Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Matematica Pura ed Applicata

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE MATEMATICHE

INDIRIZZO DI MATEMATICA COMPUTAZIONALE

XXII CICLO

PARTIAL EXCHANGEABILITY AND CHANGE DETECTION FOR HIDDEN MARKOV MODELS

Direttore della Scuola: Prof. Paolo Dai Pra

Supervisori: Dr. Lorenzo Finesso
Prof. László Gerencsér
Prof. Wolfgang Runggaldier

Dottoranda: Cecilia Prosdocimi

alle mie Assenze

Sunto

La tesi affronta lo studio dei modelli di Markov nascosti. Essi sono oggi giorno molto popolari, in quanto presentano una struttura più versatile dei processi indipendenti ed identicamente distribuiti o delle catene di Markov, ma sono tuttavia trattabili. Risulta quindi interessante cercare proprietà dei processi i.i.d. che restano valide per modelli di Markov nascosti, ed è questo l'oggetto della tesi. Nella prima parte trattiamo un problema probabilistico. In particolare ci concentriamo sui processi scambiabili e parzialmente scambiabili, trovando delle condizioni che li rendono realizzabili come processi di Markov nascosti. Per una classe particolare di processi scambiabili binari troviamo anche un algoritmo di realizzazione. Nella seconda parte affrontiamo il problema del rilevamento di un cambiamento nei parametri caratterizzanti la dinamica di un modello di Markov nascosto. Adattiamo ai modelli di Markov nascosti un algoritmo di tipo *cumulative sum* (CUSUM), introdotto inizialmente per osservazioni i.i.d. Questo ci porta a studiare la statistica CUSUM con processo di entrata L-mixing. Troviamo quindi una proprietà di perdita di memoria della statistica CUSUM, quando non ci sono cambiamenti nella triettoria, dapprima nel caso più elemenatare di processo di entrata i.i.d. (con media negativa e momenti esponenziali di qualche ordine finiti), e poi per processo di entrata L-mixing e limitato, sotto opportune ipotesi tecniche.

La rimanente parte di questo testo è redatta in lingua inglese per consentirne la fruibilità ad un numero maggiore di lettori.

Abstract

The thesis focuses on Hidden Markov Models (HMMs).They are very popular models, because they have a more versatile structure than independent identically distributed sequences or Markov chains, but they are still tractable. It is thus of interest to look for properties of i.i.d. sequences that hold true also for HHMs, and this is the object of the thesis. In the first part we concentrate on a probabilistic problem. In particular we focus on exchangeable and partially exchangeable sequences, and we find conditions to realize them as HHMs. For a special class of binary exchangeable sequences we also give a realization algorithm. In the second part we consider the problem of detecting changes in the statistical pattern of a hidden Markov process. Adapting to HHMs the so-called cumulative sum (CUSUM) algorithm, first introduced for independent observations, we are led to the study of the CUSUM statistics with L-mixing input sequence. We establish a loss of memory property of the CUSUM statistics when there is no change, first in the easier case of a i.i.d. input sequence, (with negative expectation, and finite exponential moments of some positive order), and then, under some technical conditions, for bounded and L-mixing input sequence.

Introduction

A Hidden Markov Model (HMM) is a function of a homogeneous Markov chain.
In general a HMM needs not be Markov and will exhibit long-range dependencies of some kind. This theoretical inconvenience is actually a blessing in disguise. The class of HMMs contains processes with complex dynamical behaviors and yet it admits a simple parametric description, therefore it comes as no surprise that it is extensively employed in many applications to real data. HMMs appear in such diverse fields as engineering (to model the output of stochastic automata, for automatic speech recognition and for communication networks), genetics (sequence analysis), medicine (to study neuro-transmission through ion-channels), mathematical finance (to model rating transitions, or to solve asset allocation problem), and many others.

On the theoretical side the lack of the Markov property makes the class of HMMs difficult to work with. Theoretical work on the specific class of HMMs has proceeded along two main lines, probabilistic and statistical.

Probabilistic aspects. The early contributions, inspired by the seminal papers of Blackwell and Koopmans [11], and of Gilbert [35], concentrated on the probabilistic aspects. During the sixties many authors looked for a characterization of HMMs. More specifically the problem analyzed was: *among all processes Y_n characterize, in terms of their finite dimensional distributions, those that admit a HMM representation.* This problem produced a host of false starts and partial solutions until it was finally settled by Heller [36]. To some extent Heller's result is not quite satisfactory since his methods are non-constructive. If (Y_n) is a stochastic process which satisfies the conditions to be represented as a HMM, no algorithm is proposed to produce a Markov chain (X_n) and a function f such that $Y_n = f(X_n)$. That explains why the effort to produce a constructive theory of HMMs has been pursued even after the publication of Heller's paper. In more recent years the problem has attracted the attention of workers in the area of Stochastic Realization Theory (see for example [2], [51]) but, while some of the issues have been clarified, a constructive and effective algorithm valid for general HMMs is still missing.

A niche of positive results was obtained restricting attention to special subclasses of HMMs which exhibit a particularly simple, yet probabilistically and statistically interesting structure.

Most notably, in 1964, Dharmadhikari [21] restricted his attention to the class of *exchangeable processes*, posing and solving the problem of characterizing the HMMs within that class. Part I of the thesis goes in this direction.

Exchangeable processes have been introduced by de Finetti. They are characterized by having joint distributions invariant under permutations between the random variables in the sequence. The famous theorem by de Finetti himself states that a sequence is exchangeable if and only if it is a mixture of independent identically distributed (i.i.d.) sequences.

Mixtures of i.i.d. sequences are i.i.d. sequences with an unknown and random distribution, for which we select a prior. The role played by mixture models in Bayesian statistics is well known and sufficient in itself to justify the large interest, both theoretical and practical, for the class of exchangeable processes.

If the prior of the unknown distribution of the random variables of a mixture of i.i.d. sequences is concentrated on a finite (resp. countable) set we will call it a finite (resp. countable) mixture of i.i.d. sequences. The above result by Dharmadhikari states that an exchangeable sequence of random variables is a finite (resp. countable) mixture of i.i.d. sequences if and only if it is a HHM with finite (resp. countable)-valued underlying Markov chain. The class of exchangeable HMMs with finite underlying Markov chain therefore *coincides* with that of finite mixtures of i.i.d. sequences. Although these are very special exchangeable processes, they constitute an extremely versatile class of models, usefully employed in many practical applications.

It is therefore not only of theoretical, but also of practical interest to pose and solve the problem of the construction of exchangeable HMMs from distributional data, which is the first contribution we give in Part I of the thesis. In the first part of Chapter 2 we focus on $\{0,1\}$-valued exchangeable processes. We look for verifiable criteria to establish, from the knowledge of some finite distributions of the process of interest, whether it is a finite mixtures of i.i.d. sequences, i.e. if it is a HHM with finite parametrization.

Moreover we can find the finite parametrization, when it does exist.

The problems mentioned above can be completely solved resorting to known results on the classic moment problem on the unit interval. Both the characterization of the existence and the construction of the realization can be based on the analysis of a set of Hankel matrices which are defined solely in terms of the given finite distributions.

The idea of exchangeability has ramified in many directions giving origin to a number of closely related notions of *partial exchangeability*. Without entering the intricate details which will be developed later, one can define a partially exchangeable process as one whose distributions are invariant for permutations that preserve the 1-step transition counts. de Finetti conjectures that partially exchangeable sequences are mixtures of Markov chains, i.e. Markov chains with an unknown and random distribution. His conjecture was proved many years later, under some necessary regularity conditions, by Diaconis and Freedman [22] in 1980, more than 20 years after de Finetti conjectured the result. More refined versions of this result, settling issues related to the uniqueness of the representation, have been given only recently by Fortini and al. [27].

The study of partially exchangeable processes leads us to the other two main contributions of Part I of the thesis. On one hand we have generalized the above mentioned Dharmadhikari theorem, proving that a recurrent partially exchangeable process is a finite (resp. countable) mixture of Markov chains if and only if it is a HHM with finite (resp. countable)-valued underlying Markov chain. This has been proved both in the framework of [22] and of [27]. Moreover we give some advances in the realization problem for $\{0, 1\}$-valued HMMs. We focus on partially exchangeable processes and we propose a criterion to establish whether they can be realized as a finite mixture of Markov chains, i.e. as a HMM with finite parametrization.

We now go back to the statistical aspects of HMMs.

Statistical aspects.

One of the first contributions in the statistical analysis of hidden Markov processes is [8], in which the maximum-likelihood (ML) estimation of the parameters of a finite state space and finite read-out HMM is studied. Strong consistency of the maximum-likelihood estimator for finite state space and binary read-outs has been established in [3]. The extension of these results to continuous read-outs requires new insights.

The first step in proving consistency of the maximum-likelihood method would be to show the validity of the strong law of large numbers for the log-likelihood function. This can be achieved showing the validity of the strong law of large numbers for a function of an extended Markov chain (X_n, Y_n, p_n), where (X_n) is the Markov chain driving the HHM, (Y_n) is the output and (p_n) is the predictive filter. This has been investigated in the literature basically with three different methods: using the subadditive ergodic theorem in [40], using geometric ergodicity arguments in [39], and using L-mixing processes in [34], [33]. For more recent contributions see [14] and [24].

In Part II of the thesis we focus on the statistical problem of change detection for HMM-s. Detection of changes in the statistical pattern of a hidden Markov process is of interest in a number of applications. In the case of speech processing we may wish to identify the moments of switching from one speaker to another one. In the case of distributed, aggregated sensing we may wish to identify the events of sensor failure.

The theory for the statistical analysis of HHMs developed by Gerencser and coauthors allows the construction of a Hinkley-type (or CUSUM) change detection algorithm for HHMs. The proposed change detection algorithm for HHMs naturally leads to the study of the CUSUM statistics with L-mixing input. We concentrate on the case when there are no changes at all. In this case the CUSUM statistics can be represented as a non linear stochastic system. For i.i.d. input, this system is a standard object in queuing theory and many stability properties have been proved for the system, with negative expectation input. In Part II of the thesis these results are extended in two ways: first we show that for i.i.d. inputs with negative expectation, and finite exponential moments of some positive order the output of this system is L-mixing. Then, this result is generalized to L-mixing inputs, under further technical conditions such as boundedness. The results give an upper bound for the false alarm frequency.

Acknowledgments

I would like to thank my supervisors for all they have taught me about mathematics and about life, my parents for the love of knowledge they have passed on to me, my sisters for sharing with me this love, Tommaso for supporting all my choices, Raffaele, Nicola, Francesca, Andrea, Eleonora, Alice, all the Ph.D. students and people at the Department for making every working-day a pleasant day.

Contents

Countable and Finite Mixtures of Markov Chains

Input-output properties of the Hinkley detector

Countable and Finite Mixtures of Markov Chains

Introduction

Exchangeable sequences, introduced by de Finetti, are characterized by having joint distributions invariant under permutations between the random variables in the sequence. Exchangeable sequences are identically distributed random variables which are independent, up to know some common random factor of the variables in the sequence, or, in an other way, they are independent identically distributed (i.i.d.) sequences with an unknown and random distribution (*mixture of i.i.d. sequences*). de Finetti is the first who characterizes exchangeable sequences in such a way, proving his well known theorem.

de Finetti himself generalizes exchangeability to *partial exchangeability*. Roughly speaking, a sequence is partially exchangeable if the joint distributions are invariant under permutations which keep fixed the 1-step transitions. de Finetti conjectures that partially exchangeable sequences are mixtures of Markov chains, i.e. Markov chains with an unknown and random distribution. His conjecture was proved many years later.

Dharmadhikari gives a characterization of exchangeable sequences that are a *countable* mixture of i.i.d. sequences, linking exchangeable sequences with HHMs. In the thesis we extend the result of Dharmadhikari, characterizing partially exchangeable sequences which are *countable* mixture of Markov chains, finding a connection with HHMs. The result easily implies a characterization of *finite* mixtures of Markov chains, that is actually quite difficult to check in practice.

Finite mixtures of Markov chains are an appropriate statistical model when the population is naturally divided into clusters, and the time evolution of a sample is Markovian, but with distribution dependent on which cluster the sample belongs to. These models have been applied in different contexts. For example in [13] finite mixtures of Markov chains are used to model navigation patterns on a web site, clustering together users with similar behavior. In [28] bond ratings migration is modeled using a mixture of Markov chains, where different Markov chains correspond to different health states of the market.

In Chapter 1 we introduce the basic definitions and tools we will need in the thesis.

In Chapter 2 we focus on *binary*, i.e. $\{0,1\}$-valued, sequences. We provide an original criterion to establish whether a mixture of i.i.d. sequences or of Markov chains is a *finite* mixture. Differently from the Dharmadhikari characterization, the criterion is easy to check. Moreover, for finite mixtures of i.i.d. sequences it gives the exact number of the i.i.d. sequences involved, while for finite mixtures of Markov chains it gives just a bound on the number of components of the mixture. Furthermore we give an original algorithm to compute exactly the mixing measure associated with a finite mixture of i.i.d. sequences, and thus to completely identify the model. The developed theory can be applied to solve the stochastic realization problem and the positive realization of a linear system in some special cases.

In Chapter 3 we extend the Dharmadhikari theorem characterizing the exchangeable sequences which are *countable* mixtures of i.i.d. sequences, to Markov exchangeable sequences, to k-Markov exchangeable sequences and to partially exchangeable sequences.

Chapter 1

Mathematical tools

In this chapter we recall some well known definitions and results. In Section 1.1 we introduce exchangeable and partially exchangeable sequences. In Section 1.2 we give the definition of mixture of i.i.d. sequences and of Markov chains (see Section 1.2.1), with special attention to finite and countable mixtures (see Section 1.2.2), and to binary mixtures (see Section 1.2.3), because they will play a crucial role in Chapters 2 and 3. Moreover in Section 1.2.4 we report de Finetti's characterization theorem for exchangeable sequences and its extension to partially exchangeable sequences by Diaconis and Freedman, and by Fortini *et al.*, providing also an easy original extension of the result to k-Markov exchangeable sequences, that will be proved in Chapter 3. In Section 1.3 we recall the definition of Hidden Markov Model.

1.1 Exchangeability and related notions

Let $Z = (Z_1, Z_2, \dots)$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in a measurable space $(J, \mathcal{J})$, with J countable. The set J will be called alphabet and its elements symbols or letters. Finite strings of letters will be denoted by $\sigma = \sigma_1^n = \sigma_1\sigma_2 \dots \sigma_n$, where $\sigma_i \in J$ for $i = 1, \dots n$. J^* will be the set of all finite strings of letters and we will write $\{Z_1^n = \sigma_1^n\}$ to indicate the event $\{Z_1 = \sigma_1, Z_2 = \sigma_2, \dots, Z_n = \sigma_n\}$.

Let $S(n)$ be the group of permutations of $\{1, 2, \dots, n\}$. We say that the string $\tau = \tau_1^n$ is a *permutation* of the string $\sigma = \sigma_1^n$ if there exists a permutation $\pi \in S(n)$ such that

$$\tau_1^n = \tau_1\tau_2 \dots \tau_n = \sigma_{\pi(1)}\sigma_{\pi(2)} \dots \sigma_{\pi(n)}.$$

Definition 1.1.1. *The sequence* $Z = (Z_1, Z_2, \dots)$ *is* exchangeable *if for all* n*, for all* $\sigma = \sigma_1^n$ *and all permutation* $\tau = \tau_1^n$ *of* σ *we have*

$$\mathbb{P}\{Z_1^n = \sigma_1^n\} = \mathbb{P}\{Z_1^n = \tau_1^n\}.$$

Example 1.1.1. *Let* $\delta = (\delta_1, \delta_2, \dots)$ *be a sequence of i.i.d. random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. Let* $\bar{Z}$ *be a random variable on the same probability space, not necessarily independent from* δ*. Define for any* $n \geq 1$

$$Z_n := \bar{Z} + \delta_n. \tag{1.1}$$

The sequence $Z = (Z_1, Z_2, \dots)$ *is exchangeable.*

Note that a sequence of i.i.d random variables is exchangeable. Conversely the random variables of an exchangeable sequence are identically distributed, but not necessarily independent.

We introduce below the *partially exchangeable* sequences. The term partial exchangeability could cause some misunderstanding, because it is used by different authors to indicate slightly different classes of processes [1]. The first definition of partial exchangeability goes back to de Finetti in [19]. It is reported in Definition 1.1.2 below. Other authors, among them Diaconis and Freedman ([22]) and Quintana ([47]), use the term "partial exchangeability" for a different class of processes, that we will call "Markov exchangeable" (see Definition 1.1.4 below).

[1]For the sake of clarity we list below some terms commonly used in the literature, to indicate notions connected with de Finetti's partial exchangeability

- *partial exchangeability*:
  - Diaconis and Freedman in [22], and Quintana in [47] use this term to indicate processes satisfying Definition 1.1.4
  - de Finetti in [19], Fortini *et al.* in [27] and this Thesis use this term to indicate processes with successors matrix satisfying Definition 1.1.2
- *Markov exchangeability*
  - Di Cecco in [20] and this Thesis use this term to indicate processes satisfying Definition 1.1.4
- *Freedman condition*
  - Fortini *et al.* in [27] use this term to indicate the condition required in Definition 1.1.4

Let $Z = (Z_1, Z_2, \dots)$ be a sequence of random variables. To give Definition 1.1.2, we introduce the random matrix $V = (V_{i,n})_{i \in J, n \geq 1}$ of the successors of Z: for any $i \in J$ and $n \geq 1$, $V_{i,n}$ is the value that the sequence Z takes immediately after the n-th visit to state i. To avoid rows of finite length for V we introduce an additional state to J, call it $\delta \notin J$. If X visits the state i only n times, put $V_{i,m} = \delta$ for $m > n$.

Definition 1.1.2. *V is* partially exchangeable *if its distribution is invariant under finite permutations possibly distinct within each of its rows.*

Markov exchangeability is an extension of the notion of exchangeability, given restricting the class of permitted permutations. More precisely

Definition 1.1.3. *Let $\sigma = \sigma_1 \sigma_2 \dots \sigma_n$ and $\tau = \tau_1 \tau_2 \dots \tau_n$ be finite strings from J. σ and τ are* transition equivalent*, write $\sigma \sim \tau$, if they start with the same letter and exhibit the same number of transitions from letter i to letter j, for every pair $i, j \in J$.*

For example, in the binary case, $J = \{0, 1\}$, the strings $\sigma = 1100101$ and $\tau = 1101001$ are easily seen to be transition equivalent. Note that the string $\rho = 1110001$ is a permutation of σ, but σ and ρ are not transition equivalent. For a systematic way to generate, in the binary case, all the strings transition equivalent to a given one see [47].

We recast below the Lemma (5) of [22]:

Lemma 1.1.1. *Let $\sigma \sim \tau$. Then σ and τ have the same length and end with the same letter. Furthermore, for any letter i, the sequences σ and τ contain the letter i the same number of times.*

We are now ready to give the definition of *Markov exchangeable* sequences:

Definition 1.1.4. *The sequence $Z = (Z_1, Z_2, \dots)$ is* Markov exchangeable *if for every pair of transition equivalent strings $\sigma \sim \tau$, it holds*

$$\mathbb{P}\{Z_1^n = \sigma_1^n\} = \mathbb{P}\{Z_1^n = \tau_1^n\}.$$

The sequence Z is Markov exchangeable if, for all $n \in \mathbb{N}$, the joint distribution of $Z_1, Z_2, \dots, Z_n$ is invariant under permutations that keep fixed the transition counts and the initial state.

Example 1.1.2. An exchangeable sequence is trivially Markov exchangeable according to Definition 1.1.4, since the finite-dimensional joint distributions of an exchangeable sequence must be invariant under a larger class of permutations. By Lemma 1 part (b) in [27] an exchangeable sequence has also a partially exchangeable matrix of the successors.

Example 1.1.3. An homogeneous Markov chain is partially exchangeable and Markov exchangeable.

Remark 1.1.1. Note that exchangeable sequences are stationary, but Markov exchangeable sequences are not always stationary. For example, a homogeneous Markov chain which does not start at the invariant distribution is Markov exchangeable, but not stationary.

In [27] Section 2.2. Fortini et al. study the relationship between partial exchangeability and Markov exchangeability, proving that the former is generally stronger, but that the two are equivalent under a recurrence hypothesis.

Markov exchangeability can be generalized to k-Markov exchangeability. First of all, we restrict the notion of transition equivalent strings:

Definition 1.1.5. *Let $\sigma = \sigma_1 \sigma_2 \dots \sigma_n$ and $\tau = \tau_1 \tau_2 \dots \tau_n$ be strings from J. We say that σ and τ are* transition equivalent of order k *(shortly, k-transition equivalent), and we write $\sigma \sim_k \tau$, if they start with the same string of length k and exhibit the same number of transition counts of order k.*

The analogous of Lemma 1.1.1 holds, i.e.

Lemma 1.1.2. *Let $\sigma \sim_k \tau$. Then σ and τ have the same length, and end with the same string of length k. Furthermore σ and τ contain the k-length strings $i_1 i_2 \dots i_k$, with $i_1, i_2, \dots, i_k$ varying in J, the same number of times.*

We can give the definition of Markov exchangeable sequences of order k in the obvious way:

Definition 1.1.6. *The sequence Z is* Markov exchangeable of order k *(shortly, k-Markov exchangeable), if for every pair of finite k-transition equivalent strings $\sigma \sim_k \tau$, with $\sigma = \sigma_1^n$ and $\tau = \tau_1^n$, it holds*

$$\mathbb{P}\{Z_1^n = \sigma_1^n\} = \mathbb{P}\{Z_1^n = \tau_1^n\}.$$

Trivially a k-Markov exchangeable sequence is h-Markov exchangeable for $k \leq h$.

Recall the well-known definition

Definition 1.1.7. *Z is a* homogeneus Markov chain of order k *if*

$$\mathbb{P}\{Z_n = z_n | Z_1, \dots, Z_{n-1}\} = \mathbb{P}\{Z_n = z_n | Z_{n-k}, \dots, Z_{n-1}\} \quad \textit{for each } n \geq k+1.$$

An i.i.d. sequence is a Markov chain of order 0, a usual Markov chain is a Markov chain of order 1.

Example 1.1.4. *A homogeneous Markov chain of order k is a k-Markov exchangeable sequence.*

1.2 Mixtures and representation theorems

1.2.1 Mixtures of measures

In this section we give the definition of mixtures of i.i.d. sequences and of mixtures of Markov chains.

Mixtures of i.i.d. sequences

To define mixtures of i.i.d. sequences we need the following

Definition 1.2.1. *(see Chapter 7.2 in [15]) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ-field of $\mathcal{F}$. A* regular conditional probability *on $\mathcal{F}$ given $\mathcal{G}$ is a function $P : \Omega \times \mathcal{F} \longrightarrow [0,1]$ such that*

- *$P(\omega, \cdot) : \mathcal{F} \longrightarrow [0,1]$ is a probability measure on $\mathcal{F}$ for all $\omega \in \Omega$,*
- *$P(\cdot, A)$ is $\mathcal{G}$-measurable for any $A \in \mathcal{F}$ and $P(\omega, A) = \mathbb{P}\{A \mid \mathcal{G}\}(\omega)$ a.s.*

Notation: From now on, we write $P_\omega(A)$ instead of $P(\omega, A)$.
We can give the following

Definition 1.2.2. *Z is a* mixture of i.i.d. sequences *if there exists a σ-field $\mathcal{G} \subseteq \mathcal{F}$ and a regular conditional probability $P_\omega(\cdot)$ on $\mathcal{F}$ given $\mathcal{G}$ such that*

$$P_\omega(Z_1^n = z_1^n) = \prod_{t=1}^{n} P_\omega(Z_1 = z_t) \qquad \mathbb{P}\text{-a.s.}$$

i.e. the Z_i are i.i.d. relative to $P_\omega(\cdot)$ for $\mathbb{P}$-almost all ω.

Recalling the definition of regular conditional probability the last condition becomes

$$\mathbb{P}\big\{Z_1^n = z_1^n \mid \mathcal{G}\big\} = \prod_{t=1}^{n} \mathbb{P}\big\{Z_1 = z_t \mid \mathcal{G}\big\} \qquad \mathbb{P}\text{-}a.s. \tag{1.2}$$

i.e. the variables Z_t are conditionally independent given $\mathcal{G}$. Therefore for a mixture of i.i.d. sequences Z, we can write

$$\begin{aligned}
\mathbb{P}\big\{Z_1^n = z_1^n\big\} &= \int \mathbb{P}\big\{Z_1^n = z_1^n \mid \mathcal{G}\big\}(\omega)\mathbb{P}(d\omega) \\
&= \int \prod_{t=1}^{n} \mathbb{P}\big\{Z_1 = z_t \mid \mathcal{G}\big\}(\omega)\mathbb{P}(d\omega) \\
&= \int \prod_{t=1}^{n} P_\omega(Z_1 = z_t)\mathbb{P}(d\omega).
\end{aligned} \tag{1.3}$$

Let Z take values into the measurable space $(J, \mathcal{J})$, and let $\mathcal{M}_{\mathcal{J}}$ be the set of all probability measures on $(J, \mathcal{J})$. (In the case of finite $J = \{1, 2, \dots, d\}$ the set $\mathcal{M}_{\mathcal{J}}$ coincides with the face of the simplex in $\mathbb{R}^d$ i.e. the set of probability vectors $q = (q_1, \dots q_d)$ such that $\sum_{h=1}^{d} q_h = 1$). Denote with m a generic element in $\mathcal{M}_{\mathcal{J}}$. Equip $\mathcal{M}_{\mathcal{J}}$ with the σ-field generated by the maps $m \mapsto m(A)$, varying $m \in \mathcal{M}_{\mathcal{J}}$ and $A \in \mathcal{J}$. We report below an alternative definition of mixture of i.i.d. sequences, widely used in the literature.

Definition 1.2.3. *A sequence Z of random variables with values on $(J, \mathcal{J})$ is a* mixture of i.i.d. sequences *if there exists a measure μ in $\mathcal{M}_{\mathcal{J}}$ such that*

$$\mathbb{P}\big\{Z_1^n = z_1^n\big\} = \int_{\mathcal{M}_{\mathcal{J}}} \prod_{t=1}^{n} m(z_t)\mu(dm). \tag{1.4}$$

Proposition 1.2.1. *Let Z be a mixture of i.i.d. sequences according to Definition 1.2.2. Then Z is a mixture of i.i.d. sequences according to Definition 1.2.3 as well.*

Proof. P_ω as in Definition 1.2.2 is a probability measure on $(\Omega, \mathcal{F})$ for any fixed ω. Therefore the random variable $Z_1 : (\Omega, \mathcal{F}) \longrightarrow (J, \mathcal{J})$ induces a probability measure $\bar{P}_\omega$ on $(J, \mathcal{J})$ in the usual way, i.e. for any $A \in \mathcal{J}$

$$\bar{P}_\omega(A) := P_\omega(Z_1 \in A).$$

Here $\bar{P}_\omega(\cdot)$ is a probability measure on $(J, \mathcal{J})$ for any fixed ω. Letting ω vary on Ω, $\bar{P}_{(\cdot)}$ is a random variable on $(\Omega, \mathcal{F})$ with values in the set $\mathcal{M}_\mathcal{J}$. Let $\mathcal{L}_{\bar{P}}$ be the law of $\bar{P}_{(\cdot)}$, thus $\mathcal{L}_{\bar{P}}$ is a probability measure on $\mathcal{M}_\mathcal{J}$ defined as

$$\mathcal{L}_{\bar{P}}(M) := \mathbb{P}(\bar{P}_\omega \in M) \tag{1.5}$$

for any measurable $M \subseteq \mathcal{M}_\mathcal{J}$.

By the previous observations, we can write equation (1.3) as

$$\begin{aligned}\mathbb{P}\{Z_1^n = z_1^n\} &= \int \prod_{t=1}^{n} P_\omega(Z_1 = z_t)\mathbb{P}(d\omega) = \int \prod_{t=1}^{n} \bar{P}_\omega(z_t)\mathbb{P}(d\omega) \\ &= \int_{\mathcal{M}_\mathcal{J}} \prod_{t=1}^{n} m(z_t)\mathcal{L}_{\bar{P}}(dm).\end{aligned}$$

The last expression is just equation (1.4) with $\mu = \mathcal{L}_{\bar{P}}$. □

Mixtures of Markov chains

We introduce below mixtures of Markov chains. Let Z be a sequence taking values in the countable measurable space $(J, \mathcal{J})$, and let $\mathcal{P}$ be the set of stochastic matrices on $J \times J$ with the topology of coordinate convergence. We give the following

Definition 1.2.4. *Z is a* mixture of homogeneus Markov chains *if for any fixed initial state z_1, there exists a random variable $\widetilde{P}^{z_1}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the set of stochastic matrices $\mathcal{P}$ such that*

$$\mathbb{P}\{Z_1^n = z_1^n \mid \widetilde{P}^{z_1}\} = \widetilde{P}^{z_1}_{z_1,z_2}\widetilde{P}^{z_1}_{z_2,z_3}\widetilde{P}^{z_1}_{z_{n-1},z_n} \ \mathbb{P}\text{-}a.s.\ , \tag{1.6}$$

where $\widetilde{P}^{z_1}_{i,j}$ indicates the ij-entry of the matrix $\widetilde{P}^{z_1}$.

We will often omit to indicate the initial state z_1, writing $\widetilde{P}$ instead of $\widetilde{P}^{z_1}$. An alternative definition of mixture of Markov chains is the following

Definition 1.2.5. *Z is a* mixture of homogeneous Markov chains *if there exists a probability μ on $J \times \mathcal{P}$ such that*

$$\mathbb{P}\{Z_1^n = z_1^n\} = \int_{\mathcal{P}} \prod_{t=1}^{n-1} P_{z_t,z_{t+1}}\mu(z_1, dP). \tag{1.7}$$

According to the previous definition the initial distribution for Z is given by

$$\mathbb{P}\{Z_1 = z_1\} = \int_{\mathcal{P}} \mu(z_1, dP). \tag{1.8}$$

Proposition 1.2.2. *Let Z be a mixture of Markov chains according to Definition 1.2.4. Then Z is a mixture of Markov chains according to Definition 1.2.5.*

Proof. Let $\widetilde{P}^{z_1}$ be the random matrix satisfying Definition 1.2.4 for the initial state z_1 and let $\mathcal{L}_{\widetilde{P}^{z_1}}$ the law of $\widetilde{P}^{z_1}$.

$$\mathbb{P}\{Z_1^n = z_1^n\} = \int_\Omega \mathbb{P}\{Z_1^n = z_1^n \mid \widetilde{P}^{z_1}\}(\omega)\mathbb{P}(d\omega) =$$

$$\int_\Omega \prod_{t=1}^{n-1} \widetilde{P}^{z_1}_{z_t,z_{t+1}} \mathbb{P}(d\omega) = \int_{\mathcal{P}} \prod_{t=1}^{n-1} \widetilde{P}^{z_1}_{z_t,z_{t+1}} \mathcal{L}_{\widetilde{P}^{z_1}}(d\widetilde{P}).$$

Taking $\mathcal{L}_{\widetilde{P}^{z_1}} = \mu(z_1, \cdot)$ we get equation (1.7). □

1.2.2 Finite and countable mixtures

In this work we deal with countable (finite) mixtures of i.i.d. sequences and of Markov chains. We need the following definition

Definition 1.2.6. *A regular conditional probability P is concentrated if there are a countable (finite) set K of indices and measures $p_1, p_2, \ldots, p_k, \ldots$ on $(\Omega, \mathcal{F})$, with $k \in K$, such that $\mu_k := \mathbb{P}\big(P_\omega(\cdot) = p_k(\cdot)\big) > 0$ and $\sum_k \mu_k = 1$.*

Definition 1.2.7. *A mixture of i.i.d. sequences according to Definition 1.2.2 is countable (finite) if the regular conditional probability P is concentrated, with K countable (finite).*

Definition 1.2.8. *A mixture of i.i.d. sequences Z according to Definition 1.2.3 is countable (finite) if there are a countable (finite) set K of indices, and measures $p_1, p_2, \ldots, p_k, \ldots$ in $\mathcal{M}_{\mathcal{J}}$, with $k \in K$, such that, letting $\mu_k := \mu(m_k)$, we have $\sum_k \mu_k = 1$.*

In this case equation (1.4) becomes

$$\mathbb{P}\big\{Z_1^n = z_1^n\big\} = \sum_{k \in K} \mu_k \Big(\prod_{t=1}^{n} p_k(z_t) \Big). \tag{1.9}$$

Definition 1.2.9. *A mixture of Markov chains according to Definition 1.2.4 is countable (finite) if the random variable $\widetilde{P}$ takes just a countable (finite) number of values.*

Definition 1.2.10. *A mixture of Markov chains according to Definition 1.2.5 is countable (finite), if there exist a countable (finite) set of indices K and matrices $P^1, P^2, \ldots, P^k, \ldots$, with $k \in K$, such that for any $k \in K$ we have $\mu(z, P^k) > 0$ for at least one $z \in J$, and $\sum_{z \in J} \sum_{k \in K} \mu(z, P^k) = 1$.*

For countable mixtures of Markov chains the integral in equation (1.7) becomes a sum and we get

$$\mathbb{P}\{Z_1^n = z_1^n\} = \sum_{k \in K} \mu(z_1, P^k) \prod_{t=1}^{n-1} P^k_{z_t, z_{t+1}}. \tag{1.10}$$

Applications of finite mixtures

Finite mixtures of Markov chains are a useful model in many different applications. They can model data sets coming from a heterogeneous population, i.e. from a population that is naturally divided into a finite number of groups, and such that, picking a sample from a group, this evolves with a Markovian dynamic. For example Cadez *et al.* in [13] use a finite mixture of Markov chains to analyze navigation patterns on a web site. Users are grouped into a finite number of clusters, according to their navigation patterns through web pages of different URL categories. To each cluster is associated a Markov chain, describing the dynamics of the sequence of URL categories visited by a user belonging to that cluster. Another example of a special kind of finite mixture of Markov chains, which is a generalization of the mover-stayer model both in discrete and continuous time, can be found in [28]. Frydman in [28] develops the EM algorithm for the maximum likelihood estimation of the parameters of the mixture. As an example, the author applies the estimation procedure to bond ratings migration, modeling the migration as a mixture of two Markov chains, where each Markov chain corresponds to a health state of the market.

1.2.3 Binary mixtures of i.i.d. and Markov sequences

In this section $J := \{0, 1\}$. Each measure $m \in \mathcal{M}_{\mathcal{J}}$ is characterized by a number $p \in [0, 1]$, where $p := m\{1\}$, and for each number $p \in [0, 1]$

there is a measure m in $\mathcal{M}_{\mathcal{J}}$ such that $m\{1\} = p$, so there is a one-to-one correspondence between $\mathcal{M}_{\mathcal{J}}$ and the unitary interval $I = [0,1]$: set $T : \mathcal{M}_{\mathcal{J}} \longrightarrow [0,1]$ with $T(m) = m(1)$, the function T is bijective.
Let Z be a sequence of binary i.i.d. random variables with respect to a measure $m \in \mathcal{M}_{\mathcal{J}}$, let $p := m(1)$. Define $n_1 = \sum_{t=1}^{n} z_t$ the number of 1-s in the finite string z_1^n; for any z_1^n we have

$$m(z_1^n) = \prod_{t=1}^{n} m(z_t) = p^{n_1} (1-p)^{n-n_1}. \tag{1.11}$$

Let now Z be a mixture of i.i.d. binary sequences. Combining equations (1.4) and (1.11), we get

$$\mathbb{P}\{Z_1^n = z_1^n\} = \int_0^1 p^{n_1} (1-p)^{n-n_1} d\nu(p), \tag{1.12}$$

where ν is a probability measure on the interval [0,1] such that $\nu(A) := \mu(T^{-1}(A))$ for all $A \subset \mathcal{B}([0,1])$. With a small abuse of notations we will often indicate with the same symbol measures ν on $\mathcal{B}([0,1])$ and measures μ on $\mathcal{M}_{\mathcal{J}}$. We give below an example of a mixture of six i.i.d. binary sequences.

Example 1.2.1. *Take six coins, numbered from 1 to 6, and a die with six faces. Let the probability of Head for the k-th coin be p_k and let the probability of face k of the die be μ_k, we get $\sum_{k=1}^{6} \mu_k = 1$. Throw the die* once, *then toss n times the coin corresponding to the outcome of the die. Let $\mathbb{I}_A$ be the indicator function of the event A. Define*

$$Z_t := \mathbb{I}_{\{\text{the } t\text{-th coin toss is Head}\}}.$$

For any binary sequence z_1^n, we get

$$\mathbb{P}\{Z_1^n = z_1^n\} = \sum_{k=1}^{6} \mu_k p_k^{n_1} (1-p_k)^{n-n_1}. \tag{1.13}$$

Z is a finite binary mixture of i.i.d. sequences. Given the outcome of the die, Z is an i.i.d. coin tossing.

From this example it is apparent that to make inference about mixtures models, repeated sequences of measurements are needed.

Let us now consider binary Markov chains, i.e. Markov chains with values in $J := \{0, 1\}$. Let $\mathcal{P}$ be the set of stochastic matrices P of dimension two. They are of the form

$$P = \begin{pmatrix} P_{00} & 1 - P_{00} \\ 1 - P_{11} & P_{11} \end{pmatrix}.$$

They are completely characterized by the two numbers $P_{00}, P_{11} \in [0, 1]$. Thus there is a one-to-one correspondence between $\mathcal{P}$ and the unit square $U := [0, 1] \times [0, 1]$: set $V : \mathcal{P} \longrightarrow U$ defined as $V(P) := (P_{00}, P_{11})$, the function V is bijective.

Let z_1^n be a binary string and let n_{ij} be the number of transitions from state i to state j in the string z_1^n, for $i, j = 0, 1$. Let Z be a binary mixture of Markov chains, according to Definition 1.2.5. By the previous observations we can write equation (1.7) as

$$\mathbb{P}\{Z_1^n = z_1^n\} = \int_U P_{00}^{n_{00}}(1 - P_{00})^{n_{01}} P_{11}^{n_{11}}(1 - P_{11})^{n_{10}} \nu(z_1, d(P_{00}, P_{11})),$$

where $\nu(z_1, B) := \mu(z_1, V^{-1}(B))$ for any $B \in \mathcal{B}(U)$ and for any fixed z_1.

1.2.4 Representation theorems

Before reviewing some representation theorems for mixture models, we recall two well known definitions. Let $Z = (Z_1, Z_2, \dots)$ be a sequence of random variables with values in a countable measurable set $(J, \mathcal{J})$. We indicate with $\mathbb{P}_z\{\cdot\}$ the conditional probability $\mathbb{P}\{\cdot \mid Z_1 = z\}$.

Definition 1.2.11. *A sequence Z is* recurrent *if, for any initial state $z \in J$,*

$$\mathbb{P}_z\{Z_n = z \ i.o.\} = 1.$$

Definition 1.2.12. *A sequence Z is* strongly recurrent *if*

$$\bigcap_{t=1}^{m} \{Z_t = z_t\} = \bigcap_{t=1}^{m} \big(\{Z_t = z_t\} \cap \{Z_n = z_t \ i.o.\}\big) \quad \mathbb{P} - a.s. \tag{1.14}$$

for any $m \in \mathbb{N}$ and $z_1, \dots, z_m \in J$.

Thus a strongly recurrent sequence hits each state either never or infinitely many times. We recall now de Finetti's representation theorem (for the proof of the theorem and an extensive discussion of mixtures of i.i.d. sequences see Aldous [1] and Chow and Teicher [15]).

Theorem 1.2.1. *(de Finetti) Let* $Z = (Z_1, Z_2, \dots)$ *be a sequence on* $(\Omega, \mathcal{F}, \mathbb{P})$ *taking values in* $(J, \mathcal{J})$*. The sequence* Z *is exchangeable if and only if* Z *is a mixture of i.i.d. sequences.*

The σ-field $\mathcal{G}$ of Definition 1.2.2, can be the tail σ-field $\mathcal{F}^{(\infty)}$ of Z or the σ-field of permutable events. It can be shown that the conditional mean with respect to these two σ-fields is equal *a.s.* (for definitions and details see Chapter 7.3 in [15]). By defining mixtures of i.i.d. sequences according to Definition 1.2.3, the measure μ in equation (1.4) is uniquely determined by $\mathbb{P}$.

The de Finetti theorem for mixtures of i.i.d. sequences has been generalized to mixtures of Markov chains. Diaconis and Freedman in [22] extend de Finetti's theorem connecting the class of Markov exchangeable sequences, which Diaconis and Freedman call partially exchangeable sequences, with mixtures of Markov chains. They proved the following

Theorem 1.2.2. *(See Theorem (7) in [22]) Let* $Z = (Z_1, Z_2, \dots)$ *be a recurrent sequence on* $(\Omega, \mathcal{F}, \mathbb{P})$ *taking values in* $(J, \mathcal{J})$*. The sequence* Z *is Markov exchangeable if and only if* Z *is a mixture of Markov chains according to Definition 1.2.5.*

Moreover in [22] page 127 it is proved that any stationary Markov exchangeable sequence is a mixture of stationary Markov chains, i.e. Markov chains that start with their stationary distributions.

In [27] Fortini *et al.* connect the class of partially exchangeable sequences with mixtures of Markov chains. We recast in the following Theorem 1.2.3 the precise result. In particular, in [27] Section 2.1 Fortini and coauthors prove that the matrix V of successors of Z is partially exchangeable if and only if Z is a mixture of Markov chains. They refer to Definition 1.2.4 of mixture of Markov chains. With a clever restriction of the class of possible mixing Markov chains they get the uniqueness of the mixing distribution. More precisely let $J^* = J \cup \{\delta\}$. Let $\mathcal{P}^*$ be the set of transition matrices on J^* for which δ is an absorbing state, equipped with the σ-field of coordinate convergence and let us indicate with $P_{i,j}$ the i,j-entry of the matrix $P \in \mathcal{P}^*$. Assume that Z starts in a specific state z_1.

Consider a measurable set $\mathcal{K}$ in $\mathcal{P}^*$ such that for any $P \in \mathcal{K}$ there is a set $A_P \subset J^*$ satisfying

1. $z_1 \in A_P$, $\delta \notin A_P$;

2. $P_{i,j} = 0$ if $i \in A_P$ and $j \notin A_P$;

3. $P_{i,\delta} = 1$ if $i \notin A_P$.

Theorem 1.2.3. *(see Theorem 1 in [27]) The V matrix of Z is partially exchangeable if and only if there exists a random element $\widetilde{P}$ of $\mathcal{P}^*$ such that*

- $\mathbb{P}\{Z_1^n = z_1^n \mid \widetilde{P}\} = \widetilde{P}_{z_1,z_2}\widetilde{P}_{z_2,z_3}\dots\widetilde{P}_{z_{n-1},z_n}$ *a.s.-*$\mathbb{P}$*;*
- $\mathbb{P}\{Z_n = z_1 \quad i.o.\ n \mid \widetilde{P}\} = 1$ *a.s.-*$\mathbb{P}$*;*
- $\mathbb{P}\{\widetilde{P} \in \mathcal{K}\} = 1$ *for a measurable set* $\mathcal{K}$ *satisfying conditions 1,2, and 3 above.*

Moreover, $\widetilde{P}$ is unique in distribution.

The mixing measure in Theorem 1.2.2 in general is not unique. The uniqueness in distribution of $\widetilde{P}$ in Theorem 1.2.3 is obtained requiring $\mathbb{P}\{\widetilde{P} \in \mathcal{K}\} = 1$. For any Markov chain with transition matrix in $\mathcal{K}$, there is a set A_P of states hit by the chain. With the third condition satisfied by $\mathcal{K}$, we fix the behavior of the chain on the states that we never see for that chain, getting uniqueness. For a detailed discussion see [27], in particular Example 2.

Both Diaconis and Freedman and Fortini *et al* give a characterization theorem for mixtures of Markov chains, but note that Diaconis and Freedman refer to the Definition 1.2.5 of mixture of Markov chains and show a theorem for Markov exchangeable sequences, while Fortini *et al* refer to Definition 1.2.4 and show a theorem for partially exchangeable sequences. Markov exchangeability is weaker than partial exchangeability in general, so Theorem 1.2.2 and Theorem 1.2.3 are not in contradiction thanks to Proposition 1.2.2.

In the following Theorem 1.2.4 we characterize k-Markov exchangeable sequences, generalizing the result of Diaconis and Freedman ([22]) for Markov exchangeable sequences. To state the theorem, we need the following recurrence condition

Condition 1.2.1. $\mathbb{P}\{Y_n^{n+k-1} = Y_1^k \ i.o.\ n\ \} = 1.$

The characterization theorem is the following

Theorem 1.2.4. *Let Z be a sequence satisfying the recurrence Condition 1.2.1. Z is k-Markov exchangeable if and only if it is a mixture of Markov chains of order k.*

We will prove the theorem in Section 3.3.1.

1.3 Hidden Markov Models

There are many equivalent definitions of Hidden Markov Model (HMM).
In the first part of the thesis we will refer to the oldest and probably most intuitive one, but all the results could be recast using any of the alternative equivalent definitions of HMM.

Definition 1.3.1. *Y is a HMM if there exist a homogeneous Markov chain $X = (X_1, X_2, \dots)$ taking values in a set χ and a "many to one" deterministic function $f : \chi \longrightarrow J$ such that $Y_n = f(X_n)$ for each $n \in \mathbb{N}$.*

Indicating with $\sharp S$ the cardinality of a set S, we have $\sharp\chi \geq \sharp J$. X will be called the "underlying Markov chain" of the HMM.

Definition 1.3.2. *Y is a countable HMM if J and χ in the previous definition are at most countable.*

In the thesis we will often handle HMMs with a recurrent underlying Markov chain. We will make use the acronyms RHMM to indicate this class of HMMs.
Notice that in the literature "recurrent Markov chain" often means a Markov chain with just one recurrence class. But for us a recurrent Markov chain is a Markov chain satisfying Definition 1.2.11. Trivially a Markov chain with just one recurrence class is recurrent according to Definition 1.2.11.

A Hidden Markov Model can be equivalently defined as a probabilistic function of a Markov chain as follows

Definition 1.3.3. *The sequence (Y_n) with values in J is a Hidden Markov process if there exists a homogenous Markov chain (X_n) with state space χ such that (Y_n) is conditionally independent and identically distributed given the σ-field generated by the process (X_n).*

The sets χ and J are called respectively the state space, and the observation or read-out space of the HHM. The previous definition can be extended to general read-out space J.
According to Definition 1.3.3, for finite read-out space J, the HHM is completely characterize by the transition probability matrix Q of the unobserved Markov chain (X_n)

$$Q_{ij} := P(X_{n+1} = j \mid X_n = i),$$

and by the read-out probabilities $b^x(y)$

$$b^x(y) := P(Y_n = y \mid X_n = x).$$

For continuous read-outs the read-out densities are defined as

$$P(Y_n \in dy \mid X_n = x) = b^x(y)\lambda(dy),$$

for a suitable non-negative σ-finite measure λ.

Chapter 2

Finite binary mixtures

In this chapter we restrict attention to $\{0,1\}$-valued, i.e. binary, sequences. In Section 2.1 we characterize the binary exchangeable sequences which are *finite* mixtures of i.i.d. ones. For finite mixtures, we devise an original algorithm to compute the associated de Finetti's measure. We will make use of a well known connection between the moments of the de Finetti measure and the probabilities of strings of 1s.
In Section 2.2 we use the results of Section 2.1 to solve two classical engineering problems: the stochastic realization problem and the positive realization of a linear system, in two special cases. We solve the stochastic realization problem for binary sequences which are mixtures of N i.i.d. sequences, i.e. we find a parametrization of the distribution function of the sequence knowing some finite distributions of the sequence. In particular we just need the probabilities $\mathbb{P}(1), \mathbb{P}(1^2), \dots, \mathbb{P}(1^{2N})$. This is not surprising, in fact for binary exchangeable sequences the probability of any finite string is computable from the probabilities of strings of 1-s, as recalled in the subsection 2.2.1. The results developed in Section 2.1 allow us also to find the positive realization for linear systems with impulse responses in a special class and associated Hankel matrix of finite rank.
In Section 2.3 we partially extend the results of Section 2.1 to Markov exchangeable sequences, characterizing those that are *finite* mixtures of Markov chains.

2.1 Mixing measure of i.i.d. mixtures

2.1.1 Cardinality of the mixing measure

In this section we characterize exchangeable binary sequences, whose de Finetti's mixing measure μ is concentrated on a finite set. The criterion will be given in Theorem 2.1.1 below. We first recall some well known definitions and facts.

Moments matrix and distribution function

Let μ be a probability measure on $[0,1]$. The m-th moment α_m of μ is defined as

$$\alpha_m := \int_0^1 x^m \mu(dx). \tag{2.1}$$

Assume that the moments α_m are all finite. Define

$$\mathbf{M_n} := \begin{pmatrix} \alpha_0 & \alpha_1 & \dots & \alpha_n \\ \alpha_1 & \alpha_2 & \dots & \alpha_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_n & \dots & \dots & \alpha_{2n} \end{pmatrix} \quad \text{and} \quad d_n := \det(M_n). \tag{2.2}$$

Given a real valued function f, we say that x_0 is a *point of increase* for f if $f(x_0+h) > f(x_0-h)$ for all $h > 0$.

Given a probability measure μ on $[0,1]$, the *distribution function* F^μ of μ is defined as

$$F^\mu(x) := \mu([0,x]) \text{ for any } x \in [0,1]. \tag{2.3}$$

Following Cramer [17], Chapter 12.6 it can be shown

Lemma 2.1.1. *If F^μ has N points of increase, then $d_n \neq 0$ for $n = 0,\dots,N-1$ and $d_n = 0$ for $n \geq N$.*
If F^μ has infinitely many points of increase, then $d_n \neq 0$ for any n.

The previous lemma will play a central role in the future treatment, thus we report the proof given by Cramer [17].

Proof. For any $n \geq 0$, define the quadratic form Q_n on $\mathbb{R}^{n+1}$

$$Q_n(u_0,\dots,u_n) := \int_0^1 (u_0 + u_1 x + \dots + u_n x^n)^2 dF^\mu(x) = \sum_{i,j=0}^{n} \alpha_{i+j} u_i u_j \geq 0, \tag{2.4}$$

with α_m being the m-th moments of μ. The symmetric matrix associated with the quadratic form Q_n is the moments matrix M_n.
Let F^μ have N points of increase. For $n \leq N-1$, for any $(u_0, u_1, \dots, u_n)$ there is at least one point of increase x of F^μ, that is not a root of $u_0 + u_1 x + \dots + u_n x^n$, so $Q(u_0, u_1, \dots, u_n) > 0$. Thus Q_n is positive definite for $n \leq N-1$, and $d_n = \det(M_n) > 0$ for all $n \leq N-1$. But Q_n is semi-definite positive for $n \geq N$, in fact there exist $(u_o, u_1, \dots, u_n)$, not identically 0, such that all the points of increase of F^μ are roots of $u_0 + u_1 x + \dots + u_n x^n$. Thus $Q(u_0, u_1, \dots, u_n) = 0$ for some $(u_o, u_1, \dots, u_n)$, not identically 0, and $d_n = \det(M_n) = 0$ for all $n \geq N$. This concludes the proof of the first statement of the lemma.
If F^μ has an infinite number of points of increase, then for any n and for any $(u_0, u_1, \dots, u_n)$, there are points of increase x of F^μ that are not roots of $u_0 + u_1 x + \dots + u_n x^n$. Thus the integral in equation (2.4) is strictly positive as long as the u_i are not all equal to zero. Thus the quadratic form Q_n is positive definite for all n, and $d_n = \det(M_n) > 0$ for all n. □

Remark 2.1.1. *F^μ has exactly N points of increase if and only if F^μ is a step-function, i.e. there exist a partition of the interval $[0,1]$ of points $0 = p_0 < p_1 < \dots < p_{N+1} = 1$ and numbers $0 = a_1 < \dots < a_N < a_{N+1} = 1$ such that $F^\mu(x) = \sum_{i=1}^{N+1} a_i \mathbb{I}_{[p_{i-1}, p_i[}$.*

μ gives positive probability to the points of discontinuity of F^μ. Moreover

Remark 2.1.2. *F^μ has exactly N points of increase $p_1, \dots, p_N$ if and only if μ is concentrated on $p_1, \dots, p_N$, i.e. $\mu_i := \mu(p_i) > 0$ for $i = 1, \dots, N$ and $\sum_{i=1}^N \mu_i = 1$. It holds that $a_i = \sum_{j=1}^{i-1} \mu(p_j)$.*

Concentration points of a measure

For any measure m on a measurable space $(E, \mathcal{E})$, define the set of concentration points

$$C_m := \{x \in E \mid m(x) > 0\}. \tag{2.5}$$

For absolutely continuous measures m on $\mathbb{R}$, the set C_m is empty. For measures concentrated on a finite set, C_m coincides with the set of points of increase of F^m, but it does not hold in general. For example consider the probability measure on $[0,1]$

$$m(\cdot) := \frac{1}{4}\,\delta_{1/3}(\cdot) + \frac{1}{4}\,\delta_{2/3}(\cdot) + \frac{1}{2}\lambda(\cdot),$$

where δ_c is the Dirac measure concentrated in c and λ is the Lebesgue measure on $[0,1]$. $C_m = \{1/3, 2/3\}$, but it does not coincide with the set of points of increase of F^m, which is the open interval $]0,1[$.

Moments of the de Finetti measure

Let us consider a binary exchangeable sequence $Y = (Y_1, Y_2; \dots)$. By the de Finetti theorem (see Theorem 1.2.1 in Chapter 1) for binary sequences, we can write

$$\mathbb{P}(1^m 0^n) = \int_0^1 p^m (1-p)^n d\mu(p), \tag{2.6}$$

where μ is a probability measure on the interval $[0,1]$ uniquely determined by $\mathbb{P}$. Thus

Remark 2.1.3. *The probability of a sequence of m 1s is the m-th moment α_m of the measure μ:*

$$\mathbb{P}(1^m) = \int_0^1 p^m d\mu(p) = \alpha_m.$$

Hankel matrices

We need now introduce a class of Hankel matrices. Let us first fix some notations. Let $Y = (Y_1, Y_2, \dots)$ be a binary sequence on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. With 0^k we indicate the string of k consecutive 1s, while $\{0^k\}$ indicates the event $\{Y_1^k = 0^k\}$. With $\emptyset$ we indicate the empty string, and the event $\{\emptyset\} = \{Y = \emptyset\}$ corresponds to the whole space Ω, thus $\mathbb{P}(\{\emptyset\}) = 1$.
For any $n \in \mathbb{N}$, let $H_n = (h_{ij})_{0 \le i,j \le n}$ be the $(n+1) \times (n+1)$ Hankel matrix, with entries $h_{ij} := h_{i+j} = \mathbb{P}(1^i 1^j) = \mathbb{P}(1^{i+j})$:

$$\mathbf{H_n} := \begin{pmatrix} \mathbb{P}(\{\emptyset\}) & \mathbb{P}(1) & \mathbb{P}(11) & \dots & \mathbb{P}(1^n) \\ \mathbb{P}(1) & \mathbb{P}(11) & \mathbb{P}(111) & \dots & \mathbb{P}(1^{n+1}) \\ \mathbb{P}(11) & \mathbb{P}(111) & \mathbb{P}(1111) & \dots & \mathbb{P}(1^{n+2}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbb{P}(1^n) & \mathbb{P}(1^{n+1}) & \mathbb{P}(1^{n+2}) & \dots & \mathbb{P}(1^{2n}) \end{pmatrix}. \tag{2.7}$$

H_∞ is defined in the obvious way .

The following example gives an alternative representation of the matrix H_n for a sequence of i.i.d. random variables.

Example 2.1.1. *Let Y be a sequence of i.i.d. random variables taking values on* $\{0,1\}$ *and let* $p = \mathbb{P}\{1\}$*. Then*

$$H_n = \begin{pmatrix} 1 & p & \dots & p^n \\ p & p^2 & \dots & p^{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ p^n & p^{n+1} & \dots & p^{2n} \end{pmatrix} = \begin{pmatrix} 1 \\ p \\ p^2 \\ \vdots \\ p^n \end{pmatrix} \begin{pmatrix} 1 & p & p^2 & \dots & p^n \end{pmatrix}.$$

In fact for an i.i.d. sequence

$$\mathbb{P}\{1^k\} = (\mathbb{P}\{1\})^k = p^k.$$

Remark 2.1.3 in the previous subsection leads us to the following

Remark 2.1.4. *Let Y be an exchangeable binary sequence and let* μ *be the associated de Finetti measure. The matrix* H_n *defined in equation (2.7) is the moments matrix* M_n *of the measure* μ*.*

Cardinality of the mixing measure

We are ready to state the main result of this section

Theorem 2.1.1. *Let Y be a binary exchangeable sequence and let* (H_n) *be defined as in equation (2.7). Then*

- *Y is a mixture of N i.i.d. sequences if and only if*

$$rank(H_n) = \begin{cases} n+1 & for\ n = 0, \dots, N-1 \\ N & for\ n \geq N \end{cases} \tag{2.8}$$

- *Y is a mixture of an infinite number of i.i.d. sequences if and only if*

$$rank(H_n) = n + 1\ for\ any\ n. \tag{2.9}$$

Corollary 2.1.1. *The rank of the semi-infinite Hankel matrix* H_∞ *associated to a binary exchangeable sequence Y is finite and equal to N if and only Y is a mixture of N i.i.d. sequences.*

The proof of the corollary is trivial.

Proof. of Theorem 2.1.1 Let $rank(H_n)$ be as in equation (2.8). H_n is a $(n+1)\times(n+1)$ matrix, thus

$$det(H_n)\begin{cases} \neq 0 & \text{for } n = 0,\dots,N-1 \\ = 0 & \text{for } n \geq N. \end{cases}$$

Let μ be the mixing measure associated with the exchangeable sequence Y and let F^{μ} its distribution function. $det(H_n) \neq 0$ does not hold for any n, thus by the second statement of Lemma 2.1.1 the distribution function F^{μ} has finitely many points of increase, and F^{μ} has exactly N points of increase by the first statement of Lemma 2.1.1. Thus by Remark 2.1.2 the measure μ is concentrated on N points, and Y is a mixture of N i.i.d. sequences indeed.

Let Y be a mixture of N i.i.d. sequences. μ is concentrated on N points, call them $p_1, p_2, \dots, p_N$. By Remark 2.1.2 the function F^{μ} has N points of increase, thus by Lemma 2.1.1, we have $det(H_n) \neq 0$ for $n = 0,\dots,N-1$ and $det(H_n) = 0$ for $n \geq N$. Hence H_n has full rank for $n = 0,\dots,N-1$, but the rank is not full anymore for $n \geq N$. It remains to show that $rank(H_n) = N$ for $n \geq N$. The measure μ is concentrated on N points so

$$\mathbb{P}(1^m) = \int_0^1 p^m d\mu(p) = \sum_{i=1}^{N} \mu_i p_i^m.$$

From the last equation we have

$$H_n = \begin{pmatrix} \sum_{i=1}^N \mu_i & \sum_{i=1}^N \mu_i p_i & \dots & \sum_{i=1}^N \mu_i p_i^n \\ \sum_{i=1}^N \mu_i p_i & \sum_{i=1}^N \mu_i p_i^2 & \dots & \sum_{i=1}^N \mu_i p_i^{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^N \mu_i p_i^n & \sum_{i=1}^N \mu_i p_i^{n+1} & \dots & \sum_{i=1}^N \mu_i p_i^{2n} \end{pmatrix}$$

$$= \mu_1 \begin{pmatrix} 1 \\ p_1 \\ \vdots \\ p_1^n \end{pmatrix} \begin{pmatrix} 1 & p_1 & \dots & p_1^n \end{pmatrix} + \dots + \mu_N \begin{pmatrix} 1 \\ p_N \\ \vdots \\ p_N^n \end{pmatrix} \begin{pmatrix} 1 & p_N & \dots & p_N^n \end{pmatrix}.$$

$p_i \neq p_j$ for $i \neq j$ and so the vectors $(1, p_i, \cdots, p_i^n)'$ are linearly independent for $n \geq N-1$. For $n \geq N-1$ the matrix H_n is a linear combination of the vector product of N linearly independent vectors, so by an easy argument of linear algebra H_n has rank N.
The second statement of the theorem is just an easy consequence of Remark 2.1.2 and Lemma 2.1.1. □

Given $H_0, H_1, H_2, \dots$, the previous theorem gives a criterion to establish whether the measure μ associated with an exchangeable binary sequence is concentrated on a finite set:

Criterion 1. *μ is concentrated on N points if and only if there exists an integer n such that $det(H_{n-1}) \neq 0$ and $det(H_n) = 0$, and in this case $N = n$.*

Note that if $det(H_{n-1}) \neq 0$ and $det(H_n) = 0$ then by Theorem 2.1.1 we *a priori* know that $rank(H_n) = N$ for all $n \geq N$.

2.1.2 Computation of the mixing measure

In this section we propose an original algorithm to compute the de Finetti measure associated with a binary finite mixture of i.i.d. sequences. The algorithm is a by-product of Theorem 2.1.3 below, which follows by the results of the previous section and by a well known theorem on Hankel matrices reported below (see [29])

Theorem 2.1.2. *Let $H_\infty = (h_{i+j})_{i,j \geq 0}$ be a semi-infinite Hankel matrix of rank N. Then the entries of H_∞ satisfy an N-term recurrence equation of the form:*

$$h_m = a_{N-1} h_{m-1} + a_{N-2} h_{m-2} + \dots + a_0 h_{m-N} \quad \text{for all } m \geq N, \tag{2.10}$$

for suitable $(a_0, a_1, \dots, a_{N-1})$.
Given $(a_0, a_1, \dots, a_{N-1})$, let $p_1, \dots, p_l$ be the distinct roots of the polynomial

$$q(x) := x^N - a_{N-1} x^{N-1} - a_{N-2} x^{N-2} - \dots - a_0, \tag{2.11}$$

and let m_i be the multiplicity of the root p_i, with $1 \leq i \leq l$.
Then a N-dimensional base $< v_1, \dots, v_N >$ of the space of the infinite vectors $[h_0, h_1, h_2, \dots]$ which are solutions of the equation (2.10), is given by

$$\begin{aligned}
v_1 &:= \begin{pmatrix} 1 & p_1 & p_1^2 & \dots \end{pmatrix}, \\
&\vdots \\
v_{m_1} &:= \frac{1}{(m_1-1)!} \frac{d^{m_1-1}}{dp_1^{m_1-1}} \begin{pmatrix} 1 & p_1 & p_1^2 & \dots \end{pmatrix} = \begin{pmatrix} 0 & \dots & 1 & \binom{m_1}{1} p_1 & \dots . \end{pmatrix}, \\
v_{m_1+1} &:= \begin{pmatrix} 1 & p_2 & p_2^2 & \dots \end{pmatrix} \\
&\vdots \\
v_N &:= \frac{1}{(m_l-1)!} \frac{d^{m_l-1}}{dp_i^{m_l-1}} \begin{pmatrix} 1 & p_l & p_l^2 & \dots \end{pmatrix} = \begin{pmatrix} 0 & \dots & 1 & \binom{m_l}{1} p_l & \dots . \end{pmatrix},
\end{aligned}$$

for $i = 1, \ldots, l$ *and* $\sum_{i=1}^{l} m_i = N$.

Remark 2.1.5. *If the roots of the polynomial* $q(x)$ *defined in equation (2.11) are all distinct, i.e.* $m_i = 1$ *for all* $i = 1, \ldots, l$*, then the base* $< v_1, \ldots, v_N >$ *of the solutions of equation (2.10) is given by*

$$\begin{aligned} v_1 &:= \begin{pmatrix} 1 & p_1 & p_1^2 & p_1^3 & \ldots\ldots \end{pmatrix}, \\ v_2 &:= \begin{pmatrix} 1 & p_2 & p_2^2 & p_2^3 & \ldots\ldots \end{pmatrix}, \\ &\vdots \\ v_N &:= \begin{pmatrix} 1 & p_N & p_N^2 & p_N^3 & \ldots\ldots \end{pmatrix}. \end{aligned}$$

Thus there exist linear combinators $\mu_1, \ldots, \mu_N$ *such that the entries* h_m *of the Hankel matrices* (H_n) *can be written as*

$$h_m = \sum_{i=1}^{N} \mu_i p_i^m. \tag{2.12}$$

Note that this is a general result on Hankel matrices and that (2.12) needs not be a convex combination. The following theorem adapts the result to the Hankel matrix of a finite mixture of i.i.d. processes.

Theorem 2.1.3. *Let* Y *be a binary mixture of* N *i.i.d. sequences, and let* μ *be the associated de Finetti mixing measure. Then the matrix* H_∞ *associated with* Y *has rank* N*, the roots of the polynomial* $q(x)$ *defined in equation (2.11) are all distinct, the linear combinators* $\mu_1, \ldots, \mu_N$ *in equation (2.12) are all in* $]0, 1[$ *and sum to* 1*. Moreover* μ *is concentrated on the roots* $p_1, \ldots, p_N$ *of* $q(x)$ *and* $\mu(p_i) = \mu_i$ *for* $i = 1, \ldots, N$*, up to permutations of indices.*

Proof. By Corollary 2.1.1, Y is a mixture of N i.i.d. sequences, thus μ is concentrated on N points $p_1, \ldots, p_N$, with weights $\mu(p_i) = \mu_i$ for $i = 1, \ldots, N$. By de Finetti's theorem for each $m \in \mathbb{N}$ we get

$$h_m = \mathbb{P}(1^m) = \int_0^1 p^m d\mu(p) = \sum_i \mu_i p_i^m, \tag{2.13}$$

and by the uniqueness of the de Finetti measure, up to permutations of the p_i, there is no other way to write h_m as in equation (2.13) with weights $\mu_i \in]0, 1[$. We have to show that $p_1, \ldots, p_N$ are the roots of $q(x)$. By Theorem 2.1.2

for each $m \geq N$, the vector $(h_m, h_{m-1}, \dots, h_{m-N})$ satisfies the recurrence equation (2.10), thus we get

$$h_m - a_{N-1}h_{m-1} - a_{N-2}h_{m-2} - \cdots - a_0 h_{m-N} = 0 \quad \text{for } m \geq N. \tag{2.14}$$

Combining the last equation with (2.13), we get

$$\begin{aligned} 0 &= h_m - a_{N-1}h_{m-1} - a_{N-2}h_{m-2} - \cdots - a_0 h_{m-N} \\ &= \sum_i^N \mu_i p_i^m - a_{N-1}\sum_i^N \mu_i p_i^{m-1} - a_{N-2}\sum_i^N \mu_i p_i^{m-2} - \cdots - a_0 \sum_i^N \mu_i p_i^{m-N} \\ &= \sum_i^N \mu_i p_i^{m-N}\left(p_i^N - a_{N-1}p_i^{N-1} - a_{N-2}p_i^{N-2} - \cdots - a_0\right). \end{aligned}$$

All the μ_i and all the p_i are positive thus for each $i = 1, \dots, N$ we get

$$p_i^N - a_{N-1}p_i^{N-1} - a_{N-2}p_i^{N-2} - \cdots - a_0 = 0. \tag{2.15}$$

Thus $p_1, \dots, p_N$ are roots of $q(x)$ and $q(x)$ has degree N, so $p_1, \dots, p_N$ are the N distinct roots of $q(x)$. □

Theorems 2.1.1, 2.1.2 and 2.1.3 allow us to construct an algorithm to identify the mixing measure μ associated to a mixture of N i.i.d. sequences, from the knowledge of the Hankel matrix H_N. The algorithm is presented in the next section.

The algorithm

Let Y be a binary exchangeable sequence, which is a mixture of N i.i.d. sequences. In this section we propose an algorithm to identify the de Finetti mixing measure μ associated with Y, given $\mathbb{P}(\{\emptyset\}), \mathbb{P}(1), \dots, \mathbb{P}(1^{2N})$. μ is concentrated on N points, call them $p_1, \dots, p_N$ and let $\mu_i = \mu(p_i)$, for $i = 1, \dots, N$. Thus to characterize μ we must find $(p_i, \mu_i)_{1 \leq i \leq N}$.
Y is a mixture of N i.i.d. sequences, thus by Theorem 2.1.3, the measure μ is concentrated on the N distinct roots of the polynomial $q(x)$ defined in equation (2.11). Thus to identify $p_1, \dots, p_N$, we need to determine the polynomial $q(x)$, i.e. find its coefficients $(a_0, \dots, a_N)$, (note $a_N = 1$). To this end construct the matrix H_N as defined in equation (2.7), note that

$\mathbb{P}(1), \ldots, \mathbb{P}(1^{2N})$ are given. The matrix H_N is a submatrix of H_∞, thus its entries satisfy the recurrence equation (2.10) in Theorem 2.1.2. This gives

$$\begin{aligned} h_N &= a_{N-1}h_{N-1} + a_{N-2}h_{N-2} + \cdots + a_0 h_0 \\ h_{N+1} &= a_{N-1}h_N + a_{N-2}h_{N-1} + \cdots + a_0 h_1 \\ &\cdots\cdots \\ h_{2N-1} &= a_{N-1}h_{2N-2} + a_{N-2}h_{2N-3} + \cdots + a_0 h_{N-1} \end{aligned}$$

Denoting by $h^{(N)} := (h_N, h_{N+1}, \ldots, h_{2N-1})'$ (the components of $h^{(N)}$ are the first N elements of the $(N+1)$-th column of the matrix H_N) and by $\mathbf{a} := (a_0, a_1, \ldots, a_{N-1})'$ the unknown coefficients of $q(x)$, we can write the previous set of equations in matrix form as follows

$$H_{N-1}\mathbf{a} = h^{(N)}. \tag{2.16}$$

The coefficients $(a_0, a_1, \ldots, a_{N-1})$ of the polynomial $q(x)$ are computed solving the linear system in equation (2.16) in the unknown $\mathbf{a}$. Once $\mathbf{a}$ is determined, find the roots of $q(x)$, getting the points $p_1, \ldots, p_N$ where μ is concentrated.

To completely know μ, we must find the weights μ_i. Recall that we have

$$h_m = \mathbb{P}(1^m) = \sum_i^N \mu_i p_i^m. \tag{2.17}$$

Define the matrices

$$V := \begin{pmatrix} 1 & p_1 & p_1^2 & \ldots & p_1^{N-1} \\ 1 & p_2 & p_2^2 & \ldots & p_2^{N-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & p_N & p_N^2 & \ldots & p_N^{N-1} \end{pmatrix} \quad W := \begin{pmatrix} \mu_1 & 0 & 0 & \ldots & 0 \\ 0 & \mu_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \mu_N \end{pmatrix},$$

where V is by now known and, under the hypothesis, invertible. Equation (2.17) implies

$$H_{N-1} = V'WV. \tag{2.18}$$

To find the matrix W, invert equation (2.18) to get

$$W = (V')^{-1} H_{N-1} V^{-1}. \tag{2.19}$$

The weights μ_i are the diagonal elements of W. Thus we have completely identify μ indeed.

We summarize the algorithm proposed above:

- Find the coefficients $\mathbf{a}$ of $q(x)$ solving the linear system

$$H_{N-1}\mathbf{a} = h^{(N)}$$

- find $(p_1, p_2, \ldots, p_N)$ determining the roots of the polynomial $q(x)$.
- Find $(\mu_1, \ldots, \mu_N)$ constructing the matrix V as in (2.1.2) and computing

$$W = (V^{-1})' H_{N-1} V^{-1}.$$

2.2 Two applications

2.2.1 Some preliminary results

Probability of finite strings for a binary exchangeable sequence

The following proposition shows that for an exchangeable binary sequence the knowledge of the measure of finite strings of 1s is sufficient for the knowledge of the measure of all finite strings. The result is proved in an alternative way in [26], vol.2, Chapter VII.4.

Proposition 2.2.1. *Let Y be an exchangeable binary sequence. The measure of the strings 1^k for $k = 0, 1, 2\ldots$ is sufficient to compute the measure of any finite string. In particular*

$$\mathbb{P}(1^m 0^n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \mathbb{P}(1^{m+k}). \tag{2.20}$$

Remark 2.2.1. *An exchangeable sequence gives the same probability to a string and to any of its permutations. Thus the probability of a binary string depends only on the number of 0s and 1s and not on the order of them. Thus equation (2.20) gives the probability of any permutation of the string $1^m 0^n$.*

We prove the formula in equation (2.20) by induction on (n, m), using the well-founded induction [1].

[1]**A note on well-founded induction**
We can define a total order on $\mathbb{N} \times \mathbb{N}$ in the following way

$$(m, n) \leq (m', n') \text{ if } \begin{cases} n < n' \text{ or} \\ n = n' \text{ and } m \leq m'. \end{cases} \tag{2.21}$$

Proof. By the total probabilities formula we have

$$\mathbb{P}(1) = \mathbb{P}(10) + \mathbb{P}(11), \tag{2.24}$$

and rearranging the terms

$$\mathbb{P}(10) = \mathbb{P}(1) - \mathbb{P}(11). \tag{2.25}$$

Thus equation (2.20) holds for $(m, n) = (1, 1)$. Suppose that it holds for any integer $(m, n) < (\bar{m}, \bar{n})$. In particular it holds for $(\bar{m}, \bar{n}-1)$ and $(\bar{m}+1, \bar{n}-1)$.

We write

$$(m, n) < (m', n') \text{ if } \begin{cases} n < n' \text{ or} \\ n = n' \text{ and } m < m'. \end{cases} \tag{2.22}$$

Note that any non empty subset of $\mathbb{N} \times \mathbb{N}$ has a minimal element according to the order defined in equation (2.21).
Let $S(m, n)$ be a property defined for any $(m, n) \in \mathbb{N} \times \mathbb{N}$. The following well known result holds

Theorem 2.2.1. *(well-founded induction) Assume that*

i) $S(1, 1)$ *holds,*

ii) if $S(m, n)$ *holds for all* $(m, n) < (\bar{m}, \bar{n})$*, then* $S(\bar{m}, \bar{n})$ *holds,*

then $S(m, n)$ *holds for all* $(m, n) \in \mathbb{N} \times \mathbb{N}$.

Proof. We sketch the proof arguing by contradiction. Let

$$S := \{(m, n) \in \mathbb{N} \times \mathbb{N} \text{ such that } S(m, n) \text{ holds.}\} \tag{2.23}$$

S is non empty by the first hypothesis. Let T be the complementary set of S in $\mathbb{N} \times \mathbb{N}$. The set T has a minimum element, call it $(\hat{m}, \hat{n})$, and obviously $(\hat{m}, \hat{n})$ is not in S. We have $(m, n) \in S$ for all $(m, n) < (\hat{m}, \hat{n})$. Thus by the second hypothesis $S(\hat{m}, \hat{n})$ holds, thus we find a contradiction. □

We have to show that it holds for $(\bar{m}, \bar{n})$. We have

$$
\begin{aligned}
\mathbb{P}(1^{\bar{m}}0^{\bar{n}}) &= \mathbb{P}(1^{\bar{m}}0^{\bar{n}-1}) - \mathbb{P}(1^{\bar{m}+1}0^{\bar{n}-1}) = (\text{ inductive hypothesis }) \\
&= \sum_{k=0}^{\bar{n}-1}(-1)^k\binom{\bar{n}-1}{k}\mathbb{P}(1^{\bar{m}+k}) - \sum_{k=0}^{\bar{n}-1}(-1)^k\binom{\bar{n}-1}{k}\mathbb{P}(1^{\bar{m}+k+1}) \\
&= \mathbb{P}(1^{\bar{m}}) + \left\{\sum_{k=0}^{\bar{n}-2}\left[(-1)^{k+1}\binom{\bar{n}-1}{k+1} - (-1)^k\binom{\bar{n}-1}{k}\right]\mathbb{P}(1^{\bar{m}+k+1})\right\} \\
&\quad - (-1)^{\bar{n}-1}\mathbb{P}(1^{\bar{m}+\bar{n}}) \\
&= \mathbb{P}(1^{\bar{m}}) + \left\{\sum_{k=0}^{\bar{n}-2}(-1)^{k+1}\binom{\bar{n}}{k+1}\mathbb{P}(1^{\bar{m}+k+1})\right\} + (-1)^{\bar{n}}\mathbb{P}(1^{\bar{m}+\bar{n}}) \\
&= \mathbb{P}(1^{\bar{m}}) + \left\{\sum_{k=1}^{\bar{n}-1}(-1)^{k}\binom{\bar{n}}{k}\mathbb{P}(1^{\bar{m}+k})\right\} + (-1)^{\bar{n}}\mathbb{P}(1^{\bar{m}+\bar{n}}) \\
&= \sum_{k=0}^{\bar{n}}(-1)^k\binom{\bar{n}}{k}\mathbb{P}(1^{\bar{m}+k}),
\end{aligned}
$$

where we have used in the fourth equality

$$
\begin{aligned}
(-1)^{k+1}\binom{\bar{n}-1}{k+1} - (-1)^k\binom{\bar{n}-1}{k} &= (-1)^{k+1}\binom{\bar{n}-1}{k+1} + (-1)^{k+1}\binom{\bar{n}-1}{k} \\
&= (-1)^{k+1}\binom{\bar{n}}{k+1}.
\end{aligned}
$$

□

The Hausdorff moments problem

We recall in the following Theorem 2.2.2 a classical result about the so called Hausdorff moments problem. Let us first fix some notations. Given a sequence of numbers $\alpha_0, \alpha_1, \alpha_2, \ldots$ the differencing operator Δ is defined by

$$\Delta\alpha_n := \alpha_{n+1} - \alpha_n.$$

The higher differences Δ^r are obtained recursively by the relation $\Delta^r = \Delta(\Delta^{r-1})$ where $\Delta^1 = \Delta$, (see [26] Chapter VII.1).
We say that a sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$ is *completely monotone* if for all r and all k we have $(-1)^r\Delta^r\alpha_k \geq 0$.

Theorem 2.2.2. *A sequence of numbers $\alpha_0, \alpha_1, \alpha_2, \ldots$ represents the moments sequence of some probability measure μ on $[0,1]$ if and only if*

- $\alpha_0 = 1$
- *the sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$ is completely monotone.*

For the proof of the theorem see [26], Chapter VII.3.

There is a nice relationship between completely monotone sequences and exchangeable binary sequences. As mentioned in Section 2.1.1, for a binary exchangeable sequence Y the probabilities $\mathbb{P}(1^m)$, $m = 1, 2, \ldots$, are the moments of the de Finetti measure μ associated with Y:

$$\mathbb{P}(1^m) = \int_0^1 p^m d\mu(p),$$

and thus the sequence $\left(\mathbb{P}(1^m)\right)_{m\geq 0}$ is completely monotone.
Let us consider now a completely monotone sequence $\alpha = (\alpha_m)_{m\geq 0}$. By Theorem 2.2.2 there exist a measure μ such that α is the sequence of the moments of the measure μ. We can associate to α an exchangeable sequence Y setting

$$\mathbb{P}(1^m) = \alpha_m = \int_0^1 p^m \mu(dp),$$

and then assigning the probability of any finite binary string as in the relation (2.20). But the complete monotonicity of the sequence $\left(\mathbb{P}(1^m)\right)_{m\geq 0}$ of the joint distributions of a sequence Y is not a sufficient condition for the exchangeability of Y. The following example will clarify the previous statement

Example 2.2.1. *Let for some $p \in [0,1]$*

$$\alpha_m := p^m \text{ for all } m \geq 0.$$

The sequence (α_m) is clearly completely monotone. We can construct an i.i.d sequence Y such that $\mathbb{P}(1^m) = p^m = \alpha_m$, and thus $\mathbb{P}(\sigma) = p^m(1-p)^n$, for any finite string σ with m 1s and n 0s. But we can also construct a sequence $\widetilde{Y}$ such that $\mathbb{P}^{\widetilde{Y}}(1^m) = p^m = \alpha_m$[2]*, but which is not an i.i.d. sequence. Take for example, for all $m \geq 0$*

$$\mathbb{P}^{\widetilde{Y}}(1^m) = p^m \Rightarrow \mathbb{P}^{\widetilde{Y}}(1^m 0) = p^m(1-p)$$
$$\mathbb{P}^{\widetilde{Y}}(1^m 0 \sigma) = p^m(1-p)/2^l,$$

[2] We write $\mathbb{P}^{\widetilde{Y}}(\cdot)$ just to recall we are giving the joint distribution of the sequence $\widetilde{Y}$.

where $l \geq 1$ is the length of the binary finite string σ. Thus $\widetilde{Y}$ is such that $\mathbb{P}^{\widetilde{Y}}(1^m) = p^m$, but it is not an i.i.d. sequence. For example $\mathbb{P}^{\widetilde{Y}}(01) = (1-p)/2 \neq p(1-p) = \mathbb{P}^{\widetilde{Y}}(10)$.

2.2.2 Stochastic realization

Let $Y = (Y_1, Y_2, \dots)$ be a binary exchangeable sequence. In this section we solve the *stochastic realization problem* for Y, i.e. given the joint distributions of the random variables in the sequence, we look for parameters $\theta = (\theta_1, \dots, \theta_M)$, which completely describe the probabilistic behavior of Y, if they do exist. We will make use of the results on the de Finetti mixing measure of a binary exchangeable sequence we have developed in the previous section. In particular we pose and solve the following problem.
Problem. (Stochastic realization of binary exchangeable processes)
Given the joint distributions $\mathbb{P}(1), \mathbb{P}(11), \dots, \mathbb{P}(1^n), \dots$ of a binary exchangeable sequence find, when it exists, a finite mixture of i.i.d. sequences, i.e. an integer N (possibly the smallest one) and parameters $(p_1, \dots, p_N; \mu_1, \dots, \mu_N)$, with $0 \leq p_1, \dots, p_N \leq 1$, $0 < \mu_1, \dots, \mu_N < 1$ and $\sum_{i=1}^N \mu_i = 1$ such that, for any binary string y_1^n,

$$\mathbb{P}(y_1^n) = \sum_{k=1}^{N} \mu_k p_k^{n_1} (1-p_k)^{n-n_1}, \tag{2.26}$$

with $n_1 := \sum_{t=1}^n y_t$.
Notice that we just require knowledge of $\mathbb{P}(1), \mathbb{P}(11), \dots, \mathbb{P}(1^n), \dots$, and not of all the joint distributions of Y, see Proposition 2.2.1.
Solution. To solve the problem, construct the matrices (H_n) defined in equation (2.7). Check whether Y is a finite mixture of i.i.d. sequences, looking at the rank of the matrices (H_n) (see Theorem 2.1.1). If Y is not a finite mixture of i.i.d. sequences, by the uniqueness of the de Finetti measure, Y can not be realized as in equation (2.26). If Y is a mixture of N i.i.d. sequences, compute the de Finetti mixing measure $\mu = (p_1, \dots, p_N; \mu_1, \dots, \mu_N)$ associated to Y using the algorithm developed in Section 2.1.2.

We can completely solve the stochastic realization problem for finite mixtures of binary i.i.d. sequences. It could thus be of interest to find a good approximation of a *not finite* mixture of i.i.d. sequences, i.e. of a general exchangeable sequence, with a *finite* mixture of i.i.d. sequences. We leave the problem for further investigations.

2.2.3 Positive realization of linear systems

The results of Section 2.1 enable us to solve a classical problem in linear systems theory in a special case. More precisely, in Theorem 2.2.3 we show that for transfer functions in a special class the rank of the Hankel matrix coincides with the order of positive realization. The theorem is proved constructively, exhibiting a positive realization.
Before stating the main theorem, we recall some definitions and some well known results. Let $\big(u(t)\big)_{t\in\mathbb{N}}$ be an i.i.d. sequence. Given the linear system

$$\begin{cases} x(t+1) &= Ax(t) + bu(t) \\ y(t) &= c^T x(t), \end{cases} \tag{2.27}$$

the *transfer function* associated to the system is defined as

$$G(z) := c^T(z\mathbb{I} - A)^{-1}b,$$

with $\mathbb{I}$ denoting the identity matrix.
The *realization problem* is the following: given a rational function $G(z)$

$$G(z) = \frac{q_{n-1}z^{n-1} + \cdots + q_0}{z^n + p_{n-1}z^{n-1} + \cdots + p_0} = \sum_{k\geq 0} g_k z^{-(k+1)}, \tag{2.28}$$

find the minimal N and a triple (A, b, c), where $A \in \mathbb{R}^{N\times N}$ and $b, c \in \mathbb{R}^N$ such that

$$G(z) = c^T(z\mathbb{I} - A)^{-1}b. \tag{2.29}$$

The minimal N is called the *order* of the realization. It is well known that if the polynomials $P(z) = z^n + p_{n-1}z^{n-1} + \cdots + p_0$ and $Q(z) = q_{n-1}z^{n-1} + \cdots + q_0$ are coprime, then $N = n$. The function $G(z)$ defined as in equation (2.28) satisfies equation (2.29), if and only if for any $k \geq 0$ we have

$$g_k = c^T A^k b. \tag{2.30}$$

Given a rational function $G(z)$ as in equation (2.28), to solve the *positive realization problem* means to find a triple (A, b, c) that satisfies equation (2.29) with the additional constrain that (A, b, c) have nonnegative entries.

For a linear system with rational transfer function as in equation (2.28), define the Hankel matrix

$$H_\infty := \begin{pmatrix} g_0 & g_1 & g_2 & \cdots \\ g_1 & g_2 & g_3 & \cdots \\ g_2 & g_3 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.31}$$

If (A, b, c) is a triple which realizes the system, equation (2.30) holds. Thus we have

$$H_\infty := \begin{pmatrix} c^T b & c^T A b & c^T A^2 b & \dots \\ c^T A b & c^T A^2 b & c^T A^3 b & \dots \\ c^T A^2 b & c^T A^3 b & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$
$$= \begin{pmatrix} c^T \\ c^T A \\ c^T A^2 \\ \vdots \end{pmatrix} \begin{pmatrix} b & Ab & A^2 b & \dots \end{pmatrix}.$$

Definition 2.2.1. *The impulse response (g_k) is of the relaxation type if it is a completely monotone sequence (see the definition below), with $g_1 = 1$.*

For the notion of linear systems of the relaxation type in continuous time (without positivity constraints on the realization) the reader is referred to e.g. [52]. We are not aware of previous work on discrete time *positive* systems of the relaxation type.

We can state the following

Theorem 2.2.3. *Let $G(z)$ be a rational transfer function defined as in equation (2.28) of the relaxation type. Let $H_0, H_1, \dots$ be the principal submatrices of the Hankel matrix H_∞ defined in equation (2.31). Let*

$$rank(H_n) = \begin{cases} n+1 & \text{for } n = 0, \dots, N-1 \\ N & \text{for } n \geq N. \end{cases}$$

Then there exist $0 < p_1 < \dots, < p_N < 1$ and $0 < \mu_1, \dots, \mu_N < 1$, with $\sum_{i=1}^{n} \mu_i = 1$, such that

$$g_k = \sum_{i=1}^{N} \mu_k p_i^k. \tag{2.32}$$

Moreover

$$A := \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & \dots & \dots & p_N \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad c = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix}.$$

is a positive realization for $G(z)$.

Proof. $g_0 = 1$ and the sequence $g_0, g_1, g_2, \dots$ is completely monotone, thus by Theorem 2.2.2, the sequence $g_0, g_1, g_2, \dots$ is the moments sequence of some probability measure μ on $[0, 1]$, i.e.

$$g_k = \int_0^1 p^k \mu(dp),$$

for any $k \geq 0$. Thus g_k can be interpreted as the probability $\mathbb{P}(1^k)$ of a suitable exchangeable sequence Y, with de Finetti's associated measure μ.
The matrices $H_0, H_1, \dots$ satisfy the rank condition in equation (2.8), thus by Theorem 2.1.1 the measure μ is concentrated on N points. Call them $p_1, p_2, \dots, p_N$ and let $\mu_1, \mu_2, \dots, \mu_N$ be the corresponding weights, $\mu_i = \mu(p_i)$, as computed in Section 2.1.2. We can write

$$g_k = \int_0^1 p^k \mu(dp) = \sum_{i=1}^{N} \mu_k p_i^k. \tag{2.33}$$

Define

$$A := \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & \dots & \dots & p_N \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad c = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix}.$$

(A, b, c) is a positive realization for $G(z)$. In fact, for any $k \geq 0$, we have

$$c^T A^k b = \sum_{i}^{N} \mu_i p_i^k = g_k. \tag{2.34}$$

□

A well known result in linear system theory states that the rank of the Hankel matrix H_∞ coincides with the order of a realization. This is not the case for positive realization, where the rank of the Hankel matrix is only a lower bound to the order of positive realization. Benvenuti and Farina in [9] and [10] find sufficient conditions to achieve the lower bound.
Note that linear systems as in Theorem 2.2.3 have order of positive realization coinciding with rank of the Hankel matrix H_∞.

2.3 Mixing measure of Markov mixtures

In this section we propose a criterion to check whether a binary mixture of Markov chains is a *finite* mixture of Markov chains. To state the criterion, we need some assumptions and some preliminary definitions.
Following Definition 1.2.5, for a binary mixture of Markov chains we can write

$$\mathbb{P}\{Y_1^n = y_1^n\} = \int_{\mathcal{P}} \prod_{t=1}^{n-1} P_{y_t, y_{t+1}} \mu(y_1, dP), \tag{2.35}$$

where $\mathcal{P}$ is the set of stochastic matrices of dimension 2 and μ is a probability measure on $\{0,1\} \times \mathcal{P}$. Assume that the random choice of the transition matrices in the mixture is independent from the initial condition. More precisely, assume

Condition 2.3.1. *μ factorizes as*

$$\mu(y_1, P) = \tilde{\mu}(y_1)\bar{\mu}(P),$$

where $\tilde{\mu}(\cdot)$ is a measure on $\{0,1\}$ and $\bar{\mu}(\cdot)$ is a measure on $\mathcal{P}$.

Recall that the stochastic matrices of dimension 2 are of the form

$$P = \begin{pmatrix} P_{00} & 1 - P_{00} \\ 1 - P_{11} & P_{11} \end{pmatrix},$$

thus the elements $P \in \mathcal{P}$ are completely characterized by the two numbers (P_{00}, P_{11}) and there is a one-to-one correspondence between the set $\mathcal{P}$ and the unit square $[0,1] \times [0,1]$ (see Section 1.2.3). Let us fix some notations. For any element $A \in \mathcal{B}([0,1])$, define

$$S_A^{(0)} := \{P \in \mathcal{P} \mid P_{00} \in A\}, \quad \text{and} \quad S_A^{(1)} := \{P \in \mathcal{P} \mid P_{11} \in A\}.$$

We give the following

Definition 2.3.1. *For $i = 0, 1$, let $\nu^{(i)}$ be the function on $[0,1]$ defined as*

$$\nu^{(i)}(A) := \bar{\mu}(S_A^{(i)}). \tag{2.36}$$

Lemma 2.3.1. *For $i = 0, 1$, the function $\nu^{(i)}$ is a probability measure.*

Proof. We have to check that the null set has null measure:

$$\nu^{(i)}(\emptyset) = \bar{\mu}(S_{\emptyset}^{(i)}) = \bar{\mu}(\emptyset) = 0.$$

Moreover let $A_1, A_2, \dots$ be disjoint sets, we have to check that $\nu^{(i)}(\bigcup_k A_k) = \sum_k \nu^{(i)}(A_k)$:

$$\nu^{(i)}\Big(\bigcup_k A_k\Big) = \bar{\mu}(S_{\bigcup_k A_k}^{(i)}) = \bar{\mu}\Big(\bigcup_k S_{A_k}^{(i)}\Big) = \sum_k \bar{\mu}(S_{A_k}^{(i)}) = \sum_k \nu^{(i)}(A_k).$$

□

Let $H_n^{(0)} = (h_{ij}^{(0)})_{0\le i,j\le n}$ be the $(n+1)\times(n+1)$ Hankel matrix with entries $h_{ij}^{(0)} := h_{i+j}^{(0)} = \mathbb{P}(0^i 0^j) = \mathbb{P}(0^{i+j})$

$$\mathbf{H_n}^{(0)} := \begin{pmatrix} \mathbb{P}(0) & \mathbb{P}(00) & \dots & \mathbb{P}(0^{n+1}) \\ \mathbb{P}(00) & \mathbb{P}(000) & \dots & \mathbb{P}(0^{n+2}) \\ \mathbb{P}(000) & \mathbb{P}(0000) & \dots & \mathbb{P}(0^{n+3}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{P}(0^{n+1}) & \mathbb{P}(0^{n+2}) & \dots & \mathbb{P}(0^{2n+1}) \end{pmatrix}, \tag{2.37}$$

and let $H_n^{(1)} = (h_{ij}^{(1)})_{0\le i,j\le n} = (h_{i+j}^{(1)})_{0\le i,j\le n} = \mathbb{P}(1^{i+j})$

$$\mathbf{H_n}^{(1)} := \begin{pmatrix} \mathbb{P}(1) & \mathbb{P}(11) & \dots & \mathbb{P}(1^{n+1}) \\ \mathbb{P}(11) & \mathbb{P}(111) & \dots & \mathbb{P}(1^{n+2}) \\ \mathbb{P}(111) & \mathbb{P}(1111) & \dots & \mathbb{P}(1^{n+3}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{P}(1^{n+1}) & \mathbb{P}(1^{n+2}) & \dots & \mathbb{P}(1^{2n+1}) \end{pmatrix}. \tag{2.38}$$

The next lemma establishes a connection between the moments matrix of $\nu^{(i)}$ and the matrices $H_n^{(i)}$, for a mixture of Markov chains Y, for $i = 0, 1$.

Lemma 2.3.2. *Let Y be a binary mixture of Markov chains and let the associated mixing measure μ satisfy Condition 2.3.1. Up to a multiplying constant, $\mathbf{H_n}^{(i)}$ is the moments matrix of the measure $\nu^{(i)}$, for $i = 0, 1$.*

Proof. We prove the statement for $i = 1$.

$$\begin{aligned}\mathbb{P}(1^m) &= \int_{\mathcal{P}} P_{11}^{m-1}\mu(1, dP) = \int_{\mathcal{P}} P_{11}^{m-1}\tilde{\mu}(1)\bar{\mu}(dP) \\ &= \tilde{\mu}(1)\int_{\mathcal{P}} P_{11}^{m-1}\bar{\mu}(dP) = \tilde{\mu}(1)\int_{[0,1]} q^{m-1}\nu^{(1)}(dq),\end{aligned}$$

where we have put $P_{11} = q$. Thus

$$\mathbf{H_n}^{(1)} = \tilde{\mu}(1) \begin{pmatrix} 1 & \int_{[0,1]} q\nu^{(1)}(dq) & \dots & \int_{[0,1]} q^n\nu^{(1)}(dq) \\ \int_{[0,1]} q\nu^{(1)}(dq) & \int_{[0,1]} q^2\nu^{(1)}(dq) & \dots & \int_{[0,1]} q^{n+1}\nu^{(1)}(dq) \\ \vdots & \vdots & \vdots & \vdots \\ \int_{[0,1]} q^n\nu^{(1)}(dq) & \int_{[0,1]} q^{n+1}\nu^{(1)}(dq) & \dots & \int_{[0,1]} q^{2n}\nu^{(1)}(dq) \end{pmatrix}.$$

□

We indicate with $F^{(\nu^i)}$ the distribution function of $\nu^{(i)}$, as in equation (2.3).

Lemma 2.3.3. *If $F^{(\nu^i)}$ has N points of increase, then $det(\mathbf{H}_\mathbf{n}^{(\mathbf{i})}) \neq 0$ for $n = 0, \dots, N-1$ and $det(\mathbf{H_n}^{(i)}) = 0$ for $n \geq N$.*
If $F^{(\nu^i)}$ has infinitely many points of increase, then $det(\mathbf{H_n}^{(i)}) \neq 0$ for any n.

Proof. It is just an easy consequence of Lemma 2.3.2 and of Lemma 2.1.1. □

Definition 2.3.2. *For $i = 0, 1$, let r_i be the first integer n such that $det(\mathbf{H_n}^{(i)}) = 0$. If $det(\mathbf{H_n}^{(i)}) \neq 0$ for any n, put $r_i = +\infty$.*

Lemma 2.3.4. *r_i is the number of points of increase of $F^{\nu^{(i)}}$.*

Proof. It follows trivially by Lemma 2.3.3 and by Definition 2.3.2. □

In the following lemma we figure out the relationship between the concentration point sets $C_{\nu^{(0)}}$ and $C_{\nu^{(1)}}$, and C_μ. We will make use of Lemma 2.3.4 to prove the main theorem of this section, given as Theorem 2.3.1 below.
Let $proj_0$ and $proj_1$ denote the projection from $\mathbb{R}^2$ on the first and on the second coordinate respectively. With a small abuse of notations, we will often indicate with $\bar{\mu}$ the measure on $[0,1] \times [0,1]$ induced by the measure $\bar{\mu}$ on $\mathcal{P}$.

Lemma 2.3.5. *$\bar{\mu}$ is concentrated on a finite set if and only if $\nu^{(0)}$ and $\nu^{(1)}$ are both concentrated on finite sets.*

Proof. Let $\bar{\mu}$ be concentrated on the set of points (P_{00}^k, P_{11}^k), $k = 1, \dots, K$. By definition, $\nu^{(0)}(\{P_{00}^1, \dots, P_{00}^K\}) = \bar{\mu}(S_{\{P_{00}^1,\dots,P_{00}^K\}}) = 1$ and $\nu^{(1)}(\{P_{11}^1, \dots, P_{11}^K\}) = \bar{\mu}(S_{\{P_{11}^1,\dots,P_{11}^K\}}) = 1$, thus $\nu^{(0)}$ and $\nu^{(1)}$ are concentrated.
Conversely let $\nu^{(0)}$ and $\nu^{(1)}$ be concentrated on $\{P_{00}^1, \dots, P_{00}^{N_0}\}$ and $\{P_{11}^1, \dots, P_{11}^{N_1}\}$ respectively. By definition $\nu^{(0)}(\{P_{00}^1, \dots, P_{00}^{N_0}\}) = \bar{\mu}(S^{(0)}_{\{P_{00}^1,\dots,P_{00}^{N_0}\}}) = 1$ and $\nu^{(1)}(\{P_{11}^1, \dots, P_{11}^{N_1}\}) = \bar{\mu}(S^{(1)}_{\{P_{11}^1,\dots,P_{11}^{N_1}\}}) = 1$.

Thus $\bar{\mu}\Big(S^{(0)}_{\{P^1_{00},\dots,P^{N_0}_{00}\}}\bigcap S^{(1)}_{\{P^1_{11},\dots,P^{N_1}_{11}\}}\Big) = 1$. But $S^{(0)}_{\{P^1_{00},\dots,P^{N_0}_{00}\}}\bigcap S^{(1)}_{\{P^1_{11},\dots,P^{N_1}_{11}\}}$ is constituted by a finite number of points, thus $\bar{\mu}$ is concentrated indeed. □

Lemma 2.3.6. *Let $\nu^{(0)}$ and $\nu^{(1)}$ be concentrated. Then the following relations between $C_{\nu^{(0)}}$, $C_{\nu^{(1)}}$ and $C_{\bar{\mu}}$ hold*

- $C_{\nu^{(0)}} = proj_0(C_{\bar{\mu}})$ *and* $C_{\nu^{(1)}} = proj_1(C_{\bar{\mu}})$,
- $C_{\bar{\mu}} \subseteq C_{\nu^{(0)}} \times C_{\nu^{(1)}}$.

Proof. Let $p \in C_{\nu^{(0)}}$. We have $\nu^{(0)}(p) = \bar{\mu}(S^{(0)}_p) > 0$. Notice that by Lemma 2.3.5 the measure $\bar{\mu}$ is concentrated on a finite set, and $\bar{\mu}(S^{(0)}_p) > 0$, thus there are a finite number of points $(p, P^1_{11}), \dots, (p, P^{N_1}_{11})$ such that $\bar{\mu}(p, P^i_{11}) > 0$ for $i = 1, \dots, N_1$ and $\sum_{i=1}^{N_1} \bar{\mu}(p, P^i_{11}) = \bar{\mu}(S^{(0)}_p)$. Thus $p \in proj_0(C_{\bar{\mu}})$, and we get $C_{\nu^{(0)}} \subseteq proj_0(C_{\bar{\mu}})$.
$\bar{\mu}$ is concentrated on a finite set, thus $C_{\bar{\mu}} := \{(P^1_{00}, P^1_{11}), \dots, (P^N_{00}, P^N_{11})\}$, with $\bar{\mu}((P^i_{00}, P^i_{11})) > 0$ for $i = 1, \dots, N$ and $\sum_{i=1}^N \bar{\mu}((P^i_{00}, P^i_{11})) = 1$. Thus $proj_0(C_{\bar{\mu}})$ is a finite number of points, call them $\{P^1_{00}, \dots, P^{N_1}_{00}\}$, and for $i = 1, \dots, N_1$ we have $\nu^{(0)}(P^i_{00}) = \bar{\mu}(S^{(0)}_{P^i_{00}}) = \bar{\mu}\big((P^j_{00}, P^j_{11}) \in C_{\bar{\mu}} \text{ s.t. } P^j_{00} = P^i_{00}\big)$. Thus $\nu^{(0)}(P^i_{00}) > 0$ and $P^i_{00} \in C_{\nu^{(0)}}$, and we get $proj_0(C_{\bar{\mu}}) \subseteq C_{\nu^{(0)}}$. Arguing in the same way for $\nu^{(1)}$, the first statement of the lemma is proved indeed.
Let $(P_{00}, P_{11}) \in C_{\bar{\mu}}$. Thus $\nu^{(0)}(P_{00}) = \bar{\mu}(S^{(0)}_{P_{00}}) \geq \bar{\mu}(P_{00}, P_{11}) > 0$, and $\nu^{(1)}(P_{11}) = \bar{\mu}(S^{(1)}_{P_1}) \geq \bar{\mu}(P_{00}, P_{11}) > 0$. Thus we get $P_{00} \in C_{\nu^{(0)}}$ and $P_{11} \in C_{\nu^{(1)}}$, i.e. $(P_{00}, P_{11}) \in C_{\nu^{(0)}} \times C_{\nu^{(1)}}$. This concludes the proof of the lemma. □

Theorem 2.3.1. *Y is a finite mixture of Markov chains if and only if*

$$\max\{r_0, r_1\} < +\infty.$$

If Y is a mixture of N Markov chains, then

$$\max\{r_0, r_1\} \leq N \leq r_0\, r_1. \tag{2.39}$$

Proof. We indicate with $\sharp$ the cardinality of a set. Let Y be a mixture of N Markov chains. Thus $\bar{\mu}$ is concentrated on N points, and by Lemma 2.3.5 the measure $\nu^{(0)}$ and $\nu^{(1)}$ are concentrated. The points of increase of $F^{\nu^{(i)}}$ therefore coincide with $C_{\nu^{(i)}}$ and we get for $i = 0, 1$

$$r_i = \sharp C_{\nu^{(i)}} = \sharp proj_i(C_{\bar{\mu}}) \leq \sharp C_{\bar{\mu}} = N.$$

Thus $\max\{r_0, r_1\} \le N < +\infty$.
Let $\max\{r_0, r_1\} < +\infty$. The functions $F^{\nu^{(0)}}$ and $F^{\nu^{(1)}}$ have a finite number of points of increase, and thus by Remark 2.1.2 the measures $\nu^{(0)}$ and $\nu^{(1)}$ are concentrated. By Lemma 2.3.5 the measure $\bar{\mu}$ is concentrated and Y is a finite mixture of Markov chains.

Let Y be a mixture of N Markov chains. $\bar{\mu}$ is thus concentrated on N points. We have just shown that $\max\{r_0, r_1\} \le N$. To prove the second inequality in equation (2.39), notice that by Lemma 2.3.5 the measures $\nu^{(0)}$ and $\nu^{(1)}$ are concentrated, and thus $r_0 = \sharp C_{\nu^{(0)}}$ and $r_1 = \sharp C_{\nu^{(1)}}$. By the second statement of Lemma 2.3.6 it follows

$$N = \sharp C_{\bar{\mu}} \le \sharp\{C_{\nu^{(0)}} \times C_{\nu^{(1)}}\} = r_0 r_1.$$

□

2.4 Future work

A natural extension of the results of Section 2.1 would be the generalization to exchangeable sequences (Y_n) with more general state spaces. Trying to go from binary to finite state space, we encounter the same technical difficulties we found in the extension from exchangeable to Markov exchangeable binary sequences of the algorithm identifying the mixing measure μ. The one to one correspondence between the moments matrices of the mixing measure and the Hankel matrices (H_n) is lost. For exchangeable sequences with finite state space and for Markov exchangeable sequences, the Hankel matrices (H_n) are, up to a multiplicative constant, the moments matrices of a marginalization of the mixing measure. The natural tool to attack both cases seems to be the Hausdorff moment problem on the unit hypercube.

Chapter 3

Representations of countable Markov mixtures

In Chapter 1 we have defined mixtures of Markov chains, and we have recalled the de Finetti-type representation theorems for mixtures of Markov chains, due to Diaconis and Freedman for Markov exchangeable sequences, and to Fortini *et al.* for partially exchangeable sequences. In the present chapter we look for a representation theorem for *countable* mixtures of Markov chains, i.e. for mixtures with concentrated mixing measure. In [21] Dharmadhikari solves the analogous problem for mixtures of i.i.d. sequences. More precisely, he shows that an exchangeable sequence is a mixture of a *countable* number of i.i.d. sequences if and only if it is an Hidden Markov Model. In this chapter we extend Dharmadhikari's result for exchangeable sequences to Markov exchangeable, k-Markov exchangeable and partially exchangeable sequences. To get information on the mixing measure of a mixture, looking at some characteristic of the mixture itself is a challenging mathematical problem. Moreover our result implies a characterization of *finite* mixtures of Markov chains and, as we have seen in Section 1.2.2, these are useful models in many different applications. The other side of our problem, that is to characterize mixtures of i.i.d. sequences or of Markov chains with absolutely continuous mixing measure, is still an open problem up to our knowledge.

In Section 3.3.1 we prove a slight extension of the Diaconis and Freedman result to k-Markov exchangeable sequences, stated in Section 1.2.4 as Theorem 1.2.4.

3.1 A known result

Dharmadhikari characterizes the countable mixtures of i.i.d. sequences, linking countable HHMs to the class of exchangeable sequences. More precisely in [21] he proves the following

Theorem 3.1.1. *Let $Z = (Z_1, Z_2, \dots)$ be an exchangeable sequence taking values on a countable set. The sequence Z is a* countable *HMM with stationary underlying Markov chain if and only if Z is a* countable *mixture of i.i.d. sequences.*

The main results of Chapter 3 are the generalization of Dharmadhikari's result to Markov exchangeable sequences, to k-Markov exchangeable sequences and to partially exchangeable sequences (see Sections 3.2, 3.3 and 3.4 respectively).

3.2 Countable mixtures of Markov chains

3.2.1 1-blocks

We recall the definition of the 1-block sequence Z of a given sequence Y. The introduction of the 1-block sequence, used first in [22], is a technical trick that allows to reduce the study of Markov exchangeable sequences to that of exchangeable sequences. Let J be a countable set, assume that $J \subseteq \{1, 2, 3, \dots\}$.

Definition 3.2.1. *A* 1-block *is a string of letters from J which begins with 1 and contains no further 1's.*

Let $Y = (Y_1, Y_2, \dots)$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in J, and let Y be recurrent. From now on, in this section, we suppose that the following condition is satisfied:

Condition 3.2.1. $\mathbb{P}\{Y_1 = 1\} = 1$.

Thus we could write $\mathbb{P}\{\cdot \mid Y_1 = 1\}$ instead of $\mathbb{P}\{\cdot\}$, since these probabilities coincide under Condition 3.2.1.

The sequence $Z = (Z_1, Z_2, \dots)$ of the successive 1-blocks of Y is well defined. In fact Y starts at 1 with probability 1 by Condition 3.2.1, moreover

Y is recurrent, thus its successive 1-blocks are *a.s.* of finite length. Z is a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values on J^*, which is countable.
Let $z_1^n = z_1\, z_2 \ldots z_n$ be a string of 1-blocks, $z_i \in J^*$, say $z_i = 1y_{2i}y_{3i}\ldots y_{l_i i}$, where l_i is the length of z_i and $y_{ji} \in J$. A 1-block z_i is a string of letters from J, thus a string z_1^n of 1-blocks is still a finite string of letters from J, i.e.

$$z_1^n = 1y_{21}y_{31}\ldots y_{l_1 1}1y_{22}y_{32}\ldots y_{l_2 2}\ldots\ldots y_{l_n n} =: y_1^m.$$

We indicate the same string with the two different notations z_1^n and y_1^m, where $m = \sum_{i=1}^n l_i$. The event $\{Z_1^n = z_1^n\}$ is equivalent to the event $\{Y_1^m = y_1^m;\, Y_{m+1} = 1\}$.

To derive the law of Z, we introduce the sequence $\tau_1, \tau_2, \ldots$ of random times defined recursively by

$$\tau_1 = 1, \qquad \tau_n := \inf_t\{t > \tau_{n-1} \mid Y_t = 1\}. \tag{3.1}$$

Y hits the letter 1 the n-th time at τ_n. Let $\mathcal{F}_m$ be the σ-field generated by $Y_1, Y_2, \ldots, Y_m$. The event $\{\tau_n = m\} \in \mathcal{F}_m$ for any n and m, therefore $\tau_1, \tau_2, \ldots$ are stopping times w.r.t. the sequence Y.
The law of Z is easily given as a function of the law of Y: for any 1-block z of length l

$$\{Z_m = z\} = \{Y_{\tau_m}^{\tau_m+l-1} = z;\, \tau_{m+1} = \tau_m + l\} = \{Y_{\tau_m}^{\tau_m+l-1} = z;\, Y_{\tau_m+l} = 1\},$$

and for any sequence of 1-blocks z_1^n of lengths $l_1, \ldots, l_n$ respectively, we have

$$\begin{aligned}&\{Z_m^{m+n-1} = z_1^n\} = \\ &\{Y_{\tau_m} = 1, \ldots, Y_{\tau_m+l_1-1} = y_{l_1 1}, \ldots Y_{\tau_{(m+n-1)}+l_n-1} = y_{l_n n};\\ &\qquad\qquad \tau_{m+1} = \tau_m + l_1, \ldots, \tau_{m+n} = \tau_{(m+n-1)} + l_n\}.\end{aligned}$$

For the purpose of proving the main Theorem 3.2.1 below we assume without loss of generality that Y is the coordinate process, i.e. that $\Omega = J^\infty$ and that $Y_n(\omega) = \omega_n$, with $\omega \in \Omega$. In this case, by definition, $\{Y_1^n = y_1^n\} = \{y_1^n\}$ and thus $\mathbb{P}\{Y_1^n = y_1^n\} = \mathbb{P}\{y_1^n\}$, and $\{Z_1^n = z_1^n\} = \{z_1^n\}$ and thus $\mathbb{P}\{Z_1^n = z_1^n\} = \mathbb{P}\{z_1^n\}$.

A simple, very useful, property of the 1-blocks of Markov exchangeable sequences is recalled below (see [22]).

Proposition 3.2.1. *Let Y be a Markov exchangeable recurrent sequence satisfying Condition 3.2.1. The sequence Z of the 1-blocks of Y is exchangeable.*

Proof. Permutations of 1-blocks do not change transition counts nor the initial state, therefore they produce strings that are transition equivalent. If Y is Markov exchangeable it gives the same measure to transition equivalent strings. The exchangeability of Z follows immediately.

□

Remark 3.2.1. *All the definitions of this section are given under Condition 3.2.1, but it is possible to fix the initial state of Y at any $i \in J$ and use the same arguments to introduce in the obvious way the i-blocks's sequence. Proposition 3.2.1 and the other observations are still valid for i-blocks.*

Note that exchangeability of the 1-blocks is not a sufficient condition for a sequence Y to be Markov exchangeable.

Counterexample 3.2.1. Let us consider the binary case, $J = \{0, 1\}$. If Y is Markov exchangeable the sequence of 1-blocks, as well as the sequence of 0-blocks, are exchangeable. It is not true that if the 1-blocks alone or the 0-blocks alone are exchangeable then Y is Markov exchangeable. This is due to the fact that there are strings which are transition equivalent but that are not obtained permuting just 1-blocks or just 0-blocks. As an example the strings 0110100010110111110001000 and 0100011101100001110100010 are transition equivalent, but cannot be transformed one into the other just permuting 0-blocks (for the connection between the equivalence relation $\sim$ and block-switch transformations see [22] page 124).

An example of binary sequence with exchangeable 1-blocks which is not necessarily Markov exchangeable is the HHM constructed as follows. Let X be a Markov chain with state space χ, and let $Y_n := f(X_n)$. Define the deterministic function f to be $f(x_0) = 1$ for some $x_0 \in \chi$ and $f(x) = 0$ for $x \neq x_0$. In this case the HMM Y is a binary renewal process [38]. It is easy to check that Y gives the same measure to strings obtained permuting 1-blocks, but Y is not necessarily Markov exchangeable.

3.2.2 The main result

We are now ready to state the main result of this section, extending Dharmadhikari theorem to Markov exchangeable sequences.

Theorem 3.2.1. *Let $Y = (Y_1, Y_2, \dots)$ be a Markov exchangeable recurrent sequence taking values in a countable set. The sequence Y is a* countable *RHMM if and only if Y is a* countable *mixture of recurrent Markov chains.*

Remark 3.2.2. *In [27] Fortini* et al. *prove that a recurrent Markov exchangeable sequence is strongly recurrent. So Y in Theorem 3.2.1 is actually strongly recurrent.*

Remark 3.2.3. *If Y is a mixture of Markov chains, Y is recurrent if and only if Y is a mixture of recurrent Markov chains. Moreover a countable HMM Y with recurrent underlying Markov chain X is recurrent, so the hypothesis of the recurrence of Y in Theorem 3.2.1 is actually redundant.*

Proof of the necessity

We prove first the necessity part of Theorem 3.2.1, i.e. that if a Markov exchangeable recurrent sequence Y is a *countable* RHMM, then Y is a *countable* mixture of Markov chains, assuming without loss of generality Condition 3.2.1 and that Y is the coordinate process.
In [22] Diaconis and Freedman prove that a Markov exchangeable recurrent sequence Y is a *general* mixture of Markov chains. It is useful to recall the skeleton of the proof in [22]. Let Z be the 1-blocks sequence of Y. The sequence Z is exchangeable (see Proposition 3.2.1 above) so, by de Finetti's theorem (see Theorem 1.2.1), we can write

$$\mathbb{P}\{z_1^n\} = \int P_\omega(z_1^n)\mathbb{P}(d\omega), \tag{3.2}$$

where Z is an i.i.d. sequence with respect to the regular conditional probability P_ω for $\mathbb{P}$-almost all ω. Then Diaconis and Freedman prove that, if the sequence Z of the 1-blocks is itself i.i.d. with respect to some measure, then the sequence Y is a Markov chain with respect to the same measure. The precise proposition is as follows

Proposition 3.2.2. *(see [22], Proposition (15)) Let Y be a Markov exchangeable recurrent sequence satisfying Condition 3.2.1. If the 1-blocks of Y are i.i.d., then Y is an homogeneous Markov chain.*

By definition of regular conditional probability we can write

$$\mathbb{P}\{y_1^m\} = \int P_\omega(y_1^m)\mathbb{P}(d\omega), \tag{3.3}$$

where P_ω is the same regular conditional probability which appears in equation (3.2). Z is an i.i.d. sequence with respect to P_ω for $\mathbb{P}$-almost all ω, thus by Proposition 3.2.2 the sequence Y is a Markov chain with respect to P_ω for $\mathbb{P}$-almost all ω. Letting $\tilde{P}(\omega)$ denote the corresponding transition probabilities matrix, write

$$\mathbb{P}\{y_1^m\} = \int \prod_{t=1}^{m-1} \tilde{P}_{y_t,y_{t+1}}(\omega)\mathbb{P}(d\omega), \tag{3.4}$$

which is equation (1.7) when $\mathbb{P}\{Y_1 = 1\} = 1$. Thus Diaconis and Freedman conclude that Y is a mixture of Markov chains.

For our purpose, let us suppose that Z is an exchangeable sequence that is a *countable* mixture of i.i.d. sequences. Then we can write equation (3.2) with a sum instead of the integral

$$\mathbb{P}\{z_1^n\} = \sum_{k\in K} \mu_k p_k(z_1^n), \tag{3.5}$$

where the sequence Z is i.i.d. with respect to each p_k and K is a countable set. Thus equation (3.3) becomes

$$\mathbb{P}\{y_1^m\} = \sum_{k\in K} \mu_k p_k(y_1^m). \tag{3.6}$$

By the previous arguments Y is a Markov chain with respect to p_k for each k. Letting $\tilde{P}_{i,j}^k$ denote the corresponding transition probability, we get

$$\mathbb{P}\{y_1^m\} = \sum_{k\in K} \mu_k \prod_{t=1}^{m-1} \tilde{P}_{y_t,y_{t+1}}^k, \tag{3.7}$$

which is equation (1.7) when $\mathbb{P}\{Y_1 = 1\} = 1$, with a countable sum replacing the integral. Thus if Z is a countable mixture of i.i.d. sequences, Y is a countable mixture of Markov chains.

Hence, to obtain the necessity part of Theorem 3.2.1, it remains to show that if a Markov exchangeable recurrent sequence Y is a countable RHMM, then the 1-blocks sequence Z of Y is a countable mixture of i.i.d. sequences. This is achieved in Lemma 3.2.2 and Lemma 3.2.3 below. To prove Lemma 3.2.2 we need the following

Lemma 3.2.1. *Let* $(X_n)_{n\geq 1}$ *be a Markov chain with state space* χ. *Let* $S \subseteq \chi$. *Define recursively the sequence of random times* $(\tau_n)_{n\in\mathbb{N}}$ *by*

$$\tau_1 = 1, \qquad \tau_n := \inf_t\{t > \tau_{n-1} \mid X_t \in S\}.$$

Then

$$\mathbb{P}\{X_{\tau_n} = x \mid X_{\tau_{n-1}} X_{\tau_{n-2}} \dots X_{\tau_1}\} = \mathbb{P}\{X_{\tau_n} = x \mid X_{\tau_{n-1}}\}.$$

Proof. *of Lemma 3.2.1* $(\tau_n)_{n\in\mathbb{N}}$ is a sequence of stopping times with respect to the σ-field generated by X. Noting that, given $X_{\tau_{n-1}}$, τ_n depends only on $(X_{\tau_{n-1}+1}, X_{\tau_{n-1}+2}, \dots, X_{\tau_n-1})$ and recalling Theorem 1.24 in Chapter 5 of [16] (i.e. the strong Markov property), we get

$$\begin{aligned}
&\mathbb{P}\{X_{\tau_n} = x \mid X_{\tau_{n-1}} X_{\tau_{n-2}} \dots X_{\tau_1}\} \\
&= \sum_{s=1}^{+\infty} \mathbb{P}\{\tau_n = \tau_{n-1} + s \quad X_{\tau_{n-1}+s} = x \mid X_{\tau_{n-1}} X_{\tau_{n-2}} \dots X_{\tau_1}\} \\
&= \sum_{s=1}^{+\infty} \mathbb{P}\{\tau_n = \tau_{n-1} + s \quad X_{\tau_{n-1}+s} = x \mid X_{\tau_{n-1}}\} = \mathbb{P}\{X_{\tau_n} = x \mid X_{\tau_{n-1}}\}.
\end{aligned}$$

□

Lemma 3.2.2. *Assume that* Y *is recurrent and satisfies Condition 3.2.1. If* Y *is a countable RHMM, then the 1-blocks sequence* Z *is a countable RHMM.*

Proof. Let $Y_n = f(X_n)$, where $X = (X_1, X_2, X_3, \dots)$ is a recurrent homogeneous Markov chain, with a countable state space χ, and $f : \chi \longrightarrow J$ is a deterministic function. For all $y \in J$

$$\mathcal{S}_y := f^{-1}(y) = \{x \in \chi \mid f(x) = y\}.$$

Clearly $\chi = \cup_{y\in J}\mathcal{S}_y$. Define $\hat{\chi} := \chi \setminus \mathcal{S}_1 = \cup_{y\in J, y\neq 1} \mathcal{S}_y$. We call $\hat{\chi}^*$ the set of finite strings from $\hat{\chi}$, *i.e.* $\hat{\chi}^*$ is the set of finite strings of elements of χ that do not contain elements of $\mathcal{S}_1$. Define

$$\bar{\chi} := \bigcup_{x\in\mathcal{S}_1} \left(\{x\} \times \hat{\chi}^*\right).$$

$\bar{\chi}$ is the set of finite strings of elements of χ beginning with an element in $\mathcal{S}_1$ and that do not contain further elements of $\mathcal{S}_1$. The elements of $\bar{\chi}$ will be called S_1-blocks. Y satisfies Condition 3.2.1, thus

$$\mathbb{P}\{X_1 \in S_1\} = 1. \tag{3.8}$$

By the recurrence hypothesis on X we get

$$\mathbb{P}\{X_n \in \mathcal{S}_1 \text{ for infinitely many } n\} = 1. \tag{3.9}$$

Let $W_1, W_2, \dots$ be the successive S_1-blocks of X. The sequence $W = (W_1, W_2, \dots)$ is well defined by equations (3.8) and (3.9). Let $P_{x_1,x_2} := \mathbb{P}\{X_{t+1} = x_2 \mid X_t = x_1\}$, $x_1, x_2 \in \chi$. Up to events of probability zero, W takes values in the set of finite strings $w_i \in \bar{\chi}$, with $w_i = x_{1i}x_{2i}\dots x_{l_i i}$ such that $P_{x_{ti},x_{t+1i}} > 0$ for $t = 1, \dots, l_i - 1$.
Define $\bar{f} : \bar{\chi} \longrightarrow J^*$ as $\bar{f}(x_1, x_2, \dots, x_n) := f(x_1)f(x_2)\dots f(x_n) = 1f(x_2)\dots f(x_n)$, where the last equality follows by the definition of $\bar{\chi}$. If Z_i is the i-th 1-block of Y, then $Z_i = \bar{f}(W_i)$, where W_i is the i-th S_1-block of X.
To derive the law of W, recall the definition of the stopping times τ_n in equation (3.1)

$$\tau_1 = 1, \qquad\qquad \tau_n := \inf_t\{t > \tau_{n-1} \mid Y_t = 1\}.$$

Note that

$$\tau_n = \inf_t\{t > \tau_{n-1} \mid X_t \in S_1\}.$$

Thus $\tau_1, \tau_2, \dots$ are stopping times also with respect to the σ-field generated by X.
Let $w_1^n = w_1, w_2 \dots w_n$ be a sequence of S_1-blocks, say $w_i = x_{1i}x_{2i}\dots x_{l_i i}$, $i = 1, 2, \dots, n$. Define

$$\mathbb{P}\{W_m = w_1\} := \mathbb{P}\{X_{\tau_m}^{\tau_m+l_1-1} = w_1; \tau_{m+1} = \tau_m + l_1\} = \mathbb{P}\{X_{\tau_m}^{\tau_m+l_1-1} = w_1; X_{\tau_m+l_1} \in S_1\},$$

$$\begin{aligned}\mathbb{P}\{W_m^{m+n-1} = w_1^n\} :=& \\ &\mathbb{P}\{X_{\tau_m}^{\tau_m+l_1-1} = w_1,\ X_{\tau_{m+1}}^{\tau_{m+1}+l_2-1} = w_2, \dots,\ X_{\tau_{(m+n-1)}}^{\tau_{(m+n-1)}+l_n-1} = w_n; \\ &\qquad \tau_{m+1} = \tau_m + l_1, \dots, \tau_{m+n} = \tau_{m+n-1} + l_n\}.\end{aligned}$$

The Markovianity of W follows from that of X applying Lemma 3.2.1 and noting that τ_{n+1} is a function of $(X_{\tau_n}, X_{\tau_n+1}, \dots)$:

$$\begin{aligned}&\mathbb{P}\big\{W_n = w_n \mid W_1 = w_1, W_2 = w_2, \dots, W_{n-1} = w_{n-1}\big\} = \\ &= \mathbb{P}\{X_{\tau_n}^{\tau_n+l_n-1} = w_n; \tau_{n+1} = \tau_n + l_n \mid \\ &\ X_1^{l_1} = w_1,\ \dots,\ X_{\tau_{(n-1)}}^{\tau_{(n-1)}+l_{(n-1)}} = w_{n-1}; \tau_2 = \tau_1 + l_1, \dots, \tau_n = \tau_{n-1} + l_{n-1}\} \\ &= \text{(Lemma 3.2.1)} \\ &= \mathbb{P}\{X_{\tau_n}^{\tau_n+l_n-1} = w_n; \tau_{n+1} = \tau_n + l_n \mid X_{\tau_{n-1}}^{\tau_{n-1}+l_{n-1}-1} = w_{n-1}; \tau_n = \tau_{n-1} + l_{n-1}\} \\ &= \mathbb{P}\{W_n = w_n \mid W_{n-1} = w_{n-1}\}.\end{aligned}$$

We have to show that W is recurrent, i.e. $\mathbb{P}_w\{W_n = w \ i.o.\} = 1$ for any initial state $w = x_1x_2 \dots x_l$. It is clearly true if $\mathbb{P}_{x_1x_2\dots x_l}\{X_n^{n+l-1} = x_1x_2 \dots x_l \ i.o.\} = 1$. X is recurrent, so $\mathbb{P}_{x_1}\{X_n = x_1 \ i.o.\} = 1$. Define recursively

$$\widetilde{\tau}_1 := 1, \qquad \widetilde{\tau}_n := \inf\{t > \widetilde{\tau}_{n-1} \mid X_t = x_1\}, \tag{3.10}$$

where the inf of a empty set is equal to $+\infty$. $\widetilde{\tau}_n$ is the n-th time that X hits the state x_1. By recurrence, X hits the state x_1 infinitely many times, thus the random times $\widetilde{\tau}_1, \dots, \widetilde{\tau}_n, \dots$ are all finite *a.s.*.
By the strong Markov property the events $\{X_{\widetilde{\tau}_1}^{\widetilde{\tau}_1+l-1} = w\}, \dots, \{X_{\widetilde{\tau}_n}^{\widetilde{\tau}_n+l-1} = w\}, \dots$ are independent. For any n,

$$\mathbb{P}_{x_1x_2\dots x_l}\{X_{\widetilde{\tau}_n}^{\widetilde{\tau}_n+l-1} = w\} = P_{x_1,x_2} \dots P_{x_{l-1},x_l} > 0.$$

Thus by the Borel-Cantelli Lemma,

$$\mathbb{P}_{x_1x_2\dots x_l}\{X_n^{n+l-1} = x_1x_2 \dots x_l \ i.o.\} = \mathbb{P}_{x_1x_2\dots x_l}\Big\{\limsup_{n\to+\infty}\{X_{\widetilde{\tau}_n}^{\widetilde{\tau}_n+l-1} = w\}\Big\} = 1, \tag{3.11}$$

so W has the required recurrence property. W takes values in $\bar{\chi}$, that is clearly countable. We have noted before that $Z_t = \bar{f}(W_t)$, thus Z is a countable HMM with recurrent underlying Markov chain. □

Lemma 3.2.3. *Let Y be Markov exchangeable, recurrent and satisfying Condition 3.2.1. If Y is a countable RHMM, then Z is a countable mixture of i.i.d. sequences.*

Proof. Z is exchangeable by Proposition 3.2.1 and by the previous Lemma also a countable RHMM. We want to apply Theorem 3.1.1 to Z. Notice that the stationarity of the underlying Markov chain is in the hypothesis of Theorem 3.1.1. But going back to the proof of Dharmadhikari (see [21]), it appears that just the recurrence of the underlying Markov chain is necessary, to ensure the convergence of the Cesaro limit of the transition matrix and its powers. So we can apply Theorem 3.1.1 to Z and conclude that Z is a countable mixture of i.i.d. sequences. □

This concludes the proof of the necessity.

Remark 3.2.4. Note that all the results are still valid if the word *countable* is consistently substituted with the word *finite* in all the hypotheses and conclusions.

Proof of the sufficiency

To conclude the proof of Theorem 3.2.1 we have to show

Proposition 3.2.3. *Let Y be a countable mixture of recurrent Markov chains. Then Y is a countable RHMM. Moreover if Y is also stationary, then Y is a HMM with underlying Markov chain which starts at its stationary distribution.*

Proof. Y is a countable mixture of Markov chains so we can write

$$\mathbb{P}\{Y_1^n = y_1^n\} = \sum_{k \in K} \mu(y_1, P_k) \prod_{t=1}^{n-1} P^k_{y_t, y_{t+1}}, \tag{3.12}$$

where K is a countable set. Let $\mathbf{P}$ be the direct sum of the matrices P^k:

$$\mathbf{P} := \begin{pmatrix} P^1 & \mathbb{O} & \mathbb{O} & \dots \\ \mathbb{O} & P^2 & \mathbb{O} & \dots \\ \mathbb{O} & \mathbb{O} & P^3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and let

$$\pi := (\mu(1, P^1), \mu(2, P^1), \dots, \mu(1, P^2), \mu(2, P^2), \dots, \mu(1, P^k), \mu(2, P^k) \dots\dots),$$

with k varying in K. Let $X = (X_1, X_2, \dots)$ be the Markov chain on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with state-space $J \times K$ [1], transition matrix $\mathbf{P}$ and initial distribution π, with $\pi(y, k) = \mu(y, P^k)$. X is clearly recurrent.

Let $f : J \times K \longrightarrow J$ be the projection on the first component, i.e. $f(y, k) = y$. We will show that $Y_n \overset{d}{=} f(X_n)$, that is Y is a RHMM, showing

$$\mathbb{P}\big\{Y_1^n = y_1^n\big\} = \mathbb{P}\big\{f(X_1) = y_1, \dots, f(X_n) = y_n\big\}. \tag{3.13}$$

We have

$$\mathbb{P}\big\{X_1 = (y_1, k_1), X_2 = (y_2, k_2), \dots, X_n = (y_n, k_n)\big\} = \mu(y_1, k_1) \prod_{t=1}^{n-1} \mathbf{P}_{(y_t, k_t), (y_{t+1}, k_{t+1})}.$$

Note that $\mathbf{P}_{(y_t, k_t), (y_{t+1}, k_{t+1})} = 0$ if $k_t \neq k_{t+1}$, thus the product above is positive just for $k_1 = \dots = k_n$. By the previous formula we have

[1] order the states in this way: $(1,1), (2,1), \dots, (1,2), (2,2), \dots\dots$

$$
\begin{aligned}
&\mathbb{P}\big\{f(X_1)=y_1,\dots,f(X_n)=y_n\big\} = \\
&\qquad = \mathbb{P}\Big\{\bigcup_{k\in K} X_1=(y_1,k_1), \bigcup_{k\in K} X_2=(y_2,k_2),\dots,\bigcup_{k\in K} X_n=(y_n,k_n)\Big\} \\
&\qquad = \mathbb{P}\Big\{\bigcup_{k\in K}\Big(X_1=(y_1,k),X_2=(y_2,k),\dots,X_n=(y_n,k)\Big)\Big\} \\
&\qquad = \sum_{k\in K}\mathbb{P}\Big\{X_1=(y_1,k),X_2=(y_2,k),\dots,X_n=(y_n,k)\Big\} \\
&\qquad = \sum_{k\in K}\pi(y_1,k)\prod_{t=1}^{n-1}\mathbf{P}_{(y_t,k)(y_{t+1},k)} = \sum_{k\in K}\mu(y_1,k)\prod_{t=1}^{n-1}P^k_{(y_t,k)(y_{t+1},k)}
\end{aligned}
$$

Comparing equation (3.12) with the last expression, equation (3.13) is proved. □

3.3 Countable mixtures of k-Markov chains

The main result of this section is the generalization of Theorem 3.2.1 to k-Markov exchangeable sequences:

Theorem 3.3.1. *Let Y be a k-Markov exchangeable sequence satisfying Condition 1.2.1. Y is a countable RHMM if and only if Y is a countable mixture of recurrent Markov chains of order k.*

To prove this result, we need the characterization Theorem 1.2.4 for k-Markov exchangeable sequences, which we have stated in Section 1.2.4, and that we report below

Theorem 3.3.2. *Let Z be a sequence satisfying the recurrence Condition 1.2.1 in Chapter 1.2.4. Z is k-Markov exchangeable if and only if it is a mixture of Markov chains of order k.*

We provide a proof of this result in the next section.

3.3.1 de Finetti theorem for k-Markov exchangeable sequences

The aim of this section is to prove Theorem 3.3.2. We first need to introduce the 1-blocks of order k.

1-blocks of order k

Definition 3.3.1. *A* 1-block of order k *(or 1^k-block) is a string of length $l \geq k$, whose initial substring is* $\underbrace{11\dots1}_{k\,times}$ *and which does not contain any further occurrences of* $\underbrace{11\dots1}_{k\,times}$ *from position* $k+1$ *on.*

For example 11123114 is a 1^3-block, but 11123114 are two 1^2-blocks, $\underline{11}123$ and $\underline{11}4$. The string 111123 are two 1^2-blocks, respectively $\underline{11}$ and $\underline{11}23$, but it is also one 1^3-block, $\underline{111}123$.

In what follows we assume

Condition 3.3.1. $\mathbb{P}\{Y_1^k = 1^k\} = 1$.

Let Y be a k-Markov exchangeable sequence with values in J satisfying Condition 1.2.1 in Section 1.2.4 and Condition 3.3.1. We can define the sequence $Z = (Z_1, Z_2, \dots\dots)$ of the successive 1^k-blocks of Y. The sequence Y satisfies Condition 1.2.1 and Condition 3.3.1, thus its successive 1-blocks are *a.s.* of finite length and the sequence $Z = (Z_1, Z_2, \dots)$ is well defined.
As for 1-blocks, there is a $1-1$ correspondence between strings z_1^n of the first n 1^k-blocks and strings y_1^m of single letters from J. A 1^k-block is a finite string of letters from J, thus a finite string z_1^n of 1^k-blocks is still a finite string of letters from J, write

$$z_1^n = \underbrace{11\dots1}_{k\,times} y_{(k+1)1} y_{(k+2)1} \cdots y_{l_1 1} \underbrace{11\dots1}_{k\,times} y_{(k+1)2} y_{(k+2)2} \cdots y_{l_2 2} \cdots\cdots y_{l_n n} = y_1^m.$$

The event $\{Z_1^n = z_1^n\}$ corresponds to the event $\{Y_1^m = y_1^m,\ Y_{m+1}^{m+k} = 1^k\}$, and thus they have the same measure under all measures.
To give the whole law of Z, as we have done for 1-blocks, we define the random times $(\bar{\tau}_n)_{n\geq 1}$:

$$\bar{\tau}_1 := k \qquad \bar{\tau}_n := \inf_t \{t > \bar{\tau}_{n-1} + k - 1 \mid Y_{t-k+1} = Y_{t-k+2} = \cdots = Y_t = 1\}. \tag{3.14}$$

$\bar{\tau}_n$ are stopping times with respect to the σ-field generated by Y [2]. Let $z_1^n = z_1\, z_2 \dots z_n$ be a string of 1^k-blocks, $z_i \in J^*$, say $z_i = \underbrace{11\dots1}_{k\,times} y_{(k+1)i} y_{(k+2)i} \cdots y_{l_i i}$,

[2] It would be more natural to define

$$\bar{\tau}_1 := 1 \qquad \bar{\tau}_n := \inf_t \{t > \bar{\tau}_{n-1} + k - 1 \mid Y_t = Y_{t+1} = \cdots = Y_{t+k-1} = 1\},$$

but these are not stopping times with respect to the σ-field generated by Y.

where $y_{ji} \in J$ and l_i is the length of z_i. Define

$$\mathbb{P}\{Z_m^{m+n-1} = z_1^n\} := \mathbb{P}\{Y_{\bar{\tau}_m-k+1}^{\bar{\tau}_m} = 1^k, Y_{\bar{\tau}_m+1} = y_{(k+1)1}, \dots, Y_{\bar{\tau}_m+l_1-k} = y_{l_1 1}, \dots$$
$$Y_{\bar{\tau}_{m+n-1}-k+1} = 1, \dots, Y_{\bar{\tau}_{m+n-1}+l_n-k} = y_{l_n n}; \bar{\tau}_{m+1} = \bar{\tau}_m + l_1, \dots \bar{\tau}_{m+n} = \bar{\tau}_{m+n-1} + l_n\}.$$

It can be proved that

Proposition 3.3.1. *The sequence Z of the 1^k-blocks of Y is exchangeable.*

Proof. The sequence Y is k-Markov exchangeable by assumption and permutations of 1^k-blocks do not change transition counts of order k nor the first k letters. The exchangeability of Z follows immediately. □

Remark 3.3.1. *We have forced the first k states of Y to be 1 in Condition 3.3.1. But we could have forced them to be any string $y_1 \dots y_k$ of letters from J, defining then the $y_1 \dots y_k$-blocks and deriving the same results.*

Proof of Theorem 1.2.4

We prove the necessity part of Theorem 1.2.4, since the sufficiency part is trivial. We need to extend the Proposition (15) in [22]. In particular we prove

Proposition 3.3.2. *Let Y be a k-Markov exchangeable sequence, satisfying Condition 1.2.1 and Condition 3.3.1. If the 1^k-blocks of Y are independent and identically distributed, then Y is an homogenous Markov chain of order k.*

Proof. Let $\mathbb{Q}$ be a probability measure on $\mathcal{F}$, such that the sequence Z of the 1^k-blocks of Y are independent and identically distributed with respect to $\mathbb{Q}$.
Let σ and σ' be finite strings of states which start with k 1-s and end with the same string $i_1 i_2 \dots i_k$ of length k, with $i_1, i_2, \dots, i_k \in J$. We do not assume that σ and σ' are k-transition equivalent. Let l_σ and $l_{\sigma'}$ be the length of the strings σ and σ' respectively.
The Markov property, which must be proved is

$$\mathbb{Q}\Big\{Y_{l_\sigma+1} = j \mid Y_1^{l_\sigma} = \sigma\Big\} = \mathbb{Q}\Big\{Y_{l_{\sigma'}+1} = j \mid Y_1^{l_{\sigma'}} = \sigma'\Big\}. \tag{3.15}$$

In fact the last k elements of the strings σ and σ' are the same, so, if equation (3.15) holds, the probability that the sequence takes the value j depends only on the previous k values.
To avoid division by 0 instead of (3.15) we prove

$$\mathbb{Q}\Big\{Y_1^{l_{\sigma'}} = \sigma', Y_{l_{\sigma'}+1} = j\Big\}\mathbb{Q}\Big\{Y_1^{l_\sigma} = \sigma\Big\} = \mathbb{Q}\Big\{Y_1^{l_\sigma} = \sigma, Y_{l_\sigma+1} = j\Big\}\mathbb{Q}\Big\{Y_1^{l_{\sigma'}} = \sigma'\Big\} \quad (3.16)$$

For any strings of states α and β

$$\mathbb{Q}\Big\{Y_1^{l_\alpha+l_\beta+2k} = \underbrace{11\ldots1}_{k\,times}\alpha\underbrace{11\ldots1}_{k\,times}\beta\Big\} = \quad (3.17)$$
$$= \mathbb{Q}\Big\{Y_1^{l_\alpha+2k} = \underbrace{11\ldots1}_{k\,times}\alpha\underbrace{11\ldots1}_{k\,times}\Big\}\mathbb{Q}\Big\{Y_1^{l_\beta+k} = \underbrace{11\ldots1}_{k\,times}\beta\Big\}$$

because the $\underbrace{11\ldots1}_{k\,times}$-blocks are independent and identically distributed.

Let ψ run through all the finite strings of states which do not pass through the string of k consecutive 1-s. Condition 1.2.1 implies

$$\mathbb{Q}\Big\{Y_1^{l_\sigma} = \sigma\Big\} = \sum_\psi \mathbb{Q}\Big\{Y_1^{l_\sigma+l_\psi+k} = \sigma\psi\underbrace{11\ldots1}_{k\,times}\Big\}. \quad (3.18)$$

The first and the last k states of σ and σ' are the same, thus

$$\sigma\psi\sigma' j \sim_k \sigma'\psi\sigma j.$$

By the k-Markov exchangeability of Y

$$\mathbb{Q}\Big\{Y_1^{l_\sigma+l_\psi+l_{\sigma'}+1} = \sigma\psi\sigma' j\Big\} = \mathbb{Q}\Big\{Y_1^{l_{\sigma'}+l_\psi+l_\sigma+1} = \sigma'\psi\sigma j\Big\}.$$

Recalling that σ and σ' start with k 1-s, by (3.17)

$$\mathbb{Q}\Big\{Y_1^{l_\sigma+l_\psi+k} = \sigma\psi\underbrace{11\ldots1}_{k\,times}\Big\}\mathbb{Q}\Big\{Y_1^{l_{\sigma'}} = \sigma',\, Y_{l_{\sigma'}+1} = j\Big\} = \mathbb{Q}\Big\{Y_1^{l_\sigma+l_\psi+l_{\sigma'}+1} = \sigma\psi\sigma' j\Big\} =$$
$$\mathbb{Q}\Big\{Y_1^{l_{\sigma'}+l_\psi+l_\sigma+1} = \sigma'\psi\sigma j\Big\} = \mathbb{Q}\Big\{Y_1^{l_{\sigma'}+l_\psi+k} = \sigma'\psi\underbrace{11\ldots1}_{k\,times}\Big\}\mathbb{Q}\Big\{Y_1^{l_\sigma} = \sigma,\, Y_{l_\sigma+1} = j\Big\}.$$

Sum out ψ and use (3.18) to get (3.16). □

Proof. of Theorem 1.2.4. To prove the necessity part of Theorem 1.2.4 we will use the same line of thoughts of Diaconis and Freedman in [22]. Without loss of generality let us consider Y as the coordinate sequence. By Proposition 3.3.1 the sequence Z of the 1^k-blocks of Y is exchangeable. So we can apply de Finetti's theorem (see Theorem 1.2.1) to Z writing for all n and all $z_1^n \in J^*$

$$\mathbb{P}\{z_1^n\} = \int P_\omega(z_1^n) d\mathbb{P}(\omega) \tag{3.19}$$

for a suitable regular conditional probability P. Z is an i.i.d. sequence with respect to P_ω for $\mathbb{P}$-all ω. By definition of regular conditional probability we can write

$$\mathbb{P}\{y_1^m\} = \int P_\omega(y_1^m) d\mathbb{P}(\omega), \tag{3.20}$$

where P is the regular conditional probability which appears in equation (3.19). Z is an i.i.d. sequence with respect to P_ω for $\mathbb{P}$-all ω, thus, by Proposition 3.3.2, the sequence Y is a Markov chain of order k with respect to P_ω for $\mathbb{P}$-all ω. So we have shown that a k-Markov exchangeable sequence Y satisfying Condition 1.2.1 and Condition 3.3.1 is a mixture of Markov chains of order k, thus the necessity part of Theorem 1.2.4 is proved. □

3.3.2 The main result

Proof of the necessity

To prove the necessity part of the Theorem 3.3.1 we first generalize Lemma 3.2.2 to 1^k-blocks:

Lemma 3.3.1. *Let Y satisfy Condition 1.2.1 and Condition 3.3.1. If Y is a countable HMM with recurrent underlying Markov chain, then the 1^k-blocks sequence Z is a countable HMM with recurrent underlying Markov chain.*

Proof. Let X, χ, f and $\mathcal{S}_y$ be as in the proof of Lemma 3.2.2. Define

$$\begin{aligned} \mathcal{S}_n^k := \{ &\mathcal{S}_{y_1} \times \mathcal{S}_{y_2} \times \cdots \times \mathcal{S}_{y_n}, \text{for } y_1, y_2, \dots y_n \in J \\ &\text{and } y_1 y_2 \dots y_n \text{ not containing } \underbrace{11\dots1}_{k\,times} \text{ as a substring } \} \end{aligned}$$

Define

$$\hat{\chi}_k := \{x_1 x_2 \dots x_n \in \mathcal{S}_n^k, for\, n \in \mathbb{N}\},$$

$\hat{\chi}_k$ is the set of finite strings from χ that do not contain k or more successive elements from S_1. Define

$$\bar{\chi} := \bigcup_{x_1,x_2,\dots x_k \in \mathcal{S}_1} \big(\{x_1\} \times \{x_2\} \times \dots \{x_k\} \times \hat{\chi}_k\big).$$

This is the set of finite strings of elements of χ beginning with k elements in $\mathcal{S}_1$ and that do not contain further strings of k or more elements of $\mathcal{S}_1$. The elements of $\bar{\chi}$ will be called $(\mathcal{S}_1)^k$-blocks.
Y satisfies Condition 1.2.1 and Condition 3.3.1, thus we get

$$\begin{aligned} 1 = \mathbb{P}\{Y_n^{n+k-1} = \underbrace{11\dots1}_{k\,times} \text{ for infinitely many } n\} = \\ \mathbb{P}\{X_n^{n+k-1} \in \underbrace{\mathcal{S}_1 \times \mathcal{S}_1 \times \dots \mathcal{S}_1}_{k\,times} \text{ for infinitely many } n\}. \end{aligned} \tag{3.21}$$

Let $W_1, W_2, \dots$ be the successive $(\mathcal{S}_1)^k$-blocks of X. The sequence $W = (W_1, W_2, \dots)$ is well defined by property (3.21).
Define $\bar{f} : \bar{\chi} \longrightarrow J^*$ as

$$\bar{f}(x_1, x_2, \dots, x_n) := f(x_1)f(x_2)\dots f(x_n) = \underbrace{11\dots1}_{k\,times} f(x_{k+1})\dots f(x_n),$$

where the last equality follows by the definition of W. If Z_i is the i-th 1^k-block of Y, then $Z_i = \bar{f}(W_i)$, where W_i is the i-th $(\mathcal{S}_1)^k$-block of X.
Let $w_1^n = w_1, w_2 \dots w_n$ be a sequence of elements in $\bar{\chi}$,

$$w_i = x_{1i}x_{2i}\dots x_{ki}x_{k+1i}\dots x_{l_i i}$$

with $x_{1i}, x_{2i}, \dots, x_{ki} \in \mathcal{S}_1$ and $x_{k+1i}\dots x_{l_i i} \in \hat{\chi}_k$. Recalling the definition of $(\bar{\tau}_n)_n$ in equation (3.14), we can define:

$$\mathbb{P}\{W_m = w_1\} := \mathbb{P}\{X_{\bar{\tau}_m-k+1}^{\bar{\tau}_m+l_1-k} = w_1;\, \bar{\tau}_{m+1} = \bar{\tau}_m + l_1\} \qquad \text{and}$$

$$\begin{aligned} \mathbb{P}\{W_m^{m+n-1} = w_1^n\} := \mathbb{P}\{X_{\bar{\tau}_m-k+1}^{\bar{\tau}_m+l_1-k} = w_1,\, X_{\bar{\tau}_{m+1}-k+1}^{\bar{\tau}_{m+1}+l_2-k} = w_2, \dots,\, X_{\bar{\tau}_{m+n-1}-k+1}^{\bar{\tau}_{m+n-1}+l_n-k} = w_n; \\ \bar{\tau}_{m+1} = \bar{\tau}_m + l_1, \dots, \bar{\tau}_{m+n} = \bar{\tau}_{m+n-1} + l_n\}. \end{aligned}$$

The Markovianity of W follows from the strong Markov property for X and Lemma 3.2.1:

$$
\begin{aligned}
&\mathbb{P}\big\{W_n = w_n \mid W_1 = w_1, W_2 = w_2, \dots, W_{n-1} = w_{n-1}\big\} = \\
&= \mathbb{P}\{X_{\bar\tau_{m+n-1}-k+1}^{\bar\tau_{m+n-1}+l_n-k} = w_n; \bar\tau_{m+n} = \bar\tau_{m+n-1} + l_n \mid X_{\bar\tau_m-k+1}^{\bar\tau_m+l_1-k} = w_1, \dots, \\
&\quad X_{\bar\tau_{m+n-2}-k+1}^{\bar\tau_{m+n-2}+l_{n-1}-k} = w_{n-1};\ \bar\tau_{m+1} = \bar\tau_m + l_1, \dots, \bar\tau_{m+n-1} = \bar\tau_{m+n-2} + l_{n-1}\} \\
&= (\text{Lemma 3.2.1}) = \mathbb{P}\{X_{\bar\tau_{m+n-1}-k+1}^{\bar\tau_{m+n-1}+l_n-k} = w_n; \bar\tau_{m+n} = \bar\tau_{m+n-1} + l_n \mid \\
&\qquad\qquad X_{\bar\tau_{m+n-2}-k+1}^{\bar\tau_{m+n-2}+l_{n-1}-k} = w_{n-1}; \bar\tau_{m+n-1} = \bar\tau_{m+n-2} + l_{n-1}\} \\
&\quad = \mathbb{P}\{W_n = w_n \mid W_{n-1} = w_{n-1}\}.
\end{aligned}
$$

The recurrence of W follows by the recurrence of X, as in the proof of Lemma 3.2.2. W takes values in $\bar\chi$, that is clearly countable. We have noted before that $Z_t = \bar f(W_t)$, so Z is a countable HMM with recurrent underlying Markov chain W. □

We can conclude the proof of the necessity part of Theorem 3.3.1 as in Section 3.2, using Proposition 3.3.2 instead of Proposition 3.2.2.

Proof of the sufficiency

Enlarging the state space, it can be shown that a Markov chain of order k is a Markov chain of order 1. So the sufficiency easily follows by the same arguments of Proposition 3.2.3.

3.4 Countable mixtures of Markov chains à la Fortini et al.

The aim of this Section is to prove the analog of Theorem 3.2.1 in the setting of [27]. Recalling the notations of Theorem 1.2.3 we have the following:

Theorem 3.4.1. *Let the matrix V of the successors of Y be partially exchangeable. Then Y is a countable HMM with underlying strongly recurrent Markov chain if and only if the random element $\widetilde{P}$ of $\mathcal{P}^*$ takes just a countable number of values.*

Proof of the necessity

To prove the necessity part of the theorem we need the following Lemma, which is the analogous of Lemma 3.2.2 for the matrix of the successors.

Lemma 3.4.1. *Let V the matrix of the successor of Y and let Y be a countable HMM with underlying strongly recurrent Markov chain. Then for all $i \in I$, the i-th row of V is a countable HHM with underlying recurrent Markov chain..*

Proof. of Lemma 3.4.1 Without loss of generality, let us fix i. Let X, χ, f and $\mathcal{S}_y$ be as in the proof of Lemma 3.2.2. Define inductively

$$\gamma_1^i := \inf\{t \mid Y_t = i\} = \inf\{t \mid X_t \in S_i\},$$

$$\gamma_n^i := \inf\{t > \gamma_{n-1}^i \mid Y_t = i\} = \inf\{t > \gamma_{n-1}^i \mid X_t \in S_i\}.$$

Y takes the value i for the n-th times at γ_n^i, or equivalently X enters in S_i the n-th times at γ_n^i.

Define the process $(W_{S_i,n})_{n\geq 1}$, $W_{S_i,n} : \Omega \longrightarrow \chi$, of the successors of S_i: $W_{S_i,n} = x$ if $X_{\gamma_n^i+1} = x$, $x \in \chi$. We claim that $(W_{S_i,n})_{n\geq 1}$ is a Markov chain, i.e.

$$\mathbb{P}\{W_{S_i,n} = x \mid W_{S_i,n-1} W_{S_i,n-2} \dots W_{S_i,1}\} = \mathbb{P}\{W_{S_i,n} = x \mid W_{S_i,n-1}\}$$

for all $n \geq 1$ and all $x \in \chi$. In fact, using Lemma 3.2.1

$$\begin{aligned}\mathbb{P}\{W_{S_i,n} = x \mid W_{S_i,n-1} W_{S_i,n-2} \dots W_{S_i,1}\} &= \mathbb{P}\{X_{\gamma_n^i+1} = x \mid X_{\gamma_{n-1}^i+1} X_{\gamma_{n-2}^i+1} \dots X_{\gamma_1^i+1}\} \\ &= \mathbb{P}\{X_{\gamma_n^i+1} = x \mid X_{\gamma_{n-1}^i+1}\} = \mathbb{P}\{W_{S_i,n} = x \mid W_{S_i,n-1}\}.\end{aligned}$$

To prove the recurrence of W, we have to show that for any fixed i

$$\mathbb{P}_x\big\{W_{S_i,n} = x \text{ for infinitely many } n\big\} = \mathbb{P}_x\big\{\limsup_n \{W_{\mathcal{S}_i,n} = x\}\big\} = 1.$$

Let s be the first state in S_i hit by X, i.e. $X_{\gamma_1^i} = s$. By the strongly recurrence of X,

$$\mathbb{P}\{X_n = s \text{ for infinitely many } n\} = 1.$$

Define

$$\gamma_1^s = \inf\{t \mid X_t = s\} = \gamma_1^i \quad \text{and} \quad \gamma_n^s := \inf\{t > \gamma_{n-1}^s \mid X_t = s\}.$$

$(\gamma_n^s)_n$ is a subsequence of $(\gamma_n^i)_n$ and it is infinite by the strongly recurrence of X. Let be $\omega \in \{X_{\gamma_n^s+1} = x\}$. Then there exists an $m \geq n$ such that $\omega \in \{X_{\gamma_m^i+1} = x\}$, that is

$$\{X_{\gamma_n^s+1} = x\} \subseteq \bigcup_{m\geq n} \{X_{\gamma_m^i+1} = x\}. \tag{3.22}$$

Thus

$$\limsup_n \{X_{\gamma_n^s+1} = x\} \subseteq \limsup_n \{X_{\gamma_n^i+1} = x\} = \limsup_n \{W_{\mathcal{S}_i,n} = x\}. \quad (3.23)$$

To get the recurrence of W, we have to show that $\mathbb{P}_x\big\{\limsup_n\{W_{\mathcal{S}_i,n} = x\}\big\} = 1$. By equation (3.23), it is sufficient to prove that $\mathbb{P}_x\big\{\limsup_n\{X_{\gamma_n^s+1} = x\}\big\} = 1$. The events $\{X_{\gamma_1^s+1} = x\}, \{X_{\gamma_2^s+1} = x\}, \dots, \{X_{\gamma_n^s+1} = x\}, \dots$ are independent by the Markovianity of X and they have positive probability. Thus by Borel-Cantelli Lemma we get

$$\mathbb{P}_x\big\{\limsup_n\{X_{\gamma_n^s+1} = x\}\big\} = 1. \quad (3.24)$$

Combining equations (3.23) and (3.24) we get the recurrence of W.

Notice that $V_{i,n} = f(W_{i,n})$. $(V_{i,n})_{n\geq 1}$ is indeed a countable RHMM.

□

Proof. of Theorem 3.4.1 Let i be fixed. $(V_{i,n})_{n\geq 1}$ is an exchangeable sequence and by Lemma 3.4.1 it is an HMM. So Theorem 3.1.1 applies to $(V_{i,n})_{n\geq 1}$, showing that it is a countable mixture of i.i.d. sequences. For any fixed i, let θ_i be the random measure defined in the proof of Theorem 1 in [27] as $\lim_{n\to\infty} \frac{1}{n}\sum_{r=1}^n \delta_{V_{i,r}}$. $(V_{i,n})_{n\geq 1}$ is a countable mixture of i.i.d. sequences, thus for any i, θ_i is a random measure that takes just a countable number of values. In the proof of Theorem 1 in [27], θ_i is the i-th row of the random matrix $\widetilde{P}$. Recalling that a countable union of countable sets is countable, we get that the random matrix $\widetilde{P}$ takes just a countable number of values. This concludes the proof. □

Proof of the sufficiency

Proof. Let $\widetilde{P}$ take just a countable number of values, call them $\widetilde{P}^1, \widetilde{P}^2, \dots$ and let the initial condition of Y be fixed at y_1. As in the proof of Proposition 1.2.2, we can write

$$\mathbb{P}\{Y_1^n = y_1^n\} = \int_{\mathcal{P}} \prod_{t=1}^{n-1} \widetilde{P}_{y_t,y_{t+1}} \mathcal{L}_{\widetilde{P}}(d\widetilde{P}).$$

$\widetilde{P}$ take just a countable number of values, so the last equation becomes

$$\mathbb{P}\{Y_1^n = y_1^n\} = \sum_{k\in K} \mu_k \Big(\prod_{t=1}^{n-1} \widetilde{P}^k_{y_t,y_{t+1}}\Big),$$

where K is a countable set. By the last equation we get

$$\mathbb{P}\{Y_n = y_n\} = \sum_{k \in K} \mu_k \widetilde{(P^k_{y_1,y_n})}^{n-1}.$$

Denote with Π the direct sum of matrices $\widetilde{P}^1, \widetilde{P}^2, \ldots$. Consider the Markov chain X with state space $\{(k,j),\ \text{with } k \in K \text{ and } i \in J\}$ and transition matrix Π. Recalling that the initial condition of Y is fixed, conclude as in the proof of Proposition 3.2.3. $\square$

Input-output properties of the CUSUM statistics

Introduction

Detection of changes of statistical patterns is a fundamental problem in many applications. A basic method for detecting temporal changes in an independent sequence is the Cumulative Sum (CUSUM) algorithm or Hinkley-detector, which will be presented in Section 4.1. It was first used for independent observations, but its range of applicability can be extended. We propose an adaptation of the CUSUM algorithm to HHMs. The key problem here is to generate an appropriate residual process. We will see that the proposed one, under some reasonable technical conditions, is L-mixing. (For the definition of L-mixing see 4.2).

The CUSUM test is defined via a sequence of random variables (X_n), often called residuals in the engineering literature, such as likelihood ratios, such that

$$\mathbb{E}(X_i) < 0 \quad \text{for } i \le \tau^* - 1, \quad \text{and} \quad \mathbb{E}(X_i) > 0 \quad \text{for } i \ge \tau^*,$$

with τ^* denoting the change point.

Letting $S_0 := 0$ and $S_n := \sum_{i=1}^n X_i$, the CUSUM statistics or Hinkley detector is defined for $n \ge 0$ as

$$g_n := S_n - \min_{0 \le k \le n} S_k = \max_{0 \le k \le n} (S_n - S_k). \tag{3.25}$$

An alarm is given if g_n exceeds a pre-fixed threshold $\delta > 0$. The moment of alarm is defined by

$$\hat{\tau} = \inf\{n \,|\, S_n - \min_{0 \le k \le n} S_k \ge \delta\}. \tag{3.26}$$

The CUSUM statistics (g_n) can be equivalently defined via a non-linear dynamical system as follows, with $a_+ = \max\{0, a\}$:

$$g_n = (g_{n-1} + X_n)_+ \quad \text{with} \quad g_0 = 0. \tag{3.27}$$

From a system-theoretic point of view this system is not stable in any sense. E.g., for a constant, positive input (g_n) becomes unbounded. On the other hand, for an i.i.d. input sequence (X_n), with negative expectation some stability of the output process (g_n) can be expected. The resulting stochastic system is a standard object in queuing theory. In this case the process (g_n) is clearly a homogenous Markov chain, also called a one-sided random walk. A number of stability properties of (g_n) have been established in [42] (see Section 5.2).

The purpose of the thesis is to study some input-output properties of the CUSUM statistics, extending the results for i.i.d. inputs mentioned above, in two ways: first we show that for random i.i.d. inputs with negative expectation, and finite exponential moments of some positive order the output (g_n) of this system is L-mixing. Then, this result is extended, under further technical conditions such as boundedness, to L-mixing inputs, which is the case of interest for the CUSUM algorithm proposed for HHMs.

The assumption that (X_n) is an i.i.d. sequence reflects the tacit assumption that actually there is no change at all, i.e. $\tau^* = +\infty$. The CUSUM algorithm can still be used to monitor the process, and we may occasionally get an alarm. The frequency of these false alarms is of great practical interest. Our results can be applied to give an upper bound for the almost sure *false alarm frequency* as a function of the threshold δ, defined as

$$\limsup_{N \longrightarrow +\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\{g_n \geq \delta\}} \tag{3.28}$$

with the tacit assumption that $\tau^* = +\infty$. In fact the lim sup above is tractable if we have a strong Law of Large Numbers (LLN) for $\mathbb{I}_{\{g_i \geq \delta\}}$ or for $f(g_i)$, with f smooth and $f(g_i) \geq \mathbb{I}_{\{g_i \geq \delta\}}$. This is ensured if (g_n) is L-mixing.

Chapter 4

Definitions and Preliminary Results

4.1 Change detection and the CUSUM algorithm

To state the *change detection problem*, let us consider a sequence of random variables $Y_1, Y_2, \dots Y_N$ such that $Y_1, Y_2 \dots Y_{\tau^*-1}$ are distributed according to $f_0(\cdot)$, and $Y_{\tau^*}, \dots, Y_N$ are distributed according to $f_1(\cdot)$, for an unknown $1 \leq \tau^* \leq N+1$. The change detection problem consists on finding the change point τ^*, observing $Y_1, Y_2, \dots Y_N$. (Set $\tau^* = 1$ if all the random variables are distributed according to $f_1(\cdot)$, and $\tau^* = N+1$ if all the random variables are distributed according to $f_0(\cdot)$. These are the no change cases.)

Change detection problems have been extensively studied for decades (for a survey see [7] and [12]). There are two main different approaches to change detection problems, the oldest one models the change point τ^* as a deterministic and unknown time point (see for example the works of Page, Hinkley and Lorden cited below), but τ^* can be modeled also as a random variable with known prior distribution (see the work of Sirjaev [49]).
One of the first and most used algorithm for change detection is the cumulative sum (CUSUM) algorithm, that was introduced by Page ([43], [44] and [45]) and analyzed later, among others, by Hinkley (see [37]) and Lorden (see [41]). We describe now the CUSUM algorithm in the simplest case, for independent observations and known $f_0(\cdot)$ and $f_1(\cdot)$, but it has been adapted to more general frameworks, as we will see below. Let $Y_1, \dots Y_N$ be independent

and let us suppose at first that all the observations $y_1, \dots y_N$ are available (so we are treating an *off-line* change detection problem). Let f_0 and f_1 be known. We indicate with $l_\tau(y_1^N)$ the likelihood of the observations $y_1, \dots, y_N$ under the hypothesis that the distribution changes at τ, $1 \le \tau \le N+1$. The likelihood function is thus

$$l_\tau(Y_1^N) = -\log f(Y_1, Y_2, \dots Y_N) = -\sum_{i=1}^{\tau-1} \log f_0(Y_i) - \sum_{i=\tau}^{N} \log f_1(Y_i). \quad (4.1)$$

According to the Maximum Likelihood principle, a reasonable estimator $\hat{\tau}$ for τ^* can be defined as

$$\hat{\tau}(Y_1^N) := \arg \min_{1 \le \tau \le N+1} l_\tau(Y_1^N).$$

We say that a change in the distribution has occurred at n if $\hat{\tau} = n$ for some $1 \le n \le N+1$.

A more involved problem is the *on-line* change detection, where the observations $y_1, \dots, y_N$ are received sequentially and the change is detected in real time. We now present a trick that allows to go from *off-line* to *on-line* change detection. By equation (4.1) we have

$$l_\tau(Y_1^N) - l_{\tau-1}(Y_1^N) = -\log f_0(Y_{\tau-1}) + \log f_1(Y_{\tau-1}) = \log \frac{f_1(Y_{\tau-1})}{f_0(Y_{\tau-1})}.$$

Taking a telescopic sum

$$l_\tau(Y_1^N) = l_1(Y_1^N) + \sum_{i=2}^{\tau} \left(l_i(Y_1^N) - l_{i-1}(Y_1^N) \right) = l_1(Y_1^N) + \sum_{i=1}^{\tau-1} \log \frac{f_1(Y_i)}{f_0(Y_i)}.$$

Define

$$X_i := \log \frac{f_1(Y_i)}{f_0(Y_i)} \qquad \text{and}$$

$$S_\tau(Y_1^N) := \sum_{i=1}^{\tau} X_i = l_{\tau+1}(Y_1^N) - l_1(Y_1^N). \quad (4.2)$$

Thus we have

$$l_\tau(Y_1^N) = S_{\tau-1}(Y_1^N) + l_1(Y_1^N). \quad (4.3)$$

Trivially

$$\begin{aligned}\hat{\tau}(Y_1^N) &:= \arg \min_{1 \le \tau \le N+1} l_\tau(Y_1^N) = \arg \min_{1 \le \tau \le N+1} \left(l_\tau(Y_1^N) - l_1(Y_1^N) \right) \\ &= \arg \min_{0 \le \tau \le N} S_\tau.\end{aligned}$$

Notice that

$$\begin{cases} \mathbb{E}_{f_0}(X_i) &= -\mathbb{D}(f_0||f_1) < 0 \quad \text{for } i \leq \tau^* - 1 \\ \mathbb{E}_{f_1}(X_i) &= \mathbb{D}(f_1||f_0) > 0 \qquad \text{for } i \geq \tau^* \end{cases} \tag{4.4}$$

where $\mathbb{E}_{f_i}$ denotes the expectation of random variables under the distribution function $f_i(\cdot)$ and $\mathbb{D}(p||q)$ denotes the Kullback-Leibler divergence between p and q. By equation (4.4) the terms X_i of the sum in (4.2) have negative expectation for $i \leq \tau^* - 1$ and positive expectation for $i \geq \tau^*$. Thus if the change occurs in τ^*, $S_n - \min_{0\leq k\leq n} S_k$ has zero expectation for $i \leq \tau^* - 1$ and positive expectation for $i \geq \tau^*$. Taking into account random effects, use the following algorithm to find $\hat{\tau}$

$$\hat{\tau} = \inf\{n \,|\, (S_n - \min_{0\leq k\leq n} S_k) \geq \delta\}.$$

for some $\delta > 0$. Notice that $(S_n - \min_{0\leq k\leq n} S_k)$ is a function of the observation until time n, thus the algorithm works *on-line*.

We have presented the CUSUM algorithm in the simplest case, as it was introduced at first, for independent observations and $f_0(\cdot)$ and $f_1(\cdot)$ known, but the CUSUM algorithm can be used also in more general frameworks. The key elements in a CUSUM algorithm are

- a sequence $X_1, \ldots, X_N$ function of the observations, the so-called residual process, such that

$$\mathbb{E}(X_i) < 0 \text{ for } i \leq \tau^* - 1 \quad \text{and} \quad \mathbb{E}(X_i) > 0 \text{ for } i \geq \tau^*,$$

- the sequence $S_0, \ldots, S_N$, with $S_n := \sum_{i=1}^{n} X_i$ and $S_0 := 0$,

- the CUSUM statistics

$$g_n := \max_{0\leq k\leq n}(S_n - S_k) = S_n - \min_{0\leq k\leq n} S_k \quad \text{for } n \geq 0,$$

- a stopping threshold $\delta > 0$,

- a stopping rule

$$\hat{\tau} = \inf\{n \,|\, (S_n - \min_{0\leq k\leq n} S_k) \geq \delta\}.$$

The choice of δ is tricky: for a small value the algorithm could be too sensitive to random effects, but for a big value of δ the algorithm could not detect the changes.

The CUSUM was introduced for independent observations, but its range of applicability has been extended for dependent sequences, such as stationary ergodic processes in [6]. The applicability of the CUSUM algorithm to ARMA systems, with unknown dynamics, has been demonstrated in [5]. Much later, it was adapted to Hidden Markov Models, with unknown dynamics, in [32].

In the no change case, there are two key quantities to measure the performance of a change detection algorithm: the false alarm probability and the false alarm frequency. The *false alarm probability* of a change detection algorithm is the probability that the algorithm detects a change when this change has not occurred, that is

$$\mathbb{P}_{f_0}\{g_n \geq \delta\}.$$

It is close to the idea of *first kind* error of classical statistic. The almost sure *false alarm frequency* is defined as

$$\limsup_{N \longrightarrow +\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{g_i \geq \delta\}}, \tag{4.5}$$

where $\mathbb{I}_A$ denotes the characteristic function of the set A.
The lim sup above is tractable (as we will see in Section 5.4) if we have a Law of Large Numbers (LLN) for $\mathbb{I}_{\{g_i \geq \delta\}}$ or for $f(g_i)$, with f smooth and $f(g_i) \geq \mathbb{I}_{\{g_i \geq \delta\}}$. This is ensured if (g_n) is L-mixing. In the next chapter we will show that this is indeed the case for i.i.d. input sequence and for L-mixing input sequence, under suitable technical conditions.

4.1.1 Change detection for HHMs

Detection of changes in the statistical pattern of a hidden Markov process is of interest in a number of applications. Change detection for HHMs is one of the main motivation of the investigations in the next chapter. As we have seen in the previous section, a basic method for detecting temporal changes in an independent sequence is the Cumulative Sum (CUSUM) algorithm or Hinkley-detector. The adaptation of the CUSUM algorithm to the case when the dynamics of the HMM before and after the change is unknown was considered in [32], using a number of heuristic arguments. We consider the mathematically cleaner case when the dynamics before and after the change is known. More precisely let θ^* be the true parameter driving the dynamics of a HHM (χ_n, Y_n) and let

$$\theta^* = \begin{cases} \theta_1 & \text{for } n \leq \tau^* - 1 \\ \theta_2 & \text{for } n \geq \tau^*, \end{cases} \tag{4.6}$$

for an unknown τ^*, but for given $\theta_1 \neq \theta_2$. Our goal is to estimate τ^*. To state our change-detection algorithm, first note that the negative of the log-likelihood function can be interpreted as a code-length, modulo a constant. Thus we first set for any feasible θ

$$C_n(Y_n; \theta) := -\log p(Y_n \mid Y_{n-1}, \ldots, Y_0; \theta), \tag{4.7}$$

and then define the *residual-process*

$$X_n := C_n(Y_n; \theta_1) - C_n(Y_n; \theta_2). \tag{4.8}$$

We certainly get, by the Kullback-Leibler inequality,

$$\mathbb{E}_{\theta^*}(X_n) < 0 \quad \text{for} \quad n \leq \tau^* - 1,$$

and also, in the case of $\tau^* = 0$,

$$\mathbb{E}_{\theta^*}(X_n) > 0 \quad \text{for} \quad n \geq \tau^*.$$

The CUSUM statistics defined in terms of this residual process (X_n) yields the desired change detection algorithm.

To study the probabilistic properties of the resulting CUSUM statistics, we first note, that under suitable conditions, the process $C_n(Y_n; \theta)$ is L-mixing. To state the precise result define

$$\delta(y) = \frac{\max_x b^x(y)}{\min_x b^x(y)}.$$

We will indicate with $Q^* = Q(\theta^*)$ the true transition matrix, and with $Q = Q(\theta)$ the estimated one. We have the following

Theorem 4.1.1 (see Theorem 5.2 in [34]). *Consider a hidden Markov process (χ_n, Y_n). Assume that the transition probability matrices Q^* and Q, corresponding to θ^* and θ, respectively, are primitive*[1]*, and that for all $x \in \mathcal{X}$ we have $b^x(y) > 0$ for λ almost all $y \in \mathcal{Y}$. Furthermore assume that for all $s \geq 1$ and for all $i, j \in \mathcal{X}$*

$$\int |\log b^j(y)|^s \; b^{*i}(y)\lambda(dy) < \infty, \tag{4.9}$$

and also that for all $s \geq 1$ and for all $i \in \mathcal{X}$

$$\int |\delta(y)|^s b^{*i}(y)\lambda(dy) < \infty. \tag{4.10}$$

Then the process $C_n(Y_n; \theta)$ is L-mixing.

We conclude that under the assumption of no change, i.e. $\tau^* = \infty$, the residual (X_n) defined in equation (4.8) is L-mixing. The conditions (4.9) and (4.10) are certainly satisfied for a finite read-out space $\mathcal{Y}$, assuming that $b^x(y) > 0$ for all x, y. If, in addition, Q^* and Q are positive, then (X_n) is a *bounded sequence.*

4.2 L-mixing processes

For the reader's sake we summarize a few definitions given in ([30]). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let (X_n) be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 4.2.1. *We say that (X_n) is M-*bounded *if for all $1 \leq q < +\infty$*

$$M_q(X) := \sup_{n \geq 0} \| X_n \|_q < +\infty. \tag{4.11}$$

We can also define $M_q(X)$ for $q = +\infty$ as

$$M_\infty(X) := \sup_{n \geq 1} \operatorname{ess\,sup} |X_n|.$$

Let $(\mathcal{F}_n)_{n \geq 1}$ be an increasing family of σ-fields and let $(\mathcal{F}_n^+)_{n \geq 1}$ be a decreasing family of σ-fields, $\mathcal{F}_n \subseteq \mathcal{F}$ and $\mathcal{F}_n^+ \subseteq \mathcal{F}$ for any n. Assume that $\mathcal{F}_n$ and

[1] i.e. there exist an integer r such that Q^r has positive entries.

$\mathcal{F}_n^+$ are independent for any n. Let τ be a positive integer. Define for all $1 \leq q < +\infty$

$$\gamma_q(\tau, X) = \gamma_q(\tau) := \sup_{n \geq \tau} \| X_n - \mathbb{E}(X_n | \mathcal{F}_{n-\tau}^+) \|_q, \tag{4.12}$$

$$\Gamma_q(X) := \sum_{\tau=0}^{+\infty} \gamma_q(\tau, X). \tag{4.13}$$

We can also define

$$\gamma_\infty(\tau, X) := \sup_{n \geq \tau} \operatorname{ess\,sup} |X_n - \mathbb{E}(X_n | \mathcal{F}_{n-\tau}^+)|,$$

and

$$\Gamma_\infty(X) := \sum_{\tau=0}^{+\infty} \gamma_\infty(\tau, X).$$

Definition 4.2.2. *A process (X_n) is* L-mixing *with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ if*

- *X_n is $\mathcal{F}_n$-measurable for all $n \geq 1$,*
- *(X_n) is M-bounded,*
- *$\Gamma_q(X) < +\infty$ for all $1 \leq q < +\infty$.*

The definition can be generalized for parameter dependent processes (see [30]). We give some examples of L-mixing processes.

Example 4.2.1. *Let (X_n) be an i.i.d. process. Define $\mathcal{F}_n := \sigma(X_i | i \leq n)$ and $\mathcal{F}_n^+ := \sigma(X_i | i \geq n+1)$. If the moments of (X_n) are all finite, then (X_n) is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$.*

Example 4.2.2. *Let $(e_n)_{n \geq 1}$ be an M-bounded independent sequence, $e_n \in \mathbb{R}^k$. Define the vector-valued process $(y_n)_{n \geq 1}$ by*

$$\begin{aligned} x_{n+1} &= Ax_n + Be_n \qquad x_1 = 0 \\ y_n &= Cx_n, \end{aligned}$$

with $A \in \mathbb{R}^{m \times m}$ stable, $B \in \mathbb{R}^{m \times k}$, and $C \in \mathbb{R}^{p \times m}$. Let $\mathcal{F}_n := \sigma(e_i | i \leq n-1)$ and $\mathcal{F}_n^+ := \sigma(e_i | i \geq n)$. Then $(y_n)_{n \geq 1}$ is an L-mixing process with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$.

Proof. For any n, y_n is $\mathcal{F}_n$-adapted, because so it is x_n. We have

$$x_n = \sum_{k=1}^{n-1} A^{n-1-k}\, B\, e_k$$

Thus

$$\begin{aligned}\| y_n \|_q &\leq \| C \| \| B \| \sum_{k=1}^{n-1} \| A \|^{n-1-k} \| e_k \|_q \\ &\leq M_q(e) \ \| C \| \| B \| \sum_{k=1}^{n-1} \| A \|^{n-1-k} < +\infty.\end{aligned}$$

Thus $(y_n)_{n\geq 1}$ is M-bounded. For any $n \leq \tau$ we have

$$x_n = A^\tau x_{n-\tau} + \sum_{k=0}^{\tau-1} A^k B e_{n-k-1}, \text{ and so}$$

$$\| x_n - \mathbb{E}\{x_n | \mathcal{F}^+_{n-\tau}\} \|_q = \| A^\tau x_{n-\tau} - A^\tau \mathbb{E}\{x_{n-\tau}\} \|_q .$$

We get

$$\| y_n - \mathbb{E}\{y_n | \mathcal{F}^+_{n-\tau}\} \|_q \leq \| C \| \| \| A \|^\tau \, 2M_q(x).$$

Recalling the definitions in the equations (4.12) and (4.13)

$$\begin{aligned}\gamma_q(\tau, y) &= \sup_{n\geq\tau} \| y_n - \mathbb{E}\{y_n | \mathcal{F}^+_{n-\tau}\} \|_q \leq \| C \| \| \| A \|^\tau \, 2M_q(x), \\ \Gamma_q(\tau, y) &= \sum_{\tau\geq 0} \gamma_q(\tau, y) \leq \| C \| \, 2M_q(x) \sum_{\tau\geq 0} \| A \|^\tau < +\infty.\end{aligned}$$

Thus $(y_n)_{n\geq 1}$ is L-mixing indeed. □

Counterexample 4.2.1. *The exchangeable sequence proposed in Example 1.1.1 is clearly not L-mixing.*

Observation 4.2.1. *Let (X_n) be an L-mixing process with respect to $(\mathcal{F}_n, \mathcal{F}^+_n)$. Let f be a real Lipschitz-continuous bounded function. Then $(f(X_n))_{n\geq 1}$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}^+_n)$.*

To verify L-mixing property, $\gamma_q(\tau, X)$ has to be estimated. It is in general difficult to give an explicit form of $\gamma_q(\tau, X)$. By the following Lemma 4.2.1, to get an upper bound for $\gamma_q(\tau, X)$ it is sufficient to find, for any n, a $\mathcal{F}^+_{n-\tau}$-measurable random variable, that well approximates X_n.

Lemma 4.2.1. *Let $\mathcal{F}' \subset \mathcal{F}$ be two σ-algebras. Let ξ be an M-bounded $\mathcal{F}$-measurable random variable. Then for any $1 \leq q < +\infty$ and any $\mathcal{F}'$-measurable random variable η we have*

$$\| \xi - \mathbb{E}(\xi|\mathcal{F}') \|_q \leq 2 \| \xi - \eta \|_q \, . \tag{4.14}$$

Centered L-mixing processes satisfy the strong law of large numbers, see Corollary 1.3 in [30]:

Theorem 4.2.1. *Let (X_n) be a real valued L-mixing process such that $\mathbb{E}(X_n) = 0$ for all $n \geq 0$. Then*

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} X_i = 0 \qquad \text{with probability 1.} \tag{4.15}$$

Chapter 5

L-mixing property for the CUSUM statistics

5.1 Equivalent formulations for g_n

In the following Propositions 5.1.1 and 5.1.2 we give two equivalent formulations for the CUSUM statistics (g_n), which will be useful in the forthcoming computations.

Proposition 5.1.1. *Let (X_n) be a stochastic process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let (g_n) be the CUSUM statistics defined as in equation (3.25). Then*

$$g_n = (g_{n-1} + X_n)_+ \quad \textit{with} \quad g_0 = 0 \tag{5.1}$$

where $a_+ = \max\{0, a\}$ and $n \geq 1$.

The previous proposition defines (g_n) in a recursive form, so that g_n is a deterministic function of g_{n-1} and X_n. The following Proposition 5.1.2 gives an alternative representation of g_n that will be useful for further calculations. It can be easily proved with induction arguments.

Proposition 5.1.2. *Let (X_n) be a stochastic process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let (g_n) be the CUSUM statistics defined as in equation (3.25). Then for $n \geq 1$*

$$g_n = \max_{1 \leq i \leq n} (X_i + \cdots + X_n)_+. \tag{5.2}$$

5.2 The CUSUM with i.i.d. input

In this Section we analyze properties of the process (g_n) defined by equation (3.25), when (X_n) is a sequence of independent identically distributed random variables. By Proposition 5.1.1, it is clear that (g_n) is a homogeneous Markov chain. This special case shows up in a variety of problems, like queuing theory (see [50] Chapter 1 and [4] Chapters 1.5 and 3.6), and in the theory of risk processes (see [46]).

Meyn and Tweedie (see [42]) call (g_n) a random walk on the half line. They prove stability properties for the Markov chain (g_n), under the hypothesis that $\mathbb{E}(X_1) < 0$. In particular they prove the existence of a unique invariant measure, for (X_n) i.i.d. In fact in Proposition 8.5.1. they show that (g_n) is recurrent under the hypothesis $\mathbb{E}(X_1) < 0$, and thus by Theorem 10.0.1. in [42] the homogeneous Markov chain (g_n) admits a unique invariant measure.

Moreover in [42] the so-called geometric ergodicity of (g_n) is established, using a fairly complex machinery, under the additional technical assumption that $\mathbb{E}(\exp c' X_1) < \infty$ for some $c' > 0$. More precisely in Section 16.1.3 in [42] they verify a drift condition, with Lyapunov function $V(x) = \exp(cx)$, under the condition $\mathbb{E}(\exp c' X_1) < \infty$, with $0 < c \leq c'$.

The recurrence and the geometric ergodicity of (g_n) are obtained also in [4] Chapter 1.5 Example 5.7, but for (X_n) taking values on the natural numbers.

According to Proposition 5.1.1, the sequence (g_n) can be generated in a convenient way by repeated applications of random functions. Letting $f_X(g) := (g + X)_+$ we have

$$g_n = f_{X_n} f_{X_{n-1}} f_{X_{n-2}} \dots f_{X_1}(g_0). \tag{5.3}$$

A nice source on iterated random functions is Diaconis and Freedman [23]. Consider a set $\{f_\theta, \theta \in \Theta\}$ of measurable functions of a metric space S into itself. Fix a probability measure μ on Θ, and a starting point v_0 in S, and define inductively the process

$$V_{n+1} = f_{\theta_{n+1}}(V_n) = f_{\theta_{n+1}} \dots f_{\theta_2} f_{\theta_1}(v_0), \tag{5.4}$$

where $\theta_1, \theta_2, \dots, \theta_n$ are independent draws from μ. Relation (5.4) is called the *forward iteration.* If v_0 is independent from $\theta_1, \theta_2, \dots$, then the resulting process $(V_n)_{n\geq 0}$ is a Markov chain.

Let the functions f_θ be Lipschitz-continuous with Lipschitz constant K_θ. It is proved in [23] that if

$$\int \log K_\theta \, \mu(d\theta) < 0,$$

and some simple additional technical conditions hold, then $(V_n)_{n\geq 1}$ has a unique stationary distribution (see Theorem 1.1 in [23]).
Unfortunately this result is not applicable to the analysis of (g_n) defined above by equation (5.3), although the functions f_{X_i} in (5.3) are Lipschitz-continuous, but with Lipschitz constants equal to 1. On the other hand, as it is mentioned in Paragraph 4 in [23], if (X_n) is an i.i.d. sequence, and $\mathbb{E}(X_1) < 0$, using a backward iteration it can be shown that (g_n) has an invariant measure.

5.2.1 L-mixing property

The purpose of this section is to provide a simple, but useful result complementing the above results. For its formulation we need the following notations:

$$\mathcal{F}_n := \sigma(X_i \,|\, i \leq n) \quad \text{and} \quad \mathcal{F}_n^+ := \sigma(X_i \,|\, i \geq n+1).$$

Thus $\mathcal{F}_n$ is the past, and $\mathcal{F}_n^+$ is the future of (X_n). Assume that

$$\mathbb{E}(X_1) < 0,$$

and, in addition, $E(\exp c" X_1) < \infty$ for some $c" > 0$. Then

$$\mu := \mathbb{E}(\exp c' X_1) < 1 \quad \text{for some} \quad c' > 0. \tag{5.5}$$

Theorem 5.2.1. *Let (X_n) be a sequence of i.i.d. random variables such that (5.5) holds. Then (g_n), defined by equation (3.25), is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$.*

Proof. of Theorem 5.2.1 For the proof we will use the equivalent formulation for (g_n) given in Proposition 5.1.2:

$$g_n = \max_{1\leq i\leq n} (X_i + \cdots + X_n)_+. \tag{5.6}$$

First note that for any $n \in \mathbb{N}$, $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent, and (g_n) is $\mathcal{F}_n$-adapted. Standard results of the theory of risk processes imply that for any

c such that $0 < c < c'$, we have $\mathbb{E}(\exp cg_n) < \infty$, hence (g_n) is M-bounded. (This will also follow from the arguments below.)

To show that (g_n) is L-mixing we have to show that for $1 \leq q < +\infty$

$$\gamma_q(\tau, g) := \sup_{n \geq \tau} \| g_n - \mathbb{E}(g_n | \mathcal{F}^+_{n-\tau}) \|_q$$

is summable. For this purpose we shall make use of Lemma 4.2.1 in 4.2, implying that

$$\| g_n - \mathbb{E}(g_n | \mathcal{F}^+_{n-\tau}) \|_q \leq 2 \| g_n - g^{++}_{n,n-\tau} \|_q, \tag{5.7}$$

for any $\mathcal{F}^+_{n-\tau}$-measurable random variable $g^{++}_{n,n-\tau}$. In particular, define

$$g^{++}_{n,n-\tau} := \max_{n-\tau+1 \leq i \leq n} (X_i + \cdots + X_n)_+. \tag{5.8}$$

Note that $g^{++}_{n,n-\tau}$ is $\mathcal{F}^+_{n-\tau}$ measurable, as required.
To estimate $g_n - g^{++}_{n,n-\tau}$ we use Lemma A.0.3 in the Appendix, setting

$$I_1 = \{n - \tau + 1, \ldots, n\}, \quad I_2 = \{1, \ldots, n - \tau\}, \quad A_i = (X_i + \cdots + X_n)_+\,.$$

We get

$$\begin{aligned} g_n - g^{++}_{n,n-\tau} &= \max_{1 \leq i \leq n} (X_i + \cdots + X_n)_+ - \max_{n-\tau+1 \leq i \leq n} (X_i + \cdots + X_n)_+ \\ &\leq \max_{1 \leq i \leq n-\tau} (X_i + \cdots + X_n)_+ =: g_{n,n-\tau}. \end{aligned} \tag{5.9}$$

At this point we could continue as in the proof of Theorem 5.3.1 in the next section, but for the sake of variation we follow a slightly different route. But first we have to fix some notations. Given two random variables X and Y, with distribution functions F_X and F_Y, respectively, we write $X \stackrel{\mathcal{L}}{=} Y$ to indicate that $F_X(x) = F_Y(x)$ for all x, and we write $X \stackrel{\mathcal{L}}{\leq} Y$ to indicate that $F_X(x) \geq F_Y(x)$ for all x. In this case we say that X is *stochastically smaller* than Y.

Now, exploiting the fact that (X_n) is i.i.d., (5.9) can be continued as follows:

$$g_{n,n-\tau} \stackrel{\mathcal{L}}{=} \max_{\tau+1 \leq j \leq n} (X_1 + \cdots + X_j)_+ \stackrel{\mathcal{L}}{\leq} \max_{j \geq \tau+1} (X_1 + \cdots + X_j)_+ =: g^*_{\tau+1}. \tag{5.10}$$

To complete the proof of Theorem 5.2.1, we need Lemmas 5.2.1 and 5.2.2 below, extending known result in risk theory to estimate the tail probability of default (see [48] and [46]) using exponential moments.

Lemma 5.2.1. *Let (X_n) be as in Theorem 5.2.1, and let μ and c' as in (5.5). Define for some fixed integer $\tau > 0$*

$$g^*_\tau := \sup_{i \geq \tau} \big((X_1 + \cdots + X_\tau) + \cdots + X_i\big)_+. \tag{5.11}$$

Then for any c such that $0 < c < c'$, we have

$$\mathbb{E}\big(\exp cg^*_\tau\big) \leq 1 + \Big(\frac{c}{c'-c}\Big)\frac{\mu^\tau}{1-\mu}. \tag{5.12}$$

Since $\mu < 1$, the lemma states the exponential decay of $\mathbb{E}\big(\exp cg^*_\tau\big) - 1$ with respect to τ. The lemma is a direct consequence of the following tail-probability estimates:

Lemma 5.2.2. *Let (X_n) be as in Theorem 5.2.1, and let μ and c' as in (5.5). Let g^*_τ be as in Lemma 5.2.1. Then for each $x \geq 0$ we have*

$$\mathbb{P}\big(g^*_\tau > x\big) \leq \frac{\mu^\tau}{1-\mu}\exp(-c'x). \tag{5.13}$$

Proof. of Lemma 5.2.2. We have

$$\mathbb{P}(g^*_\tau > x) \leq \sum_{i \geq \tau}^{+\infty} \mathbb{P}((X_1 + \cdots + X_i) > x) \tag{5.14}$$

$$\leq \sum_{i\geq\tau}^{+\infty} \mathbb{E}\big(\exp c'(X_1 + \cdots + X_i)\big)/\exp(c'x)$$

$$= \sum_{i\geq\tau}^{+\infty} \mu^i/\exp(c'x) = \frac{\mu^\tau}{1-\mu}\exp(-c'x).$$

□

Proof. of Lemma 5.2.1. We have

$$\mathbb{E}\big(\exp cg^*_\tau\big) = \int_0^{+\infty} \mathbb{P}(\exp cg^*_\tau > x)dx. \tag{5.15}$$

Now for $x \geq 1$ we get by Lemma 5.2.2

$$\begin{aligned}\mathbb{P}\big(\exp cg^*_\tau > x\big) &= \mathbb{P}\Big(g^*_\tau > \frac{\log x}{c}\Big) \leq \frac{\mu^\tau}{1-\mu}\exp\Big(-\frac{c'\log x}{c}\Big)\\ &\leq \frac{\mu^\tau}{1-\mu}x^{-c'/c}.\end{aligned} \tag{5.16}$$

For $x \leq 1$ we have $\mathbb{P}(\exp cg_\tau^* > x) = 1$. Combining (5.15) and (5.16) we get

$$\begin{aligned}\mathbb{E}(\exp cg_\tau^*) &= 1 + \int_1^{+\infty} \mathbb{P}(\exp cg_\tau^* > x)dx \\ &= 1 + \frac{\mu^\tau}{1-\mu}\int_1^{+\infty} x^{-c'/c}dx \leq 1 + \left(\frac{c}{c'-c}\right)\frac{\mu^\tau}{1-\mu}.\end{aligned} \tag{5.17}$$

□

Corollary 5.2.1. *Under the conditions of Lemma 5.2.2, for any integer $p \geq 1$*

$$\| g_\tau^* \|_p \leq H_p \mu^{\tau/p}, \tag{5.18}$$

where

$$H_p := \frac{1}{c}\left(\frac{c}{c'-c}\right)^{1/p}\left(\frac{p!}{1-\mu}\right)^{1/p}.$$

Proof. Using a Taylor expansion for $\exp cg_\tau^*$ we get

$$\exp cg_\tau^* \geq 1 + (c)^p \frac{(g_\tau^*)^p}{p!}. \tag{5.19}$$

Taking the expectation in the last equation, the claim follows directly from Lemma 5.2.1. □

We continue the proof of Theorem 5.2.1. To prove that (g_n) is M-bounded, we need an upper bound for the moments of (g_n). We have

$$\begin{aligned} g_n &= \max_{1\leq i\leq n}(X_i + \cdots + X_n)_+ \overset{\mathcal{L}}{=} \max_{1\leq i\leq n}(X_1 + \cdots + X_i)_+ \\ &\overset{\mathcal{L}}{\leq} \max_{i\geq 1}(X_1 + \cdots + X_i)_+ = g_1^*.\end{aligned} \tag{5.20}$$

In short we get

$$g_n \overset{\mathcal{L}}{\leq} g_1^*. \tag{5.21}$$

Fix q, $1 \leq q < +\infty$ and let $p := \lceil q \rceil$ be the first integer greater or equal to q. Using Corollary 5.2.1 we get

$$\| g_n \|_q \leq \| g_n \|_p \leq \| g_1^* \|_p \leq H_p \mu^{1/p}. \tag{5.22}$$

Thus we get that (g_n) is M-bounded (see Definition 4.2.1 in 4.2).

Recall equations (5.7), (5.9), and (5.10). To conclude the proof we need an upper bound for $\| g^*_{\tau+1} \|_q$. We get it using Corollary 5.2.1 with $p := \lceil q \rceil$:

$$\| g^*_{\tau+1} \|_q \leq \| g^*_{\tau+1} \|_p \leq H_p \mu^{(\tau+1)/p}. \tag{5.23}$$

Thus we get

$$\| g_n - \mathbb{E}(g_n | \mathcal{F}^+_{n-\tau}) \|_q \leq 2 \| g^*_{\tau+1} \|_q \leq 2 H_p \mu^{(\tau+1)/p}. \tag{5.24}$$

The right hand side is obviously summable, hence (g_n) is L-mixing indeed. □

5.3 The CUSUM with L-mixing input

Consider now the case when the input (X_n) is L-mixing. This condition is motivated by change detection problems for HHMs, as we have seen in Section 4.1.1. We will show that (g_n) is L-mixing for L-mixing input, under two additional technical assumptions. The first one is fairly mild, requiring that (X_n) is an L-mixing process with respect to $(\mathcal{F}_n, \mathcal{F}^+_n)$ such that

$$\sum_{\tau=0}^{+\infty} \tau \gamma_q(\tau, X) < +\infty \qquad \text{for all} \quad 1 \leq q < +\infty. \tag{5.25}$$

The second assumption is much more restrictive, saying that

$$M_\infty(X) < +\infty \qquad \text{and} \qquad \Gamma_\infty(X) < +\infty. \tag{5.26}$$

This condition will be discussed in the remark following Lemma 5.3.2. We define a critical exponent in terms of $M_\infty(X)$ and $\Gamma_\infty(X)$ as follows:

$$\beta^* := \varepsilon / (4 M_\infty(X) \Gamma_\infty(X)). \tag{5.27}$$

Then, for any $\beta' \leq \beta^*$ define

$$\lambda = \lambda(\beta') := \exp\left(4 M_\infty(X) \Gamma_\infty(X) (\beta')^2 - \beta' \varepsilon\right). \tag{5.28}$$

Note that for the critical value β^* we have $\lambda(\beta^*) = 1$, and for $\beta' < \beta^*$ we have $\lambda(\beta') < 1$. The main result of this section is then the following:

Theorem 5.3.1. *Let (X_n) be an L-mixing process w.r.t. $(\mathcal{F}_n, \mathcal{F}_n^+)$ such that (5.25) and (5.26) are satisfied, and*

$$\mathbb{E}(X_n) \le -\varepsilon < 0 \qquad \text{for all } n \ge 0. \tag{5.29}$$

Let (g_n) be defined as in (3.25). Then (g_n) is L-mixing w.r.t. $(\mathcal{F}_n, \mathcal{F}_n^+)$. In addition, for any β, β' such that $0 < \beta < \beta' < \beta^*$, we have with $\lambda = \lambda(\beta')$

$$\mathbb{E}\big(\exp \beta\, g_n\big) \le 1 + \Big(\frac{\beta}{\beta' - \beta}\Big)\frac{\lambda}{1-\lambda}. \tag{5.30}$$

Proof. of Theorem 5.3.1. To prove the theorem, define for $0 \le \tau \le n$ the auxiliary process

$$g_{n,n-\tau}(X) := \max_{1 \le i \le n-\tau} (X_i + \cdots + X_n)_+\,, \tag{5.31}$$

The exponential moments of $g_{n,n-\tau}(X)$ will be bounded as follows:

Lemma 5.3.1. *Let (X_n) and β, β' and λ be as in Theorem 5.3.1. Then*

$$\mathbb{E}\big(\exp \beta\, g_{n,n-\tau}(X)\big) \le 1 + \Big(\frac{\beta}{\beta' - \beta}\Big)\frac{\lambda^{\tau+1}}{1-\lambda}. \tag{5.32}$$

The result ensures an exponential decay of $\mathbb{E}\big(\exp \beta g_{n,n-\tau}(X)\big) - 1$ in τ, a property which will be used later.

Lemma 5.3.2. *Let (X_n) and β' and λ be as in Theorem 5.3.1. Then for any $x \ge 0$ we have*

$$\mathbb{P}\big(g_{n,n-\tau}(X) > x\big) \le \frac{\lambda^{\tau+1}}{1-\lambda}\exp(-\beta' x). \tag{5.33}$$

For the proof of the lemma we need an exponential inequality for partial sums of L-mixing processes. We do have such an inequality, see [31], for the case $M_\infty(X) < +\infty$ and $\Gamma_\infty(X) < +\infty$. Unfortunately, it is not clear if this inequality can be extended to unbounded processes. It seems that the boundedness of (X_n) is a common assumption for exponential inequalities for partial sums of mixing processes, see Section 1.4.2. in [25].

Proof. of Lemma 5.3.2. We follow the arguments of the proof of Theorem 3.1 in [31]. To estimate tail probabilities of partial sums a natural tool, developed in large deviation theory, is an appropriate exponential inequality. First we estimate $\mathbb{E}\big(\exp\beta'(X_i+\cdots+X_n)\big)$, $1\le i\le n-\tau$. Define

$$D_j := X_j - \mathbb{E}(X_j),$$

for all $j\ge 1$. Obviously $\mathbb{E}(D_j)=0$ for all j, $M_\infty(D)\le 2M_\infty(X)$, and $\Gamma_\infty(D)=\Gamma_\infty(X)$. By the exponential inequality, given as Theorem 5.1 in [31], applied to the process $(D_j)_{i\le j\le n}$ with weights $f_j=\beta'$ we obtain

$$\mathbb{E}\Big(\exp\big(\beta'\sum_{j=i}^{n} D_j - 2M_\infty(D)\Gamma_\infty(D){\beta'}^2(n-i+1)\big)\Big)\le 1.$$

After rearrangement and multiplication by $\exp\beta'\sum_{j=i}^{n}\mathbb{E}(X_j)$, we get

$$\mathbb{E}\Big(\exp\Big(\beta'\sum_{j=i}^{n}\big(D_j+\mathbb{E}(X_j)\big)\Big)\Big)\le\exp\Big(\alpha{\beta'}^2(n-i+1)+\beta'\sum_{j=i}^{n}\mathbb{E}(X_j)\Big),$$

with $\alpha := 2M_\infty(D)\Gamma_\infty(D)$. Noting that $D_j+\mathbb{E}(X_j)=X_j$, $\mathbb{E}(X_j)\le-\varepsilon$, and $\alpha\le 4M_\infty(X)\Gamma_\infty(X)$, we conclude that

$$\mathbb{E}\Big(\exp\beta'\sum_{j=i}^{n}X_j\Big)\le\exp\Big(4M_\infty(X)\Gamma_\infty(X){\beta'}^2(n-i+1)-\beta'\varepsilon(n-i+1)\Big).$$

Take $\beta'<\beta^*$. Recalling the definition of $\lambda(\beta')$ we get

$$\mathbb{E}\Big(\exp\beta'\sum_{j=i}^{n}X_j\Big)\le\Big(\exp\big(4M_\infty(X)\Gamma_\infty(X){\beta'}^2-\beta'\varepsilon\big)\Big)^{(n-i+1)}=\lambda(\beta')^{n-i+1}.$$

Now, for $\beta'<\beta^*$, we have $\lambda=\lambda(\beta')<1$, and thus we obtain for $x\ge 0$

$$\mathbb{P}\big(h_{n,n-\tau}(X)>x\big)\le\sum_{i=1}^{n-\tau}\mathbb{P}\Big(\big(X_i+\cdots+X_n\big)_+>x\Big)$$
$$\le\sum_{i=1}^{n-\tau}\mathbb{E}\big(\exp\beta'(X_i+\ldots X_n)\big)/\exp(\beta'x)\le\sum_{i=1}^{n-\tau}\lambda^{n-i+1}/\exp(\beta'x)$$
$$\le\sum_{l=\tau+1}^{n}\lambda^l/\exp(\beta'x)\le\sum_{l=\tau+1}^{+\infty}\lambda^l/\exp(\beta'x)=\frac{\lambda^{\tau+1}}{1-\lambda}\exp(-\beta'x).$$

□

Proof. of Lemma 5.3.1 Using Lemma 5.3.2 the proof is carried out exactly as the proof of Lemma 5.2.1. □

Corollary 5.3.1. *Under the conditions and notations of Theorem 5.3.1 we have*

$$\| g_{n,n-\tau}(X) \|_p \leq K_p \lambda^{(\tau+1)/p} \tag{5.34}$$

for any integer $p \geq 1$, where

$$K_p := \frac{1}{\beta} \left(\frac{\beta}{\beta' - \beta} \right)^{1/p} \left(\frac{p!}{1-\lambda} \right)^{1/p}.$$

Proof. See the proof of Corollary 5.2.1. □

(X_n) is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$, thus X_n is $\mathcal{F}_n$-adapted for any $n \in \mathbb{N}$. It follows that (g_n) is $\mathcal{F}_n$-adapted. First we show that (g_n) is M-bounded. By definition

$$g_n = g_{n,n}(X).$$

For any fixed q, let $p := \lceil q \rceil$. Then by Corollary 5.3.1, we have

$$\| g_n \|_q \leq \| g_n \|_p \leq K_p \lambda^{1/p}. \tag{5.35}$$

To show that $\Gamma_q(g)$ is finite for any $1 \leq q < +\infty$, we estimate $\gamma_q(\tau, g)$ using Lemma 4.2.1 in 4.2. Let

$$X^+_{i,n-\tau} := \mathbb{E}\big(X_i | \mathcal{F}^+_{n-\tau}\big).$$

Note that (X_n) is L-mixing, and thus for $i \geq n - \lceil \frac{\tau}{2} \rceil + 1$, or $i - (n-\tau) \geq \tau - \lceil \frac{\tau}{2} \rceil + 1$, $X^+_{i,n-\tau}$ is a good approximation of X_i. Approximate g_n by

$$g^{++}_{n,n-\tau} := \max_{n - \lceil \frac{\tau}{2} \rceil + 1 \leq i \leq n} (X^+_{i,n-\tau} + \cdots + X^+_{n,n-\tau})_+. \tag{5.36}$$

Note that $g^{++}_{n,n-\tau}$ is $\mathcal{F}^+_{n-\tau}$ measurable, as required. For each τ, define

$$\gamma_q^{++}(\tau) := \sup_{n \geq \tau} \| g_n - g^{++}_{n,n-\tau} \|_q \qquad \text{and} \qquad \Gamma_q^{++}(g) := \sum_{\tau=0}^{+\infty} \gamma_q^{++}(\tau). \tag{5.37}$$

By Lemma 4.2.1 in Chapter 4.2 we have

$$\Gamma_q(g) \leq 2\Gamma_q^{++}(g). \tag{5.38}$$

So to show that $\Gamma_q(g) < +\infty$, it is thus enough to prove that $\Gamma_q^{++}(g) < +\infty$.

To estimate the residual $g_n - g^{++}_{n,n-\tau}$ we shall use an *intermediate* approximation of g_n, defined as

$$\overline{g}_{n,n-\tau} := \max_{n-\lceil\frac{\tau}{2}\rceil+1\le i\le n} (X_i + \cdots + X_n)_+. \tag{5.39}$$

$\overline{g}_{n,n-\tau}$ is similar to the approximation of g_n we have used in the i.i.d. case, but $\overline{g}_{n,n-\tau}$ is not $\mathcal{F}^+_{n-\tau}$-measurable. By the triangular inequality we have

$$\| g_n - g^{++}_{n,n-\tau} \|_q \le \| g_n - \overline{g}_{n,n-\tau} \|_q + \| \overline{g}_{n,n-\tau} - g^{++}_{n,n-\tau} \|_q . \tag{5.40}$$

Define

$$\overline{\gamma}_q(\tau) := \sup_{n\ge\tau} \| g_n - \overline{g}_{n,n-\tau} \|_q \qquad \text{and} \qquad \overline{\Gamma}_q(g) := \sum_{\tau=0}^{+\infty} \overline{\gamma}_q(\tau),$$

$$\overline{\gamma}^{++}_q(\tau) := \sup_{n\ge\tau} \| \overline{g}_{n,n-\tau} - g^{++}_{n,n-\tau} \|_q \qquad \text{and} \qquad \overline{\Gamma}^{++}_q(g) := \sum_{\tau=0}^{+\infty} \overline{\gamma}^{++}_q(\tau).$$

Taking $\sup_{n\ge\tau}$ in equation (5.40) and summing over τ we get

$$\Gamma^{++}_q(g) \le \overline{\Gamma}_q(g) + \overline{\Gamma}^{++}_q(g). \tag{5.41}$$

So it is enough to show that $\overline{\Gamma}_q(g) < +\infty$ and $\overline{\Gamma}^{++}_q(g) < +\infty$.
First estimate $\| g_n - \overline{g}_{n,n-\tau} \|_q$. Taking

$$I = \{1,\dots,n\}, \quad I_1 = \{1,\dots,n-\lceil\frac{\tau}{2}\rceil\}, \quad I_2 = \{n-\lceil\frac{\tau}{2}\rceil+1,\dots,n\},$$
$$\text{and} \quad A_i = (X_i + \cdots + X_n)_+ ,$$

and applying Lemma A.0.3 we get

$$g_n - \overline{g}_{n,n-\tau} \le \max_{1\le i\le n-\lceil\frac{\tau}{2}\rceil} (X_i + \cdots + X_n)_+ = g_{n,n-\lceil\frac{\tau}{2}\rceil}(X).$$

Let $p := \lceil q \rceil$. Using Corollary 5.3.1, we have

$$\| g_n - \overline{g}_{n,n-\tau} \|_q \le \| g_n - \overline{g}_{n,n-\tau} \|_p \le \| g_{n,n-\lceil\frac{\tau}{2}\rceil}(X) \|_p \le K_p \lambda^{(\lceil\frac{\tau}{2}\rceil+1)/p}.$$

Thus we get for the intermediate approximation error the inequalities

$$\overline{\gamma}_q(\tau) = \sup_{n\ge\tau} \| g_n - \overline{g}_{n,n-\tau} \|_q \le K_p \lambda^{(\lceil\frac{\tau}{2}\rceil+1)/p}, \tag{5.42}$$

$$\overline{\Gamma}_q(g) = \sum_{\tau=0}^{+\infty} \overline{\gamma}_q(\tau) \le K_p \sum_{\tau=0}^{+\infty} \lambda^{(\lceil\frac{\tau}{2}\rceil+1)/p} < +\infty. \tag{5.43}$$

Next we need to find an upper bound for $\| \overline{g}_{n,n-\tau} - g^{++}_{n,n-\tau} \|_q$. We use Lemma A.0.2 with $\hat{a} = \overline{g}_{n,n-\tau}$ and $\hat{b} = g^{++}_{n,n-\tau}$. Thus we get

$$\| \overline{g}_{n,n-\tau} - g^{++}_{n,n-\tau} \|_q \leq \sum_{i=n-\lceil \frac{\tau}{2} \rceil +1}^{n} \| X_i - X^{+}_{i,n-\tau} \|_q$$

$$= \sum_{i=n-\lceil \frac{\tau}{2} \rceil +1}^{n} \| X_i - \mathbb{E}(X_i \mid \mathcal{F}^{+}_{n-\tau}) \|_q \leq \sum_{j=\lfloor \frac{\tau}{2} \rfloor +1}^{\tau} \gamma_q(j, X).$$

It follows that

$$\overline{\Gamma}^{++}_q(g) = \sum_{\tau=0}^{+\infty} \overline{\gamma}^{++}_q(\tau) \leq \sum_{\tau=0}^{+\infty} \sum_{j=\lfloor \frac{\tau}{2} \rfloor +1}^{\tau} \gamma_q(j, X) \leq \sum_{\tau=0}^{+\infty} \tau \gamma_q(\tau, X) < +\infty, \quad (5.44)$$

where the last sum is finite by the condition stated in equation (5.25). Combining equations (5.38), (5.41), (5.43) and (5.44) we conclude that $\Gamma_q(g) < +\infty$, as stated.

To conclude the proof, it remains to show that the exponential bound given in equation (5.30) holds. But it easily follows by Lemma 5.3.1 recalling that $g_n = g_{n,n}$. □

5.3.1 Comments

In Section 4.1.1 we have presented a CUSUM algorithm for HHMs, with known dynamics before and after the change. This has lead us to the study of the CUSUM statistics with L-mixing input. We have shown in this section that, under certain technical conditions, the output of the CUSUM statistics is L-mixing.

It should be admitted though, that this result is not directly applicable to the change detection of HMMs. Namely, the Hinkley-scores defined for HMMs are typically not bounded. A notable exception is the case of finite state and read-out space with all transition and read-out probabilities being positive. But even in this case the condition $\Gamma_\infty(X) < +\infty$ can not be guaranteed. The technical difficulty in extending this result is the apparent lack of an appropriate exponential inequality for the partial sums of unbounded L-mixing processes, given as Theorem 5.1 in [31], and used in the proof of Lemma 5.3.2.

5.4 Estimation of False Alarm frequency

As a corollary of Theorems 5.2.1 and 5.3.1 we get an upper bound for the a.s. false alarm frequency defined as

$$\limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\{g_n>\delta\}}, \tag{5.45}$$

with the tacit assumption that $\tau^* = +\infty$. In fact the lim sup above is tractable if we have a strong law of large numbers (LLN) for $\mathbb{I}_{\{g_i\geq\delta\}}$ or for $f(g_i)$, with f smooth and $f(g_i) \geq \mathbb{I}_{\{g_i\geq\delta\}}$. This is ensured if (g_n) is L-mixing (see Theorem 4.2.1).

Proposition 5.4.1. *Let* (g_n) *be as above, let the input sequence* (X_n)*,* c'*, and* μ *be as in Theorem 5.2.1. Then*

$$\limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\{g_n\geq\delta\}} \leq \frac{\mu}{1-\mu}\exp(-c'\delta) \quad \textit{for any } \delta > 0. \tag{5.46}$$

Proof. Recalling equation (5.21) in the proof of Theorem 5.2.1, we have

$$g_n \overset{\mathcal{L}}{\leq} g_1^*.$$

Thus by Lemma 5.2.2

$$\mathbb{P}\big(g_n \geq x\big) \leq \frac{\mu}{1-\mu}\exp(-c'x) \quad \text{for any } x \geq 0 \text{ and any } n \geq 1. \tag{5.47}$$

To give an estimation of the false alarm frequency we will use the strong law of large numbers for L-mixing processes, but note that $I_{\{g_i\geq\delta\}}$ itself is not L-mixing, even if (g_n) is L-mixing. We thus use an L-mixing approximation from above of $I_{\{g_i\geq\delta\}}$. To be precise let $\delta' < \delta$ and let f be a smooth Lipschitz-continuous function such that $\mathbb{I}_{\{g\geq\delta\}} \leq f(g) \leq \mathbb{I}_{\{g\geq\delta'\}}$. By Theorem 5.2.1 $(g_n)_{n\geq0}$ is L-mixing, thus $(f(g_n))_{n\geq0}$ is also L-mixing (see Observation 4.2.1 in Chapter 4.2). Using the strong law of large numbers for zero mean L-mixing processes, we get, after centering,

$$\begin{aligned} \limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{g_i\geq\delta\}} &\leq \limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{i=1}^{N} f(g_i) \\ = \limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\big(f(g_i)\big) &\leq \limsup_{N\longrightarrow+\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\big(\mathbb{I}_{\{g_i\geq\delta'\}}\big) \end{aligned} \tag{5.48}$$

Taking into account (5.47), and that δ' is arbitrary, we get the claim. □

Proposition 5.4.2. *Let (g_n) be as above, let the input sequence (X_n) be as in Theorem 5.3.1 and let β', $\lambda = \lambda(\beta')$ be as in Lemma 5.3.1. Then*

$$\limsup_{N \longrightarrow +\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{g_i \geq \delta\}} \leq \frac{\lambda}{1-\lambda} \exp(-\beta' \delta) \quad \textit{for any } \delta > 0. \tag{5.49}$$

Proof. Note that $g_n = g_{n,n}$, thus by Lemma 5.3.2

$$\mathbb{P}\big(g_n \geq x\big) \leq \frac{\lambda}{1-\lambda} \exp(-\beta' x) \quad \text{for any } x \geq 0 \text{ and any } n \geq 1. \tag{5.50}$$

By Theorem 5.3.1 (g_n) is L-mixing. Conclude as in the proof of Proposition 5.4.1. □

Appendix A

Technical Lemmas

In this appendix we state three easy lemmas useful in the proof of Theorems 5.2.1 and 5.3.1 in the previous chapter.

Lemma A.0.1. *Let* $(a_n)_{n\geq 1}$ *and* $(b_n)_{n\geq 1}$ *be two sequences of numbers and for given* m, n*, with* $1 \leq m \leq n$*, let*

$$\hat{a} := \max_{m\leq i\leq n} (a_i + \cdots + a_n)_+ \qquad \textit{and} \qquad \hat{b} := \max_{m\leq i\leq n} (b_i + \cdots + b_n)_+. \tag{A.1}$$

Then

$$\hat{a} - \hat{b} \leq \sum_{i=m}^{n} (a_i - b_i)_+. \tag{A.2}$$

Proof. Let $\hat{a} = a_r + \cdots + a_n$ for some r, $m \leq r \leq n$ and $\hat{b} = b_s + \cdots + b_n$ for some s, $m \leq s \leq n$. Let $s \leq r$, then

$$b_r \cdots + b_n \leq b_s + \cdots + b_r + \cdots + b_n.$$

By the last inequalities

$$\sum_{i=s}^{r-1} b_i \geq 0. \tag{A.3}$$

We have

$$\begin{aligned} \hat{a} - \hat{b} = a_r + \ldots a_n - b_s - \cdots - b_n &= \sum_{i=r}^{n} (a_i - b_i) - \sum_{i=s}^{r-1} b_i \\ &\leq \sum_{i=r}^{n} (a_i - b_i) \leq \sum_{i=r}^{n} (a_i - b_i)_+ \leq \sum_{i=m}^{n} (a_i - b_i)_+ \,. \end{aligned}$$

If $r \leq s$, then $\sum_r^{s-1} b_i \leq 0$, and so

$$\begin{aligned}\hat{a} - \hat{b} = a_r + \dots a_n - b_s - \dots - b_n &= \sum_{i=r}^{s-1} a_i + \sum_{i=s}^{n} (a_i - b_i) \\ &\leq \sum_{i=r}^{s-1} a_i - \sum_{i=r}^{s-1} b_i + \sum_{i=s}^{n} (a_i - b_i) \leq \sum_{i=m}^{n} (a_i - b_i)_+.\end{aligned}$$

The lemma is thus proved. □

Lemma A.0.2. *Let $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ be two stochastic processes on the same probability space. Define $\hat{a}$ and $\hat{b}$ as in equation (A.1). Then*

$$\| \hat{a} - \hat{b} \|_q \leq \sum_{i=m}^{n} \| a_i - b_i \|_q . \tag{A.4}$$

Proof. Equation (A.2) implies

$$\hat{a} - \hat{b} \leq \sum_{i=m}^{n} |a_i - b_i|.$$

Interchanging the role of a and b we get

$$\hat{b} - \hat{a} \leq \sum_{i=m}^{n} |b_i - a_i|.$$

Thus

$$|\hat{a} - \hat{b}| \leq \sum_{i=m}^{n} |a_i - b_i|.$$

The result follows taking the L_q-norm and using the triangular inequality. □

Lemma A.0.3. *Let I be a finite set of the natural numbers, and let $I = I_1 \cup I_2$ with $I_1 \cap I_2 = \emptyset$. Let $(A_i)_{i\in I}$ be a collection of non-negative real numbers. Then*

$$\max_{i\in I} A_i \leq \max_{i\in I_1} A_i + \max_{i\in I_2} A_i \tag{A.5}$$

Proof.

$$\max_{i\in I} A_i = \max\{\max_{i\in I_1} A_i, \max_{i\in I_2} A_i\}.$$

The result follows by direct inspection. □

Bibliography

[1] Aldous, D.J. (1985) *Exchangeability and related topics* in *Ecole d'Été de Probabilités de Saint-Flour XIII - 1983*, Lecture Notes in Mathematics 1117, Springer, Berlin

[2] Anderson, B.D.O. (1999) *The realization problem for hidden Markov models* in *Mathematics of Control, Signals, and Systems*, **12(1)**, 80-120

[3] Arapostathis, A. and Marcus, I. (1990) *Analysis of an identification algorithm arising in the adaptive estimation of Markov chains* in *Math. Control Signals and Systems*, Vol.3, 1-29

[4] Asmussen, S. (2003) *Applied probability and queues*, Springer-Verlag, New York

[5] Baikovicius, J. and Gerencsér, L. (1992) *Change-point detection as model selection* in *Informatica* **3**, 1, 3-20

[6] Bansal, R. and Papantoni-Kazakos, P. (1986) *An algorithm for detecting a change in a stochastic process* in *IEEE Transactions on information Theory*, **32**, 2, 227-235

[7] Basseville, M. and Nikiforov, I.V. (1993) *Detection of abrupt changes: theory and application*, Prentice-Hall

[8] Baum, L.E. and Petrie, T. (1966) *Statistical inference for probabilistic functions of finte state Markov chains* in *The Annals of Mathematical Statistics*, **37**, 1554-1563

[9] Benvenuti, L. and Farina, L. (1998) *A note on minimality of positive realization* in *IEEE Transactions of Circuits and Systems*, **45**, 676-677

[10] Benvenuti, L. and Farina, L. (2006) *Correction to "A note on minimality of positive realization"* in *IEEE Transactions of Circuits and Systems*, **53**, 215

[11] Blackwell, D. and Koopmans, L. (1957) *On the identifiability problem for functions of finite Markov chains* in *The Annals of Mathematical Statistics*, **28**, 1011-1015

[12] Brodsky, B.E. and Darkhovsky, B.S. (1993) *Nonparametric methods in change-point problems*, Kluwer Academic Publishers

[13] Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. (2003) *Model based clustering and visualization of navigation patterns on a web site* in *Data Mining and Knowledge Discovery*, **7**, 393-424.

[14] Chigansky, P. (2009) *Maximum likelihood estimator for hidden Markov models in continuous time* in *Statistical Inference for Stochastic Processes*, Vol. 12, **2**, 139-163

[15] Chow, Y.S. and Teicher, H (1997) *Probability theory: independence, interchangeability, martingales*, Springer, New York

[16] Cinlar,E. (1975) *Introduction to stochastic processes*, Prentice-Hall, Englewood Cliffs

[17] Cramer, H. (1946) *Mathematical methods of statistics*, Princeton University Press

[18] de Finetti, B. (1937) *La prévision: ses lois logiques, ses sources subjectives* in *Annales de l'Institut Henri Poincaré*, **7**, 1-68

[19] de Finetti, B. (1938) *Sur la condition d' equivalence partielle* in *Actualité Scientifiques et Industrielles*, Hermann, Paris, **739**, 5-18

[20] Di Cecco, D. (2009) *A class of models for multiple binary sequences under the hypothesis of Markov exchangeability* in *Electronic Journal of Statistics*, **3**, 1113-1132

[21] Dharmadhikari, S.W. (1964) *Exchangeable processes which are function of stationary Markov chains* in *The Annals of Mathematical Statistics*, **35**, 429-430

[22] Diaconis, P. and Freedman, D. (1980) *de Finetti's theorem for Markov chains* in The Annals of Probability, **8**, 115-130

[23] Diaconis, P. and Freedman, D. (1999) *Iterated random functions* in SIAM Review, **41**, 1, 45-76

[24] Douc,R., Moulines, E., Olsson, J. and van Handel, R. (2010) *Consistency of the maximum likelihood estimator for general hidden Markov models*, *preprint*

[25] Doukhan, P. (1994) *Mixing: properties and examples*, Springer-Verlag, New York

[26] Feller, W. (1966) *An introduction to probability theory and its applications*, Wiley

[27] Fortini, S., Ladelli, L., Petris, G. and Regazzini, E. (2002) *On mixtures of distribution of Markov chains* in Stochastic Processes and their Applications, **100**, 147-165

[28] Frydman, H. (2005) *Estimation in the mixture of Markov chains moving with different speeds* in Journal of the American Statistical Association, **100**, 471, 1046-1053

[29] Gantmacher, F.R. (1959) *The theory of matrices*, Chelsea, New York

[30] Gerencsér, L. (1989) *On a class of mixing processes* in Stochastics, **26**, 165-191

[31] Gerencsér, L. (1991) *Almost sure exponential stability of random linear differential equations* in Stochastics and Stochastics Reports, Vol.36, 91-107

[32] Gerencsér, L. and Molnár-Sáska, G. (2004) *Change detection of hidden Markov models* in Proceedings of the 43th IEEE Conference on Decision and Control, 1754-1758

[33] Gerencsér, L. and Molnár-Sáska, G. (2007) *Identification of hidden Markov models - uniform LLN-s*, in Modeling, Estimation and Control, Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday, Lecture Notes in Control and Information Sciences, Springer, **364**, 135-149

[34] Gerencsér, L., Michaletzky G. and Molnár-Sáska, G. (2007) *An improved bound for the exponential stability of predictive filters of hidden Markov models* in *Communications in Information and Systems*, **7**, 2, 133-152

[35] Gilbert, E.J. (1959) *On the identifiability problem for functions of finite Markov chains* in *The Annals of Mathematical Statistics*, **30**, 688-697

[36] Heller, A. (1965) *On stochastic processes derived from Markov chains* in *The Annals of Mathematical Statistics*, **36**, 1286-1291

[37] Hinkley, D.V. (1971) *Inference about the change-point from cumulative sum tests* in *Biometrika*, **58**, 3, 509-523

[38] Khudanpur, S. and Narayan, P. (2002) *Order estimation for a special class of hidden Markov sources and binary renewal processes* in *IEEE Transactions on information Theory*, **48**, 1704-1713

[39] LeGland, F. and Mevel, L. (2000) *Exponential forgetting and geometric ergodicity in hidden markov models* in *Mathematics of Control Signals and Systems*, **13**, 63-93

[40] Leroux, B.G. (1992) *Maximum likelihood estimation for hidden Markov models* in *Stochastic Processes and their Applications*, Vol. 40, 127-143

[41] Lorden, G. (1971) *Procedures for reacting to a change in distribution* in *The Annals Mathematical Statistics*, **42**, 6, 1897-1908

[42] Meyn, S.P. and Tweedie, R.L. (1993) *Markov chains and stochastic stability*, Springer-Verlag, London

[43] Page, E.S. (1954) *Continuous inspection schemes* in *Biometrika*, **41**, 1/2, 100-115

[44] Page, E.S. (1955) *A test for a change in a parameter occurring at an unknown point* in *Biometrika*, **42**, 3-4, 523-527

[45] Page, E.S. (1957) *On problem in which a change in a parameter occurs at an unknown point* in *Biometrika*, **44**, 248-252

[46] Panjer, H.H. and Willmot, G.E. (1992) *Insurance risk models*, Society of Actuaries

[47] Quintana, F.A. and Newton, M.A. (1998) *Assessing the order of dependence for partially exchangeable binary data* in *Journal of the American Statistical Association*, **93**, 194-202

[48] Sparre Andersen, E. (1957) *On the collective theory of risk in the case of contagion between the claims* in *Transactions of the XV-th International Congress of Actuaries*, 219-229

[49] Sirjaev, A.N. (1973) *Statistical sequential analysis: optimal stopping rules*, American Mathematical Society, Providence

[50] Takacs, L. (1962) *Introduction to the theory of queues*, Oxford University Press, New York

[51] Vidyasagar, M. *The complete realization problem for hidden Markov models: A survey and some new results* to appear in *Mathematics of Control, Signals and Systems*

[52] Willems, J.C. (1976) *Realization of systems with internal passivity and symmetry constraints*, in *Journal of the Franklin Institute*, vol. 301, 605-621