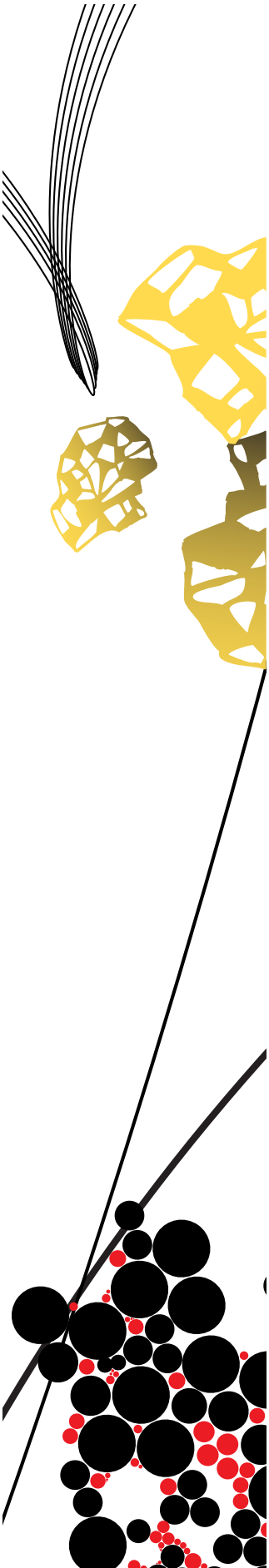

University of Twente

MSc Thesis Applied Mathematics

Obtaining bounds on the number of hidden Markov states in a tracking context

January 21, 2020

Author:
Steef Severijn

Daily Supervisors:
Dr. ir. Jasper Goseling (UT)
Dr. Pranab Mandal (UT)
Dr. ir. Martin Podt (Thales)
Dr. ir. Rienk Bakker (Thales)

Graduation Committee:
Prof. dr. Richard Boucherie (UT)
Dr. ir. Jasper Goseling (UT)
Dr. Pranab Mandal (UT)
Dr. ir. Martin Podt (Thales)
Dr. Katharina Proksch (UT)

THALES

UNIVERSITY OF TWENTE.

Abstract

For the classi cation of moving objects based on their trajectories, one wants to be able to build tracking models which are as accurate as possible. For this purpose, double hidden Markov models are used. The goal of this research is to nd a method or a combination of methods to determine the optimal number of hidden Markov states for a double hidden Markov model. The method most looked into in this thesis, called the autocovariance method in this report, relates the number of Markov states of the considered model to the orders of a vector autoregressive moving average model with the same autocovariance structure as the considered model. This method provides a theoretical lower bound on the number of hidden Markov states, although this bound is not guaranteed in practice due to the method of determining the autoregressive and moving average orders in this research.

In this report, the existing lower bounds are generalised for a wider class of models and for di erenced time series, useful in the case of cointegrated time series. The bounds of the existing literature and introduced in this report become weak if no assumptions are made on the correlation of the time series with the Markov Process. More promising bounds are available if there is no correlation.

From synthetic experiments, it appeared that these bounds perform best for processes where the hidden Markov chain stays, on average, in the same hidden Markov state for longer periods of time, for higher dimensional problems, for lower orders of the number of hidden Markov states, for a higher number of observations, for problems where the Markov states are su ciently distinctive and when the process does not contain Markov states in which the process is (highly) nonstationary. For di erenced time series, the number of di erent Markov states e ectively is the original number of Markov states squared, which can often not all be distinguished. Therefore, this method is in these cases often not useful in practice, perhaps except in the case when one wants to nd out if one needs multiple hidden Markov states at all. It is preferable to use the non-di erenced method if possible.

The autocovariance method is computationally inexpensive and therefore suitable to be incorporated in the model building proces, although it is advisible to use the estimate of the method in combination with other information, such as estimates from information criteria which can be used as upper bounds.

Preface

This thesis is the result of my last and biggest project of my student period. During the last 7 months, I have worked hard on this nal hurdle towards graduation and it certainly has been the most challenging of my studies. Since I did my graduation project, and also my internship, at Thales, I needed to get used to a standard working rhythm for the rst time and make longer days than I was used to. Luckily, I slowly started to get used to this rhythm, which will be proof useful in the upcoming phase of my life. In the end, I am happy how this project turned out.

I want to thank a few people for helping me during the graduation project. Firstly, I would like to thank my daily supervisors both at the UT and Thales, Jasper Goseling, Pranab Mandal, Martin Podt and Rienk Bakker, for guiding me in the right direction and ploughing through the though mathematical context, which was not always in the direction of their specialization. Moreover, I would like to thank my boyfriend Sander, who endured my complaints about my project if things did not work out, or if the end of the project seemed eternities away. Lastly, I would like to thank everyone who provided the necessary and enjoyable distractions from my thesis: of course my family, where I could always rest in the weekend, the people of the athletics and kayaking association and my co-workers at Thales for the lunch breaks and the board game evenings.

Contents

Chapter 1

Introduction

1.1 Problem description

Monitoring moving objects is one of the major tasks of the armed forces. For many military and defence applications, one wants to classify these di erent objects, based on the type or behaviour of the object. For example, the navy may want to know if a certain ship is a potential smuggler, airport security if an observed object is a bird or a drone and the air forces whether an incoming airplane is going to attack or will not be a threat. The estimated motion trajectories of objects, based on observations of (amongst others) radar systems, can be used as input for these classi cations. The motion patterns of one class may exhibit certain characteristics that are uncommon for other classes, while the state variables at any single point in time may not be enough to classify reliably. For example, birds may y at similar altitudes and speeds as man-made objects, however, when making local ights they may exhibit more erratic ight behaviour than man-made objects (Moon, 2002).

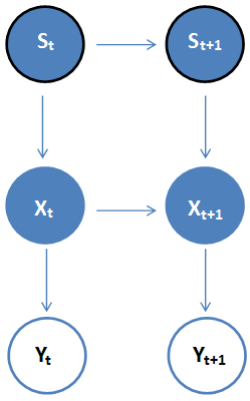

Figure 1: Structure of the Bayesian Network used for tracking.

To be able to model the trajectories of the observed objects, a certain type of Bayesian Networks are used, which can be seen as an extension of Hidden Markov Models (HMMs). The structure of the model is shown in Figure 1. The circles denote variables or vectors of variables and the arrows dependence relations. The processes in blue fS_tg and fX_tg are

unobserved, while fY_tg is observed. In the context of this thesis, Y_t are the radar observations and possibly other types of observations at time period t. These observations are a noisy representation of the real kinematic state of the object, X_t. X_t is a vector including the 2- or 3-dimensional location coordinates and the velocity in the $x;y$ and possibly z-direction. X_t changes each time step in a stochastic manner according to a speci c model k, where the value k is the outcome of the hidden discrete univariate variable S_t, which is called the mode of the system. Each mode k thys implies a speci c model k. The process fS_tg is a Markov process with domain S. We will refer to the model of Figure 1 as a Double Hidden Markov Model (DHMM).

To summarize, the DHMM $fS_t; X_t; Y_tg$ can be expressed as follows:

$$P(S_t = ijS_{t\ 1} = j) = P_{ji}; \tag{1}$$

$$X_t = f_{S_t}(X_{t\ 1}; V_t); \tag{2}$$

$$Y_t = g(X_t; W_t): \tag{3}$$

In the above equations, $P = (P_{ji})$ is the probability transition matrix of S_t, f_{S_t} are the transition functions, which can together with g have any form in the most general case, and fV_tg and fW_tg are random in uences, or more accurately white noise processes, for which a de nition will be provided later in this report (De nition 1 in Section 2.2).

Current practice is that the number of modes jSj, is determined manually. The functional form of the transition function f is also determined manually to some extent, as some parameters are estimated. A small arti cial example will be introduced to illustrate the di culties with this approach.

Suppose we have di erent types of 2-dimensional trajectories we would like to distinguish and we decide to use three di erent transition functions f_i and thus three di erent Markov states. One model corresponding to a movement forward, another function corresponding to a rightward turn and a last one corresponding to a turn to the left. Although the hidden Markov states S_t remain unknown, one can in this way fairly accurately guess in which hidden Markov state one has been, based on the observations Y_t, if the noise level is not too great. A large change in the heading angle, which is the angle of the direction of the object with the north direction, now assumed to be the y-axis, is likely to correspond to a turn model. Likewise, no change or a very tiny change in the heading angle is likely to correspond to the transition function for the movement forward. Some change in the heading angle, say , must be the threshold where the most likely model is not the movement forward anymore, but the turn model to the right or left, depending on the sign of .

Now suppose one of the trajectories we would like to classify is an 8-pattern as shown in Figure 2. This trajectory alternates between a sequence of left turns until a full circle has been made and a sequence of an equal number of right turns the same number of right turns. However, if the change in heading angle per time step is halved, while keeping the sequence of right and left turns constant, one gets a zigzag pattern as in Figure 3. If we desire to distinguish the 8-patterns and the zigzag-patterns, we are not be able to do so if the value of is too small and thus the expected sequence of the hidden Markov states as well as the

estimated probability transition matrix P would be identical. This is visualised in Figure 4 and 5, where the change in heading angle as well as the threshold values are plotted over time. For both patterns the change in heading angle are at the same time above, below or between the two thresholds and .

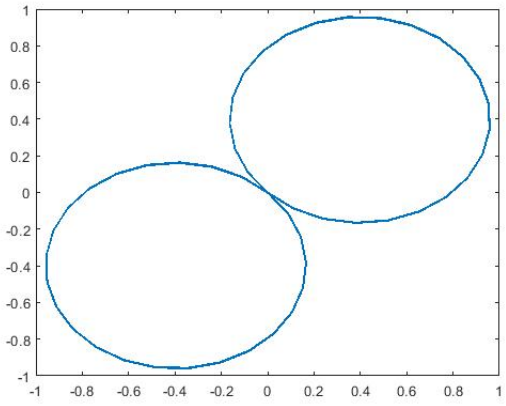

Figure 2: 8-shaped trajectory

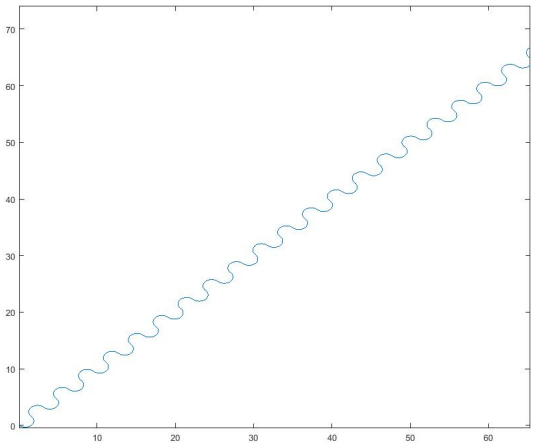

Figure 3: Zigzag-shaped trajectory

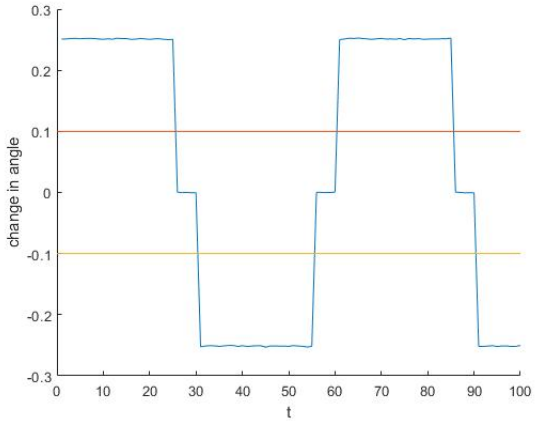

Figure 4: Change in heading angles of an 8-shaped trajectory over time and low threshold values ; .

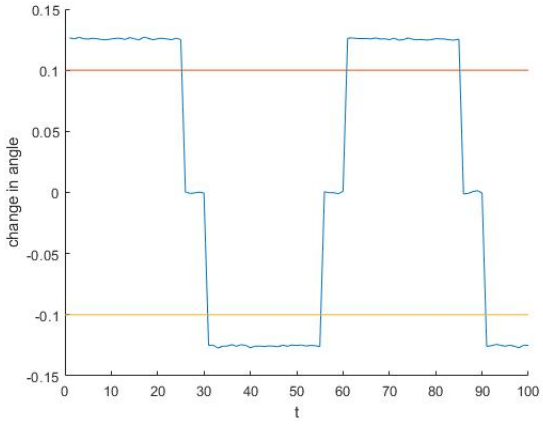

Figure 5: Change in heading angles of a zigzag-shaped trajectory over time and low threshold values ; .

If on the other hand the value of is too large, one cannot distinguish between the straight and turn models in the zigzag pattern, as shown in Figure 6 and 7. In Figure 6 we see that the change in heading angles exceeds the thresholds at certain points in time for the 8-shaped trajectory, while the change in heading angles for the zigzag-shaped trajectory never exceeds the thresholds, shown in Figure 7. We would the wrongly conclude the zig-zag pattern always stays in the same mode. Moreover, if we have trajectories of mixed zigzag

and 8-shaped patterns, three submodels cannot adequately describe the trajectory well and we would not be able to distinguish that trajectory either from the pure 8-shaped trajectory or the zigzag trajectory, depending on the value of .

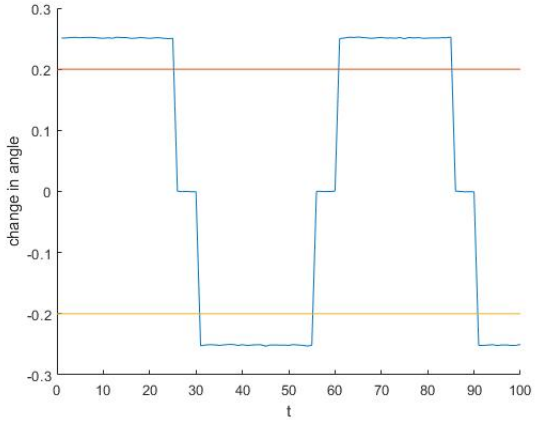

Figure 6: Change in heading angles of an 8-shaped trajectory over time and two high threshold values

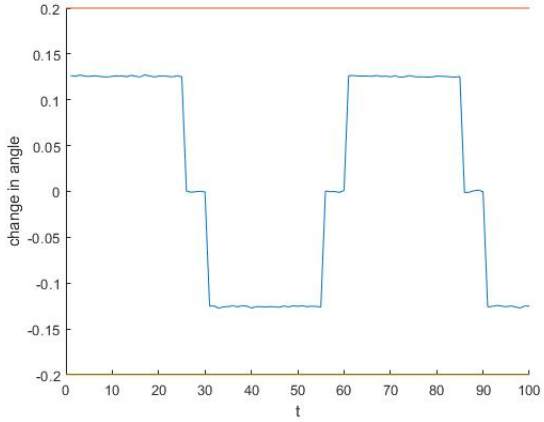

Figure 7: Change in heading angles of a zigzag-shaped trajectory over time and two high threshold values

The above example is of course arti cial and its solution is straightforward: add two extra modes corresponding to sharp turns right and left. However, for real situations it is not straightforward how many modes and which transition functions should be chosen, since the transitions may not only include change in direction, but also change in velocity, acceleration and elevation, which may make manual setting of parameters prohibitively di cult.

Ideally, the number of modes, the transition functions and all parameters should be determined simultaneously in an optimized manner without any manual in uence and this thesis aims to come closer to this ideal procedure compared to the current practice. This thesis builds further on the work of Richa (2018), who formulated a tracking and classi cation procedure. It is outside the scope of this thesis to explain the existing modelling procedure in detail, since many of its steps remain unaltered, but a bird's eye view of the procedure is given next.

Figure 8 gives a schematic overview of the steps that are taken sequentially to be able to classify an object. These steps can be decomposed in three phases, the rst being the model building part. Here, the number of submodels, or hidden Markov states, is determined. Moreover, the speci c form of the transition functions f and g of Equations (2) and (3) are determined, although parameters may still be undetermined. Others, most notably the parameters de ning the cuto between the di erent hidden Markov states, the aforementioned change in heading angle threshold , are set in advance. The model building part of the procedure is currently done in a non-systematic manner, using expert knowledge and inspection of the data. It is this phase of the procedure which will be addressed in this thesis.

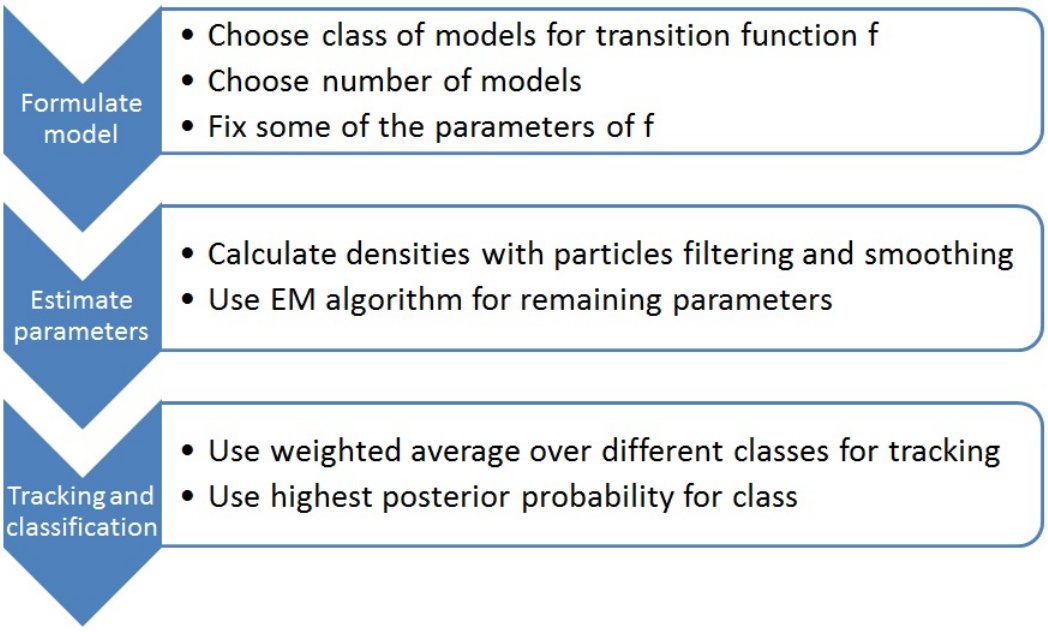

Figure 8: Existing model building procedure.

In the second phase, the non- xed parameters of the model are estimated using the maximum likelihood framework. For this the probability density functions of the observed variables Y_t are needed. Since these are too complex to calculate exactly, they are estimated using particle ltering and smoothing. The maximization is also too di cult to do analytically and therefore the Expectation-Maximization (EM) algorithm is used instead.

In the last phase, the freshly built model is applied for tracking. In short, the Bayesian statistics framework is used to obtain an estimate for the hidden continuous states X_t given the observations Y_t. In case one also needs to classify, all previous phases and steps are done for the di erent classes and one thus has built and estimated a di erent model (1) - (3) for each class i. One applies each model i separately to obtain the state estimates $\hat{X}_t^{(i)}$. The Bayesian classi er is used to choose the most probable class, by for each class i calculating the posterior probability of the observation sequence given that the class of the process is i. The nal estimate of the hidden continuous state X_t is the average of the estimates for the di erent models, weighted with the posterior probability of the classes, $\hat{X}_t = \sum_i w_i \hat{X}_t^{(i)}$.

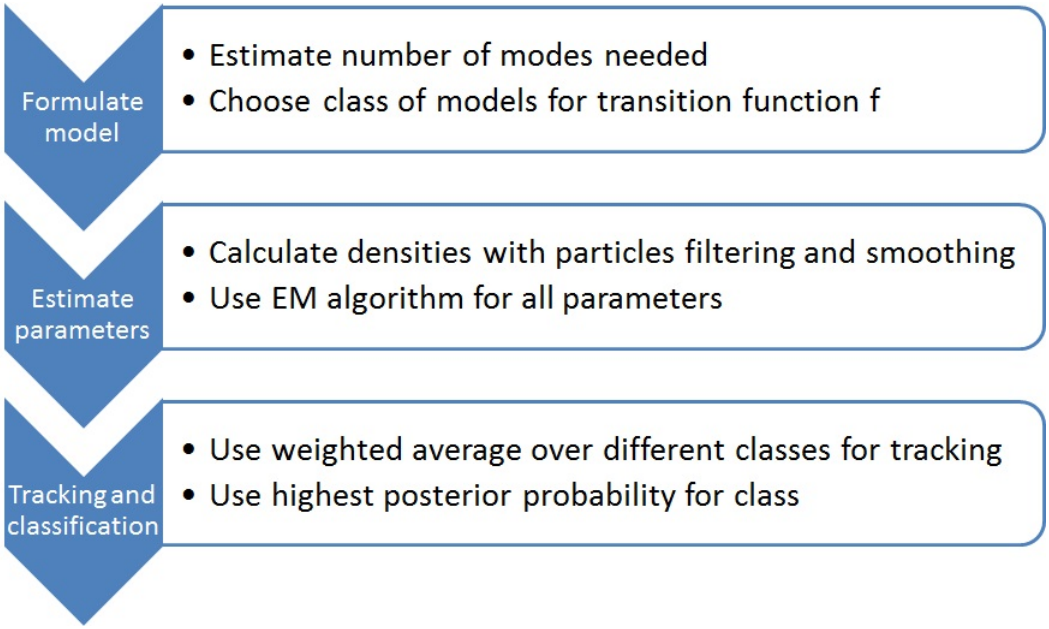

Figure 9: Proposed new model building procedure.

In the procedure proposed in this thesis, shown in crude form in Figure 9, the non-systematic part of the procedure is reduced. All the parameters are estimated in the new procedure; either by EM, or for the number of hidden Markov states beforehand using a time series procedure. Maximum likelihood estimation is not suitable to estimate the number of hidden Markov states, since the likelihood is increasing in the number of hidden Markov states. The class of models is still picked manually. Phase 2 and 3 of the existing procedure remain virtually unchanged, although it is proposed that the a wider family of the transition functions f is chosen, so that more parameters can be estimated with the EM algorithm.

1.2 Research questions

As sketched in the introduction, the ultimate purpose of this research is to enhance the procedure for tracking and classi cation in a mathematically sound manner. This thesis will focus on the model building aspect, especially the number of modes needed and we can thus summarize the main research objective as:

Construct a procedure to determine the optimal number of hidden Markov States jSj for the model of the form

$$P(S_t = ijS_{t\ 1} = j) = P_{ji}; \tag{4}$$

$$X_t = f_{S_t}(X_{t\ 1}; V_t); \tag{5}$$

$$Y_t = g(X_t; W_t): \tag{6}$$

This problem will be tackled by answering the following subquestions:

1. How can the optimal number of hidden Markov States be determined based on the data?
2. How can the above methods be included in the existing overall procedure for tracking and classi cation?

Since the model formulation in the main research question is very general, attention is restricted to a speci c class of models, where f_{S_t} is linear, which allows for more analysis.

1.3 Thesis structure

Several procedures to determine the optimal number of hidden Markov states already exist. In Chapter 2 these procedures are introduced. An overview of the various existing approaches is given in Section 2.1 and necessary background information for the two procedures used in the rest of the thesis is discussed in Section 2.2.

To answer the first research question, two approaches will be taken. In the first method, the autocovariance structure of the model (4) - (6) is related to that of a simpler model. Using this relation a lower bound of the number of hidden Markov states K can be obtained in terms of parameters of the simpler model. This method, which will be called the autocovariance method, will be discussed in Chapter 3. In Section 3.1, the bounds of K in terms of the parameters of the simpler model are obtained. In Section 3.2, methods to obtain the parameters of the simpler model are discussed. In Section 3.3, some experiments are performed to establish the performance of the autocovariance method under various circumstances.

The second approach uses penalty functions to find the number of hidden Markov states. More Markov states will lead to more flexibility and better fits of the data, but also to a greater risk of overfitting the data. Penalty functions are criterions to be maximized, such as the likelihood, while giving a penalty for the complexity of the model, such as the number of parameters. The disadvantage of these methods is that it is often difficult to find the right penalty so that it is severe enough, but not too severe. The penalty function approach will be discussed in Chapter 4. In Section 4.1, penalty functions are introduced. In Section 4.2, likelihood optimization is discussed, which is for the penalty functions considered in this thesis the optimization needed to use the penalty function approach. In Section 4.3, the EM algorithm is discussed, which is the likelihood optimization method used in this thesis.

In Chapter 5, both approaches will be applied to the tracking context. In Section 5.1, the kinematic model is introduced. In Section 5.2, the assumptions of the two method approaches are recapped. In Sections 5.3 to 5.5, some of these assumptions are discussed in detail for the specific kinematic model used. In Section 5.6, some experiments using the kinematic model are discussed. Lastly, in Chapter 6 the research questions will be answered and some recommendations are made for those wishing to put the results of this thesis to use.

This thesis contains several scientific contributions. A number of general theoretical results have been established, mostly derived in Section 3.1. If in this thesis a proof is provided for a theoretical result, this result is new. For existing results, references are made to the literature where a proof of the theorems can be found. Moreover, in this thesis the autocovariance method is to the best knowledge for the first time applied for multivariate time series in this thesis. Lastly, the autocovariance method has hitherto mostly been applied in economic contexts and in this thesis it will be applied to the tracking context for the first time.

To establish the necessary theory for especially the autocovariance method, this thesis contains some parts which are mathematically demanding, while for practical users it may not be necessary to know all theoretical details. Practice-oriented Readers may decide to skip Lemma 2 in Section 2.2.2, Theorem 3, Lemma 4 and 5, and Theorem 6 in 3.1 and all proofs.

Chapter 2

Background

In this chapter, the problem of choosing the number of hidden Markov states is put in a wider context. Moreover, the existing solution directions of the problem are discussed, both in Section 2.1. From this discussion, it becomes apparent that it is bene cial to restrict attention of the general model formulation (4) - (6) to so-called Markov Switching VARMA models. These models are discussed in Section 2.2.

2.1 Literature Review

Hidden Markov Models are a solution to the problem of time series segmentation. Various scienti c disciplines work with time series models which are not able to describe the entire time series, but can describe part of the time series. Therefore, it is desirable to segment the time series into blocks which are homogeneous within the blocks and heterogeneous between blocks, although it is not obvious how to approach this.

One intuitive approach is iteratively checking where a time series can be split. This approach is often pursued in the eld of data mining, where time series are usually segmented to be able to represent the data in a simpli ed form, for instance as a point of a line, to save on computer storage (Liu et al., 2008). Metaheuristics that are often used in this eld are the sliding window approach, where the current segment is allowed to grow in size until it reaches a certain error criterion, the top-down approach which partitions the series in a recursive manner until a certain error criterion is reached, and the bottom-up approach which merges di erent partitions until some error criterion is reached (Keogh et al., 2001). The list of approaches in other elds is extensive and varied and includes amongst others dynamic programming (Kehagias et al., 2006), Bayesian procedure (Lee and Heghinian, 1977), statistical tests (Chow, 1960), (Buishand, 1982) and branch-and-bounding (Hubert, 2000). See Basseville et al. (1993) for an extensive discussion on segmentation.

An alternative approach which is often taken in the tracking literature is the Multiple Models method, rst introduced in Magill (1965). In this method a set of models is used, which represents the various possible model patterns (modes). In the article of Magill, the model was assumed to be chosen out of a set of models, but xed over the time window. The outcomes of the models are in this situation combined by taking a weighted average based on the con-

ditional probability of the models given the data. In more recent versions of Multiple Model Estimation, the active model may switch during each time period. In these cases finding the optimal sequence of models needs to be considered, although this is for long time-lengths intractable, since the number of possible sequences grows exponentially. Several heuristics have been proposed, the first one in Ackerson and Fu (1970), often focusing on pruning or merging the various sequences. Later, the Multiple Models method was extended to the Variable Structure Multiple Models method, in which the set of models depends on the time period. See for instance Li et al. (2000).

The other approach often taken in the tracking literature are hidden Markov models, first introduced by Baum and Petrie (1966). Hidden Markov Models are used in various scientific disciplines, including biology, speech recognition, econometrics and finance. The disadvantage of this method compared to many of the other methods is that one needs to make an additional choice determining the number of hidden Markov states, which is often not straightforward to do, since the likelihood function is increasing in the number of hidden Markov states.

Beal et al. (2002) circumvent this problem by introducing an infinite HMM, based on a hierarchical Dirichlet process (Ghosal and Van der Vaart, 2017). A Dirichlet Process is a stochastic process which generates probability distributions as an outcome. In the hierarchical model a Dirichlet Process determines in which lower-level Dirichlet Process one is. This lower-level Dirichlet Process generates the transition probabilities of the hidden Markov states. Each hidden Markov state has so-called emission probabilities for the observed variables drawn from some distribution. The model is infinite because a transition happens to an already visited state with probability proportional to the number of visits to that state and to one of the infinitely many unvisited states with probability based on some parameter. Although there is an infinite number of hidden Markov states available, only a finite number will be visited. By using a suitable learning procedure, the infinite HMM can be used to estimate the suitable number of hidden Markov states in a normal HMM, by examining the number of visited hidden Markov states. Fox et al. (2010) extended this idea to linear dynamic models with continuous variables, including state-space models.

Several statistical tests for hypothesis testing on the number of hidden Markov states in Hidden Markov Models have been proposed. General tests as the likelihood ratio test and the Lagrange Multiplier test do not have their usual convergence behaviour, since the regularity conditions are unfulfilled, for example because of unidentified parameters in the transition matrix of the hidden Markov states under the null hypothesis (Lange and Rahbek, 2009). A solution proposed by Hansen (1992) is to look at the supremum of the likelihood ratio statistic over the possible parameters, which gives an upper bound of the statistic value. The same principle is also applied to the Langrange Multiplier test and the Wald Test (Altissimo and Corradi, 2002). The disadvantage of these procedures is that they only deliver upper bounds for the test statistics. An alternative approach would be to generate artificial time series by means of simulation and use those to determine critical values. The disadvantage of this approach is that it requires the estimation of many Hidden Markov Models, which has great computational cost (Franses et al., 2014). Another disadvantage depending on the

application may be the asymmetric handling of the null and alternative hypothesis due to the nature of hypothesis testing.

In contrast to these frequentist methods, Robert et al. (2000) take a Bayesian approach, where they do the intractable inferences by means of the reversible jump Markov Chain Monte Carlo method. A disadvantage of this method are the low acceptance probabilities in their Monte Carlo method they reported, slowing down the convergence behaviour.

Another solution direction, taken by MacKay (2002) is minimizing the distance between the empirical distribution function of the data and the distribution function of the Hidden Markov model, penalized by the log values of the stationary probabilities. Also other penalized objective functions have been considered, such as penalized likelihood functions by Ryden (1995), who proves that a certain class of these penalized functions asymptotically provide an upper bound on the number of hidden Markov states. This approach is often chosen in practice due to its conceptual simplicity, with selection based on the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) being the most used. The AIC, an approximation of the Kullback-Leibler (KL) divergence, often overestimates the correct number of hidden Markov states to be chosen and Smith et al. (2006) have proposed an alternative KL divergence-based criterion designed for hidden Markov regression models, the Markov Switching Criterion.

In the nancial literature, HMMs have been rst introduced by Hamilton (1989) as regime switching models, which are a generalization of Hidden Markov Models where the observable states depend not only on the hidden Markov state, but also on the observable states at previous timestamps in a linear manner. These models have been related to autoregressive moving average (ARMA) models by Zhang and Stine (2001). In their paper they derive an upper bound for the order of the autoregressive and moving average lags in the ARMA model as a function of the number of regimes. These can be used to obtain a lower bound on the number of regimes when the autoregressive and moving average orders are known. Since ARMA models are widely used in nance and econometrics, methods to obtain these orders are well investigated. Cavicchioli (2014) has obtained bounds for a related class of Markov switching time series models.

State-space models with Markov Switching can be seen as yet another generalization of HMMs and regime switching models, where the dependence between the current observable state and the previous observable states is not necessarily linear. Literature on the determination of the number of hidden discrete states in Markov Switching State Space Models (MS SSMs) is limited: apart from the aforementioned paper by Fox et al. (2010), who only considers linear MS SSMs, no literature has been found which addresses this issue for general MS SSMs.

2.2 Theory on time series

From the previous literature review, we can conclude that if we want to be able to use techniques of the existing literature, in most cases we need to restrict our attention to linear models. Since with certain penalty functions we can obtain an upper bound on the number of the number of hidden Markov states and with the method of Zhang and Stine (2001) we can obtain a lower bound, these methods are investigated in more detail. In the following section, the relevant background theory on time series will be discussed. First, in Section 2.2.1, a general introduction to Autoregressive Moving Average (ARMA) models will be given. Afterwards, in Section 2.2.2, time series with an underlying Markov process will be introduced. See for more elaborate background on multivariate time series for instance Lütkepohl (2005), which is the major source of the following subsection.

2.2.1 Background on time series

A time series is a chronological collection of M-dimensional vectors $\{y_t\}$ over some time period, where t may be continuous or discrete and the time period may be in nite. In this thesis it will be assumed that $t \in \mathbb{Z}$ and that the time series has been going on inde nitely. Moreover, it is assumed the time periods between two observations y_t and y_{t-1} are equidistant. This latter assumption may be relaxed, as long as later explicit assumptions on for example stationarity or distribution do hold. However, these are often violated if the time periods are not of equal length.

Often, it is the case there is a strong intertemporal dependence between the variables of interest y_t, which gives rise to modelling those variables y_t as being a function of past observations Y_t, the history of the time series at time t. A widely used class of time series models is the vector ARMA (VARMA) model in which the variables can be expressed as:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + B_1 \varepsilon_{t-1} + \cdots + B_q \varepsilon_{t-q} + \varepsilon_t. \tag{7}$$

The rst p terms on the right hand side of the equation are the autoregressive (AR) terms of the equation and A_i are the $M \times M$ parameter matrices of these AR terms. The M-dimensional sequence $\{\varepsilon_t\}$ is a time series of so called white noise, which means it has zero mean, constant variance and exhibits no correlation within the time series:

Definition 1 (White noise)**.** *A process $\{\varepsilon_t\}$ is a white noise process if the following hold:*

$\mathrm{E}[\varepsilon_t] = 0 \quad \forall t;$

$\mathrm{E}[\varepsilon_t \varepsilon_t^T] = \Sigma \quad \forall t$ *where* Σ *is a constant matrix,*

$\mathrm{E}[\varepsilon_t \varepsilon_s^T] = 0 \quad \forall t \neq s.$

The variable ε_t can be interpreted as a random shock at time t. In VARMA models the random shocks of past time periods, the moving average (MA) terms, have direct in uence on the variable y_t and are incorporated in the model as $B_1 \varepsilon_{t-1}, \ldots, B_q \varepsilon_{t-q}$, where the parameter matrices B_i are of size $M \times M$. In AR models, that is VARMA models without MA terms, random shocks of past time periods also have e ect on future values of the time series $\{y_t\}$, but then indirectly through the autoregressive terms.

The VARMA($p;q$) model can be written compactly as

$$A_p(L)y_t = B_q(L)\ _t; \tag{8}$$

where L is the lag operator, that is $Ly_t \quad y_{t\ 1}$ and $L^j y_t \quad y_{t\ j}$, and $A_p(L)$ and $B_q(L)$ are respectively the AR polynomial and the MA polynomial:

$$A_p(L) = I \quad A_1L \quad ::: \quad A_pL^p;$$
$$B_q(L) = I + B_1L + ::: + B_qL^q:$$

An important characteristic of time series is stationarity, of which there are multiple, related de nitions (Karlin, 2014). A stochastic process is stationary in the strict sense if the joint unconditional probability density function does not depend on the time period:

Definition 2 (Strict stationarity)**.** *A stochastic process fy_tg is strictly stationary if the joint distribution of $fy_{t_1};:::;y_{t_n}g$ is the same as the joint distribution of $fy_{t_1+\ };:::;y_{t_n+\ }g$ for all values $\ ;t_1;:::;t_n\ 2\ Z$ and for all positive integers n.*

A related, slightly weaker form of stationarity is k-th order stationarity:

Definition 3 (k-th order stationarity)**.** *A stochastic process fy_tg is k-th order stationary if the joint distribution of $fy_{t_1};:::;y_{t_n}g$ is the same as the joint distribution of $fy_{t_1+\ };:::;y_{t_n+\ }g$ for all values $\ ;t_1;:::;t_n\ 2\ Z$ and for all positive integers $n \quad k$.*

A last form of stationarity is stationarity in the wide sense:

Definition 4 (Wide sense stationarity)**.** *A stochastic process is stationary in the wide sense if:*

$\mathrm{E}[y_t] = \ _y$, *8t, where* $\ _y$ *is a constant.*

$\mathrm{E}[(y_t \quad \mathrm{E}[y_t])(y_t \quad \mathrm{E}[y_t])^T] = \ _y$, *8t, where* $\ _y$ *is constant and nite*

$\mathrm{E}[(y_t \quad \mathrm{E}[y_t])(y_{t+h} \quad \mathrm{E}[y_{t+h}])^T] = \ _y(h)$, *8h* > 0 *and 8t. That is, the autocovariance matrix of the process only depends on the time di erence h and not on the time period t.*

Stationarity in the strict sense implies k-th order stationarity for all k and k-th order stationarity implies wide sense stationarity for $k \quad 2$ if the rst two moments are nite. However, wide sense stationarity does not necessarily imply second order stationarity.

A last thing to note is that certain time series may only be stationary in the limit for $t \ ! \ 1$, since the values of y_t depend on y_0, where the strength of the dependence decreases for $t \ ! \ 1$. In that case the mean and variance of y_t may be unequal to the mean and variance of $y_{t\ 1}$ for nite t. Such a process is called asymptotically stationary. It is assumed all time series started in the in nite past and we thus do not di erentiate between asymptotic stationarity and non-asymptotic stationarity.

The roots of the AR-polynomial can be used to formulate a su cient condition for (asymptotic) wide sense stationarity:

Definition 5 (Stability). *A VARMA process is stable if the AR-polynomial satisfies:*

$$det(A_p(z)) \neq 0; \quad \forall z \in \mathbb{C} \text{ s.t. } |z| \leq 1;$$

where $A_p(z)$ is the AR-polynomial in z: $A_p(z) = I - A_1 z - \dots - A_p z^p$.

Lemma 1 (Stationarity Condition: Lütkepohl (2005), Proposition 2.1). *A stable VARMA process is wide sense stationary.*

A characteristic of stable VARMA time series is that the effects of random shocks fade over time. In case there is a unit root solution to the AR-polynomial, that is $|z| = 1$, random shocks persist over time and the process is said to exhibit a stochastic trend. A notable example of a stochastic process with a unit root is the random walk, which is the univariate AR(1) process $y_t = y_{t-1} + \varepsilon_t$. If there is a solution in the interior of the unit circle of the AR-polynomial, the influence of random shocks is growing over time. This possibility is often not considered in the literature (Franses et al., 2014). The situation where the AR-polynomial has unit roots occurs frequently and such processes are said to be integrated:

Definition 6 (Integrated process). *A stochastic VARMA process which has an AR-polynomial that has d unit roots is called a d-th order integrated ($I(d)$) random process, $d = 0, 1, 2, \dots$.*

A VARMA(p,q) process which is d-th order integrated is called a VARIMA(p,d,q) process. For some multivariate stochastic processes some of the variables might share a stochastic trend, in which case we call the multivariate process cointegrated:

Definition 7 (Cointegrated process). *A stochastic process $\{y_t\}$ is cointegrated of order (d,b) if all components of the process are $I(d)$ and there exists a linear combination $z_t = \beta' y_t$ such that z_t is $I(d-b)$. β is called the cointegration vector.*

Often when a stochastic process $\{y_t\}$ is not stationary, the stochastic process $\{y_t - y_{t-1}\} = \{\Delta y_t\}$ is stationary. Note that we can always try to factorize $(1-L)$ from the AR-polynomial $A_p(L)$ such that we obtain a new polynomial $A_p^*(L)$ for which holds $A_p(L)y_t = A_p^*(L)(1-L)y_t = B_q(L)\varepsilon_t$. In case $z = 1$ is a solution to the AR-polynomial $A_p(z) = 0$, the factorized polynomial $A_p^*(z)$ contains one fewer unit root. In many of the univariate non-stationary time series encountered in practice, there is a sole unit root equalling $z = 1$ and differencing causes the time series $\{\Delta y_t\}$ to be stable by Definition 5 and thus stationary following Lemma 1.

This idea is used in Vector Error Correction Models (VECM), introduced by Granger (1981). If we have a cointegrated VAR(p)-model in which all variables are $I(1)$ or $I(0)$, the corresponding VECM is:

$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \dots + \Gamma_{p-1} \Delta y_{t-p+1} + u_t. \tag{9}$$

In the above equation, $\{u_t\}$ is a white noise process, $\Pi = (I_M - A_1 - \dots - A_p)$ and $\Gamma_i = -(A_{i+1} + \dots + A_p)$, where I_M is the identity matrix of dimension M. Assuming the process $\{y_t\}$ is stable, the right hand side of Equation (9) must be too. Since the regressors $\Delta y_{t-1}, \dots, \Delta y_{t-p+1}$ are stable, Πy_{t-1} must be too and thus the variables y_t are cointegrated with the rows of Π as cointegration vectors (Kilian and Lütkepohl, 2017, p76). The term Πy_{t-1} is the error correction term, which indicates how short term deviations from the long run equilibrium relations, or cointegration relations, are corrected, hence the name Vector Error Correction Model.

2.2.2 Markov switching time series

A Markov Switching (MS) time series model is a time series model where the parameters vary between the time periods and are determined by an unobservable, or hidden, Markov process. A Markov Process fS_tg is a stochastic process with the property that the future is independent of the past given the present. More precisely $P(S_{t+1} = s_{t+1}jS_0 = s_0;:::;S_t = s_t) = P(S_{t+1} = s_{t+1}jS_t = s_t)$, for all states $s_0;:::;s_{t+1}$ and all time periods t. It is assumed in this thesis that the reader has basic knowledge on Markov Processes. In this thesis, we assume that the Markov Process fS_tg is irreducible, stationary and ergodic, that is aperiodic and positive recurrent. Moreover, we assume that fS_tg has a nite number of states.

If we assume that the time series fy_tg given the state S_t is linear in the past observations and past shocks, the stochastic process $fS_t;y_tg$ can be expressed as:

$$y_t = A_{1;s_t}y_{t\ 1} + ::: + A_{p;s_t}y_{t\ p} + B_{1;s_t\ t\ 1} + ::: + B_{q;s_t\ t\ q} + \ _t; \tag{10}$$

where the covariance matrix of $\ _t$ may also depend on the state s_t. The outcomes of the Markov process s_t are called the modes of the joint stochastic process $fS_t;Y_tg$. Note that the matrices $A_{p;s_t}$ may be 0 for some states s_t, so that the order $(p;q)$ may e ectively be di erent for di erent states s_t. We will call model (10) with the process fS_tg having a cardinality of K an MS(K) VARMA($p;q$) model.

Since an MS VARMA process can be seen as a mixture of K di erent VARMA processes, one may expect that the switching process is stationary if each of the K submodels is a stationary process. However, this is neither a necessary, nor a su cient condition for stationary. Francq and Zako•an (2001) have investigated stationarity conditions for the process fx_tg, where

$$x_t = \ _{s_t} + \sum_{i=1} A_{i;s_t}x_{t\ i} + \ _t + \sum_{j=1} B_{j;s_t\ t\ j}$$

is of a slightly more general model family than y_t of Equation 10, since it is an M-dimensional MS VARMA process with a mode-dependent mean $\ _{s_t}$. It is assumed the white noise process $\ _t$ can be written as $\ _t = \ _{s_t\ t}$, where $\ _{s_t}$ functions as the mode-dependent covariance matrix and $\ _t$ is i.i.d. The process is rewritten as an MS(K) VAR(1) process fz_tg as follows:

$$z_t = !_t + \ _t z_{t\ 1}: \tag{11}$$

In the above equation $!_t = \ _t + \ _t$,

$$\ _t = \begin{bmatrix} \ _{s_t} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; z_t = \begin{bmatrix} X_t \\ X_{t\ 1} \\ \vdots \\ X_{t\ p+1} \\ \ _t \\ \ _{t\ 1} \\ \vdots \\ \ _{t\ q+1}; \end{bmatrix}; \ _t = \begin{bmatrix} \ _{s_t} \\ 0 \\ \vdots \\ 0 \\ \ _{s;t} \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \ _t = \begin{bmatrix} A_{1;s_t} & A_{2;s_t} & ::: & A_{p;s_t} & B_{1;s_t} & ::: & B_{q;s_t} \\ I_k & 0 & ::: & 0 & 0 & ::: & 0 \\ \vdots & I_k & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ::: I_k & 0 & 0 & ::: & 0 \\ 0 & 0 & ::: & 0 & 0 & ::: & 0 \\ 0 & 0 & ::: & 0 & I_k & 0::: & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & I_k \ddots & \vdots \\ 0 & 0 & ::: & 0 & 0 & ::: I_k & 0 \end{bmatrix};$$

where $_t$ and z_t are $M(p+q)$-dimensional vectors, $_t$ is an $M(p+q)$ M-dimensional matrix and $_t$ is an $M(p+q)$ $M(p+q)$-dimensional matrix. The matrix $_t$ can be decomposed as a block matrix as follows:

$$_t = \begin{matrix} A_t & B_t \\ 0 & J \end{matrix};$$

with $A_t = \begin{matrix} A_{1;s_t} & A_{2;s_t} & \cdots & A_{p;s_t} \\ I_k & 0 & \cdots & 0 \\ \vdots & I_k \ddots & \ddots & \vdots \\ 0 & 0 & \cdots I_k & 0 \end{matrix}$, $B_t = \begin{matrix} B_{1;s_t} & \cdots & B_{q;s_t} \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{matrix}$ and $J_t = \begin{matrix} 0 & \cdots & 0 \\ I_k & 0 \cdots & 0 \\ \vdots & I_k \ddots & \vdots \\ 0 & \cdots I_k & 0 \end{matrix}$.

The su cient stationary condition formulated by Francq and Zako•an (2001) requires the top Lyapunov exponent, which for the time series (11) is de ned as:

$$= \inf_{t2\mathbb{N}} \mathrm{E}\frac{1}{t}\log jj \quad _t \quad _{t\ 1} \cdots \quad _1 jj: \tag{12}$$

The stationarity condition is as follows:

Lemma 2 (Francq and Zako•an (2001), Theorem 1)**.** *If the top Lyapunov exponent de ned in Equation (12) is strictly negative, the time series (11) converges a.s. and X_t is strictly stationary.*

It was shown that instead of the top Lyapunov exponent in the above lemma, using the quantity ^{0} can be used, which is easier to calculate. Where

$$^0 = \inf_{t2\mathbb{N}} \mathrm{E}\frac{1}{t}\log jjA_tA_{t\ 1} \cdots A_1 jj: \tag{13}$$

Chapter 3

Autocovariance method

In this chapter, the autocovariance method will be discussed, which allows for the estimation of a lower bound on the number of hidden Markov states for MS VARMA models. The main component of this method is are theorems tying the autocovariance of an MS VARMA($p;q$) model to that of a non-MS VARMA($p;q$) model with formulas tying the number of hidden Markov states K to the values of $p;q;p$ and q. The steps of the autocovariance method are shown in Figure 10. We assume that there is a data set coming from an MS VARMA model and that we do not know the number of modes K or any other parameters of the model, except the AR and MA orders p and q. If needed and possible, we transform the data to MS VECM(0,q) data by di erencing the observations in step 2. In step 3, we obtain the orders p and q using a so-called order selection method. Lastly, in step 4 we apply the formula, tying K to $p;q;p$ and q.

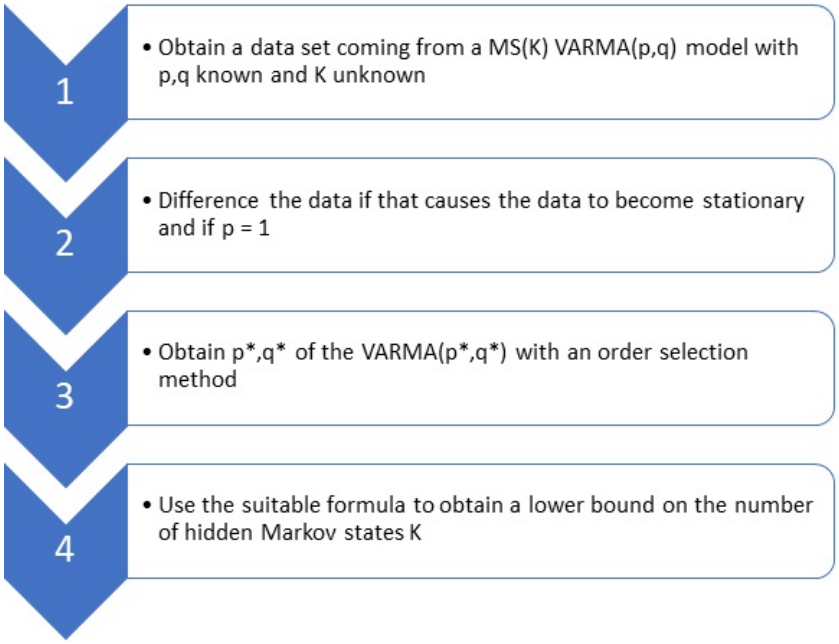

Figure 10: The practical procedure of the autocovariance method to obtain a lower bound on the number of hidden Markov states.

In this chapter, the necessary theoretical steps are taken to be able to use and fully grasp this practical procedure. For clarity, these theoretical steps are visualised in a diagram in Figure 11. To obtain the formulas of step 4 in Figure 10, we need to relate the autocovariance function of a non-MS VARMA model to that of an MS VARMA model. Step 1 of the theoretical steps, the autocovariance function of a non-MS VARMA model, is discussed in Section 3.1.1. The rest of Section 3.1 is used to obtain multiple lower bounds, step 3, by means of the autocovariance functions of the MS VARMA and MS VECM models, step 2. In Section 3.2, order selection methods to obtain the unknown parameters p and q are discussed, corresponding to step 4. In Section 3.3, some experiments are performed to establish the performance of the autocovariance method under various circumstances.

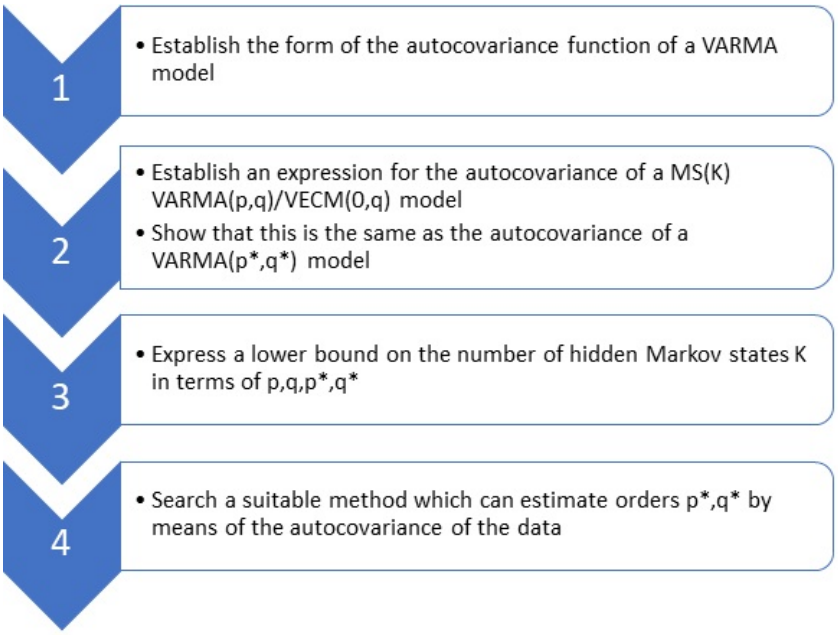

Figure 11: Theoretical steps taken to establish the autocovariance method.

3.1 Relation ARMA and MS time series

The idea to use the relation between Markov Switching time series models and ARMA time series to determine the number of hidden Markov states is not new. Several authors have proven that certain speci c MS models have the same autocovariance structures as ARMA models. Karlsen (1990) relates the MS AR(1) model with K modes to an ARMA(p,q) model with $p \quad K$ and $q \quad K \quad 1$ for scalar time series. Zhang and Stine (2001) provide conditions for which the autocovariance of general stochastic processes can be related to that of VARMA models. Moreover, they provide explicit bounds on the number of modes K for VAR(p) models. Francq and Zako•an (2001) provide stricter bounds under the same conditions. Cavicchioli (2014) provide even stricter bounds in case the regime variable is uncorrelated with the observable variable. In the next subsection the main results of the latter three papers are introduced, after which these results are extended to be able to give bounds for cointegrated time series in Section 3.1.2.

3.1.1 Stationary Markov Switching Time Series

Beguin et al. (1980) identi ed a necessary and su cient condition for stochastic processes to possess a scalar ARMA representation. Zhang and Stine (2001) extended this result to multidimensional stochastic processes:

Theorem 3 (Zhang and Stine (2001), Theorem 1)**.** *A M-dimensional second-order stationary process fY_tg with mean zero has a minimal VARMA(p,q) representation if and only if the autocovariance function $_Y(h) = \mathrm{E}[Y_t Y_t^T{}_h]; h \ 2 \ \mathbb{Z}$ is such that $(p;q)$ is the smallest pair for which there exist $M \quad M$ matrices $A_1;:::;A_p; A_p \ 6= \ 0$ such that:*

$$_Y(h) \quad A_1 \ _Y(h \quad 1) \quad ::: \quad A_p \ _Y(h \quad p) \begin{cases} = 0 & ; \text{if } h \quad q + 1; \\ 6= 0 & ; \text{if } h = q: \end{cases} \tag{14}$$

With having a minimal (V)ARMA representation, we mean in this thesis that the autocovariance function of the process $_Y(h)$ is equal to the autocovariance (h) of some (V)ARMA$(p \ ;q \ )$ process and where p is the lowest autoregressive order for which this holds and q is the lowest moving average order for which this holds, given that the autoregressive order is p .

The following two lemmas, generalizations of a result of Zhang and Stine (2001, Lemma 1) by Cavicchioli (2014) provides a tool to see whether Theorem 3 holds, as will become clear in later proofs.

Lemma 4 (Cavicchioli (2014), Theorem 4)**.** *If the autocovariance function $_Y(h)$ of an M-dimensional second-order stationary process fY_tg satis es:*

$$B(L) \ _Y(h) = a^T Q^h b \tag{15}$$

for $h \quad q \quad 0$, where $a;b \ 2 \ \mathrm{R}^{n \quad M} nf0g$, $Q \ 2 \ \mathrm{R}^{n \quad n}$ and $B(L) = \sum_{i=0}^{p} B_i L^i$ is a p-th degree lag polynomial with $B_i \ 2 \ \mathrm{R}^{M \quad M}; B_0 = I_m; B_p \ 6= \ 0$, then fY_tg has a VARMA$(p \ ;q \ )$ representation where $p \quad n + p$ and $q \quad n + q \quad 1$.

In the following lemma, a different generalization of Lemma 1 of Zhang and Stine is introduced and proved. The proof is along the lines of that of Zhang and Stine.

Lemma 5. *If the autocovariance function $\gamma_Y(h)$ of an M-dimensional second-order stationary process $\{Y_t\}$ satisfies:*

$$\gamma_Y(h) = a^T Q^{h-c} b \tag{16}$$

for $h \geq c \geq 0$, where $a, b \in \mathbb{R}^{n \times M} \setminus \{0\}$ and $Q \in \mathbb{R}^{n \times n}$ then $\{Y_t\}$ has a VARMA(p, q) representation where $p \leq n$ and $q \leq n + c - 1$.

Proof. By the Cayley-Hamilton theorem (see for instance Straubing (1983)), there exist parameters $\alpha_1, \ldots, \alpha_n$ such that

$$Q^n - \alpha_1 Q^{n-1} - \cdots - \alpha_n I_n = f(Q) = 0,$$

where $f(\cdot)$ is the characteristic polynomial of Q. It follows that

$$a^T Q^d (Q^n - \alpha_1 Q^{n-1} - \cdots - \alpha_n I_n) b = 0,$$

for $d = 0, 1, \ldots$. The above equation can be expressed as

$$\gamma_Y(n+d+c) - \alpha_1 \gamma_Y(n+d+c-1) - \cdots - \alpha_n \gamma_Y(c+d) = 0, \quad \text{for } d = 0, 1, 2, \ldots$$

It follows that Equation (14) in Theorem 3 is satisfied with $p = n$ and $h \leq n + c$, which gives us an upper bound on the smallest pair (p, q) which satisfies that equation: $p \leq n$ and $q \leq n + c - 1$. □

The above three results are useful, since they provide a tool to know whether a stochastic process can be represented as a VARMA model. If we can find an expression for the autocovariance function of the MS process and manage to write it in the form of Equation (15) or (16), it is possible to bound the orders of the corresponding non-MS VARMA models p, q from above. For the Markov Switching time series models considered this bound turns out to depend on the number of hidden Markov states of the Markov process K. If we can estimate the order of the corresponding VARMA model, we can thus gain a lower bound on the number of hidden Markov states K.

All the following lemmas and theorems provide such bounds for the order of the corresponding VARMA models, where the proofs rely on finding expressions for the autocorrelation function and using Lemmas 4 and 5. Another result of the literature we need is that of Karlsen (1990), who has provided a formula for the autocovariance function of functions of the form $Y_t = A_{s_t} Y_{t-1} + \varepsilon_{s_t} V_t$. The exact form is not of interest in this thesis, although it is important to know that the autocovariance function can be written in the form of Equation (15).

Theorem 6 (Karlsen (1990), Theorem 4.1)**.** *Let Y_t be an MS time series of the form $Y_t = U_{s_t} + A_{s_t}(Y_{t-1} - U_{s_{t-1}}) + \varepsilon_{s_t} V_t$, where U_{s_t} functions as a mode-dependent mean. If the absolute values of all eigenvalues of the matrix*

$$\mathrm{diag}\{((\alpha_i, \beta_j^{-1} + \gamma_i)(\alpha_i, \beta_j^{-1} + \gamma_i)); i = 1, \ldots, K; j = 1, \ldots, K\}(P^T \otimes I_{M^2}); \tag{17}$$

are smaller than 1, the covariance of Y_t can be expressed in the form $a^T Q^h b$.

Zhang and Stine used Lemma 4 to obtain a bound on VAR(p) processes. This bound is extended to VARMA(p,q) processes in Corollary 8.

Theorem 7 (Zhang and Stine (2001), Theorem 4)**.** *A stationary M-dimensional MS VAR(p) process fY_tg, with K modes has a VARMA(p ; q) representation, with p $K(Mp)^2$ and q $K(Mp)^2$ 1.*

Corollary 8. *A stationary M-dimensional MS VARMA(p,q) process fY_tg, with K modes has a VARMA(p ; q) representation, with p $K(M(p+q))^2$ and q $K(M(p+q))^2$ 1*

Proof. A MS VARMA(p,q) process can be written as an MS VAR(1) process with $M(p+q)$ variables. Substituting $M^0 = M(p+q)$ and $p^0 = 1$ in Theorem 7 gives the desired result. □

The above bounds are in general not tight and can be improved on without any additional assumptions:

Theorem 9 (Francq and Zako•an (2001), Section 4.3)**.** *A stationary M-dimensional MS VARMA(p,q) process fY_tg, with K modes has a VARMA(p ; q) representation, with p $KM(p+q)$ and q $KM(p+q)$.*

With an additional assumption Cavicchioli (2014) has shown that for MS VAR(p) processes of the form

$$A_{p;s_t}(L)Y_t = v_{s_t} + \quad _{s_t} \ _t; \tag{18}$$

stricter bounds are possible. v_{s_t} is an intercept vector which given the state s_t is constant. The matrix $[v_1 \quad v_K; :::; v_{k\ 1} \quad v_K]$ must be di erent from the zero matrix. This is satis ed if not all v_{s_t} are equal. In case these are all equal, a translation of the variables $\tilde{Y}_t \quad Y_t + c$, where c is some constant may be enough to obtain the correct form.

Theorem 10 (Cavicchioli (2014), Theorem 7)**.** *Let $fY_t; S_tg$ be an M-dimensional MS VAR(p) process of the form of Equation (18) with K Markov states, where fY_tg are the observed variables and fS_tg is the hidden Markov process. If the indicator $I_{S_{t+h}=k}$ is uncorrelated with Y_t for all h 0 and all k, then Y_t has a stable VARMA(p ; q) representation with p $K + p$ 1, q K 1.*

The assumption that the Markov process S_t is uncorrelated with the observed time series is counterintiative, since there is a direct causal relation between these two processes. However, correlation is neither a su cient nor a necessary condition for causation and in certain cases the condition may hold, which in Section 5.4 will be discussed further.

Above theorem can be extended to MS VARMA($p;q$) processes. The proof is along the lines of the proof of Theorem 10 of Cavicchioli.

Theorem 11. *Let $fY_t; S_tg$ be an M-dimensional stationary MS(K) VARMA(p,q) process of the form*

$$A_{p;s_t}(L)Y_t = v_{s_t} + B_{q;s_t}(L) \ _t; \tag{19}$$

with $A_{p;s_t}$ 6= 0, $B_{q;s_t}$ 6= 0, where fY_tg are the observed variables and fS_tg is the irreducible, stationary and ergodic hidden Markov process. If the indicator $1_{S_{t+h}=k}$ is uncorrelated with Y_t

for all $h \quad 0$ *and all* k*, then* Y_t *has a stable VARMA*$(p\ ;q\ )$ *representation with* $p \quad K+p \quad 1$*,* $q \quad K+q \quad 1$*.*

Proof. We will rst rewrite Equation (19) in such a way that the subscripts s_t are replaced by a variable $\ _t$, which is a vector with its i-th entry equal to 1 if $s_t = i$ and the other entries equal to zero. Afterwards we will nd an expression of the covariance of both the left hand side and the right hand side of the rewritten Equation (19) with Y_t. As a last step we apply Lemma 4.

Let the Markov Chain S_t have transition matrix $P = (p_{ji})$ and stationary probabilities $\ $. We can express the process $f\ _t g$ as a VAR(1) model

$$\ _t = P^T\ _{t\ 1} + u_t;$$

where $u_t = \ _t \quad \mathrm{E}[\ _t j\ _{t\ 1}]$. Let $\mathrm{E}[\ _t] = \ $, then $\mathrm{E}[\ _t\ _{t+h}^T] = DP^h$, with $D = \mathrm{diag}(\ _1;:::;\ _K)$. Let $A_p(L) = [A_{p;1}(L);:::;A_{p;K}(L)]$ be the vector of the K di erent AR-polynomials of the di erent Markov states and set $B_q(L) = [B_{q;1}(L);:::;B_{q;K}(L)]$ to be the vector of the MA-polynomials. Moreover, let $V = [v_1;:::;v_K]$ be the vector with the K di erent intercepts. Then we can rewrite Equation (19) as

$$A_p(L)(\ _t \quad I_M)Y_t = V\ _t + B_q(L)(\ _t \quad I_M)\ _t: \tag{20}$$

Now de ne X_t as the left hand side of this equation. We can express $\mathrm{cov}(X_{t+h};Y_t)$ in terms of the autocorrelation function of Y_t; $\ _Y(h)$

$$\begin{aligned}
\mathrm{cov}(X_{t+h};Y_t) &= \mathrm{cov}(A_p(L)(\ _{t+h} \quad I_M)Y_{t+h};Y_t) \\
&= \mathrm{E}[(A_p(L)(\ _{t+h} \quad I_M)Y_{t+h} \quad \mathrm{E}[A_p(L)(\ _{t+h} \quad I_M)Y_{t+h}])(Y_t \quad \mathrm{E}[Y_t])^T] \\
&= A_p(L)\mathrm{E}[((\ _{t+h} \quad I_M)Y_{t+h} \quad \mathrm{E}[(\ _{t+h} \quad I_M)Y_{t+h}])(Y_t \quad \mathrm{E}[Y_t])^T] \\
&= A_p(L)\mathrm{E}[((\quad I_M)Y_{t+h} \quad \mathrm{E}[(\quad I_M)Y_{t+h}])(Y_t \quad \mathrm{E}[Y_t])^T] \\
&= A_p(L)(\quad I_M)\mathrm{cov}(Y_{t+h};Y_t) \\
&= A_p(L)(\quad I_M)\ _Y(h);
\end{aligned}$$

where the fourth step can be taken since S_{t+h} and Y_t are uncorrelated.

We now examine the right hand side of Equation (20). We will rewrite this a second time, now by replacing $\ _t$ with $\ _t$, which is de ned to be $\ _t = (\ _{1;t} \quad \ _1;:::;\ _{K\ 1;t} \quad \ _{K\ 1})$. The $(K \quad 1)$-dimensional, zero mean process $f\ _t g$ can also be expressed as an AR(1) model

$$\ _t = F\ _{t\ 1} + w_t; \tag{21}$$

with $w_t = [I_{K\ 1};\ \mathbf{1}_{K\ 1}]u_t$, where $\mathbf{1}_{K\ 1}$ is a $K \quad 1$-dimensional vector of ones and

$$F = \begin{bmatrix} p_{11} \quad p_{K1} & \cdots & p_{K\ 1;1} \quad p_{K1} \\ \vdots & \ddots & \vdots \\ p_{1;K\ 1} \quad p_{K;K\ 1} & \cdots & p_{K\ 1;K\ 1} \quad p_{K;K\ 1} \end{bmatrix}:$$

It follows from the AR(1) representation of Equation (21) that $\ _{t+h} = F^h\ _t + \sum_{j=0}^{h-1} F^j w_{t+h\ \ j}$. The right hand side of Equation (20) then becomes

$$V\ _t + B_q(L)(\ _t \quad I_M)\ _t = V \quad + \Psi\ _t + B_q^{\wedge}(L)(\ _t \quad I_M)\ _t + B_q(L)(\quad I_M)\ _t: \tag{22}$$

Here $\Psi = [v_1 \quad v_K; : : : ; v_{K\ \ 1} \quad v_K]$ and $B_q^{\wedge}(L) = [B_{q;1}(L) \quad B_{q;K}(L); : : : ; B_{q;K\ \ 1}(L) \quad B_{q;K}(L)]$. Now set the right hand side of Equation (22) equal to X_t, which is naturally equal to the previously de ned X_t. We again examine the covariance of X_{t+h} with Y_t. For $h > q$

$$\begin{aligned}
\text{cov}(X_{t+h}; Y_t) &= \text{cov}(V \quad + \Psi\ _{t+h} + B_q^{\wedge}(L)(\ _{t+h} \quad I_M)\ _{t+h} + B_q(L)(\quad I_M)\ _{t+h}; Y_t) \\
&= \text{cov}(\Psi\ _{t+h}; Y_t) \\
&= \text{E}[\Psi F^h\ _t Y_t^T] \\
&= \Psi F^h \text{E}[\ _t Y_t^T]:
\end{aligned}$$

The expression for $h \quad q$ is di erent and not of interest. Equating the two expressions for $\text{cov}(X_{t+h}; Y_t)$, we get that $A_p(L)(\quad I_M)\ _Y(h) = \Psi F^h \text{E}[\ _t Y_t^T]$. This is in the right form to use Lemma 4 with q replaced by $q + 1$. We then obtain that the MS VARMA(p,q) process has a VARMA($p\ ; q\ $) representation with $p \quad K + p \quad 1$ and $q \quad K + q \quad 1$. □

3.1.2 Cointegrated time series

In this section, the results of the previous section are extended to di erenced time series. This will enable the autocovariance method to be used for cointegrated time series that become stationary when they are di erenced. We rst proof a result for cointegrated time series without the assumption that S_{t+h} is uncorrelated with Y_t for all $h \quad 0$. Afterwards we will improve on the following result, while making this additional assumption.

Theorem 12. *Let $f\ Y_t g$ be a stationary MS VECM(0,q) process, $Y_t = \ _{s_t} Y_{t\ \ 1} + B_q(L)V_t$ where the process $fS_t g$ has K modes. If the process has invertible matrices $\ _{s_t} 8 s_t$ and if the absolute values of all eigenvalues of the matrix*

$$diag f((\ _i; \ _{j}^{\ 1} + \ _i) \quad (\ _i; \ _{j}^{\ 1} + \ _i)); i = 1; : : : ; K; j = 1; : : : ; K g(P^T \quad I_{M^2}); \tag{23}$$

are smaller than 1, then $f\ Y_t g$ has a VARMA($p\ ; q\ $) representation, for which it holds that $p \quad K^2 M(1 + q) + KM$ and $q \quad \min\ K^2 M + KM + q; (KM)^2 + KM + q \quad 1\ .$

Proof. The process $f\ Y_t g$ can be expressed as:

$$\begin{aligned}
Y_t &= \ _{s_t} Y_{t\ \ 1} + B_q(L)V_t \\
Y_t &= (I + \ _{s_t}) Y_{t\ \ 1} + B_q(L)V_t \\
\ _{s_{t+1}} Y_t &= \ _{s_{t+1}} (I + \ _{s_t}) Y_{t\ \ 1} + \ _{s_{t+1}} B_q(L)V_t
\end{aligned}$$

Define $X_t = \ _{s_{t+1}} Y_t$. Then we can write the above equation as:

$$X_t = D_{s_t^0} X_{t\ \ 1} + B_q^0(L)V_t; \tag{24}$$

with $s_t^{\emptyset} = (s_{t+1}; s_t)$; $D_{s_t^{\emptyset}} = {}_{s_{t+1}\ s_t}{}^{1} + {}_{s_{t+1}}$ and $B_q^{\emptyset}(L) = B_q^{\emptyset}(L; s_t^{\emptyset}) = {}_{s_{t+1}} B_q(L)$. Since $fS_t g$ is a Markov process, the process $fS_t^{\emptyset} g$ is as well with at most K^2 modes, where all states $(S_i; S_j)$ with transition probabilities $P_{S_i;S_j} = 0$ are omitted from the Markov Chain. As a rst step, it is proven that $fS_t^{\emptyset} g$ is ergodic. For $fS_t^{\emptyset} g$ to be aperiodic, every state $(S_i; S_j)$ must be able to occur at each time period t. Since $fS_t g$ is aperiodic, S_i can happen every time period and since $fS_t g$ is a Markov Chain, the transition from S_i to S_j must happen with the same probability in each time step. If this probability is positive, the state $(S_i; S_j)$ can happen every time period, if the probability is 0, the state never occurs and was omitted from the combined Markov Chain $fS_t^{\emptyset} g$. This applies to any state $(S_i; S_j)$ of $fS_t^{\emptyset} g$ and therefore the combined Markov Chain is aperiodic. Moreover, the combined Markov Chain is positive recurrent. To see this, note that the probability to ever return from any non-omitted state $(S_i; S_j)$ to $(S_k; S_l)$ is by the Markovian property equal to $P_{S_i;S_j} r(S_j; S_k) P_{S_k;S_l}$, where $r(S_j; S_k)$ is the recurrence probability from state S_j to S_k. Since this probability is strictly positive, as the original Markov Chain is positive recurrent and the transition probabilities are strictly positive as well, since the two combined states are not omitted, $P_{S_i;S_j} r(S_j; S_k) P_{S_k;S_l}$ is strictly positive. We can thus conclude that the Markov Chain $fS_t^{\emptyset} g$ is ergodic.

Since Y_t is stationary, X_t is as well and therefore X_t is a stationary M-dimensional MS VARMA(1,q) process. It follows from Theorem 7 and 9 that X_t has a VARMA(p',q') representation with $p \quad K^2 M(1+q)$, $q \quad K^2 M(1+q)$ and $q \quad K^2 (M(1+q))^2 \quad 1$. Moreover, since the eigenvalues of the matrix of Equation (23) are inside the unit circle, the autocovariance function can be written as $a_x^T Q_x b_x$, following Theorem 6.

Since $Y_T = X_{t\ 1} + B_q(L) V_t$, we can express the covariance function of Y_t for $h > q$ as

$$\begin{aligned}
\mathrm{E}[\ Y_t\ \ Y_t^T{}_h] &= \mathrm{E}[(X_{t\ 1} + B_q(L)V_t)(X_{t\ h\ 1} + B_q(L)V_{t\ h})^T]; \\
&= \mathrm{E}[X_{t\ 1} X_{t\ h\ 1}^T] + \mathrm{E}[B_q(L)V_t X_{t\ h\ 1}^T] + \mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] \\
&\quad + \mathrm{E}[B_q(L)V_t (B_q(L)V_{t\ h})^T]; \\
&= \mathrm{E}[X_t X_t^T] + \mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T];
\end{aligned}$$

where the last step follows from the white noise property of V_t. For general q, the last term can be expressed as

$$\begin{aligned}
\mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] = \mathrm{E}[&D_{s_{t\ 1}^{\emptyset}}(D_{s_{t\ 2}^{\emptyset}}(:::(D_{s_{t\ h\ q}^{\emptyset}} X_{t\ h\ q\ 1} + B_q^{\emptyset}(L; s_{t\ h\ q}^{\emptyset}) V_{t\ h\ q}) \\
&+ B_q^{\emptyset}(L; s_{t\ h\ q+1}^{\emptyset}) V_{t\ h\ q+1}) :::) \\
&+ B_q^{\emptyset}(L; s_2^{\emptyset}) V_{t\ 2}) + B_q(L; s_1) V_{t\ 1})(B_q(L; s_{t\ h}) V_{t\ h})^T];
\end{aligned}$$

which after working out the brackets and the lag polynomials can due to lack of autocorre-

lation of the white noise V_t be simpli ed to

$$\begin{aligned}
\mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] = \mathrm{E}[(&D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h\ q+1}}\ s_{t\ h\ q+1} B_{0;s_{t\ h\ q}} V_{t\ h\ q} \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h\ q+2}}\ s_{t\ h\ q+2} B_{1;s_{t\ h\ q+1}} V_{t\ h\ q} + \cdots \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h+1}}\ s_{t\ h+1} B_{q;s_{t\ h}} V_{t\ h\ q}) V^T_{t\ h\ q} B^T_{q;s_{t\ h}}] + \\
\mathrm{E}[(&D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h\ q+2}}\ s_{t\ h\ q+2} B_{0;s_{t\ h\ q+1}} V_{t\ h\ q+1} \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h\ q+3}}\ s_{t\ h\ q+3} B_{1;s_{t\ h\ q+2}} V_{t\ h\ q+1} + \cdots \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h+2}}\ s_{t\ h+2} B_{q;s_{t\ h+1}} V_{t\ h\ q+1}) V^T_{t\ h\ q+1} B^T_{q\ 1;s_{t\ h}}] + \cdots + \\
\mathrm{E}[(&D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h+1}}\ s_{t\ h+1} B_{0;s_{t\ h}} V_{t\ h} \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h+2}}\ s_{t\ h+2} B_{1;s_{t\ h+1}} V_{t\ h} + \cdots \\
&+ D_{s^0_{t\ 1}} \cdots D_{s^0_{t\ h+q+1}}\ s_{t\ h+q+1} B_{q;s_{t\ h+q}} V_{t\ h}) V^T_{t\ h} B^T_{0;s_{t\ h}}]
\end{aligned}$$

Expressing the above expectation in terms of $s_{t\ i}$ and rearranging terms leads to

$$\begin{aligned}
\mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] = \mathrm{E}[&\ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h\ q+1}) B_{0;s_{t\ h\ q}} V_{t\ h\ q} \\
&+ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h\ q+2}) B_{1;s_{t\ h\ q+1}} V_{t\ h\ q} + \cdots \\
&+ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h+1}) B_{q;s_{t\ h}} V_{t\ h\ q}) V^T_{t\ h\ q} B^T_{q;s_{t\ h}}] + \cdots \\
+ \mathrm{E}[&\ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h+1}) B_{0;s_{t\ h}} V_{t\ h} \\
&+ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h+2}) B_{1;s_{t\ h+1}} V_{t\ h} + \cdots \\
&+ s_t(I + s_{t\ 1}) \cdots (I + s_{t\ h+q+1}) B_{q;s_{t\ h+q}} V_{t\ h}) V^T_{t\ h} B^T_{0;s_{t\ h}}]
\end{aligned}$$

Let the variance of V_t be , $\mathrm{E}[V_t V_t^T] =$, P the probability transition matrix of the stochastic process fS_tg, $P_{j;i} = P(S_t = ijS_{t\ 1} = j)$, D the diagonal matrix of the steady state probabilities of fS_tg, $D_{i;i} = (i)$, $M_l = \mathrm{diag}fB_{l;i}; i = 1;\ldots;Kg$, $M =$ $\mathrm{diag}f_{\ i}; i = 1;\ldots;Kg$ and $M_{l+} = \mathrm{diag}fI + {}_i; i = 1;\ldots;Kg$. Let by convention $\prod_{k=2}^{0} a_k = 1$ if $h < k$. Then we can formulate the expectation as

$$\begin{aligned}
\mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] = \sum_{l=0}^{h+q} \sum_{m=0}^{l} \Bigg(&\sum_{i_1=1}^{K} P(s_{t\ h+l\ m} = i_1) \sum_{i_2=1}^{K} P_{i_1;i_2} \cdots \sum_{i_{h+m\ l}=1}^{K} P_{i_{h+m\ l\ 1};i_{h+m\ l}} \\
&\prod_{k=2}^{h+q\ l} (I + {}_{i_k}) B_{l;i_1} B^T_{m;i_{m\ l\ 1}} \Bigg) :
\end{aligned}$$

This expression can be rewritten in matrix form. For $q = 0$ this results in

$$\begin{aligned}
\mathrm{E}[X_{t\ 1}(B_q(L)V_{t\ h})^T] = &(1_K \quad I_M)^T M\ (P^T \quad 1_{M\ M})\ M_{I+}\ (P^T \quad 1_{M\ M})^{h\ 1} \\
&M_0(D \quad I_M)(\quad I_k) M_0^T (1_K \quad I_M);
\end{aligned}$$

which is in the form $a^T Q^{h\ 1} b$. Letting $M_{PB} = \mathrm{diag}fi; \sum_{k=1}^{K} P_{ki} B_{0;k}(_{\ k} \quad I_M)g$, we obtain for

q = 1:

$$
\begin{aligned}
&\mathrm{E}[X_{t-1}(B_q(L)V_{t-h})^T] = \\
&(1_K \otimes I_M)^T M \{M_{I+} (P^T \otimes 1_M \otimes I_M)\}^h M_{I+} M_{PB} (\Delta \otimes I_k) M_1^T (1_K \otimes I_M) \\
&+ (1_K \otimes I_M)^T M \{(P^T \otimes 1_M \otimes I_M) M_{I+}\}^{h-1} (P^T \otimes 1_M \otimes I_M) M_1 (\Delta \otimes I_k) M_1^T (1_K \otimes I_M) \\
&+ (1_K \otimes I_M)^T M \{(P^T \otimes 1_M \otimes I_M) M_{I+}\}^{h-1} (P^T \otimes 1_M \otimes I_M) M_0 (\Delta \otimes I_k) M_0^T (1_K \otimes I_M) \\
&+ (1_K \otimes I_M)^T M \{(P^T \otimes 1_M \otimes I_M) M_{I+}\}^{h-2} \\
&M_1 (P^T \otimes 1_M \otimes I_M) (\Delta \otimes I_k) M_0^T (1_K \otimes I_M). \qquad (25)
\end{aligned}
$$

It follows that the autocovariance can be expressed as $a^T Q^{h-2} b$ for $h \geq 2$. For general integer q the expression gets more lengthy but similar in the form $a^T Q^{h-q-1} b$, with $Q = M_{I+} (P^T \otimes 1_M \otimes I_M)$ a $KM \times KM$ matrix.

Now let

$$
a_Y = a_X Q_X^{q+1}; \; a_\Delta;
$$

$$
Q_Y = \begin{matrix} Q_X & 0 \\ 0 & M_{I+} (P^T \otimes 1_M \otimes I_M) \end{matrix}; \quad \text{and}
$$

$$
b_Y = b_X; \; b_\Delta.
$$

Then $\mathrm{E}[Y_t \otimes Y_{t-h}^T] = a_Y^T Q_Y^{h-q-1} b_Y$ for $h \geq q+1$. It follows from Lemma 5 that $p \leq K^2 M(1+q) + KM$ and $q \leq \min\{K^2 M + KM + q; (KM)^2 + KM + q - 1\}$. □

The bound of Theorem 10 can also be used to obtain a bound for MS VECM(0,q) models, where the first steps of the derivation are the same as the above proof. However, the result only holds if $X_t = \Delta_{S_{t+1}} Y_t$ is uncorrelated with $\{S_{t+h}; S_{t+h+1}\}$ for $h > 0$. Since V_t is white noise and thus uncorrelated with Y_{t-1}, X_t is uncorrelated with $\{S_t; S_{t+1}\}$ if and only if Y_t is uncorrelated with $\{S_{t+h}; S_{t+h+1}\}$ for $h > 0$, or equivalently with $\{S_{t+h}\}$.

Theorem 13. *Let $\{Y_t\}$ be a stationary MS VECM(0,q) process, $Y_t = \Delta_{s_t} Y_{t-1} + B_q(L)V_t$ where the process $\{S_t\}$ has K modes. If the process has invertible parameter matrices $\Delta_{s_t}; \forall s_t$ and if the indicator $I_{S_{t+h}=k}$ is uncorrelated with $\{Y_t\}$ for all $h \geq 0$ and all $k = 1, \ldots, K$, then $\{Y_t\}$ has a VARMA(p; q) representation, with $p \leq K^2$ and $q \leq K^2 + q$.*

Proof. As in the previous proof, define $X_t = \Delta_{s_{t+1}} Y_t$, such that

$$
X_t = D_{s_t^0} X_{t-1} + B_q^0(L)V_t.
$$

Since $Y_t = X_{t-1} + B_q(L)V_t$, we have that

$$
\begin{aligned}
Y_t - B_q(L)V_t &= D_{s_{t-1}^0}(Y_{t-1} - B_q(L)V_{t-1}) + B_q^0(L)V_{t-1}; \\
Y_t &= D_{s_{t-1}^0} Y_{t-1} - (D_{s_{t-1}^0} B_q(L) - B_q^0(L))V_{t-1} + B_q(L)V_t.
\end{aligned}
$$

And thus $\{Y_t\}$ is an MS VARMA(1,q+1) process with K^2 modes. Let $\forall Y_t = Y_t - c$. It is straightforward to see this is an MS VARMA(1,1) process in the form of Equation (19) with K^2 modes as well. Applying Theorem 11 with the process $\{\forall Y_t\}$, we obtain that $p \leq K^2$ and $q \leq K^2 + q$. □

The condition that $\{Y_t\}$ is uncorrelated with $\{S_t\}$ is for instance fulfilled if $\{Y_t\}$ is uncorrelated with $\{S_{t+h}\}$, for each $h > 0$ and the mean of Y_t is stationary.

Note that the technique of the proof of Theorem 13 can also be used to obtain a bound when no assumption is made on the correlation between Y_t and S_t, but that this bound is slightly weaker than that of Theorem 12. However, one does not need the condition that the eigenvalues of the matrix in Equation (23) are inside the unit circle.

Theorem 14. *Let $\{Y_t\}$ be a stationary MS VECM(0,q) process, $Y_t = \_\_\__{s_t} Y_{t-1} + B_q(L)V_t$ where the process $\{S_t\}$ has K modes. If the process has invertible matrices $\_\_\__{s_t}$ $\forall s_t$, then $\{Y_t\}$ has a VARMA(p ; q) representation, with $p \leq K^2M(2+q)$ and $q \leq K^2M(2+q)$.*

Proof. As shown in the proof of Theorem 13, the process Y_t can be expressed as an MS VARMA(1,q+1) process. The result follows from Corollary 8. □

The above theorems provide a way to estimate a lower bound on the number of modes K, if the order p ; q of the VARMA(p ; q) model having the same autocovariance structure as the MS VARMA($p;q$) can be estimated. In the following section various order selection methods are discussed.

3.2 Order selection

3.2.1 Order selection

Often, it is not theoretically clear which orders of p and q are most suitable for a VARMA($p;q$) model. An important tool in determining these orders is the autocorrelation function (ACF), since di erent orders of p and q give rise to di erent characteristic ACFs. The ACF is de ned as follows:

$$R(k) = R_k = D^{1=2}\ {}_kD^{1=2};$$

where D is a diagonal matrix with the variances of the j-th univariate variables y_{jt} of the multivariate time series Y_t on its diagonal, $D_{jj} = \mathrm{Var}(y_{jt})$ and ${}_k$ is the k-th order autocovariance, ${}_k = \mathrm{E}[(y_t \quad \mathrm{E}[y_t])(y_{t\ k} \quad \mathrm{E}[y_{t\ k}])^T]$. A function related to the ACF is the partial autocorrelation function (PACF) where (k) is de ned as the matrix ${}_k$ in the regression:

$$y_t = A_1 y_{t\ 1} + \cdots + A_{k\ 1} y_{t\ k+1} + \ {}_k y_{t\ k} + u_t:$$

In above regression equation, $fy_{t\ i}g_{i=1}^{k}$ are the regressors and $fA_i g_{i=1}^{k}$ and ${}_k$ are the parameter matrices estimated in the regression.

For pure AR models, the order p can easily be obtained from the PACF, since ${}_k = 0$ for $k > p$. Similarly, the order of pure MA models can easily be obtained from the ACF, since $R_k = 0$ for $k > q$. Comparing the empirical PACF and ACF with the theoretical counterparts for various ARIMA models was suggested rst by Box and Jenkins (1970), as part of the Box-Jenkins method, which is a wider approach for ARIMA model building. Tiao and Box (1981) have extended this approach to multivariate time series.

However, for ARMA models where both $p;q > 0$, the relation between the (P)ACF and the ARMA order is less clear by inspection and it is often too di cult to obtain p and q exactly from the empirical ACF and PACF, although it can provide a good rst idea. Velicer and Harrop (1983) showed in an experimental setting that students having followed a course in time series analysis were more often than not unable to correctly identify generated univariate ARIMA($p;d;q$) models with low orders p, d and q, even if p or q equalled 0. The Box-Jenkins and Tiao-Box methods are, therefore, in practice a rather unreliable manner of order selection.

Although the research into systematic approaches for ARMA order identi cation is extensive, there is, unfortunately, no recent literature or experimental review on this topic. Therefore, there are no straightforward best approaches. Choi (1992) claimed that the existing methods at that time could be categorized in penalty function methods, innovation regression methods and pattern identi cation methods. Penalty function methods include selection based on criteria like AIC and BIC or other penalty functions, although this requires knowledge about the distribution of the error terms, which is a rather strong assumption. Moreover, one needs to estimate parameters for each model under consideration, which may be computationally expensive. Innovation regression methods are methods based on a regression equation of the form of Equation (7) of Section 2.2.1, where the lagged error terms are for instance replaced by estimates of a previous iteration, or where the whole equation is estimated by means of

maximum likelihood estimation.

Pattern identi cation methods choose the order based on the autovariance structure. Note that a VARMA(p,q) process can be expressed as

$$\sum_{i=0} A_i y_{t\ 1} = \sum_{i=0} B_{i\ t\ i};$$

where $A_0 = B_0 = I$. After some algebraic manipulations, this leads to the multivariate Yule-Walker equations (Pollock, 2011):

$$\sum_{i=0} A_i \quad _i = \sum_{i=} B_i \quad _i^T : \tag{26}$$

Using estimates of the covariances, several procedures to obtain the orders of the VARMA model have been suggested, such as the three pattern method (Choi, 1993), the R and S array method (Gray et al., 1978), the Corner method (Beguin et al., 1980) and the extended sample autocorrelation function (ESACF) method (Tsay and Tiao, 1984). An advantage of these pattern identi cation methods is that no assumptions need to be made about the distribution of the errors $_t$, apart from being white noise. However, not all of the aforementioned methods have been constructed for multivariate ARMA processes.

More recently, several metaheuristics have been applied to the ARMA order selection problem succesfully, such as genetic algorithms (Abo-Hammour et al., 2012), threshold acceptance (Winker, 2000), or search algorithms (Hoglund and Ostermark, 1991).

Another approach makes use of the eigenpairs of the data covariance matrix. Liang et al. (1993) developed such a technique and showed that the eigenvalues of the data covariance matrix tend to zero when the order of ARMA models is greater than or equal to the real order in idealized situations. Lardies and Larbi (2001) used this property to propose an approach for estimating the order of multivariate AR models and Cassar et al. (2010) extended this approach to multivariate ARMA models. The disadvantage of this approach is that it heavily relies on normality of the errors $_t$.

In this thesis, a method is necessary which makes use of the autocovariance structure only and which is able to extract the order from multivariate time series. Although several authors have developed theory on the autocovariance method of multivariate time series, to our best knowledge experimental work in the literature hitherto only exists for 1-dimensional time series. In particular, the order selection method they used, the three pattern method, is only formulated for univariate models. No multivariate pattern recognition methods have been found which are solely based on the autocovariance structure. Nevertheless, two order selection methods have been selected to be implemented and used in experiments, although this leaves question marks about the validity of using such methods.

We will make use of the multivariate ESACF method by Tiao and Tsay (1983), which will be discussed in the next section. This choice is motivated by the fact that the distribution

of the errors can be unspecified, the low computational complexity and that that method scored best in an empirical comparison between different pattern identification methods by Chan (1999).

The second method that is implemented is the approach making use of the eigenvalues of the data covariance matrix by Cassar et al. (2010). This method is chosen, since it relies on the covariance structure, is easy to implement and has a relatively short running time. The methods will be discussed below.

3.2.2 ESACF

In this section, the basic intuition of the ESACF method will be given and the algorithm itself will be described. For a more thorough discussion of the method, see the original paper of Tiao and Tsay (1983).

Suppose we have data Y_t coming from an ARMA(p,q) process and assume for the moment that p is known while q is unknown. We try to fit an AR(p) model on this process by means of the regression:

$$Y_t = \sum_{i=1}^{p} \phi^{(0)}_{i;(p)} Y_{t-i} + e^{(0)}_{p;t},$$

where $\phi^{(0)}_{i;(p)}$ are the parameters of the AR(p) model. If $q > 0$, this AR(p) model is not adequate and the error estimates $\hat{e}^{(0)}_{p;t}$ are not white noise, but still contain some information about Y_t. We could then do the following second regression:

$$Y_t = \sum_{i=1}^{p} \phi^{(1)}_{i;(p)} Y_{t-i} + \phi^{(1)}_{1;(p)} \hat{e}^{(0)}_{p;t-1} + e^{(1)}_{p;t}.$$

If $q \geq 2$, the regression is still not fully adequate and the errors $e^{(1)}_{p;t}$ contain further information about Y_t. However, if $q \geq 1$, the model is consistent. In particular, when $q = 0$, the estimates $\hat{\phi}^{(1)}_{1;(p)}$ would converge in probability to 0 for increasing sample size. Similarly, for general q, in the regression

$$Y_t = \sum_{i=1}^{p} \phi^{(q+1)}_{i;(p)} Y_{t-i} + \sum_{j=1}^{q+1} \phi^{(q+1)}_{j;(p)} \hat{e}^{(j-1)}_{p;t-j} + e^{(q+1)}_{p;t}; \tag{27}$$

the parameters $\hat{\phi}^{(q+1)}_{q+1;(p)}$ converge in probability to 0 and the errors $e^{(q)}_{p;t-j}$ are thus uncorrelated with Y_t. The ESACF method uses this property by performing these regressions for every p and checking whether the correlation between Y_t and the errors is significantly different from 0.

Algorithm 1 shows the pseudocode of the ESACF method. Apart from the time serie data Y_t, we need to predefine the maximum orders $p_{\max}$ and $q_{\max}$ for p and q that are being considered for the model. As a first step we do the AR(p) regressions for all values of p. Instead of doing the auxiliary regressions of Equation (27), we can calculate the same parameters recursively using the formula shown in line 10, saving computation time. In this equation,

$\hat{\phi}^{(q-1)}_{0;(p)}$ is defined to be I for all p and q. To do the recursive calculations, we need higher order AR(p) regressions, which is why initially we need to do $p_{\max} + q_{\max} + 1$ regressions. Once the parameters are obtained, we check by means of statistical testing whether there is no significant autocorrelation left in the errors of the regression, denoted $W^{(q)}_{p;t}$ in the algorithm. We fill a matrix A with the results of these statistical tests. If there is indeed no autocorrelation, a 0 is filled in. If there is autocorrelation an 1 is filled in. Koreisha and Yoshimoto (1991) advise to use three standard deviations from 0 as a threshold.

Algorithm 1 ESACF Method

Require: a (multivariate) time series Y_t, some maximum orders $p_{\max}; q_{\max}$

1: **for** $p = 0$ **to** $p_{\max} + q_{\max}$ **do**
2: Perform regression on the equation $Y_t = \sum_{i=1}^{p} \hat{\phi}^{(0)}_{i;(p)} Y_{t-i} + e^{(0)}_{p;t}$ to obtain $f\hat{\phi}^{(0)}_{i;(p)} g_{i=1}^{p}$.
3: **end for**
4: **for** $q = 1$ **to** $q_{\max}$ **do**
5: **for** $p = 0$ **to** $p_{\max} + q_{\max} - q$ **do**
6: **if** $p = 0$ **then**
7: $\hat{\rho}_p(q)$ the correlation matrix $Corr[Y_t; Y_{t-q}]$
8: **else**
9: **for** $i = 1$ **to** p **do**
10: $\hat{\phi}^{(q)}_{i;(p)} = \hat{\phi}^{(q-1)}_{i;(p+1)} - \hat{\phi}^{(q-1)}_{p+1;(p+1)} \left[\hat{\phi}^{(q-1)}_{p;(p)}\right]^{-1} \hat{\phi}^{(q-1)}_{i-1;(p)}$.
11: **end for**
12: **end if**
13: **if** $p \geq p_{\max}$ **then**
14: $W^{(q)}_{p;t} = Y_t - \sum_{i=1}^{p} \hat{\phi}^{(q)}_{i;(p)} Y_{t-i}$
15: $\hat{\rho}_p(q)$ the correlation matrix $Corr\left[W^{(q)}_{p;t}; W^{(q)}_{p;t-q}\right]$
16: **if** every entry of $|\hat{\rho}_p(q)| \geq 3/\sqrt{n-p-q}$ **then**
17: $A(p;q) = 0$
18: **else**
19: $A(p;q) = 1$
20: **end if**
21: **end if**
22: **end for**
23: **end for**
24: $(p^*; q^* + 1)$ the indices of the vertex of the maximum upper right triangle in A with only 0-entries
25: **return** $p^*; q^*$

Theoretically, the matrix A should contain a triangular pattern of zeros such that the entries $(i; j) = 0$ for $i = p^*; j \geq q^*$ and for the pairs $(i; j)$ with $i = p^* + k$, $j \geq q^* + k$ for all positive integers k, where p^* and q^* are the real orders of the ARMA process. In words: the vertex of the triangle of zeros is $(p^*; q^*)$ and extends along the diagonal and the row to

higher orders. Due to randomness, the triangular pattern may not hold exactly. Tiao and Tsay do not provide a method to distill the order from a noise matrix A, other than human judgement. However, the Statistical Analysis System (SAS) recommends to choose the vertex of the maximum triangle with only 0 entries as the chosen order (SAS institute, 1999), which is the approach being followed in this research .

3.2.3 Eigenvalue method

The method of Cassar et al. (2010) is based on the observation that for a given VARMA($p;q$) time series, the minimum value of the Bayesian Information Criterion (BIC) can be expressed in terms of the eigenvalues of the data covariance matrix. The whole derivation of the original paper will not be repeated here, but the main idea is presented in this section.

For an M-dimensional VARMA($p;q$) time series y_t with N observations, de ne the data matrix $D_{p;q}$ as the N $M(p+q+1)$ matrix with the i-th row

$$D_{p;q}^{(i)} = [y_i^T; y_{i\ 1}^T; \dots; y_{i\ p}^T; e_{i\ 1}^T; \dots; e_{i\ q}^T]$$

for each i. Assume y_i and e_i are zero vectors when i 0. Moreover, de ne the $M(p+q+1)$ M parameter matrix $_{p;q}$ as

$$_{p;q} = [I_M; A_1^T; \dots; A_p^T; \ B_1^T; \dots; \ B_q^T]^T;$$

where A_j is the parameter matrix of the j-th AR-term and B_k the parameter matrix of the k-th MA term. Then, the N M-sized vector of errors $e = [e_1; \dots; e_N]^T$ can be expressed as $e = D_{p;q\ p;q}$. The errors e_i are usually unknown, but are suggested to be estimated by tting a high order AR model to the data. Lastly, de ne the data covariance matrix as $R_{p;q} = D_{p;q}^T D_{p;q}$.

Suppose that for a given order p and q, we wish to minimize the BIC de ned as

$$BIC = \ 2 \ln f_y(y_1; \dots; y_N) + C \ln N;$$

where C is the number of parameters and f_y the probability density function of the observations. Assume the errors are normally distributed, then $f_y(y_1; \dots; y_N) = f_e(e_1; \dots; e_N)$, which can be expressed as

$$f_e(e_1; \dots; e_N) = \frac{1}{(2\ )^{MN=2}} \det(Q_e)^{\ N=2} e^{\ \frac{1}{2}\ ^T D_{p;q}^T Q_e^{\ 1} D_{p;q\ p;q}};$$

with Q_e de ned as the covariance matrix of the errors. The covariance matrix Q_e which minimizes this density function and thus the BIC can be shown to be $^T_{p;q} R_{p;q\ p;q}$. The vector which minimizes the criterion is equal to the eigenvectors corresponding to the $2M$ smallest eigenvectors of $R_{p;q}$. Substituting this in leads, after some algebraic manipulations, to the following expression[1]:

$$\frac{4}{N} BIC = \ln \left(\prod_{i=1}^{2M} \ _i \right) N^{C=N}: \tag{28}$$

[1] The formula given in Cassar et al. (2010) contains a mistake, as it states the product over the eigenvalues runs from 1 to M, which is not in line with their derivation.

In the above equation $_i$ is the i-th smallest eigenvalue of the data covariance matrix $R_{p;q}$.

Algorithm 2 Eigenvalue Method

Require: a (multivariate) time series Y_t, some maximum orders $p_{\max}; q_{\max}$, an initial threshold value T, threshold value (both recommended to be 5)

Estimate the error terms e_t from a higher order AR model
for $p = 0$ **to** $p_{\max}$ **do**
  for $q = 0$ **to** $q_{\max}$ **do**
    Calculate the eigenvalues of the data covariance matrix $R_{p;q}$
    Calculate $BIC\ (p;q)$ using Equation (29)
  end for
end for
for $p = 1$ **to** $p_{\max}$ **do**
  CR$(p;0)$ $BIC\ (p\ \ 1;0)/BIC\ (p;0)$
  for $q = 1$ **to** $q_{\max}$ **do**
    RR$(0;q)$ $BIC\ (0;q\ \ 1)/BIC\ (0;q)$
    RR$(p;q)$ $BIC\ (p;q\ \ 1)/BIC\ (p;q)$
    CR$(p;q)$ $BIC\ (p\ \ 1;q)/BIC\ (p;q)$
    PM$(p;q)$ RR$(p;q)$ CR$(p;q)$
  end for
end for
repeat
  if All $(p;q)$-pairs are explored **then**
    T T 1
    All $(p;q)$-pairs become unexplored
  end if
  $p\ ;q$ $(p;q)$-pair with highest PM-value which is still unexplored
until PM$(p\ ;q\ )$ TPM$(p\ + 1;q\ )$ **and** PM$(p\ ;q\ )$ TPM$(p\ ;q\ + 1)$
if $p\ = 1$ **and** CR(1,0) **then**
  p 0
end if
if $q\ = 1$ **and** RR(0,1) **then**
  q 0
end if
return $p\ ;q$

Above derivation provides an approach based on the BIC to come to an optimal order p and q without estimating the parameters and covariance matrix explicitely for each value p and q, but by calculating the eigenvalues of the matrix R for each p and q until some prede ned upper bounds $p_{\max}$ and $q_{\max}$. Instead of using Equation 28 to calculate the BIC, its exponent is calculated, dropping some multiplicative constants:

$$BIC = \left(\prod_{i=1}^{2NM} {}_i \right) N^{M^2(p+q)/N}. \tag{29}$$

Moreover, Cassar et al. (2010) noticed that rather than choosing the order p and q by min-

imizing the BIC criterion, in practice better results are obtained by choosing the order in such a way that there is a big drop in the BIC value, compared to lower orders of p and q. For this purpose, a column ratio (CR) matrix is constructed, where its $(p;q)$ element is the ratio $BIC\ (p\ \ 1;q)/BIC\ (p;q)$. Similarly, the row ratio (RR) matrix is constructed with element $(p;q)$ equal to $BIC\ (p;q\ \ 1)/BIC\ (p;q)$. The elementwise product of these matrices, resulting in the product matrix PM, are used to nd the largest drop in the BIC value. See Algorithm 2 for the exact procedure.

This procedure is only able to identify VARMA orders $p;q\ \ 1$. To check whether either order would need to equal 0, Camilleri (2012, p56) notes that if $p = 0$, the rst row of the CR matrix is expected to have values close to 1 and if $q = 0$, the rst column of the RR matrix is expected to have values close to 1. In contrast, if $p = 1$, the rst entry of the rst row of the CR matrix will have a high value. Likewise, the rst entry of the rst column of the RR matrix is high for $q = 1$. Therefore, if the above procedure nds that $p = 1$ or $q = 1$, the rst entry of the CR respectively RR matrix is checked. If this value is lower than some parameter , the order is chosen to be 0, otherwise the order is chosen to be 1. From experiments conducted during this nal project, it was concluded that = 5 leads to the most correct classi cations.

From the derivation of the eigenvalue method, it gets clear the method is not guaranteed to work for MS VARMA models or MS VECMs. The autocovariance of these models are equal to the autocovariance of some VARMA model, however this is generally not the case for the density function of the observations. It is therefore not certain that the order $(p\ ;q\ )$ of the VARMA model with equal autocovariance structure to the MS VARMA model, is also the order which would minimize the BIC criterion.

Another theoretical problem is that the data covariance matrix does not only consist of terms $y_t y_{t\ \ h}^T$, but also of the terms $y_t e_{t\ \ h}^T$ and $e_t e_{t\ \ h}^T$. The errors from a Markov Switching process are estimated by means of a non-Markov Switching process, which may cause large errors in the estimates of the errors and thus in the data covariance matrix. Moreover, the error covariance structure may not represent that of a VARMA process with autocovariance structure $E[y_t y_{t\ \ h}^T]$. However, also for non-Markov Switching VARMA models the error terms are estimated in a ruther crude manner and one would therefore expect the method to be rather robust to misestimations of the errors.

3.3 Synthetic experiments

In this section, several synthetic time series are generated and analyzed to determine the tightness of the methods introduced in the previous sections. First, in Section 3.3.1, to examine the strength of the methods, several stationary time series are randomly generated under di erent conditions and Theorem 11 is used to obtain a lower bound on the number of hidden Markov states. Afterwards, in Section 3.3.2, an experiment is performed where time series are di erenced to compare the lower bounds of Theorem 11 and 13.

3.3.1 Non-differenced time series

Theorem 11 has only been used very sparsely in the existing literature and, therefore, not much is known about the tightness of the lower bound in practice. In the paper of Cavicchioli (2014) which introduced the method, a synthetic experiment was performed. However, these time series seemed to violate the assumption that the Markov state S_t should be uncorrelated with the observations Y_t. In order to have a rst idea of the performance of the autocovariance method, several synthetic time series are randomly generated. Of these time series, the number of hidden Markov states is estimated using the autocovariance method, assuming the time series is uncorrelated with the hidden variable, so that the formula of Theorem 11 can be applied, which is stricter than the formulas of Corollary 8, Theorem 9 and Theorem 14. The goal of the synthetic experiments is to investigate if the method indeed results in a lower bound and which conditions lead to tighter lower bounds. For this purpose, stationary time series are generated. All time series are MS VARMA(1,0) time series, where all elements of the parameter matrix A_1 are outcomes of a uniformly distributed random variable distributed on the interval $(-1;1)$. If the resulting time series was non-stationary, another time series was generated with random parameters.

Various scenarios were investigated, where the time series were of di erent dimension $M \in \{1;\dots;4\}$, di erent number of hidden Markov states $K \in \{2;\dots;5\}$, varying amount of variance, varying amount of data available and two varying types of transition matrices. The variances of the random noise were set to be equal to σI_M for all hidden Markov states, where $\sigma \in \{0.1; 1; 5\}$ and I_M is the identity matrix of dimension M. The noise is generated as i.i.d. (multivariate) Gaussian random variables with mean 0. The time series length $N \in \{1000; 10{,}000\}$. The probability transition matrix P is either a uniform matrix, that is $P_{ij} = \frac{1}{K}$ for all $i; j$, resulting in almost constant switching, or P is a random matrix generated in such a manner that there is relatively few switching. For each $i = 1;\dots;K$, the i-th row of the matrix P is generated by generating a row vector of i.i.d. variables generated uniformly from the interval $(0; 1)$. This vector is standardized so that the sum of the elements of the vector equals 0.2. Lastly 0.8 is added to the i-th element of the i-th row vector. For all combinations, 100 time series were generated of length $N + 50$, where the rst 50 observations were discarded to decrease the in uence of the starting up of the time series generation. With these time series, the lower bounds on the number of hidden Markov states were estimated. The autoregressive and moving average orders $p = 1$ and $q = 0$ of the MS VARMA(p,q) model were assumed known, while the other parameters were assumed unknown. The orders $p\,;q$ of the VARMA$(p\,;q\,)$ model were estimated using the eigenvalue method with maximum orders considered 9, ignoring that the time series were Markov Switching time series. The

resulting values p ; q were used to obtain a lower bound on the number of hidden Markov states $\hat{K}$ by the formula $\hat{K} = \max\{q + 1; p\}$.

To examine the effects of these varying parameters in isolation, the results are aggregated for all experiments with certain values for K and for the parameters under consideration. These are shown in Tables 1 to 4. For instance in Table 1, the point estimate of the probability that a randomly generated MS VARMA(1,0) with K hidden Markov states coincides with the obtained lower bound when a sample has sample size N is available, where the other parameters vary. In parentheses, the Clopper-Pearson confidence interval is given (Clopper and Pearson, 1934). Table 2 to 4 have similar structure.

	N = 1000		N = 10000	
K = 2	0.736	(0.718,0.753)	0.800	(0.783,0.815)
K = 3	0.191	(0.176,0.207)	0.268	(0.250,0.286)
K = 4	0.031	(0.025,0.039)	0.077	(0.067,0.088)
K = 5	0.010	(0.007,0.015)	0.025	(0.019,0.032)

Table 1: Correctly classified orders with varying sample size N

	= 0.1		= 1		= 5	
K = 2	0.771	(0.751,0.792)	0.748	(0.727,0.769)	0.784	(0.763,0.804)
K = 3	0.237	(0.217,0.258)	0.218	(0.199,0.239)	0.233	(0.213,0.254)
K = 4	0.056	(0.046,0.069)	0.053	(0.043,0.065)	0.053	(0.043,0.065)
K = 5	0.020	(0.014,0.028)	0.018	(0.013,0.026)	0.015	(0.010.0.022)

Table 2: Correctly classified orders with varying standard deviation

The aspect which is most striking is that there is a strong relation between the number of hidden Markov states K and the chance the lower bound coincides with K. For $K = 2$, this amount is decent, while for $K = 4$ and $K = 5$ the number of hidden Markov states seldomly coincides with the lower bound in these experiments. For ease, if K coincides with the lower bound, this will be called a correct classification, although strictly speaking, the lower bound is also correct if any value less or equal than K is obtained. An explanation for the relatively high rate of correct classifications for $K = 2$ is that the eigenvalue method was originally not designed to classify orders p ; $q < 1$. As mentioned in Section 3.2.3, the author of this method proposed an extension of the method to be able to distinguish between the situation where p or q equals 0 and where either or both equals 1, but this method did not seem to perform well in these experiments. Part of the experiment was repeated, where if either p or q was found to be equal to 1, the ESCAF was applied to determine whether p or q equals 0 or 1 according to that method. These led to a very minor change in the results, but to a sizeable increase in runtime.

Since if $q = 1$, $\hat{K}$ is at least 2, the number of hidden Markov states was rarely estimated to be 1, even for non-MS time series models. Since one does not expect a lot of overestimations of the order, based on the theoretical results, the order is often correctly classified for $K = 2$.

A first conclusion of this experiment is that the lower bound generally becomes less tight when the number of hidden Markov states grows larger.

Examining the effect of changing the sample size in Table 1, it becomes apparent that there is a significant positive effect on the percentage of correct classifications for larger sample sizes. Curiously enough, this effect does not seem to be present for the magnitude of the noise . An explanation would be that the randomness induced by the Markov state transitions is more important than the randomness of the white noise. Another explanation is that a sample size of $N = 1000$ is already large enough for the effect of the random noise to cancel out, while this is not the case for the sequences of the hidden Markov states. Either way, one expects the hidden Markov state transitions to be important. This is in line with the results of Table 3, which shows that if the probability transition matrix P is uniform, one is almost never able to correctly classify the number of hidden Markov states K if K 3. In 98% of the generated time series with uniform matrix P, one obtains an estimate $\hat{K}$ 2, while if the time series stays in the same state for a longer time period, this percentage is only 50%. If the MS VARMA time series switches less often from hidden Markov state, there are longer stretches of the time series which evolve in distinct ways, which are easier to distinguish than many short stretches. This provides a plausible explanation for the strong effect of the structure of probability transition matrix P on the ability to classify. Interestingly enough, the algorithm also overestimates the number of hidden Markov states more often if P is nonuniform, for $K = 2$ in 21% of the cases. This may be explained by the fact that if the time series stays in a specific state for a relatively long period of time, that time series may evolve in a nonstationary manner for that period. Time series in a specific mode may be nonstationary, which other modes can compensate. However, for short periods of time, the observations may grow exponentially. Since the eigenvalue method cannot handle nonstationary time series well, the results obtained may then be less reliable.

Table 3: Correctly classified orders with varying probability transition matrix P

	Uniform P		Nonuniform P	
K = 2	0.871	(0.858,0.884)	0.664	(0.645,0.683)
K = 3	0.014	(0.010,0.020)	0.444	(0.425,0.464)
K = 4	0.000	(0.000,0.002)	0.108	(0.096,0.121)
K = 5	0.000	(0.000,0.002)	0.035	(0.028,0.043)

Table 4: Correctly classified orders with varying observation dimension M

	M = 1		M = 2		M = 3		M = 4	
K = 2	0.566	(0.538,0.594)	0.790	(0.767,0.813)	0.862	(0.842,0.881)	0.853	(0.833,0.873)
K = 3	0.210	(0.183,0.229)	0.187	(0.166,0.210)	0.248	(0.225,0.274)	0.277	(0.252,0.303)
K = 4	0.016	(0.010,0.025)	0.048	(0.038,0.062)	0.065	(0.052,0.080)	0.087	(0.072,0.104)
K = 5	0.000	(0.000,0.003)	0.013	(0.008,0.021)	0.028	(0.020,0.038)	0.031	(0.023,0.042)

A last observation is that the performance of the order classification improves significantly

with increasing dimension of the observations, when comparing for instance the con dence intervals for $M = 2$ and $M = 4$. This may be explained by the fact that the di erence (in matrix norm) between the parameter matrices A_k can be bigger for higher dimensions, which increases the distinctness between the hidden Markov states.

Above analysis gives a rst idea which hyperparameters lead to favourable order classi cations. However, even for a given set of these hyperparameters $N; M; K;$ and P, the variability in the lower bound obtained may be large. As a small experiment, 400 time series were randomly generated in the same way as described above, with $M = 4; K = 5; = 0:1; N = 50000$ and a nonuniform probability transition matrix P, as above analysis showed that these parameters would lead to favourable order classi cations. In Table 5, the classi cations of these time series are listed.

Table 5: Classi cations of 400 randomly generated time series with 5 Markov States

$\hat{K}$	Amount classi ed
1	13
2	11
3	169
4	130
5	60
6	8
7-9	9

We can see that of these 400 instances, the lower bound is now tight in more cases than for the above tabulated cases with $K = 5$, although this still remains relatively rare. Note that if the formula of Theorem 9 was used, only lower bounds of 1, 2 and in very rare cases 3 would be obtained, showing the weakness of the bound.

We would expect that the markedness of the di erent transition matrices A_k matters for the tightness of the bound. However, when analysing the correlation between the distance of the transition matrices $\sum_{i;j:i\neq j} jjA_i - A_jjj_2$ and the estimated number of Markov states $\hat{K}$, we see that the correlation is almost zero, with a point estimate of 0.018. An explanation for this phenomenon is that randomly generated matrices which are further apart from each other in norm have a higher chance of leading to non-stationary modes, which the algorithm cannot handle well. Consider for example the time series with again $M = 4; K = 5; = 0:1; N = 50000; P$ nonuniform and where A_i is a matrix with each entry $i + 5c$ equals 0.9 and the other entries equal 0, where $c = 0; 1; 2; : : :$ and matrix entries are counted from left to right and top to bottom, such that the rst row contains entries 1 to M and the second $M + 1$ until $2M$. The norm di erence between these parameter matrices is quite small, 1.45-1.69, compared to randomly generated matrices where it is more than 2 for most random matrices. However, when 100 time series were generated from these matrices A_i, 73 were classi ed correctly as having 5 hidden Markov states. If all positive entries of these matrices A_i are set equal to 1.8 instead of 0.9, the time series is still stationary, but once in a while there are short periods where the magnitude of the series explodes. The result is that the estimated number of hidden Markov states is a highly uctuating variable, which provides almost no information on the true number of hidden Markov states. If instead all positive entries of the matrices A_i are set to 0.3 instead of 0.9, the estimated order is virtually always $\hat{K} = 2$, since the di erence between the transition matrices becomes too small to detect.

Table 6: Classi cations of 100 randomly generated time series with 3 Markov States

$\hat{K}$	Amount classi ed using Y_t	Amount classi ed using Y_t
1	0	5
2	22	95
3	59	0
4	15	0
5	4	0

3.3.2 Differenced time series

The conclusions of the above experiments are not a positive sign for the strength of the lower bound of Theorem 13, the lower bound for the di erenced time series. One important conclusion was that it becomes increasingly harder to nd the correct number of hidden Markov states when the true number of hidden Markov states increases. Since for the di erenced time series, both the Markov state at time t, as well as at time $t \quad 1$ a ects the time series, the Markov Chain which corresponds to this di erenced time series has K^2 hidden Markov states instead of K which need to be recognized. Therefore, ideally one needs to estimate large lags p and q in the VARMA order estimation method. However, it is not necessary to be able to recognize all K^2 hidden Markov states. If at least $(K \quad 1)^2 + 1$ Markov states are recognized, it can be concluded that at least K hidden Markov states are present in the non-di erenced time series.

Another potential hazard is that the di erenced time series will often stay slightly shorter in the same Markov state $(S_t; S_{t \ 1})$, since a switch at time t in the Markov process fS_tg, will imply switches at t and $t + 1$ in the Markov process $fS_t; S_{t \ 1}g$. If the average time until a transition is large enough for the non-di erenced time series, there are still relatively long time periods in which the Markov Chain $(S_t; S_{t \ 1})$ remains in the same state $(k; k)$, but this Markov Chain will generally be only one time period in the state $(i; j)$ for $i \neq j$, which may be di cult to detect for the order estimation algorithm.

To examine the use of Theorem 13, again synthetic time series are generated, in the same way as the previous experiments, with $= 0{:}1$, nonuniform probability transition matrix P, $N = 50000$ and $K = 3$ and $M = 4$. Since that theorem does not require that the non-di erenced time series is non-stationary, even though that is where the theorem may be most useful, stationary time series are generated, since the bounds can then be compared with the bounds obtained by using Theorem 11. The results of this experiment are shown in Table 6. While the Markov states are estimated correctly in the majority of the cases using the non-di erenced time series, the number of Markov states is underestimated always by the order selection method, when using the di erenced time series. The orders $p ; q$ found for the di erenced time series are often even slightly lower than the orders found when using the non-di erenced time series, even though higher orders are necessary to obtain the same bound. This experiment therefore is a con rmation that it is very di cult to detect the K^2 di erent Markov states of which most only stay in their Markov state for one time period.

Chapter 4

Penalty Function Methods

In this chapter, the penalty function approach for estimating the number of hidden Markov states will be discussed. This method complements the lower bounds obtained by the auto-covariance method of the previous chapter, since for some penalty functions it provides an asymptotic upper bound on the number of hidden Markov states. In Section 4.1, penalty functions are introduced and the conditions for which they provide an asymptotic upper bound are discussed. In Section 4.2, likelihood optimization and consistency results are brie y discussed. For the penalty functions considered in this thesis, likelihood optimization is the type of optimization needed for the penalty function approach. In Section 4.3, the EM algorithm is discussed, which is the likelihood optimization method used in this thesis.

4.1 Penalty functions

Penalty functions are a widely used method for model selection. The idea behind the method is that a good model must optimize some objective function with a relatively low amount of parameters, to reduce the risk of over tting. The most used objective function is the likelihood function of the available observations, although the underlying idea may be based on other objective functions (Stoica and Selen, 2004). Notable examples include the KL divergence between the real probability distribution function (PDF) $p(x)$ and the tted PDF $\hat{p}(x)$,

$$D_{KL} = \int_{-1}^{1} p(x) \ln \frac{p(x)}{\hat{p}(x)} \, dx;$$

and the maximum a posteriori (MAP) objective function given some prior distribution $g(x)$,

$$\max_{\hat{p}} \hat{p}(x)g(x):$$

There exist ample criteria which do not use the likelihood function, but are for example based on the predictive least squared principles (Wei et al., 1992). However, in this chapter, only criteria using the likelihood function are discussed. This set consists of functions that for this context are of the form $-2L_K + \ _K$, where L_K is the log-likelihood function for the model with optimized parameters and K hidden Markov states and $\ _K$ is the penalty for K hidden Markov states, where it holds that $\ _{K+1} \ \ \ _K$. Penalty functions of these forms are

minimized by maximizing the log likelihood function over different values of K. This can be done by calculating the penalty function for all K from 1 to some predefined relatively low value or until the penalty function starts increasing.

The choices for $_K$ in the literature are varied. The most well known example is the Akaike Information Criterion, which is originally based on the KL divergence, where $_K$ equals to two times the number of parameters in the model, d_K, with K hidden Markov states, $_K = 2d_K$. For small sample sizes N, the AIC_c (AIC with correction) is proposed, where $_K = 2d_K \frac{N}{N \; d_K \; 1}$. The Generalized Information Criterion is a generalization of the AIC, with $_K = vd_K$, where v is a parameter which has been shown to lead to the best results for $v \; 2$ [2; 6]. If $v = 2$, one obtains the AIC. Another widely used criterion is the Bayesian Information Criterion, also called Schwarz Information Criterion, which is based on the MAP objective function, with $_K = \ln(N)d_K$. Other information criteria include Takeuchi's Information Criterion, the Information Complexity criterion, Hannah-Quinn Information Criterion and many others (Seghouane and Amari, 2007), (Bozdogan, 2000).

A last information criterion which will be discussed here shortly is the Markov Switching Criterion (MSC), since it was designed by Smith et al. (2006) for Markov Switching models. This criterion is, just as the AIC, derived from the KL divergence. For any number of hidden Markov states considered, the estimated parameters $\hat{}$ need to be calculated from the available data. The MSC then leads to an optimal choice K and parameter set , which includes the probability transition matrix P and the estimated stationary probability distribution of the hidden Markov states $_i$, which is the eigenvector of P such that $P \quad = \quad$. The formula for the penalty also requires for $k = 1; : : : ; K$, the quantity $\hat{f}_k = \sum_{t=1}^{N} \hat{P}(S_t = k)$, which is the sum of the estimated probabilities that the hidden Markov state is state k over all observations. The penalty is

$$_K = \sum_{k=1}^{K} \frac{\hat{f}_k(\hat{f}_k + \; _k M)}{_k \hat{f}_k \quad _k M \quad 2};$$

with $_k = E\left[\frac{_k}{\hat{}_k}\right]$ and $_k = E\left[\frac{_k}{\hat{}_k}^2\right]$. However, these latter two quantities are difficult to calculate. In practice, a lower bound of 1 can be used for $_k$. An approximation for $_k$ that performed well in Monte Carlo simulations is K, according to the founders of the method.

Under a set of technical regularity conditions on the data generating process, Ryden (1995) has shown that information criteria of the form $2L_K + \; _K$ in the limit when the sample size $N \; ! \; 1$ do not underestimate the true number of hidden Markov States if the following two conditions hold for the penalty $_K$:

1. $_{K+1} \quad _K$ for every sample size N,
2. $\limsup_{N \; ! \; 1} \frac{K}{N} = 0$ with probability 1.

It is straightforward to check that these conditions holds for the AIC, the small sample corrected AIC, the GIC and the BIC. However, for the MSC the second condition does not

hold. A small informal example will illustrate this: assume that $\hat{P}(S_t = k) = \frac{1}{K}$ is for every t and k. Then

$$\kappa = \sum_{k=1}^{K} \frac{\frac{N}{K}\left(\frac{N}{K} + c_{1;k}\right)}{c_{2;k}\frac{N}{K} + c_{3;k}} = \sum_{k=1}^{K} \frac{\frac{N}{K} + c_{1;k}}{c_{2;k} + c_{3;k}\frac{K}{N}},$$

where $c_{1;k}; c_{2;k}; c_{3;k}$ are constants. In this case, the denominator of $\lim_{N \to \infty} \frac{K}{N}$ will converge to 0, while the numerator will converge to 1, so that the fraction will grow unbounded in the limit. We can, therefore, conclude that the upper bound property of Ryden (1995) does not hold for the MSC and that the MSC-optimizing choice for K may only be used as an estimate.

4.2 Likelihood optimization

Using above information criteria requires the use of the maximum likelihood estimator (MLE). However, the MLE is not always consistent. Under a set of regularity conditions, which is a subset of the regularity conditions which causes the information criteria to be an upper bound on the number of hidden Markov states, Leroux (1992) has shown that the MLE is consistent for hidden Markov models. It is outside the scope of this thesis report to discuss all those criteria rigorously, but one criterion is highlighted, since it results in an important case where maximum likelihood estimation fails.

Let $\{S_t; Y_t\}$ be an M-dimensional HMM, where each Markov state k corresponds to a distribution $f(Y_t; \theta_k)$ for Y_t and where θ_k is an element of some parameter set Θ. Then there must exist a continuous function h such that $f(Y_t; \theta) \leq h(Y_t)$ for all θ and it must hold that $\int_{\mathbb{R}^M} p(Y_t; \theta_K) |\log h(Y_t)| \, \mu_M(dY_t) < \infty$. In this equation, θ_K are the K real parameter sets and $p(\cdot)$ is the corresponding real probability density function. This condition is violated if one tries to maximize a mixture of normally distributed variables, since one can obtain an arbitrarily high likelihood by setting the mean of one component equal to one of the data points and letting the variance decrease to 0. Therefore, the maximum likelihood estimator does not exist if the variables are normally distributed and no conditions are set on the variance and it can thus naturally also not be consistent. However, if the variance is fixed, all conditions are fulfilled and the MLE is consistent.

The maximization of the log likelihood function is not straightforward due to the presence of the hidden Markov states. The MLE can in most of these cases not be maximized algebraically. Moreover, due to the existence of many local maxima and slow convergence, also many algorithms have difficulty finding the global maximum. One popular method to maximize the likelihood function is the Expectation Maximization (EM) algorithm, which procedure will be explained in the next section and which is used in this thesis. Other methods include general purpose optimization techniques, such as the conjugate gradient method or hill climbing algorithms. The EM algorithm is relatively fast in practice, does not require the computation of the Hessian matrix and depending on the distribution the updates during its iterations can be computed analytically. However, it is only guaranteed to converge to a local maximum and there is no guaranteed convergence rate. In practice one tries several start conditions to increase the probability of finding a good solution (Ryden, 1995).

4.3 Expectation Maximization algorithm

Let the M-dimensional process fY_tg be of the form

$$Y_t = \quad _{s_t} + \sum_{i=1}^{p} a_{s_t}^{(i)} Y_{t \quad i} + \sum_{j=1}^{q} b_{s_t}^{(j)} e_{t \quad j} + e_t \quad e_t \quad N(0; \quad _{s_t}) \tag{30}$$

and assume we have observations $1 \quad p; 2 \quad p; : : : ; N$, such that we have N observations where the p lags are available. We wish to nd the parameters that maximize the log-likelihood function of Y_t of which we have a series of N observations. However, this is di cult since we do not know the values of the hidden variables S_t. To simplify matters, we instead try to maximize the expectation of the log-likelihood function. This can be done by conditioning on the values of the hidden variables S_t and weighing the log-likelihoods $\ln P(Y_1; : : : ; Y_N j S_1; : : : ; S_N)$ with the probabilities of the values of the hidden Markov variables $P(S_1; : : : ; S_N)$. However, in order to do this we need the parameters, which is what we want to estimate. The idea behind the Expectation-Maximization algorithm is that we iteratively nd estimates for the parameters and do the aforementioned weighting with the probability of the hidden variables given the parameters found in the previous iteration. New parameters are then chosen such that the expected loglikelihood function is optimized. Note that the errors e_t are also unknown.

Although the use of the EM algorithm for this purpose is not new, the details of the algorithm for this speci c purpose were not found in the literature. For the sake of completeness, these will be provided below. For a general discussion on the EM algorithm, see for instance the text book of Bishop (2006).

4.3.1 Expectation

The expectation of the log-likelihood function $Q(\ ^{\text{old}}; \ )$, which is a function of the previous parameters $\ ^{\text{old}}$ and the new parameters $\ $, can be expressed as:

$$\begin{aligned} Q(\ ^{\text{old}}; \ ) &= \mathrm{E}^{\text{old}}[\ln P(Y_1; : : : ; Y_N; S_1; : : : ; S_N j \ )] \\ &= \sum_{S_1=1}^{K} \cdots \sum_{S_N=1}^{K} P(S_1; : : : ; S_N j Y_1; : : : ; Y_N; \ ^{\text{old}}) \sum_{t=1}^{N} \ln P(Y_t; S_t j Y_1; : : : ; Y_{t \quad 1}; S_1; : : : ; S_{t \quad 1}; \ ) \\ &= \sum_{t=1}^{N} \sum_{S_{t \quad 1}=1}^{K} \sum_{S_t=1}^{K} P(S_{t \quad 1}; S_t j Y_1; : : : ; Y_N; \ ^{\text{old}}) \ln P(Y_t; S_t j Y_1; : : : ; Y_{t \quad 1}; S_{t \quad 1}; \ ); \end{aligned}$$

where we de ne $P(S_0; S_1 j Y_1; : : : ; Y_N; \ ^{\text{old}})$ as $\frac{1}{K} P(S_1 j Y_1; : : : ; Y_N; \ ^{\text{old}})$. For more compact notation the pre-sample observations $Y_{1 \quad p}; : : : ; Y_0$ are considered part of both $\ $ and $\ ^{\text{old}}$. The last equation follows since Y_t is independent of $S_1; : : : ; S_{t \quad 1}$ given S_t and S_t only depends on $S_{t \quad 1}$ since fS_tg is a Markov Process. Therefore, we can weigh the log-probability with the joint conditional probability of $S_{t \quad 1}$ and S_t, $P(S_{t \quad 1}; S_t j Y_1; : : : ; Y_N; \ ^{\text{old}})$. The probability $P(Y_t; S_t j Y_1; : : : ; Y_{t \quad 1}; S_{t \quad 1}; \ )$ can be decomposed as

$$P(Y_t; S_t j Y_1; : : : ; Y_{t \quad 1}; S_{t \quad 1}; \ ) = P(S_t j S_{t \quad 1}; \ ) P(Y_t j Y_1; : : : ; Y_{t \quad 1}; S_t; \ );$$

which is the product of two probabilities that can easily be calculated, since we know that

$$Y_t | Y_1, \ldots, Y_{t-1}, S_t, \theta^{old} \sim N\left(\mu_{S_t}^{old} + \sum_{i=1}^{p} a_{S_t}^{(i);old} Y_{t-i} + \sum_{j=1}^{q} b_{S_t}^{(j);old} e_{t-j}, \sigma_{S_t}^{old}\right)$$

and the transition probability $P(S_t | S_{t-1}, \theta^{old})$ is one of the parameters in the parameter set θ^{old}.

The probabilities $P(S_{t-1}, S_t | Y_1, \ldots, Y_N, \theta^{old})$ can be calculated using the forward-backward algorithm. In this algorithm the probability $P(S_{t-1} = k, S_t = l, Y_1, \ldots, Y_N | \theta^{old})$ is calculated by the decomposition

$$P(S_{t-1} = k, S_t = l, Y_1, \ldots, Y_N | \theta^{old}) = P(Y_{t+1}, \ldots, Y_N | S_{t-1} = k, S_t = l, Y_1, \ldots, Y_t, \theta^{old}) \cdot P(S_{t-1} = k, S_t = l, Y_1, \ldots, Y_t | \theta^{old}).$$

The desired probabilities $P(S_{t-1}, S_t | Y_1, \ldots, Y_N, \theta^{old})$ can then be calculated by normalizing the joint probabilities such that $\sum_{k=1}^{K} \sum_{l=1}^{K} P(S_{t-1} = k, S_t = l | Y_1, \ldots, Y_N, \theta^{old}) = 1$, since $P(S_{t-1} = k, S_t = l | Y_1, \ldots, Y_N, \theta^{old}) / P(S_{t-1} = k, S_t = l, Y_1, \ldots, Y_N | \theta^{old})$.

The probability $P(S_{t-1} = k, S_t = l, Y_1, \ldots, Y_t | \theta^{old})$, which we will denote by $\alpha_t(k, l)$, is calculated in a recursive manner from $t = 1$ to N (the forward pass) by conditioning on S_{t-2} and decomposing the probability:

$$\begin{aligned}
\alpha_t(k, l) &= \sum_{S_{t-2}=1}^{K} P(S_{t-2}, S_{t-1} = k, S_t = l, Y_1, \ldots, Y_t | \theta^{old}) \\
&= \sum_{S_{t-2}=1}^{K} P(Y_t | S_t = l, S_{t-1} = k, S_{t-2}, Y_1, \ldots, Y_{t-1}, \theta^{old}) P(S_t = l | S_{t-1} = k, S_{t-2}, Y_1, \ldots, Y_{t-1}, \theta^{old}) \\
&\quad P(S_{t-2}, S_{t-1} = k, Y_1, \ldots, Y_{t-1} | \theta^{old}) \\
&= \sum_{S_{t-2}=1}^{K} P(Y_t | S_t = l, Y_1, \ldots, Y_{t-1}, \theta^{old}) P(S_t = l | S_{t-1} = k, \theta^{old}) \alpha_{t-1}(S_{t-2}, k).
\end{aligned}$$

The first two terms are probabilities that can be calculated easily and $\alpha_{t-1}(S_{t-2}, k)$ was calculated in a previous iteration. The probability $\alpha_1(k, l)$ is defined as $\frac{1}{K} \pi(k | \theta^{old}) P(Y_1 | S_1 = l, \theta^{old})$, where $\pi(k)$ is the stationary probability to be in state k.

In the backward pass we calculate $P(Y_{t+1}, \ldots, Y_N | S_t = l, Y_1, \ldots, Y_t, \theta^{old})$, denoted by $\beta_{t+1}(l)$

recursively from $t = N$ to 1. This time we condition on S_{t+1}:

$$
\begin{aligned}
{}_{t+1}(l) &= \sum_{S_{t+1}=1}^{K} P(Y_{t+1}, \ldots, Y_N, S_{t+1} \mid S_t = l, Y_1, \ldots, Y_t, \theta^{\text{old}}) \\
&= \sum_{S_{t+1}=1}^{K} P(Y_{t+2}, \ldots, Y_N \mid S_{t+1}, S_t = l, Y_1, \ldots, Y_{t+1}, \theta^{\text{old}}) P(Y_{t+1} \mid S_{t+1}, S_t = l, Y_1, \ldots, Y_t, \theta^{\text{old}}) \\
&\quad P(S_{t+1} \mid S_t = l, Y_1, \ldots, Y_t, \theta^{\text{old}}) \\
&= \sum_{S_{t+1}=1}^{K} {}_{t+2}(S_{t+1}) P(Y_{t+1} \mid S_{t+1}, Y_1, \ldots, Y_t, \theta^{\text{old}}) P(S_{t+1} \mid S_t = l, \theta^{\text{old}});
\end{aligned}
$$

where the last two terms are again probabilities that can be calculated straightforwardly. The probability ${}_{N+1}(l)$ is de ned as 1 for every value l.

With above calculations we can calculate the value of the function Q for given values of θ^{old} and θ.

4.3.2 Maximization

In the maximization step of the EM-algorithm we need to nd the parameters $\theta = \{\pi_k, a_k^{(i)}, b_k^{(j)}, p_{kl}, \mu_k\}_{i=1,\ldots,p;\, j=1,\ldots,q;\, k,l=1,\ldots,K}$, where $p_{kl} = P(S_t = l \mid S_{t-1} = k)$ which maximize the expectation of the log-likelihood function Q by taking the derivatives of $Q(\theta^{\text{old}}, \theta)$ with respect to the di erent parameters and setting these to 0.

The rst derivative that we are taking, is that with respect to p_{kl}. Since an extra condition on p_{kl} is that $\sum_{j=1}^{K} p_{kj} = 1$, the Langrangian multiplier method is used:

$$
\begin{aligned}
&\frac{Q(\theta^{\text{old}}, \theta)}{p_{kl}} = \\
&\frac{1}{p_{kl}} \sum_{t=1}^{N-1} \sum_{S_1=1}^{K} \cdots \sum_{S_N=1}^{K} P(S_{t-1}, S_t \mid Y_1, \ldots, Y_N, \theta^{\text{old}}) (\ln p_{S_{t-1}, S_t} + \ln P(Y_t \mid Y_1, \ldots, Y_{t-1}, S_t, \theta)) \quad \left(\sum_{j=1}^{K} p_{kj} - 1\right) \\
&= \sum_{t=1}^{N-1} P(S_{t-1} = k, S_t = l \mid Y_1, \ldots, Y_N, \theta^{\text{old}}) \frac{1}{p_{kl}} \quad = 0, \\
&p_{kl} = \sum_{t=1}^{N-1} P(S_{t-1} = k, S_t = l \mid Y_1, \ldots, Y_N, \theta^{\text{old}}).
\end{aligned}
$$

Summing both sides over l, we obtain that $= \sum_{m=1}^{p}\sum_{t=1}^{p} P(S_{t-1}=k, S_t=m|Y_1,\ldots,Y_N,\theta^{old})$. Therefore, we obtain

$$p_{kl} = \frac{\sum_{t=1}^{p} P(S_{t-1}=k, S_t=l|Y_1,\ldots,Y_N,\theta^{old})}{\sum_{m=1}^{p}\sum_{t=1}^{p} P(S_{t-1}=k, S_t=m|Y_1,\ldots,Y_N,\theta^{old})};$$

which is a function of known quantities. We now proceed with the other derivatives, such as with respect to θ_k. Denote the probability

$$\sum_{S_{t-1}=1}^{K} P(S_{t-1}, S_t|Y_1,\ldots,Y_N,\theta^{old}) = P(S_t|Y_1,\ldots,Y_N,\theta^{old})$$

by $\gamma_t(S_t)$ for brevity. We then obtain

$$\frac{Q(\theta^{old};\theta)}{\theta_k} = \frac{\partial}{\partial \theta_k}\sum_{t=1}^{N}\sum_{S_{t-1}=1}^{K}\cdots\sum_{S_t=1}^{K} P(S_{t-1}, S_t|Y_1,\ldots,Y_N,\theta^{old})(\ln p_{s_{t-1},s_t} + \ln P(Y_t|Y_1,\ldots,Y_{t-1},S_t,\theta))$$

$$= \frac{\partial}{\partial \theta_k}\sum_{t=1}^{N}\gamma_t(k)\ln P(Y_t|Y_1,\ldots,Y_{t-1},S_t=k,\theta): \qquad (31)$$

The above formula applies also to the other partial derivatives, except the transition probabilities, where $\frac{\partial}{\partial \theta_k}$ is replaced by the parameter under investigation. It is useful to rewrite Equation (30) in matrix notation to be able to obtain the derivatives in one go. De ne the vector X_t of size $M(p+q)+1$ as

$$X_t = [1, Y_{t-1}^T, Y_{t-2}^T,\ldots,Y_{t-p}^T, e_{t-1}^T,\ldots,e_{t-q}^T]^T:$$

Then we can rewrite Equation (30) as

$$Y_t = \sum_{k=1}^{K} 1_{S_t=k}\,\theta_k X_t + e_t \quad e_t \sim N(0, \Sigma_{s_t});$$

where 1_a is the indicator function that equals 1 if a is true and 0 otherwise and $\theta_k = [\alpha_k, a_k^{(1)},\ldots,a_k^{(p)}, b_k^{(1)},\ldots,b_k^{(q)}]$.

We can now substitute the probability distribution function of $P(Y_t|Y_1,\ldots,Y_{t-1},S_t=k,\theta^{old})$ in Equation (31) and take the derivative of θ_k and all $a_k^{(i)}$ and $b_k^{(j)}$ in one go, using the results on matrix derivatives in Petersen et al. (2008):

$$\frac{Q(\theta,\theta^{old})}{\theta_k} = \sum_{t=1}^{N}\gamma_t(k)\left[-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y_t-\theta_k X_t)^T\Sigma_k^{-1}(Y_t-\theta_k X_t)\right]$$

$$= \sum_{t=1}^{N}\gamma_t(k)\Sigma_k^{-1}(Y_t-\theta_k X_t)X_t^T = 0;$$

Working out the brackets, premultiplying by $_k$ and isolating $_k$ results in the formula

$$_k = \left(\sum_{t=1}^{N} {}_t(k) Y_t X_t^T \right) \left(\sum_{t=1}^{N} {}_t(k) X_t X_t^T \right)^{1}: \tag{32}$$

The only remaining parameter that needs to be found is $_k$.

$$\begin{aligned}
\frac{Q(\; ; \; ^{\text{old}})}{_k} &= \frac{}{_k} \sum_{t=1}^{N} {}_t(k) \left[\frac{M}{2} \ln(2) \quad \frac{1}{2} \ln j \; _k j \quad \frac{1}{2} (Y_t \quad _k X_t)^T \; _k^{1} (Y_t \quad _k X_t) \right] \\
&= \frac{}{_k} \sum_{t=1}^{N} {}_t(k) \left[\frac{M}{2} \ln(2) \quad \frac{1}{2} \ln j \; _k j \quad \frac{1}{2} \mathrm{tr} \; (Y_t \quad _k X_t)^T \; _k^{1} (Y_t \quad _k X_t) \right] \\
&= \frac{}{_k} \sum_{t=1}^{N} {}_t(k) \left[\frac{M}{2} \ln(2) \quad \frac{1}{2} \ln j \; _k j \quad \frac{1}{2} \mathrm{tr} \; (Y_t \quad _k X_t)(Y_t \quad _k X_t)^T \; _k^{1} \right] \\
&= \sum_{t=1}^{N} \frac{1}{2} {}_t(k) \left[\; _k^{1} + \; _k^{1} (Y_t \quad _k X_t)(Y_t \quad _k X_t)^T \; _k^{1} \right] = 0:
\end{aligned}$$

Pre- and postmultiplying by $_k$ and isolating $_k$ leads to the equation

$$_k = \frac{\sum_{t=1}^{N} {}_t(k) (Y_t \quad _k X_t)(Y_t \quad _k X_t)^T}{\sum_{t=1}^{N} {}_t(k)}; \tag{33}$$

which is the final of the parameter updates in the maximization phase. The expectation and maximization steps are repeated until expectation of the log-likelihood improves less than a certain predefined tolerance.

4.3.3 Simulated EM

Above calculations become intractable if the MA-order q is bigger than zero, since the probability $P(Y_t | Y_1; : : : ; Y_{t \; 1}; S_t; \;)$ is in that case dependent on the previous hidden Markov states through the previous errors $e_{t \; j}; j > 0$. A solution to be still able to use EM is by simulation (Nielsen, 2000). For this algorithm, one would in every iteration simulate D sequences $\{S_t^{(d)}\}_{t=1}^{N}$ and calculate:

$$\hat{E}^{\text{old}}[\ln P(Y_1; : : : ; Y_N; S_1; : : : ; S_N | \;)] = \sum_{d=1}^{D} \ln P(Y_1; : : : ; Y_N; S_1^{(d)}; : : : ; S_N^{(d)} | \;):$$

The parameters update Equations (32) and (33) continue to hold, where $_t(k) = 1_{S_t^{(d)} = k}$ is equal to an indicator function and where the D data sets Y and X are combined in one.

Chapter 5

Application on tracking models

In this chapter, a kinematic model used in tracking contexts is introduced which will be the basis of some synthetic experiments. The characteristics of this model are examined, in particular whether the conditions of the theory established in previous chapters of this report hold. Afterwards, experiments will be done using synthetic data generated according to the kinematic model.

5.1 Model description

We wish to apply the theory of the previous chapters on tracking models. For this purpose, a kinematic model is introduced, which is an MS VARMA model where each mode k corresponds to a turn made with a certain turn rate ω_k (of which one equals $\omega_k = 0$). As described before, the model is used to describe the motion of an object such as an airplane or a drone. We assume that the target is a point object moving through a Cartesian plane. The specific two-dimensional kinematic model used, is described in more detail by Li and Jilkov (2003).

The Double Hidden Markov model being analyzed is the stochastic process $\{S_t; X_t; Y_t\}$. The discrete Markov Process process at time t, S_t, selects the hidden mode at time t and describes which model we are in. This process can take values in $[K]$, where $s_t = 1$ corresponds to the so-called Constant Velocity (CV) model and $s_t = 2, \ldots, K$ to models called Coordinated Turn (CT) models with turn rate $\omega_{s_t} \neq 0$ and $\omega_j \neq \omega_i$ if $j \neq i$. The process $\{X_t\}$ refers to the kinematic state of the system: the location coordinates x_t and y_t and the velocity in the x and y direction, $\dot{x}_t$ and $\dot{y}_t$. X_t is, therefore, the 4-dimensional vector

$$X_t = \begin{bmatrix} x_t \\ \dot{x}_t \\ y_t \\ \dot{y}_t \end{bmatrix}.$$

However, we do not observe X_t directly, but rather some noisy measurements Y_t, which is a function of X_t. We assume that location and velocity are measured directly, but inaccurately. In reality, these variables are calculated from other measurements, such as the Doppler effect. Our simplifying assumption causes Y_t to be simply equal to the sum of X_t and white noise.

The stochastic process $\{S_t; X_t; Y_t\}$ can be modelled as follows:

$$P(S_t = ij S_{t-1} = j) = P_{ji}; \tag{34}$$

$$X_t = F_{s_t} X_{t-1} + \Gamma_{x;s_t} V_{x;t}; \tag{35}$$

$$Y_t = X_t + \Gamma_y V_{y;t}. \tag{36}$$

The processes $\{V_{x;t}\}$ and $\{V_{y;t}\}$ are white noise processes with mean zero and standard deviation 1. $\Gamma_{x;s_t}$ and Γ_y are constant matrices fulfilling the role of covariance matrices. The mode-dependent matrices F_{s_t} describe the transition of X_t to X_{t+1}. For the CV model

$$F_{cv} = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{37}$$

In above equation T is the sampling time and assumed to be a fixed and known parameter. As the name suggests, the velocity variables do not change (random influences excluded). The location variables change proportional to the velocity. For the CT models

$$F_{ct} = \begin{bmatrix} 1 & \frac{\sin(\omega T)}{\omega} & 0 & -\frac{1-\cos(\omega T)}{\omega} \\ 0 & \cos(\omega T) & 0 & -\sin(\omega T) \\ 0 & \frac{1-\cos(\omega T)}{\omega} & 1 & \frac{\sin(\omega T)}{\omega} \\ 0 & \sin(\omega T) & 0 & \cos(\omega T) \end{bmatrix} \tag{38}$$

and is a function of the turn rate ω. The turn rate is the constant change in heading angle per time unit and is different for each mode. The turn rate can be assumed known or can be estimated from the data. Note that the time between two observations is T time units and the turn made between two observations is an angle of ωT. In this model, the total speed $v = \sqrt{\dot{x}_t^2 + \dot{y}_t^2}$ is constant, excluding the influence of the random terms $V_{x;t}$.

The covariance matrix $\Gamma_{x;s_t}$ for the constant velocity models are given by the following equations:

$$\Gamma_{x;CV} = \operatorname{diag}\left(\sigma_x^2 Q; \sigma_y^2 Q\right);$$

$$\text{where } Q = \begin{bmatrix} \frac{T^3}{3} & \frac{T^2}{2} \\ \frac{T^2}{2} & T \end{bmatrix}.$$

In the first of the two above equations σ_x^2 and σ_y^2 are the power spectral densities for the x- and y-direction, which are (potentially unknown) parameters. The covariance matrix for the

coordinated turn model is given by:

$$F_{x;CT} = \begin{bmatrix} \frac{2(\omega T - \sin(\omega T))}{\omega^3} & \frac{1-\cos(\omega T)}{\omega^2} & 0 & \frac{\omega T - \sin(\omega T)}{\omega^2} \\ \frac{1-\cos(\omega T)}{\omega^2} & T & \frac{\omega T - \sin(\omega T)}{\omega^2} & 0 \\ 0 & \frac{\omega T - \sin(\omega T)}{\omega^2} & \frac{2(\omega T - \sin(\omega T))}{\omega^3} & \frac{1-\cos(\omega T)}{\omega^2} \\ \frac{\omega T - \sin(\omega T)}{\omega^2} & 0 & \frac{1-\cos(\omega T)}{\omega^2} & T \end{bmatrix};$$

In the above equation ω is a parameter.

We wish to relate the autocovariance of above model to that of a (non MS-)VARMA model. To be able to do this, we first rewrite Equations (35) and (36) in matrix form:

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} F_{s_t} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \sigma_{x;s_t} & 0 \\ 0 & \sigma_y \end{bmatrix} \begin{bmatrix} V_{x;t} \\ V_{y;t} \end{bmatrix};$$

which can be rewritten in standard MS VARMA form as

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} F_{s_t} & 0 \\ F_{s_t} & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \sigma_{x;s_t} & 0 \\ \sigma_{x;s_t} & \sigma_y \end{bmatrix} \begin{bmatrix} V_{x;t} \\ V_{y;t} \end{bmatrix};$$

However, since only Y_t is observed, above equation needs to be rewritten such that X_t is substituted out of the equation. This results in:

$$Y_t = F_{s_t} Y_{t-1} - F_{s_t} \sigma_y V_{y;t-1} + \sigma_{x;s_t} V_{x;t} + \sigma_y V_{y;t}; \tag{39}$$

which can be written in standard MS VARMA form as:

$$\begin{bmatrix} Y_t \\ 0 \end{bmatrix} = \begin{bmatrix} F_{s_t} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \sigma_y & \sigma_{x;s_t} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{y;t} \\ V_{x;t} \end{bmatrix} + \begin{bmatrix} F_{s_t} \sigma_y & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{y;t-1} \\ V_{x;t-1} \end{bmatrix};$$

5.2 Assumptions

In developing the methodology of Chapter 3 and 4, several assumptions had to be made. First these assumptions, except for the assumptions on the functional form of the model and the technical assumptions on the consistency of the maximum likelihood estimation, will be summarized, after which they will be discussed shortly. In particular, some of these assumptions follow directly from the model description, which we will provide proofs for.

The eight assumptions that need to hold are:

1. The intervals between two sequential observations $t_s; t_{s+1}$ is constant.
2. The random disturbances $V_{x;t}$ and $V_{y;t}$ are draws from white noise processes.
3. The Markov Process S_t is ergodic, irreducible and stationary.
4. The transition matrices of the continuous variables are distinct for every hidden mode.
5. The random disturbances $V_{x;t}$ and $V_{y;t}$ are normally distributed.
6. The di erenced time series Y_t is a second order stationary random process.
7. The Markov process S_{t+h} is uncorrelated with the observations Y_t for every h 0.
8. The transition matrices of the continuous variables are invertible.

The rst assumption is a standard assumption for time series model. The second assumption is vital for many of the proofs, although in practice white noise assumptions often do not hold exactly. The third assumption holds for instance if every entry in the probability transition matrix is positive. The fourth assumption implies that it is not possible for models to di er only in the covariance matrix of the errors. The fth assumption is only necessary for the order selection method based on the eigenvalues of the data covariance matrix and can be relaxed if another method is used. The last three assumptions follow from the model and will be discussed in the following sections.

5.3 Stationarity

To be able to apply the results of Section 3.1.1, time series generated by the kinematic model need to be stationary. This is not the case, since nowhere in the process the e ects of random shocks fades. The process thus is integrated in each Markov state and therefore exhibits random walk-like behaviour, as shown in the following lemma:

Lemma 15. *The process $\{Y_t\}$ described in Equations (34) - (36) is not stationary.*

Proof. Since $\{V_{y;t}\}$ is a white noise process and thus stationary and independent of $\{X_t\}$, the process $\{Y_t\} = \{X_t + V_{y;t}\}$ is stationary if and only if $\{X_t\}$ is stationary. The approach that will be taken, is looking at the second moments to examine if wide sense stationarity can be ful lled. We will now investigate whether there exist stationary expressions for the second moments of the variables, starting with the velocity variables. Denote the second respectively the fourth dimension of the vector $_{x;i}V_{x;t}$ by $V_{\underline{x};t;i}$ respectively $V_{\underline{y};t;i}$. From Equation (35), (37) and (38) it follows that

$$\mathrm{E}[\underline{x}_{t+1}^2] = (1)\mathrm{E}[(\underline{x}_t + V_{\underline{x};t+1;1})^2] + \sum_{k=2}^{\infty} (k)\mathrm{E}[(\underline{x}_t \cos(\omega_k T) \quad \underline{y}_t \sin(\omega_k T) + V_{\underline{x};t+1;k})^2]; \quad (40)$$

$$\mathrm{E}[\underline{y}_{t+1}^2] = (1)\mathrm{E}[(\underline{y}_t + V_{\underline{y};t+1;1})^2] + \sum_{k=2}^{\infty} (k)\mathrm{E}[(\underline{y}_t \cos(\omega_k T) + \underline{x}_t \sin(\omega_k T) + V_{\underline{y};t+1;k})^2]: \quad (41)$$

Both equations can be rearranged as follows:

$$\mathrm{E}[\underline{x}_{t+1}^2] = \mathrm{E}[(\underline{x}_t^2] \left(1) + \sum_{k=2}^{\infty} (k)\cos^2(\omega_k T)\right) + \mathrm{E}[(\underline{y}_t^2] \sum_{k=2}^{\infty} (k)\sin^2(\omega_k T) \quad (42)$$
$$\mathrm{E}[\underline{x}_t \underline{y}_t] \left(\sum_{k=2}^{\infty} (k)\sin(2\omega_k T) + \sum_{k=1}^{\infty} (k)\ {}^2_{2;k};\right.$$

$$\mathrm{E}[\underline{y}_{t+1}^2] = \mathrm{E}[(\underline{y}_t^2] \left(1) + \sum_{k=2}^{\infty} (k)\cos^2(\omega_k T)\right) + \mathrm{E}[(\underline{x}_t^2] \sum_{k=2}^{\infty} (k)\sin^2(\omega_k T) + \quad (43)$$
$$\mathrm{E}[\underline{x}_t \underline{y}_t] \left(\sum_{k=2}^{\infty} (k)\sin(2\omega_k T) + \sum_{k=1}^{\infty} (k)\ {}^2_{4;k};\right.$$

since many terms cancel because of the white noise properties of the variables V. In above equations, ${}^2_{i;k}$ is the known variance of the errors $V_{i;t+1;k}$. Since we are trying to nd stationary solutions, we can assume $\mathrm{E}[\underline{x}_{t+1}\underline{y}_{t+1}] = \mathrm{E}[\underline{x}_t \underline{y}_t]; \mathrm{E}[\underline{x}_{t+1}^2] = \mathrm{E}[\underline{x}_t^2]$ and $\mathrm{E}[\underline{y}_{t+1}^2] = \mathrm{E}[\underline{y}_t^2]$. Moreover, let us write $c_1 = \quad (1) + \sum_{k=2}^{\infty} (k)\cos^2(\omega_k T)$; $c_2 = \sum_{k=2}^{\infty} (k)\sin^2(\omega_k T)$. Adding Equation (42) and (43) and rearranging terms we get:

$$\mathrm{E}[\underline{x}_t^2](1 \quad c_1 \quad c_2) + \mathrm{E}[\underline{y}_t^2](1 \quad c_1 \quad c_2) = \sum_{k=1}^{\infty} (k)(\ {}^2_{2;k} + \ {}^2_{4;k})$$

However, since $\sum_{k=1}^{p} (k) = 1$ and $\cos^2(!) + \sin^2(!) = 1; 8!$, we get $c_1 + c_2 = 1$ and therefore $\sum_{k=1}^{p} (k)(\ ^2_{2;k} + \ ^2_{4;k}) = 0$, which is only true if $fx_t g$ and $fy_t g$ are nonstochastic. Therefore, $fx_t g$ and $fy_t g$ are not stationary. □

The first differenced time series Y_t does appear to be stationary. In this case, we have from Equation (39)

$$Y_t = (I_M \quad F_{s_t})Y_{t\ 1} \quad F_{s_t\ y}V_{y;t\ 1} + \ _{x;s_t}V_{x;t} + \ _y V_{y;t}; \tag{44}$$

where I_M is the M-dimensional identity matrix, which in our context is I_4. Replacing $Y_{t\ 1}$ by $Y_{t\ 1}$ in Equation (44) leaves a remaining term of $Y_{t\ 2}$. Replacing this again with $Y_{t\ 2}$ and repeating this process and doing the same for the errors variables V_x and V_y, above equation can be rewritten in standard MS ARMA form as follows:

$$\begin{matrix} Y_t \\ 0 \end{matrix} = \sum_{i=1} \begin{matrix} F_{s_t} & I_M \\ 0 & 0 \end{matrix} \begin{matrix} 0 \\ 0 \end{matrix} \begin{matrix} Y_{t\ i} \\ 0 \end{matrix} + \begin{matrix} y & x;s_t \\ 0 & 0 \end{matrix} \begin{matrix} V_{y;t} \\ V_{x;t} \end{matrix} + \sum_{i=1} \begin{matrix} F_{s_t\ y} & y \\ 0 & \end{matrix} \begin{matrix} x;s_t \\ 0 \end{matrix} \begin{matrix} V_{y;t\ i} \\ V_{x;t\ i} \end{matrix} :$$

Recall that in Lemma 2, a sufficient and necessary condition for the stationarity of Markov Switching VARMA models was stated. In order to apply their results, the model needs to be rewritten in again another form, as an MS VAR(1) model:

$$z_t = \ _t z_{t\ 1} + \ _t \ _t;$$

where

$$z_t = \begin{matrix} Y_t \\ 0 \\ Y_{t\ 1} \\ 0 \\ \vdots \\ Y_{t\ l} \\ 0 \\ \ _y V_{y;t} \\ \ _{x;s_t} V_{x;t} \\ (F_{s_t\ y} \quad \ _y) V_{y;t\ 1} \\ \ _{x;st} V_{x;t\ 1} \\ \vdots \\ (F_{s_t\ y} \quad \ _y) V_{y;t\ l} \\ \ _{x;st} V_{x;t\ l} \end{matrix}, \quad _t = \begin{matrix} \ _y & \ _{x;s_t} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \ _y & \ _{x;s_t} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{matrix}, \quad _t = \begin{matrix} V_{y;t} \\ V_{x;t} \end{matrix} \text{ and}$$

$$
{}_t = \begin{bmatrix}
F_{s_t} & I_M & 0 & \cdots & & F_{s_t} & I_M & 0 & F_{s_t} & y & y & x;s_t & \cdots & & F_{s_t} & y & y & x;s_t \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
I_M & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
0 & I_M & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
\vdots & \ddots & \ddots & & & \vdots & \vdots & & \vdots & & & \vdots & \ddots & & \vdots & & \vdots \\
0 & 0 & I_M & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & I_M & & & 0 & & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & 0 & & & I_M & & & 0 & & 0 \\
\vdots & \vdots & \ddots & & & \vdots & \vdots & & \vdots & & & \ddots & \ddots & & \vdots & & \vdots \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & I_M & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & & 0 \\
0 & 0 & & & & 0 & 0 & & 0 & & & 0 & & & 0 & &
\end{bmatrix}:
$$

Recall from Section 2.2.2 that $_t$ can be partitioned as

$$
{}_t = \begin{matrix} A_t & B_t \\ 0 & J \end{matrix} :
$$

If the top Lyapunov exponent , or equivalently ^{0},

$$
{}^{0} = \inf_{t2\mathrm{N}} \mathrm{E}\frac{1}{t} \log jjA_tA_{t\ 1}:::A_1jj
$$

is strictly negative the di erenced time series is stationary.

Lemma 16. *The model described in Equation (44) is stationary.*

Proof. It is easy to see that if the state $s_t = 1$, which corresponds to the constant velocity model, the matrix $(F_1 \quad I_M)(F_{s_{t\ 1}} \quad I_M) = (F_{s_{t+1}} \quad I_M)(F_1 \quad I_M) = 0_{M \quad M}$. Moreover, simple matrix multiplications show that the only nonzero entries in the block matrix $A^n; n \quad 3$ are sums of the terms $\prod_{i=0}^{p-1}(F_{s_{t\ i}} \quad I_M)$, $p \quad 2$. And therefore, $A_nA_{n\ 1}:::A_1 = 0$ if $s_n = 1$ for any $n = 1;:::;t$ and $t \quad 3$. The norm of a zero matrix is 0 by de nition and therefore the log value of the norm is minus in nity. Now assume the stationary long term probability to be in the CV model (1) is strictly greater than 0. Then for time length $t = 3$ it holds that:

$$
\mathrm{E}\frac{1}{3} \log jjA_3A_2A_1jj = \quad 1;
$$

since the matrix norm $jjA_3A_2A_1jj$ is nite in case $! > 0$ and $s_n \neq 1$ for $n = 1;2;3$. Above derivations are independent of the MS VARMA order of l and thus also holds for $\lim_{l\ !\ 1}$. □

Note that the above proof strongly relies on the behaviour of the process in the CV model and that it is not required that a mixture of only CT models results in stationary di erenced time series. Since the in nite Markov Switching VARMA model is stationary, the corresponding Markov Switching VECM model of Equation (44) is as well and we can thus conclude the variables fY_tg are cointegrated.

5.4 Correlation Markov process and observation process

We need to show that the correlation, or equivalently the covariance of Y_t with the hidden Markov process is 0. If this holds, then this also holds for the differenced process $Y_t - Y_{t-1}$. No formal proof will be provided in this section. Instead, an intuition why this assumption can be assumed to be true is provided.

Instead of focusing on the correlation of Y_t with the hidden Markov Process, we can examine the covariance of X_t with the indicator function $1_{S_{t+h}=k}$, since Y_t equals X_t plus some white noise process, where the latter is by definition uncorrelated with the hidden Markov process for each $h \neq 0$. Define μ_Z as the mean of the variable Z. We obtain:

$$\begin{aligned}
\text{Cov}(X_t; 1_{S_{t+h}=k}) &= \text{E}[(X_t - \mu_X)(1_{S_{t+h}=k} - \mu_{1_{S_{t+h}=k}}] \\
&= \text{E}[X_t 1_{S_{t+h}=k}] - \mu_X \text{E}[1_{S_{t+h}=k}] - \text{E}[X_t]\,\mu_{1_{S_{t+h}=k}} + \mu_X\,\mu_{1_{S_{t+h}=k}} \\
&= \text{E}[X_t 1_{S_{t+h}=k}] - \mu_X\,\pi(k) - \mu_X\,\pi(k) + \mu_X\,\pi(k) \\
&= \sum_{j=1}^{K} \text{E}\left[X_t 1_{S_{t+h}=k} | S_{t+h} = j\right] \pi(j) - \mu_X\,\pi(k) \\
&= \text{E}[X_t | S_{t+h} = k]\,\pi(k) - \mu_X\,\pi(k);
\end{aligned}$$

which equals 0 if $\text{E}[X_t | S_{t+h} = k] = \mu_X$. That is, we need to argue that the expected value of the variables X_t is independent of the hidden Markov state. Moreover, since the time series X_t is non-stationary, we also need to check whether the mean $\text{E}[X_t]$ is constant for all t.

When we just examine the velocity variables in the horizontal and vertical direction, $\underline{x}$ and $\underline{y}$, the transition of the CT models between successive time periods can be expressed by the transition matrix

$$\begin{bmatrix} \cos(\omega_{S_t} T) & \sin(\omega_{S_t} T) \\ \sin(\omega_{S_t} T) & \cos(\omega_{S_t} T) \end{bmatrix},$$

which is a rotation matrix. This means that if we disregard the random error terms for the moment, the velocity variables $\underline{x}$ and $\underline{y}$ change in a circular manner like the x and y coordinates of a point moving over the perimeter of a circle with radius the constant total speed $\sqrt{\underline{x}^2 + \underline{y}^2}$. Then the average rotation made is $\sum_{k=2}^{K} \pi(k)\,\omega_k = R$, where we assumed $k = 1$ corresponds to the sole CV model. Moreover, we can define a related rotation process ρ_t as the sum of the rotations up to time t. In the long run, we expect the probability mass of each state k occurring at some rotation ω to distribute evenly over the domain of ρ_t. If $R \neq 0$ this is easy to see, since the rotation process is expected to move around the circle. However, if $R = 0$, we can expect the same, since the rotation process can be interpreted as a random walk-like process, whose variance in the limit goes to infinity, such that the influence of the starting position becomes negligible. If the domain of ρ_t is symmetric over the x and y axis, this all implies that the expected value of the velocity variables given the hidden Markov state equals zero for all hidden Markov states $k = 1; \ldots; K$. The same reasoning can be applied to the location variables. Moreover, if the expected velocity in the x and y direction equals 0, the expected value of the location variables will also be stationary.

There are some exceptions imaginable where above reasoning does not hold. For example, if the turn rate of all CT models equal $c_k \quad 2$, for some integers c_k, since then the domain of $\quad_t$ is not symmetric over the x and y axis. Another slightly less trivial case is when the transitions of the hidden Markov states are not random but all xed, since the rotation process can not be seen as a random walk anymore. An example is when the process is in each even time period t in the Markov state with $! = 0{:}2$ and in the odd time periods in the Markov state with $! = \quad 0{:}2$ with probability 1.

5.5 Invertible transition matrices

The last and most problematic assumption is the invertibility of the transition matrices of the differenced process. It is straightforward to check that the transition matrices $F_{cv} - I_4$ and $F_{ct} - I_4$ are not invertible. As a result, the conditions of Theorem 12 and 13 are not fulfilled. However, when restricting the focus to the velocity process $Y_t' = \begin{pmatrix} \underline{x}_t \\ \underline{y}_t \end{pmatrix}$, with only CT models, the transition matrices are invertible, when $\omega_k \neq c\pi$ for all Markov states k and all integers c. As a proxy for the CV models, we can thus use a CT model with small angle ω. We can therefore use the autocovariance method to obtain a lower bound estimate of the number of hidden Markov states needed for the kinematic model, by looking at the differenced time series of the velocity variables only.

From a theoretical viewpoint, using only CT models causes the differenced process to not be stationary anymore. However, in practice this effect will be mild since the acceleration of objects usually falls in a limited range. The variance of the process therefore does not increase unboundly, which could be the case when location data was taken into account. When data is simulated from the theoretical data, the results may therefore be more erratic than for real instances.

5.6 Estimation of the number of modes

The kinematic model with CT submodels of which one with small turn angle, approximating the CV model, is used as the basis of an experiment. The goal of the experiment is determining whether the autocovariance method, MLE and penalty methods can be used in the model building procedure, introduced in Figure 8 of the introduction. It will be examined whether the correct number of hidden Markov states can be distilled from a kinematic model with 1 approximate CV submodel and either 2 or 4 CT submodels using the autocovariance method as lower bound, the AIC and BIC as upper bounds and additionally the MSC.

For this experiment, a tracking model was generated with probability transition matrix $P = \frac{0.1}{K}\mathbf{1}_M\mathbf{1}_M^T + 0.9I_M$, turn rates $!_k$ equal to $\quad ; \quad 0.5 \; ; 0.5 \quad$ and $\quad$ if $K = 5$ and there are thus four CT models and equal to $\quad 0.5 \; ; 0.5 \quad$ if $K = 3$. Instead of generating from a CV model as the last submodel, a CT model with turn rate $0.1 \quad$ was used. The sampling time T is set to 0.1, the covariance matrix of the errors is in contrast to the default covariance matrices equal to the identity matrix for each Markov state, since this decreases the variability of the whole time series without decreasing the variability per time step. The number of observations N equals 100,000.

A potential problem of this setting is that the transition matrices A_k are rather similar for di erent hidden Markov states k. As mentioned in the previous section, this matrix equals for the true CV model

$$F_{cv} = \begin{bmatrix} 1 & 0.1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

while for $!_k = 0.5$

$$F_{ct} = \begin{bmatrix} 1 & 0.0996 & 0 & 0.0078 \\ 0 & 0.9877 & 0 & 0.1564 \\ 0 & 0.0078 & 1 & 0.0996 \\ 0 & 0.1564 & 0 & 0.9877 \end{bmatrix};$$

which can be expected to be too similar matrices to be distinguishable, taking the experiments of Section 3.3 as a reference. Since one often stays in the same state for a long amount of time, one may solve this problem by an approach called thinning. For this methodology, one uses only each i-th observation, since the di erence between the di erent Markov states gets more pronounced after more observations. The number of Markov states of the time series are therefore estimated several times using no thinning, thinning with $i = 5$ and with $i = 10$.

The autocovariance method is not very computationally expensive and for a fairly great number of con gurations (24), the number of hidden Markov states was estimated. Maximum Likelihood Estimation takes more time and only for a subset of the con gurations the penalty functions were applied. MLE was used only for the non-di erenced 2-dimensional observations, since this is less computationally expensive and the thinned data with $i = 10$ is

not considered. This leaves four different configurations. However, for the settings that were examined, MLE was used twice, once where apart from the other parameters, the variance was estimated, violating the assumptions leading to the consistency of the upper bounds of the penalty functions, and once where the variance was assumed known.

The maximum order K considered for MLE was 8. For each order K, three different randomly generated starting positions were used, to decrease the effect of local maxima. The entries of the matrices $\hat{A}_k$ were initialized to random numbers between 0 and 1. The probability matrix P was initialized to a uniform matrix with each entry $1/K$ and the variance for each Markov state was set to the unconditional sample variance of the observations when the variances were not assumed known.

Table 7: Estimations $\hat{K}$ of randomly generated kinematic non-differenced time series with 3 Markov States using the autocovariance method

		No thinning	Thinning $i = 5$	Thinning $i = 10$
Non-differenced	$M = 2$	3	4	3
	$M = 4$	1	2	2
Differenced	$M = 2$	2	2	2
	$M = 4$	1	1	1

Table 8: Estimations $\hat{K}$ of randomly generated kinematic non-differenced time series with 5 Markov States using the autocovariance method

		No thinning	Thinning $i = 5$	Thinning $i = 10$
Non-differenced	$M = 2$	3	5	2
	$M = 4$	1	2	2
Differenced	$M = 2$	2	2	2
	$M = 4$	1	1	2

The results of the autocovariance method in Table 7 and 8 give rise to some clear observations. Differencing leads to fairly poor results, as in the synthetic experiments before. However, for the non-differenced data, using only the velocity variables the results are better. This is interesting, since the experiments of Section 3.3 suggested that more dimensions usually led to better performance, but also well explainable since the 4-dimensional process cannot be expressed as a VARMA process due to noninvertibility of the transition matrices. Lastly, the experiments seem to suggest that thinning may increase the ability to differentiate the various Markov states as expected, although the evidence is not decisive and one should see the conclusions drawn about thinning as preliminary. A too high thinning factor can decrease performance, which was the case for $i = 10$. In the case where there are 3 Markov states, thinning with $i = 5$ led to the detection of an extra Markov state. It could be that the algorithm was confused by a mixture of Markov states, since the time periods in one thinning interval will sometimes not all be of the same Markov state. It can be concluded that thinning is a promising technique, but that the thinning factor should be chosen with care.

Table 9: Estimations $\hat{K}$ of randomly generated non-differenced velocity time series using penalty functions with unknown variance

		AIC	BIC	MSC
$K = 3$	No thinning	8	8	8
	Thinning $i = 5$	8	8	8
$K = 5$	No thinning	8	8	8
	Thinning $i = 5$	8	8	8

Table 10: Estimations $\hat{K}$ of randomly generated non-differenced velocity time series using penalty functions with known variance

		AIC	BIC	MSC
$K = 3$	No thinning	5	5	5
	Thinning $i = 5$	4	4	4
$K = 5$	No thinning	6	6	6
	Thinning $i = 5$	5	5	5

The results of the penalty functions where the variance is not fixed are tabulized in Table 9, but actually do not need to be summarized in a table: for each investigated setting, more hidden Markov states lead to a better objective function for the AIC, the BIC as well as the MSC, since multiple hidden Markov states close to one another with decreasing variance can lead to significant increases in the likelihood value, which the penalties do not counter enough. These results do therefore not convey any information on the actual number of hidden Markov states.

The results of the experiment where the variance was fixed, as shown in Table 10 are more informative. The number of hidden Markov states estimated, is most of the times an overestimation, although the extend of the overestimation is especially for $K = 5$ quite modest. Thinning seems to improve the estimates, although the number of experiments is too small to draw strong conclusions. What is striking, is that the estimation of the three penalty functions are in all cases equal, even though the MSC does not exhibit the upper bound property of the AIC and BIC. A thing to note is that although MLE does seem to provide decent estimates, it is computationally expensive and can for large sample sizes and with slow convergence take minutes, where the autocovariance method takes seconds. Moreover, MLE requires more information, since the variance should be fixed and known.

All in all, it seems that the penalty functions and the autocovariance complement each other quite well and they can be used to gain a rather accurate idea of the number of hidden Markov states. Based on the thinned 2-dimensional data, one would conclude one would need 4 hidden Markov states for the model with 3 modes and 5 hidden Markov states for the model with 5 modes. Using no thinning, the intervals are wider, but correct for both 3 modes and 5 modes. It is thus possible to use the autocovariance method and the penalty functions method to obtain an estimate of the number of hidden Markov states, but one must take the uncertainties of the methods into account.

Chapter 6

Conclusion and recommendations

6.1 Conclusion

The intention of this research was to answer the following main research question:
Construct a procedure to determine the optimal number of hidden Markov States jSj for the model of the form

$$P(S_t = ijS_{t\ 1} = j) = P_{ji};$$
$$X_t = f_{S_t}(X_{t\ 1}; V_t);$$
$$Y_t = g(X_t; W_t);$$

with the intention to incorporate this in the overall procedure for tracking and classi cation. Most attention was paid to the autocovariance method. A big advantage of this method is that it is easy to implement and does not require a lot of computational power. Therefore, there is little reason to not use the results in the process of determining the number of hidden Markov states. However, the amount of knowledge gained depends strongly on the characteristics of the hidden Markov Process. Factors which contribute to better performance of the autocovariance method are su cient distinctiveness of the di erent modes, low amount of hidden Markov states, long time periods spent in the same Markov state, high number of observations, high number of dimensions and a lack of modes which would be nonstationary on their own. Since the strength of the autocovariance method depends strongly on these factors, one can, therefore, only know how to interpret the results of the estimation of the autocovariance method, if one knows these characteristics. Although in practice one does not know, for instance, the probability transition matrix, one may have already a prior idea whether the diagonal values are close to one or not. If one does not have such a prior idea, estimates of MLE may provide an idea whether the estimated number of hidden Markov states is reliable, although it may stay di cult to determine whether the number of hidden Markov states is underestimated, or estimated accurately.

The rst subquestion,

How can the optimal number of hidden Markov States be determined based on the data,

can thus partially be answered based on this research. The autocovariance method can in some cases lead to an accurate estimate of the optimal number of hidden Markov states and

may, together with penalty functions, lead to a narrow interval the number of hidden Markov states should fall into. However, in other situations where the data generating process has unfavourable characteristics, the autocovariance method provides little guidance.

as the answer returned by the autocovariance method cannot always be simply taken for granted, the answer of the second subquestion

> *How can the above methods be included in the existing overall procedure for tracking and classi cation,*

is, therefore, not as straightforward as one may have initially hoped. Therefore, one should carefully consider if the obtained estimate is reliable by looking at the characteristics of the time series, possibly with the help of MLEs and prior, expert knowledge. However, for multivariate problems one needs to estimate too many parameters to reliably estimate the model from scratch and one should, therefore, be careful in interpreting these parameter estimates. It may for high dimensional problems also be necessary to limit the number of parameters in some way, for example by restricting attention to CT models with unknown angle.

The implication of these conclusions is that it still involves human judgement to determine the number of hidden Markov states and build the model, while one ideally works as structured as possible, using as little human judgement as possible. However, it is still an improvement over previous practice, as one has tools and estimates to base the model building on and one does not solemnly have to rely on domain knowledge.

6.2 Recommendations for further research

Although the autocovariance method was already first introduced almost 20 years ago, research into its applicability for especially multivariate problems has remained underexposed. In this thesis a number of characteristics of time series have been identified that affect the suitability of the method, however a few points remain open for further research. The experiments in Section 5.6 suggest that for random walk like time series the bounds for undifferenced time series perform relatively well, but so far, this does not have a solid theoretical basis. It is worthwhile to find out whether this a general property for random walk like time series and whether there this can theoretically be explained.

Another important suggestion for further research is whether it is possible to find a VARMA order selection method which preserves the lower bound guarantee of the autocovariance method. For univariate time series this method already exists as the three pattern method and a multivariate extension would thus be valuable.

6.3 Recommendations for application

This thesis work has lead to all sorts of insights and theoretical results, some of which can be of practical importance for the tracking model building process of Thales. In the introduction, a very crude overview of the new model formulation process was given, shown again in Figure 12. Using the conclusions of the various results in this report, we can provide some further details on this process.

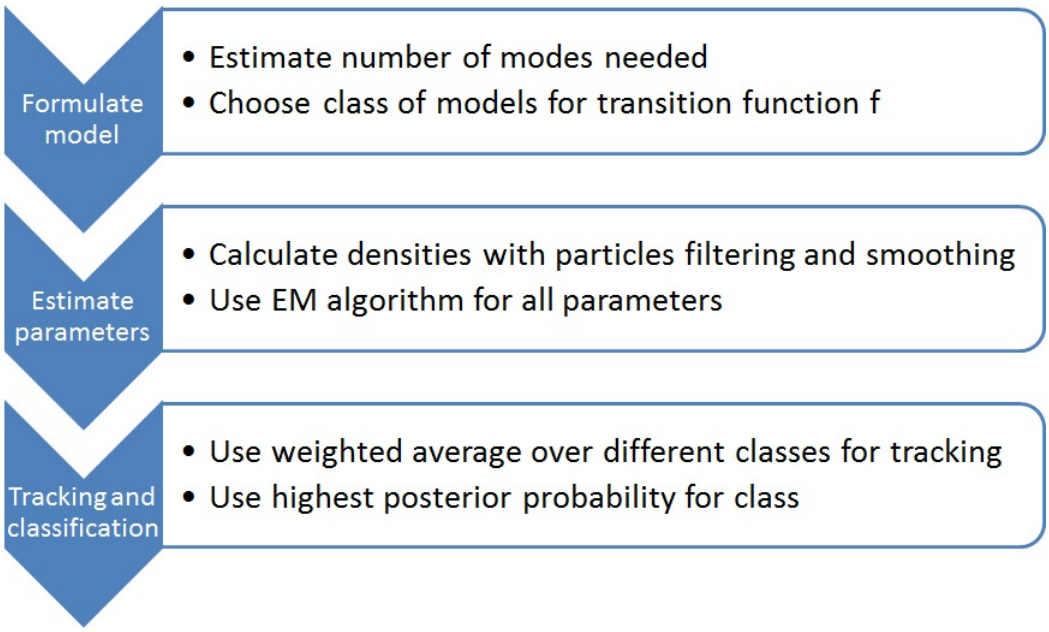

Figure 12: New model building procedure

Firstly, one needs to decide if a VAR(1) model would be a good characterisation of the time series at hand, or if we need a more general VARMA model. Moreover, if we are dealing with noisy observations of underlying kinematic states, a VARMA(1,1) model speci cation may be more appropriate for the time series, but if the observation noise is negligible, it may be wiser to approach the data as from a VAR(1) process, although not strictly correct.

When using the autocovariance method, it is advisable to only use the velocity variables, since nonstationarity issues are more modest than for the location variables. Moreover, depending on the sampling time of the observations one may need to use thinning for more reliable results. Di erencing is not recommended, as this performed poorly in all experiments of this thesis.

The reliability of the lower bounds depends strongly on the characteristics of the model and it is therefore important to take those into account. One needs to ask oneself questions such as does the hidden Markov chain stay, on average, in the same hidden Markov state for longer periods of time? In how many dimensions does the object under consideration move? How many observations do we have? What is the number of hidden Markov states we approximately expect? Does the process contain periods of erratic, nonstationary be-

haviour? Are the Markov states su ciently di erent from each other? And perhaps the most important question, to what extent do the assumptions of the methodologies hold, as were summarized in Section 5.2. Some of these questions may be di cult to answer, since one does not have enough prior information, but parameter estimates of MLE might give an idea on the direction of the answers, even though these estimates should also be handled with care due to the high dimensionality of the estimation problem and the presence of many local minima. Maximum likelihood estimation for models with hidden Markov states is rather computationally expensive, but since its results can complement the autocovariance method in two ways, by providing context for its interpretation and by obtaining an upper bound using penalty functions it is recommended to use both methods in the model building process. The maximum likelihood estimates may also provide an idea which class of models is suitable for further use. Both methods together can thus be used in a more informed model formulation process.

Appendix A

List of abbreviations

ACF: Autocorrelation Function

AIC: Akaike Information Criterion

AR: Autoregressive

ARIMA: Autoregressive Integrated Moving Average

ARMA: Autoregressive Moving Average

BIC: Bayesian Information Criterion

CR: Column Ratio

CT: Coordinated Turn

CV: Constant Velocity

DHMM: Double Hidden Markov Model

EM: Expectation Maximization

ESCAF: Extended sample autocorrelation function

HMM: Hidden Markov Model

KL: Kullback-Leibler

MA: Moving Average

MAP: Maximum A Posteriori

MLE: Maximum Likelihood Estimation/estimator

MS: Markov Switching

MSC: Markov Switching Criterion

PACF: Partial Autocorrelation Function

PM: Product Matrix

RR: Row Ratio

SSM: State Space Model

VAR: Vector Autoregressive

VARIMA: Vector Autoregressive Integrated Moving Average

VARMA: Vector Autoregressive Moving Average

VECM: Vector Error Correction Model

Appendix B

References

Abo-Hammour, Z. S., Alsmadi, O. M., Al-Smadi, A. M., Zaqout, M. I., and Saraireh, M. S. (2012). ARMA model order and parameter estimation using genetic algorithms. *Mathematical and Computer Modelling of Dynamical Systems*, 18(2):201{221.

Ackerson, G. and Fu, K. (1970). On state estimation in switching environments. *IEEE Transactions on Automatic Control*, 15(1):10{17.

Altissimo, F. and Corradi, V. (2002). Bounds for inference with nuisance parameters present only under the alternative. *The Econometrics Journal*, 5(2):494{519.

Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cli s.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of nite state Markov chains. *The annals of mathematical statistics*, 37(6):1554{1563.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The in nite hidden Markov model. In *Advances in neural information processing systems*, pages 577{584.

Beguin, J.-M., Gourieroux, C., and Monfort, A. (1980). Identi cation of a mixed autoregressive-moving average process: The corner method. *Time series*, pages 423{436.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Box, G. E. and Jenkins, G. M. (1970). *Time series analysis: forecasting and control*. San Francisco: Holden Day.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1):62{91.

Buishand, T. A. (1982). Some methods for testing the homogeneity of rainfall records. *Journal of hydrology*, 58(1-2):11{27.

Camilleri, T. (2012). *Multiple modelling of EEG data to classify di erent mental states*. University of Malta.

Cassar, T., Camilleri, K. P., and Fabri, S. G. (2010). Order estimation of multivariate ARMA models. *IEEE Journal of Selected Topics in Signal Processing*, 4(3):494{503.

Cavicchioli, M. (2014). Determining the number of regimes in Markov switching VAR and VMA models. *Journal of Time Series Analysis*, 35(2):173{186.

Chan, W.-S. (1999). A comparison of some of pattern identi cation methods for order determination of mixed ARMA models. *Statistics & Probability Letters*, 42(1):69{79.

Choi, B. (1992). *ARMA model identi cation*. Springer Science & Business Media.

Choi, B. (1993). Two chi-square statistics for determining the orders p and q of an ARMA (p, q) process. *IEEE transactions on signal processing*, 41(6):2165{2176.

Chow, G. C. (1960). Tests of equality between sets of coe cients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591{605.

Clopper, C. J. and Pearson, E. S. (1934). The use of con dence or ducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404{413.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2010). Bayesian nonparametric methods for learning Markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43{54.

Francq, C. and Zako•an, J.-M. (2001). Stationarity of multivariate Markov{switching ARMA models. *Journal of Econometrics*, 102(2):339{364.

Franses, P. H., van Dijk, D., and Opschoor, A. (2014). *Time series models for business and economic forecasting*. Cambridge university press.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model speci cation. *Journal of econometrics*, 16(1):121{130.

Gray, H. L., Kelley, G. D., and Mc Intire, D. (1978). A new approach to ARMA modeling. *Communications in Statistics-Simulation and Computation*, 7(1):1{77.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357{384.

Hansen, B. E. (1992). The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *Journal of applied Econometrics*, 7(S1):S61{S82.

Hœglund, R. and Ostermark, R. (1991). Automatic ARIMA modelling by the cartesian search algorithm. *Journal of forecasting*, 10(5):465{476.

Hubert, P. (2000). The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes. *Stochastic Environmental Research and Risk Assessment*, 14(4-5):297{304.

Karlin, S. (2014). *A rst course in stochastic processes*. Academic press.

Karlsen, H. A. (1990). *Doubly stochastic vector AR(1) processes*. University of Bergen.

Kehagias, A., Nidelkou, E., and Petridis, V. (2006). A dynamic programming segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Research and Risk Assessment*, 20(1-2):77{94.

Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2001). An online algorithm for segmenting time series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 289{296. IEEE.

Kilian, L. and Lutkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Koreisha, S. and Yoshimoto, G. (1991). A comparison among identi cation procedures for autoregressive moving average models. *International Statistical Review/Revue Internationale de Statistique*, pages 37{57.

Lange, T. and Rahbek, A. (2009). An introduction to regime switching time series models. In *Handbook of Financial Time Series*, pages 871{887. Springer.

Lardies, J. and Larbi, N. (2001). A new method for model order selection and modal parameter estimation in time domain. *Journal of Sound and Vibration*, 245(2):187{203.

Lee, A. F. and Heghinian, S. M. (1977). A shift of the mean level in a sequence of independent normal random variables A Bayesian approach. *Technometrics*, 19(4):503{506.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127{143.

Li, X.-R., Bar-Shalom, Y., and Blair, W. D. (2000). Engineer's guide to variable-structure multiple-model estimation for tracking. *Multitarget-multisensor tracking: Applications and advances.*, 3:499{567.

Li, X. R. and Jilkov, V. P. (2003). Survey of maneuvering target tracking. Part I. Dynamic models. *IEEE Transactions on aerospace and electronic systems*, 39(4):1333{1364.

Liang, G., Wilkes, D. M., and Cadzow, J. A. (1993). ARMA model order estimation based on the eigenvalues of the covariance matrix. *IEEE transactions on signal processing*, 41(10):3003{3009.

Liu, X., Lin, Z., and Wang, H. (2008). Novel online methods for time series segmentation. *IEEE Transactions on Knowledge and Data Engineering*, 20(12):1616{1626.

Lutkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

MacKay, R. J. (2002). Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30(4):573{589.

Magill, D. (1965). Optimal adaptive estimation of sampled stochastic processes. *IEEE Transactions on Automatic Control*, 10(4):434{439.

Moon, J. R. (2002). E ects of birds on radar tracking systems. *RADAR 2002*, pages 300{304.

Nielsen, S. F. (2000). On simulated EM algorithms. *Journal of Econometrics*, 96(2):267{292.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Pollock, S. G. (2011). Lecture Notes in econometrics, Chapter 10 multivariate ARMA models, Retreived from www.le.ac.uk/users/dsgp1/COURSES/THIRDMET/MYLECTURES/10MULTARMA.pdf.

Richa, E. (2018). Jump Markov nonlinear system identi cation in multi-sensor target tracking: a novel approach for multiple model joint tracking and behavior classi cation (Master thesis). *Retreived from TU Delft repository*.

Robert, C. P., Ryden, T., and Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57{75.

Ryden, T. (1995). Estimating the order of hidden Markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345{354.

SAS institute (1999). Documentation SAS Institute version 8. `https://v8doc.sas.com/sashtml/ets/chap7/sect28.htm#ariesacf`. Accessed: 2020-01-19.

Seghouane, A.-K. and Amari, S.-I. (2007). The AIC criterion and symmetrizing the Kullback{Leibler divergence. *IEEE Transactions on Neural Networks*, 18(1):97{106.

Smith, A., Naik, P. A., and Tsai, C.-L. (2006). Markov-switching model selection using Kullback{Leibler divergence. *Journal of Econometrics*, 134(2):553{577.

Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36{47.

Straubing, H. (1983). A combinatorial proof of the Cayley-Hamilton theorem. *Discrete Mathematics*, 43(2-3):273{279.

Tiao, G. C. and Box, G. E. (1981). Modeling multiple time series with applications. *journal of the American Statistical Association*, 76(376):802{816.

Tiao, G. C. and Tsay, R. S. (1983). Multiple time series modeling and extended sample cross-correlations. *Journal of Business & Economic Statistics*, 1(1):43{56.

Tsay, R. S. and Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association*, 79(385):84{96.

Velicer, W. F. and Harrop, J. (1983). The reliability and accuracy of time series model identi cation. *Evaluation Review*, 7(4):551{560.

Wei, C.-Z. et al. (1992). On predictive least squares principles. *The Annals of Statistics*, 20(1):1{42.

Winker, P. (2000). Optimized multivariate lag structure selection. *Computational Economics*, 16(1-2):87{103.

Zhang, J. and Stine, R. A. (2001). Autocovariance structure of Markov regime switching models and model selection. *Journal of Time Series Analysis*, 22(1):107{124.