Hidden Markov Models

Theory, Computation and Inference

Donald Zvadah

3 August 2026

Table of contents

Preface 6
What is covered . . . 6
How to read them . . . 7
Conventions . . . 7
Primary references . . . 8

I I. Foundations 9

1 Probability and linear algebra 10
1.1 Probability spaces and random elements . . . 10
1.2 Conditional expectation . . . 11
1.3 Conditional distributions and Markov kernels . . . 12
1.4 Modes of convergence . . . 13
1.5 The simplex and stochastic matrices . . . 13
1.6 Perron–Frobenius theory and the spectral gap . . . 14
1.7 Convexity, entropy and Kullback–Leibler divergence . . . 15
1.8 Summary . . . 16
1.9 Further reading . . . 17

2 Markov chains 18
2.1 Definition and basic structure . . . 18
2.2 Ergodic theory . . . 20
2.3 Quantifying the rate . . . 21
2.4 Time-inhomogeneous chains . . . 23
2.5 Summary . . . 23
2.6 Further reading . . . 24

II II. The model 25

3 The hidden Markov model 26
3.1 The two assumptions . . . 26
3.2 The model . . . 27
3.3 The factorisation . . . 28
3.4 The complete-data likelihood . . . 29
3.5 The observed-data likelihood . . . 29
3.6 Why evaluation order matters . . . 30
3.7 What the factors mean . . . 30
3.8 A worked check . . . 31
3.9 Symmetry and the shape of the likelihood . . . 31
3.10 The parameter space . . . 32
3.11 Summary . . . 33
3.12 Further reading . . . 33

4 Observation models **34**
4.1 Choosing an emission family . . . 34
4.2 Discrete . . . 34
4.3 Continuous . . . 35
4.4 Exponential family . . . 35
4.5 Covariate-dependent . . . 35
4.6 Identifiability . . . 37
4.7 Model checking . . . 38
4.8 Summary . . . 38
4.9 Further reading . . . 39

III III. Inference and estimation **40**

5 Inference: filtering, smoothing, decoding **41**
5.1 Filtering: the forward algorithm . . . 41
5.2 Smoothing: the forward–backward algorithm . . . 44
5.3 Decoding: the Viterbi algorithm . . . 46
5.4 Other quantities . . . 48
5.5 Summary . . . 48
5.6 Further reading . . . 48

6 Maximum likelihood and EM **50**
6.1 Maximum likelihood . . . 50
6.2 The EM algorithm . . . 51
6.3 Standard errors . . . 54
6.4 Summary . . . 55
6.5 Further reading . . . 55

7 Computation and numerical practice **56**
7.1 Underflow and the two remedies . . . 56
7.2 Complexity . . . 57
7.3 Vectorisation and hardware . . . 58
7.4 Approximate inference . . . 59
7.5 Implementation checks . . . 59
7.6 Summary . . . 60
7.7 Further reading . . . 60

8 Asymptotic theory **61**
8.1 The setting . . . 61
8.2 The limiting log-likelihood . . . 62
8.3 Consistency . . . 63
8.4 Asymptotic normality . . . 63
8.5 Misspecification . . . 64
8.6 Choosing the number of states . . . 65
8.7 Summary . . . 66
8.8 Further reading . . . 66

IV IV. Bayesian and hierarchical inference **67**

9 Bayesian inference **68**
9.1 Priors . . . 68

9.2 Posterior simulation . . . 69
9.3 Label switching . . . 71
9.4 Marginal likelihood . . . 72
9.5 Summary . . . 72
9.6 Further reading . . . 73

10 Hierarchical models **74**
10.1 Hierarchical state spaces . . . 74
10.2 Multilevel and mixed HMMs . . . 75
10.3 Hierarchical priors within a single HMM . . . 77
10.4 Summary . . . 79
10.5 Further reading . . . 79

11 The telescoping sampler **80**
11.1 Mixtures of finite mixtures . . . 81
11.2 The key computation . . . 82
11.3 The algorithm . . . 83
11.4 Telescoping for HMMs . . . 84
11.5 Where this fits . . . 86
11.6 Summary . . . 87
11.7 Further reading . . . 88

V V. Extensions **89**

12 State-space and switching models **90**
12.1 General state-space models . . . 90
12.2 The Kalman filter . . . 91
12.3 Sequential Monte Carlo . . . 92
12.4 Switching state-space models . . . 92
12.5 Other extensions . . . 93
12.6 Summary . . . 93
12.7 Further reading . . . 94

13 Online prediction and decision problems **95**
13.1 Prediction without a model . . . 95
13.2 HMMs as partially observable decision problems . . . 97
13.3 Summary . . . 98
13.4 Further reading . . . 98

VI VI. Applications and exercises **99**

14 Applications **100**
14.1 Speech recognition . . . 100
14.2 Bioinformatics . . . 101
14.3 Finance and econometrics . . . 101
14.4 Animal movement and ecology . . . 102
14.5 Epidemiology and disease progression . . . 102
14.6 A checklist . . . 103
14.7 Summary . . . 104
14.8 Further reading . . . 104

15 Exercises **105**
15.1 Probability and Markov chains . . . 105
15.2 Model structure and likelihood . . . 106
15.3 Inference algorithms . . . 106
15.4 Estimation . . . 107
15.5 Asymptotics . . . 108
15.6 Bayesian, hierarchical and telescoping . . . 109
15.7 Extensions and alternative perspectives . . . 110
15.8 Using these exercises . . . 111

Appendices **112**

A Notation **112**
A.1 Model . . . 112
A.2 Inference . . . 112
A.3 Bayesian and hierarchical . . . 113
A.4 General . . . 113
A.5 Conventions . . . 113

B Reading guide **114**
B.1 Dependencies . . . 116
B.2 Routes through the material . . . 117
B.3 A twelve-week course . . . 118
B.4 Prerequisites . . . 118

About **119**
How they are written . . . 119
What is unusual here . . . 119

FAQ **120**
What background do I need? . . . 120
Is there code? . . . 120
Where should I start? . . . 120
How do I choose the number of states? . . . 120
Why does my EM fit give different answers each run? . . . 121
My Bayesian fit gives identical posteriors for every state. Why? . . . 121
Should I use EM or MCMC? . . . 121
What if my data are several short series rather than one long one? . . . 121
Are there solutions to the exercises? . . . 121
Can I use these for teaching? . . . 122
How long does the course take? . . . 122

Corrections and contact **123**
Reporting an error . . . 123
Known limitations . . . 123
Reuse . . . 124

Bibliography **125**

Preface

These are lecture notes on hidden Markov models, written at the level of a graduate course in statistics. They cover the model, the algorithms, the asymptotic theory, and the Bayesian and hierarchical extensions, with derivations given in full rather than cited.

The organising idea is that an HMM is a statement about conditional independence, and that almost every result follows from rewriting a conditional expectation. The forward algorithm, the Viterbi recursion, the EM updates, and the samplers of Part IV are all consequences of a single factorisation, and the notes try to make that visible rather than presenting them as separate inventions.

What is covered

I. Foundations

Conditional expectation and Markov kernels, spectral theory for stochastic matrices, convexity and KL divergence. Finite-state Markov chains: ergodicity, mixing, and the contraction results that reappear as filter stability.

II. The model

The two conditional independence assumptions, the factorisation they force, and the likelihood — as a sum over K^T sequences and as a product of T matrices. Then emission families, and identifiability from three consecutive observations.

III. Inference and estimation

Filtering, smoothing and decoding with correctness proofs. Maximum likelihood and EM as coordinate ascent on a lower bound. Numerical practice and complexity. Consistency and asymptotic normality.

IV. Bayesian and hierarchical inference

Priors, data augmentation, and forward filtering with backward sampling. Hierarchical state spaces, multilevel models across series, and hierarchical priors. The telescoping sampler for an unknown number of states.

V. Extensions

State-space models and the Kalman filter as a continuous-state forward algorithm. Switching models and why they are intractable. Regret bounds, and HMMs as partially observable decision problems.

VI. Applications and exercises

Speech, genomics, finance, animal movement and epidemiology, each analysed through the same four modelling questions. Proof-based exercises grouped by chapter, with a suggested twelve-week schedule.

How to read them

Each chapter opens with what you will be able to do by the end of it, and closes with a short self-check whose answers are hidden until you want them. The chapters are ordered so that each depends only on those before it, and reading straight through works.

Three shorter routes:

A first course on the algorithms. Chapters 3 to 7. Enough to specify, fit and check an HMM; skips the measure theory and the asymptotics.

Theory. Chapters 1, 2, 3, 5 and 8 — a self-contained development from conditional expectation to asymptotic normality of the maximum likelihood estimator.

Bayesian methods. Chapter 9, Chapter 10 and Chapter 11, assuming Chapter 5 for the recursions. Chapter 11 is self-contained given Chapter 9 and can be read on its own by anyone interested in the unknown-K problem.

Appendix B sets out the dependencies in more detail, and Appendix A collects the symbols.

Conventions

Notation follows Zucchini et al. (2016) wherever possible, so that formulas here match theirs term for term: δ for the initial distribution, Γ for the transition matrix, f_i for the state-dependent densities. Where a result is stated without proof, a reference is given; where a proof is given, it is complete unless explicitly labelled a sketch.

Numbered definitions, theorems and examples are cross-referenced throughout and can be clicked. Citations are author–date and link to the bibliography.

Emphasis throughout is on which assumptions are doing the work. The conditions under which a theorem holds are usually more useful than the theorem, since applied work spends most of its time in situations where one of them fails.

Primary references

Four books cover most of what follows, and the notes are written to be read alongside them.

- Zucchini et al. (2016) — the standard applied treatment of finite-state HMMs, with R code and a wide range of worked examples. The source of the notation used here.
- Cappé et al. (2005) — the rigorous reference for inference and asymptotic theory, in general state spaces.
- Frühwirth-Schnatter (2006) — the reference for the Bayesian and mixture-model side, and the background for Part IV.
- Douc et al. (2014) — HMMs placed within the broader theory of nonlinear time series.

Full details are in the bibliography.

Part I

I. Foundations

1 Probability and linear algebra

A hidden Markov model is a statement about conditional independence, and almost every result in these notes is obtained by rewriting a conditional expectation. This chapter fixes the language for doing that. It is a reference chapter rather than a course in measure theory: the aim is to make later arguments precise, not to develop probability from the axioms.

Three bodies of material get used repeatedly. Conditional expectation and conditional distributions justify writing things like $\mathbb{P}(S_t = i \mid Y_{1:T} = y_{1:T})$ and manipulating them. Spectral theory for non-negative matrices controls how fast a Markov chain forgets its starting point, which is what makes both the ergodic theory of Chapter 8 and the filter stability results of Chapter 5 work. Convex analysis on the simplex explains why the EM algorithm of Chapter 6 increases the likelihood at every step.

By the end of this chapter you will be able to

- state what conditional expectation is and use the tower property to manipulate it;
- explain why conditioning on $\{Y_{1:T} = y_{1:T}\}$ makes sense even when that event has probability zero;
- read off the stationary distribution and convergence rate of a stochastic matrix from its spectrum;
- recognise Kullback–Leibler divergence as a Bregman divergence and use its non-negativity.

> **Prerequisites**
>
> Undergraduate probability (random variables, conditional probability, the law of large numbers), linear algebra through eigenvalues and eigenvectors, and calculus. Measure theory is used but not assumed; every measure-theoretic object here is introduced where it is needed. Williams (1991) and Kallenberg (2021) are the standard places to fill gaps.

1.1 Probability spaces and random elements

1.1.1 The underlying space

Definition 1.1 (Probability space)**.** A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ in which Ω is a set of outcomes, $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra (it contains Ω and is closed under complements and countable unions), and $\mathbb{P}\colon \mathcal{F} \to [0,1]$ is countably additive with $\mathbb{P}(\Omega) = 1$.

The σ-algebra is the set of statements to which the model assigns a probability. In a finite-state HMM the hidden states take values in $E = \{1, \dots, K\}$ with the discrete σ-algebra, so every subset is measurable and nothing can go wrong. The formalism still earns its place, for two reasons.

First, an HMM is a statement about an infinite sequence $(S_t, Y_t)_{t\geq 1}$, and the law of that sequence lives on a product space. For a measurable space $(S, \mathcal{S})$, the countable product $(S^{\mathbb{N}}, \mathcal{S}^{\otimes\mathbb{N}})$ carries the smallest σ-algebra making every coordinate projection measurable. Kolmogorov's extension theorem says that a consistent family of finite-dimensional distributions — exactly what Chapter 3 writes down — determines a unique measure on that product space. Without it, "the process $(S_t)_{t\geq 1}$" would be an abuse of notation.

Second, the observation space is usually not discrete. Once Y_t takes values in $\mathbb{R}^d$, conditioning on $\{Y_{1:T} = y_{1:T}\}$ conditions on an event of probability zero, and the naive definition of conditional probability divides by zero. Sections 1.2 and 1.3 repair this.

1.1.2 Random variables and their laws

Definition 1.2 (Random element)**.** Let $(S, \mathcal{S})$ be a measurable space. A **random element** of S is a measurable map $X\colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$. Its **law** is the pushforward measure $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ for $A \in \mathcal{S}$.

When $S = \mathbb{R}$ we say random variable, and when $S = \mathbb{R}^d$ random vector. Throughout these notes S_t denotes a random element of the finite state space E and Y_t a random element of the observation space $\mathcal{Y}$.

Two pieces of notation recur. For $1 \leq s \leq t$ we write $Y_{s:t} = (Y_s, Y_{s+1}, \dots, Y_t)$ for a contiguous block, and we write f_i for the density of the observation distribution in state i with respect to a fixed dominating measure. Taking counting measure recovers discrete emissions and Lebesgue measure recovers continuous ones, so a single set of formulas covers both. The full notation table is in Appendix A.

1.2 Conditional expectation

Conditioning on a σ-algebra rather than on an event is what makes the zero-probability problem disappear.

Definition 1.3 (Conditional expectation)**.** Let X be integrable and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. A **conditional expectation** of X given $\mathcal{G}$ is any $\mathcal{G}$-measurable integrable random variable Z satisfying

$$\int_G Z \, d\mathbb{P} = \int_G X \, d\mathbb{P} \qquad \text{for every } G \in \mathcal{G}.$$

We write $Z = \mathbb{E}[X \mid \mathcal{G}]$.

Existence follows from the Radon–Nikodym theorem, and any two versions agree almost surely, so $\mathbb{E}[X \mid \mathcal{G}]$ is well defined up to null sets. Read the defining property as: $\mathbb{E}[X \mid \mathcal{G}]$ is the best guess at X using only information in $\mathcal{G}$, in the sense that it has the same integral as X over every $\mathcal{G}$-event. When X is square integrable this is literally an orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

The following properties are used so often in later derivations that it is worth naming them once.

Proposition 1.1 (Rules for conditional expectation)**.** *Let X, Y be integrable and $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$.*

1. Linearity. $\mathbb{E}[aX + bY \mid \mathcal{G}] = a\,\mathbb{E}[X \mid \mathcal{G}] + b\,\mathbb{E}[Y \mid \mathcal{G}]$ *almost surely.*

2. Tower property. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$ *almost surely.*
3. Taking out what is known. *If Z is $\mathcal{G}$-measurable and ZX is integrable, then* $\mathbb{E}[ZX \mid \mathcal{G}] = Z\,\mathbb{E}[X \mid \mathcal{G}]$ *almost surely.*
4. Independence. *If $\sigma(X)$ is independent of $\mathcal{G}$ then* $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ *almost surely.*

Proofs are in Williams (1991, ch. 9). Every recursion in Chapter 5 is an application of the tower property combined with the conditional independence structure of the model: the forward recursion conditions on S_t and then integrates it out, the backward recursion does the same in reverse time.

Filtering and smoothing are conditional expectations. Writing $\mathcal{Y}_t = \sigma(Y_1, \dots, Y_t)$ for the σ-algebra generated by the observations up to time t, the filtering probability is

$$\mathbb{P}(S_t = i \mid \mathcal{Y}_t) = \mathbb{E}[\mathbf{1}\{S_t = i\} \mid \mathcal{Y}_t],$$

and smoothing replaces $\mathcal{Y}_t$ by $\mathcal{Y}_T$. The forward and backward algorithms are nothing more than efficient ways of evaluating these objects.

1.3 Conditional distributions and Markov kernels

Conditional expectation gives a number for each fixed function. To speak of *the conditional distribution* of one variable given another, we need those numbers to fit together into a probability measure.

Definition 1.4 (Markov kernel)**.** Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces. A **Markov kernel** from S to T is a map $Q \colon S \times \mathcal{T} \to [0, 1]$ such that $Q(x, \cdot)$ is a probability measure on $\mathcal{T}$ for each $x \in S$, and $x \mapsto Q(x, B)$ is $\mathcal{S}$-measurable for each $B \in \mathcal{T}$.

Definition 1.5 (Regular conditional distribution)**.** Let X and Y be random elements of $(S, \mathcal{S})$ and $(T, \mathcal{T})$. A **regular conditional distribution** of X given Y is a Markov kernel Q from T to S with

$$\mathbb{P}(X \in A,\, Y \in B) = \int_B Q(y, A)\, \mathbb{P}_Y(dy) \qquad \text{for all } A \in \mathcal{S},\ B \in \mathcal{T}.$$

Theorem 1.1 (Existence on standard Borel spaces)**.** *If $(S, \mathcal{S})$ is a standard Borel space — a Polish space with its Borel σ-algebra — then a regular conditional distribution of X given Y exists and is unique up to $\mathbb{P}_Y$-null sets.*

See Kallenberg (2021, ch. 8) for the proof. Every state and observation space in these notes is standard Borel: finite sets, $\mathbb{R}^d$, and countable products of these all qualify. So expressions such as $\mathbb{P}(S_t = i \mid Y_{1:T} = y_{1:T})$ denote a genuine object rather than a formal manipulation, and the freedom to choose a version explains the "for almost every y" qualifiers that appear in careful statements of the smoothing recursions.

Kernels also give the cleanest statement of the model itself. A finite-state HMM is a Markov kernel Γ from E to E together with a kernel from E to $\mathcal{Y}$; when E is finite the first is a matrix, and Chapter 12 replaces it by a genuine kernel to get continuous-state models.

1.4 Modes of convergence

Asymptotic statements in Chapter 8 distinguish three notions.

Definition 1.6 (Almost sure, in probability, in L^p)**.** For random variables X_n, X on a common space:

- $X_n \to X$ **almost surely** if $\mathbb{P}(X_n \to X) = 1$;
- $X_n \to X$ **in probability** if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for every $\varepsilon > 0$;
- $X_n \to X$ **in** L^p $(p \geq 1)$ if $\mathbb{E}|X_n - X|^p \to 0$.

Almost sure convergence and L^p convergence each imply convergence in probability, which in turn implies convergence in distribution; no other implication holds in general. The relevance to HMMs is direct. Consistency of the maximum likelihood estimator is an almost sure statement about the normalised log-likelihood $T^{-1}\ell_T(\theta)$, proved by an ergodic theorem. Asymptotic normality is a statement about convergence in distribution of $T^{-1/2}$ times the score, proved by a martingale central limit theorem. The two arguments have different shapes and are kept apart in Chapter 8.

1.5 The simplex and stochastic matrices

1.5.1 Probability vectors

Definition 1.7 (Probability simplex)**.** For $K \geq 1$, the **probability simplex** is

$$\Delta^{K-1} = \Big\{ \mu = (\mu_1, \dots, \mu_K)^\top \in \mathbb{R}^K : \mu_i \geq 0,\ \textstyle\sum_{i=1}^K \mu_i = 1 \Big\}.$$

It is a compact convex set of dimension $K - 1$. Two distances on it matter. The ℓ^1 distance $\|\mu - \nu\|_1 = \sum_i |\mu_i - \nu_i|$ equals twice the total variation distance and is the natural metric for mixing results. The ℓ^2 distance is more convenient when spectral decompositions are in play. On a finite space these are equivalent up to a factor of $\sqrt{K}$, so the choice is one of convenience rather than substance.

1.5.2 Stochastic matrices

Definition 1.8 (Row-stochastic matrix)**.** A matrix $\Gamma = (\gamma_{ij}) \in \mathbb{R}^{K \times K}$ is **row-stochastic** if $\gamma_{ij} \geq 0$ for all i, j and $\sum_{j=1}^K \gamma_{ij} = 1$ for every i.

Equivalently $\Gamma \mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the all-ones column vector. Row-stochastic matrices are exactly the Markov kernels on a K-point set: γ_{ij} is the probability of moving to j from i. If $\mu \in \Delta^{K-1}$ is the law of S_t, then $\mu^\top \Gamma$ is the law of S_{t+1}, so the matrix acts on row vectors from the right. This convention follows Zucchini et al. (2016) and is used throughout.

Products of row-stochastic matrices are row-stochastic, which is why the n-step transition matrix in Chapter 2 is simply Γ^n.

1.6 Perron–Frobenius theory and the spectral gap

How fast a Markov chain forgets its initial condition is a spectral question about Γ.

Definition 1.9 (Irreducible and primitive matrices)**.** A non-negative matrix $A \in \mathbb{R}^{K\times K}$ is **irreducible** if for every pair (i, j) there is $n \geq 1$ with $(A^n)_{ij} > 0$. It is **primitive** if there is a single $n \geq 1$ with $(A^n)_{ij} > 0$ for all i, j simultaneously.

Theorem 1.2 (Perron–Frobenius)**.** *Let A be a non-negative irreducible $K \times K$ matrix with spectral radius $\rho(A)$. Then $\rho(A)$ is a positive eigenvalue of A; it is algebraically simple; and it admits left and right eigenvectors with strictly positive entries. Up to scaling these are the only non-negative eigenvectors of A. If A is primitive, every other eigenvalue λ satisfies $|\lambda| < \rho(A)$.*

Horn and Johnson (2013, ch. 8) and Seneta (2006) give proofs. Applied to a row-stochastic Γ, the identity $\Gamma\mathbf{1} = \mathbf{1}$ shows 1 is an eigenvalue with positive right eigenvector, so $\rho(\Gamma) = 1$. Irreducibility then makes it simple, and the corresponding left eigenvector normalised to sum to one is the unique stationary distribution π:

$$\pi^\top \Gamma = \pi^\top, \qquad \pi \in \Delta^{K-1}.$$

Primitivity — equivalently, irreducibility together with aperiodicity in the language of Chapter 2 — pushes every other eigenvalue strictly inside the unit disc, and that gap is what drives convergence.

Definition 1.10 (Spectral gap)**.** Order the eigenvalues of Γ so that $1 = \lambda_1$ and $|\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_K|$. The **absolute spectral gap** is $\varrho = 1 - |\lambda_2|$.

Proposition 1.2 (Geometric convergence)**.** *Suppose Γ is primitive and diagonalisable with stationary distribution π. There is a constant C, depending on the eigenvector matrix but not on t, such that for every $\mu_0 \in \Delta^{K-1}$,*

$$\left\| \mu_0^\top \Gamma^t - \pi^\top \right\|_2 \;\leq\; C\,|\lambda_2|^t.$$

Proof. Write the spectral decomposition $\Gamma = \sum_{k=1}^{K} \lambda_k\, v_k u_k^\top$ with $u_k^\top v_l = \delta_{kl}$, ordered so that $\lambda_1 = 1$, $v_1 = \mathbf{1}$ and $u_1 = \pi$. Then $\Gamma^t = \mathbf{1}\pi^\top + \sum_{k\geq 2} \lambda_k^t\, v_k u_k^\top$. Since $\mu_0^\top \mathbf{1} = 1$, the leading term contributes exactly $\pi^\top$, and the remainder is bounded in norm by $|\lambda_2|^t \sum_{k\geq 2} \|v_k\|_2 \|u_k\|_2$. □

Diagonalisability is a convenience, not a necessity; the Jordan form gives the same conclusion with an extra polynomial factor. The reversible case in Chapter 2 is cleaner still, because self-adjointness in $L^2(\pi)$ makes the spectrum real and the eigenvector basis orthonormal, so C can be computed explicitly.

1.7 Convexity, entropy and Kullback–Leibler divergence

1.7.1 Convex functions on the simplex

Definition 1.11 (Convexity)**.** A function $f\colon \Delta^{K-1} \to \mathbb{R}$ is **convex** if $f(\theta\mu + (1-\theta)\nu) \le \theta f(\mu) + (1-\theta) f(\nu)$ for all $\mu, \nu \in \Delta^{K-1}$ and $\theta \in [0,1]$, and **strictly convex** if the inequality is strict whenever $\mu \neq \nu$ and $\theta \in (0,1)$.

Two functionals do most of the work. The Shannon entropy $H(\mu) = -\sum_i \mu_i \log \mu_i$ is strictly concave on Δ^{K-1}, and the negative entropy $\phi(\mu) = \sum_i \mu_i \log \mu_i$ is therefore strictly convex.

1.7.2 Kullback–Leibler divergence

Definition 1.12 (Kullback–Leibler divergence)**.** For $p, q \in \Delta^{K-1}$ with $q_i > 0$ whenever $p_i > 0$,

$$\mathrm{KL}(p \| q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i},$$

with the convention $0 \log 0 = 0$. For measures $P \ll Q$ on a general space, $\mathrm{KL}(P\|Q) = \int \log \frac{dP}{dQ}\, dP$.

Proposition 1.3 (KL divergence is a Bregman divergence)**.** *Let $\phi(p) = \sum_i p_i \log p_i$ on the relative interior of Δ^{K-1} and define the Bregman divergence*

$$D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla\phi(q),\, p - q \rangle.$$

Then $D_\phi(p,q) = \mathrm{KL}(p\|q)$. Consequently $\mathrm{KL}(p\|q) \ge 0$, with equality if and only if $p = q$.

Proof. Since $\partial\phi/\partial q_i = \log q_i + 1$,

$$D_\phi(p,q) = \sum_i p_i \log p_i - \sum_i q_i \log q_i - \sum_i (\log q_i + 1)(p_i - q_i).$$

The terms $\sum_i (p_i - q_i)$ vanish because both vectors sum to one, and the remaining terms collapse to $\sum_i p_i \log p_i - \sum_i p_i \log q_i = \mathrm{KL}(p\|q)$. Non-negativity and the equality case follow from strict convexity of ϕ, since a Bregman divergence is the gap between a strictly convex function and its tangent plane. $\square$

Non-negativity of KL divergence is the single fact behind the monotonicity of EM (Chapter 6), the variational bound of Chapter 7, and the identification of the pseudo-true parameter under misspecification (Chapter 8). Note that KL is not symmetric and does not satisfy a triangle inequality, so it is not a metric; the asymmetry is exactly why forward and reverse KL give different variational approximations.

1.7.3 Exponential families and convex duality

Definition 1.13 (Exponential family)**.** A family of densities with respect to a dominating measure ν is an **exponential family in natural form** if

$$f(y;\eta) = h(y)\exp\{\langle \eta,\, T(y)\rangle - A(\eta)\}, \qquad A(\eta) = \log \int h(y) e^{\langle \eta, T(y)\rangle}\, \nu(dy),$$

where T is the sufficient statistic, η the natural parameter, and A the log-partition function.

The log-partition function is convex, and on the interior of the natural parameter space it is smooth with $\nabla A(\eta) = \mathbb{E}_\eta[T(Y)]$ and $\nabla^2 A(\eta) = \mathrm{Var}_\eta(T(Y))$. Its convex conjugate

$$A^*(m) = \sup_\eta \{\langle \eta, m\rangle - A(\eta)\}$$

is the negative entropy expressed in mean parameters, which is the general form of the pairing seen above for the finite simplex, where $\phi(p) = \sum_i p_i \log p_i$ and $\phi^*(\eta) = \log \sum_i e^{\eta_i}$.

Two consequences shape later chapters. Because ∇A maps natural to mean parameters bijectively on the interior, maximising an exponential-family likelihood amounts to moment matching, and the M-step of EM in Chapter 6 reduces to setting expected sufficient statistics equal to weighted empirical ones. Because A and A^* are conjugate, entropy-regularised problems of the form

$$\max_q \{\mathbb{E}_q[\log p(Y, S)] + H(q)\}$$

have exponential-family solutions, which is the starting point for the variational treatment in Wainwright and Jordan (2008). The log-sum-exp expressions that appear all over the numerical implementations of Chapter 7 are the finite-dimensional instance of A.

Check your understanding. Why does the theory need regular conditional distributions rather than just conditional expectations, given that everything in a finite-state HMM is discrete?

Show answer

The *states* are discrete but the *observations* usually are not. Writing $\mathbb{P}(S_t = i \mid Y_{1:T} = y_{1:T})$ conditions on an event of probability zero whenever Y has a continuous distribution, so the elementary definition divides by zero.
Conditional expectation given the σ-algebra $\sigma(Y_{1:T})$ is well defined, but it produces a number for each fixed function, not a probability measure. Theorem 1.1 supplies the measure, and it is what makes the smoothing distribution a genuine object rather than a formal manipulation. The "for almost every y" qualifiers in careful statements of the recursions are the price.

1.8 Summary

The chapter has assembled four things:

- conditional expectation and regular conditional distributions, which make filtering and smoothing well-defined objects rather than notation;
- Markov kernels, which give a single description of the transition and emission mechanisms that survives the move from finite to continuous state spaces;

- Perron–Frobenius theory and the spectral gap, which quantify how fast a chain forgets its initial distribution;
- convexity, KL divergence and exponential families, which underlie EM, variational bounds, and the identification arguments in the asymptotic theory.

1.9 Further reading

Williams (1991) is the shortest rigorous route to conditional expectation and martingales; Kallenberg (2021, ch. 8) covers regular conditional distributions and kernels in full generality. For non-negative matrices, Seneta (2006) is written with Markov chains in mind and Horn and Johnson (2013, ch. 8) gives the general spectral theory. Boyd and Vandenberghe (2004) covers convex duality at the level used here, and Amari (2016) develops the information-geometric picture of exponential families. Cappé et al. (2005, ch. 1) sets up HMMs in exactly this measure-theoretic language and is the natural companion to this chapter.

2 Markov chains

The hidden layer of an HMM is a Markov chain, and most of what can be said about an HMM's long-run behaviour comes from what can be said about that chain. This chapter develops the finite-state theory to the level the rest of the notes needs: enough to state the ergodic theorem used for consistency in Chapter 8, and enough spectral machinery to quantify how quickly the chain forgets where it started.

Zucchini et al. (2016, ch. 1) covers the same ground at an applied level. The treatment here adds proofs and keeps the notation compatible.

By the end of this chapter you will be able to

- check whether a transition matrix is irreducible and aperiodic, and say what each buys;
- state the ergodic theorem and explain why it needs irreducibility but not aperiodicity;
- bound the distance to stationarity using either the spectral gap or Dobrushin's coefficient, and say when each applies;
- explain why a time-inhomogeneous chain can still forget its initial condition.

> **Where this chapter leads**
>
> The ergodic theorem (Theorem 2.2) turns time averages into expectations, and it is the engine behind consistency of the maximum likelihood estimator. The contraction results (Theorem 2.3, Proposition 2.5) reappear as filter stability: an HMM filter initialised at the wrong distribution recovers at a geometric rate, which is why misspecifying δ costs so little in practice.

2.1 Definition and basic structure

2.1.1 The Markov property

Definition 2.1 (Time-homogeneous Markov chain)**.** Let $E = \{1, \dots, K\}$. A process $(S_t)_{t \geq 1}$ with values in E is a **time-homogeneous Markov chain** with transition matrix $\Gamma = (\gamma_{ij})$ and initial distribution $\delta \in \Delta^{K-1}$ if $\mathbb{P}(S_1 = i) = \delta_i$ and, for every $t \geq 1$ and every $i_1, \dots, i_t, j$ with $\mathbb{P}(S_{1:t} = i_{1:t}) > 0$,

$$\mathbb{P}(S_{t+1} = j \mid S_{1:t} = i_{1:t}) = \mathbb{P}(S_{t+1} = j \mid S_t = i_t) = \gamma_{i_t j}.$$

The condition says the future depends on the past only through the present. It has an equivalent product form that is often easier to work with.

Proposition 2.1 (Finite-dimensional distributions)**.** *(S_t) is a Markov chain with parameters (δ, Γ) if and only if, for every $T \geq 1$ and every $i_{1:T} \in E^T$,*

$$\mathbb{P}(S_{1:T} = i_{1:T}) = \delta_{i_1} \prod_{t=1}^{T-1} \gamma_{i_t i_{t+1}}.$$

Proof. For the forward direction, apply the chain rule $\mathbb{P}(S_{1:T} = i_{1:T}) = \mathbb{P}(S_1 = i_1) \prod_{t=1}^{T-1} \mathbb{P}(S_{t+1} = i_{t+1} \mid S_{1:t} = i_{1:t})$ and replace each conditional factor using Definition 2.1. For the converse, sum the product expression over i_T to check consistency, then form the ratio of the expressions for $\mathbb{P}(S_{1:t+1})$ and $\mathbb{P}(S_{1:t})$; every factor cancels except $\gamma_{i_t i_{t+1}}$, which does not depend on $i_{1:t-1}$. □

This factorisation is the hidden-state half of the HMM likelihood in Chapter 3. The observation half multiplies in one emission factor per time point.

2.1.2 Chapman–Kolmogorov

Write $\gamma_{ij}^{(n)} = \mathbb{P}(S_{t+n} = j \mid S_t = i)$ for the n-step transition probabilities, which by homogeneity do not depend on t.

Proposition 2.2 (Chapman–Kolmogorov equations)**.** *For all* $m, n \geq 0$, $\Gamma^{(m+n)} = \Gamma^{(m)}\Gamma^{(n)}$, *and consequently* $\Gamma^{(n)} = \Gamma^n$.

Proof. Condition on the state at the intermediate time and use the Markov property:

$$\gamma_{ij}^{(m+n)} = \sum_{k \in E} \mathbb{P}(S_{m+n} = j \mid S_m = k)\, \mathbb{P}(S_m = k \mid S_0 = i) = \sum_k \gamma_{ik}^{(m)} \gamma_{kj}^{(n)}.$$

The second claim follows by induction from $\Gamma^{(1)} = \Gamma$. □

So the entire temporal structure of the chain is encoded in the powers of a single matrix, and the spectral theory of Chapter 1 applies directly.

2.1.3 Stationary distributions

Definition 2.2 (Stationary distribution)**.** A vector $\pi \in \Delta^{K-1}$ is a **stationary distribution** for Γ if $\pi^\top \Gamma = \pi^\top$.

If $S_1 \sim \pi$ then $S_t \sim \pi$ for every t, and the whole process is strictly stationary. Existence on a finite state space is automatic: Γ maps the compact convex set Δ^{K-1} into itself continuously, so Brouwer's fixed point theorem supplies a fixed vector. Uniqueness needs irreducibility, which we come to next.

2.1.4 Reversibility

Definition 2.3 (Reversibility and detailed balance)**.** A chain with transition matrix Γ and stationary distribution π is **reversible** if

$$\pi_i \gamma_{ij} = \pi_j \gamma_{ji} \qquad \text{for all } i, j \in E.$$

Detailed balance says the equilibrium flow from i to j matches the flow from j to i. Summing over i recovers stationarity, so detailed balance is the stronger condition. Its value is spectral.

Proposition 2.3 (Self-adjointness under reversibility)**.** *Equip $\mathbb{R}^E$ with the inner product $\langle f, g\rangle_\pi = \sum_i \pi_i f(i)g(i)$ and let $(\Gamma f)(i) = \sum_j \gamma_{ij} f(j)$. If the chain is reversible with $\pi_i > 0$ for all i, then Γ is self-adjoint on $L^2(\pi)$, so its eigenvalues are real and lie in $[-1, 1]$ and it admits a π-orthonormal eigenbasis.*

Proof. Expand and apply detailed balance:

$$\langle f, \Gamma g\rangle_\pi = \sum_{i,j} \pi_i \gamma_{ij} f(i)g(j) = \sum_{i,j} \pi_j \gamma_{ji} f(i)g(j) = \langle \Gamma f, g\rangle_\pi.$$

The spectral theorem for self-adjoint operators on a finite-dimensional inner product space gives the rest; the eigenvalue bound follows because $\|\Gamma\|_\infty = 1$. □

Hidden chains in HMMs are not assumed reversible, and estimated transition matrices rarely satisfy detailed balance. Reversible chains remain useful as a source of examples where mixing rates can be computed exactly.

2.2 Ergodic theory

2.2.1 Communication and irreducibility

Write $i \rightsquigarrow j$ when $\gamma_{ij}^{(n)} > 0$ for some $n \geq 0$, and $i \leftrightarrow j$ when both $i \rightsquigarrow j$ and $j \rightsquigarrow i$. The relation $\leftrightarrow$ is an equivalence relation, and its classes are the **communicating classes** of the chain.

Definition 2.4 (Irreducibility)**.** A chain is **irreducible** if E is a single communicating class, that is, every state can be reached from every other.

For an HMM, irreducibility of the hidden chain is a modelling assumption with real content. It rules out absorbing states and transient start-up states, guarantees a unique stationary distribution, and is a standing hypothesis in the identifiability and consistency results of Chapters 6 and 8.

2.2.2 Periodicity

Definition 2.5 (Period)**.** The **period** of $i \in E$ is $d(i) = \gcd\{n \geq 1 : \gamma_{ii}^{(n)} > 0\}$. A chain is **aperiodic** if $d(i) = 1$ for some, hence every, state.

Proposition 2.4 (Period is a class property)**.** *If $i \leftrightarrow j$ then $d(i) = d(j)$.*

Proof. Choose m, n with $\gamma_{ij}^{(m)} > 0$ and $\gamma_{ji}^{(n)} > 0$, so $\gamma_{ii}^{(m+n)} > 0$ and $d(i)$ divides $m + n$. For any r with $\gamma_{jj}^{(r)} > 0$ we get $\gamma_{ii}^{(m+r+n)} > 0$, so $d(i)$ divides $m + r + n$ and hence divides r. Therefore $d(i)$ divides $d(j)$, and the reverse follows by symmetry. □

A chain with period $d > 1$ cycles deterministically between d subsets of the state space, and $\mathbb{P}(S_t = \cdot)$ never settles down. Aperiodicity excludes that. Together, irreducibility and aperiodicity are exactly primitivity of Γ in the sense of Definition 1.9, which connects the probabilistic conditions to the Perron–Frobenius theorem.

2.2.3 Convergence to equilibrium

Theorem 2.1 (Convergence theorem)**.** *Let Γ be irreducible and aperiodic. Then Γ has a unique stationary distribution π, it satisfies $\pi_i > 0$ for all i, and for every initial distribution δ,*

$$\lim_{t\to\infty} \left\| \delta^\top \Gamma^t - \pi^\top \right\|_{\mathrm{TV}} = 0,$$

where $\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sum_i |\mu_i - \nu_i|$. The convergence is geometric.

Irreducibility plus aperiodicity makes Γ primitive, so Theorem 1.2 gives a simple eigenvalue 1 with strictly positive left eigenvector, every other eigenvalue strictly smaller in modulus, and the geometric rate of Proposition 1.2. Levin and Peres (2017, ch. 4) gives a coupling proof that avoids diagonalisability altogether.

2.2.4 The ergodic theorem

Theorem 2.2 (Ergodic theorem for finite chains)**.** *Let (S_t) be irreducible with stationary distribution π, started from any initial distribution. For every $f\colon E \to \mathbb{R}$,*

$$\frac{1}{T}\sum_{t=1}^{T} f(S_t) \xrightarrow[T\to\infty]{a.s.} \mathbb{E}_\pi[f(S)] = \sum_{i\in E} \pi_i f(i).$$

Proof. Fix a state i and let $\tau_1 < \tau_2 < \dots$ be the successive return times to i. By irreducibility on a finite state space these are almost surely finite with finite mean, and the strong Markov property makes the excursion blocks between consecutive returns independent and identically distributed. Applying the strong law of large numbers to the block sums of f and separately to the block lengths, the ratio converges almost surely to $\mathbb{E}_i[\sum_{t<\tau_1} f(S_t)]/\mathbb{E}_i[\tau_1]$, which is the standard renewal-theoretic expression for $\mathbb{E}_\pi[f]$. Bounding the contribution of the incomplete final block by $\|f\|_\infty \tau_{n+1}/T \to 0$ finishes the argument. See Levin and Peres (2017, ch. 4) for details. □

Note that aperiodicity is not required: time averages converge even when the distribution at a fixed time does not. This is the version needed in Chapter 8, where the normalised log-likelihood is a time average along the trajectory of an ergodic process.

2.3 Quantifying the rate

Knowing that the chain converges is rarely enough; the asymptotic theory needs a rate, because the error terms in a central limit theorem are controlled by how quickly dependence decays.

2.3.1 Mixing time

Definition 2.6 (Mixing time)**.** With $\mu_t(\delta) = \delta^\top \Gamma^t$, the **mixing time** at accuracy $\varepsilon > 0$ is

$$t_{\mathrm{mix}}(\varepsilon) = \min\Big\{ t \geq 0 : \sup_{\delta \in \Delta^{K-1}} \|\mu_t(\delta) - \pi\|_{\mathrm{TV}} \leq \varepsilon \Big\}.$$

Proposition 2.5 (Spectral bound on mixing)**.** *For a reversible, irreducible, aperiodic chain with absolute spectral gap* $\varrho = 1 - |\lambda_2|$ *and* $\pi_{\min} = \min_i \pi_i$*,*

$$\|\mu_t(\delta) - \pi\|_{\mathrm{TV}} \;\le\; \frac{(1-\varrho)^t}{2\sqrt{\pi_{\min}}}, \qquad \textit{hence} \qquad t_{\mathrm{mix}}(\varepsilon) \le \frac{1}{\varrho} \log \frac{1}{2\varepsilon\sqrt{\pi_{\min}}}.$$

The bound is proved by expanding $\delta^\top \Gamma^t$ in the π-orthonormal eigenbasis supplied by Proposition 2.3 and applying Cauchy–Schwarz to pass from $\ell^2(\pi)$ to total variation; Levin and Peres (2017, ch. 12) has the details. The shape is what matters: the relaxation time $1/\varrho$ sets the timescale, and the dependence on $\pi_{\min}$ enters only logarithmically.

2.3.2 Coupling

Coupling gives bounds without any spectral assumption.

Definition 2.7 (Coupling)**.** A **coupling** of two chains with transition matrix Γ and initial distributions δ, δ' is a process (S_t, S'_t) on a common probability space whose marginals are the two chains. Its **coalescence time** is $\tau = \inf\{t \ge 0 : S_t = S'_t\}$, and the coupling is such that $S_s = S'_s$ for all $s \ge \tau$.

Proposition 2.6 (Coupling inequality)**.** *For any such coupling and every* $t \ge 0$*,*

$$\left\|\delta^\top \Gamma^t - (\delta')^\top \Gamma^t\right\|_{\mathrm{TV}} \le \mathbb{P}(\tau > t).$$

Proof. For any $A \subseteq E$, the events $\{S_t \in A\}$ and $\{S'_t \in A\}$ agree on $\{\tau \le t\}$, so their probabilities differ by at most $\mathbb{P}(\tau > t)$. Taking the supremum over A gives the total variation distance. □

Constructing a coupling that coalesces quickly then gives a mixing bound directly. The same idea, applied to two filters started from different initial distributions, is how filter stability is proved in Chapter 5.

2.3.3 Dobrushin's coefficient

Definition 2.8 (Dobrushin contraction coefficient)**.** For a stochastic matrix Γ,

$$\kappa(\Gamma) = \frac{1}{2} \max_{i,i' \in E} \sum_{j \in E} \left|\gamma_{ij} - \gamma_{i'j}\right| \;=\; \max_{i,i'} \left\|\Gamma_{i,\cdot} - \Gamma_{i',\cdot}\right\|_{\mathrm{TV}} \in [0,1].$$

Theorem 2.3 (Dobrushin contraction)**.** *For any probability vectors* μ, ν *on* E*,*

$$\left\|\mu^\top \Gamma - \nu^\top \Gamma\right\|_{\mathrm{TV}} \;\le\; \kappa(\Gamma)\, \|\mu - \nu\|_{\mathrm{TV}},$$

and κ *is submultiplicative:* $\kappa(\Gamma_1 \Gamma_2) \le \kappa(\Gamma_1)\kappa(\Gamma_2)$*.*

Two features make this more useful than the spectral bound in the HMM setting. It needs no reversibility, no diagonalisability and no stationary distribution. And because it is submultiplicative, it applies verbatim to products of *different* transition matrices, which is exactly the situation for time-inhomogeneous chains and for the observation-dependent operators that appear inside the forward recursion. A sufficient condition for $\kappa(\Gamma) < 1$ is the Doeblin minorisation $\gamma_{ij} \ge \epsilon\, \nu_j$ for all i, j, some $\epsilon > 0$ and a probability vector ν; then $\kappa(\Gamma) \le 1 - \epsilon$.

2.4 Time-inhomogeneous chains

Some extensions replace Γ by a sequence (Γ_t), with $\mathbb{P}(S_{t+1} = j \mid S_t = i) = (\Gamma_t)_{ij}$. This covers HMMs whose transition probabilities depend on covariates, a construction used throughout Zucchini et al. (2016, ch. 12) and again in the hierarchical models of Chapter 10.

Define the composite kernel $\Gamma_{s,t} = \Gamma_s \Gamma_{s+1} \cdots \Gamma_{t-1}$ for $s < t$. The Chapman–Kolmogorov identity survives as $\Gamma_{s,u} = \Gamma_{s,t}\Gamma_{t,u}$ for $s < t < u$, but stationary distributions generally do not exist, so the questions change: instead of asking where the chain settles, we ask whether it forgets its initial condition.

Proposition 2.7 (Loss of memory)**.** *If $\kappa(\Gamma_t) \le 1 - \epsilon$ for every t and some $\epsilon > 0$, then for any initial distributions μ, ν,*

$$\left\| \mu^\top \Gamma_{1,t} - \nu^\top \Gamma_{1,t} \right\|_{\mathrm{TV}} \le (1-\epsilon)^{t-1} \|\mu - \nu\|_{\mathrm{TV}} \longrightarrow 0.$$

This follows from Theorem 2.3 by submultiplicativity. The uniform lower bound on transition probabilities is a Doeblin condition, and it is the standard hypothesis in the filter stability literature; Cappé et al. (2005, ch. 4) and Del Moral and Guionnet (2001) develop the consequences for HMMs.

Check your understanding. The chain with transition matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is irreducible. Does $\mathbb{P}(S_t = 1)$ converge? Does $T^{-1} \sum_t \mathbf{1}\{S_t = 1\}$ converge?

Show answer

The chain alternates deterministically, so it has period 2 and is not aperiodic. Started from $S_1 = 1$, $\mathbb{P}(S_t = 1)$ alternates between 1 and 0 forever and never converges — Theorem 2.1 does not apply.
The time average does converge, to $1/2 = \pi_1$. Theorem 2.2 requires only irreducibility, which is exactly why it is the version needed in Chapter 8: the normalised log-likelihood is a time average along a single trajectory, and periodicity does not disturb it.
Note also that $\kappa(\Gamma) = 1$ here, so Theorem 2.3 gives no contraction — correctly, since this chain never forgets where it started.

2.5 Summary

- A finite-state Markov chain is determined by (δ, Γ), and its finite-dimensional distributions factorise as a product of one initial term and one transition term per step (Proposition 2.1).
- Irreducibility gives a unique stationary distribution; adding aperiodicity gives convergence to it from any starting point (Theorem 2.1).
- Time averages converge almost surely to expectations under π whenever the chain is irreducible (Theorem 2.2), with no aperiodicity needed.
- Rates come from the spectral gap under reversibility (Proposition 2.5), or from coupling and Dobrushin's coefficient in general (Theorem 2.3). The latter extends to time-inhomogeneous chains and to the filter recursions of Chapter 5.

2.6 Further reading

Levin and Peres (2017) is the standard modern reference for finite-state chains and mixing times, with coupling arguments given in full. Seneta (2006) approaches the same material through non-negative matrices, and includes a detailed treatment of Dobrushin's coefficient. For general state spaces, Meyn and Tweedie (2009) is the canonical source. Cappé et al. (2005, ch. 4) specialises the theory to the operators that arise in HMM filtering, and Zucchini et al. (2016, ch. 1) gives a compact applied account with the same notation used here.

Part II

II. The model

3 The hidden Markov model

This is the chapter the rest of the notes rests on. It states the model, writes down its two conditional independence assumptions, derives the factorisation those assumptions force, and turns that factorisation into a likelihood.

The last step is where the subject becomes computational. The likelihood is a sum over K^T state sequences, which is hopeless as written, and it is identical to a product of T small matrices, which is not. Everything in Parts III and IV follows from noticing that.

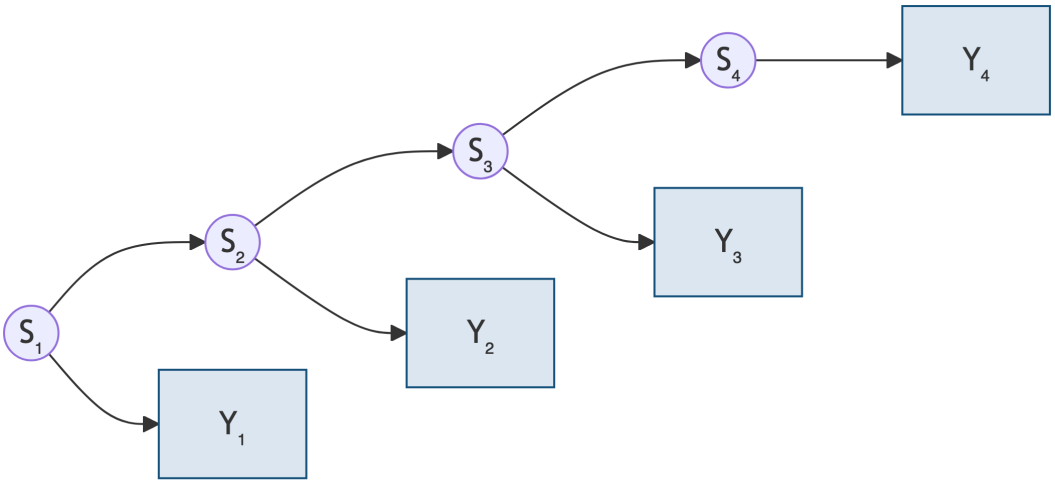
By the end of this chapter you will be able to

- state the two conditional independence assumptions that define an HMM, and say what each one rules out;
- derive the joint factorisation of states and observations, and read off which independences hold;
- write the observed-data likelihood both as a sum over sequences and as a matrix product, and explain why the second costs $\mathcal{O}(K^2T)$;
- explain why the log-likelihood has at least $K!$ global maxima, and what that does and does not matter for.

Notation follows Zucchini et al. (2016) throughout, so the formulas here match theirs term for term.

3.1 The two assumptions

An HMM over T time points has hidden states $S_1, \dots, S_T$ taking values in $E = \{1, \dots, K\}$ and observations $Y_1, \dots, Y_T$ taking values in a measurable space $\mathcal{Y}$. As a directed graphical model it has exactly two kinds of edge: state to next state, and state to its own observation.



The HMM as a directed graphical model. Shaded nodes are observed.

Definition 3.1 (The two HMM assumptions)**. (A1) Markov states.** For every t, $S_{t+1} \perp\!\!\!\perp S_{1:t-1} \mid S_t$.

(A2) Conditionally independent emissions. For every t,

$$Y_t \perp\!\!\!\perp (S_{1:t-1}, S_{t+1:T},\ Y_{1:t-1}, Y_{t+1:T}) \ \mid\ S_t.$$

They fail in different ways and are relaxed by different extensions, so it is worth keeping them apart.

(A1) is a statement about memory in the latent process. It forces the time spent in a state to be geometrically distributed, since leaving is an independent coin flip at each step. Hidden semi-Markov models replace that with an arbitrary dwell-time distribution; higher-order chains let the next state depend on several previous ones.

(A2) says the current state screens off the current observation from everything else. Autoregressive HMMs weaken it by letting Y_t depend on Y_{t-1} as well as S_t, which is what Chapter 14 needs for Markov-switching autoregressions and for modelling dinucleotide composition in genomes.

3.2 The model

Definition 3.2 (Finite-state hidden Markov model)**.** A **finite-state hidden Markov model** consists of:

- a finite **state space** $E = \{1, \dots, K\}$;
- an **observation space** $(\mathcal{Y}, \mathcal{B}_y)$ with a σ-finite dominating measure ν;
- an **initial distribution** $\delta = (\delta_1, \dots, \delta_K)^\top \in \Delta^{K-1}$;
- a **transition matrix** $\Gamma = (\gamma_{ij})_{i,j=1}^K$, row-stochastic;
- **emission densities** $f_i = dF_i/d\nu$ for $i \in E$.

The associated process $(S_t, Y_t)_{t \geq 1}$ is defined by: $S_1 \sim \delta$; $\mathbb{P}(S_{t+1} = j \mid S_{1:t}, Y_{1:t}) = \gamma_{S_t j}$; and conditionally on $S_{1:T}$ the observations $Y_1, \dots, Y_T$ are independent with $Y_t \sim F_{S_t}$.

We write $\theta = (\delta, \Gamma, \phi_1, \dots, \phi_K)$ for the full parameter, where ϕ_i parameterises f_i.

Three remarks on the definition.

The dominating measure absorbs the discrete–continuous distinction. Taking ν to be counting measure on a finite $\mathcal{Y}$ makes f_i a probability mass function; taking Lebesgue measure on $\mathbb{R}^d$ makes it a density. Every formula below holds in both cases unchanged, which is why these notes never split into separate discrete and continuous treatments.

The requirement $\mathbb{P}(S_{t+1} = j \mid S_{1:t}, Y_{1:t}) = \gamma_{S_t j}$ conditions on past observations as well as past states. That is stronger than the bare Markov property of Definition 2.1 and is the precise form of (A1): the hidden chain evolves without reference to what has been seen. It is what lets the proof below drop $y_{1:t-1}$ from the transition factor.

Nothing so far requires δ to be stationary. Setting $\delta = \pi$ makes (S_t, Y_t) strictly stationary and removes $K - 1$ free parameters. That choice is standard in the asymptotic theory of Chapter 8, where stationarity is needed to define the limiting log-likelihood, and it is often right for a long series. For a short one it forces an assumption the data may not support.

3.3 The factorisation

Theorem 3.1 (Joint factorisation)**.** *Under (A1) and (A2), for every $T \geq 1$ the joint density of $(S_{1:T}, Y_{1:T})$ with respect to counting measure on E^T times $\nu^{\otimes T}$ on $\mathcal{Y}^T$ is*

$$p(s_{1:T}, y_{1:T}) \;=\; \delta_{s_1} f_{s_1}(y_1) \prod_{t=2}^{T} \gamma_{s_{t-1} s_t} f_{s_t}(y_t).$$

Proof. Apply the chain rule in the order $S_1, Y_1, S_2, Y_2, \ldots$:

$$p(s_{1:T}, y_{1:T}) = p(s_1)\, p(y_1 \mid s_1) \prod_{t=2}^{T} p(s_t \mid s_{1:t-1}, y_{1:t-1})\, p(y_t \mid s_{1:t}, y_{1:t-1}).$$

Assumption (A1), in the form stated in Definition 3.2, reduces $p(s_t \mid s_{1:t-1}, y_{1:t-1})$ to $\gamma_{s_{t-1} s_t}$. Assumption (A2) reduces $p(y_t \mid s_{1:t}, y_{1:t-1})$ to $f_{s_t}(y_t)$. □

One line of algebra, and it is the source of every algorithm in these notes. The forward recursion sums it left to right, the backward recursion right to left, Viterbi maximises instead of summing, and EM takes expectations of its logarithm.

3.3.1 What follows, and what does not

Proposition 3.1 (Some consequences of the structure)**.** *Under (A1) and (A2):*

1. *$Y_{1:t-1} \perp\!\!\!\perp Y_{t+1:T} \mid S_t$ — the current state separates past from future observations.*
2. *$S_{t+1:T} \perp\!\!\!\perp Y_{1:t} \mid S_t$ — the current state is a sufficient summary of the past for predicting the future.*
3. *Given $S_{1:T}$, the observations $Y_1, \ldots, Y_T$ are mutually independent.*
4. *Given $Y_{1:T}$, the hidden states $S_1, \ldots, S_T$ form a Markov chain in t, generally time-inhomogeneous.*

Statement 1 is what makes the two-filter formula of Chapter 5 work: it licenses splitting $p(s_t, y_{1:T})$ into a past factor and a future factor. Statement 4 is why backward sampling in Chapter 9 can draw an entire state trajectory in one sweep.

What is *not* true is that the observations are marginally independent. Marginalising out the states induces dependence at every lag, and that dependence is the whole signal an HMM exploits. Nor are the smoothed states independent across time, which is why the pairwise quantities $\xi_t(i, j)$ have to be computed separately for EM.

Check your understanding. Someone tells you that in an HMM, Y_1 and Y_2 are independent, because emissions are conditionally independent given the states. Where is the error?

Show answer

The independence in (A2) is *given the states*. Statement 3 of Proposition 3.1 says $Y_1 \perp\!\!\!\perp Y_2 \mid S_{1:T}$, not $Y_1 \perp\!\!\!\perp Y_2$. Marginally,

$$\mathbb{P}(Y_1 \in A, Y_2 \in B) = \sum_{i,j} \delta_i \gamma_{ij} F_i(A) F_j(B),$$

which factorises into $\mathbb{P}(Y_1 \in A)\, \mathbb{P}(Y_2 \in B)$ only if $\delta_i \gamma_{ij}$ factorises — that is, only if S_2 is independent of S_1. An HMM with a non-trivial transition matrix produces dependent observations by construction. Exercise 15.8 computes the autocorrelation explicitly.

3.4 The complete-data likelihood

If the states were observed, estimation would be counting. It is worth writing that case down first, because EM in Chapter 6 works by repeatedly pretending it holds.

By Theorem 3.1 the **complete-data likelihood** is

$$L_c(\theta; s_{1:T}, y_{1:T}) \;=\; \delta_{s_1} f_{s_1}(y_1; \phi_{s_1}) \prod_{t=2}^{T} \gamma_{s_{t-1} s_t} \, f_{s_t}(y_t; \phi_{s_t}),$$

with logarithm

$$\ell_c(\theta) \;=\; \log \delta_{s_1} \;+\; \sum_{t=2}^{T} \log \gamma_{s_{t-1} s_t} \;+\; \sum_{t=1}^{T} \log f_{s_t}(y_t; \phi_{s_t}). \tag{3.1}$$

The three terms separate cleanly, and each depends on the state sequence only through counts. Writing

$$u_i = \mathbf{1}\{s_1 = i\}, \qquad n_{ij} = \sum_{t=2}^{T} \mathbf{1}\{s_{t-1} = i,\, s_t = j\},$$

3.1 becomes

$$\ell_c(\theta) = \sum_i u_i \log \delta_i + \sum_{i,j} n_{ij} \log \gamma_{ij} + \sum_i \sum_{t : s_t = i} \log f_i(y_t; \phi_i).$$

The initial-state indicator, the transition count matrix, and the observations grouped by state are complete-data sufficient statistics. Maximising over Γ subject to the row constraints gives $\hat{\gamma}_{ij} = n_{ij} / \sum_l n_{il}$, the empirical transition frequency, and the emission parameters split into K independent maximum likelihood problems. That is the whole content of the M-step; the work in Chapter 6 lies in replacing unobservable counts by their conditional expectations.

3.5 The observed-data likelihood

Definition 3.3 (Observed-data likelihood)**.** For $y_{1:T} \in \mathcal{Y}^T$,

$$L_T(\theta) \;=\; p_\theta(y_{1:T}) \;=\; \sum_{s_{1:T} \in E^T} \delta_{s_1} f_{s_1}(y_1) \prod_{t=2}^{T} \gamma_{s_{t-1} s_t} f_{s_t}(y_t),$$

and $\ell_T(\theta) = \log L_T(\theta)$.

3.5.1 Why the naive sum is hopeless

The sum runs over K^T sequences. With $K = 5$ states and $T = 100$ — modest by any standard — that is $5^{100} \approx 8 \times 10^{69}$ terms. No hardware helps; the growth is exponential in the length of the series, so evaluating the likelihood even once is impossible.

What redeems it is the structure of the summand. Each term is a product in which s_t appears in only three factors, so the sum distributes: summing over s_1 first leaves an expression in which s_1 no longer appears, and the process repeats. That is the distributive law applied systematically, and it is the whole idea behind the forward algorithm.

3.5.2 The matrix-product form

Definition 3.4 (Emission matrix)**.** For $y \in \mathcal{Y}$ set $\mathbf{Q}(y) = \operatorname{diag}(f_1(y), \dots, f_K(y))$, a $K \times K$ diagonal matrix.

Theorem 3.2 (Matrix-product representation)**.** *For every $T \geq 1$,*

$$L_T(\theta) \;=\; \delta^\top \mathbf{Q}(y_1)\, \Gamma \mathbf{Q}(y_2)\, \Gamma \mathbf{Q}(y_3) \cdots \Gamma \mathbf{Q}(y_T)\, \mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^K$ is the vector of ones.

Proof. Expand the product entrywise. Writing $\mathbf{A} = \delta^\top \mathbf{Q}(y_1) \prod_{t=2}^{T} \Gamma \mathbf{Q}(y_t)$, the s_T-th entry of the row vector $\mathbf{A}$ is obtained by summing over all intermediate indices:

$$A_{s_T} = \sum_{s_1, \dots, s_{T-1}} \delta_{s_1} f_{s_1}(y_1)\, \gamma_{s_1 s_2} f_{s_2}(y_2) \cdots \gamma_{s_{T-1} s_T} f_{s_T}(y_T),$$

because each factor $\Gamma \mathbf{Q}(y_t)$ contributes $\gamma_{s_{t-1} s_t} f_{s_t}(y_t)$ and matrix multiplication sums over the shared index. Right-multiplying by $\mathbf{1}$ sums over s_T as well, giving exactly Definition 3.3. □

This is Zucchini et al. (2016, eq. 2.12), and it repays study. The likelihood is a product of T matrices of size $K \times K$, with the observation at time t entering only through the diagonal reweighting $\mathbf{Q}(y_t)$.

3.6 Why evaluation order matters

Matrix multiplication is associative, so the product can be evaluated in any order — but the cost differs enormously.

Right to left, matrix by matrix. Each $\Gamma \mathbf{Q}(y_t)$ product is $K \times K$ times $K \times K$, costing $\mathcal{O}(K^3)$ per step and $\mathcal{O}(K^3 T)$ overall, with $\mathcal{O}(K^2)$ memory.

Left to right, vector by matrix. Start with the row vector $\delta^\top \mathbf{Q}(y_1)$ and multiply by one matrix at a time. Each step is a row vector times a $K \times K$ matrix, costing $\mathcal{O}(K^2)$, for $\mathcal{O}(K^2 T)$ overall and only $\mathcal{O}(K)$ memory.

The second is the forward algorithm. It is not a separate invention — it is the sensible evaluation order for this product, and the intermediate row vectors it produces are exactly the forward variables of Chapter 5.

3.7 What the factors mean

Read the product from the left as a belief being pushed through time.

- $\delta^\top$ is what you believe about S_1 before seeing anything.
- $\mathbf{Q}(y_1)$ reweights each state by how well it explains the observation just seen.
- Γ pushes that belief one step forward through the dynamics.
- The final $\mathbf{1}$ sums over the last state, which you do not care about.

Alternating *predict* and *update* is exactly the structure of the Kalman filter in Chapter 12, and of Bayesian filtering in general. The finite-state case is special only in that a belief is a vector of K numbers, so the operations are matrix algebra.

3.8 A worked check

Take $K = 2$, $\delta = (0.5,\ 0.5)^\top$, and

$$\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}, \qquad \mathbf{Q}(y_1) = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.1 \end{pmatrix}, \qquad \mathbf{Q}(y_2) = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.6 \end{pmatrix}.$$

Then $\delta^\top \mathbf{Q}(y_1) = (0.20,\ 0.05)$, and

$$(0.20,\ 0.05)\,\Gamma = (0.19,\ 0.06), \qquad (0.19,\ 0.06)\,\mathbf{Q}(y_2) = (0.057,\ 0.036),$$

so $L_2(\theta) = 0.057 + 0.036 = 0.093$. Summing over all $2^2 = 4$ state sequences directly gives the same number, and doing that by hand once is the fastest way to believe Theorem 3.2.

Two consequences follow immediately.

Corollary 3.1 (Smoothness)**.** *If each $\phi \mapsto f_i(y;\phi)$ is continuous, then $\theta \mapsto L_T(\theta)$ is continuous; if each is k times continuously differentiable, so is L_T.*

The likelihood is a finite sum of products of the parameters, so smoothness is inherited term by term. This licenses the Taylor expansions in Chapter 8 and the gradient-based optimisation of Chapter 7.

Corollary 3.2 (Multilinearity)**.** *$L_T(\theta)$ is a multilinear function of δ, of the rows of Γ, and of the vectors $(f_i(y_t))_i$ — linear in each when the others are held fixed.*

Multilinearity is what the tensor arguments behind Theorem 4.1 exploit. It is also a warning: linearity in each block separately says nothing about joint concavity.

3.9 Symmetry and the shape of the likelihood

3.9.1 Label switching

Proposition 3.2 (Permutation invariance)**.** *Let σ be a permutation of E and define θ^σ by $\delta^\sigma_i = \delta_{\sigma^{-1}(i)}$, $\gamma^\sigma_{ij} = \gamma_{\sigma^{-1}(i)\sigma^{-1}(j)}$ and $\phi^\sigma_i = \phi_{\sigma^{-1}(i)}$. Then θ and θ^σ induce the same law for $Y_{1:T}$, for every T.*

Proof. In the sum of Definition 3.3, reindex by $s'_t = \sigma(s_t)$. Each factor maps to the corresponding factor of θ^σ, and the sum runs over the same set E^T. □

State labels carry no meaning, so nothing scientific is lost. What follows is that the best any procedure can do is identify θ as an *unordered collection* of states. The practical consequences recur in Chapter 6 and, much more sharply, in Chapter 9.

3.9.2 Non-concavity

Proposition 3.3 (The log-likelihood is not concave)**.** *For $K \geq 2$, ℓ_T is in general a non-concave function of θ with multiple local maxima.*

Proof. By Proposition 3.2, $\ell_T(\theta) = \ell_T(\theta^\sigma)$ for each of the $K!$ permutations σ. If θ is a strict local maximum with distinct states, so is every θ^σ, and these are distinct points of Θ. A concave function on a convex set cannot have two strict local maxima, since its maximiser set is convex and any point on the segment between them would have to attain at least the common value. □

The permutation symmetry alone guarantees at least $K!$ global maxima, which is harmless — they are relabellings of each other. What is not harmless is that HMM likelihoods routinely have local maxima that are *not* relabellings, corresponding to genuinely worse fits from which local optimisation cannot escape. Any procedure reporting a single answer from a single starting point is reporting a local property of the surface.

Check your understanding. A colleague fits a three-state HMM twice from different random starts and gets log-likelihoods of -412.7 both times, but completely different parameter values. Is something wrong?

Show answer

Almost certainly not. With $K = 3$ there are $3! = 6$ relabellings of any fit, all with identical log-likelihood. Check whether the two parameter sets are permutations of one another — sort the states by their emission means and compare. If they match, both runs found the same optimum and there is nothing to fix.
Had the log-likelihoods differed, say -412.7 against -418.3, that would be the more interesting case: two genuinely different local maxima, and a sign that more starting values are needed.

3.10 The parameter space

The parameter decomposes as

$$\Theta \;=\; \Delta^{K-1} \times \mathcal{G}_K \times \Phi_1 \times \cdots \times \Phi_K,$$

where $\mathcal{G}_K = (\Delta^{K-1})^K$ is the set of row-stochastic $K \times K$ matrices.

Counting free parameters: δ contributes $K-1$, or 0 under stationarity; Γ contributes $K(K-1)$; the emissions contribute $K\dim(\Phi)$ for a common family.

Table 3.1: Free parameters in a stationary univariate Gaussian HMM

K	Transition parameters	Emission parameters	Total
2	2	4	6
3	6	6	12
5	20	10	30
10	90	20	110

The quadratic growth in K is the practical reason state numbers stay small, and the reason model selection in Chapter 8 and Chapter 11 is not an afterthought.

The constraints are simplex constraints, and optimisers dislike them. Two standard devices:

- **Reparameterise.** Write $\gamma_{ij} = \exp(\tau_{ij}) / \sum_l \exp(\tau_{il})$ with $\tau_{ii} = 0$ fixed for identification, and optimise the unconstrained τ. This is what Zucchini et al. (2016) do to hand the problem to a general-purpose optimiser, and it extends naturally to covariate-dependent transitions.
- **Solve the constrained problem in closed form.** The M-step of EM maximises $\sum_j n_{ij} \log \gamma_{ij}$ subject to $\sum_j \gamma_{ij} = 1$, and a Lagrange multiplier gives $\gamma_{ij} \propto n_{ij}$ directly. This is why EM never needs a constrained optimiser.

3.11 Summary

- An HMM is defined by two conditional independence assumptions (Definition 3.1): a Markov latent chain, and observations conditionally independent given the states.
- Those assumptions force the joint factorisation of Theorem 3.1, from which every algorithm in these notes follows.
- With complete data, estimation is counting: the complete-data log-likelihood depends on the states only through initial, transition and emission sufficient statistics.
- The observed-data likelihood is a sum over K^T sequences and equals a product of T matrices of size $K \times K$ (Theorem 3.2). Evaluating that product left to right costs $\mathcal{O}(K^2 T)$ and is the forward algorithm.
- The likelihood is smooth and multilinear in its blocks, but not concave: label switching alone forces $K!$ global maxima (Proposition 3.2, Proposition 3.3).

3.12 Further reading

Zucchini et al. (2016, ch. 2) is the closest match to this chapter and derives the matrix representation in the same notation. Cappé et al. (2005, chs. 1–2) give the general state-space construction, where Γ becomes a Markov kernel and the matrix product a composition of operators; the algebra is identical. Rabiner (1989) states the three canonical problems — evaluation, decoding, learning — that organise Chapter 5 and Chapter 6. Murphy (2012, ch. 17) and Bishop (2006, ch. 13) place HMMs in the wider graphical model framework, which is the cleanest way to see which independences are structural and which are incidental.

4 Observation models

Chapter 3 treated the emission densities f_i as given. This chapter is about choosing them, and about what that choice determines.

More rides on it than is obvious. The emission family fixes what a state can mean, how expensive the M-step of Chapter 6 is, whether the likelihood is bounded, and — through the identifiability results at the end — whether the parameters can be recovered at all.

By the end of this chapter you will be able to

- match an emission family to a data type, and say what goes wrong when the match is poor;
- recognise an exponential-family emission model and write down the M-step it implies;
- build covariate dependence into either the emissions or the transitions;
- distinguish label switching from structural non-identifiability, and state a checkable condition under which an HMM is identified.

4.1 Choosing an emission family

The state-dependent distributions $\{F_i\}_{i \in E}$, with densities f_i against a common dominating measure ν, are where the modelling happens.

4.2 Discrete

When $\mathcal{Y} = \{1, \dots, M\}$ is finite, each state carries a probability vector,

$$b_i(y) = \mathbb{P}(Y_t = y \mid S_t = i), \qquad \sum_{y=1}^{M} b_i(y) = 1,$$

collected into a $K \times M$ emission matrix $\mathbf{B}$.

This is the setting of the original HMM literature — Baum et al. (1970) for the algorithm, Rabiner (1989) for speech, Durbin et al. (1998) for biological sequences — and it remains the cleanest case for identifiability theory, because the model is then a finite-dimensional algebraic object and the tensor methods of Allman et al. (2009) apply directly.

The parameter count is $K(M-1)$, which grows quickly in the alphabet size. For large M that is the binding constraint, and it is why speech systems moved from free tables to parametric or neural emission scores.

4.3 Continuous

When $\mathcal{Y} \subseteq \mathbb{R}^d$, each state carries a density against Lebesgue measure. Gaussian emissions $f_i = \mathcal{N}(\cdot\,; \mu_i, \Sigma_i)$ are the default and give the state a direct reading as a regime with its own level and variability.

Mixtures of Gaussians within each state add flexibility at the cost of a second layer of label ambiguity. Heavy-tailed alternatives such as the Student t guard against a well-known failure: an outlier can otherwise capture a state of its own, driving that state's variance towards zero.

That degeneracy is not a numerical accident but a property of the model — see the warning below.

4.4 Exponential family

Most standard families fit Definition 1.13, with state i carrying its own natural parameter:

$$f_i(y) = h(y) \exp\{\langle \eta_i, T(y) \rangle - A(\eta_i)\}.$$

The payoff is computational. The M-step of EM reduces to

$$\nabla A(\eta_i) \;=\; \frac{\sum_{t=1}^{T} \gamma_t(i)\, T(y_t)}{\sum_{t=1}^{T} \gamma_t(i)},$$

a weighted moment-matching problem with the same form as the complete-data problem. Only weighted sufficient statistics need accumulating, so the E-step and M-step communicate through a fixed, small amount of information regardless of T.

4.5 Covariate-dependent

Nothing forces emission parameters to be constant. Writing $f_i(y_t; \phi_i, x_t)$ for a state-dependent regression on covariates turns the HMM into a Markov-switching regression: states index regimes, each with its own coefficient vector.

Transitions can depend on covariates in the same way, through a multinomial logit,

$$\gamma_{ij}(x_t) = \frac{\exp(\beta_{ij}^\top x_t)}{\sum_{l=1}^{K} \exp(\beta_{il}^\top x_t)},$$

with one column constrained for identification. The chain is then time-inhomogeneous in the sense of Chapter 2, but the recursions of Chapter 5 are unaffected: they never used homogeneity, only the factorisation of Theorem 3.1.

4.5.1 What to use when

Table 4.1: Emission families in common use, following Zucchini et al. (2016, ch. 8)

Data type	Emission family	Typical use
Counts	Poisson, negative binomial	event counts per interval; overdispersion via the latter
Binary or categorical	Bernoulli, multinomial	sequence labelling, symbol streams
Continuous, unbounded	Gaussian, Student t	returns, sensor readings
Continuous, positive	gamma, log-normal, Weibull	durations, step lengths
Angular	von Mises, wrapped Cauchy	turning angles in movement models
Multivariate	multivariate normal, copula constructions	joint sensor or biomarker panels

The rule behind the table is that the support of the family should match the support of the data. A Gaussian fitted to a strictly positive variable puts mass below zero, and the model compensates by manufacturing a state to handle the boundary — a spurious state that then gets interpreted as real.

⚠ Unconstrained Gaussian emissions make the likelihood unbounded

Put $\mu_i = y_1$ for some state i and let $\sigma_i \to 0$. The term $f_i(y_1)$ diverges while every other factor stays bounded away from zero, so $L_T(\theta) \to \infty$.
The supremum is therefore not attained at any sensible parameter value, and "the maximum likelihood estimator" has to mean a suitable *local* maximum. In practice: bound the variances away from zero, use a prior that does it for you (Chapter 9), or choose a family whose scale cannot collapse. The same problem afflicts finite Gaussian mixtures and is not specific to HMMs.

Example 4.1 (The occasionally dishonest casino)**.** A dealer alternates between a fair die and a loaded one. Hidden states are $E = \{\text{fair}, \text{loaded}\}$; observations are face values $\mathcal{Y} = \{1, \dots, 6\}$; emissions are $b_{\text{fair}}(y) = 1/6$ and $b_{\text{loaded}}(6) = 1/2$ with $b_{\text{loaded}}(y) = 1/10$ otherwise; and the transition matrix has large diagonal entries, so switches are rare.

Every inference question has a reading here. Filtering asks whether the die in play *right now* is loaded. Smoothing asks the same with hindsight. Viterbi decoding asks for the single most plausible sequence of switches. Durbin et al. (1998, ch. 3) use this example throughout, and it is the cheapest way to build intuition for why the three answers differ.

Check your understanding. You are modelling daily rainfall totals, which are zero on many days and positive otherwise. Why is a Gaussian emission a poor choice, and what would you do instead?

Show answer

Two problems. The support is wrong — a Gaussian puts mass on negative rainfall — and the data have an atom at zero, which no continuous density can represent.
The usual fix is a hurdle construction: model $\mathbb{P}(Y_t = 0 \mid S_t = i) = p_i$ and, conditional on being positive, use a gamma or log-normal density. That is still an emission distribution in the sense of Definition 3.2 — the dominating measure is counting measure at $\{0\}$

plus Lebesgue measure on $(0, \infty)$ — so every algorithm in these notes applies unchanged. This is precisely what the dominating-measure convention in Chapter 3 buys.

4.6 Identifiability

Two parameter values that induce the same distribution for the observed process cannot be distinguished by any amount of data. For HMMs this happens for one unavoidable reason, covered in Chapter 3, and several avoidable ones.

4.6.1 The unavoidable part

Proposition 3.2 established that permuting state labels leaves the law of $Y_{1:T}$ unchanged. Identifiability can therefore only ever be claimed *up to permutation*, and the parameter is properly thought of as an unordered set of states. Nothing scientific is lost, since labels mean nothing.

4.6.2 The avoidable part

Beyond permutation, identifiability can fail for reasons that signal a badly posed model.

Duplicated states. If $f_i = f_j$ and rows i and j of Γ agree, the two states are observationally identical and can be merged, leaving a continuum of parameter values with the same likelihood.

Empty states. If $\pi_i = 0$ then ϕ_i never enters the likelihood and is unidentified — the data say nothing about a state that is never visited.

Overparameterised emissions. If the f_i are linearly dependent as functions, for instance if one state's density is a mixture of the others', then transition and emission parameters can trade off against one another.

The first two are symptoms of fitting too many states. That is exactly the question Chapter 11 addresses, by treating K as unknown and separating the number of states in the model from the number the data actually occupy.

4.6.3 When identifiability does hold

The positive results are sharper than one might expect.

Theorem 4.1 (Identifiability from three consecutive observations)**.** *Let (S_t, Y_t) be a stationary HMM with K states, transition matrix Γ of full rank K, and stationary distribution π with $\pi_i > 0$ for all i. If the emission distributions $F_1, \dots, F_K$ are linearly independent, then the law of (Y_1, Y_2, Y_3) determines Γ and $\{F_i\}$ up to permutation of the states.*

The result is due to Allman et al. (2009). The joint law of three consecutive observations is written as a three-way tensor whose factors are the emission distributions, and Kruskal's uniqueness theorem for trilinear decompositions (Kruskal 1977) gives the conclusion.

What makes it striking is that three time points suffice. No asymptotic argument is needed, and both conditions are checkable in practice. Linear independence is also much weaker than distinctness: it fails only when one emission distribution is an affine combination of the others.

The same tensor structure is what Hsu et al. (2012) turn into a spectral estimation algorithm with polynomial sample complexity, and it extends to fully nonparametric emissions, where the F_i are left unspecified and only linear independence is assumed (Gassiat et al. 2016). Petrie (1969) gave the first identifiability results for finite emissions, long before the tensor machinery existed.

Check your understanding. A fitted two-state model has $\hat{\mu}_1 = 2.01$, $\hat{\mu}_2 = 2.03$, $\hat{\sigma}_1 = \hat{\sigma}_2 = 1.4$, and nearly identical rows of $\hat{\Gamma}$. Which hypothesis of Theorem 4.1 is in trouble, and what should you do?

Show answer

Linear independence of the emission distributions is close to failing: F_1 and F_2 are almost the same distribution, so the tensor in the Kruskal argument is near-singular and the parameters are weakly identified. The near-identical transition rows compound it — this is the duplicated-states case above, approached from close by.
Expect huge standard errors, EM iterates that drift for many iterations while the likelihood barely moves, and fits that vary wildly across starting values. The right response is to fit $K = 1$ and compare, not to try harder to estimate a model the data do not support.

4.7 Model checking

Choosing a family is a modelling decision, so it should be checked rather than assumed. Two tools carry most of the load.

Pseudo-residuals. With $u_t = F(y_t \mid y_{1:t-1})$ the one-step predictive CDF evaluated at the observation, the transformed values $\Phi^{-1}(u_t)$ are approximately independent standard normal under correct specification. Quantile plots of these detect a mismatched emission family directly, and the predictive densities they need are already produced by the forward recursion (Proposition 5.1). Zucchini et al. (2016, ch. 6) develop this in detail.

Posterior predictive checks. Simulate series from the fitted model and compare features the model was *not* fitted to — the marginal distribution, the autocorrelation function, the distribution of run lengths within a state. Run lengths are especially informative, because assumption (A1) forces them to be geometric, and a mismatch there points to a hidden semi-Markov model rather than to a different emission family.

4.8 Summary

- The emission family determines what a state can mean, how expensive the M-step is, and whether the likelihood is bounded.

- Match the support of the family to the support of the data; a mismatch makes the model manufacture states to cover the boundary.
- Exponential-family emissions reduce the M-step to weighted moment matching, which is why they dominate in practice.
- Covariate dependence can enter through the emissions, the transitions, or both, and costs nothing algorithmically.
- Identifiability holds up to permutation and no better (Proposition 3.2). Under full rank of Γ and linear independence of the emissions, three consecutive observations suffice (Theorem 4.1).

4.9 Further reading

Zucchini et al. (2016, chs. 6, 8) catalogue emission families with worked examples and develop pseudo-residual diagnostics. Cappé et al. (2005, ch. 1) gives the measure-theoretic construction. For identifiability, read Allman et al. (2009) first and then Gassiat et al. (2016) for the nonparametric extension; Hsu et al. (2012) turns the same algebra into an estimator. Murphy (2012, ch. 17) and Bishop (2006, ch. 13) place the model within the wider graphical model framework.

Part III

III. Inference and estimation

5 Inference: filtering, smoothing, decoding

Given a fully specified HMM and a stretch of data, three questions arise. What is the state now, given everything seen so far? What was the state at time t, given the whole series? And which single sequence of states best explains the data? The three have different answers, computed by different algorithms, and confusing them is a common source of error.

All three algorithms are dynamic programming on a chain, all three cost $\mathcal{O}(K^2 T)$, and all three come out of the factorisation in Theorem 3.1. The treatment follows Zucchini et al. (2016, ch. 4) and Rabiner (1989), with the derivations written out and the numerical issues treated properly rather than deferred.

By the end of this chapter you will be able to

- derive the forward and backward recursions from the model assumptions, and prove they compute what they claim;
- implement either of them without underflow, and explain what the scale factors mean;
- say precisely how Viterbi decoding differs from taking the marginal mode at each time, and construct a case where they disagree;
- explain why the choice of initial distribution stops mattering after a few dozen observations.

The three problems

	Question	Object	Algorithm
Filtering	state now, given the past	$\mathbb{P}(S_t = i \mid Y_{1:t})$	forward
Smoothing	state then, given everything	$\mathbb{P}(S_t = i \mid Y_{1:T})$	forward–backward
Decoding	best whole path	$\arg\max_{s_{1:T}} \mathbb{P}(S_{1:T} = s_{1:T} \mid Y_{1:T})$	Viterbi

Throughout, θ is fixed and known, and we suppress it from the notation. Observations $y_{1:T}$ are fixed.

5.1 Filtering: the forward algorithm

5.1.1 Forward variables

Definition 5.1 (Forward variables)**.** For $t = 1, \dots, T$ and $i \in E$, the **unnormalised forward variable** is

$$\tilde{\alpha}_t(i) \;=\; p(S_t = i,\, Y_{1:t} = y_{1:t}),$$

and the **filtering probability** is $\alpha_t(i) = \mathbb{P}(S_t = i \mid Y_{1:t} = y_{1:t}) = \tilde{\alpha}_t(i) / \sum_j \tilde{\alpha}_t(j)$.

Working with the unnormalised version first is deliberate: it satisfies an exact linear recursion, whereas the normalised version satisfies a ratio of linear recursions. The normalising constants turn out to be useful in their own right.

Theorem 5.1 (Forward recursion)**.** *The variables of Definition 5.1 satisfy*

$$\tilde{\alpha}_1(i) = \delta_i f_i(y_1), \qquad \tilde{\alpha}_{t+1}(j) = f_j(y_{t+1}) \sum_{i=1}^{K} \tilde{\alpha}_t(i)\, \gamma_{ij} \quad (1 \le t \le T-1),$$

and $L_T(\theta) = \sum_{i=1}^{K} \tilde{\alpha}_T(i)$.

Proof. The initial case is the definition of a joint density: $p(S_1 = i, Y_1 = y_1) = \mathbb{P}(S_1 = i)\, p(y_1 \mid S_1 = i) = \delta_i f_i(y_1)$.

For the induction, marginalise over the state at time t and factorise:

$$\begin{aligned}\tilde{\alpha}_{t+1}(j) &= \sum_{i=1}^{K} p(S_t = i,\, S_{t+1} = j,\, Y_{1:t+1} = y_{1:t+1}) \\ &= \sum_{i=1}^{K} p(S_t = i, Y_{1:t} = y_{1:t})\, \mathbb{P}(S_{t+1} = j \mid S_t = i, Y_{1:t} = y_{1:t})\, p(y_{t+1} \mid S_{t+1} = j, S_t = i, Y_{1:t} = y_{1:t}).\end{aligned}$$

The second factor equals γ_{ij} by assumption (A1) in the form stated in Definition 3.2, and the third equals $f_j(y_{t+1})$ by assumption (A2). Since $f_j(y_{t+1})$ does not depend on i it comes outside the sum, giving the stated recursion.

Summing $\tilde{\alpha}_T(i)$ over i marginalises out S_T and leaves $p(y_{1:T}) = L_T(\theta)$. □

5.1.2 Matrix form

Collecting $\tilde{\alpha}_t = (\tilde{\alpha}_t(1), \dots, \tilde{\alpha}_t(K))$ as a row vector and using $\mathbf{Q}(y)$ from Definition 3.4, the recursion reads

$$\tilde{\alpha}_1 = \delta^\top \mathbf{Q}(y_1), \qquad \tilde{\alpha}_{t+1} = \tilde{\alpha}_t\, \Gamma\, \mathbf{Q}(y_{t+1}),$$

so that $\tilde{\alpha}_T \mathbf{1}$ reproduces Theorem 3.2. The forward algorithm *is* the left-to-right evaluation of that matrix product; nothing new has been introduced, only an order of operations. Each step costs one vector–matrix multiplication, $\mathcal{O}(K^2)$, for $\mathcal{O}(K^2 T)$ overall, and only the current vector need be retained if the likelihood is all that is wanted.

5.1.3 Prediction

Filtering also delivers forecasts. The one-step-ahead predictive distribution is

$$p(y_{t+1} \mid y_{1:t}) \;=\; \sum_{j=1}^{K} \Big(\sum_{i=1}^{K} \alpha_t(i) \gamma_{ij} \Big) f_j(y_{t+1}) \;=\; \alpha_t^\top \Gamma \left(f_j(y_{t+1}) \right)_j,$$

a mixture of the emission densities with weights given by the predicted state distribution. Iterating Γ gives h-step forecasts, and by Theorem 2.1 those weights converge to π as h grows: long-horizon forecasts from an ergodic HMM revert to the stationary mixture, regardless of the current state. Forecast pseudo-residuals built from these predictive distributions are the standard HMM diagnostic in Zucchini et al. (2016, ch. 6).

5.1.4 Numerical stability

The recursion is exact and, implemented naively, useless. Each step multiplies by an emission density and a transition probability, so $\tilde{\alpha}_t(i)$ decays roughly geometrically in t. With typical per-step factors of order 10^{-2}, double precision underflows to zero somewhere around $t \approx 150$, and the computed likelihood becomes exactly 0 with logarithm $-\infty$.

Two standard remedies, both exact in principle.

Scaling. Normalise at every step. Set $c_t = \sum_i \tilde{\alpha}_t(i)$ computed from the current unnormalised vector, and propagate the normalised vector $\hat{\alpha}_t = \tilde{\alpha}_t / c_t$ instead. Concretely, with $\hat{\alpha}_0^\top := \delta^\top$,

$$\mathbf{v}_t = \hat{\alpha}_{t-1}^\top \Gamma \mathbf{Q}(y_t), \qquad c_t = \mathbf{v}_t \mathbf{1}, \qquad \hat{\alpha}_t^\top = \mathbf{v}_t / c_t.$$

Proposition 5.1 (Scaling identity)**.** *With the above definitions, $\hat{\alpha}_t(i) = \mathbb{P}(S_t = i \mid Y_{1:t} = y_{1:t})$, the scale factor is the one-step predictive density $c_t = p(y_t \mid y_{1:t-1})$, and*

$$L_T(\theta) = \prod_{t=1}^{T} c_t, \qquad \ell_T(\theta) = \sum_{t=1}^{T} \log c_t.$$

Proof. By induction, $\mathbf{v}_t = \tilde{\alpha}_t / \prod_{s<t} c_s$, so $c_t = (\sum_i \tilde{\alpha}_t(i)) / \prod_{s<t} c_s = p(y_{1:t})/p(y_{1:t-1})$, which is the predictive density, and the telescoping product of the c_t is $p(y_{1:T})$. Normalising $\mathbf{v}_t$ removes the common factor and leaves the conditional distribution. □

The log-likelihood therefore accumulates as a sum of logarithms of quantities of order one. This is the implementation used throughout Zucchini et al. (2016), and it has the pleasant side effect that the quantities you must compute anyway are exactly the predictive densities needed for residual diagnostics.

Log domain. Alternatively, store $a_t(i) = \log \tilde{\alpha}_t(i)$ and compute

$$a_{t+1}(j) = \log f_j(y_{t+1}) + \operatorname{logsumexp}_i \{a_t(i) + \log \gamma_{ij}\},$$

using the stable form $\operatorname{logsumexp}(z) = m + \log \sum_i e^{z_i - m}$ with $m = \max_i z_i$. This handles a wider dynamic range than scaling and copes with structural zeros in Γ without special cases, at the cost of K^2 exponentials and logarithms per step. Scaling is faster; the log domain is more robust. Chapter 7 compares them in detail.

5.1.5 Filter stability

A practical question: how much does the answer depend on δ, which is often the least well-determined part of the model?

Theorem 5.2 (Forgetting of the initial distribution)**.** *Let α_t^δ and $\alpha_t^{\delta'}$ denote the filtering distributions at time t obtained from the same observations $y_{1:t}$ but different initial distributions. If the transition matrix satisfies the Doeblin condition $\gamma_{ij} \geq \epsilon \nu_j$ for some $\epsilon > 0$ and probability vector ν, then*

$$\left\| \alpha_t^\delta - \alpha_t^{\delta'} \right\|_{\mathrm{TV}} \leq (1-\epsilon)^{t-1} \left\| \delta - \delta' \right\|_{\mathrm{TV}}.$$

The proof runs exactly as Proposition 2.7. The filter update is the map $\mu \mapsto \mu^\top \Gamma \mathbf{Q}(y_t)/(\text{normalisation})$, and the diagonal reweighting by $\mathbf{Q}(y_t)$ does not increase the Dobrushin coefficient, so the composite map contracts total variation by at least $(1-\epsilon)$ per step. Cappé et al. (2005, ch. 4) and Del Moral and Guionnet (2001) give the general treatment.

The practical reading: two analysts who disagree about δ but agree on everything else will produce filtering distributions that are indistinguishable after a few dozen observations, provided the chain mixes. It also explains why the same misspecification barely moves ℓ_T for large T, and hence why δ is estimated poorly from a single long series.

5.2 Smoothing: the forward–backward algorithm

Filtering uses only the past. For anything retrospective — state estimates for interpretation, or the expected counts needed by EM — the whole series should be used.

5.2.1 Backward variables

Definition 5.2 (Backward variables)**.** For $t = 1, \dots, T$ and $i \in E$,

$$\beta_t(i) \;=\; p(Y_{t+1:T} = y_{t+1:T} \mid S_t = i),$$

with the convention $\beta_T(i) = 1$ (an empty product).

Note the asymmetry with Definition 5.1: $\tilde{\alpha}_t$ is a joint density and β_t is a conditional one. The asymmetry is what makes their product come out right.

Theorem 5.3 (Backward recursion)**.** *For $t = T-1, T-2, \dots, 1$,*

$$\beta_t(i) \;=\; \sum_{j=1}^{K} \gamma_{ij}\, f_j(y_{t+1})\, \beta_{t+1}(j), \qquad \textit{i.e.} \qquad \beta_{t} = \Gamma \mathbf{Q}(y_{t+1}) \beta_{t+1}$$

with β_t a column vector.

Proof. Condition on the next state and split the future observations at time $t+1$:

$$\beta_t(i) = \sum_j \mathbb{P}(S_{t+1} = j \mid S_t = i)\; p(y_{t+1} \mid S_{t+1} = j)\; p(y_{t+2:T} \mid S_{t+1} = j, S_t = i, y_{t+1}).$$

The first two factors are γ_{ij} and $f_j(y_{t+1})$. In the third, assumptions (A1) and (A2) let us drop S_t and y_{t+1}: the future after $t+1$ depends on the past only through S_{t+1}. What remains is $\beta_{t+1}(j)$. □

5.2.2 Combining the two passes

Theorem 5.4 (Two-filter formula)**.** *For every t and i,*

$$p(S_t = i,\, Y_{1:T} = y_{1:T}) \;=\; \tilde{\alpha}_t(i)\,\beta_t(i),$$

and hence the ***smoothing probability*** *is*

$$\gamma_t(i) \;:=\; \mathbb{P}(S_t = i \mid Y_{1:T} = y_{1:T}) \;=\; \frac{\tilde{\alpha}_t(i)\beta_t(i)}{L_T(\theta)}, \qquad L_T(\theta) = \sum_j \tilde{\alpha}_t(j)\beta_t(j) \;\; \textit{for any } t.$$

Proof. Factorise the joint density by splitting the observations at t:

$$p(S_t = i, y_{1:T}) = p(S_t = i, y_{1:t})\; p(y_{t+1:T} \mid S_t = i, y_{1:t}).$$

The first factor is $\tilde{\alpha}_t(i)$. In the second, part 1 of Proposition 3.1 says future observations are conditionally independent of past observations given S_t, so it equals $p(y_{t+1:T} \mid S_t = i) = \beta_t(i)$. □

That the likelihood can be recovered at *any* t is a useful implementation check: if $\sum_j \tilde{\alpha}_t(j)\beta_t(j)$ drifts with t, there is a bug or a scaling inconsistency.

5.2.3 Pairwise probabilities

EM needs expected transition counts, which requires the joint smoothed distribution of consecutive states.

Proposition 5.2 (Pairwise smoothing probabilities)**.** *For $t = 1, \dots, T-1$,*

$$\xi_t(i,j) \;:=\; \mathbb{P}(S_t = i, S_{t+1} = j \mid Y_{1:T} = y_{1:T}) \;=\; \frac{\tilde{\alpha}_t(i)\,\gamma_{ij}\,f_j(y_{t+1})\,\beta_{t+1}(j)}{L_T(\theta)}.$$

Moreover $\sum_j \xi_t(i,j) = \gamma_t(i)$ and $\sum_i \xi_t(i,j) = \gamma_{t+1}(j)$.

The derivation is the two-filter argument with the split made between t and $t+1$, and the four factors have a clean reading: everything up to and including t, the transition, the emission at $t+1$, and everything after $t+1$. The marginal consistency conditions are the second standard implementation check.

5.2.4 Scaled smoothing

Combining scaled forward variables with backward variables requires care, since the two passes must be normalised compatibly. The cleanest route is to scale the backward variables by the *same* constants used in the forward pass: set $\hat{\beta}_T = \mathbf{1}$ and

$$\hat{\beta}_t = \Gamma \mathbf{Q}(y_{t+1})\hat{\beta}_{t+1} / c_{t+1}.$$

Then $\tilde{\alpha}_t(i)\beta_t(i)/L_T = \hat{\alpha}_t(i)\hat{\beta}_t(i)$ exactly, so

$$\gamma_t(i) = \hat{\alpha}_t(i)\hat{\beta}_t(i), \qquad \xi_t(i,j) = \hat{\alpha}_t(i)\gamma_{ij}f_j(y_{t+1})\hat{\beta}_{t+1}(j)/c_{t+1},$$

with no further normalisation needed. The alternative is to normalise $\hat{\beta}_t$ to sum to one at each step and renormalise γ_t and ξ_t afterwards; that is simpler to reason about and equally stable, but loses the exact identity above.

5.3 Decoding: the Viterbi algorithm

Smoothing gives a marginal distribution at each time separately. If a single trajectory is wanted, there are two natural definitions and they do not agree.

The algorithm below was introduced by Viterbi (1967) for decoding convolutional codes, and Forney (1973) identified it as dynamic programming on a trellis — the formulation used here.

Definition 5.3 (Local and global decoding)**. Local (posterior) decoding** takes $\hat{s}_t = \arg\max_i \gamma_t(i)$ independently at each t.

Global (MAP) decoding takes

$$\hat{s}_{1:T} \in \arg\max_{s_{1:T} \in E^T} \mathbb{P}(S_{1:T} = s_{1:T} \mid Y_{1:T} = y_{1:T}) = \arg\max_{s_{1:T}} p(s_{1:T}, y_{1:T}),$$

the denominator being constant in $s_{1:T}$.

Local decoding maximises the expected number of correctly assigned states. Global decoding maximises the probability of getting the entire sequence right. Neither dominates, and they answer different questions.

Example 5.1 (They can differ, and the local path can be impossible)**.** Take $K = 3$ and suppose $\gamma_{13} = 0$. Local decoding might return $\hat{s}_t = 1$ and $\hat{s}_{t+1} = 3$ if state 1 and state 3 each happen to be the marginal mode at their respective times. That path has probability zero under the model. Global decoding cannot produce it, because the transition factor $\gamma_{13} = 0$ annihilates any path containing that step.

The example is not pathological: whenever Γ has structural zeros — common in profile HMMs and in duration-constrained models — local decoding can and does return infeasible paths. When the states are to be read as a trajectory, use Viterbi.

5.3.1 The recursion

Theorem 5.5 (Viterbi algorithm)**.** *Define*

$$m_t(j) = \max_{s_{1:t-1} \in E^{t-1}} p(S_{1:t-1} = s_{1:t-1},\, S_t = j,\, Y_{1:t} = y_{1:t}).$$

Then

$$m_1(j) = \delta_j f_j(y_1), \qquad m_t(j) = f_j(y_t) \max_{i \in E} \{m_{t-1}(i)\gamma_{ij}\} \quad (t \geq 2),$$

and with backpointers $\psi_t(j) \in \arg\max_i \{m_{t-1}(i)\gamma_{ij}\}$*, the path defined by*

$$\hat{s}_T \in \arg\max_j m_T(j), \qquad \hat{s}_t = \psi_{t+1}(\hat{s}_{t+1}) \quad (t = T-1, \dots, 1)$$

is a global maximiser in the sense of Definition 5.3.

Proof. The recursion first. Splitting the maximisation over $s_{1:t-1}$ into the value of s_{t-1} and the rest,

$$m_t(j) = \max_i \max_{s_{1:t-2}} p(s_{1:t-2}, S_{t-1} = i, S_t = j, y_{1:t}) = \max_i \Big[\max_{s_{1:t-2}} p(s_{1:t-2}, S_{t-1} = i, y_{1:t-1})\Big] \gamma_{ij} f_j(y_t),$$

where the factorisation of Theorem 3.1 lets the last two factors be pulled out of the inner maximisation because they do not involve $s_{1:t-2}$. The bracket is $m_{t-1}(i)$.

For the backtracking, note that $\max_{s_{1:T}} p(s_{1:T}, y_{1:T}) = \max_j m_T(j)$, so $\hat{s}_T$ is the last state of some optimal path. Suppose inductively that $\hat{s}_{t+1}$ is the state at time $t+1$ of an optimal path. By the recursion, the maximum over paths ending in $\hat{s}_{t+1}$ at time $t+1$ is attained by extending a path that attains $m_t(\psi_{t+1}(\hat{s}_{t+1}))$, so setting $\hat{s}_t = \psi_{t+1}(\hat{s}_{t+1})$ preserves optimality. This is the Bellman principle of optimality: every suffix-optimal path has prefix-optimal prefixes. □

In practice one runs the recursion in logarithms, turning products into sums:

$$\log m_t(j) = \log f_j(y_t) + \max_i \{\log m_{t-1}(i) + \log \gamma_{ij}\}.$$

No log-sum-exp is needed, because max commutes with log. Viterbi is therefore numerically better behaved than the forward algorithm, and structural zeros are handled by $-\infty$ entries without special cases.

5.3.2 The semiring view

Compare the two recursions:

$$\tilde{\alpha}_t(j) = f_j(y_t) \sum_i \tilde{\alpha}_{t-1}(i)\gamma_{ij}, \qquad m_t(j) = f_j(y_t) \max_i m_{t-1}(i)\gamma_{ij}.$$

They are the same computation over different semirings: $(+, \times)$ for the forward algorithm and $(\max, \times)$ for Viterbi. Taking logarithms turns the latter into $(\max, +)$, and negating gives $(\min, +)$, the shortest-path algebra. So Viterbi decoding is literally a shortest-path problem on the trellis graph whose nodes are (t, i) pairs and whose edge weights are $-\log \gamma_{ij} - \log f_j(y_t)$.

The abstraction earns its keep: the same code, with the semiring swapped, computes the likelihood, the MAP path, the n best paths, expected counts, and the entropy of the posterior over paths. Murphy (2012, ch. 17) develops the general message-passing formulation.

5.3.3 Cost

Time is $\mathcal{O}(K^2 T)$, matching forward–backward. Memory is $\mathcal{O}(KT)$, because the backpointers ψ_t must all be retained for the backward pass — worse than the $\mathcal{O}(K)$ of a likelihood-only forward pass. For long sequences, checkpointing recomputes segments of the trellis on demand and trades a constant factor of time for $\mathcal{O}(K\sqrt{T})$ memory.

5.4 Other quantities

Once forward and backward variables are available, most other quantities of interest are one line away.

- **State occupancy.** $\mathbb{E}[\#\{t : S_t = i\} \mid y_{1:T}] = \sum_{t=1}^{T} \gamma_t(i)$.
- **Expected transitions.** $\mathbb{E}[\#\{t : S_{t-1} = i, S_t = j\} \mid y_{1:T}] = \sum_{t=2}^{T} \xi_{t-1}(i,j)$. These two are exactly the expected sufficient statistics that EM requires in Chapter 6.
- **Conditional simulation.** The state sequence can be drawn from $p(s_{1:T} \mid y_{1:T})$ by sampling $S_T \sim \gamma_T(\cdot)$ and then, backwards, $\mathbb{P}(S_t = i \mid S_{t+1} = j, y_{1:T}) \propto \tilde{\alpha}_t(i)\gamma_{ij}$. This is the forward filtering–backward sampling algorithm, developed in Chapter 9, where it is the workhorse of every MCMC scheme for HMMs.
- **Pseudo-residuals.** With $u_t = F(y_t \mid y_{1:t-1})$ the predictive CDF evaluated at the observation, the $\Phi^{-1}(u_t)$ are approximately standard normal under correct specification, giving the diagnostic plots of Zucchini et al. (2016, ch. 6).

Check your understanding. You compute $\sum_i \tilde{\alpha}_t(i)\beta_t(i)$ at $t = 10$ and get 3.2×10^{-41}, and at $t = 400$ you get 0. What has happened, and is the model wrong?

Show answer

Nothing is wrong with the model; the implementation has underflowed. Theorem 5.4 says that sum equals $L_T(\theta)$ for *every* t, so a value that changes with t — let alone collapses to zero — is a numerical failure, not a statistical one.
The forward variables decay geometrically, so by $t = 400$ they are below the smallest representable double. Fix it with the scaling of Proposition 5.1 or by working in the log domain; Chapter 7 compares the two. That this invariant catches the bug immediately is why it is the first implementation check listed there.

5.5 Summary

- Filtering, smoothing and decoding answer three different questions and are computed by three different recursions, all costing $\mathcal{O}(K^2T)$.
- The forward recursion (Theorem 5.1) is the left-to-right evaluation of the matrix product for the likelihood; its scale factors are the one-step predictive densities and their logarithms sum to ℓ_T (Proposition 5.1).
- Forward and backward variables multiply to give smoothed marginals (Theorem 5.4) and, with one transition and one emission factor inserted, pairwise probabilities (Proposition 5.2).
- Viterbi maximises instead of summing (Theorem 5.5). Its output can differ from the sequence of marginal modes, and unlike that sequence it is always a feasible path (Example 5.1).
- Under a Doeblin condition, filtering forgets its initial distribution geometrically fast (Theorem 5.2), which is why δ matters so little in long series.

5.6 Further reading

Zucchini et al. (2016, chs. 4–5) give the same algorithms with scaling handled exactly as here, plus R implementations. Rabiner (1989) remains the most readable algorithmic account and

fixed the notation most of the field still uses. Cappé et al. (2005, ch. 3) treats the recursions as operator equations on measures, which is the right level of generality for Chapter 12. Murphy (2012, ch. 17) presents them as message passing on a factor graph and makes the semiring structure explicit. Forney (1973) remains the clearest account of the trellis view of decoding. For filter stability, Del Moral and Guionnet (2001) and Cappé et al. (2005, ch. 4).

6 Maximum likelihood and EM

Chapter 5 assumed the parameters were known. This chapter drops that assumption. The likelihood is available in $\mathcal{O}(K^2T)$ from the forward recursion, so in principle any optimiser will do; in practice the dedicated algorithm — EM, known in this context as Baum–Welch — has properties worth understanding, and its failure modes are as instructive as its guarantees.

Two things are worth stating at the outset. EM never decreases the likelihood, and that is a theorem with a two-line proof. EM does not find the global maximum, and no amount of care changes that. Both facts follow from the same decomposition.

By the end of this chapter you will be able to

- derive the EM algorithm as coordinate ascent on a lower bound, and prove it cannot decrease the likelihood;
- write down the Baum–Welch updates and explain why two of them are closed-form;
- diagnose the standard failure modes: local maxima, symmetric initialisation, collapsing states, slow convergence;
- choose between EM and direct numerical maximisation for a given problem, and obtain standard errors either way.

6.1 Maximum likelihood

Definition 6.1 (Maximum likelihood estimator)**.** Given $y_{1:T}$, a **maximum likelihood estimator** is any

$$\hat{\theta}_T \in \arg\max_{\theta \in \Theta} \ell_T(\theta), \qquad \ell_T(\theta) = \log p_\theta(y_{1:T}),$$

with Θ the parameter space of Chapter 3.

The indefinite article is deliberate. By Proposition 3.2 the maximiser is never unique: each of the $K!$ relabellings attains the same value. And by Proposition 3.3 the surface has local maxima that are not relabellings of one another.

6.1.1 Direct numerical maximisation

The forward recursion with scaling returns $\ell_T(\theta)$ for any θ, so a general-purpose quasi-Newton optimiser can be pointed at it. The constraints are handled by reparameterisation, as in Chapter 3: transition probabilities through a multinomial logit, variances through logarithms, and so on. Zucchini et al. (2016, ch. 3) treat this as the primary method and give it careful practical treatment, which is a useful corrective to literature that presents EM as the only option.

Direct maximisation has real advantages. It converges superlinearly near the optimum where EM crawls, it produces a numerical Hessian and hence standard errors as a by-product, and it accommodates constraints and covariate effects that would complicate an M-step. Its disadvantages are that it can step outside the parameter space if the reparameterisation is

careless, and that it offers no monotonicity guarantee, so a bad step can move the iterate to a worse point.

6.1.2 The gradient

Even when using EM, it is worth knowing that the score has a closed form in terms of quantities already computed.

Proposition 6.1 (Fisher's identity)**.** *Under regularity conditions permitting differentiation under the integral sign,*

$$\nabla_\theta \, \ell_T(\theta) \;=\; \mathbb{E}_\theta\big[\nabla_\theta \log p_\theta(S_{1:T}, Y_{1:T}) \;\big|\; Y_{1:T} = y_{1:T}\big].$$

Proof. Write $p_\theta(y) = \int p_\theta(s, y)\, ds$ and differentiate:

$$\nabla_\theta \log p_\theta(y) = \frac{\nabla_\theta p_\theta(y)}{p_\theta(y)} = \int \frac{\nabla_\theta p_\theta(s, y)}{p_\theta(y)}\, ds = \int \frac{\nabla_\theta p_\theta(s, y)}{p_\theta(s, y)} \cdot \frac{p_\theta(s, y)}{p_\theta(y)}\, ds,$$

which is the conditional expectation of $\nabla_\theta \log p_\theta(s, y)$ under $p_\theta(s \mid y)$. □

The complete-data log-likelihood 3.1 is a sum of terms depending on one or two states at a time, so its conditional expectation involves only $\gamma_t(i)$ and $\xi_t(i, j)$ — the smoothed quantities from Chapter 5. A single forward–backward pass therefore yields both the likelihood and its gradient, which is what makes gradient-based optimisation of HMMs cheap.

6.2 The EM algorithm

6.2.1 General form

Definition 6.2 (EM algorithm)**.** With observed data Y, latent data S and current iterate $\theta^{(k)}$:

E-step. Compute $\;Q(\theta \mid \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}\big[\log p_\theta(S, Y) \;\big|\; Y\big]$.

M-step. Set $\;\theta^{(k+1)} \in \arg\max_{\theta \in \Theta} Q(\theta \mid \theta^{(k)})$.

The algorithm is due to Dempster et al. (1977) in general and, for HMMs specifically, predates it: Baum et al. (1970) derived the recursions and proved monotonicity for probabilistic functions of Markov chains.

6.2.2 Why it works

The mechanism is a single identity. For any distribution q over the latent states,

$$\log p_\theta(y) \;=\; \underbrace{\mathbb{E}_q\big[\log p_\theta(S, y)\big] + H(q)}_{\mathcal{F}(q, \theta)} \;+\; \mathrm{KL}\big(q \,\big\|\, p_\theta(\cdot \mid y)\big), \tag{6.1}$$

where $H(q) = -\mathbb{E}_q[\log q(S)]$. To verify it, substitute $p_\theta(s, y) = p_\theta(s \mid y) p_\theta(y)$ into $\mathcal{F}$ and observe that the $\log p_\theta(y)$ term comes out of the expectation as a constant.

Since $\mathrm{KL} \geq 0$ by Proposition 1.3, $\mathcal{F}(q, \theta)$ is a lower bound on the log-likelihood — the *evidence lower bound* — and it is tight exactly when $q = p_\theta(\cdot \mid y)$.

Theorem 6.1 (EM increases the likelihood)**.** *The EM iterates satisfy* $\ell(\theta^{(k+1)}) \geq \ell(\theta^{(k)})$ *for every* k*, with equality only if* $\theta^{(k+1)}$ *leaves* $Q(\cdot \mid \theta^{(k)})$ *at the same value.*

Proof. Set $q^{(k)} = p_{\theta^{(k)}}(\cdot \mid y)$. The E-step makes the bound tight: $\mathcal{F}(q^{(k)}, \theta^{(k)}) = \ell(\theta^{(k)})$, because the KL term vanishes. The M-step maximises $\mathcal{F}(q^{(k)}, \cdot)$, since $\mathcal{F}(q^{(k)}, \theta) = Q(\theta \mid \theta^{(k)}) + H(q^{(k)})$ and the entropy does not involve θ. Therefore

$$\ell(\theta^{(k+1)}) \;\geq\; \mathcal{F}(q^{(k)}, \theta^{(k+1)}) \;\geq\; \mathcal{F}(q^{(k)}, \theta^{(k)}) \;=\; \ell(\theta^{(k)}),$$

the first inequality because $\mathcal{F}$ bounds ℓ below and the second because $\theta^{(k+1)}$ maximises it. □

Read as coordinate ascent on $\mathcal{F}(q, \theta)$ — E-step over q, M-step over θ — the algorithm loses its air of mystery, and the generalisations follow immediately. Restricting q to a tractable family gives variational EM; improving rather than maximising in either step gives generalised EM, and Theorem 6.1 survives unchanged since only the inequality $\mathcal{F}(q^{(k)}, \theta^{(k+1)}) \geq \mathcal{F}(q^{(k)}, \theta^{(k)})$ was used. This is the view of Neal and Hinton (1998), and it is the bridge to the approximate methods of Chapter 7.

6.2.3 The E-step for HMMs

Taking the expectation of 3.1 term by term under $p_{\theta^{(k)}}(\cdot \mid y_{1:T})$,

$$Q(\theta \mid \theta^{(k)}) \;=\; \underbrace{\sum_{i=1}^{K} \gamma_1^{(k)}(i) \log \delta_i}_{\text{initial}} \;+\; \underbrace{\sum_{t=2}^{T} \sum_{i,j} \xi_{t-1}^{(k)}(i,j) \log \gamma_{ij}}_{\text{transitions}} \;+\; \underbrace{\sum_{t=1}^{T} \sum_{i} \gamma_t^{(k)}(i) \log f_i(y_t; \phi_i)}_{\text{emissions}},$$

where $\gamma_t^{(k)}$ and $\xi_t^{(k)}$ are the smoothed and pairwise probabilities of Chapter 5 evaluated at $\theta^{(k)}$. So the E-step is exactly one forward–backward pass.

The structure to notice is that the three groups of parameters appear in three separate terms. The M-step therefore decouples into three independent maximisations.

6.2.4 The M-step

Proposition 6.2 (Baum–Welch updates)**.** *Write* $\hat{n}_{ij} = \sum_{t=2}^{T} \xi_{t-1}^{(k)}(i,j)$ *for the expected transition counts and* $w_t(i) = \gamma_t^{(k)}(i)$ *for the smoothed weights. The M-step gives*

$$\delta_i^{(k+1)} = \gamma_1^{(k)}(i), \qquad \gamma_{ij}^{(k+1)} = \frac{\hat{n}_{ij}}{\sum_{l=1}^{K} \hat{n}_{il}},$$

and $\phi_i^{(k+1)}$ *maximises the weighted log-likelihood* $\sum_{t=1}^{T} w_t(i) \log f_i(y_t; \phi_i)$ *separately for each* i*.*

Proof. The three terms of Q involve disjoint parameter blocks, so each is maximised separately. For the initial distribution, maximise $\sum_i \gamma_1^{(k)}(i) \log \delta_i$ subject to $\sum_i \delta_i = 1$. The Lagrangian gives $\gamma_1^{(k)}(i)/\delta_i = \lambda$, so $\delta_i \propto \gamma_1^{(k)}(i)$, and since the $\gamma_1^{(k)}(i)$ already sum to one, $\delta_i = \gamma_1^{(k)}(i)$. The same argument applied row by row to $\sum_j \hat{n}_{ij} \log \gamma_{ij}$ subject to $\sum_j \gamma_{ij} = 1$ gives the transition update. The emission term is already in the stated form. □

Both closed-form updates are instances of the same fact: maximising $\sum_j c_j \log p_j$ over the simplex returns the normalised weights $p_j = c_j / \sum_l c_l$. The complete-data problem of Chapter 3 was counting; the incomplete-data problem is the same counting with expected counts in place of observed ones.

For Gaussian emissions the weighted problem has the expected solution,

$$\mu_i^{(k+1)} = \frac{\sum_t w_t(i) y_t}{\sum_t w_t(i)}, \qquad (\sigma_i^2)^{(k+1)} = \frac{\sum_t w_t(i) (y_t - \mu_i^{(k+1)})^2}{\sum_t w_t(i)},$$

and for Poisson emissions $\lambda_i^{(k+1)} = \sum_t w_t(i) y_t / \sum_t w_t(i)$. More generally, for an exponential family in natural form the M-step solves $\nabla A(\eta_i) = \sum_t w_t(i) T(y_t) / \sum_t w_t(i)$, matching model moments to weighted empirical ones. Zucchini et al. (2016, ch. 4) tabulate the updates for the families in common use.

> ⚠ Stationarity breaks the closed form
>
> If δ is constrained to equal the stationary distribution of Γ, the initial and transition terms of Q are no longer separate, since δ becomes a function of Γ. The M-step then requires numerical optimisation over Γ. Some implementations simply ignore the constraint during the M-step and re-impose it afterwards; this breaks the monotonicity guarantee of Theorem 6.1. Direct numerical maximisation of ℓ_T is the cleaner option when stationarity matters.

6.2.5 Convergence

Theorem 6.2 (Limit points are stationary points)**.** *Suppose ℓ is continuous on Θ, $Q(\cdot \mid \cdot)$ is continuous in both arguments, and the level set $\{\theta : \ell(\theta) \geq \ell(\theta^{(0)})\}$ is compact. Then the sequence $\ell(\theta^{(k)})$ converges monotonically to some ℓ^*, and every limit point of $\{\theta^{(k)}\}$ is a stationary point of ℓ.*

This is Wu (1983), correcting the convergence claims of Dempster et al. (1977). Three qualifications matter more than the theorem.

The likelihood *values* converge; the *iterates* need not. Convergence of the sequence $\theta^{(k)}$ itself requires extra conditions, and on a ridge — which occurs whenever two states are nearly indistinguishable — the iterates can drift for a long time while ℓ barely moves.

A stationary point is not a maximum. Saddle points are stationary, and EM initialised at a symmetric configuration will sit on one indefinitely: if two states start with identical parameters, their E-step weights are identical and the M-step keeps them identical, forever. Random perturbation of the starting values is not a nicety.

The rate is linear, with a factor governed by the ratio of missing to complete information. When states are well separated the smoothed probabilities $\gamma_t(i)$ are near 0 or 1, little information is missing, and EM converges quickly. When states overlap, convergence is slow — and this is exactly the regime where the answer is most uncertain. Switching to a quasi-Newton method after a few dozen EM iterations is a standard remedy.

6.2.6 Initialisation

Since the algorithm finds a local maximum near where it starts, the starting point is part of the method rather than an implementation detail. What works in practice:

- Run from many dispersed starts and keep the best result. This is the only reliable protection against local maxima, and the number of starts should scale with K.
- Initialise emission parameters by clustering the observations, ignoring time order, and set Γ to a diagonally dominant matrix reflecting expected persistence.
- Fit K states by starting from a good $K-1$ state fit with one state split, rather than from scratch.
- Watch for degeneracy. A state whose expected occupancy $\sum_t \gamma_t(i)$ falls towards zero, or whose variance collapses, signals either too many states or the unboundedness discussed in Chapter 4.

Check your understanding. You run EM and the log-likelihood goes -812.4, -806.1, -806.9, -804.2. What do you conclude?

Show answer

There is a bug. Theorem 6.1 says the log-likelihood cannot decrease at any iteration, so the drop from -806.1 to -806.9 is impossible for a correctly implemented EM.
The usual culprits, in order of likelihood: a stationarity constraint being imposed after the M-step rather than inside it, which breaks the monotonicity guarantee (see the warning above); an M-step that maximises the wrong objective, often by using filtered $\alpha_t(i)$ where smoothed $\gamma_t(i)$ is required; or a mismatch between the scaling used in the forward and backward passes, so that $\xi_t(i,j)$ is wrong by a t-dependent factor. Check the marginal consistency conditions of Proposition 5.2 first — they localise the last of these immediately.

6.3 Standard errors

The observed information matrix $-\nabla^2 \ell_T(\hat{\theta}_T)$ inverts to an approximate covariance, justified by the asymptotic theory of Chapter 8. EM does not produce it, which is a genuine drawback. Three ways round it:

- **Numerical differentiation** of ℓ_T, cheap to implement since the forward recursion evaluates ℓ_T at any θ, but sensitive to step size in high dimensions.
- **Louis's identity**, which expresses the observed information as the complete-data information minus the conditional variance of the complete-data score. Both pieces are computable from forward–backward quantities.
- **The parametric bootstrap**, simulating series from $\hat{\theta}_T$ and refitting. Expensive, but it also captures the finite-sample behaviour of the estimator, which is what Zucchini et al. (2016, ch. 3) recommend when the asymptotics look doubtful.

On the boundary — a transition probability estimated at zero, say — none of these are valid, because the asymptotic normality argument assumes an interior maximum. Bootstrap intervals are the pragmatic choice there, with the caveat that they too degrade.

6.4 Summary

- The likelihood is computable in $\mathcal{O}(K^2T)$, so both direct numerical maximisation and EM are viable; the choice is a trade-off between monotonicity and speed near the optimum.
- EM is coordinate ascent on the evidence lower bound 6.1, which is why it cannot decrease the likelihood (Theorem 6.1).
- The E-step is one forward–backward pass; the M-step decouples into three independent problems, two of which have closed-form solutions obtained by normalising expected counts (Proposition 6.2).
- Limit points are stationary points, not maxima (Theorem 6.2). Multiple dispersed starts are the only protection against local optima, and symmetric initialisation is a trap.
- Standard errors require extra work: numerical Hessians, Louis's identity, or the bootstrap.

6.5 Further reading

Zucchini et al. (2016, chs. 3–4) cover both direct maximisation and EM with practical guidance and R code. Cappé et al. (2005, chs. 10–11) give the rigorous treatment, including Louis's identity and recursive variants. Dempster et al. (1977) is the original EM paper and Wu (1983) the definitive convergence analysis; Neal and Hinton (1998) is the source of the lower-bound view used here and the shortest route to understanding the generalisations. Rydén (2008) compares EM against MCMC on the same problems and is a useful reality check before committing to either.

7 Computation and numerical practice

The algorithms of Chapter 5 are exact in exact arithmetic. In floating point they are not, and a naive implementation of the forward recursion fails on sequences of moderate length. This chapter covers what has to be done to make the mathematics survive contact with a computer, together with the complexity facts that determine what is feasible.

Zucchini et al. (2016) give these matters unusual and welcome attention; most of what follows makes their advice explicit and adds the error analysis.

By the end of this chapter you will be able to

- explain why the forward recursion underflows, and fix it by scaling or by working in logarithms;
- state the time and memory cost of each algorithm, and exploit sparsity in Γ when it is there;
- vectorise the recursions so the inner loop is linear algebra rather than scalar arithmetic;
- test an implementation using four invariants that hold regardless of the data.

7.1 Underflow and the two remedies

7.1.1 The problem

The unnormalised forward variable $\tilde{\alpha}_t(i) = p(S_t = i, y_{1:t})$ is a joint density over t observations. Each step multiplies by a transition probability and an emission density, so its magnitude decays geometrically. If the typical per-step factor is 10^{-2}, then $\tilde{\alpha}_t \sim 10^{-2t}$.

Double precision holds normalised numbers down to about 10^{-308}. Underflow therefore begins around $t \approx 154$ and is complete shortly after: the vector becomes exactly zero, the likelihood is reported as zero, and the log-likelihood as $-\infty$. Sequences of length a few hundred are entirely ordinary, so this is not an edge case.

Note that the problem is intrinsic to the quantity, not to the algorithm. The true value of $\tilde{\alpha}_{500}(i)$ genuinely is around 10^{-1000}, and no rearrangement of the arithmetic will make it representable. Both remedies work by never forming it.

7.1.2 Scaling

Normalise the vector at each step, as in Proposition 5.1: propagate $\hat{\alpha}_t$ summing to one and accumulate $\ell_T = \sum_t \log c_t$.

Cost per step is one vector–matrix product (K^2 multiply–adds), K multiplications for the emission reweighting, K additions for the sum, K divisions, and one logarithm. The overhead beyond the unavoidable K^2 work is $\mathcal{O}(K)$, so scaling is close to free.

Accuracy is good. Each c_t is a predictive density of order one, so its logarithm is computed with full relative precision, and the total error in ℓ_T accumulates as $\mathcal{O}(T\varepsilon_{\text{mach}})$ in absolute terms. For $T = 10^6$ that is around 10^{-10} — negligible against any statistical uncertainty.

Scaling is the default. It is what Zucchini et al. (2016) implement, and it produces the one-step predictive densities needed for residual diagnostics as a by-product.

7.1.3 The log domain

Store logarithms throughout and combine with log-sum-exp:

$$a_{t+1}(j) = \log f_j(y_{t+1}) + \operatorname{logsumexp}_i\{a_t(i) + \log \gamma_{ij}\}, \qquad \operatorname{logsumexp}(z) = m + \log \sum_i e^{z_i - m}, \;\; m = \max_i z_i.$$

Subtracting the maximum guarantees the largest exponential is $e^0 = 1$; terms that underflow to zero were negligible anyway.

Cost per step is K^2 additions, K maxima over K elements, K^2 exponentials and K logarithms. Transcendental functions are perhaps an order of magnitude more expensive than multiplication, so this is several times slower than scaling.

What it buys is robustness in cases where scaling struggles: emission densities that vary over an enormous dynamic range within one time step, and structural zeros in Γ, which become $-\infty$ and propagate correctly without special-casing. It is also the natural representation when the same code must run Viterbi, since that already works additively.

> **Which to use**
>
> Scaling for routine work: faster, accurate, and it yields the predictive densities you want anyway. The log domain when the model has structural zeros, when emissions can be extremely small under some states, or when robustness matters more than speed. Implementing both and checking they agree to a few digits on a test problem is a good use of an afternoon.

7.1.4 Backward and Viterbi

The backward variables need the same treatment, and the cleanest option is the synchronised scaling of Chapter 5: divide $\widehat{\beta}_t$ by the *forward* scale factor c_{t+1}, so that $\gamma_t(i) = \widehat{\alpha}_t(i)\widehat{\beta}_t(i)$ exactly with no further normalisation.

Viterbi needs no special handling beyond working in logarithms, since max commutes with log and no log-sum-exp arises. It is the numerically best-behaved of the three recursions.

7.2 Complexity

7.2.1 Single sequence

Table 7.1: Cost of exact inference

Operation	Time	Memory
Forward, likelihood only	$\mathcal{O}(K^2T)$	$\mathcal{O}(K)$
Forward–backward, smoothing	$\mathcal{O}(K^2T)$	$\mathcal{O}(KT)$
Viterbi	$\mathcal{O}(K^2T)$	$\mathcal{O}(KT)$
FFBS, one trajectory draw	$\mathcal{O}(K^2T)$	$\mathcal{O}(KT)$

The K^2 factor is the vector–matrix product and the T factor is the sweep; both are unavoidable for exact inference in a dense model. The memory asymmetry matters: computing the likelihood needs only the current vector, but anything requiring a backward pass must retain the forward quantities.

For very long sequences, checkpointing stores the forward vector every $\sqrt{T}$ steps and recomputes intermediate segments during the backward pass, giving $\mathcal{O}(K\sqrt{T})$ memory for roughly twice the time. Rarely needed, but the right answer when it is.

7.2.2 Estimation

One EM iteration is one forward–backward pass plus M-step updates that cost $\mathcal{O}(K^2T)$ for the transition counts and $\mathcal{O}(KT)$ for typical emission families. So an iteration is $\mathcal{O}(K^2T)$, and a full fit is that times the number of iterations — which, as Chapter 6 noted, can be large when states overlap.

With N independent sequences of average length T, cost per iteration is $\mathcal{O}(NK^2T)$, and the sequences are independent so the E-step parallelises exactly. This is the easiest speedup available and should be the first one taken.

The quadratic growth in K is what keeps state numbers small. Going from $K = 5$ to $K = 20$ multiplies the cost sixteenfold, and the parameter count — $K(K-1)$ transition parameters alone — grows just as fast, so the statistical problem worsens in step with the computational one.

7.2.3 Exploiting sparsity

If Γ has many structural zeros, as in left-to-right models for speech or profile HMMs for sequence alignment, the recursion touches only the non-zero entries and the cost falls to $\mathcal{O}(|\mathcal{E}|T)$ where $|\mathcal{E}|$ is the number of permitted transitions. For a banded matrix allowing transitions only to nearby states this is $\mathcal{O}(KT)$ — linear in K. Sparse representations are worth the implementation effort whenever the topology is constrained.

7.3 Vectorisation and hardware

The forward recursion is inherently sequential in t: step $t+1$ needs the output of step t. Within a step, however, everything is a dense linear algebra operation, and that is where the performance is.

- **Vectorise over states.** Write the step as $\hat{\alpha}_t^\top = \hat{\alpha}_{t-1}^\top \Gamma \mathbf{Q}(y_t)$ and let a BLAS routine do it, rather than writing loops over i and j. In R or Python this alone can be a hundredfold improvement over the naive double loop.

- **Batch over sequences.** With N sequences, stack their forward vectors into an $N \times K$ matrix and advance all of them with one matrix–matrix product. This converts a memory-bound operation into a compute-bound one and is where GPUs help.
- **Precompute emission matrices.** The values $f_i(y_t)$ for all (i, t) can be computed once as a $T \times K$ matrix before the recursion begins. For expensive emission densities this is a large saving, since EM recomputes them at every iteration otherwise.

7.4 Approximate inference

Exact inference is cheap enough that approximation is needed only outside the finite-state, moderate-K setting.

Beam search for decoding. Retain only the B highest-scoring partial paths at each step, giving $\mathcal{O}(BKT)$. The result is not guaranteed optimal, and the error is unbounded in principle, but with a generous beam the loss is usually negligible. Standard in large-vocabulary speech recognition, where K runs to millions.

Pruning. Drop states whose forward probability falls below a threshold relative to the current maximum. Cheaper than beam search and easier to implement, with the same caveat.

Variational approximation. Restrict q in the bound 6.1 to a tractable family and maximise. For a plain HMM this is unnecessary — the exact posterior over states is already tractable — but for the factorial and switching models of Chapter 12 it is the main option (Ghahramani and Hinton 2000; Wainwright and Jordan 2008).

Sequential Monte Carlo. For continuous state spaces, or discrete spaces too large for $\mathcal{O}(K^2)$ steps.

Online EM. For streaming data, update sufficient statistics recursively with a decreasing gain η_t rather than sweeping the whole series each iteration. Convergence rests on the ergodicity assumed in Chapter 8. Cappé et al. (2005, ch. 11) develops the theory.

7.5 Implementation checks

Bugs in HMM code are easy to write and hard to see, because wrong answers still look like probabilities. Four checks catch most of them.

1. **Likelihood consistency across t.** By Theorem 5.4, $\sum_i \tilde{\alpha}_t(i)\beta_t(i)$ equals L_T for *every* t. Compute it at several t and confirm agreement. This catches scaling errors immediately.
2. **Marginal consistency of pairwise probabilities.** Proposition 5.2 requires $\sum_j \xi_t(i, j) = \gamma_t(i)$ and $\sum_i \xi_t(i, j) = \gamma_{t+1}(j)$.
3. **Monotonicity of EM.** By Theorem 6.1 the log-likelihood cannot decrease. If it does, the M-step is wrong, or a constraint such as stationarity is being imposed incorrectly — see the warning in Chapter 6.
4. **Recovery on simulated data.** Simulate from known parameters with T large and confirm the estimates converge to them, up to a permutation of the state labels. This is the only check that tests the model as a whole, and it is worth automating.

A fifth, for samplers: run with no data and confirm the output reproduces the prior.

7.6 Summary

- Forward variables underflow within a few hundred time steps. Scaling (fast, accurate, gives predictive densities) and log-domain arithmetic (slower, more robust, handles structural zeros) both fix it exactly.
- Exact inference costs $\mathcal{O}(K^2T)$ time; memory is $\mathcal{O}(K)$ for the likelihood alone and $\mathcal{O}(KT)$ once a backward pass is needed.
- Sparsity in Γ reduces the cost to $\mathcal{O}(|\mathcal{E}|T)$ and is worth exploiting when the topology is constrained.
- The recursion is sequential in t but vectorises perfectly over states and batches over sequences.
- Four cheap invariants — likelihood consistency across t, marginal consistency of ξ_t, EM monotonicity, and recovery on simulated data — catch most implementation errors.

7.7 Further reading

Zucchini et al. (2016, chs. 3–5 and appendices) give practical guidance and working R code with scaling handled exactly as described here. Cappé et al. (2005, chs. 5, 11) cover numerical considerations and online algorithms rigorously. Murphy (2012, ch. 17) discusses the numerical issues alongside the algorithms. For approximate methods, Wainwright and Jordan (2008) on variational inference and Chopin and Papaspiliopoulos (2020) on sequential Monte Carlo.

8 Asymptotic theory

Does the maximum likelihood estimator of Chapter 6 converge to the truth, and if so how fast? For independent data the answers are classical. For HMMs they are not, because the observations are dependent, the latent states are never seen, and the parameter is identified only up to permutation. Each of these breaks a step of the standard argument, and repairing them took the field roughly thirty years — from Baum and Petrie (1966) to Bickel et al. (1998).

This chapter states the results and sketches the arguments. The technical machinery lives in Cappé et al. (2005, chs. 12–13) and Douc et al. (2014); what follows is the structure of the proofs and, more usefully, the conditions under which they hold.

By the end of this chapter you will be able to

- explain why the normalised log-likelihood is not an average of stationary terms, and how the infinite-past construction repairs that;
- state the consistency and asymptotic normality results, and identify which assumption each one leans on;
- say what changes under misspecification, and why the sandwich covariance replaces the inverse information;
- explain why AIC and BIC are heuristics for choosing the number of states rather than justified criteria.

8.1 The setting

Assume throughout:

(C1) The data are generated by an HMM with parameter $\theta^\star$ in the interior of Θ.

(C2) The hidden chain under $\theta^\star$ is irreducible and aperiodic, with stationary distribution $\pi^\star$ having all entries positive, and the process is started from stationarity, so $(S_t, Y_t)_{t \geq 1}$ is strictly stationary and ergodic.

(C3) Θ is compact, and for every $\theta \in \Theta$ the transition matrix satisfies $\gamma_{ij}(\theta) \geq \epsilon > 0$ for all i, j.

(C4) The emission densities $f_i(y; \phi)$ are positive, continuous in ϕ, and satisfy the integrability condition $\mathbb{E}_{\theta^\star}[\sup_\theta |\log f_i(Y_1; \phi_i)|] < \infty$.

(C5) The model is identifiable up to permutation, in the sense of Theorem 4.1.

> ⚠ These conditions are strong
>
> (C3) is the most restrictive. A uniform lower bound on every transition probability is a Doeblin condition, and it fails for exactly the sparse, structurally constrained transition matrices that motivate many applications — left-to-right models in speech, profile HMMs in bioinformatics. It is assumed because it delivers uniform forgetting via Theorem 2.3,

which is what makes the log-likelihood behave like a sum of weakly dependent terms. Douc et al. (2004) relax it substantially at the cost of considerable technical work. Compactness of Θ in (C1) also excludes the unbounded Gaussian likelihood of Chapter 4, and is typically enforced by bounding variances away from zero.

Because of the permutation symmetry, statements about consistency must be about the equivalence class

$$\mathcal{E}(\theta^\star) = \{\theta^{\star\sigma} : \sigma \in \mathfrak{S}_K\},$$

and convergence means $d(\hat{\theta}_T, \mathcal{E}(\theta^\star)) \to 0$ for a metric d on Θ.

8.2 The limiting log-likelihood

The standard consistency argument compares $T^{-1}\ell_T(\theta)$ with its expectation and shows the limit is uniquely maximised at the truth. For independent data $T^{-1}\ell_T(\theta)$ is an average of i.i.d. terms and the law of large numbers applies at once. Here the terms are dependent, and worse, the natural decomposition

$$\ell_T(\theta) = \sum_{t=1}^{T} \log p_\theta(y_t \mid y_{1:t-1})$$

has summands that are not stationary: the t-th term conditions on a history of growing length.

The fix is to compare with a stationary process that conditions on the *infinite* past.

Theorem 8.1 (Existence of the limit)**.** *Under (C1)–(C4), for each $\theta \in \Theta$ there is a constant $\ell_\infty(\theta)$ with*

$$\bar{\ell}_T(\theta) := \frac{1}{T}\ell_T(\theta) \xrightarrow[T\to\infty]{\text{a.s. and } L^1} \ell_\infty(\theta),$$

and the limit is

$$\ell_\infty(\theta) = \mathbb{E}_{\theta^\star}\big[\log p_\theta(Y_0 \mid Y_{-\infty:-1})\big].$$

Moreover ℓ_∞ is continuous on Θ and the convergence is uniform.

The argument has two halves. First, extend the observation process to a two-sided stationary sequence $(Y_t)_{t\in\mathbb{Z}}$ and show that the conditional density $p_\theta(y_0 \mid y_{-m:-1})$ converges as $m \to \infty$ to a limit $p_\theta(y_0 \mid y_{-\infty:-1})$, uniformly and geometrically fast. This is precisely filter forgetting: by (C3) and Theorem 5.2, the predictive distribution at time 0 depends on observations far in the past only through an error decaying like $(1-\epsilon)^m$. Second, the resulting sequence $\log p_\theta(Y_t \mid Y_{-\infty:t-1})$ is stationary and ergodic, so Birkhoff's ergodic theorem applies, and the geometric approximation error is summable, so replacing the finite past by the infinite past changes the average by $o(1)$.

The result is due to Leroux (1992), who obtained it via Kingman's subadditive ergodic theorem, an elegant route that avoids the forgetting argument entirely: ℓ_T is almost superadditive in T.

The interpretation of $\ell_\infty(\theta)$ is worth stating plainly. It is the expected log predictive density of one new observation given the infinite past, when the data come from $\theta^\star$ but the prediction is made using θ. It is a scoring rule, evaluated in expectation.

8.3 Consistency

Theorem 8.2 (Strong consistency of the MLE)**.** *Under (C1)–(C5),*

$$\ell_\infty(\theta) \le \ell_\infty(\theta^\star) \quad \text{for all } \theta \in \Theta,$$

with equality if and only if $\theta \in \mathcal{E}(\theta^\star)$*. Consequently* $d(\widehat{\theta}_T, \mathcal{E}(\theta^\star)) \to 0$ *almost surely.*

Proof. The inequality is an information inequality. Writing $p^\star$ for the true conditional density,

$$\ell_\infty(\theta^\star) - \ell_\infty(\theta) = \mathbb{E}_{\theta^\star}\left[\log \frac{p^\star(Y_0 \mid Y_{-\infty:-1})}{p_\theta(Y_0 \mid Y_{-\infty:-1})}\right] = \mathbb{E}_{\theta^\star}\big[\mathrm{KL}(p^\star(\cdot \mid Y_{-\infty:-1}) \,\|\, p_\theta(\cdot \mid Y_{-\infty:-1}))\big],$$

which is non-negative by Proposition 1.3, and zero only if the two predictive densities agree almost surely. Agreement for all conditioning histories means the two parameters induce the same law for the observed process, and (C5) then forces $\theta \in \mathcal{E}(\theta^\star)$.

For the conclusion, note $\bar{\ell}_T(\widehat{\theta}_T) \ge \bar{\ell}_T(\theta^\star)$ by definition of the maximiser. Uniform convergence in Theorem 8.1 lets us pass to the limit on both sides, giving $\limsup_T \ell_\infty(\widehat{\theta}_T) \ge \ell_\infty(\theta^\star)$. Since ℓ_∞ is continuous, Θ compact, and the maximiser set of ℓ_∞ is exactly $\mathcal{E}(\theta^\star)$, any limit point of $\widehat{\theta}_T$ lies in $\mathcal{E}(\theta^\star)$. □

The shape of the argument is the classical Wald one; all the HMM-specific difficulty is buried in Theorem 8.1. Note also what the theorem does not say: it is a statement about the *global* maximiser, and Chapter 6 showed that EM returns a local one. Consistency of the global MLE is not consistency of what an algorithm produces.

8.4 Asymptotic normality

Definition 8.1 (Score and information)**.** The score is $U_T(\theta) = \nabla_\theta \ell_T(\theta)$ and the Fisher information per observation is

$$I(\theta^\star) = \lim_{T\to\infty} \frac{1}{T} \mathrm{Var}_{\theta^\star}(U_T(\theta^\star)),$$

assumed finite and non-singular.

Theorem 8.3 (Central limit theorem for the score)**.** *Under (C1)–(C5) and suitable smoothness,* $T^{-1/2} U_T(\theta^\star) \xrightarrow{d} \mathcal{N}(0, I(\theta^\star))$.

The proof is a martingale argument. Decompose the score as a telescoping sum of increments,

$$U_T(\theta^\star) = \sum_{t=1}^{T} \nabla_\theta \log p_{\theta^\star}(Y_t \mid Y_{1:t-1}),$$

and note each increment has conditional mean zero given $\mathcal{Y}_{t-1}$, so the sum is a martingale with respect to the observation filtration. Replacing the finite conditioning history by the infinite one — legitimate by the same geometric forgetting used before — makes the increments stationary and ergodic, and the martingale central limit theorem applies. The information $I(\theta^\star)$ is the variance of a single increment of the stationary version, which is why it can be written as an infinite sum over lags of score autocovariances. Bickel et al. (1998) give the argument for general state spaces; Douc et al. (2004) extend it to autoregressive regime-switching models.

Theorem 8.4 (Asymptotic normality of the MLE)**.** *Under the above conditions, with $\hat{\theta}_T$ a consistent sequence of local maximisers converging to a fixed representative $\theta^\star$ of its equivalence class,*

$$\sqrt{T}(\hat{\theta}_T - \theta^\star) \xrightarrow{d} \mathcal{N}(0,\, I(\theta^\star)^{-1}).$$

Proof. Since $\theta^\star$ is interior and $\hat{\theta}_T \to \theta^\star$, eventually $U_T(\hat{\theta}_T) = 0$. A first-order Taylor expansion about $\theta^\star$ gives

$$0 = U_T(\theta^\star) + \nabla^2 \ell_T(\tilde{\theta}_T)(\hat{\theta}_T - \theta^\star)$$

for some $\tilde{\theta}_T$ on the segment between them. Dividing by $\sqrt{T}$,

$$\sqrt{T}(\hat{\theta}_T - \theta^\star) = -\Big(\tfrac{1}{T}\nabla^2 \ell_T(\tilde{\theta}_T)\Big)^{-1} \cdot \tfrac{1}{\sqrt{T}} U_T(\theta^\star).$$

A uniform law of large numbers for the Hessian gives $T^{-1}\nabla^2 \ell_T(\tilde{\theta}_T) \to -I(\theta^\star)$ in probability, using consistency to handle $\tilde{\theta}_T$. Slutsky's theorem combined with Theorem 8.3 yields the stated limit. □

Interiority is essential and is often violated in practice. If a transition probability is genuinely zero, $\theta^\star$ sits on the boundary of Θ, the expansion is invalid, and the limiting distribution becomes a projection of a Gaussian onto a cone rather than a Gaussian. Confidence intervals computed from the inverse Hessian will then be wrong, typically too wide.

8.4.1 Computing the information

Three routes, in increasing order of reliability and cost.

The **observed information** $-\nabla^2 \ell_T(\hat{\theta}_T)$, computed by numerical differentiation of the forward recursion, is the usual choice and is what Zucchini et al. (2016) recommend. **Louis's identity** expresses the same quantity in terms of complete-data quantities available from forward–backward, avoiding numerical differentiation. The **parametric bootstrap** sidesteps the asymptotics entirely and remains informative when the regularity conditions are doubtful, at the cost of refitting the model many times.

8.5 Misspecification

If the data-generating process is not an HMM in the assumed class, there is no $\theta^\star$. The theory still says something.

Definition 8.2 (Pseudo-true parameter)**.** Let $\mathbb{P}^{\text{true}}$ be the law of the observed process. A **pseudo-true parameter** is

$$\theta^\circ \in \arg\max_{\theta \in \Theta} \mathbb{E}^{\text{true}}\big[\log p_\theta(Y_0 \mid Y_{-\infty:-1})\big],$$

equivalently a minimiser of the KL divergence rate from $\mathbb{P}^{\text{true}}$ to $\mathbb{P}_\theta$.

Under conditions parallel to (C1)–(C4) — now imposed on the true process rather than on a member of the model class — Theorem 8.1 still holds and $\hat{\theta}_T \to \theta^\circ$ almost surely. Asymptotic normality also survives, but with the sandwich covariance

$$\sqrt{T}(\hat{\theta}_T - \theta^\circ) \xrightarrow{d} \mathcal{N}(0,\, H^{-1} V H^{-1}),$$

where H is the limiting Hessian and V the limiting score variance at θ°. Under correct specification the information equality $H = V$ collapses this to I^{-1}; under misspecification it does not, and using I^{-1} understates the uncertainty. This is White (1982) adapted to dependent data.

The practical message is that a fitted HMM should be read as the best available approximation in KL divergence, not as a recovered truth, and that reported standard errors should be of sandwich form unless there is reason to trust the specification.

8.6 Choosing the number of states

K is a discrete parameter and the theory above says nothing about it. It is also the hardest choice in practice.

The difficulty is that the models are nested in a degenerate way. A K-state model sits inside a $(K+1)$-state model on a boundary — either one state has zero stationary probability, or two states have identical parameters — and in both cases the parameter is not identified at that point. The regularity conditions for likelihood ratio tests fail, and the χ^2 asymptotics do not apply. Nor is there a simple replacement: the limiting distribution depends on the geometry of the non-identified set.

Definition 8.3 (Information criteria)**.** For a fitted model with d free parameters,

$$\mathrm{AIC} = -2\ell_T(\hat{\theta}_T) + 2d, \qquad \mathrm{BIC} = -2\ell_T(\hat{\theta}_T) + d \log T.$$

Both are used routinely and both should be treated as heuristics here, since the derivations of Akaike (1974) and Schwarz (1978) assume regularity conditions this problem violates. Empirically, AIC tends to select too many states and BIC fewer, and neither is reliable when the model is misspecified — which, for a model as stylised as an HMM, is the normal case. Pohle et al. (2017) document the pattern carefully and argue against relying on either in isolation.

More honest alternatives:

- **Cross-validated likelihood**, holding out blocks of the series and scoring one-step predictions (Celeux and Durand 2008). Slower, but it measures what is usually wanted.
- **Model checking** through pseudo-residuals and posterior predictive summaries, asking whether a fitted model reproduces the features of the data that matter, rather than which model has the best score.
- **Treating K as a parameter** with a prior and reporting its posterior. This is the Bayesian route, and it is the subject of Chapter 11, where the telescoping sampler makes it computationally routine. It also shifts the question in a useful way: from "how many states are in the model" to "how many states are occupied by the data", a distinction that Chapter 11 takes seriously.

A modelling remark that carries more weight than any criterion: states are chosen to be interpretable as well as to fit. A three-state movement model whose states correspond to resting, foraging and travelling is more useful than a seven-state model with a better BIC and no interpretation. Pohle et al. (2017) make this argument at length.

8.7 Summary

- The normalised log-likelihood converges almost surely to a limit that is the expected log predictive density under the truth (Theorem 8.1), proved by combining filter forgetting with the ergodic theorem.
- That limit is maximised exactly on the permutation class of $\theta^\star$, giving strong consistency of the global MLE (Theorem 8.2).
- The score is a martingale, so a martingale CLT gives asymptotic normality with the inverse Fisher information as covariance (Theorem 8.4) — provided $\theta^\star$ is interior.
- Under misspecification the MLE converges to the KL-closest parameter and the covariance takes sandwich form.
- The number of states falls outside this theory. Information criteria are heuristics here; cross-validation, model checking, and putting a prior on K (Chapter 11) are the alternatives.

8.8 Further reading

Cappé et al. (2005, chs. 12–13) is the definitive treatment of consistency and asymptotic normality for HMMs. Leroux (1992) is the original consistency proof and remains readable; Bickel et al. (1998) established asymptotic normality for general state spaces and Douc et al. (2004) extended it to regime-switching autoregressions. Douc et al. (2014) place HMM asymptotics inside the broader theory of nonlinear time series. Vaart (1998) supplies the general machinery — Wald consistency, martingale CLTs, sandwich covariances — that these papers specialise. For the state-number problem, Pohle et al. (2017) and Celeux and Durand (2008).

Part IV

IV. Bayesian and hierarchical inference

9 Bayesian inference

Maximum likelihood returns a point and, with work, a standard error. A Bayesian analysis returns a distribution, and for HMMs it also solves several problems that are awkward otherwise: unbounded likelihoods, states that empty out, boundary parameters, and — the subject of Chapter 11 — an unknown number of states.

The computational story is unusually clean. The latent states are the natural augmentation variable, and conditional on them the model becomes a collection of independent standard problems. The one algorithm that makes this work, forward filtering with backward sampling, is a small modification of the recursions already derived in Chapter 5.

This chapter sets up priors, derives the sampler, and confronts label switching, which is more consequential in the Bayesian setting than the classical one. Frühwirth-Schnatter (2006, chs. 10–11) is the standard reference throughout.

By the end of this chapter you will be able to

- specify conjugate priors for an HMM and explain what the Dirichlet hyperparameters encode;
- derive forward filtering with backward sampling, and say why the states must be drawn as a block;
- write out a full Gibbs sweep and compare each step with the corresponding EM update;
- recognise label switching in MCMC output and choose an appropriate response.

9.1 Priors

Write $\theta = (\delta, \Gamma, \phi_{1:K})$ as before. A prior must be placed on each block.

9.1.1 Transition matrix

Rows of Γ live on the simplex, and the Dirichlet distribution is conjugate to the multinomial counts that arise from an observed state sequence.

Definition 9.1 (Independent Dirichlet rows)**.** Take the rows independently,

$$\Gamma_{j,\cdot} \sim \mathrm{Dir}(e_{j1}, \dots, e_{jK}), \qquad j = 1, \dots, K,$$

with hyperparameters $e_{jk} > 0$. The symmetric choice $e_{jk} = e_0$ for all j, k is the default.

The hyperparameters act as prior transition counts, and their size controls how strongly the posterior is pulled towards uniformity. Two choices are worth distinguishing. Setting $e_0 = 1$ gives a uniform prior on each row, which sounds neutral but in fact favours transition matrices with little persistence — a problem when the states are meant to represent regimes that last. Putting extra mass on the diagonal, $e_{jj} = e_0 + \varkappa$ with $\varkappa > 0$, encodes persistence directly;

this is the "sticky" prior that Fox et al. (2011) introduced for the nonparametric case and it applies equally to finite models.

Small values $e_0 < 1$ favour sparse rows, pushing mass towards transition matrices with near-zero entries. That behaviour is exploited deliberately in Chapter 11.

The initial distribution takes a Dirichlet prior in the same way, $\delta \sim \mathrm{Dir}(d_1, \dots, d_K)$, unless stationarity is imposed, in which case δ is determined by Γ and carries no prior of its own.

9.1.2 Emission parameters

The natural choice is the conjugate prior for the emission family, applied independently across states:

$$\phi_i \overset{\text{iid}}{\sim} G_0, \qquad i = 1, \dots, K.$$

For Gaussian emissions with unknown mean and variance, G_0 is normal–inverse-gamma; for Poisson emissions, gamma; for categorical emissions, Dirichlet. Conjugacy is not required, but it makes the corresponding sampler step a direct draw rather than a Metropolis update.

> **The prior fixes the unbounded likelihood**
>
> Chapter 4 noted that a Gaussian HMM has an unbounded likelihood: sending $\sigma_i \to 0$ with μ_i at an observation drives $\ell_T \to \infty$. A proper inverse-gamma prior on σ_i^2 has a density vanishing at zero, so the posterior remains proper and the degeneracy disappears. This is a real advantage of the Bayesian formulation, not a technicality — it is why Bayesian HMM fits do not need variance floors.

Independence across states in G_0 says that knowing state 1's parameters tells you nothing about state 2's. When there are several related series, that assumption is usually wrong and wasteful; Chapter 10 replaces it with a hierarchical prior that lets the states borrow strength.

9.2 Posterior simulation

The posterior $p(\theta \mid y_{1:T}) \propto p(\theta) L_T(\theta)$ is not tractable, because L_T is a sum over K^T sequences. Data augmentation makes it tractable by treating the states as unknowns to be sampled rather than summed out.

9.2.1 The augmented target

Target the joint posterior of parameters *and* states,

$$p(\theta, s_{1:T} \mid y_{1:T}) \;\propto\; p(\theta)\, \delta_{s_1} f_{s_1}(y_1; \phi_{s_1}) \prod_{t=2}^{T} \gamma_{s_{t-1} s_t} f_{s_t}(y_t; \phi_{s_t}),$$

and read off the two conditionals. Given θ, the states have the distribution that FFBS below can sample from. Given the states, the complete-data structure of Chapter 3 applies, and the parameter blocks separate exactly as they did in the M-step of Proposition 6.2 — with conjugate priors turning "maximise" into "sample".

9.2.2 Forward filtering, backward sampling

The states must be drawn jointly, not one at a time. Sampling S_t from its full conditional given all other states, a single-site Gibbs move, is correct but mixes disastrously: consecutive states are strongly dependent, so the chain moves through state space one time point at a time and takes $\mathcal{O}(T^2)$ sweeps to traverse it. Drawing the whole trajectory in one block removes the problem.

Theorem 9.1 (Forward filtering, backward sampling)**.** *Let $\alpha_t(i) = \mathbb{P}_\theta(S_t = i \mid y_{1:t})$ be the filtering probabilities from Chapter 5. Then*

$$p_\theta(s_{1:T} \mid y_{1:T}) \;=\; p_\theta(s_T \mid y_{1:T}) \prod_{t=T-1}^{1} p_\theta(s_t \mid s_{t+1}, y_{1:t}),$$

where

$$\mathbb{P}_\theta(S_t = i \mid S_{t+1} = j,\, y_{1:t}) \;=\; \frac{\alpha_t(i)\,\gamma_{ij}}{\sum_{l=1}^{K} \alpha_t(l)\gamma_{lj}}.$$

Proof. For the factorisation, apply the chain rule backwards and use part 2 of Proposition 3.1: conditional on S_{t+1}, the state S_t is independent of $S_{t+2:T}$ and of $y_{t+1:T}$, so the conditioning set collapses from $(s_{t+1:T}, y_{1:T})$ to $(s_{t+1}, y_{1:t})$.

For the conditional itself, Bayes' theorem gives

$$\mathbb{P}(S_t = i \mid S_{t+1} = j, y_{1:t}) \propto \mathbb{P}(S_t = i \mid y_{1:t})\, \mathbb{P}(S_{t+1} = j \mid S_t = i, y_{1:t}) = \alpha_t(i)\gamma_{ij},$$

the second factor reducing to γ_{ij} by the Markov property. Normalising over i gives the stated expression. □

So one forward pass storing the filtered vectors $\alpha_1, \dots, \alpha_T$, followed by a backward sweep drawing $S_T \sim \alpha_T$ and then each S_t from a K-point distribution, produces an exact draw from $p_\theta(s_{1:T} \mid y_{1:T})$. Cost is $\mathcal{O}(K^2 T)$ time and $\mathcal{O}(KT)$ memory, the same as forward–backward. The algorithm is due to Frühwirth-Schnatter (1994) and Carter and Kohn (1994) for state-space models and Chib (1996) for the discrete case; no backward *variables* are needed, only backward *sampling*.

9.2.3 The Gibbs sampler

Definition 9.2 (Data augmentation sampler for a finite HMM)**.** Given the current $\theta^{(m)}$, one sweep produces $\theta^{(m+1)}$ as follows.

1. **States.** Draw $s_{1:T}^{(m+1)} \sim p_{\theta^{(m)}}(\cdot \mid y_{1:T})$ by FFBS (Theorem 9.1).
2. **Transitions.** Form the counts $N_{jk} = \#\{t \geq 2 : s_{t-1} = j, s_t = k\}$ and draw the rows independently,
$$\Gamma_{j,\cdot}^{(m+1)} \sim \mathrm{Dir}(e_{j1} + N_{j1}, \dots, e_{jK} + N_{jK}).$$
3. **Initial distribution.** Draw $\delta^{(m+1)} \sim \mathrm{Dir}(d_1 + \mathbf{1}\{s_1 = 1\}, \dots, d_K + \mathbf{1}\{s_1 = K\})$.
4. **Emissions.** For each i, draw $\phi_i^{(m+1)}$ from its conjugate posterior given the observations assigned to state i, namely $\{y_t : s_t = i\}$.

Each step is a standard conjugate update. Step 2 is Dirichlet–multinomial conjugacy applied row by row, which works because rows are a priori independent and the complete-data likelihood factorises across them. Step 4 is K separate conjugate updates, each involving only the observations currently assigned to that state — for Gaussian emissions, a normal–inverse-gamma update using the within-state sample mean and sum of squares.

Comparing with Proposition 6.2 is instructive. The M-step normalised expected counts; the Gibbs step draws from a Dirichlet centred near the normalised realised counts. EM propagates a soft assignment $\gamma_t(i)$; the sampler propagates a hard assignment drawn from the same distribution. The two algorithms have the same skeleton, and Rydén (2008) compares their behaviour on identical problems in detail.

9.2.4 Mixing and diagnostics

Blocking the states is what makes the sampler work at all, but it does not make it fast in every regime. Mixing degrades when states overlap heavily, since the assignments then change little from sweep to sweep, and when the transition matrix is very persistent, since a long run in one state resists being reassigned. Standard remedies are to monitor the number of occupied states and the log-likelihood rather than individual parameters, to run several chains from dispersed starts, and — where the emission model permits — to integrate out ϕ analytically and sample the states from the collapsed target, which usually improves mixing at the cost of losing the conditional independence that makes step 4 cheap.

9.3 Label switching

Definition 9.3 (The problem)**.** By Proposition 3.2 the likelihood is invariant under relabelling. If the prior is also invariant — which it is under symmetric Dirichlet rows and exchangeable G_0 — then the posterior has $K!$ identical modes, and any state-specific posterior summary is meaningless: $\mathbb{E}[\phi_1 \mid y]$, $\mathbb{E}[\phi_2 \mid y], \ldots$ are all equal to the same average over all states.

This is worse than the classical version of the problem. An optimiser converges to one mode and stays there, so the symmetry is invisible. A sampler that mixes *properly* must visit all $K!$ modes, and then the marginal posterior of ϕ_1 is a K-fold symmetric mixture. Paradoxically, a sampler that appears to give clean, unimodal state-specific posteriors is often one that has failed to mix.

Three responses, in increasing order of respectability.

Identifiability constraints. Impose an ordering, such as $\mu_1 < \mu_2 < \cdots < \mu_K$, and permute each draw to satisfy it. Simple and often adequate, but the choice of ordering functional is arbitrary and a poor choice can cut through a mode rather than separating modes, producing artefacts. Jasra et al. (2005) give examples.

Relabelling algorithms. Post-process the output by choosing, for each draw, the permutation that minimises a loss relative to a reference — a Kullback–Leibler criterion in Stephens (2000), or a clustering of the draws in the space of allocation vectors. More work, but it adapts to the geometry of the posterior instead of assuming it.

Report only invariant quantities. Many questions do not require labelled states. The number of occupied states, the posterior predictive distribution, the co-clustering probabilities $\mathbb{P}(S_t = S_{t'} \mid y_{1:T})$, and the log marginal likelihood are all permutation-invariant and need no relabelling at all. When the scientific question can be phrased this way, it should be.

Deliberately *randomising* the labels — applying a uniformly random permutation after each sweep, the random permutation sampler of Frühwirth-Schnatter (2006) — is a useful diagnostic. It forces the sampler to explore the full symmetric posterior, so the resulting output can be checked for the balance that a well-mixing chain should exhibit, and it provides a clean input to a relabelling step.

Check your understanding. Your sampler for a three-state model gives clean, well-separated posteriors for μ_1, μ_2, μ_3, with narrow credible intervals and no multimodality. Should you be pleased?

Show answer

Probably not. The posterior is invariant under the $3! = 6$ relabellings (Definition 9.3), so a chain that is exploring it properly should show each marginal as the *same* six-fold symmetric mixture. Three distinct unimodal posteriors mean the chain has stayed in one labelling for the whole run — it has not mixed over the symmetric modes.
That may still be fine for practical purposes, and it is what an identifiability constraint deliberately produces. But it is not evidence of good mixing, and it should not be reported as though the sampler had converged. Run the random permutation sampler as a diagnostic: if the chain then fails to explore all six modes, the mixing problem is real and affects more than the labels.

9.4 Marginal likelihood

Bayesian model comparison across different K needs the marginal likelihood $p(y_{1:T} \mid K) = \int p_\theta(y_{1:T}) p(\theta \mid K)\, d\theta$. It is not available from the Gibbs output directly, and naive estimators such as the harmonic mean are notoriously unstable.

The standard approach is Chib (1996), which uses the basic marginal identity

$$\log p(y_{1:T}) = \log p_{\theta^*}(y_{1:T}) + \log p(\theta^*) - \log p(\theta^* \mid y_{1:T})$$

at a fixed high-density point θ^*. The first term comes from the forward recursion, the second from the prior, and the third from a sequence of reduced Gibbs runs that estimate the posterior ordinate blockwise. The method works well for HMMs precisely because the likelihood is available in closed form; its weakness is that the posterior ordinate estimate must account for label switching, or it will be wrong by a factor of $K!$.

Where the goal is to choose K rather than to report evidence for a fixed K, putting a prior on K and sampling it is usually better than computing marginal likelihoods for a grid of values. That is the subject of Chapter 11.

9.5 Summary

- Conjugate priors — Dirichlet on the rows of Γ, a conjugate family for the emissions — make every parameter update a direct draw, and a proper prior on the variances removes the unbounded-likelihood pathology of Chapter 4.
- Data augmentation targets the joint posterior of parameters and states. Conditional on the states the model decomposes exactly as the complete-data problem of Chapter 3.
- States must be drawn as a block. Forward filtering with backward sampling (Theorem 9.1) does this exactly in $\mathcal{O}(K^2 T)$, using only the filtered probabilities.

- Label switching is sharper here than in the classical setting: a well-mixing chain has $K!$ symmetric modes. Report invariant quantities where possible, and relabel deliberately where not.

9.6 Further reading

Frühwirth-Schnatter (2006, chs. 10–11) is the reference treatment of Bayesian HMMs, including priors, samplers and label switching, and Frühwirth-Schnatter et al. (2019) collects the current state of the mixture-model literature it belongs to. Robert and Casella (2004) is the general reference for the Monte Carlo methods used here. Scott (2002) is a readable applied account of recursive Bayesian computation for HMMs. Chib (1996) introduces both the FFBS scheme in this setting and the marginal likelihood estimator. On label switching, Stephens (2000) and Jasra et al. (2005). Rydén (2008) compares EM and MCMC directly and is worth reading before choosing between them. Robert et al. (2000) handle unknown K by reversible jump, the approach that Chapter 11 is designed to avoid.

10 Hierarchical models

"Hierarchical HMM" means at least three different things in the literature, and the three have almost nothing in common beyond the word. This chapter separates them, because conflating them causes real confusion:

1. **Hierarchical state spaces**, where states are nested — a state can itself be an HMM, recursively. This is the hierarchical HMM of Fine et al. (1998).
2. **Multilevel models**, where several series share structure and the parameters of each series are drawn from a population distribution. These are the mixed HMMs of Altman (2007).
3. **Hierarchical priors**, where a Bayesian model places a prior on the hyperparameters governing the states of a single HMM. This is the structure that Chapter 11 builds on, and the same idea taken to its infinite limit gives the HDP-HMM.

The second and third are where most applied work happens, and they are the focus here.

By the end of this chapter you will be able to

- tell the three senses of "hierarchical" apart and say which one a given problem needs;
- specify a multilevel HMM for several series, and explain how it interpolates between pooling and not pooling;
- diagnose the funnel geometry that breaks samplers for hierarchical models, and fix it by reparameterising;
- explain how a shared base measure ties the rows of Γ together, and what happens in the infinite limit.

> **i** The common thread
>
> All three add a layer of structure *above* the parameters of the basic model, and all three buy the same thing: a way to fit a model with many parameters without estimating each one independently. Sharing information across states, across subjects, or across levels is what makes the parameter count manageable. Gelman and Hill (2007) make the general argument; nothing about it is specific to HMMs.

10.1 Hierarchical state spaces

10.1.1 The model

Definition 10.1 (Hierarchical HMM)**.** A **hierarchical HMM** of depth D has states arranged in a tree. Each internal state at level $d < D$ is itself an HMM over its children at level $d + 1$. Production states, at level D, emit observations. When an internal state is entered, control passes to its sub-HMM; when that sub-HMM reaches its designated end state, control returns to the parent, which then makes its own transition.

The recursive structure captures phenomena with genuine multi-scale organisation. In speech, a sentence is a sequence of words, each word a sequence of phones, each phone a sequence of acoustic sub-states. In behavioural ecology, an animal alternates between coarse activity modes, each of which decomposes into finer movement patterns. A flat HMM can represent any of these only by taking the product of all levels, which multiplies the number of states and, more damagingly, discards the constraint that transitions respect the hierarchy.

10.1.2 Inference

The original algorithm of Fine et al. (1998) is a generalisation of the inside–outside algorithm for probabilistic context-free grammars, and it costs $\mathcal{O}(T^3)$ — cubic in the length of the series, because a sub-HMM may be entered and exited at any pair of times.

Murphy and Paskin (2001) observed that a hierarchical HMM can be flattened into an equivalent dynamic Bayesian network with one state variable per level plus indicator variables signalling when each level has finished. Inference in the flattened model is linear in T, at the cost of a state space that is the product across levels. Their construction is the practical route: express the hierarchy, flatten it, and run the standard $\mathcal{O}(K^2T)$ recursions of Chapter 5 on the product chain.

The reason this class sees less use than the other two is that the hierarchy has to be specified in advance. When the levels are known — as in speech, where the phonetic hierarchy is given — it is exactly right. When they are not, the multilevel models of the next section usually answer the real question better.

10.2 Multilevel and mixed HMMs

10.2.1 The setting

Suppose there are N series rather than one: N animals with movement tracks, N patients with symptom histories, N households with consumption records. Write $y_{n,1:T_n}$ for series n.

Two extreme analyses suggest themselves and both are bad.

Complete pooling fits one HMM to all series, treating them as replicates of a single process. It ignores between-series variation entirely. If one animal forages more than another, the pooled model attributes the difference to noise.

No pooling fits a separate HMM to each series. It respects heterogeneity but estimates N times as many parameters, and for short series the individual fits are unstable — some will have states that empty out, others will land on spurious local maxima.

The multilevel model interpolates. Each series gets its own parameters, but those parameters are themselves drawn from a population distribution to be estimated.

Definition 10.2 (Mixed HMM)**.** For series $n = 1, \dots, N$, let $(S_{n,t}, Y_{n,t})$ be an HMM with transition matrix Γ_n and emission parameters $\phi_{n,1:K}$, and let these be drawn from a population distribution:

$$(\Gamma_n, \phi_{n,1:K}) \mid \psi \overset{\text{iid}}{\sim} H(\cdot \mid \psi), \qquad \psi \sim p(\psi),$$

with the series conditionally independent given the individual parameters. The states are labelled consistently across series, so that state i means the same thing for every n.

The last requirement is essential and easy to overlook. If state 1 means "resting" for animal n and "travelling" for animal n', then averaging over n is meaningless. In practice consistency is enforced through the population distribution itself: a tight H keeps the series' state definitions close together, and identifiability constraints or relabelling (Chapter 9) handle what remains.

10.2.2 Random effects on transitions

The most common specification puts subject-level random effects on the multinomial-logit scale, which respects the simplex constraints automatically:

$$\gamma_{n,ij} = \frac{\exp(\beta_{ij} + b_{n,ij} + \lambda_{ij}^\top x_{n,t})}{\sum_{l=1}^{K} \exp(\beta_{il} + b_{n,il} + \lambda_{il}^\top x_{n,t})}, \qquad b_{n,i\cdot} \sim \mathcal{N}(0, \Sigma_i),$$

with $\beta_{ii} = 0$ and $b_{n,ii} = 0$ fixed for identification. Here β carries the population-level transition structure, b_n the individual departure from it, and λ the effect of covariates. Setting $\Sigma_i = 0$ recovers complete pooling; letting $\Sigma_i \to \infty$ recovers no pooling. The data choose where between the extremes to sit, which is the entire point.

Random effects on the emission parameters work the same way — $\phi_{n,i} = \phi_i + u_{n,i}$ with $u_{n,i} \sim \mathcal{N}(0, \Omega_i)$ — and allow each subject its own baseline level within a shared state structure.

10.2.3 Fitting

The likelihood for series n conditional on its random effects is the usual $\mathcal{O}(K^2 T_n)$ forward recursion. The marginal likelihood integrates over them:

$$L(\psi) = \prod_{n=1}^{N} \int \left[\delta_n^\top \prod_{t=1}^{T_n} \mathbf{Q}_n(y_{n,t}) \Gamma_n \, \mathbf{1} \right] H(d\Gamma_n, d\phi_n \mid \psi).$$

The integral has no closed form, and it is over a space whose dimension grows with the number of random effects. Three approaches:

- **Numerical integration.** Gauss–Hermite quadrature is feasible for one or two random effects per subject and hopeless beyond that, since the cost grows exponentially in the dimension.
- **Monte Carlo EM.** Treat the random effects as additional missing data and approximate the E-step by simulation. Straightforward to implement given a working HMM E-step, but slow and with a stopping rule that requires care.
- **MCMC.** Extend the Gibbs sampler of Definition 9.2 with steps for the random effects and hyperparameters. Given b_n, the FFBS step is unchanged; given the states, the random effects update is a standard generalised linear mixed model step, typically requiring Metropolis or Hamiltonian moves since the logit link breaks conjugacy.

MCMC is the usual choice, and it is worth knowing one implementation detail. Hierarchical models with small group-level variances produce a posterior geometry — Neal's funnel — in which the scale parameter and the random effects are strongly and non-linearly dependent, and samplers move through it badly. Reparameterising from the centred form $b_n \sim \mathcal{N}(0, \Sigma)$ to the non-centred form $b_n = \Sigma^{1/2} z_n$ with $z_n \sim \mathcal{N}(0, I)$ decouples them and usually fixes it. Papaspiliopoulos et al. (2007) give the general theory of when to prefer each parameterisation, and Betancourt and Girolami (2015) analyse the geometry directly.

10.2.4 What multilevel HMMs buy

Shrinkage. Individual estimates are pulled towards the population mean by an amount that depends on how much information each series carries. Short series shrink more, long series less. The result is better prediction for individuals than no pooling gives, and honest heterogeneity that complete pooling suppresses.

Population-level inference. The hyperparameters ψ answer questions that no individual fit can: how much do subjects differ in their tendency to persist in a state, and does a covariate shift that tendency systematically?

Stability. Series too short to support their own K-state fit become fittable, because the population distribution supplies what the data cannot. This is what makes multilevel HMMs the default in longitudinal studies with many short records.

Altman (2007) introduced the framework and derived the likelihood carefully; Maruotti (2011) surveys the variants and their estimation.

10.3 Hierarchical priors within a single HMM

The third sense of "hierarchical" adds a level to the prior of a single-series model. Instead of fixing the hyperparameters of Chapter 9, give them a prior.

10.3.1 Sharing across states

The independent Dirichlet rows of Definition 9.1 treat the K rows as unrelated. A hierarchical version ties them together through a shared base measure:

$$\beta \sim \mathrm{Dir}(\alpha/K, \ldots, \alpha/K), \qquad \Gamma_{j,\cdot} \mid \beta \overset{\text{iid}}{\sim} \mathrm{Dir}(\alpha_0 \beta).$$

Now β is a shared profile of how popular each destination state is overall, and each row is a perturbation of it with concentration α_0. The consequence is that a state that is rarely visited from anywhere gets small probability in every row simultaneously, rather than having to be discovered independently K times. Emission parameters can be tied in the same way by giving G_0 its own hyperprior, so that the states share information about scale and location.

Learning the concentration parameters matters more than it might seem. Fixing e_0 or α_0 fixes how sparse the transition matrix is allowed to be, which is a strong assumption dressed up as a default. Placing a gamma prior on them and sampling lets the data decide, and Chapter 11 makes this step explicit.

Check your understanding. You have 60 patients with between 8 and 15 observations each. Fitting a separate three-state HMM per patient fails for about a third of them — states empty out or variances collapse. Fitting one pooled model works but gives implausibly wide state-dependent distributions. What is going on?

Show answer

This is the no-pooling and complete-pooling dilemma in its purest form. Ten observations cannot support a model with $3 \times 2 + 6 = 12$ free parameters, so the individual fits are hitting the degeneracies of Chapter 4. Pooling gives enough data but forces every patient to share one set of parameters, so genuine between-patient variation is absorbed into the emission variances, inflating them.

A multilevel model (Definition 10.2) is the answer: each patient gets their own parameters, drawn from a population distribution estimated from all 60. Patients with 8 observations shrink heavily towards the population mean; those with 15 shrink less. The population-level parameters then answer the question the study is actually about, and the emission variances stop absorbing heterogeneity that belongs in the random effects.

10.3.2 The infinite limit

Letting $K \to \infty$ in the construction above, with β replaced by a stick-breaking measure, gives the hierarchical Dirichlet process HMM.

Definition 10.3 (HDP-HMM)**.** Let $\beta \sim \mathrm{GEM}(\alpha)$ be a stick-breaking distribution on the positive integers, and for each state j let

$$\Gamma_{j,\cdot} \mid \beta \sim \mathrm{DP}(\alpha_0, \beta), \qquad \phi_j \overset{\text{iid}}{\sim} G_0.$$

The state space is countably infinite, but any finite data set occupies finitely many states.

The shared β is what makes this work, and it is the reason the model is *hierarchical* rather than simply a collection of Dirichlet processes. Independent DP priors on the rows would, with probability one, give each row a disjoint set of destination states, and the chain could never return anywhere. The common base measure forces the rows to share atoms — the same states — while allowing different transition weights.

Two refinements are standard. The sticky variant of Fox et al. (2011) adds mass $\varkappa$ to the self-transition, correcting a well-documented tendency of the plain HDP-HMM to create redundant, rapidly-switching states. Beam sampling (Van Gael et al. 2008) introduces auxiliary slice variables that truncate the infinite state space to a finite random set at each iteration, so that exact FFBS can be run; this avoids the fixed truncation that variational and blocked-Gibbs schemes require.

The construction originates with Beal et al. (2002) and was given its proper hierarchical formulation by Teh et al. (2006).

10.3.3 Finite or infinite?

The HDP-HMM sidesteps choosing K by making the state space infinite. Chapter 11 keeps it finite and puts a prior on K. Which is preferable depends on what the states are meant to be.

If states are a device for flexible density estimation, and no particular number is scientifically meaningful, the nonparametric route is natural, and its behaviour is well understood: the number of occupied states grows logarithmically with T. That growth is a feature when the model is a flexible approximation and a defect when the states are supposed to be a fixed set of real regimes, since it means the answer to "how many states are there" depends on how long you observed.

If the states correspond to something real and finite — a fixed set of behaviours, disease stages, market regimes — then a finite model with a prior on K asks the question you actually want answered, and gives a posterior on K that does not drift with sample size. Miller and Harrison (2018) and Argiento and De Iorio (2022) examine the relationship between the two frameworks carefully; the finite-mixture-with-random-K construction is not merely a truncation of the nonparametric one, and its posterior for the number of components behaves differently.

10.4 Summary

- Three unrelated things are called hierarchical: nested state spaces (Definition 10.1), multilevel models across series (Definition 10.2), and hierarchical priors within one model.
- Hierarchical state spaces capture genuine multi-scale structure and are best handled by flattening to a dynamic Bayesian network, restoring $\mathcal{O}(T)$ inference (Murphy and Paskin 2001).
- Multilevel HMMs interpolate between complete pooling and no pooling, shrinking individual estimates towards a population distribution. Non-centred parameterisation is usually necessary for the sampler to work.
- Hierarchical priors tie the rows of Γ through a shared base measure, which lets states share information about which destinations matter. Taking $K \to \infty$ gives the HDP-HMM; keeping K finite with a prior gives the models of Chapter 11.

10.5 Further reading

Fine et al. (1998) and Murphy and Paskin (2001) for hierarchical state spaces. Altman (2007) introduced mixed HMMs and Maruotti (2011) surveys the field; Zucchini et al. (2016, ch. 13) covers random effects in the same notation used here. Gelman and Hill (2007) and Gelman et al. (2013, chs. 5, 15) are the general references for multilevel modelling, and Papaspiliopoulos et al. (2007) and Betancourt and Girolami (2015) for the parameterisation issues that decide whether the sampler works. For nonparametric versions, Teh et al. (2006) is the foundational paper, Fox et al. (2011) the practical refinement, and Van Gael et al. (2008) the sampler that makes them usable.

11 The telescoping sampler

Every chapter so far has taken K as given. Chapter 8 explained why choosing it is awkward — the classical theory breaks down at exactly the boundary where a K-state model sits inside a $(K+1)$-state one — and left the question open. This chapter closes it by treating K as a parameter with a prior and sampling it along with everything else.

The obstacle has always been computational. Changing K changes the dimension of θ, which normally means reversible jump MCMC (Green 1995) and the design of dimension-matching proposals that actually get accepted. Richardson and Green (1997) did this for univariate normal mixtures, with split-and-merge moves constructed by hand; the effort involved is a fair measure of the difficulty. For HMMs, where a new state needs both an emission distribution and a full row and column of the transition matrix, it is worse. Allocation-based samplers (Nobile 2004; Nobile and Fearnside 2007) avoid some of that work by operating on the partition, and the telescoping sampler can be seen as completing that programme.

The telescoping sampler of Frühwirth-Schnatter et al. (2021) removes the obstacle. It samples K explicitly, needs no reversible jump machinery, no Jacobians and no bespoke proposals, and uses only the standard updates of a fixed-K model. The trick is to marginalise the component weights before updating K, so that the move is a draw from a closed-form discrete distribution rather than a trans-dimensional jump.

By the end of this chapter you will be able to

- explain why changing K normally requires reversible jump, and how marginalising the weights avoids it;
- derive the closed-form conditional posterior of K given the partition, and identify where the dimension change is absorbed;
- distinguish static from dynamic mixtures of finite mixtures, and say what each implies about K versus K_+;
- write out the telescoping sweep for an HMM and say which steps differ from an ordinary Gibbs sweep.

> **Two different numbers**
>
> The chapter turns on a distinction that is easy to miss and does most of the work.
> K is the number of components **in the model** — how many states the sampler currently carries.
> K_+ is the number of components **occupied by the data** — how many states are actually visited by the current state sequence.
> Always $K_+ \leq K$. The components that make up the difference are empty: they have parameters, drawn from the prior, but no observations. Both have posteriors, and they answer different questions. $p(K_+ \mid y)$ is usually the one of scientific interest.

11.1 Mixtures of finite mixtures

We develop the method for finite mixtures first, following Frühwirth-Schnatter et al. (2021) exactly, and then transfer it to HMMs in Section 11.4.

11.1.1 The model

Definition 11.1 (Mixture of finite mixtures)**.** Observations $y_1, \dots, y_N$ are generated by

$$K \sim p(K), \qquad \eta \mid K \sim \mathrm{Dir}_K(\gamma_K, \dots, \gamma_K), \qquad \theta_k \mid K \overset{\text{iid}}{\sim} G_0 \quad (k = 1, \dots, K),$$

$$\mathbb{P}(S_i = k \mid \eta) = \eta_k, \qquad y_i \mid S_i = k, \theta \sim f(\cdot \mid \theta_k),$$

independently across $i = 1, \dots, N$.

The only unusual feature is that the Dirichlet parameter γ_K is allowed to depend on K. This apparently minor freedom generates two qualitatively different model classes.

Definition 11.2 (Static and dynamic MFMs)**.** The MFM is **static** if $\gamma_K = \alpha$ does not depend on K, and **dynamic** if $\gamma_K = \alpha/K$.

The distinction controls what happens to the weights as K grows. Under the static specification the total Dirichlet mass $K\gamma_K = K\alpha$ grows with K, so extra components are pushed towards having comparable weight; small α makes them sparse, and this is the sparse finite mixture of Malsiner-Walli et al. (2016). Under the dynamic specification the total mass $K\gamma_K = \alpha$ is constant, so the prior on the weights stabilises as K increases — the behaviour of a Dirichlet process, and the reason the dynamic class connects to Bayesian nonparametrics.

The consequence is a different relationship between K and K_+. Under a static MFM the prior on K_+ tracks K fairly closely when α is not small, so the number of components and the number of clusters are close. Under a dynamic MFM the two separate: K can be large while K_+ stays small and grows only slowly with N. Neither is right in the abstract; the choice is a modelling statement about whether "component" and "cluster" are meant to be the same thing.

11.1.2 The prior on the number of components

Common choices for $p(K)$ on $\{1, 2, 3, \dots\}$ are a shifted Poisson, $K - 1 \sim \mathrm{Poi}(\lambda)$; a shifted geometric or negative binomial, which has heavier tails and is less informative; or a uniform prior on $\{1, \dots, K_{\max}\}$, which is only defensible if $K_{\max}$ is genuinely a hard bound. Miller and Harrison (2018) use the shifted Poisson and give the resulting properties of K_+; the beta-negative-binomial prior favoured in Frühwirth-Schnatter et al. (2021) has heavier tails and is the safer default when little is known.

The posterior of K is more sensitive to $p(K)$ than one would like, and considerably more sensitive than the posterior of K_+. This is an argument for reporting $p(K_+ \mid y)$ as the primary summary, and for checking sensitivity to $p(K)$ explicitly. Grün et al. (2022) work through the sensitivity in detail on a benchmark data set.

11.2 The key computation

Everything rests on being able to sample K from its full conditional, given the current partition of the data but with the weights and the empty components' parameters integrated out.

11.2.1 Marginalising the weights

Proposition 11.1 (Allocation likelihood with weights integrated out)**.** *Let* $\eta \sim \mathrm{Dir}_K(\gamma_K, \dots, \gamma_K)$ *and let* $S_1, \dots, S_N$ *be conditionally i.i.d. with* $\mathbb{P}(S_i = k) = \eta_k$*. Writing* $N_k = \#\{i : S_i = k\}$*,*

$$p(S_{1:N} \mid K) \;=\; \frac{\Gamma(K\gamma_K)}{\Gamma(N + K\gamma_K)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)}.$$

Proof. Integrate the multinomial likelihood against the Dirichlet density:

$$p(S_{1:N} \mid K) = \int \prod_{k=1}^{K} \eta_k^{N_k} \cdot \frac{\Gamma(K\gamma_K)}{\Gamma(\gamma_K)^K} \prod_{k=1}^{K} \eta_k^{\gamma_K - 1}\, d\eta,$$

and recognise the remaining integral as the normalising constant of a $\mathrm{Dir}_K(\gamma_K + N_1, \dots, \gamma_K + N_K)$ density, namely $\prod_k \Gamma(N_k + \gamma_K)/\Gamma(N + K\gamma_K)$. □

The expression involves K only through the two Gamma functions and the product, and empty components contribute $\Gamma(0 + \gamma_K)/\Gamma(\gamma_K) = 1$. So the product effectively runs over the K_+ occupied components only, and the dependence on K is entirely explicit and cheap to evaluate.

11.2.2 From allocations to partitions

The allocation vector $S_{1:N}$ carries labels; the partition $\mathcal{C}$ — the grouping of observations, without labels — does not. Since the model is label-invariant, we work with the partition.

Proposition 11.2 (Partition likelihood)**.** *Let* $\mathcal{C}$ *be a partition of* $\{1, \dots, N\}$ *into* K_+ *non-empty blocks of sizes* $N_1, \dots, N_{K_+}$*. Then for* $K \geq K_+$*,*

$$p(\mathcal{C} \mid K) \;=\; \frac{K!}{(K - K_+)!} \cdot \frac{\Gamma(K\gamma_K)}{\Gamma(N + K\gamma_K)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)},$$

and $p(\mathcal{C} \mid K) = 0$ *for* $K < K_+$*.*

Proof. Each labelled allocation consistent with $\mathcal{C}$ assigns the K_+ blocks to K_+ distinct labels out of K, in some order. There are $K(K-1)\cdots(K-K_++1) = K!/(K-K_+)!$ such assignments, and by Proposition 11.1 each has the same probability, since the expression there depends on the block sizes but not on which labels they carry. Summing gives the falling factorial times the common value. □

The falling factorial is where the dimension change is absorbed. It counts the ways of embedding K_+ occupied components into a model of size K, and it is a *combinatorial* factor, not a Jacobian: no continuous change of variables is involved, which is exactly why reversible jump is unnecessary.

11.2.3 The conditional posterior of K

Theorem 11.1 (Full conditional for the number of components)**.** *Under Definition 11.1, with the weights η and the parameters of empty components integrated out,*

$$p(K \mid \mathcal{C}, \gamma_K) \;\propto\; p(K) \cdot \frac{K!}{(K-K_+)!} \cdot \frac{\Gamma(K\gamma_K)}{\Gamma(N+K\gamma_K)} \cdot \frac{1}{\Gamma(\gamma_K)^{K_+}} \prod_{k=1}^{K_+} \Gamma(N_k+\gamma_K)$$

for $K \geq K_+$, and zero otherwise.

Proof. Bayes' theorem applied to Proposition 11.2, noting that the parameters θ_k of empty components integrate out to one against their prior G_0 — they enter no likelihood term — and that the parameters of occupied components have the same posterior contribution whatever K is, so they cancel from the ratio. □

The distribution is supported on $\{K_+, K_+ + 1, \dots\}$ and, with a prior having light enough tails, can be normalised by direct evaluation over a truncation range. Sampling from it is a single categorical draw.

Note what has *not* happened. No proposal was designed, no acceptance probability computed, no Jacobian evaluated. The move is a Gibbs step in a model where the awkward continuous parameters have been integrated away, leaving a discrete quantity whose conditional distribution is available in closed form. The word "telescoping" refers to the model expanding and contracting through this step: components are created empty and destroyed when empty, and the number of them simply follows the draw.

11.3 The algorithm

Definition 11.3 (Telescoping sampler for an MFM)**.** Given the current $(K, \eta, \theta_{1:K}, S_{1:N}, \alpha)$, one sweep is:

1. **Allocate.** For each i, draw S_i from $\mathbb{P}(S_i = k) \propto \eta_k \, f(y_i \mid \theta_k)$, $k = 1, \dots, K$.
2. **Count.** Determine the occupied components, set K_+ and the sizes $N_1, \dots, N_{K_+}$, and relabel so that the occupied components come first.
3. **Update occupied parameters.** For $k = 1, \dots, K_+$, draw θ_k from its posterior given $\{y_i : S_i = k\}$ and G_0.
4. **Update the hyperparameter.** Draw α by a Metropolis–Hastings step targeting $p(\alpha \mid K, \mathcal{C}) \propto p(\alpha)\, p(\mathcal{C} \mid K, \gamma_K)$, using Proposition 11.2.
5. **Sample K.** Draw K from Theorem 11.1, on $\{K_+, K_+ + 1, \dots, K_{\max}\}$.
6. **Refill.** For $k = K_+ + 1, \dots, K$, draw $\theta_k \sim G_0$ — fresh empty components.
7. **Weights.** Draw

$$\eta \sim \mathrm{Dir}_K(\gamma_K + N_1, \dots, \gamma_K + N_{K_+}, \gamma_K, \dots, \gamma_K).$$

Steps 1, 3 and 7 are the ordinary updates of a fixed-K finite mixture, unchanged. Steps 2 and 6 are bookkeeping. The whole of the trans-dimensional problem is contained in steps 4 and 5, and neither requires anything beyond a Metropolis step on a scalar and a categorical draw.

Step 4 deserves a comment. Learning α rather than fixing it matters, because α controls how readily the model creates small components, and fixing it silently fixes the answer to the question being asked. Frühwirth-Schnatter et al. (2021) use a random-walk proposal on $\log \alpha$ with a moderately diffuse prior.

11.3.1 Reading the output

The sampler produces draws of both K and K_+, and they should be reported separately.

$p(K_+ \mid y)$ estimates how many clusters the data support. It is the summary to lead with, it is comparatively robust to $p(K)$, and it is permutation-invariant, so none of the relabelling machinery of Chapter 9 is needed to compute it.

$p(K \mid y)$ estimates the number of components in the generating mechanism, including any that happen to be unoccupied in this sample. It is more prior-sensitive, and under a dynamic MFM it can be substantially larger than K_+ without that indicating anything is wrong.

For state-specific summaries, the label switching problem returns in full force, and relabelling is complicated by the fact that the number of components changes between draws. The usual approach is to condition on the modal value of K_+, discard the other draws, and relabel within that subsample.

11.4 Telescoping for HMMs

The construction transfers to hidden Markov models. What changes is that the single weight vector η becomes the K rows of the transition matrix, and the independent allocations become a Markov chain — so the counts are transition counts and the allocation step becomes FFBS.

> **!** Status of this section
>
> Frühwirth-Schnatter et al. (2021) develop and validate the telescoping sampler for finite mixtures, and the `telescope` package (Malsiner-Walli et al. 2026) implements that case. The HMM version below follows the same construction and is derived here rather than quoted from a published HMM implementation. It is presented as the natural extension, and anyone using it should validate it on simulated data before trusting it. Established alternatives for HMMs with unknown K include reversible jump (Robert et al. 2000), overfitted models with sparse priors (Rousseau and Mengersen 2011; Havre et al. 2015), and the nonparametric route of Chapter 10.

11.4.1 The model

Take the finite HMM of Definition 3.2 with a prior on K:

$$K \sim p(K), \qquad \Gamma_{j,\cdot} \mid K \overset{\text{iid}}{\sim} \text{Dir}_K(\gamma_K, \dots, \gamma_K), \qquad \phi_k \mid K \overset{\text{iid}}{\sim} G_0,$$

with $\gamma_K = \alpha$ (static) or $\gamma_K = \alpha/K$ (dynamic), and a Dirichlet prior on δ or the stationarity constraint.

Given a state sequence $s_{1:T}$, define the transition counts and row totals

$$N_{jk} = \#\{t \geq 2 : s_{t-1} = j,\ s_t = k\}, \qquad N_{j\cdot} = \sum_{k=1}^{K} N_{jk}.$$

Call state j **occupied** if it is visited by $s_{1:T}$, and let K_+ be the number of occupied states.

11.4.2 Marginalising the transition matrix

Proposition 11.3 (State-sequence likelihood with Γ integrated out)**.** *With independent symmetric Dirichlet rows,*

$$p(s_{1:T} \mid K, \delta) \;=\; \delta_{s_1} \prod_{j=1}^{K} \left[\frac{\Gamma(K\gamma_K)}{\Gamma(N_{j\cdot} + K\gamma_K)} \prod_{k=1}^{K} \frac{\Gamma(N_{jk} + \gamma_K)}{\Gamma(\gamma_K)} \right].$$

Proof. The complete-data likelihood factorises across rows, $\prod_{j,k} \gamma_{jk}^{N_{jk}}$, and the rows are a priori independent, so the integral separates into K independent Dirichlet–multinomial integrals, each evaluated as in Proposition 11.1. □

The structure that made the mixture case work survives. A row j belonging to an unoccupied state has $N_{j\cdot} = 0$ and $N_{jk} = 0$ for all k, so its bracket equals $\Gamma(K\gamma_K)/\Gamma(K\gamma_K) = 1$ and it drops out. Within an occupied row, columns corresponding to unoccupied states also contribute 1. So the product runs effectively over the $K_+ \times K_+$ block of occupied states, and the dependence on K is once again explicit and cheap.

11.4.3 The conditional posterior of K

Theorem 11.2 (Full conditional for K in an HMM)**.** *With Γ and the emission parameters of unoccupied states integrated out, and states relabelled so the occupied ones come first,*

$$p(K \mid s_{1:T}, \alpha) \;\propto\; p(K) \cdot \frac{K!}{(K-K_+)!} \cdot \prod_{j=1}^{K_+} \left[\frac{\Gamma(K\gamma_K)}{\Gamma(N_{j\cdot} + K\gamma_K)} \prod_{k=1}^{K_+} \frac{\Gamma(N_{jk} + \gamma_K)}{\Gamma(\gamma_K)} \right]$$

for $K \geq K_+$, and zero otherwise.

Proof. Apply Bayes' theorem to Proposition 11.3, using the argument of Proposition 11.2 to convert from labelled state sequences to their unlabelled equivalence class: the K_+ occupied states can be assigned to K available labels in $K!/(K-K_+)!$ ways, each giving the same value of Proposition 11.3. Rows of unoccupied states contribute a factor of 1, and the emission parameters of unoccupied states integrate to 1 against G_0. □

Compare with Theorem 11.1. The single Dirichlet–multinomial factor has become a product of K_+ of them, one per occupied row, because a mixture has one weight vector and an HMM has one per state. Everything else — the falling factorial, the support on $K \geq K_+$, the closed form — is unchanged.

11.4.4 The sampler

Definition 11.4 (Telescoping sampler for a finite HMM)**.** Given $(K, \delta, \Gamma, \phi_{1:K}, s_{1:T}, \alpha)$, one sweep is:

1. **States.** Draw $s_{1:T}$ by forward filtering, backward sampling (Theorem 9.1) using the current K-state model.
2. **Count.** Identify the occupied states, set K_+, form the transition counts N_{jk}, and relabel so occupied states come first.
3. **Emissions.** For each occupied k, draw ϕ_k from its conjugate posterior given $\{y_t : s_t = k\}$.

4. **Hyperparameter.** Update α by a Metropolis step targeting $p(\alpha)\, p(s_{1:T} \mid K, \alpha)$ from Proposition 11.3.
5. **Number of states.** Draw K from Theorem 11.2.
6. **Refill.** For $k = K_+ + 1, \dots, K$, draw $\phi_k \sim G_0$.
7. **Transitions.** Draw the rows independently,

$$\Gamma_{j,\cdot} \sim \mathrm{Dir}_K(\gamma_K + N_{j1}, \dots, \gamma_K + N_{jK}), \qquad j = 1, \dots, K,$$

   where $N_{jk} = 0$ whenever j or k is unoccupied.
8. **Initial distribution.** Draw $\delta \sim \mathrm{Dir}_K(d + \mathbf{1}\{s_1 = \cdot\})$, or recompute the stationary distribution of the new Γ.

Steps 1, 3, 7 and 8 are exactly the sampler of Definition 9.2. The only additions are the two steps that move K, and they cost a scalar Metropolis update and one categorical draw per sweep — negligible beside the $\mathcal{O}(K^2 T)$ FFBS pass.

⚠ Practical cautions

Unoccupied states are visitable. Because unoccupied states have parameters drawn from G_0 and rows drawn from the prior, FFBS in step 1 can assign observations to them. This is the mechanism by which new states are born, and it is why G_0 must be a sensible prior over plausible emission parameters rather than something vague: a wildly diffuse G_0 makes empty states unattractive and the sampler will never populate them.
Mixing is harder than for mixtures. Persistence in the transition matrix makes it costly to reassign a run of consecutive observations to a new state, so births of genuinely useful states can be slow. Running several chains from different K and comparing the posteriors of K_+ is the minimum diagnostic.
Truncate honestly. Step 5 requires evaluating Theorem 11.2 over a finite range. The upper limit $K_{\max}$ must be large enough that the tail mass beyond it is negligible; check this rather than assuming it.
Report K_+. As in the mixture case, the posterior of the number of *occupied* states is the more stable and more interpretable summary.

Check your understanding. Your sampler reports a posterior for K concentrated on $\{6, 7, 8\}$ but a posterior for K_+ concentrated on 3. Is the model overfitting?

Show answer

No — that pattern is expected, particularly under a dynamic MFM. K counts components carried by the model, including empty ones created afresh from G_0 at every sweep; K_+ counts states the data actually occupy. Empty components cost nothing and contribute nothing to the fit, which is exactly what makes the sampler's dimension changes cheap. Report $p(K_+ \mid y)$ as the substantive answer: the data support three states. A large K with small K_+ says the prior on K has heavy tails, not that the fit is bloated. If K and K_+ were both concentrated on 8 while a three-state fit looked adequate, *that* would warrant concern.

11.5 Where this fits

Four ways of handling an unknown number of states, and what distinguishes them:

Table 11.1: Handling an unknown number of states

Approach	Mechanism	Cost
Information criteria (Chapter 8)	fit each K, compare scores	regularity conditions fail; heuristic only
Reversible jump (Robert et al. 2000)	trans-dimensional MH moves	proposals must be designed per model; tuning is delicate
Overfitted / sparse priors (Rousseau and Mengersen 2011)	fix K large, let the prior empty states	K is never sampled; asymptotic justification only
Telescoping (Frühwirth-Schnatter et al. 2021)	marginalise weights, sample K directly	needs conjugate weights; otherwise standard steps

The overfitted-mixture approach is the closest competitor and deserves a fair comparison. Setting K deliberately too large with $e_0 < 1$ causes superfluous states to empty out, and Rousseau and Mengersen (2011) show that the posterior concentrates on the correct number asymptotically. It is simpler to implement than telescoping, since it is just a fixed-K sampler. What it does not give is a posterior distribution over K: the result is one number, sensitive to the chosen K and to e_0, with no accompanying statement of uncertainty. Telescoping gives the distribution, at a small additional cost.

The requirement that limits telescoping is conjugacy of the weight prior. The marginalisation in Proposition 11.1 is what makes the K move a closed-form draw, and it depends on the Dirichlet–multinomial pair. A logistic-normal prior on the rows — natural when transition probabilities depend on covariates, as in Chapter 10 — has no such conjugacy, and the method as stated does not apply.

11.6 Summary

- Treating K as unknown is the honest response to the failure of classical model selection theory at the boundary between K and $K + 1$ states.
- The telescoping sampler marginalises the component weights before updating K, turning a trans-dimensional move into a categorical draw from a closed-form distribution (Theorem 11.1). No reversible jump, no Jacobians.
- The falling factorial $K!/(K - K_+)!$ absorbs the change in dimension combinatorially, by counting the ways of embedding K_+ occupied components into K available slots.
- Static ($\gamma_K = \alpha$) and dynamic ($\gamma_K = \alpha/K$) specifications differ in how K and K_+ relate; the choice is a modelling statement about whether components and clusters are the same thing.
- The construction transfers to HMMs by replacing the single weight vector with the rows of Γ, giving Theorem 11.2; the sampler is the standard HMM Gibbs sweep plus two cheap steps.
- Report $p(K_+ \mid y)$, the posterior on occupied states, as the primary summary.

11.7 Further reading

Frühwirth-Schnatter et al. (2021) is the source for everything in this chapter and should be read directly; the exposition is careful about exactly the points that matter. Malsiner-Walli et al. (2016) introduced sparse finite mixtures and is the natural precursor. Miller and Harrison (2018) develop the MFM framework and its relationship to Dirichlet process mixtures, and Argiento and De Iorio (2022) examine that relationship from the nonparametric side. Grün et al. (2022) walk through a full applied analysis with sensitivity checks and is the best model for how to report results. For HMM-specific alternatives, Robert et al. (2000) for reversible jump and Rousseau and Mengersen (2011) for the asymptotics of overfitted models. Software for the mixture case is in Malsiner-Walli et al. (2026).

Part V

V. Extensions

12 State-space and switching models

The finite state space has been doing a lot of work. Dropping it — letting the hidden state be continuous, or letting discrete and continuous states coexist — produces the models in this chapter. The conditional independence structure of Definition 3.1 survives intact; what changes is that sums over states become integrals, and exact inference survives only in special cases.

The organising question is when the recursions of Chapter 5 remain computable. Two cases keep them: a finite state space, giving the HMM, and a linear-Gaussian model, giving the Kalman filter. Everything else requires approximation.

By the end of this chapter you will be able to

- write the filtering recursion for a general state-space model and identify the predict and update steps;
- derive the Kalman filter and match each of its pieces to the corresponding piece of the forward algorithm;
- explain precisely why switching linear-Gaussian models are intractable, and what the standard approximations control;
- place factorial, semi-Markov and autoregressive HMMs relative to the base model.

12.1 General state-space models

Definition 12.1 (State-space model)**.** A **state-space model** consists of a hidden Markov process $(X_t)_{t\geq 1}$ on a measurable space $\mathcal{X}$ with initial density μ and transition density $p(x_t \mid x_{t-1})$, together with observations (Y_t) satisfying $Y_t \mid X_t = x_t \sim g(\cdot \mid x_t)$, conditionally independent across t given the states.

The joint density is

$$p(x_{1:T}, y_{1:T}) = \mu(x_1) g(y_1 \mid x_1) \prod_{t=2}^{T} p(x_t \mid x_{t-1})\, g(y_t \mid x_t),$$

identical in form to Theorem 3.1 with $\delta_{s_1} \to \mu(x_1)$ and $\gamma_{s_{t-1}s_t} \to p(x_t \mid x_{t-1})$. A finite-state HMM is the case $\mathcal{X} = \{1, \dots, K\}$ with counting measure.

The filtering recursion carries over exactly, with integration in place of summation:

$$\underbrace{p(x_t \mid y_{1:t-1}) = \int p(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{1:t-1})\, dx_{t-1}}_{\text{predict}}, \qquad \underbrace{p(x_t \mid y_{1:t}) \propto g(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})}_{\text{update}}.$$

Prediction propagates the current belief through the dynamics; the update reweights it by the likelihood of the new observation. In the finite case these are a vector–matrix product and a diagonal reweighting, which is precisely $\hat{\alpha}_{t-1}^{\top} \Gamma \mathbf{Q}(y_t)$ from Chapter 5.

The difficulty is that the integral is generally intractable, and the filtering distribution generally has no finite-dimensional representation. The finite-state case escapes because a distribution on K points is a vector of K numbers. The linear-Gaussian case escapes for a different reason.

12.2 The Kalman filter

Definition 12.2 (Linear-Gaussian state-space model)**.**

$$X_t = FX_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, Q), \qquad Y_t = HX_t + V_t, \quad V_t \sim \mathcal{N}(0, R),$$

with $X_1 \sim \mathcal{N}(m_0, P_0)$ and all noise terms independent.

Theorem 12.1 (Kalman filter)**.** *For the model of Definition 12.2, $X_t \mid Y_{1:t} \sim \mathcal{N}(m_t, P_t)$, where with $m_{t|t-1} = Fm_{t-1}$ and $P_{t|t-1} = FP_{t-1}F^\top + Q$,*

$$K_t = P_{t|t-1}H^\top (HP_{t|t-1}H^\top + R)^{-1},$$

$$m_t = m_{t|t-1} + K_t(y_t - Hm_{t|t-1}), \qquad P_t = (I - K_tH)P_{t|t-1}.$$

Proof. By induction. Suppose $X_{t-1} \mid y_{1:t-1} \sim \mathcal{N}(m_{t-1}, P_{t-1})$. The prediction step applies a linear map plus independent Gaussian noise, so the predictive law is Gaussian with the stated moments. For the update, (X_t, Y_t) given $y_{1:t-1}$ is jointly Gaussian with

$$\mathrm{Cov}(X_t, Y_t) = P_{t|t-1}H^\top, \qquad \mathrm{Var}(Y_t) = HP_{t|t-1}H^\top + R,$$

and the standard formula for the conditional distribution of one Gaussian block given another yields exactly the stated mean and covariance. □

The filter is closed because the Gaussian family is closed under linear maps, convolution and conditioning — the same reason the finite-state filter is closed under its operations. The gain K_t interpolates between trusting the model (R large, K_t small) and trusting the observation (Q large, K_t large).

The parallels with Chapter 5 are exact and worth listing. Filtering is the forward recursion. The Rauch–Tung–Striebel smoother is forward–backward. The prediction error decomposition $\ell_T = \sum_t \log \mathcal{N}(y_t; Hm_{t|t-1}, HP_{t|t-1}H^\top + R)$ is Proposition 5.1, with the scale factors c_t again equal to one-step predictive densities. And simulation smoothing is FFBS (Theorem 9.1). The algorithms are the same algorithm; only the representation of the filtering distribution differs. Kalman (1960) is the original; Durbin et al. (1998) and Douc et al. (2014) give modern treatments.

Nonlinear or non-Gaussian variants — the extended and unscented Kalman filters — retain a Gaussian representation by linearising or by propagating a deterministic set of points. They are cheap and often adequate, but they are approximations with no error guarantee, and they fail on genuinely multimodal filtering distributions.

12.3 Sequential Monte Carlo

When neither escape route is available, represent the filtering distribution by a weighted sample.

Definition 12.3 (Bootstrap particle filter)**.** Maintain particles $\{x_t^{(n)}\}_{n=1}^N$ with weights $\{w_t^{(n)}\}$ approximating $p(x_t \mid y_{1:t}) \approx \sum_n w_t^{(n)} \delta_{x_t^{(n)}}$. At each step: propagate each particle through the dynamics, $x_t^{(n)} \sim p(\cdot \mid x_{t-1}^{(n)})$; reweight by $w_t^{(n)} \propto g(y_t \mid x_t^{(n)})$; and resample when the effective sample size falls below a threshold.

The estimator of the likelihood, $\hat{L}_T = \prod_t (N^{-1} \sum_n g(y_t \mid x_t^{(n)}))$, is unbiased. That fact is more consequential than it appears: an unbiased likelihood estimator can be substituted into a Metropolis–Hastings acceptance ratio without disturbing the stationary distribution, which is the basis of particle MCMC and makes Bayesian inference feasible for models where the likelihood is unavailable. Chopin and Papaspiliopoulos (2020) is the modern reference.

For finite-state HMMs particle filters are pointless, since exact inference costs $\mathcal{O}(K^2 T)$. They matter when $\mathcal{X}$ is continuous, or when K is so large that K^2 is prohibitive.

12.4 Switching state-space models

Definition 12.4 (Switching linear-Gaussian model)**.** A discrete regime $S_t \in \{1, \dots, K\}$ follows a Markov chain with transition matrix Γ, and given the regime the continuous state evolves linearly:

$$X_t = F_{S_t} X_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, Q_{S_t}), \qquad Y_t = H_{S_t} X_t + V_t, \quad V_t \sim \mathcal{N}(0, R_{S_t}).$$

The model combines the interpretability of discrete regimes with the flexibility of continuous dynamics, and it is the natural formalisation of a system that changes mode: an economy entering recession, a machine developing a fault, an aircraft changing flight regime.

Exact inference is intractable, and the reason is specific. Conditional on the regime sequence $s_{1:T}$, the model is linear-Gaussian and the Kalman filter applies exactly. But the filtering distribution of X_t unconditionally is a mixture over all K^t regime histories, and the number of components multiplies by K at every step. This is exponential growth in the *representation*, not in the arithmetic, and no reorganisation of the recursion avoids it.

The standard responses all control that growth:

- **Collapsing.** After each step, approximate the K-component mixture by a single Gaussian matching its first two moments, keeping one Gaussian per regime. This is generalised pseudo-Bayes, cheap and surprisingly effective when regimes are well separated.
- **Rao–Blackwellised particle filters.** Sample the regime sequence and run an exact Kalman filter conditional on it. Only the discrete part is approximated, which is the right division of labour and gives far lower variance than sampling both.
- **Variational methods.** Approximate the posterior by a product of a discrete chain and a linear-Gaussian chain, and optimise the bound of 6.1. Ghahramani and Hinton (2000) develop this in detail.

In econometrics the special case where X_t is absent and Y_t follows a regime-dependent autoregression is Hamilton's Markov-switching model (Hamilton 1989), and there exact filtering *is* available: with $Y_t = \mu_{S_t} + \varphi_{S_t} Y_{t-1} + \varepsilon_t$, the only latent variable is discrete, so the standard forward recursion applies with the emission density replaced by the conditional density given y_{t-1}. The intractability of Definition 12.4 comes entirely from the continuous latent state.

12.5 Other extensions

Factorial HMMs. Several independent chains $S_t^{(1)}, \dots, S_t^{(M)}$ jointly generate one observation stream. The state space is the product, of size K^M, so exact inference costs $\mathcal{O}(K^{2M}T)$ and is infeasible beyond small M. The representation is far more compact than a flat HMM with K^M states, which is what makes the model attractive and approximate inference worth developing (Ghahramani and Jordan 1997).

Hidden semi-Markov models. In an HMM the time spent in a state is geometric, since leaving is an independent coin flip at each step. That is often wrong — behaviours and regimes have characteristic durations with modes away from zero. A hidden semi-Markov model specifies the dwell-time distribution directly. It can be approximated arbitrarily well by an HMM with a block of states per semi-Markov state, which keeps the standard algorithms at the cost of a larger K; Zucchini et al. (2016, ch. 12) give the construction.

Nonparametric emissions. Leaving the f_i unspecified and estimating them alongside the discrete structure is possible, and Gassiat et al. (2016) show identifiability holds under linear independence alone. Estimation uses spectral methods or nonparametric mixtures.

Autoregressive and covariate-dependent HMMs. Allowing Y_t to depend on Y_{t-1} as well as S_t weakens assumption (A2). All the algorithms survive: the emission factor $f_{s_t}(y_t)$ simply becomes $f_{s_t}(y_t \mid y_{t-1})$, and nothing in the derivations of Chapter 5 used the absence of that dependence.

12.6 Summary

- State-space models keep the HMM conditional independence structure and replace the finite state space with a general one; the filtering recursion is unchanged apart from integration replacing summation (Definition 12.1).
- Exact closed-form filtering survives in exactly two cases: finite state spaces, and linear-Gaussian models where the Kalman filter (Theorem 12.1) plays the role of the forward algorithm.
- Sequential Monte Carlo handles the general case, and its unbiased likelihood estimator supports exact Bayesian inference through particle MCMC.
- Switching models combining discrete regimes with continuous states are intractable because the filtering distribution is a mixture whose component count multiplies by K each step; collapsing, Rao–Blackwellisation and variational bounds all control that growth.

12.7 Further reading

Douc et al. (2014) covers state-space models, Kalman filtering and switching models with the same notation used here. Cappé et al. (2005, chs. 3–4) treat general state spaces rigorously. Chopin and Papaspiliopoulos (2020) is the standard modern reference for sequential Monte Carlo, and Doucet et al. (2001) remains a useful collection. Hamilton (1989) for regime-switching in econometrics, Ghahramani and Jordan (1997) and Ghahramani and Hinton (2000) for factorial and switching models with variational inference, and Zucchini et al. (2016, chs. 10, 12) for semi-Markov and covariate extensions in the finite-state setting.

13 Online prediction and decision problems

Everything so far has assumed the data were generated by an HMM, or at least that a best-approximating HMM exists in the sense of Chapter 8. This chapter drops that assumption in two directions.

The first is adversarial: suppose nothing at all about how the data arise, and ask how well a forecaster using HMMs performs relative to the best HMM chosen with hindsight. The answer is a regret bound, and it holds for every sequence, including sequences constructed to defeat the forecaster.

The second is decision-theoretic: suppose the point is not to infer the state but to act on it. An HMM is then a partially observable Markov decision process without actions, and the filtering distribution acquires a new role as a sufficient statistic for control.

Neither perspective is standard in the HMM literature, and both clarify what the model is doing.

By the end of this chapter you will be able to

- express cumulative log-loss as a negative log-likelihood, and define regret against a model class;
- prove that a Bayesian mixture over HMMs has logarithmic regret, with no assumption about the data;
- say what that guarantee licenses and what it does not;
- explain why the belief state is a sufficient statistic, and how filtering becomes the belief update in a POMDP.

13.1 Prediction without a model

13.1.1 The protocol

At each round $t = 1, 2, \ldots$ a forecaster issues a distribution q_t over the next observation, based only on $y_{1:t-1}$; then y_t is revealed and the forecaster suffers the logarithmic loss $-\log q_t(y_t)$. Cumulative loss after T rounds is

$$L_T(q) = \sum_{t=1}^{T} -\log q_t(y_t).$$

No probabilistic assumption is made about the sequence $y_1, y_2, \ldots$. It may be generated by an adversary who sees the forecaster's strategy.

Log-loss is the natural currency here. By the chain rule, $L_T(q) = -\log q(y_{1:T})$ for the joint distribution implied by the sequence of conditionals, so cumulative log-loss is exactly negative log-likelihood, and a forecaster is nothing other than a probabilistic model of the sequence. The connection to coding is immediate: $L_T(q)$ is the code length in nats of the sequence under an optimal code for q.

13.1.2 Regret

Definition 13.1 (Regret against a model class)**.** For a class $\{p_\theta : \theta \in \Theta\}$ of HMMs, the **regret** of a forecaster q is

$$R_T(q) = L_T(q) - \inf_{\theta\in\Theta} L_T(p_\theta) = \log \frac{\sup_\theta p_\theta(y_{1:T})}{q(y_{1:T})}.$$

The benchmark is the best HMM in the class *for this particular sequence*, selected with full knowledge of it. Sublinear regret — $R_T/T \to 0$ — means the forecaster's average loss matches that unattainable benchmark asymptotically, whatever the sequence.

13.1.3 Bayesian mixtures achieve it

Definition 13.2 (Mixture forecaster)**.** Given a prior Π on Θ, define $q(y_{1:T}) = \int_\Theta p_\theta(y_{1:T})\,\Pi(d\theta)$, with conditionals $q_t(\cdot) = q(\cdot \mid y_{1:t-1})$.

Proposition 13.1 (Regret of the mixture forecaster)**.** *For any set $A \subseteq \Theta$ with $\Pi(A) > 0$,*

$$R_T(q) \;\le\; \log \frac{1}{\Pi(A)} \;+\; \sup_{\theta\in A} \log \frac{\sup_{\theta'} p_{\theta'}(y_{1:T})}{p_\theta(y_{1:T})}.$$

Proof. Restricting the integral to A and bounding the integrand below by its infimum,

$$q(y_{1:T}) \ge \int_A p_\theta(y_{1:T})\Pi(d\theta) \ge \Pi(A) \inf_{\theta\in A} p_\theta(y_{1:T}).$$

Taking negative logarithms and comparing with $\sup_{\theta'} p_{\theta'}(y_{1:T})$ gives the bound. □

The two terms are a prior penalty and an approximation term. Choosing A to be a small neighbourhood of the best θ makes the second term small; the first then measures how much prior mass that neighbourhood carries. For a d-dimensional parameter space with a prior having a positive continuous density, a ball of radius $T^{-1/2}$ has mass of order $T^{-d/2}$, giving the classical rate

$$R_T(q) \;\lesssim\; \frac{d}{2}\log T + \mathcal{O}(1).$$

So a Bayesian mixture over HMMs, with any reasonable prior, has regret growing logarithmically in T against the best HMM chosen in hindsight — for every sequence, with no assumption that the data are generated by an HMM or by anything at all. The $\frac{d}{2}\log T$ term is the same quantity that appears in BIC, which is not a coincidence: both come from integrating a likelihood against a prior over a d-dimensional space. Merhav and Feder (1998) survey the universal prediction literature and Cesa-Bianchi and Lugosi (2006) give the general theory of prediction with expert advice.

> **i** What this justifies, and what it does not
>
> The result licenses using an HMM as a *forecasting device* without believing the model. Whatever the world does, you will predict about as well as the best HMM would have. It does not license interpreting the fitted states. A guarantee about predictive loss says nothing about whether state 2 corresponds to anything real. The interpretive claims of Chapter 14 rest on the model being approximately correct, which is a separate matter

entirely.

13.2 HMMs as partially observable decision problems

13.2.1 The setup

Definition 13.3 (Partially observable Markov decision process)**.** A POMDP consists of states $S_t \in E$, actions $A_t \in \mathcal{A}$, observations $Y_t \in \mathcal{Y}$, transition probabilities $p(s' \mid s, a)$, observation probabilities $p(y \mid s)$, a reward $r(s, a)$, and a discount factor $\lambda \in (0, 1]$.

An HMM is a POMDP with a single action and no reward. The extension matters because in most applications the states are inferred in order to do something: inspect a machine, treat a patient, adjust a portfolio.

13.2.2 The belief state

Definition 13.4 (Belief state)**.** The **belief** at time t is $b_t(i) = \mathbb{P}(S_t = i \mid y_{1:t}, a_{1:t-1}) \in \Delta^{K-1}$.

Theorem 13.1 (Beliefs form a Markov decision process)**.** *The belief state is a sufficient statistic for the history: the optimal action at time t depends on $(y_{1:t}, a_{1:t-1})$ only through b_t, and (b_t) is itself a Markov decision process on Δ^{K-1} with transition law induced by the filtering update.*

This is the result of Åström (1965). Its consequence is a reduction: a control problem with a growing history becomes a control problem on a fixed continuous state space, and dynamic programming applies. The value function satisfies

$$V(b) = \max_{a \in \mathcal{A}} \Big\{ r(b, a) + \lambda \, \mathbb{E}\big[V(b') \;\big|\; b, a\big] \Big\}, \qquad r(b, a) = \sum_i b(i)\, r(i, a),$$

where b' is the belief after taking a and observing the next y — computed by exactly the filtering update of Chapter 5. Filtering is the belief update; there is no separate algorithm.

The remarkable structural fact is that for finite horizons the value function is piecewise linear and convex in b, representable as a maximum of finitely many linear functions. Convexity has a clean reading: uncertainty is costly, so a belief spread over several states is worth less than a mixture of the certain beliefs it averages. This underpins exact solution methods, though the number of linear pieces can grow doubly exponentially in the horizon, and solving POMDPs exactly is intractable in general (Kaelbling et al. 1998).

13.2.3 Where this arises

- **Condition-based maintenance.** Hidden state is equipment health, observations are sensor readings, actions are inspect, repair or do nothing. The decision problem is when the belief about degradation justifies the cost of intervention.
- **Medical decision making.** Hidden state is disease stage, observations are test results, actions are test, treat or wait. Tests carry cost and information value simultaneously, and the belief formulation prices both.

- **Sequential design.** Choosing which measurement to take next to reduce uncertainty about a latent state is a POMDP whose reward is information gain.

In all three, the value of the HMM is that it converts a history into a belief, and the belief is what the decision rule acts on.

13.2.4 Beyond expected reward

Two variants matter when the criterion is not expected total reward.

Risk sensitivity. Maximising the exponential utility $-\lambda^{-1} \log \mathbb{E}[e^{-\lambda \sum_t r_t}]$ penalises variance as well as mean. Expanding in λ gives $\mathbb{E}[\sum r_t] - \frac{\lambda}{2} \operatorname{Var}(\sum r_t) + \ldots$, so λ is a risk-aversion parameter. The resulting recursions involve a KL-divergence penalty and connect directly to the variational structures of Chapter 1.

Robustness. Instead of trusting a single model, optimise the worst case over a set of models within some divergence of the nominal one. The resulting minimax problems are again KL-regularised, and the dual of the robust problem is a risk-sensitive one — a duality that recurs throughout this literature and gives the two variants a common origin.

13.3 Summary

- Under log-loss, a forecaster is a probability model and cumulative loss is negative log-likelihood. Regret compares a forecaster with the best HMM chosen in hindsight.
- A Bayesian mixture over HMMs has regret $\mathcal{O}(\frac{d}{2} \log T)$ against that benchmark, for every sequence and with no generative assumption (Proposition 13.1). This justifies HMMs as forecasting devices independently of whether they are true.
- Adding actions and rewards turns an HMM into a POMDP. The belief state is a sufficient statistic (Theorem 13.1), the filtering recursion is the belief update, and dynamic programming applies on the simplex.
- Value functions are piecewise linear and convex in the belief; exact solution is intractable in general, but the structure supports good approximations.

13.4 Further reading

Cesa-Bianchi and Lugosi (2006) is the standard reference for prediction with expert advice and regret bounds; Merhav and Feder (1998) survey universal prediction and its coding-theoretic side, and Cover and Thomas (2006, ch. 13) gives the information-theoretic background. For POMDPs, Åström (1965) established the sufficiency of the belief state and Kaelbling et al. (1998) is the standard modern account with algorithms. Murphy (2012, ch. 17) connects HMM inference to reinforcement learning.

Part VI

VI. Applications and exercises

14 Applications

An HMM is a modelling template, and applying it means answering four questions: what are the states, what generates the observations, what constraints does the transition matrix carry, and which inference problem is actually being solved. This chapter works through five domains with those four questions in view, and pays attention to where the standard assumptions strain.

The point is not a survey. It is that the same $\mathcal{O}(K^2T)$ recursion answers questions in speech, genomics, finance, ecology and epidemiology, and that the modelling decisions — not the algorithms — are what differ.

By the end of this chapter you will be able to

- specify an HMM for a new problem by answering four questions: states, emissions, transition constraints, inference task;
- recognise where a standard assumption is knowingly violated, and judge whether that matters;
- exploit structural constraints on Γ for both interpretation and speed;
- follow a workflow that ends in model checking rather than a fitted object.

14.1 Speech recognition

States. Sub-phonetic units. A phone is typically modelled by three states — onset, steady portion, offset — and context-dependent triphones multiply this substantially, so that K runs to tens of thousands even before a language model is composed in.

Observations. Short-time spectral features extracted from overlapping windows of about 25 ms at 10 ms intervals, classically mel-frequency cepstral coefficients in $\mathbb{R}^{39}$. Emissions were Gaussian mixtures for decades; in modern systems a neural network supplies the state-conditional scores and the HMM contributes the temporal structure.

Transitions. Strictly left-to-right: $\gamma_{ij} = 0$ unless $j \in \{i, i+1\}$. A phone is entered, dwelt in, and left, never revisited. This is where the sparsity of Chapter 7 pays off — the recursion touches $\mathcal{O}(K)$ rather than $\mathcal{O}(K^2)$ entries per step.

Problem. Decoding, on an enormous scale. Beam search rather than exact Viterbi, for the reasons in Chapter 7.

Two model failures are worth naming, because they are instructive. Geometric dwell times are wrong for phones, which have characteristic durations — the hidden semi-Markov extension of Chapter 12 is the principled fix, and duration models bolted onto the decoder are the practical one. And conditionally independent emissions are badly wrong for speech, where successive frames overlap by construction; the assumption is retained because it is what makes decoding tractable, and the resulting likelihood is treated frankly as a score rather than a probability.

Rabiner (1989) and Levinson et al. (1983) are the classical references; Jurafsky and Martin (2025) covers the modern hybrid architectures.

14.2 Bioinformatics

14.2.1 Segmenting a genome

States. Compositional classes of a DNA sequence — the standard example being CpG islands, regions unusually rich in the dinucleotide CG, against background.

Observations. Nucleotides, $\mathcal{Y} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

Emissions. State-dependent categorical distributions. Because the signal is a *dinucleotide* frequency, the useful model conditions on the previous base: $\mathbb{P}(Y_t = y \mid S_t = i, Y_{t-1} = y')$, an autoregressive HMM in the sense of Chapter 12, giving each state a 4×4 conditional table.

Transitions. Two states with strong self-transition; the diagonal entries set the expected island length directly, since a geometric dwell time with parameter $1-\gamma_{ii}$ has mean $1/(1-\gamma_{ii})$.

Problem. Decoding — which positions are island. Posterior decoding is often preferred to Viterbi here, because per-position confidence is more useful than a single global segmentation, and the caveat of Example 5.1 does not bite when the transition matrix has no structural zeros.

14.2.2 Profile HMMs

Profile HMMs model a family of related sequences rather than a single one, and their topology is the model. Each column of a multiple alignment gets a match state, plus an insert state for extra residues and a delete state for skipped ones. Transitions between these encode gap penalties, and the analogy is exact: log transition probabilities play the role of the affine gap costs in classical alignment, so a scoring scheme chosen by hand becomes a probability model that can be estimated.

The consequence is a proper statistical framework for homology detection. Aligning a query to the profile is Viterbi; scoring family membership is the forward algorithm; building the profile from unaligned sequences is Baum–Welch. Krogh et al. (1994) introduced the construction, Eddy (1998) made it practical, and Durbin et al. (1998) is the standard textbook.

14.3 Finance and econometrics

States. Market regimes — most commonly a low-volatility and a high-volatility regime, sometimes with a distinct crisis state.

Observations. Returns $Y_t = \log(P_t/P_{t-1})$.

Emissions. $Y_t \mid S_t = i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, with σ_i differing sharply across states and μ_i much less so.

Transitions. Highly persistent, with diagonal entries around 0.95–0.99 at daily frequency.

The model earns its place by reproducing stylised facts that a single Gaussian cannot. Marginally, a mixture of normals with different variances is leptokurtic, matching the fat tails of returns. Volatility clustering follows from regime persistence. And both emerge from a model with a handful of parameters and an interpretable state.

Problems. Filtering gives the current regime probability for risk management; smoothing gives a retrospective chronology of regimes; the predictive distribution feeds value-at-risk calculations that account for regime uncertainty rather than conditioning on a point estimate.

Hamilton's Markov-switching autoregression (Hamilton 1989) extends this to the conditional mean,

$$Y_t = \mu_{S_t} + \sum_{r=1}^{p} \varphi_{r,S_t} Y_{t-r} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{S_t}),$$

with states interpreted as expansion and recession. As noted in Chapter 12, exact filtering is available because the only latent variable remains discrete.

The honest caveat: regimes fitted this way are not economic regimes, they are variance regimes. A two-state fit on almost any return series finds a calm state and a turbulent one, and interpreting these as business-cycle phases requires evidence beyond the fit.

14.4 Animal movement and ecology

States. Behavioural modes — resting, foraging, travelling.

Observations. Bivariate, from successive GPS relocations: step length $\ell_t > 0$ and turning angle $\psi_t \in (-\pi, \pi]$.

Emissions. Gamma or Weibull for step length, von Mises or wrapped Cauchy for turning angle, taken as conditionally independent given the state. The states are then characterised by the joint pattern: short steps with undirected turns for resting, long steps with directed turns for travelling.

Transitions. Often covariate-dependent through a multinomial logit, with time of day, habitat type or reproductive status shifting the propensity to switch. This is the covariate-dependent construction of Chapter 4.

Problem. State decoding to produce a behavioural time series, and inference on the covariate coefficients, which is usually the scientific question — does habitat quality change foraging propensity?

The field also supplies the clearest use of the hierarchical machinery of Chapter 10, because studies routinely track many individuals with short and irregular records. A multilevel HMM with animal-level random effects on the transition parameters estimates both the population-level behavioural structure and individual variation around it, which is exactly what a study of a population wants. Langrock et al. (2012) set out the framework and Zucchini et al. (2016, ch. 16) work through examples in full.

14.5 Epidemiology and disease progression

States. Latent health states — disease stages, or infection status not directly observable.

Observations. Imperfect diagnostics: test results with known sensitivity and specificity, symptom scores, biomarkers.

Emissions. The emission matrix *is* the diagnostic's error profile. For a binary test in a two-state model, $b_i(y)$ has entries given by sensitivity and specificity, which are often known from validation studies and can therefore be fixed rather than estimated — a case where genuine external information pins down part of the model.

Transitions. Progression probabilities, frequently constrained. Irreversible disease gives an upper-triangular Γ; a treatment covariate shifts the progression rate, which is typically the estimand.

Problem. Estimating transition intensities in the presence of misclassification, and predicting state occupancy over time. Jackson et al. (2003) develop multistate models with classification error, which is the epidemiological name for this construction.

Two complications are structural. Observation times are irregular and often informative — patients attend when unwell — so the discrete-time chain must be replaced by a continuous-time one with transition matrix $\exp(\mathbf{A}\Delta_t)$ for the elapsed time Δ_t and generator $\mathbf{A}$. And panel data means many short series, which again makes the multilevel structure of Chapter 10 the natural formulation rather than an optional refinement.

Check your understanding. A colleague fits a two-state Gaussian HMM to daily equity returns, finds a low-variance and a high-variance state, and writes that the model has identified expansion and recession regimes. What would you say?

Show answer

That the model has identified a low-variance state and a high-variance state, which is what it was asked to do. Almost any return series yields that fit, because returns exhibit volatility clustering; the states are variance regimes, and calling them business-cycle phases is an interpretation the fit does not support.
To make the economic claim, you would need external evidence: do the smoothed state probabilities line up with dated recessions? Does the model beat a simple volatility model out of sample at something other than volatility? The general point is the one in Chapter 13 — a model that predicts well need not have interpretable internals, and the two claims require separate arguments.

14.6 A checklist

The five applications share a common workflow, and it is worth stating explicitly.

1. **Choose K for interpretability first, fit second.** States that map onto something real are more valuable than states that improve a criterion. Use Chapter 8 and Chapter 11 to inform the choice, not to make it.
2. **Match the emission family to the data type.** Positive quantities get positive-support families, angles get circular ones, counts get count families. A Gaussian on a strictly positive variable will manufacture states to handle the boundary.
3. **Constrain the transition matrix where the science does.** Irreversibility, left-to-right structure and forbidden transitions are information; imposing them improves both estimation and speed.
4. **Fit from many starting values.** Local maxima are the norm, not the exception (Chapter 6).
5. **Check the fit, not just the score.** Pseudo-residuals, posterior predictive comparisons of features the model was not fitted to, and — where states are interpreted — whether the decoded sequence makes sense to someone who knows the domain.
6. **Report uncertainty in the states, not only the parameters.** A decoded state sequence with no accompanying $\gamma_t(i)$ overstates what is known.

14.7 Summary

- The same recursions serve speech, genomics, finance, ecology and epidemiology; what changes is the interpretation of the states, the emission family, and the constraints on Γ.
- Structural constraints are common and valuable: left-to-right in speech, profile topologies in sequence analysis, irreversibility in disease progression. They encode science and reduce cost.
- Some standard assumptions are known to be wrong in specific domains — geometric dwell times in speech, conditional independence of overlapping acoustic frames — and are retained because the alternative is intractable. Knowing which assumption is being abused is part of using the model well.
- Multilevel structure (Chapter 10) is the norm rather than the exception in ecology and epidemiology, where studies produce many short series rather than one long one.

14.8 Further reading

Rabiner (1989) and Jurafsky and Martin (2025) for speech; Durbin et al. (1998) and Eddy (1998) for biological sequence analysis. Hamilton (1989) and Frühwirth-Schnatter (2006, ch. 12) for regime switching in economics. Langrock et al. (2012) and Zucchini et al. (2016, chs. 15–16) for animal movement and ecological applications, which are the best-documented worked examples in the literature. Jackson et al. (2003) for multistate models with misclassification in epidemiology.

15 Exercises

The exercises below are proof-based and grouped by the chapter they follow. Difficulty is marked: (*) routine given the chapter, (**) requires assembling several results, (***) open-ended or research-adjacent. No solutions are provided.

Exercises marked with a computer icon call for implementation. Anything that can be implemented should be, since the invariants of Chapter 7 make HMM code unusually easy to check.

15.1 Probability and Markov chains

Exercise 15.1 (Conditional expectation (*))**.** Let X be integrable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra.

(a) Show any two versions of $\mathbb{E}[X \mid \mathcal{G}]$ agree almost surely.

(b) Prove the tower property (Proposition 1.1, part 2) from the defining property alone.

(c) Show that if $X \in L^2$, then $\mathbb{E}[X \mid \mathcal{G}]$ is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$, and deduce $\operatorname{Var}(X) = \mathbb{E}[\operatorname{Var}(X \mid \mathcal{G})] + \operatorname{Var}(\mathbb{E}[X \mid \mathcal{G}])$.

Exercise 15.2 (Stochastic matrices and their spectra (**))**.**

(a) Show that every eigenvalue λ of a row-stochastic matrix satisfies $|\lambda| \le 1$.

(b) Give a 3×3 irreducible stochastic matrix with an eigenvalue of modulus 1 other than 1 itself. What is its period?

(c) Show that a stochastic matrix with all entries strictly positive is primitive, and bound $|\lambda_2|$ in terms of $\min_{ij} \gamma_{ij}$.

Exercise 15.3 (The ergodic theorem (**))**.** Prove Theorem 2.2 for an irreducible finite-state chain, using the regeneration structure at returns to a fixed state. Where exactly does the argument require finiteness of the state space, and where does it require irreducibility? Show by example that the conclusion can fail without irreducibility.

Exercise 15.4 (Dobrushin's coefficient (**))**.**

(a) Prove Theorem 2.3: $\|\mu^\top \Gamma - \nu^\top \Gamma\|_{\mathrm{TV}} \le \kappa(\Gamma) \|\mu - \nu\|_{\mathrm{TV}}$.

(b) Prove submultiplicativity, $\kappa(\Gamma_1 \Gamma_2) \le \kappa(\Gamma_1)\kappa(\Gamma_2)$.

(c) Show the Doeblin condition $\gamma_{ij} \ge \epsilon \nu_j$ implies $\kappa(\Gamma) \le 1 - \epsilon$.

(d) Compare the bound from (c) with the spectral bound of Proposition 2.5 for a two-state chain with $\gamma_{12} = \gamma_{21} = p$. Which is sharper, and for which values of p?

Exercise 15.5 (KL divergence (*))**.**

(a) Verify Proposition 1.3 by direct computation.

(b) Show KL is jointly convex in (p, q).

(c) Give an example with $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$ and one where the triangle inequality fails, confirming KL is not a metric.

15.2 Model structure and likelihood

Exercise 15.6 (Conditional independence (*))**.**

(a) Prove Theorem 3.1 from assumptions (A1) and (A2), being explicit about where each is used.

(b) Prove parts 1 and 4 of Proposition 3.1.

(c) Give a counterexample showing Y_1 and Y_2 are not in general independent, and compute $\mathrm{Cov}(Y_1, Y_2)$ for a two-state HMM with Gaussian emissions in terms of π, Γ and the state means.

Exercise 15.7 (Matrix representation (*))**.**

(a) Prove Theorem 3.2 by induction on T.

(b) Show that if $\delta = \pi$ then L_T is invariant under time reversal in the sense that reversing $y_{1:T}$ and replacing Γ by the time-reversed matrix $\tilde{\gamma}_{ij} = \pi_j \gamma_{ji} / \pi_i$ leaves the likelihood unchanged.

(c) Deduce that a reversible hidden chain gives a time-reversible observed process, and explain why the converse fails.

Exercise 15.8 (What HMMs can represent (**))**.** Consider a stationary two-state HMM with Poisson emissions, means $\lambda_1 \neq \lambda_2$.

(a) Compute the marginal distribution of Y_t and show it is overdispersed relative to a Poisson with the same mean.

(b) Compute the autocorrelation function of (Y_t) and show it decays geometrically at rate $\lambda_2(\Gamma)$, the second eigenvalue.

(c) Deduce which autocorrelation functions a two-state Poisson HMM can and cannot produce. Can it produce a negative lag-1 autocorrelation?

15.3 Inference algorithms

Exercise 15.9 (Forward and backward recursions (*))**.**

(a) Prove Theorem 5.1 by induction, stating where each model assumption enters.

(b) Prove Theorem 5.3.

(c) Prove Proposition 5.2 and verify the marginal consistency conditions stated there.

Exercise 15.10 (Scaling (*))**.**

(a) Prove Proposition 5.1: the scale factors are the one-step predictive densities and their logarithms sum to ℓ_T.

(b) Show the synchronised backward scaling of Chapter 5 gives $\gamma_t(i) = \hat{\alpha}_t(i)\hat{\beta}_t(i)$ exactly.

(c) Implement the forward algorithm in both scaled and log-domain form. Verify agreement to machine precision on a test problem, then find a model where the unscaled version underflows and confirm both scaled versions still work.

Exercise 15.11 (Viterbi (**))**.**

(a) Prove Theorem 5.5, stating the optimal substructure property precisely and showing the recursion computes it.

(b) Construct an explicit two-state HMM and an observation sequence of length at most 5 for which the sequence of marginal modes differs from the Viterbi path. Give the numbers.

(c) Construct an example where the sequence of marginal modes has probability zero under the model.

(d) Generalise the recursion to return the n best paths. What is the complexity?

Exercise 15.12 (The semiring view (**))**.** Both the forward algorithm and Viterbi have the form $v_t(j) = f_j(y_t) \otimes \bigoplus_i (v_{t-1}(i) \otimes \gamma_{ij})$ for a commutative semiring $(\oplus, \otimes)$.

(a) Identify the semiring for each.

(b) What is computed by the semiring $(\max, \min)$? By $(\oplus, \otimes) = (+, \times)$ over the expectation semiring used to accumulate expected counts?

(c) Show that the posterior entropy $H(S_{1:T} \mid y_{1:T})$ can be computed in $\mathcal{O}(K^2 T)$ by a suitable choice, and give the recursion.

Exercise 15.13 (Filter stability (**))**.**

(a) Prove Theorem 5.2 under the Doeblin condition, by showing the filter update map does not increase the Dobrushin coefficient.

(b) Verify numerically: fit a model, then run the filter from two very different initial distributions on the same data and plot the total variation distance against t on a log scale. Compare the observed rate with the bound.

(c) Explain why this result makes δ hard to estimate from a single long series.

15.4 Estimation

Exercise 15.14 (EM (*))**.**

(a) Verify identity 6.1.

(b) Prove Theorem 6.1, and identify exactly the conditions under which the inequality is strict.

(c) Prove Proposition 6.2, deriving both closed-form updates by Lagrange multipliers.

(d) Show that the generalised EM algorithm, which only increases Q rather than maximising it, retains the monotonicity property.

Exercise 15.15 (Fixed points and failure modes (**))**.**

(a) Show that if two states are initialised with identical emission parameters and identical rows of Γ, EM leaves them identical forever. Identify the corresponding point of the likelihood surface.

(b) For a Gaussian HMM with unconstrained variances, show $\sup_\theta \ell_T(\theta) = +\infty$ and exhibit a sequence attaining it.

(c) Implement Baum–Welch. On simulated data, run from 100 random starts and tabulate the distribution of the resulting log-likelihoods. How often is the best value found?

Exercise 15.16 (Score and information (**))**.**

(a) Prove Fisher's identity (Proposition 6.1).

(b) Write out $\nabla_\theta \ell_T$ explicitly in terms of $\gamma_t(i)$ and $\xi_t(i,j)$ for the transition parameters, and confirm one forward–backward pass suffices.

(c) State Louis's identity and use it to express the observed information in terms of forward–backward quantities.

Exercise 15.17 (Identifiability (**))**.**

(a) Prove Proposition 3.2.

(b) Show that if two states have identical emission distributions and identical rows of Γ, the parameter is not identified, and describe the resulting non-identified set.

(c) Construct two parameter values, not related by permutation, inducing the same law for $Y_{1:T}$ for all T. Which hypothesis of Theorem 4.1 fails?

(d) (***) Read Allman et al. (2009) and reconstruct the tensor argument behind Theorem 4.1, stating Kruskal's condition and verifying it in the HMM setting.

15.5 Asymptotics

Exercise 15.18 (The limiting log-likelihood (***))**.**

(a) Explain why $T^{-1}\ell_T(\theta)$ is not an average of stationary terms, and why the infinite-past construction of Theorem 8.1 repairs this.

(b) Show that under the Doeblin condition the approximation error from replacing the finite past by the infinite past decays geometrically, and that the errors are summable.

(c) (***) Read Leroux (1992) and reconstruct the subadditive ergodic theorem argument. Compare with the forgetting-based proof.

Exercise 15.19 (Consistency and normality (**))**.**

(a) Prove the information inequality in Theorem 8.2 and identify the equality case.

(b) Complete the Wald-style argument, stating where compactness and uniform convergence are used.

(c) Show the score is a martingale with respect to the observation filtration, and explain why this makes a martingale CLT the natural tool.

(d) Explain what goes wrong when $\theta^\star$ lies on the boundary of Θ, and describe the limiting distribution in a simple two-state example where one transition probability is zero.

Exercise 15.20 (Number of states (**))**.**

(a) Describe precisely the two ways in which a K-state HMM is embedded in a $(K+1)$-state model, and show the parameter is not identified at either.

(b) Explain why the likelihood ratio statistic does not have a χ^2 limit.

(c) Simulate from a three-state model and fit $K = 1, \dots, 6$. Tabulate AIC and BIC over many replications and report how often each selects the truth.

15.6 Bayesian, hierarchical and telescoping

Exercise 15.21 (Forward filtering, backward sampling (*))**.**

(a) Prove Theorem 9.1, identifying which conditional independence is used at each step.

(b) Explain why single-site Gibbs updating of the states mixes badly, and estimate how the mixing time scales with T for a highly persistent chain.

(c) Implement FFBS and verify it by comparing the empirical frequency of sampled state sequences with exact smoothed probabilities on a short series where enumeration is feasible.

Exercise 15.22 (Label switching in MCMC (**))**.**

(a) Show that if both likelihood and prior are permutation-invariant, then so is the posterior, and deduce that all marginal posteriors of ϕ_i coincide.

(b) Explain why a chain producing clean unimodal state-specific posteriors may indicate poor mixing.

(c) Give three permutation-invariant functionals of the posterior that can be reported without relabelling.

(d) Run the sampler of Definition 9.2 with and without random permutation of labels after each sweep, and compare the resulting marginal posteriors.

Exercise 15.23 (Multilevel HMMs (**))**.**

(a) Write out the full hierarchical model for N series with random effects on the transition logits, and give the complete-data likelihood.

(b) Derive the conditional posterior of the random effects given the states and hyperparameters. Is it conjugate?

(c) Explain the funnel geometry that arises when the group-level variance is small, and show how the non-centred parameterisation removes it.

(d) Show that as the group-level variance tends to zero and to infinity, the model recovers complete pooling and no pooling respectively.

Exercise 15.24 (The telescoping sampler (**))**.**

(a) Prove Proposition 11.1.

(b) Prove Proposition 11.2, explaining carefully where the falling factorial comes from.

(c) Derive Theorem 11.1, and verify that the parameters of empty components genuinely drop out.

(d) Show that under the dynamic specification $\gamma_K = \alpha/K$, the total Dirichlet mass is constant in K, and describe the effect on the prior for K_+.

Exercise 15.25 (Telescoping for HMMs (***))**.**

(a) Prove Proposition 11.3, and verify that rows corresponding to unoccupied states contribute a factor of exactly 1.

(b) Derive Theorem 11.2 and compare its structure term by term with Theorem 11.1.

(c) Explain the mechanism by which a new state becomes occupied during the FFBS step, and why a diffuse G_0 suppresses it.

(d) Implement the sampler of Definition 11.4. Simulate from a three-state model and check that the posterior of K_+ concentrates on 3. Then vary the prior $p(K)$ and report how the posteriors of K and K_+ respond.

(e) (***) Compare against an overfitted model with K fixed at 10 and $e_0 = 0.1$, in the manner of Rousseau and Mengersen (2011). Which recovers the truth more reliably, and how do the two summaries differ in what they express?

15.7 Extensions and alternative perspectives

Exercise 15.26 (The Kalman filter as a forward algorithm (**))**.**

(a) Derive Theorem 12.1 from the predict–update recursion.

(b) Show the prediction error decomposition gives ℓ_T as a sum of Gaussian log-densities, and identify the analogue of the scale factors c_t.

(c) State the analogue of FFBS for the linear-Gaussian model, and compare its structure with Theorem 9.1.

Exercise 15.27 (Intractability of switching models (**))**.**

(a) Show that in Definition 12.4 the exact filtering distribution of X_t given $y_{1:t}$ is a mixture of K^t Gaussians.

(b) Explain why no reorganisation of the recursion avoids this, in contrast with the finite-state case.

(c) Describe the collapsing approximation and state precisely what it preserves.

Exercise 15.28 (Regret bounds (**))**.**

(a) Prove Proposition 13.1.

(b) Derive the $\frac{d}{2}\log T$ rate by choosing A to be a ball of radius $T^{-1/2}$ and computing its prior mass.

(c) Explain the relationship between this term and the penalty in BIC.

(d) Explain precisely why a regret guarantee licenses using an HMM for prediction but not for interpreting its states.

Exercise 15.29 (Belief states (**))**.**

(a) Prove Theorem 13.1: beliefs form a Markov decision process on the simplex.

(b) Show the belief update is exactly the filtering step of Chapter 5.

(c) Show the finite-horizon value function is piecewise linear and convex in b, and interpret the convexity.

15.8 Using these exercises

A one-semester course can be built as follows.

Table 15.1: A twelve-week schedule

Weeks	Chapters	Exercises
1–2	Chapter 1, Chapter 2	Exercise 15.1 – Exercise 15.5
3–4	Chapter 3, Chapter 4	Exercise 15.6 – Exercise 15.8
5–6	Chapter 5	Exercise 15.9 – Exercise 15.13
7–8	Chapter 6, Chapter 7	Exercise 15.14 – Exercise 15.17
9	Chapter 8	Exercise 15.18 – Exercise 15.20
10–11	Chapter 9, Chapter 10, Chapter 11	Exercise 15.21 – Exercise 15.25
12	Chapter 12, Chapter 13, Chapter 14	Exercise 15.26 – Exercise 15.29

The implementation exercises are worth more than the proofs for most students, and Exercise 15.15 (b) and (c) in particular tend to change how people think about maximum likelihood.

A Notation

Symbols used throughout, following Zucchini et al. (2016) where a convention exists.

A.1 Model

Table A.1: Model notation

Symbol	Meaning	Introduced
$E = \{1, \dots, K\}$	hidden state space	Definition 3.2
K	number of states, or components in the model	Definition 3.2
K_+	number of states occupied by the data	Chapter 11
$\mathcal{Y}$	observation space, with dominating measure ν	Definition 3.2
S_t	hidden state at time t	Definition 2.1
Y_t	observation at time t	Definition 3.2
$Y_{s:t}$	the block $(Y_s, Y_{s+1}, \dots, Y_t)$	Chapter 1
T	length of the observed series	Definition 3.3
δ	initial distribution, $\delta_i = \mathbb{P}(S_1 = i)$	Definition 3.2
Γ	transition matrix, $\gamma_{ij} = \mathbb{P}(S_{t+1} = j \mid S_t = i)$	Definition 3.2
π	stationary distribution, $\pi^\top \Gamma = \pi^\top$	Definition 2.2
$f_i(\cdot)$, F_i	emission density and distribution for state i	Definition 3.2
ϕ_i	emission parameter for state i	Definition 3.2
$\mathbf{Q}(y)$	$\mathrm{diag}(f_1(y), \dots, f_K(y))$	Definition 3.4
θ	full parameter $(\delta, \Gamma, \phi_{1:K})$	Definition 3.2
Θ	parameter space	Chapter 3

A.2 Inference

Table A.2: Inference notation

Symbol	Meaning	Introduced
$\tilde{\alpha}_t(i)$	unnormalised forward variable, $p(S_t = i, y_{1:t})$	Definition 5.1
$\alpha_t(i)$, $\hat{\alpha}_t(i)$	filtering probability $\mathbb{P}(S_t = i \mid y_{1:t})$	Definition 5.1
$\beta_t(i)$	backward variable, $p(y_{t+1:T} \mid S_t = i)$	Definition 5.2
c_t	scale factor, equal to $p(y_t \mid y_{1:t-1})$	Proposition 5.1
$\gamma_t(i)$	smoothing probability $\mathbb{P}(S_t = i \mid y_{1:T})$	Theorem 5.4
$\xi_t(i, j)$	pairwise smoothing probability	Proposition 5.2
$m_t(j)$, $\psi_t(j)$	Viterbi value and backpointer	Theorem 5.5
$L_T(\theta)$, $\ell_T(\theta)$	likelihood and log-likelihood	Definition 3.3
$Q(\theta \mid \theta')$	EM objective	Definition 6.2

Symbol	Meaning	Introduced
$\mathcal{F}(q, \theta)$	evidence lower bound	Chapter 6
$U_T(\theta)$, $I(\theta)$	score and Fisher information per observation	Definition 8.1

A.3 Bayesian and hierarchical

Table A.3: Bayesian notation

Symbol	Meaning	Introduced
e_0, e_{jk}	Dirichlet hyperparameters for rows of Γ	Definition 9.1
G_0	prior for emission parameters	Chapter 9
N_{jk}, $N_{j\cdot}$	transition counts and row totals	Definition 9.2
η	mixture weights	Definition 11.1
γ_K	Dirichlet parameter, α or α/K	Definition 11.2
$\mathcal{C}$	partition of the data into occupied components	Proposition 11.2
β	shared base measure over states	Chapter 10
Σ_i, Ω_i	random effect covariances	Chapter 10

A.4 General

Table A.4: General notation

Symbol	Meaning
Δ^{K-1}	probability simplex in $\mathbb{R}^K$
$\mathbf{1}$	column vector of ones
$\lVert\cdot\rVert_{\mathrm{TV}}$	total variation distance, $\frac{1}{2}\sum_i \lvert\mu_i - \nu_i\rvert$
$\mathrm{KL}(p\lVert q)$	Kullback–Leibler divergence
$H(\cdot)$	Shannon entropy
$\kappa(\Gamma)$	Dobrushin contraction coefficient
ϱ	absolute spectral gap, $1 - \lvert\lambda_2\rvert$
$\perp\!\!\!\perp$	conditional independence
$\mathrm{Dir}_K(\cdot)$	Dirichlet distribution on Δ^{K-1}
$\mathfrak{S}_K$	symmetric group on K elements

A.5 Conventions

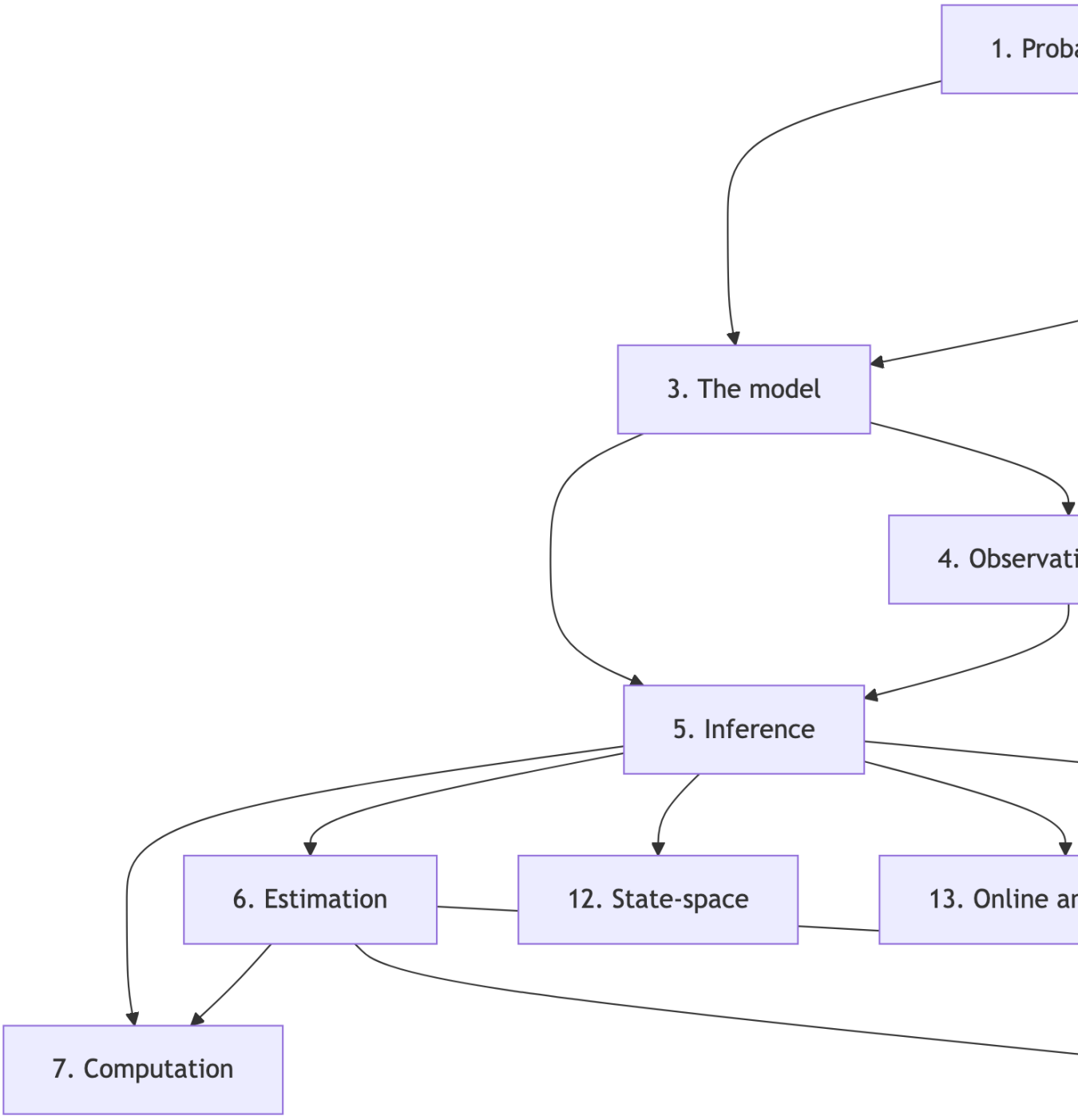
Row vectors multiply transition matrices from the left, so that $\mu^\top \Gamma$ is the distribution one step on. Densities are taken with respect to a fixed dominating measure, so the same formulas cover discrete and continuous observations. Indices i, j, k run over states, t over time, n over series or particles, and m over MCMC iterations.

A superscript star, $\theta^\star$, denotes a true parameter; a circle, θ°, a pseudo-true one under misspecification; a hat, $\widehat{\theta}$, an estimator.

B Reading guide

Which chapters depend on which, and where to start depending on what you want.

B.1 Dependencies



Two dependencies worth flagging. Chapter 11 needs Chapter 9 for forward filtering with backward sampling and Dirichlet conjugacy, and benefits from Chapter 10 for the prior structure, but does not require Chapter 8. And Chapter 7 sits after Chapter 6 because it discusses the cost and the monotonicity check of EM, though its first half — underflow and scaling — can be read straight after Chapter 5.

B.2 Routes through the material

B.2.1 A first course on HMMs

Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7.

Enough to specify a model, fit it, check it and interpret the output. Skip the measure-theoretic material in Chapter 1, referring back only for the KL divergence results used by EM. Follow with Chapter 14 for worked settings and the exercises marked (*) in Chapter 15.

B.2.2 Theory

Chapter 1, Chapter 2, Chapter 3, Chapter 5, Chapter 8.

A self-contained development from conditional expectation to asymptotic normality. Chapter 2 supplies the contraction results, Chapter 5 supplies filter stability, and Chapter 8 combines them. Read Cappé et al. (2005, chs. 12–13) in parallel.

B.2.3 Bayesian methods

Chapter 5 for the recursions, then Chapter 9, Chapter 10, Chapter 11.

Chapter 9 is the prerequisite for the other two. Someone interested only in the unknown-K problem can read Chapter 9 and Chapter 11 and skip Chapter 10, at the cost of missing the connection to the HDP-HMM.

B.2.4 Applied modelling

Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 10, Chapter 14.

Weighted towards specification and diagnosis rather than proof. Chapter 10 is included because most applied problems involve several series rather than one, and the multilevel formulation is usually the right default.

B.3 A twelve-week course

The schedule at the end of Chapter 15 maps chapters to weeks and exercises. In outline: two weeks on foundations, two on the model, three on inference, estimation and computation, one on asymptotics, two on Bayesian and hierarchical methods, and two on extensions and applications.

Two adjustments are worth considering. If students are strong in probability, compress the first two weeks into one and spend the time on Chapter 11, which rewards a slow reading. If the course is applied, drop Chapter 8 to a single lecture and expand Chapter 14, using the animal movement examples in Zucchini et al. (2016, chs. 15–16), which are the best documented in the literature.

B.4 Prerequisites

Assumed: undergraduate probability, linear algebra through eigenvalues, calculus, and some exposure to likelihood-based inference.

Not assumed: measure theory, stochastic processes, Bayesian computation. Each is introduced where needed — measure-theoretic language in Chapter 1, Markov chain theory in Chapter 2, MCMC in Chapter 9.

Implementation exercises assume the ability to write a loop and call a linear algebra routine. Language is immaterial; the notes use no code.

About

Lecture notes on hidden Markov models at the level of a graduate course in statistics, covering the model, the algorithms, the asymptotic theory, and the Bayesian and hierarchical extensions.

They exist because the two standard references rarely meet in the middle. Zucchini et al. (2016) is applied and practical but leaves proofs to citations; Cappé et al. (2005) is rigorous but assumes more measure theory than most readers arrive with. These notes derive the algorithms in full, state the theory honestly, and say what to do when a model fits badly.

How they are written

Proofs are given, not cited — except where the proof is a paper in itself. The identifiability theorem of Allman et al. (2009) and the asymptotic normality result of Bickel et al. (1998) are the main exceptions, and both are flagged where they appear.

Assumptions are named. The conditions under which a theorem holds are more useful than the theorem, because applied work happens where one of them fails.

Failure modes get equal billing. Unbounded likelihoods, spurious local maxima, label switching, empty states and boundary parameters are part of the subject, not caveats.

Notation follows Zucchini et al. (2016), so formulas here match theirs. See Appendix A.

What is unusual here

Two chapters most treatments omit.

Chapter 10 separates three unrelated things that all get called hierarchical — nested state spaces, multilevel models across several series, and hierarchical priors within one model. The multilevel case is what most applied problems actually need.

Chapter 11 covers the telescoping sampler of Frühwirth-Schnatter et al. (2021), which samples the number of states directly without reversible jump machinery. The mixture case follows the original paper; the HMM extension is derived here and marked as such.

Corrections are welcome — see contact.

FAQ

What background do I need?

Undergraduate probability, linear algebra through eigenvalues and eigenvectors, calculus, and some exposure to likelihood-based inference. Measure theory is not assumed — Chapter 1 introduces the pieces that get used, and nothing beyond them is needed.

If Markov chains are unfamiliar, read Chapter 2 carefully rather than skimming it. The contraction results there are used repeatedly.

Is there code?

No. The notes are language-agnostic and describe algorithms in enough detail to implement them in anything. Chapter 7 covers the numerical issues that matter — scaling, log-domain arithmetic, complexity — and lists four invariants that catch most implementation bugs.

Zucchini et al. (2016) provide R code for everything in Parts II and III, and the `telescope` package (Malsiner-Walli et al. 2026) implements the mixture case of Chapter 11.

Where should I start?

Depends on what you want. Appendix B sets out four routes with the dependencies. In short: for the algorithms, start at Chapter 4; for the theory, start at Chapter 1; for Bayesian methods, read Chapter 5 then jump to Chapter 9.

Reading straight through also works — the chapters are ordered so that each depends only on those before it.

How do I choose the number of states?

The honest answer is that this is the hardest part of using an HMM, and the notes give it three separate treatments.

Chapter 8 explains why the classical theory fails: a K-state model sits on the boundary of a $(K+1)$-state one, the parameter is not identified there, and the likelihood ratio statistic has no χ^2 limit. AIC and BIC are used routinely but are heuristics in this setting.

Chapter 11 puts a prior on K and samples it, giving a posterior distribution rather than a point choice. It also separates the number of states in the model from the number occupied by the data, which turns out to be the more stable and more interpretable quantity.

Chapter 14 makes the modelling argument: choose states you can interpret, and use the criteria to inform that choice rather than to make it. A three-state model whose states mean something usually beats a seven-state model with a better score.

Why does my EM fit give different answers each run?

Because it is finding local maxima, and which one it finds depends on where it starts. This is expected behaviour, not a bug — see Chapter 6 and Proposition 3.3.

Run from many dispersed starting values and keep the best. If the emissions are Gaussian and unconstrained, also check that no state's variance is collapsing towards zero; the likelihood is unbounded in that direction and the fit is chasing a singularity rather than a mode.

My Bayesian fit gives identical posteriors for every state. Why?

Label switching. The posterior is invariant under permuting the state labels, so if the sampler is mixing properly it visits all $K!$ symmetric modes and every state-specific marginal is the same average. See Definition 9.3.

Counterintuitively, clean, distinct state-specific posteriors often indicate the sampler has *failed* to mix. Either report permutation-invariant quantities, or relabel the output deliberately using one of the methods in Chapter 9.

Should I use EM or MCMC?

EM is faster and gives a point estimate. MCMC gives uncertainty, handles the unbounded-likelihood problem through the prior, and extends naturally to unknown K and hierarchical structure.

Rydén (2008) compares them on identical problems and is worth reading before deciding. A reasonable default is to fit by EM from many starts to understand the likelihood surface, then run MCMC if uncertainty quantification matters.

What if my data are several short series rather than one long one?

Then the multilevel models of Chapter 10 are the right formulation, not an optional refinement. Fitting each series separately wastes information and is unstable for short records; pooling them into one model ignores real heterogeneity. A hierarchical model interpolates between the two, and the data determine where.

Are there solutions to the exercises?

No. The exercises in Chapter 15 are intended for coursework, qualifying exam preparation and reading groups, and are graded (*) to (***).

The implementation exercises are more valuable than the proofs for most readers. Exercise 15.15 in particular tends to change how people think about maximum likelihood.

Can I use these for teaching?

Yes. Chapter 15 includes a twelve-week schedule mapping chapters to weeks, and Appendix B suggests adjustments for stronger or more applied cohorts.

How long does the course take?

Twelve weeks at two to three hours a week for the full set. An intensive reading course covering Parts I to IV compresses into eight weeks for well-prepared students. Parts V and VI work as a second-pass or as optional reading.

Corrections and contact

Notes of this length contain errors. Reports are welcome.

Reporting an error

The most useful reports say where and what:

- the chapter and the numbered environment or equation — "Theorem 5.3, second line of the proof" is enough to locate anything;
- what is wrong, and if you know it, what it should say.

Mathematical errors, missing hypotheses and unclear proofs are the most valuable. Notation inconsistencies between chapters are worth reporting too, since they are easy to introduce and hard to spot from inside.

If the notes are hosted from a version-controlled repository, the issue tracker is the right channel, and corrections can be submitted directly as pull requests by anyone comfortable editing Quarto markdown.

Known limitations

Stated plainly, so that nobody spends time discovering them independently.

The HMM telescoping sampler is derived, not cited. Frühwirth-Schnatter et al. (2021) develop and validate the method for finite mixtures. The extension to HMMs in Section 11.4 follows the same construction and is worked out in these notes. It has not been validated against a published HMM implementation, and Chapter 11 says so. Validate on simulated data before relying on it.

Some proofs are sketched. The identifiability theorem of Allman et al. (2009), the asymptotic normality result of Bickel et al. (1998), and the ergodic theorem for finite chains are stated with the argument outlined rather than given in full. Each is flagged where it appears, with a reference to the complete proof.

Continuous-time models are not covered. Chapter 14 notes that irregular observation times in epidemiology require a continuous-time chain, but the notes do not develop the theory.

No code. By design, but it does mean the implementation exercises in Chapter 15 are unsupported.

Reuse

The notes are intended to be used and adapted. If you build a course on them, attribution to the primary references in Appendix B matters more than attribution here — Zucchini et al. (2016), Cappé et al. (2005), Frühwirth-Schnatter (2006) and Douc et al. (2014) did the work these notes organise.

Response times vary. This is a resource, not a service.

Bibliography

Works cited throughout the notes, in author–date order. Citations in the text link here.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Allman, Elizabeth S., Catherine Matias, and John A. Rhodes. 2009. "Identifiability of Parameters in Latent Structure Models with Many Observed Variables." *The Annals of Statistics* 37 (6A): 3099–132. https://doi.org/10.1214/09-AOS689.

Altman, Rachel MacKay. 2007. "Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting." *Journal of the American Statistical Association* 102 (477): 201–10. https://doi.org/10.1198/016214506000001086.

Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*. Springer. https://doi.org/10.1007/978-4-431-55978-8.

Argiento, Raffaele, and Maria De Iorio. 2022. "Is Infinity That Far? A Bayesian Nonparametric Perspective of Finite Mixture Models." *The Annals of Statistics* 50 (5): 2641–63. https://doi.org/10.1214/22-AOS2201.

Åström, Karl Johan. 1965. "Optimal Control of Markov Processes with Incomplete State Information." *Journal of Mathematical Analysis and Applications* 10 (1): 174–205. https://doi.org/10.1016/0022-247X(65)90154-X.

Baum, Leonard E., and Ted Petrie. 1966. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." *The Annals of Mathematical Statistics* 37 (6): 1554–63. https://doi.org/10.1214/aoms/1177699147.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics* 41 (1): 164–71. https://doi.org/10.1214/aoms/1177697196.

Beal, Matthew J., Zoubin Ghahramani, and Carl E. Rasmussen. 2002. "The Infinite Hidden Markov Model." *Advances in Neural Information Processing Systems* 14: 577–84.

Betancourt, Michael, and Mark Girolami. 2015. "Hamiltonian Monte Carlo for Hierarchical Models." *Current Trends in Bayesian Methodology with Applications*, 79–101. https://doi.org/10.1201/b18502-5.

Bickel, Peter J., Ya'acov Ritov, and Tobias Rydén. 1998. "Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models." *The Annals of Statistics* 26 (4): 1614–35. https://doi.org/10.1214/aos/1024691255.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Boyd, Stephen, and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804441.

Cappé, Olivier, Éric Moulines, and Tobias Rydén. 2005. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer. https://doi.org/10.1007/0-387-28982-8.

Carter, Chris K., and Robert Kohn. 1994. "On Gibbs Sampling for State Space Models." *Biometrika* 81 (3): 541–53. https://doi.org/10.1093/biomet/81.3.541.

Celeux, Gilles, and Jean-Baptiste Durand. 2008. "Selecting Hidden Markov Model State Number with Cross-Validated Likelihood." *Computational Statistics* 23 (4): 541–64. https://doi.org/10.1007/s00180-007-0097-1.

Cesa-Bianchi, Nicolò, and Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546921.

Chib, Siddhartha. 1996. "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models." *Journal of Econometrics* 75 (1): 79–97. https://doi.org/10.1016/0304-4076(95)01770-4.

Chopin, Nicolas, and Omiros Papaspiliopoulos. 2020. *An Introduction to Sequential Monte Carlo*. Springer. https://doi.org/10.1007/978-3-030-47845-2.

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory*. 2nd ed. Wiley. https://doi.org/10.1002/047174882X.

Del Moral, Pierre, and Alice Guionnet. 2001. "On the Stability of Interacting Processes with Applications to Filtering and Genetic Algorithms." *Annales de l'Institut Henri Poincaré (B) Probability and Statistics* 37 (2): 155–94. https://doi.org/10.1016/S0246-0203(00)01064-5.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39 (1): 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Douc, Randal, Éric Moulines, and Tobias Rydén. 2004. "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime." *The Annals of Statistics* 32 (5): 2254–304. https://doi.org/10.1214/009053604000000021.

Douc, Randal, Éric Moulines, and David S. Stoffer. 2014. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Chapman; Hall/CRC. https://doi.org/10.1201/b16331.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon, eds. 2001. *Sequential Monte Carlo Methods in Practice*. Springer. https://doi.org/10.1007/978-1-4757-3437-9.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790492.

Eddy, Sean R. 1998. "Profile Hidden Markov Models." *Bioinformatics* 14 (9): 755–63. https://doi.org/10.1093/bioinformatics/14.9.755.

Fine, Shai, Yoram Singer, and Naftali Tishby. 1998. "The Hierarchical Hidden Markov Model: Analysis and Applications." *Machine Learning* 32 (1): 41–62. https://doi.org/10.1023/A:1007469218079.

Forney, G. David. 1973. "The Viterbi Algorithm." *Proceedings of the IEEE* 61 (3): 268–78. https://doi.org/10.1109/PROC.1973.9030.

Fox, Emily B., Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. 2011. "A Sticky HDP-HMM with Application to Speaker Diarization." *The Annals of Applied Statistics* 5 (2A): 1020–56. https://doi.org/10.1214/10-AOAS395.

Frühwirth-Schnatter, Sylvia. 1994. "Data Augmentation and Dynamic Linear Models." *Journal of Time Series Analysis* 15 (2): 183–202. https://doi.org/10.1111/j.1467-9892.1994.tb00184.x.

Frühwirth-Schnatter, Sylvia. 2006. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer. https://doi.org/10.1007/978-0-387-35768-3.

Frühwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P. Robert, eds. 2019. *Handbook of Mixture Analysis*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429055911.

Frühwirth-Schnatter, Sylvia, Gertraud Malsiner-Walli, and Bettina Grün. 2021. "Generalized Mixtures of Finite Mixtures and Telescoping Sampling." *Bayesian Analysis* 16 (4): 1279–307. https://doi.org/10.1214/21-BA1294.

Gassiat, Elisabeth, Alice Cleynen, and Stéphane Robin. 2016. "Inference in Finite State Space Non Parametric Hidden Markov Models and Applications." *Statistics and Computing* 26 (1): 61–71. https://doi.org/10.1007/s11222-014-9523-8.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Chapman; Hall/CRC. https://doi.org/10.1201/b16018.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942.

Ghahramani, Zoubin, and Geoffrey E. Hinton. 2000. "Variational Learning for Switching State-Space Models." *Neural Computation* 12 (4): 831–64. https://doi.org/10.1162/089976600300015619.

Ghahramani, Zoubin, and Michael I. Jordan. 1997. "Factorial Hidden Markov Models." *Machine Learning* 29 (2): 245–73. https://doi.org/10.1023/A:1007425814087.

Green, Peter J. 1995. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika* 82 (4): 711–32. https://doi.org/10.1093/biomet/82.4.711.

Grün, Bettina, Gertraud Malsiner-Walli, and Sylvia Frühwirth-Schnatter. 2022. "How Many Data Clusters Are in the Galaxy Data Set? Bayesian Cluster Analysis in Action." *Advances in Data Analysis and Classification* 16 (2): 325–49. https://doi.org/10.1007/s11634-021-00461-8.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–84. https://doi.org/10.2307/1912559.

Havre, Zoe van, Nicole White, Judith Rousseau, and Kerrie Mengersen. 2015. "Overfitting Bayesian Mixture Models with an Unknown Number of Components." *PLOS ONE* 10 (7): e0131739. https://doi.org/10.1371/journal.pone.0131739.

Horn, Roger A., and Charles R. Johnson. 2013. *Matrix Analysis*. 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511810817.

Hsu, Daniel, Sham M. Kakade, and Tong Zhang. 2012. "A Spectral Algorithm for Learning Hidden Markov Models." *Journal of Computer and System Sciences* 78 (5): 1460–80. https://doi.org/10.1016/j.jcss.2011.12.025.

Jackson, Christopher H., Linda D. Sharples, Simon G. Thompson, Stephen W. Duffy, and Elisabeth Couto. 2003. "Multistate Markov Models for Disease Progression with Classification Error." *Journal of the Royal Statistical Society, Series D* 52 (2): 193–209. https://doi.org/10.1111/1467-9884.00351.

Jasra, Ajay, Chris C. Holmes, and David A. Stephens. 2005. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science* 20 (1): 50–67. https://doi.org/10.1214/088342305000000016.

Jurafsky, Daniel, and James H. Martin. 2025. *Speech and Language Processing*. 3rd (draft). https://web.stanford.edu/~jurafsky/slp3/.

Kaelbling, Leslie Pack, Michael L. Littman, and Anthony R. Cassandra. 1998. "Planning and Acting in Partially Observable Stochastic Domains." *Artificial Intelligence* 101 (1–2): 99–134. https://doi.org/10.1016/S0004-3702(98)00023-X.

Kallenberg, Olav. 2021. *Foundations of Modern Probability*. 3rd ed. Springer. https://doi.org/10.1007/978-3-030-61871-1.

Kalman, Rudolph E. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering* 82 (1): 35–45. https://doi.org/10.1115/1.3662552.

Krogh, Anders, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. 1994. "Hidden Markov Models in Computational Biology: Applications to Protein Modeling." *Journal of Molecular Biology* 235 (5): 1501–31. https://doi.org/10.1006/jmbi.1994.1104.

Kruskal, Joseph B. 1977. "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics." *Linear Algebra and Its Applications* 18 (2): 95–138. https://doi.org/10.1016/0024-3795(77)90069-6.

Langrock, Roland, Ruth King, Jason Matthiopoulos, Len Thomas, Daniel Fortin, and Juan

M. Morales. 2012. “Flexible and Practical Modeling of Animal Telemetry Data: Hidden Markov Models and Extensions.” *Ecology* 93 (11): 2336–42. https://doi.org/10.1890/11-2241.1.

Leroux, Brian G. 1992. “Maximum-Likelihood Estimation for Hidden Markov Models.” *Stochastic Processes and Their Applications* 40 (1): 127–43. https://doi.org/10.1016/0304-4149(92)90141-C.

Levin, David A., and Yuval Peres. 2017. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society. https://doi.org/10.1090/mbk/107.

Levinson, Stephen E., Lawrence R. Rabiner, and Man Mohan Sondhi. 1983. “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition.” *Bell System Technical Journal* 62 (4): 1035–74. https://doi.org/10.1002/j.1538-7305.1983.tb03114.x.

Malsiner-Walli, Gertraud, Sylvia Frühwirth-Schnatter, and Bettina Grün. 2016. “Model-Based Clustering Based on Sparse Finite Gaussian Mixtures.” *Statistics and Computing* 26 (1): 303–24. https://doi.org/10.1007/s11222-014-9500-2.

Malsiner-Walli, Gertraud, Bettina Grün, and Sylvia Frühwirth-Schnatter. 2026. *Telescope: Bayesian Mixtures with an Unknown Number of Components*. https://CRAN.R-project.org/package=telescope.

Maruotti, Antonello. 2011. “Mixed Hidden Markov Models for Longitudinal Data: An Overview.” *International Statistical Review* 79 (3): 427–54. https://doi.org/10.1111/j.1751-5823.2011.00160.x.

Merhav, Neri, and Meir Feder. 1998. “Universal Prediction.” *IEEE Transactions on Information Theory* 44 (6): 2124–47. https://doi.org/10.1109/18.720534.

Meyn, Sean, and Richard L. Tweedie. 2009. *Markov Chains and Stochastic Stability*. 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511626630.

Miller, Jeffrey W., and Matthew T. Harrison. 2018. “Mixture Models with a Prior on the Number of Components.” *Journal of the American Statistical Association* 113 (521): 340–56. https://doi.org/10.1080/01621459.2016.1255636.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murphy, Kevin P., and Mark A. Paskin. 2001. “Linear-Time Inference in Hierarchical HMMs.” *Advances in Neural Information Processing Systems* 14: 833–40.

Neal, Radford M., and Geoffrey E. Hinton. 1998. “A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants.” In *Learning in Graphical Models*, edited by Michael I. Jordan. Springer. https://doi.org/10.1007/978-94-011-5014-9_12.

Nobile, Agostino. 2004. “On the Posterior Distribution of the Number of Components in a Finite Mixture.” *The Annals of Statistics* 32 (5): 2044–73. https://doi.org/10.1214/009053604000000788.

Nobile, Agostino, and Alastair T. Fearnside. 2007. "Bayesian Finite Mixtures with an Unknown Number of Components: The Allocation Sampler." *Statistics and Computing* 17 (2): 147–62. https://doi.org/10.1007/s11222-006-9014-7.

Papaspiliopoulos, Omiros, Gareth O. Roberts, and Martin Sköld. 2007. "A General Framework for the Parametrization of Hierarchical Models." *Statistical Science* 22 (1): 59–73. https://doi.org/10.1214/088342307000000014.

Petrie, Ted. 1969. "Probabilistic Functions of Finite State Markov Chains." *The Annals of Mathematical Statistics* 40 (1): 97–115. https://doi.org/10.1214/aoms/1177697812.

Pohle, Jennifer, Roland Langrock, Floris M. van Beest, and Niels Martin Schmidt. 2017. "Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement." *Journal of Agricultural, Biological and Environmental Statistics* 22 (3): 270–93. https://doi.org/10.1007/s13253-017-0283-8.

Rabiner, Lawrence R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77 (2): 257–86. https://doi.org/10.1109/5.18626.

Richardson, Sylvia, and Peter J. Green. 1997. "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society, Series B* 59 (4): 731–92. https://doi.org/10.1111/1467-9868.00095.

Robert, Christian P., and George Casella. 2004. *Monte Carlo Statistical Methods*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4757-4145-2.

Robert, Christian P., Tobias Rydén, and D. Michael Titterington. 2000. "Bayesian Inference in Hidden Markov Models Through the Reversible Jump Markov Chain Monte Carlo Method." *Journal of the Royal Statistical Society, Series B* 62 (1): 57–75. https://doi.org/10.1111/1467-9868.00219.

Rousseau, Judith, and Kerrie Mengersen. 2011. "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models." *Journal of the Royal Statistical Society, Series B* 73 (5): 689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x.

Rydén, Tobias. 2008. "EM Versus Markov Chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective." *Bayesian Analysis* 3 (4): 659–88. https://doi.org/10.1214/08-BA326.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Scott, Steven L. 2002. "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century." *Journal of the American Statistical Association* 97 (457): 337–51. https://doi.org/10.1198/016214502753479464.

Seneta, Eugene. 2006. *Non-Negative Matrices and Markov Chains*. Revised. Springer. https://doi.org/10.1007/0-387-32792-4.

Stephens, Matthew. 2000. "Dealing with Label Switching in Mixture Models." *Journal of*

the Royal Statistical Society, Series B 62 (4): 795–809. https://doi.org/10.1111/1467-9868.00265.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–81. https://doi.org/10.1198/016214506000000302.

Vaart, Aad W. van der. 1998. *Asymptotic Statistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

Van Gael, Jurgen, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. "Beam Sampling for the Infinite Hidden Markov Model." *Proceedings of the 25th International Conference on Machine Learning*, 1088–95. https://doi.org/10.1145/1390156.1390293.

Viterbi, Andrew J. 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory* 13 (2): 260–69. https://doi.org/10.1109/TIT.1967.1054010.

Wainwright, Martin J., and Michael I. Jordan. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1 (1–2): 1–305. https://doi.org/10.1561/2200000001.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50 (1): 1–25. https://doi.org/10.2307/1912526.

Williams, David. 1991. *Probability with Martingales*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813658.

Wu, C. F. Jeff. 1983. "On the Convergence Properties of the EM Algorithm." *The Annals of Statistics* 11 (1): 95–103. https://doi.org/10.1214/aos/1176346060.

Zucchini, Walter, Iain L. MacDonald, and Roland Langrock. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. 2nd ed. Chapman; Hall/CRC. https://doi.org/10.1201/b20790.