Olivier Cappé, Eric Moulines and Tobias Rydén

Inference in Hidden Markov Models

May 4, 2005

Springer
Berlin Heidelberg NewYork
HongKong London
Milan Paris Tokyo

Preface

Hidden Markov models—most often abbreviated to the acronym "HMMs"—are one of the most successful statistical modelling ideas that have came up in the last forty years: the use of hidden (or unobservable) states makes the model generic enough to handle a variety of complex real-world time series, while the relatively simple prior dependence structure (the "Markov" bit) still allows for the use of efficient computational procedures. Our goal with this book is to present a reasonably complete picture of statistical inference for HMMs, from the simplest finite-valued models, which were already studied in the 1960's, to recent topics like computational aspects of models with continuous state space, asymptotics of maximum likelihood, Bayesian computation and model selection, and all this illustrated with relevant running examples. We want to stress at this point that by using the term *hidden Markov model* we do not limit ourselves to models with finite state space (for the hidden Markov chain), but also include models with continuous state space; such models are often referred to as *state-space models* in the literature.

We build on the considerable developments that have taken place during the past ten years, both at the foundational level (asymptotics of maximum likelihood estimates, order estimation, etc.) and at the computational level (variable dimension simulation, simulation-based optimization, etc.), to present an up-to-date picture of the field that is self-contained from a theoretical point of view and self-sufficient from a methodological point of view. We therefore expect that the book will appeal to academic researchers in the field of HMMs, in particular PhD students working on related topics, by summing up the results obtained so far and presenting some new ideas. We hope that it will similarly interest practitioners and researchers from other fields by leading them through the computational steps required for making inference in HMMs and/or providing them with the relevant underlying statistical theory.

The book starts with an introductory chapter which explains, in simple terms, what an HMM is, and it contains many examples of the use of HMMs in fields ranging from biology to telecommunications and finance. This chapter also describes various extension of HMMs, like models with autoregression or hierarchical HMMs. Chapter 2 defines some basic concepts like transition kernels and Markov chains. The remainder of the book is divided into three parts: *State Inference*, *Parameter Inference* and *Background and Complements*; there are also three appendices.

Part I of the book covers inference for the unobserved state process. We start in Chapter 3 by defining smoothing, filtering and predictive distributions and describe the forward-backward decomposition and the corresponding recursions. We do this in a general framework with no assumption on finiteness of the hidden state space. The special cases of HMMs with finite state space and Gaussian linear state-space models are detailed in Chapter 5. Chapter 3 also introduces the idea that the conditional distribution of the hidden Markov chain, given the observations, is Markov too, although non-homogeneous, for both ordinary and time-reversed index orderings. As a result, two alternative algorithms for smoothing are obtained. A major theme of Part I is simulation-based methods for state inference; Chapter 6 is a brief introduction to Monte Carlo simulation, and to Markov chain Monte Carlo and its applications to HMMs in particular, while Chapters 7 and 8 describe, starting from scratch, so-called sequential Monte Carlo (SMC) methods for approximating filtering and smoothing distributions in HMMs with continuous state space. Chapter 9 is devoted to asymptotic analysis of SMC algorithms. More specialized topics of Part I include recursive computation of expectations of functions with respect to smoothed distributions of the hidden chain (Section 4.1), SMC approximations of such expectations (Section 8.3) and mixing properties of the conditional distribution of the hidden chain (Section 4.3). Variants of the basic HMM structure like models with autoregression and hierarchical HMMs are considered in Sections 4.2, 6.3.2 and 8.2.

Part II of the book deals with inference for model parameters, mostly from the maximum likelihood and Bayesian points of views. Chapter 10 describes the expectation-maximization (EM) algorithm in detail, as well as its implementation for HMMs with finite state space and Gaussian linear state-space models. This chapter also discusses likelihood maximization using gradient-based optimization routines. HMMs with continuous state space do not generally admit exact implementation of EM, but require simulation-based methods. Chapter 11 covers various Monte Carlo algorithms like Monte Carlo EM, stochastic gradient algorithms and stochastic approximation EM. In addition to providing the algorithms and illustrative examples, it also contains an in-depth analysis of their convergence properties. Chapter 12 gives an overview of the framework for asymptotic analysis of the maximum likelihood estimator, with some applications like asymptotics of likelihood-based tests. Chapter 13 is about Bayesian inference for HMMs, with the focus being on models with finite state space. It covers so-called reversible jump MCMC algorithms for choosing between models of different dimensionality, and contains detailed examples illustrating these as well as simpler algorithms. It also contains a section on multiple imputation algorithms for global maximization of the posterior density.

Part III of the book contains a chapter on discrete and general Markov chains, summarizing some of the most important concepts and results and applying them to HMMs. The other chapter of this part focuses on order estimation for HMMs with both finite state space and finite output alphabet; in particular it describes how concepts from information theory are useful for elaborating on this subject.

Various parts of the book require different amounts of, and also different kinds of, prior knowledge from the reader. Generally we assume familiarity with probability and statistical estimation at the levels of Feller (1971) and Bickel and Doksum (1977), respectively. Some prior knowledge of Markov chains (discrete and/or general) is very helpful, although Part III does contain a primer on the topic; this chapter should however be considered more a brush-up than a comprehensive treatise of the subject. A reader with that knowledge will be able to understand most parts of the book. Chapter 13 on Bayesian estimation features a brief introduction to the subject in general but, again, some previous experience with Bayesian statistics will undoubtedly be of great help. The more theoretical parts of the book (Section 4.3, Chapter 9, Sections 11.2–11.3, Chapter 12, Sections 14.2–14.3 and Chapter 15) require knowledge of probability theory at the measure-theoretic level for a full understanding, even though most of the results as such can be understood without it.

There is no need to read the book in linear order, from cover to cover. Indeed, this is probably the wrong way to read it! Rather we encourage the reader to first go through the more algorithmic parts of the book, to get an overall view of the subject, and then, if desired, later return to the theoretical parts for a fuller understanding. Readers with particular topics in mind may of course be even more selective. A reader interested in the EM algorithm, for instance, could start with Chapter 1, have a look at Chapter 2, and then proceed to Chapter 3 before reading about the EM algorithm in Chapter 10. Similarly a reader interested in simulation-based techniques could go to Chapter 6 directly, perhaps after reading some of the introductory parts, or even directly to Section 6.3 if he/she is already familiar with MCMC methods. Each of the two chapters entitled “Advanced Topics in...” (Chapters 4 and 8) is really composed of three disconnected complements to Chapters 3 and 7, respectively. As such, the sections that compose Chapters 4 and 8 may be read independently of one another. Most chapters end with a section entitled “Complements” whose reading is not required for understanding other parts of the book—most often, this section mostly contains bibliographical notes—although in some chapters (9 and 11 in particular) it also features elements needed to prove the results stated in the main text.

Even in a book of this size, it is impossible to include all aspects of hidden Markov models. We have focused on the use of HMMs to model long, potentially stationary, time series; we call such models *ergodic HMMs*. In other applications, for instance speech recognition or protein alignment, HMMs are used to represent short variable-length sequences; such models are often called *left-to-right HMMs* and are hardly mentioned in this book. Having said that we stress that the computational tools for both classes of HMMs are virtually the same. There are also a number of generalizations of HMMs which we do not consider. In Markov random fields, as used in image processing applications, the Markov chain is replaced by a graph of dependency which may be represented as a two-dimensional regular lattice. The numerical techniques that can be used for inference in hidden Markov random fields are similar to some of the methods studied in this book but the statistical side is very different. Bayesian networks are even more general since the dependency structure is allowed to take any form representable by a (directed or undirected) graph. We do not consider Bayesian networks in their generality although some of the concepts developed in the Bayesian networks literature (the graph representation, the sum-product algorithm) are used. Continuous-time HMMs may also be seen as a further generalization of the models considered in this book. Some of these "continuous-time HMMs", and in particular partially observed diffusion models used in mathematical finance, have recently received considerable attention. We however decided this topic to be outside the range of the book; furthermore, the stochastic calculus tools needed for studying these continuous-time models are not appropriate for our purpose.

We acknowledge the help of Stéphane Boucheron, Randal Douc, Gersende Fort, Elisabeth Gassiat, Christian P. Robert, and Philippe Soulier, who participated in the writing of the text and contributed the two chapters that compose Part III (see next page for details of the contributions). We are also indebted to them for suggesting various forms of improvement in the notations, layout, etc., as well as helping us tracking typos and errors. We thank François Le Gland and Catherine Matias for participating in the early stages of this book project. We are grateful to Christophe Andrieu, Arnaud Doucet, Hans Künsch, Steve Levinson, Ya'acov Ritov and Mike Titterington, who provided various helpful inputs and comments. Finally, we thank John Kimmel of Springer for his support and enduring patience.

Paris, France
& Lund, Sweden,
March 2005

Olivier Cappé
Eric Moulines
Tobias Rydén

Contributors

We are grateful to

Randal Douc
Ecole Polytechnique
Christian P. Robert
CREST INSEE & Université Paris-Dauphine

for their contributions to Chapters 9 (Randal) and 6, 7, and 13 (Christian) as well as for their help in proofreading these and other parts of the book

Chapter 14 was written by

Gersende Fort
CNRS & LMC-IMAG
Philippe Soulier
Université Paris-Nanterre

with Eric Moulines

Chapter 15 was written by

Stéphane Boucheron
Université Paris VII-Denis Diderot
Elisabeth Gassiat
Université d'Orsay, Paris-Sud

Contents

Part I State Inference

Part II Parameter Inference

Part III Background and Complements

Part IV Appendices

1

Introduction

1.1 What Is a Hidden Markov Model?

A *hidden Markov model* (abbreviated HMM) is, loosely speaking, a Markov chain observed in noise. Indeed, the model comprises a Markov chain, which we will denote by $\{X_k\}_{k\geq 0}$, where k is an integer index. This Markov chain is often assumed to take values in a finite set, but we will not make this restriction in general, thus allowing for a quite arbitrary state space. Now, the Markov chain is *hidden*, that is, it is not observable. What is available to the observer is another stochastic process $\{Y_k\}_{k\geq 0}$, linked to the Markov chain in that X_k governs the distribution of the corresponding Y_k. For instance, Y_k may have a normal distribution, the mean and variance of which is determined by X_k, or Y_k may have a Poisson distribution whose mean is determined by X_k. The underlying Markov chain $\{X_k\}$ is sometimes called the *regime*, or *state*. All statistical inference, even on the Markov chain itself, has to be done in terms of $\{Y_k\}$ only, as $\{X_k\}$ is not observed. There is also a further assumption on the relation between the Markov chain and the observable process, saying that X_k must be the only variable of the Markov chain that affects the distribution of Y_k. This is expressed more precisely in the following formal definition.

A hidden Markov model is a bivariate discrete time process $\{X_k, Y_k\}_{k\geq 0}$, where $\{X_k\}$ is a Markov chain and, conditional on $\{X_k\}$, $\{Y_k\}$ is a sequence of independent random variables such that the conditional distribution of Y_k only depends on X_k. We will denote the state space of the Markov chain $\{X_k\}$ by X and the set in which $\{Y_k\}$ takes its values by Y.

The dependence structure of an HMM can be represented by a *graphical model* as in Figure 1.1. Representations of this sort use a directed graph without loops to describe dependence structures among random variables. The nodes (circles) in the graph correspond to the random variables, and the edges (arrows) represent the structure of the joint probability distribution, with the interpretation that the latter may be factored as a product of the conditional distributions of each node given its “parent” nodes (those that are directly

2

Main Definitions and Notations

We now formally describe hidden Markov models, setting the notations that will be used throughout the book. We start by reviewing the basic definitions and concepts pertaining to Markov chains.

2.1 Markov Chains

2.1.1 Transition Kernels

Definition 2.1.1 (Transition Kernel). *Let* $(\mathsf{X}, \mathcal{X})$ *and* $(\mathsf{Y}, \mathcal{Y})$ *be two measurable spaces. An* unnormalized transition kernel *from* $(\mathsf{X}, \mathcal{X})$ *to* $(\mathsf{Y}, \mathcal{Y})$ *is a function* $Q : \mathsf{X} \times \mathcal{Y} \to [0, \infty]$ *that satisfies*

(i) for all $x \in \mathsf{X}$, $Q(x, \cdot)$ *is a positive measure on* $(\mathsf{Y}, \mathcal{Y})$*;*
(ii) for all $A \in \mathcal{Y}$*, the function* $x \mapsto Q(x, A)$ *is measurable.*

If $Q(x, \mathsf{Y}) = 1$ *for all* $x \in \mathsf{X}$*, then* Q *is called a* transition kernel*, or simply a* kernel*. If* $\mathsf{X} = \mathsf{Y}$ *and* $Q(x, \mathsf{X}) = 1$ *for all* $x \in \mathsf{X}$*, then* Q *will also be referred to as a* Markov transition kernel *on* $(\mathsf{X}, \mathcal{X})$*.*

An (unnormalized) transition kernel Q *is said to* admit a density *with respect to the positive measure* μ *on* Y *if there exists a non-negative function* $q : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$*, measurable with respect to the product* σ*-field* $\mathcal{X} \otimes \mathcal{Y}$*, such that*

$$Q(x, A) = \int_A g(x, y)\, \mu(dy) \;, \qquad A \in \mathcal{Y} \;.$$

The function q *is then referred to as an (unnormalized)* transition density function*.*

When X *and* Y *are countable sets it is customary to write* $Q(x, y)$ *as a shorthand notation for* $Q(x, \{y\})$*, and* Q *is generally referred to as a* transition matrix *(whether or not* X *and* Y *are finite sets).*

We summarize below some key properties of transition kernels, introducing important pieces of notation that are used in the following.

3

Filtering and Smoothing Recursions

This chapter deals with a fundamental issue in hidden Markov modeling: given a fully specified model and some observations $Y_0, \ldots, Y_n$, what can be said about the corresponding unobserved state sequence $X_0, \ldots, X_n$? More specifically, we shall be concerned with the evaluation of the conditional distributions of the state at index k, X_k, given the observations $Y_0, \ldots, Y_n$, a task that is generally referred to as *smoothing*. There are of course several options available for tackling this problem (Anderson and Moore, 1979, Chapter 7) and we focus, in this chapter, on the *fixed-interval smoothing* paradigm in which n is held fixed and it is desired to evaluate the conditional distributions of X_k for all indices k between 0 and n. Note that only the general mechanics of the smoothing problem are dealt with in this chapter. In particular, most formulas will involve integrals over X. We shall not, for the moment, discuss ways in which these integrals can be effectively evaluated, or at least approximated, numerically. We postpone this issue to Chapter 5, which deals with some specific classes of hidden Markov models, and Chapters 6 and 7, in which generally applicable Markov chain Monte Carlo methods or sequential importance sampling techniques are reviewed.

The driving line of this chapter is the existence of a variety of smoothing approaches that involve a number of steps that only increase linearly with the number of observations. This is made possible by the fact (to be made precise in Section 3.3) that conditionally on the observations $Y_0, \ldots, Y_n$, the state sequence still is a Markov chain, albeit a non-homogeneous one.

Readers already familiar with the field could certainly object that as the probabilistic structure of any hidden Markov model may be represented by the generic probabilistic network drawn in Figure 1.1 (Chapter 1), the fixed interval smoothing problem under consideration may be solved by applying the general principle known as probability propagation or sum-product—see Cowell *et al.* (1999) or Frey (1998) for further details and references. As patent however from Figure 1.1, the graph corresponding to the HMM structure is so simple and systematic in its design that efficient instances of the probability propagation approach are all based on combining two systematic phases:

4

Advanced Topics in Smoothing

This chapter covers three distinct complements to the basic smoothing relations developed in the previous chapter.

In the first section, we provide recursive smoothing relations for computing smoothed expectations of general functions of the hidden states. In many respects, this technique is reminiscent of the filtering recursion detailed in Section 3.2.2, but somewhat harder to grasp because the quantity that needs to be updated recursively is less directly interpretable.

In the second section, it is shown that the filtering and smoothing approaches discussed so far (including those of Section 4.1) may be applied, with minimal adaptations, to a family of models that is much broader than simply the hidden Markov models. We consider in some detail the case of hierarchical HMMs (introduced in Section 1.3.4) for which marginal filtering and smoothing formulas are still available, despite the fact that the hierarchic component of the state process is not *a posteriori* Markovian.

The third section is different in nature and is devoted to the so-called *forgetting* property of the filtering and smoothing recursions, which are instrumental in the statistical theory of HMMs (see Chapter 12). Forgetting refers to the fact that observations that are either far back in the past or in the remote future (relative to the current time index) have little impact on the posterior distribution of the current state. Although this section is written to be self-contained, its content is probably better understood after some exposure to the stability properties of Markov chains as can be found in Chapter 14.

4.1 Recursive Computation of Smoothed Functionals

Chapter 3 mostly dealt with *fixed-interval smoothing*, that is, computation of $\phi_{k|n}$[1] for a fixed value of the observation horizon n and for all indices

[1] Note that we omit the dependence with respect to the initial distribution ν, which is not important in this section.

5

Applications of Smoothing

Remember that in the previous two chapters, we basically considered that integration over X was a feasible operation. This is of course not the case in general, and numerical evaluation of the integrals involved in the smoothing recursions turns out to be a difficult task. In Chapters 6 and 7, generally applicable methods for approximate smoothing, based on Monte Carlo simulations, will be considered. Before that, we first examine two very important particular cases in which an exact numerical evaluation is feasible: models with finite state space in Section 5.1 and Gaussian linear state-space models in Section 5.2. Most of the concepts to be used below have already been introduced in Chapters 3 and 4, and the current chapter mainly deals with computational aspects and algorithms. It also provides concrete examples of application of the methods studied in the previous chapters.

Note that we do not yet consider examples of application of the technique studied in Section 4.1, as the nature of functionals that can be computed recursively will only become more explicit when we discuss the EM framework in Chapter 10. Corresponding examples will be considered in Section 10.2.

5.1 Models with Finite State Space

We first consider models for which the state space X of the hidden variables is finite, that is, when the unobservable states may only take a finite number of distinct values. In this context, the smoothing recursions discussed in Chapter 3 take the familiar form described in the seminal paper by Baum *et al.* (1970) as well as Rabiner's (1989) tutorial (which also covers scaling issues). Section 5.1.2 discusses a technique that is of utmost importance in many applications, for instance digital communications and speech processing, by which one can determine the maximum *a posteriori* sequence of hidden states given the observations.

6

Monte Carlo Methods

This chapter takes a different path to the study of hidden Markov models in that it abandons the pursuit of closed-form formulas and exact algorithms to cover instead simulation-based techniques. This change of perspective allows for a much broader coverage of HMMs, which is not restricted to the specific cases discussed in Chapter 5. In this chapter, we consider *sampling* the unknown sequence of states $X_0, \ldots, X_n$ *conditionally on the observed sequence* $Y_0, \ldots Y_n$. In subsequent chapters, we will also use simulation to do inference about the parameters of HMMs, either using simulation-based stochastic algorithms that optimize the likelihood (Chapter 11) or in the context of Bayesian joint inference on the states and parameters (Chapter 13). But even the sole simulation of the missing states may prove itself a considerable challenge in complex settings like continuous state-space HMMs. Therefore, and although these different tasks are presented in separate chapters, simulating hidden states in a model whose parameters are assumed to be known is certainly not disconnected from parameter estimation to be discussed in Chapters 11 and 13.

6.1 Basic Monte Carlo Methods

Although we will not go into a complete description of simulation methods in this book, the reader must be aware that recent developments of these methods have offered new opportunities for inference in complex models like hidden Markov models and their generalizations. For a more in-depth covering of these simulation methods and their implications see, for instance, the books by Chen and Shao (2000), Evans and Swartz (2000), Liu (2001), and Robert and Casella (2004).

7

Sequential Monte Carlo Methods

The use of Monte Carlo methods for non-linear filtering can be traced back to the pioneering contributions of Handschin and Mayne (1969) and Handschin (1970). These early attempts were based on sequential versions of the *importance sampling* paradigm, a technique that amounts to simulating samples under an instrumental distribution and then approximating the target distributions by weighting these samples using appropriately defined *importance weights*. In the non-linear filtering context, importance sampling algorithms can be implemented sequentially in the sense that, by defining carefully a sequence of instrumental distributions, it is not needed to regenerate the population of samples from scratch upon the arrival of each new observation. This algorithm is called *sequential importance sampling*, often abbreviated SIS. Although the SIS algorithm has been known since the early 1970s, its use in non-linear filtering problems was rather limited at that time. Most likely, the available computational power was then too limited to allow convincing applications of these methods. Another less obvious reason is that the SIS algorithm suffers from a major drawback that was not clearly identified and properly cured until the seminal paper by Gordon *et al.* (1993). As the number of iterations increases, the importance weights tend to degenerate, a phenomenon known as *sample impoverishment* or *weight degeneracy*. Basically, in the long run most of the samples have very small normalized importance weights and thus do not significantly contribute to the approximation of the target distribution. The solution proposed by Gordon *et al.* (1993) is to allow rejuvenation of the set of samples by duplicating the samples with high importance weights and, on the contrary, removing samples with low weights.

The *particle filter* of Gordon *et al.* (1993) was the first successful application of sequential Monte Carlo techniques to the field of non-linear filtering. Since then, sequential Monte Carlo (or SMC) methods have been applied in many different fields including computer vision, signal processing, control, econometrics, finance, robotics, and statistics (Doucet *et al.*, 2001a; Ristic *et al.*, 2004). This chapter reviews the basic building blocks that are needed to implement a sequential Monte Carlo algorithm, starting with concepts re-

8

Advanced Topics in Sequential Monte Carlo

This chapter deals with three disconnected topics that correspond to variants and extensions of the sequential Monte Carlo framework introduced in the previous chapter. Remember that we have already examined in Section 7.2 a first and very important degree of freedom in the application of sequential Monte Carlo methods, namely the choice of the instrumental kernel R_k used to simulate the trajectories of the particles. We now consider solutions that depart, more or less significantly, from the sequential importance sampling with resampling (SISR) method of Algorithm 7.3.4.

The first section covers a far-reaching revision of the principles behind the SISR algorithm in which sequential Monte Carlo is interpreted as a repeated sampling task. This reinterpretation suggests several other sequential Monte Carlo schemes that differ, sometimes significantly, from the SISR approach. Section 8.2 reviews methods that exploit the specific hierarchical structure found in some hidden Markov models, and in particular in conditionally Gaussian linear state-space models (CGLSSMs). The algorithms to be considered there combine the sequential simulation approach presented in the previous chapter with the Kalman filtering recursion discussed in Chapter 5. Finally, Section 8.3 discusses the use of sequential Monte Carlo methods for approximating smoothed quantities of the form introduced in Section 4.1.

8.1 Alternatives to SISR

We first present a reinterpretation of the objectives of the sequential importance sampling with resampling (SISR) algorithm in Section 7.3. This new interpretation suggests a whole range of different approaches that combines more closely the sampling (trajectory update) and resampling (weight reset) operators involved in the SISR algorithm.

In the basic SISR approach (Algorithm 7.3.4), we expect that after a resampling step, say at index k, the particle trajectories $\xi^1_{0:k}, \ldots, \xi^N_{0:k}$ approximately form an i.i.d. sample of size N from the distribution $\phi_{0:k|k}$. We will

9

Analysis of Sequential Monte Carlo Methods

The previous chapters have described many algorithms to approximate prediction, filtering, and smoothing distributions. The development of these algorithms was motivated mainly on heuristic grounds, and the validity of these approximations is of course a question of central interest. In this chapter, we analyze these methods, mainly from an asymptotic perspective. That is, we study the behavior of the estimators in situations where the number of particles gets large. Asymptotic analysis provides approximations that in many circumstances have proved to be relatively robust. Most importantly, asymptotic arguments provide insights in the sampling methodology by verifying that the procedures are sensible, providing a framework for comparing competing procedures, and providing understanding of the impact of different options (choice of importance kernel, etc.) on the overall performance of the samplers.

9.1 Importance Sampling

9.1.1 Unnormalized Importance Sampling

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. Define on $(\mathsf{X}, \mathcal{X})$ two probability distributions: the *target distribution* μ and the *instrumental distribution* ν.

Assumption 9.1.1. *The target distribution μ is absolutely continuous with respect to the instrumental distribution ν, $\mu \ll \nu$, and $d\mu/d\nu > 0$ ν-a.s.*

Let f be a real-valued measurable function on X such that $\mu(|f|) = \int |f| \, d\mu < \infty$. Denote by $\xi^1, \xi^2, \ldots$ an i.i.d. sample from ν and consider the estimator

$$\tilde{\mu}^{\mathrm{IS}}_{\nu,N}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi^i) \frac{d\mu}{d\nu}(\xi^i) \; . \tag{9.1}$$

10

Maximum Likelihood Inference, Part I: Optimization Through Exact Smoothing

In previous chapters, we have focused on structural results and methods for HMMs, considering in particular that the models under consideration were always perfectly known. In most situations, however, the model cannot be fully specified beforehand, and some of its parameters need to be calibrated based on observed data. Except for very simplistic instances of HMMs, the structure of the model is sufficiently complex to prevent the use of direct estimators such as those provided by moment or least squares methods. We thus focus in the following on computation of the *maximum likelihood estimator*.

Given the specific structure of the likelihood function in HMMs, it turns out that the key ingredient of any optimization method applicable in this context is the ability to compute smoothed functionals of the unobserved sequence of states. Hence the methods discussed in the second part of the book for evaluating smoothed quantities are instrumental in devising parameter estimation strategies.

This chapter only covers the class of HMMs discussed in Chapter 5, for which the smoothing recursions described in Chapters 3 and 4 may effectively be implemented on computers. For such models, the likelihood function is computable, and hence our main task will be to optimize a possibly complex but entirely known function. The topic of this chapter thus relates to the more general field of numerical optimization. For models that do not allow for exact numerical computation of smoothing distributions, this chapter provides a framework from which numerical approximations can be built. Those will be discussed in Chapter 11.

10.1 Likelihood Optimization in Incomplete Data Models

To describe the methods as concisely as possible, we adopt a very general viewpoint in which we only assume that the likelihood function of interest may be written as the marginal of a higher dimensional function. In the terminology introduced by Dempster *et al.* (1977), this higher dimensional function is

11

Maximum Likelihood Inference, Part II: Monte Carlo Optimization

This chapter deals with maximum likelihood parameter estimation for models in which the smoothing recursions of Chapter 3 cannot be implemented. The task is then considerably more difficult, as it is not even possible to evaluate the likelihood to be maximized. Most of the methods applicable in such cases are reminiscent of the iterative optimization procedures (EM and gradient methods) discussed in the previous chapter but rely on approximate smoothing computations based on some form of Monte Carlo simulation. In this context, the methods covered in Chapters 6 and 7 for simulating the unobservable sequence of states conditionally on the observations play a prominent role.

It is important to distinguish the topic of this chapter with a distinct—although not entirely disconnected—problem. The methods discussed in the previous chapters were all based on local exploration (also called hill-climbing strategies) of the likelihood function. Such methods are typically unable to guarantee that the point reached at convergence is a global maximum of the function; indeed, it may well be a local maximum only or even a saddle point—see Section 10.5 for details regarding the EM algorithm. Many techniques have been proposed to overcome this significant difficulty, and most of them belong to a class of methods that Geyer (1996) describes as random search optimization. Typical examples are the so-called genetic and simulated annealing algorithms that both involve simulating random moves in the parameter space (see also Section 13.3, which describes a technique related to simulated annealing). In these approaches, the main motivation for using simulations (in parameter space and/or hidden variable space) is the hope to design more robust optimization rules that can avoid local maxima.

The focus of the current chapter is different, however, as we examine below methods that can be considered as simulation-based extensions of approaches introduced in the previous chapter. The primary objective is here to provide tools for maximum likelihood inference also for the class of HMMs in which exact smoothing is not available.

12

Statistical Properties of the Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is one of the backbones of statistics, and as we have seen in previous chapters, it is very much appropriate also for HMMs, even though numerical approximations are required when the state space is not finite. A standard result in statistics says that, except for "atypical cases", the MLE is consistent, asymptotically normal with asymptotic (scaled) variance equal to the inverse Fisher information matrix, and efficient. The purpose of the current chapter is to show that these properties are indeed true for HMMs as well, provided some conditions of rather standard nature hold. We will also employ the asymptotic results obtained to verify the validity of certain likelihood-based tests.

Recall that the distribution (law) P of $\{Y_k\}_{k\geq 0}$ depends on a parameter θ that lies in a parameter space Θ, which we assume is a subset of $\mathbb{R}^{d_\theta}$ for some d_θ. Commonly, θ is a vector containing some components that parameterize the transition kernel of the hidden Markov chain—such as the transition probabilities if the state space X is finite—and other components that parameterize the conditional distributions of the observations given the states. Throughout the chapter, it is assumed that the HMM model is, for all θ, fully dominated in the sense of Definition 2.2.3 and that the underlying Markov chain is positive (see Definition 14.2.26).

Assumption 12.0.1.

(i) There exists a probability measure λ *on* $(\mathsf{X}, \mathcal{X})$ *such that for any* $x \in \mathsf{X}$ *and any* $\theta \in \Theta$, $Q_\theta(x, \cdot) \ll \lambda$ *with transition density* q_θ. *That is,* $Q_\theta(x, A) = \int q_\theta(x, x')\,\lambda(dx')$ *for* $A \in \mathcal{X}$.

(ii) There exists a probability measure μ *on* $(\mathsf{Y}, \mathcal{Y})$ *such that for any* $x \in \mathsf{X}$ *and any* $\theta \in \Theta$, $G_\theta(x, \cdot) \ll \mu$ *with transition density function* g_θ. *That is,* $G_\theta(x, A) = \int g_\theta(x, y)\,\mu(dy)$ *for* $A \in \mathcal{Y}$.

(iii) For any $\theta \in \Theta$, Q_θ *is positive, that is,* Q_θ *is phi-irreducible and admits a (necessarily unique) invariant distribution denoted by* π_θ.

13

Fully Bayesian Approaches

Some previous chapters have already mentioned MCMC and conditional (or posterior) distributions, especially in the set-up of posterior state estimation and simulation. The spirit of this chapter is obviously different in that it covers the fully Bayesian processing of HMMs, which means that, besides the hidden states and their conditional (or parameterized) distributions, the model parameters are assigned probability distributions, called *prior distributions*, and the inference on these parameters is of Bayesian nature, that is, conditional on the observations (or the *data*). Because more advanced Markov chain Monte Carlo methodology is also needed for this fully Bayesian processing, additional covering of MCMC methods, like reversible jump techniques, will be given in this chapter (Section 13.2). The emphasis is put on HMMs with finite state space (X is finite), but some facts are general and the case of continuous state space is addressed at some points.

13.1 Parameter Estimation

13.1.1 Bayesian Inference

Although the whole apparatus of modern Bayesian inference cannot be discussed here (we refer the reader to, e.g., Robert, 2001, or Gelman *et al.*, 1995), we briefly recall the basics of a Bayesian analysis of a statistical model, and we also introduce some notation not used in earlier chapters.

Given a general parameterized model

$$Y \sim p(y|\theta), \quad \theta \in \Theta ,$$

where $p(y|\theta)$ thus denotes a parameterized density, the idea at the core of Bayesian analysis is to provide an inferential assessment (on θ) *conditional on the realized value of* Y, which we denote (as usual) by y. Obviously, to give a proper probabilistic meaning to this conditioning, θ itself must be embedded with a probability distribution called the *prior distribution*, which

14

Elements of Markov Chain Theory

14.1 Chains on Countable State Spaces

We review the key elements of the mathematical theory developed for studying the limiting behavior of Markov chains. In this first section, we restrict ourselves to the case where the state space X is countable, which is conceptually simpler. On our way, we will also meet a number of important concepts to be used in the next section when dealing with Markov chains on general state spaces.

14.1.1 Irreducibility

Let $\{X_k\}_{k\geq 0}$ be a Markov chain on a countable state space X with transition matrix Q. For any $x \in X$, we define the first hitting time σ_x on x and the return time τ_x to x respectively as

$$\sigma_x = \inf\{n \geq 0 : X_n = x\} \ , \tag{14.1}$$

$$\tau_x = \inf\{n \geq 1 : X_n = x\} \ , \tag{14.2}$$

where, by convention, $\inf \emptyset = +\infty$. The successive hitting times $\sigma_x^{(n)}$ and return times $\tau_x^{(n)}$, $n \geq 0$, are defined inductively by

$$\sigma_x^{(0)} = 0,\ \sigma_x^{(1)} = \sigma_x,\ \sigma_x^{(n+1)} = \inf\{k > \sigma_x^{(n)} :\ X_k = x\} \ ,$$

$$\tau_x^{(0)} = 0,\ \tau_x^{(1)} = \tau_x,\ \tau_x^{(n+1)} = \inf\{k > \tau_x^{(n)} :\ X_k = x\} \ .$$

For two states x and y, we say that state x *leads to* state y, which we write $x \to y$, if $\mathrm{P}_x(\sigma_y < \infty) > 0$. In words, x leads to y if the state y can be reached from x. An alternative, equivalent definition is that there exists some integer $n \geq 0$ such that the n-step transition probability $Q^n(x, y) > 0$. If both x leads to y and y leads to x, then we say that the x and y *communicate*, which we write $x \leftrightarrow y$.

15

An Information-Theoretic Perspective on Order Estimation

Statistical inference in hidden Markov models with finite state space X has to face a serious problem: order identification. The order of an HMM $\{Y_k\}_{k\geq 1}$ over Y (in this chapter, we let indices start at 1) is the minimum size of the hidden state space X of an HMM over (X, Y) that can generate $\{Y_k\}_{k\geq 1}$. In many real-life applications of HMM modeling, no hints about this order are available. As order misspecification is an impediment to parameter estimation, consistent order identification is a prerequisite to HMM parameter estimation.

Furthermore, HMM order identification is a distinguished representative of a family of related problems that includes Markov order identification. In all those problems, a nested family of models is given, and the goal is to identify the smallest model that contains the distribution that has generated the data. Those problems differ in an essential way according to whether identifiability does or does not depend on correct order specification.

Order identification problems are related to composite hypothesis testing. As the performance of generalized likelihood ratio testing in this framework is still a matter of debate, order identification problems constitute benchmarks for which the performance of generalized likelihood ratio testing can be investigated (see Zeitouni *et al.*, 1992). As a matter of fact, analyzing order identification issues boils down to understanding the simultaneous behavior of (possibly infinitely) many maximum likelihood estimators. When identifiability depends on correct order specification, universal coding arguments have proved to provide very valuable insights into the behavior of likelihood ratios. This is the main reason why source coding concepts and techniques have become a standard tool in the area.

This chapter presents four kinds of results: first, in a Bayesian setting, a general consistency result provides hints about the ideal penalties that could be used in penalized maximum likelihood order estimation. Then universal coding arguments are shown to provide a general construction of strongly consistent order estimators. Afterwards, a general framework for analyzing the Bahadur efficiency of order estimation procedures is presented, following the lines of Gassiat and Boucheron (2003). Consistency and efficiency results