NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

MASTER THESIS

A Modified EM Algorithm for Shrinkage Estimation in Multivariate Hidden Markov Models

---

Author:
Efstratios Manifavas

Supervisor:
Dr. Samis Trevezas
Assistant Professor
Department of Mathematics

Department of Mathematics

January 13, 2023

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Department of Mathematics

Master In Statistical And Operational Research

A Modified EM Algorithm for Shrinkage Estimation in Multivariate Hidden Markov Models

by Efstratios Manifavas

Hidden Markov models are used in a wide range of applications due to their construction that renders them mathematically tractable and allows for the use of efficient computational techniques. There are methods for the estimation of the model's parameters, such as the EM algorithm, but also for the estimation of the hidden states of the underlying Markov chain, such as the Viterbi algorithm.

In applications where the dimension of the data is comparable to the sample size, the sample covariance matrix is known to be ill-conditioned, which directly affects the maximisation step (M-step) of the EM algorithm, where its inverse is involved in the computations. This problem might be amplified if there are rarely visited states resulting in a small sample size for the estimation of the corresponding parameters. Therefore, the direct implementation of these methods can be proved to be troublesome, as many computational problems might occur in the estimation of the covariance matrix and its inverse, further affecting the estimation of the one-step transition probability matrix and the reconstruction of the hidden Markov chain.

In this paper, a modified version of the EM algorithm is studied, both theoretically and computationally, in order to obtain the shrinkage estimator of the covariance matrix during the maximisation step. This is achieved by maximising a penalised log-likelihood function, which is also used in the estimation step (E-step). A variant of this modified version, where the penalised log-likelihood function is only used in the maximisation step (M-step), is also studied computationally.

Περίληψη

Ένας Τροποποιημένος EM Αλγόριθμος για Εκτίμηση με Συρρίκνωση σε Πολυδιάστατα Κρυμμένα Μαρκοβιανά Μοντέλα

Τα κρυμμένα Μαρκοβιανά μοντέλα χρησιμοποιούνται σε ένα ευρύ πεδίο εφαρμογών, λόγω της κατασκευής τους που τα καθιστά μαθηματικώς διαχειρίσιμα και επιτρέπει τη χρήση αποτελεσματικών υπολογιστικών τεχνικών. Έχουν αναπτυχθεί μέθοδοι για την εκτίμηση των παραμέτρων του μοντέλου, όπως ο αλγόριθμος EM, αλλά και για την εύρεση των κρυμμένων καταστάσεων της Μαρκοβιανής αλυσίδας, όπως ο αλγόριθμος Viterbi.

Σε εφαρμογές στις οποίες η διάσταση των δεδομένων είναι συγκρίσιμη με το μέγεθος του δείγματος, είναι γνωστό πως ο δειγματικός πίνακας συνδιακύμανσης είναι αριθμητικά ασταθής, γεγονός που επηρεάζει άμεσα το βήμα μεγιστοποίησης (M-step) του αλγορίθμου EM, στο οποίο εμπλέκεται ο υπολογισμός του αντιστρόφου του. Το πρόβλημα αυτό μπορεί να ενταθεί λόγω ενδεχόμενης ύπαρξης καταστάσεων οι οποίες εμφανίζονται σπάνια, με αποτέλεσμα το μέγεθος δείγματος για την εκτίμηση των αντίστοιχων παραμέτρων να είναι μικρό. Επομένως, η άμεση χρήση αυτών των μεθόδων είναι πιθανό να οδηγήσει σε αριθμητικά προβλήματα, όσον αφορά στην εκτίμηση του πίνακα συνδιακύμανσης και του αντιστρόφου του, επηρεάζοντας επιπλέον την εκτίμηση του πίνακα πιθανοτήτων μετάβασης και την ανακατασκευή της κρυμμένης Μαρκοβιανής αλυσίδας.

Στη συγκεκριμένη εργασία μελετάται θεωρητικά και αλγοριθμικά μία τροποποίηση του αλγορίθμου EM, έτσι ώστε ο εκτιμήτης που προκύπτει για τον πίνακα συνδιακύμανσης, κατά το βήμα μεγιστοποίησης, να είναι αυτός που απορρέει από τη χρήση της μεθόδου συρρίκνωσης (shrinkage). Για τον σκοπό αυτό, στη συνάρτηση της λογαριθμικής πιθανοφάνειας ενσωματώνονται κάποιες ποινές, ώστε να κανονικοποιηθεί το αντίστοιχο πρόβλημα μεγιστοποίησης. Η συνάρτηση αυτή, χρησιμοποιείται και στο βήμα εκτίμησης (E-step). Επίσης, μελετάται αλγοριθμικά και μία παραλλαγή αυτής της μεθόδου, στην οποία η συνάρτηση με τις ποινές χρησιμοποιείται μόνο κατά το βήμα μεγιστοποίησης (M-step).

Acknowledgements

As this Master's program comes to an end, I would like to express my deepest gratitude towards my supervisor Assistant Professor Samis Trevezas for his valuable guidance during the writing of this thesis. Despite his workload, he always found the time to help me overcome any problems that I encountered. His contribution is invaluable to me. His work ethic and all the effort and time he puts into improving the quality of education of the students in the Department, are inspiring. The course "Statistics for Stochastic Processes" he taught during the second semester of the academic year 2020-2021 was the main inspiration for this thesis. I am also grateful for his support when I was in search of a Master's program.

I would like to thank my committee members Professor Apostolos Burnetas and Assistant Professor Fotios Siannis. I am particularly grateful to Professor Apostolos Burnetas for all our discussions during the course "Deterministic Models in Operational Research" and our discussion when I was searching for potential subjects for my thesis. I am also grateful to Assistant Professor Fotios Siannis for all the interesting discussions during the courses "Linear and Nonlinear Models" and "Simulation".

I would like to extend my sincere thanks to Professor Antonios Oikonomou for all the inspiring discussions during the courses "Nonlinear Programming" and "Stochastic Models in Operational Research" and his support when I was searching for a Master's program and, later, when I was in search of potential subjects of my thesis. A debt of gratitude is also owed to PhD Candidate Giannis Oikonomidis for his significant contribution to the Sector of Statistics and Operational Research of the Department, which in many cases is voluntary. I would also like to thank all my professors during the course of my studies both as an undergraduate and as a graduate student.

Many thanks to all my friends for their emotional support. Especially, I would like to express my gratitude towards Dimitris Mpavelas for discussing with me two of the problems I encountered while writing the code for this thesis. I am extremely grateful to Yannis Lamprakis for executing parts of the code on his computer, which saved me a lot of valuable time.

Lastly, I could not have undertaken this journey without my parents, Stamatia Katsarou and Christos Manifavas. Words cannot express my gratitude for their care, support and sacrifices all the years of my life, to help me grow as a person and go after my dreams.

Contents

Abstract 3

Acknowledgements 5

Introduction 8

1 Finite-State Markov Chains 12

1.1 Dynamical Systems and Stochastic Processes . . . 12

1.2 Discrete-time Markov Chains . . . 14

1.2.1 Definition and Characterisation . . . 14

1.2.2 Transient Distributions . . . 18

1.2.3 Classification of States . . . 20

1.2.3.1 Irreducibility . . . 20

1.2.3.2 Recurrence, Transience and Periodicity . . . 22

1.2.4 Stationary Distribution . . . 24

1.2.4.1 Empirical Estimation of the Stationary Distribution . . . 28

2 Hidden Markov Models and the EM Algorithm 29

2.1 Hidden Markov Models . . . 29

2.2 The EM Algorithm . . . 34

2.2.1 Description of the EM algorithm . . . 35

2.2.2 The EM Algorithm for HMMs . . . 39

2.2.2.1 E-Step . . . 42

2.2.2.2 M-Step . . . 46

2.3 The Data-Generating Model . . . 50

2.3.1 Definition of the Data-Generating Model . . . 50

2.3.2 EM Algorithm for the Data-Generating Model . . . 52

2.3.2.1 E-Step . . . 56

2.3.2.2 M-Step . . . 56

3 EM Algorithm and Shrinkage **59**

3.1 Penalised Likelihood For Independent And Identically Distributed Random Vectors . 60

3.2 Shrinkage under Regime Switching 71

3.3 Incorporating Shrinkage into the EM Algorithm 77

4 Applications **82**

4.1 Simulations 82

4.1.1 Simulation 1 84

4.1.2 Simulation 2 86

4.1.3 Simulation 3 88

4.1.4 Simulation 4 89

4.2 U.S. Industry Portfolio 90

A Auxiliary Material **1**

A.1 Probability And Measure Theory 1

A.2 Statistical Analysis 7

A.3 Multivariate Statistical Analysis 11

A.3.1 Basic Elements of Multivariate Statistical Analysis 11

A.3.2 Multivariate Normal Distribution 16

A.4 Time Series 16

A.5 Orders of Magnitude and Rate of Convergence 19

A.6 Linear Algebra 24

B Technical Result of Subsection 2.3.2 **27**

C Proofs and Technical Results of Chapter 3 **30**

C.1 Technical Result 1 30

C.2 Proofs and Technical Lemmas of Section 3.2 30

C.2.1 Technical Lemma 1 32

C.2.2 Technical Lemma 2 36

C.2.3 Proof of Lemma 3.9 39

C.2.4 Technical Lemma 3 58

C.2.5 Proof of Theorem 3.3 64

D The Viterbi Algorithm **66**

Introduction

A hidden Markov model (abbreviated HMM) is a statistical model in which it is assumed that an observed stochastic process is linked to another stochastic process, a (discrete-time) Markov chain, which is not observable. Simply put, if we have a sequence of observations at different points in time (i.e. a time series), we assume that the distribution of an observation at a specific time point is governed by the corresponding "hidden" state (regime) of the unobserved Markov chain.

HMMs were generally introduced in Baum and Petrie (1966). The authors defined a finite state Markov process $\{X_t\}$ and a stochastic finite state process $\{Y_t\}$ as a probabilistic function of $\{X_t\}$, where the value of Y_t depends (probabilistically) on the value of X_t. It was assumed that both the transition probability matrix of the Markov process $\{X_t\}$ and the emission matrix are unknown and it was proved that these quantities can be recovered from a sample of observations $\{Y_1, \ldots, Y_T\}$. Consistency and asymptotic normality of the maximum likelihood estimators was also proved under suitable conditions. Petrie (1969) weakened the conditions for consistency and provided sufficient conditions for identifiability of the model. Baum et al. (1970) and Baum (1972) developed forward–backward recursions for calculating the conditional probability of a state given an observation sequence from a general HMM. A computationally efficient iterative procedure for maximum likelihood estimation of the parameters of the model was also developed, using the above-mentioned recursions. This algorithm is often referred to as the Baum–Welch algorithm. The authors also established local convergence of the algorithm in the same papers, using results from Baum and Eagon (1967) and Baum and Sell (1968).

The notion of an underlying hidden mechanism (the hidden Markov chain) which controls the data generating process, makes HMMs versatile enough to be implemented in a plethora of complex real-world problems. The Markov property of the hidden process makes these models mathematically tractable and allows for the use of efficient computational techniques (see, e.g. Cappé et al. 2005, or Zucchini et al. 2016). Since their first introduction, HMMs have been studied broadly (see, e.g., Ephraim and Merhav 2002). Many extensions and generalisations have been developed, such as higher-order HMMs, hierarchical HMMs, continuous-time HMMs, non-homogeneous HMMs and layered HMMs to name a few (see, e.g., Cappé et al. 2005; Mor et al. 2021).

HMMs were initially used in signal-processing applications, especially in the context of automatic speech recognition (Levinson et al., 1983; Rabiner, 1989; Juang and Rabiner, 1991; Jelinek, 1998). Apart from their use in signal-processing, HMMs and their variations have been applied in a wide range of fields, some of which are:

- pattern recognition
  - face (Samaria, 1995; Nefian and Hayes, 1998; Alhadi et al., 2005)
  - gesture (Rigoll et al., 1997; Chen et al., 2003)
  - handwriting (Gilloux, 1994; Bunke et al., 1995; Hu et al., 1996)
  - signature (Yang et al., 1995; Dolfing et al., 2002; Daramola and Ibiyemi, 2010)
- bioinformatics
  - sequence analysis (Eddy et al., 1995; Eddy, 1998; Lunter, 2007)
  - modeling of proteins (Karplus et al., 1999; Krogh et al., 1994, 2001)
  - identification of coding regions in genomes (Lukashin and Borodovsky, 1998; Antonov and Borodovsky, 2010)
- environment
  - earthquakes (Wang and Bebbington, 2011; Avesani et al., 2012; Chambers et al., 2014)
  - rainfall (Sansom, 1998; Betrò et al., 2008; Stoner and Economou, 2020)
  - modeling wind times series and short-horizon wind forecasting (Hocaoğlu et al., 2010; Jafarzadeh et al., 2010; Barber et al., 2010; Ailliot et al., 2015)
- finance
  - portfolio analysis and management (e.g. asset allocation) (Elliott and Van der Hoek, 1997; Erlwein et al., 2011)
  - modelling financial time series such as:
    - financial returns (Rydén et al., 1998; Mary R. Hardy A.S.A., 2001)
    - interest rates (Smith, 2002; Wilson and Elliott, 2014)
    - exchange rates (Engel, 1994; Caporale and Spagnolo, 2004; Chen, 2006; Ali et al., 2014)

Recently HMMs have also been used in astronomy, in the search for gravitational waves (Suvorova et al., 2016, 2017; Sun and Melatos, 2019; Middleton et al., 2020) and for pulsar glitch detection (Melatos et al., 2020).

A very popular and extensively studied method to estimate the parameters of HMMs, is the expectation-maximisation (EM) algorithm (Dempster et al., 1977) which is presented in Section 2.2. Besides parameter estimation, it is also possible to estimate the most probable sequence of the hidden states, given the observed data, using the Viterbi algorithm (Viterbi, 1967), which is a dynamic programming algorithm, specially designed for this purpose (see Appendix D).

In this paper the modified EM algorithm proposed by Fiecas et al. (2017) is presented. These authors implement an HMM to analyse high-dimensional time-series data from finance and specifically for portfolio analysis. A usual assumption when modeling such data (e.g. the vector of returns of all assets in a stock portfolio) is that they are independent random vectors from the multivariate Normal distribution with covariance matrix Σ and volatility matrix $\Sigma^{1/2}$. Such models fail to describe potential changes in the market environment, such as a move to a more volatile state, which leads to a change of the covariance matrix. HMMs are able to model these changes, as the hidden states can represent the states of the market (see the introductions in Tadjuidje Kamgaing 2013 and Fiecas et al. 2017). They are also able to describe some particular features of this kind of data (also called *stylised facts*, see Rydén et al. 1998), such as departures from the normality assumption and dependence between the observations, due to the dependence structure of the underlying Markov chain.

A crucial part of many multivariate statistical analysis applications is the estimation of the covariance matrix and/or its inverse (see, e.g., Johnstone 2001, Ledoit and Wolf 2004, Bickel and Levina 2008 and Cai et al. 2011). However, when the data are of high dimension, as it is usually the case in many real-world problems, such as the ones mentioned above, this task can be challenging. The natural estimator, the sample covariance matrix, is known to perform poorly in high-dimensional situations (see, e.g., Ledoit and Wolf 2004, Bickel and Levina 2008, or Fan et al. 2008). Especially when the dimension of the data is comparable to the sample size, the sample covariance matrix may be numerically ill-conditioned or even not invertible at all. When regime switching models, such as HMMs, are implemented, this problem can be amplified by the fact that there might be some rarely visited states, which results in a very small corresponding sample size to estimate the covariance matrices. Any estimation technique involving the inverse of the covariance matrix can also be affected by the instability of the estimator. For instance, when the EM algorithm is utilised to estimate the model parameters of an HMM, if it is assumed that the data come from the multivariate Normal distribution, the inverses of the covariance matrices are involved in the calculations. This can have an adverse effect not only on the estimates of the parameters of the state-dependent distributions, but also on the estimation of the transition probability matrix and the reconstruction of the hidden Markov chain (see the Introduction in Fiecas et al. 2017).

To combat this problem Fiecas et al. (2017) incorporate the method of shrinkage into the EM algorithm, in order to obtain stable estimates of the covariance matrix during the M-step. This in an extension of the method of Sancetta (2008), who proposed a shrinkage estimator of the covariance matrix of data that exhibit time-series dependence. Sancetta (2008) also extended the work of Ledoit and Wolf (2004), who introduced a shrinkage estimator of the covariance matrix of independent and identically distributed random vectors.

This paper is organised as follows. In Chapter 1 a brief overview of the Markov chain theory is presented. It mostly serves as an introduction to some basic notions, that play an important role in the following chapters. Chapter 2 is dedicated to HMMs and the EM algorithm. In Section 2.1 hidden Markov models are defined and discussed quite extensively. In Section 2.2, the EM algorithm and its general application to HMMs is presented in detail. In Section 2.3 the model of Fiecas et al. (2017) is introduced and the EM algorithm is applied to this specific model. Chapter 3 contains the necessary ingredients to derive the modified EM algorithm of Fiecas et al. (2017). In Section 3.1 the work of Ledoit and Wolf (2004) is presented. It is also shown how their estimator can be obtained using the regularisation method of Yuan and Huang (2009) for the likelihood of independent and identically distributed random variables. In Section 3.2 the framework of Section 3.1 is extended to models with regime switching, where it is assumed that the states are known. In Section 3.3 the modified version of the EM algorithm is presented. Finally, in Chapter 4 we report the results of the application of the method of Fiecas et al. (2017) to simulated and real data. A variant of this method, where the penalised likelihood is only used during the M-step of the algorithm, is also applied to the same data sets. The results are also reported in Chapter 4. This method has also been used by Chen et al. (2014), to estimate multivariate Gaussian parameters in models with missing data. Some additional content that complements the main body of this work can be found in the Appendices.

Chapter 1

Finite-State Markov Chains

In this chapter we briefly describe dynamical systems and stochastic processes. The main body of this chapter consists of a presentation of the basic theory of finite-state discrete-time Markov chains, which will be a vital part of the content of the following chapters. We mainly follow the presentation of Kulkarni (2011, 2017) and Fakinos (2012).

1.1 Dynamical Systems and Stochastic Processes

A dynamical system is a system that evolves over time. At any given point in time, the system is in a state, which is represented by an element of an appropriately chosen set, called the *state space*. When the function that describes the evolution of the system is deterministic, the system is characterised as a deterministic dynamical system, or, simply, a deterministic system. On the other hand, when the evolution of the system is affected by random factors, the system is characterised as a stochastic (or random) dynamical system or, simply, a stochastic (or random) system.

Stochastic processes are mathematical models used to analyse the evolution of stochastic systems. If the system is observed at a set of discrete times, for instance at the beginning of every hour or every day, we have a *discrete-time* stochastic process. Otherwise, if the system is observed continuously, then we have a *continuous-time* stochastic process. (see Section 1.1 in Kulkarni 2017).

We now give a formal definition of a stochastic process, which can be found in Section 4.1 in Fakinos (2012).

Definition 1.1 (Stochastic Process)**.** *A family of random variables* $\{X(t) : t \in T\}$*, defined on the same probability space* $(\Omega, \mathcal{A}, P)$[1]*, is called a stochastic process. The parameter* t *takes its values in the parameter space* T *and is often referred to as the time parameter. If the parameter space* T *is a countable set (usually the set* $\mathbb{N}$*), the process is referred to as a discrete-time stochastic process and is usually denoted by* $\{X_n : n \in \mathbb{N}\}$*. On the other hand, if the parameter space* T *is an uncountable set, the process is referred to as a continuous-time*

[1] see Definition A.9 in Appendix A

stochastic process and is denoted by $\{X(t) : t \in T\}$. *Each possible value of the random variables* $X(t)$ *is called a state of the process. The state space* S *is referred to as discrete if it is a countable set (finite or countably infinite) and as continuous if it is an uncountable set. Hence, there are four categories of stochastic processes, discrete/continuous-time stochastic processes with a discrete/continuous state space.*

Every stochastic process $\{X(t) : t \in T\}$, defined on a probability space $(\Omega, \mathcal{A}, P)$, is a function

$$X : T \times \Omega \ni (t, \omega) \to X(t, \omega) \in S.$$

For fixed values $t = t_0$, $\omega = \omega_0$, $X(t_0, \omega_0)$ corresponds to the state of the stochastic process at time t_0 for a particular outcome ω_0. For a fixed value $t = t_0$, the function

$$X(t_0, \cdot) : \Omega \ni \omega \to X(t_0, \omega) \in S,$$

is the random variable $X(t_0)$. For a fixed value $\omega = \omega_0$, the function

$$X(\cdot, \omega_0) : T \ni t \to X(t, \omega_0) \in S,$$

represents a specific evolution (or trajectory) of the stochastic process and it is called a *realisation* or a *sample path* (see e.g. Section 1.1 in Kulkarni 2017 or Section 4.1 in Fakinos 2012).

We now define a specific class of stochastic processes, which possess an important property, called *stationarity*, that will be needed in the following.

Definition 1.2 (Stationary Stochastic Process)**.** *A stochastic process* $\{X(t) : t \in T\}$, *is characterised as strongly or strictly stationary if, for all* $n \in \mathbb{N}$, $t_1, t_2, \ldots, t_n \in T$, $s > 0$, *the n-dimensional random variables*

$$[X(t_1), X(t_2), \ldots, X(t_n)], \ [X(t_1 + s), X(t_2 + s), \ldots, X(t_n + s)],$$

have the same joint distribution.

Alternatively, the stochastic processes $\{X(t) : t \in T\}$ and $\{X(t + s) : t \in T\}$ are *stochastically equivalent*, which means that any moment can be deemed as the beginning of time. A process with this property is said to be in a state of *statistical equilibrium* or *stationarity* (see Section 4.2 in Fakinos 2012).

Remark. *The term "stationarity" refers to the statistical properties of the stochastic process and not the process itself. The process evolves over time in a manner that its statistical properties remain constant (e.g., parameters such as its mean are independent of time).*

1.2 Discrete-time Markov Chains

In this section we introduce the basic concepts of the theory of Markov chains. Our focus is restricted to first-order finite-state discrete-time Markov chains which are necessary for a better understanding of the mathematical properties that are presented in the following chapters. There are many textbooks on the topic of Markov chains. Some recommendations for the interested reader are Ross (1995), Norris (1997), Grimmett and Stirzaker (2001), Kulkarni (2011, 2017), and Fakinos (2012).

1.2.1 Definition and Characterisation

Consider a discrete-time stochastic process $\{X_n : n \in \mathbb{N}\}$, used to model a stochastic system, with a countable state space S, for simplicity $S = \mathbb{N}$, and let us pick a fixed value $n \in \mathbb{N}$, which shall be called the *present*. The random variable X_n is called the *present state* of the system, $(X_0, X_1, \ldots, X_{n-1})$ is said to be the *past* of the system and $(X_{n+1}, X_{n+2}, \ldots)$ is said to be the *future* of the system. If $X_n = i$ and $X_{n+1} = j$, where $i, j \in S$, then it is said that the system has jumped, or made a transition from state i to state j from time n to time $(n+1)$, or at the $(n+1)$-th step (see Section 2.1 in Kulkarni 2017).

A stochastic system is said to possess the Markov (or Markovian) property if, given its present state, the future of the system is independent of its past. This property was named after the Russian mathematician Andrey Andreyevich Markov (1856-1922), who was the first to introduce this notion, generalising the concept of independence for a collection of random variables (see, e.g., Section 2.1 in Kulkarni 2011, or the introduction of Chapter 6 in Fakinos 2012). A stochastic process that models such a system is called a Markov process. Markov processes with a discrete state space (finite or countably infinite), are also referred to as *Markov chains* (see Fakinos 2012).

The dependence structure of a Markov chain is depicted in Figure 1.1 below, using a *directed graphical model*. Directed acyclic graphs, like this one, represent the joint probability of a set of random variables. The nodes (circles) in the graph correspond to the random variables, whereas the directed edges (arrows) represent the dependency relation between them. Any random variable is conditionally independent of all the other random variables with which it is not directly connected, given the values of its parent nodes[2]. Hence, the joint probability distribution can be expressed as a product of the conditional distribution of each node given its parents. Some recommendations for further reading on graphical models are Lauritzen (1996), Jordan (2004), Cowell et al. (2007), Koller and Friedman (2009) and Sucar (2015).

[2] In a directed acyclic graph, if two nodes, say X_i and X_j are connected by a directed edge (arrow) starting from X_i and leading to X_j ($X_i \to X_j$), then X_i is said to be a *parent* of X_j and X_j is called a *child* of X_i (see, e.g. Subsection 2.1.1 in Lauritzen 1996).

FIGURE 1.1: Probabilistic graphical model for the dependence structure of a Markov chain.

We now give a formal definition of discrete-time Markov chains.

Definition 1.3 (Discrete-Time Markov Chain). *A stochastic process* $\{X_n : n \in \mathbb{N}\}$ *with discrete state space* S*, is called a discrete-time Markov chain if*

$$\mathrm{P}(X_{n+1} = j \mid X_n = i, X_{n-1}, \ldots, X_0) = \mathrm{P}(X_{n+1} = j \mid X_n = i), \tag{1.1}$$

for all $n \in \mathbb{N}$ *and* $i, j \in S$*.*

Equivalently, the process $\{X_n : n \in \mathbb{N}\}$ is a Markov chain if and only if, for any fixed value $n \in \mathbb{N}$, the random variable X_{n+1} is conditionally independent of the random variables $(X_0, X_1, \ldots, X_{n-1})$, given the value of the random variable X_n (see Definition 1 of Section 6.1 in Fakinos 2012). Hence, given the present state of the system, X_n, the future $(X_{n+1}, X_{n+2}, \ldots)$ is independent of its past $(X_0, X_1, \ldots, X_{n-1})$, as has already been stated.

Remark. *For all* $n \in \mathbb{N}$*, the random variable* X_{n+1} *depends on all its previous states* $(X_0, X_1, \ldots, X_n)$*. However, the value of the random variable* X_n*, provides all the information needed to determine the probability of* X_{n+1} *taking a specific value. In other words, given the value of* X_n*, the values of the random variables* $(X_0, X_1, \ldots, X_{n-1})$ *are no longer informative for the distribution of* X_{n+1}*.*

In cases where there is an integer d*, such that for any* $n > d$

$$\mathrm{P}(X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathrm{P}(X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_{n-d} = i_{n-d}),$$

then the stochastic process is said to be a Markov chain of order d*.*

For $d = 0$*, the sequence of random variables* $\{X_n : n \in \mathbb{N}\}$*, is simply a case of a family of mutually independent random variables. For* $d = 1$*,* $\{X_n : n \in \mathbb{N}\}$ *is a "classic" Markov chain.*

Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$*, of order* $d > 1$*. It is easy to check that by defining a process* $\{Y_n : n \in \mathbb{N}\}$*, such that*

$$Y_n = (X_n, X_{n+1}, \ldots, X_{n+d-1}),$$

then $\{Y_n : n \in \mathbb{N}\}$ *is a Markov chain of order* 1 *(see Section 6 in Billingsley 1961).*

For any fixed value $n \in \mathbb{N}$ and any $i, j \in S$, the conditional probability $\mathrm{P}(X_{n+1} = j \mid X_n = i)$ is called the *one-step transition probability*, or simply the *transition probability*, from state i to state j,

from time n to time $(n+1)$, or at the $(n+1)$-th step (jump) of the chain and is usually denoted by $p_{ij}(n, n+1)$. Generally, the transition probabilities are functions of both the pair of states i, j and the time at which the transition happens. In case the transition probabilities are independent of time, the Markov chain is characterised as *time homogeneous* (see, e.g., the discussion following Definition 2.1 in Kulkarni 2011, or Definition 1 of Section 6.1 in Fakinos 2012). We give the following definition.

Definition 1.4 (Time-Homogeneous Discrete-Time Markov Chain)**.** *A discrete-time Markov chain* $\{X_n : n \in \mathbb{N}\}$ *with discrete state space* S *is said to be time homogeneous if*

$$p_{ij} = \mathrm{P}(X_{n+1} = j \mid X_n = i) = \mathrm{P}(X_1 = j \mid X_0 = i), \tag{1.2}$$

for all $n \in \mathbb{N}$ *and* $i, j \in S$.

In the following we focus on time-homogeneous discrete-time Markov chains, as it is the only class needed for our purpose. We will refer to them, simply, as Markov chains.

Let $\boldsymbol{P} = (p_{ij})_{i,j\in S}$ denote the matrix whose (i,j)-th entry contains the transition probability p_{ij}, for all $i, j \in S$. Such a matrix is called the *one-step transition probability matrix* or simply *transition probability matrix*. When the state space is finite, say $S = \{1, 2, \ldots, N\}$, the transition probability matrix can be displayed as

$$\boldsymbol{P} = (p_{ij})_{i,j\in S} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}$$

(see the discussion following Definition 2.2 in Kulkarni 2017). The transition probability matrix of a discrete-time Markov chain is always a square matrix. An important property of square matrices is given below.

Definition 1.5 (Stochastic Matrix)**.** *A square matrix* $\boldsymbol{P} = (p_{ij})_{i,j\in S}$ *is called stochastic if it satisfies the following conditions:*

(i) $p_{ij} \geq 0$, *for all* $i, j \in S$,

(ii) $\sum_{j\in S} p_{ij} = 1$, *for all* $i, j \in S$.

The next theorem shows the relevance of this definition (see, e.g. Theorem 2.1 in Kulkarni 2017).

Theorem 1.1. *The one-step transition probability matrix of a Markov chain is stochastic.*

Proof. Statement (i) is obvious, since every element of the transition probability matrix, p_{ij} is a (conditional) probability, thus it is non-negative. Statement (ii) is proved as follows:

$$\sum_{j\in S} p_{ij} = \sum_{j\in S} \mathrm{P}(\, X_{n+1} = j \mid X_n = i) = \mathrm{P}(\, X_{n+1} \in S \mid X_n = i) = 1,$$

since by definition X_{n+1} must take some value in S, for all $n \in \mathbb{N}$, regardless of the value of X_n. □

We can completely describe a Markov chain when we are able to specify the probability of any finite sample path $(i_0, i_1, \ldots, i_n)$. For this to be possible, the distribution of the initial random variable X_0 is needed (see Section 6.1 in Fakinos 2012). The distribution of X_0 cannot be determined by the transition probability matrix, as its elements are conditional probabilities (see the discussion following the proof of Theorem 2.1 in Kulkarni 2017). Let us denote the probability of the event $\{X_0 = i\}$ by $p_i^{(0)}$, that is

$$p_i^{(0)} = \mathrm{P}(X_0 = i) \qquad (i \in S)\,.$$

The vector $\boldsymbol{p}^{(0)} = \left(p_i^{(0)}\right)_{i\in S}$ containing all these probabilities is called the *initial distribution* of the Markov chain.

We now give the following theorem for the characterisation of a Markov chain (see Theorem 2.2 in Kulkarni 2017).

Theorem 1.2. *A Markov chain $\{X_n : n \in \mathbb{N}\}$ is completely described by its initial distribution $\boldsymbol{p}^{(0)} = \left(p_i^{(0)}\right)_{i\in S}$ and the transition probability matrix $\boldsymbol{P}$.*

Proof. For all $n \in \mathbb{N}$ and $i_0, i_1, \ldots, i_{n-1}, i_n \in S$, by the general multiplication rule for the joint probability of the sample path $(i_0, i_1, \ldots, i_{n-1}, i_n)$, we get

$$\mathrm{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) =$$

$$\mathrm{P}(X_0 = i_0) \cdot \prod_{k=1}^{n} \mathrm{P}(\, X_k = i_k \mid X_0 = i_0, \ldots, X_{k-1} = i_{k-1}) =$$

(Markov property)

$$\mathrm{P}(X_0 = i_0) \cdot \prod_{k=1}^{n} \mathrm{P}(\, X_k = i_k \mid X_{k-1} = i_{k-1}) =$$

(Time homogeneity)

$$p_{i_0}^{(0)} \cdot p_{i_0 i_1} \cdot p_{i_1 i_2} \cdot \ldots \cdot p_{i_{n-1} i_n}.$$

□

1.2.2 Transient Distributions

Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S, transition probability matrix $\boldsymbol{P}$ and initial distribution $\boldsymbol{p}^{(0)} = \left(p_i^{(0)}\right)_{i \in S}$. The probability of a transition from a state i to a state j $(i, j \in S)$ in m steps is called the *m-step transition probability* of the Markov chain from state i to state j. For all $n, m \in \mathbb{N}$ and $i, j \in S$ we use the notation:

$$p_{ij}^{(m)} = \mathrm{P}(\, X_{n+m} = j \mid X_n = i) = \mathrm{P}(\, X_m = j \mid X_0 = i)\,,$$

where in the last equality we have used the property of time homogeneity (see Section 2.3 in Kulkarni 2011).

We have the following two special cases:

(i) $m = 0$:

$$p_{ij}^{(0)} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \qquad (i, j \in S)\,, \tag{1.3}$$

(ii) $m = 1$:

$$p_{ij}^{(1)} = p_{ij} \qquad (i, j \in S)\,. \tag{1.4}$$

Analogous to the one-step transition probability matrix, we define the *m-step transition probability matrix* as $\boldsymbol{P}^{(m)} = \left(p_{ij}^{(m)}\right)_{i,j \in S}$, which is also stochastic. As in the case of the one-step transition probability matrix, when the state space is finite, say $S = \{1, 2, \ldots, N\}$, the m-step transition probability matrix can be displayed as

$$\boldsymbol{P}^{(m)} = \left(p_{ij}^{(m)}\right)_{i,j \in S} = \begin{pmatrix} p_{11}^{(m)} & p_{12}^{(m)} & \cdots & p_{1N}^{(m)} \\ p_{21}^{(m)} & p_{22}^{(m)} & \cdots & p_{2N}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1}^{(m)} & p_{N2}^{(m)} & \cdots & p_{NN}^{(m)} \end{pmatrix}.$$

By Equations (1.3) and (1.4), it follows that

$$\boldsymbol{P}^{(0)} = \boldsymbol{I}_N,$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and

$$\boldsymbol{P}^{(1)} = \boldsymbol{P},$$

respectively.

We now turn our attention towards a way of expressing the m-step transition probabilities $p_{ij}^{(m)}$ as functions of the one-step transition probabilities. This is possible thanks to the Markov property and the time homogeneity. Notice that the m-step transition probability $p_{ij}^{(m)}$ is equal to the summation of the probabilities of all the possible finite paths, that start from state i and lead to state j in m steps. More formally, by the law of total probability we get

$$\begin{aligned}
p_{ij}^{(m)} &= \mathrm{P}(X_m = j \mid X_0 = i) \\
&= \sum_{k \in S} \mathrm{P}(X_m = j, X_{m-1} = k \mid X_0 = i) \\
&= \sum_{k \in S} \mathrm{P}(X_{m-1} = k \mid X_0 = i)\, \mathrm{P}(X_m = j \mid X_{m-1} = k, X_0 = i) \qquad \text{(Markov property)} \\
&= \sum_{k \in S} \mathrm{P}(X_{m-1} = k \mid X_0 = i)\, \mathrm{P}(X_m = j \mid X_{m-1} = k) \qquad \text{(time homogeneity)} \\
&= \sum_{k \in S} p_{ik}^{(m-1)} p_{kj} \qquad (m \in \mathbb{N},\ i, j \in S)\,.
\end{aligned}$$

The expression above can also be written in matrix form as

$$\boldsymbol{P}^{(m)} = \boldsymbol{P}^{(m-1)} \boldsymbol{P} \qquad (m \in \mathbb{N})\,,$$

hence we have

$$\boldsymbol{P}^{(m)} = \boldsymbol{P}^{(m-1)} \cdot \boldsymbol{P} = \boldsymbol{P}^{(m-2)} \cdot \boldsymbol{P} \cdot \boldsymbol{P} = \ldots = \boldsymbol{P}^{(1)} \cdot \boldsymbol{P}^{m-1} = \boldsymbol{P}^{m} \qquad (m \in \mathbb{N})\,. \tag{1.5}$$

Equation (1.5) shows that the m-step transition probabilities are the elements of the m-th power of the one-step transition probability matrix $\boldsymbol{P}$. We can also deduce that for fixed values $m, n \in \mathbb{N}$ it holds that

$$\boldsymbol{P}^{(m+n)} = \boldsymbol{P}^{(m)} \cdot \boldsymbol{P}^{(n)},$$

or in scalar form

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} \qquad (i, j \in S)\,. \tag{1.6}$$

The Equations (1.6) are called the Chapman-Kolmogorov equations and express the intuitive idea that for a transition from state i to state j to take place in $(m + n)$ steps, the chain needs to jump from state i to an intermediate state k in m steps and then from state k to state j in the remaining n steps (see Section 6.2 in Fakinos 2012).

We are now able to define the *transient distribution*, that is the distribution of the random variable X_n for some fixed value $n \in \mathbb{N}$ (see Section 2.3 in Kulkarni 2011). Let us use the notation

$$\boldsymbol{p}^{(n)} = \left(p_j^{(n)}\right)_{j \in S},$$

where

$$p_j^{(n)} = \mathrm{P}(X_n = j) \qquad (n \in \mathbb{N},\ j \in S)\,.$$

For $n = 0$, we have the initial distribution, $\boldsymbol{p}^{(0)} = \left(p_j^{(0)}\right)_{j \in S}$. By the law of total probability we get

$$p_j^{(n)} = \mathrm{P}(X_n = j) = \sum_{i \in S} \mathrm{P}(X_0 = i)\,\mathrm{P}(\,X_n = j \mid X_0 = i) = \sum_{i \in S} p_i^{(0)} p_{ij}^{(n)} \qquad (n \in \mathbb{N},\ j \in S)\,.$$

Working in a similar fashion, we get the general case,

$$p_j^{(m+n)} = \sum_{i \in S} p_i^{(m)} p_{ij}^{(n)} \qquad (m, n \in \mathbb{N},\ j \in S)\,,$$

and a special case

$$p_j^{(n+1)} = \sum_{i \in S} p_i^{(n)} p_{ij} \qquad (n \in \mathbb{N},\ j \in S)\,.$$

The equations above can also be expressed as

$$\begin{aligned} \boldsymbol{p}^{(n)} &= \boldsymbol{p}^{(0)} \boldsymbol{P}^{(n)} \qquad (n \in \mathbb{N})\,, \\ \boldsymbol{p}^{(m+n)} &= \boldsymbol{p}^{(m)} \boldsymbol{P}^{(n)} = \boldsymbol{p}^{(m)} \boldsymbol{P}^{n} \qquad (m, n \in \mathbb{N})\,, \\ \boldsymbol{p}^{(n+1)} &= \boldsymbol{p}^{(n)} \boldsymbol{P} \qquad (n \in \mathbb{N})\,. \end{aligned} \tag{1.7}$$

1.2.3 Classification of States

In this subsection the notions of *communicating class*, *irreducibility*, *transience*, *recurrence*, *null recurrence*, *positive recurrence* and *periodicity* are introduced. These concepts are of paramount importance for the analysis of Markov chains.

1.2.3.1 Irreducibility

We begin with a series of definitions which will enable the classification of states, sets of states, or whole chains, according to their properties.

Definition 1.6 (Accessibility). *A state j is said to be accessible from a state i if and only if there is an $n \in \mathbb{N}$ such that $p_{ij}^{(n)} > 0$.*

The inequality $p_{ij}^{(n)} > 0$ implies that there is a sequence of states $(i = i_0, i_1, \ldots, i_{n-1}, i_n = j)$ such that $p_{i_k i_{k+1}} > 0$ for $k = 0, 1, \ldots, n-1$. Thus, an intuitive interpretation of this definition is that a state j is accessible from i, if there is a directed path $i = i_0 \to i_1 \to \ldots \to i_{n-1} \to i_n = j$, for a value $n \in \mathbb{N}$ (see the discussion following Definition 4.1 in Kulkarni 2017).

We write $i \to j$ if state j is accessible from i. Since $p_{ii}^{(0)} = 1$, it follows that every state is accessible from itself, that is $i \to i$.

Definition 1.7 (Communication)**.** *States i and j are said to communicate if and only if j is accessible from i and i is accessible from j ($i \to j$ and $j \to i$).*

If i and j communicate, we write $i \leftrightarrow j$. The relation of communication possesses the following properties:

(i) $i \leftrightarrow i$ (reflexivity),

(ii) $i \leftrightarrow j \Leftrightarrow j \leftrightarrow i$ (symmetry),

(iii) $i \leftrightarrow j$, $j \leftrightarrow k \Rightarrow i \leftrightarrow k$ (transitivity).

Proof. Properties (i) and (ii) derive directly from the definition of communication. To prove the property of transitivity (iii), note that:

$$i \leftrightarrow j \Rightarrow i \to j, \quad j \leftrightarrow k \Rightarrow j \to k, \tag{1.8}$$

and

$$j \leftrightarrow i \Rightarrow j \to i, \quad k \leftrightarrow j \Rightarrow k \to j. \tag{1.9}$$

By (1.8) and (1.9) we get that $i \to k$ and $k \to i$, respectively. Hence it is clear that $i \leftrightarrow k$. □

Consequently, communication defines an *equivalence relation* in the state space S, which implies that S can be partitioned in *equivalence classes*. These classes consist of the states that communicate with each other, hence they are called *communicating classes*.

Definition 1.8 (Communicating Class)**.** *A set $C \subset S$ is called a communicating class if and only if*

(i) $i \in C$, $j \in C \Rightarrow i \leftrightarrow j$,

(ii) $i \in C$, $i \leftrightarrow j \Rightarrow j \in C$.

Property (i) assures what was mentioned above, that all the states of a communicating class communicate with one another. By property (ii) it follows that C is a maximal set, which means that

there is no strict superset of C that can be a communicating class. It is possible for a state j, that does not belong to C, to be accessible from a state i inside C, but in this case, i cannot be accessible from j (otherwise j would belong to C). Accordingly, it is possible for a state i, that belongs to C, to be accessible from a state j outside C, but in this case, j cannot be accessible from i (see the discussion following Definition 4.3 in Kulkarni 2017).

Definition 1.9 (Closed Communicating Class)**.** *A communicating class C is called closed if and only if $p_{ij} = 0$, for all $i \in C$ and $j \notin C$. Otherwise, the class C is called non-closed.*

If a closed class C is a singleton, say $C = \{i\}$, then the state i is referred to as an *absorbing state* (see Definition 2 of Section 6.4 Fakinos 2012). We are now able to define *irreducibility*.

Definition 1.10 (Irreducibility)**.** *A Markov chain with state space S is called irreducible if and only if all its states communicate with each other, equivalently if and only if the whole state space S is a communicating class. A Markov chain that is not irreducible is called reducible.*

1.2.3.2 Recurrence, Transience and Periodicity

Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S and transition probability matrix $\boldsymbol{P}$. Assume that $X_0 = i$, for some $i \in S$. Let us denote by T_{ij} the number of transitions that are needed for the Markov chain to visit for the first time the state $j \in S$, given that $X_0 = i$, that is

$$T_{ij} = \min\{n \in \mathbb{N} : X_n = j \mid X_0 = i\}.$$

If such a transition is impossible, we set $T_{ij} = \infty$. The random variable T_{ij} is called the *first passage time* into state j starting from state i (see Section 6.3 in Fakinos 2012). The probability of the first passage time to state j starting from i being equal to n, is given by

$$f_{ij}^{(n)} = \mathrm{P}(T_{ij} = n) = \mathrm{P}(X_n = j, X_k \neq j,\ k = 1, 2, \ldots, n-1 \mid X_0 = i) \qquad (n \in \mathbb{N}).$$

The probability of eventually visiting state j, starting from state i is

$$f_{ij} = \mathrm{P}(T_{ij} < +\infty) = \sum_{n=1}^{+\infty} f_{ij}^{(n)} \leq 1.$$

In the special case where $j = i$, the random variable T_{ii} is called the *first return time* to state i. The probability of eventually returning to state i is

$$f_{ii} = \mathrm{P}(T_{ii} < +\infty) = \sum_{n=1}^{+\infty} f_{ii}^{(n)} \leq 1.$$

The *mean return time* to state i is $\mathrm{E}(T_{ii})$. For a state i we have the following two cases.

(i) It is certain that a return to state i will occur ($f_{ii} = 1$):

$$\mathrm{E}(T_{ii}) = \sum_{n=1}^{+\infty} n f_{ii}^{(n)} \leq +\infty.$$

(ii) It is not certain that a return to state i will occur ($f_{ii} < 1$ equivalently $\mathrm{P}(T_{ii} = +\infty) = 1 - f_{ii} > 0$):

$$\mathrm{E}(T_{ii}) = +\infty.$$

We are now able to give the following definitions.

Definition 1.11 (Recurrence and Transience)**.** *A state i is called:*

(i) *recurrent if $f_{ii} = 1$,*

(ii) *transient if $f_{ii} < 1$.*

Definition 1.12 (Positive and Null Recurrence)**.** *A recurrent state i is called:*

(i) *positive recurrent if* $\mathrm{E}(T_{ii}) < +\infty$,

(ii) *null recurrent if* $\mathrm{E}(T_{ii}) = +\infty$.

We now proceed to define the notion of *periodicity*.

Definition 1.13 (Period)**.** *The period of a positive recurrent state i is the number d_i such that*

$$d_i = gcd\left\{n \in \mathbb{N} : f_{ii}^{(n)} > 0\right\} = gcd\left\{n \in \mathbb{N} : p_{ii}^{(n)} > 0\right\}.$$

Definition 1.14 (Periodicity)**.** *A positive recurrent state i is called:*

(i) *aperiodic, if $d_i = 1$,*

(ii) *periodic with period d_i, if $d_i > 1$.*

A return to a periodic state i with period d_i is possible only at times that are integer multiples of d_i (see the discussion following Definition 4.10 in Kulkarni 2017, or the discussion following Definition 2 of Section 6.3 in Fakinos 2012). Consequently, for all $k = 1, 2, \ldots, d_i - 1$ we have

$$p_{ii}^{(nd_i+k)} = 0 \qquad (n \in \mathbb{N}).$$

If the one-step transition probability from a state i to itself is positive $(p_{ii} > 0)$, then state i is aperiodic.

We now give a series of theorems and corollaries that are useful for classifying communicating classes and Markov chains. The proofs are classical and a reference is given for the interested reader.

Theorem 1.3. *Every pair of states that communicate with each other are of the same type, that is, they are both positive recurrent or null recurrent or transient. Additionally, if they are positive recurrent, then they are both aperiodic or periodic with the same period.*

Proof. See the proofs of Theorems 4.5, 4.6 and 4.8 in Kulkarni (2017). □

Corollary 1.3.1. *All states inside a communicating class are of the same type. Especially, all the states of an irreducible Markov chain are of the same type (see Corollary 1 of Section 6.4 in Fakinos 2012).*

Theorem 1.4. *All states in a finite closed communicating class are positive recurrent.*

Proof. See the proof of Theorem 4.9 in Kulkarni (2017). □

Corollary 1.4.1. *Every finite Markov chain possesses at least one positive recurrent state. Additionally, if the chain is irreducible, then all its states are positive recurrent (see Corollary 2 of Section 6.4 in Fakinos 2012).*

Theorem 1.5. *All states in a non-closed communicating class are transient.*

Proof. See the proof of Theorem 4.10 in Kulkarni (2017). □

Remark. *A communicating class is called positive recurrent or null recurrent or transient (aperiodic or periodic with period d) if all the states in it are positive recurrent or null recurrent or transient (aperiodic or periodic with period d), respectively (see Definition 4.8 and Theorem 4.8 in Kulkarni 2017).*

An irreducible Markov chain is called positive recurrent or null recurrent or transient (aperiodic or periodic with period d) if all its states are positive recurrent or null recurrent or transient (aperiodic or periodic with period d), respectively (see Definition 4.9 in Kulkarni 2017).

An irreducible positive recurrent aperiodic Markov chain is also called ergodic (see Subsection 4.5.5 in Kulkarni 2017).

These statements show that in order to completely classify a communicating class or an irreducible Markov chain, it suffices to classify only one of its states.

1.2.4 Stationary Distribution

Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S, transition probability matrix $\boldsymbol{P}$ and initial distribution $\boldsymbol{p}^{(0)}$. We wish to examine the limiting behaviour of the chain, in other words, we

want to examine the properties of the distribution $\boldsymbol{p}^{(n)}$, after a long period of time (as n tends to infinity).

Definition 1.15 (Limiting Distribution)**.** *Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S, transition probability matrix $\boldsymbol{P}$ and initial distribution $\boldsymbol{p}^{(0)}$. If for all $i \in S$ there exists the limit*

$$\lim_{n\to+\infty} (p_i^n) = \lim_{n\to+\infty} [\mathrm{P}(X_n = i)] = p_i,$$

then the vector $\boldsymbol{p} = (p_i)_{i\in S}$ is called the limiting or steady-state distribution of the Markov chain $\{X_n : n \in \mathbb{N}\}$ (see Section 2.5 in Kulkarni 2011).

We have the following theorem for the limiting distribution of a Markov chain.

Theorem 1.6. *If a Markov chain $\{X_n : n \in \mathbb{N}\}$ has a limiting distribution $\boldsymbol{p}$, then it satisfies the following equations:*

$$p_i = \sum_{k\in S} p_k p_{ki} \qquad (i \in S)\,, \tag{1.10}$$

$$\sum_{i\in S} p_i = 1 \qquad (i \in S)\,. \tag{1.11}$$

Proof. See the proof of Theorem 2.5 in Kulkarni (2011). □

For a specific state i, Equation (1.10) is referred to as the *balance equation*, because it balances the probability of entering state i with the probability of transitioning from state i to any state of the state space (including i). Equation (1.11) is known as the *normalising equation* (see the discussion following the proof of Theorem 4.19 in Kulkarni 2017).

The equations given by (1.10) can also be expressed in matrix form as

$$\boldsymbol{p} = \boldsymbol{p}\boldsymbol{P}. \tag{1.12}$$

In case the limiting distribution of the Markov chain (if it exists) is chosen as its initial distribution, then the limiting distribution is called *stationary* and is usually denoted by $\boldsymbol{\pi} = (\pi_i)_{i\in S}$. By Equation (1.12) we have

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}. \tag{1.13}$$

It follows that

$$\boldsymbol{\pi}\boldsymbol{P}^2 = (\boldsymbol{\pi}\boldsymbol{P})\,\boldsymbol{P} = \boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi},$$

and by induction we get

$$\boldsymbol{\pi}\boldsymbol{P}^n = \boldsymbol{\pi} \qquad (n \in \mathbb{N})\,. \tag{1.14}$$

As was shown in Subsection 1.2.2, by Equation (1.7), the transient distribution $\left\{\boldsymbol{p}^{(n)} : n \in \mathbb{N}\right\}$ of a Markov chain $\{X_n : n \in \mathbb{N}\}$ is a function of both the initial distribution $\boldsymbol{p}^{(0)}$ and the n-step transition probability matrix $\boldsymbol{P}^{(n)}$, therefore for all $n \in \mathbb{N}$, $\boldsymbol{p}^{(n)}$ depends on the specific value of n. Choosing the stationary distribution as the initial, that is $\boldsymbol{\pi} = \boldsymbol{p}^{(0)}$, by (1.7) and (1.14), we get

$$\boldsymbol{p}^{(n)} = \boldsymbol{\pi} \qquad (n \in \mathbb{N}),$$

which is equivalent to

$$\pi_i = \mathrm{P}(X_n = i) \qquad (n \in \mathbb{N},\ i \in S).$$

As a result, the transient distribution $\left\{\boldsymbol{p}^{(n)} : n \in \mathbb{N}\right\}$ becomes constant (independent of time n) and the same holds for the evolution of the Markov chain (not the process itself, but the manner in which it evolves). This justifies the term *stationary*.

We now give a formal definition of the stationary distribution of a Markov chain.

Definition 1.16 (Stationary Distribution)**.** *Consider a Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S and transition probability matrix $\boldsymbol{P}$. The distribution $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ is said to be the stationary distribution of the chain $\{X_n : n \in \mathbb{N}\}$ if, for all $n \in \mathbb{N}$ and $i \in S$, it holds*

$$\mathrm{P}(X_0 = i) = \pi_i \Rightarrow \mathrm{P}(X_n = i) = \pi_i.$$

For the stationary distribution of a Markov chain we have the following theorem.

Theorem 1.7. *The distribution $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ is stationary if and only if it satisfies the balance and normalising equations*

$$\pi_i = \sum_{k \in S} \pi_k p_{ki} \qquad (i \in S), \tag{1.15}$$

$$\sum_{i \in S} \pi_i = 1 \qquad (i \in S). \tag{1.16}$$

Proof. See the proof of Theorem 2.6 in Kulkarni (2011). □

Remark. *Any vector that is a non-negative solution of Equation (1.15) can be transformed into a stationary distribution, as long as the sum of its elements is a finite number. This can be achieved by normalising the vector, i.e. by dividing all its elements by their sum. For infinite-state Markov chains the existence of such solutions is not certain, in which case there is no stationary distribution. There are also some cases where a Markov chain has multiple stationary distributions. In such cases there is an infinite number of stationary*

distributions, that correspond to the Markov chain, since every convex combination of them is also a stationary distribution (see the discussion following Definition 1 of Section 6.5 in Fakinos 2012).

Corollary 1.7.1. *A limiting distribution, when it exists, is also a stationary distribution.*

Proof. See the proof of Corollary 2.3 in Kulkarni (2011). □

If a Markov chain $\{X_n : n \in \mathbb{N}\}$ has a stationary distribution, then for all $k, n \in \mathbb{N}$ and $i_0, i_1, \ldots, i_n \in S$ it holds

$$\mathrm{P}(X_k = i_0, X_{k+1} = i_1, \ldots, X_{k+n} = i_n) =$$

(General multiplication rule and Markov property)

$$p_{i_0}^{(k)} p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-1} i_n} = \pi_{i_0} p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-1} i_n} =$$

(Time homogeneity)

$$\mathrm{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n)\,,$$

which means that the processes $\{X_n : n \in \mathbb{N}\}$ and $\{X_{k+n} : n \in \mathbb{N}\}$ are stochastically equivalent. It follows that the Markov chain $\{X_n : n \in \mathbb{N}\}$ becomes a stationary stochastic process (see Definition 1.2) when $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ (see Section 6.5 in Fakinos 2012).

The following theorem is very useful as it provides a necessary and sufficient condition for the positive recurrence of irreducible Markov chains (see Subsection 4.5.7 in Kulkarni 2017).

Theorem 1.8. *An irreducible Markov chain is positive recurrent if and only if it has a unique stationary distribution $\boldsymbol{\pi} = (\pi_i)_{i \in S}$.*

Proof. See the proof of Theorem 4.22 in Kulkarni (2017). □

It follows that to determine whether an irreducible Markov chain is positive recurrent, one can directly try to solve Equations (1.15) and (1.16). If there is a unique solution, then the positive recurrence of the chain is assured (see the discussion following Theorem 4.22 in Kulkarni 2017).

We also have a very useful result for finite-state irreducible Markov chains (see Theorem 2.8 in Kulkarni 2011).

Theorem 1.9. *A finite-state irreducible Markov chain has a unique stationary distribution.*

For finite-state irreducible Markov chains which are also aperiodic, we have the following theorem (see Theorem 2.10 in Kulkarni 2011).

Theorem 1.10. *A finite-state irreducible aperiodic Markov chain has a unique limiting distribution.*

The following result derives directly from Corollary 1.7.1 and Theorem 1.10.

Corollary 1.10.1. *A finite-state irreducible aperiodic Markov chain has a unique limiting distribution, which is also its stationary distribution.*

1.2.4.1 Empirical Estimation of the Stationary Distribution

Consider an irreducible finite-state Markov chain $\{X_n : n \in \mathbb{N}\}$ with state space S, transition probability matrix $\boldsymbol{P}$, initial distribution $\boldsymbol{p}^{(0)}$ and stationary distribution $\boldsymbol{\pi}$, where $\boldsymbol{P}$ and $\boldsymbol{\pi}$ are unknown parameters.

Suppose that we wish to estimate the stationary distribution $\boldsymbol{\pi} = (\pi)_{i \in S}$ using a set of observations of the random variables $(X_0, \ldots, X_N)$. We define the sample estimator of π_i as

$$\hat{\pi}_i = \frac{1}{N+1} \sum_{n=0}^{N} \mathbf{1}_{\{X_n = i\}} \qquad (n \in \{0, \ldots, N\}, \, i \in S), \tag{1.17}$$

where

$$\mathbf{1}_{\{X_n = i\}} = \begin{cases} 1, & \text{if } X_n = i \\ 0, & \text{if } X_n \neq i \end{cases} \qquad (n \in \{0, \ldots, N\}, \, i \in S).$$

As $\sum_{n=0}^{N} \mathbf{1}_{\{X_n = i\}}$ is the total number of visits in state i in the first $(N+1)$ steps of the chain, $\hat{\pi}_i$ is the ratio of visits in state i (in the first $(N+1)$ steps of the chain).

If a Markov chain $\{X_n : n \in \mathbb{N}\}$ is stationary, for all $n \in \{0, \ldots, N\}$ and $i \in S$, we have that

$$\mathrm{E}(\mathbf{1}_{\{X_n = i\}}) = \mathrm{P}(X_n = i) = \pi_i,$$

and by the linearity of the expectation, it follows that

$$\mathrm{E}(\hat{\pi}_i) = \mathrm{E}\left(\frac{1}{N+1} \sum_{n=0}^{N} \mathbf{1}_{\{X_n = i\}} \right) = \frac{1}{N+1} \sum_{n=0}^{N} \mathrm{E}(\mathbf{1}_{\{X_n = i\}}) = \frac{1}{N+1} \sum_{n=0}^{N} \pi_i = \pi_i,$$

therefore $\hat{\pi}_i$ is an unbiased estimator of π_i, equivalently $\hat{\boldsymbol{\pi}}$ is an unbiased[3] estimator of $\boldsymbol{\pi}$. $\hat{\boldsymbol{\pi}}$ is also a consistent[4] estimator of $\boldsymbol{\pi}$ and under some appropriate conditions, it is $\sqrt{n}$-consistent[5] and asymptotically normal[6] (see Section 2 in Athreya and Majumdar 2003).

[3] See Definition A.31 in Appendix A
[4] See Definition A.33 in Appendix A
[5] See Definition A.34 in Appendix A
[6] See Definition A.35 in Appendix A

Chapter 2

Hidden Markov Models and the EM Algorithm

We start this chapter by introducing hidden Markov models in Section 2.1. We, then, proceed to present in detail the EM algorithm and its application to HMMs in Section 2.2. In Section 2.3, the model of Fiecas et al. (2017) is defined and the EM algorithm is applied to this model.

2.1 Hidden Markov Models

Hidden Markov models comprise a class of models which consist of two stochastic processes. An observable stochastic process $\{Y_t : t \geq 1\}$ and a Markov chain $\{X_t : t \geq 1\}$ (to keep the notation simple in the sequel, from now on we will assume that the initial moment is $t = 1$, instead of $t = 0$), which is "hidden" (not observable). The former process depends on the Markov chain in that X_t governs the distribution of the corresponding Y_t. For example, given X_t, the random variable Y_t may have a Poisson distribution whose parameter differs for each possible value of X_t. An important assumption of the simplest class of HMMs is that each observation Y_t depends on the Markov chain only through the random variable X_t.

Models, such as HMMs, where it is assumed that there are unobserved random variables, are called *latent variable models*. Another way to interpret the unobserved random quantities is as missing data, hence such models are also referred to as *missing data models*, or *models with incomplete data* (see Section 1.1 in Cappé et al. 2005). Thus, HMMs may also be useful in cases where there are indeed some data that are missing.

The dependence structure of HMMs is depicted in Figure 2.1, using a directed graphical model, similar, in spirit, to the one that was presented in Figure 1.1.

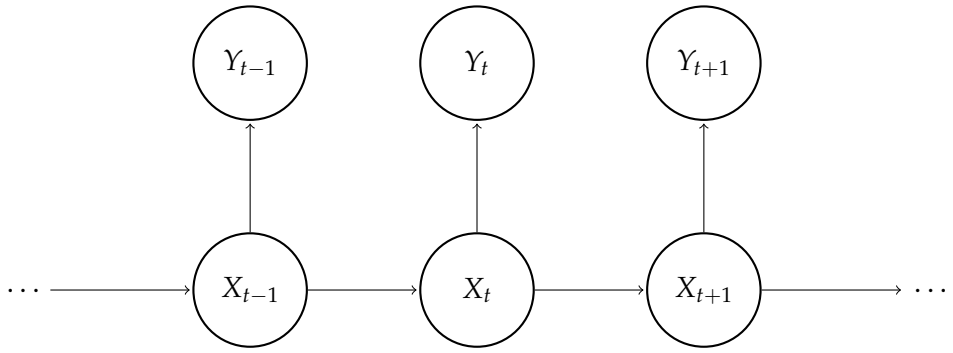

FIGURE 2.1: Probabilistic graphical model for the dependence structure of a hidden Markov model.

In Figure 2.1, $\{Y_t : t \geq 1\}$ is the sequence of observations and $\{X_t : t \geq 1\}$ is the underlying Markov chain. As it can be deduced by the graph, for any $t \geq 1$, given the current state X_t, the random variable Y_t is conditionally independent of all the previous observations $(Y_1, \ldots, Y_{t-1})$ and the past of the chain $(X_1, \ldots, X_{t-1})$. This is exactly the property that we mentioned above, that the distribution of Y_t can be determined by the corresponding state X_t only.

We now give a formal definition of HMMs.

Definition 2.1 (Hidden Markov Model)**.** *A hidden Markov model is a bivariate process* $\{(X_t, Y_t) : t \geq 1\}$ *such that*

(i) $\{X_t : t \geq 1\}$ *is a Markov chain with initial distribution* $\boldsymbol{p}^{(1)}$ *and transition probability matrix* $\boldsymbol{P}$*,*

(ii) *conditionally on the state process* $\{X_t : t \geq 1\}$*, the observations* $\{Y_t : t \geq 1\}$ *are independent, and for each* t *the conditional distribution of* Y_t *depends on* X_t *only.*

A more general definition of a hidden Markov model can be found in Cappé et al. (2005) (see Definition 2.2.1), but the one we gave above will suffice for our purposes.

From now on, for an HMM, we will denote by $\{Y_t : t \geq 1\}$ and $\{X_t : t \geq 1\}$ the observable stochastic process and the hidden Markov chain, respectively. We will also denote by $\boldsymbol{y} = (y_1, \ldots, y_T)$, $\boldsymbol{x} = (x_1, \ldots, x_T)$ a realisation of the stochastic processes $\{Y_t : t \geq 1\}$ and $\{X_t : t \geq 1\}$, until time T, respectively. The probability mass or density function of a random variable (depending on whether it is discrete or continuous) will be denoted by $f(\cdot)$.

Consider an HMM with state space $S_X = \{1, \ldots, K\}$, initial distribution $\boldsymbol{p}^{(1)}$ and transition probability matrix $\boldsymbol{P}$ for the hidden Markov chain. We consider two cases of interest for the observable stochastic process $\{Y_t : t \geq 1\}$.

- The random variable $(Y_t | X_t = k)$ follows a distribution with parameter ϕ_k (ϕ_k can either be unidimensional or multidimensional, depending on the distribution), for all $t \geq 1$ and $k \in$

$S_X = \{1, \ldots, K\}$. We denote by $\boldsymbol{\phi}$ the vector containing all ϕ_k's, that is, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$. Then, the parameter of the HMM is $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{\phi})$, in cases where the initial distribution $\boldsymbol{p}^{(1)}$ is assumed to be known.

- The state space of the stochastic process $\{Y_t : t \geq 1\}$ is a finite set, say $S_Y = \{1, \ldots, M\}$. Then we define the probabilities

$$r_{km} = \mathrm{P}(Y_t = m \mid X_t = k) \qquad (k \in S_X,\ m \in S_Y)\,, \tag{2.1}$$

also known as the *emission probabilities*. These probabilities can also be arranged in a matrix form, $\boldsymbol{R} = (r_{km})_{k \in S_X, m \in S_Y}$, known as the *emission matrix*, which is stochastic (see Definition 4.3 in Trevezas 2021). In accordance with the previous case, the parameter of the HMM is $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{R}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{R})$.

There are three basic problems that arise when an HMM of the form described above is implemented (see e.g., Section II in Rabiner 1989).

The evaluation problem: For a sequence of observations $\boldsymbol{y} = (y_1, \ldots, y_T)$ and a specific HMM $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{R})$, we need to compute the joint probability (or density) of the observed vector given the model, that is, the likelihood $f(\boldsymbol{y}; \boldsymbol{\theta})$ of the observations.

The decoding problem: For the sequence of observations $\boldsymbol{y} = (y_1, \ldots, y_T)$ and the model $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{R})$, we need to estimate the optimal corresponding sample path $\boldsymbol{x} = (x_1, \ldots, x_T)$ of the hidden Markov chain, that is, we need to determine the most likely state sequence that produced the observations.

The learning (or training) problem: The model parameters $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{R})$, need to be adjusted so as to maximise the probability $f(\boldsymbol{y}; \boldsymbol{\theta})$. For this purpose, a proportion of the data is used to "train" the HMM, i.e., to tune the parameters in order to describe, as best as possible, how the observations arise (hence the name *training problem*).

All statistical inference on the model parameters $\boldsymbol{\theta} = \left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right)$, or $\boldsymbol{\theta} = (\boldsymbol{P}, \boldsymbol{R})$, must be achieved using only the observations $\boldsymbol{y} = (y_1, \ldots, y_T)$, since the Markov chain $\{X_t : t \geq 1\}$ is not available to the observer.

In the following, we focus our attention on time-homogeneous HMMs, which means that the underlying Markov chain $\{X_t : t \geq 1\}$ is time-homogeneous (see Definition 1.4) and that the conditional law of Y_t given X_t is also independent of time t. We will also assume that the state space of the hidden chain is finite.

Suppose that $\{X_t : t \geq 1\}$ is a Markov chain with state space $S_X = \{1, \ldots, K\}$ and that the distribution Y_t, given $X_t = k$, also referred to as the *state-dependent distribution* (see Section 2.2 in Zucchini et al. 2016), is Poisson with parameter λ_k, $k \in S_X$. Then, for the conditional random variables $(Y_t | X_t = k)$ we have that

$$(Y_t | X_t = k) \sim Poisson\,(\lambda_k)\,, \qquad (k \in S_X)\,.$$

We could also assume that, instead of a Poisson distribution, the state-dependent distribution was Normal, therefore

$$(Y_t | X_t = k) \sim N\left(\mu_k, {\sigma_k}^2\right), \qquad (k \in S_X)\,,$$

or another family of distributions, say Gamma.

Let us denote by $f_k\,(y)$ the state-dependent probability mass or density function of Y_t (depending on whether the random variables $\{Y_t : t \geq 1\}$ are discrete or continuous, respectively), given that $X_t = k$, for all $t \in \{1, \ldots, T\}$ and $k \in S_X = \{1, \ldots, K\}$. That is,

$$f_k\,(y) = \mathrm{P}(\,Y_t = y \mid X_t = k)\,,$$

in the discrete case, or

$$f_k\,(y) = f_{Y_t|X_t}\,(y | X_t = k)\,,$$

in the continuous case. By the law of total probability, the marginal distribution of Y_t is:

$$f_{Y_t}\,(y) = \sum_{k=1}^{K} \mathrm{P}(X_t = k)\, f_k\,(y). \tag{2.2}$$

Equation (2.2) shows that the marginal distribution of Y_t is a mixture of the distributions of the conditional random variables $(Y_t | X_t = k)$, for all $t \in \{1, \ldots, T\}$ and $k \in S_X = \{1, \ldots, K\}$.

Consider a sample $\boldsymbol{y} = (y_1, \ldots, y_T)$. It becomes evident by the *evaluation* and *learning* problems above, that an essential part of the statistical inference in an HMM is the calculation of the likelihood, that is the joint probability of the observations given the model parameters $\mathrm{P}(\,\boldsymbol{y} \mid \boldsymbol{\theta})$. One way to achieve this, is through enumerating every possible sample path of the hidden Markov chain, with the same length as the observation sequence, T. Consider a specific sample path $\boldsymbol{x} = (x_1, \ldots, x_T)$. The probability of the observation sequence, given this state sequence and the parameters of the model is

$$L_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{\theta}) = f(\,\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) = f(\,y_1, \ldots, y_T \mid x_1, \ldots, x_T; \boldsymbol{\theta}) = \prod_{t=1}^{T} f_{x_t}(y_t; \boldsymbol{\theta}), \tag{2.3}$$

where we have used the fact that for any $t \geq 1$, given the random variable X_t, the random variable Y_t is conditionally independent of every Y_s and X_s, where $s \geq 1$ and $s \neq t$. As we saw in the proof of Theorem 1.2, the probability of the state sequence is

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta}) = p^{(1)}_{x_1} \prod_{t=2}^{T} p_{x_{t-1}x_t}. \tag{2.4}$$

By the general multiplication rule for the joint probability of $\boldsymbol{y}$ and $\boldsymbol{x}$, which is the probability that $\boldsymbol{y}$ and $\boldsymbol{x}$ occur simultaneously, we get

$$L_{\boldsymbol{y},\boldsymbol{x}}(\boldsymbol{\theta}) = f(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta}) = \left(p^{(1)}_{x_1} \prod_{t=2}^{T} p_{x_{t-1}x_t} \right) \left(\prod_{t=1}^{T} f_{x_t}(y_t;\boldsymbol{\theta}) \right). \tag{2.5}$$

It is easy to see that in order to obtain the probability of the observation sequence, we need to sum the joint probability given in Equation (2.5), over all possible sample paths, which gives

$$\begin{aligned} L_{\boldsymbol{y}}(\boldsymbol{\theta}) = f(\boldsymbol{y};\boldsymbol{\theta}) &= \sum_{\text{all possible } \boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta}) \\ &= \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{K} \left[\left(p^{(1)}_{x_1} \prod_{t=2}^{T} p_{x_{t-1}x_t} \right) \left(\prod_{t=1}^{T} f_{x_t}(y_t;\boldsymbol{\theta}) \right) \right]. \end{aligned} \tag{2.6}$$

As it has already been mentioned, the variables $\{X_t : t \geq 1\}$ of the hidden chain can be interpreted as missing data. Generally, in models with incomplete data, the joint probability mass or density function of the observation sequence $f(\boldsymbol{y};\boldsymbol{\theta})$, is referred to as the *observed-data likelihood*, or *observed likelihood* while the joint probability $f(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})$ of $\boldsymbol{y}$ and $\boldsymbol{x}$, is called the *complete-data likelihood* (see e.g., Cappé et al. 2005, or Zucchini et al. 2016).

In the context of HMMs the incomplete data are the actual observations $\boldsymbol{y} = (y_1, \ldots, y_T)$, while the complete data consist of the observations $\boldsymbol{y} = (y_1, \ldots, y_T)$, along with the hidden state sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$. We will denote by $L_{\boldsymbol{y}}(\boldsymbol{\theta})$ the observed-data likelihood and by $L_{\boldsymbol{y},\boldsymbol{x}}(\boldsymbol{\theta})$ the complete-data likelihood. We will also denote by $L_{\boldsymbol{y}|\boldsymbol{x}}(\theta)$ the conditional probability of observing $\boldsymbol{y} = (y_1, \ldots, y_T)$, given the sample path $\boldsymbol{x} = (x_1, \ldots, x_T)$, that is $L_{\boldsymbol{y}|\boldsymbol{x}}(\theta) = f(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta})$, which we shall call the *conditional likelihood*. Under the additional assumption that the parameters of the conditional distributions of the random variables $(Y_t|X_t)$ are independent of the parameters corresponding to the initial distribution $\boldsymbol{p}^{(1)}$ and the transition probability matrix $\boldsymbol{P}$ of the Markov chain, the complete-data likelihood is written as

$$L_{\boldsymbol{y},\boldsymbol{x}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right) = L_{\boldsymbol{x}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) L_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{\phi}). \tag{2.7}$$

Taking the logarithm of both sides of Equation (2.7) gives us the complete-data log-likelihood

$$\begin{aligned}\ell_{y,x}\left(\boldsymbol{p}^{(1)},\boldsymbol{P},\boldsymbol{\phi}\right) &= \log\left[L_{y,x}\left(\boldsymbol{p}^{(1)},\boldsymbol{P},\boldsymbol{\phi}\right)\right] = \log\left[L_x\left(\boldsymbol{p}^{(1)},\boldsymbol{P}\right)\right] + \log\left[L_{y|x}(\boldsymbol{\phi})\right] \\ &= \ell_x\left(\boldsymbol{p}^{(1)},\boldsymbol{P}\right) + \ell_{y|x}(\boldsymbol{\phi})\,. \end{aligned} \tag{2.8}$$

By Equation (2.6) it becomes clear that it is practically impossible to maximise the likelihood of the model $L_y(\boldsymbol{\theta})$, using this expression, even for a small number of states and observations (see Section III in Rabiner 1989). In Section 2.2 we will see how this objective can be achieved via the use of the EM algorithm.

2.2 The EM Algorithm

The EM algorithm was presented, in full generality, and named by Dempster et al. (1977). Their paper includes a description of the method, the necessary theory to prove that the likelihood increases with successive iterations and a convergence analysis of the algorithm. More specifically, they provided conditions under which the EM algorithm converges to a stationary point[1] of the likelihood and also examined the rate of convergence close to a stationary point. However, their proof of convergence of EM sequences contained errors (Theorems 2 and 3), as pointed out by Wu (1983), who established convergence of the EM and GEM[2] sequences and provided conditions under which the likelihood sequence converges to a global or local maximum, or a saddle point.

The EM algorithm had already been used, in special cases, many times in earlier works. For instance, Hartley (1958) presented three multinomial examples similar to the first example in Dempster et al. (1977), as the authors state in their paper. Many of the basic ideas of the EM algorithm had been presented, also in special cases, in Baum et al. (1970), Hartley and Hocking (1971), Orchard and Woodbury (1972), Sundberg (1974), Sundberg (1976) (see e.g., Dempster et al. 1977, Meng and Van Dyk 1997, or Wu 1983). However, the work of Dempster et al. (1977) remains very important, as it generalised the method and rendered it a very popular statistical tool, by presenting many examples of application.

The EM algorithm is an iterative procedure for calculating maximum-likelihood estimators for the parameters of incomplete-data models. As it has already been mentioned in the description of HMMs, these models consist of not only the observations and some unknown parameters, but also

[1] A point in which the derivative of a differentiable function of one variable, is equal to zero, is called *stationary*. In case of a differentiable function of more than one variable, a stationary point is a point where all the partial derivatives of the function equal zero (equivalently, the gradient of the function equals zero). The stationary points of a function are candidates for local extrema (minima or maxima).

[2] The GEM (Generalised Expectation Maximisation) algorithm is a more general method (than the EM) of iterative computation of maximum-likelihood estimators, also presented in Dempster et al. (1977).

latent variables. These latent variables can either be actual missing data, or they can be non-existent and added to the model, in order to simplify inference.

To obtain maximum-likelihood estimators for the parameters of incomplete-data models, all the unknown quantities need to be taken into account, which means that the likelihood function needs to be maximised with respect to both the unknown parameters and the latent variables (remember that in HMMs, the initial distribution and the transition probability matrix of the hidden Markov chain are parameters of the model).

There is no guarantee that the solution produced by the EM algorithm is the one that globally maximises the likelihood. The method might converge towards a local maximum, or a saddle point, as the surface of the likelihood of HMMs is generally multimodal. Therefore, one should use several initial values for the parameter (see the last paragraph of Section 1 in Bickel et al. 1998).

2.2.1 Description of the EM algorithm

We begin by defining two families of functions that are necessary to establish the monotonic increase of the sequence of the observed-data likelihood values. We follow Section 10.1 in Cappé et al. (2005) and Section 4.1 in Trevezas (2021).

Definition 2.2 (Intermediate Quantity of the EM Algorithm)**.** *The intermediate quantity of EM is the family* $\{\mathcal{Q}(\cdot;\boldsymbol{\theta}') : \boldsymbol{\theta}' \in \Theta\}$ *of real-valued functions on* Θ*, indexed by* $\boldsymbol{\theta}'$ *and defined by*

$$\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') = \int \log[f(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})]\, f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}')\, \lambda(dx) = \mathrm{E}_{\boldsymbol{\theta}'}(\ell_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}),$$

where λ *is a* σ*-finite measure*[3] *on the measurable space*[4] *on which the hidden Markov chain is defined.*

The following assumptions are needed to ensure that $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}')$ is well-defined for all values of the pair $(\boldsymbol{\theta},\boldsymbol{\theta}')$ (see Assumption 10.1.3 in Cappé et al. 2005):

Assumptions 1. (i) *The parameter set* Θ *is an open subset of* $\mathbb{R}^{d_{\boldsymbol{\theta}}}$ *(for some integer* $d_{\boldsymbol{\theta}}$*).*

(ii) *For all* $\boldsymbol{\theta} \in \Theta$*, the observed-data likelihood* $L_{\boldsymbol{y}}(\boldsymbol{\theta})$ *is positive and finite.*

(iii) *For all pairs* $(\boldsymbol{\theta},\boldsymbol{\theta}') \in \Theta \times \Theta$*,*

$$\int ||\nabla_{\boldsymbol{\theta}} \log[f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta})]||\, f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}')\, \lambda(dx) = \mathrm{E}_{\boldsymbol{\theta}'}(||\nabla_{\boldsymbol{\theta}} \log[f(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})]|| \mid \boldsymbol{Y} = \boldsymbol{y}) < +\infty.$$

[3] See Definition A.5 in Appendix A.
[4] See Definition A.3 in Appendix A.

Definition 2.3. $\{\mathcal{H}(\cdot;\boldsymbol{\theta}') : \boldsymbol{\theta}' \in \Theta\}$ *is a family of real-valued functions on* Θ*, indexed by* $\boldsymbol{\theta}'$ *and defined by*

$$\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') = -\int \log[f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta})]\, f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}')\, \lambda(d\boldsymbol{x}) = \mathrm{E}_{\boldsymbol{\theta}'}(-\log[f(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})] \mid \boldsymbol{Y} = \boldsymbol{y})$$

We use the notation $\ell_{\boldsymbol{y}}(\boldsymbol{\theta})$ for the log-likelihood of the observed data, that is

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log\left[L_{\boldsymbol{y}}(\boldsymbol{\theta})\right] = \log[f(\boldsymbol{y};\boldsymbol{\theta})]\,.$$

Lemma 2.1. *For the complete-data log-likelihood it holds*

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') + \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}')\,. \tag{2.9}$$

Proof. For all $\boldsymbol{x}$ such that $f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}) \neq 0$ the following holds

$$f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} \Leftrightarrow f(\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta})}.$$

By letting $\boldsymbol{x}$ become a random variable, we get

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})}. \tag{2.10}$$

Taking the logarithm of both sides of Equation (2.10), we get

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log[f(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})] - \log[f(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})]\,,$$

and by taking the conditional expectations given $\boldsymbol{y}$, we get

$$\begin{gathered}\mathrm{E}_{\boldsymbol{\theta}'}\left(\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y}\right) = \mathrm{E}_{\boldsymbol{\theta}'}(\log[f(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})] \mid \boldsymbol{Y} = \boldsymbol{y}) + \mathrm{E}_{\boldsymbol{\theta}'}(-\log[f(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})] \mid \boldsymbol{Y} = \boldsymbol{y}) \Leftrightarrow \\ \ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') + \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}')\,.\end{gathered}$$

□

We are now able to present a fundamental inequality which guarantees the increase of the observed-data likelihood with successive iterations of the algorithm.

Proposition 2.1. *Under* Assumptions 1*, for all pairs* $(\boldsymbol{\theta},\boldsymbol{\theta}') \in \Theta \times \Theta$*, it holds*

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) - \ell_{\boldsymbol{y}}(\boldsymbol{\theta}') \geq \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}';\boldsymbol{\theta}')\,, \tag{2.11}$$

where the inequality is strict unless $f(\cdot;\boldsymbol{\theta})$ and $f(\cdot;\boldsymbol{\theta}')$ are equal almost everywhere[5].

Proof. By Equation (2.9) we have that

$$\begin{aligned}\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) - \ell_{\boldsymbol{y}}(\boldsymbol{\theta}') &= \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') + \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}';\boldsymbol{\theta}') - \mathcal{H}(\boldsymbol{\theta}';\boldsymbol{\theta}') \Leftrightarrow \\ \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{H}(\boldsymbol{\theta}';\boldsymbol{\theta}') &= \ell_{\boldsymbol{y}}(\boldsymbol{\theta}) - \ell_{\boldsymbol{y}}(\boldsymbol{\theta}') - \left[\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}';\boldsymbol{\theta}')\right].\end{aligned}$$

Therefore, in order to show that

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) - \ell_{\boldsymbol{y}}(\boldsymbol{\theta}') \geq \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{Q}(\boldsymbol{\theta}';\boldsymbol{\theta}'),$$

it suffices to show that

$$\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{H}(\boldsymbol{\theta}';\boldsymbol{\theta}') \geq 0.$$

For all $(\boldsymbol{\theta},\boldsymbol{\theta}') \in \Theta \times \Theta$ we have

$$\begin{aligned}\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{H}(\boldsymbol{\theta}';\boldsymbol{\theta}') &= \mathrm{E}_{\boldsymbol{\theta}'}\left(-\log[f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta})] \mid \boldsymbol{Y}=\boldsymbol{y}\right) - \mathrm{E}_{\boldsymbol{\theta}'}\left(-\log\left[f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta}')\right] \mid \boldsymbol{Y}=\boldsymbol{y}\right) \\ &= \mathrm{E}_{\boldsymbol{\theta}'}\left(-\log\left[\frac{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta}')}\right] \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right). \qquad (2.12)\end{aligned}$$

Since $\log(x)$ is a concave function, $-\log(x)$ is a convex function. By Jensen's inequality we get

$$\mathrm{E}_{\boldsymbol{\theta}'}\left(-\log\left[\frac{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta}')}\right] \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) \geq -\log\left[\mathrm{E}_{\boldsymbol{\theta}'}\left(\frac{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta}')} \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right)\right], \qquad (2.13)$$

but

$$\mathrm{E}_{\boldsymbol{\theta}'}\left(\frac{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{X}\mid\boldsymbol{y};\boldsymbol{\theta}')} \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) = \int \frac{f(\boldsymbol{x}\mid\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{x}\mid\boldsymbol{y};\boldsymbol{\theta}')} f(\boldsymbol{x}\mid\boldsymbol{y};\boldsymbol{\theta}')\,d\boldsymbol{x} = \int f(\boldsymbol{x}\mid\boldsymbol{y};\boldsymbol{\theta}')\,d\boldsymbol{x} = 1 \qquad (2.14)$$

By relations (2.12), (2.13) and (2.14), we have

$$\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{H}(\boldsymbol{\theta}';\boldsymbol{\theta}') \geq 0$$

and the proof is complete. □

The basic concept of the EM algorithm is to take advantage of the fact that the complete-data log-likelihood may be much simpler to compute, therefore it is this function that will be maximised, instead of the observed-data log-likelihood $\ell_{\boldsymbol{y}}(\boldsymbol{\theta})$. Since the latent variables are not observable, the

[5] See Definition A.7 in Apppendix A

same follows for the complete-data log-likelihood, thus we use, instead, its conditional expectation given the observations $\boldsymbol{y}$, under the current value $\boldsymbol{\theta}'$ of the model's parameter $\boldsymbol{\theta}$, which is exactly the quantity $\mathcal{Q}(\cdot;\boldsymbol{\theta}')$. By choosing with each iteration a value $\boldsymbol{\theta} \in \Theta$, such that

$$\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') \geq \mathcal{Q}(\boldsymbol{\theta}';\boldsymbol{\theta}'), \tag{2.15}$$

where $\boldsymbol{\theta}'$ is the current estimation for $\boldsymbol{\theta}$, Inequality (2.11) ensures that the new value of the observed-data log-likelihood will not decrease.

The EM algorithm as presented by Dempster et al. (1977) produces a sequence $\left\{\boldsymbol{\theta}^{(m)} : m \geq 1\right\}$ of estimated values of the parameter $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^{(0)}$ denote an initially chosen value for $\boldsymbol{\theta}$. In the E-step of the first iteration, the quantity

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(0)}\right) = \mathrm{E}_{\boldsymbol{\theta}^{(0)}}\left(\ell_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}\right).$$

is calculated. In the M-step $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(0)}\right)$ is maximised with respect to $\boldsymbol{\theta}$ over the parameter space Θ, therefore a value $\boldsymbol{\theta}^{(1)}$ is chosen, such that for all $\boldsymbol{\theta} \in \Theta$

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(0)}\right).$$

In the second iteration, the E-step and the M-step are carried out as before, with the difference that $\boldsymbol{\theta}^{(0)}$ is replaced by the current estimation $\boldsymbol{\theta}^{(1)}$. Generally, suppose that after m iterations of the algorithm, the current estimation for $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^{(m)}$. Then the next iteration is broken into two steps, as follows:

E-step: Calculate the conditional expectation of the complete-data log-likelihood given the observations $\boldsymbol{y}$, under the current value $\boldsymbol{\theta}^{(m)}$,

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right) = \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}\right).$$

M-step: Choose $\boldsymbol{\theta}^{(m+1)}$ to be any value $\boldsymbol{\theta} \in \Theta$ that maximises $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$, that is, for all $\boldsymbol{\theta} \in \Theta$,

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(m+1)};\boldsymbol{\theta}^{(m)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right).$$

This procedure continues by alternating between the E-step and the M-step until a stopping criterion is met. Typically, the algorithm stops when the difference between two successive values of the observed-data likelihood, or the estimated parameters, is smaller than a predetermined nonnegative

value. By Inequality (2.11), for all $m \geq 0$ we get

$$\ell_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m+1)}\right) - \ell_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta}^{(m+1)}; \boldsymbol{\theta}^{(m)}\right) - \mathcal{Q}\left(\boldsymbol{\theta}^{(m)}; \boldsymbol{\theta}^{(m)}\right) \geq 0,$$

therefore, with each cycle of the algorithm, the value of the observed-data likelihood never decreases. Hence convergence must be obtained with a sequence of likelihood values that are bounded above. We will not get into details about the convergence of the algorithm. The interested reader is referred to Wu (1983), Chapter 3 of McLachlan and Krishnan (2007), or Section 10.5 of Cappé et al. (2005).

Remark. *In the GEM algorithm the only difference is that at each iteration, in the M-step, instead of choosing a value* $\boldsymbol{\theta}^{(m+1)}$ *that maximises* $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ *over all* $\boldsymbol{\theta} \in \Theta$*, the new estimate for* $\boldsymbol{\theta}$ *is chosen so as to increase the value of* $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ *over the current estimate* $\boldsymbol{\theta}^{(m)}$*. That is,* $\boldsymbol{\theta}^{(m+1)}$ *is chosen such that*

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(m+1)}; \boldsymbol{\theta}^{(m)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta}^{(m)}; \boldsymbol{\theta}^{(m)}\right).$$

The monotonicity of the observed-data log-likelihood still holds, which means that at each iteration of the GEM, we have

$$\ell_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m+1)}\right) \geq \ell_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m)}\right).$$

The GEM algorithm is very useful in situations where there is no solution of the M-step in closed form, hence global maximisation of $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ *is not feasible (see* Subsection 1.5.5 *in* McLachlan and Krishnan 2007*).*

2.2.2 The EM Algorithm for HMMs

In Section 2.1 we examined the likelihood functions of both the observations and the complete data of an HMM. Instead of expressing the observed-data likelihood in the form of Equation (2.6), we take a different approach. The indicator functions will be helpful in this direction.

Equation (2.4) can be written as

$$L_{\boldsymbol{x}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) = \left(\prod_{k=1}^{K} \left(p_k^{(1)}\right)^{\mathbf{1}_{\{x_1=k\}}}\right) \cdot \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{l=1}^{K} p_{kl}^{\mathbf{1}_{\{x_{t-1}=k, x_t=l\}}}\right). \tag{2.16}$$

Letting X be a random variable, we get

$$L_{X}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) = \left(\prod_{k=1}^{K} \left(p_k^{(1)}\right)^{\mathbf{1}_{\{X_1=k\}}}\right) \cdot \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{l=1}^{K} p_{kl}^{\mathbf{1}_{\{X_{t-1}=k, X_t=l\}}}\right). \tag{2.17}$$

Taking the logarithm of both sides of Equation (2.17) gives

$$\begin{aligned}\ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) &= \log\left[L_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right)\right] \\ &= \sum_{k=1}^{K} \mathbf{1}_{\{X_1=k\}} \log\left(p_k^{(1)}\right) + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbf{1}_{\{X_{t-1}=k, X_t=l\}} \log(p_{kl}).\end{aligned} \tag{2.18}$$

Similarly the conditional likelihood can be expressed as

$$L_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\phi}) = \prod_{t=1}^{T} [f_k(y_t; \boldsymbol{\phi})]^{\mathbf{1}_{\{X_t=k\}}} \tag{2.19}$$

and the conditional log-likelihood as

$$\ell_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\phi}) = \log\left[L_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\phi})\right] = \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \log[f_k(y_t; \boldsymbol{\phi})]. \tag{2.20}$$

Thus, by Equations (2.7), (2.16) and (2.19), the complete-data likelihood is

$$\begin{aligned}L_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) = &\left(\prod_{k=1}^{K}\left(p_k^{(1)}\right)^{\mathbf{1}_{\{X_1=k\}}}\right) \cdot \left(\prod_{t=2}^{T}\prod_{k=1}^{K}\prod_{l=1}^{K} p_{kl}^{\mathbf{1}_{\{X_{t-1}=k, X_t=l\}}}\right) \\ &\cdot \left(\prod_{t=1}^{T}\prod_{k=1}^{K} [f_k(y_t; \boldsymbol{\phi})]^{\mathbf{1}_{\{X_t=k\}}}\right).\end{aligned} \tag{2.21}$$

By Equations (2.8), (2.18), and (2.20), or by taking the logarithm of both sides of Equation (2.21), we get the complete-data log-likelihood as

$$\begin{aligned}\ell_{\boldsymbol{y},\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right) = &\sum_{k=1}^{K} \mathbf{1}_{\{X_1=k\}} \log\left(p_k^{(1)}\right) + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbf{1}_{\{X_{t-1}=k, X_t=l\}} \log(p_{kl}) \\ &+ \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} \log[f_k(y_t; \boldsymbol{\phi})].\end{aligned} \tag{2.22}$$

Then for any $m \geq 0$, $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ can be expressed as

$$\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right) &= \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_{\boldsymbol{y},\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= \sum_{k=1}^{K} \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\mathbf{1}_{\{X_1=k\}} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left[\left(p_k^{(1)}\right)^{(m)}\right] \\
&\quad + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\mathbf{1}_{\{X_{t-1}=k, X_t=l\}} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left(p_{kl}^{(m)}\right) \\
&\quad + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\mathbf{1}_{\{X_t=k\}} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left[f_k\left(y_t; \boldsymbol{\phi}^{(m)}\right)\right] \\
&= \sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\theta}^{(m)}}\left(X_1 = k \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left[\left(p_k^{(1)}\right)^{(m)}\right] \\
&\quad + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}^{(m)}}\left(X_{t-1} = k, X_t = l \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left(p_{kl}^{(m)}\right) \\
&\quad + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\theta}^{(m)}}\left(X_t = k \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \log\left[f_k\left(y_t; \boldsymbol{\phi}^{(m)}\right)\right].
\end{aligned} \tag{2.23}$$

Using a similar notation to the one presented in Rabiner (1989), we define the variables

$$\begin{aligned}
\gamma_k(t) &= \mathrm{E}_{\boldsymbol{\theta}}\left(\mathbf{1}_{\{X_t=k\}} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= \mathrm{P}_{\boldsymbol{\theta}}\left(X_t = k \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \qquad (t \geq 1,\ k \in S_X = \{1, \ldots, K\}),
\end{aligned} \tag{2.24}$$

and

$$\begin{aligned}
\xi_{kl}(t) &= \mathrm{E}_{\boldsymbol{\theta}}\left(\mathbf{1}_{\{X_t=k, X_{t+1}=l\}} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= \mathrm{P}_{\boldsymbol{\theta}}\left(X_t = k, X_{t+1} = l \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \qquad (t \geq 1,\ k, l \in S_X = \{1, \ldots, K\}).
\end{aligned} \tag{2.25}$$

For each $k \in S_X = \{1, \ldots, K\}$, $\gamma_k(t)$ is the conditional probability of the hidden Markov chain being in state k at time t, given the observations $\boldsymbol{y}$ and the parameter $\boldsymbol{\theta}$. For each $k, l \in S_X = \{1, \ldots, K\}$, $\xi_{kl}(t)$ is the conditional probability of the hidden Markov chain being in state k at time t and in state l at time $(t+1)$, given the observations $\boldsymbol{y}$ and the parameter $\boldsymbol{\theta}$.

By Equations (2.23) to (2.25), we get

$$\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right) &= \sum_{k=1}^{K} \gamma_k^{(m)}(1) \log\left[\left(p_k^{(1)}\right)^{(m)}\right] + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \xi_{kl}^{(m)}(t-1) \log\left(p_{kl}^{(m)}\right) \\
&\quad + \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k^{(m)}(t) \log\left[f_k\left(y_t; \boldsymbol{\phi}^{(m)}\right)\right].
\end{aligned} \tag{2.26}$$

2.2.2.1 E-Step

In the E-step of the EM algorithm we compute the variables $\gamma_k^{(m)}(t)$ and $\xi_{kl}^{(m)}(t)$. To do so, we implement the forward-backward algorithm (see e.g., Rabiner 1989, or McLachlan and Krishnan 2007).

We define the *forward variables*

$$\alpha_k(t) = f(y_1, \ldots, y_t, X_t = k; \boldsymbol{\theta}) \qquad (t \in \{1, \ldots, T\}, \; k \in \{1, \ldots, K\}) \tag{2.27}$$

which are computed as follows:

1. **Initialisation**
$$\alpha_k(1) = p_k^{(1)} f_k(y_1; \boldsymbol{\phi}) \qquad (k \in \{1, \ldots, K\}). \tag{2.28}$$

2. **Induction**
$$\alpha_k(t) = \left(\sum_{i=1}^{K} \alpha_i(t-1) \, p_{ik} \right) f_k(y_t; \boldsymbol{\phi}), \qquad (t \in \{2, \ldots, T\}, \; k \in \{1, \ldots, K\}). \tag{2.29}$$

3. **Termination**
$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k(T). \tag{2.30}$$

In the procedure above, the variables $\alpha_k(t)$ are computed inductively starting from $t = 1$ and moving forward, until $t = T$, hence the name forward variables.

We now proceed to prove Equations (2.28), (2.29) and (2.30).

Proof. To prove Equation (2.28), we use the definition of $\alpha_k(t)$ for $t = 1$ and the general multiplication rule.

$$\alpha_k(1) = f(y_1, X_1 = k; \boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(X_1 = k) \cdot f(y_1 \mid X_1 = k; \boldsymbol{\phi}) = p_k^{(1)} f_k(y_1; \boldsymbol{\phi}).$$

To prove Equation (2.29), we use the definition of $\alpha_k(t)$ and the law of total probability, utilising the random variable X_{t-1}. For all $t \in \{2, \ldots, T\}$ and $k \in \{1, \ldots, K\}$ we have that

$$\begin{aligned}
\alpha_k(t) &= f(y_1, \ldots, y_{t-1}, y_t, X_t = k; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{K} f(y_1, \ldots, y_{t-1}, y_t, X_t = k, X_{t-1} = i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{K} \Big[f(y_1, \ldots, y_{t-1}, X_{t-1} = i; \boldsymbol{\theta}) \, \mathrm{P}_{\boldsymbol{\theta}}(X_t = k \mid y_1, \ldots, y_{t-1}, X_{t-1} = i) \\
&\qquad \cdot f(y_t \mid y_1, \ldots, y_{t-1}, X_{t-1} = i, X_t = k; \boldsymbol{\theta}) \Big] \\
&= \sum_{i=1}^{K} f(y_1, \ldots, y_{t-1}, X_{t-1} = i; \boldsymbol{\theta}) \, \mathrm{P}_{\boldsymbol{\theta}}(X_t = k \mid X_{t-1} = i) \, f(y_t \mid X_t = k; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{K} \alpha_i(t-1) \, p_{ik} f_k(y_t; \boldsymbol{\phi}).
\end{aligned}$$

To prove Equation (2.30), we use the definition of $L_{\boldsymbol{y}}(\boldsymbol{\theta})$ and the law of total probability, utilising the random variable X_T.

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = f(y_1, \ldots, y_T; \boldsymbol{\theta}) = \sum_{k=1}^{K} f(y_1, \ldots, y_T, X_T = k; \boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k(T).$$

□

We also define the *backward variables*

$$\beta_k(t) = f(y_{t+1}, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) \qquad (t \in \{1, \ldots, T-1\}, \, k \in \{1, \ldots, K\}). \tag{2.31}$$

Since there is no observation for $t > T$, $\beta_k(t)$ cannot be defined as in Equation (2.31), for $t = T$. Thus, we set $\beta_k(T) = 1$, for all $k \in \{1, \ldots, K\}$ and the computation of the backward variables is carried out as follows:

1. **Initialisation**

$$\beta_k(T) = 1 \qquad (k \in \{1, \ldots, K\}). \tag{2.32}$$

2. **Induction**

$$\beta_k(t) = \sum_{l=1}^{K} p_{kl} f_l(y_{t+1}; \boldsymbol{\phi}) \, \beta_l(t+1) \qquad (t \in \{1, \ldots, T-1\}, \, k \in \{1, \ldots, K\}). \tag{2.33}$$

The computation scheme of the backward variables $\beta_k(t)$ justifies their name, since they are calculated inductively starting from $t = T$ and moving backwards, until $t = 1$.

Equation (2.33) is proved below.

Proof. By the definition of $\beta_k(t)$ and the law of total probability, utilising the random variable X_{t+1}, for all $t \in \{1, \ldots, T-1\}$ and $k \in \{1, \ldots, K\}$, we have

$$\begin{aligned}
\beta_k(t) = f(y_{t+1}, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) &= \sum_{l=1}^{K} f(y_{t+1}, \ldots, y_T, X_{t+1} = l \mid X_t = k; \boldsymbol{\theta}) = \\
&\sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k)\, f(y_{t+1}, \ldots, y_T \mid X_t = k, X_{t+1} = l; \boldsymbol{\theta}) = \\
&\sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k)\, f(y_{t+1}, y_{t+2}, \ldots, y_T \mid X_{t+1} = l; \boldsymbol{\theta}) = \\
\sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k)\, f(y_{t+1} \mid X_{t+1} = l; \boldsymbol{\theta})\, &f(y_{t+2}, \ldots, y_T \mid y_{t+1}, X_{t+1} = l; \boldsymbol{\theta}) = \\
\sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k)\, f(y_{t+1} \mid X_{t+1} = l; \boldsymbol{\phi})\, &f(y_{t+2}, \ldots, y_T \mid X_{t+1} = l; \boldsymbol{\theta}) = \\
&\sum_{l=1}^{K} p_{kl} f_l(y_{t+1}; \boldsymbol{\phi})\, \beta_l(t+1).
\end{aligned}$$

□

We also have the following equation, which relates the forward and backward variables to the observed-data likelihood.

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k(t)\, \beta_k(t) \qquad (t \in \{1, \ldots, T\})\,. \tag{2.34}$$

Proof. For all $t \in \{1, \ldots, T\}$, we have that

$$\begin{aligned}
\sum_{k=1}^{K} \alpha_k(t)\, \beta_k(t) &= \sum_{k=1}^{K} f(y_1, \ldots, y_t, X_t = k; \boldsymbol{\theta})\, f(y_{t+1}, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) \\
&= \sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_t = k)\, f(y_1, \ldots, y_t \mid X_t = k; \boldsymbol{\theta})\, f(y_{t+1}, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) \\
&= \sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_t = k)\, f(y_1, \ldots, y_t, y_{t+1}, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) \\
&= \sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_t = k)\, f(y_1, \ldots, y_T \mid X_t = k; \boldsymbol{\theta}) \\
&= f(y_1, \ldots, y_T; \boldsymbol{\theta}) = L_{\boldsymbol{y}}(\boldsymbol{\theta})\,,
\end{aligned}$$

where in the third equality we have used the fact that the random variables $(Y_1, \ldots, Y_t)$ and $(Y_{t+1}, \ldots, Y_T)$ are conditionally independent, given the state of the hidden Markov chain at time t, X_t. □

Notice that by Equations (2.32) and (2.34) we get Equation (2.30).

We are now able to express the variables $\gamma_k(t)$ and $\xi_{kl}(t)$ as functions of the forward and backward variables. We have the following relations:

$$\gamma_k(t) = \frac{\alpha_k(t)\,\beta_k(t)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)} \qquad (t \in \{1,\dots,T\},\ k \in \{1,\dots,K\}) \tag{2.35}$$

$$\xi_{kl}(t) = \frac{\alpha_k(t)\,p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\,\beta_l(t+1)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)} \qquad (t \in \{1,\dots,T-1\},\ k,l \in \{1,\dots,K\}) \tag{2.36}$$

Proof. For Equation (2.35), for all $t \in \{1,\dots,T\}$ and $k \in \{1,\dots,K\}$, we have

$$\begin{aligned}\gamma_k(t) = \mathrm{P}_{\boldsymbol{\theta}}(X_t = k \mid \boldsymbol{Y} = \boldsymbol{y}) &= \frac{f(X_t = k, y_1,\dots,y_t;\boldsymbol{\theta})\, f(y_{t+1},\dots,y_T \mid X_t = k, y_1,\dots,y_t;\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &= \frac{f(X_t = k, y_1,\dots,y_t;\boldsymbol{\theta})\, f(y_{t+1},\dots,y_T \mid X_t = k;\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} = \frac{\alpha_k(t)\,\beta_k(t)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)}.\end{aligned}$$

For Equation (2.36), for all $t \in \{1,\dots,T-1\}$ and $k,l \in \{1,\dots,K\}$, we have

$$\begin{aligned}\xi_{kl}(t) &= \mathrm{P}_{\boldsymbol{\theta}}(X_t = k, X_{t+1} = l \mid \boldsymbol{Y} = \boldsymbol{y}) \\ &= \frac{f(X_t = k, X_{t+1} = l, y_1,\dots,y_t, y_{t+1}, y_{t+2},\dots,y_T;\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &= \frac{f(X_t = k, y_1,\dots,y_t;\boldsymbol{\theta})\,\mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k, y_1,\dots,y_t)}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &\quad \cdot f(y_{t+1} \mid X_{t+1} = l, X_t = k, y_1,\dots,y_t;\boldsymbol{\theta}) \\ &\quad \cdot f(y_{t+2},\dots,y_T \mid X_{t+1} = l, X_t = k, y_1,\dots,y_t;\boldsymbol{\theta}) \\ &= \frac{f(X_t = k, y_1,\dots,y_t;\boldsymbol{\theta})\,\mathrm{P}_{\boldsymbol{\theta}}(X_{t+1} = l \mid X_t = k)}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &\quad \cdot f(y_{t+1} \mid X_{t+1} = l;\boldsymbol{\theta})\, f(y_{t+2},\dots,y_T \mid X_{t+1} = l;\boldsymbol{\theta}) \\ &= \frac{\alpha_k(t)\,p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\,\beta_l(t+1)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)}.\end{aligned}$$

□

We can also relate $\gamma_k(t)$ and $\xi_{kl}(t)$ as follows:

$$\gamma_k(t) = \sum_{l=1}^{K} \xi_{kl}(t) \qquad (t \in \{1,\dots,T\},\ k \in \{1,\dots,K\}), \tag{2.37}$$

$$\xi_{kl}(t) = \frac{\gamma_k(t)\, p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\, \beta_l(t+1)}{\beta_k(t)} \qquad (t \in \{1,\dots,T-1\},\ k,l \in \{1,\dots,K\}). \tag{2.38}$$

Proof. For Equation (2.37), we utilise the random variable X_{t+1} and the law of total probability. For all $t \in \{1,\dots,T\}$ and $k \in \{1,\dots,K\}$, we get

$$\gamma_k(t) = \mathrm{P}_{\boldsymbol{\theta}}(X_t = k \mid \boldsymbol{Y} = \boldsymbol{y}) = \sum_{l=1}^{K} \mathrm{P}_{\boldsymbol{\theta}}(X_t = k, X_{t+1} = l \mid \boldsymbol{y}) = \sum_{l=1}^{K} \xi_{kl}(t).$$

By Equations (2.35) and (2.36), for all $t \in \{1,\dots,T-1\}$ and $k,l \in \{1,\dots,K\}$, we get

$$\begin{aligned}
\xi_{kl}(t) &= \frac{\alpha_k(t)\, p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\, \beta_l(t+1)}{\sum_{i=1}^{K} \alpha_i(t)\, \beta_i(t)} \\
&= \frac{\alpha_k(t)\, \beta_k(t)}{\sum_{i=1}^{K} \alpha_i(t)\, \beta_i(t)} \cdot \frac{p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\, \beta_l(t+1)}{\beta_k(t)} \\
&= \gamma_k(t) \cdot \frac{p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\, \beta_l(t+1)}{\beta_k(t)}.
\end{aligned}$$

□

2.2.2.2 M-Step

In the M-step we maximise $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$ subject to:

$$\sum_{k=1}^{K} p_k^{(1)} = 1$$

and

$$\sum_{l=1}^{K} p_{kl} = 1 \qquad (k \in \{1,\dots,K\}).$$

By implementing the method of Lagrange multipliers, we get the Lagrangian function for the maximisation problem above, as:

$$\mathcal{L}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}, \lambda_1, \dots, \lambda_K, \lambda\right) = \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right) - \sum_{k=1}^{K} \left[\lambda_k \left(\sum_{l=1}^{K} p_{kl} - 1\right)\right] - \lambda\left(\sum_{k=1}^{K} p_k^{(1)} - 1\right) \tag{2.39}$$

and by Equation (2.26), we get

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}, \lambda_1, \ldots, \lambda_K, \lambda\right) = &\sum_{k=1}^{K} \gamma_k(1) \log\left(p_k^{(1)}\right) + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \xi_{kl}(t-1)\log(p_{kl}) \\ &+ \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t)\log[f_k(y_t;\boldsymbol{\phi})] - \sum_{k=1}^{K}\left[\lambda_k\left(\sum_{l=1}^{K} p_{kl} - 1\right)\right] \\ &- \lambda\left(\sum_{k=1}^{K} p_k^{(1)} - 1\right).\end{aligned} \tag{2.40}$$

To get the updated values for the initial distribution $\boldsymbol{p}^{(1)} = \left(p_k^{(1)}\right)_{k\in S_X}$ and the transition probabilities p_{kl} $(k, l \in S_X)$, we compute the partial derivative of the Lagrangian function with respect to $p_k^{(1)}$, for all $k \in S_X$ and p_{kl}, for all $k, l \in S_X$, respectively and set them equal to zero. We get the following relations:

$$\frac{\partial \mathcal{L}}{\partial p_k^{(1)}} = 0 \Leftrightarrow \frac{\gamma_k(1)}{p_k^{(1)}} - \lambda = 0 \Leftrightarrow p_k^{(1)} = \frac{\gamma_k(1)}{\lambda}. \tag{2.41}$$

$$\frac{\partial \mathcal{L}}{\partial p_{kl}} = 0 \Leftrightarrow \frac{\sum_{t=2}^{T} \xi_{kl}(t-1)}{p_{kl}} - \lambda_k = 0 \Leftrightarrow p_{kl} = \frac{\sum_{t=2}^{T} \xi_{kl}(t-1)}{\lambda_k} \Leftrightarrow p_{kl} = \frac{\sum_{t=1}^{T-1} \xi_{kl}(t)}{\lambda_k} \tag{2.42}$$

Utilising the first constraint and Equation (2.41), or, equivalently, by computing the partial derivative of the Lagrangian function with respect to λ and setting it equal to zero, for all $k \in S_X$, we get

$$\sum_{k=1}^{K} p_k^{(1)} = 1 \Leftrightarrow \sum_{k=1}^{K} \frac{\gamma_k(1)}{\lambda} = 1 \Leftrightarrow \lambda = \sum_{k=1}^{K} \gamma_k(1). \tag{2.43}$$

Utilising the second constraint and Equation (2.42), or, equivalently, by computing the partial derivative of the Lagrangian function with respect to λ_k and setting it equal to zero, for all $k \in S_X$, we get

$$\sum_{l=1}^{K} p_{kl} = 1 \Leftrightarrow \sum_{l=1}^{K}\left(\frac{\sum_{t=1}^{T-1} \xi_{kl}(t)}{\lambda_k}\right) = 1 \Leftrightarrow \lambda_k = \sum_{l=1}^{K}\left(\sum_{t=1}^{T-1} \xi_{kl}(t)\right) \Leftrightarrow \lambda_k = \sum_{t=1}^{T-1}\left(\sum_{l=1}^{K} \xi_{kl}(t)\right).$$

By (2.36) we get

$$\lambda_k = \sum_{t=1}^{T-1}\left(\sum_{l=1}^{K} \frac{\alpha_k(t)\, p_{kl} f_l(y_{t+1};\boldsymbol{\phi})\, \beta_l(t+1)}{\sum_{i=1}^{K} \alpha_i(t)\, \beta_i(t)}\right),$$

which, equivalently, is written as

$$\lambda_k = \sum_{t=1}^{T-1} \left(\frac{\alpha_k(t)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)} \sum_{l=1}^{K} p_{kl} f_l(y_{t+1}; \boldsymbol{\phi})\,\beta_l(t+1) \right).$$

It follows that

$$\lambda_k = \sum_{t=1}^{T-1} \left(\frac{\alpha_k(t)}{\sum_{i=1}^{K} \alpha_i(t)\,\beta_i(t)} \beta_k(t) \right),$$

and by Equation (2.35)

$$\lambda_k = \sum_{t=1}^{T-1} \gamma_k(t). \tag{2.44}$$

As a result, by Relations (2.41) and (2.43), we get the updated values for the initial probabilities as

$$\left(\widehat{p_k^{(1)}}\right)^{(m+1)} = \frac{\gamma_k^{(m)}(1)}{\sum_{i=1}^{K} \gamma_i^{(m)}(1)} \qquad (k \in S_X),$$

and by Relations (2.42) and (2.44), we get the updated values for the transition probabilities as

$$\hat{p}_{kl}^{(m+1)} = \frac{\sum_{t=1}^{T-1} \xi_{kl}^{(m)}(t)}{\sum_{t=1}^{T-1} \gamma_k^{(m)}(t)} \qquad (k, l \in S_X).$$

By computing the partial derivative of the Lagrangian function with respect to ϕ_k and setting it equal to zero, for all $k \in S_X$, we get the estimates $\hat{\phi}_k^{(m+1)}$ (see e.g., Section III.C in Rabiner 1989).

Remark. *As we can see by Equation (2.8) the complete-data log-likelihood consists of two summands, $\ell_X\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right)$ and $\ell_{\boldsymbol{y}|X}(\boldsymbol{\phi})$, thus $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ can be written as*

$$\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right) &= \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_{\boldsymbol{y},X}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}, \boldsymbol{\phi}\right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) \\
&= \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_X\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) + \ell_{\boldsymbol{y}|X}(\boldsymbol{\phi}) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) \\
&= \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_X\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) + \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_{\boldsymbol{y}|X}(\boldsymbol{\phi}) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) \\
&= \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_X\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) + \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left[\ell_{\boldsymbol{y}|X}(\boldsymbol{\phi}) \right].
\end{aligned}$$

Therefore, in order to maximise $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$, we can maximise $\mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\ell_X\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right)$ and $\mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left[\ell_{\boldsymbol{y}|X}(\boldsymbol{\phi}) \right]$ separately. Thus, we get the following two optimisation problems:

(1) the constrained maximisation problem

$$\begin{aligned} \max \quad & \left\{ \mathrm{E}_{\boldsymbol{\theta}^{(m)}} \left(\ell_{\boldsymbol{X}} \left(\boldsymbol{p}^{(1)}, \boldsymbol{P} \right) \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right) \right\} \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k^{(1)} = 1, \\ & \sum_{l=1}^{K} p_{kl} = 1 \qquad (k \in \{1, \ldots, K\}) \end{aligned}$$

and

(2) the unconstrained maximisation problem

$$\max \quad \left\{ \mathrm{E}_{\boldsymbol{\theta}^{(m)}} \left[\ell_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\phi}) \right] \right\}$$

Below we summarise the procedures that were described for the E-step and M-step.

EM Algorithm

---

(1) Initialise with a proper value $\boldsymbol{\theta}^{(0)} = \left[\left(\boldsymbol{p}^{(1)} \right)^{(0)}, \boldsymbol{P}^{(0)}, \boldsymbol{\phi}^{(0)} \right]$ of the parameter.

(2) At the m-th iteration proceed as follows:

(2.1) **E-Step**: Compute $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)} \right)$ as a function of $\boldsymbol{\theta}$, in the following manner:

(2.1.1) For all $k \in \{1, \ldots, K\}$, compute the forward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.1.1) $\alpha_k^{(m)}(1) = \left(p_k^{(1)} \right)^{(m)} f_k \left(y_1; \phi_k^{(m)} \right)$.

(2.1.1.2) $\alpha_k^{(m)}(t) = \left(\sum_{i=1}^{K} \alpha_i^{(m)}(t-1)\, p_{ik}^{(m)} \right) f_k \left(y_t; \phi_k^{(m)} \right)$, for all $t \in \{2, \ldots, T\}$.

(2.1.2) For all $k \in \{1, \ldots, K\}$, compute the backward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.2.1) $\beta_k^{(m)}(T) = 1$.

(2.1.2.2) $\beta_k^{(m)}(t) = \sum_{l=1}^{K} p_{kl}^{(m)} f_l \left(y_{t+1}; \phi_l^{(m)} \right) \beta_l^{(m)}(t+1)$, for all $t \in \{T-1, \ldots, 1\}$.

(2.1.3) For all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$, compute the variables:

$$\gamma_k^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\, \beta_k^{(m)}(t)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\, \beta_i^{(m)}(t)}.$$

(2.1.4) For all $t \in \{1, \ldots, T-1\}$ and $k, l \in \{1, \ldots, K\}$, compute the variables:

$$\xi_{kl}^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\, p_{kl}^{(m)} f_l \left(y_{t+1}; \phi_l^{(m)} \right) \beta_l^{(m)}(t+1)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\, \beta_i^{(m)}(t)}.$$

(2.1.5) Calculate the observed-data likelihood for the current value $\boldsymbol{\theta}^{(m)}$ of the parameter as:

$$L_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m)}\right) = \sum_{k=1}^{K} \alpha_k^{(m)}(T).$$

(2.2) **M-Step**: Using the current values $\gamma_k^{(m)}(t)$ and $\xi_{kl}^{(m)}(t)$, that were calculated in the E-step, determine the value $\boldsymbol{\theta}^{(m+1)} = \left[\left(\boldsymbol{p}^{(1)}\right)^{(m+1)}, \boldsymbol{P}^{(m+1)}, \boldsymbol{\phi}^{(m+1)}\right]$ that maximises $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$. For all $k, l \in \{1, \ldots, K\}$, the values of the initial and transition probabilities of the hidden Markov chain are updated as follows:

$$\left(\widehat{p_k^{(1)}}\right)^{(m+1)} = \frac{\gamma_k^{(m)}(1)}{\sum_{i=1}^{K} \gamma_i^{(m)}(1)},$$

$$\hat{p}_{kl}^{(m+1)} = \frac{\sum_{t=1}^{T-1} \xi_{kl}^{(m)}(t)}{\sum_{t=1}^{T-1} \gamma_k^{(m)}(t)}.$$

(3) Alternate between steps (2.1) and (2.2) until a stopping criterion is met.

2.3 The Data-Generating Model

In this section we introduce the model that will be used to generate the data, proposed by Fiecas et al. (2017), along with the necessary basic assumptions. In contrast with our previous models, from now on, we consider a hidden Markov chain $\{X_t : t \in \mathbb{Z}\}$ and an observable stochastic process $\{Y_t : t \in \mathbb{Z}\}$, where, for all $t \in \mathbb{Z}$, Y_t is assumed to be a p-dimensional random vector.

2.3.1 Definition of the Data-Generating Model

Let $\{X_t : t \in \mathbb{Z}\}$ be a Markov chain with finite state space $S_X = \{1, \ldots, K\}$, and transition probability matrix $\boldsymbol{P}$. Following the notation up to this point, we define the transition probabilities:

$$p_{kl} = \mathrm{P}(X_{t+1} = l \mid X_t = k, X_{t-1}, \ldots, X_1) = \mathrm{P}(X_t = l \mid X_{t-1} = k) \qquad (k, l \in S_X = \{1, \ldots, K\}).$$

The model for the data-generating mechanism is given by the following equation:

$$Y_t = \sum_{k=1}^{K} \mathbf{1}_{\{X_t = k\}} \left(\mu_k + \Sigma_k^{1/2} \epsilon_t\right), \tag{2.45}$$

where ϵ_t, $t \in \mathbb{Z}$ are independent and identically distributed (abbreviated iid) p-dimensional random vectors with:

$$\mathrm{E}(\epsilon_t) = \mathbf{0}_p,$$
$$\mathrm{Var}(\epsilon_t) = \boldsymbol{I}_p,$$

where $\mathbf{0}_p$ is a p-dimensional vector of zeroes and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. At any moment t, there is only one possible value of $k \in \{1, \ldots, K\}$, so that

$$X_t = k \Leftrightarrow \mathbf{1}_{\{X_t = k\}} = 1,$$

which implies that

$$Y_t = \mu_k + \Sigma_k^{1/2} \epsilon_t,$$

for that specific value of k. We observe a sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$, but not the state sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$ of the hidden Markov chain. Fiecas et al. (2017) make three basic assumptions for this model.

(A1) The hidden Markov chain $\{X_t : t \in \mathbb{Z}\}$ is irreducible, aperiodic and stationary,

(A2) ϵ_t, $t \in \mathbb{Z}$, are independent of X_t, $t \in \mathbb{Z}$,

(A3) $\mathrm{E}\left(\epsilon_{t,i}^4\right) - 3 \leq \kappa_\epsilon$, $t \in \mathbb{Z}$, $i \in \{1, \ldots, p\}$, for some constant $\kappa_\epsilon < +\infty$ and $\mathrm{E}\left(\epsilon_{t,i}^8\right) < +\infty$, $t \in \mathbb{Z}$, $i \in \{1, \ldots, p\}$.

By Assumption (A1) it follows that the Markov chain is *α-mixing with exponentially decreasing rate*[6] (see e.g., Lemma 1 in Francq and Roussignol 1997, or Statements (a), (b) and (c) of Theorem 3.1 and the paragraph following it in Bradley 2005). Since the hidden Markov chain of the model is assumed to be irreducible and aperiodic with a finite state space, it follows that is has a unique stationary distribution $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ (see Corollary 1.10.1). Hence, the assumption of stationarity is not necessary. Even if the initial distribution were not also the stationary distribution of the chain, all the results would hold, as the process would reach stationarity after a sufficient amount of time (see the discussion following the Assumptions (A) in Section 2.1 in Fiecas et al. 2017).

Assumptions (A1) and (A2) guarantee that the observed sequence $\{Y_t : t \in \mathbb{Z}\}$ inherits the properties of stationarity and α-mixing with exponentially decreasing rate from the hidden process $\{X_t : t \in \mathbb{Z}\}$ (see e.g., proof of Lemma 1 in Francq and Roussignol 1997, or Section IV-C in Ephraim and Merhav

[6] See Definition A.20 in Apppendix A and the discussion following it. For more information on the topic of α-mixing (and mixing in general) the interested reader is referred to Bradley (2005).

2002). These two properties are necessary to ensure that the maximum likelihood estimates are consistent and asymptotically normal (see the paragraph preceding Proposition 3.1 in Tadjuidje Kamgaing 2013).

The first part of Assumption (A3) states that $\mathrm{E}\left(\epsilon_{t,i}^4\right)$ is uniformly bounded by $\kappa_\epsilon + 3$, for all $t \in \mathbb{Z}$ and $i \in \{1, \ldots, p\}$. This property is stronger than that ensured by the second part, as $\mathrm{E}\left(\epsilon_{t,i}^8\right) < +\infty$ only implies that $\mathrm{E}\left(\epsilon_{t,i}^4\right) < +\infty$. Moreover, the existence of the m-th moment of ϵ_t assures that the m-th moment of Y_t also exists.

Tadjuidje Kamgaing (2013) used the same model as the one given in Equation (2.45) to prove that the maximum likelihood estimates of the parameters are consistent and asympotically normal, under some suitable conditions (see Assumptions 2.1 and 3.1 to 3.3 in Tadjuidje Kamgaing 2013). As the author states in his paper (see Section 3), under these conditions, this model satisfies the assumptions of Douc et al. (2004), whose results were used to derive the asymptotic properties of the parameter estimators. He assumed that ϵ_t, $t \in \mathbb{Z}$ are iid random variables from $N_p(\mathbf{0}_p, \boldsymbol{I}_p)$ and independent of the underlying Markov chain (Assumption 2.1 in Tadjuidje Kamgaing 2013). However, as he mentions (see Sections 2 and 4), the assumption of the normality of the innovations (ϵ_t's) is used as an example and his theory could also be utilised in different settings for the distribution of the innovations.

2.3.2 EM Algorithm for the Data-Generating Model

In this subsection we present a classical approach of the EM algorithm for the data-generating model given by Equation (2.45). We assume that the state-dependent distributions are multivariate Normal[7] with parameters (μ_k, Σ_k), for all $k \in S_X = \{1, \ldots, K\}$, that is $(Y_t | X_t = k) \sim N_p(\mu_k, \Sigma_k)$, for all $t \in \mathbb{Z}$ and $k \in S_X = \{1, \ldots, K\}$. Thus we have

$$f_k(y_t; \phi_k) = \frac{1}{(2\pi)^{p/2} \left[\det(\Sigma_k)\right]^{1/2}} \exp\left\{-\frac{1}{2}(y_t - \mu_k)' \Sigma_k^{-1} (y_t - \mu_k)\right\}, \tag{2.46}$$

where $\phi_k = (\mu_k, \Sigma_k)$, for all $k \in S_X = \{1, \ldots, K\}$.

As we saw in Section 2.2, a vital part of the EM algorithm is the calculation of $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$. To do so, we need to compute the logarithm of the state-dependent distributions (see Equation (2.26)).

[7] See Subsection A.3.2 in Appendix A.

Taking the logarithm of both sides of Equation (2.46), we get

$$\begin{aligned}\log[f_k(y_t;\phi_k)] &= \log\left[\frac{1}{(2\pi)^{p/2}\left[\det(\Sigma_k)\right]^{1/2}}\exp\left\{-\frac{1}{2}\left(y_t-\mu_k\right)'\Sigma_k^{-1}\left(y_t-\mu_k\right)\right\}\right] \\ &= -\frac{p}{2}\log(2\pi)-\frac{1}{2}\log[\det(\Sigma_k)]-\frac{1}{2}\left(y_t-\mu_k\right)'\Sigma_k^{-1}(y_t-\mu_k). \end{aligned} \tag{2.47}$$

We have that

$$-\frac{1}{2}\left(y_t-\mu_k\right)'\Sigma_k^{-1}(y_t-\mu_k) = -\frac{1}{2}y_t'\Sigma_k^{-1}y_t+\frac{1}{2}y_t'\Sigma_k^{-1}\mu_k+\frac{1}{2}\mu_k'\Sigma_k^{-1}y_t-\frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k. \tag{2.48}$$

Notice that the quantity $\mu_k'\Sigma_k^{-1}y_t$ is scalar, which means that it is equal to its transpose, therefore we get

$$\mu_k'\Sigma_k^{-1}y_t = \left(\mu_k'\Sigma_k^{-1}y_t\right)' = \left(\Sigma_k^{-1}y_t\right)'\left(\mu_k'\right)' = y_t'\left(\Sigma_k^{-1}\right)'\mu_k = y_t'\left(\Sigma_k'\right)^{-1}\mu_k = y_t'\Sigma_k^{-1}\mu_k. \tag{2.49}$$

By Equations (2.48) and (2.49), it follows that

$$-\frac{1}{2}\left(y_t-\mu_k\right)'\Sigma_k^{-1}(y_t-\mu_k) = -\frac{1}{2}y_t'\Sigma_k^{-1}y_t+y_t'\Sigma_k^{-1}\mu_k-\frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k \tag{2.50}$$

By (2.47) and (2.50) we get the following relation

$$\begin{aligned}\log[f_k(y_t;\phi_k)] = &-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log[\det(\Sigma_k)]-\frac{1}{2}y_t'\Sigma_k^{-1}y_t \\ &+y_t'\Sigma_k^{-1}\mu_k-\frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k.\end{aligned} \tag{2.51}$$

As has already been mentioned, in the context of HMMs, the hidden path of the Markov chain $\{X_t : t \in \mathbb{Z}\}$, is considered as missing data. Had an *oracle* unveiled the hidden states $(X_1,\ldots,X_T)$, by Equations (2.8), (2.22), (2.47) and (2.51), we would get the following two expressions for the complete-data log-likelihood

$$
\begin{aligned}
\ell_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) &= \ell_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\phi}) + \ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) = \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} \log[f_k(y_t;\phi_k)] + \ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) \\
&= -\frac{1}{2}\sum_{t=1}^{T}\left(\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} \log[\det(\Sigma_k)]\right) - \frac{1}{2}\sum_{t=1}^{T}\left(\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} \left(y_t-\mu_k\right)' \Sigma_k^{-1}(y_t-\mu_k)\right) \\
&\quad + \ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) + c \\
&= -\frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \log[\det(\Sigma_k)]\right) - \frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t-\mu_k\right)' \Sigma_k^{-1}(y_t-\mu_k)\right) \\
&\quad + \ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) + c \qquad (2.52) \\
&= -\frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \log[\det(\Sigma_k)]\right) - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t{}'\Sigma_k^{-1} y_t\right)\right] \\
&\quad + \sum_{k=1}^{K}\left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t{}'\Sigma_k^{-1}\mu_k\right)\right] - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(\mu_k{}'\Sigma_k^{-1}\mu_k\right)\right] \\
&\quad + \ell_{\boldsymbol{X}}\left(\boldsymbol{p}^{(1)}, \boldsymbol{P}\right) + c, \qquad (2.53)
\end{aligned}
$$

where

$$
c = \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} \left[-\frac{p}{2}\log(2\pi)\right] = -\frac{p}{2}\log(2\pi)\sum_{t=1}^{T} 1 = -\frac{p}{2}\log(2\pi)\, T,
$$

as for any fixed value $t \in \mathbb{Z}$, $\mathbf{1}_{\{X_t=k\}} = 1$ for some fixed value $k \in \{1,\ldots,K\}$, and $\mathbf{1}_{\{X_t=l\}} = 0$, for all $l \in \{1,\ldots,K\}$, with $l \neq k$, therefore $\sum_{k=1}^{K} \mathbf{1}_{\{X_t=k\}} = 1$.

Maximising the complete-data likelihood with respect to μ_k and Σ_k, for all $k \in \{1,\ldots,K\}$, yields what we shall call the *oracle estimates*.

To determine the oracle estimators of the means, we set the partial derivative of the complete-data log-likelihood, given by (2.53), with respect to μ_k, equal to the zero vector $\mathbf{0}_p$. Utilising Equations (A.2) and (A.3) (see Appendix A), for all $k \in S_X = \{1,\ldots,K\}$, we have that

$$
\frac{\partial \ell_{\boldsymbol{y},\boldsymbol{X}}(\mu_k,\Sigma_k)}{\partial \mu_k} = \mathbf{0}_p \Leftrightarrow \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t{}'\Sigma_k^{-1}\right) - \frac{1}{2}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(2\mu_k{}'\Sigma_k^{-1}\right) = \mathbf{0}_p,
$$

therefore,

$$
\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(\mu_k{}'\Sigma_k^{-1}\right) = \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t{}'\Sigma_k^{-1}\right). \qquad (2.54)
$$

Multiplying both sides of (2.54) by Σ_k, from the right, gives

$$
\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(\mu_k{}'\right)\boldsymbol{I}_p = \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t{}'\right)\boldsymbol{I}_p \Leftrightarrow \left(\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}\mu_k\right)' = \left(\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} y_t\right)',
$$

which is equivalent to

$$\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \mu_k = \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} y_t.$$

As a result, the oracle estimators of the means are

$$\mu_k^{\mathrm{o}} = \frac{1}{\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}} \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} y_t \qquad (k \in S_X = \{1, \ldots, K\}). \tag{2.55}$$

To determine the oracle estimators of the covariance matrices, we set the partial derivative of the complete-data log-likelihood, given by (2.52), with respect to Σ_k, equal to the zero matrix $\mathbf{0}_{p\times p}$. Utilising Equations (A.4) and (A.5) (see Appendix A), for all $k \in S_X = \{1, \ldots, K\}$, we have that

$$\frac{\partial \ell_{\boldsymbol{y},\boldsymbol{X}}\left(\mu_k^{\mathrm{o}}, \Sigma_k\right)}{\partial \Sigma_k} = \mathbf{0}_{p\times p} \Leftrightarrow$$
$$-\frac{1}{2}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(\Sigma_k{}'\right)^{-1} - \frac{1}{2}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left[-\left(\Sigma_k{}'\right)^{-1} \left(y_t - \mu_k^{\mathrm{o}}\right) \left(y_t - \mu_k^{\mathrm{o}}\right)' \left(\Sigma_k{}'\right)^{-1}\right] = \mathbf{0}_{p\times p}.$$

As Σ_k is symmetric, we get

$$\left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}\right] \Sigma_k^{-1} = \Sigma_k^{-1} \left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t - \mu_k^{\mathrm{o}}\right) \left(y_t - \mu_k^{\mathrm{o}}\right)'\right] \Sigma_k^{-1}. \tag{2.56}$$

Multiplying both sides of (2.56) by Σ_k, both from the left and from the right, yields

$$\left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}\right] \Sigma_k \boldsymbol{I}_p = \boldsymbol{I}_p \left[\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t - \mu_k^{\mathrm{o}}\right) \left(y_t - \mu_k^{\mathrm{o}}\right)'\right] \boldsymbol{I}_p$$

Therefore, the oracle estimators of the covariance matrices are

$$\tilde{\Sigma}_k^{\mathrm{o}} = \frac{1}{\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}} \sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t - \mu_k^{\mathrm{o}}\right) \left(y_t - \mu_k^{\mathrm{o}}\right)' \qquad (k \in S_X = \{1, \ldots, K\}). \tag{2.57}$$

Notice that the oracle estimators μ_k^{o} and $\tilde{\Sigma}_k^{\mathrm{o}}$ are the sample estimators of the parameters μ_k and Σ_k, respectively, as for any fixed $k \in \{1, \ldots, K\}$, the sample size is $\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}$. In case $\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} = 0$, that is in case there is a state $k \in \{1, \ldots, K\}$ which is not observed, Fiecas et al. (2017) set $\mu_k^{\mathrm{o}} = \mathbf{0}_p$ and $\tilde{\Sigma}_k^{\mathrm{o}} = \mathbf{0}_{p\times p}$. As they also state in their paper (see the discussion following Equation (6) in Subsection 2.2), this happens with exponentially decreasing probability and it is asymptotically negligible. However, in case there are some rarely visited states, such that $\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}$ is very small, some numerical problems may arise, hence they choose to work with the biased estimator of the covariance

matrix (see also Equation (A.1) in Appendix A)

$$\Sigma_k^{\mathrm{o}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(y_t - \mu_k^{\mathrm{o}}\right)\left(y_t - \mu_k^{\mathrm{o}}\right)' = \pi_k^{\mathrm{o}}\tilde{\Sigma}_k^{\mathrm{o}} \qquad (k \in S_X = \{1,\ldots,K\}), \tag{2.58}$$

where

$$\pi_k^{\mathrm{o}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}}, \tag{2.59}$$

is the oracle (empirical) estimate of the stationary probability $\pi_k = \mathrm{P}(X_t = k)$ (see Subsection 1.2.4.1).

2.3.2.1 E-Step

As the E-step is not affected by the state-dependent distribution, it is carried out as was presented in Paragraph 2.2.2.1. The only difference is that in Equations (2.28), the initial probability $p_k^{(1)}$ is replaced by the stationary probability π_k, that is

$$\alpha_k(1) = \pi_k f_k(y_1;\boldsymbol{\phi}) \qquad (k \in \{1,\ldots,K\}),$$

since we have assumed that the hidden Markov chain $\{X_t : t \in \mathbb{Z}\}$ is stationary.

2.3.2.2 M-Step

Utilising the variables $\gamma_k(t)$ and $\xi_{k,l}(t)$ obtained in the E-step, we get the updating formulas for π_k, p_{kl}, μ_k and Σ_k, as

$$\hat{\pi}_k^{(m+1)} = \frac{1}{T}\sum_{t=1}^{T} \gamma_k^{(m)}(t) \qquad (k \in \{1,\ldots,K\}), \tag{2.60}$$

$$\hat{p}_{kl}^{(m+1)} = \frac{\sum_{t=1}^{T-1} \xi_{kl}^{(m)}(t)}{\sum_{t=1}^{T-1} \gamma_k^{(m)}(t)} \qquad (k,l \in \{1,\ldots,K\}), \tag{2.61}$$

$$\hat{\mu}_k^{(m+1)} = \frac{\sum_{t=1}^{T} \gamma_k^{(m)}(t)\, y_t}{\sum_{t=1}^{T} \gamma_k^{(m)}(t)} \qquad (k \in \{1,\ldots,K\}), \tag{2.62}$$

$$\hat{\Sigma}_k^{(m+1)} = \frac{1}{\hat{\pi}_k^{(m+1)}}\hat{\Sigma}_k^{\mathrm{o}(m+1)} \qquad (k \in \{1,\ldots,K\}), \tag{2.63}$$

where

$$\hat{\Sigma}_k^{\mathrm{o}(m+1)} = \frac{1}{T}\sum_{t=1}^{T} \gamma_k^{(m)}(t) \left(y_t - \hat{\mu}_k^{(m+1)}\right)\left(y_t - \hat{\mu}_k^{(m+1)}\right)' \qquad (k \in \{1,\ldots,K\}). \tag{2.64}$$

The EM algorithm for the data generating model is summarised as follows.

EM Algorithm for the Data Generating Model

(1) Initialise with a proper value $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\pi}^{(0)}, \boldsymbol{P}^{(0)}, \boldsymbol{\phi}^{(0)}\right)$ of the parameter.

(2) At the m-th iteration proceed as follows:

(2.1) **E-Step**: Compute $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ as a function of $\boldsymbol{\theta}$, in the following manner:

(2.1.1) For all $k \in \{1, \ldots, K\}$, compute the forward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.1.1) $\alpha_k^{(m)}(1) = \pi_k^{(m)} f_k\left(y_1; \phi_k^{(m)}\right)$.

(2.1.1.2) $\alpha_k^{(m)}(t) = \left(\sum_{i=1}^{K} \alpha_i^{(m)}(t-1)\, p_{ik}^{(m)}\right) f_k\left(y_t; \phi_k^{(m)}\right)$, for all $t \in \{2, \ldots, T\}$.

(2.1.2) For all $k \in \{1, \ldots, K\}$, compute the backward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.2.1) $\beta_k^{(m)}(T) = 1$.

(2.1.2.2) $\beta_k^{(m)}(t) = \sum_{l=1}^{K} p_{kl}^{(m)} f_l\left(y_{t+1}; \phi_l^{(m)}\right) \beta_l^{(m)}(t+1)$, for all $t \in \{T-1, \ldots, 1\}$.

(2.1.3) For all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$, compute the variables:

$$\gamma_k^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\, \beta_k^{(m)}(t)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\, \beta_i^{(m)}(t)}.$$

(2.1.4) For all $t \in \{1, \ldots, T-1\}$ and $k, l \in \{1, \ldots, K\}$, compute the variables:

$$\xi_{kl}^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\, p_{kl}^{(m)} f_l\left(y_{t+1}; \phi_l^{(m)}\right) \beta_l^{(m)}(t+1)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\, \beta_i^{(m)}(t)}.$$

(2.1.5) Calculate the observed-data likelihood for the current value $\boldsymbol{\theta}^{(m)}$ of the parameter as:

$$L_{\boldsymbol{y}}\left(\boldsymbol{\theta}^{(m)}\right) = \sum_{k=1}^{K} \alpha_k^{(m)}(T).$$

(2.2) **M-Step**: Using the current values $\gamma_k^{(m)}(t)$ and $\xi_{kl}^{(m)}(t)$, that were calculated in the E-step, determine the value $\boldsymbol{\theta}^{(m+1)} = \left(\boldsymbol{\pi}^{(m+1)}, \boldsymbol{P}^{(m+1)}, \boldsymbol{\phi}^{(m+1)}\right)$ that maximises $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$. For all $k, l \in \{1, \ldots, K\}$, the values of the initial and transition probabilities of the hidden Markov

chain are updated as follows:

$$\hat{\pi}_k^{(m+1)} = \frac{1}{T}\sum_{t=1}^{T}\gamma_k^{(m)}(t),$$
$$\hat{p}_{kl}^{(m+1)} = \frac{\sum_{t=1}^{T-1}\xi_{kl}^{(m)}(t)}{\sum_{t=1}^{T-1}\gamma_k^{(m)}(t)}.$$

For all $k \in \{1,\ldots,K\}$, the parameters of the state-dependent distributions $\boldsymbol{\phi}^{(m+1)} = \left(\phi_k^{(m+1)}\right)_{k\in S_X}$, with $\phi_k = (\mu_k, \Sigma_k)$, are updated as follows:

$$\hat{\mu}_k^{(m+1)} = \frac{\sum_{t=1}^{T}\gamma_k^{(m)}(t)\, y_t}{\sum_{t=1}^{T}\gamma_k^{(m)}(t)},$$
$$\widehat{\Sigma}_k^{(m+1)} = \frac{1}{\hat{\pi}_k^{(m+1)}}\widehat{\Sigma}_k^{\mathrm{o}(m+1)},$$

where

$$\widehat{\Sigma}_k^{\mathrm{o}(m+1)} = \frac{1}{T}\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t - \hat{\mu}_k^{(m+1)}\right)\left(y_t - \hat{\mu}_k^{(m+1)}\right)'.$$

(3) Alternate between steps (2.1) and (2.2) until a stopping criterion is met.

Remark. *Following* Fiecas et al. (2017), *instead of maximising* $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$ *with respect to* μ_k *and* Σ_k, *to determine their updating formulas, we initially calculated the maximum likelihood estimators of these parameters, as if the states were known and then replaced the unknown quantities* $\mathbf{1}_{\{X_t=k\}}$ *by their estimates* $\gamma_k(t)$. *The reason why this framework was adopted will become clear in* Chapter 3. *However, had we maximised* $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$ *with respect to* μ_k *and* Σ_k, *the same formulas as the ones given by* Equations (2.62) *and* (2.63) *would have emerged (see* Appendix B*).*

It should also be noted that instead of using the empirical estimator of the stationary distribution in the EM algorithm, in the case of a stationary Markov chain, another method can be implemented, presented in Subsection 4.2.5 *in* Zucchini et al. (2016).

Chapter 3

EM Algorithm and Shrinkage

Estimating the covariance matrix of a sample of random variables is an essential part in many areas of multivariate statistical analysis, such as principal component analysis (PCA), linear and quadratic discrimination analysis (LDA and QDA), or graphical models (see, e.g., the introductions in Bickel and Levina 2008 and Cai et al. 2011). Many real-world applications, where these methods are implemented (for instance genetic data, brain imaging and spectroscopic imaging), deal with high-dimensional data (see the Introduction in Bickel and Levina 2008). In other problems, such as portfolio risk assessment and optimal portfolio allocation, mean–variance portfolio optimisation for a large number of assets (see, e.g., the introductions in Ledoit and Wolf 2004; Sancetta 2008; Fan et al. 2008), where the data can also be of high dimension, it may be crucial to also estimate the inverse of the covariance matrix, also known as the *precision matrix* (see, for instance, the introductions in Bickel and Levina 2008 and Cai et al. 2011). However, in situations like these, estimation of a large covariance matrix and/or its inverse, can be very challenging. The natural estimator of a covariance matrix, the sample covariance matrix, is well-known that is not a good estimate when the dimension p is comparable to the sample size T. As Ledoit and Wolf (2004) point out in their paper, when the ratio p/T is larger than one, the sample covariance matrix is singular (see also Cai et al. 2011). When the ratio p/T is less than one but not negligible, even though the sample covariance matrix is invertible, it is numerically ill-conditioned[1]. In cases where the dimension p is very large, it may be impossible to find enough observations, so that the ratio p/T becomes negligible. This problem may be even more pronounced in the context of HMMs, where the sample size for some states might be minute, if they are rarely visited.

[1] A problem of solving a system of linear equations is characterised as *ill-conditioned* when small perturbations in the data (for instance, in the elements of a matrix), result in large changes in the result, otherwise it is referred to as *well-conditioned* (see the Introduction of Chapter 2 in Ciarlet 1989).

If a matrix is ill-conditioned for the problem of its inversion, approximating its inverse will probably result in an inaccurate solution due to its sensitivity to the values of the coefficients (see Chapter 4 in Westlake 1968).

A popular method to combat this problem is shrinkage. Shrinkage is a regularisation[2] technique not only used to estimate large covariance matrices, but also in other areas of statistics, such as regression analysis (Copas 1983; Tibshirani 1996; Zou 2006) and in spectral analysis of time-series data, where it has been used to estimate the spectral density matrix (Böhm and von Sachs 2009; Fiecas and Ombao 2011). In covariance matrix estimation, the idea is to obtain an estimator that is a weighted average between the sample covariance matrix, which is known to be unbiased (see the Introduction in Ledoit and Wolf 2004) and a properly chosen matrix, called the *shrinkage target* or *target matrix* (see, e.g., Subsection 2.1 in Ledoit and Wolf 2004, Subsection 2.1 in Yuan and Huang 2009, or Appendix A in Hausser and Strimmer 2008), so that the resulting estimator, asymptotically, possesses some desirable properties. This is possible when the weight on the target matrix, which is called the *shrinkage intensity* (see the same references as for the shrinkage target), is chosen optimally according to some loss function[3] . As Ledoit and Wolf (2004) state, under standard asymptotics, where the dimension of the random variables p is finite and fixed and the sample size T goes to infinity, the sample covariance matrix is well-conditioned in the limit. This assumption is not realistic in many situations where the dimension p is of the same order of magnitude as the sample size T, or even larger.

Chapter 3 is organised as follows. In Section 3.1, we give a brief overview of the work of Ledoit and Wolf (2004) and show how their estimator can be obtained in the case of iid Normal random vectors, using a penalised likelihood, as defined by Yuan and Huang (2009). In Section 3.2, the framework of Section 3.1 is implemented in the case where there are several regimes revealed by an *oracle*, similarly to Subsection 2.3.2. Finally, in Section 3.3 the modified version of the EM algorithm of Fiecas et al. (2017) is presented.

3.1 Penalised Likelihood For Independent And Identically Distributed Random Vectors

As we mentioned above, in this section we introduce the shrinkage estimator of Ledoit and Wolf (2004) for iid data and show how it can be obtained using the penalised log-likelihood of Yuan and Huang (2009).

Ledoit and Wolf (2004) proposed an estimator that, asymptotically, as the dimension p and the sample size T go to infinity simultaneously, is the optimal convex linear combination of the sample covariance matrix with a multiple of the identity matrix. According to the authors, a constructed

[2](see Demmel 1997) The term *regularisation* is used to describe the process of turning an ill-conditioned problem into a well-conditioned one, by imposing some additional conditions on the solution (see the introduction of Section 3.5 in Demmel 1997).

[3]See Definition A.27 in Appendix A.

estimator with equal variances and null covariances (that is, a diagonal matrix with equal diagonal entries) is well-conditioned. Therefore, by taking the weighted average of the sample covariance matrix and the shrinkage target (i.e. a multiple of the identity matrix) and choosing the weights optimally according to a quadratic loss function, the resulting estimator inherits the desired properties of both matrices.

Firstly, the following definitions are needed.

Definition 3.1. *Consider a square $p \times p$ matrix $\boldsymbol{A}$. The Frobenius norm is defined as*

$$||\boldsymbol{A}||_{\mathrm{F}} = \sqrt{\frac{\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^{'}\right)}{p}}. \tag{3.1}$$

We can also define an associated inner product of the Frobenius norm.

Definition 3.2. *Consider two square $p \times p$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The associated inner product of the Frobenius norm is defined as*

$$\langle \boldsymbol{A}\ ,\ \boldsymbol{B} \rangle = \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{B}^{'}\right). \tag{3.2}$$

Lemma 3.1. *Consider a $T \times p$ matrix $\boldsymbol{Y}$ of T observations of a family of p-dimensional iid random vectors $\{Y_1, \ldots, Y_T\}$ with mean zero and covariance matrix Σ. Then the sample covariance matrix given by*

$$\boldsymbol{S} = \frac{1}{T}\boldsymbol{Y}^{'}\boldsymbol{Y}, \tag{3.3}$$

is an unbiased estimator of the true covariance matrix Σ.

Proof. For all $i, j \in \{1, \ldots, p\}$ we have that

$$\left(\boldsymbol{Y}^{'}\boldsymbol{Y}\right)_{ij} = \sum_{t=1}^{T} Y_{t,i}Y_{t,j}.$$

Taking the expectation of the elements $\boldsymbol{S}_{ij}$, for all $i, j \in \{1, \ldots, p\}$, yields

$$\mathrm{E}\left(\boldsymbol{S}_{ij}\right) = \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,i}Y_{t,j}\right) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(Y_{t,i}Y_{t,j}\right) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{Cov}\left(Y_{t,i}, Y_{t,j}\right) = \mathrm{Cov}\left(Y_{t,i}, Y_{t,j}\right),$$

where we have used the fact that $\mathrm{E}(Y_t) = 0$, for all $t \in \{1, \ldots, T\}$ (see also Definition A.37 in Appendix A). Therefore,

$$\mathrm{E}(\boldsymbol{S}) = \Sigma,$$

(see Definition A.38 in Appendix A) and the proof is complete. □

In the following definition, four scalar quantities, which are crucial for the analysis of Ledoit and Wolf (2004), are presented.

Definition 3.3. Ledoit and Wolf (2004) *define the following four scalar quantities:*

(i) $\mu = \langle \Sigma \, , \, \boldsymbol{I}_p \rangle = \frac{1}{p} \mathrm{tr}\left(\Sigma \boldsymbol{I}_p{}'\right) = \frac{1}{p} \mathrm{tr}(\Sigma)$,

(ii) $\alpha^2 = ||\Sigma - \mu \boldsymbol{I}_p||_{\mathrm{F}}^2$,

(iii) $\beta^2 = \mathrm{E}\left(||\boldsymbol{S} - \Sigma||_{\mathrm{F}}^2\right)$,

(iv) $\delta^2 = \mathrm{E}\left(||\boldsymbol{S} - \mu \boldsymbol{I}_p||_{\mathrm{F}}^2\right)$.

The following lemma shows the relation between α^2, β^2 and δ^2.

Lemma 3.2. *For the scalar quantities α^2, β^2 and δ^2 it holds that*

$$\delta^2 = \alpha^2 + \beta^2.$$

Proof. See the proof of Lemma 2.1 in Ledoit and Wolf (2004). □

The shrinkage estimator of Ledoit and Wolf (2004) is given by the following theorem as the solution of a quadratic programming problem under a linear constraint.

Theorem 3.1. *Consider the optimisation problem:*

$$\begin{aligned} \min_{\rho_1, \rho_2} \quad & \left\{ \mathrm{E}\left(||\Sigma^* - \Sigma||_{\mathrm{F}}^2\right) \right\} \\ \text{s.t.} \quad & \Sigma^* = \rho_1 \boldsymbol{S} + \rho_2 \boldsymbol{I}_p, \end{aligned}$$

where the coefficients ρ_1 and ρ_2 are nonrandom. Its solution verifies

$$\Sigma^* = \frac{\alpha^2}{\delta^2} \boldsymbol{S} + \frac{\beta^2}{\delta^2} \mu \boldsymbol{I}_p, \tag{3.4}$$

and

$$\mathrm{E}\left(||\Sigma^* - \Sigma||_{\mathrm{F}}^2\right) = \frac{\alpha^2 \beta^2}{\delta^2}. \tag{3.5}$$

Proof. See the proof of Theorem 2.1 in Ledoit and Wolf (2004). □

We now proceed to present four interpretations of Theorem 3.1, above, which can be found in Subsection 2.2 of Ledoit and Wolf (2004).

(i) Consider the Hilbert space of $p \times p$ symmetric random matrices $\boldsymbol{A}$, such that $\mathrm{E}\left(||\boldsymbol{A}||_{\mathrm{F}}^2\right) < +\infty$, whose associated norm is $\sqrt{\mathrm{E}\left(||\cdot||_{\mathrm{F}}^2\right)}$ and the inner product of two random matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ is $\mathrm{E}(\langle \boldsymbol{A}_1 \, , \, \boldsymbol{A}_2 \rangle)$. Then Theorem 3.1 can be interpreted as a projection theorem on this Hilbert space. In this context Lemma 3.2 can be thought of as a restatement of the Pythagorean Theorem (see the first interpretation of Theorem 2.1 in Ledoit and Wolf 2004). The interpretation of Equation (3.4) is that the true covariance matrix Σ is projected onto the space spanned by the identity matrix $\boldsymbol{I}_p$ and the sample covariance matrix $\boldsymbol{S}$. To achieve this, Σ is initially projected onto the line spanned by $\boldsymbol{I}_p$, which results in the shrinkage target $\mu \boldsymbol{I}_p$, and then it is projected onto the line that joins the shrinkage target $\mu \boldsymbol{I}_p$ to the sample covariance matrix $\boldsymbol{S}$. The relation between the distance of the true covariance matrix Σ to the shrinkage target $\mu \boldsymbol{I}_p$ and the distance of Σ to the sample covariance matrix $\boldsymbol{S}$, determines whether the shrinkage estimator Σ^* ends up being closer to $\mu \boldsymbol{I}_p$, or to $\boldsymbol{S}$.

(ii) The second interpretation is that of a trade-off between bias and variance. The objective is to minimise the mean squared error of the shrinkage estimator Σ^*, which can be expressed as the sum of its variance and its squared bias (see Equation (10) in Ledoit and Wolf 2004 and Proposition A.3 in Appendix A), that is

$$\mathrm{E}\left(||\Sigma^* - \Sigma||_{\mathrm{F}}^2\right) = \mathrm{E}\left(||\Sigma^* - \mathrm{E}(\Sigma^*)||_{\mathrm{F}}^2\right) + ||\mathrm{E}(\Sigma^*) - \Sigma||_{\mathrm{F}}^2 .$$

For the shrinkage target $\mu \boldsymbol{I}_p$ the mean squared error is equal to its squared bias, while for the sample covariance matrix $\boldsymbol{S}$ the mean squared error is equal to its variance. Therefore, the shrinkage estimator Σ^* is the quantity that optimally balances the error coming from bias and the error coming from variance. It should be noted that the trade-off between bias and variance is a central idea of shrinkage (see, e.g. the paragraph following Equation (10) in Ledoit and Wolf 2004, the Introduction and the discussion following Proposition 1 in Sancetta 2008, or the Introduction in DeMiguel et al. 2013).

(iii) The third interpretation is Bayesian. Let us assume that the prior information for the true covariance matrix Σ is that it lies on the sphere whose centre is the shrinkage target $\mu \boldsymbol{I}_p$ with radius α (see (ii) in Definition 3.3). The information obtained from the data is that the true covariance Σ lies on the sphere with centre the sample covariance matrix $\boldsymbol{S}$ and radius β (see (iii) in Definition 3.3). Then Σ^* can be deemed as the combination of the prior information and the information provided by the data. Combining these, the true covariance matrix should lie on the intersection of the two spheres, which is a circle with centre the shrinkage estimator

Σ^*. The relation between the accuracy of the shrinkage target $\mu \boldsymbol{I}_p$ and the accuracy of the sample covariance matrix $\boldsymbol{S}$ determines the relative importance given to prior versus sample information for the resulting shrinkage estimator Σ^*. As Ledoit and Wolf (2004) point out, for a complete Bayesian approach, both the support and the distribution of the true covariance matrix Σ would need to be specified. Therefore, Σ^* should be seen as the centre of mass of the support of Σ, instead of as the expectation of the posterior distribution, as is usually the case.

(iv) Let us denote by λ_i and l_i, $i \in \{1, \dots, p\}$, the eigenvalues of the true covariance matrix Σ and the eigenvalues of the sample covariance matrix $\boldsymbol{S}$, respectively. The scalar quantity μ, defined in (i) of Definition 3.3 can be expressed as (see Equation (11) in Ledoit and Wolf 2004)

$$\mu = \frac{1}{p}\text{tr}(\Sigma) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i = \text{E}\left(\frac{1}{p}\sum_{i=1}^{p} l_i\right).$$

The last equality is proved as follows

$$\frac{1}{p}\text{tr}(\Sigma) = \frac{1}{p}\text{tr}[\text{E}(\boldsymbol{S})] = \frac{1}{p}\text{E}[\text{tr}(\boldsymbol{S})] = \frac{1}{p}\text{E}\left(\sum_{i=1}^{p} l_i\right),$$

(for the second equality see (C.48) in the proof of Lemma 3.9 in Subsection C.2.3 of Appendix C). Therefore, μ represents the mean of the eigenvalues of both the true and the sample covariance matrix (the eigenvalues of the sample covariance matrix are also called *sample eigenvalues*, see, e.g. Subsection 2.2 in Ledoit and Wolf 2004, or the Introduction in Mestre 2008). Similarly, Lemma 3.2 can be written as (see Equation (12) Ledoit and Wolf 2004)

$$\frac{1}{p}\text{E}\left[\sum_{i=1}^{p}(l_i - \mu)^2\right] = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - \mu)^2 + \text{E}\left(||\boldsymbol{S} - \Sigma||_{\text{F}}^2\right), \tag{3.6}$$

(for a proof of Equation 3.6, see Section C.1 of Appendix C). Equation 3.6 shows that the sample eigenvalues are more dispersed around their mean than the true ones, and the excess dispersion equals the risk[4] of the sample covariance matrix $\boldsymbol{S}$. This means that the largest sample eigenvalues are biased upwards (they overestimate the corresponding true eigenvalues) and the smallest are biased downwards (they underestimate the corresponding true eigenvalues), which is a well-known issue (see, e.g., Section 2 in Muirhead 1987, the Introduction in Johnstone 2001, Subsections 2.2 and 2.3 in Ledoit and Wolf 2004, or the Introduction in Mestre 2008).

[4]See Definition A.29 in Appendix A.

Shrinking the sample eigenvalues by taking

$$\lambda_i^* = \frac{\alpha^2}{\delta^2} l_i + \frac{\beta^2}{\delta^2} \mu \qquad (i \in \{1, \dots, p\}),$$

is a way of improving upon the sample covariance matrix. These are exactly the eigenvalues of the shrinkage estimator Σ^* and their dispersion is even lower than the one of the true eigenvalues, as it is shown by the following lemma.

Lemma 3.3. *The eigenvalues of the shrinkage estimator Σ^*, defined as*

$$\lambda_i^* = \frac{\alpha^2}{\delta^2} l_i + \frac{\beta^2}{\delta^2} \mu \qquad (i \in \{1, \dots, p\}),$$

are less dispersed around their mean, compared to the eigenvalues of the true covariance matrix Σ.

Proof. For all $i \in \{1, \dots, p\}$, by Lemma 3.2, we get

$$\lambda_i^* - \mu = \frac{\alpha^2}{\delta^2} l_i + \frac{\beta^2}{\delta^2} \mu - \frac{\alpha^2 + \beta^2}{\delta^2} \mu = \frac{\alpha^2}{\delta^2} (l_i - \mu).$$

Therefore, for the dispersion of the eigenvalues of the shrinkage estimator Σ^*, by (C.1), we get

$$\begin{aligned} \frac{1}{p} \mathrm{E}\left[\sum_{i=1}^{p} (\lambda_i^* - \mu)^2 \right] &= \frac{1}{p} \mathrm{E}\left(\sum_{i=1}^{p} \left[\frac{\alpha^2}{\delta^2} (l_i - \mu) \right]^2 \right) = \frac{1}{p} \mathrm{E}\left[\left(\frac{\alpha^2}{\delta^2} \right)^2 \sum_{i=1}^{p} (l_i - \mu)^2 \right] \\ &= \left(\frac{\alpha^2}{\delta^2} \right)^2 \frac{1}{p} \mathrm{E}\left[\sum_{i=1}^{p} (l_i - \mu)^2 \right] = \left(\frac{\alpha^2}{\delta^2} \right)^2 \delta^2 = \frac{\left(\alpha^2\right)^2}{\delta^2}. \end{aligned}$$

Since $\delta^2 = \alpha^2 + \beta^2$,

$$\frac{\left(\alpha^2\right)^2}{\delta^2} = \frac{\left(\alpha^2\right)^2}{\alpha^2 + \beta^2} < \frac{\left(\alpha^2\right)^2}{\alpha^2} = \alpha^2,$$

and, by Equation (C.2), the proof is complete. □

Shrinking the sample eigenvalues towards their mean results to an estimator whose condition number[5] is closer to one, therefore, the shrinkage estimator Σ^* is a well-conditioned matrix.

[5]The *condition number* for the problem of solving a system of linear equations is indicative of whether the problem is well- or ill-conditioned. A small condition number indicates a well-conditioned problem, while a large condition number indicates an ill-conditioned problem (see, e.g. Subsection 2.2 in Ciarlet 1989, Theorem 2.1 in Demmel 1997, or Chapter 4 in Westlake 1968).

For a positive definite matrix, whose eigenvalues are all positive (see Proposition A.10), the condition number can be expressed as the ratio of its largest to its smallest eigenvalue (see the third statement of Theorem 2.2-3 in Ciarlet 1989), that is

$$\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}.$$

A matrix whose condition number is close to one, is well-conditioned for solving linear equations, hence it is also well-conditioned for the problem of its inversion (see Chapter 4 in Westlake 1968).

By Definition 3.3 and Equation (3.4) it becomes apparent that Σ^* depends on the true covariance matrix Σ, which is unknown, therefore Σ^* is not a feasible, or, as Ledoit and Wolf (2004) call it, a *bona fide* estimator. Replacing the quantities μ, α^2, β^2 and δ^2 in the definition of Σ^*, by consistent estimators, results in a bona fide estimator of Σ.

Following Ledoit and Wolf (2004), let us rewrite the four quantities of Definition (3.3) as:

(i) $\mu_T = \langle \Sigma_T \, , \, \boldsymbol{I}_p \rangle = \frac{1}{p}\text{tr}\left(\Sigma_T \boldsymbol{I}_p{}'\right) = \frac{1}{p}\text{tr}(\Sigma_T)$,

(ii) $\alpha_T^2 = \left|\left|\Sigma_T - \mu_T \boldsymbol{I}_p\right|\right|_{\text{F}}^2$,

(iii) $\beta_T^2 = \text{E}\left(\left|\left|\boldsymbol{S}_T - \Sigma_T\right|\right|_{\text{F}}^2\right)$,

(iv) $\delta_T^2 = \text{E}\left(\left|\left|\boldsymbol{S}_T - \mu_T \boldsymbol{I}_p\right|\right|_{\text{F}}^2\right)$,

where the subscript T is used to emphasise the dependence of the following results on the sample size. What follows holds asymptotically as $T \to +\infty$ (see the first paragraph of Subsection 3.3 in Ledoit and Wolf 2004).

We now give a series of lemmas (see Lemmas 3.2 to 3.5 in Ledoit and Wolf 2004), before we state the main result in Theorem 3.2.

Lemma 3.4. *If* $m_T = \langle \boldsymbol{S}_T \, , \, \boldsymbol{I}_p \rangle$, *then* $\text{E}(m_T) = \mu_T$, *for all* T *and* $m_T - \mu_T$ *converges to zero in quartic mean, that is* $\text{E}\left[(m_T - \mu_T)^4\right] \xrightarrow[T \to +\infty]{} 0$.

Proof. See the proof of Lemma 3.2 in Appendix A in Ledoit and Wolf (2004). □

A direct consequence of Lemma 3.4 is that $m_T^2 - \mu_T^2 \xrightarrow{\text{q.m.}} 0$ and $m_T - \mu_T \xrightarrow{\text{q.m.}} 0$, where $\xrightarrow{\text{q.m.}}$ is used to denote convergence in quadratic mean (see the paragraph following Lemma 3.2 in Ledoit and Wolf 2004).

Lemma 3.5. *If* $d_T^2 = \left|\left|\boldsymbol{S}_T - m_T \boldsymbol{I}_p\right|\right|_{\text{F}}^2$, *then* $d_T^2 - \delta_T^2 \xrightarrow{q.m.} 0$.

Proof. See the proof of Lemma 3.3 in Appendix A in Ledoit and Wolf (2004). □

Lemma 3.6. *Let*

$$\bar{b}_T^2 = \frac{1}{T^2}\sum_{t=1}^{T}\left|\left|Y_{t.}Y_{t.}{}' - \boldsymbol{S}_T\right|\right|_{\text{F}}^2,$$

and $b_T^2 = \min\left\{\bar{b}_T^2, d_T^2\right\}$, *where* $Y_{t.}$ *denotes the* t*-th row of the matrix* $\boldsymbol{Y}$ *that contains the observations. Then* $\bar{b}_T^2 - \beta_T^2 \xrightarrow{q.m.} 0$ *and* $b_T^2 - \beta_T^2 \xrightarrow{q.m.} 0$.

Proof. See the proof of Lemma 3.4 in Appendix A in Ledoit and Wolf (2004). □

Lemma 3.7. *If* $\mathsf{a}_T^2 = d_T^2 - b_T^2$, *then* $\mathsf{a}_T^2 - \alpha_T^2 \xrightarrow{q.m.} 0$.

Proof. See the proof of Lemma 3.5 in Appendix A in Ledoit and Wolf (2004). □

We are now able to present the bona fide shrinkage estimator of the true covariance matrix, which is a consistent estimator of Σ_T^*.

Theorem 3.2. *If*

$$\boldsymbol{S}_T^* = \frac{\mathsf{a}_T^2}{d_T^2}\boldsymbol{S}_T + \frac{b_T^2}{d_T^2} m_T \boldsymbol{I}_p,$$

then $||\boldsymbol{S}_T^* - \Sigma_T^*||_{\mathrm{F}} \xrightarrow{q.m.} 0$. *Consequently, the shrinkage estimator* $\boldsymbol{S}_T^*$ *is a consistent estimator of* Σ_T^*. *Additionally,* $\boldsymbol{S}_T^*$ *and* Σ_T^* *have, asymptotically, the same risk as estimators of the true covariance matrix* Σ_T, *that is*

$$\mathrm{E}\left(||\boldsymbol{S}_T^* - \Sigma_T||_{\mathrm{F}}^2\right) - \mathrm{E}\left(||\Sigma_T^* - \Sigma_T||_{\mathrm{F}}^2\right) \xrightarrow[T\to+\infty]{} 0.$$

Proof. See the proof of Theorem 3.2 in Appendix A in Ledoit and Wolf (2004). □

We now proceed to show how the shrinkage estimator of Ledoit and Wolf (2004) can be obtained, using the method of Yuan and Huang (2009).

Consider a set of observations $\boldsymbol{y} = (y_1, \ldots, y_T)$ that are a realisation of T iid random vectors $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ from $N_p(\mu, \Sigma)$. This is equivalent to assuming that the data come from model (2.45), where the state space of the underlying Markov chain is a singleton ($K = 1$) and the random vectors ϵ_t, $t \in \{1, \ldots, T\}$ are iid from $N_p(\mathbf{0}_p, \boldsymbol{I}_p)$. For all $t \in \{1, \ldots, T\}$ we have

$$f(y_t; \mu, \Sigma) = (2\pi)^{-p/2} \left[\det(\Sigma)\right]^{-1/2} \exp\left\{-\frac{1}{2}(y_t - \mu)' \Sigma^{-1}(y_t - \mu)\right\}.$$

The likelihood function is

$$L_{\boldsymbol{y}}(\mu, \Sigma) = \prod_{t=1}^{T} f(y_t; \mu, \Sigma) = (2\pi)^{-Tp/2} \left[\det(\Sigma)\right]^{-T/2} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(y_t - \mu)' \Sigma^{-1}(y_t - \mu)\right\},$$

and the log-likelihood is

$$\ell_{\boldsymbol{y}}(\mu, \Sigma) = -\frac{Tp}{2}\log(2\pi) - \frac{T}{2}\log[\det(\Sigma)] - \frac{1}{2}\sum_{t=1}^{T}(y_t - \mu)' \Sigma^{-1}(y_t - \mu).$$

Similarly to Yuan and Huang (2009) (see Subsection 2.1), we give the penalised log-likelihood in the following definition.

Definition 3.4. *The penalised log-likelihood function is defined as*

$$p_{\boldsymbol{y}}(\mu, \Sigma) = \frac{2}{T}\ell_{\boldsymbol{y}}(\mu, \Sigma) - \lambda J(\Sigma), \tag{3.7}$$

where

$$J(\Sigma) = \log[\det(\Sigma)] + \operatorname{tr}\left(\Sigma^{-1}\Omega\right). \tag{3.8}$$

According to Yuan and Huang (2009), the penalty function given by Equation (3.8) is strictly convex over the cone of positive definite matrices, is chosen so that it is ensured that the penalised log-likelihood function, defined in Equation (3.7), has a unique maximiser and its maximisation is computationally tractable. The penalty has a unique minimiser Ω, which is an appropriately chosen, well-conditioned matrix and shrinkage target of the resulting shrinkage estimator (see the discussion following Equation (3) in Subsection 2.1 of their paper).

The coefficient λ is called the *regularisation parameter* (see Fan et al. 2006).

Lemma 3.8. *Consider the maximisation problem*

$$\max_{\mu,\, \Sigma>0} \{p_y(\mu,\Sigma)\} = \max_{\mu,\, \Sigma>0} \left\{ \frac{2}{T}\ell_y(\mu,\Sigma) - \lambda J(\Sigma) \right\},$$

where $\Sigma > 0$ denotes that Σ is a positive definite matrix. The estimators for μ and Σ, denoted by $\hat{\mu}$, Σ^s, respectively that maximise the objective function are

$$\hat{\mu} = \bar{y},$$

and

$$\Sigma^s = \frac{1}{1+\lambda}S + \frac{\lambda}{1+\lambda}\Omega.$$

Proof. The penalised log-likelihood, given by (3.7), is

$$\begin{aligned} p_y(\mu,\Sigma) &= \frac{2}{T}\left(-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log[\det(\Sigma)] - \frac{1}{2}\sum_{t=1}^{T}(y_t-\mu)'\Sigma^{-1}(y_t-\mu) \right) \\ &\quad - \lambda\left[\log[\det(\Sigma)] + \operatorname{tr}\left(\Sigma^{-1}\Omega\right)\right] \\ &= -\log(2\pi) - (1+\lambda)\log[\det(\Sigma)] - \frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)'\Sigma^{-1}(y_t-\mu) - \lambda\cdot\operatorname{tr}\left(\Sigma^{-1}\Omega\right). \end{aligned} \tag{3.9}$$

We have

$$(y_t-\mu)'\Sigma^{-1}(y_t-\mu) = y_t'\Sigma^{-1}y_t - y_t'\Sigma^{-1}\mu - \mu'\Sigma^{-1}y_t + \mu'\Sigma^{-1}\mu = y_t'\Sigma^{-1}y_t - 2y_t'\Sigma^{-1}\mu + \mu'\Sigma^{-1}\mu,$$

thus, we get

$$
\begin{aligned}
-\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)^{'}\Sigma^{-1}(y_t-\mu) &= -\frac{1}{T}\sum_{t=1}^{T}y_t{}^{'}\Sigma^{-1}y_t+\frac{2}{T}\sum_{t=1}^{T}y_t{}^{'}\Sigma^{-1}\mu-\frac{1}{T}\sum_{t=1}^{T}\mu^{'}\Sigma^{-1}\mu \\
&= -\frac{1}{T}\sum_{t=1}^{T}y_t{}^{'}\Sigma^{-1}y_t+2\left(\bar{y}\right)^{'}\Sigma^{-1}\mu-\mu^{'}\Sigma^{-1}\mu \\
&= -\frac{1}{T}\sum_{t=1}^{T}y_t{}^{'}\Sigma^{-1}y_t+2\left(\Sigma^{-1}\bar{y}\right)^{'}\mu-\mu^{'}\Sigma^{-1}\mu, \qquad (3.10)
\end{aligned}
$$

where $\bar{y}=\frac{1}{T}\sum_{t=1}^{T}y_t$. In the last relation we have used the fact that Σ^{-1} is symmetric.

Setting the partial derivative of $p_{\boldsymbol{y}}(\mu,\Sigma)$, with respect to μ, equal to the zero vector $\mathbf{0}_p$ and utilising Equations (3.10), (A.2) and (A.3), yields

$$
2\left(\Sigma^{-1}\bar{y}\right)^{'}-\mu^{'}\left[\Sigma^{-1}+\left(\Sigma^{-1}\right)^{'}\right]=\mathbf{0}_p,
$$

which results to

$$
2\mu^{'}\Sigma^{-1}=2\left(\bar{y}\right)^{'}\Sigma^{-1}\Leftrightarrow\mu^{'}\Sigma^{-1}\Sigma=\left(\bar{y}\right)^{'}\Sigma^{-1}\Sigma\Leftrightarrow\mu^{'}\boldsymbol{I}_p=\left(\bar{y}\right)^{'}\boldsymbol{I}_p\Leftrightarrow\mu=\bar{y}.
$$

Therefore, the estimator of μ, that maximises the penalised log-likelihood, is

$$
\hat{\mu}=\bar{y}.
$$

Setting the partial derivative of $p_{\boldsymbol{y}}(\mu,\Sigma)$, with repsect to Σ, equal to the zero matrix $\mathbf{0}_{p\times p}$, by (3.9) and (A.4) to (A.6) and using the fact $\mathrm{tr}\left(\Sigma^{-1}\Omega\right)=\mathrm{tr}\left(\boldsymbol{I}_p\Sigma^{-1}\Omega\right)$, we get

$$
-\left(1+\lambda\right)\Sigma^{-1}+\frac{1}{T}\Sigma^{-1}\sum_{t=1}^{T}\left(y_t-\bar{y}\right)\left(y_t-\bar{y}\right)^{'}\left(\Sigma^{-1}\right)^{'}+\lambda\left(\Sigma^{-1}\boldsymbol{I}_p\Omega\Sigma^{-1}\right)^{'}=\mathbf{0}_{p\times p}. \qquad (3.11)
$$

Let us denote by $\boldsymbol{S}$ the biased sample estimator of Σ

$$
\boldsymbol{S}=\frac{1}{T}\sum_{t=1}^{T}\left(y_t-\bar{y}\right)\left(y_t-\bar{y}\right)^{'},
$$

(see Definition A.42 in Appendix A and compare with Equation (2.58)). Then, as Σ^{-1} is symmetric, (3.11) becomes

$$
-\left(1+\lambda\right)\Sigma^{-1}+\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}+\lambda\Sigma^{-1}\Omega\Sigma^{-1}=\mathbf{0}_{p\times p}. \qquad (3.12)
$$

Multiplying both sides of (3.12) by Σ, both from the left and from the right, results in

$$-(1+\lambda)\,\Sigma\Sigma^{-1}\Sigma+\Sigma\Sigma^{-1}\boldsymbol{S}\Sigma^{-1}\Sigma+\lambda\Sigma\Sigma^{-1}\Omega\Sigma^{-1}\Sigma=\mathbf{0}_{p\times p}\Leftrightarrow\ -(1+\lambda)\,\Sigma+\boldsymbol{S}+\lambda\Omega=\mathbf{0}_{p\times p}.$$

Consequently, the maximum penalised log-likelihood estimator of Σ, is given by

$$\Sigma^{s}=\frac{1}{1+\lambda}\boldsymbol{S}+\frac{\lambda}{1+\lambda}\Omega. \tag{3.13}$$

□

Setting $w=\dfrac{\lambda}{1+\lambda}\in[0,1]$, Equation (3.13) is written as

$$\Sigma^{s}=(1-w)\,\boldsymbol{S}+w\Omega.$$

Setting

(i) $\Omega=\nu\boldsymbol{I}_p$, with $\nu=\langle\Sigma\,,\,\boldsymbol{I}_p\rangle$,

(ii) $w=\dfrac{\beta^2}{\delta^2}$, where $\beta^2=\mathrm{E}\left(||\boldsymbol{S}-\Sigma||_{\mathrm{F}}^2\right)$ and $\delta^2=\mathrm{E}\left(\left|\left|\boldsymbol{S}-\nu\boldsymbol{I}_p\right|\right|_{\mathrm{F}}^2\right)$,

yields the optimal linear shrinkage estimator proposed by Ledoit and Wolf (2004),

$$\Sigma^{*}=\frac{\alpha^2}{\delta^2}\boldsymbol{S}+\frac{\beta^2}{\delta^2}\nu\boldsymbol{I}_p,$$

where $\alpha^2=\left||\Sigma-\nu\boldsymbol{I}_p\right||_{\mathrm{F}}^2$.

Then a consistent estimator of the optimal weight w is

$$\widehat{w}=\frac{b^2}{d^2},$$

where

$$d^2=\left||\boldsymbol{S}-\hat{\nu}\boldsymbol{I}_p\right||_{\mathrm{F}}^2,$$

$$\hat{\nu}=\left\langle\boldsymbol{S}\,,\,\boldsymbol{I}_p\right\rangle,$$

and

$$b^2=\min\left\{\frac{1}{T}\sum_{t=1}^{T}\left|\left|\bar{y}\bar{y}'-\boldsymbol{S}\right|\right|_{\mathrm{F}}^2,\ d^2\right\},$$

which gives the following consistent estimator of Σ^{*},

$$\widehat{\Sigma}^{*}=\frac{\mathsf{a}^2}{d^2}\boldsymbol{S}+\frac{b^2}{d^2}\hat{\nu}\boldsymbol{I}_p,$$

where $\mathsf{a}^2 = d^2 - b^2$.

Since $\lambda = \dfrac{w}{1-w}$, it becomes evident that to derive the shrinkage estimator Σ^*, the regularisation parameter λ should be chosen by taking into account the underlying distribution of the data.

3.2 Shrinkage under Regime Switching

As we saw in Subsection 2.3.2, an important part in the implementation of the EM algorithm for the hidden Markov model given by Equation (2.45), was the complete-data log-likelihood

$$\ell_{\boldsymbol{y},\boldsymbol{X}}[(\boldsymbol{\pi},\boldsymbol{P},\boldsymbol{\phi})] = \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{1}_{\{X_t=k\}}f_k(y_t;\phi_k) + \ell_{\boldsymbol{X}}[(\boldsymbol{\pi},\boldsymbol{P})]\,.$$

Assuming that an oracle reveals the hidden states $(X_1,\ldots,X_T)$ leads to the oracle estimates μ_k^{o} and $\tilde{\Sigma}_k^{\mathrm{o}}$ given by Equations (2.55) and (2.57), respectively. To get well-conditioned estimates for the covariance matrices, following the framework of Section 3.1, we incorporate K different penalty terms, one for each state, into the complete-data log-likelihood, to get a penalised complete-data log-likelihood, defined as

$$p_{\boldsymbol{y},\boldsymbol{X}}[(\boldsymbol{\pi},\boldsymbol{P},\boldsymbol{\phi})] = \frac{2}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{1}_{\{X_t=k\}}f_k(y_t;\phi_k) - \sum_{k=1}^{K}\lambda_k J_k(\Sigma_k) + \frac{2}{T}\ell_{\boldsymbol{X}}[(\boldsymbol{\pi},\boldsymbol{P})]\,, \tag{3.14}$$

where $J_k(\Sigma_k) = \log[\det(\Sigma_k)] + \tilde{\nu}_k\mathrm{tr}\left(\Sigma_k^{-1}\right)$ and $\tilde{\nu}_k = \dfrac{1}{p}\mathrm{tr}(\Sigma_k)$, where Σ_k denotes the true covariance matrix under state k, for all $k \in \{1,\ldots,K\}$. Following the frameworks of Subsection 2.3.2 and Section 3.1, the maximum penalised log-likelihood estimators of the covariance matrices are the shrunken versions of the oracle estimators $\tilde{\Sigma}_k^{\mathrm{o}}$, that is

$$\tilde{\Sigma}_k^{s} = (1-w_k)\,\tilde{\Sigma}_k^{\mathrm{o}} + w_k\tilde{\nu}_k\boldsymbol{I}_p, \tag{3.15}$$

where $w_k = \dfrac{\lambda_k}{1+\lambda_k} \in [0,1]$, for all $k \in \{1,\ldots,K\}$. As has already been mentioned in Subsection 2.3.2, Fiecas et al. (2017) chose to work with $\Sigma_k^{\mathrm{o}} = \pi_k^{\mathrm{o}}\tilde{\Sigma}_k^{\mathrm{o}}$ (see Equation (2.58)). To derive the corresponding shrinkage estimators, the scaling factors $\tilde{\nu}_k$, $k \in \{1,\ldots,K\}$, need to be adapted accordingly. Using the optimal weights of Ledoit and Wolf (2004), we get the shrinkage estimators of the covariance matrices, based on Σ_k^{o}, as

$$\Sigma_k^{s} = (1-w_k)\,\Sigma_k^{\mathrm{o}} + w_k\nu_k\boldsymbol{I}_p \qquad (k \in \{1,\ldots,K\})\,, \tag{3.16}$$

where $\nu_k = \dfrac{1}{p}\mathrm{tr}(\pi_k\Sigma_k)$, $k \in \{1,\ldots,K\}$.

Σ_k^s is not a bona fide estimator, as w_k and ν_k are unknown and need to be estimated from the data. To do so, Fiecas et al. (2017) use the framework of Sancetta (2008), who extended the work of Ledoit and Wolf (2004) for data that exhibit time-series dependence. As the author points out in his paper (see the Introduction and the remark on Condition 1 in Subsection 2.2), his results are weaker than the ones of Ledoit and Wolf (2004), but they cover more cases, therefore the two papers should be viewed as complementary. Sancetta's shrinkage estimator is of the form

$$\Sigma^s = (1-\alpha)\, \boldsymbol{S} + \alpha \boldsymbol{F},$$

where $\alpha \in [0,1]$, $\boldsymbol{S}$ is the sample covariance matrix and $\boldsymbol{F}$ is a constrained version of the true covariance matrix Σ. The shrinkage target $\boldsymbol{F}$ is chosen appropriately in order to impose strict conditions on Σ and it is unknown, so it needs to be estimated.

The first step to determining an estimator for Σ_k^s, is to replace the scaling factor ν_k, which relies on the unknown covariance matrix Σ_k and the unknown stationary probability π_k, with an estimate. Fiecas et al. (2017) show, using Lemma 3.9, that a proper choice is $\nu_k^o = \frac{1}{p}\mathrm{tr}(\Sigma_k^o)$ (for the proof of Lemma 3.9 see Appendix C).

Lemma 3.9. *Let the data be generated from model (2.45), either for fixed p, or for $p \to +\infty$ with increasing T, satisfying* Assumptions (A) *(see Subsection 2.3.1). Consider the following assumptions:*

(B1) $\max_{1\le i\le p}\{|\mu_{k,i}|\}$ *and* $\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}}^2 = p^{-1}\mathrm{tr}(\Sigma_k)$ *are uniformly bounded in p,*

(B2) $\pi_k > 0$, $\mathrm{tr}(\Sigma_k) \ge cp^r$, *for some $c > 0$, $0 \le r \le 1$ and all p.*

It follows for some $0 < \beta < 1$ and all $k \in \{1,\dots,K\}$,

(i)
$$\mathrm{E}(\mu_k^o) = \mu_k + \mu_k \mathrm{O}\left(\beta^T\right) = \mu_k + \mathrm{O}\left(p^{1/2}\beta^T\right),$$
$$\mathrm{E}\left(\pi_k^o\, ||\mu_k^o - \mu_k||_2^2\right) = T^{-1}\pi_k \mathrm{tr}(\Sigma_k) = \mathrm{O}(p/T),$$

and

$$\mathrm{E}(\Sigma_k^o) = (\pi_k - 1/T)\,\Sigma_k + \mathrm{O}\left(\beta^T/T\right) = \pi_k\Sigma_k + \mathrm{O}(\sqrt{p}/T).$$

(ii) *If* (B1) *holds, then the normalised trace estimator ν_k^o is mean-square consistent, that is,*

$$\mathrm{E}\left[(\nu_k^o - \nu_k)^2\right] = \mathrm{O}\left(T^{-1}\right).$$

Additionally,

$$\mathrm{E}(\nu_k^o) = \nu_k + \mathrm{O}\left(T^{-1}\right),$$

where

$$\nu_k = \frac{1}{p}\text{tr}(\pi_k \Sigma_k)\,, \tag{3.17}$$

and

$$\nu_k^{\text{o}} = \frac{1}{p}\text{tr}(\Sigma_k^{\text{o}})\,. \tag{3.18}$$

(iii) *If both* (B1) *and* (B2) *hold, then*

$$\frac{\text{Var}\left[\text{tr}\left(\Sigma_k^{\text{o}}\right)\right]}{\left|\text{tr}(\pi_k \Sigma_k)\right|^2} = \text{O}\left(\frac{p^{2(1-r)}}{T}\right),$$

that is, the relative error of $\text{tr}\left(\Sigma_k^{\text{o}}\right)$ *as an estimate of* $\text{tr}(\pi_k \Sigma_k)$ *converges to zero if* $p^{1-r} = \text{o}\left(\sqrt{T}\right)$.

We present some comments on Assumptions (B1) and (B2) that can be found in the discussion following Lemma 1 in Fiecas et al. (2017).

If p is fixed and the true covariance matrix Σ_k is nonsingular (all its eigenvalues are different than zero), Assumptions (B1) and (B2) are automatically satisfied. If the largest eigenvalue of Σ_k, say $\lambda_{\max}(\Sigma_k)$, is uniformly bounded in p, then the second part of Assumption (B1) is automatically satisfied, as $\text{tr}(\Sigma_k) \leq p\lambda_{\max}(\Sigma_k)$. However, this is not a necessary condition. Even if $\lambda_{\max}(\Sigma_k)$ is growing with a rate up to p, Assumption (B1) is still satisfied, as long as all the eigenvalues of Σ_k, except a finite number of them, are uniformly bounded in p. In case the smallest eigenvalue of Σ_k, say $\lambda_{\min}(\Sigma_k)$, satisfies the inequality $\lambda_{\min}(\Sigma_k) \geq cp^{r-1} = \frac{c}{p^{1-r}}$, which means that it does not converge too fast to zero, as the dimension p increases, then Assumption (B2) is automatically satisfied, since $\text{tr}(\Sigma_k) \geq p\lambda_{\min}(\Sigma_k)$. However, Assumption (B2) includes the case of Σ_k being singular, that is at least one of its eigenvalues equals zero, which implies that $\lambda_{\min}(\Sigma_k) = 0$.

For the optimal weights we use the formula given by Proposition 1 of Sancetta (2008), setting $\boldsymbol{F} = \nu_k \boldsymbol{I}_p$. It should be noted that Sancetta (2008) uses the unscaled Frobenius norm, that is $||A|| = \sqrt{\text{tr}(AA')}$, where $A \in \mathbb{R}^{p\times p}$.

Lemma 3.10. *Consider the shrinkage estimators* $\Sigma_k^s = (1 - w_k)\,\Sigma_k^{\text{o}} + w_k \nu_k \boldsymbol{I}_p$, $k \in \{1, \ldots, K\}$, *where* Σ_k^{o} *and* ν_k *are given by* (2.58) *and* (3.17), *respectively. The optimal weights* w_k, $k \in \{1, \ldots, K\}$, *which minimise the risk*

$$\min_{w_k \in [0,1]} \left\{ \text{E}\left(\left|\left| (1 - w_k)\,\Sigma_k^{\text{o}} + w_k \nu_k \boldsymbol{I}_p - \pi_k \Sigma_k \right|\right|_{\text{F}}^2 \right) \right\},$$

are given by

$$w_k = \min\left\{ \frac{\text{E}\left(\left|\left| \Sigma_k^{\text{o}} - \pi_k \Sigma_k \right|\right|_{\text{F}}^2 \right)}{\text{E}\left(\left|\left| \Sigma_k^{\text{o}} - \nu_k \boldsymbol{I}_p \right|\right|_{\text{F}}^2 \right)}, 1 \right\},$$

assuming that all the relevant moments exist.

Proof. See the proof of Proposition 1 in Section 4 in Sancetta (2008). □

As the optimal weight of Lemma 3.10 depends on the true stationary probability π_k and the true covariance matrix Σ_k, the optimal shrinkage estimator Σ_k^s is not a bona fide estimator. To estimate w_k, we need to find consistent estimators for both the nominator and the denominator. For the denominator, ν_k is substituted for its oracle estimate ν_k^o (see Equation (3.18) in Lemma 3.9). This follows the framework of Ledoit and Wolf (2004), who replace the expected loss in the denominator of the optimal weight with its sample counterpart, by substituting the scalar quantity μ for its sample estimator m (see Definition 3.3, Lemmas 3.4 and 3.5 and Theorems 3.1 and 3.2).

To determine an estimator for the nominator, we are going to follow the framework of Sancetta (2008).

By Statement (ii) of Lemma C.3 and Equation (C.79) (in the proof of Lemma C.3), we get

$$\mathrm{E}\left(||\Sigma_k^o - \pi_k\Sigma_k||_F^2\right) = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{X_t=k\}}\left(Y_{t,i}-\mu_{k,i}\right)\left(Y_{t,j}-\mu_{k,j}\right)\right] + \mathrm{O}\left(\frac{p}{T^2}\right). \tag{3.19}$$

Setting

$$\boldsymbol{Z}_{k,ij}(t) = \mathbf{1}_{\{X_t=k\}}\left(Y_{t,i}-\mu_{k,i}\right)\left(Y_{t,j}-\mu_{k,j}\right),$$

Equation (3.19) can be written as

$$\mathrm{E}\left(||\Sigma_k^o - \pi_k\Sigma_k||_F^2\right) = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_{k,ij}(t)\right) + \mathrm{O}\left(\frac{p}{T^2}\right).$$

This time series is stationary with exponentially decreasing autocovariances $c_{k,ij}(s)$ and a continuous spectral density $f_{k,ij}(\omega)$ (see Lemma C.1 in Appendix C). We define the corresponding sample autocovariances[6] as

$$\bar{c}_{k,ij}^o(s) = \frac{1}{T}\sum_{t=1}^{T-s}\left(\boldsymbol{Z}_{k,ij}(t) - \frac{1}{T}\sum_{\tau=1}^{T}\boldsymbol{Z}_{k,ij}(\tau)\right)\left(\boldsymbol{Z}_{k,ij}(t+s) - \frac{1}{T}\sum_{\tau=1}^{T}\boldsymbol{Z}_{k,ij}(\tau)\right) \qquad (s\in\{0,\dots,T-1\}),$$

where $\bar{c}_{k,ij}^o(s) = \bar{c}_{k,ij}^o(-s)$. As in the proof of Lemma 3.9 (see Subsection C.2.3) for the variance of π_k^o, by Remark 1 to Theorem 7.1.1 in Brockwell and Davis (1991), we get

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_{k,ij}(t)\right) \to \frac{1}{T}f_{k,ij}(0),\ T\to+\infty \qquad (k\in\{1,\dots,K\},\ i,j\in\{1,\dots,p\}).$$

[6] See Definition A.49 in Appendix A.

As in Sancetta (2008), the spectral densities at frequency zero,

$$f_{k,ij}(0) = \sum_{n=-\infty}^{+\infty} \bar{c}^{\mathrm{o}}_{k,ij}(n) \qquad (k \in \{1,\dots,K\}\,,\ i,j \in \{1,\dots,p\})\,,$$

are going to be estimated via kernel smoothing[7]. We consider a kernel $K(\cdot)$, with the following properties:

(i) $K(u) \geq 0$, for all $u \in \mathbb{R}$,

(ii) $K(u) = K(-u)$, for all $u \in \mathbb{R}$,

(iii) $K(0) = 1$.

For some bandwidth (or smoothing parameter) $b > 0$, the estimators of the spectral densities at frequency zero are given by

$$\bar{f}^{b}_{k,ij}(0) = \sum_{s=-T+1}^{T-1} K\Big(\frac{s}{b}\Big)\, \bar{c}^{\mathrm{o}}_{k,ij}(s) \qquad (k \in \{1,\dots,K\}\,,\ i,j \in \{1,\dots,p\})\,.$$

Using the same kind of arguments that were used to prove Lemma C.3, replacing the unknown means μ_k by their oracle estimates μ^{o}_k, asymptotically, has no significant impact. Therefore, we can utilise the following bona fide estimators, defined for all $t \in \{1,\dots,T\}$, $s \in \{0,\dots,T-1\}$, $k \in \{1,\dots,K\}$ and $i,j \in \{1,\dots,p\}$, as:

$$\widehat{\mathbf{Z}}_{k,ij}(t) = \mathbf{1}_{\{X_t=k\}}\left(Y_{t,i} - \mu^{\mathrm{o}}_{k,i}\right)\left(Y_{t,j} - \mu^{\mathrm{o}}_{k,j}\right),$$

$$\hat{c}^{\mathrm{o}}_{k,ij}(s) = \frac{1}{T}\sum_{t=1}^{T-s}\left(\widehat{\mathbf{Z}}_{k,ij}(t) - \frac{1}{T}\sum_{\tau=1}^{T}\widehat{\mathbf{Z}}_{k,ij}(\tau)\right)\left(\widehat{\mathbf{Z}}_{k,ij}(t+s) - \frac{1}{T}\sum_{\tau=1}^{T}\widehat{\mathbf{Z}}_{k,ij}(\tau)\right),$$

and

$$\hat{f}^{b}_{k,ij}(0) = \sum_{s=-T+1}^{T-1} K\Big(\frac{s}{b}\Big)\, \hat{c}^{\mathrm{o}}_{k,ij}(s).$$

Therefore, we get the following oracle estimators of the optimal weights

$$w^{\mathrm{o}}_k = \min\left\{\frac{(pT)^{-1}\sum_{i=1}^{p}\sum_{j=1}^{p}\hat{f}^{b}_{k,ij}(0)}{\left|\left|\Sigma^{\mathrm{o}}_k - \nu^{\mathrm{o}}_k \boldsymbol{I}_p\right|\right|^2_{\mathrm{F}}}, 1\right\} \qquad (k \in \{1,\dots,K\})\,, \tag{3.20}$$

[7] *Kernel density estimation* or *kernel smoothing* is a technique for nonparametric estimation of density functions (probability or spectral). In this method, some functions with specific properties, called *kernels*, and a *bandwidth* or *smoothing parameter* are utilised, hence the name (see, e.g., Andrews 1991, or Wand and Jones 1994).

which yields the bona fide optimal shrinkage estimator

$$\widehat{\Sigma}_k^s = (1 - w_k^o)\,\Sigma_k^o + w_k^o \nu_k^o \boldsymbol{I}_p \qquad (k \in \{1, \ldots, K\})\,. \tag{3.21}$$

Notice that Equation (3.20) is the exact same formula as the one for $\hat{\alpha}_T$ given after Equation (4) in Sancetta (2008). The term p^{-1} in the nominator of w_k^o is justified by the fact that we use the scaled Frobenius norm, while, as has already been mentioned, Sancetta (2008) uses the unscaled version.

Theorem 3.3. *Under* Assumptions (A1), (A2) *and* (A3) *on the Markov chain* $\{X_t : t \in \mathbb{Z}\}$ *and the random vectors* ϵ_t, $t \in \mathbb{Z}$, *and under* Assumption (B1) *of* Lemma 3.9, *let the kernel* $K(u)$ *be continuous, symmetric, nonnegative and, for* $u > 0$, *decreasing with* $K(0) = 1$ *and* $\int_0^{+\infty} [K(u)]^2\, du < +\infty$. *Additionally, choose the bandwidth* $b = b_T \to +\infty$, *such that* $b_T / \sqrt{T} \to 0$, *as* $T \to +\infty$. *Moreover, assume either:*

(C1) p *fixed,* $\nu_k \boldsymbol{I}_p \neq \pi_k \Sigma_k$, *or,*

(C2) $p \to +\infty$, $p^{1-\gamma} \left|\left|\nu_k \boldsymbol{I}_p - \pi_k \Sigma_k\right|\right|_{\mathrm{F}}^2 \to c > 0$ *for some* $\gamma \in (0, 2]$, *such that* $p^{2-\gamma}/T \to 0$.

Then, with $r_T = T$ *in case* (C1), *or* $r_T = T/p^{2-\gamma}$ *in case* (C2), *we have the following results:*

(i) $w_k^o \asymp \frac{1}{r_T}$,

(ii) $r_T \left(w_k^o - w_k\right) = \mathrm{o}_p(1)$,

(iii) $\left|\left|\widehat{\Sigma}_k^s - \pi_k \Sigma_k\right|\right|_{\mathrm{F}} = \left|\left|\Sigma_k^s - \pi_k \Sigma_k\right|\right|_{\mathrm{F}} \left[1 + \mathrm{o}_p(1/\sqrt{r_t})\right]$.

Before we proceed to show how the methodology, presented in this section, is incorporated into the EM algorithm, we make some comments on Theorem 3.3, that can be found in Subsections 2.1 and 2.1 in Sancetta (2008) and in the discussion following Theorem 1 in Fiecas et al. (2017).

Assumptions (C1) and (C2) come from classical assumptions for shrinkage (see Condition 1 (3) in Sancetta 2008). The target matrix needs to be quite different than the true parameter $\pi_k \Sigma_k$, so that the number of the elements of the shrinkage target to be estimated will be relatively small (see the paragraph following Example 3 in Subsection 2.2 in Sancetta 2008). Also, if the shrinkage target $\nu_k \boldsymbol{I}_p$ were equal to the true value of the parameter $\pi_k \Sigma_k$, there would be no point of using shrinkage. The parameter γ quantifies how different the target matrix $\nu_k \boldsymbol{I}_p$ is from $\pi_k \Sigma_k$, under the Frobenius norm (see the paragraph following Example 2 in Subsection 2.2 in Sancetta 2008, keeping in mind that the Frobenius norm used by Sancetta 2008 is not scaled).

Theorem 3.3 assures that the optimal weight w_k and its oracle estimate w_k^o are, asymptotically, equivalent and that the error of the bona fide estimator $\widehat{\Sigma}_k^s$, is in probability asymptotically the same as the error of the non bona fide shrinkage estimator Σ_k^s. This means that using $\widehat{\Sigma}_k^s$ as an estimator

of $\pi_k\Sigma_k$, is in probability asymptotically, equivalent to using Σ_k^s. The rate of convergence of the error of $\widehat{\Sigma}_k^s$, as an estimator of $\pi_k\Sigma_k$, in the corresponding error of Σ_k^s is also provided. Consistency of the estimator $\widehat{\Sigma}_k^s$ is not guaranteed by Theorem 3.3, as it is not the main concern. What is of interest, is that $\widehat{\Sigma}_k^s$ is a better estimator than Σ_k^o under the Frobenius norm (see Subsection 2.1 in Sancetta 2008).

3.3 Incorporating Shrinkage into the EM Algorithm

As has been discussed throughout this chapter, the sample covariance matrix performs poorly in situations where the sample size T is relatively small and the data dimension p is high. As we saw in Subsection 2.3.2, the estimator $\tilde{\Sigma}_k^o$, which is the sample covariance matrix for the hidden Markov model given by Equation (2.45), under state k, $k \in \{1, \ldots, K\}$, faces the same problems. Even in cases where the dimension p is low, if there are some states for which there is only a small number of observations, then the corresponding matrices might be numerically ill-conditioned, or even singular.

To overcome this problem Fiecas et al. (2017) propose to incorporate the shrinkage estimator of Section 3.2 into the EM algorithm. The first step is to substitute the complete-data log-likelihood $\ell_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta})$, in the function $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}')$, for its penalised counterpart, given by Equation (3.14). Therefore, we define

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') &= \mathrm{E}_{\boldsymbol{\theta}'}\left(p_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) = \mathrm{E}\left(\frac{2}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{1}_{\{X_t=k\}}f_k(y_t;\boldsymbol{\theta}) - \sum_{k=1}^{K}\lambda_k J_k(\Sigma_k) + \frac{2}{T}\ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) \\
&= \frac{2}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left(\mathbf{1}_{\{X_t=k\}} \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) f_k(y_t;\boldsymbol{\theta}) - \sum_{k=1}^{K}\lambda_k J_k(\Sigma_k) + \mathrm{E}\left(\frac{2}{T}\ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right) \\
&= \frac{2}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\gamma_k(t)\, f_k(y_t;\boldsymbol{\theta}) - \sum_{k=1}^{K}\lambda_k J_k(\Sigma_k) + \mathrm{E}\left(\frac{2}{T}\ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right). \qquad (3.22)
\end{aligned}$$

The E-step is carried out as was described in Paragraph 2.3.2.1 with $\mathcal{Q}\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}^{(m)}\right)$ as defined in Equation (3.22). It should be noted that the penalty terms $J_k(\Sigma_k)$ are independent of $\boldsymbol{\theta}^{(m)}$, except for the scaling factors $\nu_k = p^{-1}\mathrm{tr}(\Sigma_k)$, which are linear in the corresponding covariance matrices. As was shown in Paragraph 2.3.2.1, in the E-step, the conditional expectations, of the random variables $\mathbf{1}_{\{X_t=k\}}$ and $\mathbf{1}_{\{X_t=k,X_{t+1}=l\}}$, given the observations $(y_1, \ldots, y_T)$, $\gamma_k(t)$ and $\xi_{k,l}(t)$, respectively, are obtained.

In the M-step, as was presented in Paragraph 2.3.2.2, we derived the maximum likelihood estimates based on the formulas of the oracle estimates, where the random variables $\mathbf{1}_{\{X_t=k\}}$ were replaced by their conditional expectations $\gamma_k(t)$ from the E-step. Working in a similar fashion, now the maximum penalised likelihood estimates are going to be utilised, where a data-adaptive regularisation parameter has been employed, in order to obtain them.

As was shown in Section 3.1, the penalty term has an effect only on the formula of the estimator of the covariance matrix, therefore the updating formulas for the transition probabilities p_{kl} and the means μ_k are the same as the ones that were given by Equations (2.61) and (2.62), respectively. In Subsection 2.3.2, we started from the maximiser of the complete-data log-likelihood $\tilde{\Sigma}_k^{\mathrm{o}}$ and using the sample estimate of the stationary probability π_k, we derived the biased oracle estimator of the covariance matrix $\Sigma_k^{\mathrm{o}} = \pi_k^{\mathrm{o}} \tilde{\Sigma}_k^{\mathrm{o}}$. Then, we substituted the random variables $\mathbf{1}_{\{X_t=k\}}$ for their conditional expectations $\gamma_k(t)$, derived from the E-step, to obtain the updating formulas for the covariance matrices (see Equations (2.63) and (2.64)). Following that framework, in Section 3.2 we started from the shrinkage estimator $\tilde{\Sigma}_k^s$ which maximises the penalised complete-data log-likelihood. Then, we obtained Σ_k^s by replacing $\tilde{\Sigma}_k^{\mathrm{o}}$ and $\tilde{\nu}_k$ with Σ_k^{o} and ν_k, respectively, in the definition of $\tilde{\Sigma}_k^s$ (see Equations (3.15) and (3.16)). Finally, by estimating the unknown quantities in the definition of Σ_k^s, we obtained the bona fide optimal shrinkage estimator $\widehat{\Sigma}_k^s$ given by Equation (3.21). Same as before, replacing the random variables $\mathbf{1}_{\{X_t=k\}}$ with their conditional expectations $\gamma_k(t)$, yields the estimators for the covariance matrices given by

$$\widehat{\Sigma}_k = \frac{1}{\hat{\pi}_k} \widehat{\Sigma}_k^s = \frac{1}{\hat{\pi}_k} \left[(1 - \widehat{w}_k^{\mathrm{o}}) \, \widehat{\Sigma}_k^{\mathrm{o}} + \widehat{w}_k^{\mathrm{o}} \hat{v}_k^{\mathrm{o}} \boldsymbol{I}_p \right] \qquad (k \in \{1, \ldots, K\}), \tag{3.23}$$

where

$$\widehat{\Sigma}_k^{\mathrm{o}} = \frac{1}{T} \sum_{t=1}^{T} \gamma_k(t) \, (y_t - \hat{\mu}_k) \, (y_t - \hat{\mu}_k)' \qquad (k \in \{1, \ldots, K\}),$$

$$\hat{v}_k^{\mathrm{o}} = \frac{1}{p} \mathrm{tr}\left(\widehat{\Sigma}_k^{\mathrm{o}} \right) \qquad (k \in \{1, \ldots, K\}),$$

and $\widehat{w}_k^{\mathrm{o}}$ defined as w_k^{o} given by Equation (3.20), where the random variables $\mathbf{1}_{\{X_t=k\}}$, again, are replaced by their conditional expectation $\gamma_k(t)$ in all the quantities that are involved in the definition of w_k^{o}. That is, defining $\widehat{\boldsymbol{Z}}_{k,ij}^*(t)$, $\hat{c}_{k,ij}^*(s)$ and $\hat{f}_{k,ij}^{b*}(0)$ as

$$\widehat{\boldsymbol{Z}}_{k,ij}^*(t) = \gamma_k(t) \, (y_{t,i} - \hat{\mu}_{k,i}) \, (y_{t,j} - \hat{\mu}_{k,j}),$$

$$\hat{c}_{k,ij}^*(s) = \frac{1}{T} \sum_{t=1}^{T-s} \left(\widehat{\boldsymbol{Z}}_{k,ij}^*(t) - \frac{1}{T} \sum_{\tau=1}^{T} \widehat{\boldsymbol{Z}}_{k,ij}^*(\tau) \right) \left(\widehat{\boldsymbol{Z}}_{k,ij}^*(t+s) - \frac{1}{T} \sum_{\tau=1}^{T} \widehat{\boldsymbol{Z}}_{k,ij}^*(\tau) \right),$$

and

$$\hat{f}_{k,ij}^{b*}(0) = \sum_{s=-T+1}^{T-1} K\left(\frac{s}{b} \right) \hat{c}_{k,ij}^*(s),$$

then $\widehat{w}_k^{\mathrm{o}}$ is given by

$$\widehat{w}_k^{\mathrm{o}} = \min\left\{ \frac{\dfrac{1}{pT}\sum_{i=1}^{p}\sum_{j=1}^{p}\hat{f}_{k,ij}^{b*}(0)}{\left|\left|\widehat{\Sigma}_k^{\mathrm{o}} - \hat{v}_k^{\mathrm{o}}\boldsymbol{I}_p\right|\right|_{\mathrm{F}}^2}, 1 \right\} \qquad (k \in \{1,\ldots,K\})\,.$$

Finally, let us summarise the process of the modified EM algorithm.

Modified EM Algorithm for the Data-Generating Model

---

(1) Initialise with a proper value $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\pi}^{(0)}, \boldsymbol{P}^{(0)}, \boldsymbol{\phi}^{(0)}\right)$ of the parameter.

(2) At the m-th iteration proceed as follows:

(2.1) **E-Step**: Compute the modified $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right)$ defined by Equation (3.22) as a function of $\boldsymbol{\theta}$, in the following manner:

(2.1.1) For all $k \in \{1,\ldots,K\}$, compute the forward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.1.1) $\alpha_k^{(m)}(1) = \pi_k^{(m)} f_k\left(y_1; \phi_k^{(m)}\right)$.

(2.1.1.2) $\alpha_k^{(m)}(t) = \left(\sum_{i=1}^{K} \alpha_i^{(m)}(t-1)\, p_{ik}^{(m)}\right) f_k\left(y_t; \phi_k^{(m)}\right)$, for all $t \in \{2,\ldots,T\}$.

(2.1.2) For all $k \in \{1,\ldots,K\}$, compute the backward variables under the value $\boldsymbol{\theta}^{(m)}$ as:

(2.1.2.1) $\beta_k^{(m)}(T) = 1$.

(2.1.2.2) $\beta_k^{(m)}(t) = \sum_{l=1}^{K} p_{kl}^{(m)} f_l\left(y_{t+1}; \phi_l^{(m)}\right) \beta_l^{(m)}(t+1)$, for all $t \in \{T-1,\ldots,1\}$.

(2.1.3) For all $t \in \{1,\ldots,T\}$ and $k \in \{1,\ldots,K\}$, compute the variables:

$$\gamma_k^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\,\beta_k^{(m)}(t)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\,\beta_i^{(m)}(t)}.$$

(2.1.4) For all $t \in \{1,\ldots,T-1\}$ and $k, l \in \{1,\ldots,K\}$, compute the variables:

$$\xi_{kl}^{(m)}(t) = \frac{\alpha_k^{(m)}(t)\, p_{kl}^{(m)} f_l\left(y_{t+1}; \phi_l^{(m)}\right) \beta_l^{(m)}(t+1)}{\sum_{i=1}^{K} \alpha_i^{(m)}(t)\,\beta_i^{(m)}(t)}.$$

(2.1.5) Calculate the conditional expectation of the penalised complete-data log-likelihood, given the observations $(y_1, \ldots, y_T)$ for the current values $\boldsymbol{\theta}^{(m)}$, $\lambda_k^{(m)}$ and $J_k\left(\widehat{\Sigma}_k^{(m)}\right)$ as:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(p_{\boldsymbol{y},\boldsymbol{X}}(\boldsymbol{\pi},\boldsymbol{P},\boldsymbol{\phi}) \,\middle|\, y_1,\ldots,y_T\right) = \frac{2}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\gamma_k^{(m)}(t)\, f_k\left(y_t;\phi_k^{(m)}\right) - \sum_{k=1}^{K}\lambda_k^{(m)} J_k\left(\widehat{\Sigma}_k^{(m)}\right) \\ + \mathrm{E}_{\boldsymbol{\theta}^{(m)}}\left(\frac{2}{T}\ell_{\boldsymbol{X}}(\boldsymbol{\pi},\boldsymbol{P}) \,\middle|\, y_1,\ldots,y_T\right).\end{aligned}$$

where

$$\lambda_k^{(m)} = \frac{\widehat{w}_k^{\mathrm{o}(m)}}{1-\widehat{w}_k^{\mathrm{o}(m)}},$$

and

$$J_k\left(\widehat{\Sigma}_k^{(m)}\right) = \log\left[\det\left(\widehat{\Sigma}_k^{(m)}\right)\right] + \widehat{v}_k^{\mathrm{o}(m)}\mathrm{tr}\left[\left(\widehat{\Sigma}_k^{(m)}\right)^{-1}\right],$$

with

$$\widehat{v}_k^{\mathrm{o}(m)} = \frac{1}{p}\mathrm{tr}\left(\widehat{\Sigma}_k^{\mathrm{o}(m)}\right).$$

(2.2) **M-Step**: Using the current values $\gamma_k^{(m)}(t)$ and $\xi_{kl}^{(m)}(t)$, that were calculated in the E-step, determine the value $\boldsymbol{\theta}^{(m+1)} = \left(\boldsymbol{\pi}^{(m+1)}, \boldsymbol{P}^{(m+1)}, \boldsymbol{\phi}^{(m+1)}\right)$ that maximises $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$. For all $k, l \in \{1,\ldots,K\}$, the values of the initial and transition probabilities of the hidden Markov chain are updated as follows:

$$\begin{aligned}\hat{\pi}_k^{(m+1)} &= \frac{1}{T}\sum_{t=1}^{T}\gamma_k^{(m)}(t), \\ \hat{p}_{kl}^{(m+1)} &= \frac{\sum_{t=1}^{T-1}\xi_{kl}^{(m)}(t)}{\sum_{t=1}^{T-1}\gamma_k^{(m)}(t)}.\end{aligned}$$

For all $k \in \{1,\ldots,K\}$, the parameters of the state-dependent distributions $\boldsymbol{\phi}^{(m+1)} = \left(\phi_k^{(m+1)}\right)_{k\in S_X}$, with $\phi_k = (\mu_k, \Sigma_k)$, are updated as follows:

$$\begin{aligned}\hat{\mu}_k^{(m+1)} &= \frac{\sum_{t=1}^{T}\gamma_k^{(m)}(t)\, y_t}{\sum_{t=1}^{T}\gamma_k^{(m)}(t)}, \\ \widehat{\Sigma}_k^{(m+1)} &= \frac{1}{\hat{\pi}_k^{(m+1)}}\left[\left(1-\widehat{w}_k^{\mathrm{o}(m+1)}\right)\widehat{\Sigma}_k^{\mathrm{o}(m+1)} + \widehat{w}_k^{\mathrm{o}(m+1)}\widehat{v}_k^{\mathrm{o}(m+1)}\boldsymbol{I}_p\right],\end{aligned}$$

where

$$\widehat{\Sigma}_k^{\mathrm{o}(m+1)} = \frac{1}{T}\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t - \hat{\mu}_k^{(m+1)}\right)\left(y_t - \hat{\mu}_k^{(m+1)}\right)',$$

$$\hat{v}_k^{\mathrm{o}(m+1)} = \frac{1}{p}\mathrm{tr}\left(\widehat{\Sigma}_k^{\mathrm{o}(m+1)}\right),$$

$$\hat{w}_k^{\mathrm{o}(m+1)} = \min\left\{\frac{\frac{1}{pT}\sum_{i=1}^{p}\sum_{j=1}^{p}\hat{f}_{k,ij}^{b*(m+1)}(0)}{\left|\left|\widehat{\Sigma}_k^{\mathrm{o}(m+1)} - \hat{v}_k^{\mathrm{o}(m+1)}\boldsymbol{I}_p\right|\right|_{\mathrm{F}}^2}, 1\right\},$$

with

$$\hat{f}_{k,ij}^{b*(m+1)}(0) = \sum_{s=-T+1}^{T-1} K\left(\frac{s}{b}\right)\hat{c}_{k,ij}^{*(m+1)}(s),$$

for all $i, j \in \{1, \ldots, p\}$,

$$\hat{c}_{k,ij}^{*(m+1)}(s) = \frac{1}{T}\sum_{t=1}^{T-s}\left(\widehat{\boldsymbol{Z}}_{k,ij}^{*(m+1)}(t) - \frac{1}{T}\sum_{\tau=1}^{T}\widehat{\boldsymbol{Z}}_{k,ij}^{*(m+1)}(\tau)\right)\left(\widehat{\boldsymbol{Z}}_{k,ij}^{*(m+1)}(t+s) - \frac{1}{T}\sum_{\tau=1}^{T}\widehat{\boldsymbol{Z}}_{k,ij}^{*(m+1)}(\tau)\right),$$

for all $s \in \{0, \ldots, T-1\}$ and $i, j \in \{1, \ldots, p\}$, and

$$\widehat{\boldsymbol{Z}}_{k,ij}^{*(m+1)}(t) = \gamma_k^{(m)}(t)\left(y_{t,i} - \hat{\mu}_{k,i}^{(m+1)}\right)\left(y_{t,j} - \hat{\mu}_{k,j}^{(m+1)}\right)',$$

for all $t \in \{1, \ldots, T\}$ and $i, j \in \{1, \ldots, p\}$.

(3) Alternate between steps (2.1) and (2.2) until a stopping criterion is met.

Fiecas et al. (2017) point out (see the last paragraph of Section 3.3 of their paper) that this modified version of the EM algorithm does not possess the monotonicity property of the original algorithm (see Proposition 2.1). The authors state that the reason for the loss of monotonicity is not the incorporation of the penalty term itself, but the fact that the regularisation parameter is adapted according to the data. If both λ_k and $\tilde{v}_k$ were fixed, then the modified algorithm would still possess the property of monotonicity with respect to the penalised observed-data likelihood.

Chapter 4

Applications

In this chapter we report our results from the application of the classical EM algorithm, the modified version of Fiecas et al. (2017) and a variant of the latter, where, in the E-step, we estimate the complete-data log-likelihood function, and, in the M-step, we maximise the penalised complete-data log-likelihood, to derive the shrinkage estimators for the covariance matrices. This framework has also been adopted by Chen et al. (2014).

We used both simulated data and real data. The results from the simulations are presented in Section 4.1, while the results from the analysis of the real data set are given in Section 4.2.

The code for this paper was written and executed using R Statistical Software R Core Team (2022). For the implementation of all three algorithms, we modified the function `em.mvn` that can be found inside the function `em.hmm` of the R package Xu and Liu (2018).

4.1 Simulations

In this section we follow the simulation study of Fiecas et al. (2017).

The simulated data are a multivariate time series of dimension $p = 20$ or $p = 50$ generated from the model defined by Equation (2.45), either with two or three states ($K = 2$ or $K = 3$), with sample size $T = 128$ or $T = 256$.

For the cases with two states ($K = 2$), the stationary distribution of the underlying Markov chain is $\pi = (1/2\,,\ 1/2)$ and the one step transition probability matrix is

$$\begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix},$$

while for the three-state cases ($K = 3$), the stationary distribution is $\pi = (1/3\,,\ 1/3\,,\ 1/3)$ and the one step transition probability matrix is

$$\begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix}.$$

Fiecas et al. (2017) state that their choice of construction of the transition probability matrices is a way to overcome the problem of identifiability of the states and to make the hidden Markov chain more stable and balanced. This is not too restrictive, as the relation between the dimension, p, of the data and the sample size, T, and the fact that, due to the construction of the model, the sample size for some states might be smaller than $T/2$ or $T/3$ (when $K = 2$ or $K = 3$, respectively), makes the settings challenging enough to assess the performance of the methods (see Section 4 in their paper).

The construction of the different covariance matrices is as follows.

- The covariance matrix under state 1 is defined as $(\Sigma_1)_{ij} = 0.7^{|i-j|}$, $i, j \in \{1, \ldots, p\}$, with $p = 20$ or $p = 50$.
- The covariance matrix under state 2 is a block-diagonal matrix, obtained from the following construction scheme. Initially, a 5×5 correlation matrix $R_5^{(2)}$ is constructed, whose all off-diagonal entries are equal to 0.25, and all its diagonal entries are, of course, equal to 1. Then a $p \times p$ ($p = 20$ or $p = 50$) block-diagonal correlation matrix $R_p^{(2)}$ is constructed with $R_5^{(2)}$ along the diagonals. Setting $C_p^{(2)} = 1.5\boldsymbol{I}_p$ ($p = 20$ or $p = 50$), the covariance matrix under state 2 is defined as $\Sigma_2 = C_p^{(2)} R_p^{(2)} C_p^{(2)}$.
- The covariance matrix under state 3 is also a block-diagonal matrix, obtained from a similar construction scheme to the one used for Σ_2. We begin by constructing a $(p/2) \times (p/2)$ ($p = 20$ or $p = 50$) correlation matrix $R_{p/2}^{(3)}$, whose all off-diagonal entries are equal to 0.5 and all its diagonal entries are equal to 1. Then a $p \times p$ block-diagonal correlation matrix $R_p^{(3)}$ is created with two diagonal blocks, where each block is the matrix $R_{p/2}^{(3)}$. Setting $C_p^{(3)} = 1.5\boldsymbol{I}_p$ ($p = 20$ or $p = 50$), the covariance matrix under state 3 is defined as $\Sigma_2 = C_p^{(3)} R_p^{(3)} C_p^{(3)}$.

To determine the square root of the covariance matrices, which is necessary for the generating procedure of the data (see Equation (2.45)), the Cholesky decomposition is implemented.

To compare the performance of the modified EM algorithm of Fiecas et al. (2017), and its variant, to the classical EM algorithm, where the sample covariance matrix is used, the percentage relative

improvement in average loss (abbreviated PRIAL) is used, defined as

$$PRIAL\left(\widehat{\Sigma}_k^s\right) = 100 \times \frac{\mathrm{E}\left(\left|\left|\widehat{\Sigma}_k - \Sigma_k\right|\right|_{\mathrm{F}}^2\right) - \mathrm{E}\left(\left|\left|\widehat{\Sigma}_k^s - \Sigma_k\right|\right|_{\mathrm{F}}^2\right)}{\mathrm{E}\left(\left|\left|\widehat{\Sigma}_k - \Sigma_k\right|\right|_{\mathrm{F}}^2\right)}. \tag{4.1}$$

The expected values involved in (4.1) are estimated by their sample analogs using 100 Monte Carlo samples.

In the sequel, for the results that were obtained using the method of Fiecas et al. (2017), we use the abbreviation SPEM (Shrinkage Penalised EM), while the abbreviation SEM (Shrinkage EM) is used for the results that came from the method where the penalised complete-data log-likelihood is involved only in the M-step.

It should be noted that in Simulations 2 to 4, due to long execution times, only one set of initial values was used in the implementation of the algorithms, for each Monte Carlo sample. The initial values for the stationary distribution and the transition probability matrix were random. For the vector of means, the sample mean of the data was used for all states. Accordingly, we used the sample covariance matrix of the data, as the initial value for the covariance matrices of all states.

As Fiecas et al. (2017) point out (see the paragraph following Equation (19)), there is a label-switching problem[1] which impedes the evaluation of the performance of the two methods in estimating the covariance matrix under each state. To overcome this, following their framework, the performance of the estimates is assessed by estimating the PRIAL of the transition probability matrix, which is not affected by the label-switching problem due to its symmetry, and the marginal covariance matrix, defined as $\Sigma = \sum_{k=1}^{K} \pi_k \Sigma_k$.

4.1.1 Simulation 1

In Simulation 1 it is assumed that the states are known, which allows us to evaluate the performance of the estimators for the true covariance and precision matrices, under each state, without the "burden" of the estimation of the hidden state variables. The means under each state are assumed to be equal to the zero vector $\mathbf{0}_p$ ($p = 20$ or $p = 50$).

The results from the first simulation are summarised in Table 4.1 below. The estimators of the true precision matrices are the inverses of the corresponding shrinkage estimators and the corresponding

[1] Consider a finite mixture model with parameters $[(w_1, \phi_1), \ldots, (w_K, \phi_K)]$, where w_i and ϕ_i are the weight and the parameter of the distribution of component i, $i \in \{1, \ldots, K\}$, respectively. If the value of the likelihood function is invariant to permutations of the parameters, that is if the value of the likelihood function remains unchanged when we interchange the indices of (w_i, ϕ_i) and $\left(w_j, \phi_j\right)$, for any $i.j \in \{1, \ldots, K\}$, then it is said that there is a label-switching problem (see, e.g., Section 3 in Redner and Walker 1984, or Subsection 2.2.1 in Stephens 2000).

sample covariance matrices. The PRIALs for the precision matrices are defined as in Equation (4.1), where each matrix has been substituted for its inverse. The same method is also implemented to compute the PRIALs of the marginal covariance matrices.

Number of States K	Dimension p	Sample Size T	State	Covariance Matrix	Precision Matrix	Marginal Covariance Matrix
Two States	20	128	1	7.76	—	32.73
			2	44.01	—	
	20	256	1	5.48	91.89	21.16
			2	27.33	88.63	
	50	128	1	23.45	—	52.21
			2	64.59	—	
	50	256	1	13.91	—	40.35
			2	52.72	—	
Three States	20	128	1	14.02	—	14.23
			2	49.20	—	
			3	15.76	—	
	20	256	1	3.49	97.76	9.53
			2	39.03	98.66	
			3	9.89	87.85	
	50	128	1	34.90	—	18.74
			2	71.39	—	
			3	16.41	—	
	50	256	1	20.95	—	10.60
			2	59.99	—	
			3	9.76	—	

TABLE 4.1: Simulation 1 PRIALs for the covariance and the precision matrices, under each known state and PRIALs for the marginal covariance matrices. The hyphen (—) denotes that some of the sample covariance matrix estimates were singular so that the corresponding PRIALs could not be calculated.

It becomes clear from the results that in every case, shrinkage improves upon the estimates for the true covariance matrices under each state. Notice that, while keeping the value of K fixed, the PRIALs

decrease with decreasing values of the ratio p/T. This should not come as a surprise, since the larger the sample size T compared to the dimension p, the better the sample covariance matrix performs as an estimator of the true covariance matrix, which means that the benefit of using shrinkage decreases.

As was discussed in Chapter 3, it is important to have well-conditioned estimates of the precision matrices. In our case the precision matrix is needed for the computation of the likelihood function. As it can be seen by the corresponding PRIALs, the shrinkage estimator has a substantial advantage, when used to estimate the true precision matrix, over the sample covariance matrix. In our simulations, some of ththe shrinkage estimates were not invertible in the following cases:

- case $K = 2$, $p = 20$, $T = 128$: some of the shrinkage estimates under state 1, $\widehat{\Sigma}_1^s$,
- case $K = 3$, $p = 20$, $T = 128$: some of the shrinkage estimates under state 1, $\widehat{\Sigma}_1^s$,
- case $K = 3$, $p = 50$, $T = 128$: some of the shrinkage estimates under state 3, $\widehat{\Sigma}_3^s$.

In all these cases, for one or more samples, the corresponding state of the Markov chain did not occur in the sample paths. As a result the corresponding sample covariance matrices were equal to the zero matrix (see Eqaution (2.58)) therefore, the corresponding shrinkage estimates were also equal to the zero matrix (see Equations (3.18) and (3.21)). In any other case where the PRIAL is missing, the shrinkage estimates were always invertible, while some of the corresponding sample covariance matrices were singular. These results indicate that the shrinkage estimates are, indeed, well-conditioned.

It should be noted that we also examined if and how the choice of a kernel function and the bandwidth, which are involved in the estimation of the optimal shrinkage weights, affects the results. In particular, we used the Bartlett kernel, the Parzen kernel, the Tukey-Hanning kernel and their renormalised versions (see Section 2 in Andrews 1991). For the bandwidth, we tested values ranging from 10^{-8}, up to $T^{1/3}$, such that $b_T = o\left(\sqrt{T}\right)$ (see Theorem 3.3). The difference in the PRIALs was insignificant, for the different choices of a kernel function. The choice of bandwidth seemed to have a somewhat bigger impact on the results, but, again, the differences were not significant (the largest differences were about $2-3\%$). However, it was observed that smaller values of the bandwidth led to a slight increase in the PRIALs, in most cases. For values of the bandwidth below 0.5, the values of the PRIALs were identical. The results presented in Table 4.1 were obtained using the renormalised Parzen kernel and bandwidth $b = 10^{-2}$.

4.1.2 Simulation 2

Simulation 2 is the same as Simulation 1, with the only difference being that the states are unknown, so that the corresponding variables are estimated using the EM algorithm. Our results are

summarised in Table 4.2 below, where the renormalised Parzen kernel has been used and the value of the bandwidth was set equal to $T^{1/3}$.

Number of States K	Dimension p	Sample Size T	Transition Probability Matrix		Marginal Covariance Matrix	
			SPEM	SEM	SPEM	SEM
Two States	20	128	-5.20	31.66	26.57	27.99
	20	256	22.24	39.75	19.34	20.29
	50	128	39.67	20.32	45.45	45.81
	50	256	-55.44	30.31	32.01	34.77
Three States	20	128	-29.42	1.66	17.52	21.06
	20	256	-13.08	10.29	8.71	13.44
	50	128	-114.08	-4.49	21.51	29.21
	50	256	-14.72	10.54	12.82	19.79

TABLE 4.2: Simulation 2 PRIALs for the transition probability matrix and the marginal covariance matrix.

The estimates for the transition probability matrices, produced by the method of Fiecas et al. (2017), are much less accurate than the ones obtained using the classical EM algorithm, in most cases. The only settings where the SPEM algorithm produces better estimates are when $(K = 2, p = 20, T = 256)$ and $(K = 2, p = 50, T = 128)$. This contradicts the results presented by Fiecas et al. (2017) (see Table 2 in Subsection 4.2 of their paper). On the other hand, the estimates for the transition probability matrices that derive from the SEM algorithm, are far more accurate than the corresponding estimates of the classical EM, in all cases except when $(K = 3, p = 50, T = 128)$.

In every setting the shrinkage estimator of both the SPEM and the SEM algorithms improved upon the estimates for the covariance matrix. From the PRIALs of the marginal covariance matrix, once again, it becomes evident that the benefit of using shrinkage is greater for larger values of the ratio p/T, when the number of states K remains fixed. We also observe that the marginal PRIALs corresponding to the estimates produced by SEM are slightly and, in some cases, quite significantly larger than their counterparts coming from SPEM.

4.1.3 Simulation 3

Following Fiecas et al. (2017), to assess the performance of the modified EM algorithms when the means differ between states, the data in Simulation 3 are generated as in the previous settings, but with different means under each state, with equal entries across all dimensions. Under state $k \in \{1, \ldots, K\}$ ($K = 2$ or $K = 3$), the elements of the vector of means are $2\,(-1)^k / \sqrt{pk}$, for all $i \in \{1, \ldots, p\}$ ($p = 20$ or $p = 50$). Again, we report the PRIALs that emerged using the renormalised Parzen kernel and setting the bandwidth equal to $T^{1/3}$.

Number of States K	Dimension p	Sample Size T	Transition Probability Matrix		Marginal Covariance Matrix	
			SPEM	SEM	SPEM	SEM
Two States	20	128	5.07	46.51	39.75	44.82
	20	256	-7.22	41.61	23.78	24.31
	50	128	39.48	23.71	51.57	53.82
	50	256	-45.74	36.90	37.88	43.20
Three States	20	128	-22.20	6.97	19.48	23.80
	20	256	-17.30	11.65	8.52	11.65
	50	128	-112.08	-0.05	21.73	27.83
	50	256	-44.89	8.25	13.02	19.42

TABLE 4.3: Simulation 3 PRIALs for the transition probability matrix and the marginal covariance matrix.

The results are similar to the ones from Simulation 2. Once again, our implementation of the SPEM algorithm, performs worse than the classical EM in the estimation of the transition probability matrix, which contradicts the corresponding results of Fiecas et al. 2017 (see Table 3 in Subsection 4.2). The SEM algorithm improves upon estimates of the transition probability matrices in almost all cases.

The estimates for the covariance matrices are far more accurate when shrinkage has been used (both from SPEM and SEM methods), compared to the sample estimates. Again, there is a slight advantage of using the SEM algorithm, over the SPEM, to estimate the covariance matrices. The relation between the ratio p/T and the benefit of shrinkage, is confirmed once more.

4.1.4 Simulation 4

In Simulation 4 the data are generated from the multivariate Student distribution with 15 degrees of freedom and using the covariance matrices Σ_1 and Σ_3, for the two-state scenarios, and Σ_1, Σ_2 and Σ_3 for the three-state scenarios. Simulation 4 is used to evaluate the performance of the shrinkage estimators for the cases where the underlying distribution of the data has heavy tails (see Section 4 in Fiecas et al. 2017). As in the previous two simulations (2 and 3), we used the renormalised Parzen kernel and set the value of the bandwidth equal to $T^{1/3}$.

Number of States K	Dimension p	Sample Size T	Transition Probability Matrix		Marginal Covariance Matrix	
			SPEM	SEM	SPEM	SEM
Two States	20	128	-12.29	5.33	28.94	31.41
	20	256	0.52	27.83	23.57	24.79
	50	128	36.07	14.93	30.15	28.69
	50	256	-60.70	22.20	23.87	23.93
Three States	20	128	-47.43	-6.61	37.94	44.60
	20	256	-18.92	-1.12	33.23	38.73
	50	128	-135.49	-11.46	45.39	44.13
	50	256	-61.80	-0.32	38.56	32.90

TABLE 4.4: Simulation 4 PRIALs for the transition probability matrix and the marginal covariance matrix.

Once again, the estimates of the transition probability matrices produced by the SPEM algorithm, are far less accurate than the ones derived from the classical EM, except in the cases $(K = 2, p = 20, T = 256)$ and $(K = 2, p = 50, T = 128)$. The SEM algorithm also failed to improve upon the estimates for the transition probability matrices in the three-state scenarios, however they were far more accurate than the ones obtained from the SPEM.

The shrinkage estimators of both SPEM and SEM, produced more accurate estimates for the covariance matrices than the sample estimators in all settings. As in Simulations 2 and 3, the benefit of using shrinkage is greater for larger values of the ratio p/T.

Our findings indicate that shrinkage always leads to more stable and accurate estimates for the covariance matrix, compared with the sample estimator, under challenging conditions. The benefit

of using shrinkage becomes more pronounced for higher data dimensions and lower sample sizes. From the results of Simulation 1, it becomes clear that, especially, for the estimation of precision matrices, the method yields stable estimates even in cases where the classical approach would simply fail.

Contrary to what was expected, our implementation of the modified EM algorithm proposed by Fiecas et al. (2017) (SPEM), produced far less accurate estimates of the transition probability matrices, in most cases, compared to the classical EM algorithm. Further investigation is needed from our part, to determine the reasons behind this contradiction. However, the implementation of the variant of this method (SEM), where the penalised complete-data log-likelihood is used only in the M-step, improved quite significantly upon these estimates, in most cases, and even when the classical approach proved superior, the differences were not as dramatic as the ones that were observed when SPEM was used.

4.2 U.S. Industry Portfolio

The data set used for the analysis is a publicly available collection of U.S. industry portfolio data, that can be found here. It is composed of monthly returns from $p = 30$ industry sectors taken from NYSE, AMEX and NASDAQ. A description for every industry sector is available here.

Following Fiecas et al. (2017), we used the returns starting from January 2000 and ending at December 2011, resulting in a 30-dimensional time series with $T = 144$ observations. The log-transform of these returns was taken, so that a light tail is induced in the underlying distribution, in order for the moment conditions of model (2.45) to be satisfied (see the discussion following Assumptions (A) in Subsection 2.3.1 and the second paragraph of Section 5 in Fiecas et al. 2017). The analysis was performed using the mean-centred log-returns.

We used 1000 random sets of initial values for the parameters, for both the SPEM and the SEM algorithms and report the results from the initialisation that yielded the largest value of the penalised complete-data log-likelihood and the largest value of the complete-data log-likelihood, respectively. We examined two different cases for the number of states of the hidden Markov chain, $K = 2$ and $K = 3$. To determine the number of states of the model that best fits the data, we used Akaike's Information Criterion (abbreviated AIC) (see, e.g. Section 6.1 in Zucchini et al. 2016, or Section 4.4 in Trevezas 2021). To compute the value of AIC we used the value of the penalised complete-data log-likelihood when SPEM was used, while for the SEM, we used the value of the complete-data log-likelihood. The model with $K = 2$ states resulted in the lowest value[2] of AIC for both methods.

[2]When AIC is used for model selection,, the model with the lowest value of the criterion is chosen as the one that best fits the data (see Section 6.1 in Zucchini et al. 2016).

Fiecas et al. (2017) considered models with up to $K = 5$ states, using 1000 random sets of initial values. Due to long execution times, we examined these cases using only 200 sets of initial values. Again, the model with $K = 2$ states yielded the lowest value of AIC for both methods.

Before we proceed to present the results, we should note that it is likely that there are more than two states. For multivariate time series, the sample size $T = 144$ is not sufficiently large for AIC to be able to capture fine features in the data structure that might indicate the existence of more states (see the last paragraph of Section 5 in Fiecas et al. 2017).

We begin with the results produced using the modified algorithm of Fiecas et al. (2017) (SPEM). The estimated optimal weights are 0.602 for state 1 and 0.078 for state 2. The estimated transition probability matrix is

$$\widehat{\boldsymbol{P}}_{SPEM} = \begin{pmatrix} 0.289 & 0.711 \\ 0.007 & 0.993 \end{pmatrix}.$$

State 1 represents a state of higher volatility (variance) for all industry sectors, while in state 2 the correlations between the sectors are stronger. By the estimate of the transition probability matrix $\widehat{\boldsymbol{P}}_{SPEM}$, we deduce that the industry portfolios are more likely to move from the more volatile state 1, towards the less volatile state 2 and stay there. In Figure 4.1, below, the values of the centred log-returns of the 30 industry sectors are depicted, with the different colours in the background representing the underlying states. For the hidden path reconstruction we used the Viterbi algorithm (see Appendix D), using the estimates of the parameters yielded by the SPEM algorithm.

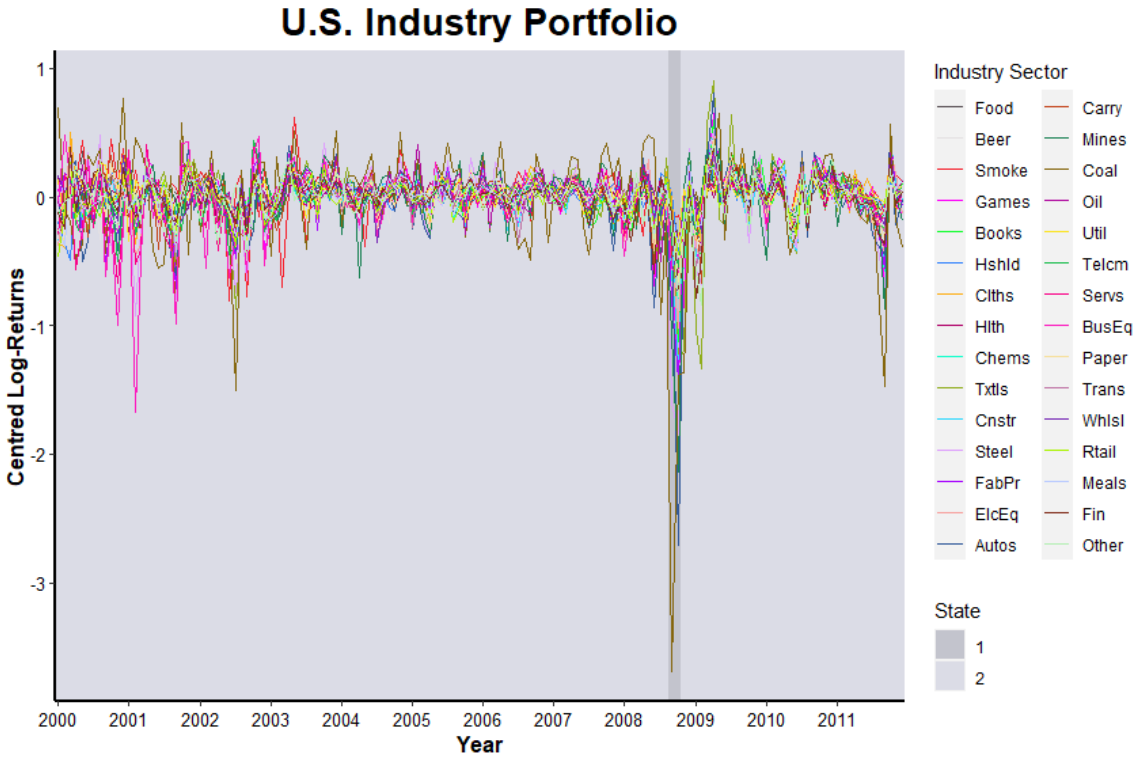

FIGURE 4.1: The time series of the centred log-returns of the 30 industry sectors and the estimated state sequence using the SPEM algorithm.

It is evident that there exists heteroscedasticity in the data. There are periods of higher and lower volatility. For example within the first quarter of 2001, or the third quarter of 2002, there are large spikes in the values of the centred log-returns, which implies that during these periods, the market was probably under a state of high risk. The more volatile state 1 is probably the underlying state during these periods, but this was not captured by the estimated state sequence. In the path produced by the estimates of SPEM, state 1 only occurs within the second half of 2008, where the highest volatility is observed.

We now report the results of the SEM algorithm. The optimal weight for state 1 is 0.058, and for state 2 is 0.403. The estimated transition probability matrix is

$$\widehat{\boldsymbol{P}}_{SEM} = \begin{pmatrix} 0.941 & 0.059 \\ 0.321 & 0.679 \end{pmatrix}.$$

Here the labels are switched, compared to the results of SPEM. Under state 1 the correlations between the industry sectors are stronger, while under state 2 the variance of each industry sector is higher. As was done for the results of the SPEM algorithm, in Figure 4.2, below, the time series of the centred log-returns of the 30 industry sectors is depicted, along the estimated state sequence of the Markov chain, represented by the colours in the background.

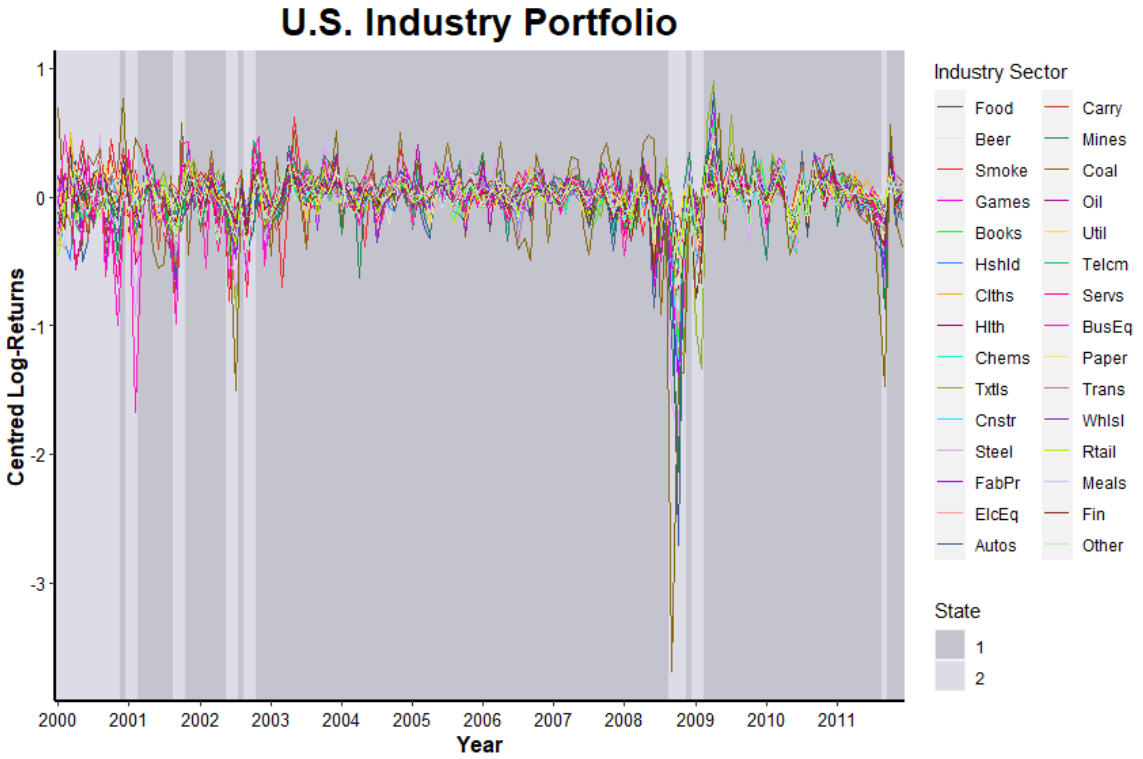

FIGURE 4.2: The time series of the centred log-returns of the 30 industry sectors and the estimated state sequence using the SEM algorithm.

We can see that using the parameter estimates derived from the SEM algorithm, the resulting estimated path is very well adapted to the changes in the variance of the 30 industry sectors. This suggests that, using the SEM method, we were able to capture quite well the changes in the market, that might have happened during the period under consideration. These results are quite similar to the ones reported in Section 5 of Fiecas et al. (2017).

Below we present the corresponding figure which was yielded using the classical EM algorithm. The resulting estimates, led to a state sequence with too many transitions between the two states. Even though many potential changes in the market are captured, the estimated path is less interpretable than the corresponding path that was obtained using the SEM. This implies that the estimates of the parameters produced by the classical EM algorithm are less stable than their counterparts from the SEM (see Section 5 in Fiecas et al. 2017).

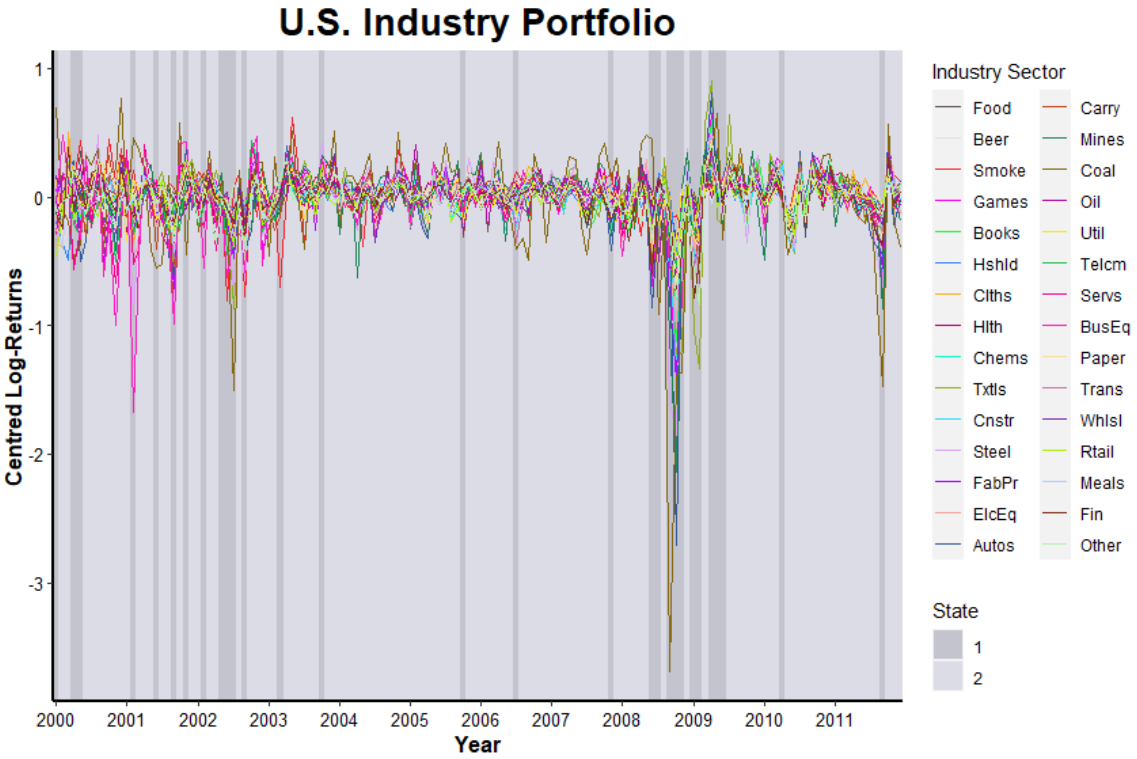

FIGURE 4.3: The time series of the centred log-returns of the 30 industry sectors and the estimated state sequence using the EM algorithm.

Bibliography

L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, The Annals of Mathematical Statistics 37 (1966) 1554–1563. doi:10.1214/aoms/1177699147.

T. Petrie, Probabilistic functions of finite state Markov chains, The Annals of Mathematical Statistics 40 (1969) 97 – 115. URL: https://doi.org/10.1214/aoms/1177697807. doi:10.1214/aoms/1177697807.

L. E. Baum, T. Petrie, G. Soules, N. Weiss, A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, The Annals of Mathematical Statistics 41 (1970) 164 – 171. URL: https://doi.org/10.1214/aoms/1177697196. doi:10.1214/aoms/1177697196.

L. E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes, Inequalities 3 (1972) 1–8.

L. E. Baum, J. A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bulletin of the American Mathematical Society 73 (1967) 360 – 363. URL: http://dx.doi.org/10.1090/S0002-9904-1967-11751-8. doi:bams/1183528841.

L. E. Baum, G. Sell, Growth transformations for functions on manifolds, Pacific Journal of Mathematics 27 (1968) 211–227. doi:http://dx.doi.org/10.2140/pjm.1968.27.211.

O. Cappé, E. Moulines, T. Rydén, Inference in Hidden Markov Models, in: Proceedings of EUSFLAT conference, 2005. doi:https://doi.org/10.1007/0-387-28982-8.

W. Zucchini, I. MacDonald, R. Langrock, Hidden Markov Models for Time Series: An Introduction Using R, 2nd ed., Chapman and Hall/CRC, 2016. doi:https://doi.org/10.1201/b20790.

Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Transactions on Information Theory 48 (2002) 1518–1569. doi:http://dx.doi.org/10.1109/TIT.2002.1003838.

B. Mor, S. Garhwal, A. Kumar, A systematic review of hidden markov models and their applications, Archives of Computational Methods in Engineering 28 (2021) 1429–1448. doi:https://doi.org/10.1007/s11831-020-09422-4.

S. E. Levinson, L. R. Rabiner, M. M. Sondhi, An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, The Bell System Technical Journal 62 (1983) 1035–1074. doi:10.1002/j.1538-7305.1983.tb03114.x.

L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–286. doi:10.1109/5.18626.

B. H. Juang, L. R. Rabiner, Hidden Markov models for speech recognition, Technometrics 33 (1991) 251–272. doi:10.2307/1268779.

F. Jelinek, Statistical Methods for Speech Recognition, MIT press, 1998.

F. S. Samaria, Face recognition using hidden Markov models, 1995. doi:https://doi.org/10.17863/CAM.14051.

A. Nefian, M. Hayes, Face detection and recognition using hidden Markov models, in: Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), volume 1, 1998, pp. 141–145. doi:10.1109/ICIP.1998.723445.

F. Alhadi, W. Fakhr, A. F. Seddik, Hidden Markov models for face recognition., in: Computational Intelligence, 2005, pp. 409–413.

G. Rigoll, A. Kosmala, S. Eickeler, High performance real-time gesture recognition using hidden Markov models, in: International Gesture Workshop, Springer, 1997, pp. 69–80. doi:http://dx.doi.org/10.1007/BFb0052990.

F.-S. Chen, C.-M. Fu, C.-L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, Image and Vision Computing 21 (2003) 745–758. doi:http://dx.doi.org/10.1016/S0262-8856(03)00070-2.

M. Gilloux, Hidden Markov models in handwriting recognition, in: Fundamentals in Handwriting Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, 1994, pp. 264–288. doi:http://dx.doi.org/10.1007/978-3-642-78646-4_15.

H. Bunke, M. Roth, E. Schukat-Talamazzini, Off-line cursive handwriting recognition using hidden Markov models, Pattern Recognition 28 (1995) 1399–1413. URL: https:

//www.sciencedirect.com/science/article/pii/003132039500013P. doi:https://doi.org/10.1016/0031-3203(95)00013-P.

J. Hu, M. Brown, W. Turin, HMM based online handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 1039–1045. doi:10.1109/34.541414.

L. Yang, B. Widjaja, R. Prasad, Application of hidden Markov models for signature verification, Pattern Recognition 28 (1995) 161–170. URL: https://www.sciencedirect.com/science/article/pii/003132039400092Z. doi:https://doi.org/10.1016/0031-3203(94)00092-Z.

J. G. A. Dolfing, E. H. L. Aarts, J. J. G. M. van Oosterhout, On-line signature verification with hidden Markov models, in: Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170), IEEE Computer Society, 2002. doi:http://dx.doi.org/10.1109/ICPR.1998.711942.

S. A. Daramola, T. S. Ibiyemi, Offline signature recognition using hidden Markov model (HMM), International Journal of Computer Applications 10 (2010) 17–22. doi:http://dx.doi.org/10.5120/1454-1967.

S. R. Eddy, et al., Multiple alignment using hidden Markov models., in: Intelligent Systems for Molecular Biology, volume 3, 1995, pp. 114–120.

S. R. Eddy, Profile hidden Markov models., Bioinformatics 14 (1998) 755–763. doi:https://doi.org/10.1093/bioinformatics/14.9.755.

G. Lunter, Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes, Bioinformatics 23 (2007) i289–i296. doi:https://doi.org/10.1093/bioinformatics/btm185.

K. Karplus, C. Barrett, M. Cline, M. Diekhans, L. Grate, R. Hughey, Predicting protein structure using only sequence information, Proteins: Structure, Function, and Bioinformatics 37 (1999) 121–125. doi:https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<121::AID-PROT16>3.0.CO;2-Q.

A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology: applications to protein modeling, Journal of Molecular Biology 235 (1994) 1501–1531. doi:https://doi.org/10.1006/jmbi.1994.1104.

A. Krogh, B. Larsson, G. Von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, Journal of Molecular Biology 305 (2001) 567–580. doi:https://doi.org/10.1006/jmbi.2000.4315.

A. V. Lukashin, M. Borodovsky, GeneMark.hmm: new solutions for gene finding, Nucleic Acids Research 26 (1998) 1107–1115. doi:10.1093/nar/26.4.1107.

I. Antonov, M. Borodovsky, Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm, Journal of Bioinformatics and Computational Biology 8 (2010) 535–551. doi:10.1142/s0219720010004847.

T. Wang, M. Bebbington, Identifying anomalous signals in GPS data using HMMs: an increased likelihood of earthquakes?, Computational Statistics & Data Analysis 58 (2011). doi:10.1016/j.csda.2011.09.019.

R. Avesani, A. Azzoni, M. Bicego, M. Orozco-Alzate, Automatic classification of volcanic earthquakes in HMM-induced vector spaces, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 640–647. doi:https://doi.org/10.1007/978-3-642-33275-3_79.

D. Chambers, J. Baglivo, J. Ebel, A. Kafka, Earthquake forecasting using hidden Markov models, Pure and Applied Geophysics 169 (2014) 625–639. doi:10.1007/s00024-011-0315-1.

J. Sansom, A hidden markov model for rainfall using breakpoint data, Journal of Climate 11 (1998) 42 – 53. URL: https://journals.ametsoc.org/view/journals/clim/11/1/1520-0442_1998_011_0042_ahmmfr_2.0.co_2.xml. doi:10.1175/1520-0442(1998)011<0042:AHMMFR>2.0.CO;2.

B. Betrò, A. Bodini, Q. A. Cossu, Using a hidden Markov model to analyse extreme rainfall events in central-east sardinia, Environmetrics 19 (2008) 702–713. doi:https://doi.org/10.1002/env.944.

O. Stoner, T. Economou, An advanced hidden Markov model for hourly rainfall time series, Computational Statistics & Data Analysis 152 (2020) 107045. URL: https://www.sciencedirect.com/science/article/pii/S0167947320301365. doi:https://doi.org/10.1016/j.csda.2020.107045.

F. O. Hocaoğlu, Ömer Nezih Gerek, M. Kurban, A novel wind speed modeling approach using atmospheric pressure observations and hidden Markov models, Journal of Wind Engineering and Industrial Aerodynamics 98 (2010) 472–481. URL: https://www.sciencedirect.com/science/article/pii/S0167610510000255. doi:https://doi.org/10.1016/j.jweia.2010.02.003.

S. Jafarzadeh, S. Fadali, C. Y. Evrenosoglu, H. Livani, Hour-ahead wind power prediction for power systems using hidden Markov models and Viterbi algorithm, in: IEEE Power & Energy Society General Meeting, 2010, pp. 1–6. doi:10.1109/PES.2010.5589844.

C. Barber, J. Bockhorst, P. Roebber, Auto-regressive HMM inference with incomplete data for short-horizon wind forecasting, in: Advances in Neural Information Processing Systems, volume 23, Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/file/242c100dc94f871b6d7215b868a875f8-Paper.pdf.

P. Ailliot, J. Bessac, V. Monbet, F. Pène, Non-homogeneous hidden Markov-switching models for wind time series, Journal of Statistical Planning and Inference 160 (2015) 75–88. URL: https://www.sciencedirect.com/science/article/pii/S0378375814002018. doi:https://doi.org/10.1016/j.jspi.2014.12.005.

R. J. Elliott, J. Van der Hoek, An application of hidden Markov models to asset allocation problems, Finance and Stochastics 1 (1997) 229–238. doi:https://doi.org/10.1007/s007800050022.

C. Erlwein, R. Mamon, M. Davison, An examination of HMM-based investment strategies for asset allocation, Applied Stochastic Models in Business and Industry 27 (2011) 204–221. doi:https://doi.org/10.1002/asmb.820.

T. Rydén, T. Teräsvirta, S. Åsbrink, Stylized facts of daily return series and the hidden Markov model, Journal of Applied Econometrics 13 (1998) 217–244. doi:https://doi.org/10.1002/(SICI)1099-1255(199805/06)13:3<217::AID-JAE476>3.0.CO;2-V.

F. Mary R. Hardy A.S.A., A regime-switching model of long-term stock returns, North American Actuarial Journal 5 (2001) 41–53. URL: https://doi.org/10.1080/10920277.2001.10595984. doi:10.1080/10920277.2001.10595984.

D. R. Smith, Markov-switching and stochastic volatility diffusion models of short-term interest rates, Journal of Business & Economic Statistics 20 (2002) 183–197. URL: https://doi.org/10.1198/073500102317351949. doi:10.1198/073500102317351949.

C. A. Wilson, R. J. Elliott, Stochastic Volatility or Stochastic Central Tendency: Evidence from a Hidden Markov Model of the Short-Term Interest Rate, Springer US, Boston, MA, 2014, pp. 33–53. URL: https://doi.org/10.1007/978-1-4899-7442-6_2. doi:10.1007/978-1-4899-7442-6_2.

C. Engel, Can the Markov switching model forecast exchange rates?, Journal of International Economics 36 (1994) 151–165. URL: https://www.sciencedirect.com/science/article/pii/0022199694900620. doi:https://doi.org/10.1016/0022-1996(94)90062-0.

G. M. Caporale, N. Spagnolo, Modelling east asian exchange rates: a Markov-switching approach, Applied Financial Economics 14 (2004) 233–242. URL: https://doi.org/10.1080/0960310042000201192. doi:10.1080/0960310042000201192.

S.-S. Chen, Revisiting the interest rate–exchange rate nexus: A Markov-switching approach, Journal of Development Economics 79 (2006) 208–224. doi:10.1016/j.jdeveco.2004.11.003.

F. M. Ali, F. Spagnolo, N. Spagnolo, Exchange rates and net portfolio flows: a Markov-switching approach, in: Hidden Markov Models in Finance, Springer, 2014, pp. 117–132. doi:10.1007/978-1-4899-7442-6_2.

S. Suvorova, L. Sun, A. Melatos, W. Moran, R. J. Evans, Hidden Markov model tracking of continuous gravitational waves from a neutron star with wandering spin, Physical Review D 93 (2016) 123009. URL: https://link.aps.org/doi/10.1103/PhysRevD.93.123009. doi:10.1103/PhysRevD.93.123009.

S. Suvorova, P. Clearwater, A. Melatos, L. Sun, W. Moran, R. J. Evans, Hidden Markov model tracking of continuous gravitational waves from a binary neutron star with wandering spin. ii. binary orbital phase tracking, Physical Review D 96 (2017) 102006. URL: https://link.aps.org/doi/10.1103/PhysRevD.96.102006. doi:10.1103/PhysRevD.96.102006.

L. Sun, A. Melatos, Application of hidden markov model tracking to the search for long-duration transient gravitational waves from the remnant of the binary neutron star merger gw170817, Physical Review D 99 (2019) 123003. URL: https://link.aps.org/doi/10.1103/PhysRevD.99.123003. doi:10.1103/PhysRevD.99.123003.

H. Middleton, P. Clearwater, A. Melatos, L. Dunn, Search for gravitational waves from five low mass x-ray binaries in the second advanced LIGO observing run with an improved hidden Markov model, Physical Review D 102 (2020) 023006. URL: https://link.aps.org/doi/10.1103/PhysRevD.102.023006. doi:10.1103/PhysRevD.102.023006.

A. Melatos, L. Dunn, S. Suvorova, W. Moran, R. Evans, Pulsar glitch detection with a hidden Markov model, The Astrophysical Journal 896 (2020) 78. doi:https://doi.org/10.3847/1538-4357/ab9178.

A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1977) 1–22. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x. doi:https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on Information Theory 13 (1967) 260–269. doi:10.1109/TIT.1967.1054010.

M. Fiecas, J. Franke, R. von Sachs, J. T. Kamgaing, Shrinkage estimation for multivariate hidden Markov models, Journal of the American Statistical Association 112 (2017) 424–435. URL: https://doi.org/10.1080/01621459.2016.1148608. doi:10.1080/01621459.2016.1148608.

J. Tadjuidje Kamgaing, Maximum Likelihood Estimators for Multivariate Hidden Markov Mixture Models (2013). URL: https://kluedo.ub.uni-kl.de/frontdoor/deliver/index/docId/3480/file/145_joseph15April2013.pdf, Technical Report, Report in Wirtschaftsmathematik 146, Department of Mathematics, University of Kaiserslautern.

I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, The Annals of Statistics 29 (2001) 295 – 327. URL: https://doi.org/10.1214/aos/1009210544. doi:10.1214/aos/1009210544.

O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis 88 (2004) 365–411. URL: https://www.sciencedirect.com/science/article/pii/S0047259X03000964. doi:https://doi.org/10.1016/S0047-259X(03)00096-4.

P. J. Bickel, E. Levina, Covariance regularization by thresholding, The Annals of Statistics 36 (2008) 2577 – 2604. URL: https://doi.org/10.1214/08-AOS600. doi:10.1214/08-AOS600.

T. Cai, W. Liu, X. Luo, A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation, Journal of the American Statistical Association 106 (2011). doi:10.1198/jasa.2011.tm10155.

P. J. Bickel, E. Levina, Regularized estimation of large covariance matrices, The Annals of Statistics 36 (2008) 199 – 227. URL: https://doi.org/10.1214/009053607000000758. doi:10.1214/009053607000000758.

J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, Journal of Econometrics 147 (2008) 186–197. URL: https://www.sciencedirect.com/science/article/pii/S0304407608001346. doi:https://doi.org/10.1016/j.jeconom.2008.09.017, econometric modelling in finance and risk management: An overview.

A. Sancetta, Sample covariance shrinkage for high dimensional dependent data, Journal of Multivariate Analysis 99 (2008) 949–967. URL: https://www.sciencedirect.com/science/article/pii/S0047259X07000851. doi:https://doi.org/10.1016/j.jmva.2007.06.004.

M. Yuan, J. Z. Huang, Regularized parameter estimation of high dimensional t distribution, Journal of Statistical Planning and Inference 139 (2009) 2284–2292. URL: https://www.sciencedirect.com/science/article/pii/S0378375808004163. doi:https://doi.org/10.1016/j.jspi.2008.10.014.

L. S. Chen, R. L. Prentice, P. Wang, A penalized em algorithm incorporating missing data mechanism for gaussian parameter estimation, Biometrics 70 (2014) 312–322. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12149. doi:https://doi.org/10.1111/biom.12149. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12149.

V. G. Kulkarni, Introduction to modeling and analysis of stochastic systems, Springer, 2011. doi:http://dx.doi.org/10.1007/978-1-4419-1772-0.

V. G. Kulkarni, Modeling and analysis of stochastic systems, Chapman and Hall/CRC, 2017. doi:https://doi.org/10.1201/9781315367910.

D. Fakinos, ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΤΙΣ ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ, ΕΚΔΟΣΕΙΣ ΣΥΜΜΕΤΡΙΑ, 2012.

S. M. Ross, Stochastic processes, Wiley New York, 1995.

J. R. Norris, Markov Chains, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997. doi:10.1017/CBO9780511810633.

G. Grimmett, D. Stirzaker, Probability and random processes, Oxford university press, 2001.

S. L. Lauritzen, Graphical models, volume 17, Clarendon Press, 1996.

M. I. Jordan, Graphical models, Statistical science 19 (2004) 140–155. doi:10.1214/088342304000000026.

R. G. Cowell, P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer Science & Business Media, 2007. doi:https://doi.org/10.1007/b97670.

D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT press, 2009.

L. E. Sucar, Probabilistic Graphical Models, volume 10, Springer, 2015. doi:https://doi.org/10.1007/978-1-4471-6699-3.

P. Billingsley, Statistical Methods in Markov Chains, The Annals of Mathematical Statistics 32 (1961) 12 – 40. URL: https://doi.org/10.1214/aoms/1177705136. doi:10.1214/aoms/1177705136.

K. Athreya, M. Majumdar, Estimating the stationary distribution of a markov chain, Economic Theory 21 (2003) 729–742. doi:10.1007/s00199-002-0292-9.

S. Trevezas, Σημειώσεις από το Μάθημα Στατιστική για Στοχαστικές Διαδικασίες [ΣΕΕ 31], 2021. URL: https://eclass.uoa.gr/modules/document/file.php/MATH429/Statistics_for_Stochastic_Processes_12_05.pdf.

C. F. J. Wu, On the Convergence Properties of the EM Algorithm, The Annals of Statistics 11 (1983) 95 – 103. URL: https://doi.org/10.1214/aos/1176346060. doi:10.1214/aos/1176346060.

H. O. Hartley, Maximum likelihood estimation from incomplete data, Biometrics 14 (1958) 174–194. URL: http://www.jstor.org/stable/2527783. doi:https://doi.org/10.2307/2527783.

H. O. Hartley, R. R. Hocking, The analysis of incomplete data, Biometrics 27 (1971) 783–823. URL: http://www.jstor.org/stable/2528820. doi:https://doi.org/10.2307/2528820.

T. Orchard, M. A. Woodbury, A MISSING INFORMATION PRINCIPLE: THEORY AND APPLICATIONS, University of California Press, 1972, pp. 697–716. URL: https://doi.org/10.1525/9780520325883-036. doi:doi:10.1525/9780520325883-036.

R. Sundberg, Maximum likelihood theory for incomplete data from an exponential family, Scandinavian Journal of Statistics 1 (1974) 49–58. URL: http://www.jstor.org/stable/4615553.

R. Sundberg, An iterative method for solution of the likelihood equations for incomplete data from exponential families, Communications in Statistics - Simulation and Computation 5 (1976) 55–64. URL: https://doi.org/10.1080/03610917608812007. doi:10.1080/03610917608812007.

X.-L. Meng, D. Van Dyk, The EM algorithm—an old folk-song sung to a fast new tune, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59 (1997) 511–567. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00082. doi:https://doi.org/10.1111/1467-9868.00082.

P. J. Bickel, Y. Ritov, T. Rydén, Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, The Annals of Statistics 26 (1998) 1614 – 1635. URL: https://doi.org/10.1214/aos/1024691255. doi:10.1214/aos/1024691255.

G. J. McLachlan, T. Krishnan, The EM algorithm and Extensions, volume 382, 2nd ed., John Wiley & Sons, 2007. doi:http://dx.doi.org/10.1002/9780470191613.

R. C. Bradley, Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions, Probability Surveys 2 (2005) 107 – 144. URL: https://doi.org/10.1214/154957805100000104. doi:10.1214/154957805100000104.

C. Francq, M. Roussignol, On white noises driven by hidden Markov chains, Journal of Time Series Analysis 18 (1997) 553–578. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00068. doi:https://doi.org/10.1111/1467-9892.00068.

R. Douc, Éric Moulines, T. Rydén, Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime, The Annals of Statistics 32 (2004) 2254 – 2304. URL: https://doi.org/10.1214/009053604000000021. doi:10.1214/009053604000000021.

P. G. Ciarlet, Introduction to Numerical Linear Algebra and Optimisation, Cambridge Texts in Applied Mathematics, Cambridge University Press, 1989. doi:10.1017/9781139171984.

J. Westlake, A HANDBOOK OF NUMERICAL MATRIX INVERSION AND SOLUTION OF LINEAR EQUATIONS, John Wiley & Sons, 1968.

J. W. Demmel, Applied numerical linear algebra, SIAM, 1997.

J. B. Copas, Regression, prediction and shrinkage, Journal of the Royal Statistical Society: Series B (Methodological) 45 (1983) 311–335. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1983.tb01258.x. doi:https://doi.org/10.1111/j.2517-6161.1983.tb01258.x.

R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1996) 267–288. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x. doi:https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429. URL: https://doi.org/10.1198/016214506000000735. doi:10.1198/016214506000000735.

H. Böhm, R. von Sachs, Shrinkage estimation in the frequency domain of multivariate time series, Journal of Multivariate Analysis 100 (2009) 913–935. URL: https://www.sciencedirect.com/science/article/pii/S0047259X08001942. doi:https://doi.org/10.1016/j.jmva.2008.09.009.

M. Fiecas, H. Ombao, The generalized shrinkage estimator for the analysis of functional connectivity of brain signals, The Annals of Applied Statistics 5 (2011) 1102 – 1125. URL: https://doi.org/10.1214/10-AOAS396. doi:10.1214/10-AOAS396.

J. Hausser, K. Strimmer, Entropy inference and the james-stein estimator, with application to nonlinear gene association networks, J. Mach. Learn. Res. 10 (2008). doi:10.1145/1577069.1755833.

V. DeMiguel, A. Martin-Utrera, F. J. Nogales, Size matters: Optimal calibration of shrinkage estimators for portfolio selection, Journal of Banking & Finance 37 (2013) 3018–3034. URL:

https://www.sciencedirect.com/science/article/pii/S0378426613002161. doi:https://doi.org/10.1016/j.jbankfin.2013.04.033.

X. Mestre, Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates, IEEE Transactions on Information Theory 54 (2008) 5113–5129. doi:10.1109/TIT.2008.929938.

R. J. Muirhead, Developments in Eigenvalue Estimation, Springer Netherlands, Dordrecht, 1987, pp. 277–288. URL: https://doi.org/10.1007/978-94-017-0653-7_14. doi:10.1007/978-94-017-0653-7_14.

J. Fan, P. Bickel, B. Li, A. Tsybakov, S. van de Geer, B. Yu, T. Valdés, C. Rivero, A. Vaart, Regularization in statistics, TEST: An Official Journal of the Spanish Society of Statistics and Operations Research 15 (2006) 271–344. doi:https://doi.org/10.1007/BF02607055.

P. J. Brockwell, R. A. Davis, Time series: Theory and Methods, 2nd ed., Springer New York, NY, 1991. doi:https://doi.org/10.1007/978-1-4419-0320-4.

D. W. K. Andrews, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, Econometrica 59 (1991) 817—858. URL: http://www.jstor.org/stable/2938229. doi:https://doi.org/10.2307/2938229.

M. Wand, M. Jones, Kernel Smoothing, 1st ed., Chapman and Hall/CRC, 1994. doi:https://doi.org/10.1201/b14876.

R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022. URL: https://www.R-project.org/, R version 4.2.1 (2022-06-23 ucrt).

Z. Xu, Y. Liu, rarhsmm: Regularized Autoregressive Hidden Semi Markov Model, 2018. URL: https://www.rdocumentation.org/packages/rarhsmm/versions/1.0.7, R package version 1.0.7.

R. A. Redner, H. F. Walker, Mixture densities, maximum likelihood and the em algorithm, SIAM Review 26 (1984) 195–239. URL: https://doi.org/10.1137/1026034. doi:10.1137/1026034.

M. Stephens, Dealing with label switching in mixture models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62 (2000) 795–809. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00265. doi:https://doi.org/10.1111/1467-9868.00265.

P. Billingsley, Probability and measure, 3rd ed., John Wiley & Sons, 1995.

A. W. van der Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998. doi:10.1017/CBO9780511802256.

E. L. Lehmann, G. Casella, Theory of Point Estimation, Springer Texts in Statistics, 2nd ed., Springer New York, NY, 1998. doi:https://doi.org/10.1007/b98854.

E. L. Lehmann, G. Casella, Mathematical Statistics: Basic Ideas and Selected Topics, 2nd ed., Prentice-Hall, 2001.

G. Casella, R. Berger, Statistical Inference Second Edition, Duxbur y advanced series, 2002.

E. L. Lehmann, Elements of Large-Sample Theory, Springer Texts in Statistics, 1st ed., Springer New York, NY, 1999. doi:https://doi.org/10.1007/b98855.

W. K. Härdle, L. Simar, Applied Multivariate Statistical Analysis, 2nd ed., Springer Berlin, Heidelberg, 2007. doi:https://doi.org/10.1007/978-3-540-72244-1.

T. W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., Wiley & Suns, 2003.

R. A. Johnson, D. W. Wichern, et al., Applied multivariate statistical analysis, volume 5, Prentice hall Upper Saddle River, NJ, 2002.

R. Shumway, D. Stoffer, Time Series Analysis and Its Applications, Springer New York, NY, 2011.

O. Bamdorff-Nielsen, D. Cox, Asymptotic Techniques for Use in Statistics, Chapman and Hall, 1989.

N. G. de Bruijn, Asymptotic methods in analysis, North-Holland Publishing Co., Amsterdam, 1958.

R. A. Horn, C. R. Johnson, Matrix Analysis, 2nd ed., Cambridge University Press, 2012. doi:10.1017/9781139020411.

K. B. Petersen, M. S. Pedersen, The Matrix Cookbook, 2012. URL: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf, version: November 15, 2012.

P. Doukhan, S. Louhichi, A new weak dependence condition and applications to moment inequalities, Stochastic Processes and their Applications 84 (1999) 313–342. URL: https://www.sciencedirect.com/science/article/pii/S0304414999000551. doi:https://doi.org/10.1016/S0304-4149(99)00055-1.

Appendix A

Auxiliary Material

A.1 Probability And Measure Theory

In this section some basic definitions and results from Probability and Measure theory are presented. This section is useful for someone who is not familiar with some of the notions presented throughout this paper and it mostly serves as a way to provide a sense of completeness. For a proper introduction to Probability and Measure theory the interested reader is referred to Billingsley (1995). We mainly follow Billingsley (1995) and Section 2.4 in Trevezas (2021).

Definition A.1 (Sample Space)**.** *Let us assume an experiment whose outcome, after each trial, is not predictable with certainty, but all its possible outcomes, say $\omega_1, \ldots, \omega_n$, are known. The set $\Omega = \{\omega_1, \ldots, \omega_n\}$ of all possible outcomes is called the sample space of the experiment. After each trial only one possible outcome occurs, that is if ω_i takes place, then no other ω_j can occur, for all $i, j \in \Omega = \{\omega_1, \ldots, \omega_n\}$, with $i \neq j$.*

Definition A.2 (σ-algebra)**.** *Consider a non-empty set Ω. A σ-algebra or sigma-field $\mathcal{A}$ on Ω is a collection of subsets of Ω, satisfying the following conditions:*

(i) $\Omega \in \mathcal{A}$,

(ii) *if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ ($\mathcal{A}$ is closed under complementation),*

(iii) *if $\{A_n : n \geq 1\}$ is a sequence of sets in $\mathcal{A}$, then $A = \bigcup_{n=1}^{+\infty} A_n \in \mathcal{A}$ ($\mathcal{A}$ is closed under countable unions).*

Since Ω is an element of $\mathcal{A}$ (Condition (i)), by Condition (ii) its complement, the empty set $\emptyset$, is also an element of $\mathcal{A}$. Additionally, by Condition (iii) for a sequence of sets in $\mathcal{A}$, $\{A_n : n \geq 1\}$, we have that $\bigcup_{n=1}^{+\infty} A_n \in \mathcal{A}$. By de Morgan's laws we have that $\bigcup_{n=1}^{+\infty} A_n = \left(\bigcap_{n=1}^{+\infty} A_n^c \right)^c$ and by Condition (ii) A_n^c is an element of $\mathcal{A}$ for all $n \geq 1$. It follows that $\mathcal{A}$ is also closed under countable intersections.

Definition A.3 (Measurable Space)**.** *A measurable space is a pair $(\Omega, \mathcal{A})$, where Ω is a non-empty set and $\mathcal{A}$ is a σ-algebra on Ω.*

Definition A.4 (Measure). *A measure is a set function μ on a σ-algebra $\mathcal{A}$ in Ω, satisfying the following conditions:*

(i) $\mu(A) \in [0, +\infty)$, *for any set* $A \in \mathcal{A}$,

(ii) $\mu(\emptyset) = 0$,

(iii) *for any sequence* $\{A_n : n \geq 1\}$ *of pairwise disjoint sets of* $\mathcal{A}$ ($A_i \cap A_j = \emptyset$, *for all* $i, j \geq 1$ *and* $i \neq j$), *it holds*

$$\mu\left(\bigcup_{n=1}^{+\infty} A_n\right) = \sum_{n=1}^{+\infty} \mu(A_n).$$

Definition A.5 (Finite, Infinite, σ-finite Measure). *Consider a measurable space* $(\Omega, \mathcal{A})$.

(i) *A measure* μ *is said to be finite on* $(\Omega, \mathcal{A})$, *if* $\mu(\Omega) < +\infty$.

(ii) *A measure* μ *is said to be infinite on* $(\Omega, \mathcal{A})$, *if* $\mu(\Omega) = +\infty$.

(iii) *A measure* μ *is said to be* σ*-finite on* $(\Omega, \mathcal{A})$, *if there exists a sequence* $\{A_n : n \geq 1\}$ *of sets in* $\mathcal{A}$, *with* $\mu(A_n) < +\infty$, *for all* $n \geq 1$, *such that* $\Omega = \bigcup_{n=1}^{+\infty} A_n$.

Definition A.6 (Measure Space). *A measure space is a triplet* $(\Omega, \mathcal{A}, \mu)$, *where* Ω *is a non-empty set,* $\mathcal{A}$ *is a* σ*-algebra on* Ω *and* μ *is a probability measure defined on* $\mathcal{A}$.

Definition A.7 (Almost Everywhere). *Consider a measure space* $(\Omega, \mathcal{A}, \mu)$. *A property* P *is said to hold almost everywhere (abbreviated a.e.) on* Ω, *or for almost every* x, *if for all* $x \in \Omega$ *which satisfy* P, *it holds that* $x \in A^c = \Omega \setminus A$, *where* $A \in \mathcal{A}$, *such that* $\mu(A) = 0$.

Definition A.8 (Probability Measure). *Consider a measurable space* $(\Omega, \mathcal{A})$. *A probability measure set function* $\mathrm{P} : \mathcal{A} \to [0, 1]$, *satisfying the following conditions:*

(i) $\mathrm{P}(A) \in [0, +\infty)$, *for any set* $A \in \mathcal{A}$,

(ii) P *is normalised, that is the measure of the entire sample space* Ω *equals 1,* $\mathrm{P}(\Omega) = 1$,

(iii) P *is countably additive or* σ*-additive, which means that for any countable sequence* $\{A_n : n \geq 1\}$ *of pairwise disjoint sets of* $\mathcal{A}$, *it holds*

$$\mathrm{P}\left(\bigcup_{n=1}^{+\infty} A_n\right) = \sum_{n=1}^{+\infty} \mathrm{P}(A_n).$$

In accordance with the definition of *measurespaces* we have the following definition.

Definition A.9 (Probability Space). *A probability space is a triplet* $(\Omega, \mathcal{A}, \mathrm{P})$*, where* Ω *is a non-empty set,* $\mathcal{A}$ *is a* σ*-algebra on* Ω *and* P *is a probability measure defined on* $\mathcal{A}$*.*

Definition A.10 (σ-algebra Generated by an Arbitrary Family). *Consider a non-empty set* Ω *and a collection* $\mathfrak{D}$ *consisting of subsets of* Ω*. A* σ*-algebra* $\mathcal{A}$ *which satisfies the following conditions:*

(i) $\mathfrak{D} \subseteq \mathcal{A}$,

(ii) *for any* σ*-algebra* $\mathcal{A}'$ *on* Ω *such that* $\mathfrak{D} \subseteq \mathcal{A}'$*, it follows* $\mathcal{A} \subseteq \mathcal{A}'$,

is called the σ*-algebra generated by* $\mathfrak{D}$ *and is denoted by* $\sigma(\mathfrak{D})$*.*

For any set Ω and a collection $\mathfrak{D}$, as defined in Definition A.10, the σ-algebra generated by $\mathfrak{D}$, $\sigma(\mathfrak{D})$, always exists and it is unique. It is the intersection of all the σ-algebras on Ω containing $\mathfrak{D}$ (see Section 2 in Billingsley 1995).

Definition A.11 (Borel σ-algebra). *Consider a metric space*[1] (Ω, d)*. The* σ*-algebra generated by the open sets*[2] *of* (Ω, d) *is called the Borel* σ*-algebra (on* Ω*) and is denoted by* $\mathcal{B}(\Omega)$*. The members of* $\mathcal{B}(\Omega)$ *are called the Borel sets of* Ω*.*

Let us denote by $\mathcal{O}_\Omega$ the family of all open sets of the metric space (Ω, d). Then Definition A.11 implies that $\mathcal{B}(\Omega) = \sigma(\mathcal{O}_\Omega)$.

Definition A.12 (Measurable Function). *Consider two measurable spaces* $(\Omega_1, \mathcal{A}_1)$ *and* $(\Omega_2, \mathcal{A}_2)$*. A function* $f : \Omega_1 \to \Omega_2$ *is called* $\mathcal{A}_1/\mathcal{A}_2$*-measurable (or* $\mathcal{A}_1$*-measurable when* $\mathcal{A}_2$ *is obvious, or just measurable when both* $\mathcal{A}_1$ *and* $\mathcal{A}_2$ *are obvious) if for every set* $B \subseteq \mathcal{A}_2$ *the inverse image of* B *under* f *is in* $\mathcal{A}_1$*, that is*

$$f^{-1}(B) = \{\omega \in \Omega_1 : f(x) \in B\} \in \mathcal{A}_1, \quad \textit{for all } B \in \mathcal{A}_2.$$

Equivalently, f *is* $\mathcal{A}_1/\mathcal{A}_2$*-measurable if* $f^{-1}(\mathcal{A}_2) \subseteq \mathcal{A}_1$*.*

Definition A.13 (Random Variable). *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$*. A real valued function* X *defined on* Ω*,* $X : \Omega \to \mathbb{R}$*, is called a random variable if it is* $\mathcal{A}/\mathcal{B}(\mathbb{R})$*-measurable, that is*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}, \quad \textit{for all } B \in \mathcal{B}(\mathbb{R})\,.$$

[1] A metric space is an ordered pair (Ω, d) where Ω is a non-empty set and d is a metric on Ω, that is, a nonnegative function $d : \Omega \times \Omega \to \mathbb{R}$, satisfying the following conditions:

(i) $d(x, y) = 0 \Leftrightarrow x = y, \forall x, y \in \Omega$,

(ii) $d(x, y) = d(y, x), \forall x, y \in \Omega$,

(iii) $d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in \Omega$.

[2] A subset B of a metric space (Ω, d) is called open if, for any point $x \in B$, there exists a real number $\epsilon > 0$ such that for any point $y \in \Omega$ with $d(x, y) < \epsilon$, it follows that $y \in B$.

Definition A.14 (Random Vector)**.** *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$*. A random vector* $\boldsymbol{X}$ *on* $(\Omega, \mathcal{A}, \mathrm{P})$ *is a p-tuple* $\boldsymbol{X} = (X_1, \ldots, X_p)$ *of random variables.*

Definition A.15 (Convergence of Random Variables)**.** *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$ *and a sequence of random variables* $\{X_n : n \geq 1\}$ *defined on* $(\Omega, \mathcal{A}, \mathrm{P})$*. The sequence* $\{X_n : n \geq 1\}$ *is said to converge*

(i) *with probability 1, or almost surely to a random variable X if and only if*

$$\mathrm{P}\left(\left\{\omega \in \Omega : \exists \lim_{n\to+\infty} [X_n(\omega)] \text{ and } \lim_{n\to+\infty} [X_n(\omega)] = X(\omega)\right\}\right) = 1,$$

or simply

$$\mathrm{P}\left(\lim_{n\to+\infty} (X_n) = X\right) = 1$$

and it is denoted by

$$X_n \xrightarrow{a.s.} X \Leftrightarrow |X_n - X| \xrightarrow{a.s.} 0,$$

(ii) *in probability to a random variable X if and only if for all* $\epsilon > 0$

$$\mathrm{P}(|X_n - X| > \epsilon) \xrightarrow[n\to+\infty]{} 0,$$

or equivalently

$$\mathrm{P}(|X_n - X| \leq \epsilon) \xrightarrow[n\to+\infty]{} 1,$$

and it is denoted by

$$X_n \xrightarrow{p} X \Leftrightarrow |X_n - X| \xrightarrow{p} 0,$$

(iii) *in distribution to a random variable X if and only if for all* $x \in \mathbb{R}$ *such that* $F(x^-) = F(x) \Leftrightarrow \lim_{y\to x^-} [F(y)] = F(x)$

$$F_n(x) \xrightarrow[n\to+\infty]{} F(x),$$

where $F_n(\cdot)$ *is the cumulative distribution function of* X_n *and* $F(x)$ *is the corresponding function of X, and it is denoted by*

$$X_n \xrightarrow{d} X.$$

In accordance to the univariate case, we can define the convergence of random vectors.

Definition A.16 (Convergence of Random Vectors)**.** *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$ *and a sequence of random vectors* $\{\boldsymbol{X}_n : n \geq 1\}$ *defined on* $(\Omega, \mathcal{A}, \mathrm{P})$*, where* $\boldsymbol{X}_n \in \mathbb{R}^p$*, for all* $n \geq 1$*. The sequence* $\{\boldsymbol{X}_n : n \geq 1\}$ *is said to converge*

(i) *with probability 1, or almost surely to a random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *if and only if*

$$||\boldsymbol{X}_n - \boldsymbol{X}||_2 \xrightarrow{a.s.} 0,$$

where $||\cdot||_2$ *denotes the Euclidean norm of a vector, or equivalently*

$$\mathrm{P}\left(\lim_{n\to+\infty}(\boldsymbol{X}_n) = \boldsymbol{X}\right) = 1$$

and it is denoted by

$$\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{X},$$

(ii) *in probability to a random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *if and only if*

$$||\boldsymbol{X}_n - \boldsymbol{X}||_2 \xrightarrow{p} 0,$$

and it is denoted by

$$\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X},$$

(iii) *in distribution to a random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *if and only if for all* $B \in \mathcal{B}(\mathbb{R}^p)$ *such that* $\mathrm{P}(\boldsymbol{X} \in \partial B) = 0$

$$\mathrm{P}(\boldsymbol{X}_n \in B) \xrightarrow[n\to+\infty]{} \mathrm{P}(\boldsymbol{X} \in B),$$

where $\partial B = \overline{B} \setminus B^o$ *is the topological boundary of* B *(with* B^o *denoting the interior of* B*), and it is denoted by*

$$\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}.$$

For the convergence in distribution both for random variables and random vectors, the sequences need not be defined in the same probability space with their corresponding limits (see the paragraph following Definition 2.3 in Trevezas 2021).

The following proposition relates the convergence with probability 1 and in probability of random vectors to the corresponding convergence of random variables (see Proposition 2.10 Trevezas 2021).

Proposition A.1. *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$*, a sequence of random vectors* $\{\boldsymbol{X}_n : n \geq 1\}$*, where* $\boldsymbol{X}_n \in \mathbb{R}^p$ *for all* $n \geq 1$*, and a random vector* $\boldsymbol{X} \in \mathbb{R}^p$*, all defined on* $(\Omega, \mathcal{A}, \mathrm{P})$*. Then*

$$\boldsymbol{X}_n \xrightarrow{a.s./p} \boldsymbol{X} \Leftrightarrow X_{n,i} \xrightarrow{a.s./p} X_i,$$

for all $i \in \{1, \ldots, p\}$.

For the convergence in distribution of random vectors we have the following result (see Theorem 2.1 in Trevezas 2021).

Theorem A.1. *A sequence of random vectors* $\{\boldsymbol{X}_n : n \geq 1\}$, *where* $\boldsymbol{X}_n \in \mathbb{R}^p$ *for all* $n \geq 1$ *converges to a random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *if and only if*

(i) *for all* $\boldsymbol{x} \in \mathbb{R}^p$, *such that* $F(\boldsymbol{x}^-) = F(\boldsymbol{x})$

$$F_n(\boldsymbol{x}) \xrightarrow[n \to +\infty]{} F(\boldsymbol{x}),$$

where $F_n(\cdot)$ *and* $F(\cdot)$ *are the cumulative distribution functions of* $\boldsymbol{X}_n$ *and* $\boldsymbol{X}$, *respectively,*

(ii) *any linear combination of the elements of* $\boldsymbol{X}_n$ *converges in distribution to the corresponding linear combination of the elements of* $\boldsymbol{X}$, *that is for any vector* $u \in \mathbb{R}^p$

$$u^{'}\boldsymbol{X}_n \xrightarrow{d} u^{'}\boldsymbol{X},$$

(iii) *for any continuous and bounded function* $f : \mathbb{R}^p \to \mathbb{R}$

$$\mathrm{E}[f(\boldsymbol{X}_n)] \xrightarrow[n \to +\infty]{} \mathrm{E}[f(\boldsymbol{X})].$$

The following proposition gives the relation between the three ways of convergence for random vectors. The same relation also holds for the corresponding ways of convergence for random variables (see Proposition 2.1 in Trevezas 2021).

Proposition A.2. *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$, *a sequence of random vectors* $\{\boldsymbol{X}_n : n \geq 1\}$, *where* $\boldsymbol{X}_n \in \mathbb{R}^p$ *for all* $n \geq 1$, *and a random vector* $\boldsymbol{X} \in \mathbb{R}^p$, *all defined on* $(\Omega, \mathcal{A}, \mathrm{P})$. *Then*

$$\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}.$$

In case $\boldsymbol{X} = \boldsymbol{c}$, *where* $\boldsymbol{c} \in \mathbb{R}^p$ *is a constant vector, then*

$$\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{c} \Leftrightarrow \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}.$$

Definition A.17 (Independence of Random Variables)**.** *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$ *and a sequence of random variables* $(X_1, \ldots, X_n)$. *The random variables* $(X_1, \ldots, X_n)$ *are called independent if and only if*

$$\mathrm{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathrm{P}(X_1 \in A_1) \cdot \ldots \cdot \mathrm{P}(X_n \in A_n),$$

for all $A_i \in \mathcal{B}(\mathbb{R})$*, with* $i \in \{1, \ldots, n\}$.

Definition A.18 (Independence of Random Vectors)**.** *Consider a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$ *and a sequence of random vectors* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$*, where* $\boldsymbol{X}_i \in \mathbb{R}^{p_i}$*, with* $i \in \{1, \ldots, n\}$*. The random vectors* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *are called independent if and only if*

$$\mathrm{P}(\boldsymbol{X}_1 \in A_1, \ldots, \boldsymbol{X}_n \in A_n) = \mathrm{P}(\boldsymbol{X}_1 \in A_1) \cdot \ldots \cdot \mathrm{P}(\boldsymbol{X}_n \in A_n)\,,$$

for all $A_i \in \mathcal{B}(\mathbb{R}^{p_i})$*, with* $i \in \{1, \ldots, n\}$.

Definition A.19 (σ-algebra Generated by a Random Variable)**.** *Consider a random variable* X *on a probability space* $(\Omega, \mathcal{A}, \mathrm{P})$*. The smallest* σ*-algebra with respect to which the random variable* X *is measurable, is called the* σ*-algebra generated by* X *and is denoted by* $\sigma(X)$*. It is the intersection of all* σ*-algebras with respect to which* X *is measurable. Generally, for any set* I *of indices the* σ*-algebra generated by the sequence* $(X_i : i \in I)$*, is the smallest* σ*-algebra with respect to which each* X_i *is measurable and is denoted by* $\sigma(X_i : i \in I)$.

Definition A.20 (α-mixing)**.** *Consider a probability space a* $(\Omega, \mathcal{A}, \mathrm{P})$*, a stochastic process* $\{X_t : -\infty < t < +\infty\}$ *on* $(\Omega, \mathcal{A}, \mathrm{P})$ *and a sequence* $\{a_n : n \geq 1\}$ *such that*

$$a_n = sup\left\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\,\mathrm{P}(B)| : A \in \sigma(X_t : t \leq k)\,,\ B \in \sigma(X_t : t \geq k + n)\right\}.$$

The elements a_n *are called the* α*-mixing coefficients (or strong mixing coefficients) of the process* $\{X_t : -\infty < t < +\infty\}$*. The process itself is said to be* α*-mixing or strongly mixing if*

$$a_n \xrightarrow[n \to +\infty]{} 0.$$

A stochastic process $\{X_t : -\infty < t < +\infty\}$ is *α-mixing with exponential decay* or *exponential mixing rate* if $a_n \xrightarrow[n \to +\infty]{} 0$ exponentially fast, where a_n are the α-mixing coefficients of $\{X_t : -\infty < t < +\infty\}$, as defined above (see the discussion following Equation (2.11) and the paragraph following Theorem 3.1 in Bradley 2005).

A.2 Statistical Analysis

In this section we provide some basic elements of Statistical analysis that mainly concern point estimation. The interested reader is referred to van der Vaart (1998), Lehmann and Casella (1998), Lehmann and Casella (2001), Casella and Berger (2002) and Lehmann (1999).

Definition A.21 (Statistical Model)**.** *Consider a random experiment with sample space Ω and a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined on Ω. If the outcome of the experiment is $\omega \in \Omega$, then $\boldsymbol{X}(\omega)$ is referred to as the observations or data. Only $\boldsymbol{X}$ is observable, thus we only need to consider its probability distribution. The family of distributions $\mathcal{P}$ defined on $\mathbb{R}^n$, in which this probability distribution belongs to, is a statistical model.*

Definition A.22 (Parameterisation)**.** *A function $\theta \to \mathrm{P}_\theta$, from a space of labels, called a parameter space, Θ to a model $\mathcal{P}$, used to describe $\mathcal{P}$, is called a parameterisation and it is expressed as $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$.*

Definition A.23 (Identifiable Model)**.** *A model $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ is called identifiable if the parameterisation $\theta \to \mathrm{P}_\theta$ is a $1-1$ function, that is if*

$$\theta_1 \neq \theta_2 \Rightarrow \mathrm{P}_{\theta_1} \neq \mathrm{P}_{\theta_2},$$

otherwise the model $\mathcal{P}$ is referred to as unidentifiable.

Definition A.24 (Statistic)**.** *A statistic T is a function from a sample space $\mathcal{X}$ to a space of values $\mathcal{T}$, such that if we observe $\boldsymbol{X} = \boldsymbol{x}$, then $T(\boldsymbol{x})$ is the quantity that we can compute.*

Definition A.25 (Estimator)**.** *Consider a parameter $\theta \in \Theta$ and a real-valued function $g(\theta)$ defined over Θ. Any real-valued function $\delta(\boldsymbol{X})$ defined over the sample space $\mathcal{X}$, used to estimate $g(\theta)$, is called an estimator of $g(\theta)$. If $\boldsymbol{x}$ is the observed value of $\boldsymbol{X}$, then the value $\delta(\boldsymbol{x})$ is the estimate of $g(\theta)$.*

Definition A.26 (Action Space)**.** *A set $\mathcal{A}$ of possible actions or decisions one can make is called action space.*

Definition A.27 (Loss Function)**.** *A function $L : \mathcal{P} \times \mathcal{A} \to \mathbb{R}^+$ is called loss function. $L(\mathrm{P}, \alpha)$, or $L(\theta, \alpha)$, if $\mathcal{P}$ is parameterised, where $\mathrm{P} \in \mathcal{P}$, $\alpha \in \mathcal{A}$ and $\theta \in \Theta$, represents the nonnegative loss incurred by taking action α when the probability distribution producing the data is P, or the true value of the parameter is θ, respectively.*

Definition A.28 (Decision Rule or Procedure)**.** *The decision rule or procedure is any function $\delta : \mathcal{X} \to \mathcal{A}$, such that, if a point from the sample space $\mathcal{X}$, $\boldsymbol{X} = \boldsymbol{x}$, is observed, then the action taken is $\delta(\boldsymbol{x})$.*

Consider a setting where $\boldsymbol{X} = \boldsymbol{x}$ is the outcome of an experiment, $\mathrm{P} \in \mathcal{P}$ is the probability distribution that generated the data, the loss function used is L and the decision rule is δ, thus the loss is $L[\mathrm{P}, \delta(\boldsymbol{x})]$. Since P is unknown, the same holds for $L[\mathrm{P}, \delta(\boldsymbol{x})]$. It is desirable for decision rules to possess good properties not only at one specific $\boldsymbol{x}$, but for a range of possible values over the sample space. This motivates the following definition.

Definition A.29 (Risk Function)**.** *For the outcome of an experiment $\boldsymbol{X} = \boldsymbol{x}$, where $\mathrm{P} \in \mathcal{P}$ is the probability distribution that generated the data, with loss function $L[\mathrm{P}, \delta(\boldsymbol{x})]$, for a decision rule δ, the risk function is the expectation of the loss $L[\mathrm{P}, \delta(\boldsymbol{x})]$, regarding the latter as a random variable. That is*

$$R(\mathrm{P}, \delta) = \mathrm{E}_{\mathrm{P}}(L[\mathrm{P}, \delta(\boldsymbol{X})])\,.$$

In case the model $\mathcal{P}$ is parameterised and $\theta \in \Theta$ is the true value of the parameter, then the risk function is defined, accordingly, as

$$R(\theta, \delta) = \mathrm{E}_{\theta}(L[\theta, \delta(\boldsymbol{X})])\,.$$

Definition A.30 (Bias)**.** *Consider a parameter $\theta \in \Theta$, a function $g : \Theta \to \mathbb{R}$ and an estimator $\delta(\boldsymbol{X})$ of $g(\theta)$. The bias of $\delta(\boldsymbol{X})$ as an estimator of $g(\theta)$ is defined as*

$$\mathrm{Bias}[\delta(\boldsymbol{X})] = \mathrm{E}_{\theta}[\delta(\boldsymbol{X}) - g(\theta)] = \mathrm{E}_{\theta}[\delta(\boldsymbol{X})] - g(\theta)\,.$$

Definition A.31 (Unbiased Estimator)**.** *Consider a parameter $\theta \in \Theta$ and a function $g : \Theta \to \mathbb{R}$. An estimator $\delta(\boldsymbol{X})$ of $g(\theta)$ is characterised as unbiased if and only if*

$$\mathrm{Bias}[\delta(\boldsymbol{X})] = 0 \Leftrightarrow \mathrm{E}_{\theta}[\delta(\boldsymbol{X})] = g(\theta)\,,$$

otherwise $\delta(\boldsymbol{X})$ is called biased.

Definition A.32 (Mean Squared Error)**.** *Consider a parameter $\theta \in \Theta$, a function $g : \Theta \to \mathbb{R}$ and an estimator $\delta(\boldsymbol{X})$ of $g(\theta)$. The mean squared error of $\delta(\boldsymbol{X})$ as an estimator of $g(\theta)$ is defined as*

$$\mathrm{MSE}[\delta(\boldsymbol{X})] = \mathrm{E}_{\theta}\Big([\delta(\boldsymbol{X}) - g(\theta)]^2\Big)\,.$$

The following proposition relates the mean squared error of an estimator to its bias.

Proposition A.3. *Consider a parameter $\theta \in \Theta$, a function $g : \Theta \to \mathbb{R}$ and an estimator $\delta(\boldsymbol{X})$ of $g(\theta)$. The mean squared error and the bias of $\delta(\boldsymbol{X})$ as an estimator of $g(\theta)$ are related via the following equation*

$$\mathrm{MSE}[\delta(\boldsymbol{X})] = (\mathrm{Bias}[\delta(\boldsymbol{X})])^2 + \mathrm{Var}[\delta(\boldsymbol{X})]\,.$$

Proof.

$$\begin{aligned}
&\mathrm{MSE}[\delta(\boldsymbol{X})] = \mathrm{E}\Big([\delta(\boldsymbol{X}) - g(\theta)]^2\Big) = \mathrm{E}\Big([\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})] + \mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)]^2\Big) = \\
&\mathrm{E}\Big[(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])^2 + 2\,(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])\,[\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)] + [\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)]^2\Big] = \\
&\mathrm{E}\Big[(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])^2\Big] + 2\mathrm{E}(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])\,[\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)] + [\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)]^2 = \\
&\mathrm{E}\Big[(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])^2\Big] + 2\,(\mathrm{E}[\delta(\boldsymbol{X})] - \mathrm{E}[\delta(\boldsymbol{X})])\,[\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)] + [\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)]^2 = \\
&\mathrm{E}\Big[(\delta(\boldsymbol{X}) - \mathrm{E}[\delta(\boldsymbol{X})])^2\Big] + [\mathrm{E}[\delta(\boldsymbol{X})] - g(\theta)]^2 = \mathrm{Var}[\delta(\boldsymbol{X})] + (\mathrm{Bias}[\delta(\boldsymbol{X})])^2\,.
\end{aligned}$$

□

Definition A.33 (Consistent Estimator). *Consider a parameter $\theta \in \Theta$, a function $g : \Theta \to \mathbb{R}$ and a sequence of estimators $\delta_n = \delta_n(X_1, \ldots, X_n)$, $n \in \{1, 2, \ldots\}$ of $g(\theta)$. $\{\delta_n : n \geq 1\}$ is characterised as a consistent sequence of estimators of $g(\theta)$ if for every $\theta \in \Theta$*

$$\delta_n \xrightarrow{p} g(\theta)\,, \text{ as } n \to +\infty.$$

The following theorem gives some ways to prove consistency for a sequence of estimators (see Theorem 8.2 in Lehmann and Casella 1998).

Theorem A.2. *Consider a parameter $\theta \in \Theta$, a function $g : \Theta \to \mathbb{R}$ and a sequence of estimators $\{\delta_n : n \geq 1\}$.*

(i) *If for all $\theta \in \Theta$ it holds that*

$$\mathrm{MSE}(\delta_n) = \mathrm{E}_\theta\Big([\delta_n - g(\theta)]^2\Big) \xrightarrow[n \to +\infty]{} 0,$$

then $\{\delta_n : n \geq 1\}$ is a consistent sequence of estimators for $g(\theta)$.

(ii) *If for all $\theta \in \Theta$ it holds that*

$$\mathrm{Bias}(\delta_n) \xrightarrow[n \to +\infty]{} 0 \Leftrightarrow \mathrm{E}(\delta_n) \xrightarrow[n \to +\infty]{} g(\theta)\,,$$

and

$$\mathrm{Var}_\theta(\delta_n) \xrightarrow[n \to +\infty]{} 0,$$

then $\{\delta_n : n \geq 1\}$ is a consistent sequence of estimators for $g(\theta)$.

(iii) *As a direct result of* (ii)*, if δ_n is unbiased for all $n \geq 1$ and*

$$\mathrm{Var}_\theta(\delta_n) \xrightarrow[n \to +\infty]{} 0,$$

then $\{\delta_n : n \geq 1\}$ *is a consistent sequence of estimators for* $g(\theta)$.

Definition A.34 ($\sqrt{n}$-consistent Estimator). *Consider a parameter* $\theta \in \Theta$, *a function* $g : \Theta \to \mathbb{R}$ *and a sequence of estimators* $\{\delta_n : n \geq 1\}$. $\{\delta_n : n \geq 1\}$ *is said to be a* $\sqrt{n}$*-consistent sequence of estimators for* $g(\theta)$ *if* $\sqrt{n}\,[\delta_n - g(\theta)]$ *is bounded in probability, that is, if*

$$\sqrt{n}\,[\delta_n - g(\theta)] = \mathrm{O}_p(1) \Leftrightarrow [\delta_n - g(\theta)] = \mathrm{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

Definition A.35 (Asymptotically Normal Estimator). *Consider a parameter* $\theta \in \Theta$, *a function* $g : \Theta \to \mathbb{R}$ *and a sequence of estimators* $\{\delta_n : n \geq 1\}$. $\{\delta_n : n \geq 1\}$ *is said to be an asymptotically normal sequence of estimators for* $g(\theta)$ *if there exists a function* $r(\theta) \in (0, +\infty)$, *such that*

$$\sqrt{n}\,[\delta_n - g(\theta)] \xrightarrow{d} N[0, r(\theta)]\,.$$

The function $r(\theta)$ *is referred to as the asymptotic variance of* $\{\delta_n : n \geq 1\}$.

A.3 Multivariate Statistical Analysis

In this section some basic definitions of multivariate statistical analysis are introduced and the multivariate Normal distribution is presented, that is used in Chapters 2 and 3. We mainly follow Härdle and Simar (2007) and Trevezas (2021). The interested reader is also referred to Anderson (2003) and Johnson et al. (2002).

A.3.1 Basic Elements of Multivariate Statistical Analysis

Definition A.36 (Cumulative Distribution Function of Random Vectors). *The cumulative distribution function of a random vector* $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)'$, *where for all* $i \in \{1, \ldots, p\}$, X_i *is a random variable, is the joint cumulative distribution function of its elements, that is*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathrm{P}\left(X_1 \leq x_1, \ldots, X_p \leq x_p\right).$$

Definition A.37 (Expectation of Random Vectors). *Consider a random vector* $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)'$. *If* $\mathrm{E}(|X_i|) < +\infty$, *for all* $i \in \{1, \ldots, p\}$, *then the expectation of the vector* $\boldsymbol{X}$ *is defined as*

$$\mathrm{E}(\boldsymbol{X}) = \mathrm{E}\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix},$$

where for all $i \in \{1, \ldots, p\}$

$$\mathrm{E}(X_i) = \begin{cases} \sum\limits_{x_i \in S} x_i \mathrm{P}(X_i = x_i), & \text{if } X_i \text{ is a discrete random variable taking its values in } S \\ \int\limits_{-\infty}^{+\infty} x_i f_{X_i}(x_i)\, dx_i, & \text{if } X_i \text{ is a continuous random variable with probability density function } f_{X_i}(\cdot) \end{cases}$$

Definition A.38 (Expectation of Random Matrices)**.** *Consider random matrix* $\boldsymbol{X} \in \mathbb{R}^{T \times p}$*, that is a matrix whose entries* X_{ij} *are random variables, for all* $i \in \{1, \ldots, T\}$ *and* $j \in \{1, \ldots, p\}$*. If* $\mathrm{E}\left(\left|X_{ij}\right|\right) < +\infty$*, for all* $i \in \{1, \ldots, T\}$ *and* $i \in \{1, \ldots, p\}$*, then the expectation of the matrix* $\boldsymbol{X}$ *is defined as*

$$\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{E}(X_{11}) & \mathrm{E}(X_{12}) & \ldots & \mathrm{E}(X_{1p}) \\ \mathrm{E}(X_{21}) & \mathrm{E}(X_{22}) & \ldots & \mathrm{E}(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(X_{T1}) & \mathrm{E}(X_{T2}) & \ldots & \mathrm{E}(X_{Tp}) \end{pmatrix},$$

where for all $i \in \{1, \ldots, T\}$ *and* $j \in \{1, \ldots, p\}$

$$\mathrm{E}(X_{ij}) = \begin{cases} \sum\limits_{x_{ij} \in S} x_{ij} \mathrm{P}(X_{ij} = x_{ij}), & \text{if } X_{ij} \text{ is a discrete random variable taking its values in } S \\ \int\limits_{-\infty}^{+\infty} x_{ij} f_{X_{ij}}(x_{ij})\, dx_{ij}, & \text{if } X_{ij} \text{ is a continuous random variable with probability density function } f_{X_{ij}}(\cdot) \end{cases}$$

Definition A.39 (Variance - Covariance Matrix)**.** *Consider a random vector* $\boldsymbol{X} \in \mathbb{R}^p$*,* $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)'$*. If* $\mathrm{E}\left(X_i^2\right) < +\infty$*, for all* $i \in \{1, \ldots, p\}$*, then the variance - covariance, or simply the covariance matrix of the vector* $\boldsymbol{X}$ *is defined as*

$$\mathrm{Var}(\boldsymbol{X}) = \left[\mathrm{Cov}\left(X_i, X_j\right)\right]_{i,j \in \{1,\ldots,p\}} = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}\left(X_1, X_p\right) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \ldots & \mathrm{Cov}\left(X_2, X_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(X_p, X_1\right) & \mathrm{Cov}\left(X_p, X_2\right) & \ldots & \mathrm{Cov}\left(X_p, X_p\right) \end{pmatrix},$$

where for all $i, j \in \{1, \ldots, p\}$

$$\mathrm{Cov}\left(X_i, X_j\right) = \mathrm{E}\left(\left[X_i - \mathrm{E}(X_i)\right]\left[X_j - \mathrm{E}\left(X_j\right)\right]\right).$$

The diagonal elements of $\mathrm{Var}(\boldsymbol{X})$ *are the variances* $\mathrm{Var}(X_i)$*, for all* $i \in \{1, \ldots, p\}$*. The covariance matrix of a random vector* $\boldsymbol{X}$ *is positive semidefinite (see Definition A.59).*

Definition A.40 (Covariance Matrix)**.** *Consider the random vectors* $\boldsymbol{X} \in \mathbb{R}^p$*,* $\boldsymbol{X} = (X_1, \ldots, X_p)'$ *and* $\boldsymbol{Y} \in \mathbb{R}^s$*,* $\boldsymbol{Y} = (Y_1, \ldots, Y_s)'$*. If* $\mathrm{E}(X_i^2) < +\infty$*, for all* $i \in \{1, \ldots, p\}$*, and* $\mathrm{E}(Y_j^2) < +\infty$*, for all* $j \in \{1, \ldots, s\}$*, then the covariance matrix of the vectors* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *is defined as*

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \left[\mathrm{Cov}(X_i, Y_j)\right]_{i \in \{1,\ldots,p\},\, j \in \{1,\ldots,s\}} = \begin{pmatrix} \mathrm{Cov}(X_1, Y_1) & \mathrm{Cov}(X_1, Y_2) & \ldots & \mathrm{Cov}(X_1, Y_s) \\ \mathrm{Cov}(X_2, Y_1) & \mathrm{Cov}(X_2, Y_2) & \ldots & \mathrm{Cov}(X_2, Y_s) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, Y_1) & \mathrm{Cov}(X_p, Y_2) & \ldots & \mathrm{Cov}(X_p, Y_s) \end{pmatrix}.$$

If $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}_{p \times s}$*, then the random vectors* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are uncorrelated.*

In the following propositions the properties of the expectation and the covariance matrix are presented (see e.g. Section 4.2 in Härdle and Simar 2007, or Section 2.1 in Trevezas 2021).

Proposition A.4. *Consider two random vectors* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^p$*, for which the corresponding expectations exist. For a matrix* $\boldsymbol{A}$ *and a vector* $\boldsymbol{b}$ *of appropriate dimensions, we have the following relations:*

(i) $\mathrm{E}(\boldsymbol{X}') = [\mathrm{E}(\boldsymbol{X})]'$,

(ii) $\mathrm{E}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{X}) + \boldsymbol{b}$,

(iii) $\mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y})$.

The same relations also hold in case $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are random matrices.*

Proposition A.5. *Consider two random vectors* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^p$*, for which the corresponding covariance matrices exist. For a matrix* $\boldsymbol{A}$ *and a vector* $\boldsymbol{b}$ *of appropriate dimensions, we have the following relations:*

(i) $\mathrm{Var}(\boldsymbol{X}) = \mathrm{E}\left([\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})]\,[\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})]'\right) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}') - \mathrm{E}(\boldsymbol{X})\,\mathrm{E}(\boldsymbol{X}')$,

(ii) $\mathrm{Var}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\,\boldsymbol{A}'$,

(iii) $\mathrm{Var}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{X}) + \mathrm{Var}(\boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{X})$.

Proposition A.6. *Consider two random vectors* $\boldsymbol{X} \in \mathbb{R}^p$*,* $\boldsymbol{Y} \in \mathbb{R}^s$*, for which the corresponding covariance matrices exist. For tow matrices* $\boldsymbol{A}, \boldsymbol{B}$ *and two vectors* $\boldsymbol{b}, \boldsymbol{c}$ *of appropriate dimensions, we have the following relations:*

(i) $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}\left([\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})]\,[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]'\right) = \mathrm{E}(\boldsymbol{X}\boldsymbol{Y}') - \mathrm{E}(\boldsymbol{X})\,\mathrm{E}(\boldsymbol{Y}')$,

(ii) $\text{Cov}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}, \boldsymbol{B}\boldsymbol{Y} + \boldsymbol{c}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X}, \boldsymbol{Y})\,\boldsymbol{B}^{'}$.

Proposition A.7. *Consider two independent random vectors $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{Y} \in \mathbb{R}^s$, for which the corresponding covariance matrices exist. We have the following relations:*

(i) $\text{E}\left(\boldsymbol{X}\boldsymbol{Y}^{'}\right) = \text{E}(\boldsymbol{X})\,\text{E}\left(\boldsymbol{Y}^{'}\right)$,

(ii) $\text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}_{p\times s}$, *i.e. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are also uncorrelated,*

(iii) $\text{Var}(\boldsymbol{X} + \boldsymbol{Y}) = \text{Var}(\boldsymbol{X}) + \text{Var}(\boldsymbol{Y})$, *if $p = s$.*

We now proceed to define the sample analogs of the expectation and the covariance matrix for random vectors.

Definition A.41 (Sample Mean of Random Vectors)**.** *Consider a sample of T observations of the random vectors $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$, where $\boldsymbol{X}_t \in \mathbb{R}^p$, for all $t \in \{1, \ldots, T\}$. The sample mean is defined as*

$$\overline{\boldsymbol{X}}_T = \begin{pmatrix} \overline{\boldsymbol{X}}_1 \\ \overline{\boldsymbol{X}}_2 \\ \vdots \\ \overline{\boldsymbol{X}}_p \end{pmatrix},$$

where for all $i \in \{1, \ldots, p\}$

$$\overline{\boldsymbol{X}}_i = \frac{1}{T}\sum_{t=1}^{T} X_{t,i}.$$

As in the univariate case, for a sample of T i.i.d. random vectors, the sample mean is an unbiased estimator of the expectation (see Example 4.15 in Härdle and Simar 2007).

Definition A.42 (Sample Covariance Matrix)**.** *Consider a sample of T observations of the random vectors $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$, where $\boldsymbol{X}_t \in \mathbb{R}^p$, for all $t \in \{1, \ldots, T\}$. The sample covariance matrix is defined as*

$$\boldsymbol{S}_T = \left(\boldsymbol{S}_{X_i X_j}\right)_{i,j\in\{1,\ldots,p\}},$$

where for all $i, j \in \{1, \ldots, p\}$

$$\boldsymbol{S}_{X_i X_j} = \frac{1}{T}\sum_{t=1}^{T}\left(X_{t,i} - \overline{\boldsymbol{X}}_i\right)\left(X_{t,j} - \overline{\boldsymbol{X}}_j\right).$$

The diagonal elements of the sample covariance matrix are the sample variances

$$\boldsymbol{S}_{X_i X_i} = \frac{1}{T}\sum_{t=1}^{T}\left(X_{t,i} - \overline{\boldsymbol{X}}_i\right)^2.$$

The sample covariance matrix is also positive semidefinite (see Equation (3.20) *in* Härdle and Simar 2007*).*

Definition A.43 (Sample Covariance Matrix of Two Random Vectors)**.** *Consider two samples of T observations of the random vectors $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$ and $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$, where $\boldsymbol{X}_t \in \mathbb{R}^p$ and $\boldsymbol{Y}_t \in \mathbb{R}^s$, for all $t \in \{1, \ldots, T\}$. The sample covariance matrix is defined as*

$$\boldsymbol{S} = \left(\boldsymbol{S}_{X_i Y_j}\right)_{i \in \{1,\ldots,p\},\, j \in \{1,\ldots,s\}},$$

where for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, s\}$

$$\boldsymbol{S}_{X_i Y_j} = \frac{1}{T} \sum_{t=1}^{T} \left(X_{t,i} - \overline{\boldsymbol{X}}_i\right) \left(Y_{t,j} - \overline{\boldsymbol{Y}}_j\right).$$

In case of small samples of T i.i.d. random variables, to correct the bias of the sample estimators given in Definitions A.42 and A.43, the denominator T in the formulas of the sample variance and covariance is replaced by $(T-1)$ (see the paragraph following Equation (3.3) in Härdle and Simar 2007).

A sample of T observations of the random vectors $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$, where $\boldsymbol{X}_t \in \mathbb{R}^p$, for all $i \in \{1, \ldots, p\}$ can be displayed in matrix form as

$$\mathcal{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & \ldots & X_{1,p} \\ X_{2,1} & X_{2,2} & \ldots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{T,1} & X_{T,2} & \ldots & X_{T,p} \end{pmatrix},$$

where the t-th row of the matrix is the t-th observation of the p-dimensional random vector X_t.

The sample mean and the sample covariance matrix of the random vectors $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$ can be expressed in matrix form as (see Equations (3.17) to (3.19) in Härdle and Simar 2007)

(i)
$$\overline{\boldsymbol{X}}_T = \frac{1}{T} \mathcal{X}' \mathbf{1}_T,$$

where $\mathbf{1}_T$ is a T-dimensional vector of ones,

(ii)
$$\boldsymbol{S}_T = \frac{1}{T} \mathcal{X}' \mathcal{X} - \overline{\boldsymbol{X}}_T \overline{\boldsymbol{X}}_T{}' = \frac{1}{T} \left(\mathcal{X}' \mathcal{X} - \frac{1}{T} \mathcal{X}' \mathbf{1}_T \mathbf{1}_T{}' \mathcal{X} \right),$$

which is equivalent to

$$\boldsymbol{S}_T = \frac{1}{T} \sum_{t=1}^{T} \left(X_t - \overline{\boldsymbol{X}}_T\right) \left(X_t - \overline{\boldsymbol{X}}_T\right)'. \tag{A.1}$$

As in the univariate case, for small samples of T i.i.d. random vectors the unbiased estimator of the covariance matrix is $\boldsymbol{S} = \frac{T}{T-1}\boldsymbol{S}_T$ (see Example 4.15 in Härdle and Simar 2007).

A.3.2 Multivariate Normal Distribution

Definition A.44 (Multivariate Normal Distribution). *A random vector $\boldsymbol{X} \in \mathbb{R}^p$ that follows the multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, where Σ is a positive definite matrix (see Definition A.59), denoted by $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, has the following probability density function*

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}\left[\det(\Sigma)\right]^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

The following theorem shows how the multivariate Normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$ is related to the standard multivariate Normal distribution $N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$.

Theorem A.3. *Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{Y} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, where $\Sigma^{-1/2} = \left(\Sigma^{1/2}\right)^{-1}$ and $\Sigma^{1/2}$ is the square root of Σ (see Definition A.60). Then $\boldsymbol{Y} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$, which means that the elements Y_i are independent random variables from the univariate standard Normal distribution $N(0,1)$, for all $i \in \{1, \ldots, p\}$.*

Proof. See the proof of Theorem 4.5 in Härdle and Simar (2007). □

We also have the following result for affine transformations (see, e.g., Result 4.3 in Johnson et al. 2002, or Theorem 4.6 in Härdle and Simar 2007).

Theorem A.4. *Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ and a vector $\boldsymbol{b} \in \mathbb{R}^p$ and a random vector $\boldsymbol{Y} \in \mathbb{R}^p$, such that $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, it holds that $\boldsymbol{Y} \sim N_p\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}'\right)$.*

Proof. See the proof of Result 4.3 in Johnson et al. (2002). □

More information about the multivariate Normal distribution can be found in, e.g., Chapter 4 of Johnson et al. (2002), Chapter 2 of Anderson (2003), Section 4.4 and Chapter 5 of Härdle and Simar (2007), or Section 2.3 of Trevezas (2021).

A.4 Time Series

This section contains some basic notions of univariate time series theory that are used in Chapter 3. The main purpose of this presentation is to provide a sense of completeness and not a proper introduction on the topic of time series. The interested reader is referred to the works of Brockwell and Davis (1991) and Shumway and Stoffer (2011), which we follow in this section.

Definition A.45 (The Autocovariance Function). *If $\{X_t : t \in T\}$ is a stochastic process such that $\mathrm{Var}(X_t) < +\infty$ for each $t \in T$, then the autocovariance function $c(\cdot,\cdot)$ of $\{X_t : t \in T\}$ is defined as*

$$c(t,s) = \mathrm{Cov}(X_t, X_s) = \mathrm{E}([X_t - \mathrm{E}(X_t)]\,[X_s - \mathrm{E}(X_s)]) \qquad (t,s \in T)\,.$$

Definition A.46 (Strictly Stationary Time Series). *The time series $\{X_t : t \in \mathbb{Z}\}$ is said to strictly stationary if the joint distributions of the random variables $(X_{t_1}, \ldots, X_{t_k})$ and $\left(X_{t_{1+h}}, \ldots, X_{t_{k+h}}\right)$ are the same for all positive integers k and all $t_1, \ldots, t_k \in \mathbb{Z}$.*

Definition A.47 (Weakly Stationary Time Series). *The time series $\{X_t : t \in \mathbb{Z}\}$ is said to be weakly stationary if*

(i) $\mathrm{E}\left(|X_t|^2\right) < +\infty$, *for all* $t \in \mathbb{Z}$,

(ii) $\mathrm{E}(X_t) = \mu$, *for all* $t \in \mathbb{Z}$,

(iii) $c(t,s) = c(t+h, s+h)$, *for all* $t,s,h \in \mathbb{Z}$.

Remark. *If a time series is strictly stationary with finite second moments, then it is also weakly stationary. The converse is not true in general (see the discussion preceding* Definition 1.3.4 *in* Brockwell and Davis 1991, *or the discussion following* Definition 1.7 *in* Shumway and Stoffer 2011*).*

We have the following result for the autocovariance of a (strictly or weakly) stationary time series (see Remark 2 in Chapter 1 of Brockwell and Davis 1991, or Shumway and Stoffer 2011).

Remark. *If $\{X_t : t \in \mathbb{Z}\}$ is a (strictly or weakly) stationary time series, then its autocovariance $c(t,s)$ depends on t and s only through their difference $|t-s|$, and not on the specific moments t and s, for all $t,s \in \mathbb{Z}$. More formally, it holds that*

$$c(t,s) = c(t-s, 0) \qquad (t,s \in \mathbb{Z})\,.$$

As a result, we have the following definition for the autocovariance of a (strictly or weakly) stationary time series.

Definition A.48 (The Autocovariance Function For Stationary Time Series). *The autocovariance function of a (strictly or weakly) stationary time series $\{X_t : t \in \mathbb{Z}\}$ is defined as*

$$c(h) = c(h,0) = \mathrm{Cov}(X_{t+h}, X_t) = \mathrm{E}([X_{t+h} - \mathrm{E}(X_{t+h})]\,[X_t - \mathrm{E}(X_t)]) \qquad (t,h \in \mathbb{Z})\,.$$

The quantity $c(h)$ is referred to as the value of the autocovariance function of $\{X_t : t \in \mathbb{Z}\}$ at "lag" h.

The autocovariance function of a (strictly or weakly) stationary time series $\{X_t : t \in \mathbb{Z}\}$ possesses three properties given by the following proposition (see Proposition 1.5.1 in Brockwell and Davis 1991, or the discussion following Example 1.20 in Shumway and Stoffer 2011).

Proposition A.8. *If $c(\cdot)$ is the autocovariance of a (strictly or weakly) stationary time series $\{X_t : t \in \mathbb{Z}\}$, then it satisfies the following relations:*

(i) $c(0) \geq 0$,

(ii) $|c(h)| \leq c(0)$, *for all* $h \in \mathbb{Z}$,

(iii) $c(h) = c(-h)$, *for all* $h \in \mathbb{Z}$, *which means that it is an even function.*

Proof. (i) For the first property, for all $t \in \mathbb{Z}$, we have

$$c(0) = \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t) \geq 0.$$

(ii) The second property is a direct implication of the Cauchy–Schwarz inequality,

$$|c(h)| = |\mathrm{Cov}(X_{t+h}, X_t)| \leq [\mathrm{Var}(X_{t+h})]^{1/2} [\mathrm{Var}(X_t)]^{1/2} = c(0)\,.$$

(iii) The third and last property is proved as follows:

$$c(-h) = \mathrm{Cov}(X_{t-h}, X_t) = \mathrm{Cov}(X_t, X_{t+h}) = c(h)\,.$$

□

Definition A.49 (The Sample Autocovariance Function)**.** *Consider a (strictly or weakly) stationary time series $\{X_t : t \in \mathbb{Z}\}$ with autocovariance function $c(\cdot)$. For a sequence of observations $(x_1, \ldots, x_T)$ of the process $\{X_t : t \in \mathbb{Z}\}$, the sample autocovariance function is defined as*

$$\hat{c}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (x_{t+h} - \bar{x})\,(x_t - \bar{x}),$$

with $\hat{c}(-h) = \hat{c}(h)$, for $h \in \{0, \ldots, T-1\}$, where $\bar{x}$ is the sample mean, $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$.

We now give a brief definition of the spectral density of a univariate time series. More information on the topic of spectral analysis of univariate or multivariate time series can be found in Chapters 4, 10 and 11 of Brockwell and Davis (1991), or in Chapter 4 and Appendix C of Shumway and Stoffer (2011).

Definition A.50 (Spectral Density). *Consider a (strictly or weakly) stationary process* $\{X_t : t \in \mathbb{Z}\}$ *whose autocovariance function* $c(\cdot)$ *is absolutely summable, that is*

$$\sum_{h=-\infty}^{+\infty} |c(h)| < +\infty.$$

The spectral density of $\{X_t : t \in \mathbb{Z}\}$ *is defined as*

$$f(\omega) = \sum_{h=-\infty}^{+\infty} c(h)\, e^{-2\pi i \omega h} \qquad (\omega \in [-1/2, 1/2])\,.$$

If the autocovariance function $c(\cdot)$ of a (strictly or weakly) stationary process $\{X_t : t \in \mathbb{Z}\}$ is absolutely summable, then it has the following representation, which is an inverse transform of the spectral density function $f(\cdot)$ (see Property 4.2 in Shumway and Stoffer 2011)

$$c(h) = \int_{-1/2}^{1/2} e^{2\pi i \omega h} f(\omega)\, d\omega \qquad (h \in \mathbb{Z})\,.$$

For the spectral density of a (strictly or weakly) stationary process $\{X_t : t \in \mathbb{Z}\}$ we have that (see the discussion following Property 4.2 in Shumway and Stoffer 2011)

(i) $f(\omega) \geq 0$, for all $\omega \in [-1/2, 1/2]$,

(ii) $f(\omega) = f(-\omega)$ and $f(\omega) = f(1-\omega)$, for all $\omega \in [-1/2, 1/2]$,

(iii) $c(0) = \mathrm{Var}(X_t) = \int_{-1/2}^{1/2} f(\omega)\, d\omega$.

A.5 Orders of Magnitude and Rate of Convergence

In this section we present the notions of boundedness and order of magnitude along with the big-O and little-o notation, as well as their stochastic analogs. We mainly follow Sections 1.4 and 2.1 in Lehmann (1999), but also Section 2.3 in Bamdorff-Nielsen and Cox (1989) and Section 2.2 in van der Vaart (1998). The interested reader is also referred to Sections 1.2 and 1.3 in de Bruijn (1958).

Definition A.51 (Boundedness). *A real-valued sequence* $\{a_n : n \in \mathbb{N}\}$ *is said to be bounded if there exist a constant* $M > 0$ *and a positive integer* n_0*, such that for all* $n > n_0$

$$|a_n| \leq M.$$

Definition A.52 (Bachmann–Landau Notation or Big-O Notation). *Consider two real-valued sequences* $\{a_n : n \in \mathbb{N}\}$ *and* $\{b_n : n \in \mathbb{N}\}$. *We write:*

(i) $a_n = \mathrm{O}(1)$, *if* $\{a_n : n \in \mathbb{N}\}$ *is a bounded sequence,*

(ii) $a_n = \mathrm{O}(b_n)$, *if and only if there exist a constant* $M > 0$ *and a positive integer* n_0, *such that for all* $n > n_0$

$$|a_n| \leq M\,|b_n|\,.$$

If $b_n \neq 0$ *for all* $n \in \mathbb{N}$, *then* $a_n = \mathrm{O}(b_n)$ *means that the sequence* $\left\{\dfrac{a_n}{b_n} : n \in \mathbb{N}\right\}$ *is bounded, that is, there exist a constant* $M > 0$ *and a positive integer* n_0, *such that for all* $n > n_0$

$$\left|\frac{a_n}{b_n}\right| \leq M \Leftrightarrow \frac{a_n}{b_n} = \mathrm{O}(1)\,.$$

It is said that the order of $\{a_n : n \in \mathbb{N}\}$ *is smaller than or equal to the order of* $\{b_n : n \in \mathbb{N}\}$.

In the following lemma we give some properties of big-O notation (see Lemma 1.4.1 and Problem 4.17 of Chapter 1 in Lehmann 1999).

Lemma A.1. (i) *If* $a_n = \mathrm{O}(b_n)$ *and* $b_n = \mathrm{O}(c_n)$, *then* $a_n = \mathrm{O}(c_n)$.

(ii) *If* $a_n = \mathrm{O}(b_n)$, *then* $ca_n = \mathrm{O}(b_n)$, *for any constant* $c \neq 0$.

(iii) *For any real-valued sequence* $\{c_n : n \in \mathbb{N}\}$, *with* $c_n \neq 0$ *for all* $n \in \mathbb{N}$, *if* $a_n = \mathrm{O}(b_n)$, *then* $c_n a_n = \mathrm{O}(c_n b_n)$.

(iv) *If* $d_n = \mathrm{O}(b_n)$ *and* $e_n = \mathrm{O}(c_n)$, *then* $d_n e_n = \mathrm{O}(b_n c_n)$.

Definition A.53 (Sequences of the Same Order)**.** *Consider two real-valued sequences* $\{a_n : n \in \mathbb{N}\}$ *and* $\{b_n : n \in \mathbb{N}\}$, *with* $b_n \neq 0$ *for all* $n \in \mathbb{N}$. *The sequences* $\{a_n : n \in \mathbb{N}\}$ *and* $\{b_n : n \in \mathbb{N}\}$ *are said to be of the same order if there exist constants* m, M, *with* $0 < m < M < +\infty$, *and a positive integer* n_0, *such that for all* $n > n_0$

$$m < \left|\frac{a_n}{b_n}\right| < M$$

and it is denoted by $a_n \asymp b_n$.

The sequences $\{a_n : n \in \mathbb{N}\}$ and $\{b_n : n \in \mathbb{N}\}$ are of the same order if and only if $a_n = \mathrm{O}(b_n)$ and $b_n = \mathrm{O}(a_n)$ (see Problem 4.16 Lehmann 1999).

For two sequences of the same order we have the following lemma (see Lemma 1.4.2 in Lehmann 1999).

Lemma A.2. *If* $a_n \asymp b_n$, *then* $ca_n \asymp b_n$, *for any constant* $c \neq 0$.

Definition A.54 (Little-o Notation)**.** *Consider two real-valued sequences* $\{a_n : n \in \mathbb{N}\}$ *and* $\{b_n : n \in \mathbb{N}\}$, *with* $b_n \neq 0$ *for all* $n \in \mathbb{N}$. *We write:*

(i) $a_n = o(1)$, *if*

$$a_n \xrightarrow[n\to+\infty]{} 0,$$

(ii) $a_n = o(b_n)$, *if*

$$\frac{a_n}{b_n} \xrightarrow[n\to+\infty]{} 0.$$

It is said that the order of $\{a_n : n \in \mathbb{N}\}$ *is smaller than the order of* $\{b_n : n \in \mathbb{N}\}$.

An interpretation of Definition A.54 is that in case both $\{a_n : n \in \mathbb{N}\}$ and $\{b_n : n \in \mathbb{N}\}$ tend to infinity, then the rate at which $\{a_n : n \in \mathbb{N}\}$ tends to infinity is slower than the corresponding rate of $\{b_n : n \in \mathbb{N}\}$, while if both sequences tend to zero, then $\{a_n : n \in \mathbb{N}\}$ tends to zero faster than $\{b_n : n \in \mathbb{N}\}$ (see the paragraph following Definition 1.4.1 in Lehmann 1999).

In the following lemma we give some properties of little-o notation (see Lemma 1.4.1 in Lehmann 1999).

Lemma A.3. (i) *If* $a_n = o(b_n)$ *and* $b_n = o(c_n)$, *then* $a_n = o(c_n)$.

(ii) *If* $a_n = o(b_n)$, *then* $ca_n = o(b_n)$, *for any constant* $c \neq 0$.

(iii) *For any real-valued sequence* $\{c_n : n \in \mathbb{N}\}$, *with* $c_n \neq 0$ *for all* $n \in \mathbb{N}$, *if* $a_n = o(b_n)$, *then* $c_n a_n = o(c_n b_n)$.

(iv) *If* $d_n = o(b_n)$ *and* $e_n = o(c_n)$, *then* $d_n e_n = o(b_n c_n)$.

We now present the stochastic analogs of the theory presented above.

Definition A.55 (Boundedness in Probability)**.** *A sequence of random variables (or random vectors)* $\{X_n : n \in \mathbb{N}\}$ *is said to be bounded in probability if for any* $\epsilon > 0$, *there exist a constant* $M := M(\epsilon)$ *and a positive integer* $n_0 := n_0(\epsilon)$, *such that for all* $n > n_0$

$$P(|X_n| \leq M) \geq 1 - \epsilon.$$

For a sequence of random vectors, the absolute value $|\cdot|$ *in the inequality above, is replaced by the Euclidean norm* $||\cdot||_2$ *(see the paragraph preceding* 2.4 Theorem *in* van der Vaart 1998*).*

Definition A.56 (Stochastic Big-O Notation)**.** *Consider two sequences of random variables (or random vectors)* $\{X_n : n \in \mathbb{N}\}$ *and* $\{Y_n : n \in \mathbb{N}\}$. *We write:*

(i) $X_n = O_p(1)$, *if* $\{X_n : n \in \mathbb{N}\}$ *is a bounded in probability sequence,*

(ii) $X_n = \mathrm{O}(Y_n)$, *if for any* $\epsilon > 0$, *there exist a constant* $M := M(\epsilon)$ *and a positive integer* $n_0 := n_0(\epsilon)$, *such that for all* $n > n_0$

$$\mathrm{P}(|X_n| \leq M\,|Y_n|) \geq 1 - \epsilon.$$

If $Y_n \neq 0$ *for all* $n \in \mathbb{N}$ *almost surely, then* $X_n = \mathrm{O}(Y_n)$ *means that the sequence* $\left\{\frac{X_n}{Y_n} : n \in \mathbb{N}\right\}$ *is bounded in probability, that is, for any* $\epsilon > 0$, *there exist a constant* $M := M(\epsilon)$ *and a positive integer* $n_0 := n_0(\epsilon)$, *such that for all* $n > n_0$

$$\mathrm{P}\left(\left|\frac{X_n}{Y_n}\right| \leq M\right) \geq 1 - \epsilon \Leftrightarrow \frac{X_n}{Y_n} = \mathrm{O}_p(1)\,.$$

It is said that the order of $\{X_n : n \in \mathbb{N}\}$ *is smaller than or equal to the order of* $\{Y_n : n \in \mathbb{N}\}$ *in probability.*

For sequences of random vectors, the absolute values $|\cdot|$ *in the inequalities above, are replaced by the Euclidean norm* $||\cdot||_2$.

The following lemma is the stochastic analog of Lemma A.1 (see Lemma 1.4.1 and Problem 1.11 (ii) of Chapter 2 in Lehmann 1999).

Lemma A.4. (i) *If* $X_n = \mathrm{O}_p(Y_n)$ *and* $Y_n = \mathrm{O}_p(Z_n)$, *then* $X_n = \mathrm{O}_p(Z_n)$.

(ii) *If* $X_n = \mathrm{O}_p(Y_n)$, *then* $cX_n = \mathrm{O}_p(Y_n)$, *for any constant* $c \neq 0$.

(iii) *For any real valued sequence* $\{c_n : n \in \mathbb{N}\}$, *with* $c_n \neq 0$ *for all* $n \in \mathbb{N}$, *if* $X_n = \mathrm{O}_p(Y_n)$, *then* $c_n X_n = \mathrm{O}_p(c_n Y_n)$.

(iv) *For any sequence of random variables (or random vectors)* $\{Z_n : n \in \mathbb{N}\}$, *with* $Z_n \neq 0$ *for all* $n \in \mathbb{N}$ *almost surely, if* $X_n = \mathrm{O}_p(Y_n)$, *then* $Z_n X_n = \mathrm{O}_p(Z_n Y_n)$.

(v) *If* $U_n = \mathrm{O}_p(Y_n)$ *and* $V_n = \mathrm{O}_p(Z_n)$, *then* $U_n V_n = \mathrm{O}_p(Y_n Z_n)$.

Definition A.57 (Sequences of Same Order in Probability)**.** *Consider two sequences of random variables (or random vectors)* $\{X_n : n \in \mathbb{N}\}$ *and* $\{Y_n : n \in \mathbb{N}\}$, *with* $Y_n \neq 0$ *for all* $n \in \mathbb{N}$ *almost surely. The sequences* $\{X_n : n \in \mathbb{N}\}$ *and* $\{Y_n : n \in \mathbb{N}\}$ *are said to be of the same order in probability if for any* $\epsilon > 0$, *there exist constants* $m := m(\epsilon)$ *and* $M := M(\epsilon)$, *with* $0 < m < M < +\infty$, *and a positive integer* $n_0 := n_0(\epsilon)$, *such that for all* $n > n_0$

$$\mathrm{P}\left(m < \left|\frac{X_n}{Y_n}\right| < M\right) \geq 1 - \epsilon,$$

and it is denoted by $X_n \asymp_p Y_n$.

For sequences of random vectors, the absolute value $|\cdot|$ in the inequality above, is replaced by the Euclidean norm $||\cdot||_2$.

Definition A.58 (Stochastic Little-o Notation)**.** *Consider two sequences of random variables (or random vectors) $\{X_n : n \in \mathbb{N}\}$ and $\{Y_n : n \in \mathbb{N}\}$, with $Y_n \neq 0$ for all $n \in \mathbb{N}$ almost surely. We write:*

(i) $X_n = o_p(1)$, *if and only if*

$$X_n \xrightarrow{p} 0,$$

(ii) $X_n = o(Y_n)$, *if and only if*

$$\frac{X_n}{Y_n} \xrightarrow{p} 0.$$

It is said that the order of $\{X_n : n \in \mathbb{N}\}$ is smaller than the order of $\{Y_n : n \in \mathbb{N}\}$ in probability.

The following lemma is the stochastic analog of Lemma A.3 (see the discussion following Definition 2.1.3 and Problem 1.11 (ii) of Chapter 2 in Lehmann 1999).

Lemma A.5. (i) *If $X_n = o_p(Y_n)$ and $Y_n = o_p(Z_n)$, then $X_n = o_p(Z_n)$.*

(ii) *If $X_n = o_p(Y_n)$, then $cX_n = o_p(Y_n)$, for any constant $c \neq 0$.*

(iii) *For any real-valued sequence$\{c_n : n \in \mathbb{N}\}$, with $c_n \neq 0$ for all $n \in \mathbb{N}$, if $X_n = o_p(Y_n)$, then $c_n X_n = o_p(c_n Y_n)$.*

(iv) *For any sequence of random variables (or random vectors) $\{Z_n : n \in \mathbb{N}\}$, with $Z_n \neq 0$ for all $n \in \mathbb{N}$ almost surely, if $X_n = o_p(Y_n)$, then $Z_n X_n = o_p(Z_n Y_n)$.*

(v) *If $U_n = o_p(Y_n)$ and $V_n = o_p(Z_n)$, then $U_n V_n = o_p(Y_n Z_n)$.*

(vi) *For any real valued sequence $\{c_n : n \in \mathbb{N}\}$, with $c_n \neq 0$ for all $n \in \mathbb{N}$, $c_n\, o_p(X_n) = o_p(c_n X_n)$.*

(vii) *For any sequence of random variables (or random vectors) $\{Z_n : n \in \mathbb{N}\}$, with $Z_n \neq 0$ for all $n \in \mathbb{N}$ almost surely, $Z_n\, o_p(X_n) = o_p(Z_n X_n)$.*

For the last two statements see (2.1.19) *following* Definition 2.1.3 *in* Lehmann (1999).

We now give some further relations involving O, o, O_p and o_p that can be found in Exercise 2.6 in Further Results and Exercises in Bamdorff-Nielsen and Cox (1989) and in Section 2.2 in van der Vaart (1998). Some of these relations are special cases of the properties presented in Lemmas A.1, A.3, A.4 and A.5.

Lemma A.6. (i) $o(1) + o(1) = o(1)$,

(ii) $o(1) + O(1) = O(1)$,

(iii) $\mathrm{O}(1)\,\mathrm{o}(1) = \mathrm{o}(1)$,

(iv) $\dfrac{1}{1+\mathrm{o}(1)} = \mathrm{O}(1)$,

(v) $\mathrm{o}[\mathrm{O}(1)] = \mathrm{o}(1)$,

(vi) $\mathrm{o}_p(1) + \mathrm{o}_p(1) = \mathrm{o}_p(1)$,

(vii) $\mathrm{o}_p(1) + \mathrm{O}_p(1) = \mathrm{O}_p(1)$,

(viii) $\mathrm{O}_p(1)\,\mathrm{o}_p(1) = \mathrm{o}_p(1)$,

(ix) $\dfrac{1}{1+\mathrm{o}_p(1)} = \mathrm{O}_p(1)$,

(x) $\mathrm{o}_p\big[\mathrm{O}_p(1)\big] = \mathrm{o}_p(1)$,

(xi) $\mathrm{O}(n^{-a})\,\mathrm{O}\big(n^{-b}\big) = \mathrm{O}\big(n^{-a-b}\big)$,

(xii) $\mathrm{O}(n^{-a})\,\mathrm{o}\big(n^{-b}\big) = \mathrm{o}\big(n^{-a-b}\big)$,

(xiii) $\mathrm{o}(n^{-a})\,\mathrm{o}\big(n^{-b}\big) = \mathrm{o}\big(n^{-a-b}\big)$,

(xiv) $\mathrm{O}_p(n^{-a})\,\mathrm{O}_p\big(n^{-b}\big) = \mathrm{O}_p\big(n^{-a-b}\big)$,

(xv) $\mathrm{O}_p(n^{-a})\,\mathrm{o}_p\big(n^{-b}\big) = \mathrm{o}_p\big(n^{-a-b}\big)$,

(xvi) $\mathrm{o}_p(n^{-a})\,\mathrm{o}_p\big(n^{-b}\big) = \mathrm{o}_p\big(n^{-a-b}\big)$,

(xvii) $\mathrm{O}_p(n^{-a})\,\mathrm{O}\big(n^{-b}\big) = \mathrm{O}_p\big(n^{-a-b}\big)$,

(xviii) $\mathrm{O}_p(n^{-a})\,\mathrm{o}\big(n^{-b}\big) = \mathrm{o}_p\big(n^{-a-b}\big)$,

(xix) $\mathrm{o}_p(n^{-a})\,\mathrm{O}\big(n^{-b}\big) = \mathrm{o}_p\big(n^{-a-b}\big)$,

(xx) $\mathrm{o}_p(n^{-a})\,\mathrm{o}\big(n^{-b}\big) = \mathrm{o}_p\big(n^{-a-b}\big)$.

A.6 Linear Algebra

Proposition A.9. *For any square matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, it holds that*

$$\mathrm{tr}\Big(\boldsymbol{A}\boldsymbol{A}^{'}\Big) = \sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_{ij}^2,$$

where α_{ij}, $i, j \in \{1,\ldots,p\}$ is the (i,j)-th entry of $\boldsymbol{A}$.

Proof. The diagonal elements of the matrix $\boldsymbol{A}\boldsymbol{A}^{'}$ are

$$\boldsymbol{A}\boldsymbol{A}^{'} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \dots & \alpha_{pp} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{21} & \dots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \dots & \alpha_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \dots & \alpha_{pp} \end{pmatrix} =$$

$$\begin{pmatrix} \alpha_{11}^2 + \alpha_{12}^2 + \ldots + \alpha_{1p}^2 & \dots & \dots & \dots \\ \dots & \alpha_{21}^2 + \alpha_{22}^2 + \ldots + \alpha_{2p}^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & \alpha_{p1}^2 + \alpha_{p2}^2 + \ldots + \alpha_{pp}^2 \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{j=1}^{p} \alpha_{1j}^2 & \dots & \dots & \dots \\ \dots & \sum_{j=1}^{p} \alpha_{2j}^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & \sum_{j=1}^{p} \alpha_{pj}^2 \end{pmatrix}.$$

Therefore, $\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^{'}\right) = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_{ij}^2$. □

Definition A.59 (Positive Definite and Semidefinite Matrix)**.** *A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is characterised as*

(i) *positive definite if and only if for all $\boldsymbol{x} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$*

$$\boldsymbol{x}^{'}\boldsymbol{A}\boldsymbol{x} > 0,$$

(ii) *positive semidefinite if and only if for all $\boldsymbol{x} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$*

$$\boldsymbol{x}^{'}\boldsymbol{A}\boldsymbol{x} \geq 0.$$

It is clear that if $\boldsymbol{A}$ is positive definite, then it is also positive semidefinite.

Proposition A.10. *All the eigenvalues of a positive definite matrix are positive. All the eigenvalues of a positive semidefinite matrix are nonnegative.*

Proof. See the proof of Observation 7.1.4 in Horn and Johnson (2012). □

Corollary A.4.1. *The trace and the determinant of a positive definite matrix are positive. The trace and the determinant of a positive semidefinite matrix are nonnegative.*

Proof. Corollary A.4.1 is a direct result of Proposition A.10, as the trace of a matrix is equal to the sum of its eigenvalues and the determinant is equal to the product of its eigenvalues. □

Definition A.60 (Square Root of a Positive (Semi)Definite Matrix). *If $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is a positive (semi)definite matrix, then it has a unique positive (semi)definite square root, denote by $\boldsymbol{A}^{1/2}$, such that*

$$\left(\boldsymbol{A}^{1/2}\right)' \boldsymbol{A}^{1/2} = \boldsymbol{A}^{1/2}\left(\boldsymbol{A}^{1/2}\right)' = \left(\boldsymbol{A}^{1/2}\right)^2 = \boldsymbol{A}.$$

Proposition A.11. *For any positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^p$ it holds that*

$$||\boldsymbol{A}||_{\mathrm{F}}^2 \leq \frac{1}{p}\left[\mathrm{tr}(\boldsymbol{A})\right]^2.$$

Proof. Consider a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^p$ with eigenvalues λ_i, $i \in \{1,\ldots,p\}$. We have

$$||\boldsymbol{A}||_{\mathrm{F}}^2 = \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}^2\right) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i^2 = \frac{1}{p}\left[\left(\sum_{i=1}^{p}\lambda_i\right)^2 - \sum_{i=1}^{p}\sum_{\substack{j=1\\ j\neq i}}^{p}\lambda_i\lambda_j\right] \leq \frac{1}{p}\left(\sum_{i=1}^{p}\lambda_i\right)^2 = \frac{1}{p}\left[\mathrm{tr}(\boldsymbol{A})\right]^2,$$

since, by Proposition A.10, $\lambda_i \geq 0$, for all $i \in \{1,\ldots,p\}$, therefore $\lambda_i\lambda_j \geq 0$, for all $i,j \in \{1,\ldots,p\}$. □

Below are presented some relations from matrix calculus that are used in Subsection 2.3.2 and Section 3.1, to derive the formulas of μ_k^{o}, $\tilde{\Sigma}_k^{\mathrm{o}}$, $\hat{\mu}$ and Σ^s (see Equations (69), (81), (57), (61) and (124), respectively, in Petersen and Pedersen 2012).

$$\frac{\partial \boldsymbol{x}'\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}', \tag{A.2}$$

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}'\left(\boldsymbol{A}+\boldsymbol{A}'\right), \tag{A.3}$$

$$\frac{\partial \log(|\det(\boldsymbol{X})|)}{\partial \boldsymbol{X}} = \left(\boldsymbol{X}^{-1}\right)' = \left(\boldsymbol{X}'\right)^{-1}, \tag{A.4}$$

$$\frac{\partial \boldsymbol{a}'\boldsymbol{X}^{-1}\boldsymbol{b}}{\partial \boldsymbol{X}} = -\left(\boldsymbol{X}^{-1}\right)'\boldsymbol{a}\boldsymbol{b}'\left(\boldsymbol{X}^{-1}\right)' = -\left(\boldsymbol{X}'\right)^{-1}\boldsymbol{a}\boldsymbol{b}'\left(\boldsymbol{X}'\right)^{-1}, \tag{A.5}$$

$$\frac{\partial \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}\right)}{\partial \boldsymbol{X}} = -\left(\boldsymbol{X}^{-1}\boldsymbol{B}\boldsymbol{A}\boldsymbol{X}^{-1}\right)' = -\left(\boldsymbol{X}'\right)^{-1}\boldsymbol{A}'\boldsymbol{B}'\left(\boldsymbol{X}'\right)^{-1}. \tag{A.6}$$

Appendix B

Technical Result of Subsection 2.3.2

By Equations (2.26), (2.47) and (2.51) we get the two following expressions for $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right)$, where we have omitted the factor $-\frac{p}{2}\log(2\pi)$, since it does not depend on any of the model's parameters.

$$\begin{aligned}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}\right) &= \left[\sum_{k=1}^{K}\gamma_k^{(m)}(1)\log\left[\left(p_k^{(0)}\right)^{(m)}\right]\right] + \left[\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{kl}^{(m)}(t-1)\log\left(p_{kl}^{(m)}\right)\right] \\ &\quad - \frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\gamma_k^{(m)}(t)\log[\det(\Sigma_k)]\right) \\ &\quad - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t-\mu_k\right)'{\Sigma_k}^{-1}(y_t-\mu_k)\right] \qquad \text{(B.1)} \\ &= \left[\sum_{k=1}^{K}\gamma_k^{(m)}(1)\log\left[\left(p_k^{(0)}\right)^{(m)}\right]\right] + \left[\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{kl}^{(m)}(t-1)\log\left(p_{kl}^{(m)}\right)\right] \\ &\quad - \frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\gamma_k^{(m)}(t)\log[\det(\Sigma_k)]\right) - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t'{\Sigma_k}^{-1}y_t\right)\right] \\ &\quad + \sum_{k=1}^{K}\left[\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t'{\Sigma_k}^{-1}\mu_k\right)\right] - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(\mu_k'{\Sigma_k}^{-1}\mu_k\right)\right]. \qquad \text{(B.2)}\end{aligned}$$

By Equation (2.40), the Lagrangian function can be written in the following two forms:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{p}^{(0)},\boldsymbol{P},\boldsymbol{\phi},\lambda_1,\ldots,\lambda_K,\lambda\right) &= \sum_{k=1}^{K}\gamma_k(1)\log\left(p_k^{(0)}\right) + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{kl}(t-1)\log(p_{kl}) \\ &\quad - \frac{1}{2}\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\gamma_k^{(m)}(t)\log[\det(\Sigma_k)]\right) \\ &\quad - \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{t=1}^{T}\gamma_k^{(m)}(t)\left(y_t-\mu_k\right)'{\Sigma_k}^{-1}(y_t-\mu_k)\right] \\ &\quad - \sum_{l=1}^{K}\left[\lambda_k\left(\sum_{l=1}^{K}p_{kl}-1\right)\right] - \lambda\left(\sum_{k=1}^{K}p_k^{(0)}-1\right), \qquad \text{(B.3)}\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{p}^{(0)}, \boldsymbol{P}, \boldsymbol{\phi}, \lambda_1, \ldots, \lambda_K, \lambda\right) = & \sum_{k=1}^{K} \gamma_k(1) \log\left(p_k^{(0)}\right) + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{kl}(t-1) \log(p_{kl}) \\ & - \frac{1}{2} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \gamma_k^{(m)}(t) \log[\det(\Sigma_k)] \right) - \frac{1}{2} \sum_{k=1}^{K} \left[\sum_{t=1}^{T} \gamma_k^{(m)}(t) \left({y_t}' {\Sigma_k}^{-1} y_t \right) \right] \\ & + \sum_{k=1}^{K} \left[\sum_{t=1}^{T} \gamma_k^{(m)}(t) \left({y_t}' {\Sigma_k}^{-1} \mu_k \right) \right] - \frac{1}{2} \sum_{k=1}^{K} \left[\sum_{t=1}^{T} \gamma_k^{(m)}(t) \left({\mu_k}' {\Sigma_k}^{-1} \mu_k \right) \right] \\ & - \sum_{l=1}^{K} \left[\lambda_k \left(\sum_{l=1}^{K} p_{kl} - 1 \right) \right] - \lambda \left(\sum_{k=1}^{K} p_k^{(0)} - 1 \right). \end{aligned} \tag{B.4}$$

To determine the estimates for μ_k and Σ_k, we calculate the partial derivatives of the Lagrangian function with respect to these parameters, for all $k \in S_X = \{1, \ldots, K\}$ and set them equal to zero.

To determine the estimates for μ_k, we utilise Equation (B.4) (see also Equations (A.2) and (A.3) in Appendix A). For all $k \in S_X = \{1, \ldots, K\}$, we have that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_k} = \mathbf{0}_p \Leftrightarrow \\ \sum_{t=1}^{T} \gamma_k(t) \left({y_t}' {\Sigma_k}^{-1} \right) - \frac{1}{2} \sum_{t=1}^{T} \gamma_k(t) \left(2 {\mu_k}' {\Sigma_k}^{-1} \right) = \mathbf{0}_p \Leftrightarrow \\ \sum_{t=1}^{T} \gamma_k(t) \left({\mu_k}' {\Sigma_k}^{-1} \right) = \sum_{t=1}^{T} \gamma_k(t) \left({y_t}' {\Sigma_k}^{-1} \right) \Leftrightarrow \\ \sum_{t=1}^{T} \gamma_k(t) \left({\mu_k}' {\Sigma_k}^{-1} \right) \Sigma_k = \sum_{t=1}^{T} \gamma_k(t) \left({y_t}' {\Sigma_k}^{-1} \right) \Sigma_k \Leftrightarrow \\ \sum_{t=1}^{T} \gamma_k(t) \left({\mu_k}' \right) \boldsymbol{I}_p = \sum_{t=1}^{T} \gamma_k(t) \left({y_t}' \right) \boldsymbol{I}_p \Leftrightarrow \\ \left(\sum_{t=1}^{T} \gamma_k(t) \, \mu_k \right)' = \left(\sum_{t=1}^{T} \gamma_k(t) \, y_t \right)' \Leftrightarrow \\ \sum_{t=1}^{T} \gamma_k(t) \, \mu_k = \sum_{t=1}^{T} \gamma_k(t) \, y_t \Leftrightarrow \\ \mu_k = \frac{1}{\sum_{t=1}^{T} \gamma_k(t)} \sum_{t=1}^{T} \gamma_k(t) \, y_t. \end{aligned}$$

To determine the estimates for Σ_k, we utilise Equation (B.3) (see also Equations (A.4) and (A.5) in Appendix A). For all $k \in S_X = \{1, \ldots, K\}$, we have that

$$\begin{gathered}
\frac{\partial l\left(\boldsymbol{y}, \boldsymbol{X} | \boldsymbol{\theta}\right)}{\partial \Sigma_k} = \mathbf{0}_{p \times p} \Leftrightarrow \\
-\frac{1}{2} \sum_{t=1}^{T} \gamma_k(t) \left(\Sigma_k{}'\right)^{-1} - \frac{1}{2} \sum_{t=1}^{T} \gamma_k(t) \left[- \left(\Sigma_k{}'\right)^{-1} \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)' \left(\Sigma_k{}'\right)^{-1} \right] = \mathbf{0}_{p \times p} \Leftrightarrow \\
\sum_{t=1}^{T} \gamma_k(t) \, {\Sigma_k}^{-1} = \sum_{t=1}^{T} \gamma_k(t) \left[{\Sigma_k}^{-1} \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)' {\Sigma_k}^{-1} \right] \Leftrightarrow \\
\left[\sum_{t=1}^{T} \gamma_k(t) \right] {\Sigma_k}^{-1} = {\Sigma_k}^{-1} \left[\sum_{t=1}^{T} \gamma_k(t) \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)' \right] {\Sigma_k}^{-1} \Leftrightarrow \\
\Sigma_k \left[\sum_{t=1}^{T} \gamma_k(t) \right] {\Sigma_k}^{-1} \Sigma_k = \Sigma_k {\Sigma_k}^{-1} \left[\sum_{t=1}^{T} \gamma_k(t) \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)' \right] {\Sigma_k}^{-1} \Sigma_k \Leftrightarrow \\
\left[\sum_{t=1}^{T} \gamma_k(t) \right] \Sigma_k \boldsymbol{I}_p = \boldsymbol{I}_p \left[\sum_{t=1}^{T} \gamma_k(t) \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)' \right] \boldsymbol{I}_p \Leftrightarrow \\
\Sigma_k = \frac{1}{\sum_{t=1}^{T} \gamma_k(t)} \sum_{t=1}^{T} \gamma_k(t) \left(y_t - \mu_k\right) \left(y_t - \mu_k\right)'.
\end{gathered}$$

Therefore we get the estimates for μ_k and Σ_k as

$$\hat{\mu}_k^{(m+1)} = \frac{1}{\sum_{t=1}^{T} \gamma_k^{(m)}(t)} \sum_{t=1}^{T} \gamma_k^{(m)}(t) \, y_t \qquad \left(k \in S_X = \{1, \ldots, K\}\right) \tag{B.5}$$

and

$$\hat{\Sigma}_k^{(m+1)} = \frac{1}{\sum_{t=1}^{T} \gamma_k^{(m)}(t)} \sum_{t=1}^{T} \gamma_k^{(m)}(t) \left(y_t - \hat{\mu}_k^{(m+1)}\right) \left(y_t - \hat{\mu}_k^{(m+1)}\right)' \tag{B.6}$$

Appendix C

Proofs and Technical Results of Chapter 3

C.1 Technical Result 1

In this section we prove Equation (3.6) of Section 3.1.

Proof. We have that

$$||\boldsymbol{S}-\mu\boldsymbol{I}_p||_{\mathrm{F}}^2=\frac{1}{p}\mathrm{tr}\Big[\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)^{'}\Big],$$

and

$$\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)^{'}=\boldsymbol{S}\boldsymbol{S}^{'}-\mu\boldsymbol{S}\boldsymbol{I}_p{}^{'}-\mu\boldsymbol{I}_p\boldsymbol{S}^{'}+\mu^2\boldsymbol{I}_p\boldsymbol{I}_p{}^{'}=\boldsymbol{S}\boldsymbol{S}^{'}-\mu\boldsymbol{S}-\mu\boldsymbol{S}^{'}+\mu^2\boldsymbol{I}_p.$$

Therefore,

$$\begin{gathered}\mathrm{tr}\Big[\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)\left(\boldsymbol{S}-\mu\boldsymbol{I}_p\right)^{'}\Big]=\mathrm{tr}\Big(\boldsymbol{S}\boldsymbol{S}^{'}-\mu\boldsymbol{S}-\mu\boldsymbol{S}^{'}+\mu^2\boldsymbol{I}_p\Big)=\\ \mathrm{tr}\Big(\boldsymbol{S}\boldsymbol{S}^{'}\Big)-\mu\mathrm{tr}(\boldsymbol{S})-\mu\mathrm{tr}\Big(\boldsymbol{S}^{'}\Big)+\mu^2\mathrm{tr}(\boldsymbol{I}_p)=\mathrm{tr}\Big(\boldsymbol{S}\boldsymbol{S}^{'}\Big)-2\mu\mathrm{tr}(\boldsymbol{S})+\mu^2p=\\ \mathrm{tr}\Big(\boldsymbol{S}^2\Big)-2\mu\mathrm{tr}(\boldsymbol{S})+\mu^2p=\sum_{i=1}^{p}l_i^2-2\mu\sum_{i=1}^{p}l_i+\sum_{i=1}^{p}\mu=\sum_{i=1}^{p}\left(l_i-\mu\right)^2.\end{gathered}$$

Hence,

$$\delta^2=\mathrm{E}\Big(||\boldsymbol{S}-\mu\boldsymbol{I}_p||_{\mathrm{F}}^2\Big)=\mathrm{E}\left(\frac{1}{p}\sum_{i=1}^{p}\left(l_i-\mu\right)^2\right). \tag{C.1}$$

Similarly it is proved that

$$\alpha^2=||\Sigma-\mu\boldsymbol{I}_p||_{\mathrm{F}}^2=\frac{1}{p}\sum_{i=1}^{p}\left(\lambda_i-\mu\right)^2, \tag{C.2}$$

and the proof of Equation (3.6) is complete. □

C.2 Proofs and Technical Lemmas of Section 3.2

This section contains some technical lemmas used in Chapter 3.2, along with the proofs of Lemma 3.9 and Theorem 3.3.

To simplify notation in what follows, let us define $R_t = \mathbf{1}_{\{X_t=k\}}$ and

$$\boldsymbol{Z}_k(t) = \mathbf{1}_{\{X_t=k\}} \left(Y_t - \mu_k\right) \left(Y_t - \mu_k\right)' \tag{C.3}$$

for any fixed $k \in \{1, \ldots, K\}$. We will also use $\boldsymbol{c}_m^n$ to denote a vector $(c_m, \ldots, c_n)$, where $m, n \in \mathbb{N}$ and $m < n$.

As $\{X_t : t \in \mathbb{Z}\}$ is a Markov chain, the conditional random variables $(X_t \mid X_{t-1})$ are independent of the random variables $(X_{t-2}, X_{t-3}, \ldots)$, for all $t \in \mathbb{Z}$. The same holds for the random variables $\left(\mathbf{1}_{\{X_t=k\}} \mid \mathbf{1}_{\{X_{t-1}=k\}}\right)$, that is $\left(\mathbf{1}_{\{X_t=k\}} \mid \mathbf{1}_{\{X_{t-1}=k\}}\right)$ are independent of the random variables $\left(\mathbf{1}_{\{X_{t-2}=k\}}, \mathbf{1}_{\{X_{t-3}=k\}}, \ldots\right)$. Therefore, $\{R_t : t \in \mathbb{Z}\}$ is a Markov chain with state space $\{0, 1\}$. We have that

$$\begin{aligned}
\mathrm{P}(R_{t+1} = 1 \mid R_t = 0) &= \mathrm{P}\left(\mathbf{1}_{\{X_{t+1}=k\}} = 1 \,\middle|\, \mathbf{1}_{\{X_t=k\}} = 0\right) \\
&= \mathrm{P}(X_{t+1} = k \mid X_t \neq k) \\
&= \sum_{l \in S_X \setminus \{k\}} \mathrm{P}(X_{t+1} = k \mid X_t = l) \\
&= \sum_{l \in S_X \setminus \{k\}} p_{lk},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(R_{t+1} = 1 \mid R_t = 1) &= \mathrm{P}\left(\mathbf{1}_{\{X_{t+1}=k\}} = 1 \,\middle|\, \mathbf{1}_{\{X_t=k\}} = 1\right) \\
&= \mathrm{P}(X_{t+1} = k \mid X_t = k) \\
&= p_{kk}.
\end{aligned}$$

We also have that

$$\mathrm{P}(R_{t+1} = 0 \mid R_t = 0) = 1 - \mathrm{P}(R_{t+1} = 1 \mid R_t = 0) = 1 - \sum_{l \in S_X \setminus \{k\}} p_{lk}$$

and

$$\mathrm{P}(R_{t+1} = 0 \mid R_t = 1) = 1 - \mathrm{P}(R_{t+1} = 1 \mid R_t = 1) = 1 - p_{kk}.$$

Let us denote by $\boldsymbol{B}$ the transition probability matrix of $\{R_t : t \in \mathbb{Z}\}$. As we showed above, we have

$$\boldsymbol{B} = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix} = \begin{pmatrix} 1 - \sum\limits_{l \in S_X \setminus \{k\}} p_{lk} & \sum\limits_{l \in S_X \setminus \{k\}} p_{lk} \\ 1 - p_{kk} & p_{kk} \end{pmatrix}.$$

$$\mathrm{P}(R_t = 1) = \mathrm{P}\big(\mathbf{1}_{\{X_t=k\}} = 1\big) = \mathrm{P}(X_t = k) = \pi_k \qquad (k \in \{1, \ldots, K\})\,. \tag{C.4}$$

A useful relationship in what follows is

$$R_t^n = R_t, \qquad (t \in \mathbb{Z},\ n \in \mathbb{N}) \tag{C.5}$$

as,

$$R_t = \mathbf{1}_{\{X_t=k\}} = \begin{cases} 1, & \text{if } X_t = k \\ 0, & \text{if } X_t \neq k \end{cases} \qquad (t \in \mathbb{Z}).$$

For the time series $\boldsymbol{Z}_k(t)$, we have

$$\begin{aligned}
\boldsymbol{Z}_k(t) &= \mathbf{1}_{\{X_t=k\}} \left(Y_t - \mu_k\right) \left(Y_t - \mu_k\right)' \\
&= \mathbf{1}_{\{X_t=k\}} \begin{pmatrix} Y_{t,1} - \mu_{k,1} \\ Y_{t,2} - \mu_{k,2} \\ \vdots \\ Y_{t,p} - \mu_{k,p} \end{pmatrix} \left(Y_{t,1} - \mu_{k,1}, Y_{t,2} - \mu_{k,2}, \ldots, Y_{t,p} - \mu_{k,p}\right) \\
&= \mathbf{1}_{\{X_t=k\}} \begin{pmatrix} \left(Y_{t,1} - \mu_{k,1}\right)^2 & \left(Y_{t,1} - \mu_{k,1}\right)\left(Y_{t,2} - \mu_{k,2}\right) & \ldots & \left(Y_{t,1} - \mu_{k,1}\right)\left(Y_{t,p} - \mu_{k,p}\right) \\ \left(Y_{t,2} - \mu_{k,2}\right)\left(Y_{t,1} - \mu_{k,1}\right) & \left(Y_{t,2} - \mu_{k,2}\right)^2 & \ldots & \left(Y_{t,2} - \mu_{k,2}\right)\left(Y_{t,p} - \mu_{k,p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(Y_{t,p} - \mu_{k,p}\right)\left(Y_{t,1} - \mu_{k,1}\right) & \left(Y_{t,p} - \mu_{k,p}\right)\left(Y_{t,2} - \mu_{k,2}\right) & \ldots & \left(Y_{t,p} - \mu_{k,p}\right)^2 \end{pmatrix}.
\end{aligned}$$

Therefore, the (i, j)-th entry of the matrix $\boldsymbol{Z}_k(t)$ is

$$[\boldsymbol{Z}_k(t)]_{ij} = \boldsymbol{Z}_{k,ij}(t) = \mathbf{1}_{\{X_t=k\}} \left(Y_{t,i} - \mu_{k,i}\right) \left(Y_{t,j} - \mu_{k,j}\right)$$

C.2.1 Technical Lemma 1

Lemma C.1. *For any $1 \leq k \leq K$, $1 \leq i, j \leq p$, the univariate time series*

$$z_t = \boldsymbol{Z}_{k,ij}(t) = [\boldsymbol{Z}_k(t)]_{ij} = \mathbf{1}_{\{X_t=k\}} \left(Y_{t,i} - \mu_{k,i}\right) \left(Y_{t,j} - \mu_{k,j}\right) \qquad (t \in \mathbb{Z})\,, \tag{C.6}$$

is stationary with autocovariances[1]

$$c_{k,ij}(n) = \mathrm{Cov}(z_t, z_{t+n}) = \pi_k \Sigma_{k,ij}^{1/2} \left[(\boldsymbol{B}^n)_{11} - \pi_k \right] \to 0,$$

exponentially fast, as $n \to +\infty$, *where* $\Sigma_{k,ij}^{1/2} = \left(\Sigma_k^{1/2} \right)_{ij}$, *and having a continuous spectral density*[2]

$$f_{k,ij}(\omega) = \sum_{n=-\infty}^{+\infty} c_{k,ij}(n)\, e^{in\omega}.$$

Proof. Let $k \in \{1, \ldots, K\}$ and $i, j \in \{1, \ldots, p\}$ be given and fixed. We define $q_t = \left(\Sigma_k^{1/2} \epsilon_t \right)_i \left(\Sigma_k^{1/2} \epsilon_t \right)_j$. Let $\rho_{ij} = \left(\Sigma_k^{1/2} \right)_{ij}$, for all $i, j \in \{1, \ldots, p\}$. We have that

$$\Sigma_k^{1/2} \epsilon_t = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix} \cdot \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \vdots \\ \epsilon_{t,p} \end{pmatrix} = \begin{pmatrix} \sum_{\alpha=1}^p \rho_{1\alpha} \epsilon_{t,\alpha} \\ \sum_{\alpha=1}^p \rho_{2\alpha} \epsilon_{t,\alpha} \\ \vdots \\ \sum_{\alpha=1}^p \rho_{p\alpha} \epsilon_{t,\alpha} \end{pmatrix},$$

therefore,

$$\begin{aligned} q_t &= \left(\Sigma_k^{1/2} \epsilon_t \right)_i \left(\Sigma_k^{1/2} \epsilon_t \right)_j \\ &= \left(\sum_{\alpha=1}^p \rho_{i\alpha} \epsilon_{t,\alpha} \right) \left(\sum_{\alpha=1}^p \rho_{j\alpha} \epsilon_{t,\alpha} \right) \\ &= \sum_{\alpha=1}^p \rho_{i\alpha} \rho_{j\alpha} \epsilon_{t,\alpha}^2 + \sum_{\alpha=1}^p \left(\sum_{\substack{\beta=1, \\ \beta \neq \alpha}}^p \rho_{i\alpha} \rho_{j\beta} \epsilon_{t,\alpha} \epsilon_{t,\beta} \right). \end{aligned}$$

The expectation of q_t is

$$\begin{aligned} \mathrm{E}(q_t) &= \mathrm{E} \left[\sum_{\alpha=1}^p \rho_{i\alpha} \rho_{j\alpha} \epsilon_{t,\alpha}^2 + \sum_{\alpha=1}^p \left(\sum_{\substack{\beta=1, \\ \beta \neq \alpha}}^p \rho_{i\alpha} \rho_{j\beta} \epsilon_{t,\alpha} \epsilon_{t,\beta} \right) \right] \\ &= \sum_{\alpha=1}^p \rho_{i\alpha} \rho_{j\alpha} \mathrm{E}\left(\epsilon_{t,\alpha}^2 \right) + \sum_{\alpha=1}^p \left(\sum_{\substack{\beta=1, \\ \beta \neq \alpha}}^p \rho_{i\alpha} \rho_{j\beta} \mathrm{E}\left(\epsilon_{t,\alpha} \epsilon_{t,\beta} \right) \right). \end{aligned}$$

[1] See Definition A.45 in Appendix A.
[2] See Definition A.50 in Appendix A.

For the model (2.45), we have assumed that ϵ_t, $t \in \mathbb{Z}$ are iid with

$$\begin{aligned} \mathrm{E}(\epsilon_t) &= \mathbf{0}_p, \\ \mathrm{Var}(\epsilon_t) &= \boldsymbol{I}_p. \end{aligned}$$

By Definitions A.37 and A.39, for all $t \in \mathbb{Z}$, we have

$$\mathrm{E}(\epsilon_{t,i}) = 0 \qquad (i \in \{1, \ldots, p\}), \tag{C.7}$$

$$\mathrm{Var}(\epsilon_{t,i}) = \mathrm{E}\big(\epsilon_{t,i}^2\big) - [\mathrm{E}(\epsilon_{t,i})]^2 = \mathrm{E}\big(\epsilon_{t,i}^2\big) = 1 \qquad (i \in \{1, \ldots, p\}), \tag{C.8}$$

$$\mathrm{Cov}\big(\epsilon_{t,i}, \epsilon_{t,j}\big) = \mathrm{E}\big(\epsilon_{t,i}\epsilon_{t,j}\big) - \mathrm{E}(\epsilon_{t,i})\,\mathrm{E}\big(\epsilon_{t,j}\big) = \mathrm{E}\big(\epsilon_{t,i}\epsilon_{t,j}\big) = 0 \qquad (i, j \in \{1, \ldots, p\} \text{ and } i \neq j). \tag{C.9}$$

By Equations (C.8) and (C.9) for the expectation of q_t we get

$$\mathrm{E}(q_t) = \sum_{\alpha=1}^{p} \rho_{i\alpha}\rho_{j\alpha} = \sum_{\alpha=1}^{p} \rho_{i\alpha}\rho_{\alpha j} = \left(\left[\Sigma_k^{1/2}\right]^2\right)_{ij} = (\Sigma_k)_{ij} = \Sigma_{k,ij},$$

where we have used the fact that the matrix $\Sigma_k^{1/2}$ is symmetric, which implies that $\rho_{ij} = \rho_{ji}$, for all $i, j \in \{1, \ldots, p\}$.

Since ϵ_t, $t \in \mathbb{Z}$ are iid and independent of X_s, for all $t, s \in \mathbb{Z}$, q_t, $t \in \mathbb{Z}$ are also iid and independent of R_s, for all t, s. It also holds that $z_t = R_t q_t$, which follows from Equations (2.45) and (C.6) and the fact that

$$R_t\,(Y_t - \mu_k) = \begin{cases} \Sigma_k^{1/2}\epsilon_t, & \text{if and only if } R_t = 1 \ (\Leftrightarrow X_t = k) \\ \mathbf{0}_p, & \text{otherwise} \end{cases}.$$

Thus, we have that

$$\begin{aligned} \mathrm{E}(z_t z_{t+n}) &= \mathrm{E}(R_t q_t R_{t+n} q_{t+n}) \\ &= \mathrm{E}(R_t R_{t+n})\,\mathrm{E}(q_t)\,\mathrm{E}(q_{t+n}) \\ &= \mathrm{E}(R_t R_{t+n})\,\Sigma_{k,ij}\Sigma_{k,ij} \\ &= \mathrm{E}(R_t R_{t+n})\,\Sigma_{k,ij}^2, \end{aligned} \tag{C.10}$$

where $\Sigma_{k,ij}^2 = \Big[(\Sigma_k)_{ij}\Big]^2$.

$$\begin{aligned}\mathrm{E}(R_t R_{t+n}) &= 0 \cdot \mathrm{P}(R_t R_{t+n} = 0) + 1 \cdot \mathrm{P}(R_t R_{t+n} = 1) \\ &= \mathrm{P}(R_t = 1, R_{t+n} = 1) = \mathrm{P}(R_t = 1)\,\mathrm{P}(R_{t+n} = 1 \mid R_t = 1) \\ &= \pi_k b_{11}^{(n)} = \pi_k \left(B^{(n)}\right)_{11} = \pi_k \left(B^n\right)_{11}. \end{aligned} \tag{C.11}$$

By Equations (C.10) and (C.11) we get

$$\mathrm{E}(z_t z_{t+n}) = \pi_k \left(B^n\right)_{11} \Sigma_{k,ij}^2. \tag{C.12}$$

For all $t \in \mathbb{Z}$, it holds that

$$\mathrm{E}(z_t) = \mathrm{E}(R_t q_t) = \mathrm{E}(R_t)\,\mathrm{E}(q_t) = \mathrm{P}(R_t = 1)\,\Sigma_{k,ij} = \pi_k \Sigma_{k,ij}. \tag{C.13}$$

It follows from Equations (C.12) and (C.13) that the autocovariance of the univariate time series z_t is

$$\begin{aligned} c_{k,ij}(n) &= \mathrm{Cov}(z_t, z_{t+n}) \\ &= \mathrm{E}(z_t z_{t+n}) - \mathrm{E}(z_t)\,\mathrm{E}(z_{t+n}) \\ &= \pi_k \left(B^n\right)_{11} \Sigma_{k,ij}^2 - \pi_k \Sigma_{k,ij} \cdot \pi_k \Sigma_{k,ij} \\ &= \pi_k \Sigma_{k,ij}^2 \left[\left(B_{11}^n - \pi_k\right)\right] \qquad (k \in \{1, \ldots, K\},\ i, j \in \{1, \ldots, p\},\ n \in \mathbb{Z}). \end{aligned} \tag{C.14}$$

We have that

$$\mathrm{E}\left(\mathbf{1}_{\{X_t = k\}}\right) = P(X_t = k) = \pi_k,$$

and by Equation (C.11), we get

$$\mathrm{E}\left(\mathbf{1}_{\{X_t = k\}} \mathbf{1}_{\{X_{t+n} = k\}}\right) = \pi_k \left(B^n\right)_{11},$$

thus

$$\mathrm{Cov}\left(\mathbf{1}_{\{X_t = k\}}, \mathbf{1}_{\{X_{t+n} = k\}}\right) = \pi_k \left(B^n\right)_{11} - \pi_k \pi_k = \pi_k \left[\left(B^n\right)_{11} - \pi_k\right]. \tag{C.15}$$

It follows from Equations (C.14) and (C.15) that

$$c_{k,ij}(n) = \Sigma_{k,ij}^2 \mathrm{Cov}\left(\mathbf{1}_{\{X_t = k\}}, \mathbf{1}_{\{X_{t+n} = k\}}\right) \qquad (k \in \{1, \ldots, K\},\ i, j \in \{1, \ldots, p\},\ n \in \mathbb{Z}),$$

which implies that $c_{k,ij}(n)$ decreases exponentially fast to 0, as will be shown in Step 1 in the proof of

Statement (ii) of Lemma C.2. Consequently, the spectral density converges pointwise and is continuous. □

C.2.2 Technical Lemma 2

Lemma C.2. *Let $\{X_t : t \in \mathbb{Z}\}$ be stationary, aperiodic and irreducible. Then,*

(i) *$\{X_t : t \in \mathbb{Z}\}$ is (γ, L^∞, ψ)-weakly dependent with an exponentially decreasing sequence $\gamma = \{\gamma_r : r \in \mathbb{N}\}$ and $\psi(h,k,u,v) \leq 4\,||h||_\infty\,||k||_\infty$, that is*

$$|\mathrm{Cov}[h(X_{t_1},\ldots,X_{t_u}),k(X_{\tau_1},\ldots,X_{\tau_v})]| \leq 4\,||h||_\infty\,||k||_\infty\,\gamma_r,$$

for all $t_1 < \ldots < t_u \leq t_u + r < \tau_1 < \ldots < \tau_v$ ($t_1 \leq \ldots \leq t_u < t_u + r \leq \tau_1 \leq \ldots \leq \tau_v$), $r,u,v \geq 1$, $h,k \in L^\infty$, where the norm $||\cdot||_\infty$ is defined as

$$||f||_\infty = \inf\{c \geq 0 : |f(x)| \leq c,\ \textit{almost everywhere}\},$$

for any function $f \in L^\infty$.

(ii) *There is some constant $c > 0$ such that,*

$$\left|\mathrm{Cov}\left(Y_{t_1,i_1}\cdot\ldots\cdot Y_{t_u,i_u}, Y_{\tau_1,j_1}\cdot\ldots\cdot Y_{\tau_v,j_v}\right)\right| \leq c\gamma_r,$$

for all $t_1 < \ldots < t_u \leq t_u + r < \tau_1 < \ldots < \tau_v$ ($t_1 \leq \ldots \leq t_u < t_u + r \leq \tau_1 \leq \ldots \leq \tau_v$), $1 \leq u,v \leq 4$ and $r \geq 1$.

Proof. (i) We have assumed that $\{X_t : t \in \mathbb{Z}\}$ is stationary, aperiodic and irreducible, and it is also finite-state, thus from the Statements (a), (b) and (c) of Theorem 3.1 and the paragraph following it in Bradley (2005), it follows that $\{X_t : t \in \mathbb{Z}\}$ is strongly mixing with exponentially decreasing mixing coefficients, say γ_r (see also Subsection 2.1 in Bradley 2005). Since $\{X_t : t \in \mathbb{Z}\}$ is strongly mixing, it follows from Lemma 6 in Doukhan and Louhichi (1999) that it is also (γ, L^∞, ψ)-weakly dependent with $\psi(h,k,u,v) = 4\,||h||_\infty\,||k||_\infty$, which implies Statement (i).

(ii) Statement (ii) will be proved in three steps.

Step 1: We choose

$$h(w_1,\ldots,w_u) = \begin{cases} w_1\cdot\ldots\cdot w_u, & \text{if } |w_1|,\ldots,|w_u| \leq 1 \\ 0, & \text{otherwise} \end{cases},$$

and

$$k(w_1,\ldots,w_v)=\begin{cases} w_1\cdot\ldots\cdot w_v, & \text{if } |w_1|,\ldots,|w_v|\leq 1\\ 0, & \text{otherwise}\end{cases}.$$

We have that $h,k\in L^\infty$ and $||h||_\infty||k||_\infty\leq 1$, therefore it follows from Statement (i) that

$$\left|\mathrm{Cov}\left(\mathbf{1}_{\{X_{t_1}=i_1\}}\cdot\ldots\cdot\mathbf{1}_{\{X_{t_u}=i_u\}},\mathbf{1}_{\{X_{\tau_1}=j_1\}}\cdot\ldots\cdot\mathbf{1}_{\{X_{\tau_v}=j_v\}}\right)\right|\leq 4\gamma_r, \tag{C.16}$$

for all $t_1<\ldots<t_u\leq t_u+r<\tau_1<\ldots<\tau_v$ ($t_1\leq\ldots\leq t_u<t_u+r\leq\tau_1\leq\ldots\leq\tau_v$), $i_1,\ldots,i_u,\ j_1,\ldots,j_v\in\{1,\ldots,K\}$, $u,v,r\geq 1$.

Step 2: For a set of real-valued random variables U_1,U_2,V_1,V_2, where the pair (U_1,U_2), V_1 and V_2 are independent, we have

$$\begin{aligned}\mathrm{Cov}(U_1V_1,U_2V_2)&=\mathrm{E}(U_1V_1\cdot U_2V_2)-\mathrm{E}(U_1V_1)\,\mathrm{E}(U_2V_2)\\&=\mathrm{E}(U_1U_2)\,\mathrm{E}(V_1)\,\mathrm{E}(V_2)-\mathrm{E}(U_1)\,\mathrm{E}(V_1)\,\mathrm{E}(U_2)\,\mathrm{E}(V_2)\\&=[\mathrm{E}(U_1U_2)-\mathrm{E}(U_1)\,\mathrm{E}(U_2)]\,\mathrm{E}(V_1)\,\mathrm{E}(V_2)\\&=\mathrm{Cov}(U_1,U_2)\,\mathrm{E}(V_1)\,\mathrm{E}(V_2).\end{aligned}$$

Step 3: Consider the random variables

$$Y_{s,m}=\sum_{k=1}^{K}\mathbf{1}_{\{X_s=k\}}\left(\Sigma_k^{1/2}\epsilon_s\right)_m\qquad(s\in\mathbb{Z},\ m\in\{1,\ldots,p\}).$$

Using the bilinearity of the covariance, we get that

$$\begin{aligned}&\mathrm{Cov}\left(Y_{t_1,i_1}\cdot\ldots\cdot Y_{t_u,i_u},Y_{\tau_1,j_1}\cdot\ldots\cdot Y_{\tau_v,j_v}\right)=\\&\mathrm{Cov}\left(\prod_{n=1}^{u}\left[\sum_{k=1}^{K}\mathbf{1}_{\{X_{t_n}=k\}}\left(\Sigma_k^{1/2}\epsilon_{t_n}\right)_{i_n}\right],\prod_{m=1}^{v}\left[\sum_{l=1}^{K}\mathbf{1}_{\{X_{\tau_m}=l\}}\left(\Sigma_l^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right)=\\&\mathrm{Cov}\left(\sum_{k_1,\ldots,k_u=1}^{K}\left[\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right],\sum_{l_1,\ldots,l_v=1}^{K}\left[\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right)=\\&\sum_{\substack{k_1,\ldots,k_u=1\\l_1,\ldots,l_v=1}}^{K}\mathrm{Cov}\left(\left[\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}}\right]\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right],\left[\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right]\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right).\end{aligned}\tag{C.17}$$

Since ϵ_t, $t\in\mathbb{Z}$ are iid and independent of the Markov chain $\{X_t:t\in\mathbb{Z}\}$, the random variables $\left(\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}},\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right)$, $\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]$ and $\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]$

are independent, thus from Step 2, it follows that

$$\operatorname{Cov}\left(\left[\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}}\right]\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right],\left[\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right]\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right)=$$
$$\operatorname{Cov}\left(\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}},\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right)\mathrm{E}\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]\mathrm{E}\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right],$$

and by Equation (C.16) in Step 1, we get

$$\left|\operatorname{Cov}\left(\left[\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}}\right]\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right],\left[\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right]\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right)\right|\leq$$
$$4\gamma_r\left|\mathrm{E}\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]\mathrm{E}\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right|. \quad \text{(C.18)}$$

By Equations (C.17) and (C.18), we get

$$\left|\operatorname{Cov}\left(Y_{t_1,i_1}\cdot\ldots\cdot Y_{t_u,i_u},Y_{\tau_1,j_1}\cdot\ldots\cdot Y_{\tau_v,j_v}\right)\right|=$$
$$\left|\sum_{\substack{k_1,\ldots,k_u=1\\l_1,\ldots,l_v=1}}^{K}\operatorname{Cov}\left(\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}},\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right)\mathrm{E}\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]\mathrm{E}\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right|\leq$$
$$\sum_{\substack{k_1,\ldots,k_u=1\\l_1,\ldots,l_v=1}}^{K}\left|\operatorname{Cov}\left(\prod_{n=1}^{u}\mathbf{1}_{\{X_{t_n}=k_n\}},\prod_{m=1}^{v}\mathbf{1}_{\{X_{\tau_m}=l_m\}}\right)\mathrm{E}\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]\mathrm{E}\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right|\leq$$
$$4\gamma_r\sum_{\substack{k_1,\ldots,k_u=1\\l_1,\ldots,l_v=1}}^{K}\left|\mathrm{E}\left[\prod_{n=1}^{u}\left(\Sigma_{k_n}^{1/2}\epsilon_{t_n}\right)_{i_n}\right]\mathrm{E}\left[\prod_{m=1}^{v}\left(\Sigma_{l_m}^{1/2}\epsilon_{\tau_m}\right)_{j_m}\right]\right|.$$

Since $K,\Sigma_1,\ldots,\Sigma_K$ are fixed and the 4-th moments of ϵ_t's are bounded (see Assumption (A3) in Subsection 2.3.1), it follows that there exists some constant $c>0$, such that

$$\left|\operatorname{Cov}\left(Y_{t_1,i_1}\cdot\ldots\cdot Y_{t_u,i_u},Y_{\tau_1,j_1}\cdot\ldots\cdot Y_{\tau_v,j_v}\right)\right|\leq c\gamma_r.$$

□

C.2.3 Proof of Lemma 3.9

Proof. We have

$$\begin{aligned}
||\mu_k||_2^2 &= \sum_{i=1}^{p} \mu_{k,i}^2 \\
&\leq p \left(\max_{1 \leq i \leq p} \{|\mu_{k,i}|\} \right)^2 \qquad \text{[Assumption (B1)]} \\
&= p \left[\mathrm{O}(1) \right]^2 \\
&= p \, \mathrm{O}(1) \\
&= \mathrm{O}(p) \,,
\end{aligned}$$

and, since Σ_k is positive semidefinite, by Proposition A.11, we get

$$\begin{aligned}
||\Sigma_k||_{\mathrm{F}}^2 &\leq \frac{1}{p} \left[\mathrm{tr}(\Sigma_k) \right]^2 \\
&= p \left(\frac{1}{p} \mathrm{tr} \left[\Sigma_k^{1/2} \left(\Sigma_k^{1/2} \right)' \right] \right)^2 \\
&= p \left|\left| \Sigma_k^{1/2} \right|\right|_{\mathrm{F}}^2 \qquad \text{[Assumption (B1)]} \\
&= p \, \mathrm{O}(1) \\
&= \mathrm{O}(p) \,.
\end{aligned}$$

As a result, we get the following equations:

$$\frac{1}{p} ||\mu_k||_2^2 = \mathrm{O}(1) \,, \tag{C.19}$$

and

$$||\Sigma_k||_{\mathrm{F}}^2 = \mathrm{O}(p) \,. \tag{C.20}$$

Let us denote by R the total number of visits of the Markov chain $\{Y_t : t \in \mathbb{Z}\}$ to state k over $\{1, \ldots, T\}$, that is

$$R = \sum_{t=1}^{T} R_t. \tag{C.21}$$

(i) As was mentioned in Subsection 2.3.2 (see the discussion following Equation (2.57)), if $R = 0$, then μ_k^{o} is set equal to $\mathbf{0}_p$, therefore, it holds that

$$\mu_k^{\mathrm{o}} = \mu_k^{\mathrm{o}} \cdot \mathbf{1}_{\{R>0\}}. \tag{C.22}$$

Using Equation (C.22), we get

$$\mathrm{E}(\mu_k^{\mathrm{o}}) = \mathrm{E}\big(\mu_k^{\mathrm{o}} \cdot \mathbf{1}_{\{R>0\}}\big) = \mathrm{E}\left(\frac{1}{R}\sum_{t=1}^{T} R_t Y_t \cdot \mathbf{1}_{\{R>0\}}\right) = \mathrm{E}\left[\mathrm{E}\left(\frac{1}{R}\sum_{t=1}^{T} R_t Y_t \cdot \mathbf{1}_{\{R>0\}} \,\middle|\, \boldsymbol{X}_1^T\right)\right] =$$
$$\mathrm{E}\left(\frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t \mathrm{E}\Big(Y_t \mid \boldsymbol{X}_1^T\Big)\right) = \mathrm{E}\left(\frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t \mathrm{E}(Y_t \mid X_t)\right).$$

In the last relation we used the fact that, by the definition of HMMs, Y_t is independent of X_s, given X_t, for all $s \neq t$.

Since

$$R_t = \mathbf{1}_{\{X_t=k\}} = \begin{cases} 1, & \text{if } X_t = k \\ 0, & \text{otherwise} \end{cases},$$

it follows that

$$R_t \mathrm{E}(Y_t \mid X_t) = R_t \mathrm{E}(Y_t \mid X_t = k) = R_t \mu_k,$$

consequently

$$\mathrm{E}(\mu_k^{\mathrm{o}}) = \mathrm{E}\left(\frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t \mu_k\right) = \mathrm{E}\big(\mathbf{1}_{\{R>0\}}\mu_k\big) = \mu_k\left[0 \cdot \mathrm{P}\big(\mathbf{1}_{\{R>0\}} = 0\big) + 1 \cdot \mathrm{P}\big(\mathbf{1}_{\{R>0\}} = 1\big)\right] =$$
$$\mu_k \mathrm{P}(R>0) = \mu_k\left[1 - \mathrm{P}(R=0)\right] = \mu_k\left[1 - \mathrm{P}(R_1 = 0, \ldots, R_T = 0)\right] =$$
$$\mu_k\left[1 - \mathrm{P}(R_1 = 0)\,\mathrm{P}(R_2 = 0 \mid R_1 = 0) \cdot \ldots \cdot \mathrm{P}(R_T = 0 \mid R_{T-1} = 0)\right] =$$
$$\mu_k\left[1 - (1-\pi_k)\, b_{00} \cdot \ldots \cdot b_{00}\right] = \mu_k - \mu_k\,(1-\pi_k)\, b_{00}^{T-1}.$$

By the irreducibility of the Markov chain $\{Y_t : t \in \mathbb{Z}\}$, it holds that $0 < b_{00} < 1$. Setting as β the maximum of those K values, by (C.19), the first part of Statement (i) is assured.

Let us use the abbreviations

$$\eta_t = \Sigma_k^{1/2} \epsilon_t, \tag{C.23}$$

and

$$\delta_k = \mu_k^{\mathrm{o}} - \mu_k. \tag{C.24}$$

Since ϵ_t, $t \in \{1, \ldots, T\}$ are iid zero-mean random vectors, independent of the state variables X_t, $t \in \{1, \ldots, T\}$, the same holds for η_t. Hence, by (ii) in Proposition A.4 and (ii) in Proposition A.5 in Appendix A, for all $t \in \{1, \ldots, T\}$, we get

$$\mathrm{E}(\eta_t) = \mathrm{E}\Big(\Sigma_k^{1/2}\epsilon_t\Big) = \Sigma_k^{1/2}\,\mathrm{E}(\epsilon_t) = \Sigma_k^{1/2}\,\mathbf{0}_p = \mathbf{0}_p, \tag{C.25}$$

and

$$\mathrm{Var}(\eta_t) = \mathrm{Var}\left(\Sigma_k^{1/2}\epsilon_t\right) = \Sigma_k^{1/2}\,\mathrm{Var}(\epsilon_t)\left(\Sigma_k^{1/2}\right)' = \Sigma_k^{1/2}\,\boldsymbol{I}_p\,\left(\Sigma_k^{1/2}\right)' = \Sigma_k. \tag{C.26}$$

Since $Y_t = \mu_k + \Sigma_k^{1/2}\epsilon_t = \mu_k + \eta_t$, if and only if $R_t = 1$, it follows that $R_t Y_t = R_t\,(\mu_k + \eta_t)$. Thus, using Equation (C.22), we get the following relation for δ_k:

$$\begin{aligned}
\delta_k = \mu_k^{\mathrm{o}} - \mu_k &= \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t Y_t - \mu_k = \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\,(\mu_k + \eta_t) - \mu_k \\
&= \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\mu_k + \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\eta_t - \mu_k = \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\eta_t + \mathbf{1}_{\{R>0\}}\mu_k - \mu_k \\
&= \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\eta_t + \left(\mathbf{1}_{\{R>0\}} - 1\right)\mu_k = \frac{\mathbf{1}_{\{R>0\}}}{R}\sum_{t=1}^{T} R_t\eta_t - \mathbf{1}_{\{R=0\}}\mu_k,
\end{aligned} \tag{C.27}$$

where in the last relation we used the fact that $\mathbf{1}_{\{R>0\}} - 1 = -\mathbf{1}_{\{R=0\}}$, since

$$\mathbf{1}_{\{R>0\}} - 1 = \begin{cases} 0, & \text{if } R > 0 \\ -1, & \text{if } R = 0 \end{cases}.$$

Equation (2.59) can be written as

$$\pi_k^{\mathrm{o}} = \frac{1}{T}\sum_{t=1}^{T} R_t = \frac{R}{T}, \tag{C.28}$$

therefore, by Equations (C.28) and (C.27), we get

$$\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 = \begin{cases} \pi_k^{\mathrm{o}}\left|\left|\dfrac{\mathbf{1}_{\{R>0\}}}{R}\sum\limits_{t=1}^{T} R_t\eta_t\right|\right|_2^2, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases},$$

thus

$$\begin{aligned}
\mathrm{E}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2\right) &= \mathrm{E}\left[\mathrm{E}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\pi_k^{\mathrm{o}}\left(\frac{\mathbf{1}_{\{R>0\}}}{R}\right)^2 \left|\left|\sum_{t=1}^{T} R_t\eta_t\right|\right|_2^2 \,\middle|\, \boldsymbol{X}_1^T\right)\right] \\
&= \mathrm{E}\left(\pi_k^{\mathrm{o}}\frac{\mathbf{1}_{\{R>0\}}}{R^2}\mathrm{E}\left[\left(\sum_{t=1}^{T} R_t\eta_t\right)'\left(\sum_{t=1}^{T} R_t\eta_t\right) \,\middle|\, \boldsymbol{X}_1^T\right]\right) \\
&= \mathrm{E}\left(\pi_k^{\mathrm{o}}\frac{\mathbf{1}_{\{R>0\}}}{R^2}\mathrm{E}\left[\sum_{t=1}^{T}\left(\sum_{s=1}^{T} R_t\eta_t{}' R_s\eta_s\right) \,\middle|\, \boldsymbol{X}_1^T\right]\right) \\
&= \mathrm{E}\left(\pi_k^{\mathrm{o}}\frac{\mathbf{1}_{\{R>0\}}}{R^2}\sum_{t=1}^{T}\left[\sum_{s=1}^{T} R_tR_s\mathrm{E}\left(\eta_t{}'\eta_s \,\middle|\, \boldsymbol{X}_1^T\right)\right]\right) \\
&= \mathrm{E}\left(\pi_k^{\mathrm{o}}\frac{\mathbf{1}_{\{R>0\}}}{R^2}\sum_{t=1}^{T}\left[\sum_{s=1}^{T} R_tR_s\mathrm{E}\left(\eta_t{}'\eta_s\right)\right]\right).
\end{aligned} \tag{C.29}$$

As we mentioned, η_t are iid, with

$$\mathrm{E}(\eta_t) = \mathbf{0}_p \Leftrightarrow \mathrm{E}(\eta_{t,i}) = 0 \qquad (t \in \mathbb{Z},\ i \in \{1,\ldots,p\}),$$

therefore

$$\mathrm{E}\left(\eta_t{}'\eta_s\right) = \mathrm{E}\left(\sum_{i=1}^{p}\eta_{t,i}\eta_{s,i}\right) = \sum_{i=1}^{p}\mathrm{E}(\eta_{t,i}\eta_{s,i}) = \begin{cases} \sum_{i=1}^{p}\mathrm{E}\left(\eta_{t,i}^2\right), & \text{if } t = s \\ 0, & \text{if } t \neq s \end{cases}. \tag{C.30}$$

By Equation (C.25) we get

$$\sum_{i=1}^{p}\mathrm{E}(\eta_{t,i}^2) = \sum_{i=1}^{p}\mathrm{Var}(\eta_{t,i}^2) = \sum_{i=1}^{p}\mathrm{Var}\left[\left(\Sigma_k^{1/2}\epsilon_t\right)_i\right] = \sum_{i=1}^{p}\mathrm{Var}\left(\sum_{j=1}^{p}\rho_{ij}\epsilon_{t,j}\right). \tag{C.31}$$

As a result, by Equations (C.8), (C.9) and (C.31) (see also Proposition A.9 in Appendix A), we get

$$\begin{aligned}
\sum_{i=1}^{p}\mathrm{E}(\eta_{t,i}^2) &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}(\rho_{ij}\epsilon_{t,j}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\rho_{ij}^2\mathrm{Var}(\epsilon_{t,j}) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\rho_{ij}^2 = \mathrm{tr}\left[\Sigma_k^{1/2}\left(\Sigma_k^{1/2}\right)'\right] = \mathrm{tr}(\Sigma_k).
\end{aligned} \tag{C.32}$$

Combining (C.5), (C.30) and (C.32), we get

$$\sum_{t=1}^{T}\left[\sum_{s=1}^{T} R_t R_s \mathrm{E}\left(\eta_t{}' \eta_s\right)\right] = \sum_{t=1}^{T} R_t^2 \mathrm{E}\left(\eta_t{}' \eta_t\right) = \sum_{t=1}^{T} R_t \mathrm{tr}(\Sigma_k) = R\mathrm{tr}(\Sigma_k)\,.$$

By Equation (C.28), directly follows that

$$\pi_k^{\mathrm{o}} = \pi_k^{\mathrm{o}} \cdot \mathbf{1}_{\{R>0\}}, \tag{C.33}$$

hence, (C.29) becomes

$$\begin{aligned}\mathrm{E}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2\right) &= \mathrm{E}\left[\pi_k^{\mathrm{o}} \frac{\mathbf{1}_{\{R>0\}}}{R^2} R\mathrm{tr}(\Sigma_k)\right] = \mathrm{E}\left[\pi_k^{\mathrm{o}} \frac{1}{R}\mathrm{tr}(\Sigma_k)\right] \\ &= \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T} R_t \frac{1}{R}\mathrm{tr}(\Sigma_k)\right] = \frac{1}{T}\mathrm{tr}(\Sigma_k) = \mathrm{O}\left(\frac{p}{T}\right),\end{aligned} \tag{C.34}$$

which concludes the proof of the second part of Statement (i).

By the definition of HMMs and the construction of the model given by Equation (2.45) (see also (i) in Proposition A.6 in Appendix A), we get

$$\mathrm{E}\left(Y_t - \mu_k \mid \boldsymbol{X}_1^T = k, \ldots, X_T\right) = \mathrm{E}(Y_t - \mu_k \mid X_t = k) = \mathbf{0}_p,$$

and

$$\mathrm{E}\left[(Y_t - \mu_k)\,(Y_s - \mu_k)' \,\middle|\, \boldsymbol{X}_1^T\right] = \begin{cases} \mathrm{E}[(Y_t - \mu_k) \mid X_t] \cdot \mathrm{E}\left[(Y_s - \mu_k)' \,\middle|\, X_s\right], & \text{if } s \neq t \\ \mathrm{Var}(Y_t \mid X_t)\,, & \text{if } s = t \end{cases}.$$

It follows that

$$R_t \mathrm{E}\left(Y_t - \mu_k \mid \boldsymbol{X}_1^T\right) = \mathbf{0}_p, \tag{C.35}$$

and

$$R_t \mathrm{E}\left[(Y_t - \mu_k)\,(Y_s - \mu_k)' \,\middle|\, \boldsymbol{X}_1^T\right] = \begin{cases} R_t \Sigma_k, & \text{if } s = t \\ \mathbf{0}_{p\times p}, & \text{if } s \neq t \end{cases}. \tag{C.36}$$

For the conditional expectation of Σ_k^{o}, given the state sequence $(X_1, \ldots, X_T)$, we have

$$\begin{aligned}
\mathrm{E}\left(\Sigma_k^{\mathrm{o}} \mid \boldsymbol{X}_1^T\right) &= \frac{1}{T}\sum_{t=1}^{T} R_T \mathrm{E}\left[\left(Y_t - \mu_k^{\mathrm{o}}\right)\left(Y_t - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} R_T \mathrm{E}\left[\left(Y_t - \mu_k + \mu_k - \mu_k^{\mathrm{o}}\right)\left(Y_t - \mu_k + \mu_k - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T\right],
\end{aligned}$$

and

$$\begin{aligned}
&\left(Y_t - \mu_k + \mu_k - \mu_k^{\mathrm{o}}\right)\left(Y_t - \mu_k + \mu_k - \mu_k^{\mathrm{o}}\right)' = \\
&\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)' + \left(Y_t - \mu_k\right)\left(\mu_k - \mu_k^{\mathrm{o}}\right)' + \left(\mu_k - \mu_k^{\mathrm{o}}\right)\left(Y_t - \mu_k\right)' + \left(\mu_k - \mu_k^{\mathrm{o}}\right)\left(\mu_k - \mu_k^{\mathrm{o}}\right)'.
\end{aligned}$$

Using the linearity of the expectation, we get

$$\begin{aligned}
\mathrm{E}\left(\Sigma_k^{\mathrm{o}} \mid \boldsymbol{X}_1^T\right) = &\frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left[\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)' \,\middle|\, \boldsymbol{X}_1^T\right] \\
&+ \frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left[\left(Y_t - \mu_k\right)\left(\mu_k - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T\right] \\
&+ \frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left[\left(\mu_k - \mu_k^{\mathrm{o}}\right)\left(Y_t - \mu_k\right)' \,\middle|\, \boldsymbol{X}_1^T\right] \\
&+ \frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left[\left(\mu_k - \mu_k^{\mathrm{o}}\right)\left(\mu_k - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T\right].
\end{aligned} \tag{C.37}$$

For the first term of (C.37), by Equation (C.36), we get

$$\frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left[\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)' \,\middle|\, \boldsymbol{X}_1^T\right] = \frac{1}{T}\sum_{t=1}^{T} R_t \Sigma_k = \pi_k^{\mathrm{o}}\, \Sigma_k. \tag{C.38}$$

Since

$$\mathbf{1}_{\{R>0\}} = \begin{cases} 1, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases},$$

it holds that $R \cdot \mathbf{1}_{\{R>0\}} = R$, thus,

$$\mu_k = \frac{\mathbf{1}_{\{R>0\}}}{R \cdot \mathbf{1}_{\{R>0\}}} \sum_{t=1}^{T} R_t \mu_k = \frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{t=1}^{T} R_t \mu_k. \tag{C.39}$$

Using (C.22) and (C.39) we get

$$
\begin{aligned}
R_t \mathrm{E}\Big[(Y_t - \mu_k) (\mu_k - \mu_k^{\mathrm{o}})' \,\Big|\, \boldsymbol{X}_1^T \Big] = \\
R_t \mathrm{E}\left[(Y_t - \mu_k) \left(\frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s \mu_k - \frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s Y_s \right)' \,\middle|\, \boldsymbol{X}_1^T \right] = \\
R_t \mathrm{E}\left[(Y_t - \mu_k) \left[\frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s (\mu_k - Y_s) \right]' \,\middle|\, \boldsymbol{X}_1^T \right] = \\
-\frac{R_t \mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s \mathrm{E}\Big[(Y_t - \mu_k) (Y_s - \mu_k)' \,\Big|\, \boldsymbol{X}_1^T \Big].
\end{aligned}
$$

Thus, using (C.5) and (C.36), the second term of Equation (C.37) can be written as

$$
\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{E}\Big[(Y_t - \mu_k) (\mu_k - \mu_k^{\mathrm{o}})' \,\Big|\, \boldsymbol{X}_1^T \Big] = \\
-\frac{\mathbf{1}_{\{R>0\}}}{TR} \sum_{t=1}^{T} \sum_{s=1}^{T} R_t R_s \mathrm{E}\Big[(Y_t - \mu_k) (Y_s - \mu_k)' \,\Big|\, \boldsymbol{X}_1^T \Big] = \\
-\frac{\mathbf{1}_{\{R>0\}}}{TR} \left(\sum_{t=1}^{T} R_t R_t \mathrm{E}\Big[(Y_t - \mu_k) (Y_t - \mu_k)' \,\Big|\, \boldsymbol{X}_1^T \Big] + \sum_{t=1}^{T} \sum_{\substack{s=1 \\ s \neq t}}^{T} R_t R_s \mathrm{E}\Big[(Y_t - \mu_k) (Y_s - \mu_k)' \,\Big|\, \boldsymbol{X}_1^T \Big] \right) = \\
-\frac{\mathbf{1}_{\{R>0\}}}{TR} \sum_{t=1}^{T} R_t R_t \Sigma_k = -\frac{\mathbf{1}_{\{R>0\}}}{TR} \sum_{t=1}^{T} R_t \Sigma_k = -\frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k.
\end{aligned}
\tag{C.40}
$$

For the third term of Equation (C.37) (see also (i) in Proposition A.4 in Appendix A) we have

$$
\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{E}\Big[(\mu_k - \mu_k^{\mathrm{o}}) (Y_t - \mu_k)' \,\Big|\, \boldsymbol{X}_1^T \Big] &= \frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{E}\left(\Big[(Y_t - \mu_k) (\mu_k - \mu_k^{\mathrm{o}})' \Big]' \,\middle|\, \boldsymbol{X}_1^T \right) \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \left(\mathrm{E}\Big[(Y_t - \mu_k) (\mu_k - \mu_k^{\mathrm{o}})' \,\Big|\, \boldsymbol{X}_1^T \Big] \right)' \\
&= \left(\frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{E}\Big[(Y_t - \mu_k) (\mu_k - \mu_k^{\mathrm{o}})' \,\Big|\, \boldsymbol{X}_1^T \Big] \right)' \\
&= \left(-\frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k \right)' \\
&= -\frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k,
\end{aligned}
\tag{C.41}
$$

as Σ_k is symmetric.

Combining (C.5), (C.22), (C.36) and (C.39) and the fact that $\left(\mathbf{1}_{\{R>0\}}\right)^2 = \mathbf{1}_{\{R>0\}}$, we get

$$
\begin{aligned}
&R_t \mathrm{E}\left[\left(\mu_k - \mu_k^{\mathrm{o}}\right) \left(\mu_k - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T \right] = R_t \mathrm{E}\left[\left(\mu_k^{\mathrm{o}} - \mu_k\right) \left(\mu_k^{\mathrm{o}} - \mu_k\right)' \,\middle|\, \boldsymbol{X}_1^T \right] \\
&= R_t \mathrm{E}\left[\left(\frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s \left(Y_t - \mu_k\right) \right) \left(\frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{s=1}^{T} R_s \left(Y_t - \mu_k\right) \right)' \,\middle|\, \boldsymbol{X}_1^T \right] \\
&= R_t \left(\frac{\mathbf{1}_{\{R>0\}}}{R} \right)^2 \sum_{s_1=1}^{T} \sum_{s_2=1}^{T} R_{s_1} R_{s_2} \mathrm{E}\left[\left(Y_{s_1} - \mu_k\right) \left(Y_{s_2} - \mu_k\right)' \,\middle|\, \boldsymbol{X}_1^T \right] \\
&= R_t \frac{\mathbf{1}_{\{R>0\}}}{R^2} \sum_{s=1}^{T} R_{s_1} R_{s_1} \Sigma_k = \frac{R_t \mathbf{1}_{\{R>0\}}}{R} \Sigma_k
\end{aligned}
$$

Therefore, the fourth term of Equation (C.37) can be expressed as

$$
\frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{E}\left[\left(\mu_k - \mu_k^{\mathrm{o}}\right) \left(\mu_k - \mu_k^{\mathrm{o}}\right)' \,\middle|\, \boldsymbol{X}_1^T \right] = \frac{1}{T} \sum_{t=1}^{T} \frac{R_t \mathbf{1}_{\{R>0\}}}{R} \Sigma_k = \frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k. \tag{C.42}
$$

As a result, by Equations (C.37), (C.38) and (C.40) to (C.42), we get

$$
\mathrm{E}\left(\Sigma_k^{\mathrm{o}} \mid \boldsymbol{X}_1^T \right) = \pi_k^{\mathrm{o}} \Sigma_k - \frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k - \frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k + \frac{\mathbf{1}_{\{R>0\}}}{T} \Sigma_k = \left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T} \right) \Sigma_k. \tag{C.43}
$$

Therefore,

$$
\mathrm{E}(\Sigma_k^{\mathrm{o}}) = \mathrm{E}\left[\mathrm{E}\left(\Sigma_k^{\mathrm{o}} \mid \boldsymbol{X}_1^T \right) \right] = \mathrm{E}\left[\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T} \right) \Sigma_k \right] = \mathrm{E}(\pi_k^{\mathrm{o}}) \Sigma_k - \frac{1}{T} \mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right) \Sigma_k.
$$

We have that

$$
\mathrm{E}(\pi_k^{\mathrm{o}}) = \mathrm{E}\left(\frac{1}{T} \sum_{t=1}^{T} R_t \right) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}(R_t) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{P}(R_t = 1) = \frac{1}{T} \sum_{t=1}^{T} \pi_k = \pi_k, \tag{C.44}
$$

and

$$
\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right) = \mathrm{P}(R > 0) = 1 - \mathrm{P}(R = 0)\,. \tag{C.45}
$$

As a result, the expectation of Σ_k^{o} is

$$
\mathrm{E}(\Sigma_k^{\mathrm{o}}) = \pi_k \Sigma_k - \frac{1}{T} \Sigma_k + \frac{\mathrm{P}(R = 0)}{T} \Sigma_k. \tag{C.46}
$$

By the definition of β, $\mathrm{P}(R=0) = \mathrm{O}(\beta^T)$ (see the end of the proof of the first part of Statement (i)) and by Equation (C.20), we get $||\Sigma_k||_{\mathrm{F}} = \mathrm{O}(\sqrt{p})$. Consequently, (C.46) becomes

$$\mathrm{E}(\Sigma_k^{\mathrm{o}}) = \left(\pi_k - \frac{1}{T}\right)\Sigma_k + \mathrm{O}\left(\frac{\beta^T}{T}\right) = \pi_k\Sigma_k + \mathrm{O}\left(\frac{\sqrt{p}}{T}\right), \tag{C.47}$$

as $\sqrt{p} > \beta^T$, which concludes the proof of the third part of Statement (i).

(ii) By Assumption (B1) we have

$$\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}} = \mathrm{O}(1) \Leftrightarrow \frac{1}{p}\mathrm{tr}\left[\Sigma_k^{1/2}\left(\Sigma_k^{1/2}\right)'\right] = \mathrm{O}(1) \Leftrightarrow \mathrm{tr}(\Sigma_k) = \mathrm{O}(p)\,.$$

Using the fact that for any matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, it holds

$$\mathrm{E}[\mathrm{tr}(\boldsymbol{A})] = \mathrm{E}\left(\sum_{i=1}^{p}\alpha_{ii}\right) = \sum_{i=1}^{p}\mathrm{E}(\alpha_{ii}) = \mathrm{tr}[\mathrm{E}(\boldsymbol{A})]\,, \tag{C.48}$$

we get

$$\begin{aligned}
\mathrm{E}(\nu_k^{\mathrm{o}} - \nu_k) = \mathrm{E}(\nu_k^{\mathrm{o}}) - \nu_k = \mathrm{E}\left[\frac{1}{p}\mathrm{tr}(\Sigma_k^{\mathrm{o}})\right] - \nu_k = \frac{1}{p}\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] - \nu_k = \\
\frac{1}{p}\mathrm{tr}[\mathrm{E}(\Sigma_k^{\mathrm{o}})] - \nu_k = \frac{1}{p}\mathrm{tr}\left[\pi_k\Sigma_k - \frac{1}{T}\Sigma_k + \mathrm{O}\left(\frac{\beta^T}{T}\right)\right] - \nu_k = \\
\frac{1}{p}\pi_k\mathrm{tr}(\Sigma_k) - \frac{1}{pT}\mathrm{tr}(\Sigma_k) + \mathrm{O}\left(\frac{\beta^T}{pT}\right) - \frac{1}{p}\pi_k\mathrm{tr}(\Sigma_k) = \\
-\frac{1}{pT}\mathrm{O}(p) + \mathrm{O}\left(\frac{\beta^T}{pT}\right) = \mathrm{O}\left(\frac{1}{T}\right) + \mathrm{O}\left(\frac{\beta^T}{pT}\right) = \mathrm{O}\left(\frac{1}{T}\right),
\end{aligned}$$

as $\frac{1}{T} > \frac{1}{pT}$ and $\beta < 1$, which concludes the second part of Statement (ii).

For the first part we have

$$\begin{aligned}
\mathrm{E}\left[(\nu_k^{\mathrm{o}} - \nu_k)^2\right] &= \mathrm{E}\left(\left[\frac{1}{p}\mathrm{tr}(\Sigma_k^{\mathrm{o}}) - \frac{1}{p}\mathrm{tr}(\pi_k\Sigma_k)\right]^2\right) = \frac{1}{p^2}\mathrm{E}\left([\mathrm{tr}(\Sigma_k^{\mathrm{o}}) - \mathrm{tr}(\pi_k\Sigma_k)]^2\right) \\
&= \frac{1}{p^2}\mathrm{E}\left([\mathrm{tr}(\Sigma_k^{\mathrm{o}})]^2 - 2\mathrm{tr}(\Sigma_k^{\mathrm{o}})\,\mathrm{tr}(\pi_k\Sigma_k) + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right) \\
&= \frac{1}{p^2}\left(\mathrm{E}\left([\mathrm{tr}(\Sigma_k^{\mathrm{o}})]^2\right) - 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right) \\
&= \frac{1}{p^2}\left(\mathrm{Var}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] + (\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})])^2 - 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right).
\end{aligned}$$

Using Equation (C.48) and the third part of Statement (i), we get

$$\frac{1}{p^2}\left[\left(\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})]\right)^2 - 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{E}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right] =$$
$$\frac{1}{p^2}\left[\left(\mathrm{tr}[\mathrm{E}(\Sigma_k^{\mathrm{o}})]\right)^2 - 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{tr}[\mathrm{E}(\Sigma_k^{\mathrm{o}})] + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right] =$$
$$\frac{1}{p^2}\left(\left[\mathrm{tr}(\pi_k\Sigma_k) + \mathrm{O}\left(\frac{\sqrt{p}}{T}\right)\right]^2 - 2\mathrm{tr}(\pi_k\Sigma_k)\left[\mathrm{tr}(\pi_k\Sigma_k) + \mathrm{O}\left(\frac{\sqrt{p}}{T}\right)\right] + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right) =$$
$$\frac{1}{p^2}\left([\mathrm{tr}(\pi_k\Sigma_k)]^2 + 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{O}\left(\frac{\sqrt{p}}{T}\right) + \mathrm{O}\left(\frac{p}{T^2}\right) - 2\,[\mathrm{tr}(\pi_k\Sigma_k)]^2 - 2\mathrm{tr}(\pi_k\Sigma_k)\,\mathrm{O}\left(\frac{\sqrt{p}}{T}\right) + [\mathrm{tr}(\pi_k\Sigma_k)]^2\right) =$$
$$\frac{1}{p^2}\,\mathrm{O}\left(\frac{p}{T^2}\right) = \mathrm{O}\left(\frac{1}{pT^2}\right) = \mathrm{O}\left(\frac{1}{T^2}\right),$$

as $\frac{1}{pT^2} < \frac{1}{T^2}$.

Consequently,

$$\mathrm{E}\left[\left(\nu_k^{\mathrm{o}} - \nu_k\right)^2\right] = \frac{1}{p^2}\mathrm{Var}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] + \mathrm{O}\left(\frac{1}{T^2}\right). \tag{C.49}$$

To prove that $\mathrm{E}\left[\left(\nu_k^{\mathrm{o}} - \nu_k\right)^2\right] = \mathrm{O}(T^{-1})$, it suffices to show that $\mathrm{Var}\left[\mathrm{tr}\left(\Sigma_k^{\mathrm{o}}\right)\right] = \mathrm{O}(p^2/T)$, or, equivalently, using the law of total variance,

$$\mathrm{E}\left[\mathrm{Var}\left(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\right)\right] = \mathrm{O}\left(\frac{p^2}{T}\right). \tag{C.50}$$

For the second term of the left-hand side of (C.50), by (C.43) and (C.48), we get

$$\mathrm{E}\left(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\right) = \mathrm{tr}\left[\mathrm{E}\left(\Sigma_k^{\mathrm{o}} \mid \boldsymbol{X}_1^T\right)\right] = \mathrm{tr}\left[\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right)\Sigma_k\right] = \left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right)\mathrm{tr}(\Sigma_k),$$

therefore,

$$\mathrm{Var}\left[\mathrm{E}\left(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\right)\right] = \mathrm{Var}\left[\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right)\mathrm{tr}(\Sigma_k)\right] = [\mathrm{tr}(\Sigma_k)]^2\,\mathrm{Var}\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right).$$

We have

$$\mathrm{Var}\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right) = \mathrm{Var}(\pi_k^{\mathrm{o}}) + \frac{1}{T^2}\mathrm{Var}(\mathbf{1}_{\{R>0\}}) - 2\mathrm{Cov}\left(\pi_k^{\mathrm{o}}, \mathbf{1}_{\{R>0\}}\right). \tag{C.51}$$

For the first term of Equation (C.51) , by Remark 1 to Theorem 7.1.1 in Brockwell and Davis (1991) (see also Corollary 4.3.2) and the mixing assumption on R_t, we get

$$T\mathrm{Var}(\pi_k^{\mathrm{o}}) = T\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} R_t\right) = T\mathrm{Var}(\bar{R}_t) \to 2\pi f(0)\,, \tag{C.52}$$

where $f(0)$ is the spectral density of $\{R_t : t \in \mathbb{Z}\}$ at frequency 0. It follows that $\mathrm{Var}\left(\pi_k^{\mathrm{o}}\right) = \mathrm{O}(1/T)$.

For the second term of Equation (C.51) , by (C.45) and the fact that $\left(\mathbf{1}_{\{R>0\}}\right)^2 = \mathbf{1}_{\{R>0\}}$, we get

$$\begin{aligned}\frac{1}{T^2}\mathrm{Var}\left(\mathbf{1}_{\{R>0\}}\right) = \frac{1}{T^2}\left(\mathrm{E}\left[\left(\mathbf{1}_{\{R>0\}}\right)^2\right] - \left[\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right]^2\right) = \frac{1}{T^2}\left(\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right) - \left[\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right]^2\right) = \\ \frac{1}{T^2}\,\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\left[1 - \mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right] = \frac{1}{T^2}\left[1 - \mathrm{P}(R=0)\right]\left(1 - \left[1 - \mathrm{P}(R=0)\right]\right) = \\ \frac{1}{T^2}\left(\mathrm{P}(R=0) - \left[\mathrm{P}(R=0)\right]^2\right) = \frac{1}{T^2}\,\mathrm{O}\left(\beta^T\right) = \mathrm{O}\left(\frac{\beta^T}{T^2}\right),\end{aligned}$$

since by the proof of the first part of Statement (i), we have $\mathrm{P}(R=0) = \mathrm{O}\left(\beta^T\right)$, which implies $\left[\mathrm{P}(R=0)\right]^2 = \mathrm{O}\left(\beta^{2T}\right)$, and $\beta^T > \beta^{2T}$, as $0 < \beta < 1$.

Combining (C.33), (C.44) and (C.45), the third term of Equation (C.51) , becomes

$$\begin{aligned}-2\mathrm{Cov}\left(\pi_k^{\mathrm{o}}, \frac{\mathbf{1}_{\{R>0\}}}{T}\right) &= -\frac{2}{T}\left[\mathrm{E}\left(\pi_k^{\mathrm{o}}\mathbf{1}_{\{R>0\}}\right) - \mathrm{E}(\pi_k^{\mathrm{o}})\,\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right] \\ &= -\frac{2}{T}\left[\mathrm{E}(\pi_k^{\mathrm{o}}) - \mathrm{E}(\pi_k^{\mathrm{o}})\,\mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right] \\ &= -\frac{2}{T}\mathrm{E}(\pi_k^{\mathrm{o}})\left[1 - \mathrm{E}\left(\mathbf{1}_{\{R>0\}}\right)\right] = -\frac{2}{T}\pi_k\left[1 - \mathrm{P}(R>0)\right] \\ &= -\frac{2}{T}\pi_k\mathrm{P}(R=0) = -\frac{2}{T}\pi_k\,\mathrm{O}\left(\beta^T\right) = \mathrm{O}\left(\frac{\beta^T}{T}\right).\end{aligned}$$

Hence, (C.51) becomes

$$\mathrm{Var}\left(\pi_k^{\mathrm{o}} - \frac{\mathbf{1}_{\{R>0\}}}{T}\right) = \mathrm{O}\left(\frac{1}{T}\right) + \mathrm{O}\left(\frac{\beta^T}{T^2}\right) + \mathrm{O}\left(\frac{\beta^T}{T}\right) = \mathrm{O}\left(\frac{1}{T}\right),$$

therefore,

$$\mathrm{Var}\left[\mathrm{E}\left(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\right)\right] = \left[\mathrm{tr}(\Sigma_k)\right]^2\mathrm{O}\left(\frac{1}{T}\right) = \mathrm{O}(p^2)\,\mathrm{O}\left(\frac{1}{T}\right) = \mathrm{O}\left(\frac{p^2}{T}\right). \tag{C.53}$$

As was mentioned in Subsection 2.3.2 (see the discussion following Equation (2.57)), if $R = 0$, then $\tilde{\Sigma}_k^{\mathrm{o}}$ is set equal to $\mathbf{0}_{p\times p}$ and it follows that $\Sigma_k^{\mathrm{o}} = \pi_k^{\mathrm{o}}\tilde{\Sigma}_k^{\mathrm{o}} = \mathbf{0}_{p\times p}$. Consequently, if $R = 0$, then $\mathrm{tr}\left(\Sigma_k^{\mathrm{o}}\right) = 0$. Therefore, to calculate the conditional variance of the first term of Equation (C.50),

let us assume for the moment that $R > 0$. As has already been mentioned, η_t, $t \in \{1, \ldots, T\}$ are iid random vectors, independent of the state variables $\{X_1, \ldots, X_T\}$, with

$$\mathrm{E}(\eta_t) = \mathbf{0}_p$$

and

$$\mathrm{Var}(\eta_t) = \Sigma_k.$$

Notice that if $R_t = 1$, then

$$Y_t = \mu_k + \Sigma_k \, \epsilon_t = \mu_k + \eta_t$$

which implies that

$$R_t \left(Y_t - \mu_k^{\mathrm{o}}\right) \left(Y_t - \mu_k^{\mathrm{o}}\right)' = R_t \left(\mu_k + \eta_t - \mu_k^{\mathrm{o}}\right) \left(\mu_k + \eta_t - \mu_k^{\mathrm{o}}\right)' = R_t \left(\eta_t - \delta_k\right) \left(\eta_t - \delta_k\right)'.$$

As for any matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$,

$$\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}\left(\boldsymbol{A}'\right),$$

we get

$$\begin{aligned}
\mathrm{tr}(\Sigma_k^{\mathrm{o}}) &= \frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{tr}\left[\left(\eta_t - \delta_k\right) \left(\eta_t - \delta_k\right)'\right] \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \mathrm{tr}\left[\left(\eta_t - \delta_k\right)' \left(\eta_t - \delta_k\right)\right] \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \left(\eta_t - \delta_k\right)' \left(\eta_t - \delta_k\right) \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \left(\eta_t{}' \eta_t - \eta_t{}' \delta_k - \delta_k{}' \eta_t + \delta_k{}' \delta_k\right) \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \eta_t{}' \eta_t - \frac{2}{T} \sum_{t=1}^{T} R_t \eta_t{}' \delta_k + \frac{1}{T} \sum_{t=1}^{T} R_t \delta_k{}' \delta_k \\
&= \frac{1}{T} \sum_{t=1}^{T} R_t \left|\left|\eta_t\right|\right|_2^2 + \frac{1}{T} \sum_{t=1}^{T} R_t \left|\left|\delta_k\right|\right|_2^2 - \frac{2}{T} \sum_{t=1}^{T} R_t \eta_t{}' \delta_k,
\end{aligned}$$

where we have used the facts that

$$\mathrm{tr}\left[\left(\eta_t - \delta_k\right)' \left(\eta_t - \delta_k\right)\right] = \left(\eta_t - \delta_k\right)' \left(\eta_t - \delta_k\right), \tag{C.54}$$

as $\left(\eta_t - \delta_k\right)' \left(\eta_t - \delta_k\right)$ is scalar and $\delta_k{}' \eta_t = \eta_t{}' \delta_k$.

Since we have assumed that $R > 0$, by Equation (C.27), we get

$$\delta_k = \frac{1}{R}\sum_{t=1}^{T} R_t \eta_t,$$

hence,

$$-\frac{2}{T}\sum_{t=1}^{T} R_t {\eta_t}' \delta_k = -\frac{2}{T} R \frac{1}{R}\sum_{t=1}^{T} R_t {\eta_t}' \delta_k = -2\frac{R}{T}\left(\frac{1}{R}\sum_{t=1}^{T} R_t \eta_t\right)' \delta_k = -2\pi_k^{\mathrm{o}} {\delta_k}' \delta_k = -2\pi_k^{\mathrm{o}} ||\delta_k||_2^2.$$

As a result, we get

$$\mathrm{tr}(\Sigma_k^{\mathrm{o}}) = \frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||_2^2 - \pi_k^{\mathrm{o}} ||\delta_k||_2^2, \tag{C.55}$$

consequently,

$$\begin{aligned}
\mathrm{Var}\Big(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\Big) &= \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||_2^2 - \pi_k^{\mathrm{o}} ||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) \\
&= \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) + \mathrm{Var}\Big(\pi_k^{\mathrm{o}} ||\delta_k||_2^2 \,\Big|\, \boldsymbol{X}_1^T\Big) \\
&\quad - 2\mathrm{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||_2^2,\ \pi_k^{\mathrm{o}} ||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right).
\end{aligned} \tag{C.56}$$

As has already been mentioned, η_t, $t \in \{1, \ldots, T\}$ are iid and independent of the Markov chain, and since, by Equation (C.5), $R_t^2 = R_t$, we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) &= \frac{1}{T^2}\sum_{t=1}^{T} R_t^2 \mathrm{Var}\Big(||\eta_t||_2^2 \,\Big|\, \boldsymbol{X}_1^T\Big) \\
&= \frac{1}{T^2}\sum_{t=1}^{T} R_t \mathrm{Var}\Big(||\eta_1||_2^2\Big) = \frac{1}{T}\pi_k^{\mathrm{o}} \mathrm{Var}\Big(||\eta_1||_2^2\Big) \\
&\le \frac{1}{T}\pi_k^{\mathrm{o}} \mathrm{E}\Big(||\eta_1||_2^4\Big) = \pi_k^{\mathrm{o}}\, \mathrm{O}\left(\frac{p^2}{T}\right).
\end{aligned} \tag{C.57}$$

To prove the last relation of (C.57), recall that by $\mathrm{Var}(\epsilon_{t,i}) = 1$ and by Assumption (A3) (see Subsection 2.3.1), $\mathrm{E}\Big(\epsilon_{t,i}^4\Big) \le \kappa_\epsilon + 3$, for all $t \in \mathbb{Z}$ and $i \in \{1, \ldots, p\}$. Let us denote the i-th row of the matrix $\Sigma_k^{1/2}$ by $\rho(i) = \big(\rho_{i1}, \ldots, \rho_{ip}\big)$. Since $\Sigma_k^{1/2}$, it also holds $\rho(i) = \big(\rho_{1i}, \ldots, \rho_{pi}\big)$. Then

$$
\begin{aligned}
\mathrm{E}\left(||\eta_t||_2^4\right) &= \mathrm{E}\left(\left|\left|\Sigma_k^{1/2}\epsilon_t\right|\right|_2^4\right) \\
&\leq \kappa_\epsilon \sum_{k=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\rho_{ik}^2\rho_{jk}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\rho_{il}^2\rho_{jk}^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\rho_{il}\rho_{jl}\rho_{ik}\rho_{jk} \\
&= \kappa_\epsilon \sum_{k=1}^{p}||\rho(k)||_2^4 + p^2\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}}^4 + 2\sum_{k=1}^{p}\sum_{l=1}^{p}\left(\rho(l)\,\rho(k)^{'}\right)^2 \\
&\leq \kappa_\epsilon \sum_{k=1}^{p}\sum_{l=1}^{p}||\rho(l)||_2^2\,||\rho(k)||_2^2 + p^2\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}}^4 + 2\sum_{k=1}^{p}\sum_{l=1}^{p}||\rho(k)||_2^2\,||\rho(l)||_2^2 \\
&= (\kappa_\epsilon + 3)\,p^2\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}}^4 = \mathrm{O}(p^2)\,,
\end{aligned} \tag{C.58}
$$

since, by Assumption (B1), $\left|\left|\Sigma_k^{1/2}\right|\right|_{\mathrm{F}}^2 = \mathrm{O}(1)$.

For the second term of Equation (C.56), by Equation (C.27), we get

$$
\begin{aligned}
\mathrm{Var}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) &= (\pi_k^{\mathrm{o}})^2\,\mathrm{Var}\left(||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) \\
&= (\pi_k^{\mathrm{o}})^2\,\mathrm{Var}\left[\left(\frac{1}{R}\sum_{t=1}^{T}R_t\eta_t\right)^{'}\left(\frac{1}{R}\sum_{t=1}^{T}R_t\eta_t\right)\,\middle|\, \boldsymbol{X}_1^T\right] \\
&= \frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^4}\mathrm{Var}\left[\left(\sum_{t=1}^{T}R_t\eta_t\right)^{'}\left(\sum_{t=1}^{T}R_t\eta_t\right)\,\middle|\, \boldsymbol{X}_1^T\right],
\end{aligned} \tag{C.59}
$$

and

$$\begin{aligned}
\operatorname{Var}\left[\left(\sum_{t=1}^{T} R_t\eta_t\right)'\left(\sum_{t=1}^{T} R_t\eta_t\right)\,\middle|\,\boldsymbol{X}_1^T\right] &= \\
\operatorname{Var}\left[\sum_{i=1}^{p}\left[\left(\sum_{t=1}^{T} R_t\eta_t\right)_i\right]^2\,\middle|\,\boldsymbol{X}_1^T\right] &= \\
\operatorname{Var}\left[\sum_{i=1}^{p}\left(\sum_{t=1}^{T} R_t\eta_{t,i}\right)^2\,\middle|\,\boldsymbol{X}_1^T\right] &= \\
\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\operatorname{Cov}\left[\left(\sum_{t=1}^{T} R_t\eta_{t,i_1}\right)^2,\left(\sum_{t=1}^{T} R_t\eta_{t,i_2}\right)^2\,\middle|\,\boldsymbol{X}_1^T\right] &= \\
\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\operatorname{Cov}\left(\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} R_{t_1}\eta_{t_1,i_1}R_{t_2}\eta_{t_2,i_1},\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{s_1}\eta_{s_1,i_2}R_{s_2}\eta_{s_2,i_2}\,\middle|\,\boldsymbol{X}_1^T\right) &= \\
\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\operatorname{Cov}\left(\eta_{t_1,i_1}\eta_{t_2,i_1}\,,\,\eta_{s_1,i_2}\eta_{s_2,i_2}\mid\boldsymbol{X}_1^T\right) &= \\
\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\operatorname{Cov}(\eta_{t_1,i_1}\eta_{t_2,i_1},\eta_{s_1,i_2}\eta_{s_2,i_2}) &= \\
\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\operatorname{Cov}\left(\sum_{i_1=1}^{T}\eta_{t_1,i_1}\eta_{t_2,i_1},\sum_{i_2=1}^{T}\eta_{s_1,i_2}\eta_{s_2,i_2}\right) &= \\
\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\operatorname{Cov}\left({\eta_{t_1}}'\eta_{t_2},{\eta_{s_1}}'\eta_{s_2}\right). &\qquad \text{(C.60)}
\end{aligned}$$

As η_t, $t \in \{1,\dots,T\}$ are iid with zero mean,

$$\operatorname{Cov}(\eta_{t_1,i_1}\eta_{t_2,i_1},\eta_{s_1,i_2}\eta_{s_2,i_2}) \neq 0,$$

if $t_1 = t_2 = s_1 = s_2$, or $t_1 = s_1 \neq t_2 = s_2$, or $t_1 = s_2 \neq t_1 = s_1$. In any other case there is at least one factor say $\eta_{a,b}$, with $(a,b) \in \{t,s\} \times \{i,j\}$, independent of the rest. Let us consider the case where t_1 differs from t_2, s_1 and s_2, while the rest can either be equal or not. We have

$$\begin{aligned}
\operatorname{Cov}(\eta_{t_1,i_1}\eta_{t_2,i_1},\eta_{s_1,i_2}\eta_{s_2,i_2}) &= \operatorname{E}(\eta_{t_1,i_1}\eta_{t_2,i_1}\eta_{s_1,i_2}\eta_{s_2,i_2}) - \operatorname{E}(\eta_{t_1,i_1}\eta_{t_2,i_1})\operatorname{E}(\eta_{s_1,i_2}\eta_{s_2,i_2}) \\
&= \operatorname{E}(\eta_{t_1,i_1})\operatorname{E}(\eta_{t_2,i_1}\eta_{s_1,i_2}\eta_{s_2,i_2}) - \operatorname{E}(\eta_{t_1,i_1})\operatorname{E}(\eta_{t_2,i_1})\operatorname{E}(\eta_{s_1,i_2}\eta_{s_2,i_2}) \\
&= 0,
\end{aligned}$$

as $\operatorname{E}(\eta_{t,i}) = 0$, for all $t \in \{1,\dots,T\}$ and $i \in \{1,\dots,p\}$. The same can be proved for any other index and any other case where there is more than one inequality between the indices t_1, t_2, s_1 and s_2.

Let us, now examine the cases, where

$$\mathrm{Cov}(\eta_{t_1,i_1}\eta_{t_2,i_1}, \eta_{s_1,i_2}\eta_{s_2,i_2}) \neq 0.$$

- Case $t_1 = t_2 = s_1 = s_2$:

  In this case we have

$$\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\mathrm{Cov}\left(\eta_{t_1}{}'\eta_{t_2}, \eta_{s_1}{}'\eta_{s_2}\right) = \sum_{t_1=1}^{T} R_t^4\mathrm{Cov}\left(\eta_t{}'\eta_t, \eta_t{}'\eta_t\right)$$
$$= \sum_{t_1=1}^{T} R_t^4\mathrm{Var}\left(\eta_t{}'\eta_t\right).$$

  Combining (C.28), (C.57) (C.59) and (C.60), we get

$$\mathrm{Var}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^4}\sum_{t_1=1}^{T} R_t^4\mathrm{Var}\left(\eta_t{}'\eta_t\right) = \frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^4}\sum_{t_1=1}^{T} R_t\mathrm{Var}\left(||\eta_1||_2^2\right)$$
$$\leq \frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^3}\mathrm{O}(p^2) = \left(\frac{R}{T}\right)^2\frac{1}{R^3}\mathrm{O}(p^2)$$
$$= \frac{1}{RT^2}\mathrm{O}(p^2) = \mathrm{O}\left(\frac{p^2}{T^2}\right), \tag{C.61}$$

  since, by Equation (C.5), $R_t^4 = R_t$ and we have assumed, temporarily, that $R > 0$, which implies that $R \geq 1$.

- Cases $t_1 = s_1 \neq t_2 = s_2$, $t_1 = s_2 \neq t_2 = s_1$:

  Notice that these cases are equivalent due to symmetry. Let us present the first one.

  Since $t_1 = s_1 \neq t_2 = s_2$, Equation (C.60) becomes

$$\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1}R_{t_2}R_{s_1}R_{s_2}\mathrm{Cov}\left(\eta_{t_1}{}'\eta_{t_2}, \eta_{s_1}{}'\eta_{s_2}\right) = \sum_{t=1}^{T}\sum_{\substack{s=1\\ s\neq t}}^{T} R_tR_s\mathrm{Var}\left(\eta_t{}'\eta_s\right)$$

  Since η_t, $t \in \{1, \ldots, T\}$ are iid , it holds

$$f\left(\eta_{t_1}{}'\eta_{s_1}, \eta_{t_2}{}'\eta_{s_2}\right) = f\left(\eta_{t_1,1}\eta_{s_1,1} + \ldots + \eta_{t_1,p}\eta_{s_1,p}, \eta_{t_2,1}\eta_{s_2,1} + \ldots + \eta_{t_2,p}\eta_{s_2,p}\right)$$
$$= f\left(\eta_{t_1,1}\eta_{s_1,1} + \ldots + \eta_{t_1,p}\eta_{s_1,p}\right) f\left(\eta_{t_2,1}\eta_{s_2,1} + \ldots + \eta_{t_2,p}\eta_{s_2,p}\right)$$
$$= f\left(\eta_{t_1}{}'\eta_{s_1}\right) f\left(\eta_{t_2}{}'\eta_{s_2}\right),$$

  where by $f(\cdot)$ we denote the corresponding probability density functions of each of these random variables. The above relation shows that $\eta_t{}'\eta_s$, $t, s \in \{1, \ldots, T\}$ are also iid . Thus,

we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{\substack{s=1\\s\neq t}}^{T} R_t R_s \mathrm{Var}\left(\eta_t{}'\eta_s\right) &= \sum_{t=1}^{T}\sum_{\substack{s=1\\s\neq t}}^{T} R_t R_s \mathrm{Var}\left(\eta_1{}'\eta_2\right)\\ &\leq \sum_{t=1}^{T}\sum_{\substack{s=1\\s\neq t}}^{T} R_t R_s \mathrm{E}\left[\left(\eta_1{}'\eta_2\right)^2\right]\\ &\leq \sum_{t=1}^{T}\sum_{s=1}^{T} R_t R_s \mathrm{E}\left[\left(\eta_1{}'\eta_2\right)^2\right]\\ &= R^2\mathrm{E}\left[\left(\eta_1{}'\eta_2\right)^2\right], \end{aligned} \tag{C.62}$$

since $R_t R_s \mathrm{E}\left[\left(\eta_1{}'\eta_2\right)^2\right] \geq 0$, for all $t, s \in \{1, \ldots, T\}$.

By the Cauchy–Schwarz inequality and Jensen's inequality, we get

$$\begin{aligned}\mathrm{E}\left[\left(\eta_1{}'\eta_2\right)^2\right] &\leq \mathrm{E}\left(||\eta_1||_2^2\,||\eta_2||_2^2\right) = \mathrm{E}\left(||\eta_1||_2^2\right)\mathrm{E}\left(||\eta_2||_2^2\right)\\ &= \left[\mathrm{E}\left(||\eta_1||_2^2\right)\right]^2 \leq \mathrm{E}\left(||\eta_1||_2^4\right). \end{aligned} \tag{C.63}$$

Combining (C.58), (C.59), (C.62) and (C.63), we get

$$\begin{aligned}\mathrm{Var}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) &= \frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^4}\sum_{t=1}^{T}\sum_{\substack{s=1\\s\neq t}}^{T} R_t R_s \mathrm{Var}\left(\eta_t{}'\eta_s\right)\\ &\leq \left(\frac{R}{T}\right)^2 \frac{1}{R^4} R^2 \mathrm{E}\left(||\eta_1||_2^4\right)\\ &\leq \frac{1}{T^2}\,\mathrm{O}(p^2) = \mathrm{O}\left(\frac{p^2}{T^2}\right). \end{aligned}$$

Consequently, we proved that it always holds that

$$\mathrm{Var}\left(\pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \mathrm{O}\left(\frac{p^2}{T^2}\right). \tag{C.64}$$

For the third term of Equation (C.56) we have

$$
\begin{aligned}
&\operatorname{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} R_t\,||\eta_t||_2^2\,,\ \pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\operatorname{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} R_t\,||\eta_t||_2^2\,,\ \frac{R}{T}\left|\left|\frac{1}{R}\sum_{s=1}^{T} R_s\eta_s\right|\right|_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{R}{T^2}\operatorname{Cov}\left(\sum_{t=1}^{T} R_t\eta_t{}'\eta_t\,,\ \frac{1}{R^2}\left(\sum_{s=1}^{T} R_s\eta_s\right)'\left(\sum_{s=1}^{T} R_s\eta_s\right) \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\operatorname{Cov}\left(\sum_{t=1}^{T}\left(R_t\sum_{i=1}^{p}\eta_{t,i}^2\right)\,,\ \sum_{j=1}^{p}\left(\sum_{s=1}^{T} R_s\eta_{s,j}\right)^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\operatorname{Cov}\left(\sum_{i=1}^{p}\sum_{t=1}^{T} R_t\eta_{t,i}^2\,,\ \sum_{j=1}^{p}\left(\sum_{s=1}^{T} R_s\eta_{s,j}\right)^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\sum_{t=1}^{T} R_t\eta_{t,i}^2\,,\ \left(\sum_{s=1}^{T} R_s\eta_{s,j}\right)^2 \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\sum_{t=1}^{T} R_t\eta_{t,i}^2\,,\ \sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{s_1}\eta_{s_1,j}R_{s_2}\eta_{s_2,j} \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{t=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_tR_{s_1}R_{s_2}\operatorname{Cov}\left(\eta_{t,i}^2\,,\ \eta_{s_1,j}\eta_{s_2,j} \mid \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\sum_{t=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_tR_{s_1}R_{s_2}\operatorname{Cov}\left(\sum_{i=1}^{p}\eta_{t,i}^2\,,\ \sum_{j=1}^{p}\eta_{s_1,j}\eta_{s_2,j} \,\middle|\, \boldsymbol{X}_1^T\right) = \\
&\frac{1}{RT^2}\sum_{t=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_tR_{s_1}R_{s_2}\operatorname{Cov}\left(\eta_t{}'\eta_t\,,\ \eta_{s_1}{}'\eta_{s_2} \,\middle|\, \boldsymbol{X}_1^T\right). \qquad \text{(C.65)}
\end{aligned}
$$

Similarly as before,

$$
\operatorname{Cov}\left(\eta_t{}'\eta_t\,,\ \eta_{s_1,j}{}'\eta_{s_2,j} \,\middle|\, \boldsymbol{X}_1^T\right) \neq 0,
$$

if $t = s_1 = s_2$.

By (C.57) and (C.65), we get

$$
\begin{aligned}
\operatorname{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} R_t\,||\eta_t||_2^2\,,\ \pi_k^{\mathrm{o}}\,||\delta_k||_2^2 \,\middle|\, \boldsymbol{X}_1^T\right) &= \frac{1}{RT^2}\sum_{t=1}^{T} R_t^3\operatorname{Cov}\left(\eta_t{}'\eta_t\,,\ \eta_t{}'\eta_t \,\middle|\, \boldsymbol{X}_1^T\right) \\
&= \frac{1}{RT^2}\sum_{t=1}^{T} R_t\operatorname{Var}\left(\eta_t{}'\eta_t\right) = \frac{1}{RT^2}\sum_{t=1}^{T} R_t\operatorname{Var}\left(||\eta_t||_2^2\right) \\
&= \frac{1}{RT^2}\sum_{t=1}^{T} R_t\operatorname{Var}\left(||\eta_1||_2^2\right) = \frac{1}{T^2}\operatorname{Var}\left(||\eta_1||_2^2\right) \\
&\leq \frac{1}{T^2}\,\mathrm{O}(p^2) = \mathrm{O}\left(\frac{p^2}{T^2}\right) \qquad \text{(C.66)}
\end{aligned}
$$

Combining (C.56), (C.57), (C.64) and (C.66) we get

$$\mathrm{Var}\Big(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\Big) = \mathrm{O}\left(\frac{p^2}{T^2}\right) + \mathrm{O}\left(\frac{p^2}{T^2}\right) - 2\,\mathrm{O}\left(\frac{p^2}{T^2}\right) = \mathrm{O}\left(\frac{p^2}{T^2}\right) + \mathrm{O}\left(\frac{p^2}{T^2}\right) + \mathrm{O}\left(\frac{p^2}{T^2}\right) = \mathrm{O}\left(\frac{p^2}{T^2}\right).$$

It follows that

$$\mathrm{E}\Big[\mathrm{Var}\Big(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\Big)\Big] = \mathrm{O}\left(\frac{p^2}{T^2}\right). \tag{C.67}$$

Finally, Equation (C.50) is ensured by (C.53) and (C.67), in the case where $R > 0$, that is

$$\begin{aligned}\mathrm{Var}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] &= \mathrm{E}\Big[\mathrm{Var}\Big(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\Big)\Big] + \mathrm{Var}\Big[\mathrm{E}\Big(\mathrm{tr}(\Sigma_k^{\mathrm{o}}) \mid \boldsymbol{X}_1^T\Big)\Big] \\ &= \mathrm{O}\left(\frac{p^2}{T^2}\right) + \mathrm{O}\left(\frac{p^2}{T}\right) = \mathrm{O}\left(\frac{p^2}{T}\right).\end{aligned}$$

As we have already mentioned, if $R = 0$, then $\mathrm{tr}\big(\Sigma_k^{\mathrm{o}}\big) = 0$, which implies that $\mathrm{Var}\big[\mathrm{tr}\big(\Sigma_k^{\mathrm{o}}\big)\big] = 0$ and since $0 < p^2/T$, in any case, it holds

$$\mathrm{Var}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] = \mathrm{O}\left(\frac{p^2}{T}\right),$$

which by Equation (C.49) concludes the proof of Statement (ii).

(iii) From the proof of Statement (ii) we have

$$\mathrm{Var}[\mathrm{tr}(\Sigma_k^{\mathrm{o}})] = \mathrm{O}\left(\frac{p^2}{T}\right).$$

Under Assumption (B2) there exist some constants $c > 0$ and $0 \leq r \leq 1$, such that

$$\mathrm{tr}(\Sigma_k) \geq c p^r.$$

Thus, we get

$$\frac{\mathrm{Var}\big[\mathrm{tr}\big(\Sigma_k^{\mathrm{o}}\big)\big]}{\big|\mathrm{tr}(\pi_k \Sigma_k)\big|^2} = \frac{\mathrm{Var}\big[\mathrm{tr}\big(\Sigma_k^{\mathrm{o}}\big)\big]}{\pi_k^2\,\big|\mathrm{tr}(\Sigma_k)\big|^2} \leq \frac{\mathrm{Var}\big[\mathrm{tr}\big(\Sigma_k^{\mathrm{o}}\big)\big]}{\pi_k^2 (c p^r)^2} = \frac{1}{\pi_k^2 c^2}\frac{1}{p^{2r}}\,\mathrm{O}\left(\frac{p^2}{T}\right) = \mathrm{O}\left(\frac{p^{2(1-r)}}{T}\right), \tag{C.68}$$

since, by Assumption (A1), $\pi_k > 0$, for all $k \in \{1, \ldots, K\}$. This completes the proof of Statement (iii).

□

C.2.4 Technical Lemma 3

Lemma C.3, below, shows that substituting μ_k^{o} for the true unknown value μ in the definition of Σ_k^{o} (see Equation (2.58)), asymptotically has a negligible effect. Therefore, we define $\Sigma_k^{\text{o}*}$ as

$$\Sigma_k^{\text{o}*} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{X_t=k\}} \left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)^{'}. \tag{C.69}$$

Notice that for all $i, j \in \{1, \ldots, p\}$, the (i,j)-th element of $\Sigma_k^{\text{o}*}$, $\Sigma_{k,ij}^{\text{o}*}$ is the sample mean of the univariate time series $\boldsymbol{Z}_{k,ij}(t) = z_t = \mathbf{1}_{\{X_t=k\}} \left(X_{t,i} - \mu_{k,i}\right)\left(X_{t,j} - \mu_{k,j}\right)$.

Lemma C.3. *Under the conditions of* Lemma 3.9

(i)
$$\mathrm{E}\left(||\Sigma_k^{\text{o}*} - \Sigma_k^{\text{o}}||_{\mathrm{F}}^2\right) = \mathrm{O}\left(\frac{p}{T^2}\right),$$

(ii)
$$\mathrm{E}\left(||\Sigma_k^{\text{o}} - \pi_k\Sigma_k||_{\mathrm{F}}^2\right) = \mathrm{E}\left(||\Sigma_k^{\text{o}*} - \pi_k\Sigma_k||_{\mathrm{F}}^2\right) + \mathrm{O}\left(\frac{p}{T^2}\right).$$

Proof. (i) Taking the difference between $\Sigma_k^{\text{o}*}$ and Σ_k^{o}, by Equations (2.58) and (C.69), we get

$$\begin{aligned}
\Sigma_k^{\text{o}*} - \Sigma_k^{\text{o}} &= \frac{1}{T}\sum_{t=1}^{T} R_t \left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)^{'} - \frac{1}{T}\sum_{t=1}^{T} R_t \left(Y_t - \mu_k^{\text{o}}\right)\left(Y_t - \mu_k^{\text{o}}\right)^{'} \\
&= \frac{1}{T}\sum_{t=1}^{T} R_t \left[\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)^{'} - \left(Y_t - \mu_k^{\text{o}}\right)\left(Y_t - \mu_k^{\text{o}}\right)^{'}\right].
\end{aligned}$$

We have that

$$\begin{gathered}
\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)^{'} - \left(Y_t - \mu_k^{\text{o}}\right)\left(Y_t - \mu_k^{\text{o}}\right)^{'} = \\
Y_t Y_t^{'} - Y_t \mu_k^{'} - \mu_k Y_t^{'} + \mu_k \mu_k^{'} - Y_t Y_t^{'} + Y_t \left(\mu_k^{\text{o}}\right)^{'} + \mu_k^{\text{o}} Y_t^{'} - \mu_k^{\text{o}}\left(\mu_k^{\text{o}}\right)^{'} = \\
Y_t\left[\left(\mu_k^{\text{o}}\right)^{'} - \mu_k^{'}\right] + \left(\mu_k^{\text{o}} - \mu_k\right) Y_t^{'} + \mu_k \mu_k^{'} - \mu_k^{\text{o}}\left(\mu_k^{\text{o}}\right)^{'}.
\end{gathered}$$

Hence, using Equations (2.55), (C.24) and (C.28), we get

$$\begin{aligned}
\Sigma_k^{o*} - \Sigma_k^o &= \frac{1}{T}\sum_{t=1}^{T} R_t \left[Y_t \left[\left(\mu_k^o\right)' - \mu_k' \right] + \left(\mu_k^o - \mu_k\right) Y_t' + \mu_k \mu_k' - \mu_k^o \left(\mu_k^o\right)' \right] \\
&= \frac{1}{T}\sum_{t=1}^{T} R_t Y_t \left[\left(\mu_k^o\right)' - \mu_k' \right] + \frac{1}{T}\sum_{t=1}^{T} R_t \left(\mu_k^o - \mu_k\right) Y_t' + \frac{1}{T}\sum_{t=1}^{T} R_t \mu_k \mu_k' - \frac{1}{T}\sum_{t=1}^{T} R_t \mu_k^o \left(\mu_k^o\right)' \\
&= \frac{1}{T}\sum_{t=1}^{T} R_t Y_t \left[\left(\mu_k^o\right)' - \mu_k' \right] + \left(\mu_k^o - \mu_k\right) \left(\frac{1}{T}\sum_{t=1}^{T} R_t Y_t \right)' + \frac{1}{T}\sum_{t=1}^{T} R_t \mu_k \mu_k' - \frac{1}{T}\sum_{t=1}^{T} R_t \mu_k^o \left(\mu_k^o\right)' \\
&= \frac{R}{T}\mu_k^o \left[\left(\mu_k^o\right)' - \mu_k' \right] + \frac{R}{T}\left(\mu_k^o - \mu_k\right)\left(\mu_k^o\right)' + \frac{R}{T}\mu_k \mu_k' - \frac{R}{T}\mu_k^o \left(\mu_k^o\right)' \\
&= \frac{R}{T}\mu_k^o \left(\mu_k^o\right)' - \frac{R}{T}\mu_k^o \mu_k' + \frac{R}{T}\mu_k^o \left(\mu_k^o\right)' - \frac{R}{T}\mu_k \left(\mu_k^o\right)' + \frac{R}{T}\mu_k \mu_k' - \frac{R}{T}\mu_k^o \left(\mu_k^o\right)' \\
&= \frac{R}{T}\left(\mu_k^o \left(\mu_k^o\right)' - \mu_k^o \mu_k' - \mu_k \left(\mu_k^o\right)' + \mu_k \mu_k' \right) = \frac{R}{T}\left[\mu_k^o \left(\mu_k^o - \mu_k\right)' - \mu_k \left(\mu_k^o - \mu_k\right)' \right] \\
&= \frac{R}{T}\left(\mu_k^o - \mu_k\right)\left(\mu_k^o - \mu_k\right)' = \frac{R}{T}\delta_k \delta_k' = \pi_k^o \delta_k \delta_k'. \qquad \text{(C.70)}
\end{aligned}$$

As we have already mentioned, by the definition of π_k^o, we have $\pi_k^o \cdot \mathbf{1}_{\{R=0\}} = 0$ and $\pi_k^o \cdot \mathbf{1}_{\{R>0\}} = \pi_k^o$. Therefore, using (C.27), Equation (C.70) can also be expressed as

$$\Sigma_k^{o*} - \Sigma_k^o = \pi_k^o \frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{t=1}^{T} R_t \eta_t \left(\frac{\mathbf{1}_{\{R>0\}}}{R} \sum_{t=1}^{T} R_t \eta_t \right)' = \frac{\pi_k^o}{R^2} \sum_{t=1}^{T}\sum_{s=1}^{T} R_t R_s \eta_t \eta_s'. \qquad \text{(C.71)}$$

By the law of total expectation we have

$$\mathrm{E}\left[\left(\Sigma_{k,ij}^{o*} - \Sigma_{k,ij}^o \right)^2 \right] = \mathrm{E}\left(\mathrm{E}\left[\left(\Sigma_{k,ij}^{o*} - \Sigma_{k,ij}^o \right)^2 \middle| \boldsymbol{X}_1^T \right] \right). \qquad \text{(C.72)}$$

As η_t, $t \in \{1, \ldots, T\}$ are iid and independent of the Markov chain, and $\left(\eta_t \eta_s'\right)_{ij} = \eta_{t,i}\eta_{s,j}$, by (C.71) and the linearity of the expectation, we get

$$\begin{aligned}
\mathrm{E}\left[\left(\Sigma_{k,ij}^{o*} - \Sigma_{k,ij}^o \right)^2 \middle| \boldsymbol{X}_1^T \right] &= \mathrm{E}\left(\left[\left(\frac{\pi_k^o}{R^2} \sum_{t=1}^{T}\sum_{s=1}^{T} R_t R_s \eta_t \eta_s' \right)_{ij} \right]^2 \middle| \boldsymbol{X}_1^T \right) \\
&= \mathrm{E}\left[\left(\frac{\pi_k^o}{R^2} \sum_{t=1}^{T}\sum_{s=1}^{T} R_t R_s \eta_{t,i}\eta_{s,j} \right)^2 \middle| \boldsymbol{X}_1^T \right] \\
&= \left(\frac{\pi_k^o}{R^2} \right)^2 \mathrm{E}\left(\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1} R_{t_2} R_{s_1} R_{s_2} \eta_{t_1,i}\eta_{t_2,i}\eta_{s_1,j}\eta_{s_2,j} \middle| \boldsymbol{X}_1^T \right) \\
&= \frac{\left(\pi_k^o\right)^2}{R^4} \sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T} R_{t_1} R_{t_2} R_{s_1} R_{s_2} \mathrm{E}\left(\eta_{t_1,i}\eta_{t_2,i}\eta_{s_1,j}\eta_{s_2,j} \right). \qquad \text{(C.73)}
\end{aligned}$$

Similarly as in the case of $\mathrm{Cov}(\eta_{t_1,i_1}\eta_{t_2,i_1}, \eta_{s_1,i_2}\eta_{s_2,i_2})$, in the proof of Statement (ii) of Lemma 3.9 (see Subsection C.2.3), we have that

$$\mathrm{E}\left(\eta_{t_1,i}\eta_{t_2,i}\eta_{s_1,j}\eta_{s_2,j}\right) \neq 0,$$

if $t_1 = t_2 = s_1 = s_2$, or $t_1 = t_2 \neq s_1 = s_2$, or $t_1 = s_1 \neq t_2 = s_2$, or $t_1 = s_2 \neq t_2 = s_1$. Due to symmetry, the last two cases are equivalent.

Before we proceed to examine these four cases, we will prove a useful relationship.

Recall that ϵ_t, $t \in \{1, \ldots, T\}$ are iid with

$$\mathrm{E}(\epsilon_t) = \mathbf{0}_p,$$

and

$$\mathrm{Var}(\epsilon_t) = \boldsymbol{I}_p.$$

It follows that

$$\begin{aligned}
\mathrm{E}\left(\eta_t\eta_t{}'\right) &= \mathrm{E}\left[\Sigma_k^{1/2}\epsilon_t\left(\Sigma_k^{1/2}\epsilon_t\right)'\right] = \mathrm{E}\left[\Sigma_k^{1/2}\epsilon_t\epsilon_t{}'\left(\Sigma_k^{1/2}\right)'\right] \\
&= \Sigma_k^{1/2}\mathrm{E}\left(\epsilon_t\epsilon_t{}'\right)\left(\Sigma_k^{1/2}\right)' = \Sigma_k^{1/2}\mathrm{Var}\left(\epsilon_t\epsilon_t{}'\right)\left(\Sigma_k^{1/2}\right)' \\
&= \Sigma_k^{1/2}\boldsymbol{I}_p\left(\Sigma_k^{1/2}\right)' = \Sigma_k. \qquad \text{(C.74)}
\end{aligned}$$

As $\left(\eta_t\eta_s{}'\right)_{ij} = \eta_{t,i}\eta_{s,j}$ (see also Definition A.38), we have that

$$\mathrm{E}\left(\eta_{t,i}\eta_{t,j}\right) = \mathrm{E}\left[\left(\eta_t\eta_t{}'\right)_{ij}\right] = \left[\mathrm{E}\left(\eta_t\eta_t{}'\right)\right]_{ij},$$

consequently, for all $t \in \{1, \ldots, T\}$ and $i, j \in \{1, \ldots, p\}$, it holds

$$\mathrm{E}\left(\eta_{t,i}\eta_{t,j}\right) = \Sigma_{k,ij}. \qquad \text{(C.75)}$$

Let us, now, examine these four cases.

- Case $t_1 = t_2 = s_1 = s_2$:

$$\mathrm{E}\left(\eta_{t,i}^2\eta_{t,j}^2\right) = \mathrm{E}\left[\left(\eta_{t,i}\eta_{t,j}\right)^2\right].$$

- Case $t_1 = t_2 \neq s_1 = s_2$:

Since η_t, $t \in \{1,\dots,T\}$ are iid , for all $t,s \in \{1,\dots,T\}$, with $t \neq s$ and for all $i,j \in \{1,\dots,p\}$, $\eta_{t,i}$ and $\eta_{s,j}$ are independent random variables, hence $\eta_{t,i}^2$ and $\eta_{s,j}^2$ are also independent. By Equation (C.75), we get

$$\mathrm{E}\left(\eta_{t,i}^2\eta_{s,j}^2\right) = \mathrm{E}\left(\eta_{t,i}^2\right)\mathrm{E}\left(\eta_{s,j}^2\right) = \Sigma_{k,ii}\Sigma_{k,jj}.$$

- Cases $t_1 = s_1 \neq t_2 = s_2$ and $t_1 = s_2 \neq t_2 = s_1$:

$$\mathrm{E}\left(\eta_{t_1,i}\eta_{t_2,i}\eta_{t_1,j}\eta_{t_2,j}\right) = \mathrm{E}\left(\eta_{t_1,i}\eta_{t_1,j}\right)\mathrm{E}\left(\eta_{t_2,i}\eta_{t_2,j}\right) = \Sigma_{k,ij}\Sigma_{k,ij} = \Sigma_{k,ij}^2.$$

Therefore, Equation (C.73) is written as

$$\begin{aligned}
&\mathrm{E}\left[\left(\Sigma_{k,ij}^{\mathrm{o}*} - \Sigma_{k,ij}^{\mathrm{o}}\right)^2 \,\middle|\, \boldsymbol{X}_1^T\right] = \\
&\frac{\left(\pi_k^{\mathrm{o}}\right)^2}{R^4}\left(\sum_{t=1}^{T} R_t^4\mathrm{E}\left(\eta_{t,i}^2\eta_{t,j}^2\right) + \sum_{t=1}^{T}\sum_{s=1}^{T} R_t^2R_s^2\Sigma_{k,ii}\Sigma_{k,jj} + 2\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} R_{t_1}^2R_{t_2}^2\Sigma_{k,ij}^2\right) = \\
&\left(\frac{R}{T}\right)^2\frac{1}{R^4}\left(\sum_{t=1}^{T} R_t\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \sum_{t=1}^{T}\sum_{s=1}^{T} R_tR_s\Sigma_{k,ii}\Sigma_{k,jj} + 2\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} R_{t_1}R_{t_2}\Sigma_{k,ij}^2\right) = \\
&\frac{1}{R^2T^2}\left(R\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + R^2\Sigma_{k,ii}\Sigma_{k,jj} + 2R^2\Sigma_{k,ij}^2\right) = \frac{1}{RT^2}\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \frac{1}{T^2}\Sigma_{k,ii}\Sigma_{k,jj} + \frac{2}{T^2}\Sigma_{k,ij}^2 \leq \\
&\frac{1}{T^2}\left(\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \Sigma_{k,ii}\Sigma_{k,jj} + 2\Sigma_{k,ij}^2\right), \quad \text{(C.76)}
\end{aligned}$$

as $\pi_k^{\mathrm{o}} > 0$ if and only if $R > 0 \Leftrightarrow R \geq 1$, which implies that $1/RT^2 \leq 1/T^2$.

Therefore, by Equations (C.72) and (C.76), we get

$$\begin{aligned}
\mathrm{E}\left[\left(\Sigma_{k,ij}^{\mathrm{o}*} - \Sigma_{k,ij}^{\mathrm{o}}\right)^2\right] &\leq \mathrm{E}\left[\frac{1}{T^2}\left(\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \Sigma_{k,ii}\Sigma_{k,jj} + 2\Sigma_{k,ij}^2\right)\right] \\
&\leq \frac{1}{T^2}\left(\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \Sigma_{k,ii}\Sigma_{k,jj} + 2\Sigma_{k,ij}^2\right). \quad \text{(C.77)}
\end{aligned}$$

By the definition of the scaled Frobenius norm and Proposition A.9, we get

$$||\Sigma_k^{\mathrm{o}*} - \Sigma_k^{\mathrm{o}}||_{\mathrm{F}}^2 = \frac{1}{p}\mathrm{tr}\left[\left(\Sigma_k^{\mathrm{o}*} - \Sigma_k^{\mathrm{o}}\right)\left(\Sigma_k^{\mathrm{o}*} - \Sigma_k^{\mathrm{o}}\right)'\right] = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\left(\Sigma_k^{\mathrm{o}*} - \Sigma_k^{\mathrm{o}}\right)_{ij}\right]^2 = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\Sigma_{k,ij}^{\mathrm{o}*} - \Sigma_{k,ij}^{\mathrm{o}}\right)^2.$$

Using the linearity of the expectation and Inequality (C.77), we get

$$\begin{aligned}
\mathrm{E}\left(||\Sigma_k^{o*} - \Sigma_k^{o}||_F^2\right) &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[\left(\Sigma_{k,ij}^{o*} - \Sigma_{k,ij}^{o}\right)^2\right] \\
&\leq \frac{1}{pT^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \Sigma_{k,ii}\Sigma_{k,jj} + 2\Sigma_{k,ij}^2\right] \\
&= \frac{1}{pT^2}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\Sigma_{k,ii}\Sigma_{k,jj} + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\Sigma_{k,ij}^2\right).
\end{aligned}$$

By Equation (C.58) in the proof of Lemma 3.9 in Subsection C.2.3 we have

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left(\eta_{1,i}^2\eta_{1,j}^2\right) = \mathrm{E}\left(\sum_{i=1}^{p}\eta_{1,i}^2\sum_{j=1}^{p}\eta_{1,j}^2\right) = \mathrm{E}\left(\eta_t{}'\eta_t\eta_t{}'\eta_t\right) = \mathrm{E}\left(||\eta_t||_2^2\,||\eta_t||_2^2\right) = \mathrm{E}\left(||\eta_t||_2^4\right) = \mathrm{O}(p^2)\,.$$

By the definition of the trace we have

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\Sigma_{k,ii}\Sigma_{k,jj} = \sum_{i=1}^{p}\Sigma_{k,ii}\sum_{j=1}^{p}\Sigma_{k,jj} = \mathrm{tr}(\Sigma_k)\,\mathrm{tr}(\Sigma_k) = \mathrm{O}(p^2)\,,$$

since by Assumption (B1), $\mathrm{tr}(\Sigma_k) = \mathrm{O}(p)$.

By Proposition A.9, the proof of Proposition A.11 and the fact that Σ_k is symmetric, we get

$$2\sum_{i=1}^{p}\sum_{j=1}^{p}\Sigma_{k,ij}^2 = 2\mathrm{tr}\left(\Sigma_k\left(\Sigma_k\right)'\right) = 2\mathrm{tr}\left(\Sigma_k^2\right) \leq 2\left[\mathrm{tr}(\Sigma_k)\right]^2 = 2\,\mathrm{O}(p^2) = \mathrm{O}(p^2)\,.$$

Finally, we have

$$\mathrm{E}\left(||\Sigma_k^{o*} - \Sigma_k^{o}||_F^2\right) \leq \frac{1}{pT^2}\left[\mathrm{O}(p^2) + \mathrm{O}(p^2) + \mathrm{O}(p^2)\right] = \frac{1}{pT^2}\,\mathrm{O}(p^2) = \mathrm{O}\left(\frac{p}{T^2}\right),$$

and the proof of Statement (i) is complete.

(ii) Since by the construction of model (2.45)

$$Y_t = \mu_k + \Sigma_k\epsilon_t,$$

if and only if $R_t = \mathbf{1}_{\{X_t=k\}} = 1$, we have that

$$R_t\left(Y_t - \mu_k\right)\left(Y_t - \mu_k\right)' = R_t\left(\Sigma_k\epsilon_t\right)\left(\Sigma_k\epsilon_t\right)' = R_t\eta_t\eta_t{}',$$

Then, Equation (C.69) is written as

$$\Sigma_k^{o*} = \frac{1}{T}\sum_{t=1}^{T} R_t \eta_t \eta_t{}'.$$

Using the linearity of the expectation and Equations (C.28) and (C.74), we get

$$\mathrm{E}\left(\Sigma_k^{o*} \mid \boldsymbol{X}_1^T\right) = \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T} R_t \eta_t \eta_t{}' \,\middle|\, \boldsymbol{X}_1^T\right) = \frac{1}{T}\sum_{t=1}^{T} R_t \mathrm{E}\left(\eta_t \eta_t{}'\right) = \frac{1}{T}\sum_{t=1}^{T} R_t \Sigma_k = \pi_k^{o} \Sigma_k.$$

By the law of total expectation and Equation (C.4), we get

$$\begin{aligned}
\mathrm{E}(\Sigma_k^{o*}) &= \mathrm{E}\left[\mathrm{E}\left(\Sigma_k^{o*} \mid \boldsymbol{X}_1^T\right)\right] = \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T} R_t \Sigma_k\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(R_t)\Sigma_k = \frac{1}{T}\sum_{t=1}^{T} \mathrm{P}(R_t = 1)\Sigma_k \\
&= \frac{1}{T}\sum_{t=1}^{T} \pi_k \Sigma_k = \pi_k \Sigma_k.
\end{aligned} \tag{C.78}$$

A direct consequence of (C.78) is

$$\begin{aligned}
\mathrm{E}\left(||\Sigma_k^{o*} - \pi_k\Sigma_k||_{\mathrm{F}}^2\right) &= \mathrm{E}\left(||\Sigma_k^{o*} - \mathrm{E}(\Sigma_k^{o*})||_{\mathrm{F}}^2\right) \\
&= \mathrm{E}\left[\frac{1}{p}\mathrm{tr}\left(\left[\Sigma_k^{o*} - \mathrm{E}(\Sigma_k^{o*})\right]\left[\Sigma_k^{o*} - \mathrm{E}(\Sigma_k^{o*})\right]'\right)\right] \\
&= \frac{1}{p}\mathrm{E}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\left[\Sigma_k^{o*} - \mathrm{E}(\Sigma_k^{o*})\right]_{ij}\right)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[\left(\Sigma_{k,ij}^{o*} - \left[\mathrm{E}(\Sigma_k^{o*})\right]_{ij}\right)^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}\left(\Sigma_{k,ij}^{o*}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{X_t=k\}}\left(Y_{t,i} - \mu_{k,i}\right)\left(Y_{t,j} - \mu_{k,j}\right)\right].
\end{aligned} \tag{C.79}$$

We have that

$$||\Sigma_k^{o} - \pi_k\Sigma_k||_{\mathrm{F}}^2 = ||\Sigma_k^{o} - \Sigma_k^{o*} + \Sigma_k^{o*} - \pi_k\Sigma_k||_{\mathrm{F}}^2 =$$
$$||\Sigma_k^{o} - \Sigma_k^{o*}||_{\mathrm{F}}^2 + ||\Sigma_k^{o*} - \pi_k\Sigma_k||_{\mathrm{F}}^2 - 2\left\langle \Sigma_k^{o} - \Sigma_k^{o*}\,,\, \Sigma_k^{o*} - \pi_k\Sigma_k \right\rangle.$$

Therefore,

$$\mathrm{E}\Big(||\Sigma_k^{\mathrm{o}} - \pi_k\Sigma_k||_{\mathrm{F}}^2\Big) = \mathrm{E}\Big(||\Sigma_k^{\mathrm{o}} - \Sigma_k^{\mathrm{o}*} + \Sigma_k^{\mathrm{o}*} - \pi_k\Sigma_k||_{\mathrm{F}}^2\Big) =$$
$$\mathrm{E}\Big(||\Sigma_k^{\mathrm{o}} - \Sigma_k^{\mathrm{o}*}||_{\mathrm{F}}^2 + ||\Sigma_k^{\mathrm{o}*} - \pi_k\Sigma_k||_{\mathrm{F}}^2 - 2\left\langle \Sigma_k^{\mathrm{o}} - \Sigma_k^{\mathrm{o}*}\,,\, \Sigma_k^{\mathrm{o}*} - \pi_k\Sigma_k\right\rangle\Big) =$$
$$\mathrm{E}\Big(||\Sigma_k^{\mathrm{o}} - \Sigma_k^{\mathrm{o}*}||_{\mathrm{F}}^2\Big) + \mathrm{E}\Big(||\Sigma_k^{\mathrm{o}*} - \pi_k\Sigma_k||_{\mathrm{F}}^2\Big) - 2\mathrm{E}[\left\langle \Sigma_k^{\mathrm{o}} - \Sigma_k^{\mathrm{o}*}\,,\, \Sigma_k^{\mathrm{o}*} - \pi_k\Sigma_k\right\rangle]$$

□

C.2.5 Proof of Theorem 3.3

Proof. As was mentioned earlier in Section 3.2, Sancetta (2008) defines a shrinkage estimator of the form:

$$\Sigma^s = (1-\alpha)\,\boldsymbol{S} + \alpha\boldsymbol{F},$$

where $\alpha \in [0,1]$, $\boldsymbol{S}$ is the sample covariance matrix and $\boldsymbol{F}$ is the shrinkage target. In our case, $F = \nu_k\boldsymbol{I}_p$ and $\hat{F}_T = \nu_k^{\mathrm{o}}\boldsymbol{I}_p$. We check that Conditions 1 to 4 of Sancetta (2008) are satisfied.

Condition 1 (1): We have that

$$\mathrm{E}\Big(\left|\hat{F}_{T,ij} - F_{ij}\right|^2\Big) = \mathrm{E}\Big(\left|\left(\nu_k^{\mathrm{o}}\boldsymbol{I}_p\right)_{ij} - \left(\nu_k\boldsymbol{I}_p\right)_{ij}\right|^2\Big) = \mathrm{E}\Big(\left|\left(\nu_k^{\mathrm{o}} - \nu_k\right)\left(\boldsymbol{I}_p\right)_{ij}\right|^2\Big) =$$
$$\mathrm{E}\Big[\left(\nu_k^{\mathrm{o}} - \nu_k\right)^2\left(\boldsymbol{I}_p\right)_{ij}^2\Big] = \left(\boldsymbol{I}_p\right)_{ij}^2\mathrm{E}\Big[\left(\nu_k - \nu_k^{\mathrm{o}}\right)^2\Big] = \mathrm{O}\Big(T^{-1}\Big),$$

as

$$\left(\boldsymbol{I}_p\right)_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases},$$

and

$$\mathrm{E}\Big[\left(\nu_k - \nu_k^{\mathrm{o}}\right)^2\Big] = \mathrm{O}\Big(T^{-1}\Big),$$

by Statement (ii) of Lemma 3.9. Therefore, Condition 1 (1) is satisfied.

Condition 1 (2): Since $F = \nu_k\boldsymbol{I}_p$ and $\hat{F}_T = \nu_k^{\mathrm{o}}\boldsymbol{I}_p$, are diagonal matrices, for $\beta = 1 \in [0,2)$, we have that

$$\#\left\{1 \leq i,j \leq p : F_{ij} \neq \hat{F}_{T,ij}\right\} \leq p^{\beta},$$

where # is used to denote the cardinality of a countable set.

Condition 1 (3): From Assumption (C1), $\nu_k \boldsymbol{I}_p \neq \pi_k \Sigma_k$, therefore there exists some $\gamma > 0$, such that

$$\left|\left|\nu_k \boldsymbol{I}_p - \pi_k \Sigma_k\right|\right|_{\mathrm{F}}^2 \asymp p^{\gamma},$$

which means that Condition 1 (3) is satisfied.

Condition 1 (4): From Assumption (C2), $p \to +\infty$ such that $\dfrac{p^{2-\gamma}}{T} \to 0$, as $T \to +\infty$, for some $\gamma \in (0, 2]$, hence $p^{1-\gamma/2} = \mathrm{o}\left(\sqrt{T}\right)$. Since in our case $\beta = 1$, for any $\gamma \in (0, 2]$, we have that

$$\max\left\{p^{\beta-\gamma}, p^{1-\gamma/2}\right\} = p^{1-\gamma/2} = \mathrm{o}\left(\sqrt{T}\right),$$

therefore Condition 1 (4) is satisfied.

Condition 2: Condition 2 derives directly from Statement (ii) of Lemma C.2.

Condition 3: Condition 3 (1) is a direct result of the assumptions on the kernel $K(u)$, which are slightly stronger, than those of Sancetta (2008). Condition 3 (2) for the bandwidth, is exactly the same as the one given in the statement of Theorem 3.3.

Condition 4: From Assumptions (A1) to (A3), the process $\{Y_t : t \in \mathbb{Z}\}$ is stationary with finite eighth moments (see the discussion following Assumptions (A1) to (A3) in Subsection 2.3.1). The moment condition of Condition 4 follows from $\mathrm{E}\left(||\epsilon_t||_2^8\right) < +\infty$ and the boundedness of $\{X_t : t \in \mathbb{Z}\}$ by a similar argument as in proving Statement (ii) of Lemma C.2.

□

Appendix D

The Viterbi Algorithm

The Viterbi algorithm is a technique, based on dynamic programming methods, to estimate the single best state sequence $\{X_1, \ldots, X_T\}$ that maximises the conditional probability

$$\mathrm{P}_{\boldsymbol{\theta}}(X_1, \ldots, X_T \,|\, y_1, \ldots, y_T)\,,$$

which is equivalent to maximising the joint probability (see Section III.B in Levinson et al. 1983)

$$\mathrm{P}_{\boldsymbol{\theta}}(X_1, \ldots, X_T, y_1, \ldots, y_T)\,.$$

The parameter $\boldsymbol{\theta}$ is assumed to be known and in practice it is substituted by the estimate of the EM algorithm. We follow Appendix A in Fiecas et al. 2017.

For all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$, let us define

$$\delta_k(t) = \max_{k_1, \ldots, k_{t-1}} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_t = k, y_1, \ldots, y_t)]\}.$$

$\delta_k(t)$ is the highest log-probability along a single path, at time t, which accounts for the first t observations and ends in state k.

By induction, for all $t \in \{1, \ldots, T-1\}$ and $k, l \in \{1, \ldots, K\}$, we get

$$\begin{aligned}
\delta_l(t+1) &= \max_{k_1,\ldots,k_t} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_t = k_t, X_{t+1} = l, y_1, \ldots, y_t, y_{t+1})]\} \\
&= \max_{k_1,\ldots,k_t} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_t = k_t, y_1, \ldots, y_t)\, \mathrm{P}(X_{t+1} = l \mid X_t = k_t)\, f_l(y_{t+1}; \theta)]\} \\
&= \max_{k_t} \left\{ \max_{k_1,\ldots,k_{t-1}} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_t = k_t, y_1, \ldots, y_t)\, \mathrm{P}(X_{t+1} = l \mid X_t = k_t)\, f_l(y_{t+1}; \theta)]\} \right\} \\
&= \max_{k} \left\{ \max_{k_1,\ldots,k_{t-1}} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_t = k, y_1, \ldots, y_t)\, \mathrm{P}(X_{t+1} = l \mid X_t = k)\, f_l(y_{t+1}; \theta)]\} \right\} \\
&= \max_{k} \left\{ \max_{k_1,\ldots,k_{t-1}} \{\log[\mathrm{P}_{\boldsymbol{\theta}}(X_1 = k_1, \ldots, X_t = k, y_1, \ldots, y_t)] + \log[\mathrm{P}(X_{t+1} = l \mid X_t = k)]\} \right\} \\
&\quad + \log[f_l(y_{t+1}; \theta)] \\
&= \max_{k} \{\delta_k(t) + \log(p_{kl})\} + \log[f_l(y_{t+1}; \theta)]\,.
\end{aligned}$$

To determine the single best state sequence, it is necessary to keep track of the maximisers of the terms

$$\delta_k(t) + \log(p_{kl})$$

for all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$. Let

$$\psi_k(t) = \underset{i \in \{1,\ldots,K\}}{\arg\max} \{\delta_i(t-1) + \log(p_{ik})\} \qquad (t \in \{1, \ldots, T\},\ k \in \{1, \ldots, K\})\,.$$

The method is illustrated below.

Viterbi Algorithm

---

(1) Initialisation:

For all $k \in \{1, \ldots, K\}$

$$\begin{aligned}
\delta_k(1) &= \log(\pi_k) \log[f_k(y_1; \boldsymbol{\theta})]\,, \\
\psi_k(1) &= 0.
\end{aligned}$$

(2) Recursion:

For all $t \in \{2, \ldots, T\}$ and $k \in \{1, \ldots, K\}$

$$\begin{aligned}
\delta_k(t) &= \max_{i \in \{1,\ldots,K\}} \{\delta_i(t-1) + \log(p_{ik})\} + \log[f_k(y_t; \boldsymbol{\theta})]\,, \\
\psi_k(t) &= \underset{i \in \{1,\ldots,K\}}{\arg\max} \{\delta_i(t-1) + \log(p_{ik})\}.
\end{aligned}$$

(3) Termination:

$$x_T^* = \underset{k \in \{1,\ldots,K\}}{\arg\max} \{\delta_k(T)\}.$$

(4) Path backtracking:

For $t \in \{T-1, T-2, \ldots, 1\}$

$$x_t^* = \psi_{x_{t+1}^*}(t+1)\,.$$
