

Chapter 2

Order Estimation and Model Selection

S. Boucheron¹ and E. Gassiat²

This text is supposed to become a chapter in a forthcoming book on *Inference in Hidden Markov Models* edited by O. Cappé and T. Rydén.

Roadmap

Statistical inference in Hidden Markov Models has to face a serious problem: order identification. The order of a HMM $(Y_t)_{t \in \mathbb{N}}$ over \mathcal{Y} is the minimum size of the hidden state space \mathcal{X} of a HMM over $(\mathcal{X}, \mathcal{Y})$ that can generate $(Y_t)_t$. 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 which 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 the identifiability depends 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 where the performances of generalized likelihood ratio testing can be investigated. As a matter of fact, analyzing order identification issues boils down to understanding the simultaneous behavior of (possibly infinitely) many ML 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: a first very general consistency result in a Bayesian setting provides hints about the ideal penalties that could be used in penalized maximum likelihood order estimation. Then we provide a general construction for strongly consistent order estimators based on universal coding arguments. The third main result reports a recent *tour de force* by Csiszár and Shields (2000) who show that the Bayesian Information Criterion provides a strongly consistent Markov order estimator. We conclude by presenting a general framework for analyzing the Bahadur efficiency of order estimation procedures following the line Gassiat and Boucheron (to appear).

¹LRI UMR 8623 CNRS, Université Paris-Sud

²Mathématiques, Université Paris-Sud

2.1 Model Order Identification: what is it about ?

In the preceding chapters, we have been concerned with inference problems in HMMs where the hidden state space is known in advance: it might be either finite with known cardinality or compact under restrictive conditions, see ???. In this chapter, unless otherwise specified, we focus on HMMs with finite state space of unknown cardinality. Henceforth the observed alphabet \mathcal{Y} is assumed to be fixed. In this chapter \mathcal{M}^r denote the set distributions of \mathcal{Y} -valued processes $(Y_t)_{t \in \mathbb{N}}$ that can be generated by a HMM with hidden state space \mathcal{X} of cardinality r .

The parameter space associated with \mathcal{M}^r is Θ^r . Note that even if all the finite-dimensional distributions of $(Y_t)_{t \in \mathbb{N}}$ are known, deciding whether the distribution of $(Y_t)_{t \in \mathbb{N}}$ belongs to \mathcal{M}^r or even to $\cup_r \mathcal{M}^r$ is not trivial (Finesso, 1991, Chapter 1). Elementary arguments show that $\mathcal{M}^r \subseteq \mathcal{M}^{r+1}$, further reflexion allows to check that this inclusion is strict. Hence, for a fixed observation alphabet, the sequence $(\mathcal{M}^r)_{r \in \mathbb{N}}$ defines a nested sequence of models. We may now define the main topic of this chapter: the order of a HMM.

HMM order

Let $(Y_t)_{t \in \mathbb{N}}$ denote a HMM over \mathcal{Y} , the order of $(Y_t)_{t \in \mathbb{N}}$ is the smallest integer r such that the distribution of $(Y_t)_{t \in \mathbb{N}}$ belongs to \mathcal{M}^r .

Henceforth, when dealing with HMM $(Y_t)_{t \in \mathbb{N}}$, the order of $(Y_t)_{t \in \mathbb{N}}$ will be denoted by r^* and θ^* will denote a parameterization of this distribution in Θ^{r^*} . The distribution of the process will be denoted by \mathbf{P}^* .

Assume for a moment that we are given an infinite sequence of observations of a HMM $(Y_t)_{t \in \mathbb{N}}$: y_1, \dots, y_t, \dots , that we are told that the order of $(Y_t)_{t \in \mathbb{N}}$ is less than some r_0 , and that we are asked to estimate a parameterization of the distribution of $(Y_t)_{t \in \mathbb{N}}$. It might seem that Maximum Likelihood estimates in Θ^{r_0} would perform well in such a situation. Unfortunately, if the order of $(Y_t)_{t \in \mathbb{N}}$ is strictly smaller than r_0 , Maximum Likelihood estimation will run into trouble, see Chapter ???. As a matter of fact, if $r^* < r_0$, then θ^* is not identifiable in Θ^{r_0} . Hence, when confronted with such an estimation problem, it is highly reasonable to first estimate r^* , and then to proceed to Maximum Likelihood estimation of θ^* .

The *order estimation* question is then the following: given an outcome of the process $(y_t)_{t \leq n}$ with distribution in $\cup_r \mathcal{M}^r$, can we identify r^* ?

Order estimator

An order estimation procedure is a sequence of estimators $\hat{r}_1, \dots, \hat{r}_t, \dots$ that, given input sequences of length $1, \dots, t, \dots$ outputs estimates $\hat{r}_t(y_{1:t})$ of r^* .

A sequence of estimators is strongly consistent on θ^* if θ^* -almost surely, the sequence $\hat{r}_1, \dots, \hat{r}_t, \dots$ eventually converges toward r^* .

2.2 Order estimation in perspective

The ambition of this chapter is not only to provide with a state-of-the-art exposition of order estimation in HMMs but also to provide a perspective. There are actually many other order estimation problems in the statistical or the information-theoretical literature. All pertain to the estimation of the dimension of a model. We may quote for example:

- Models for sequences on a finite alphabet .
 - Estimating the order of a Markov process on a finite alphabet. In that case, the order should be understood as the Markov order of the process. Finesso et al. (1996), Csiszár and Shields (2000), Csiszár (2002), see Section 2.7 for precise definitions and recent advances on this topic.

- Estimating the order of semi-Markov models, which have proved to be valuable tools in telecommunication engineering.
- Estimating the order in stochastic context-free grammars, which are currently considered in genomics, Durbin et al. (1998).
- Models for sequences in general sets.
 - Estimating the number of populations in a mixture, Dacunha-Castelle and Gassiat (1997b), Dacunha-Castelle and Gassiat (1997a), Dacunha-Castelle and Gassiat (1999) and Gassiat (2002).
 - Estimating the number of change points in detection problems.
 - Estimating the number of functions in a regression.
 - Blind deconvolution of communication channels.
 - Estimating the order of ARMA models, Azencott and Dacunha-Castelle (1984), Dacunha-Castelle and Gassiat (1999).

Hence HMM order estimation is both interesting *per se* and as a paradigm of a rich family of statistical problems. Equipped with those strong motivations, we may raise our two technical questions:

1. Does there exist (strongly) consistent HMM order estimators? Is it possible to design generic order estimation procedures?
2. How efficient are the putative consistent HMM order estimators?

The analysis of order estimation problems is currently influenced by the theory of *universal coding* from Information Theory and by the theory of *composite hypothesis testing* from plain old Statistics. The first perspective provides a convenient framework for designing consistent order estimators, while the second provides guidelines in the analysis of the performance of order estimators. As a matter of fact code-based order estimators turn out to be analyzed as penalized Maximum Likelihood estimators.

Our current understanding of Markov order estimation will provide insights into the HMM order estimation problem. Though this order estimation problem is apparently very similar to the HMM order estimation problem, this resemblance should be taken cautiously. The two problems actually exhibit sharply distinct features:

- Whatever the order r^* of the Markov chain $(X_t)_{t \in \mathbb{N}}$, Maximum Likelihood estimation using orders $r > r^*$ remains strongly consistent and the Maximum Likelihood estimator retains the asymptotic normality property (see Section 2.7).
- Whatever the value of r , the maximum likelihood estimator is uniquely defined, and it can be computed easily from a (r -dependent) finite-dimensional sufficient statistic. This is in sharp contrast with the hardness of the computation of the maximum likelihood in HMM (see Chapter ??).

This example illustrates the fact that analyzing order estimation problems depends on whether the parameter describing the distribution of the observations *is or is not* identifiable when taken in a model with bigger order. When the parameter is still identifiable in larger models, stochastic behavior of the maximum likelihood statistic is well understood and can be cast into the old framework created by Wilks, Wald and Chernoff. When the parameter is no longer identifiable in larger models, stochastic description of the maximum likelihood statistic has to be investigated on an ad hoc basis. Indeed, for general HMMs, the likelihood ratio statistic is stochastically unbounded even for bounded parameters, see Kéribin and Gassiat (2000).

We will start the technical exposition by describing the relationship between order estimation and hypothesis testing.

2.3 Order estimation and composite hypothesis testing

If we have a consistent order estimation procedure, we should be able to manufacture a sequence of consistent tests for the following questions: is the true order larger than $1, \dots, r, \dots$? We may indeed phrase the following composite hypothesis testing problem:

H0: The source belongs to \mathcal{M}^{r_0} ;

H1: The source belongs to $(\cup_r \mathcal{M}^r) \setminus \mathcal{M}^{r_0}$.

To put things in perspective, in this paragraph, we will focus on testing whether some probability distribution P belongs to some subset \mathcal{M}^0 (H0) of some set \mathcal{M} of distributions over \mathcal{Y}^∞ . Hypothesis H1 corresponds to $P \in \mathcal{M}^1 = \mathcal{M} \setminus \mathcal{M}^0$.

A test on samples of length t is a function T_t that maps \mathcal{Y}^t on $\{0, 1\}$. If $T_t(y_{1:t}) = 1$, the test rejects H0 in favor of H1, otherwise the test accepts H0. The region K_t on which the test rejects hypothesis H0 is called the critical region. The power function of the test π_t , maps distributions P or rather marginal distributions P_t toward the probability of the critical region:

$$\pi_t(P) \stackrel{\text{def}}{=} P_t \left\{ Y_{1:t} \in K_t \right\}.$$

If $\pi_t(P) \leq \alpha$ for all $P \in \mathcal{M}^0$, the test T_t is said to be of *level* α . The goal of test design is to achieve high power at prescribed level. In many settings of interest, the determination of the highest achievable power at a given level for a given sample size t is beyond our possibilities. This motivates asymptotic analysis. A sequence of tests T_t is asymptotically of level α if for all $P \in \mathcal{M}^0$, then

$$\lim_t P_t \{K_t\} \leq \alpha.$$

A sequence of tests T_t with power functions π_t is consistent at level α , if all but finitely many T_t have level α , and if for all $P \in \mathcal{M}^1 \setminus \mathcal{M}^0$, $\pi_t(P) \rightarrow 1$.

When comparing two simple hypothesis, the question is solved by the Neyman-Pearson's Lemma. The latter asserts that it is enough to compare the likelihoods of observations according to the two hypotheses with a threshold. When dealing with composite hypotheses, things turn out to be more difficult. In the context of nested models, the generalized likelihood ratio test is defined in the following way:

Generalized likelihood ratio test

Let \mathcal{M}^0 and \mathcal{M} denote two sets of distributions on \mathcal{Y}^∞ , with $\mathcal{M}^0 \subseteq \mathcal{M}$, the t -th likelihood ratio test between \mathcal{M}^0 and \mathcal{M} has critical region:

$$K_t \stackrel{\text{def}}{=} \left\{ y_{1:t} : \sup_{P \in \mathcal{M}^0} \log P\{y_{1:t}\} \geq \sup_{P \in \mathcal{M}} \log P\{y_{1:t}\} - \text{pen}(t) \right\},$$

where the penalty $\text{pen}(t)$ defines a t -dependent threshold.

Increasing the penalty $\text{pen}(t)$, enlarges the critical region, and tends to diminish the level of the test. As a matter of fact, in order to get a non-trivial level, $\text{pen}(t)$ should be > 0 . The definition of the generalized likelihood ratio test raises two questions:

1. How should $\text{pen}(t)$ be chosen to warrant strong consistency?
2. Is generalized likelihood ratio testing the best way to design a consistent test?

It turns out that the answer to those two questions is highly dependent on the properties of maximum likelihood in models \mathcal{M}^0 and \mathcal{M} . Moreover the way to get the answers depends on the models under consideration. In order to answer the first question, we need to understand the behavior of

$$\sup_{P \in \mathcal{M}} \log P\{y_{1:t}\} - \sup_{P \in \mathcal{M}^0} \log P\{y_{1:t}\},$$

under the two alternate hypotheses.

A process $(Y_t)_t$ with distribution P is said to be Markov of order r if for every $y_{1:t+1} \in \mathcal{Y}^{t+1}$,

$$P^* \{y_{t+1} \mid y_{1:t}\} = P^* \{y_{t+1} \mid y_{t-r+1:t}\}.$$

Let \mathcal{M}^0 denote Markov chains of order r and \mathcal{M} denote Markov chains of order $r+1$. If P^* defines a Markov chain of order r , then as t tends toward infinity, $\sup_{P \in \mathcal{M}} \log P\{y_{1:t}\} - \sup_{P \in \mathcal{M}^0} \log P\{y_{1:t}\}$ converges in distribution toward a χ_2 random variable with $|\mathcal{S}_{r+1}| - |\mathcal{S}_r|$ degrees of freedom where \mathcal{S}_r denotes the subset of patterns from $|\mathcal{Y}|^r$ that have non-null stationary probability. As a consequence of the Law of the Iterated Logarithm, P^* -almost surely, it should remain of order $\log \log t$ as t tends toward infinity (see Finesso (1991) and Section 2.7). Hence in such a case, a good understanding of the behavior of maximum likelihood estimates provides hints for designing consistent testing procedures. As already pointed out, such a knowledge is not available for HMMs. As of this writing, the best and most useful insights into the behavior of $\sup_{P \in \mathcal{M}} \log P\{y_{1:t}\} - \sup_{P \in \mathcal{M}^0} \log P\{y_{1:t}\}$ when \mathcal{M} denotes HMMs of order r and \mathcal{M}^0 denotes HMMs of order $r' < r$, can be found in the universal coding literature.

2.4 Code-based identification

2.4.1 Definitions

The pervasive influence of concepts originating from universal coding theory in the literature dedicated to Markov order or HMM order estimation should not be a surprise. Recall that by the Kraft-McMillan inequality Cover and Thomas (1991), a uniquely-decodable code on \mathcal{Y}^t defines a (sub)-probability on \mathcal{Y}^t , and conversely, for any probability distribution P^t on \mathcal{Y}^t , there exists a uniquely-decodable code for \mathcal{Y}^t such that the length of the codeword associated with $y_{1:t}$ is upper-bounded by $\lceil \log P\{y_{1:t}\} \rceil + 1$. Henceforth, the probability associated with a code will be called the *coding probability*, and the logarithm of the coding probability will represent the *ideal codeword length* associated with the coding probability.

For each t , let Q^t denote a coding probability for \mathcal{Y}^t . The redundancy of Q_t with respect to $P \in \mathcal{M}$ is defined as

$$D(P_t \mid Q_t).$$

The (possibly non-consistent) family of coding probabilities $(Q^t)_t$ is a universal coding probability for model \mathcal{M} if and only if,

$$\sup_{P \in \mathcal{M}} \liminf_t \frac{1}{t} D(P_t \mid Q_t) = 0.$$

The quantity $\sup_{P \in \mathcal{M}} D(P_t \mid Q_t)$ is called the redundancy rate of the sequence (Q_t) with respect to \mathcal{M} . Let r denote a function from \mathbb{N} toward \mathbb{R}^+ . We are typically interested in sub-linear functions. The sequence $(Q_t)_t$ is said to achieve redundancy rate $r(\cdot)$ over \mathcal{M} , if

$$\limsup_t \sup_{P \in \mathcal{M}} \frac{1}{r(t)} D(P_t \mid Q_t) < \infty.$$

The assessment of the existence of universal coding probabilities for large classes of models, for example for the whole class of stationary ergodic sources is one of the fundamental results of Information Theory. Strictly sub-linear redundancy rates are said to be non-trivial. We will see in the sequel that non trivial redundancy rates play an important role in order estimation.

The following coding probability has played a distinguished role in the areas of universal coding and prediction of individual sequences.

Normalized Maximal Likelihood (NML)

Given a model \mathcal{M} of probability distributions over \mathcal{Y}^t , the Normalized Maximum Likelihood coding probability induced by \mathcal{M} on \mathcal{Y}^t is defined by:

$$\text{NML}\{y_{1:t}\} = \frac{\sup_{P \in \mathcal{M}} P\{y_{1:t}\}}{\mathcal{C}_t},$$

where

$$\mathcal{C}_t \stackrel{\text{def}}{=} \sum_{y_{1:t} \in \mathcal{Y}^t} \sup_{P \in \mathcal{M}} P\{y_{1:t}\}.$$

The maximum point-wise regret of a coding probability Q_t with respect to model \mathcal{M} is defined as

$$\max_{y_{1:t} \in \mathcal{Y}^t} \sup_{P \in \mathcal{M}} \log \frac{P\{y_{1:t}\}}{Q_t\{y_{1:t}\}}.$$

Note that NML_t achieves the same regret $\log \mathcal{C}_t$ over all strings from \mathcal{Y}^t . No coding probability can achieve a smaller maximum point-wise regret. This is why NML coders are said to achieve minimax point-wise regret over \mathcal{M} .

Significant progresses in universal coding theory during the last two decades can be related to the determination of precise bounds on \mathcal{C}_t for different kinds of models, notably for the class of product distributions (memoryless sources), for the class of Markov chains of order r (Markov sources), and for the class of Hidden Markov sources of order r .

The relevance of bounds on \mathcal{C}_t to our problem is immediate. Let \mathcal{C}_t be defined with respect to \mathcal{M} , let P denote the true distribution which is assumed to belong to \mathcal{M} , then

$$\sup_{P' \in \mathcal{M}} \log P'\{y_{1:t}\} - \log P\{y_{1:t}\} = \log \text{NML}\{y_{1:t}\} - \log P\{y_{1:t}\} + \log \mathcal{C}_t.$$

On the right-hand-side of this inequality, the two quantities that show up refer to two fixed probabilities. After exponentiation, those two quantities may take part into summations over $y_{1:t}$.

The NML coding probability is one among many universal coding probabilities that have been investigated in the literature. For models like HMMs with fixed order r , the parameter space Θ^r can be endowed with a probability space structure. A prior probability π can be defined on that Θ^r , and under mild measurability assumptions, this in turn defines a probability distribution over \mathcal{Y}^∞ . Such coding probabilities are called mixture coders. Historically, several prior probabilities over Θ have been considered. Uniform (or Laplace) priors have been considered first but Dirichlet distributions soon gained much attention.

Dirichlet distribution

A Dirichlet $\mathcal{D}_r(\alpha_1, \dots, \alpha_r)$ distribution is a distribution on the simplex of \mathbb{R}^r , given by the density

$$\pi(q_1, \dots, q_r | \alpha_1, \dots, \alpha_r) = \frac{\Gamma(\alpha_1 + \dots + \alpha_r)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_r)} q_1^{\alpha_1-1} \dots q_r^{\alpha_r-1} \mathbf{1}_{q_1+\dots+q_r=1},$$

where the α_i 's are all positive.

The induced coding probability is just a mixture of probabilities from \mathcal{M} , see also Section 3.2.2 in Chapter ???. Though the Dirichlet prior has a venerable history in Bayesian inference, in this Chapter we will stick to the Information-theoretical tradition and call the resulting coding probability the Krichevsky-Trofimov mixture.

Krichevsky-Trofimov mixture

The Krichevsky-Trofimov mixture is defined by providing Θ^r with a product of Dirichlet- $(1/2, \dots, 1/2)$ distributions. Each element in Θ^r is defined by a prior probability $\mu_\theta(\cdot)$ that is by an element of the simplex of \mathbb{R}^r , r emission probabilities $G_\theta(\cdot, \cdot)$, that is by r elements of the simplex of \mathbb{R}^d , where $d \stackrel{\text{def}}{=} |\mathcal{Y}|$ and r transition probabilities $Q_\theta(\cdot, \cdot)$, that is by r elements of the simplex of \mathbb{R}^r .

$$\begin{aligned} \pi(d\theta) \stackrel{\text{def}}{=} & \left[\frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{1}{2})^r} \prod_{i=1}^r \mu_\theta(i)^{-1/2} \right] \\ & \times \prod_{i=1}^r \left[\frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{1}{2})^r} \prod_{j=1}^r Q_\theta(i, j)^{-1/2} \right] \times \left[\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})^d} \prod_{j=1}^d G_\theta(i, j)^{-1/2} \right] \end{aligned} \quad (2.1)$$

Krichevsky-Trofimov mixtures define a consistent family of probability distributions over \mathcal{Y}^t for $t \in \mathbb{N}$. This is in sharp contrast with NML distributions, and is part of the reasons why Krichevsky-Trofimov (KT) mixtures became so popular in source coding theory.

The relationship between NML and KT coders for memoryless sources over a fixed alphabet are now quite well understood. Even though KT coders fail to achieve asymptotically minimax point-wise regret by a constant, they achieve asymptotically maxi-min redundancy, that is:

$$\lim_t \left[\max_\omega \min_Q E_\omega \left[D(P_\theta^t | Q^t) \right] \right] - E_\pi \left[D(P_\theta^t | \text{KT}^t) \right] = 0. \quad (2.2)$$

The gap between minimax point-wise regret and maxi-min redundancy has been quantified for memoryless sources and for Markov sources of any order

Resorting to coding-theoretical concepts provides a framework for defining an order estimation procedure known as Minimum Description Length (MDL) order estimation. MDL has been introduced and popularized by J. Rissanen in the late 1970s. Although MDL has often been promoted by borrowing material from medieval philosophy, we will see later that it can be justified using some non-trivial mathematics for Markov order estimation.

CHECK this, whether it follows from Jacquet and Szpankowski.

MDL order estimator

Assume μ is a probability distribution on the set of possible orders, and that for each order r , for each $t \in \mathbb{N}$, Q_r^t defines a coding probability for \mathcal{Y}^t with respect to \mathcal{M}^r , then the Minimum Description Length order estimator is defined by:

$$\hat{r} \stackrel{\text{def}}{=} \arg \max_r \left(\log Q_r^t \{y_{1:t}\} + \log \mu \{r\} \right).$$

Note that if the coding probability Q_r^t turns out to be the Normalized Maximum Likelihood distribution, the MDL order estimator is a special kind of penalized maximum likelihood order estimator.

Penalized Maximum Likelihood order estimator

Let $(\text{pen}(t, r))_{t,r}$ denote a family of non-negative quantities, a penalized maximum likelihood order estimator is defined by:

$$\hat{r} \stackrel{\text{def}}{=} \arg \max_r \left(\sup_{P \in \mathcal{M}^r} \log P \{y_{1:t}\} - \text{pen}(t, r) \right).$$

The Bayesian Information Criterion (BIC) order estimator is nothing but another distinguished member of the family of penalized maximum likelihood order estimators. It is closely related but different from the MDL order estimator derived from the NML coding probability.

BIC order estimator

Let $\dim(r)$ be the dimension of the parameter space Θ^r in \mathcal{M}^r :

$$\hat{r} \stackrel{\text{def}}{=} \arg \max_r \left(\sup_{P \in \mathcal{M}^r} \log P\{y_{1:t}\} - \frac{\dim(r)}{2} \log t \right) .$$

Schwarz introduced the BIC in the late 1970s using Bayesian reasoning, and using Laplace's trick to simplify high-dimensional integrals. The validity of this trick and the relevance of Bayesian reasoning to minimax framework has to be checked on an ad hoc basis.

2.4.2 Information divergence rates

The order estimators we have in mind (MDL, BIC, PML) are related to generalized likelihood ratio testing. In order to prove their consistency, we need strong laws of large numbers concerning logarithms of likelihood ratios. In the stationary independent case, those laws of large numbers reduce to the classical laws of large numbers for sums of independent random variables. Such strong laws have proved to be fundamental tools both in Statistics and Information Theory. In general (that is not necessarily i.i.d. settings), the laws of large numbers we are looking for have been called Asymptotic Equipartition Principles for Information in Information Theory or Shannon-Breiman-McMillan Theorems in Ergodic Theory Barron (1985b).

Before formulating SBM Theorems in a convenient form, let us recall some basic facts about likelihood ratios. Let P and P' denote two probabilities over \mathcal{Y}^∞ , such that for every t , P'_t is absolutely continuous with respect to P_t , then under P , the ratio P'_t/P_t is an (\mathcal{F}_t) -adapted martingale with expectation less or equal than 1. By monotonicity and concavity of the logarithm, $\log P'_t/P_t$ is a super-martingale with non-positive expectation. By Doob's Theorem, this super-martingale converges almost surely toward an integrable random variable. If the expectation of the latter random variable is infinite, P is singular with respect to P' . In such a setting, the rate of growth of $\log P'_t/P_t$ is a matter of concern. If the two distributions are product probabilities, the log-likelihood ratio is a sum of independent random variables and it grows linearly with t , if the factors are identical. Moreover, the strong law of large numbers tells us that $1/t \log P'_t/P_t$ converges almost surely towards a fixed value which is called the information divergence rate between the two distributions.

How robust is this observation? This is precisely the topic of SBM Theorems.

General AEP

A set \mathcal{M} of process laws over \mathcal{Y} satisfies a generalized Shannon-Breiman-McMillan if and only if

1. For every pair of laws P and P' from \mathcal{M} , the relative entropy rate between P and P' :

$$\lim_{t \rightarrow \infty} \frac{1}{t} D(P_t | P'_t)$$

exists. It is denoted by $D_\infty(P | P')$

2. Furthermore, if P and P' are stationary ergodic, then P -almost everywhere:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \ln \frac{P\{Y_{1:t}\}}{P'\{Y_{1:t}\}} = D_\infty(P | P') .$$

The cases of Markov models and Hidden Markov models can be dealt with using Barron's generalized Shannon-McMillan-Breiman Theorem:

Theorem 2.4.1. *Let \mathcal{Y} denote a standard Borel space, let Y_1, Y_2, \dots, Y_t be a \mathcal{Y} -valued stochastic ergodic process distributed according to P . Let P' denote a distribution over \mathcal{Y}^∞ which is assumed to be Markovian of order r , and such that for each t , P_t has a density f with respect to P'_t . Then P -almost surely,*

$$\frac{1}{t} \log \frac{dP}{dP'}(Y_1, \dots, Y_t)$$

converges toward the relative entropy rate between the two distributions:

$$D_\infty(P | P') = \lim_t \frac{1}{t} D(P_t | P'_t) = \sup_t \frac{1}{t} D(P_t | P'_t) .$$

From Barron's Theorem, it is immediate that the collection of Markov models satisfies the generalized SBM property. The status of HMMs is less straightforward. There are actually several proofs that HMM's satisfies the generalized AEP (see Finesso (1991)). The argument we present here simply resorts to the extended chain device.

Theorem 2.4.2. *The collections of HMMs over some finite observation alphabet \mathcal{Y} satisfies the generalized AEP.*

Proof. Let P and P' denote two HMMs over some finite observation alphabet \mathcal{Y} . Let Φ_t and Φ'_t denote the associated prediction filters. Then under P and P' , the sequence $(Y_t, \Phi_t, \Phi'_t)_t$ is a Markov chain over $\mathcal{Y} \times \mathbb{R}^r \times \mathbb{R}^{r'}$, which may be regarded as a standard Borel set. Moreover

$$\log P\{y_{1:t}\} = \log P\{y_{1:t}, \Phi_{1:t}, \Phi'_{1:t}\} .$$

Applying Theorem 2.4.1 to the sequence $(Y_t, \Phi_t, \Phi'_t)_t$ terminates the proof of the Theorem. \square

Lemma 2.4.3. *If P is a stationary HMM with pairwise distinct stationary ergodic components P_i with $i \in \{1, \dots, d\}$ having disjoint supports on $\mathcal{X} \times \mathcal{Y}$:*

$$P = \sum_{i=1}^d \lambda_i P_i,$$

where $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$, and if P' is a stationary ergodic HMM then

$$D_\infty(P | P') = \sum_i \lambda_i D_\infty(P_i | P') .$$

For a proof of Lemma 2.4.3, we refer to Gassiat and Boucheron (to appear).

Knowing that some collection of models satisfies the generalized AEP allows to test between two elements picked from the collection. When performing order estimation, we need more than that. If ML estimation is consistent, we need to have for every $P^* \in \mathcal{M}^{r^*} \setminus \mathcal{M}^{r^*-1}$

$$P^* - \text{a.s.} \quad \limsup_t \sup_{P' \in \mathcal{M}^{r^*-1}} \frac{1}{t} \log \frac{P'_t}{P_t} < 0 .$$

If the collection of models satisfies the generalized AEP, this should at least imply:

$$\inf_{P' \in \mathcal{M}^{r^*-1}} D_\infty(P | P') > 0 .$$

That this is actually the case for our primary concerns is summarized by the following Lemma.

Lemma 2.4.4. *Let P denote the distribution of a stationary process over \mathcal{Y} , and let \mathcal{M} denote a set of distributions of processes over \mathcal{Y} , that may be either the set of Markov chains of order r or the set of HMMs of order r . Then*

$$\lim_t \frac{1}{t} \inf_{P' \in \mathcal{M}} D(P_t | P'_t)$$

has a limit which equals

$$\inf_{P' \in \mathcal{M}} D_\infty(P | P').$$

The limit may be null if and only if P belongs to \mathcal{M} .

Proof. The proof follows the same pattern for HMMs and Markov chains. Let us consider HMMs. The proof follows from the following observations:

1. From Barron's Theorem, $D_\infty(P | P') = \sup_t \frac{1}{t} D(P_t | P'_t)$.
2. For any fixed t , if the set of probability distributions on \mathcal{Y}^t is endowed with the topology of weak convergence, the function $P' \mapsto \frac{1}{t} D(P_t | P'_t)$, is lower-semi-continuous with respect to P' , Dupuis and Ellis (1997)
3. The mapping from Θ^r toward probabilities on \mathcal{Y}^t , is continuous. Hence the mapping $\theta \mapsto \frac{1}{t} D(P_t | P_{\theta,t})$ is lower-semi-continuous. As the parameter space Θ^r is compact, this mapping achieves its infimum at some point $\hat{\theta} \in \Theta^r$.
4. If two HMMs P and P' of order not greater than r induce the same distributions on \mathcal{Y}^{2r} , then they are equal Finesso (1991).

Combining the first and the second observation, the function on Θ^r defined by

$$\theta \mapsto D_\infty(P | P_\theta)$$

is lower-semi-continuous on a compact set and thus achieves its infimum on some $\hat{\theta} \in \Theta^r$. Let \hat{P} denote the corresponding HMM. This infimum is larger than $\frac{1}{t} D(P_t | \hat{P}_t)$ for all t . If it is null, P equals \hat{P} , by the fourth observation. \square

The previous lemma will prove to be valuable tool while checking consistency of ML order estimators. The next lemma will prove useful when assessing the efficiency of ML order estimators for HMM order estimations.

Lemma 2.4.5. *Let P denote the distribution of a stationary HMM over \mathcal{Y} , and let \mathcal{M} denote a set of distributions of HMM of order r over \mathcal{Y} . Then*

$$\lim_t \frac{1}{t} \inf_{P' \in \mathcal{M}} D(P_t | P'_t)$$

has a limit which equals

$$\inf_{P' \in \mathcal{M}} D_\infty(P' | P).$$

The limit may be null if and only if P belongs \mathcal{M} .

Proof. The proof parallels the Proof of the previous Lemma and uses the lower-semi-continuity of relative entropy with respect to its first argument. \square

Finally, when dealing with efficiency issues in HMM order estimation, we will use the following property of information divergence rates.

Lower semi-continuity property of information divergence rates
 If $(P^n)_n$ and $(P^{prime,n})_n$ denote two sequences of ergodic elements of $\cup_r \mathcal{M}^r$, converging toward P and P' , then

$$D_\infty(P | P') \leq \liminf D_\infty(P^n | P'^n) .$$

Information divergence rates between Markov chains and HMMs satisfy the lower semi-continuity property. Indeed, if P and P' denote two ergodic Markov chains (or ergodic HMMs), from Barron's Theorem, the sequence $\frac{1}{t}D(P_t | P'_t)$ is non-decreasing with respect to t . Each function $\frac{1}{t}D(P_t | P'_t)$ is lower semi-continuous with respect to P and P' , and with respect to parameterizations and recalling that the supremum of a family of lower semi-continuous functions is lower semi-continuous, we get:

Lemma 2.4.6. *If $(P^n)_n$ and $(P'^n)_n$ denote two sequences of stationary ergodic HMMs converging respectively (in parameter space) toward P and P' , then*

$$D_\infty(P | P') \leq \liminf D_\infty(P^n | P'^n) .$$

2.5 MDL order estimators in Bayesian settings

Under mild but non-trivial conditions on universal redundancy rates, the above-described order estimators are strongly consistent in a minimax setting. In this Section, we will present a result that might seem to be a definitive result.

Recall that two probability distributions Q and Q' are orthogonal or mutually singular if there exists a set A such that $Q\{A\} = 1 = Q'\{A^c\}$.

Theorem 2.5.1. *Let $(\Theta^r)_{r \in \mathbb{N}}$ denote a collection of models, and let $(Q_r)_{r \in \mathbb{N}}$ denote coding probabilities associated with prior probabilities ω_r . Let $L(r)$ denote the length of a prefix binary encoding of the integer r . Assume that the mixture coding probabilities Q_r are mutually singular on the asymptotic σ -algebra. If the order estimator is defined as:*

$$\hat{r}_t \stackrel{\text{def}}{=} \arg \min_r \left[-\log_2 Q_r\{y_{1:t}\} + L(r) \right],$$

then for all r^* , ω_{r^*} -almost surely, θ -almost surely, \hat{r}_t converges toward r^* .

Proof. Let us define Q^* as a double mixture:

$$Q^* \stackrel{\text{def}}{=} C \sum_{r \neq r^*} 2^{-L(r)} Q_r,$$

where C is a normalization factor ($C \geq 1$). Under the assumptions of the Theorem, Q^* and Q_{r^*} are mutually singular on the asymptotic σ -algebra. Moreover for all $y_{1:t}$:

$$Q^*\{y_{1:t}\} \geq C \sup_{r \neq r^*} \left[2^{-L(r)} Q_r\{y_{1:t}\} \right],$$

which is equivalent to

$$-\log_2 Q^*\{y_{1:t}\} \leq -\log_2 C + \inf_{r \neq r^*} \left\{ L(r) + -\log_2 Q_r\{y_{1:t}\} \right\}.$$

On the other hand, a standard martingale argument tells us that Q_{r^*} -almost surely

$$\log_2 \frac{Q_{r^*}\{y_{1:t}\}}{Q^*\{y_{1:t}\}}$$

converges towards a limit, and the fact that \mathbf{Q}_{r^*} and \mathbf{Q}^* are mutually singular entails that this limit is \mathbf{Q}_{r^*} -almost surely infinite. Hence \mathbf{Q}_{r^*} -almost surely, for all sufficiently large t

$$-\log_2 \mathbf{Q}_{r^*}\{y_{1:t}\} + L(r^*) < \inf_{r \neq r^*} \left\{ L(r) - \log_2 \mathbf{Q}_r\{y_{1:t}\} \right\}.$$

This entails that \mathbf{Q}_{r^*} -almost surely, for all sufficiently large t , $\hat{r}_t = r^*$, which is the desired result. \square

Remark 2.1. Theorem 2.5.1 should not be misinterpreted. It does not prevent the fact that for some θ s in a set with null ω_{r^*} probability, the order estimator might be inconsistent. Neither does the Theorem give a way to identify those θ s for which the order estimator is consistent.

2.6 Strongly consistent penalized maximum likelihood estimators for HMM order estimation

In this section, we give general results concerning order estimation in the framework of nested sequences of models; we then state their application to HMMs.

Let $(\mathcal{M}^r)_{r \in \mathbb{N}}$ define a nested sequence of models for sequences $(Y_t)_{t \in \mathbb{N}}$ on the finite alphabet \mathcal{Y} .

Let \mathbf{P}^* denote the true distribution of $(Y_t)_{t \in \mathbb{N}}$ and the order r^* is as usual the smallest integer r such that \mathbf{P}^* belongs to \mathcal{M}^r . We shall consider penalized ML estimators \hat{r}_t .

Assumption 2.1.

1. The sequence of models satisfies the generalized AEP (see Section 2.4.2).
2. Whenever \mathbf{P}^* is stationary ergodic of order r^* and $r < r^*$,

$$\inf_{P \in \mathcal{M}^r} D_\infty(P^* | P) > 0;$$

3. For any $\epsilon > 0$, any r , there exists a sieve $(P_i)_{i \in I_\epsilon^r}$, that is a finite set I_ϵ^r such that $P_i \in \mathcal{M}^r$, and all P_i are stationary ergodic, and a t_ϵ^r , such that:

$$\forall P \in \mathcal{M}^r, \exists i \in I_\epsilon^r, \forall t \geq t_\epsilon^r, \forall y_{1:t}, \left| \frac{1}{t} \log P\{y_{1:t}\} - \log P_i\{y_{1:t}\} \right| \leq \epsilon.$$

Non-trivial upper-bounds on point-wise minimax regret for the different models at hand will enable to build strongly consistent code-based order estimators.

Lemma 2.6.1. *Let the penalty function $\text{pen}(t, r)$ be non-decreasing with respect to r and such that $\text{pen}(t, r)/t \rightarrow 0$. Let (\hat{r}_t) denote the sequence of penalized maximum likelihood order estimators defined by $\text{pen}(\cdot)$. Then under Assumption 2.1, \mathbf{P}^* -almost surely, $\hat{r}_t \geq r^*$ eventually.*

Proof of Lemma 2.6.1. Throughout “infinitely often” will be abbreviated to *i.o.*

$$\left(\hat{r}_t > r^* \text{ i.o.} \right) = \bigcup_{r < r^*} \left(\hat{r}_t = r \text{ i.o.} \right).$$

$$\begin{aligned} \left(\hat{r}_t = r \text{ i.o.} \right) &\subseteq \left(\sup_{P \in \mathcal{M}^r} \log P\{y_{1:t}\} \geq \log \mathbf{P}^*\{y_{1:t}\} - \text{pen}(t, r^*) \text{ i.o.} \right) \\ &\subseteq \left(\max_{i \in I_\epsilon^r} \log P_i\{y_{1:t}\} \geq \log \mathbf{P}^*\{y_{1:t}\} - t\epsilon - \text{pen}(t, r^*) \text{ i.o.} \right) \\ &\subseteq \bigcup_{i \in I_\epsilon^r} \left(\limsup \frac{1}{t} (\log P_i\{y_{1:t}\} - \log \mathbf{P}^*\{y_{1:t}\}) \geq -\epsilon \right), \end{aligned}$$

where $(\mathbf{P}_i)_{i \in I_t^r}$ is the sieve given by Assumption 2.1.3 for \mathcal{M}^r . Now, by Assumption 2.1.1, $\frac{1}{t}(\log \mathbf{P}_i\{y_{1:t}\} - \log \mathbf{P}^*\{y_{1:t}\})$ converges toward $-D_\infty(\mathbf{P}^* | \mathbf{P}_i)$ \mathbf{P}^* -almost surely and by Assumption 2.1.2, as soon as

$$\epsilon < \min_{r < r^*} \inf_{\mathbf{P} \in \mathcal{M}^r} D_\infty(\mathbf{P}^* | \mathbf{P}),$$

one gets $\mathbf{P}^*\{\widehat{r} = r \text{ i.o.}\} = 0$. \square

A possibly very conservative way of choosing penalties may be justified in a straightforward way by universal coding argument. Let \mathcal{C}_t^r denote the normalization constant in the definition of the NML coding probability induced by \mathcal{M}^r on \mathcal{Y}^t .

Lemma 2.6.2. *Let the penalty function be defined by $\text{pen}(t, r) = \sum_{r'=0}^r (\ln \mathcal{C}_t^{r'} + 2 \ln t)$. Let (\widehat{r}_t) denote the sequence of penalized maximum likelihood order estimators defined by $\text{pen}(\cdot)$.*

Then \mathbf{P}^ -almost surely,*

$$\widehat{r}_t \leq r^* \quad \text{eventually.}$$

Proof of Lemma 2.6.2. Let r denote an integer that is larger than r^* .

$$\begin{aligned} \mathbf{P}^*\{\widehat{r}_t = r\} &\leq \mathbf{P}^*\left\{\log \mathbf{P}^*\{Y_{1:t}\} \leq \sup_{\mathbf{P} \in \mathcal{M}^r} \log \mathbf{P}\{Y_{1:t}\} - \text{pen}(t, r) + \text{pen}(t, r^*)\right\} \\ &\leq \mathbf{P}^*\left\{\log \mathbf{P}^*\{Y_{1:t}\} \leq \ln \text{NML}_t^r\{Y_{1:t}\} - \sum_{r'=r^*+1}^{r-1} \ln \mathcal{C}_t^{r'} - 2(r - r^*) \ln t\right\} \\ &\leq \sum_{y_{1:t}} \exp(\log \mathbf{P}^*\{y_{1:t}\}) \mathbf{1}_{\{\log \mathbf{P}^*\{y_{1:t}\} \leq \ln \text{NML}_t^r\{y_{1:t}\} - \sum_{r'=r^*+1}^{r-1} \ln \mathcal{C}_t^{r'} - 2(r - r^*) \ln t\}} \\ &\leq \sum_{y_{1:t}} \text{NML}_t^r\{y_{1:t}\} \exp\left(-\sum_{r'=r^*+1}^{r-1} \ln \mathcal{C}_t^{r'} - 2(r - r^*) \ln t\right) \\ &\leq \exp\left(-\sum_{r'=r^*+1}^{r-1} \ln \mathcal{C}_t^{r'} - 2(r - r^*) \ln t\right) \\ &\leq t^{-2(r - r^*)}, \end{aligned}$$

since $\sum_{r'=r^*+1}^{r-1} \ln \mathcal{C}_t^{r'} = 0$ for $r = r^* + 1$.

By the union bound:

$$\begin{aligned} \mathbf{P}^*\{\widehat{r}_t > r^*\} &\leq \sum_{r > r^*} \mathbf{P}^*\{\widehat{r}_t = r\} \\ &\leq \frac{t^{-2}}{1 - t^{-2}}. \end{aligned}$$

$$\begin{aligned} \sum_t \mathbf{P}^*\{\widehat{r}_t > r^*\} &\leq \sum_t 1 \wedge \frac{t^{-2}}{1 - t^{-2}} \\ &< \infty. \end{aligned}$$

Applying the first Borel-Cantelli Lemma, we may now conclude that \mathbf{P}^* -almost surely, order over-estimation occurs only finitely many times. \square

In order to show the existence of strongly consistent order estimators for HMMs, it remains to check that Assumption 2.1 holds and that the penalties used in the statement of Lemma 2.6.2 satisfy the conditions stated in Lemma 2.6.1, that is, for all $r \in \mathbb{N}$:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{r' \leq r} (\ln \mathcal{C}_t^{r'} + 2 \ln t) = 0.$$

This last point follows immediately from the following result from universal coding theory:

Lemma 2.6.3. *For all r , for all t larger than r , for $y_{1:t}$:*

$$\ln C_t^r = \ln \frac{\sup_{\theta \in \mathcal{M}^r} P_\theta \{y_{1:t}\}}{NML_t^r \{y_{1:t}\}} \leq \frac{r(r+d-2)}{2} \ln t + c_{r,d}(t),$$

where for $t \geq 4$, $c_{r,d}(t)$ may be chosen as:

$$c_{r,d}(t) = \ln r + r \left(-\ln \frac{\Gamma(\frac{r}{2}) \Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2}) \Gamma(\frac{1}{2})} + \frac{r^2 + d^2}{4t} + \frac{1}{6t} \right).$$

Concerning Assumption 2.1, points 1. and 2. hold (see Section 2.4.2). Now, for any positive δ , let us denote by Θ_δ^r , the set of HMM parameters in Θ^r such that each coordinate is lower bounded by δ .

Now, for any $\theta \in \Theta^r$, there exists $\theta_\delta \in \Theta_\delta^r$ such that for any t and any $y_{1:t}$

$$\frac{1}{t} \left| \log P_\theta \{y_{1:t}\} - \log P_{\theta_\delta} \{y_{1:t}\} \right| \leq \frac{r^2 + d^2}{2} \delta.$$

A glimpse at the proof of this fact in Liu and Narayan (1994) reveals that this statement still holds when θ_δ is constrained to lie in a sieve for Θ_δ^r , defined as a finite subset $(\theta_i)_{i \in I}$ such that $\theta_i \in \Theta_\delta^r$, and $\forall \theta \in \Theta^r$, at least one θ_i in the sieve is within L_∞ -distance smaller than δ from θ .

This may be summarized in the following way:

Corollary 2.6.4. *let P^* be a HMM with order r^* , and (\hat{r}_t) be the sequence of penalized ML order estimators defined in Lemma 2.6.2. Then P^* -almost surely, $\hat{r}_t = r^*$ eventually.*

Remark 2.2. Resorting to universal coding arguments to cope with our poor understanding of the Maximum Likelihood in misspecified HMM models provides us with a Janus-faced result: on the one hand, it allows to describe a family of strongly consistent order estimators that will prove to be optimal as far as under-estimation is concerned; on the other hand, the question raised by J. Kieffer in Kieffer (1993) about the consistency of BIC and MDL for HMM order estimation remains open.

Before describing recent advances on the consistency of BIC estimators for the Markov order estimation problem, we will first mention that conditions on universal redundancy rates are non trivial.

Viewing Lemmas 2.6.2 and 2.6.1, a natural question is: how large can the models \mathcal{M}^r be? Would it be for example possible to test stationary renewal processes against stationary ergodic sources? The last question turns out to be equivalent to: do the class of renewal sources and the class of all stationary ergodic sources over some finite alphabet enjoy a non-trivial point-wise minimax regret? The latter does not as stated in the following theorem due to Shields (1993).

Theorem 2.6.5. *Let P' denote a coding probability on \mathcal{Y} . Let $\rho(t)$ denote any non-negative function such that $\lim_{t \rightarrow \infty} \rho(t)/t = 0$. Then there exists a stationary ergodic process with distribution P on \mathcal{Y} such that:*

$$\lim_t \frac{1}{\rho(t)} E_P \left[\log \frac{P\{Y_{1:t}\}}{P'\{Y_{1:t}\}} \right] = \infty.$$

As the minimax point-wise regret is always at least as large as the minimax average redundancy rate, Theorem 2.6.5 proves that there is no non-trivial upper-bound on the point-wise minimax regret for stationary ergodic sources.

Note however, that a large class of renewal processes has been shown to have a non-trivial redundancy rate. A discrete renewal process over $\mathcal{Y} = \{0, 1\}$ is completely defined by a probability distribution over \mathbb{N} . Discrete renewal processes over the binary alphabet may thus be parameterized by non-negative unit vectors from $\ell_1(\mathbb{N})$. In contrast with what happened with Markov sources or HMMs, such a parameter set is intrinsically infinite-dimensional. The following has been proved in Csiszár and Shields (2000) and further refined in Flajolet and Szpankowski (2002).

Theorem 2.6.6. *Let \mathcal{M} denote the class of stationary renewal processes over $\mathcal{Y} = \{0, 1\}$, the redundancy rate over \mathcal{M} belongs to $\Theta(\sqrt{n})$.*

2.7 Consistency of the BIC order estimator in the Markov order estimation problem

Though the consistency of BIC estimator for HMM order is still far from being established, recent progress concerning the Markov order estimation problem raise great expectations. As a matter of fact, the following was established by Csizsár and Shields (2000) and recently refined by Csizsár (2002).

Let us recall that a process $((Y_t)_{t \in \mathbb{N}})_{t \in \mathbb{N}}$ over \mathcal{Y} with distribution \mathbf{P} is Markov of order less or equal than r if and only if for every $t > r$, every $a_{1:t}$ such that $\mathbf{P}\{a_{1:t}\} > 0$,

$$\mathbf{P}\{a_t \mid a_{1:t-1}\} = \mathbf{P}\{a_t \mid a_{t-r:t-1}\}.$$

$((Y_t)_{t \in \mathbb{N}})_t$ is of order exactly r if it is of order less or equal than r but not less or equal than $r - 1$. The set of Markov chains of order less or equal than r may be parameterized by

$$\mathcal{M}_1(\mathcal{Y})^{|\mathcal{Y}|^r},$$

that is, a process of Markov order less or equal than r , is defined by $(|\mathcal{Y}| - 1) \times |\mathcal{Y}|^r$ parameters.

Theorem 2.7.1. *For any stationary irreducible Markov process with distribution \mathbf{P}^* over the finite set \mathcal{Y} and order r^* , \mathbf{P}^* -almost surely, the BIC order estimator converges toward r^* .*

The proofs of this remarkable theorem follow from a series of technical Lemmas concerning the behavior of maximum likelihood estimators in models \mathcal{M}^r for $r \geq r^*$. In the Markov order estimation problem, such precise results can be obtained at a reasonable price thanks to the fact that maximum likelihood estimates coincide with empirical measures. Here we follow the argument presented in Csizsár (2002).

Note first that underestimation issues are dealt with using Lemma 2.6.1. Theorem 2.7.1 actually follows almost directly from the following theorem. Recall that $\widehat{\mathbf{P}}_t^r$ denotes the Maximum Likelihood estimate in \mathcal{M}^r on sample $y_{1:t}$.

Theorem 2.7.2. *For any stationary irreducible Markov process with distribution \mathbf{P}^* of order r^* over the finite set \mathcal{Y} , \mathbf{P}^* -almost surely*

$$\sup_{r \geq r^*} \left\{ \frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \right\} \rightarrow 0.$$

To emphasize the power of this Theorem, let us first derive Theorem 2.7.1 from Theorem 2.7.2.

Proof of Theorem 2.7.1. The event

$$\widehat{r}_t > r^* \quad \text{i.o.}$$

equals the event

$$\left(\exists r > r^* : \log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \widehat{\mathbf{P}}_t^{r^*} \{y_{1:t}\} \geq \text{pen}(t, r) - \text{pen}(t, r^*) \right) \quad \text{i.o.}$$

which is included in

$$\left(\exists r > r^* : \log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \geq \text{pen}(t, r) - \text{pen}(t, r^*) \right) \quad \text{i.o.}$$

But from Theorem 2.7.2, it follows that for any $\eta > 0$, \mathbf{P}^* -almost surely

$$\sup_{r \geq r^*} \left\{ \frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \right\} < \eta.$$

But for the BIC criterion, $\text{pen}(t, r) \geq \frac{|\mathcal{S}_r| \times (|\mathcal{Y}| - 1)}{2} \log t$. □

Remark 2.3. Viewing the proof of the strong consistency of the BIC Markov order estimator, one may wonder whether an analogous results holds for MDL order estimators derived from NML coding probabilities or KT coding probabilities. If no a priori restriction on the order is enforced, the answer is negative: there exists at least one stationary ergodic Markov chain (the uniform memoryless source) for which unrestricted MDL order estimators over-estimate the order infinitely often with probability one Csiszár and Shields (2000).

But if the search for r in

$$\max_r \left(-\log Q_t^r \{y_{1:t}\} - \log \mu \{r\} \right)$$

is restricted to some interval $(0, \dots, \alpha \log t)$ where α does not depend on t , then the MDL order estimator derived by taking NML_t^r as the r -th coding probability turns to be strongly consistent. The reason why this holds is that in order to prove strong consistency, we need to control

$$\log \mathcal{C}_t^r - \frac{|\mathcal{S}_{r+1}| - |\mathcal{S}_r|}{2} \log t$$

over a large range of values of r for all sufficiently large t . Sharp estimates of the minimax point-wise regret of NML for Markov sources of order r have recently been obtained Jacquet and Szpankowski (2002). It is not clear whether such precise estimates can be obtained for models like HMM where maximum likelihood is not as well-behaved as in the Markov chain setting.

Throughout this section, \mathbf{P}^* denotes the distribution of a stationary irreducible Markov chain of order r^* over \mathcal{Y} .

For all r , and all $a_{1:r} \in \mathcal{Y}^r$:

$$N_t(a_{1:r}) \stackrel{\text{def}}{=} \sum_{i=1}^{t+1-r} \mathbf{1}_{\bigwedge_{j=1}^r Y_{i+j-1} = a_j}.$$

The ML estimator in \mathcal{M}^r equals:

$$\hat{\mathbf{P}}_t^r \{a_{r+1} \mid a_{1:r}\} = \frac{N_t(a_{1:r+1})}{N_{t-1}(a_{1:r})}$$

for all $a_{1:r+1} \in \mathcal{Y}^{r+1}$, whenever $N_{t-1}(a_{1:r}) > 0$.

The proof of Theorem 2.7.2 is decomposed into two main parts. The easiest part allows to relate the

$$\left[\log \hat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right]$$

and a χ^2 distance between empirical transition distributions $\hat{\mathbf{P}}_t^r \{\cdot \mid \cdot\}$ and $\mathbf{P}^* \{\cdot \mid \cdot\}$, under conditions that aver to be almost surely satisfied by sample paths of irreducible Markov chains. This relationship (Lemma 2.7.3) is a quantitative version of the asymptotic equivalence between relative entropy and χ^2 distance (see Csiszár (1990) for more information on this topic). The most original part actually proves that the almost sure convergence of $\hat{\mathbf{P}}_t^r$ toward \mathbf{P}^* is uniform over all $r \geq r^*$.

Lemma 2.7.3. *Let P and P' be two probability distributions on $\{1, \dots, m\}$.*

If for all $i \in \{1, \dots, m\}$, $\frac{P'\{i\}}{2} \leq P\{i\} \leq 2P'\{i\}$, then $D(P \mid P') \leq \sum_{i=1}^m \frac{(P\{i\} - P'\{i\})^2}{P'\{i\}} = \chi^2(P, P')$.

Here follows a simple corollary of this Lemma:

Corollary 2.7.4. *Let r be an integer such that $r \geq r^*$. If $y_{1:t}$ is such that for all $a_{1:r+1} \in \mathcal{S}_{r+1}$*

$$\frac{1}{2} \mathbf{P}^* \{a_{r+1} \mid a_{1:r}\} \leq \frac{N_t(a_{1:r+1})}{N_{t-1}(a_{1:r})} \leq 2 \mathbf{P}^* \{a_{r+1} \mid a_{1:r}\}$$

then

$$\log \hat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \leq \sum_{a_{1:r} \in \mathcal{S}_r} N_t(a_{1:r}) \chi^2 \left(\hat{\mathbf{P}}_t^r \{\cdot \mid a_{1:r}\}, \mathbf{P}^* \{\cdot \mid a_{1:r}\} \right).$$

2.7.1 Some martingale tools

The proof of Theorem 2.7.2 relies on martingale arguments. The basic tools of martingale theory we need are gathered here.

In the sequel, ϕ denotes the convex function $\phi(x) \stackrel{\text{def}}{=} \exp(x) - x - 1$ and ϕ^* its convex dual, $\phi^*(y) = \sup_x [yx - \phi(x)] = (y + 1) \log(y + 1) - y$, for $y \geq -1$ and ∞ otherwise. We will use repeatedly the classical inequality

$$\phi^*(x) \geq \frac{x^2}{1 + x/3} \quad \text{for } x \geq 0.$$

The following Lemma is classically considered as an extension of Bennett Inequality to Martingale with bounded increments. Various proofs may be found in textbooks on Probability Theory, see Neveu (1975); Dacunha-Castelle and Duflo (1983).

Lemma 2.7.5. *Let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ denote an increasing filtration and $(Z_t)_{t \in \mathbb{N}}$ denote an $(\mathcal{F}_t)_t$ -adapted centered square-integrable martingale with associated increasing process $(\langle Z \rangle_t)_{t \in \mathbb{N}}$ and increments bounded by 1, then for all λ , the sequence*

$$\exp\left(\lambda Z_t - \phi(\lambda) \langle Z \rangle_t\right)$$

is an (\mathcal{F}_t) -adapted super-martingale.

Let us now recall Kolmogorov maximal inequality and the optional sampling principle.

Kolmogorov maximal inequality asserts that if $(Z_t)_t$ is an $(\mathcal{F}_t)_t$ -adapted super-martingale, then for every t_0 , every $x > 0$:

$$\mathbf{P}\left\{\sup_{t \geq t_0} Z_t \geq x\right\} \leq \frac{\mathbf{E}[(Z_{t_0})_+]}{x}. \quad (2.3)$$

Recall that the random variable T is a stopping time with respect to $(\mathcal{F}_t)_t$ if and only if the event $T \leq t$ is \mathcal{F}_t -measurable.

The optional sampling theorem asserts that if $T_1, T_2, \dots, T_k, \dots$ form an increasing sequence of $(\mathcal{F}_t)_t$ -adapted stopping times, then the sequence $(Z_{T_i})_i$ is a $(\mathcal{F}_{T_i})_i$ -adapted super-martingale.

Considering a stopping time T , and the increasing sequence of stopping times $(T \vee n)_n$, it follows from Lemma 2.7.5, Kolmogorov maximal inequality, and the optional sampling Theorem that if $(Z_t)_t$ is a martingale with increments bounded by 1:

$$\text{For any stopping time } T: \quad \mathbf{P}\left\{\exists t \geq T : |Z_t| > \frac{\phi(\lambda)}{\lambda} \langle Z \rangle_t + \alpha\right\} \leq 2 \exp(-\alpha \lambda). \quad (2.4)$$

Let B_1 and B_2 be two numbers such that $B_1 \leq B_2$. If the stopping times T_1 and T_2 are defined by $T_1 = \inf\{t : \langle Z \rangle_t \geq B_1\}$, and $T_2 = \inf\{t : \langle Z \rangle_t \geq B_2\}$, fixing $x > 0$, Inequality (2.4), entails

$$\begin{aligned} \mathbf{P}\left\{\exists t \in \{T_1, \dots, T_2\} : |Z_t| > x\right\} &\leq 2 \exp\left(-B_2 \sup_{\lambda} \left[\lambda \frac{x}{B_2} - \phi(\lambda)\right]\right) \\ &= 2 \exp\left(-B_2 \phi^*\left(\frac{x}{B_2}\right)\right) \\ &\leq 2 \exp\left(-\frac{x^2}{2(B_2 + x/3)}\right). \end{aligned} \quad (2.5)$$

Inequality (2.5) will aver to be the workhorse in the proof of Theorem 2.7.2.

2.7.2 A martingale approach to $\widehat{\mathbf{P}}_t^r$

The following observation has proved to be crucial in the developments that started with Finesso (1991) and culminated in Csiszár (2002): for each $r > r^*$, for each $a_{1:r} \in \mathscr{Y}^r$, the random variables $(Z_t(a_{1:r}))_t$ defined by:

$$Z_t(a_{1:r}) \stackrel{\text{def}}{=} N_t(a_{1:r}) - N_{t-1}(a_{1:r-1}) \times \mathbf{P}^*\{a_r \mid a_{1:r-1}\},$$

is an \mathcal{F}_t -adapted martingale. Moreover this martingale has increments bounded by 1 and -1 , and the associated increasing process (or conditional variance process)

$$\langle Z_t(a_{1:r}) \rangle \stackrel{\text{def}}{=} \sum_{s=1}^t \mathbb{E} \left[\left(Z_s(a_{1:r}) - Z_{s-1}(a_{1:r}) \right)^2 \mid \mathcal{F}_{s-1} \right]$$

has the following form:

$$\langle Z(a_{1:r}) \rangle_t = N_{t-1}(a_{1:r-1}) \mathbf{P}^*\{a_r \mid a_{1:r-1}\} \left(1 - \mathbf{P}^*\{a_r \mid a_{1:r-1}\} \right). \quad (2.6)$$

Note that $|Z_t(a_{1:r})| < x$ entails that

$$\left| \widehat{\mathbf{P}}_t^{r-1}\{a_r \mid a_{1:r-1}\} - \mathbf{P}^*\{a_r \mid a_{1:r-1}\} \right| < \frac{x}{N_{t-1}(a_{1:r-1})}.$$

Hence bounds on the deviations of the martingales $Z_t(a_{1:r})$ for $a_{1:r} \in \mathcal{S}_r$ are of immediate relevance to the characterization of $\widehat{\mathbf{P}}_t^{r-1}$.

The following Lemma will be the fundamental bridging block in the proof of the large-scale typicality Theorem.

Lemma 2.7.6. *Let ξ, η be two positive reals. Let $r > r^*$, and let $a_{1:r} \in \mathcal{S}^r$. Let Z_t denote the martingale associated with $a_{1:r}$. Let $\theta > 1$,*

$$\begin{aligned} \mathbf{P}^* \left\{ \exists t : \theta^m \leq \langle Z \rangle_t \leq \theta^{m+1}, \quad |Z_t| \geq \sqrt{\langle Z \rangle_t \max[\xi r, \eta \log \log(\langle Z \rangle_t)]} \right\} \\ \leq 2 \exp \left(- \frac{\max[\xi r, \eta \log \log(\theta^m)]}{2\theta(1 + (1/3)\sqrt{\max[\xi r, \eta \log \log(\theta^m)]/\theta^{m+2}}} \right). \end{aligned} \quad (2.7)$$

Proof of Lemma 2.7.6. For each integer m , let the stopping time T_m be defined as the first instant t such that $\langle Z \rangle_t \geq \theta^m$. Note that for t between T_m and T_{m+1} , $\langle Z \rangle_t \geq \theta_m$, hence we may take $x = \sqrt{\theta^m \max[\xi r, \eta \log \log \theta^m]}$ and $B_2 = \theta^{m+1}$ in Inequality (2.5). \square

Remark 2.4. If $a_{1:r} \in \mathcal{S}_r$, ergodicity implies \mathbf{P}^* -almost surely $\langle Z(a_{1:r}) \rangle_t$ converges toward infinity. Choosing $\xi = 0$, taking $\eta = 2\theta(1 + \alpha)$ with $\alpha > 0$, the previous Lemma asserts that

$$\begin{aligned} \mathbf{P}^* \left\{ \exists t : \theta^m \leq \langle Z \rangle_t \leq \theta^{m+1}, \quad |Z_t| \geq \sqrt{2\theta(1 + \alpha)\langle Z \rangle_t \log \log(\langle Z \rangle_t)} \right\} \\ \leq 2 \exp \left(- \frac{(1 + \alpha) \log \log \theta^m}{\left(1 + \frac{1}{3} \sqrt{\frac{2(1 + \alpha) \log \log \theta^m}{\theta^{m+1}}}\right)} \right). \end{aligned}$$

The sum over m of right-hand-side terms is finite, hence by the Borel-Cantelli Lemma, \mathbf{P}^* -almost surely, the event $\exists t : \theta^m \leq \langle Z \rangle_t \leq \theta^{m+1}, \quad |Z_t| \geq \sqrt{2\theta(1 + \alpha)\langle Z \rangle_t \log \log(\langle Z \rangle_t)}$ may only occur for finitely many m s. Combining the two observations, and taking θ toward 1 and α toward 0 completes the proof that \mathbf{P}^* -almost surely

$$\limsup_t \frac{|Z_t|}{\sqrt{2\langle Z \rangle_t \log \log \langle Z \rangle_t}} \leq 1. \quad (2.8)$$

Note that by Corollary 2.7.4, this entails that for some fixed $r > r^*$, \mathbf{P}^* -almost surely, eventually for all $a_{1:r} \in \mathcal{S}_r$:

$$\frac{N_{t-1}(a_{1:r})}{|\mathcal{Y}|} \chi^2 \left(\widehat{\mathbf{P}}_t^r \{ \cdot \mid a_{1:r} \}, \mathbf{P}^* \{ \cdot \mid a_{1:r} \} \right) \leq 2 \log \log N_{t-1}(a_{1:r}),$$

and

$$\frac{1}{|\mathcal{Y}| |\mathcal{S}_r|} \left[\log \widehat{\mathbf{P}}_t^r \{ y_{1:t} \} - \log \mathbf{P}^* \{ y_{1:t} \} \right] \leq 2 \log \log t.$$

If we were ready to assume that r^* is smaller than some given upper-bound on the true order, this would be enough to ensure almost sure consistency of penalized maximum likelihood order estimators by taking

$$\text{pen}(t, r) = 2 |\mathcal{Y}|^{r+1} \log \log t.$$

2.7.3 The union bound meets martingale inequalities

The following Lemma will allow us to control $\sup_{r, r^* \leq r \leq \alpha \log t} [\log \widehat{\mathbf{P}}_t^r \{ y_{1:t} \} - \log \mathbf{P}^* \{ y_{1:t} \}]$ for a suitable range of α .

Lemma 2.7.7. *For every $\delta > 0$, there exists $\alpha > 0$ (depending in \mathbf{P}^*) such that eventually almost surely as $t \rightarrow \infty$, for all $a_{1:r}$ in \mathcal{S}_r with $r^* < r \leq \alpha \log t$*

$$|Z_t(a_{1:r})| \leq \sqrt{\delta \langle Z(a_{1:r}) \rangle_t \log \langle Z(a_{1:r}) \rangle_t}.$$

Let the event $D_t^{\xi, c, \eta}(a_{1:r})$ be defined by

$$D_t^{\xi, c, \eta}(a_{1:r}) \stackrel{\text{def}}{=} \left\{ y_{1:t} : \langle Z(a_{1:r}) \rangle_t > cr, \right. \\ \left. |Z_t(a_{1:r})| \geq \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log(\langle Z(a_{1:r}) \rangle_t)]} \right\}.$$

Lemma 2.7.8. *Let ξ, η, c be chosen in such a way that there exists $\theta > 1$ such that*

$$\xi > 2 \log |\mathcal{Y}| \left(\theta + \frac{\sqrt{\xi}}{3} \max \left[\frac{1}{\sqrt{c}}, 1 \right] \right), \quad (2.9)$$

and

$$\eta > \frac{\xi}{\frac{\xi}{2(\theta + \sqrt{\xi}/3 \max[\frac{1}{\sqrt{c}}, 1])} - \log |\mathcal{Y}|}. \quad (2.10)$$

Then,

$$\limsup_t \sum_{r \geq r^*} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{D_t^{\xi, c, \eta}(a_{1:r})} = 0 \quad \mathbf{P}^* \text{-almost surely.}$$

Proof of Lemma 2.7.8. Fix $\theta > 1$ in such a way that Inequalities (2.9) and (2.10) are satisfied. For each integer m , let the event $E_m^{\xi, c, \eta}(a_{1:r})$ be defined by

$$E_m^{\xi, c, \eta}(a_{1:r}) \stackrel{\text{def}}{=} \left\{ y_{1:\infty} : \theta^m > cr, \exists a_{1:r}, \exists t \in \{T_m(a_{1:r}), \dots, T_{m+1}(a_{1:r})\}, \right. \\ \left. |Z_t(a_{1:r})| \geq \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log(\langle Z(a_{1:r}) \rangle_t)]} \right\}.$$

The Lemma will be proved in two steps. We will first check that \mathbf{P}^* -almost surely, only finitely many events $E_m^{\xi, c, \eta}(a_{1:r})$ occur. Then we will check that on a set of sample paths which has probability 1, this entails that only finitely many events $D_t^{\xi, c, \eta}(a_{1:r})$ occur.

Note that

$$\max[\xi r, \eta \log \log(\theta^m)] = \begin{cases} \xi r & \text{if } r \geq \frac{\eta}{\xi} \log \log \theta^m \\ \eta \log \log(\theta^m) & \text{otherwise.} \end{cases}$$

To alleviate notations, let μ be defined as

$$\mu \stackrel{\text{def}}{=} \frac{\xi}{2\left(\theta + \frac{\sqrt{\xi}}{3} \max[1/\sqrt{c}, 1]\right)} - \log|\mathcal{Y}|.$$

In the sequel, summation is over all m such that $\eta \log \log \theta^m / \theta^m < 1$.

$$\begin{aligned} \mathbb{E}\left[\sum_m \sum_r \sum_{a_{1:r}} \mathbf{1}_{E_m^{\xi,c,\eta}(a_{1:r})}\right] &\leq \sum_m \left\{ \sum_{\frac{\eta}{\xi} \log \log \theta^m \leq r \leq \theta^m/c} |\mathcal{Y}|^r \exp\left(-\frac{\xi r}{2\left(\theta + \frac{1}{3}\sqrt{\frac{\xi r}{\theta^m}}\right)}\right) \right. \\ &\quad \left. + \sum_{r^* < r \leq \frac{\eta}{\xi} \log \log \theta^m} |\mathcal{Y}|^r \exp\left(-\frac{\eta \log \log \theta^m}{2\left(\theta + \frac{1}{3}\sqrt{\frac{\eta \log \log \theta^m}{\theta^m}}\right)}\right) \right\} \\ &\leq \sum_m \exp\left(-\frac{\mu \eta}{\xi} \log \log \theta^m\right) \times \left[\frac{1}{|\mathcal{Y}| - 1} + \frac{1}{1 - \exp(-\mu)}\right] \end{aligned}$$

Note that as by Inequality (2.9) $\mu \eta > \xi$, the last sum is finite which shows that our first goal is attained.

As \mathbb{P}^* is assumed to be ergodic, with probability 1, for all $r > r^*$, for all $a_{1:r} \in \mathcal{S}_r$, $\langle Z(a_{1:r}) \rangle_t$ tends toward infinity. Let us consider a sample path such that for all $r > r^*$, for all $a_{1:r} \in \mathcal{S}_r$, $\langle Z(a_{1:r}) \rangle_t$ tends toward infinity. Then if infinitely many events of the form $D_t^{\xi,c,\eta}(a_{1:r})$ occur for a fixed pattern $a_{1:r}$, then infinitely many events of the form $E_m^{\xi,c,\eta}(a_{1:r})$ occur for the same fixed pattern.

If there exists an infinite sequence $(a_{1:r_t})_t$ of patterns such that $D_t^{\xi,c,\eta}(a_{1:r_t})$ occurs for infinitely many values of t , then infinitely many events of the form $E_{m_t}^{\xi,c,\eta}(a_{1:r_t})$ occur. \square

In order to prove Lemma 2.7.7, we will need lower bounds on $\mathbb{P}^*\{a_{1:r}\}$ for all $r \geq r^*$, and all $a_{1:r} \in \mathcal{S}_r$. As \mathbb{P} has Markov order r^* , we have

$$\mathbb{P}^*\{a_{1:r}\} = \mathbb{P}^*\{a_{1:r^*}\} \prod_{j=r^*+1}^r \mathbb{P}^*\{a_j \mid a_{j-1:j-r^*}\}.$$

Now let $\gamma \stackrel{\text{def}}{=} \min_{a_{1:r^*} \in \mathcal{S}_{r^*}} \mathbb{P}^*\{a_{1:r^*}\}$ and let $\kappa \stackrel{\text{def}}{=} \min_{a_{1:r^*+1} \in \mathcal{S}_{r^*+1}} \mathbb{P}^*\{a_{r^*+1} \mid a_{1:r^*}\}$ then

$$\min_{a_{1:r} \in \mathcal{S}_r} \mathbb{P}^*\{a_{1:r}\} \geq \gamma \kappa^{r-r^*}. \quad (2.11)$$

Proof of Lemma 2.7.7. We will rely on Lemma 2.7.8. In this proof, we will fix η, ξ and c so as to satisfy the conditions in the latter Lemma. The challenge will consist in checking that for every $\delta > 0$, we will be able to find some $\alpha > 0$ such that

1. \mathbb{P}^* -almost surely all the ‘‘clocks’’ associated with patterns $\in \cup_{r \in \{r^*, \dots, \alpha \log t\}} \mathcal{S}_r$, move sufficiently fast, that is, for all sufficiently large t :

$$\langle Z(a_{1:r}) \rangle_t > r \quad \text{for all } a_{1:r} \in \cup_{r \in \{r^*, \dots, \alpha \log t\}} \mathcal{S}_r,$$

2. for all sufficiently large t :

$$\max[\xi r, \eta \log \log \langle Z(a_{1:r}) \rangle_t] \leq \delta \log t \quad \text{for all } a_{1:r} \in \cup_{r \in \{r^*, \dots, \alpha \log t\}} \mathcal{S}_r.$$

Let us first make a few observations. Assume for a moment, that $(\epsilon_r)_r$ is a non-decreasing sequence of small positive constants.

$$\text{If } 1 - \epsilon_{r-1} < \left| \frac{N_{t-1}(a_{1:r-1})}{(t-r+1)\mathbb{P}^*\{a_{1:r-1}\}} \right| < 1 + \epsilon_{r-1},$$

$$\text{and } |Z_t(a_{1:r})| < \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log \langle Z(a_{1:r}) \rangle_t]},$$

we have

$$(1 - \epsilon_{r-1})(t - r + 1)\mathbf{P}^*\{a_{1:r}\} \leq \langle Z(a_{1:r}) \rangle_t \leq (1 + \epsilon_{r+1})(t - r + 1)\mathbf{P}^*\{a_{1:r}\}.$$

Then

$$\begin{aligned} N_t(a_{1:r}) &> N_{t-1}(a_{1:r-1})\mathbf{P}^*\{a_r \mid a_{1:r-1}\} - \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log \langle Z(a_{1:r}) \rangle_t]} \\ &> (t - r - 1)\mathbf{P}^*\{a_{1:r}\} \left[1 - \epsilon_{r-1} - \frac{\sqrt{(1 + \epsilon_{r-1}) \max[\xi r, \eta \log \log (2(t - r + 1)\mathbf{P}^*\{a_{1:r}\})]}}{\sqrt{(t - r + 1)\mathbf{P}^*\{a_{1:r}\}}} \right] \\ &> (t - r + 1)\mathbf{P}^*\{a_{1:r}\} \left[1 - \epsilon_{r-1} - \frac{2\sqrt{\max[\xi r, \eta \log \log (2t)]}}{\sqrt{t\gamma\kappa^{r-r^*}}} \right], \end{aligned}$$

and

$$\begin{aligned} N_t(a_{1:r}) &< N_{t-1}(a_{1:r-1})\mathbf{P}^*\{a_r \mid a_{1:r-1}\} + \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log \langle Z(a_{1:r}) \rangle_t]} \\ &< (t - r + 1)\mathbf{P}^*\{a_{1:r}\} \left[1 + \epsilon_{r-1} + \frac{2\sqrt{\max[\xi r, \eta \log \log (2t)]}}{\sqrt{t\gamma\kappa^{r-r^*}}} \right]. \end{aligned}$$

If we agree on

$$\epsilon_r = \epsilon_{r-1} + \frac{2\sqrt{\max[\xi r, \eta \log \log (2t)]}}{\sqrt{t\gamma\kappa^{r-r^*}}}, \quad (2.12)$$

this translates into:

$$(1 - \epsilon_r) < N_t(a_{1:r}) < (1 + \epsilon_r)$$

Let us agree on

$$\epsilon_{r^*}(t) < 1/2 - \frac{\eta}{\xi} \log \log (2t) \frac{2\sqrt{\eta \log \log 2t}}{\sqrt{t\gamma\kappa^{\eta/\xi \log \log (2t)}}} + \alpha \log (2t) \frac{\sqrt{\xi \alpha \log t}}{\sqrt{\gamma t}}.$$

Then \mathbf{P}^* -almost surely, for t large enough for all $a_{1:r^*} \in \mathcal{S}_{r^*}$,

$$1 - \epsilon_{r^*} < \left| \frac{N_t(a_{1:r^*})}{(t - r + 1)\mathbf{P}^*\{a_{1:r^*}\}} \right| < 1 + \epsilon_{r^*}.$$

Let α be such that $\alpha < \frac{1}{2 \log(1/\kappa)}$. Then for all r satisfying $r^* \leq r < \alpha \log t$, if we agree on Relation (2.12), for sufficiently large t , we have $\epsilon_r(t) \leq 1/2$.

But this implies that \mathbf{P}^* -almost surely for all sufficiently large t , for all $r \leq \alpha \log t$, for all $a_{1:r} \in \mathcal{S}_r$:

$$\langle Z(a_{1:r}) \rangle_t \geq \frac{1}{2}(t - r + 1)\gamma\kappa^r \geq \frac{\gamma(t - r + 1)}{\sqrt{t}} > cr.$$

By Lemma 2.7.8, this implies that \mathbf{P}^* -almost surely, for all sufficiently large t , for all $r \leq \alpha \log t$, for all $a_{1:r} \in \mathcal{S}_r$:

$$|Z_t(a_{1:r})| \leq \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, \eta \log \log \langle Z(a_{1:r}) \rangle_t]}.$$

If α is sufficiently small, the right-hand-side is smaller than $\sqrt{\delta \langle Z(a_{1:r}) \rangle_t \log \langle Z(a_{1:r}) \rangle_t}$ in the range of r considered. \square

The next Lemma will prove crucial when checking the most delicate part of the BIC consistency Theorem. It will allow us to rule out (almost surely) the possibility that the BIC order estimator errs around $\log t$ for infinitely many values of t .

For any $\xi > 0$ and any $c > 0$, and any $a_{1:r}$, define the event $B_t^{\xi,c}(a_{1:r})$ as:

$$B_t^{\xi,c}(a_{1:r}) \stackrel{\text{def}}{=} \left\{ y_{1:t} : \langle Z(a_{1:r}) \rangle_t > cr \text{ and } |Z_t(a_{1:r})| \geq \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, 4 \log \log \langle Z(a_{1:r}) \rangle_t]} \right\}.$$

Lemma 2.7.9. *Let $\xi > 0, c > 0$, be such that $\sqrt{\xi} < 3/2$. Then,*

$$\limsup_t \sup_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{B_t^{\xi,c}(a_{1:r})} = 0 \text{ P}^* \text{-almost surely.}$$

Proof of Lemma 2.7.9. Let us choose $\theta > 1$, so that for sufficiently large m , $(\theta + \frac{1}{3})\sqrt{\frac{4 \log \log \theta^m}{\theta^m}} \leq 3/2$.

$$C_m^{\xi,c}(a_{1:r}) \stackrel{\text{def}}{=} \left\{ y_{1:\infty} : \exists t \theta^m \leq \langle Z(a_{1:r}) \rangle_t \leq \theta^{m+1} \text{ and } \theta^m > cr \right. \\ \left. \text{and } |Z_t(a_{1:r})| \geq \sqrt{\langle Z(a_{1:r}) \rangle_t \max[\xi r, 4 \log \log \langle Z(a_{1:r}) \rangle_t]} \right\}.$$

The proof of Lemma 2.7.9 is carried in two steps:

1. Proving that P*-almost surely,

$$\limsup_M \sum_{m > M} \sum_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})} = 0. \quad (2.13)$$

2. Proving that this entails

$$\limsup_t \sup_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{B_t^{\xi,c}(a_{1:r})} = 0. \quad (2.14)$$

Note that when dealing with $\frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})}$, we adapt the time-scale at which we analyze $Z_t(a_{1:r})$ to the pattern. This allows to formulate a rather strong statement: not only does

$$u_m \stackrel{\text{def}}{=} \sum_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})}$$

tend toward 0 as m tends towards infinity, but the series $\sum_m u_m$ is convergent.

Let us start with the first step. Thanks to our assumptions on the values of ξ and m .

$$\begin{aligned} \mathbb{E} \left[\sum_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})} \right] &\leq 2 \sum_{\frac{4}{\xi} \log \log \theta^m < r < \frac{\theta^m}{c}} \exp \left(- \frac{\xi r}{2(\theta + \frac{\sqrt{\xi/c}}{3})} \right) \\ &\quad + 2 \sum_{r \leq \frac{4}{\xi} \log \log \theta^m} \exp \left(- \frac{4 \log \log \theta^m}{2(\theta + \frac{1}{3})\sqrt{\frac{4 \log \log \theta^m}{\theta^m}}} \right) \\ &< \infty. \end{aligned}$$

Hence

$$\sum_{m > M} \mathbb{E} \left[\sum_{r > r^*} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})} \right] < \infty,$$

which entails that P*-almost surely, Relation (2.13) holds.

Let us now proceed to the second step. As \mathbf{P}^* is assumed to be ergodic, it is enough to consider sequences $y_{1:\infty}$ such that for all $a_{1:r}$, $\langle Z(a_{1:r}) \rangle_t$ tends toward infinity.

Assume that, there exists a sequence $(r_t)_t$ such that for some $\alpha > 0$, for infinitely many t :

$$\frac{1}{|\mathcal{S}_{r_t}|} \sum_{a_{1:r_t} \in \mathcal{S}_{r_t}} \mathbf{1}_{B_t^{\xi,c}(a_{1:r_t})} > \alpha.$$

If the sequence r_t has an accumulation point r , then there exists some $a_{1:r}$ such that $B_t^{\xi,c}(a_{1:r_t})$ occurs for infinitely many t , but this entails that infinitely many events $C_m^{\xi,c}(a_{1:r})$ occur, which means that whatever M

$$\sum_{m > M} \frac{1}{|\mathcal{S}_r|} \mathbf{1}_{C_m^{\xi,r}(a_{1:r})} = \infty.$$

If the sequence r_t is increasing, then for each t such that

$$\frac{1}{|\mathcal{S}_{r_t}|} \sum_{a_{1:r_t} \in \mathcal{S}_{r_t}} \mathbf{1}_{B_t^{\xi,c}(a_{1:r_t})} > \alpha$$

is realized,

$$\frac{1}{|\mathcal{S}_{r_t}|} \sum_{a_{1:r_t}} \sum_{m < \log_\theta(c r_t)} \mathbf{1}_{C_m^{\xi,r}(a_{1:r_t})} > \alpha.$$

Hence, whatever M :

$$\sum_{m > M} \sum_{r > \theta^m/c} \frac{1}{|\mathcal{S}_r|} \sum_{a_{1:r} \in \mathcal{S}_r} \mathbf{1}_{C_m^{\xi,c}(a_{1:r})} > \alpha.$$

□

Remark 2.5. Lemma 2.7.8 and 2.7.9 are proved in a very similar way, they have similar form, but convey a different message. In Lemma 2.7.8, the constant η may be taken rather close to 2, and the constants in the Lemma may be considered as trade-offs between the constants that show up in the Law of the Iterated Logarithm and the constants that may be obtained if the union bound has to be used repeatedly. Note that if conditions in Lemma 2.7.8 are to be met, for a given ξ , we cannot look for arbitrarily small c .

This is sharp contrast with the setting of Lemma 2.7.9. There the constant η was deliberately set to 4, and the freedom provided by this convention, as well as by the normalizing factors $1/|\mathcal{S}_r|$, allows to consider arbitrarily small values for c .

Proof of Theorem 2.7.2. It is enough to prove that for every $\delta > 0$, \mathbf{P}^* -almost surely, eventually

$$\sup_{r \geq r^*} \left\{ \frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \right\} < \delta.$$

Note first that if $|\mathcal{S}_r|$ does not grow exponentially fast with r , then the Markov chain has zero entropy rate, it is a deterministic process, and the likelihood ratios of interest are equal to 1, there is nothing to do.

Let us thus assume that there exists some $h > 0$ such that for all sufficiently large r , $\log |\mathcal{S}_r| \geq hr$.

Let κ and γ be defined as on page 20, we have

$$\frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}_t^* \{y_{1:t}\} \right] \leq e^{-hr} \log \frac{1}{\gamma \kappa^t}.$$

Hence for all sufficiently large t , for every $r \geq \frac{2}{h} \log t$, the condition is automatically enforced.

It remains to prove that for every $\delta > 0$:

$$\sup_{\substack{r \geq r^* \\ r < \frac{2}{h} \log t}} \left\{ \frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \right\} \geq \delta.$$

occurs only finitely many times.

Assume $\delta < 1/4$. Then by Lemma 2.7.7, there exists some $\alpha > 0$ depending on \mathbf{P}^* and δ , such that for all sufficiently large t , for all $r : r^* < r < \alpha \log t$, for all $a_{1:r} \in \mathcal{S}_r$

$$|Z_t(a_{1:r})| < \sqrt{\delta \langle Z(a_{1:r}) \rangle_t}. \quad (2.15)$$

But this inequality entails:

$$\left| \widehat{\mathbf{P}}_t^r \{a_r \mid a_{1:r-1}\} - \mathbf{P}^* \{a_r \mid a_{1:r-1}\} \right| \leq \sqrt{\delta \frac{\mathbf{P}^* \{a_r \mid a_{1:r-1}\} \log N_{t-1}(a_{1:r-1})}{N_{t-1}(a_{1:r-1})}}.$$

Hence \mathbf{P}^* -almost surely, for all sufficiently large t , for all $r : r^* < r < \alpha \log t$:

$$\frac{N_{t-1}(a_{1:r-1})}{|\mathcal{Y}|} \chi^2 \left(\widehat{\mathbf{P}}_t^r \{ \cdot \mid a_{1:r-1} \}, \mathbf{P}^* \{ \cdot \mid a_{1:r-1} \} \right) \leq \delta \log t. \quad (2.16)$$

On the other hand, notice that if

$$|Z_t(a_{1:r})| \leq \frac{1}{2} \langle Z(a_{1:r}) \rangle_t,$$

then

$$\left| \widehat{\mathbf{P}}_t^r \{a_r \mid a_{1:r-1}\} - \mathbf{P}^* \{a_r \mid a_{1:r-1}\} \right| \leq \frac{1}{2} \mathbf{P}^* \{a_r \mid a_{1:r-1}\}.$$

Hence, by Corollary 2.7.4, as $\delta \log u < u/4$, \mathbf{P}^* -almost surely, for all sufficiently large t , for all $r : r^* < r < \alpha \log t$

$$\frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \leq \delta.$$

Hence, \mathbf{P}^* -almost surely, for sufficiently large t

$$\sup_{r < r^* < \alpha \log t} \frac{1}{|\mathcal{S}_r|} \frac{1}{\log t} \left[\log \widehat{\mathbf{P}}_t^r \{y_{1:t}\} - \log \mathbf{P}^* \{y_{1:t}\} \right] \leq \delta.$$

Let us now consider those r such that $\alpha \log t \leq r \leq \frac{2}{h} \log t$. Let us choose ξ_2, c_2 such that for some (irrelevant) $\eta > 2$, the conditions in Lemma 2.7.8 are satisfied. Note that for t sufficiently large, for all r such that $\alpha \log t \leq r \leq \frac{1}{h} \log t$, $\max[\xi_2 r, \eta \log \log t] = \xi_2 r$.

Let $\xi_1 > 0$ and $c_1 > 0$ be chosen in such a way that $c_1 + \xi_1 < h\delta/2$. We will use Lemma 2.7.9 with those constants. Recall that c_1 and ξ_1 may be chosen arbitrarily close to 0 (see remark following Proof of Lemma).

Let $G_1^{r,t}, G_2^{r,t}, G_3^{r,t}, G_4^{r,t}$ be defined by:

$$\begin{aligned} G_1^{r,t} &\stackrel{\text{def}}{=} \left\{ a_{1:r-1} : N_{t-1}(a_{1:r-1}) < c_1 r \right\} \cap \mathcal{S}_{r-1} \\ G_2^{r,t} &\stackrel{\text{def}}{=} \left\{ a_{1:r-1} : c_1 r \leq N_{t-1}(a_{1:r-1}) \right. \\ &\quad \left. \text{and } \forall a \in \mathcal{Y}, |Z_t(a_{1:r-1}, a)| < \sqrt{\xi_1 r \langle Z(a_{1:r-1}, a) \rangle_t} \right\} \\ G_3^{r,t} &\stackrel{\text{def}}{=} \left\{ a_{1:r-1} : c_1 r > N_{t-1}(a_{1:r-1}) < c_2 r \right. \\ &\quad \left. \text{and } \exists a \in \mathcal{Y}, |Z_t(a_{1:r-1}, a)| < \sqrt{\xi_1 r \langle Z(a_{1:r-1}, a) \rangle_t} \right\} \\ G_4^{r,t} &\stackrel{\text{def}}{=} \left\{ a_{1:r-1} : c_2 r < N_{t-1}(a_{1:r-1}) \right. \\ &\quad \left. \text{and } \forall a \in \mathcal{Y}, |Z_t(a_{1:r-1}, a)| < \sqrt{\xi_2 r \langle Z(a_{1:r-1}, a) \rangle_t} \right\} \setminus G_2^{r,t}. \end{aligned}$$

By Lemma 2.7.8, \mathbf{P}^* -almost surely, for sufficiently large t , for all r such that $\alpha \log t \leq r \leq \frac{1}{h} \log t$,

$$G_1^{r,t} \cup G_2^{r,t} \cup G_3^{r,t} \cup G_4^{r,t} = \mathcal{S}_{r-1}.$$

And by Lemma 2.7.9, \mathbf{P}^* -almost surely, for sufficiently large t , for all r such that $\alpha \log t \leq r \leq \frac{1}{h} \log t$,

$$\frac{|G_3^{r,t}| + |G_4^{r,t}|}{|\mathcal{S}_{r-1}|} < \delta$$

As by the definition of $G_2^{r,t}$ and $G_4^{r,t}$, we are in a position to use Lemma 2.7.3 and Corollary 2.7.4

$$N_{t-1}(a_{1:r-1})D\left(\widehat{\mathbf{P}}_t\{\cdot \mid a_{1:r-1}\} \mid \mathbf{P}^*\{\cdot \mid a_{1:r-1}\}\right) \leq \begin{cases} \xi_1 r & \text{if } a_{1:r-1} \in G_2^{r,t} \\ \xi_2 r & \text{if } a_{1:r-1} \in G_4^{r,t}. \end{cases} \quad (2.17)$$

\mathbf{P}^* -almost surely, for sufficiently large t , for all r such that $\alpha \log t \leq r \leq \frac{1}{h} \log t$,

$$\begin{aligned} \log \widehat{\mathbf{P}}_t^r\{y_{1:t}\} - \log \mathbf{P}^*\{y_{1:t}\} & \leq \sum_{i=1}^4 \sum_{a_{1:r-1} \in G_i^{r,t}} N_{t-1}(a_{1:r-1})D\left(\widehat{\mathbf{P}}_t\{\cdot \mid a_{1:r-1}\} \mid \mathbf{P}^*\{\cdot \mid a_{1:r-1}\}\right) \\ & \leq |G_1^{r,t}|c_1 r \log \frac{1}{\kappa} + |G_2^{r,t}|\xi_1 r + |G_3^{r,t}|c_2 r \log \frac{1}{\kappa} + |G_4^{r,t}|\xi_2 r. \end{aligned}$$

Now, dividing both sides by $|\mathcal{S}_r| \log t$, we get for the range of r of interest

$$\frac{1}{|\mathcal{S}_r| \log t} \left[\log \widehat{\mathbf{P}}_t^r\{y_{1:t}\} - \log \mathbf{P}^*\{y_{1:t}\} \right] \leq \frac{2}{h} \left[c_1 + \xi_1 + c_2 \frac{|G_3^{r,t}|}{|\mathcal{S}_r|} + \frac{|G_4^{r,t}|}{|\mathcal{S}_r|} \xi_2 \right]$$

As we have agreed on $c_1 + \xi_1 \leq h\delta/2$, \mathbf{P}^* -almost surely, for sufficiently large t ,

$$\limsup_t \sup_{r: \alpha \log t \leq r \leq \frac{2}{h} \log t} \frac{1}{|\mathcal{S}_r| \log t} \left[\log \widehat{\mathbf{P}}_t^r\{y_{1:t}\} - \log \mathbf{P}^*\{y_{1:t}\} \right] \leq \delta.$$

□

2.8 Efficiency Issues

How efficient are the aforementioned order estimation procedures? The notions of efficiency that have been considered in the order estimation literature have been shaped on the testing theory setting. As a matter of fact, the classical efficiency notions have emerged from the analysis of the simple hypotheses testing problem. Determining how those notions could be tailored to the nested composite hypothesis testing problem is still a subject of debate.

Among the several notions of efficiency, or even of asymptotic relative efficiency that are regarded as relevant in testing theory, Pitman's efficiency focuses on the minimal sample size that is required to achieve simultaneously a given level and a given power at alternatives. Up to our knowledge, Pitman's efficiency for Markov order or HMM order estimation related problems has not been investigated. This is due to our lack of non-asymptotic results concerning estimation procedures for HMM and Markov chains.

The notion of efficiency that has been assessed in the order estimation literature is rather called Bahadur relative efficiency in the Statistical literature and error exponents in the Information-theoretical literature. When testing a simple hypothesis against another simple hypothesis in the memoryless setting, a classical result by Chernoff tells us that comparing likelihood ratios to a fixed threshold, both level and power may decline exponentially fast with respect to the number

of observations. In that setting Bahadur-efficient testing procedures are those ones that achieve the largest exponents. Viewing that set of circumstances, there have been several attempts to generalize those results to the composite hypothesis setting. Part of the difficulty lies in stating the proper questions.

Although consistency issues concerning BIC and MDL criterion for HMM order estimation have not yet been clarified, our understanding of efficiency issues concerning HMM order identification recently underwent significant progress.

2.8.1 Variations on Stein's Lemma

The next theorems are extensions of Stein's Lemma to the order estimation problem. Theorem 2.8.1 aims at determining the best underestimation exponent for a class of order estimators that ultimately overestimate the order with a probability bounded away from 1. Theorem 2.8.2 aims at proving that the best overestimation exponent should be trivial in most cases of interest.

Assumption 2.2. 1. *The sequence of models satisfies the general AEP (see Section 2.4.2) .*

2. *For any r , there exists $\mathcal{M}_0^r \subset \mathcal{M}^r$ such that any P in \mathcal{M}_0^r is stationary ergodic and has true order r , and such that for any $P^* \in \mathcal{M}_0^{r^*}$*

$$\inf_{P \in \mathcal{M}^r} D_\infty(P | P^*) = \inf_{P \in \mathcal{M}_0^r} D_\infty(P | P^*) .$$

Versions of the following Theorem have been proved in Finesso et al. (1996) for Markov chains and in Gassiat and Boucheron (to appear) for HMMs.

Theorem 2.8.1. *Let the sequence of nested models $(\mathcal{M}^r)_{r \in \mathbb{N}}$ satisfy Assumption 2.2. Let $(\hat{r}_t)_{t \in \mathbb{N}}$ denote a sequence of order estimators such that for some $\alpha < 1$, for all r^* , for all $P^* \in \mathcal{M}_0^{r^*}$*

$$P^* \{ \hat{r}_t(Y_{1:t}) > r^* \} < \alpha$$

for all $t \geq T_1(P^*, \alpha, r^*)$. Then, for all r^* , for all $P^* \in \mathcal{M}_0^{r^*}$,

$$\liminf_{t \rightarrow +\infty} \frac{1}{t} \log P^* \{ \hat{r}_t(Y_{1:t}) < r^* \} \geq - \min_{r' < r^*} \inf_{P \in \mathcal{M}^{r'}} D_\infty(P | P^*) .$$

Proof. Fix $P^* \in \mathcal{M}_0^{r^*}$. Let $P' \in \mathcal{M}_0^{r'}$ with $r' < r^*$, define

$$\begin{aligned} A_t(P') &\stackrel{\text{def}}{=} \{ y_{t:1} : \hat{r}_t(y_{1:t}) \leq k' \} \\ B_t(P') &\stackrel{\text{def}}{=} \left\{ y_{1:t} : \frac{1}{t} \log \frac{P'\{y_{1:t}\}}{P^*\{y_{1:t}\}} \leq D_\infty(P' | P^*) + \epsilon \right\} . \end{aligned}$$

For $t > T_1(P', \alpha, r')$,

$$P'\{A_t(P')\} > 1 - \alpha,$$

and as $\cup_r \mathcal{M}^r$ is assumed to satisfy the generalized Shannon-Breiman-McMillan Theorem, for all $t > T_3(\epsilon, P', P^*)$,

$$P'\{B_t(P')\} > 1 - \epsilon. \tag{2.18}$$

If $t > T_2(\alpha, \epsilon, P') = \max(T_1(\alpha, r'), T_3(\epsilon, P', P^*))$:

$$\begin{aligned}
\mathbb{P}^* \{ \widehat{r}_t(Y_{1:t}) < r^* \} &= \mathbb{E}_{\mathbb{P}^*} [\mathbf{1}_{\widehat{r}_t < r^*}] \\
&\text{boils down to equality if } \mathbb{P}^* \text{ and } \mathbb{P}' \text{ have the same support set for finite marginals} \\
&\geq \mathbb{E}_{\mathbb{P}'} \left[\frac{\mathbb{P}^* \{ Y_{1:t} \}}{\mathbb{P}' \{ Y_{1:t} \}} \mathbf{1}_{\widehat{r}_t < r^*} \right] \\
&\quad \text{as } r' < r^* \\
&\geq \mathbb{E}_{\mathbb{P}'} \left[\frac{\mathbb{P}^* \{ Y_{1:t} \}}{\mathbb{P}' \{ Y_{1:t} \}} \mathbf{1}_{A_t(\mathbb{P}')} \right] \\
&\quad \text{from the definition of } B_t(\mathbb{P}') \\
&\geq \mathbb{E}_{\mathbb{P}'} \left[\mathbf{1}_{A_t(\mathbb{P}')} \mathbf{1}_{B_t(\mathbb{P}')} e^{-n[D(\mathbb{P}'|\mathbb{P}^*)+\epsilon]} \right] \\
&\geq \mathbb{E}_{\mathbb{P}'} \left[\mathbf{1}_{A_t(\mathbb{P}')} \mathbf{1}_{B_t(\mathbb{P}')} \right] e^{-n[D(\mathbb{P}'|\mathbb{P}^*)+\epsilon]} \\
&\quad \text{from the union bound, and by the SBM Theorem} \\
&\geq (1 - \alpha - \epsilon) e^{-n[D(\mathbb{P}'|\mathbb{P}^*)+\epsilon]} .
\end{aligned}$$

Now optimizing with respect to θ' and r' , and taking ϵ to zero, the theorem follows. \square

Remark 2.6. Assessing that the upper-bound on under-estimation exponent is positive amounts to check properties of relative entropy rates.

Another Stein-like argument provides an even more clear-cutting statement concerning possible over-estimation exponents. Such a statement seems to be a hallmark of a family of embedded composite testing problems. It shows that in many circumstances of interest, we cannot hope to achieve both non-trivial under- and over-estimation exponents. Versions of this Theorem have been proved in Finesso et al. (1996) for Markov chains, and in Gassiat and Boucheron (to appear) for HMMs.

Theorem 2.8.2. *Let the sequence of nested models $(\mathcal{M}^r)_{r \in \mathbb{N}}$ satisfy Assumption 2.2. Assume also that for $P \in \mathcal{M}_0^r \subseteq \mathcal{M}^r$, there exists a sequence $(P^n)_n$ of elements of $\mathcal{M}_0^{r+1} \setminus \mathcal{M}^r$ such that*

$$\lim_n D_\infty (P^n | P) = 0,$$

Assume that $(\widehat{r}_t)_t$ is a consistent order estimation procedure. Then for all $P \in \mathcal{M}_0^{r^}$,*

$$\liminf_t \frac{1}{t} \log P \{ \widehat{r}_t > r^* \} = 0 .$$

The change of measure argument that proved effective in the proof of Theorem 2.8.1 can now be performed for each $P \in \mathcal{M}_0^r$.

Proof. Let P denote a distribution in $\mathcal{M}_0^{r^*}$. Let (P^n) denote a sequence of distributions from $\mathcal{M}_0^{r^*+1} \setminus \mathcal{M}^{r^*}$ such that

$$\liminf_n D_\infty (P^n | P) = 0.$$

Let ϵ denote a small positive real. For n and t sufficiently large $D_\infty (P^n | P) \leq \epsilon$, and

$$P_t^n \left\{ \frac{1}{t} \ln \frac{dP_t^n}{dP_t} \geq D_\infty (P^n | P) + \epsilon \right\} \leq \epsilon,$$

while

$$P_t^n \{ \widehat{r}_t \geq r^* + 1 \} \geq 1 - \epsilon.$$

We may now lower bound the over-estimation probability.

$$\begin{aligned}
\mathbb{P}\{\hat{r}_t > r\} &\geq \mathbb{P}\{\hat{r}_t = r^* + 1\} \\
&\geq \mathbb{E}_{\mathbb{P}^n} \left[\frac{d\mathbb{P}}{d\mathbb{P}^n} \mathbf{1}_{\hat{r}_t = r^* + 1} \right] \\
&\geq \mathbb{E}_{\mathbb{P}^n} \left[\frac{d\mathbb{P}_t}{d\mathbb{P}_t^n} \mathbf{1}_{\hat{r}_t = r^* + 1} \right] \\
&\geq \mathbb{E}_{\mathbb{P}^n} \left[\exp \left(-\ln \frac{d\mathbb{P}_t^n}{d\mathbb{P}_t} \right) \mathbf{1}_{\hat{r}_t = r^* + 1} \right] \\
&\geq \exp(-2t\epsilon)(1 - 2\epsilon).
\end{aligned}$$

Hence:

$$\lim_t \frac{1}{t} \ln \mathbb{P}_t \{\hat{r}_t > r^*\} \geq -2\epsilon$$

As ϵ may be arbitrarily close to 0. This terminates the proof of Theorem 2.8.2. \square

To check whether the conditions of Theorem 2.8.2 are satisfied in the different order estimation problems described in Section 2.2, we refer to Finesso et al. (1996) and Gassiat and Boucheron (to appear).

The message of this Section is rather straightforward: in order estimation problems like HMM order estimation problem, underestimation corresponds to large deviations of the likelihood process, while overestimation corresponds to moderate deviations of the likelihood process. In the Markov order estimation problem, the large-scale typicality theorem of Csizsár and Shields allows to assign a quantitative meaning to this statement.

2.8.2 The maximum likelihood and mixture estimators achieve optimal underestimation exponent

Stein-like Theorems (Theorem 2.8.1 and 2.8.2) provide a strong incentive to investigate the underestimation exponents of the consistent order estimators that have been described in Section 2.6. As those estimators turn out to be penalized maximum likelihood estimators, what is at stake here, is the (asymptotic) optimality of generalized likelihood ratio testing. In some situations, generalized likelihood ratio testing fails to be optimal. We will show that this is not the case in the order estimation problems we have in mind.

As will become clear from the proof, as soon as the NML normalizing constant $\ln C_t^r/t$ tends toward 0 as t tends toward infinity, NML code-based order estimators exhibit the same property.

Assumption 2.3. 1. *The sequence of models satisfies the AEP.*

2. *Each model \mathcal{M}^r can be endowed with a topology under which it is sequentially compact.*
3. *Relative entropy rates satisfy the semi-continuity property: if $\mathbb{P}^n \rightarrow \mathbb{P}$ and $\mathbb{P}'^n \rightarrow \mathbb{P}'$ then $D_\infty(\mathbb{P} | \mathbb{P}') \leq \liminf^n D_\infty(\mathbb{P}^n | \mathbb{P}'^n)$ (see 2.4.2).*
4. *For any $\epsilon > 0$, any r , there exists a sieve $(\mathbb{P}^i)_{i \in I_\epsilon^r}$, that is, a finite set I_ϵ^r such that $\mathbb{P}^i \in \mathcal{M}^r$, all \mathbb{P}^i are ergodic and a t_ϵ^r , such that:*

a)

$$\forall \mathbb{P} \in \mathcal{M}^r, \exists i \in I_\epsilon^r, \forall t \geq t_\epsilon^r, \forall y_{1:t}, \frac{1}{t} \left| \log P\{y_{1:t}\} - \log P_i\{y_{1:t}\} \right| \leq \epsilon.$$

b) *For each stationary ergodic distribution $\mathbb{P}^* \in \cup_r \mathcal{M}^r$, with order r^* , for every finite subset \mathcal{P} of the union of the collection of all sieves $\mathcal{P} \subseteq \cup_\epsilon \{\mathbb{P}^i : i \in I_\epsilon^r\} \subseteq \mathcal{M}^{r^*}$, the likelihood process $(\log P\{Y_{1:t}\})_{\mathbb{P} \in \mathcal{P}}$ satisfies a large deviation principle with good rate function $J_{\mathcal{P}}$ and*

rate t .

Moreover, any sample path of the log-likelihood process indexed by \mathcal{P} , $(u(P))_{P \in \mathcal{P}}$ that satisfies $J_{\mathcal{P}}(u) < \infty$ enjoys a representation property: there exists a distribution $P_u \in \mathcal{M}^{r^*}$ such that

$$\begin{aligned} \forall P \in \mathcal{P}, u(P) &= \lim \frac{1}{t} E_{P_u} \left[\log P\{Y_{1:t}\} \right] \\ J_{\mathcal{P}}(u) &\geq D_{\infty}(P_u | P^*). \end{aligned}$$

5. For any $r_1 < r_2$, if $P_1 \in \mathcal{M}^{r_1}$ and $P_2 \in \mathcal{M}^{r_2}$ satisfy $D_{\infty}(P_2 | P_1) = 0$ then $P_2 = P_1 \in \mathcal{M}^{r_1}$
6. If $P \in \mathcal{M}^{r^*}$ is not stationary ergodic, it can be represented by a bounded mixture of ergodic components $(P_i)_{i \leq i(r^*)}$ (where $i(r^*)$ depends only on r^*) from \mathcal{M}^{r^*} , $\sum_i \lambda_i P_i = P$ and for all ergodic P' from \mathcal{M} :

$$D_{\infty}(P | P') = \sum_{i \leq i(r^*)} \lambda_i D_{\infty}(P_i | P').$$

Remark 2.7. Assumption 2.3 holds for HMM. This is not obvious at all and follows from available LDP for additive functionals of Markov chains, the extended chain device and ad hoc considerations. The interested reader may find complete proofs and relevant information in Gassiat and Boucheron (to appear).

Theorem 2.8.3. Assume that the sequence of nested models (\mathcal{M}^r) satisfies Assumptions 2.2 and 2.3. If $\text{pen}(t, r)$ is non-negative and for each r , $\text{pen}(t, r)/t \rightarrow 0$ as $t \rightarrow \infty$, the penalized maximum likelihood order estimators achieve the optimal underestimation exponent:

$$\min_{r < r^*} \inf_{P \in \mathcal{M}^r} D_{\infty}(P | P^*).$$

The optimality of this exponent comes from Theorem 2.8.1 which holds under Assumption 2.2. Hence the proof of Theorem 2.8.3 consists in proving the achievability of the exponent.

Proof. An immediate application of the union bound entails that:

$$\limsup \frac{1}{t} \log P^* \{ \hat{r}_t < r^* \} \leq \max_{r < r^*} \limsup \frac{1}{t} \log P^* \{ \hat{r}_t = r \}.$$

Hence the problem reduces to check that for each $r < r^*$,

$$\limsup \frac{1}{t} \log P^* \{ \hat{r}_t = r \} \leq - \inf_{P \in \mathcal{M}^r} D_{\infty}(P | P^*).$$

Let us fix $r < r^*$. The proof will be organized in two steps, we will first check that for each $\epsilon > 0$, we are able to find some $\hat{P}_{\epsilon} \in I'_{\epsilon}$, and some P_{ϵ} such that:

$$\begin{aligned} D_{\infty}(P_{\epsilon} | \hat{P}_{\epsilon}) &\leq 3\epsilon \\ \limsup_t \frac{1}{t} \log P^* \{ \hat{r}_t = r \} &\leq -D_{\infty}(P_{\epsilon} | P^*). \end{aligned}$$

In the second step we take ϵ toward 0 to check that there exists some \bar{P} in \mathcal{M}^r such that

$$\lim_t \frac{1}{t} \log P^* \{ \hat{r}_t = r \} \leq -D_{\infty}(\bar{P} | P^*).$$

Let us also choose $\epsilon > 0$, and t_{ϵ} large enough so that $\text{pen}(t, r^*) \leq \epsilon t$, for $t \geq t_{\epsilon}$.

Under Assumption 2.3 4.a, we get for $t \geq t_\epsilon \vee t_\epsilon^r$:

$$\begin{aligned} \frac{1}{t} \log \mathbf{P}^* \left\{ \widehat{r}_t = r \right\} &\leq \frac{1}{t} \log \mathbf{P}^* \left\{ \sup_{\mathbf{P} \in \mathcal{M}^r} \log \mathbf{P} \{Y_{1:t}\} - \sup_{\mathbf{P} \in \mathcal{M}^{r^*}} \log \mathbf{P} \{Y_{1:t}\} \geq \text{pen}(t, r) - \text{pen}(t, r^*) \right\} \\ &\leq \frac{1}{t} \log \mathbf{P}^* \left\{ \max_{i \in I_\epsilon^r} \frac{1}{t} \log \mathbf{P}_i \{Y_{1:t}\} - \max_{i \in I_\epsilon^{r^*}} \frac{1}{t} \log \mathbf{P}_i \{Y_{1:t}\} \geq -2\epsilon \right\}. \end{aligned}$$

We may take the limsup of the two expressions as t tends toward infinity and use Assumption 2.3 4.b, to get:

$$\limsup \frac{1}{t} \log \mathbf{P}^* \left\{ \widehat{r}_t = r \right\} \leq -\inf \left\{ J_{\mathcal{P}}(u) : \sup_{i \in I_\epsilon^r} u(\mathbf{P}_i) - \sup_{i \in I_\epsilon^{r^*}} u(\mathbf{P}_i) \geq -2\epsilon \right\}$$

with

$$\mathcal{P} \stackrel{\text{def}}{=} \left\{ \mathbf{P}_i : i \in I_\epsilon^r \right\} \cup \left\{ \mathbf{P}_i : i \in I_\epsilon^{r^*} \right\}.$$

The infimum at the right-hand side of the inequality is attained at some path u_ϵ . Hence, using again Assumption 2.3 4.b:

$$\limsup \frac{1}{t} \log \mathbf{P}^* \left\{ \widehat{r}_t = r \right\} \leq -D_\infty(\mathbf{P}_\epsilon | \mathbf{P}^*), \quad (2.19)$$

where $\mathbf{P}_\epsilon \in \mathcal{M}^{r^*}$,

$$u_\epsilon(\mathbf{P}) = \lim \frac{1}{t} \mathbf{E}_{\mathbf{P}_\epsilon} \left[\log \mathbf{P} \{Y_{1:t}\} \right] \quad \text{for any } \mathbf{P} \in \mathcal{P} \quad (2.20)$$

and

$$\sup_{i \in I_\epsilon^r} u_\epsilon(\mathbf{P}_i) - \sup_{i \in I_\epsilon^{r^*}} u_\epsilon(\mathbf{P}_i) \geq -2\epsilon \quad (2.21)$$

We would have almost completed our first step if we could assume that \mathbf{P}_ϵ is ergodic. We actually do not need to make such an assumption since it is enough to approximate \mathbf{P}_ϵ by an element $\tilde{\mathbf{P}}_\epsilon$ in the sieve $I_\epsilon^{r^*}$ which is necessarily ergodic. Let us thus pick $\tilde{\mathbf{P}}_\epsilon \in \{\mathbf{P}_i, i \in I_\epsilon^{r^*}\}$ such that for $t \geq t_\epsilon^r$

$$\frac{1}{t} \left| \log \tilde{\mathbf{P}}_\epsilon \{y_{1:t}\} - \log \mathbf{P}_\epsilon \{y_{1:t}\} \right| \leq \epsilon$$

and $\widehat{\mathbf{P}}_\epsilon$ such that

$$\sup_{i \in I_\epsilon^r} u_\epsilon(\mathbf{P}_i^r) = u_\epsilon(\widehat{\mathbf{P}}_\epsilon). \quad (2.22)$$

One has

$$\begin{aligned} \limsup \frac{1}{t} \log \mathbf{P}_\epsilon \{Y_{1:t}\} &\leq \limsup \frac{1}{t} \log \tilde{\mathbf{P}}_\epsilon \{Y_{1:t}\} + \epsilon \\ &= u_\epsilon(\tilde{\mathbf{P}}_\epsilon) + \epsilon && \text{using (2.20)} \\ &\leq u_\epsilon(\widehat{\mathbf{P}}_\epsilon) + 3\epsilon && \text{using (2.22) and (2.21)} \\ &\leq \limsup \frac{1}{t} \log \widehat{\mathbf{P}}_\epsilon \{Y_{1:t}\} + 3\epsilon && \text{using (2.20) again.} \end{aligned}$$

Using Assumption 2.3.1, one thus finally obtains:

$$D_\infty(\mathbf{P}_\epsilon | \widehat{\mathbf{P}}_\epsilon) \leq 3\epsilon.$$

Let us now proceed to the second step. It remains to check, that if we take ϵ toward 0, the sequence $(\mathbf{P}_\epsilon)_\epsilon$ obtained in (2.19) has an accumulation point lying in \mathcal{M} .

Note that $\widehat{\mathbf{P}}_\epsilon$ is ergodic, and let $\sum_i \lambda_{i,\epsilon} \mathbf{P}_{i,\epsilon}$ denote the ergodic decomposition of \mathbf{P}_ϵ . Extract a subsequence of $(\lambda_{i,\epsilon})$ and $(\mathbf{P}_{i,\epsilon})$ converging to (λ_i) and (\mathbf{P}_i) such that $\bar{\mathbf{P}} = \sum_i \lambda_i \mathbf{P}_i$, while $\widehat{\mathbf{P}}$ is the corresponding accumulation point of the sequence $\widehat{\mathbf{P}}_\epsilon$. One has

$$D_\infty(\mathbf{P}_\epsilon | \widehat{\mathbf{P}}_\epsilon) = \sum_i \lambda_{i,\epsilon} D_\infty(\mathbf{P}_{i,\epsilon} | \widehat{\mathbf{P}}_\epsilon)$$

we may then apply the semi-continuity property (see Section 2.4.2) to obtain

$$\sum_i \lambda_i D_\infty(\mathbf{P}_i | \widehat{\mathbf{P}}) = 0,$$

which leads, using Assumption 2.3 6.) to

$$\sum_i \lambda_i \mathbf{P}_i = \widehat{\mathbf{P}},$$

that is $\bar{\mathbf{P}} = \widehat{\mathbf{P}} \in \mathcal{M}^r$. Using again the semi-continuity property,

$$\begin{aligned} \lim_\epsilon D_\infty(\mathbf{P}_\epsilon | \mathbf{P}^*) &= \lim_\epsilon \sum_i \lambda_{i,\epsilon} D_\infty(\mathbf{P}_{i,\epsilon} | \mathbf{P}^*) \\ &\geq D_\infty(\bar{\mathbf{P}} | \mathbf{P}^*), \end{aligned}$$

so that we get

$$\limsup \frac{1}{t} \mathbf{P}^* \{\widehat{r}_t = r\} \leq - \inf_{\mathbf{P} \in \mathcal{M}^r} D_\infty(\mathbf{P} | \mathbf{P}^*).$$

□

2.9 Bibliographical remarks

The order estimation problem for HMMs and Markov processes became an active topic in the Information Theory literature in the late nineteen-eighties. Early references can be found in Finesso (1991); Ziv and Merhav (1992). Other versions of the order estimation problem, most notably order estimation in mixture models, had been tackled even earlier, see Henna (1985); Haughton (1988). We refer to Chapter 7 in Chambaz (2003) for a brief history of order identification. Useful references on order estimation problems in mixture modeling can be found in Kéribin and Gassiat (2000). An early discussion of order estimation issues in ARMA modeling is presented in Azencott and Dacunha-Castelle (1984). Model selection in ARMA modeling is further discussed in Gerencsér (1994).

The definition of HMM order used in this Chapter is classical. A general discussion concerning HMM order and related notions like rank can be found in Finesso (1991). Finesso credits Azencott and Dacunha-Castelle (1984) for major influence on his work on Markov order estimation, see Finesso (1991). The connections between the performance of generalized likelihood ratio testing and the behavior of maximum likelihood ratios was outlined in Finesso (1991). Using the law of iterated logarithms for the empirical measure of Markov chains in order to identify small penalties warranting consistency in Markov order estimation also goes back to Finesso (1991)

The connections between order estimation and hypothesis testing has been emphasized in the work of Merhav and collaborators Zeitouni and Gutman (1991); Zeitouni et al. (1992); Ziv and Merhav (1992); Feder and Merhav (2002). Those papers present various settings for composite hypothesis testing in which generalized likelihood ratio testing may or may not be asymptotically optimal. Fast introductions to Bahadur efficiency and Stein's Lemma and their connections to Large Deviations theory can be found in Dembo and Zeitouni (1999); Cover and Thomas (1991); van der Vaart (1998). An elementary and self-contained account of nested composite testing in the independent case is presented in Csiszár (1998).

Though the use of universal coding arguments in order identification is already present in Finesso (1991); Zeitouni and Gutman (1991); Ziv and Merhav (1992). Kieffer (1993) provides the most striking exposition of the connections between order identification and universal coding. Versions of Lemmas 2.6.1 and 2.6.2 are at least serendipitous in Kieffer (1993). Results of Section 2.6 can be regarded as elaboration of the ideas exposed in Kieffer (1993).

Normalized Maximum Likelihood codes were introduced in Shtarkov (1987). The relevance of Dirichlet mixtures to universal coding has been pointed out in Krichevsky and Trofimov (1981) and Davisson et al. (1981). A thorough survey about maximinity and minimaxity of several universal codes can be found in Catoni (2001). Recent refined results concerning the asymptotic behavior of Dirichlet mixtures can be found in Yang and Barron (1998); Xie and Barron (2000). MDL was introduced in Rissanen (1978) and received deserved attention during the last twenty-five years Rissanen (1981, 1983, 1984, 1986); barron et al. (1998); Gruenwald (2000).

The role of information divergence rates in order estimation problems was underlined in Kieffer (1993) and Liu and Narayan (1994). Gentle introductions to the AEP can be found in Cover and Thomas (1991); Shields (1996). The generalized AEP which is used to characterize the asymptotic behavior of log-likelihood ratios is presented in Barron (1985b). Alternate versions of Lemmas 2.4.4 and Lemma 2.4.5 can be found in Kieffer (1993). Lemma 2.4.6 was used in Gassiat and Boucheron (to appear).

The almost-everywhere asymptotic consistency of MDL order estimation in Bayesian settings was pointed out in Barron (1985a). The fact that consistency of MDL does not hold everywhere in Markov order estimation was proved in Csiszár and Shields (2000).

Section ?? was mainly inspired by Kieffer (1993) and Gassiat and Boucheron (to appear). The proof of the first inequality in Lemma 2.6.3 goes back to Shtarkov (1987). The proof of the second inequality for HMMs goes back to Csiszár (1990). Variants of the result have been used in Finesso (1991); Liu and Narayan (1994).

Section 2.7 is mainly borrowed from Csiszár (2002) although the results presented here were already contained in Csiszár and Shields (2000) but justified with a different proof. The use of non-asymptotic tail inequalities (concentration inequalities) during the analysis of model selection procedure has become a standard approach in modern statistics (see Bartlett et al. (2002) and references therein for more examples on this topic).

Section 2.8 is largely inspired from Gassiat and Boucheron (to appear), further results in this direction can be found in Chambaz (2003). Early results concerning error exponents in the Markov order estimation problems can be found in Finesso et al. (1996). Recent results concerning order estimation in HMMs can be found in Khudanpur and Narayan (2002).

Bibliography

- R. Azencott and D. Dacunha-Castelle. *Séries d'observations irrégulières*. Masson, 1984.
- A. Barron. *Logically smooth density estimation*. PhD thesis, Stanford, 1985a.
- A. Barron. The strong ergodic theorem for densities; generalized Shannon-McMillan-Breiman theorem. *Annals of Probability*, 13:1292–1303, 1985b.
- A. Barron, J. Rissanen, and B. Yu. The Minimum Description Length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, 1998.
- P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.
- O. Cappé, C. Robert, and T. Rydén. Reversible jump, birth-and-death and more general continuous time MCMC samplers. Submitted, 2001.
- O. Catoni. *Ecole de Probabilités de Saint-Flour, 2001*, chapter Statistical learning theory and stochastic optimization. LNM. Springer-Verlag, 2001.
- A. Chambaz. *Segmentation spatiale et sélection de modèle*. PhD thesis, Université Paris-Sud, Mathématiques, 2003.
- T. Cover and J. Thomas. *Elements of information theory*. John Wiley, 1991.
- I. Csiszár. Class notes on information theory and statistics. University of Maryland, 1990.
- I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44:2505–2523, 1998.
- I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48:1616–1628, 2002.
- I. Csiszár and P. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28:1601–1619, 2000.
- D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques*, volume 2. Masson, 1983.
- D. Dacunha-Castelle and E. Gassiat. The estimation of the order of a mixture model. *Bernoulli*, 3:279–299, 1997a.
- D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models and application to mixture models. *ESAIM Probability & Statistics*, 1:285–317, 1997b.
- D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *Annals of Statistics*, 27:1178–1209, 1999.
- L.D. Davisson, R.J. McEliece, M.B. Pursley, and M.S. Wallace. Efficient universal noiseless source codes. *IEEE Trans. Inform. Theory*, 27:269–279, 1981.

- A. Dembo and O. Zeitouni. *Large deviations*. Springer-Verlag, 1999.
- P. Dupuis and R.S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics, 1997.
- R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson. *Biological Sequence Analysis*. Cambridge University Press, 1998.
- M. Feder and N. Merhav. Universal composite hypothesis testing: a competitive minimax and its applications. *IEEE Trans. Inform. Theory*, 48:1504–1517, 2002.
- L. Finesso. *Consistent estimation of the order for Markov and hidden Markov Chains*. PhD thesis, Maryland University, 1991.
- L. Finesso, C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42:1488–1497, 1996.
- P. Flajolet and W. Szpankowski. Analytic variations on redundancy rates of renewal processes. *IEEE Trans. Inform. Theory*, 48(11):2911–2921, 2002.
- E. Gassiat. Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Inst. Henri Poincaré*, 38:887–906, 2002.
- E. Gassiat and S. Boucheron. Optimal error exponents in hidden markov model order estimation. *IEEE Trans. Inform. Theory*, to appear.
- L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes. *J. Stat. Plann. Inference*, 41(3):303–325, 1994.
- P. Gruenwald. Model selection based on minimum description length. *J. Math. Psychol.*, 44(1):133–152, 2000.
- D.M. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355, 1988.
- J. Henna. On estimating the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Stat. Math.*, 37:235–240, 1985.
- P. Jacquet and W. Szpankowski. A combinatorial problem arising in information theory: precise minimax redundancy for markov sources. In *Colloquium on Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probabilities*, page MISSING. Birkhauser, 2002.
- C. Kéribin and E. Gassiat. The likelihood ratio test for the number of components in a mixture with markov regime. *ESAIM Probability and Statistics*, 4:25–52, 2000.
- S. Khudanpur and P. Narayan. Order estimation for a special class of hidden Markov sources and binary renewal processes. *IEEE Trans. Inform. Theory*, 48:1704–1713, 2002.
- J.C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, 39:893–902, 1993.
- R.E. Krichevsky and V.K. Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, 27:199–207, 1981.
- C. Liu and P. Narayan. Order estimation and sequential universal data compression of a hidden markov source by the method of mixtures. *IEEE Trans. Inform. Theory*, 40:1167–1180, 1994.
- J. Neveu. *Discrete-time martingales*. North-Holland, 1975.
- J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

- J. Rissanen. Order estimation in Box-Jenkins model for time series. *Methods Oper. Res.*, 44:143–150, 1981.
- J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29:656–664, 1983.
- J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, 30:629–636, 1984.
- J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.
- C. P. Robert and M. Titterton. Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics & Computing*, 8(2):145–158, 1998.
- C.P. Robert, T. Rydén, and M. Titterton. Bayesian inference in hidden Markov models through reversible jump Markov chain Monte Carlo. *J. Royal Statist. Soc. Ser. B*, 62:57–75, 2000.
- P. Shields. Universal redundancy rates do not exist. *IEEE Trans. Inform. Theory*, 39:520–524, 1993.
- P. Shields. *The ergodic theory of discrete sample paths*. AMS, 1996.
- Y.M. Shtarkov. Universal sequential coding of messages. *Probl. Inform. Transmission*, 23:3–17, 1987.
- A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.
- Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*, 46:431–445, 2000.
- Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, 44:95–116, 1998.
- O. Zeitouni and M. Gutman. On universal hypothesis testing via large deviations. *IEEE Trans. Inform. Theory*, 37:285–290, 1991.
- O. Zeitouni, J. Ziv, and N. Merhav. When is generalized likelihood ratio test optimal? *IEEE Trans. Inform. Theory*, 38:1597–1602, 1992.
- J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Trans. Inform. Theory*, 38:61–65, 1992.