\tilde{X}^i,1)\}_{i=1}^N$ also targeting ν. The SIR algorithm is summarized below.

10.3.2 Resampling techniques

Denoting by M^i is the number of times that the particle X^i is resampled. With these notations, we get

$$\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}^i) = \sum_{i=1}^{N} \frac{M^i}{N} f(X^i) \,.$$

Assume that the weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is adapted to $\mathcal{F}^N$. A resampling procedure is said to be *unbiased* if

$$\mathbb{E}\left[M^i \mid \mathcal{F}^N\right] = N\omega^i/\Omega^N \,, \quad i = 1,\ldots,N \,. \tag{10.40}$$

It is easily seen that this condition implies that the conditional expectation of $N^{-1}\sum_{i=1}^N f(\tilde{X}^i)$, with respect to the weighted sample $\{(X^i, \omega^i)\}_{i=1}^N$, is equal to the importance sampling estimator,

$$\mathbb{E}\left[N^{-1}\sum_{i=1}^{N} f(\tilde{X}^i) \mid \mathcal{F}^N\right] = \sum_{i=1}^{N} \frac{\omega^i}{\Omega^N} f(X^i) \,.$$

As a consequence, the mean square error of the estimator $(1/N)\sum_{i=1}^N f(\tilde{X}^i)$ after resampling is always larger than that of the importance sampling estimator (10.3):

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}^i) - \mu(f)\right)^2 \\
&\quad = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}^i) - \sum_{i=1}^{N} \frac{\omega^i}{\Omega^N} f(X^i)\right)^2 + \mathbb{E}\left(\sum_{i=1}^{N} \frac{\omega^i}{\Omega^N} f(X^i) - \mu(f)\right)^2 \,.
\end{aligned} \tag{10.41}$$

There are several different ways to construct an unbiased sampling procedure, the most obvious approach being sampling with replacement with probability of sampling each X^i equal to the normalized importance weight ω^i/Ω^N. In this case, the $\{M^i\}_{i=1}^N$ is multinomial

$$\{M^i\}_{i=1}^N | \{X^i, \omega^i\}_{i=1}^N \sim \mathsf{Mult}\left(N, \left\{\frac{\omega^i}{\Omega^N}\right\}_{i=1}^N\right) \,. \tag{10.42}$$

Another possible solution is the *deterministic plus residual multinomial resampling*, introduced in Liu and Chen (1995). Denote by $\lfloor x \rfloor$ the integer part of x and by $\langle x \rangle$ the fractional part of x, $\langle x \rangle := x - \lfloor x \rfloor$. This scheme consists of retaining $\lfloor N\omega^i/\Omega^N \rfloor$, $i = 1,\ldots,N$ copies of the particles and then reallocating the remaining particles by applying the multinomial resampling procedure with the residual importance weights defined as $\langle N\omega^i/\Omega^N \rangle$. In this case, M^i may be decomposed as $M^i = \lfloor N\omega^i/\Omega^N \rfloor + H_i$ where $\{H_i\}_{i=1}^N$ are multinomial

$$\{H_i\}_{i=1}^N | \{X^i, \omega^i\}_{i=1}^N \sim \mathsf{Mult}\left(\sum_{i=1}^{N} \left\langle \frac{N\omega^i}{\Omega^N} \right\rangle, \left\{ \frac{\langle N\omega^i/\Omega^N \rangle}{\sum_{i=1}^N \langle N\omega^i/\Omega^N \rangle} \right\}_{i=1}^N\right) \,. \tag{10.43}$$

Assuming that the weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is consistent for ν, it is a natural question to ask whether the uniformly weighted sample $\{(\tilde{X}^{N,i}, 1)\}_{i=1}^N$ is still consistent for ν.

Theorem 10.9. *Assume that the weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is adapted to $\mathcal{F}^N$ and consistent for ν. Then, the uniformly weighted sample $\{(\tilde{X}^{N,i}, 1)\}_{i=1}^N$ obtained using either* (10.42) *or* (10.43) *is consistent for ν.*

Proof. See Exercise 10.14 ∎

Similarly, the resampling procedures (10.42) and (10.43) transform an asymptotically normal weighted sample for ν into an asymptotically normal sample for ν. We will discuss only the multinomial sampling case; the corresponding result for the deterministic plus multinomial residual sampling is given in Chopin (2004) and Douc and Moulines (2008).

Theorem 10.10. *Assume that $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is adapted to $\mathcal{F}^N$, consistent for ν, and asymptotically normal for (ν, σ, ζ). Then the equally weighted particle system $\{(\tilde{X}^{N,i}, 1)\}_{i=1}^N$ obtained using* (10.42) *is asymptotically normal for $(\nu, \tilde{\sigma}, \tilde{\zeta})$ with $\tilde{\sigma}^2(f) = \mathrm{Var}_\nu(f) + \sigma^2(f)$ and $\tilde{\zeta} = \nu$.*

Proof. See Exercise 10.15. ∎

10.4 Particle filter

10.4.1 Sequential importance sampling and resampling

The resampling step can be introduced in the sequential importance sampling framework outlined in Section 10.2. As shown in (10.41), the one-step effect of resampling seems to be negative, as it increases the variance, but, as we will show later, resampling is required to guarantee that the particle approximation does not degenerate in the long run. This remark suggests that it may be advantageous to restrict the use of resampling to cases where the importance weights are unbalanced. The criteria defined in (10.36) and (10.37), are of course helpful for that purpose. The resulting algorithm, which is generally known under the name of *sequential importance sampling with resampling* (SISR), is summarized in Algorithm 10.3.

Example 10.11 (Example 10.7, cont.). The histograms shown in Figure 10.7 are the counterparts of those shown in Figure 10.5. In this case, the resampling is applied whenever the coefficient of variation, (10.36), of the normalized weights exceeds one. The histograms of the normalized importance weights displayed in Figure 10.7 show that the weight degeneracy is avoided. ◇

The SISR algorithm combines importance sampling steps with resampling steps. By applying iteratively the one-step consistency results Theorem 10.4 and Theorem 10.9, a straightforward induction shows that, starting from a weighted sample $\{(X_0^{N,i}, \omega_0^{N,i})\}_{i=1}^N$ consistent for ϕ_0, then the SISR algorithm produces at each iteration a weighted sample $\{(X_t^{N,i}, \omega_t^{N,i})\}_{i=1}^N$ consistent for ϕ_t. Similarly, if the weighted sample $\{(X_0^{N,i}, \omega_0^{N,i})\}_{i=1}^N$ is asymptotically normal for ϕ_0, then by again

Algorithm 10.3 (SISR: Sequential Importance Sampling with Resampling)

Sampling: • Draw $\{\tilde{X}_t^i\}_{i=1}^N$ conditionally independently given $\{\{(X_s^j,\omega_s^j)\}_{j=1}^N, s \le t-1\}$ from the proposal kernel $\tilde{X}_t^i \sim R_t(X_{t-1}^i,\cdot)$, $i=1,\ldots,N$.

• Compute the updated importance weights

$$\tilde{\omega}_t^i = \omega_{t-1}^i w_t(X_{t-1}^i,\tilde{X}_t^i)\,, \qquad i=1,\ldots,N\,.$$

where w_t is defined in (10.26).

Resampling (Optional):

• Draw, conditionally independently given $\{(X_s^i,\omega_s^i)\}_{i=1}^N, s\le t-1\}$ and $\{\tilde{X}_t^i\}_{i=1}^N$, a multinomial trial $\{I_t^i\}_{i=1}^N$ with probabilities of success $\{\tilde{\omega}_t^i/\widetilde{\Omega}_t^N\}_{i=1}^N$ and set $X_t^i=\tilde{X}_t^{I_t^i}$ and $\omega_t^i=1$ for $i=1,\ldots,N$.

applying iteratively the one-step asymptotic normality results, the weighted sample $\{(X_t^{N,i},\omega_t^{N,i})\}_{i=1}^N$ is asymptotically normal for ϕ_t, because both the importance sampling step and the resampling step preserve asymptotic normality. The limiting variance can be computed iteratively.

As an illustration, we may consider a special instance of Algorithm 10.3 for which the resampling procedure is triggered when the coefficient of variation exceeds a threshold κ, i.e., $\mathrm{CV}(\{\tilde{\omega}_t^{N,i}\}_{i=1}^N)>\kappa$, we draw $\{I_t^{N,i}\}_{i=1}^N$ conditionally independently given $\mathcal{F}_{t-1}^N \vee \sigma\big(\{\tilde{X}_t^{N,i},\tilde{\omega}_t^{N,i}\}_{i=1}^N\big)$

$$\mathbb{P}\left(I_t^{N,i}=j\,\middle|\,\mathcal{F}_t^N\right)=\tilde{\omega}_t^{N,j}/\widetilde{\Omega}_t^N\,,\quad i=1,\ldots,N\,,j=1,\ldots,N \tag{10.44}$$

and we set $X_t^{N,i}=\tilde{X}_t^{N,I_t^{N,i}}$ and $\omega_t^{N,i}=1$ for $i=1,\ldots,N$. If $\mathrm{CV}(\tilde{\omega}_t^{N,i})\le\kappa$, we simply keep the updated particles and weights, i.e., we set $(X_t^{N,i},\omega_t^{N,i})=(\tilde{X}_t^{N,i},\tilde{\omega}_t^{N,i})$ for $i=1,\ldots,N$. We have only described the multinomial resampling, but the deterministic plus residual sampling, or even more sophisticated alternatives, can be considered as well; see Douc and Moulines (2008).

Theorem 10.12. *Assume that the equally weighted sample $\{(X_0^{N,i},1)\}_{i=1}^N$ is consistent for ϕ_0 and asymptotically normal for (ϕ_0,σ_0,ϕ_0). Assume in addition that for all $(x,y)\in\mathsf{X}\times\mathsf{Y}$, $g(x,y)>0$, $\sup_{x\in\mathsf{X}} g(x,y)<\infty$ for all $y\in\mathsf{Y}$, and $\sup_{0\le t\le n}|w_t|_\infty<\infty$.*

Then, for any $1\le t\le n$, $\{(X_t^{N,i},\omega_t^{N,i})\}_{i=1}^N$ is consistent for ϕ_t and asymptotically normal for (ϕ_t,σ_t) where the function σ_t is defined by the recursion

$$\sigma_t^2(f)=\mathrm{Var}_{\phi_t}(f)+\frac{\sigma_{t-1}^2(Q_t[f-\phi_t(f)])}{\{\phi_{t-1}Q_t(\mathbf{1})\}^2}+$$
$$\iint\frac{\phi_{t-1}(\mathrm{d}x)R_t(x,\mathrm{d}x')w_t^2(x,x')\{f(x')-\phi_t(f)\}^2-\phi_{t-1}(\{Q_t[f-\phi_t(f)]\}^2)}{\{\phi_{t-1}Q_t(\mathbf{1})\}^2}\,.$$

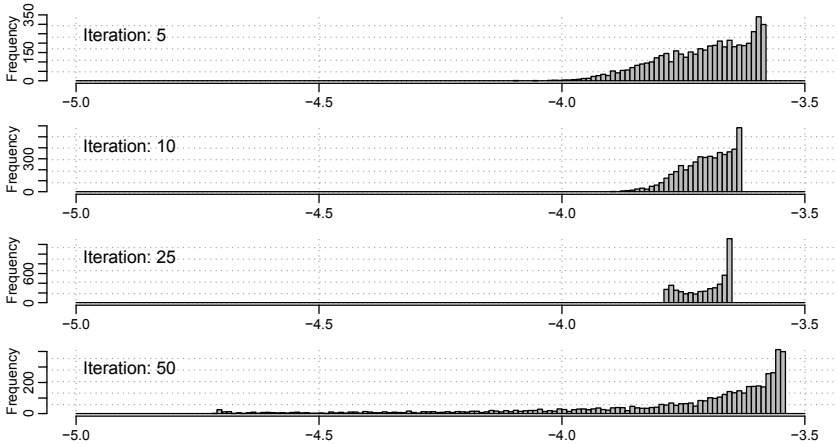


Figure 10.7 *Histograms of the base 10 logarithm of the normalized importance weights after (from top to bottom) 5, 10, 25, and 50 iterations in the noisy ARCH model of Example 10.7. Same model and data as in Figure 10.5. Resampling occurs when the coefficient of variation gets larger than 1.*

Proof. The proof is by repeated applications of Theorem 10.6 and Theorem 10.10; see Exercise 10.12. ∎

10.4.2 Auxiliary sampling

In this section, we introduce the *auxiliary particle filter* (APF) proposed by Pitt and Shephard (1999), which has proven to be one of the most useful implementations of the SMC methodology. The APF enables us to design a set of *adjustment multiplier weights* involved in the selection procedure. Assume that we at time $t-1$ have a weighted sample $\{(X_{t-1}^i, \omega_{t-1}^i)\}_{i=1}^N$ providing an approximation $\phi_{t-1}^N = \sum_{i=1}^N (\omega_{t-1}^i / \Omega_{t-1}^N)\delta_{X_{t-1}^i}$ of the filtering distribution ϕ_{t-1}. When the observation Y_t becomes available, an approximation of the filtering distribution ϕ_t may be obtained by plugging the empirical measure ϕ_{t-1}^N into the recursion (10.23), yielding, for $A \in \mathcal{X}$,

$$\phi_t^{N,\mathrm{tar}}(A) = \frac{\phi_{t-1}^N Q_t(A)}{\phi_{t-1}^N Q_t(\mathsf{X})} = \sum_{i=1}^N \frac{\omega_{t-1}^i Q_t(X_{t-1}^i, \mathsf{X})}{\sum_{j=1}^N \omega_{t-1}^j Q_t(X_{t-1}^j, \mathsf{X})} P_t^\star(X_{t-1}^i, A)\,, \tag{10.45}$$

where Q_t and $P_t^\star$ are defined in (10.22) and (10.30), respectively. Now, since we want to form a new weighted sample approximating ϕ_t, we need to find a convenient mechanism for sampling from $\phi_t^{N,\mathrm{tar}}$ given $\{(X_{t-1}^i, \omega_{t-1}^i)\}_{i=1}^N$. In most cases it is possible—but generally computationally expensive—to simulate from $\phi_t^{N,\mathrm{tar}}$ directly using *auxiliary accept-reject sampling* (see Hürzeler and Künsch, 1998, Künsch, 2005). A computationally cheaper solution consists of producing a weighted sample

approximating $\phi_t^{N,\mathrm{tar}}$ by using an importance sampling procedure. Following Pitt and Shephard (1999), this may be done by considering the *auxiliary* target distribution

$$\phi_t^{N,\mathrm{aux}}(\{i\}\times A) := \frac{\omega_{t-1}^i Q_t(X_{t-1}^i, A)}{\sum_{\ell=1}^N \omega_{t-1}^\ell Q_t(X_{t-1}^\ell, \mathsf{X})}, \quad i \in \{1,\ldots,N\}, A \in \mathcal{X}, \tag{10.46}$$

on the product space $\{1,\ldots,N\}\times\mathsf{X}$. By construction, the target distribution $\phi_t^{N,\mathrm{tar}}$ is the marginal distribution with respect to the particle index of the auxiliary distribution $\phi_t^{N,\mathrm{aux}}$. Therefore, we may construct a weighted sample targeting $\phi_t^{N,\mathrm{tar}}$ on $(\mathsf{X},\mathcal{X})$ by sampling from the auxiliary distribution, computing the associated importance weights and then discarding the indices.

To sample from $\phi_t^{N,\mathrm{aux}}$, we use an importance sampling strategy on the product space $\{1,\ldots,N\}\times\mathsf{X}$. To do this, we first draw conditionally independently pairs $\{(I_t^i, X_t^i)\}_{i=1}^N$ of indices and particles from the proposal distribution

$$\phi_t^{N,\mathrm{prop}}(\{i\}\times A) = \frac{\omega_{t-1}^i \vartheta_t(X_{t-1}^i)}{\sum_{j=1}^N \omega_{t-1}^j \vartheta_t(X_{t-1}^j)} R_t(X_{t-1}^i, A), \quad A \in \mathcal{X}, \tag{10.47}$$

on the product space $\{1,\ldots,N\}\times\mathsf{X}$, where $x \mapsto \vartheta_t(x)$ is the *adjustment multiplier weight* function and R_t is the proposal kernel. We will discuss later the choice of ϑ_t. For each draw $\{(I_t^i, X_t^i)\}_{i=1}^N$, we compute the importance weight

$$\omega_t^i = \frac{w_t(X_{t-1}^{I_t^i}, X_t^i)}{\vartheta_t(X_{t-1}^{I_t^i})}, \tag{10.48}$$

where w_t is the importance function defined in (10.26), and associate it to the corresponding particle position X_t^i. Finally, the indices $\{I_t^i\}_{i=1}^N$ are discarded. The weighted sample $\{(X_t^i, \omega_t^i)\}_{i=1}^N$ is taken as an approximation of ϕ_t. The simplest choice, yielding to the *bootstrap particle filter algorithm* proposed by Gordon et al. (1993), consists of setting, for all $x \in \mathsf{X}$, $\vartheta_t \equiv 1$ and $R_t(x,\cdot) \equiv M(x,\cdot)$. A more appealing—but often computationally costly—choice consists of using the adjustment weights $\vartheta_t(x) = \vartheta_t^\star(x) := Q_t(x,\mathsf{X})$, $x \in \mathsf{X}$, and the proposal transition kernel $P_t^\star(x,\cdot) := Q_t(x,\cdot)/Q_t(x,\mathsf{X})$. In this case, $\omega_t^i = 1$ for all $i \in \{1,\ldots,N\}$ and the auxiliary particle filter is said to be *fully adapted*. Except in some specific models, the implementation of a fully adapted sampler is computationally impractical. Heuristically, the adjustment multiplier weight function $x \mapsto \vartheta_t(x)$ should be an easy to compute proxy of $x \mapsto \vartheta_t^\star(x)$. Pitt and Shephard (1999) suggest that the adjustment multiplier weight function be set as the likelihood of the mean of the predictive distribution corresponding to each particle,

$$\vartheta_t(x) = g\left(\int x' M(x, \mathrm{d}x'), Y_{t+1}\right). \tag{10.49}$$

Other constructions are discussed in Douc et al. (2009b) and Cornebise et al. (2008).

Let n be an integer. Consider the following assumptions:

Assumption A10.13.

(a) For all $(x,y) \in \mathsf{X} \times \mathsf{Y}$ and $0 \le t \le n$, $g_t(x,y) > 0$ and $\sup_{0\le t\le n} |g_t|_\infty < \infty$.

(b) $\sup_{0\le t\le n} |\vartheta_t|_\infty < \infty$ and $\sup_{0\le t\le n} \sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} w_t(x,x')/\vartheta_t(x) < \infty$.

Theorem 10.14. *Assume A10.13. Then, for all $t \in \{0,\ldots,n\}$, the weighted sample $\{(X_t^{N,i}, \omega_t^{N,i})\}$ is consistent for ϕ_t and asymptotically normal for $(\phi_t, \sigma_t, \zeta_t)$ where*

$$\zeta_t(f) = \{\phi_{t-1}Q_t\mathbf{1}\}^{-2}\phi_{t-1}(\vartheta_t) \iint \phi_{t-1}(\mathrm{d}x) R_t(x,\mathrm{d}x') \frac{w_t^2(x,x')}{\vartheta_t(x)} f(x') \,, \tag{10.50a}$$

$$\sigma_t^2(f) = \{\phi_{t-1}Q_t\mathbf{1}\}^{-2}\sigma_{t-1}^2(Q_t[f-\phi_t(f)]) + \zeta_t([f-\phi_t(f)]^2) \,. \tag{10.50b}$$

Proof. See Exercise 10.16 ∎

Remark 10.15. The asymptotic variance is minimized if the auxiliary weights are chosen to be equal to (see Exercise 10.19)

$$\vartheta_t^{\mathrm{opt}}(x) = \left[\int R_t(x,\mathrm{d}x') w_t^2(x,x')[f(x') - \phi_t(f)]^2\right]^{1/2} . \tag{10.51}$$

As shown in Douc et al. (2009b), this choice of the adjustment weight can be related to the choice of the sampling weights of strata for stratified sampling estimators; see Exercise 10.21. The use of the optimal adjustment weights (10.51) provides, for a given sequence $\{R_t\}_{t\ge0}$ of proposal kernels, the most efficient of all auxiliary particle filters. However, exact computation of the optimal weights is in general infeasible for two reasons: firstly, they depend (via $\phi_t(f)$) on the expectation $\phi_t(f)$, that is, the quantity that we aim to estimate, and, secondly, they involve the evaluation of a complicated integral; see Douc et al. (2009b) and Cornebise et al. (2008) for a discussion.

10.5 Convergence of the particle filter

10.5.1 Exponential deviation inequalities

Non-asymptotic deviation inequality provides an explicit bound on the probability that the particle estimator $\phi_t^N(h)$ deviates from its targeted value $\phi_t(h)$ by $\varepsilon > 0$: $\mathbb{P}(|\phi_t^N(h) - \phi_t(h)| > \varepsilon) \le r(N,\varepsilon,h)$, where the rate function r is explicit. It is possible to derive such nonasymptotic bounds for the particle approximation. To make the derivations short, we concentrate in this chapter only the auxiliary particle filter introduced in Section 10.4.2 and exponential deviation inequality for bounded functions h. We preface the proof by an elementary Lemma:

Lemma 10.16. *Assume that A_N, B_N and B are random variables such that there exist positive constants β, c_1, c_2, c_3 such that*

$$|A_N/B_N| \le c_1, \ \mathbb{P}\text{-a.s. and } B \ge \beta, \ \mathbb{P}\text{-a.s.} \tag{10.52a}$$

$$\text{For all } \varepsilon > 0 \text{ and } N \ge 1, \ \mathbb{P}(|B_N - B| > \varepsilon) \le 2\mathrm{e}^{-Nc_2\varepsilon^2} \,, \tag{10.52b}$$

$$\text{For all } \varepsilon > 0 \text{ and } N \ge 1, \ \mathbb{P}(|A_N| > \varepsilon) \le 2\mathrm{e}^{-Nc_3(\varepsilon/c_1)^2}, \tag{10.52c}$$

Then,

$$\mathbb{P}(|A_N/B_N| > \varepsilon) \le 4\mathrm{e}^{-N(c_2\wedge c_3)(\varepsilon\beta/2c_1)^2)} .$$

Proof. See Exercise 10.20. ∎

Theorem 10.17. *Assume A 10.13. For any $t \in \{0,\dots,n\}$ there exist constants $0 < c_{1,t},\ c_{2,t} < \infty$ such that, for all $N \in \mathbb{N}$, $\varepsilon > 0$, and $h \in \mathbb{F}_b(\mathsf{X},\mathcal{X})$,*

$$\mathbb{P}\left[\left|N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i) - \frac{\phi_{t|t-1}(g_t h)}{\phi_{t-1}(\vartheta_t)}\right| \ge \varepsilon\right] \le c_{1,t}\mathrm{e}^{-c_{2,t}N\varepsilon^2/|h|_\infty^2} , \tag{10.53}$$

$$\mathbb{P}\left[\left|\phi_t^N(h) - \phi_t(h)\right| \ge \varepsilon\right] \le c_{1,t}\mathrm{e}^{-c_{2,t}N\varepsilon^2/\mathrm{osc}^2(h)} , \tag{10.54}$$

where the weighted sample $\{(X_t^i,\omega_t^i)\}_{i=1}^N$ is defined in (10.48)*.*

Proof. We prove (10.53) and (10.54) together by induction on $t \ge 0$. First note that, by construction, the random variables $\{(X_t^i,\omega_t^i)\}_{1\le i\le N}$ are i.i.d. conditionally to the σ-field

$$\mathcal{F}_{t-1}^N := \sigma\{(X_s^i,\omega_s^i); 0 \le s \le t-1, 1 \le i \le N\} . \tag{10.55}$$

The Hoeffding inequality implies

$$\mathbb{P}\left[\left|N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i) - \mathbb{E}\left[N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i)\,\middle|\,\mathcal{F}_{t-1}^N\right]\right| > \varepsilon\right] \le 2\mathrm{e}^{-N\varepsilon^2/(2|w_t/\vartheta_t|_\infty^2|h|_\infty^2)} . \tag{10.56}$$

For $t = 0$, we have

$$\mathbb{E}\left[N^{-1}\sum_{i=1}^N \omega_0^i h(X_0^i)\,\middle|\,\mathcal{F}_{-1}^N\right] = \mathbb{E}\left[\omega_0^1 h(X_0^1)\,\middle|\,\mathcal{F}_{-1}^N\right] = \xi(g_0 h) = \phi_{0|-1}(g_0 h) .$$

Thus, (10.54) follows by Lemma 10.16 applied with $A_N = N^{-1}\sum_{i=1}^N \omega_0^i h(X_0^i)$, $B_N = N^{-1}\sum_{i=1}^N \omega_0^i$, and $B = \beta = \xi(g_0)$ ((10.52a), (10.52b) and (10.52c) are obviously satisfied). For $t \ge 1$, we prove (10.53) by deriving an exponential inequality for $\mathbb{E}\left[N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i)\,\middle|\,\mathcal{F}_{t-1}^N\right]$ thanks to the induction assumption. It follows from the definition that

$$\mathbb{E}\left[N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i)\,\middle|\,\mathcal{F}_{t-1}^N\right] = \frac{\sum_{i=1}^N \omega_{t-1}^i Q_t h(X_{t-1}^i)}{\sum_{\ell=1}^N \omega_{t-1}^\ell \vartheta_t(X_{t-1}^\ell)} .$$

We apply Lemma 10.16 by successively checking conditions (10.52a), (10.52b) and (10.52c) with

$$\begin{cases} A_N := \phi_{t-1}^N(Q_t h) - \frac{\phi_{t-1}(Q_t h)}{\phi_{t-1}(\vartheta_t)}\phi_{t-1}^N(\vartheta_t) \\ B_N := \phi_{t-1}^N(\vartheta_t) \\ B := \beta := \phi_{t-1}(\vartheta_t) \end{cases}$$

Note that

$$\left|\frac{\phi_{t-1}^N(Q_t h)}{\phi_{t-1}^N(\vartheta_t)}\right| = |\mathbb{E}\left[\,\omega_t^1 h(X_t^1)\,\middle|\,\mathcal{F}_{t-1}^N\right]| \le |w_t/\vartheta_t|_\infty\,|h|_\infty\,,$$
$$\left|\frac{\phi_{t-1}(Q_t h)}{\phi_{t-1}(\vartheta_t)}\right| = \left|\phi_{t-1}\left[\vartheta_t(\cdot)\int\frac{w_t(\cdot,x)}{\vartheta_t(\cdot)}R_t(\cdot,\mathrm{d}x)h(x)\right]\Big/\,\phi_{t-1}(\vartheta_t)\right| \le |w_t/\vartheta_t|_\infty\,|h|_\infty\,.$$

Thus, condition (10.52a) is satisfied with $c_1 = 2|w_t/\vartheta_t|_\infty\,|h|_\infty$. Now, assume that the induction assumption (10.54) holds where t is replaced by $t-1$. Then, $A_N = \phi_{t-1}^N(H_t)$ where

$$H_t(x) := Q_t h(x) - \frac{\phi_{t-1}(Q_t h)}{\phi_{t-1}(\vartheta_t)}\vartheta_t(x)\,.$$

By noting that $\phi_{t-1}(H_t) = 0$, exponential inequalities for A_N and $B_N - B$ are then directly derived from the induction assumption under A 10.13. Thus Lemma 10.16 applies and finally (10.53) is proved for $t \ge 1$.

Finally, we must show that (10.53) implies (10.54). Without loss of generality, we assume that $\phi_t(h) = 0$. We then apply Lemma 10.16 with $A_N := N^{-1}\sum_{i=1}^N \omega_t^i h(X_t^i)$, $B_N := N^{-1}\sum_{i=1}^N \omega_t^i$, and $B := \beta := \phi_{t-1}(Q_t\mathbf{1})/\phi_{t-1}(\vartheta_t)$. But, $\phi_t(h) = 0$ implies $\phi_{t-1}(Q_t h) = 0$, so that conditions (10.52a), (10.52b), and (10.52c) follow from (10.53). ∎

10.5.2 Time-uniform bounds

The results above establish the convergence, as the number of particles N tends to infinity, of the particle filter for a *finite* time horizon $t \in \mathbb{N}$. For *infinite time horizons*, i.e., when t tends to infinity, the convergence is less obvious. Indeed, each recursive update of the weighted particles $\{(X_t^{N,i}, \omega_t^{N,i})\}_{i=1}^N$ is based on the implicit assumption that the empirical measure ϕ_{t-1}^N associated with the ancestor sample approximates perfectly well the filtering distribution ϕ_{t-1} at the previous time step; however, since the ancestor sample is marred by a sampling error itself, one may expect that the errors induced at the different updating steps accumulate and, consequently, that the total error propagated through the algorithm increases with t; this would for example imply that the asymptotic variance $\sigma_t^2(f)$ grows to infinity as $t \to \infty$, which would make the algorithm useless in practice.

Fortunately, as we will show below, the convergence of particle filters can be shown to be *uniform* in time under rather general conditions. To make the presentation simple, we will derive in this Section such stability results under very stringent conditions, but stability can be established for NLSS models under much milder assumptions.

We may decompose the error $\phi_t^N(h) - \phi_t(h)$ as follows

$$\phi_t^N(h) - \phi_t(h) = \underbrace{\phi_t^N(h) - \frac{\phi_{t-1}^N(Q_t h)}{\phi_{t-1}^N(Q_t\mathbf{1})}}_{\text{sampling error}} + \underbrace{\frac{\phi_{t-1}^N(Q_t h)}{\phi_{t-1}^N(Q_t\mathbf{1})} - \frac{\phi_{t-1}(Q_t h)}{\phi_{t-1}(Q_t\mathbf{1})}}_{\text{initialization error}}\,, \tag{10.57}$$

where we have used that $\phi_t(h) = \phi_{t-1}(Q_t h)/\phi_{t-1}(Q_t \mathbf{1})$. According to (10.57), the error $\phi_t^N(h) - \phi_t(h)$ may be decomposed into the *sampling error* introduced by replacing $\phi_{t-1}(Q_t h)/\phi_{t-1}(Q_t \mathbf{1})$ by its sampling estimate $\phi_t^N(h)$ and the *propagation error* originating from the discrepancy between empirical measure ϕ_{t-1}^N associated with the ancestor particles and the true filter ϕ_{t-1}.

By iterating the decomposition (10.57), the error $\phi_t^N(h) - \phi_t(h)$ may be written as a telescoping sum

$$\phi_t^N(h) - \phi_t(h) = \sum_{s=1}^{t} \left(\frac{\phi_s^N(Q_{s,t}h)}{\phi_s^N(Q_{s,t}\mathbf{1})} - \frac{\phi_{s-1}^N(Q_{s-1,t}h)}{\phi_{s-1}^N(Q_{s-1,t}\mathbf{1})} \right) + \frac{\phi_0^N(Q_{0,t}h)}{\phi_0^N(Q_{0,t}\mathbf{1})} - \frac{\phi_0(Q_{0,t}h)}{\phi_0(Q_{0,t}\mathbf{1})} , \quad (10.58)$$

where $Q_{t,t} = \mathrm{I}$ and for $0 \le s < t$,

$$Q_{s,t}h = Q_{s+1}\, Q_{s+2} \ldots Q_t h . \quad (10.59)$$

To prove such uniform-in-time deviation inequality, we assume that the Markov kernel M satisfies the following *strong mixing condition*.

Assumption A10.18.

(a) For all $(x,y) \in \mathsf{X} \times \mathsf{Y}$, $g(x,y) > 0$ and $\sup_{(x,y)\in \mathsf{X}\times\mathsf{Y}} g(x,y) < \infty$.

(b) $\sup_{t\ge 0} \sup_{x\in\mathsf{X}} \vartheta_t(x) < \infty$ and $\sup_{t\ge 0} \sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} w_t(x,x')/\vartheta_t(x) < \infty$.

(c) There exist constants $\sigma^+ > \sigma^- > 0$ and a probability measure ν on $(\mathsf{X}, \mathcal{X})$ such that for all $x \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\sigma^- \nu(A) \le M(x,A) \le \sigma^+ \nu(A) . \quad (10.60)$$

(d) There exists a constant $c_- > 0$ such that, $\xi(g_0) \ge c_-$ and for all $t \ge 1$,

$$\inf_{x\in\mathsf{X}} Q_t \mathbf{1}(x) \ge c_- > 0 . \quad (10.61)$$

Remark 10.19. A10.18-(b) is mild. It holds in particular under A10.18-(a) for the bootstrap filter: in this case, $\vartheta_t(x) \equiv 1$ and

$$w_t(x,x') = g_t(x') \text{ for all } x \in \mathsf{X} \text{ and } t \ge 0.$$

It automatically holds also for the fully adapted auxiliary particle filter: in this case, $\vartheta_t(x) = \int M(x,\mathrm{d}x') g_t(x') \le \sup_{t\ge 0} \sup_{x'\in\mathsf{X}} g_t(x')$ and $w_t(x,x') \equiv 1$, for all $t > 0$ and all $(x,x') \in \mathsf{X}\times\mathsf{X}$.

The key result to prove the uniform in time stability is the following uniform forgetting property.

Proposition 10.20. *Assume A10.18-*(c)*. Then, for all distributions ξ, $\xi' \in \mathbb{M}_1(\mathcal{X})$ and for all $s \le t$ and any bounded measurable functions $h \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,*

$$\left| \frac{\xi Q_{s,t} h}{\xi Q_{s,t}\mathbf{1}} - \frac{\xi' Q_{s,t} h}{\xi' Q_{s,t}\mathbf{1}} \right| \le \rho^{t-s} \operatorname{osc}(h) , \quad (10.62)$$

where $\rho := 1 - \sigma_-/\sigma_+$.

Proof. Consider the Markov kernel defined for $x_s \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\bar{Q}_{s,t}(x_s, A) = \frac{Q_{s,t}(x_s, A)}{Q_{s,t}\mathbf{1}(x_s)} \,. \tag{10.63}$$

This kernel is obtained by normalizing the kernel $Q_{s,t}$. By construction, for any $h \in \mathbb{F}(\mathsf{X}, \mathsf{X})$, we get

$$\frac{\xi Q_{s,t} h}{\xi Q_{s,t}\mathbf{1}} = \frac{\int \xi(\mathrm{d}x_s) Q_{s,t}\mathbf{1}(x_s) \bar{Q}_{s,t} h(x_s)}{\int \xi(\mathrm{d}x_s) Q_{s,t}\mathbf{1}(x_s)} = \xi_{s,t} \bar{Q}_{s,t} h \,, \tag{10.64}$$

where $\xi_{s,t}$ is the probability measure defined as

$$\xi_{s,t}(A) = \frac{\int_A \xi(\mathrm{d}x_s) Q_{s,t}\mathbf{1}(x_s)}{\int_{\mathsf{X}} \xi(\mathrm{d}x_s) Q_{s,t}\mathbf{1}(x_s)} \,. \tag{10.65}$$

We have

$$\begin{aligned} \bar{Q}_{s,t}(x_s, A) &= \frac{Q_{s,t}(x_s, A)}{Q_{s,t}\mathbf{1}(x_s)} = \frac{\int Q_{s+1}(x_s, \mathrm{d}x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1}) \bar{Q}_{s+1,t}(x_{s+1}, A)}{Q_{s,t}\mathbf{1}(x_s)} \\ &= R_{s,t} \bar{Q}_{s+1,t}(x_{s+1}, A) \,, \end{aligned}$$

where the Markov kernel $R_{s,t}$ is defined, for any $x_s \in \mathsf{X}$ and $A \in \mathcal{X}$, by

$$R_{s,t}(x_s, A) = \frac{\int_A Q_{s+1}(x_s, \mathrm{d}x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1})}{Q_{s,t}\mathbf{1}(x_s)} \,. \tag{10.66}$$

By iterating this decomposition, we may represent the Markov kernel $\bar{Q}_{s,t}$ as the product of kernels

$$\bar{Q}_{s,t} = R_{s,t} R_{s+1,t} \ldots R_{t-1,t} \,. \tag{10.67}$$

Using A10.18-(c) the kernel $R_{s,t}$ is uniformly Doeblin: for any $x_s \in \mathsf{X}$ and $A \in \mathcal{X}$, we get

$$\begin{aligned} R_{s,t}(x_s, A) &= \frac{\int_A Q_{s+1}(x_s, \mathrm{d}x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1})}{\int_{\mathsf{X}} Q_{s+1}(x_s, \mathrm{d}x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1})} \\ &\geq \frac{\sigma_-}{\sigma_+} \nu_{s,t}(A) \,, \end{aligned}$$

where $\nu_{s,t}$ is the probability on $(\mathsf{X}, \mathcal{X})$ given by

$$\nu_{s,t}(A) = \frac{\int_A \nu(\mathrm{d}x_{s+1}) g_{s+1}(x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1})}{\int_{\mathsf{X}} \nu(\mathrm{d}x_{s+1}) g_{s+1}(x_{s+1}) Q_{s+1,t}\mathbf{1}(x_{s+1})} \,. \tag{10.68}$$

Therefore, using Lemma 6.10, the Dobrushin coefficient $\Delta_{\mathrm{TV}}(R_{s,t}) \leq \rho$ of the Markov kernel $R_{s,t}$ is bounded by $\rho = 1 - \sigma_-/\sigma_+ < 1$ (see Definition 6.4). The

submultiplicativity of the Dobrushin coefficient (6.8) and the decomposition (10.67) imply that

$$\Delta_{\mathrm{TV}}\left(\bar{Q}_{s,t}\right) \leq \Delta_{\mathrm{TV}}\left(R_{s,t}\right) \Delta_{\mathrm{TV}}\left(R_{s+1,t}\right) \ldots \Delta_{\mathrm{TV}}\left(R_{t-1,t}\right) \leq \rho^{t-s} .$$

For any probability $\xi, \xi' \in \mathbb{M}_1(\mathcal{X})$, (10.64) and Lemma 6.5 imply that

$$\begin{aligned} \left\| \frac{\xi Q_{s,t}}{\xi Q_{s,t}\mathbf{1}} - \frac{\xi' Q_{s,t}}{\xi' Q_{s,t}\mathbf{1}} \right\|_{\mathrm{TV}} &= \left\| \xi_{s,t}\bar{Q}_{s,t} - \xi'_{s,t}\bar{Q}_{s,t} \right\|_{\mathrm{TV}} \\ &\leq \left\| \xi_{s,t} - \xi'_{s,t} \right\|_{\mathrm{TV}} \Delta_{\mathrm{TV}}\left(\bar{Q}_{s,t}\right) \leq \rho^{t-s} \left\| \xi_{s,t} - \xi'_{s,t} \right\|_{\mathrm{TV}} . \end{aligned}$$

The proof follows. ∎

Theorem 10.21. *Assume A 10.18. Then, the filtering distribution satisfies a time-uniform exponential deviation inequality, i.e., there exist constants c_1 and c_2 such that, for all integers N and $t \geq 0$, all measurable functions h and all $\varepsilon > 0$,*

$$\mathbb{P}\left[\left| N^{-1} \sum_{i=1}^{N} \omega_t^i h(X_t^{N,i}) - \frac{\phi_{t|t-1}(g_t h)}{\phi_{t-1}(\vartheta_t)} \right| \geq \varepsilon \right] \leq c_1 \mathrm{e}^{-c_2 N \varepsilon^2 / |h|_\infty^2} , \tag{10.69}$$

$$\mathbb{P}\left[\left| \phi_t^N(h) - \phi_t(h) \right| \geq \varepsilon \right] \leq c_1 \mathrm{e}^{-c_2 N \varepsilon^2 / \operatorname{osc}^2(h)} . \tag{10.70}$$

Proof. We first prove (10.70). Without loss of generality, we will assume that $\phi_t(h) = 0$. Similar to (Del Moral, 2004, Eq. (7.24)), the quantity $\phi_t^N(h)$ is decomposed as,

$$\phi_t^N(h) = \sum_{s=1}^{t} \left(\frac{B_{s,t}(h)}{B_{s,t}(\mathbf{1})} - \frac{B_{s-1,t}(h)}{B_{s-1,t}(\mathbf{1})} \right) + \frac{B_{0,t}(h)}{B_{0,t}\mathbf{1}} , \tag{10.71}$$

where

$$B_{s,t}(h) = N^{-1} \sum_{i=1}^{N} \omega_s^i \frac{Q_{s,t} h(X_s^i)}{|Q_{s,t}\mathbf{1}|_\infty} . \tag{10.72}$$

We first establish an exponential inequality for $B_{0,t}(h)/B_{0,t}\mathbf{1}$ where the dependence in t will be explicitly expressed. For that purpose, we will apply Lemma 10.16 by successively checking Conditions (10.52a), (10.52b), and (10.52c), with $A_N := B_{0,t}(h)$, $B_N := B_{0,t}\mathbf{1}$, and

$$B := \int \xi(\mathrm{d}x_0) g_0(x_0) \frac{Q_{0,t}\mathbf{1}(x_0)}{|Q_{0,t}\mathbf{1}|_\infty} , \quad \beta := \frac{\sigma_-}{\sigma_+} \int \xi(\mathrm{d}x_0) g_0(x_0) .$$

Under the strong mixing condition Equation 10.60, for any $0 \leq s < t$ and $(x, x') \in \mathsf{X} \times \mathsf{X}$, we have

$$\frac{Q_{s,t}\mathbf{1}(x)}{Q_{s,t}\mathbf{1}(x')} = \frac{\int \cdots \int Q_{s+1}(x, \mathrm{d}x_{s+1}) \prod_{r=s+2}^{t} Q_r(x_{r-1}, \mathrm{d}x_r)}{\int \cdots \int Q_{s+1}(x', \mathrm{d}x_{s+1}) \prod_{r=s+2}^{t} Q_r(x_{r-1}, \mathrm{d}x_r)} \geq \frac{\sigma_-}{\sigma_+} .$$

Therefore, for any $x \in \mathsf{X}$ and $0 \leq s < t$, it holds that

$$\frac{\sigma_-}{\sigma_+} \leq \frac{Q_{s,t}\mathbf{1}(x)}{|Q_{s,t}\mathbf{1}|_\infty} \leq 1 . \tag{10.73}$$

which implies that $B \geq \beta$. Since $\phi_t(h) = 0$, the forgetting condition (10.62) implies

$$\left|\frac{A_N}{B_N}\right| = \left|\frac{B_{0,t}(h)}{B_{0,t}\mathbf{1}} - \phi_t(h)\right| = \left|\frac{\sum_{i=1}^N \omega_0^i Q_{0,t}h(X_0^i)}{\sum_{i=1}^N \omega_0^i Q_{0,t}\mathbf{1}(X_0^i)} - \frac{\int \xi(\mathrm{d}x_0) g_0(x_0) Q_{0,t}h(x_0)}{\int \xi(\mathrm{d}x_0) g_0(x_0) Q_{0,t}\mathbf{1}(x_0)}\right| \leq \rho^t \operatorname{osc}(h) \ . \quad (10.74)$$

This shows condition (10.52a) with $c_1 = \rho^t \operatorname{osc}(h)$. We now turn to condition (10.52b). We have

$$B_N - B = N^{-1} \sum_{i=1}^N \omega_0^i \frac{Q_{0,t}\mathbf{1}(X_0^i)}{|Q_{0,t}\mathbf{1}|_\infty} - \int r_0(\mathrm{d}x_0) w_0(x_0) \frac{Q_{0,t}\mathbf{1}(x_0)}{|Q_{0,t}\mathbf{1}|_\infty} \ .$$

Since $\omega_0^i Q_{0,t}\mathbf{1}(X_0^i)/|Q_{0,t}\mathbf{1}|_\infty \leq |w_0|_\infty$, we have by Hoeffding's inequality

$$\mathbb{P}\left[|B_N - B| \geq \varepsilon\right] \leq 2\exp\left(-2N\varepsilon^2/|w_0|_\infty^2\right) \ .$$

We finally check condition (10.52c). We have

$$A_N = N^{-1} \sum_{i=1}^N \omega_0^i \frac{Q_{0,t}h(X_0^i)}{|Q_{0,t}\mathbf{1}|_\infty} \ .$$

Since $\phi_t(h) = 0$ implies $\int \xi(\mathrm{d}x) g_0(x) Q_{0,t}h(x) = 0$, it holds that $\mathbb{E}\left[A_N\right] = 0$. Moreover,

$$\left|\omega_0^i \frac{Q_{0,t}h(X_0^i)}{|Q_{0,t}\mathbf{1}|_\infty}\right| \leq |w_0|_\infty \left|\frac{Q_{0,t}\mathbf{1}(X_0^i)}{|Q_{0,t}\mathbf{1}|_\infty}\left(\frac{Q_{0,t}h(X_0^i)}{Q_{0,t}\mathbf{1}(X_0^i)} - \phi_t(h)\right)\right| \leq |w_0|_\infty \left|\frac{\delta_{X_0^i}Q_{0,t}h}{\delta_{X_0^i}Q_{0,t}\mathbf{1}} - \frac{\phi_0 Q_{0,t}h}{\phi_0 Q_{0,t}\mathbf{1}}\right| \leq |w_0|_\infty \rho^t \operatorname{osc}(h) \ ,$$

using (10.62) and (10.52c) follows from Hoeffding's inequality. Then, Lemma 10.16 implies

$$\mathbb{P}\left[|B_{0,t}(h)/B_{0,t}\mathbf{1}| > \varepsilon\right] \leq b\mathrm{e}^{-cN\varepsilon^2/(\rho^t \operatorname{osc}(h))^2} \ ,$$

where the constants b and c do not depend on t. We now consider for $1 \leq s \leq t$ the difference $B_{s,t}(h)/B_{s,t}(\mathbf{1})) - B_{s-1,t}(h)/B_{s-1,t}(\mathbf{1})$, where $B_{s,t}$ is defined in (10.72). We again use Lemma 10.16 with $\mathbb{P}(\cdot) = \mathbb{P}\left(\cdot \,|\, \mathcal{F}_{s-1}^N\right)$ where $\mathcal{F}_{s-1}^N$ is defined in (10.55) and

$$\begin{cases} A_N = B_{s,t}(h) - \frac{B_{s-1,t}(h)}{B_{s-1,t}(\mathbf{1})} B_{s,t}(\mathbf{1}) \\ B_N = B_{s,t}(\mathbf{1}) \\ B = \frac{\sum_{i=1}^N \omega_{s-1}^i \int Q_s(X_{s-1}^i, \mathrm{d}x) Q_{s,t}\mathbf{1}(x)}{|Q_{s,t}\mathbf{1}|_\infty \sum_{\ell=1}^N \omega_{s-1}^\ell \vartheta_s(X_{s-1}^\ell)} \end{cases}$$

Eq. (10.73) and (10.60) show that

$$B \geq \beta := \frac{c_- \sigma_-}{\sigma_+ |\vartheta_s|_\infty} \ ,$$

where σ_- and c_- are defined in (10.60) and (10.61), respectively. In addition, using the forgetting condition (10.62),

$$\left|\frac{A_N}{B_N}\right| = \left|\frac{\sum_{i=1}^N \omega_s^i Q_{s,t} h(X_s^i)}{\sum_{i=1}^N \omega_s^i Q_{s,t}\mathbf{1}(X_s^i)} - \frac{\sum_{i=1}^N \omega_{s-1}^i Q_{s-1,t} h(X_{s-1}^i)}{\sum_{i=1}^N \omega_{s-1}^i Q_{s-1,t}\mathbf{1}(X_{s-1}^i)}\right| \leq \rho^{t-s}\operatorname{osc}(h) \ , \quad (10.75)$$

showing condition (10.52a) with $c_1 = \rho^{t-s}\operatorname{osc}(h)$. We must now check condition (10.52b). By (10.56), we have

$$B_N - B = N^{-1}\sum_{i=1}^N \omega_s^i \frac{Q_{s,t}\mathbf{1}(X_s^i)}{|Q_{s,t}\mathbf{1}|_\infty} - \mathbb{E}\left[\omega_s^1 \frac{Q_{s,t}\mathbf{1}(X_s^1)}{|Q_{s,t}\mathbf{1}|_\infty}\,\middle|\,\mathcal{F}_{s-1}^N\right] ,$$

where $\mathcal{F}_{s-1}^N$ is defined in (10.55) Thus, since $\left|\omega_s^i Q_{s,t}\mathbf{1}(X_s^i)/|Q_{s,t}\mathbf{1}|_\infty\right| \leq \sup_t |w_t/\vartheta_t|_\infty$, we have by conditional Hoeffding's inequality

$$\mathbb{P}\left(|B_N - B| > \varepsilon \,|\, \mathcal{F}_{s-1}^N\right) \leq 2\mathrm{e}^{-Nc_2\varepsilon^2} ,$$

showing condition (10.52b) with $c_2 = (2\sup_t |w_t/\vartheta_t|_\infty)^{-2}$. Moreover, write $A_N = N^{-1}\sum_{\ell=1}^N \eta_s^\ell$ where

$$\eta_{s,t}^\ell(h) := \omega_s^\ell \frac{Q_{s,t}h(X_s^\ell)}{|Q_{s,t}\mathbf{1}|_\infty} - \frac{\phi_{s-1}^N(Q_{s-1,t}h)}{\phi_{s-1}^N(Q_{s-1,t}\mathbf{1})}\left(\omega_s^\ell \frac{Q_{s,t}\mathbf{1}(X_s^\ell)}{|Q_{s,t}\mathbf{1}|_\infty}\right) . \quad (10.76)$$

Since $\{(X_s^\ell, \omega_s^\ell)\}_{\ell=1}^N$ are i.i.d. conditionally to the σ-field $\mathcal{F}_{s-1}^N$, we have that $\{\eta^\ell\}_{\ell=1}^N$ are also i.i.d. conditionally to $\mathcal{F}_{s-1}^N$. Moreover, it can be easily checked using (10.56) that $\mathbb{E}\left[\eta_{s,t}^1(h)\,\middle|\,\mathcal{F}_{s-1}^N\right] = 0$. In order to apply the conditional Hoeffding inequality, we need to check that η_s^1 is bounded. This follows from (10.62),

$$|\eta_{s,t}^\ell(h)| = \omega_s^\ell \frac{Q_{s,t}\mathbf{1}(X_s^\ell)}{|Q_{s,t}\mathbf{1}|_\infty}\left|\frac{Q_{s,t}h(X_s^\ell)}{Q_{s,t}\mathbf{1}(X_s^\ell)} - \frac{\sum_{i=1}^N \omega_{s-1}^i Q_{s-1,t}h(X_{s-1}^i)}{\sum_{i=1}^N \omega_{s-1}^i Q_{s-1,t}\mathbf{1}(X_{s-1}^i)}\right|$$
$$\leq \sup_t |w_t/\vartheta_t|_\infty \rho^{t-s}\operatorname{osc}(h) .$$

Consequently,

$$\mathbb{P}\left(|A_N| > \varepsilon \,|\, \mathcal{F}_{s-1}^N\right) = \mathbb{P}\left(\left|N^{-1}\sum_{\ell=1}^N \eta_{s,t}^\ell(h)\right| > \varepsilon \,\middle|\, \mathcal{F}_{s-1}^N\right)$$
$$\leq 2\exp\left\{-Nc_3\left(\frac{\varepsilon}{\rho^{t-s}\operatorname{osc}(h)}\right)^2\right\} ,$$

with $c_3 = (2\sup_t |w_t/\vartheta_t|_\infty)^{-2}$. This shows condition (10.52c). Finally by Lemma 10.16,

$$\mathbb{P}\left(\left|\frac{B_{s,t}(h)}{B_{s,t}(\mathbf{1})} - \frac{B_{s-1,t}(h)}{B_{s-1,t}(\mathbf{1})}\right| > \varepsilon \,\middle|\, \mathcal{F}_{s-1}^N\right) \leq 4\exp\left\{-c_3 N\left(\frac{\varepsilon}{\rho^{t-s}\operatorname{osc}(h)}\right)^2\right\} .$$

The proof is concluded by using Lemma 10.22. ∎

Lemma 10.22. *Let $\{Y_{n,i}\}_{i=1}^{n}$ be a triangular array of random variables such that there exist constants $b > 0$, $c > 0$ and ρ, $0 < \rho < 1$ such that, for all n, $i \in \{1, \ldots, n\}$ and $\varepsilon > 0$, $\mathbb{P}(|Y_{n,i}| \geq \varepsilon) \leq b\mathrm{e}^{-c\varepsilon^2\rho^{-2i}}$. Then, there exists $\bar{b}$ and $\bar{c}$ such that, for any n and $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_{n,i}\right| \geq \varepsilon\right) \leq \bar{b}\mathrm{e}^{-\bar{c}\varepsilon^2} .$$

Proof. Denote by $S := \sum_{i=1}^{\infty} \sqrt{i}\rho^i$. It is plain to see that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_{n,i}\right| \geq \varepsilon\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(|Y_{n,i}| \geq \varepsilon S^{-1}\sqrt{i}\rho^i\right) \leq b\sum_{i=1}^{n} \mathrm{e}^{-cS^{-2}\varepsilon^2 i} .$$

Set $\varepsilon_0 > 0$. The proof follows by noting that, for any $\varepsilon \geq \varepsilon_0$,

$$\sum_{i=1}^{n} \mathrm{e}^{-cS^{-2}i\varepsilon^2} \leq (1 - \mathrm{e}^{-cS^{-2}\varepsilon_0^2})^{-1}\mathrm{e}^{-cS^{-2}\varepsilon^2} .$$

■

10.6 Endnotes

Importance sampling was introduced by Hammersley and Handscomb (1965) and has since been used in many different fields; see Glynn and Iglehart (1989), Geweke (1989), Evans and Swartz (1995), or Robert and Casella (2004), and the references therein.

Although the Sequential Importance Sampling (SIS) algorithm has been known since the early 1970s (Handschin and Mayne, 1969 and Handschin, 1970), its use in nonlinear filtering problems remained largely unnoticed until the early 1990s. Clearly, the available computational power was too limited to allow convincing applications of these methods. Another less obvious reason is that the SIS algorithm suffers from a major drawback that was not overcome and properly cured until the seminal papers of Gordon et al. (1993) and Kitagawa (1996). As the number of iterations increases, the importance weights degenerate: most of the particles have very small normalized importance weights and thus do not significantly contribute to the approximation of the target distribution. The solution proposed by Gordon et al. (1993) and Kitagawa (1996) is to rejuvenate the particles by replicating the particles with high importance weights while removing the particles with low weights.

Early applications of the particle filters are described in the book by Kitagawa and Gersch (1996), which included applications of spectral estimation and change points analysis. The collection of papers Doucet et al. (2001) provide a large number of methods and applications of the particle filters. The methodological papers Liu and Chen (1998) [see also the book by Liu (2001)] and Doucet et al. (2000) introduced variants of particle filters; these papers also showed that particle approximations can go far beyond filtering and smoothing problems for time series. The book Ristic et al. (2004) is devoted to the application of tracking. Recent methodological advances are covered in the survey papers by Cappé et al. (2007), Creal (2009) [presenting applications in economics and finance] and Doucet and Johansen (2009).

The convergence of the particle filter (and more generally of the interacting particle approximations of the Feynman-Kac semigroup) have been studied in a series of papers by P. Del Moral and co-authors. Early versions of the central limit theorems have been given in Del Moral and Guionnet (1999). Deviation inequalities are reported in Del Moral and Guionnet (1998). The book Del Moral (2004), which extends the survey paper Del Moral and Miclo (2000), provides a thorough coverage of the theoretical properties of sequential Monte Carlo algorithms. Recent theoretical results are presented in the survey papers Del Moral et al. (2011) and Del Moral et al. (2010). More elementary approaches of the convergence of the particle filter are presented in Chopin (2004), Künsch (2005), Cappé et al. (2005) and Douc and Moulines (2008).

The auxiliary particle filter was introduced in the work by Pitt and Shephard (1999). The consistency and asymptotic normality of the auxiliary particle filter is discussed in Douc et al. (2009b) and Johansen and Doucet (2008), which shows that the auxiliary particle filter can be seen as a particular instance of the Feynman-Kac formula. The concentration properties of interacting particle systems are studied in Del Moral et al. (2010). Non-asymptotic bounds for the auxiliary particle filter are given in Douc et al. (2010).

Exercises

10.1 (Some properties of the IS estimator). Let $\mu \in \mathbb{M}_1(\mathcal{X})$ be a target distribution and $\nu \in \mathbb{M}_1(\mathcal{X})$ be a proposal distribution. Assume that $\mu \ll \nu$ and denote $w = \mathrm{d}\mu/\mathrm{d}\nu$. Let $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$ be a function such that $\mu(|f|) < \infty$ and $\int f^2(x) w(x) \mu(\mathrm{d}x) < \infty$. Let $\{X^i\}_{i=1}^N$ be a sequence of i.i.d. random variables from ν. Denote by $\hat{\mu}_N(f) = N^{-1} \sum_{i=1}^N f(X^i) w(X^i)$ the importance sampling estimator.

(a) Show that $\hat{\mu}_N(f)$ is an unbiased estimator of $\mu(f)$.

(b) Show that $\hat{\mu}_N(f)$ is strongly consistent.

(c) Show that

$$\mathrm{Var}_\nu(fw) = [\mu(f)]^2 \, \nu\left[\left(\frac{|f|w}{\mu(|f|)} - 1\right)^2\right] .$$

(d) Under which condition does the importance sampling estimator have lower variance than the naive Monte Carlo estimator?

(e) If we choose $\nu(\mathrm{d}x) = |f(x)|\mu(\mathrm{d}x)/\mu(|f|)$, show that the variance of the IS estimator is always smaller than the variance of the naive Monte Carlo estimator.

(f) Assume that f is nonnegative. Show that the proposal distribution may be chosen in such a way that $\mathrm{Var}_\nu(fw) = 0$.

(g) Explain why this choice of the proposal distribution is only of theoretical interest.

(h) Assume that $\mathrm{Var}_\nu(fw) > 0$. Show that $\hat{\mu}_N(f)$ is asymptotically Gaussian,

$$\sqrt{N}(\hat{\mu}_N(f) - \mu(f)) \xRightarrow{\mathbb{P}} \mathrm{N}(0, \mathrm{Var}_\nu(fw)) \quad \text{as } N \to \infty .$$

10.2. Importance sampling is relevant to approximating a tail probability $\mathbb{P}(X \geq x)$. If the random variable X with density $g(x)$ has cumulant generating function $\kappa_X(t)$, then tilting by t gives a new density $h_t(x) = e^{xt-\kappa_X(t)}g(x)$.

(a) Show that the importance weight associated with an observation X drawn from $h_t(x)$ is $w_t(x) = e^{-xt+\kappa_X(t)}$.

(b) If $\mathbb{E}_t$ denotes expectation with respect to $h_t(x)$, show that the optimal tilt minimizes the second moment

$$\mathbb{E}_t[\mathbb{1}_{\{X\geq x\}}e^{-2Xt+2\kappa_X(t)}] = \mathbb{E}_0[\mathbb{1}_{\{X\geq x\}}e^{-Xt+\kappa_X(t)}] \leq e^{-xt+\kappa_X(t)} .$$

(c) It is far simpler to minimize the displayed bound than the second moment. Show that the minimum of the upper bound is attained when $K'(t) = x$.

(d) Assume that X is normally distributed with mean μ and variance σ^2. Show that the cumulant generating function is $\kappa_X(t) = \mu t + \frac{1}{2}\sigma^2 t^2$.

(e) For a given x, show that a good tilt is therefore $t = (x-\mu)/\sigma^2$.

(f) Calculate the cumulant generating functions $\kappa_X(t)$ of the exponential and Poisson distributions. Solve the equation $\kappa_X(t) = x$ for t.

(g) Suppose X follows a standard normal distribution. Write and test a program to approximate the right-tail probability $\mathbb{P}(X \geq x)$ by tilted importance sampling. Assume that x is large and positive.

10.3. Assume that the target $\mu = \mathrm{C}(0,1)$ is a standard Cauchy distribution, and the instrumental distribution $\nu = \mathrm{N}(0,1)$ is a standard Gaussian distribution.

(a) Show that the importance weight function is given by

$$w(x) = \sqrt{2\pi}\,\frac{\exp(x^2/2)}{\pi(1+x^2)} .$$

(b) Show that

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} w^2(x)\exp(-x^2/2)\,\mathrm{d}x = \infty ,$$

and conclude that the IS estimator is consistent but does not converge at an asymptotic rate $\sqrt{n}$.

(c) Plot the sample quantiles of the estimator of $\mu(f)$ versus the quantiles of a standard normal distribution when $f(x) = \exp(-|x|)$.

10.4. Assume that the target distribution is a standard normal $\mu = \mathrm{N}(0,1)$ and that the proposal distribution is Cauchy $\nu = \mathrm{C}(0,1)$.

(a) Show that the importance weight is bounded by $\sqrt{2\pi/\mathrm{e}}$.

(b) Using Exercise 10.1, show that $\sqrt{N}(\hat{\mu}_N(f) - \mu(f))$ is asymptotically normal, where $f(x) = \exp(-|x|)$ and $\hat{\mu}_N(f)$ is the importance sampling estimator of $\mu(f)$. can be applied.

(c) Plot the sample quantiles of the IS estimator $\hat{\mu}_N(f)$ versus the quantile of the standard Gaussian distribution with $(N = 50, 100, 1000)$.

(d) Assume now that $\nu = \mathrm{C}(0,\sigma)$ where $\sigma > 0$ is the scale parameter. Show that the importance weight function is bounded by $(\sqrt{2\pi}/\mathrm{e}\sigma)\mathrm{e}^{\sigma^2/2}$, $\sigma < \sqrt{2}$, $\sigma\sqrt{\pi/2}$, $\sigma \geq \sqrt{2}$.

(e) Show that the upper bound on the importance weight has a minimum at $\sigma = 1$.

(f) Argue that the choice $\sigma = 1$ leads to estimators that are better behaved than for $\sigma = 0.1$ and $\sigma = 10$.

10.5. Let f be a measurable function such that $\mu(|f|) < \infty$. Assume that $\mu \ll \nu$ and let $X^1, X^2, \ldots,$ be an i.i.d. sequence with distribution ν.

(a) Show that the self-normalized IS estimator $\hat{\mu}_N(f)$ given by (10.3) is a strongly consistent sequence of estimators of $\mu(f)$.

(b) Assume in addition that f satisfies $\int [1+f^2]w^2 \mathrm{d}\nu < \infty$. Show that the self-normalized IS estimator is asymptotically Gaussian: $\sqrt{N}\,(\hat{\mu}_N(f) - \mu(f)) \stackrel{\mathbb{P}}{\Longrightarrow} \mathrm{N}\left(0, \sigma^2(\nu, f)\right)$ where $\sigma^2(\nu,f) = \int w^2 \{f - \mu(f)\}^2 \mathrm{d}\nu$.

(c) Show that the empirical variance $\hat{\sigma}_N^2(\nu,f)$ given by

$$\hat{\sigma}_N^2(\nu,f) = N \frac{\sum_{i=1}^N (f(X^i) - \hat{\mu}_N(f))^2 w^2(X^i)}{\left(\sum_{i=1}^N w(X^i)\right)^2},$$

is a consistent sequence of estimators of $\sigma^2(\nu,f)$.

(d) Construct an asymptotically valid confidence interval for the self-normalized IS.

10.6. Let μ (known up to a normalizing constant) and ν, the proposal distribution, be probability distributions on $(\mathsf{X}, \mathcal{X})$. Suppose that for some function $w \in \mathrm{L}^1(\nu)$, we have

$$\mu(f) = \nu(wf)/\nu(w)\,. \tag{10.77}$$

Consider the weighted sample $\{(X^{N,i}, w(X^{N,i}))\}_{i=1}^N$, where for each N, $\{X^{N,i}\}_{i=1}^N$ are i.i.d. distributed according to ν.

(a) Show that, for any $\varepsilon, C > 0$,

$$N^{-1}\mathbb{E}\left[\max_{1\leq i\leq N} \omega^{N,i}\right] \leq \varepsilon + N^{-1}\sum_{i=1}^N \mathbb{E}\left[\omega^{N,i}\mathbb{1}\{\omega^{N,i} \geq \varepsilon N\}\right] \leq \varepsilon + \nu\left(w\,\mathbb{1}\{w \geq C\}\right)$$

for all $N \geq C/\varepsilon$.

(b) Show that $\nu\left(w\,\mathbb{1}\{w \geq C\}\right)$ converges to zero as $C \to \infty$.

(c) Show that the weighted sample $\{(X^{N,i}, w(X^{N,i}))\}_{i=1}^N$ is consistent for μ.

10.7 (Importance sampling; Exercise 10.6, cont.). We use the notations and the assumptions of Exercise 10.6. Assume in addition that $\nu(w^2) < \infty$. For $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$, define $S_N(f) := \sum_{i=1}^N (w(X^{N,i})/\Omega^N)[f(X^{N,i}) - \mu(f)]$.

(a) Show that $\Omega^N/N \to_{\mathbb{P}} \nu(w)$ and $N^{1/2} S_N(f) \stackrel{\mathbb{P}}{\Longrightarrow} S$, where S is Gaussian random variable with zero-mean and variance

$$\sigma^2(f) := \nu\left\{w^2[f - \mu(f)]^2\right\}\,. \tag{10.78}$$

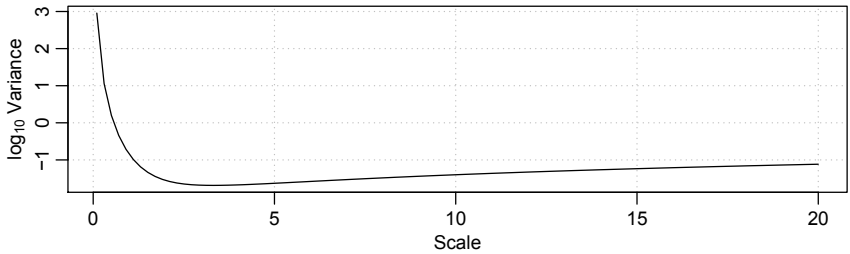


Figure 10.8 *Variance of the importance sampling estimator of the mean as a function of the scale of the t-distribution.*

(b) Show that for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,

$$N \sum_{i=1}^{N} \left(\frac{\omega^{N,i}}{\Omega^N} \right)^2 f(X^{N,i}) = N \sum_{i=1}^{N} \left(\frac{w(X^{N,i})}{\Omega^N} \right)^2 f(X^{N,i}) \xrightarrow{\mathbb{P}} \zeta(f) = \nu\left[w^2 f\right] ,$$

(c) Show that, for any $\varepsilon, C > 0$,

$$N^{-1}\mathbb{E}\left[\max_{1\le i\le N} \left(\omega^{N,i}\right)^2\right] \le \varepsilon^2 + N^{-1}\sum_{i=1}^{N} \mathbb{E}\left[\left(\omega^{N,i}\right)^2 \mathbb{1}\{\left(\omega^{N,i}\right)^2 \ge \varepsilon^2 N\}\right]$$
$$\le \varepsilon^2 + \nu\left(w^2 \mathbb{1}\{w^2 \ge C\}\right)$$

for $N \ge C/\varepsilon$.

(d) Show that $\nu\left(w^2 \mathbb{1}\{w \ge C\}\right)$ goes to zero as $C \to \infty$.

(e) Show that $(N^{1/2}/\Omega^N) \max_{1\le i\le N} \omega^{N,i} \to_{\mathbb{P}} 0$.

(f) Show that $\{(X^{N,i}, w(X^{N,i}))\}_{i=1}^{N}$ is asymptotically normal for μ.

10.8 (Variance of the IS; Exercise 10.7, cont.). (a) Compute the variance of the importance sampling estimator of the mean of a Gaussian mixture using a Student-t distribution with 4 degrees of freedom for different values of the scale.

(b) Explain Figure 10.8.

10.9 (Noisy AR(1) model). Consider an AR(1) model observed in additive noise

$$X_t = \phi X_t + \sigma_W W_t , \qquad W_t \sim \mathrm{N}(0,1) ,$$
$$Y_t = X_t + \sigma_V V_t , \qquad V_t \sim \mathrm{N}(0,1) ,$$

where $|\phi| < 1$ and $\{W_t,\, t \in \mathbb{N}\}$ and $\{V_t,\, t \in \mathbb{N}\}$ are independent Gaussian white noise processes. The initial distribution ξ is the stationary distribution of the Markov chain $\{X_t,\, t \in \mathbb{N}\}$, that is, normal with zero mean and variance $\sigma_W^2/(1-\phi^2)$.

(a) Implement the particle filter with the prior kernel.

(b) Show that the optimal kernel has a density given by

$$\mathrm{N}\left(\frac{\sigma_W^2 \sigma_V^2}{\sigma_W^2 + \sigma_V^2} \left\{ \frac{\phi x}{\sigma_W^2} + \frac{Y_t}{\sigma_V^2} \right\}, \frac{\sigma_W^2 \sigma_V^2}{\sigma_W^2 + \sigma_V^2} \right) .$$

(c) Show that the weight function is

$$Q_t(x,\mathsf{X}) \propto \exp\left[-\frac{1}{2}\frac{(Y_t-\phi x)^2}{\sigma_W^2+\sigma_V^2}\right] .$$

(d) Implement a particle filter with systematic resampling and the optimal kernel.

(e) Compare the prior and the optimal kernels when $\sigma_W \gg \sigma_V$ and $\sigma_V \ll \sigma_W$.

(f) Illustrate your conclusions by a numerical example.

10.10 (Stochastic volatility). Consider the following stochastic volatility model

$$\begin{aligned} X_t &= \phi X_{t-1} + \sigma W_t \,, \quad W_t \sim \mathrm{N}(0,1) \,, \\ Y_t &= \beta \exp(X_t/2) V_t \,, \quad V_t \sim \mathrm{N}(0,1) \,. \end{aligned}$$

(a) Show that

$$\begin{aligned} m(x,x') &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x'-\phi x)^2}{2\sigma^2}\right] , \\ g(x',Y_t) &= \frac{1}{\sqrt{2\pi\beta^2}} \exp\left[-\frac{Y_t^2}{2\beta^2}\exp(-x') - \frac{1}{2}x'\right] . \end{aligned}$$

(b) Determine the optimal kernel $p_t^\star$ and the associated importance weight.

(c) Show that the function $x' \mapsto \ln(m(x,x')g(x',Y_t))$ is concave.

(d) Show that the mode $m_t(x)$ of $x' \mapsto p_t^\star(x,x')$ is the unique solution of the non-linear equation

$$-\frac{1}{\sigma^2}(x'-\phi x) + \frac{Y_t^2}{2\beta^2}\exp(-x') - \frac{1}{2} = 0 \,.$$

(e) Propose and implement a numerical method to find this maximum.

10.11 (Stochastic volatility, cont.). We consider proposal kernel t-distribution with $\eta = 5$ degrees of freedom, with location $m_{t-1}(x)$ and scale $\sigma_t(x)$ set as square-root of minus the inverse of the second-order derivative of $x' \mapsto (\ln m(x,x')g(x',Y_t))$ evaluated at the mode $m_t(x)$.

(a) Show that

$$\sigma_{t-1}^2(x) = \left\{\frac{1}{\sigma^2} + \frac{Y_t^2}{2\beta^2}\exp\left[-m_{t-1}(x)\right]\right\}^{-1} .$$

(b) Show that the incremental importance weight is given by

$$\frac{\exp\left[-\frac{(x'-\phi x)^2}{2\sigma^2} - \frac{Y_t^2}{2\beta^2}\exp(-x') - \frac{x'}{2}\right]}{\sigma_{t-1}^{-1}(x)\left\{\eta + \frac{[x'-m_{t-1}(x)]^2}{\sigma_{t-1}^2(x)}\right\}^{-(\eta+1)/2}} .$$

(c) Implement in R a particle filter with systematic resampling using this proposal kernel.

(d) Compare numerically this implementation and the particle filter with systematic resampling and the prior kernel.

10.12. We adopt the notation of Theorem 10.12. We decompose the bootstrap filter into two steps. Starting from the equally weighted sample $\{(X_{t-1}^{N,i},1)\}_{i=1}^N$ we first form an intermediate sample $\{(\tilde{X}_t^{N,i}, w(X^{N,i},\tilde{X}^{N,i}))\}_{i=1}^N$ where $\{\tilde{X}_t^{N,i}\}_{i=1}^N$ are sampled conditionally independently from the proposal kernel $\tilde{X}_t^{N,i} \sim R_t(X_{t-1}^{N,i},\cdot)$.

(a) Show that $N^{-1}\sum_{i=1}^N w(X_{t-1}^{N,i},X_t^{N,i}) \to_{\mathbb{P}} \phi_{t-1}(Q_t 1)$.

(b) Show that

$$N^{-1/2}\sum_{i=1}^N w(X_{t-1}^{N,i},\tilde{X}_t^{N,i}) f(\tilde{X}_t^{N,i}) = N^{-1/2}\sum_{i=1}^N w(X_{t-1}^{N,i},\tilde{X}_t^{N,i}) f(\tilde{X}_t^{N,i}) - Q_t f(X_{t-1}^{N,i}) + N^{-1/2}\sum_{i=1}^N Q_t f(X_{t-1}^{N,i}) .$$

(c) Show that $\phi_{t-1}Q_t[f-\phi_t(f)]=0$ and that

$$N^{-1/2}\sum_{i=1}^N\sum_{i=1}^N Q_t[f(X_{t-1}^{N,i})-\phi_t(f)] \stackrel{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,\sigma_{t-1}^2(Q_t[f-\phi_t(f)])) .$$

(d) Show that the weighted sample $\{(\tilde{X}_t^{N,i}, w(X_{t-1}^{N,i},\tilde{X}_t^{N,i}))\}_{i=1}^N$ is asymptotically normal. Compute the asymptotic variance.

(e) Prove Theorem 10.12.

10.13 (Multinomial sampling). The *ball-in-urn* method of generating multinomial variables proceeds in two stages. In an initialization phase, the cumulative probability vector is generated, where $p_*^i = \sum_{j=1}^i \omega^i/\Omega^N$, for $i \in \{1,\ldots,N\}$. In the generation phase, N uniform random variables $\{U^i\}_{i=1}^N$ on $[0,1]$ are generated and the indices $I_i = \sup\{j \in \{1,\ldots,N\}, U^i \geq p_*^j\}$ are then computed.

(a) Show that the average number of comparisons to sample N multinomial is $N\log_2 N$ comparisons.

(b) Denote by $\{U^{(i)}\}_{i=1}^N$ the ordered uniforms. Prove that starting from $\{U^{(i)}\}_{i=1}^N$ the number of comparisons required to generate $\{I_i\}_{i=1}^N$ is at most N, worst case, only N comparisons.

(c) Show that the increments $S^i = U^{(i)} - U^{(i-1)}$, $i \in \{1,\ldots,N\}$, (where by convention $S_1 = U_{(1)}$), referred to as the *uniform spacings*, are distributed as

$$\frac{E^1}{\sum_{i=1}^{N+1} E^i},\ldots,\frac{E^N}{\sum_{i=1}^{N+1} E^i} ,$$

where $\{E^i\}_{i=1}^N$ is a sequence of i.i.d. exponential random variables.

(d) Propose an algorithm to sample N multinomial with a complexity growing only linearly with N.

Of course, much better algorithms are available; see Davis (1993). An efficient algorithm is implemented in R with a call to the built-in function `rmultinom()`, wrapped to repeat the indices.

10.14. (a) Show that the resampling procedures (10.42) and (10.43) are unbiased, i.e., for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,

$$N^{-1} \sum_{i=1}^{N} \mathbb{E}\left[f(\tilde{X}^{N,i}) \,\middle|\, \mathcal{F}^N \right] = \sum_{i=1}^{N} \omega^{N,i} / \Omega^N f(X^{N,i}) \,.$$

(b) Show that

$$N^{-1} \sum_{i=1}^{N} \mathbb{E}\left[|f(\tilde{X}^{N,i})| \mathbb{1}_{\{|f(\tilde{X}^{N,i})| \geq C\}} \,\middle|\, \mathcal{F}^N \right] = \sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} |f(X^{N,i})| \mathbb{1}_{\{|f(X^{N,i})| \geq C\}} \xrightarrow{\mathbb{P}} \nu(|f| \mathbb{1}_{\{|f| \geq C\}}) \,. \quad (10.79)$$

(c) For any $i = 1, \ldots, N$, put $U_{N,i} := N^{-1} f(\tilde{X}^{N,i})$. Show that

$$\sum_{i=1}^{N} \mathbb{E}\left[|U_{N,i}| \,|\, \mathcal{F}^N \right] = N^{-1} \sum_{i=1}^{N} \mathbb{E}\left[|f(\tilde{X}^{N,i})| \,\middle|\, \mathcal{F}^N \right] \xrightarrow{\mathbb{P}} \nu(|f|) < \infty \,,$$

(d) Show that for any $\varepsilon > 0$ and $C < \infty$, we have for all sufficiently large N,

$$\sum_{i=1}^{N} \mathbb{E}\left[|U_{N,i}| \mathbb{1}_{\{|U_{N,i}| \geq \varepsilon\}} \,\middle|\, \mathcal{F}^N \right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[|f(\tilde{X}^{N,i})| \mathbb{1}_{\{|f|(\tilde{X}^{N,i}) \geq \varepsilon N\}} \,\middle|\, \mathcal{F}^N \right] \leq N^{-1} \sum_{i=1}^{N} \mathbb{E}\left[|f(\tilde{X}^{N,i})| \mathbb{1}_{\{|f|(\tilde{X}^{N,i}) \geq C\}} \,\middle|\, \mathcal{F}^N \right] \xrightarrow{\mathbb{P}} \mu\left(|f| \mathbb{1}_{\{|f| \geq C\}} \right) \,.$$

(e) Conclude.

10.15. Assume that $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is adapted to $\mathcal{F}^N$, consistent for ν, and asymptotically normal for (ν, σ, ζ). Let $\{(\tilde{X}^{N,i}, 1)\}_{i=1}^N$ denote the equally weighted sample obtained by multinomial resampling; see (10.42). Let $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$.

(a) Show that $N^{-1} \sum_{i=1}^N f(\tilde{X}^{N,i}) - \nu(f) = A_N + B_N$ where

$$A_N = \sum_{i=1}^{N} \omega^{N,i} / \Omega^N \{ f(X^{N,i}) - \nu(f) \} \,,$$

$$B_N = N^{-1} \sum_{i=1}^{N} \left\{ f(\tilde{X}^{N,i}) - \mathbb{E}\left[f(\tilde{X}^{N,i}) \,\middle|\, \mathcal{F}^N \right] \right\} \,.$$

Prove that $N^{1/2} A_N \xRightarrow{\mathbb{P}} \mathrm{N}\left(0, \sigma^2(f)\right)$.

(b) Set $U_{N,i} := N^{-1/2} f(\tilde{X}^{N,i})$. Show that

$$\sum_{j=1}^{N} \{\mathbb{E}\left[U_{N,j}^2 \,\middle|\, \mathcal{F}^N\right] - (\mathbb{E}\left[U_{N,j} \,\middle|\, \mathcal{F}^N\right])^2\} =$$
$$N^{-1}\left(\sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} f^2(X^{N,i}) - \left\{\sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} f(X^{N,i})\right\}^2\right) \xrightarrow{\mathbb{P}} (\nu(f^2) - \{\nu(f)\}^2) .$$

(c) Pick $\varepsilon > 0$. For any $C > 0$, and N sufficiently large, show that

$$\sum_{j=1}^{N} \mathbb{E}\left[U_{N,j}^2 \mathbb{1}_{\{|U_{N,j}| \geq \varepsilon\}} \,\middle|\, \mathcal{F}^{N,j-1}\right] \leq \sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} f^2(X^{N,i}) \mathbb{1}\{|f(X^{N,i})| \geq C\}$$
$$\xrightarrow{\mathbb{P}} \nu(f^2 \mathbb{1}\{|f| \geq C\}) .$$

(d) Show that $\{(\tilde{X}^{N,i}, 1)\}_{i=1}^N$ obtained using (10.42) is asymptotically normal for $(\nu, \tilde{\sigma}, \tilde{\zeta})$ with $\tilde{\sigma}^2(f) = \mathrm{Var}_\nu(f) + \sigma^2(f)$ and $\tilde{\zeta} = \nu$.

10.16 (Proof of Theorem 10.14). Assume A10.13. Let $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$. Without loss of generality, put $\phi_t(f) = 0$, and let $N^{-1}\sum_{i=1}^N \omega_t^i f(X_t^i) = \frac{1}{N}(A_t^N + B_t^N)$, where

$$A_t^N = \mathbb{E}\left[\omega_t^1 f(X_t^1) \mid \mathcal{F}_{t-1}^N\right] ,$$
$$B_t^N = \frac{1}{N}\sum_{i=1}^{N} \left\{\omega_t^i f(X_t^i) - \mathbb{E}\left[\omega_t^i f(X_t^i) \mid \mathcal{F}_{t-1}^N\right]\right\} ,$$

where $\mathcal{F}_{t-1}^N$ is defined in (10.55).

(a) Show that

$$A_t^N = \sum_{i=1}^{N} \frac{\omega_{t-1}^i}{\sum_{j=1}^N \omega_{t-1}^j \vartheta_t(X_{t-1}^i)} Q_t f(X_{t-1}^i) .$$

(b) By applying the induction assumption, show that

$$\sqrt{N}\sum_{i=1}^{N} \frac{\omega_{t-1}^i}{\Omega_{t-1}^N} Q_t f(X_{t-1}^i) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0, \sigma_{t-1}^2(Q_t f)) .$$

(c) Show that

$$\sum_{i=1}^{N} \frac{\omega_{t-1}^i}{\Omega_{t-1}^N} \vartheta_t(X_{t-1}^i) \xrightarrow{\mathbb{P}} \phi_{t-1}(\vartheta_t) .$$

(d) Deduce that $\sqrt{N} A_t^N \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0, \{\phi_{t-1}(\vartheta_t)\}^{-2} \sigma_{t-1}^2(Q_t f))$.

(e) Show that

$$N\mathbb{E}\left[\{B_t^N\}^2 \mid \mathcal{F}_{t-1}^N\right] = \mathbb{E}\left[(\omega_t^1 f(X_t^1))^2 \mid \mathcal{F}_{t-1}^N\right] - \left(\mathbb{E}\left[\omega_t^1 f(X_t^1) \mid \mathcal{F}_{t-1}^N\right]\right)^2 .$$

(f) Show that

$$\mathbb{E}\left[(\omega_t^1 f(X_t^1))^2 \mid \mathcal{F}_{t-1}^N\right] \xrightarrow{\mathbb{P}} \alpha_t^2 = \iint \frac{\phi_{t-1}(\mathrm{d}x)}{\phi_{t-1}(\vartheta_t)} \frac{w_t^2(x,x')}{\vartheta_t(x)} Q_t(x,\mathrm{d}x') f^2(x') \,.$$

(g) Show that $\mathbb{E}\left[\omega_t^1 f(X_t^1) \mid \mathcal{F}_{t-1}^N\right] \xrightarrow{\mathbb{P}} 0$, $\sqrt{N}B_t^N \xRightarrow{\mathbb{P}} \mathrm{N}(0,\alpha_t^2)$, and that $\sqrt{N}A_t^N$ and $\sqrt{N}B_t^N$ are asymptotically independent.

(h) Show that $\Omega_t^N/N \xrightarrow{\mathbb{P}} \phi_{t-1}(Q_t 1)/\phi_{t-1}(\vartheta_t)$ and conclude.

10.17. Assume A10.13 and that $\vartheta_s \equiv 1$ for all $t \in \{1,\ldots,n\}$ and $Q_t \equiv M$. Show that for any $t \in \{0,\ldots,n\}$ and any $f \in \mathbb{F}_b(\mathsf{X},\mathcal{X})$, $\sqrt{N}(\phi_t^N(f) - \phi_t(f))$ is asymptotically normal with zero mean and variance

$$\sigma_t^2(f) = \sum_{s=1}^{t} \frac{\mathrm{Var}_{\phi_s}\left\{Q_{s,t}[f - \phi_t(f)]\right\}}{\left[\phi_s Q_{s,t}(\mathbf{1})\right]^2} + \frac{\sigma_0^2\left\{Q_{0,t}[f-\phi_t(f)]\right\}}{[\phi_0 Q_{0,t}(\mathbf{1})]^2} \,,$$

where $Q_{s,t}$ is defined in (10.59). Assume A10.18. Show that $\sup_{t\geq 0}\sigma_t^2(f) < \infty$.

10.18. In this exercise, we provide an alternate proof of the central limit theorem for the auxiliary particle filter, based on the decomposition (10.58). This proof allows us to obtain an explicit expression of the asymptotic variance. Assume A10.13. Set for $s \in \{0,\ldots,t\}$, $W_{s,t}^N(h) = N^{-1/2}\sum_{\ell=1}^N \eta_{s,t}^{N,\ell}(h)$ where, for $s \in \{1,\ldots,t\}$,

$$\eta_{s,t}^{N,\ell}(h) := \omega_s^\ell \left\{ Q_{s,t}h(X_s^\ell) - \frac{\phi_{s-1}^N(Q_{s-1,t}h)}{\phi_{s-1}^N(Q_{s-1,t}\mathbf{1})} Q_{s,t}\mathbf{1}(X_s^\ell) \right\} \,,$$

and for $s = 0$,

$$\eta_{0,t}^N := \omega_0^\ell \left\{ Q_{0,t}h(X_s^\ell) - \frac{\phi_0(Q_{0,t}h)}{\phi_0^N(Q_{0,t}\mathbf{1})} Q_{0,t}\mathbf{1}(X_0^\ell) \right\} \,.$$

Denote $\alpha_{s,t}^N = \left\{N^{-1}\sum_{\ell=1}^N Q_{s,t}\mathbf{1}(X_s^\ell)\right\}^{-1}$.

(a) Show that $\sqrt{N}\{\phi_t^N(h) - \phi_t(h)\} = \sum_{s=1}^t \alpha_{s,t}^N W_{s,t}^N(h)$.

(b) Show that, for all $s \in \{0,\ldots,t\}$, $\{\eta_{s,t}^\ell(h)\}_{\ell=1}^N$ are zero-mean and i.i.d. conditionally on $\mathcal{F}_{s-1}^N$, defined in (10.55).

(c) Show that, for all $s \in \{0,\ldots,t\}$, $W_{s,t}^N(h)$ converges in distribution to a zero-mean Gaussian variable, conditionally independent of $\mathcal{F}_{s-1}^N$.

(d) Show that the $[W_{0,t}^N(h), W_{1,t}^N(h), \ldots, W_{t,t}^N(h)]$ converges in distribution to a zero-mean random vector with diagonal covariance matrix.

(e) Show that, for $s \in \{1,\ldots,t\}$, $\alpha_{s,t}^N \xrightarrow{\mathbb{P}} \phi_{s-1}(\vartheta_s)/\phi_s(\mathbf{1})$.

(f) Show that, $\sqrt{N}[(\phi_0^N(h) - \phi_0(h)), \ldots, (\phi_n^N(h) - \phi_n(h))]$ converges to a multivariate Gaussian distribution with zero mean and covariance matrix Γ_n.

10.19. Show that, for any $\nu \in \mathbb{M}_1(\mathcal{X})$ and any functions $f, g \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$,

$$\left(\int \nu(\mathrm{d}x) f^{1/2}(x) \right)^2 \leq \int \nu(\mathrm{d}x) g(x) \int \nu(\mathrm{d}x) \frac{f(x)}{g(x)} \, .$$

Establish Eq. (10.51).

10.20. We use the notations and definitions of Lemma 10.16. Show that: $|A_N/B_N| \leq B^{-1}|A_N/B_N||B - B_N| + B^{-1}|A_N|$. Conclude.

10.21. We consider the problem of estimating, for some given measurable target function f, the expectation $\pi(f)$, where $\pi = \sum_{i=1}^d \omega_i \mu_i$. $\omega_i \geq 0$, $\sum_{i=1}^d \omega_i = 1$ and $\{\mu_i\}_{i=1}^N \in \mathbb{M}_1(\mathcal{X})$. In order to relate this to the particle filtering paradigm, we will make use of the following algorithm. Let $\{\nu_i\}_{i=1}^d \subset \mathbb{M}_1(\mathcal{X})$ be probability measures such that $\mu_i(A) = \int_A w_i(x) \nu_i(\mathrm{d}x)$ for some $w_i \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$.

For $k = 1$ to N,

(i) draw an index $I^{N,k}$ multinomially with probability proportional to the weights $\omega_i \tau_i$, $1 \leq i \leq d$;

(ii) simulate $X^k \sim \nu_{I^{N,k}}$.

Subsequently, having at hand the sample $\{X^{N,i}\}_{i=1}^N$, we use

$$\widehat{\pi}^N(f) = \frac{\sum_{k=1}^N \tau_{I^{N,k}}^{-1} w_{I^{N,k}}(X^{N,k}) f(X^{N,k})}{\sum_{k=1}^N \tau_{I^{N,k}}^{-1} w_{I^{N,k}}(X^{N,k})} \, .$$

(a) Show that

$$N^{1/2} \left[\widehat{\pi}^N(f) - \pi(f) \right] \stackrel{\mathbb{P}}{\Longrightarrow} \mathrm{N}\left(0, \sum_{i=1}^d \frac{\omega_i \alpha_i(f)}{\tau_i} \right) \, , \tag{10.80}$$

where $\alpha_i(f) := \int_{\mathsf{X}} [w_i(x)]^2 [f(x) - \pi(f)]^2 \, \nu_i(\mathrm{d}x)$.

(b) Show that the adjustment weights $\tau_i^* := \alpha_i^{1/2}(f)$, $i = 1, \ldots, d$ minimize the asymptotic variance of the limiting distribution (10.80).