Preface

This work is aimed at mathematics students in the area of stochastic dynamical systems and at engineering graduate students in signal processing and control systems. First-year graduate-level students with some background in systems theory and probability theory can tackle much of this material, at least once the techniques of Chapter 2 are mastered (with reference to the Appendices and some tutorial help). Even so, most of this work is new and would benefit more advanced graduate students. Familiarity with the language of the general theory of random processes and measure-theoretic probability will be a help to the reader. Well-known results such as the Kalman filter and Wonham filter, and also H^2, H^∞ control, emerge as special cases. The motivation is from advanced signal processing applications in engineering and science, particularly in situations where signal models are only partially known and are in noisy environments. The focus is on optimal processing, but with a counterpoint theme in suboptimal, adaptive processing to achieve a compromise between performance and computational effort.

The central theme of the book is the exploitation, in novel ways, of the so-called reference probability methods for optimal estimation and control. These methods supersede, for us at least, the more familiar innovation and martingale representation methods of earlier decades. They render the theory behind the very general and powerful estimation and control results accessible to the first-year graduate student. We claim that these reference probability methods are powerful and, perhaps, comprehensive in the context of discrete-time stochastic systems; furthermore, they turn out to be relevant for systems control. It is in the nature of mathematics that these methods were first developed for the technically more demanding area of continuous time stochastic systems, starting with the theorems of Cameron and Martin (1944), and Girsanov (1960). The reference probability approach to optimal filtering was introduced in continuous-time in Duncan (1967), Mortensen (1966) and Zakai (1969). This material tends to be viewed as inaccessible to graduate students in engineering. However, apart from contributions in Boel (1976), Brémaud and van Schuppen (1976), di Masi and Runggaldier (1982), Segall (1976b), Kumar and Varaiya (1986b) and Campillo and le Gland (1989), there has been little work on discrete-time filtering and control using the measure change approach.

An important feature of this book is the systematic introduction of new, equivalent probability measures. Under the new measure the variables of the observation process, and at times the state process, are independent, and the computations are greatly simplified, being no more difficult than processing for linear models. An inverse change of measure returns the variables to the "real world" where the state influences the observations. Our methods also apply in continuous time, giving simpler proofs of known theorems together with new results. However, we have chosen to concentrate on models whose state is a noisily observed Markov chain. We thus avoid much of the delicate mathematics associated with continuous-time diffusion processes.

The signal models discussed in this text are, for the main part, in discrete time and, in the first instance, with states and measurements in a discrete set. We proceed from discrete time to continuous time, from linear models to nonlinear ones, from completely known models to partially known models, from one-dimensional signal processing to two-dimensional processing, from white noise environments to colored noise environments, and from general formulations to specific applications.

Our emphasis is on recent results, but at times we cannot resist the temptation to provide "slicker" derivations of known theorems.

This work arose from a conversation two of the authors had at a conference twenty years ago. We talked about achieving adaptive filter stability and performance enhancement using martingale theory. We would have been incredulous then at what we have recently achieved and organized as this book. Optimal filtering and closed-loop control objectives have been attained for quite general nonlinear signal models in noisy environments. The optimal algorithms are simply stated. They are derived in a systematic manner with a minimal number of steps in the proofs.

Of course, twenty years ago we would have been absolutely amazed at the power of supercomputers and, indeed, desktop computers today, and so would not have dreamt that *optimal* processing could actually be implemented in applications except for the simplest examples. It is still true that our simply formulated optimal algorithms can be formidable to implement, but there are enough applications areas where it is possible to proceed effectively from the foundations laid here, in spite of the dreaded curse of dimensionality.

Our work starts with discrete-time signal models and with states and measurements belonging to a discrete set. We first apply the change-of-measure technique so that the observations under a probability measure are independent and uniformly distributed. We then achieve our optimization objectives, and, in a final step, translate these results back to the real world. Perhaps at first glance, the work looks too mathematical for the engineers of today, but all the results have engineering motivation, and our pedagogical style should allow an engineer to build the mathematical tools without first taking numerous mathematics courses in probability theory and stochastic systems. The advanced mathematics student may find later chapters immediately accessible and see earlier chapters as special cases. However, we believe many of the key insights are right there in the first technical chapter. For us, these first results were the key to most of what follows, but it must be admitted that only

by tackling the harder, more general problems did we develop proofs which we now use to derive the first results.

Actually, it was just two years ago that we got together to work on hidden Markov model (HMM) signal processing. One of us (JBM) had just developed exciting application studies for such models in biological signal processing. It turns out that ionic channel currents in neuron cell membranes can now be observed using Nobel prize winning apparatus measuring femto (10^{-15}) amps. The noise is white and Gaussian but dominates the signals. By assuming that the signals are finite-state Markov chains, and adaptively estimating transition probability and finite state values, much information can be obtained about neural synapses and the synaptic response to various new drug formulations. We believed that the on-line biological signal processing techniques which we developed could be applied to communication systems involving fading channels, such as mobile radio communications.

The key question for us, two years ago, was how could we do all this signal processing, with uncertain models in noisy environments, *optimally*? Then, if this task was too formidable for implementation, how could we achieve a reasonable compromise between computational effort and performance? We believed that the martingale approach would be rewarding, and it was, but it was serendipitous to find just how powerful were the reference probability methods for discrete-time stochastic systems. This book has emerged somewhat as a surprise.

In our earlier HMM studies, work with Ph.D. student Vikram Krishnamurthy and postdoctoral student Dr. Lige Xia set the pace for adaptive HMM signal processing. Next, work with Ph.D. student Hailiang Yang helped translate some continuous-time domain filtering insights to discrete time. The work of some of our next generation of Ph.D. students, including Iain Collings, features quite significantly in our final manuscript. Also, discussions with Matt James, Alain Bensoussan, and John Baras have been very beneficial in the development of the book. We wish to acknowledge to seminal thinking of Martin Clarke in the area of nonlinear filtering and his influence on our work. Special thanks go to René Boel for his review of the first version of the book and to N. Krylov for supplying corrections to the first printing.

The support of the Cooperative Research Centre for Robust and Adaptive Systems, the Boeing Commercial Airplane Company, and the NSERC Grant A7964 are gratefully acknowledged. We acknowledge the typing support of Shelley Hey, and Marita Rendina, and LaTeX programming support of James Ashton.

Chapter 2
Discrete States and Discrete Observations

2.1 Introduction

In this chapter, we deal with signals denoted by $\{X_k\}$, $k \in \mathbb{N}$ in *discrete time*. These signals are further restricted to a discrete set and are thus termed *discrete-state* signals. They transit between elements in this set with transition probabilities dependent only on the previous state, and so are *Markov chains*. The *transition probabilities* are independent of time, and so the Markov chains are said to be *homogeneous*. The Markov chain is not observed directly; rather there is a discrete-time, finite-state observation process $\{Y_k\}, k \in \mathbb{N}$, which is a noisy function of the chain. Consequently, the Markov chain is said to be *hidden* in the observations.

Our objective is to estimate the state of the chain, given the observations. Our preference is to achieve such estimation on-line in an optimal recursive manner, using what we term optimal estimators. The term *estimator* covers the special cases of *on-line filters*, where the estimates are calculated as the measurements are received, *on-line predictors* where there is a prediction at a fixed number of discrete time instants in the future, and *on-line smoothers* where there is improved estimation achieved by using a fixed number of future measurements as well as the previous ones. We also seek recursive filters and smoothers for the number of jumps from one state to another, for the occupation time of a state, and for a process related to the observations.

In the first instance, we assume that the equations describing the HMM are known. However, if this is not the case, it is possible to estimate the parameters also on-line and so achieve *adaptive* (or *self-tuning*) estimators. Unfortunately, it is usually not practical to achieve optimal adaptive estimators. In seeking practical suboptimal schemes, a *multipass scheme* is to update the parameters estimates only after processing a large data set, perhaps the entire data set. At the end of each pass through this data set, the parameter estimates are updated, to yield improved parameter estimates; see, for example, the so-called expectation maximization (EM) scheme; see Dempster, Laird and Rubin (1977). Our approach requires only a forward pass through the data to achieve parameter updates, in contrast to earlier so-called *forward-backward* algorithms of the Baum-Welch type (Baum and Petrie 1966).

Hidden Markov models have been found useful in many areas of probabilistic modeling, including speech processing; see Rabiner (1989). We believe our model is of wide applicability and generality. Many state and observation processes of the form (2.14) arise in the literature. In addition, certain time-series models can be approximated by HMMs.

As mentioned in the introduction, one of the fundamental techniques employed throughout this book is the discrete-time *change of measure*. This is a version of *Girsanov's Theorem* (see Theorem A.1.2). It is developed for the discrete-state HMM in Section 2.3 of this chapter.

A second basic observation is the *idempotent property* of the indicator functions for the state space of the Markov chain. With X one of the unit (column) vectors e_i, $1 \le i \le N$, prime denoting transpose, and using the inner product notation $\langle a, b\rangle = a'b$, this idempotent property allows us to write the square XX' as $\sum_{i=1}^{N} \langle X, e_i\rangle e_i e_i'$ and so obtain *closed (finite-dimensional), recursive filters* in Sections 2.4–2.9. More generally, any real function $f(X)$ can be expressed as a linear functional $f(X) = \langle f, X\rangle$ where $\langle f, e_i\rangle = f(e_i) = f_i$ and $f = (f_1, \ldots, f_N)$. Thus with $X^i = \langle X, e_i\rangle$,

$$f(X) = \sum_{i=1}^{N} f(e_i) X^i = \sum_{i=1}^{N} f_i X^i. \tag{1.1}$$

For the vector of indicator functions X, note that from the definition of expectations of a simple random variable, as in Appendix A,

$$E[\langle X, e_i\rangle] = \sum_{j=1}^{N} \langle e_j, e_i\rangle P(X = e_j) = P(X = e_i). \tag{1.2}$$

Section 2.10 of this chapter discusses similar estimation problems for a discrete-time, discrete-state hidden Markov model in the case where the noise terms in the Markov chain X and observation process Y are not independent. A test for independence is given. This section may be omitted on a first reading.

2.2 Model

All processes are defined initially on a *probability space* $(\Omega, \mathcal{F}, P)$. Below, a new probability measure $\overline{P}$ is defined. See Appendix A for related background in probability theory.

A system is considered whose state is described by a finite-state, homogeneous, discrete-time Markov chain X_k, $k \in \mathbb{N}$. We suppose X_0 is given, or its distribution known. If the state space of X_k has N elements it can be identified without loss of generality, with the set

$$S_X = \{e_1, \ldots, e_N\}, \tag{2.1}$$

where e_i are unit vectors in $\mathbb{R}^N$ with unity as the ith element and zeros elsewhere.

Write $\mathscr{F}_k^0 = \sigma\{X_0, \ldots, X_k\}$, for the σ-*field* generated by $X_0, \ldots, X_k$, and $\{\mathscr{F}_k\}$ for the *complete filtration* generated by the $\mathscr{F}_k^0$; this augments $\mathscr{F}_k^0$ by including all subsets of events of probability zero. Again, see Appendix A for related background in probability theory. The *Markov property* implies here that

$$P(X_{k+1} = e_j \mid \mathscr{F}_k) = P(X_{k+1} = e_j \mid X_k).$$

Write

$$a_{ji} = P(X_{k+1} = e_j \mid X_k = e_i), \qquad A = (a_{ji}) \in \mathbb{R}^{N \times N} \tag{2.2}$$

so that using the property (1.2), then

$$E[X_{k+1} \mid \mathscr{F}_k] = E[X_{k+1} \mid X_k] = AX_k. \tag{2.3}$$

Define

$$V_{k+1} := X_{k+1} - AX_k. \tag{2.4}$$

So that

$$X_{k+1} = AX_k + V_{k+1}. \tag{2.5}$$

This can be referred to as a *state equation*.

Now observe that taking the *conditional expectation* and noting that $E[AX_k \mid X_k] = AX_k$, we have

$$E[V_{k+1} \mid \mathscr{F}_k] = E[X_{k+1} - AX_k \mid X_k] = AX_k - AX_k = 0,$$

so $\{V_k\}$, $k \in \mathbb{N}$, is a sequence of martingale increments.

The state process X is not observed directly. We suppose there is a function $c(.,.)$ with finite range and we observe the values

$$Y_{k+1} = c(X_k, w_{k+1}), \qquad k \in \mathbb{N}. \tag{2.6}$$

The w_k in (2.6) are a sequence of independent, identically distributed (i.i.d.) random variables, with V_k, w_k being mutually independent.

$\{\mathscr{G}_k^0\}$ will be the σ-field on Ω generated by $X_0, X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$, and $\mathscr{G}_k$ its completion. Also $\{\mathscr{Y}_k^0\}$ will be the σ-field on Ω generated by $Y_1, \ldots, Y_k$ and $\mathscr{Y}_k$ its completion. Note $\mathscr{G}_k \subset \mathscr{G}_{k+1} \subset \cdots$ and $\mathscr{Y}_k \subset \mathscr{Y}_{k+1} \subset \cdots$. The increasing family of σ-fields is called a *filtration*. A function is $\mathscr{G}_k^0$-measurable if and only if it is a function of $X_0, X_1, \ldots, X_k, Y_1, \ldots, Y_k$. Similarly, for $\mathscr{Y}_k^0$, $\mathscr{Y}_k$. See also Appendix A.

The w_k in (2.6) are a sequence of independent, identically distributed (i.i.d.) random variables, with V_k, w_k being mutually independent. The pair of processes $(X_k, Y_k), k \in \mathbb{N}$, provides our first, basic example of a hidden Markov model, or HMM. This term is appropriate because the Markov chain is not observed directly but, instead, is hidden in the noisy observations Y. In this HMM the time parameter is discrete and the state spaces of both X and Y are finite (and discrete). Note that there is a unit delay between the state X at time k and its measurement Y at time $k+1$. A *zero delay observation model* is discussed later in this chapter.

Suppose the range of $c(.,.)$ consists of M points. Then we can identify the range of $c(.,.)$ with the set of unit vectors

$$S_Y = \{f_1, \ldots, f_M\}, \qquad f_j = (0, \ldots, 1, \ldots, 0)' \in \mathbb{R}^M, \tag{2.7}$$

where the unit element is the jth element.

We have assumed that $c(.,.)$ is independent of the time parameter k, but the results below are easily extended to the case of a nonhomogeneous chain X and a time-dependent $c(.,.)$.

Now (2.6) implies

$$P(Y_{k+1} = f_j \mid X_0, X_1, \ldots, X_k, Y_1, \ldots, Y_k) = P(Y_{k+1} = f_j \mid X_k).$$

Write

$$C = (c_{ji}) \in \mathbb{R}^{M \times N}, \qquad c_{ji} = P(Y_{k+1} = f_j \mid X_k = e_i) \tag{2.8}$$

so that $\sum_{j=1}^{M} c_{ji} = 1$ and $c_{ji} \geq 0$, $1 \leq j \leq M$, $1 \leq i \leq N$. We have, therefore,

$$E[Y_{k+1} \mid X_k] = CX_k. \tag{2.9}$$

If $W_{k+1} := Y_{k+1} - CX_k$, then taking the conditional expectation and noting $E[CX_k \mid X_k] = CX_k$ we have

$$\begin{aligned} E[W_{k+1} \mid \mathscr{G}_k] &= E[Y_{k+1} - CX_k \mid X_k] \\ &= CX_k - CX_k = 0, \end{aligned}$$

so W_k is a $(P, \mathscr{G}_k)$ martingale increment and

$$Y_{k+1} = CX_k + W_{k+1}. \tag{2.10}$$

Equation (2.10) can be thought of as an *observation equation*. The case where, given $\mathscr{G}_k$, the noise terms W_k in the observations Y_k are possibly correlated with the noise terms V_k in the Markov chain will be considered in Section 2.10.

Notation 2.1 *Write* $Y_k^i = \langle Y_k, f_i \rangle$ *so* $Y_k = \left(Y_k^1, \ldots, Y_k^M\right)'$, $k \in \mathbb{N}$. *For each* $k \in \mathbb{N}$, *exactly one component is equal to* 1, *the remainder being* 0.

Note $\sum_{i=1}^{M} Y_k^i = 1$. Write $c_{k+1}^i = E\left[Y_{k+1}^i \mid \mathscr{G}_k\right] = \sum_{j=1}^{N} c_{ij} \langle e_j, X_k \rangle$ and $c_{k+1} = (c_{k+1}^1, \ldots, c_{k+1}^M)'$. Then

$$c_{k+1} = E[Y_{k+1} \mid \mathscr{G}_k] = CX_k. \tag{2.11}$$

We shall suppose initially that $c_k^i > 0$, $1 \leq i \leq M$, $k \in \mathbb{N}$. (See, however, the construction of P from $\overline{P}$ in Section 2.3). Note $\sum_{i=1}^{M} c_k^i = 1$, $k \in \mathbb{N}$. We shall need the following result in the sequel.

Lemma 2.2 *With* $\operatorname{diag}(z)$ *denoting the diagonal matrix with vector* z *on its diagonal, we have*

$$V_{k+1}V_{k+1}' = \operatorname{diag}(AX_k) + \operatorname{diag}(V_{k+1}) - A\operatorname{diag}X_kA' - AX_kV_{k+1}' - V_{k+1}(AX_k)' \tag{2.12}$$

and

$$\begin{aligned}\langle V_{k+1}\rangle &:= E\left[V_{k+1}V_{k+1}' \mid \mathscr{F}_k\right] \\ &= E\left[V_{k+1}V_{k+1}' \mid X_k\right] \\ &= \operatorname{diag}(AX_k) - A\operatorname{diag}X_kA'.\end{aligned} \tag{2.13}$$

Proof From (2.4)

$$X_{k+1}X_{k+1}' = AX_k(AX_k)' + AX_kV_{k+1}' + V_{k+1}(AX_k)' + V_{k+1}V_{k+1}'.$$

However, $X_{k+1}X_{k+1}' = \operatorname{diag}(X_{k+1}) = \operatorname{diag}(AX_k) + \operatorname{diag}(V_{k+1})$. Equation (2.12) follows. The terms on the right side of (2.12) involving V_{k+1} are martingale increments; conditioning on X_k we see

$$\langle V_{k+1}\rangle = E\left[V_{k+1}V_{k+1}' \mid X_k\right] = \operatorname{diag}(AX_k) - A\operatorname{diag}X_kA'.$$

□

Similarly, we can show that

$$\langle W_{k+1}\rangle := E\left[W_{k+1}W_{k+1}' \mid \mathscr{G}_k\right] = \operatorname{diag}(CX_k) - C\operatorname{diag}X_kC'.$$

In summary then, we have the following state space signal model for a Markov chain hidden in noise with discrete measurements.

Discrete HMM *The discrete HMM under P has the* state space equations

$$\boxed{\begin{aligned}X_{k+1} &= AX_k + V_{k+1}, \\ Y_{k+1} &= CX_k + W_{k+1}, \qquad k \in \mathbb{N},\end{aligned}} \tag{2.14}$$

where $X_k \in S_X$, $Y_k \in S_Y$, A and C are matrices of transition probabilities given in (2.2) *and* (2.8). *The entries satisfy*

$$\sum_{j=1}^{N} a_{ji} = 1, \qquad a_{ji} \geq 0, \tag{2.15}$$

$$\sum_{j=1}^{M} c_{ji} = 1, \qquad c_{ji} \geq 0. \tag{2.16}$$

V_k and W_k are martingale increments satisfying

$$\begin{aligned}E\left[V_{k+1} \mid \mathscr{F}_k\right] = 0, \qquad E\left[W_{k+1} \mid \mathscr{G}_k\right] = 0, \\ \langle V_{k+1}\rangle := E\left[V_{k+1}V_{k+1}' \mid X_k\right] = \operatorname{diag}(AX_k) - A\operatorname{diag}X_kA', \\ \langle W_{k+1}\rangle := E\left[W_{k+1}W_{k+1}' \mid X_k\right] = \operatorname{diag}(CX_k) - C\operatorname{diag}X_kC'.\end{aligned}$$

2.3 Change of Measure

The idea of introducing new probability measures, as outlined in the previous chapter, is now discussed for the observation process Y. This measure change concept is the key to many of the results in this and the following chapters.

We assume, for this measure change, $c_\ell^i > 0$, $1 \le i \le M$, $\ell \in \mathbb{N}$. This assumption, in effect, is that given any $\mathcal{G}_k$, the observation noise is such that there is a nonzero probability that $Y_{k+1}^i > 0$ for all i. This assumption is later relaxed to achieve the main results of this section. Define

$$\lambda_\ell = \sum_{i=1}^{M} \left(\frac{M^{-1}}{c_\ell^i} \right) \langle Y_\ell, f_i \rangle, \tag{3.1}$$

and

$$\Lambda_k = \prod_{\ell=1}^{k} \lambda_\ell. \tag{3.2}$$

Note that $Y_\ell^i = 1$ for only one i at each ℓ, and $Y_\ell^i = 0$ otherwise, so that λ_ℓ is merely the product of unity terms and one nonunity term. Consequently, since λ_k is a nonlinear function of Y_k, then property (1.1) tells us that $\lambda_k = \lambda_k(Y_k) = \sum_{i=1}^{M} Y_k^i / M c_k^i$.

Lemma 3.1 *With the above definitions*

$$E[\lambda_{k+1} \mid \mathcal{G}_k] = 1. \tag{3.3}$$

Proof Applying the properties (1.1) and (1.2),

$$\begin{aligned} E[\lambda_{k+1} \mid \mathcal{G}_k] &= E\left[\sum_{i=1}^{M} \frac{1}{M c_{k+1}^i} Y_{k+1}^i \,\middle|\, \mathcal{G}_k \right] \\ &= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{c_{k+1}^i} P\left(Y_{k+1}^i = 1 \mid \mathcal{G}_k \right) \\ &= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{c_{k+1}^i} \cdot c_{k+1}^i = 1. \end{aligned}$$

Here as in many places, we interchange expectations and summations, for a simple random variable. This is permitted, of course, by a special case of Fubini's Theorem; see Loève (1978) and Appendix A. □

We now define a new probability measure $\overline{P}$ on $(\Omega, \bigvee_{\ell=1}^{\infty} \mathcal{G}_\ell)$ by putting the restriction of the Radon-Nikodym derivative $d\overline{P}/dP$ to the σ-field $\mathcal{G}_k$ equal to Λ_k. Thus

$$\left. \frac{d\overline{P}}{dP} \right|_{\mathcal{G}_k} = \Lambda_k. \tag{3.4}$$

[The existence of $\overline{P}$ follows from *Kolmogorov's Extension Theorem* (Kolmogorov 1933)]; see also Appendix A. This means that, for any set $B \in \mathcal{G}_k$,

$$\overline{P}(B) = \int_B \Lambda_k \, dP.$$

Equivalently, for any $\mathcal{G}_k$-*measurable* random variable ϕ

$$\overline{E}[\phi] = \int \phi \, d\overline{P} = \int \phi \frac{d\overline{P}}{dP} dP = \int \phi \Lambda_k \, dP = E[\Lambda_k \phi], \tag{3.5}$$

where $\overline{E}$ and E denote expectations under $\overline{P}$ and P, respectively. In the discrete-state case under consideration, $d\overline{P}/dP$ reduces to the ratio $\overline{P}/P$ and the integrations reduce to sums. This equation exhibits the basic idea of the change of measure; for most of the results in this book a big challenge is to determine the appropriate forms for λ and Λ. It is not straightforward to give insight into this process other than to illustrate by examples and present hindsight proofs. Perhaps the measure changes of Chapter 3 are the most transparent, and more discussion is given for these.

We now give a conditional form of *Bayes' Theorem* which is fundamental for the results that follow. The result relates conditional expectations under two different measures. Recall that ϕ is *integrable* if $E|\phi| < \infty$. First we shall consider a simple case.

Consider the experiment of throwing a die. The set of outcomes is $\Omega = \{1, 2, \ldots, 6\}$. Suppose the die is not necessarily balanced, so that the probability of i showing is $P(i) = p_i$, $p_1 + \cdots + p_6 = 1$.

The σ-field $\mathcal{F}$ associated with this experiment is the collection of all subsets of Ω, including the empty set ϕ. The sets in $\mathcal{F}$ are the *events*. (See also Appendix A.) The probability of the event "odd number," for instance, is $P\{1,3,5\} = p_1 + p_3 + p_5$. Consider the sub-$\sigma$-field $\mathcal{G}$ of $\mathcal{F}$ defined by $\mathcal{G} = \{\Omega, \phi, \{1,3,5\}, \{2,4,6\}\}$.

Now suppose ϕ is a real random variable on $(\Omega, \mathcal{F})$, that is, $\phi(i) \in \mathbb{R}$ for $i = 1, 2, \ldots, 6$. The mean, or expected, value of ϕ is then $E[\phi] = \sum_{i=1}^{6} \phi(i) p_i$.

The conditional expected value of ϕ, given $\mathcal{G}$, $E[\phi \mid \mathcal{G}]$, is then a function which is constant on the smallest, nonempty sets of $\mathcal{G}$. That is,

$$E[\phi \mid \mathcal{G}](i) = \frac{\phi(1) p_1 + \phi(3) p_3 + \phi(5) p_5}{p_1 + p_3 + p_5}, \qquad \text{if } i \in \{1,3,5\},$$

$$E[\phi \mid \mathcal{G}](i) = \frac{\phi(2) p_2 + \phi(4) p_4 + \phi(6) p_6}{p_2 + p_4 + p_6}, \qquad \text{if } i \in \{2,4,6\}$$

We note that $\psi = E[\phi \mid \mathcal{G}]$ can be considered a function on $(\Omega, \mathcal{F})$ and that then $E[E[\phi \mid \mathcal{G}]] = E[\phi]$.

Suppose we now rebalance the die by introducing weights $\Lambda(i)$ on the different faces. Note that Λ is itself, therefore, a random variable on $(\Omega, \mathcal{F})$.

Write $\overline{p}_i = \Lambda(i) p_i = \overline{P}(i)$, $i = 1, \ldots, 6$, for the new balance proportion assigned to the ith face. Then, because $\overline{P}$ is to be a probability measure, $E[\Lambda] = \overline{p}_1 + \cdots + \overline{p}_6 = \Lambda(1) p_1 + \cdots + \Lambda(6) p_6 = 1$.

We have the following expressions:

$$E[\Lambda\phi \mid \mathscr{G}](i) = \frac{\phi(1)\Lambda(1)p_1 + \phi(3)\Lambda(3)p_3 + \phi(5)\Lambda(5)p_5}{p_1+p_3+p_5}, \qquad \text{if } i \in \{1,3,5\},$$
$$E[\Lambda\phi \mid \mathscr{G}](i) = \frac{\phi(2)\Lambda(2)p_2 + \phi(4)\Lambda(4)p_4 + \phi(6)\Lambda(6)p_6}{p_2+p_4+p_6}, \qquad \text{if } i \in \{2,4,6\}.$$

Similarly,

$$E[\Lambda \mid \mathscr{G}](i) = \frac{\Lambda(1)p_1 + \Lambda(3)p_3 + \Lambda(5)p_5}{p_1+p_3+p_5}, \qquad \text{if } i \in \{1,3,5\},$$
$$E[\Lambda \mid \mathscr{G}](i) = \frac{\Lambda(2)p_2 + \Lambda(4)p_4 + \Lambda(6)p_6}{p_2+p_4+p_6}, \qquad \text{if } i \in \{2,4,6\}.$$

However, with $\overline{E}$ denoting expectation under the new probability $\overline{P}$:

$$\overline{E}[\phi \mid \mathscr{G}](i) = \frac{\phi(1)\overline{p}_1 + \phi(3)\overline{p}_3 + \phi(5)\overline{p}_5}{\overline{p}_1+\overline{p}_3+\overline{p}_5}, \qquad \text{if } i \in \{1,3,5\},$$
$$\overline{E}[\phi \mid \mathscr{G}](i) = \frac{\phi(2)\overline{p}_2 + \phi(4)\overline{p}_4 + \phi(6)\overline{p}_6}{\overline{p}_2+\overline{p}_4+\overline{p}_6}, \qquad \text{if } i \in \{2,4,6\}.$$

Consequently, $\overline{E}[\phi \mid \mathscr{G}] = E[\Lambda\phi \mid \mathscr{G}]/E[\Lambda \mid \mathscr{G}]$.

We now prove this result in full generality. For background on conditional expectation see Elliott (1982b).

Theorem 3.2 (Conditional Bayes Theorem) *Suppose $(\Omega, \mathscr{F}, P)$ is a probability space and $\mathscr{G} \subset \mathscr{F}$ is a sub-σ-field. Suppose $\overline{P}$ is another probability measure absolutely continuous with respect to P and with Radon-Nikodym derivative $d\overline{P}/dP = \Lambda$. Then if ϕ is any $\overline{P}$ integrable random variable*

$$\overline{E}[\phi \mid \mathscr{G}] = \psi \quad \textit{where} \quad \psi = \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]} \quad \textit{if } E[\Lambda \mid \mathscr{G}] > 0$$
$$\textit{and} \quad \psi = 0 \quad \textit{otherwise.}$$

Proof Suppose B is any set in $\mathscr{G}$. We must show

$$\int_B \overline{E}[\phi \mid \mathscr{G}]\, d\overline{P} = \int_B \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]} d\overline{P}.$$

Define $\psi = E[\Lambda\phi \mid \mathscr{G}]/E[\Lambda \mid \mathscr{G}]$ if $E[\Lambda \mid \mathscr{G}] > 0$ and $\psi = 0$ otherwise. Then $\overline{E}[\phi \mid \mathscr{G}] = \psi$.

Suppose A is any set in $\mathscr{G}$. We must show $\int_A \overline{E}[\phi \mid \mathscr{G}]\, d\overline{P} = \int_A \psi\, d\overline{P}$. Write $G = \{\omega : E[\Lambda \mid \mathscr{G}] = 0\}$, so $G \in \mathscr{G}$. Then $\int_G E[\Lambda \mid \mathscr{G}]\, dP = 0 = \int_G \Lambda\, dP$ and $\Lambda \geq 0$ a.s.

So either $P(G) = 0$, or the restriction of Λ to G is 0 a.s. In either case, $\Lambda = 0$ a.s. on G.

Now $G^c = \{\omega : E[\Lambda \mid \mathscr{G}] > 0\}$. Suppose $A \in \mathscr{G}$; then $A = B \cup C$ where $B = A \cap G^c$ and $C = A \cap G$. Further,

$$\begin{aligned}\int_A \overline{E}[\phi \mid \mathscr{G}]\, d\overline{P} &= \int_A \phi\, d\overline{P} = \int_A \phi\Lambda\, dP \\ &= \int_B \phi\Lambda\, dP + \int_C \phi\Lambda\, dP. \end{aligned} \tag{3.6}$$

Of course, $\Lambda = 0$ a.s. on $C \subset G$, so

$$\int_C \phi\Lambda\, dP = 0 = \int_C \psi\, d\overline{P}, \tag{3.7}$$

by definition.

Now

$$\begin{aligned}\int_B \psi d\overline{P} &= \int_B \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]} d\overline{P} \\ &= \overline{E}\left[I_B \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]}\right] \\ &= E\left[I_B \Lambda \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]}\right] \\ &= E\left[E\left[I_B \Lambda \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]} \mid \mathscr{G}\right]\right] \\ &= E\left[I_B E[\Lambda \mid \mathscr{G}] \frac{E[\Lambda\phi \mid \mathscr{G}]}{E[\Lambda \mid \mathscr{G}]}\right] \\ &= E[I_B E[\Lambda\phi \mid \mathscr{G}]] \\ &= E[I_B \Lambda\phi]. \end{aligned}$$

That is

$$\int_B \Lambda\phi\, dP = \int_B \psi\, d\overline{P}. \tag{3.8}$$

From (3.6), adding (3.7) and (3.8) we see that

$$\begin{aligned}&\int_C \Lambda\phi\, dP + \int_B \Lambda\phi\, dP = \int_A \Lambda\phi\, dP \\ &= \int_A \overline{E}[\phi \mid \mathscr{G}]\, d\overline{P} = \int_A \psi\, d\overline{P}, \end{aligned}$$

and the result follows. □

A sequence $\{\phi_k\}$ is said to be $\mathscr{G}$-*adapted* if ϕ_k is $\mathscr{G}_k$-measurable for every k.

Applying Theorem 3.2 result to the P and $\overline{P}$ of (3.4) we have the following:

Lemma 3.3 *If $\{\phi_k\}$ is a $\mathscr{G}$-adapted integrable sequence of random variables, then*

$$\overline{E}\left[\phi_k \mid \mathscr{Y}_k\right] = \frac{E\left[\Lambda_k \phi_k \mid \mathscr{Y}_k\right]}{E\left[\Lambda_k \mid \mathscr{Y}_k\right]}.$$

Lemma 3.4 *Under $\overline{P}$, $\{Y_k\}$, $k \in \mathbb{N}$, is a sequence of i.i.d. random variables each having the uniform distribution that assigns probability $\frac{1}{M}$ to each point f_i, $1 \le i \le M$, in its range space.*

Proof With $\overline{E}$ denoting expectation under $\overline{P}$, using Lemma 3.1, Theorem 3.2 and properties (1.1) and (1.2), then

$$\begin{aligned}
\overline{P}\left(Y_{k+1}^j = 1 \mid \mathscr{G}_k\right) &= \overline{E}\left[\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right] \\
&= \frac{E\left[\Lambda_{k+1}\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right]}{E\left[\Lambda_{k+1} \mid \mathscr{G}_k\right]} \\
&= \frac{\Lambda_k E\left[\lambda_{k+1}\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right]}{\Lambda_k E\left[\lambda_{k+1} \mid \mathscr{G}_k\right]} \\
&= E\left[\lambda_{k+1}\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right] \\
&= E\left[\prod_{i=1}^{M}\left(\frac{1}{Mc_{k+1}^i}\right)^{Y_{k+1}^i}\left\langle Y_{k+1}, f_j\right\rangle \,\middle|\, \mathscr{G}_k\right] \\
&= E\left[\sum_{i=1}^{M}\left(\frac{1}{Mc_{k+1}^i}\right) Y_{k+1}^i Y_{k+1}^j \,\middle|\, \mathscr{G}_k\right] \\
&= \frac{1}{Mc_{k+1}^j} E\left[Y_{k+1}^j \mid \mathscr{G}_k\right] \\
&= \frac{1}{Mc_{k+1}^j} c_{k+1}^j = \frac{1}{M} = \overline{P}\left(Y_{k+1}^j = 1\right),
\end{aligned}$$

a quantity independent of $\mathscr{G}_k$ which finishes the proof. □

Now note that $\overline{E}\left[X_{k+1} \mid \mathscr{G}_k\right] = E\left[\Lambda_{k+1}X_{k+1} \mid \mathscr{G}_k\right] / E\left[\Lambda_{k+1} \mid \mathscr{G}_k\right] = E\left[\lambda_{k+1}X_{k+1} \mid \mathscr{G}_k\right] = AX_k$ so that under $\overline{P}$, X remains a Markov chain with transition matrix A.

A Reverse Measure Change

What we wish to do now is start with a probability measure $\overline{P}$ on $(\Omega, \bigvee_{n=1}^{\infty} \mathscr{G}_n)$ such that

1. the process X is a finite-state Markov chain with transition matrix A and
2. $\{Y_k\}$, $k \in \mathbb{N}$, is a sequence of i.i.d. random variables and

$$\overline{P}\left(Y_{k+1}^j = 1 \mid \mathscr{G}_k\right) = \overline{P}\left(Y_{k+1}^j = 1\right) = \frac{1}{M}.$$

Suppose $C = (c_{ji})$, $1 \le j \le M$, $1 \le i \le N$ is a matrix such that $c_{ji} \ge 0$ and $\sum_{j=1}^{M} c_{ji} = 1$.

We shall now construct a new measure P on $(\Omega, \bigvee_{n=1}^{\infty} \mathcal{G}_n)$ such that under P, (2.14) still holds and $E[Y_{k+1} \mid \mathcal{G}_k] = CX_k$. We again write

$$c_{k+1} = CX_k$$

and $c_{k+1}^i = \langle c_{k+1}, f_i \rangle = \langle CX_k, f_i \rangle$, so that

$$\sum_{i=1}^{M} c_{k+1}^i = 1. \tag{3.9}$$

The construction of P from $\overline{P}$ is inverse to that of $\overline{P}$ from P. Write

$$\overline{\lambda}_\ell = \prod_{i=1}^{M} \left(M c_\ell^i\right)^{Y_\ell^i}, \qquad \ell \in \mathbb{N}, \tag{3.10}$$

and

$$\overline{\Lambda}_k = \prod_{\ell=1}^{k} \overline{\lambda}_\ell. \tag{3.11}$$

Lemma 3.5 *With the above definitions*

$$\overline{E}\left[\overline{\lambda}_{k+1} \mid \mathcal{G}_k\right] = 1. \tag{3.12}$$

Proof Following the proof of Lemma 3.5

$$\begin{aligned}
\overline{E}\left[\overline{\lambda}_{k+1} \mid \mathcal{G}_k\right] &= \overline{E}\left[\prod_{i=1}^{M} \left(M c_{k+1}^i\right)^{Y_{k+1}^i} \mid \mathcal{G}_k\right] \\
&= M \sum_{i=1}^{M} c_{k+1}^i \overline{P}\left(Y_{k+1}^i = 1 \,\middle|\, \mathcal{G}_k\right) \\
&= M \sum_{i=1}^{M} \frac{c_{k+1}^i}{M} = \sum_{i=1}^{M} c_{k+1}^i = 1,
\end{aligned}$$

□

This time set

$$\left.\frac{dP}{d\overline{P}}\right|_{\mathcal{G}_k} = \overline{\Lambda}_k. \tag{3.13}$$

[The existence of P follows from Kolmogorov's Extension Theorem (Kolmogorov 1933); see also Appendix A.]

Lemma 3.6 *Under P,*

$$E[Y_{k+1} \mid \mathcal{G}_k] = CX_k.$$

Proof Using Theorem 3.2 and the now familiar properties (1.1) and (1.2), then

$$\begin{aligned} P\left(Y_{k+1}^j = 1 \mid \mathscr{G}_k\right) &= E\left[\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right] \\ &= \frac{\overline{E}\left[\overline{\Lambda}_{k+1}\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right]}{\overline{E}\left[\overline{\Lambda}_{k+1} \mid \mathscr{G}_k\right]} \qquad (\text{case } \overline{\Lambda} \neq 0) \\ &= \frac{\overline{E}\left[\overline{\lambda}_{k+1}\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right]}{\overline{E}\left[\overline{\lambda}_{k+1} \mid \mathscr{G}_k\right]} \\ &= \overline{E}\left[\prod_{i=1}^{M}\left(Mc_{k+1}^i\right)^{Y_{k+1}^i}\left\langle Y_{k+1}, f_j\right\rangle \,\middle|\, \mathscr{G}_k\right] \\ &= M\overline{E}\left[c_{k+1}^j\left\langle Y_{k+1}, f_j\right\rangle \mid \mathscr{G}_k\right] = c_{k+1}^j. \end{aligned}$$

In case $\overline{\Lambda}_{k+1} = 0$ we take $\frac{0}{0} = 1$, and the result follows. □

2.4 Unnormalized Estimates and Bayes' Formula

Recall our discrete HMM of Section 2.2; recall also that $\mathscr{Y}_k$ is the complete σ-field generated by knowledge of $Y_1, \ldots, Y_k$ and $\mathscr{G}_k$ is the complete σ-field generated by knowledge of $X_0, X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$. We suppose there is a probability $\overline{P}$ on $(\Omega, \bigvee_{n=1}^{\infty} \mathscr{G}_n)$ such that, under $\overline{P}$, $X_{k+1} = AX_k + V_{k+1}$, where V_k is a $\left(\overline{P}, \mathscr{G}_k\right)$ martingale increment. That is, $\overline{E}\left[V_{k+1} \mid \mathscr{G}_k\right] = 0$ and the $\{Y_k\}$ are i.i.d. with $\overline{P}\left(Y_k^j = 1\right) = \frac{1}{M}$, and the Y_k are conditionally independent of V_k, given $\mathscr{G}_k$, under both P and $\overline{P}$. We also have via the double expectation property listed in Appendix A,

$$\begin{aligned} \overline{E}\left[V_{k+1} \mid \mathscr{Y}_{k+1}\right] &= \overline{E}\left[\overline{E}\left[V_{k+1} \mid \mathscr{G}_k, \mathscr{Y}_{k+1}\right] \mid \mathscr{Y}_{k+1}\right] \\ &= \overline{E}\left[\overline{E}\left[V_{k+1} \mid \mathscr{G}_k\right] \mid \mathscr{Y}_{k+1}\right] = 0. \end{aligned} \tag{4.1}$$

The measure P is then defined using (3.13). Recall from Lemma 3.3 that for a $\mathscr{G}$-adapted sequence $\{\phi_k\}$,

$$E\left[\phi_k \mid \mathscr{Y}_k\right] = \frac{\overline{E}\left[\overline{\Lambda}_k \phi_k \mid \mathscr{Y}_k\right]}{\overline{E}\left[\overline{\Lambda}_k \mid \mathscr{Y}_k\right]}. \tag{4.2}$$

Remark 4.1 This identity indicates why the unnormalized conditional expectation $\overline{E}\left[\overline{\Lambda}_k \phi_k \mid \mathscr{Y}_k\right]$ is investigated.

■

Write $q_k(e_r)$, $1 \leq r \leq N$, $k \in \mathbb{N}$, for the unnormalized, conditional probability distribution such that

$$\overline{E}\left[\overline{\Lambda}_k \left\langle X_k, e_r\right\rangle \mid \mathscr{Y}_k\right] = q_k(e_r).$$

Note that an alternative standard notation for this unnormalized conditional distribution is α; this is used in later chapters for a related distribution.

Now $\sum_{i=1}^{N} \langle X_k, e_i \rangle = 1$, so

$$\sum_{i=1}^{N} q_k(e_i) = \overline{E}\left[\overline{\Lambda}_k \sum_{i=1}^{N} \langle X_k, e_i \rangle \mid \mathscr{Y}_k\right] = \overline{E}\left[\overline{\Lambda}_k \mid \mathscr{Y}_k\right].$$

Therefore, from (4.2) the normalized conditional probability distribution

$$p_k(e_r) = E\left[\langle X_k, e_r \rangle \mid \mathscr{Y}_k\right]$$

is given by

$$p_k(e_r) = \frac{q_k(e_r)}{\sum_{j=1}^{k} q_k(e_j)}.$$

To conclude this section with a basic example we obtain a recursive expression for q_k. Recursive estimates for more general processes will be obtained in Section 2.5.

Notation 4.2 *To simplify the notation we write* $c_j(Y_k) = M \prod_{i=1}^{M} c_{ij}^{Y_k^i}$.

Theorem 4.3 *For* $k \in \mathbb{N}$ *and* $1 \leq r \leq N$, *the recursive filter for the unnormalized estimates of the states is given by*

$$\boxed{q_{k+1} = A \operatorname{diag} c(Y_{k+1}) \cdot q_k.} \tag{4.3}$$

Proof Using the independence assumptions under $\overline{P}$ and the fact that $\sum_{j=1}^{N} \langle X_k, e_j \rangle = 1$, as well as properties (1.1) and (1.2), we have

$$\begin{aligned} q_k(e_r) &= \overline{E}\left[\langle X_{k+1}, e_r \rangle \overline{\Lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] \\ &= \overline{E}\left[\langle AX_k + V_{k+1}, e_r \rangle \overline{\Lambda}_k \prod_{i=1}^{M} \left(M c_{k+1}^i\right)^{Y_{k+1}^i} \,\middle|\, \mathscr{Y}_{k+1}\right] \\ &= M\overline{E}\left[\langle AX_k, e_r \rangle \overline{\Lambda}_k \prod_{i=1}^{M} \left(\langle CX_k, f_i \rangle\right)^{Y_{k+1}^i} \,\middle|\, \mathscr{Y}_{k+1}\right] \end{aligned}$$

[because V_{k+1} is a martingale increment with (4.1) holding]

$$\begin{aligned} &= M \sum_{j=1}^{N} \overline{E}\left[\langle X_k, e_j \rangle a_{rj} \overline{\Lambda}_k \mid \mathscr{Y}_{k+1}\right] \prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i} \\ &= M \sum_{j=1}^{N} \overline{E}\left[\langle X_k, e_j \rangle a_{rj} \overline{\Lambda}_k \mid \mathscr{Y}_k\right] \prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i} \end{aligned}$$

(because y_k is i.i.d. under $\overline{P}$)

$$= M \sum_{j=1}^{N} q_k(e_j) a_{rj} \prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i}.$$

Using Notation 4.2 the result follows. □

Remark 4.4 This unnormalized recursion is a discrete-time form of Zakai's Theorem (Zakai 1969). This recursion is linear.

■

2.5 A General Unnormalized Recursive Filter

We continue to work under measure $\overline{P}$ so that

$$X_{k+1} = AX_k + V_{k+1} \tag{5.1}$$

and the Y_k are independent random variables, uniformly distributed over $f_1, \ldots, f_M$.

Notation 5.1 *If $\{H_k\}$, $k \in \mathbb{N}$, is any integrable sequence of random variables we shall write*

$$\gamma_k(H_k) = \overline{E}\left[\overline{\Lambda}_k H_k \mid \mathscr{Y}_k\right]. \tag{5.2}$$

Note this makes sense for vector processes H.

Using Lemma 3.3 we see that

$$E\left[H_k \mid \mathscr{Y}_k\right] = \frac{\overline{E}\left[\overline{\Lambda}_k H_k \mid \mathscr{Y}_k\right]}{\overline{E}\left[\overline{\Lambda}_k \mid \mathscr{Y}_k\right]} = \frac{\gamma_k(H_k)}{\gamma_k(1)}. \tag{5.3}$$

Consequently $\gamma_k(H_k)$ is an unnormalized conditional expectation of H_k given $\mathscr{Y}_k$. We shall take $\gamma_0(X_0) = E[X_0]$; this provides the initial value for later recursions.

Now suppose $\{H_k\}$, $k \in \mathbb{N}$, is an integrable (scalar) sequence. With $\Delta H_{k+1} = H_{k+1} - H_k$, $H_{k+1} = H_k + \Delta H_{k+1}$, then

$$\gamma_{k+1}(H_{k+1}) = \overline{E}\left[\overline{\Lambda}_{k+1} H_k \mid \mathscr{Y}_{k+1}\right] + \overline{E}\left[\overline{\Lambda}_{k+1} \Delta H_{k+1} \mid \mathscr{Y}_{k+1}\right].$$

Consider the first term on the right. Then, using the now familiar properties (1.1) and (1.2),

$$\begin{aligned}
\overline{E}\left[\overline{\Lambda}_{k+1} H_k \mid \mathscr{Y}_{k+1}\right] &= \overline{E}\left[\overline{\Lambda}_k H_k \overline{\lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&= \overline{E}\left[\overline{\Lambda}_k H_k M \prod_{i=1}^{M} \langle CX_k, f_i\rangle^{Y_{k+1}^i} \;\middle|\; \mathscr{Y}_{k+1}\right] \\
&= \sum_{j=1}^{N} \overline{E}\left[\overline{\Lambda}_k H_k \langle X_k, e_j\rangle \mid \mathscr{Y}_k\right] M \prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i} \\
&= \sum_{j=1}^{N} c_j(Y_{k+1}) \langle \gamma_k(H_k X_k), e_j\rangle.
\end{aligned}$$

In this way the estimate for $\gamma_{k+1}(H_{k+1})$ introduces $\gamma_k(H_kX_k)$. A technical trick is to investigate the recursion for $\gamma_{k+1}(H_{k+1}X_{k+1})$. A similar discussion to that above then introduces the term $\gamma_k\left(H_kX_kX_k'\right)$; this can be written $\sum_{i=1}^N \langle \gamma_k(H_kX_k), e_i\rangle e_ie_i'$. Therefore, the estimates for $\gamma_{k+1}(H_{k+1}X_{k+1})$ can be recursively expressed in terms of $\gamma_k(H_kX_k)$ (together with other terms). Writing $\underline{1}$ for the vector $(1,1,\ldots,1)' \in \mathbb{R}^N$ we see $\langle X_k, \underline{1}\rangle = \sum_{i=1}^N \langle X_k, e_i\rangle = 1$, so

$$\langle \gamma_k(H_kX_k), \underline{1}\rangle = \gamma_k(H_k\langle X_k, \underline{1}\rangle) = \gamma_k(H_k). \tag{5.4}$$

Consequently, the unnormalized estimate $\gamma_k(H_k)$ is obtained by summing the components of $\gamma_k(H_kX_k)$. Furthermore, taking $H_k = 1$ in (5.4) we see

$$\gamma_k(1) = \langle \gamma_k(X_k), \underline{1}\rangle = \overline{E}\left[\overline{\Lambda}_k \mid \mathscr{Y}_k\right] = \sum_{i=1}^N q_k(e_i)$$

using the notation of Section 2.4. Therefore, the normalizing factor $\gamma_k(1)$ in (5.3) is obtained by summing the components of $\gamma_k(X_k)$.

We now make the above observations precise by considering a more specific, though general, process H.

Suppose, for $k \geq 1$, H_k is a scalar process of the form

$$\begin{aligned} H_{k+1} &= \sum_{\ell=1}^{k+1} (\alpha_\ell + \langle \beta_\ell, V_\ell\rangle + \langle \delta_\ell, Y_\ell\rangle) \\ &= H_k + \alpha_{k+1} + \langle \beta_{k+1}, V_{k+1}\rangle + \langle \delta_{k+1}, Y_{k+1}\rangle. \end{aligned} \tag{5.5}$$

Here $V_\ell = X_\ell - AX_{\ell-1}$ and α_ℓ, β_ℓ, δ_ℓ are $\mathscr{G}$-predictable processes of appropriate dimensions, that is, α_ℓ, β_ℓ, δ_ℓ are $\mathscr{G}_{\ell-1}$ measurable, α_ℓ is scalar, β_ℓ is N-dimensional, and δ_ℓ is M-dimensional.

Notation 5.2 *For any process ϕ_k, $k \in \mathbb{N}$, write*

$$\gamma_{m,k}(\phi_m) = \overline{E}\left[\overline{\Lambda}_k\phi_mX_k \mid \mathscr{Y}_k\right]. \tag{5.6}$$

Theorem 5.3 *For $1 \leq j \leq M$ write $c_j = Ce_j = \left(c_{1j},\ldots,c_{Mj}\right)'$ for the jth column of $C = (c_{ij})$ and $a_j = Ae_j = \left(a_{1j},\ldots,a_{Nj}\right)'$ for the jth column of $A = (a_{ij})$. Then*

$$\begin{aligned} &\gamma_{k+1,k+1}(H_{k+1}) \\ &\quad = \sum_{j=1}^N c_j(Y_{k+1})\left\{\left\langle \gamma_{k,k}(H_k) + \gamma_{k+1,k}(\alpha_{k+1} + \langle \delta_{k+1}, Y_{k+1}\rangle), e_j\right\rangle a_j\right. \\ &\qquad \left. + \left[\operatorname{diag}(a_j) - a_ja_j'\right]\overline{E}\left[\left\langle \overline{\Lambda}_kX_k, e_j\right\rangle \beta_{k+1} \mid \mathscr{Y}_{k+1}\right]\right\}. \end{aligned} \tag{5.7}$$

Proof

$$\begin{aligned} &\gamma_{k+1,k+1}(H_{k+1}) \\ &\quad = \overline{E}\left[X_{k+1}H_{k+1}\overline{\Lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] \end{aligned}$$

$$
\begin{aligned}
&= \overline{E}\left[(AX_k + V_{k+1}) (H_k + \alpha_{k+1} + \langle \beta_{k+1}, V_{k+1} \rangle + \langle \delta_{k+1}, Y_{k+1} \rangle) \right. \\
&\qquad \left. \times \overline{\Lambda}_k \overline{\lambda}_{k+1} \mid \mathscr{Y}_{k+1} \right] \\
&= \overline{E}\left[((H_k + \alpha_{k+1} + \langle \delta_{k+1}, Y_{k+1} \rangle) AX_k + \langle V_{k+1} \rangle \beta_{k+1}) \right. \\
&\qquad \left. \times \overline{\Lambda}_k \overline{\lambda}_{k+1} \mid \mathscr{Y}_{k+1} \right],
\end{aligned}
$$

{because, as in Lemma 2.2,

$$
\begin{aligned}
&\overline{E}\left[\overline{\Lambda}_k \overline{\lambda}_{k+1} V_{k+1} V'_{k+1} \mid \mathscr{Y}_k \right] \\
&\qquad = \overline{E}\left[\overline{E}\left[\overline{\Lambda}_k \overline{\lambda}_{k+1} V_{k+1} V'_{k+1} \mid X_0, X_1, \ldots, X_k, \mathscr{Y}_k \right] \mid \mathscr{Y}_k \right] \\
&\qquad = \overline{E}\left[\left\langle \overline{\Lambda}_k \overline{\lambda}_{k+1} V_{k+1} \right\rangle \mid \mathscr{Y}_k \right] \}
\end{aligned}
$$

$$
\begin{aligned}
= \sum_{j=1}^{N} c_j (Y_{k+1}) \overline{E}\left[\left((H_k + \alpha_{k+1} + \langle \delta_{k+1}, Y_{k+1} \rangle) a_j \right. \right. \\
\left. \left. + \langle V_{k+1} \rangle \beta_{k+1} \right) \overline{\Lambda}_k \left\langle X_k, e_j \right\rangle \mid \mathscr{Y}_{k+1} \right].
\end{aligned}
$$

Finally, because the Y are i.i.d. this final conditioning is the same as conditioning on $\mathscr{Y}_k$. Using Lemma 2.2 and Notation 5.2 the desired result follows. □

2.6 States, Transitions, and Occupation Times

Estimators for the State

Take $H_{k+1} = H_0 = \alpha_0 = 1$, $\alpha_\ell = 0$, $\ell \geq 1$, $\beta_\ell = 0$, $\ell \geq 0$ and $\delta_\ell = 0$, $\ell \geq 0$. Applying Theorem 5.3 we have again the *unnormalized filter* Equation (4.3) for $q_k = (q_k(e_1), \ldots, q_k(e_N))$ in vector form:

$$
\boxed{q_{k+1} = \sum_{j=1}^{N} c_j (Y_{k+1}) \left\langle q_k, e_j \right\rangle a_j.} \tag{6.1}
$$

with normalized form

$$
p_k = q_k \langle q_k, \underline{1} \rangle^{-1}. \tag{6.2}
$$

This form is similar to that given by Aström (1965) and Stratonovich (1960). We can also obtain a recursive form for the unnormalized conditional expectation of $\left\langle X_m, e_p \right\rangle$ given $\mathscr{Y}_{k+1}$, $m < k+1$. This is the *unnormalized smoother*. For this we take $H_{k+1} = H_m = \left\langle X_m, e_p \right\rangle$, $m < k+1$, $1 \leq p \leq N$, $\alpha_\ell = 0$, $\beta_\ell = 0$ and $\delta_\ell = 0$. Applying Theorem 5.3 we have

$$
\boxed{\overline{E}\left[\overline{\Lambda}_{k+1} \langle X_m, e_p \rangle \mid \mathscr{Y}_{k+1} \right] = \sum_{j=1}^{N} c_j (Y_{k+1}) \left\langle \gamma_{m,k} (\langle X_m, e_p \rangle), e_j \right\rangle a_j.} \tag{6.3}
$$

We see that Equation (6.3) is indeed a recursion in k; this is why we consider $H_k X_k$. Taking the inner product with $\underline{1}$ and using Notation 5.1 gives the smoothed, unnormalized estimate

$$\gamma_k\left(\langle X_m, e_p\rangle\right) = \overline{E}\left[\overline{\Lambda}_k \langle X_m, e_p\rangle \mid \mathscr{Y}_k\right].$$

Estimators for the Number of Jumps

The number of jumps from state e_r to state e_s in time k is given by

$$\mathscr{J}_k^{rs} = \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_r\rangle \langle X_\ell, e_s\rangle.$$

Using $X_\ell = AX_{\ell-1} + V_\ell$ this is

$$\begin{aligned}
&= \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_r\rangle \langle AX_{\ell-1}, e_s\rangle + \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_r\rangle \langle V_\ell, e_s\rangle \\
&= \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_1 r\rangle a_{sr} + \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_r\rangle \langle V_\ell, e_s\rangle.
\end{aligned}$$

Applying Theorem 5.3 with $H_{k+1} = \mathscr{J}_{k+1}^{rs}$, $H_0 = 0$, $\alpha_\ell = \langle X_{\ell-1}, e_r\rangle a_{sr}$, $\beta_\ell = \langle X_{\ell-1}, e_r\rangle e_s$, $\delta_\ell = 0$ we have

$$\begin{aligned}
&\gamma_{k+1,k+1}\left(\mathscr{J}_{k+1}^{rs}\right) \\
&\quad = M \sum_{j=1}^{N} \left(\prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i}\right) \Big\{ \left\langle \gamma_{k,k}\left(\mathscr{J}_k^{rs}\right) + \gamma_{k,k}\left(\langle X_k, e_r\rangle a_{sr}\right), e_j\right\rangle a_j \\
&\qquad\qquad + \left[\operatorname{diag}(a_j) - a_j a_j'\right] \\
&\qquad\qquad \times \overline{E}\left[\left\langle \overline{\Lambda}_k X_k, e_j\right\rangle \langle X_k, e_r\rangle e_s \mid \mathscr{Y}_{k+1}\right] \Big\} \\
&\quad = M \sum_{j=1}^{N} \left(\prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i}\right) \left\langle \gamma_{k,k}\left(\mathscr{J}_k^{rs}\right), e_j\right\rangle a_j \\
&\qquad + M \langle q_k, e_r\rangle \left(\prod_{i=1}^{M} c_{ir}^{Y_{k+1}^i}\right) \left[a_{sr} a_r + e_s \operatorname{diag}(a_r) - e_s\left(a_r a_r'\right)\right]
\end{aligned}$$

that is, using Notation 4.2,

$$\boxed{\begin{aligned}
\gamma_{k+1,k+1}\left(\mathscr{J}_{k+1}^{rs}\right) = &\sum_{j=1}^{N} c_j\left(Y_{k+1}\right) \left\langle \gamma_{k,k}\left(\mathscr{J}_k^{rs}\right), e_j\right\rangle a_j \\
&+ c_r\left(Y_{k+1}\right) \langle q_k, e_r\rangle a_{sr} e_s.
\end{aligned}} \tag{6.4}$$

Together with the recursive Equation (6.1) for q_k we have in (6.4) a recursive estimator for $\gamma_{k,k}\left(\mathscr{J}_k^{rs}\right)$. Taking its inner product with $\underline{1}$, that is, summing its components, we obtain $\gamma_k\left(\mathscr{J}_k^{rs}\right) = \overline{E}\left[\overline{\Lambda}_k \mathscr{J}_k^{rs} \mid \mathscr{Y}_k\right]$.

Taking $H_{k+1} = H_m = \mathscr{J}_m^{rs}$, $\alpha_\ell = 0$, $\ell > m$, $\beta_\ell = 0$, $\ell \geq 0$, $\delta_\ell = 0$, $\ell \geq 0$, and applying Theorem 5.3 we obtain for $k > m$, the unnormalized smoothed estimate $\overline{E}\left[\overline{\Lambda}_{k+1} \mathscr{J}_m^{rs} X_{k+1} \mid \mathscr{Y}_{k+1}\right]$

$$\boxed{\gamma_{m,k+1}\left(\mathscr{J}_m^{rs}\right) = \sum_{j=1}^{N} c_j\left(Y_{k+1}\right)\left\langle \gamma_{m,k}\left(\mathscr{J}_m^{rs}\right), e_j\right\rangle a_j.} \tag{6.5}$$

Again, by considering the product $\mathscr{J}_m^{rs} X_k$ a recursive form has been obtained. Taking the inner product with $\underline{1}$ gives the smoothed unnormalized estimate $\overline{E}\left[\overline{\Lambda}_k \mathscr{J}_m^{rs} \mid \mathscr{Y}_k\right]$.

Estimators for the Occupation Time

The number of occasions up to time k for which the Markov chain X has been in state e_r, $1 \leq r \leq N$, is

$$\mathscr{O}_{k+1}^r = \sum_{\ell=1}^{k+1} \langle X_{\ell-1}, e_r\rangle .$$

Taking $H_{k+1} = \mathscr{O}_{k+1}^r$, $H_0 = 0$, $\alpha_\ell = \langle X_{\ell-1}, e_r\rangle$, $\beta_\ell = 0$, $\delta_\ell = 0$ and applying Theorem 5.3 we have

$$\begin{aligned}\gamma_{k+1,k+1}\left(\mathscr{O}_{k+1}^r\right) = M \sum_{j=1}^{N}\prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i}\Big(&\left\langle \gamma_{k,k}\left(\mathscr{O}_k^r\right), e_j\right\rangle \\ &+ \left\langle \gamma_{k,k}\left(\langle X_k, e_r\rangle\right), e_j\right\rangle\Big) a_j.\end{aligned}$$

That is

$$\begin{aligned}\gamma_{k+1,k+1}\left(\mathscr{O}_{k+1}^r\right) = &\sum_{j=1}^{N} c_j\left(Y_{k+1}\right)\left\langle \gamma_{k,k}\left(\mathscr{O}_k^r\right), e_j\right\rangle a_j \\ &+ c_r\left(Y_{k+1}\right)\langle q_k, e_r\rangle a_r.\end{aligned} \tag{6.6}$$

Together with (6.1) for q_k this equation gives a recursive expression for $\gamma_{k,k}\left(\mathscr{O}_k^r\right)$. Taking the inner product with $\underline{1}$ gives $\gamma_k\left(\mathscr{O}_k^r\right) = \overline{E}\left[\mathscr{O}_k^r \mid \mathscr{Y}_k\right]$. For the related smoother take $k > m$, $H_{k+1} = H_m = \mathscr{O}_m^r$, $\alpha_\ell = 0$, $\beta_\ell = 0$, $\delta_\ell = 0$ and apply Theorem 5.3 to obtain

$$\boxed{\gamma_{m,k+1}\left(\mathscr{O}_m^r\right) = \sum_{j=1}^{N} c_j\left(Y_{k+1}\right)\left\langle \gamma_{m,k}\left(\mathscr{O}_m^r\right), e_j\right\rangle a_j.} \tag{6.7}$$

Estimators for State to Observation Transitions

In estimating the parameters of our model in the next section we shall require estimates and smoothers of the process

$$\mathscr{T}_k^{rs} = \sum_{\ell=1}^{k} \langle X_{\ell-1}, e_r \rangle \langle Y_\ell, f_s \rangle$$

which counts the number of times up to time k that the observation process is in state f_s given the Markov chain at the preceding time is in state e_r, $1 \le r \le N$, $1 \le s \le M$. Taking $H_{k+1} = \mathscr{T}_{k+1}^{rs}$, $H_0 = 0$, $\alpha_\ell = 0$, $\beta_\ell = 0$, $\delta_\ell = \langle X_{\ell-1}, e_r \rangle f_s$ and applying Theorem 5.3

$$\gamma_{k+1,k+1}\left(\mathscr{T}_{k+1}^{rs}\right) = M \sum_{j=1}^{N} \prod_{i=1}^{M} c_{ij}^{Y_{k+1}^i} \left(\left\langle \gamma_{k,k}\left(\mathscr{T}_k^{rs}\right), e_j \right\rangle + \left\langle \gamma_{k,k}\left(\langle X_k, e_r \rangle \langle Y_{k+1}, f_s \rangle \right), e_j \right\rangle \right) a_j.$$

That is, using Notation 4.2,

$$\boxed{\gamma_{k+1,k+1}\left(\mathscr{T}_{k+1}^{rs}\right) = \sum_{j=1}^{N} c_j\left(Y_{k+1}\right) \left\langle \gamma_{k,k}\left(\mathscr{T}_k^{rs}\right), e_j \right\rangle a_j + M \langle q_k, e_r \rangle \langle Y_{k+1}, f_s \rangle c_{sr} a_r.}$$

Together with Equation (6.1) for q_k we have a recursive expression for $\gamma_{k,k}\left(\mathscr{T}_k^{rs}\right)$. To obtain the related smoother take $k+1 > m$, $H_{k+1} = H_m = \mathscr{T}_m^{rs}$, $\alpha_\ell = 0$, $\beta_\ell = 0$, $\delta_\ell = 0$ and apply Theorem 5.3 to obtain

$$\boxed{\gamma_{m,k+1}\left(\mathscr{T}_m^{rs}\right) = \sum_{j=1}^{N} c_j\left(Y_{k+1}\right) \left\langle \gamma_{m,k}\left(\mathscr{T}_m^{rs}\right), e_j \right\rangle a_j.} \tag{6.9}$$

This is recursive in k.

Remark 6.1 Note the similar form of the recursions (6.1), (6.4), (6.6), and (6.8). ■

2.7 Parameter Reestimation

In this section we show how, using the expectation maximization (EM) algorithm, the parameters of the model can be estimated. In fact, it is a conditional pseudo *log-likelihood* that is maximized, and the new parameters are expressed in terms of the recursive estimates obtained in Section 2.6. We begin by describing the EM algorithm.

The basic idea behind the EM algorithm is as follows (Baum and Petrie 1966). Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathscr{F})$ all absolutely continuous with respect to a fixed probability measure P_0 and let $\mathscr{Y} \subset \mathscr{F}$. The likelihood function for computing an estimate of the parameter θ based on the information available in $\mathscr{Y}$ is

$$L(\theta) = E_0\left[\frac{dP_\theta}{dP_0} \,\middle|\, \mathscr{Y}\right],$$

and the maximum likelihood estimate (MLE) is defined by

$$\hat{\theta} \in \operatorname*{argmax}_{\theta\in\Theta} L(\theta).$$

The reasoning is that the most likely value of the parameter θ is the one that maximizes this conditional expectation of the density.

In general, the MLE is difficult to compute directly, and the EM algorithm provides an iterative approximation method:

Step 1. Set $p = 0$ and choose $\hat{\theta}_0$.
Step 2. (E-step) Set $\theta^* = \hat{\theta}_p$ and compute $Q(\cdot, \theta^*)$, where

$$Q(\theta, \theta^*) = E_{\theta^*}\left[\log\frac{dP_\theta}{dP_{\theta^*}} \,\middle|\, \mathscr{Y}\right].$$

Step 3. (M-step) Find

$$\hat{\theta}_{p+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta, \theta^*).$$

Step 4. Replace p by $p+1$ and repeat beginning with Step 2 until a stopping criterion is satisfied.

The sequence generated $\left\{\hat{\theta}_p, p \geq 0\right\}$ gives nondecreasing values of the likelihood function to a local maximum of the likelihood function: it follows from Jensen's Inequality, see Appendix A, that

$$\log L\left(\hat{\theta}_{p+1}\right) - \log L\left(\hat{\theta}_p\right) \geq Q\left(\hat{\theta}_{p+1}, \hat{\theta}_p\right),$$

with equality if $\hat{\theta}_{p+1} = \hat{\theta}_p$. We call $Q(\theta, \theta^*)$ a conditional pseudo-log-likelihood. Finding a set of parameters which gives a (local) maximum of the expected log-likelihood function gives an optimal estimate.

Our model (2.14) is determined by the set of *parameters*

$$\theta := (a_{ji},\ 1 \leq i, j \leq N,\ c_{ji},\ 1 \leq j \leq M,\ 1 \leq i \leq N)$$

which are also subject to the constraints (2.15) and (2.16). Suppose our model is determined by such a set θ and we wish to determine a new set

$$\hat{\theta} = (\hat{a}_{ji}(k),\ 1 \leq i, j \leq N,\ \hat{c}_{ji}(k),\ 1 \leq j \leq M,\ 1 \leq i \leq N)$$

which maximizes the conditional pseudo-log-likelihoods defined below. Recall $\mathscr{F}_k$ is the complete σ-field generated by $X_0, X_1, \ldots, X_k$. Consider first the parameters a_{ji}. To replace the parameters a_{ji} by $\hat{a}_{ji}(k)$ in the Markov chain X we define

$$\Lambda_k = \prod_{\ell=1}^{k} \left(\sum_{r,s=1}^{N} \left[\frac{\hat{a}_{sr}(k)}{a_{sr}} \right] \langle X_\ell, e_s \rangle \langle X_{\ell-1}, e_r \rangle \right).$$

In case $a_{ji} = 0$, take $\hat{a}_{ji}(k) = 0$ and $\hat{a}_{ji}(k)/a_{ji} = 0$. Set

$$\left. \frac{dP_{\hat{\theta}}}{dP_\theta} \right|_{\mathscr{F}_k} = \Lambda_k.$$

To justify this we establish the following result.

Lemma 7.1 *Under the probability measure $P_{\hat{\theta}}$ and assuming $X_k = e_r$, then*

$$E_{\hat{\theta}} \left[\langle X_{k+1}, e_s \rangle \mid \mathscr{F}_k \right] = \hat{a}_{sr}(k).$$

Proof

$$\begin{aligned} E_{\hat{\theta}} \left[\langle X_{k+1}, e_s \rangle \mid \mathscr{F}_k \right] &= \frac{E \left[\langle X_{k+1}, e_s \rangle \Lambda_{k+1} \mid \mathscr{F}_k \right]}{E \left[\Lambda_{k+1} \mid \mathscr{F}_k \right]} \\ &= \frac{E \left[\langle X_{k+1}, e_s \rangle \frac{\hat{a}_{sr}(k)}{a_{sr}} \mid \mathscr{F}_k \right]}{E \left[\sum_{r=1}^{N} \left[\frac{\hat{a}_{sr}(k)}{a_{sr}} \right] \langle X_{k+1}, e_s \rangle \mid \mathscr{F}_k \right]} \\ &= \frac{\frac{\hat{a}_{sr}(k)}{a_{sr}} a_{sr}}{\sum_{r=1}^{N} \frac{\hat{a}_{sr}(k)}{a_{sr}} a_{sr}} \\ &= \hat{a}_{sr}(k). \end{aligned}$$

□

Notation 7.2 *For any process ϕ_k, $k \in \mathbb{N}$, write $\hat{\phi}_k = E \left[\phi_k \mid \mathscr{Y}_k \right]$ for its $\mathscr{Y}$-optional projection. In discrete time this conditioning defines the $\mathscr{Y}$-optional projection.*

Theorem 7.3 *The new estimates of the parameter $\hat{a}_{sr}(k)$ given the observations up to time k are given, when defined, by*

$$\boxed{\hat{a}_{sr}(k) = \frac{\hat{\mathscr{J}}_k^{rs}}{\hat{\mathscr{O}}_k^r} = \frac{\gamma_k \left(\mathscr{J}_k^{rs} \right)}{\gamma_k \left(\mathscr{O}_k^r \right)}.} \tag{7.1}$$

We take $\frac{0}{0}$ to be 0.

Proof

$$\begin{aligned}\log \Lambda_k &= \sum_{r,s=1}^{N} \sum_{\ell=1}^{k} \langle X_\ell, e_s \rangle \langle X_{\ell-1}, e_r \rangle [\log \hat{a}_{sr}(k) - \log a_{sr}] \\ &= \sum_{r,s=1}^{N} \mathscr{J}_k^{rs} \log \hat{a}_{sr}(k) + R(a),\end{aligned}$$

where $R(a)$ is independent of $\hat{a}$. Therefore,

$$E[\log \Lambda_k \mid \mathscr{Y}_k] = \sum_{r,s=1}^{N} \hat{\mathscr{J}}_k^{rs} \log \hat{a}_{sr}(k) + \hat{R}(a). \tag{7.2}$$

Now the $\hat{a}_{sr}(k)$ must also satisfy the analog of (2.15)

$$\sum_{s=1}^{N} \hat{a}_{sr}(k) = 1. \tag{7.3}$$

Observe that

$$\sum_{s=1}^{N} \mathscr{J}_k^{rs} = \mathscr{O}_k^r \tag{7.4}$$

and in conditional form

$$\sum_{s=1}^{N} \hat{\mathscr{J}}_k^{rs} = \hat{\mathscr{O}}_k^r. \tag{7.5}$$

We wish, therefore, to choose the $\hat{a}_{sr}(k)$ to maximize (7.2) subject to the constraint (7.3). Write λ for the Lagrange multiplier and put

$$L(\hat{a}, \lambda) = \sum_{r,s=1}^{N} \hat{\mathscr{J}}_k^{rs} \log \hat{a}_{sr}(k) + \hat{R}(a) + \lambda \left(\sum_{s=1}^{N} \hat{a}_{sr}(k) - 1 \right).$$

Differentiating in λ and $\hat{a}_{sr}(k)$, and equating the derivatives to 0, we have the optimum choice of $\hat{a}_{sr}(k)$ is given by the equations

$$\frac{1}{\hat{a}_{sr}(k)} \hat{\mathscr{J}}_k^{rs} + \lambda = 0, \tag{7.6}$$

$$\sum_{s=1}^{N} \hat{a}_{sr}(k) = 1. \tag{7.7}$$

From (7.5)–(7.7) we see that $\lambda = -\hat{\mathscr{O}}_k^r$ so the optimum choice of $\hat{a}_{sr}(k)$, $1 \le s, r \le N$, is

$$\hat{a}_{sr}(k) = \frac{\hat{\mathscr{J}}_k^{rs}}{\hat{\mathscr{O}}_k^r} = \frac{\gamma_k(\mathscr{J}_k^{rs})}{\gamma_k(\mathscr{O}_k^r)}. \tag{7.8}$$

□

Note that the unnormalized conditional expectations in (7.8) are given by the inner product with $\underline{1}$ of (6.4) and (6.6).

Consider now the parameters c_{ji} in the matrix C. To replace the parameters c_{sr} by $\hat{c}_{sr}(k)$ we must now consider the Radon-Nikodym derivative

$$\widetilde{\Lambda}_k = \prod_{\ell=1}^{k} \left(\sum_{r=1}^{N} \sum_{s=1}^{M} \left[\frac{\hat{c}_{sr}(k)}{c_{sr}} \right] \langle X_{\ell-1}, e_r \rangle \langle Y_\ell, f_s \rangle \right).$$

By analogy with Lemma 3.1 we introduce a new probability by setting

$$\left. \frac{dP_{\hat{\theta}}}{dP_\theta} \right|_{\mathscr{G}_k} = \widetilde{\Lambda}_k.$$

Then $E_{\hat{\theta}} [\langle Y_{k+1}, f_s \rangle \mid X_k = e_r] = \hat{c}_{sr}(k)$.

Then

$$E \left[\log \widetilde{\Lambda}_k \mid \mathscr{Y}_k \right] = \sum_{r=1}^{N} \sum_{s=1}^{M} \mathscr{T}_k^{rs} \log \hat{c}_{sr}(k) + \widetilde{R}(c), \tag{7.9}$$

where $\widetilde{R}(c)$ is independent of $\hat{c}$. Now the $\hat{c}_{sr}(k)$ must also satisfy

$$\sum_{s=1}^{M} \hat{c}_{sr}(k) = 1. \tag{7.10}$$

Observe that

$$\sum_{s=1}^{M} \mathscr{T}_k^{rs} = \mathscr{O}_k^r$$

and conditional form

$$\sum_{s=1}^{M} \hat{\mathscr{T}}_k^{rs} = \hat{\mathscr{O}}_k^r. \tag{7.11}$$

We wish, therefore, to choose the $\hat{c}_{sr}(k)$ to maximize (7.9) subject to the constraint (7.11). Following the same procedure as above we obtain:

Theorem 7.4 *The maximum log likelihood estimates of the parameters $\hat{c}_{sr}(k)$ given the observation up to time k are given, when defined, by*

$$\boxed{\hat{c}_{sr}(k) = \frac{\gamma_k\left(\mathscr{T}_k^{rs}\right)}{\gamma_k\left(\mathscr{O}_k^r\right)}.} \tag{7.12}$$

We take $\dfrac{0}{0}$ *to be* 0.

Together with the estimates for $\gamma_k\left(\mathscr{T}_k^{rs}\right)$ given by the inner product with $\underline{1}$ of Equation (6.8) and the estimates for $\gamma_k\left(\mathscr{O}_k^r\right)$ given by taking the inner product with $\underline{1}$

of Equation (6.6) we can determine the optimal choice for $\hat{c}_{sr}(k)$, $1 \le s \le M-1$, $1 \le r \le N$. However, $\sum_{s=1}^{M} \hat{c}_{sr}(k) = 1$ for each r, so the remaining $\hat{c}_{Mr}(k)$ can also be found.

Remarks 7.5 The revised parameters $\hat{a}_{sr}(k)$, $\hat{c}_{sr}(k)$ determined by (7.8) and (7.12) give new probability measures for the model. The quantities $\gamma_k\left(\mathscr{J}_k^{rs}\right)$, $\gamma_k\left(\mathscr{T}_k^{rs}\right)$, $\gamma_k\left(\mathscr{O}_k^{r}\right)$ can then be reestimated using the new parameters and perhaps new data, together with smoothing equations.

■

2.8 Recursive Parameter Estimation

In Section 2.7 we obtained estimates for the a_{ji} and the c_{ji}. However, these are not recursive, that is, the estimate at time k is not expressed as the estimate at time $(k-1)$ plus a correction based on new information. In this section we derive *recursive estimates* for the parameters. Unfortunately, these recursions are not in general finite-dimensional. Recall our discrete HMM signal model (2.14) is parametrized in terms of a_{ji}, c_{ji}. Let us collect these parameters into a parameter vector θ, so that we can write $A = A(\theta)$, $C = C(\theta)$. Suppose that θ is not known a priori. Let us estimate θ in a recursive manner, given the observations $\mathscr{Y}_k$. We assume that θ will take values in some set $\Theta \in \mathbb{R}^p$.

Let us now write $\mathscr{G}_k$ for the complete σ-field generated by knowledge of $X_0, X_1, \ldots, X_k, Y_1, \ldots, Y_k$, together with θ. Again $\mathscr{Y}_k$ will be the complete σ-field generated by knowledge of $Y_1, \ldots, Y_k$. With this enlarged $\mathscr{G}_k$ the results of Sections 2.2 and 2.3 still hold. We suppose there is a probability $\overline{P}$ on $(\Omega \times \Theta, \bigvee_{\ell=1}^{\infty} \mathscr{G}_\ell)$ such that, under $\overline{P}$, the $\{Y_\ell\}$ are i.i.d. with $\overline{P}\left(Y_\ell^j = 1\right) = \frac{1}{M}$, and $X_{k+1} = AX_k + V_{k+1}$, where V_k is a $\left(\overline{P}, \mathscr{G}_k\right)$ martingale increment. Write $q_k^r(\theta)$, $1 \le r \le N$, $k \in \mathbb{N}$, for an unnormalized, conditional density such that

$$\overline{E}\left[\overline{\Lambda}_k \langle X_k, e_r \rangle I(\theta \in d\theta) \mid \mathscr{Y}_k\right] = q_k^r(\theta)\, d\theta.$$

Where $d\theta$ is Lebesgue measure on $\Theta \in \mathbb{R}^p$.

Here, $I(A)$ is the indicator function of the set A, that is, the function that is 1 on A and 0 otherwise. The existence of $q_k^r(\theta)$ will be discussed below. Equalities in the variable θ can be interpreted almost surely.

The normalized conditional density $p_k^r(\theta)$, such that

$$p_k^r(\theta)\, d\theta = E\left[\langle X_k, e_r \rangle I(\theta \in d\theta) \mid \mathscr{Y}_k\right],$$

is then given by

$$p_k^r(\theta) = \frac{q_k^r(\theta)}{\sum_{j=1}^{N} \int_\Theta q_k^j(u)\, du}.$$

We suppose an initial distribution $p_0(.) = \left(p_0^1(.), \ldots, p_0^N(.)\right)$ is given. This is further discussed in Remark 8.2. A recursive expression for $q_k^r(\theta)$ is now obtained:

Theorem 8.1 *For $k \in \mathbb{N}$, and $1 \leq r \leq N$, then the recursive estimates of an unnormalized joint conditional distribution of X_k and θ are given by*

$$\boxed{q_{k+1}^r(\theta) = a'_{r,(.)} \operatorname{diag}(q_k(\theta))\, c_{(.)}(Y_{k+1}).} \tag{8.1}$$

Proof Suppose g is any real-valued *Borel function* on Θ. Then

$$\begin{aligned}
&\overline{E}\left[\langle X_{k+1}, e_r\rangle g(\theta)\overline{\Lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&\quad = \int_\Theta q_{k+1}^r(u) g(u)\, du \qquad (8.2) \\
&\quad = \overline{E}\left[\langle AX_k + V_{k+1}, e_r\rangle g(\theta)\overline{\Lambda}_k \sum_{i=1}^M M\langle CX_k, f_i\rangle \langle Y_{k+1}, f_i\rangle \,\middle|\, \mathscr{Y}_{k+1}\right] \\
&\quad = M\overline{E}\left[\langle AX_k, e_r\rangle g(\theta)\overline{\Lambda}_k \sum_{i=1}^M \langle CX_k, f_i\rangle \langle Y_{k+1}, f_i\rangle \,\middle|\, \mathscr{Y}_{k+1}\right] \\
&\quad = M\sum_{s=1}^N \overline{E}\left[\langle X_k, e_s\rangle a_{rs} g(\theta)\overline{\Lambda}_k \mid \mathscr{Y}_k\right] \prod_{i=1}^M c_{is}^{Y_{k+1}^i} \\
&\quad = M\int_\Theta \sum_{s=1}^N a_{rs} q_k^s(u) g(u)\, du \prod_{i=1}^M c_{is}^{Y_{k+1}^i}. \qquad (8.3)
\end{aligned}$$

As g is arbitrary, from (8.2) and (8.3) we see

$$q_{k+1}^r(u) = M\sum_{s=1}^N \left(a_{rs} q_k^s(u) \prod_{i=1}^M c_{is}^{Y_{k+1}^i}\right).$$

Using Notation 4.2 the result follows. □

Compared with Theorem 4.3 the new feature of Theorem 8.1 is that it updates recursively the estimate of the parameter.

Remark 8.2 Suppose $\pi = (\pi_1, \ldots, \pi_N)$, where $\pi_i = P(X_0 = e_i)$ is the initial distribution for X_0 and $h(\theta)$ is the prior density for θ. Then

$$q_0^r(\theta) = \pi_r h(\theta),$$

and the updated estimates are given by (8.1).

■

If the prior information about X_0 is that, say, $X_0 = e_i$, then the dynamics of X, (2.4) will move the state around and the estimate is given by (8.1). If the prior information about θ is that θ takes a particular value, then $h(\theta)$ (or a factor of h)

is a delta function at this value. No noise or dynamics enters into θ, so the equations (8.1) just continue to give the delta function at this value. This is exactly to be expected. The prior distribution h taken for θ must represent the a priori information about θ; it is not an initial guess for the value of θ.

Time-varying dynamics for θ could be incorporated in our model. Possibly $\theta_{k+1} = A_\theta \theta_k + v_{k+1}$, where v_{k+1} is the noise term. However, the problem then arises of estimating the terms of the matrix A_θ.

Finally, we note the equations (8.1) are really just a family of equations parametrized by θ. In particular, if θ can take one of finitely many values $\theta_1, \theta_2, \ldots, \theta_p$ we obtain p equations (8.1) for each possible θ_i. The prior for θ is then just a distribution over $\theta_1, \ldots, \theta_p$.

2.9 Quantized Observations

Suppose now the signal process $\{x_\ell\}$ is of the form

$$x_{k+1} = Ax_k + v_{k+1},$$

where $x_k \in \mathbb{R}^d$, $A = (a_{ji})$ is a $d \times d$ matrix and $\{v_\ell\}$, $\ell \in \mathbb{N}$, is a sequence of i.i.d. random variables with density function ψ. (Time-varying densities or nonlinear equations for the signal can be considered.) We suppose x_0, or its distribution, is known. The observation process is again denoted by Y_ℓ, $\ell \in \mathbb{N}$. However, the observations are *quantized*, so that the range space of Y_ℓ is finite. Here, also, we shall identify the range of Y_ℓ with the unit vectors $f_1, \ldots, f_M$, $f_j = (0, \ldots, 1, \ldots, 0)' \in \mathbb{R}^M$, for some M. Again suppose some parameters $\theta \in \Theta$ in the model are not known. Write $\mathscr{G}_k$ for the complete σ-field generated by $x_0, x_1, \ldots, x_k$, $Y_1, \ldots, Y_k$ and θ; $\mathscr{Y}_k$ is the complete σ-field generated by $Y_1, \ldots, Y_k$. If $Y_\ell^i = \langle Y_\ell, f_i \rangle$, $1 \le i \le M$, then $Y_\ell = \left(Y_\ell^i, \ldots, Y_\ell^M\right)'$ and $\sum_{i=1}^M Y_\ell^i = 1$. Write

$$c_\ell^i = E\left[\langle Y_\ell, f_i \rangle \mid \mathscr{G}_{\ell-1}\right] = P\left(Y_\ell = f_i \mid \mathscr{G}_{\ell-1}\right).$$

We shall suppose

$$P\left(Y_\ell = f_i \mid \mathscr{G}_{\ell-1}\right) = P\left(Y_\ell = f_i \mid x_{\ell-1}\right), \qquad 1 \le i \le M, \ell \in \mathbb{N}.$$

In this case we write $c_\ell^i(x_{\ell-1})$. Suppose $c_\ell^i(x_{\ell-1}) > 0$, $1 \le i \le M$, $\ell \in \mathbb{N}$. Write

$$\Lambda_k = \prod_{\ell=1}^k \left(\sum_{i=1}^M \left[\frac{1}{M c_\ell^i(x_{\ell-1})} \right] \langle Y_\ell, f_i \rangle \right).$$

Defining $\overline{P}$ by setting

$$\left. \frac{d\overline{P}}{dP} \right|_{\mathscr{G}_k} = \Lambda_k$$

gives a measure such that

$$\overline{E}\left[\langle Y_\ell, f_i\rangle \mid \mathscr{G}_{\ell-1}\right] = \frac{1}{M}.$$

Suppose the parameter θ takes values in $\mathbb{R}^d$ and is random and unknown.

Suppose we start with a measure $\overline{P}$ on $\left(\Omega \times \mathbb{R}^d, \bigvee_{\ell=1}^{\infty} \mathscr{G}_\ell\right)$ such that

$$\overline{E}\left[\langle Y_\ell, f_i\rangle \mid \mathscr{G}_{\ell-1}\right] = \frac{1}{M}$$

and $x_{k+1} = Ax_k + v_{k+1}$. Write

$$\overline{\Lambda}_k = \prod_{\ell=1}^{k}\left(\sum_{i=1}^{M} M\left[c_\ell^i(x_{\ell-1})\right]\langle Y_\ell, f_i\rangle\right).$$

[Note this no longer requires $c_{k+1}^i(x_k) > 0$.]

Introduce P by putting

$$\left.\frac{dP}{d\overline{P}}\right|_{\mathscr{G}_k} = \overline{\Lambda}_k.$$

Suppose f is any Borel function on $\mathbb{R}^d$ and g is any Borel function on Θ, and write $q_k(z,\theta)$ for an unnormalized conditional density such that

$$\overline{E}\left[\overline{\Lambda}_k I(x_k \in dz) I(\theta \in d\theta) \mid \mathscr{Y}_k\right] = q_k(z,\theta)\, dz d\theta.$$

Then

$$\overline{E}\left[f(x_{k+1}) g(\theta) \overline{\Lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] = \iint f(\xi) g(u) q_{k+1}(\xi, u)\, d\xi\, d\lambda(u). \tag{9.1}$$

The right-hand side is also equal to

$$= M\overline{E}\left[f(Ax_k + v_{k+1}) g(\theta) \overline{\Lambda}_k \prod_{i=1}^{M} c_{k+1}^i(x_k)^{Y_{k+1}^i} \mid \mathscr{Y}_{k+1}\right]$$

$$= M \iiint f(Az+v) g(u) \left[\prod_{i=1}^{M} c_{k+1}^i(z)^{Y_{k+1}^i}\right] \psi(v) q_k(z,u)\, dv\, dz\, d\lambda(u).$$

Write $\xi = Az + v$, so $v = \xi - Az$. The above is

$$= M \iiint f(\xi) g(u) \left(\prod_{i=1}^{M} c_{k+1}^i(z)^{Y_{k+1}^i}\right) \psi(\xi - Az) q_k(z,u)\, dz\, d\xi\, d\lambda(u). \tag{9.2}$$

Comparing (9.1) and (9.2) and denoting

$$c_{k+1}(Y_{k+1},z) = M\sum_{i=1}^{M} c_{k+1}^{i}(z)\,\langle Y_{k+1}, f_i\rangle$$

we have the following result:

Theorem 9.1 *The recursive estimate of an unnormalized joint conditional density of the signal x and the parameter θ satisfies*:

$$\boxed{q_{k+1}(\xi,u) = \int_{\mathbb{R}^d} c_{k+1}(Y_{k+1},z)\,\psi(\xi - Az)\,q_k(z,u)\,dz.}$$

Example

In Kulhavy (1990) the following simple situation is considered. Suppose $\theta \in \mathbb{R}$ is unknown. $\{v_\ell\}$, $\ell \in \mathbb{N}$, is a sequence of i.i.d. $N\left(0,\sigma^2\right)$ random variables. The real line is partitioned into M disjoint intervals,

$$I_1 = (-\infty,\alpha_1), I_2 = [\alpha_1,\alpha_2), \ldots, I_{M-1} = [\alpha_{M-2},\alpha_{M-1}), I_M = [\alpha_M,\infty).$$

The signal process is $x_\ell = \theta + v_\ell$, $\ell \in \mathbb{N}$. The observation process Y_ℓ is an M-dimensional unit vector such that $Y_\ell^i = 1$ if $x_\ell \in I_i$. Then

$$\begin{aligned} c_\ell^i &= P\left(Y_\ell^i = 1 \mid \mathscr{G}_{\ell-1}\right) \\ &= P\left(Y_\ell^i = 1 \mid \theta\right) = P\left(\alpha_{i-1} \le Y_\ell^i < \alpha_i \mid \theta\right) \\ &= \left(2\pi\sigma^2\right)^{-1/2} \int_{\alpha_{i-1}-\theta}^{\alpha_i-\theta} \exp\left(-x^2/2\sigma^2\right) dx \\ &= c_\ell^i(\theta), \qquad 1 \le i \le M. \end{aligned}$$

Measure $\overline{P}$ is now introduced. Write $q_k(\theta)$ for the unnormalized conditional density such that

$$\overline{E}\left[\overline{\Lambda}_k I(\theta \in d\theta) \mid \mathscr{Y}_k\right] = q_k(\theta)\,d\theta.$$

Then, for an arbitrary Borel function g,

$$\begin{aligned} \overline{E}\left[g(\theta)\,\overline{\Lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] &= \int_{\mathbb{R}} g(\lambda)\,q_{k+1}(\lambda)\,d\lambda \\ &= M\overline{E}\left[g(\theta)\,\overline{\Lambda}_k \sum_{i=1}^{M} c_{k+1}^i(\theta)\,\langle Y_{k+1}, f_i\rangle \mid \mathscr{Y}_{k+1}\right] \\ &= M\int_{\mathbb{R}} g(\lambda)\left[\sum_{i=1}^{M} c_{k+1}^i(\lambda)\,\langle Y_{k+1}, f_i\rangle\right] q_k(\lambda)\,d\lambda. \end{aligned}$$

We, therefore, have the following recursion formula for the unnormalized conditional density of θ:

$$\boxed{q_{k+1}(\lambda) = \left(\prod_{i=1}^{M} c_{k+1}^{i}(\lambda)^{Y_{k+1}^{i}} \right) q_k(\lambda).} \tag{9.3}$$

The conditional density of θ given $\mathscr{Y}_k$ is then

$$p_k(\lambda) = \frac{q_k(\lambda)}{\int_{\mathbb{R}} q_k(\xi)\, d\xi}.$$

2.10 The Dependent Case

The situation considered in this section, (which may be omitted on a first reading), is that of a hidden Markov Model for which the "noise" terms in the state and observation processes are possibly dependent. An elementary prototype of this situation, for which the observation process is a single point process, is discussed in Segall (1976b). The filtrations $\{\mathscr{F}_k\}$, $\{\mathscr{G}_k\}$ and $\{\mathscr{Y}_k\}$ are as defined in Section 1.2. The semimartingale form of the Markov chain is, as in Section 2.2,

$$X_{k+1} = AX_k + V_{k+1}, \qquad k \in \mathbb{N},$$

where V_k is an $\{\mathscr{F}_k\}$ martingale increment, $a_{ji} = P(X_{k+1} = e_j \mid X_k = e_i)$ and $A = (a_{ji})$. Again the Markov chain is not observed directly; rather we suppose there is a finite-state observation process Y. The relation between X and Y can be given as $P(Y_{k+1} = f_r \mid \mathscr{G}_k) = P(Y_{k+1} = f_r \mid X_k)$ so that

$$Y_{k+1} = CX_k + W_{k+1}, \qquad k \in \mathbb{N},$$

where W_k is an $\{\mathscr{G}_k\}$ martingale increment, $c_{ji} = P(Y_{k+1} = f_j \mid X_k = e_i)$ and $C = (c_{ji})$. We initially assume c_{ji} positive for $1 \le i \le N$ and $1 \le j \le M$.

However, the noise, or martingale increment, terms V_k and W_k are not independent. In fact, the joint distribution of Y_k and X_k is supposed, given by

$$Y_{k+1}X_{k+1}' = SX_k + \Gamma_{k+1}, \qquad k \in \mathbb{N},$$

where $S = (s_{rji})$ denotes a $MN \times N$ matrix, or tensor, mapping $\mathbb{R}^N$ into $\mathbb{R}^M \times \mathbb{R}^N$ and

$$s_{rji} = P(Y_k = f_r,\, X_k = e_j \mid X_{k-1} = e_i) \qquad 1 \le r \le M,\ 1 \le i, j \le N.$$

Again Γ_{k+1} is a martingale increment, so $E[\Gamma_{k+1} \mid \mathscr{G}_k] = 0$.

If the terms are independent

$$SX_k = CX_k (AX_k)'.$$

In this dependent case, recursive estimates are derived for the state of the chain, the number of jumps from one state to another, the occupation time of the chain in any state, the number of transitions of the observation process into a particular state, and the number of joint transitions of the chain and the observation process. Using the expectation maximization algorithm optimal estimates are obtained for the elements a_{ji}, c_{ji} and s_{rji} of the matrices A, C, and S, respectively. Our model is again, therefore, adaptive or "self-tuning." In the independent case our results specialize to those of Section 2.5.

Dependent Dynamics

We shall suppose

$$P(Y_{k+1} = f_r, X_{k+1} = e_j \mid \mathscr{G}_k) = P(Y_{k+1} = f_r, X_{k+1} = e_j \mid X_k) \tag{10.1}$$

and write

$$s_{rji} = P(Y_{k+1} = f_r, X_{k+1} = e_j \mid X_k = e_i), \qquad 1 \le r \le M,\ 1 \le i, j \le N.$$

Then $S = (s_{rji})$ denotes a $MN \times N$ matrix, or tensor, mapping $\mathbb{R}^N$ into $\mathbb{R}^M \times \mathbb{R}^N$. From this hypothesis we have immediately:

$$Y_{k+1}X'_{k+1} = SX_k + \Gamma_{k+1}, \qquad k \in \mathbb{N}, \tag{10.2}$$

where Γ_{k+1} is a $(P, \mathscr{G}_k), \mathbb{R}^M \times \mathbb{R}^N$ martingale increment.

Remark 10.1 Our model, therefore, involves the three sets of parameters (a_{ji}), (c_{ri}), and (s_{rji}).

■

Write $\underline{1} = (1, 1, \ldots, 1)'$ for the vector, in $\mathbb{R}^M$ or $\mathbb{R}^N$ according to context, all components of which are 1.

Lemma 10.2 *For* $\underline{1} \in \mathbb{R}^M$, *then*

$$\langle \underline{1}, SX_k \rangle = AX_k. \tag{10.3}$$

For $\underline{1} \in \mathbb{R}^N$, *then*

$$\langle SX_k, \underline{1} \rangle = CX_k. \tag{10.4}$$

Proof In each case $\langle \underline{1}, \Gamma_k \rangle$ and $\langle \Gamma_k, \underline{1} \rangle$ are martingale increments. Taking the inner product of (10.2) with $\underline{1}$ the left side is, respectively, either $\langle \underline{1}, Y_{k+1}X'_{k+1} \rangle = X_k$ or $\langle Y_{k+1}X'_{k+1}, \underline{1} \rangle = Y_{k+1}$. Therefore, the result follows from the unique decompositions of the special semimartingales X_k and Y_k. □

In contrast to the independent situation, we have here $P\left[X_{k+1}=e_j \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] = P\left[X_{k+1}=e_j \mid X_k, \mathscr{Y}_{k+1}\right]$. This is not, in general, equal to $P\left[X_{k+1}=e_j \mid X_k\right]$ so that knowledge of $\mathscr{Y}_k$, or in particular Y_k, now gives extra information about X_k.

Write

$$\alpha_{jir} = \frac{s_{rji}}{c_{ri}};$$

(recall the c_{ri} are positive). We then have the following:

Lemma 10.3 *With $\tilde{A}$ the $N \times (N \times M)$ matrix (α_{jir}), $1 \le i, j \le N$, $1 \le r \le M$,*

$$X_{k+1} = \tilde{A}\left(X_k Y'_{k+1}\right) + \tilde{V}_{k+1},$$

where

$$E\left[\tilde{V}_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] = 0. \tag{10.5}$$

Proof

$$\begin{aligned} &P\left[X_{k+1}=e_j \mid X_k=e_i, Y_{k+1}=f_r\right] \\ &\quad = \frac{P\left[Y_{k+1}=f_r, X_{k+1}=e_j \mid X_k=e_i\right]}{P\left[Y_{k+1}=f_r \mid X_k=e_i\right]} \\ &\quad = \frac{s_{rji}}{c_{ri}} = \alpha_{jir}. \end{aligned}$$

With $\tilde{A} = (\alpha_{jir})$, $1 \le i, j \le N$, $1 \le r \le M$, we define $\tilde{V}_k$ by putting

$$X_{k+1} = \tilde{A}\left(X_k Y'_{k+1}\right) + \tilde{V}_{k+1}. \tag{10.6}$$

Then

$$\begin{aligned} E\left[\tilde{V}_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] &= E\left[X_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] - \tilde{A}\left(X_k Y'_{k+1}\right) \\ &= \tilde{A}\left(X_k Y'_{k+1}\right) - \tilde{A}\left(X_k Y'_{k+1}\right) = 0. \end{aligned}$$

□

In summary then, we have the following.

Dependent Discrete HMM The dependent discrete HMM is

$$\boxed{\begin{aligned} X_{k+1} &= \tilde{A}\left(X_k Y'_{k+1}\right) + \tilde{V}_{k+1} \\ Y_{k+1} &= CX_k + W_{k+1}, \qquad k \in \mathbb{N}, \end{aligned}} \tag{10.7}$$

where $X_k \in S_X$, $Y_k \in S_Y$, $\tilde{A}$ and C are matrices of transition probabilities given in Lemmas 10.3 and (2.8). The entries of $\tilde{A}$ satisfy

$$\sum_{j=1}^{N} \alpha_{jir} = 1, \qquad \alpha_{jir} \geq 0. \tag{10.8}$$

$\tilde{V}_k$ is a martingale increment satisfying

$$E\left[\tilde{V}_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] = 0.$$

Next, we derive filters and smoothers for various processes.

The State Process

We shall be working under a probability measure $\overline{P}$ as discussed in Sections 2.3 and 2.4, so that the observation process is a sequence of i.i.d. random variables, uniformly distributed over the set of standard unit vectors $\{f_1, \ldots, f_M\}$ of $\mathbb{R}^M$.

Here Λ_k is as defined in Section 2.3. Using Bayes' Theorem we see that

$$\begin{aligned}
\overline{P}\left[X_{k+1} = e_j \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] &= \overline{E}\left[\left\langle X_{k+1}, e_j\right\rangle \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] \\
&= \frac{E\left[\left\langle X_{k+1}, e_j\right\rangle \Lambda_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right]}{E\left[\Lambda_{k+1} \mid \mathscr{G}_k, \mathscr{Y}_{k+1}\right]} \\
&= \frac{\Lambda_{k+1} E\left[\left\langle X_{k+1}, e_j\right\rangle \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right]}{\Lambda_{k+1}} \\
&= P\left[X_{k+1} = e_j \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] \\
&= P\left[X_{k+1} = e_j \mid X_k, Y_{k+1}\right].
\end{aligned}$$

Therefore under $\overline{P}$, the process X satisfies (10.7). Write $\tilde{q}_k$, $k \in \mathbb{N}$, for the unnormalized conditional probability distribution such that

$$\overline{E}\left[\overline{\Lambda}_k X_k \mid \mathscr{Y}_k\right] := \tilde{q}_k.$$

Also write

$$\tilde{A}\left(e_j f_r'\right) = \alpha_{\cdot jr} = \left(\alpha_{1jr}, \alpha_{2jr}, \ldots, \alpha_{Njr}\right) \text{ and } s_{r\cdot j} = \left(s_{r1j}, \ldots, s_{rNj}\right).$$

Lemma 10.4 *A recursive formula for $\tilde{q}_{k+1}$ is given by*

$$\boxed{\tilde{q}_{k+1} = M \sum_{r=1}^{M} \sum_{j=1}^{N} \left\langle \tilde{q}_k, e_j \right\rangle \left\langle Y_{k+1}, f_r \right\rangle s_{r\cdot j} = MS\tilde{q}_k Y_{k+1}'.} \tag{10.9}$$

Proof

$$\begin{aligned}
\tilde{q}_{k+1} &= \overline{E}\left[\overline{\Lambda}_{k+1}X_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&= \overline{E}\left[\overline{\Lambda}_k \prod_{r=1}^{M} \left(M\left\langle CX_k, f_r\right\rangle\right)^{Y_{k+1}^r} \overline{E}\left[X_{k+1} \mid \mathscr{F}_k, \mathscr{Y}_{k+1}\right] \mid \mathscr{Y}_{k+1}\right] \\
&= \overline{E}\left[\overline{\Lambda}_k \prod_{r=1}^{M} \left(M\left\langle CX_k, f_r\right\rangle\right)^{Y_{k+1}^r} \tilde{A}X_k Y_{k+1}' \mid \mathscr{Y}_{k+1}\right] \\
&= M\sum_{r=1}^{M}\sum_{j=1}^{N}\left\langle \tilde{q}_k, e_j\right\rangle \left\langle Y_{k+1}, f_r\right\rangle c_{rj}\alpha_{\cdot jr} \\
&= M\sum_{r=1}^{M}\sum_{j=1}^{N}\left\langle \tilde{q}_k, e_j\right\rangle \left\langle Y_{k+1}, f_r\right\rangle s_{r\cdot j} \\
&= MS\tilde{q}_k Y_{k+1}'.
\end{aligned}$$

□

Remark 10.5 If the noise terms in the state X and observation Y are independent, then

$$\begin{aligned}
SX_k &= E\left[Y_{k+1}X_{k+1}' \mid \mathscr{G}_k\right] \\
&= CX_k\left(AX_k\right)' \\
&= \sum_{i=1}^{N}\left\langle X_k, e_i\right\rangle c_i a_i',
\end{aligned}$$

where $c_i = Ce_i$ and $a_i = Ae_i$.

■

A General Recursive Filter

Suppose H_k is a scalar $\mathscr{G}$-adapted process such that H_0 is $\mathscr{F}_0$ measurable. With $\Delta H_{k+1} = H_{k+1} - H_k$, $H_{k+1} = H_k + \Delta H_{k+1}$. For any $\mathscr{G}$-adapted process ϕ_k, $k \in \mathbb{N}$, write $\tilde{\gamma}_{m,k}(\phi_m) = \overline{E}\left[\overline{\Lambda}_k \phi_m X_k \mid \mathscr{Y}_k\right]$. Then

$$\begin{aligned}
&\tilde{\gamma}_{k+1,k+1}(H_{k+1}) \\
&\quad = \overline{E}\left[\overline{\Lambda}_{k+1}H_k X_{k+1} \mid \mathscr{Y}_{k+1}\right] + \overline{E}\left[\overline{\Lambda}_{k+1}\Delta H_{k+1}X_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&\quad = \overline{E}\left[\overline{\Lambda}_k H_k \tilde{A}\left(X_k Y_{k+1}'\right)\overline{\lambda}_{k+1} \mid \mathscr{Y}_{k+1}\right] + \overline{E}\left[\overline{\Lambda}_{k+1}\Delta H_{k+1}X_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&\quad = M\sum_{r=1}^{M}\sum_{j=1}^{N}\left\langle \tilde{\gamma}_{k,k}(H_k), e_j\right\rangle \left\langle Y_{k+1}, f_r\right\rangle s_{r\cdot j} \\
&\qquad\qquad + \overline{E}\left[\overline{\Lambda}_{k+1}\Delta H_{k+1}X_{k+1} \mid \mathscr{Y}_{k+1}\right] \\
&\quad = MS\tilde{\gamma}_{k,k}(H_k)Y_{k+1}' + \overline{E}\left[\overline{\Lambda}_{k+1}\Delta H_{k+1}X_{k+1} \mid \mathscr{Y}_{k+1}\right]. \qquad (10.10)
\end{aligned}$$

For the smoother at time $m < k+1$, we have

$$\begin{aligned}\tilde{\gamma}_{m,k+1}(H_m) &= M\sum_{r=1}^{M}\sum_{j=1}^{N}\left\langle \tilde{\gamma}_{m,k}(H_m), e_j\right\rangle \langle Y_{k+1}, f_r\rangle s_{r\cdot j}\\ &= MS\tilde{\gamma}_{m,k}(H_m)Y'_{k+1}. \end{aligned} \tag{10.11}$$

Remark 10.6 The use of the product $H_{k+1}X_{k+1}$ and H_mX_{k+1} is explained in Section 2.5. Specializing (10.10) and (10.11), estimates and smoothers for various processes of interest are now obtained.

■

The State Process

Here $H_{k+1} = H_0 = 1$ and $\Delta H_{k+1} = 0$. Denoting $\tilde{\gamma}_{k,k}(1)$ by $\tilde{q}_k$ we have from (10.10) and (10.11)

$$\boxed{\tilde{q}_{k+1} = MS\tilde{q}_kY'_{k+1}} \tag{10.12}$$

which we have already obtained in Lemma 10.4. For $m < k+1$ we have the smoothed estimate

$$\boxed{\tilde{\gamma}_{m,k+1}\left(\left\langle X_m, e_p\right\rangle\right) = MS\tilde{\gamma}_{m,k}\left(\left\langle X_m, e_p\right\rangle\right)Y'_{k+1}.} \tag{10.13}$$

The Number of Jumps

Here $H_{k+1} = \mathscr{J}^{pq}_{k+1} = \sum_{n=1}^{k+1}\left\langle X_{n-1}, e_q\right\rangle\left\langle X_n, e_p\right\rangle$ and $\Delta H_{k+1} = \left\langle X_k, e_p\right\rangle \times \left\langle X_{k+1}, e_q\right\rangle$. Substitution of these quantities in (10.10) and (10.11) gives the estimates and smoothers for the number of jumps:

$$\boxed{\tilde{\gamma}_{k+1,k+1}\left(\mathscr{J}^{pq}_{k+1}\right) = M\left(S\tilde{\gamma}_{k,k}\left(\mathscr{J}^{pq}_{k}\right)Y'_{k+1} + \left\langle \tilde{q}_k, e_p\right\rangle\left\langle Y_{k+1}, s_{\cdot qp}\right\rangle e_q\right)} \tag{10.14}$$

and for $m < k+1$ we have the smoothed estimate

$$\boxed{\tilde{\gamma}_{m,k+1}\left(\mathscr{J}^{pq}_{m}\right) = MS\tilde{\gamma}_{m,k}\left(\mathscr{J}^{pq}_{m}\right)Y'_{k+1}.} \tag{10.15}$$

The Occupation Time

Here $H_{k+1} = \mathscr{O}^{p}_{k+1} = \sum_{n=1}^{k+1}\left\langle X_n, e_p\right\rangle$ and $\Delta H_{k+1} = \left\langle X_k, e_p\right\rangle$. Using again (10.10) and (10.11) we have the estimates

$$\boxed{\tilde{\gamma}_{k+1,k+1}\left(\mathscr{O}^{p}_{k+1}\right) = M\left(S\tilde{\gamma}_{k,k}\left(\mathscr{O}^{p}_{k}\right)Y'_{k+1} + \left\langle \tilde{q}_k, e_p\right\rangle\left\langle Y_{k+1}, s_{\cdot\cdot p}\right\rangle\right),} \tag{10.16}$$

where $\langle Y_{k+1}, s_{\cdot\cdot p}\rangle = \sum_{r=1}^{M} \langle Y_{k+1}, f_r\rangle s_{r\cdot p}$, and the smoothers for $m < k+1$

$$\boxed{\tilde{\gamma}_{m,k+1}(\mathscr{O}_m^p) = MS\tilde{\gamma}_{m,k}(\mathscr{O}_m^p)Y'_{k+1}.} \tag{10.17}$$

The Process Related to the Observations

Here $H_{k+1} = \mathscr{T}_{k+1}^{ps} = \sum_{\ell=1}^{k+1} \langle X_{\ell-1}, e_p\rangle \langle Y_\ell, f_s\rangle$ and $\Delta H_{k+1} = \langle X_k, e_p\rangle \langle Y_{k+1}, f_s\rangle$. Again, substitution in (10.10) and (10.11) gives

$$\boxed{\tilde{\gamma}_{k+1,k+1}(\mathscr{T}_{k+1}^{ps}) = M\left(S\tilde{\gamma}_{k,k}(\mathscr{T}_k^{ps})Y'_{k+1} + \langle \tilde{q}_k, e_p\rangle \langle Y_{k+1}, f_s\rangle s_{s\cdot p}\right)} \tag{10.18}$$

and for $m < k+1$ we have the smoothed estimate

$$\boxed{\tilde{\gamma}_{m,k+1}(\mathscr{T}_m^{ps}) = MS\tilde{\gamma}_{m,k}(\mathscr{T}_m^{ps})Y'_{k+1}.} \tag{10.19}$$

The Joint Transition

In the dependent situation a new feature is the joint transition probabilities. Here $H_{k+1} = \mathscr{L}_{k+1}^{tqp} = \sum_{\ell=1}^{k+1} \langle Y_\ell, f_t\rangle \langle X_\ell, e_q\rangle \langle X_{\ell-1}, e_p\rangle$ and $\Delta H_{k+1} = \langle Y_{k+1}, f_t\rangle \langle X_{k+1}, e_q\rangle \times \langle X_k, e_p\rangle$. Estimates and smoothers for the joint transitions are obtained using again (10.10) and (10.11). These are:

$$\boxed{\tilde{\gamma}_{k+1,k+1}(\mathscr{L}_{k+1}^{tqp}) = M\left(S\tilde{\gamma}_{k,k}(\mathscr{L}_k^{tqp})Y'_{k+1} + \langle \tilde{q}_k, e_p\rangle \langle Y_{k+1}, f_t\rangle s_{tqp}e_q\right)} \tag{10.20}$$

and

$$\boxed{\tilde{\gamma}_{m,k+1}(\mathscr{L}_m^{tqp}) = MS\tilde{\gamma}_{m,k}(\mathscr{L}_m^{tqp})Y'_{k+1}.} \tag{10.21}$$

Parameter Estimation

Our hidden Markov model is described by the equations:

$$\begin{aligned} X_{k+1} &= AX_k + V_{k+1} \\ Y_{k+1} &= CX_k + W_{k+1} \\ Y_{k+1}X'_{k+1} &= SX_k + \Gamma_{k+1}, \qquad k \in \mathbb{N}. \end{aligned}$$

The parameters in the model are, therefore, given in a set

$$\begin{aligned} \theta = \big\{ & a_{ji}, 1 \le i, j \le N; \\ & c_{ji}, 1 \le j \le M,\ 1 \le i \le N; \\ & s_{rji}, 1 \le r \le M,\ 1 \le i, j \le N \big\}. \end{aligned}$$

These satisfy

$$\sum_{j=1}^{N} a_{ji} = 1, \qquad \sum_{j=1}^{M} c_{ji} = 1, \qquad \sum_{r=1}^{M}\sum_{j=1}^{N} s_{rji} = 1. \tag{10.22}$$

Suppose such a set θ is given and we wish to determine a new set $\hat{\theta} = \left\{(\hat{a}_{ji}(k)), (\hat{c}_{ji}(k)), (\hat{s}_{rji}(k))\right\}$ which maximizes the log-likelihood function defined below. Consider the parameters $(s_{rji}, 1 \le r \le M, 1 \le i, j \le N)$. To replace the joint transitions s_{rji} by $\hat{s}_{rji}(k)$ consider the Radon-Nikodym derivatives

$$\left.\frac{d\hat{P}}{dP}\right|_{\mathscr{G}_k} = \prod_{\ell=1}^{k}\prod_{r=1}^{M}\prod_{i,j=1}^{N}\left[\frac{\hat{s}_{rji}(k)}{s_{rji}(k)}\right]^{\langle Y_\ell, f_r\rangle\langle X_\ell, e_j\rangle\langle X_{\ell-1}, e_i\rangle}.$$

Therefore

$$E\left[\log\left.\frac{d\hat{P}}{dP}\right|_{\mathscr{G}_k} \mid \mathscr{Y}_k\right] = \sum_{r=1}^{M}\sum_{i,j=1}^{N} \hat{\mathscr{L}}_k^{rji} \log \hat{s}_{rji}(k) + \hat{R}(s), \tag{10.23}$$

where $\hat{R}(s)$ is independent of $\hat{s}$. Now observe that

$$\sum_{r=1}^{M}\sum_{j=1}^{N} \mathscr{L}_k^{rji} = \mathscr{O}_k^i. \tag{10.24}$$

Conditioning (10.24) on $\mathscr{Y}_k$ we have:

$$\sum_{r=1}^{M}\sum_{j=1}^{N} \hat{\mathscr{L}}_k^{rji} = \hat{\mathscr{O}}_k^i. \tag{10.25}$$

Now the $\hat{s}_{rji}(k)$ must also satisfy:

$$\sum_{r=1}^{M}\sum_{i=1}^{N} \hat{s}_{rji}(k) = 1. \tag{10.26}$$

We wish, therefore, to choose the $\hat{s}_{rji}(k)$ to maximize the conditional log-likelihood (10.23) subject to the constraint (10.26). Write λ for the *Lagrange multiplier* and put

$$F(\hat{s}, \lambda) = \sum_{r=1}^{M}\sum_{i,j=1}^{N} \hat{\mathscr{L}}_k^{rji} \log \hat{s}_{rji}(k) + \hat{R}(s) + \lambda\left(\sum_{r=1}^{M}\sum_{j=1}^{N} \hat{s}_{rji}(k) - 1\right).$$

Equating the derivatives of F in $\hat{s}_{rji}(k)$ and λ to zero we have that the optimum choice of $\hat{s}_{rji}(k)$ is given, when defined, by

$$\boxed{\hat{s}_{rji}(k) = \frac{\hat{\mathscr{L}}_k^{rji}}{\hat{\mathscr{O}}_k^i} = \frac{\tilde{\gamma}_k\left(\mathscr{L}_k^{rji}\right)}{\tilde{\gamma}_k\left(\mathscr{O}_k^i\right)}.} \tag{10.27}$$

Similarly, as in Section 2.7 the optimal choice for $\hat{a}_{ji}(k)$ and $\hat{c}_{ji}(k)$ given the observations are, respectively, when defined

$$\boxed{\hat{a}_{ji}(k) = \frac{\tilde{\gamma}_k\left(\mathscr{J}_k^{ij}\right)}{\tilde{\gamma}_k\left(\mathscr{O}_k^i\right)}} \tag{10.28}$$

and

$$\boxed{\hat{c}_{ji}(k) = \frac{\tilde{\gamma}_k\left(\mathscr{T}_k^{ij}\right)}{\tilde{\gamma}_k\left(\mathscr{O}_k^i\right)}.} \tag{10.29}$$

Remark 10.7 We have found recursive expressions for $\tilde{\gamma}_k(\mathscr{O}_k^i)$, $\tilde{\gamma}_k(\mathscr{L}_k^{rji})$, $\tilde{\gamma}_k(\mathscr{J}_k^{ij})$ and $\tilde{\gamma}_k(\mathscr{T}_k^{ij})$. The revised parameters $\hat{\theta} = ((\hat{a}_{ji}(k)), (\hat{c}_{ji}(k)), (\hat{s}_{rji}(k)))$, are then determined by (10.27), (10.28), and (10.29). This procedure can be iterated and an increasing sequence of likelihood ratios obtained.

■

A Test for Independence

Taking inner products with $\underline{1} \in \mathbb{R}^N$, (10.16) and (10.20) provide estimates for $\tilde{\gamma}_k\left(\mathscr{O}_k^i\right)$ and $\tilde{\gamma}_k\left(\mathscr{L}_k^{rji}\right)$, respectively; an optimal estimate for $\hat{s}_{rji}(k)$ is then obtained from (10.27). However, if the noise terms in the state X and observation Y are independent we have

$$SX_k = C \operatorname{diag} X_k A'.$$

Taking $X_k = e_i$ and considering

$$\left\langle Se_i, f_r e_j' \right\rangle = \langle Ce_i, f_r \rangle \left\langle Ae_i, e_j \right\rangle$$

we see that if the noise terms are independent:

$$s_{rji} = c_{ri} a_{ji}$$

for $1 \le r \le M$, $1 \le i, j \le N$. If the noise terms are independent $\gamma_{k,k}\left(\mathscr{J}_k^{ij}\right)$, $\gamma_{k,k}\left(\mathscr{O}_k^i\right)$, and $\gamma_{k,k}\left(\mathscr{T}_k^{ij}\right)$ are given in Section 2.6. Taking inner products with $\underline{1} \in \mathbb{R}^N$ gives estimates for $\gamma_k\left(\mathscr{J}_k^{ij}\right)$, $\gamma_k\left(\mathscr{O}_k^i\right)$, and $\gamma_k\left(\mathscr{T}_k^{ij}\right)$, and substituting in (10.28) and (10.29) gives estimates for $\hat{a}_{ji}(k)$ and $\hat{c}_{ji}(k)$. Consequently, a test for independence is to check whether

$$\hat{s}_{rji}(k) = \hat{c}_{ri}(k) \cdot \hat{a}_{ji}(k).$$

Modification of our model and this test will give other tests for independence. For example, by enlarging the state space, so the state at time k is in fact (X_{k+1}, X_k) a test can be devised to check whether either the process X_k is Markov, or (X_{k+1}, X_k) is Markov, in a hidden Markov model situation. Alternatively, models can be considered where X_{k+1} and Y_{k+1} depend also on Y_k.

2.11 Problems and Notes

Problems

1. Show that $\overline{\Lambda}_k$ defined in Section 2.3 is a $(\overline{P}, \mathscr{G}_k)$-martingale, and Λ_k defined in Section 2.7 is a $(P, \mathscr{G}_k)$-martingale.
2. Fill in the details in the proof of Theorem 5.3.
3. Write $\rho_{m,k}(e_r) = \overline{E}\left[\langle X_m, e_r\rangle \overline{\Lambda}_k \mid \mathscr{Y}_k\right]$, $\overline{\Lambda}_{m,k} = \prod_{\ell=m}^{k} \overline{\gamma}_\ell$ and $\beta_{m,k}(e_r) = \overline{E}[\overline{\Lambda}_{m+2,k} \mid X_m = e_r, \mathscr{Y}_k]$. Show that $\beta_{m,k}$ satisfies the following backward recursive equation

$$\boxed{\beta_{m,k}(e_r) = M \sum_{\ell=1}^{N} \prod_{i=1}^{M} d_{i\ell}^{Y_{m+2}^i} \beta_{m,k}(e_\ell)\, p_{r\ell}}$$

   and $\beta_{m,k}(\cdot) = \beta_{n-1,k}(\cdot) = 1$. Then verify that:

$$\boxed{\rho_{m,k}(e_r) = q_m(e_r)\, \beta_{m,k}(e_r) \prod_{i=1}^{M} d_{ir}^{Y_m^i},}$$

   where $q_m(\cdot)$ is given recursively by (4.3).
4. Prove Theorem 7.4.
5. It is pointed out in Section 2.10 that alternatively, the transitions at time k of the processes X and Y could also depend on Y_{k-1}. Describe the dynamics of this model and define a new probability measure under which the observed process Y is a sequence of i.i.d. random variables uniformly distributed.
6. Using a "double change of measure" changing both processes X and Y into i.i.d. uniform random variables, rederive the recursions of Sections 2.4 to 2.6.

Notes

Hidden Markov models, HMMs, have found applications in many areas. The survey by Rabiner (1989) describes their role in speech processing. Stratonovich (1960) describes some similar models in Stratonovich (1960). The results of Aström (1965) are obtained using Bayes' rule, and the recursion he obtained is related to Theorem 4.3.

The expectation maximization, EM, algorithm was first introduced by Baum and Petrie (1966) and further developed by Dempster et al. (1977).

Our formulation, in terms of filters which estimate the number of jumps from one state to another $\mathscr{J}$, the occupation time $\mathscr{O}$, and the $\mathscr{T}$ process, avoids use of the forward-backward algorithm and does not require so much memory. However, it requires a larger number of calculations that can be done in parallel.

Related contributions can be found in Boel (1976) and Segall (1976b). The latter discusses only a single counting observation process. Boel has considered multidimensional point processes, but has not introduced Zakai equations or the change of measure.

The continuous-time versions of these results are presented in Chapters 7 and 8.