IDENTIFICATION OF LINEAR SYSTEMS IN CANONICAL FORM WITH THE EXPECTATION MAXIMIZATION ALGORITHM

By
Pavlos V. Papadopoulos

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
TECHNICAL UNIVERSITY OF CRETE
CHANIA, GREECE
DECEMBER 2009

TECHNICAL UNIVERSITY OF CRETE
DEPARTMENT OF
ELECTRONICS AND COMPUTER ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "**Identification of Linear Systems in Canonical Form with the Expectation Maximization Algorithm**" by **Pavlos V. Papadopoulos** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: December 2009

Supervisor: _____
Prof. Vasilios Digalakis

Readers: _____
Prof. Athanasios Liavas

_____
Assis. Prof. Michael Lagoudakis

TECHNICAL UNIVERSITY OF CRETE

Date: **December 2009**

Author: **Pavlos V. Papadopoulos**

Title: **Identification of Linear Systems in Canonical Form with the Expectation Maximization Algorithm**

Department: **Electronics and Computer Engineering**

Degree: **M.Sc.** Convocation: **December** Year: **2009**

Signature of Author

I have yet to see any problem, however complicated, which, when you looked at it in the right way, did not become still more complicated.

Poul Anderson

Table of Contents

List of Figures

Abstract

System Identification (SI) has been a hot research topic for over 30 years since it can find applications in various fields such as biosciences, economics and control system design to name few. This very fact leads to the development of a general mathematical theory that can deal with the problem of SI independent of the field of application.

In this thesis we examine the family of linear state-space models and aim to find the system parameters from a series of observation data. Contrary to the dominant approaches of Least-Squares (LS) estimation of the system parameters we have developed a new procedure that solves the problem through the Expectation Maximization (EM) algorithm.

Regardless of the estimation algorithm, if there are no restrictions on the form of the matrices we want to estimate, the matrices can be determined up to within a linear transformation and thus the result may be different than the true solution. Moreover, the convergence of iterative algorithms may be affected by iterating in a neighborhood of the true solution. To overcome this problem one must constrain the system matrices to follow structures which are commonly known as *canonical forms*.

We examine a family of such canonical forms and apply the EM algorithm. First we form the auxiliary function that is used in the EM algorithm and then by use of the Kalman Filter and the Rauch-Tung-Striebel (RTS) Smoother we collect the sufficient statistics that appear in the auxiliary function. Finally based on those statistics we reestimate the system matrices. We examine the EM behavior both in the general form of a system and its Innovation Representation (IR).

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Vasilios Digalakis for his guidance and inspiration. This is the second time we cooperate, the first was in my diploma thesis, i feel that through our collaboration i have come closer to the understandings and works of a scientific procedure.

I would also like to thank professors Athanasios Liavas and Michael Lagoudakis for their comments on my work and for participation the examining committee.

I could not forget of course to thank all the guys in the lab. Hlias, Orfeas, Maria, Nikos, Theodosis, Vaso, Giannis, Michael for all the good times we had in and out of the lab and for always cheering me up in my bad "moods".

I feel very fortunate for having met my friends from the rest of telecom lab. Dimitri, Gianni, Despoina, Taso thanks for all the good times we had in our common breaks.

I would also like to thank all my friends back in Edessa for not making me feel nostalgic of home and for their hearty welcome the times i returned back where we tried to compensate for the "lost" time we spent apart.

Finally nothing could have happened without the support and love of my family. This thesis is dedicated to them.

Introduction

A large variety of papers on system identification have appeared over the last 40 years. Though there was a substantial progress in the theory of stochastic processes and multivariable statistical analysis during 1950s, it is widely recognized that the theory of system identification started only in the mid-1960s with the publication of two important papers. The first was published from Åström and Bohlin [9], in which the Maximum Likelihood (ML) method was extended to a serially correlated time series to estimate Autoregressive Moving Average model with exogenous inputs model (ARMAX) models, while the second was published from Ho and Kalman [46], in which the deterministic state-space realization problem was solved for the first time by forming a Hankel matrix in terms of impulse responses. The (deterministic) realization problem as stated in [54] is to find the state dimension and system matrices (up to similarity transforms) from a sequence of impulse responses $\{G_t, t = 0, 1, \cdots\}$ or a transfer matrix $G(z)$. These two papers [9],[46] gave birth to the future developments of system identification theory and techniques [35].

The work of Ho and Kalman [46] laid the foundation for the development of Subspace Methods for System Identification. Ho and Kalman dealt with the realization problem of deterministic systems which does not consider any noise. The realization problem for stochastic systems is to find all Markov models whose outputs simulate given covariance data or spectral density matrix [54] and was first addressed by Faurre [30] and Akaike [1]. A key step in stochastic realization is either to apply the deterministic realization theory to a certain Hankel matrix constructed with sample estimates of the process covariances, or to apply the canonical correlation analysis (CCA) [48] to the future and past of the observed process. Stochastic realization theory suffered from the same drawback as deterministic realization theory, up to the early 1990s stochastic realization supported modeling of stochastic processes, or time series, only. Thus the results of realization theory could not be applied to a system in which both a deterministic test input and a stochastic disturbance are involved. These realization theory based techniques have led to a development of various so-called subspace identification methods. The most well known algorithms for subspace identification are Numerical algorithms for Subspace State Space System IDentification (N4SID) [93] and Multivariable Output-Error State sPace (MOESP) [96], [97] which apply the realization theory along with linear algebra tools like LQ decomposition and Singular Value Decomposition (SVD).

Astrom and Bohlin [9] attacked the problem from a different angle. In their work,

they attempted to build single-input, single-output (SISO) ARMAX models from observed input-output data sequences applying ML estimation on the system parameters. This work laid the foundation for many statistical identification techniques which have been developed in the literature, most of which are now comprised under the label of Prediction Error Methods (PEM), the common characteristic of these methods is that they usually attempt to minimize a cost criterion. This has lead from the work of Eykhoff [29] to perhaps the most successful books dealing with the subject of PEM of Södesrstörm-Stoica [84] and of Ljung [60], both of which adopted the same clear distinction between choice of model structure and choice of criterion. Södesrstörm and Stoica gave more emphasis on analysis and alternative criteria (i.e correlation methods, Instrumental Variables) and less on design issues. Ljung's book has become the standard reference book in System Identification (SI). The impact of his book was greatly amplified by the simultaneous production of the MATLAB identification toolbox, which enabled further research in the field of SI. At this moment we can say that theory of system identification for SISO systems is established, and the various identification algorithms have been well tested, and are now available as MATLAB programs. Identification of multiple-input multiple-output (MIMO) systems though is quite more complex and the focus of much work among of which is ours too. The issues that arise in MIMO SI will be addressed in detail in this work and we will present our proposal on how to resolve them.

One of the main implications we face when dealing with the identification of MIMO systems is that if there is no restriction on the form of the matrices we want to estimate, the procedure can determine these matrices up to a linear transformation or affect the convergence of iterative algorithms [36], [40]. To resolve this problem we have to adopt some structural constraints which result on some specific forms on the matrices of the system. These forms are known as *Canonical Forms* and the exact parametrization of which has been extensively under study for some time [14], [63]. After the form of the system matrices is established classical PEM methods involve an iterative procedure aiming to minimize some cost function estimating in each step the new system matrices. The estimation usually applied in this family of algorithms is Least Squares Estimation and its variants. In our work we have managed to apply the Expectation Maximization (EM) Algorithm [25] to a class of identifiable models presented in [91] by formulating the auxiliary function that appears in the EM. To ensure the convergence to the true system matrices we also examine the the convergence of the Steady-State Kalman Gain (SSKG). We also move on to examine the identifiability of a state-space model in its original form as well as its forward innovations representation.

The rest of the thesis is organized as follows: Chapter 1 introduces some basic theory on linear state-space models. We will present the equations that describe a linear state-space model, we will formulate the identification problem and discuss the criteria that are typically used, while also we will give a brief overview of the Kalman Filter. Chapter 2 describes the EM algorithm and we will explain how it is applied in Hidden-state model identification . Chapter 3 examines Canonical Forms presented in various texts and how they affect the identifiability. Our identification algorithm is discussed in Chapter 4 and we will present our experimental results. Chapter 5 outlines interesting directions for future work.

Chapter 1

Linear State-Space Model Identification

1.1 Introduction

This chapter introduces some basic concepts of linear state-space models and their identification. We will describe analytically the formulation of state-space models and explain why we will deal with this class of models instead of other mathematical models that exist. Then we define the problem of identification and present the criteria that are used for identification procedures. We will also make an overview of the Discrete Kalman Filter, which we use for our identification method as it will be shown in the following chapters. Finally, we will describe the basic concept behind PEM and Subspace Methods for Identification.

1.2 State-Space Model Description

It is well established that there is an infinite collection of mathematical systems. In this thesis, we have restricted ourselves to discrete time, linear, time-invariant, state space models. This might seem like a highly restricted class of models (especially the fact they are linear), but, surprisingly enough, many processes can be described very accurately by this type of models and a large variety of scientific areas employ them for modeling, i.e. economics, engineering, biosciences to name a few. Moreover, the number of control system design tools that are available to build a controller based on this type of models, is almost without bound ([18], [32]). For this reason, this model class is a very interesting one.

Mathematically, these models are described by the following set of difference equations:

$$x_{k+1} = Fx_k + Bu_k + w_k \tag{1.2.1a}$$

$$y_k = Hx_k + Du_k + v_k \tag{1.2.1b}$$

$$\mathbf{E}[\begin{pmatrix} w_p \\ v_p \end{pmatrix}(w_p^T v_p^T)] = \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \delta_{pq} \geq 0 \tag{1.2.1c}$$

where $\mathbf{E}$ denotes the expected value operator,A^T denotes the transpose of a matrix and δ_{pq} is the Kronecker delta.

In this model we have([94], [78]):

vectors: The vectors $u_k \in \mathbb{R}^l$ and $y_k \in \mathbb{R}^m$ are the measurements at time instant k of the l inputs and m outputs of the process, respectively. The vector $x_k \in \mathbb{R}^n$ is the state vector of the process at discrete-time instant k and contains the numerical values of n states. Of course, if the system states would have some physical meaning, one could always find a similarity transformation of the state space model to convert the states to physically meaningful ones. Both $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are unmeasurable vector signals, most commonly known as state and measurement noise, respectively. It is assumed that they are zero mean, stationary, white, gaussian noise vector sequences.

matrices: $F \in \mathbb{R}^{n\times n}$ is called the (dynamical) system matrix. It describes the dynamics of the system (as completely characterized by its eigenvalues). $B \in \mathbb{R}^{n\times l}$ is the input matrix which represents the linear transformation by which the deterministic inputs influence the next state. $H \in \mathbb{R}^{m\times n}$ is the output matrix, which describes how the internal state is transferred to the outside world in the measurements y_k. The term with the matrix $D \in \mathbb{R}^{m\times l}$ is called the direct feedthrough term.The matrices $Q \in \mathbb{R}^{n\times n}$, $S \in \mathbb{R}^{n\times m}$ and $R \in \mathbb{R}^{m\times m}$ are the covariance matrices of the noise sequences w_k and v_k. The matrix pair $\{F, H\}$ is assumed to be observable (see Theorem 1.2.2), which implies that all modes in the system can be observed in the output y_k and can thus be identified. The matrix pair $\{F, [B\ Q^{1/2}]\}$ is assumed to be controllable (see Theorem 1.2.1), which in its turn implies that all modes of the system are excited by either the deterministic input u_k and/or the stochastic input w_k.

A graphical representation of the system described by equations 1.2.1 can be found in the following figure.

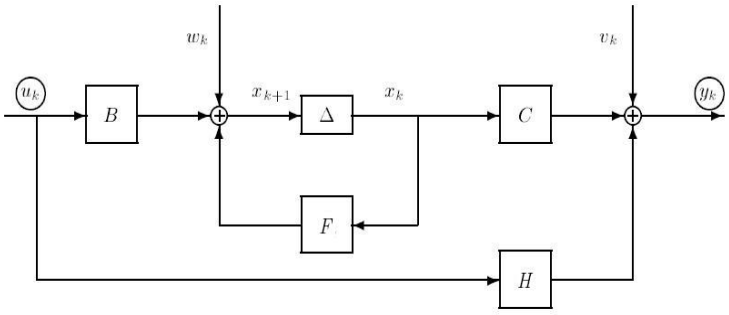

Figure 1.1: Finite Dimension discrete time, linear, time-invariant, state space models ([94]).

Next we present two theorems for checking controllability and observability of a pair of matrices, the proof of which can be found in [64].

Theorem 1.2.1. *For two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ we say that the pair $\{A, B\}$ is completely controllable if and only if the extended $n \times nm$ matrix*

$$M = [B, AB, \cdots, A^{n-1}B]$$

has rank n

Theorem 1.2.2. *For two matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$ we say that the pair $\{A, C\}$ is completely observable if and only if the extended $pn \times n$ matrix*

$$M = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

has rank n

Of course, there are more elaborate and robust methods to check for controllability and observability [69], but for the purposes of this thesis the above theorems are adequate.

In the systems we have examined we have made the assumption that we don't have external input, thus $B = D = 0$ and the cross-covariance of the state and measurement noise is zero, $S = 0$, though the extension is straightforward. Moreover we assume that we don't have degenerate Gaussian noise neither in the state nor the measurement equation (meaning that their respective covariance matrices are positive definite). Hence, the equations 1.2.1 describing the system become:

$$x_{k+1} = Fx_k + w_k \tag{1.2.2a}$$

$$y_k = Hx_k + v_k \tag{1.2.2b}$$

$$\mathbf{E}[\begin{pmatrix} w_p \\ v_p \end{pmatrix}(w_q^T v_q^T)] = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}\delta_{pq} > 0 \tag{1.2.2c}$$

For the rest of the thesis this is the form of the state-space system we have adopted.

1.3 Kalman Filtering

The celebrated *K*alman Filter [52] is an estimator for what is called the *linear-quadratic problem*, which is the problem of estimating the state at time k of a linear dynamical system which is affected by white noise by using measurements linearly related to the state which are also corrupted by white noise. Kalman Filter is an optimal estimator with respect to any quadratic function of error estimation [39]. Moreover, the solution is recursive in the sense that each updated estimate of the state is computed from the previous estimate and the new input data, so only the previous estimate requires storage. In addition to eliminating the need for storing the entire past observed data, the Kalman Filter is computationally more efficient than computing the estimate directly

from the entire past observed data at each step of the filtering process. The original work of Kalman, where the derivation of the filter equations is described, can be found in [52]. Though many different derivations which lead to the same result have been developed over the years, many textbooks choose to follow the original derivation of Kalman due to its elegancy [45], [59]. Kalman Filter is the succesor of the Wiener Filter [99], which introduced the idea of statistically representing signals and Levinson Filter [58], which simplifies some computational aspects of the Wiener theory.

1.3.1 Kalman Filter and RTS-Smoother

Kalman filtering is a recursive procedure that consists of two steps. In the first, the model makes predictions of the state mean and covariance $\hat{x}_{k|k-1}$ and $\Sigma_{k|k-1}$ (prediction step), then in the second, these predictions are updated by projecting them into the observation space by computing the error e_k of the true measurement y_k and the estimated $\hat{y}_k$, and adjusted to give the new estimates $\hat{x}_{k|k}$ and $\Sigma_{k|k}$ (update step). This process provides a means of updating the state distribution as new observations are made.

As we have mentioned the Kalman Filter solves the optimum linear filtering problem but it has been shown in [76] that smoothing the estimates of Kalman Fitler greatly improves its performance. Smoothing is a non-real-time operation in that it involves estimation of the state x_k for $0 < k \leq N$, using all the available data, past as well as future. In fact in [76] the optimum linear smoothing problem was solved and the well known Rauch-Tung-Striebel (RTS) Smoother was introduced. The complete set of equations of Kalman Filter (Forward Recursions) and RTS-Smoother (Backward Recursions) for a state-space model described by 1.2.2 are summarized below:

Forward Recursions

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \left(\Sigma_{k|k-1} H^T \Sigma_{e_k}^{-1}\right) e_k \tag{1.3.1a}$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} \tag{1.3.1b}$$

$$\hat{y}_{k+1|k} = H\hat{x}_{k+1|k} \tag{1.3.1c}$$

$$e_k = y_k - H\hat{x}_{k|k-1} \tag{1.3.1d}$$

$$\Sigma_{e_k} = H\Sigma_{k|k-1}H^T + R \tag{1.3.1e}$$

$$\Sigma_{k|k} = \Sigma_{k|k-1} - \Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}H\Sigma_{k|k-1} \tag{1.3.1f}$$

$$\Sigma_{k,k-1|k} = \left(I - \left(\Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}\right)H\right)F\Sigma_{k-1|k-1} \tag{1.3.1g}$$

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + Q \tag{1.3.1h}$$

Backward Recursions

$$\hat{x}_{k-1|N} = \hat{x}_{k-1|k-1} + A_k \left[\hat{x}_{k|N} - \hat{x}_{k|k-1}\right] \tag{1.3.2a}$$

$$\Sigma_{k-1|N} = \Sigma_{k-1|k-1} + A_k \left[\Sigma_{k|N} - \Sigma_{k|k-1}\right] A_k^T \tag{1.3.2b}$$

$$A_k = \Sigma_{k-1|k-1} F^T \Sigma_{k|k-1}^{-1} \tag{1.3.2c}$$

$$\Sigma_{k,k-1|N} = \Sigma_{k,k-1|k} + \left[\Sigma_{k|N} - \Sigma_{k|k}\right] \Sigma_{k|k}^{-1} \Sigma_{k,k-1|k} \tag{1.3.2d}$$

The quantity $\Sigma_{k,k-1|k}$ is not included in the standard Kalman Filter equations. It was derived in [26] and we present it here since it is involved in the auxiliary function of the EM algorithm, on which we have based our proposed identification method. The term $\hat{y}_{k+1|k}$ is the filtered output and is useful in some applications [57].

The term e_k is called innovations process and a special case of state-space model can be constructed from it, as we will see in Subection 1.3.3. The quantity defined below:

$$K_k = F \Sigma_{k|k-1} H^T \Sigma_{e_k}^{-1} \tag{1.3.3}$$

is called Kalman Gain. In many texts (i.e. [82], [26], [31], [34]) Kalman Gain is defined by the quantity

$$K_k = \Sigma_{k|k-1} H^T \Sigma_{e_k}^{-1} \tag{1.3.4}$$

which appears in the forward and backward recursions in the set of equations 1.3.1, 1.3.2. The benefit of this definition is that it is directly involved in the Kalman Filter equations and is in fact the optimal gain in the sense that minimizes the mean square error of the estimates. We choose to follow the definition of equation 1.3.3 to avoid confusion when we will discuss about the innovations model in Subsection 1.3.3.

There are numerous different implementations of the Kalman Filter. This is attributed to the fact that when the Kalman Filter was implemented, it was discovered that it is prone to numerical instabilities due to short word lengths of computers ([80]) and very sensitive to roundoff errors (more details about roundoff errors can be found in [37]). The state covariance matrix Σ computed in the Kalman Filter should theoretically always be a symmetric positive semi-definite matrix, but numerical problems in computer implementations sometimes led to matrices that became indefinite or nonsymmetric. Hence many alternative implementations of the Kalman Filter were developed to deal with these problems, the most important of which are presented below:

- *Information Filtering.* This is an implementation of the Kalman Filter that propagates the inverse of the state covariance matrix Σ (which is called information matrix). It is used when the dimension of the measurement vector is much larger than the dimension of the state vector. More details about information filtering can be found in ([33]).

- *Sequential Processing.* Here the measurement vector is processed one component at a time, as it is shown in [22]. This implementation is used mostly when the noise of the output measurement vector is block-diagonal or constant.

- *Square-root Filtering.* This implementation was first developed by James Potter for systems without state noise and scalar measurement [53] and later extended for state noise and vector measurements in [6], [13], [28]. Square-root Filtering propagates the root of the state covariance Σ and improves the numerical characteristics of the Kalman Filter (results in twice as much precision) at the cost of computational requirements.

Now that computers have become so much more capable, we do not have to worry about numerical problems as often and the original equations of Kalman Filter are usually implemented [82].

There are variations of the Kalman Filter to non-linear filtering too. The most common variations are :

- *Extended Kalman Filter* (EKF). The main idea of EKF is that we linearize the nonlinear system around the Kalman filter estimate, and the Kalman filter estimate is based on the linearized system. EKF was originally proposed by Stanley Schmidt so that the Kalman filter could be applied to nonlinear spacecraft navigation problems [13].

- *Unscented Kalman Filter* (UKF). UKF is an extension of the Kalman filter that reduces the linearization errors of the EKF by a deterministic sampling approach resulting in a better estimation of the mean and covariance of the state. UKF was first proposed in [49] and further developed in [98].

Despite the existence of non-linear filters, in many applications it is assumed that the system is linear and a version of the original Kalman Filter is implemented. This is attributed to the fact that linear time-invariant (LTI) systems are the simplest and most important class of dynamic systems used in practice and in the literature. Though they are nothing but idealized models, experience shows that they can approximate well many industrial processes [60].

1.3.2 Time-Invariant Kalman Filter

In this section, we are interested in determining the conditions for which that the optimal filter for a model described by equations (1.2.2) is time invariant, or asymptotically time invariant and asymptotically stable, simultaneously. Time invariance or asymptotic time invariance, arises when there is a constant, or asymptotically constant solution of the variance equation :

$$\Sigma_{k+1|k} = F\left[\Sigma_{k|k-1} - \Sigma_{k|k-1}H^T\left(H\Sigma_{k|k-1}H^T + R\right)^{-1}H\Sigma_{k|k-1}\right] + Q \tag{1.3.5}$$

if $\bar{\Sigma}$ is a constant or asymptotically constant solution of the equation (1.3.5) then the associated Kalman Gain is

$$K = F\bar{\Sigma}H^T\left(H\bar{\Sigma}H^T + R\right)^{-1} \tag{1.3.6}$$

If the signal process is stationary then $\Sigma_{k+1|k}$ has a limiting solution [4], however it is not obvious that the associated filter is also asymptotically stable. This question

was answered in [56] and the filter is indeed asymptotically stable and the following conclusions were drawn:

If a state-space model is time-invariant and asymptotically stable, i.e. $|\lambda_i(F)| < 1$ (where $\lambda_i(F)$ are the eigenvalues of F) then:

1. For any nonnegative symmetric initial condition $\Sigma_{k_0|k_{0-1}}$ there exists

$$\lim_{k\to\infty} \Sigma_{k+1|k} = \bar{\Sigma} \tag{1.3.7}$$

with $\bar{\Sigma}$ independent of $\Sigma_{k_0|k_{0-1}}$ and satisfies the steady-state version of 1.3.5:

$$\bar{\Sigma} = F\left[\bar{\Sigma} - \bar{\Sigma}H^T\left(H\bar{\Sigma}H^T + R\right)^{-1}H\bar{\Sigma}\right] + Q \tag{1.3.8}$$

Equation 1.3.8 is known as Discrete Algebraic Riccati Equation (DARE).

2. It is true that

$$|\lambda_i\left(F - KH\right)| < 1 \tag{1.3.9}$$

where $\lambda_i\left(F - KH\right)$ are the eigenvalues of $F - KH$ and K is given by 1.3.6 and is called Steady-State Kalman Gain (SSKG) and is the steady-state version of 1.3.3.

In [8] it was proved that the solution of (1.3.8) is symmetric positive definite matrix if the model is time-invariant and asymptotically stable and vice versa and an efficient algorithm for solving (1.3.8) was implemented.

It is easily deductable that for a time-invariant and asymptotically stable model we can formulate a steady-state version of the Kalman Filter (SSKF) by substituting with the steady-state version of the state covariance on the set of equations (1.3.1), (1.3.2).

1.3.3 Innovations Representation

It is shown in [4] that there is a collection of state-space models with the same Kalman Filter so there is a many-to-one nature of state-space models to Kalman Filter mapping. So the question arises as to whether there is one particular model, among the collection of state-space models, with one-to-one mapping to a Kalman Filter. Indeed there is, this model is called *i*nnovations model, so-called because its input white noise process is identical with the innovations process of the associated filter.

The most important properties of the innovations model are [4]:

1. It is determinable from the covariance data only and is unique
2. The input to the innovations model can be determined from its output.
3. The Kalman Filter can estimate the state of the innovations model with zero error, and the Kalman Filter innovations sequence is identical with the input noise sequence of the innovations model

The equations that describe the innovations model are:

$$x_{k+1} = Fx_k + K_k e_k \qquad x_0 = 0 \tag{1.3.10a}$$

$$y_k = Hx_k + e_k \tag{1.3.10b}$$

$$\mathbf{E}\left[e_p e_q^T\right] = \Lambda\delta_{pq} > 0 \qquad \mathbf{E}\left[e_p\right] = 0 \tag{1.3.10c}$$

where K_k is the Kalman Gain. A method for transforming a state-space model to its innovations representation can be found in [92].

In the previous Subsection 1.3.2 we discussed about time-invariant systems, the conclusions drawn there can be extended in the case of innovations representation too. In fact it is proven ([4]) that a model of the form:

$$x_{k+1} = Fx_k + Ke_k \qquad x_0 = 0 \tag{1.3.11a}$$

$$y_k = Hx_k + e_k \tag{1.3.11b}$$

$$\mathbf{E}\left[e_p e_q^T\right] = \Lambda\delta_{pq} > 0 \qquad \mathbf{E}\left[e_p\right] = 0 \tag{1.3.11c}$$

is an innovations representation if and only if $|\lambda_i(F)| < 1$ and $|\lambda_i(F - KH)| \leq 1$.

The importance of the innovations representation of a model lies to the fact that it is unique given the the covariance Λ and the steady-state Kalman Gain K. This is the reason that many identification methods are applied in the innovations representation of the model instead of its original form [60].

1.4 System Identification

System Identification (SI) is a methodology developed mainly in the area of automatic control, by which we can choose the best model(s) from a given model set based on the observed input-output data from the system. Thus the problem of System Identification is specified by three elements [60] :

- A data set $\mathcal{D}$ obtained by input-output measurements.
- A model set $\mathcal{M}$, or a model structure, containing candidate models.
- A criterion, or loss, function $\mathcal{L}$ to select the best model(s), or a rule to evaluate candidate models, based on the data.

The input-output data $\mathcal{D}$ are collected through experiment. In this case, we must design the experiment by deciding input signals, output signals to be measured, the sampling interval, etc., thereby systems characteristics are well reflected in the observed data. Thus, to obtain useful data for system identification, we should have some *a priori* information, or some physical knowledge, about the system.

A choice of model set $\mathcal{M}$ is a difficult issue in system identification, but usually several classes of discrete-time linear time-invariant (LTI) systems are used. Since these models do not necessarily reflect the knowledge about the structure of the system, they are referred to as *black-box models*. One of the most difficult problems is to find a good model structure, or to fix the order of the model, based on the given input-output data.

A solution to this problem is given by the Akaike Information Criterion (AIC) [1], [2]. Also, by using some physical principles, we can construct models that contain several unknown parameters. These models are called *gray-box* models because some basic laws from physics are employed to describe the dynamics of a system or a phenomenon.

The next step is to find a model in the model set $\mathcal{M}$, by which the experimental data is best explained. Therefore, we need a criterion to measure the distance between a model and a real system, so that the criterion should be of physical meaning and simple enough to be handled mathematically In terms of the input u, the output y of a real system, and the model output y_M, the criterion is usually defined as :

$$V_N = \sum_{n=1}^{N} l\left(y(n), y_M(n), u(n)\right) \tag{1.4.1}$$

where $l(\cdot)$ is a nonnegative loss function, and N the number of data. If the model set is parametrized as $\mathcal{M} = M(\theta), \theta \in \Theta$, then the identification reduces to an optimization problem minimizing the criterion V_N with respect to θ. The following Figure depicts a graphical representation of the System Identification procedure.

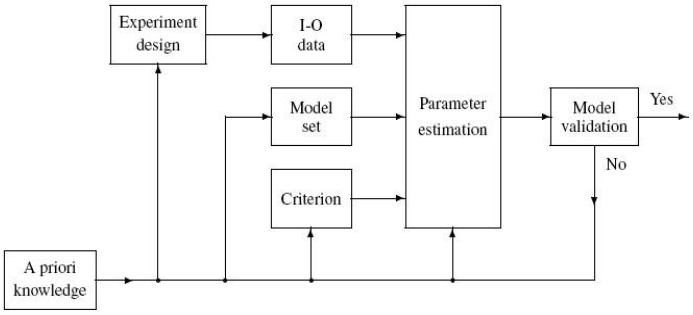

Figure 1.2: A flow diagram of System Identification [84], [60].

It should be emphasized in this point that system identification is a technique of approximating real systems by means of our models since there is no "true" system in practical applications [3].

As we have already mentioned, our work deals with the identification of linear state-space models and we assume that we have a well-defined data set. Hence our interest focuses on the identification criteria, the most common of which we present in the following section.

1.5 Identifiability Criteria

This far we have given a formal and thorough definition of System Identification. Loosely speaking though, the goal of System Identification is to find a unique model that is equal to the "true" system. To this end we have to choose a criterion that expresses the relationship of the estimated model with the true system. The reason we did not included the concept of the identification criterion to the formal definition

is that it is a tool used in experiments when a new identification method is proposed in order to check the validation and robustness of the proposed method. In general there are two such criteria based on different scientific areas, one being closer to control theory and the other closer to pattern recognition area.

Assume we have a model described by equations (1.2.2) which represents the true system:

$$x_{k+1} = Fx_k + w_k$$
$$y_k = Hx_k + v_k$$

with w_k, v_k zero-mean, white, Gaussian noises with covariances Q, R respectively and N the number of our measurements. Let θ be the set of parameters we want to estimate which are incorporated in the system matrices and $F(\hat{\theta}_N), H(\hat{\theta}_N), Q(\hat{\theta}_N), R(\hat{\theta}_N)$ be our estimates of the matrices based on our data. A choice of criterion (from a statistician's perspective) could be that an identification method results uniquely in the true system if [91]:

$$F(\hat{\theta}_N) \to F \quad H(\hat{\theta}_N) \to H \quad Q(\hat{\theta}_N) \to Q \quad R(\hat{\theta}_N) \to R$$
$$\text{as} \quad N \to \infty$$

meaning that $\hat{\theta}_N$ must converge in probability to θ as $N \to \infty$.

Another approach (closer to control theory) for examining if the estimated system have resulted in the true system is by comparing the transfer functions of the estimated system with the true system [60]. Let $H(z, \theta)$ be the transfer function of the true system above and $H(z, \hat{\theta})$ the transfer function of the estimated system, then if:

$$H(z, \theta) \equiv H(z, \hat{\theta})$$
$$\text{for almost all } z \text{ then } \theta = \hat{\theta}$$

So when someone is testing an identification procedure one may use one of the above criterions to observe how well the procedure approximates the true system. Identification criteria are closely related to the concept of identifiability which we will examine at Chapter 3.

In the following sections we will present the basic idea of the two main approaches that dominate System Identification, PEM and Subspace Identification. Though in our work we examine systems without external inputs we will include external inputs, to emphasize that these methods generalize in ARMAX models.

1.6 Prediction Error Methods

Consider an innovations representation of a discrete-time LTI system of the form:

$$x_{k+1} = Fx_k + Bu_k + Ke_k$$
$$y_k = Hx_k + Du_k + e_k$$
$$\mathbf{E}\left[e_p e_q^T\right] = \Lambda\delta_{pq} \qquad \mathbf{E}\left[e_p\right] = 0$$

where $y_k \in \mathbb{R}^m$ is the output vector,$u_k \in \mathbb{R}^p$ is the input vector, $x_k \in \mathbb{R}^n$ is the state vector, $e_k \in \mathbb{R}^m$ is the innovation vector with mean zero and a positive definite covariance matrix $\Lambda > 0$ and F, H, B, D, K are matrices of appropriate dimensions. The unknown parameters in the state space model are contained in these system matrices and covariance matrix Λ of the innovations process. Applying the Kalman Filter on the system we can find the state estimations $\hat{x}_k$ and compute the prediction error $\varepsilon_k(\theta)$ by a linear state-space model of the form

$$\hat{x}_{k+1}(\theta) = [F(\theta) - K(\theta)H(\theta)]\,\hat{x}_k(\theta) + [B(\theta) - K(\theta)D(\theta)]\,u_k + K(\theta)y_k(\theta) \tag{1.6.1a}$$

$$\varepsilon_k(\theta) = -H(\theta)\hat{x}_k(\theta) - D(\theta)u_k + y_k(\theta) \tag{1.6.1b}$$

with initial condition $\hat{x}_0(\theta) = 0$. The formulation of equations 1.6.1 is known as *Whitening Filter* [4]. So, in terms of $\varepsilon_k(\theta)$, the performance index is given by:

$$V_N(\theta) = \sum_{k=0}^{N-1} \|\varepsilon_k(\theta)\|^2$$

where N is the number of data.

Thus the PEM estimates are obtained by minimizing $V_N(\theta)$ with respect to θ, and the covariance matrix Λ of e is estimated by computing the sample covariance matrix of $\varepsilon_k, k \in [0, 1, 2, \ldots, N-1]$.

If we can evaluate the gradient $\frac{\partial V_N}{\partial \theta}$, we can in principle compute a (local) minimum of the criterion $V_N(\theta)$ by utilizing a gradient method. Usually an iterative procedure is followed where based on the new estimates of the system matrices we predict again the states $x_k(\theta^{new})$ and based on minimizing the new cost function V_N^{new} we re-estimate the system matrices using again optimization methods to determine the minimum of the cost function [61]. Optimization methods though need canonical parameterizations and it may be difficult to guess a suitable canonical parametrization from the outset. Since no single continuous parametrization covers all possible multi-variable linear systems, it may be necessary to change parametrization in the course of the optimization routine. Moreover it is well known that for a triplet (m, n, p) there does not exist a unique MIMO state-space model which will be the result of the optimization routine if some form of canonical or pseudo-canonical parametrization is not applied on the system matrices [40], [63], [36]. In Chapter 3 we will examine in more detail why canonical parameterizations are necessary for PEM methods and the effect they have on the identifiability of the systems. Prediction Error Methods are examined in detail in [84] and [60].

1.7 Subspace Identification Methods

Subspace Identification Methods differ from PEM in the sense that they do not attempt to minimize a cost function but instead they enlist linear algebra and geometrical tools to estimate the system matrices [24]. Subspace identification methods are based on the following idea. Suppose that an estimate of a sequence of state vectors of the state

space model (see below) are somehow constructed from the observed input-output data. Then for $k = 0, 1, \cdots, N-1$

$$\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} = \begin{bmatrix} F & B \\ H & D \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \begin{bmatrix} \eta_k \\ \nu_k \end{bmatrix}$$

where $x \in \mathbb{R}^n$ is the estimate of the state vector, $y \in \mathbb{R}^m$ the output vector, $u \in \mathbb{R}^l$ the input vector, F, H, B, D matrices of the appropriate dimensions and η, ν are the residuals. Since we have assumed that we know the state estimates we can find a Least-Squares estimate of $\Theta := \begin{bmatrix} F & B \\ H & D \end{bmatrix}$ from the following formula:

$$\hat{\Theta}_{LS} = \left(\sum_{k=0}^{N-1} \begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} \begin{bmatrix} x_k^T & u_k^T \end{bmatrix} \right) \left(\sum_{k=0}^{N-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k^T & u_k^T \end{bmatrix} \right)^{-1}$$

This class of estimates uniquely exists if the following rank condition is satisfied [38]:

$$rank \begin{bmatrix} x_0 & x_1 & \cdots & x_{N-1} \\ u_0 & u_1 & \cdots & u_{N-1} \end{bmatrix} = n + l$$

Also the covariance matrices of the residuals are given by:

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \frac{1}{N} \sum_{k=0}^{N-1} \begin{bmatrix} \eta_k \\ \nu_k \end{bmatrix} \begin{bmatrix} \eta_k & \nu_k \end{bmatrix}$$

The question that remains to be answered is how we compute the state estimates. A possible answer to this question is by applying LQ Decomposition on block Hankel Matrices defined by the input and output data [94]:

$$U_{0|k-1} = \begin{bmatrix} u_0 & u_1 & \cdots & u_{N-1} \\ u_1 & u_2 & \cdots & u_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k-1} & u_k & \cdots & u_{N+k-2} \end{bmatrix} \in \mathbb{R}^{kp \times N}$$

and

$$Y_{0|k-1} = \begin{bmatrix} y_0 & y_1 & \cdots & uy_{N-1} \\ y_1 & y_2 & \cdots & y_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k-1} & y_k & \cdots & y_{N+k-2} \end{bmatrix} \in \mathbb{R}^{kp \times N}$$

where $k > n$ and N is sufficiently large. For simplicity, let p and f denote the past and future, respectively. Then, we define the past data as $U_p := U_{0|k-1}$ and $Y_p := Y_{0|k-1}$ and

the joint past $W_p^T := [U_p^T \quad Y_p^T]$. Similarly, we define the future data as $U_f := U_{k|2k-1}$ and $Y_f := Y_{k|2k-1}$. We have the following LQ decomposition

$$\begin{bmatrix} U_f \\ W_p \\ Y_f \end{bmatrix} = \begin{bmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \end{bmatrix}$$

where $R_{11} \in \mathbb{R}^{kl \times kl}$,$R_{22} \in \mathbb{R}^{k(m+l) \times k(m+l)}$ and $R_{33} \in \mathbb{R}^{km \times km}$ are upper triangular matrices while $Q_i, i = 1, 2, 3$ are orthogonal matrices. The oblique projection of the future Y_f onto the joint past W_p along the future U_f is given by [94]:

$$\xi = \hat{E}_{\|U_f}\{Y_f|W_p\} = R_{32}R_{22}^{\dagger}W_p$$

where $(\cdot)^{\dagger}$ denotes the pseudo-inverse of a matrix. $\hat{E}_{\|Z}\{x|Y\}$ denotes the oblique projection of x onto Y along Z. Moreover it can be shown ([54]) that ξ can be factored to the extended observability matrix $\mathcal{O}_k$ and the future state vector $X_f := [x_k, x_{k+1}, \cdots, x_{k+N-1}] \in \mathbb{R}^{n \times N}$ thus we have:

$$\xi = \mathcal{O}_k X_f = R_{32}R_{22}^{\dagger}W_p$$

Let the SVD of ξ be given by $\xi = U\Sigma V^T$ with $rank(\Sigma) = n$. The extended observability matrix is $\mathcal{O} = U\Sigma^{1/2}$ [94]. Hence we have $X_f = \mathcal{O}^{\dagger}\xi = \Sigma^{1/2}V^T$ which is the state estimates we wanted to calculate. Of course there are other ways to calculate the state estimates in ARMAX models, in [24] the problem was solved using principal angles while Aoki in [7] used Canonical Correlation Analysis (CCA) to estimate the states. Subspace Identification methods are examined exclusively in [54] and [94].

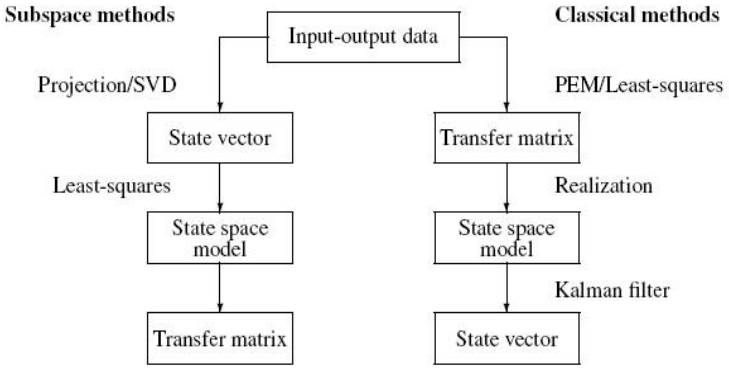

Figure 1.3: Subspace and classical methods of system identification, [54].

In Figure 1.3, we see some differences in the classical and subspace methods of system identification, where the left-hand side is the subspace method, and the right-hand side is the classical optimization-based method. It is interesting to observe the difference in the flow of two approaches; in the classical method, a transfer function model is

first identified, and then a state-space model is obtained by using some realization technique; from the state-space model, we can compute state vectors, or the Kalman Filter state vectors. In subspace methods, we first construct the state estimates from given input-output data by using a procedure based on tools of numerical linear algebra, and a state-space model is obtained by solving a least-squares problem.

1.8 Summary

In this chapter we described general state-space model and their relation to Kalman Filter. We presented the equations of the filter and some alternative implementations while we also talked about the innovations representation and emphasized that it is unique for every state-space system. Also we described the steady-state case of model and how it affects the corresponding Kalman Filter. We mentioned the concept of identification criterion and finally presented the basic idea behind the two approaches that govern System Identification Theory.

Chapter 2

Identification through the EM Algorithm

2.1 Introduction

In this chapter we will view the Expectation Maximization (EM) Algorithm and how it is used to maximize a likelihood function. The main difference of EM from classical Maximum Likelihood (ML) optimization is that EM is an iterative procedure that increases the likelihood function at each iteration by maximizing an auxiliary function. Auxiliary functions are categorized as *strong-sense* or *weak-sense* and we will explain this distinction. Furthermore we will demonstrate analytically how we can use the EM Algorithm for linear state-space model identification and the role of Kalman Filter in this procedure.

2.2 Expectation Maximization Algorithm

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood estimate when we have missing or hidden data and was first introduced in its current context in [25], though it had appeared in many forms previously. In Maximum Likelihood estimation, we wish to estimate the model parameters for which the observed data are the most "likely".

There are two main applications of the EM algorithm. The first occurs when there are missing (or hidden) values from the data. The second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of values for additional but missing (or hidden) parameters [16]. In both cases, maximization of the likelihood function is very complicated or not feasible at all, thus an auxiliary function $\mathcal{Q}$ is attempted to be maximized. The basic idea in the expectation maximization or EM algorithm, is to iteratively estimate the likelihood given the data that is present and consists of two steps: The Expectation step (E-step), and the Maximization step (M-step). During the first step, the expectation (E) step, the expected log-likelihood of the complete data (by complete we mean both the observed and the missing components of the data) is calculated based on the observed data and the current parameter estimates, thus the

auxiliary function is formed which serves as a new likelihood function. In the M-step, this likelihood function is maximized with respect to the parameters we want to estimate. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

We will give a description of the algorithm as given in [27] on an example of corrupted data. Consider a dataset $D = \{x_1, \ldots, x_N\}$ drawn from a single $d-dimensional$ distribution. Suppose that some features are corrupted thus any sample point can be written as $x_k = \{x_{kg}, x_{kb}\}$, a combination of the "good" features and the missing or "bad" ones. We separate these features into two sets, D_g and D_b with $D = Dg \cup Db$ being the union of such features. Next we form the function

$$\mathcal{Q}(\theta; \theta^i) = \mathbf{E}_{D_b}[\ln p(D_g, D_b; \theta | D_g; \theta^i)] \tag{2.2.1}$$

where $\mathcal{Q}$ represents the auxiliary function of the EM and $\mathbf{E}$ the expectation operator. The use of the semicolon denotes, for instance on the left hand side, that $\mathcal{Q}(\theta; \theta_i)$ is a function of θ with θ_i assumed fixed, on the right hand side it denotes that the expected value is over the missing features assuming θ_i are the true parameters describing the distribution. Simply stated, the parameter vector θ_i is the current best estimate for the distribution, θ is a candidate vector for an improved estimate. Given such a candidate θ, the right hand side of 2.2.1 calculates the likelihood of the data, including the unknown feature D_b *marginalized* with respect to the current best distribution, which is described by θ^i. Different candidate θs will of course lead to different such likelihoods. The EM algorithm will select the best such candidate θ and call it θ^{i+1} which is the one corresponding to the greatest $\mathcal{Q}(\theta; \theta^i)$. In general the steps of the EM algorithm with i representing the iteration counter and T a convergence threshold are described in [27] as:

1. begin initialize $\theta_0, T, i = 0$
2. do $i \leftarrow i + 1$
3. **E-Step** compute $\mathcal{Q}(\theta; \theta^i)$
4. **M-Step** $\theta^{i+1} \leftarrow \arg\max_{\theta} \mathcal{Q}(\theta; \theta^i)$
5. until $\mathcal{Q}(\theta^{i+1}; \theta^i) - \mathcal{Q}(\theta^i; \theta^{i-1}) \leq T$
6. return $\hat{\theta} \leftarrow \theta^{i+1}$
7. end

It can be shown that the successive estimates θ^i never decrease the likelihood function, the likelihood function keeps increasing until a maximum (local or global) is reached and the EM algorithm converges [25], [100], [19]. Theoretical results as well as practical experimentation confirm that the convergence is slower than the quadratic convergence of Newton-type searching algorithms, although near the optimum a speedup may be possible. However, the great advantage of the algorithm is that its convergence is smooth and is not vulnerable to instabilities. Furthermore, it is computationally more

attractive than Newton-like methods, which require the computation of the Hessian matrix [87].

Two of the most important problems in statistical estimation are solved through the EM algorithm (which indicates its importance). The first is the estimation of the parameters of a Gaussian Mixture Model (GMM), the weights, mean vectors and covariance matrices [87], [16], [27]. Though of course there are methods to estimate GMM parameters,, EM is the dominant estimation method. The second problem is the estimation of the parameters of a Hidden Markov Model (HMM), the transition Matrix and the output probabilities (which often expressed as a GMM) and this is the well known Baum-Welch Algorithm which is actually an implementation of the EM Algorithm [75], [47].

One might wonder about the reason behind EM's properties of convergence and robustness. The answer is that the auxiliary function that is used in the EM is a strong-sense auxiliary function. If a function $\mathcal{F}(\theta)$ is to be maximized, then $\mathcal{Q}(\theta, \theta^i)$ is a strong-sense auxiliary function for $\mathcal{F}(\theta)$ if and only if:

$$\mathcal{Q}(\theta, \theta^i) - \mathcal{Q}(\theta^i, \theta^i) \leq \mathcal{F}(\theta) - \mathcal{F}(\theta^i) \tag{2.2.2}$$

as stated in [73], [74].

This property holds for the auxiliary function that is used in the EM Algorithm. The idea is illustrated in Figure 2.1-(a). A maximum with respect to θ of the function $\mathcal{Q}(\theta, \theta^i)$ is found indicated by the arrow. If this increases $\mathcal{Q}$ (the lower line) then it will also increase $\mathcal{F}$ and if $\mathcal{Q}$ is at a local maximum then $\mathcal{F}$ is at a local maximum too. These conditions follow from (2.2.2) and imply that repeated maximization of the auxiliary function $\mathcal{Q}$ is guaranteed to reach a local maximum of $\mathcal{F}$ which makes EM (who uses a strong-sense auxiliary function) such an attractive choice.

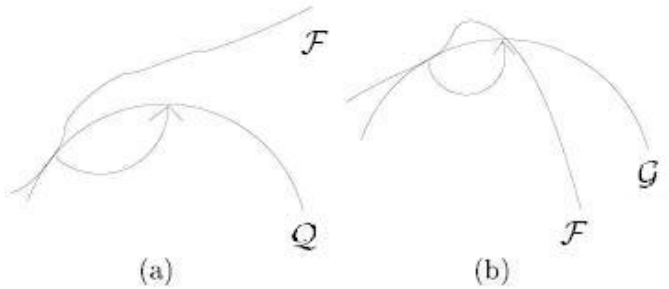

Figure 2.1: Use of (a) strong-sense and (b) weak-sense auxiliary functions for function optimization, [73].

The difference of a weak-sense with a strong-sense auxiliary function is that a weak-sense auxiliary function for $\mathcal{F}(\theta)$ around θ^i is a smooth function $\mathcal{G}(\theta, \theta^i)$ such that:

$$\frac{\partial}{\partial\theta}\mathcal{G}(\theta, \theta^i)\Big|_{\theta=\theta^i} = \frac{\partial}{\partial\theta}\mathcal{F}(\theta)\Big|_{\theta=\theta^i} \tag{2.2.3}$$

The idea is shown in Figure 2.1-(b). The gradients of the two functions are the same around the point θ^i. Maximization of the function $\mathcal{G}$ with respect to θ though does

not guarantee an increase in $\mathcal{F}$. However, if there is no change in θ after maximization of a particular iteration, this implies that we have reached a local maximum of $\mathcal{F}$ (the gradient is zero at that point). If the update converges it will be to a local maximum. The weak-sense auxiliary function condition of (2.2.3) can be considered a minimum condition for an auxiliary function used for optimization [73]. The usefulness of weak-sense auxiliary functions lie to the fact that they can be used to modify procedures based on strong-sense auxiliary functions as the EM Algorithm.

In the next section we will see how we can apply the EM Algorithm to a linear time-invariant state-space model in order to estimate its system parameters.

2.3 Hidden-state model identification using the EM algorithm

Assume that a sequence of observations $Y = [y_0, y_1, \ldots, y_N]$ is generated by the finite-dimensional linear state-space model:

$$x_{k+1} = Fx_k + w_k \tag{2.3.1a}$$

$$y_k = Hx_k + v_k \tag{2.3.1b}$$

$$E[w_p w_q^T] = Q\delta_{pq} \tag{2.3.1c}$$

$$E[v_p v_q^T] = R\delta_{pq} \tag{2.3.1d}$$

where δ_{qp} is the Kronecker delta, the state x is a $n \times 1$ vector, the observation y is $m \times 1$ vector and w_k,v_k are uncorrelated zero-mean Gaussian vectors noise vectors with covariances defined by (2.3.1c) and (2.3.1d) respectively. We further assume that the initial state x_0 is Gaussian with known mean and covariance x_0 $N(\mu_0\Sigma_0)$. Maximum likelihood estimates of the unknown parameters θ in F, H, Q, R can be obtained by minimizing the negative log likelihood or equivalently the quantity [41]:

$$J(Y,\theta) = -L(Y,\theta) = \sum_{k=0}^{N} \{\log \Sigma_{e_k}(\theta) + e_k^T(\theta)\Sigma_{e_k}^{-1} e_k(\theta)\} + \text{constant} \tag{2.3.2}$$

where A^T, A^{-1} denotes the transpose and inverse of the matrix A respectively. The terms $e_k(\theta), \Sigma_{e_k}$ are the prediction error and its covariance and can be obtained from the Kalman Filter equations (see 1.3.1). The minimization of (2.3.2) with respect to θ requires the computation of the gradient and perhaps the Hessian. The quantities that must be computed for this purpose are the state sensitivities with respect to each one of the system parameters which is a complex procedure as shown in [41].

An alternative approach through the EM framework can indeed simplify the problem. Consider for the moment the following slightly modified estimation problem, where we assume that the state of the system described by equations (2.3.1) is not hidden and we want to find the ML estimates of the system parameters θ given $Y = [y_0, y_1, \ldots, y_N]$ and $X = [x_0, x_1, \ldots, x_N]$. In this case the ML estimates of θ

are obtained by maximizing:

$$\begin{aligned} L(X,Y,\theta) = &-\sum_{k=1}^{N}\left\{log|Q| + (x_k - Fx_{k-1})^T Q^{-1}(x_k - Fx_{k-1})\right\} \\ &-\sum_{k=0}^{N}\left\{log|R| + (y_k - Hx_k)^T R^{-1}(y_k - Hx_k)\right\} + \text{constant} \end{aligned} \tag{2.3.3}$$

since, without loss of generality, w_k and v_k were assumed uncorrelated white Gaussian noise sources and the base of log is e.

A critical observation here that enables us to apply the EM Algorithm is that the original problem can be treated as one with incomplete data with the state vector playing the role of missing observations. In this case, the auxiliary function of the EM becomes [26]:

$$\mathcal{Q}(\theta^{i+1},\theta) = \mathbf{E}\left\{L(X,Y,\theta^{i+1})|Y,\theta^i\right\} \tag{2.3.4}$$

which is the conditional expectation of $L(X,Y,\theta)$ defined by equation (2.3.3) given the observed data Y and the current parameter estimations θ^i. It can be shown [25] that the EM algorithm for the exponential family, as is our case under the Gaussian assumption, reduces to computing the conditional expectations of the complete data sufficient statistics during the E-step and using these in place of the complete-data sufficient statistics in the M-step in order to compute the ML estimates.

At this point we give the definition of differentiation of a scalar function with respect to a matrix and some basic properties of such derivatives as presented in [42]. Consider the $n \times m$ matrix $X = [x_{ij}]$ and a scalar function $f(x)$, then the derivative of $f(x)$ with respect to the matrix $X = [x_{ij}]$ is a matrix with elements the partial derivatives of $f(x)$ with respect to the matrix elements of corresponding position.

$$\frac{\partial f(x)}{\partial X} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_{11}} & \frac{\partial f(x)}{\partial x_{12}} & \cdots & \frac{\partial f(x)}{\partial x_{1m}} \\ \frac{\partial f(x)}{\partial x_{21}} & \frac{\partial f(x)}{\partial x_{22}} & \cdots & \frac{\partial f(x)}{\partial x_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x)}{\partial x_{n1}} & \frac{\partial f(x)}{\partial x_{n2}} & \cdots & \frac{\partial f(x)}{\partial x_{nm}} \end{bmatrix}$$

Using this definition, it is easy to prove that $\frac{\partial(f(x)+g(x))}{\partial X} = \frac{\partial f(x)}{\partial X} + \frac{\partial g(x)}{\partial X}$ (addition property)[42]. It also follows that for vectors a, b and matrices X, D (where D is symmetric) of appropriate dimensions the following equations are true

$$\begin{aligned} \frac{\partial a^T X b}{\partial X} &= ab^T \qquad \frac{\partial a^T X^T b}{\partial X} = ba^T \\ \frac{\partial a^T X^T D X b}{\partial X} &= D^T X a b^T + D X b a^T \qquad \frac{\partial \log|X|}{\partial X} = X^{-1} \end{aligned} \tag{2.3.5}$$

More formulas of matrix derivatives can be found in [70].

Assuming that there are no constraints on the structure of the system matrices we take the partial derivatives of (2.3.3) with respect to the matrices of interest and set to zero. Below we will present the derivation of $\hat{H}$ and $\hat{R}$ and by the same manner the

estimates of $\hat{F}$ and $\hat{Q}$ are computed. Since the first sum and the constant of (2.3.3) are not dependent of H we neglect those terms (their derivatives with respect to H is zero).

$$\frac{\partial L(X,Y,\theta)}{\partial H} = 0 \Rightarrow -\frac{\partial}{\partial H}\sum_{k=0}^{N}\left\{\log|R| + (y_k - Hx_k)^T R^{-1}(y_k - Hx_k)\right\} = 0 \Rightarrow$$

$$\frac{\partial}{\partial H}\sum_{k=0}^{N}\left\{\log|R| + y_k^T R^{-1} y_k - y_k^T R^{-1} H x_k - x_k^T H^T R^{-1} y_k + x_k^T H^T R^{-1} H x_k\right\} = 0 \Rightarrow$$

$$\sum_{k=0}^{N}\left\{-\frac{\partial}{\partial H} y_k^T R^{-1} H x_k - \frac{\partial}{\partial H} x_k^T H^T R^{-1} y_k + \frac{\partial}{\partial H} x_k^T H^T R^{-1} H x_k\right\} = 0 \Rightarrow$$

$$\sum_{k=0}^{N}\left\{-R^{-1} y_k x_k^T - R^{-1} y_k x_k^T + 2R^{-1} H x_k x_k^T\right\} = 0 \Rightarrow$$

$$-2R^{-1}\sum_{k=0}^{N} y_k x_k^T + 2R^{-1} H \sum_{k=0}^{N} x_k x_k^T = 0 \Rightarrow$$

$$\hat{H} = \left[\sum_{k=0}^{N} y_k x_k^T\right]\left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1}$$

To obtain this result, we have used the addition property, the formulae in (2.3.5), the fact that R^{-1} is a positive definite symmetric matrix (since R is the covariance matrix of a gaussian random variable) and the observation that $\frac{\partial}{\partial H}\log|R| = 0$. We have also assumed that we have enough data N for $\sum_{k=0}^{N} x_k x_k^T$ to be full rank (thus becoming invertible).

Now that we have find $\hat{H}$ we move on to find an estimate for R. Again, we will take partial derivative of (2.3.3) neglecting the first sum and the constant. Maximizing (2.3.3) with respect to the elements of R is equivalent to maximizing with respect to the elements of its inverse [26]. Thus, we have:

$$\frac{\partial L(X,Y,\theta)}{\partial R^{-1}} = 0 \Rightarrow -\frac{\partial}{\partial R}\sum_{k=0}^{N}\left\{\log|R| + (y_k - \hat{H}x_k)^T R^{-1}(y_k - \hat{H}x_k)\right\} = 0 \Rightarrow$$

$$\frac{\partial}{\partial R^{-1}}\sum_{k=0}^{N}\log|R|^{-1} - \frac{\partial}{\partial R^{-1}}\sum_{k=0}^{N}\left\{(y_k - \hat{H}x_k)^T R^{-1}(y_k - \hat{H}x_k)\right\} = 0 \Rightarrow$$

$$N\frac{\partial}{\partial R^{-1}}\sum_{k=0}^{N}\log|R^{-1}| - \sum_{k=0}^{N}\left\{(y_k - \hat{H}x_k)(y_k - \hat{H}x_k)^T\right\} = 0 \Rightarrow$$

$$(N+1)R = \left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right]\hat{H}^T - \hat{H}\left[\sum_{k=0}^{N} x_k y_k^T\right] + \hat{H}\left[\sum_{k=0}^{N} x_k x_k^T\right]\hat{H}^T = 0 \Rightarrow$$

$$(N+1)R = \left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right] \hat{H}^T - \hat{H}\left[\sum_{k=0}^{N} x_k y_k^T\right] + \hat{H}\left[\sum_{k=0}^{N} x_k x_k^T\right] \hat{H}^T = 0 \Rightarrow$$

$$\begin{aligned}(N+1)R = &\left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} y_k x_k^T\right]^T \\ &- \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} x_k y_k^T\right] \\ &+ \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} x_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} y_k x_k^T\right]^T = 0 \Rightarrow\end{aligned}$$

$$\begin{aligned}(N+1)R = &\left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} y_k x_k^T\right]^T \\ &+ \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} y_k x_k^T\right]^T \\ &- \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} x_k y_k^T\right]\end{aligned}$$

$$(N+1)R = \left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1} \left[\sum_{k=0}^{N} x_k y_k^T\right]$$

$$(N+1)R = \left[\sum_{k=0}^{N} y_k y_k^T\right] - \left[\sum_{k=0}^{N} y_k x_k^T\right] \left(\left[\sum_{k=0}^{N} y_k x_k^T\right] \left[\sum_{k=0}^{N} x_k x_k^T\right]^{-1}\right)^T$$

$$R = \frac{1}{N+1}\left[\sum_{k=0}^{N} y_k y_k^T\right] - \frac{1}{N+1}\left[\sum_{k=0}^{N} y_k x_k^T\right] \hat{H}^T$$

To reach this result first we used the logarithm property $a \log x = \log x^a$, we also exploited the fact that $\log |R|^{-1} = \log |R^{-1}|$, this is easy to prove[1] since $\log |R^{-1}| = \log \frac{1}{|R|} = -\log |R| = \log |R|^{-1}$. Moreover, we facilitated the properties 2.3.5 and observed that (under the assumption that N is large to satisfy invertibility) that $\sum_{k=0}^{N} x_k x_K^T$ is symmetric. Finally, we used some properties of a transpose of a matrix such as $(A+B)^T = A^T + B^T$ and $(AB)^T = B^T A^T$, as well as the fact that the transpose of a symmetric matrix is the matrix itself.

The derivation of F, Q are made in a similar manner. Assuming that there are no

[1] we remind that $|AA^{-1}| = |I| \Rightarrow |A||A^{-1}| = 1 \Rightarrow |A^{-1}| = \frac{1}{|A|}$

constraints on the structure of the matrices F, H, Q, R the estimates are:

$$\hat{F} = \Gamma_4 \Gamma_3^{-1} \tag{2.3.7a}$$

$$\hat{Q} = \Gamma_2 - \Gamma_4 \Gamma_3^{-1} \Gamma_4^T = \Gamma_2 - \Gamma_4 \hat{F}^T \tag{2.3.7b}$$

$$\hat{H} = \Gamma_6 \Gamma_1^{-1} \tag{2.3.7c}$$

$$\hat{R} = \Gamma_5 - \Gamma_6 \Gamma_1^{-1} \Gamma_6^T = \Gamma_5 - \Gamma_6 \hat{H}^T \tag{2.3.7d}$$

where the sufficient statistics are [26]:

$$\Gamma_1 = \frac{1}{N+1} \sum_{k=0}^{N} x_k x_k^T \tag{2.3.8a}$$

$$\Gamma_2 = \frac{1}{N} \sum_{k=1}^{N} x_k x_k^T \tag{2.3.8b}$$

$$\Gamma_3 = \frac{1}{N} \sum_{k=1}^{N} x_{k-1} x_{k-1}^T \tag{2.3.8c}$$

$$\Gamma_4 = \frac{1}{N} \sum_{k=1}^{N} x_k x_{k-1}^T \tag{2.3.8d}$$

$$\Gamma_5 = \frac{1}{N+1} \sum_{k=0}^{N} y_k y_k^T \tag{2.3.8e}$$

$$\Gamma_5 = \frac{1}{N+1} \sum_{k=0}^{N} y_k x_k^T \tag{2.3.8f}$$

As we have mentioned earlier in the text, the EM algorithm for the exponential family as is our case under the Gaussian assumption, reduces to computing the conditional expectations of the complete data sufficient statistics during the E-step and using these in place of the complete-data sufficient statistics in the M-step in order to compute the ML estimates. In simpler words, this means that the EM estimates are given by 2.3.7 but we have to find the expected values of the sufficient statistics (2.3.8). Thus, the problem of maximizing the auxiliary function (2.3.4) has reduced to calculating the quantities at each iteration i:

$$\mathbf{E}\left\{ y_k x_k^T | Y, \theta^i \right\} = y_k E\left\{ x_k^T | Y, \theta^i \right\} \tag{2.3.9a}$$

$$\mathbf{E}\left\{ y_k y_k^T | Y, \theta^i \right\} = y_k y_k^T \tag{2.3.9b}$$

$$\mathbf{E}\left\{ x_k x_k^T | Y, \theta^i \right\} \tag{2.3.9c}$$

$$\mathbf{E}\left\{ x_k x_{k-1}^T | Y, \theta^i \right\} \tag{2.3.9d}$$

Now, since the input process is Gaussian, then the state process will also be Gaussian (since we have assumed that the initial state x_0 follows a Gaussian distribution and that the sum of Gaussian random variables is also a Gaussian random variable), furthermore

the conditional distribution of the state of the system given the observations on a fixed interval is Gaussian [23]:

$$x_k \sim N(\hat{x}_{k|N}; \Sigma_{k|N})$$

Thus, the statistics in (2.3.9) at iteration i become:

$$\mathbf{E}\left\{y_k x_k^T | Y, \theta^i\right\} = y_k E\left\{x_k^T | Y, \theta^i\right\} \tag{2.3.10a}$$

$$\mathbf{E}\left\{y_k y_k^T | Y, \theta^i\right\} = y_k y_k^T \tag{2.3.10b}$$

$$\mathbf{E}\left\{x_k | Y, \theta^i\right\} = \hat{x}_{k|N} \tag{2.3.10c}$$

$$\mathbf{E}\left\{x_k x_k^T | Y, \theta^i\right\} = \Sigma_{k|N} + \hat{x}_{k|N}\hat{x}_{k|N}^T \tag{2.3.10d}$$

$$\mathbf{E}\left\{x_k x_{k-1}^T | Y, \theta^i\right\} = \Sigma_{k,k-1|N} + \hat{x}_{k|N}\hat{x}_{k-1|N}^T \tag{2.3.10e}$$

The fixed interval smoothing form of the Kalman Filter, the RTS Smoother, can be applied here to compute the required statistics. It consists of a backward pass that follows the standard Kalman Filter forward recursions. The Kalman Filter and RTS Smoother are described by equations (1.3.1) and (1.3.2) in Subsection 1.3.1. For convenience, we present them again here.

Forward Recursions

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \left(\Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}\right)e_k \tag{2.3.11a}$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} \tag{2.3.11b}$$

$$e_k = y_k - H\hat{x}_{k|k-1} \tag{2.3.11c}$$

$$\Sigma_{e_k} = H\Sigma_{k|k-1}H^T + R \tag{2.3.11d}$$

$$\Sigma_{k|k} = \Sigma_{k|k-1} - \Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}H\Sigma_{k|k-1} \tag{2.3.11e}$$

$$\Sigma_{k,k-1|k} = \left(I - \left(\Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}\right)H\right)F\Sigma_{k-1|k-1} \tag{2.3.11f}$$

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + Q \tag{2.3.11g}$$

Backward Recursions

$$\hat{x}_{k-1|N} = \hat{x}_{k-1|k-1} + A_k\left[\hat{x}_{k|N} - \hat{x}_{k|k-1}\right] \tag{2.3.12a}$$

$$\Sigma_{k-1|N} = \Sigma_{k-1|k-1} + A_k\left[\Sigma_{k|N} - \Sigma_{k|k-1}\right]A_k^T \tag{2.3.12b}$$

$$A_k = \Sigma_{k-1|k-1}F^T\Sigma_{k|k-1}^{-1} \tag{2.3.12c}$$

$$\Sigma_{k,k-1|N} = \Sigma_{k,k-1|k} + \left[\Sigma_{k|N} - \Sigma_{k|k}\right]\Sigma_{k|k}^{-1}\Sigma_{k,k-1|k} \tag{2.3.12d}$$

We remind that the term $\Sigma_{k,k-1|k}$ and its smoothed form $\Sigma_{k,k-1|N}$ are not involved neither in the standard Kalman Filter equations nor the RTS Smoother. The term was first calculated in [26] and its derivation can be found there.

To summarize, the EM Algorithm involves at each iteration the computation of the sufficient statistics described in equations (2.3.10) using the recursions of the Kalman Filter (2.3.11) and the RTS Smoother (2.3.12) and the old estimates of the model parameters (E-step). The new estimates for the system parameters can then be obtained from these statistics as the simple multivariate regression coefficients given in (2.3.7) (M-step) as described in [26].

2.4 Summary

In this Chapter, we presented the Expectation Maximization Algorithm and analyzed its basic properties as well as the importance of the (strong-sense) auxiliary function it is involved in the procedure, its formulation on the E-step and then its maximization on the M-step. Then we showed how the EM Algorithm is applied in identification of a general state-space model when its matrices are unconstrained. We showed how we can get the ML estimates of the parameters through the EM by computing the sufficient statistics through the Kalman Filter and the RTS Smoother. In the next chapter, we will deal with systems which matrices that are in Canonical Form, examine the concept of identifiability based on these forms, while in Chapter 4 we show how we can apply the EM to a family of systems whose matrices are in a specific canonical form.

Chapter 3

Canonical Forms and Identifiability

3.1 Introduction

In this chapter we deal with the matter of identifiability and its relation to canonical forms. We will present some canonical forms that were developed in various works and examine under which identification criteria these forms are identifiable. We also show how to construct canonical forms based on canonical parameter sets and how to extract those parameters from a state-space model in its general form.

3.2 Canonical Parameter Sets,Forms and Pseudo-Canonical Forms

We have already mentioned in Introduction that if there is no restriction on the form of the matrices we want to estimate, the procedure can determine these matrices up to a linear transformation. Kalman in [51] showed that for a deterministic system of the form:

$$x_{k+1} = Fx_k + Bu_k \tag{3.2.1a}$$

$$y_k = Hx_k \tag{3.2.1b}$$

there are many different triplets (F, H, B) which can produce a given data set. Two systems of the form 3.2.1 with F, B controllable pair and F, H observable pair have the same behavior (transfer function, impulse response) if and only if there exists an invertible matrix T such that:

$$F_1 = TF_2T^{-1} \quad B_1 = TB_2 \quad H_1 = H_2T^{-1} \tag{3.2.2}$$

Therefore, external measurements determine an equivalence class of systems in form (3.2.1). (F_1, H_1, B_1) and (F_2, H_2, B_2) are equivalent if and only if (3.2.2) is satisfied. In general (if the matrices do not have a specific structure) the equivalence class includes more than one member and identification of F, H, B is impossible.

However, if F, H are constrained to specific forms then the equivalence class will have only one member (or equivalently $T = I$, the identity matrix) [14], [63], [40]. Moreover in [65], [91] it is proven that for a deterministic system of the form (3.2.1), constraining only F, H and not B to follow specific forms, is a sufficient condition for the equivalence class to contain only one member. These constrained forms are commonly known as canonical forms, when constructed from a canonical parameter set, or pseudo-canonical forms otherwise, though in general the term canonical form is most commonly used even if it is not based on canonical parameter sets describing the structure of the matrices [91].

According to [91], a Canonical Parameter Set (CPS) $\mathcal{S}$ for the pair (F, H), is an ordered set of numbers, uniquely determined by (F, H), with the following properties:

- **Invariance**: $\mathcal{S}$ remains unchanged if (F, H) is replaced by (TFT^{-1}, HT^{-1}), where T is a nonsingular matrix.
- **Independence**: for any $\mathcal{S}$ there is an observable pair (F, H) whose cps is equal to $\mathcal{S}$
- **Completeness**: if (F_1, H_1) and (F_2, H_2) have the same cps $\mathcal{S}$ there exists a nonsingular matrix T such that $F_1 = TF_2T^{-1}$,$H_1 = H_2T^{-1}$

The independence property implies that there are no fixed relations among the elements of $\mathcal{S}$. There are many different types of canonical parameter sets associated with linear systems. Two of these will be described in the following subsections.

3.2.1 Canonical Parameter Set Type A

This canonical parameter set was first developed in [72] by Popov. Consider a controllable and observable system with state vector of dimension $n \times 1$, observation vector of dimension $m \times 1$ and the ordered set of vectors:

$$h_1, h_2, \ldots, h_m, h_1F, h_2F, \ldots, h_mF, h_1F^2, h_2F^2, \ldots, h_mF^2, \ldots \tag{3.2.3}$$

where h_i is the i^{th} row of H. For $i = 1, 2, \ldots, m$ let n_i be the smallest non-negative integer such that $c_iF^{n_i}$ is linearly dependent on its antecedents (vectors to its left) in (3.2.3). A vector c_jF^k is called a regular vector if and only if $k < n_j$. It is clear that every nonsingular vector in (3.2.3) is a linear combination of its regular antecedents and that regular vectors are linearly dependent. This fact implies that there is a unique set of numbers $\{\alpha_{ijk}\}$ such that, for $i = 1, 2, \ldots, m$:

$$h_iF^{n_i} = \sum_{j=1}^{i-1} \sum_{k=0}^{\min(n_i, n_j-1)} \alpha_{ijk} h_j F^k + \sum_{j=1}^{m} \sum_{k=0}^{\min(n_i, n_j)-1} \alpha_{ijk} h_j F^k \tag{3.2.4}$$

Obviously if $n_i = 0$ 3.2.4 becomes:

$$h_i = \sum_{j \in J} \alpha_{ij0} h_j, \quad J = \{j : 1 \leq j \leq i-1 \text{ and } n_j > 0\} \tag{3.2.5}$$

Moreover the observability condition implies that

$$n_1 + n_2 + \ldots + n_m = n \tag{3.2.6}$$

In [72], [91] the ordered set $\{n_i, \alpha_{ijk}\}$ is proved to be a canonical parameter set for F, H. An example follows of how we can compute the the CPS for a specific matrix pair (F, H).

Consider the following matrices:

$$F = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad H = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

In this case, it easy to observe that h_2 is linearly dependent on h_1 thus $n_2 = 0$. Moreover $h_1 = h_2 = [1\ 1\ 0]$, $h_1F = h_2F = [1\ -1\ 1]$, $h_1F^2 = h_2F^2 = [0\ 0\ -1]$ and $h_1F^3 = [-1\ -1\ 0]$ which makes it the first linear dependent vector for $i = 1$ on its antecedents in (3.2.3) thus $n_1 = 3$. Since we have found $n_1 = 3$ and $n_2 = 0$ equations (3.2.4), (3.2.5) become:

$$\begin{aligned} h_1F^3 &= \alpha_{110}h_1 + \alpha_{111}h_1F + \alpha_{112}h_1F^2 \\ h_2 &= \alpha_{210}h_1 \end{aligned}$$

Now, all that is left is to solve the linear system and since $h_1F^3 = \begin{bmatrix} -1 & -1 & 0 \end{bmatrix}$ the following equations determine α_{ijk}:

$$\begin{aligned} -1 &= \alpha_{110} + \alpha_{111} \\ -1 &= \alpha_{110} - \alpha_{111} \\ 0 &= \alpha_{111} - \alpha_{112} \\ 1 &= \alpha_{210} \end{aligned}$$

Thus, $\alpha_{110} = -1, \alpha_{111} = 0, \alpha_{112} = 0, \alpha_{210} = 1$.

3.2.2 Canonical Parameter Set Type B

This canonical parameter set was developed in [91] and is similar to the canonical parameter set we described in Subsection 3.2.1. Consider the vectors in (3.2.3) in a different order and name:

$$h_1, h_1F, \ldots, h_1F^{p_1-1}, h_2, h_2F, \ldots, h_2F^{p_2-1}, \ldots, h_m, h_mF, \ldots, h_mF^{p_m-1} \tag{3.2.7}$$

where $p_i,\ i = 1, 2, \ldots, m$ is the smallest non-negative integer such that $h_iF^{p_i}$ is linearly dependent on its antecedents in (3.2.7).

As before, every non-regular vector in (3.2.7) can be uniquely expressed as linear combination of its regular antecedents. Thus, there is a unique ordered set of numbers

β_{ijk} for $i = 1, 2, \ldots, m$ such that:

$$\begin{aligned} h_i F^{p_i} &= \sum_{j=1}^{i} \sum_{k=0}^{p_j - 1} \beta_{ijk} h_i F^k \text{ if } p_i > 0 \\ h_i &= \sum_{j=1}^{i-1} \sum_{k=0}^{p_j - 1} \beta_{ijk} h_j F^k \text{ if } p_i = 0 \end{aligned} \tag{3.2.8}$$

Again, the observability condition ensures that $p_1 + p_2 + \ldots + p_m = n$

Of course, one may find other canonical parameter sets. Based on canonical parameter sets canonical forms are constructed, a more formal definition of canonical form can be found in [91] and states:

A canonical form is a pair of matrices (F, H) which is expressed only in terms of a canonical parameter set $\mathcal{S}$ and which has $\mathcal{S}$ as its canonical parameter set.

In literature the term *canonical form* has been used to describe specific structures of matrices which are not necessarily constructed from canonical parameter sets. We will make the distinction and separate these structured forms by the name of *pseudo-canonical forms*. For the pseudo-canonical forms we will present, it is proved in [91], that for a certain structure of a pseudo-canonical form the number of free parameters is equal or greater than the number of free parameters the structure would have if it had been constructed based on canonical parameter set.

3.2.3 Pseudo-Canonical Forms

Here we will present some of the most common pseudo canonical forms.

Type I

The following canonical form was introduced by Luenberger in [63] and examined further by Rosenbrock in [78]. Consider the following structure:

$$F = \begin{bmatrix} F_{11} & \cdots & F_{1m} \\ \vdots & \ddots & \vdots \\ F_{m1} & \cdots & F_{mm} \end{bmatrix} \tag{3.2.9a}$$

$$F_{ii} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \times & \times & \cdots & \times & \times \end{bmatrix}, F_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ \times & \cdots & \times \end{bmatrix} \tag{3.2.9b}$$

$$h_i = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}, n_i > 0 \tag{3.2.9c}$$

where F_{ii} is a $n_i \times n_i$ matrix filled with zeros and has ones in its superdiagonal while we let the last row be filled with free parameters $\times$, F_{ij} is $n_i \times n_j$ matrix filled with zeros and we let its last row be filled with free parameters $\times$ and h_i has 1 in column $1 + n_1 + \cdots + n_{i-1}$. If $n_i = 0$ then h_i has a free parameter $\times$ in every column in which $h_1, h_2, \ldots, h_{i-1}$ have non-zero entries and zeros elsewhere. We will present two examples of construction of canonical form type I based on a canonical parameter set A ([91]).

Consider the canonical parameter set n_i, α_{ijk} with values

$$n_1 = 3, n_2 = 0$$
$$\alpha_{110} = -1, \alpha_{111} = 0, \alpha_{112} = 0, \alpha_{210} = 0.7$$

For this canonical parameter set canonical form type I is:

$$F = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 \\ 0.7 & 0 & 0 \end{bmatrix}$$

Now consider that $n_1 = 1, n_2 = 2$ then in terms of $\{\alpha_{ijk}\}$ the canonical form type I is:

$$F = \begin{bmatrix} \alpha_{110} & \alpha_{120} & 0 \\ 0 & 0 & 1 \\ \alpha_{210} & \alpha_{220} & \alpha_{221} \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Type II

The following pseudo-canonical form was developed in [20]. The matrices have the following forms:

$$F = \begin{bmatrix} F_{11} & \cdots & F_{1m} \\ \vdots & \ddots & \vdots \\ F_{m1} & \cdots & F_{mm} \end{bmatrix} \tag{3.2.10a}$$

$$F_{ii} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \times \\ 1 & 0 & \cdots & 0 & \times \\ 0 & 1 & \cdots & 0 & \times \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \times \end{bmatrix}, F_{ij} = \begin{bmatrix} 0 & \cdots & 0 & \times \\ \vdots & \ddots & \vdots & \times \\ 0 & \cdots & 0 & \times \end{bmatrix} \tag{3.2.10b}$$

$$H = \begin{bmatrix} H_1 & H_2 & \cdots & H_m \end{bmatrix}, H_i = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & \times \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \times \end{bmatrix} \tag{3.2.10c}$$

where $F_i i$ is a $n_i \times n_1$ matrix filled with zeros with ones in its subdiagonal and free parameters $\times$ in its last column, F_{ij} is a $n_i \times n_j$ matrix filled with zeros except its last column which is filled with free parameters $\times$. H_i is a $m \times n_i$ matrix with 1 appearing in the i^{th} row. We will present an example of construction of a canonical form type II based on canonical a parameter set type A ([91]).

Consider the canonical parameter set n_i, α_{ijk} with values

$$n_1 = 3, n_2 = 0$$
$$\alpha_{110} = -1, \alpha_{111} = 0, \alpha_{112} = 0, \alpha_{210} = 0.7$$

For this canonical parameter set canonical form type II is:

$$F = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, H = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0.7 \end{bmatrix}$$

Type III

The next pseudo-canonical form was developed by Luenberger in [63] and Bucy in [21] in the special case where $p_i > 0$ and by Mayne in [65] for the general case:

$$F = \begin{bmatrix} F_{11} & 0 & \cdots & 0 \\ F_{21} & F_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mm} \end{bmatrix} \tag{3.2.11a}$$

$$F_{ii} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ \times & \times & \times & \times & \times \end{bmatrix}, F_{ij} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \times & \times & \cdots & \times \end{bmatrix} \tag{3.2.11b}$$

$$h_i = [0, \dots, 0, 1, 0, \dots, 0] \text{ if } p_i > 0 \tag{3.2.11c}$$

$$h_i = [\times, \times, \dots, \times, 0, \dots, 0] \text{ if } p_i = 0 \tag{3.2.11d}$$

where F is a lower triangular block matrix, F_{ii} is $p_i \times p_i$ matrix filled with zeros with ones in its superdiagonal and its last row filled with free parameters $\times$ while F_{ij} is a $p_i \times p_j$ matrix filled with zeros and its last row is filled with free parameters $\times$, as for h_i in the first case where $p_i > 0$ the 1 is in the $1 + p_1 + \ldots + p_{i-1}$ column and in the second case where $p_i = 0$ the free parameters occupy the first $p_1 + \ldots + p_{i-1}$ columns. An example will follow to clarify the construction of this form based on a canonical parameter set type B ([91]).

Consider the following canonical parameter set p_i, β_{ijk} with values

$$p_1 = 2, p_2 = 0, p_3 = 0$$
$$\beta_{110} = -0.3, \beta_{111} = 0.7, \beta_{310} = 0.3, \beta_{311} = 0.5, \beta_{330} = \beta_{331} = 0.4$$
$$\beta_{210} = \beta_{221} = 0.6$$

For this canonical parameter set canonical form type III is:

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.3 & 0.5 & 0.4 & 0.4 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.6 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

This form holds some special properties, which will examine in following sections and in fact this is the form we have employed for our identification procedure.

In [91] there are details on how to construct canonical (and not pseudo-canonical) forms for each of the types we have presented.

It is mentioned in [90], [91] and proven in [50] that for a general stochastic system of the form:

$$x_{k+1} = Fx_k + w_k \tag{3.2.12}$$
$$y_k = Hx_k + v_k \tag{3.2.13}$$

constraining the matrices F, H is not enough to ensure that the equivalence class will have only one member due to the effect of state and observation noise (which in general case are not the same). To overcome this problem, most identification algorithms are applied on the innovations representation form of the system, since it is unique for every system [4]. Moreover, an equivalence class can be defined for the innovations representation ([50]), for which matrix structures exist that ensure that the class will have only one member. In the next section, we present one well-known form for the innovations representation developed in [60] .

3.3 Identifiable Forms for Transfer Function Identifiability

As we have already mentioned in Section 1.5 there are different approaches in checking when a system is identifiable. For each of these approaches in order for the system to be identifiable (to ensure in other words that we will determine uniquely the system parameters and these parameters are equal to the "true" system) the matrices of the system must be constrained to specific forms. Consider a steady-state innovation representation of the form (3.3.2), it is shown in [50] there exists an invertible matrix T such that:

$$\begin{aligned} F(\theta_1) &= TF(\theta_2)T^{-1} \qquad B(\theta_1) = TB(\theta_2) \\ K(\theta_1) &= TK(\theta_2) \qquad H(\theta_1) = H(\theta_2)T^{-1} \end{aligned} \tag{3.3.1}$$

This means that there exists a linear transformation of the system matrices for which if we replace the matrices $F(\theta_1), H(\theta_1), K(\theta_1), B(\theta_1)$ with $TF(\theta_2)T^{-1}, H(\theta_2)T^{-1}, TK(\theta_2)$, $TB(\theta_2)$ we have the same output or differently stated the equivalence class of the system contain more than one member in the general case. This creates the problem that an identification procedure can yield not the true system matrices but a linear transformation of those. Thus the need arises to impose some form on the matrices in order the identification procedure to result to the true solution (equivalently to ensure that $T = I$, the identity matrix) or in simpler words make the system "identifiable". The forms we have presented in the previous section ensure identifiability (under certain identification procedures).

Transfer function identifiability, which was studied excessively in [60], is the focus of this section. We remind that under this rule a system is identifiable if and only if $H(z, \hat{\theta}) = H(z, \theta)$ meaning that the transfer function of the system with the parameters we have estimated is the same as the transfer function of the true system [60]. A pseudo-canonical form was developed, which we will describe below.

Consider a multivariate state-space model in the steady-state innovation representation form

$$x_{k+1} = F(\theta)x_k + B(\theta)u_k + Ke_k \tag{3.3.2a}$$

$$y_k = H(\theta)x_k + e_k \tag{3.3.2b}$$

$$\mathbf{E}\left[e_p e_q^T\right] = \Lambda\delta_{pq} > 0 \tag{3.3.2c}$$

where δ_{pq} is the Kronecker delta and $\mathbf{E}$ is the expectation operator. The following pseudo-canonical form was introduced by Ljung in [60]:

$$F(\theta) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \end{bmatrix}, \qquad B(\theta) = \begin{bmatrix} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{bmatrix}$$

$$K(\theta) = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}, \quad H(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.3.3}$$

In the example presented here, the state dimension is $n = 9$, the input dimension is $m = 2$ and the observation dimension is $p = 3$. The number of rows with $\times$ in $F(\theta)$ equals the number of outputs. The structure in general can be defined as ([60]):

Let $F(\theta)$ initially be a matrix filled with zeros and with ones along the superdiagonal. Let then row numbers $r_1, r_2, \ldots, r_p$, where $r_p = n$, be filled with parameters. Take $r_0 = 0$ and let $H(\theta)$ be filled with zeros, and then let row i have a one in column $r_{i-1}+1$. Let $B(\theta)$ and $K(\theta)$ be filled with parameters.

The parametrization is uniquely characterized by the p numbers r_i that are to be chosen by the user. Moreover, in [60] it is proven that the structure we have presented

is identifiable if and only if the matrix pair $(F(\theta), [B(\theta)\ K(\theta)])$ is controllable pair. The input can be omitted, eliminating the input matrix B, in which case the structure of F, H remain exactly the same and the controllability condition must hold for the pair (F, K).

3.4 Identification of Innovations Representation through the EM Algorithm

We have already mentioned that we cannot apply an identification procedure straight on a linear state-space model but rather on its innovation representation. In Chapter 2 we stated that for the exponential family the EM reduces to computing the conditional expectations of the complete data sufficient statistics during the E-step and using these in place of the complete data sufficient statistics in the M-step. In [26], [5] it is proven that if $[e_1(\theta)\ e_2(\theta)\ \cdots\ e_N(\theta)]$ has full rank and θ, Σ have no common parameters then the quantity:

$$L = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{k=1}^{N} e_k^T(\theta)\Sigma^{-1}e_k(\theta) \tag{3.4.1}$$

is maximized by

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta} \left| \left[\frac{1}{N}\sum_{k=1}^{N} e_k(\theta)e_k^T(\theta) \right] \right| \\ \hat{\Sigma} &= \frac{1}{N}\sum_{k=1}^{N} e_k(\hat{\theta})e_k^T(\hat{\theta}) \end{aligned} \tag{3.4.2}$$

We will now examine if we can determine uniquely the parameters of an innovation representation by the above equations.

Consider the following transformation of an innovation representation:

$$\begin{aligned} x_{k+1} &= Fx_k + Ke_k \\ y_k &= Hx_k + e_k \end{aligned} \quad \Rightarrow \quad \begin{aligned} x_{k+1} &= Fx_k + K\left(y_k - Hx_k\right) \\ e_k &= -Hx_k + y_k \end{aligned}$$

then by some manipulation we can easily see that:

$$\begin{aligned} e_{k+1} &= \left[-HF + HKH\right]x_k - HKy_k + y_{k+1} \\ &= y_{k+1} - \begin{bmatrix} HF - HKH & HK \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \end{aligned} \tag{3.4.3}$$

The joint log-likelihood of the complete (observed and unobserved) data (x_k, y_k) can be calculated by substituting 3.4.3 in equation 3.4.1.

We will now examine for some pseudo-canonical forms, if we can determine the parameters uniquely based on equations (3.4.2) with e_k given by (3.4.3). Consider the

following form constructed by Ljung's method presented in Section 3.3:

$$F = \begin{bmatrix} 0 & 1 & 0 \\ f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \end{bmatrix} \text{ then } \begin{matrix} r_0 = 0 \\ r_1 = 2 \\ r_3 = 3 \end{matrix} \text{ thus } \begin{matrix} h_1 = r_0 + 1 = 1 \\ h_2 = r_1 + 1 = 3 \end{matrix} \Rightarrow$$

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \\ k_5 & k_6 \end{bmatrix}$$

by substituting in (3.4.3) we have:

$$HF = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ f_4 & f_5 & f_6 \end{bmatrix}$$

$$HK = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \\ k_5 & k_6 \end{bmatrix} = \begin{bmatrix} k_1 & k_2 \\ k_5 & k_6 \end{bmatrix}$$

$$HKH = \begin{bmatrix} k_1 & k_2 \\ k_5 & k_6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} k_1 & 0 & k_2 \\ k_5 & 0 & k_6 \end{bmatrix}$$

It easy to see that the parameters f_1, f_2, f_3, k_3, k_4 are eliminated, which means that we cannot get unique solutions for these terms which in turn, renders the system unidentifiable through the Expectation-Maximization framework. This was expected since Ljung's form contains $2nm$ parameters and the parameters of the system of equations in (3.4.3) are at most $nm + m^2$ since $m <= n$. The only case where someone can find unique solution for the equation system is when $m = n$, which makes $H = I$ (the identity matrix) and F filled with free parameters in all its rows.

Now let us examine what would happen if the system matrices were constrained in Type III form. Consider the following Type III form:

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 \\ f_1 & f_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ f_3 & f_4 & f_5 & f_6 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ h_1 & h_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

by substituting in 3.4.3 we have:

$$HF = \begin{bmatrix} 1 & 0 & 0 & 0 \\ h_1 & h_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ f_1 & f_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ f_3 & f_4 & f_5 & f_6 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ h_2 f_1 & h_1 + h_2 f_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.4.4}$$

It is obvious that we cannot find unique solutions for the elements of F.

It has now become clear, that the innovations representation is not recommended for identification through an EM framework and canonical forms that would give unique solutions under other identification procedures cannot yield results that converge to the true system. In the next section we describe how we overcome this problem.

3.5 Identifiability of the Steady-State Kalman Filter

In Subsection 1.3.3 we referred to the nature of mapping of a state-space model to a Kalman Filter, which is many-to-one [4]. If we could find some form of the system matrices that could reduce the mapping to one-to-one and estimate uniquely the parameters of the Kalman Filter, then we would have identified the parameters of the model. Consider the linear time-invariant system given by:

$$x_{k+1} = Fx_k + w_k \tag{3.5.1a}$$

$$y_k = Hx_k + v_k \tag{3.5.1b}$$

$$\mathbf{E}[w_p w_q^T] = Q\delta_{pq} \quad \mathbf{E}[v_p v_q^T] = R\delta_{pq} \tag{3.5.1c}$$

where w_k and v_k are zero-mean Gaussian noises with covariance matrices Q and R respectively. We define $\theta = \{F, H, Q, R\}$ and make the following assumptions:

- F is stable
- (F, H) is observable pair
- (F, K) is controllable pair, where K is the steady-state Kalman Gain given by 3.5.2b.

A compact subset of θ with the above properties will be denoted $\mathcal{R}_c$. Now, let Σ be the steady-state state covariance given by the Discrete Algebraic Riccati Equation (DARE):

$$\Sigma = F\left[\Sigma - \Sigma H^T \left(H\Sigma H^T + R\right)^{-1} H\Sigma\right] + Q \tag{3.5.2a}$$

$$K = F\Sigma H^T \left(H\Sigma H^T + R\right)^{-1} \tag{3.5.2b}$$

Furthermore, consider $\bar{y}$ the random variable which express the "filtered" output of a steady-state Kalman Filter. Since the system (3.5.1) is linear and the noises are Gaussian, $p(\bar{y}_k|Y^{k-1}, \theta)$ is Gaussian with mean $\hat{y}_{k|k-1}$ (see subsection 1.3.1) and covariance $H\Sigma H^T + R$ as $k \to \infty$ (steady-state). According to [85]:
Two parameters $\theta_1, \theta_2 \in \mathcal{R}_c, \theta_1 \neq \theta_2$ are said to be CML (Constraint Maximum Likelihood) unresolvable if the quality:

$$p(y_k|Y^{k-1}, \theta_1) = p(y_k|Y^{k-1}, \theta_2)$$

holds with probability 1 with respect to θ_1 and θ_2 as $k \to \infty$, which means that we cannot determine uniquely the parameters of the steady-state filter through Maximum Likelihood estimation. In [85] the equivalence class (3.5.3) is presented and it is proven

that $\theta_1, \theta_2 \in \mathcal{R}_c$ are CML unresolvable if and only if there exists a nonsingular matrix T such that:

$$\begin{aligned} F_1 &= TF_2T^{-1} \\ H_1 &= H_2T^{-1} \\ K_1 &= TK_2 \\ H_1\Sigma_1H_1^T + R_1 &= H_2\Sigma_2H_2^T + R_2 \end{aligned} \tag{3.5.3}$$

As long as (3.5.3) is satisfied, the equivalence class has more than one member. This means that the two steady-state Kalman Filters have the same impulse response, which implies that a steady-state Kalman Filter in this case cannot determine uniquely a linear system. Alternatively the steady-state Kalman Filter can be associated with a number of different systems.

Two critical observations are made and proven in [91]:

1. If the matrices F, H of system of the form (3.5.1) follow the structure of pseudo-canonical form type III, then there is only member in the equivalence class which in turn means that the mapping of the system with the Kalman Filter is one-to-one.

2. The parameters (F, H, Σ_{e_k}, K) of the steady-state Kalman Filter are uniquely determined by the output measurements, and the filter is therefore identifiable, as long as F, H are in pseudo-canonical form type III. Where $\Sigma_{e_k} = H\Sigma H^T + R$ is the covariance of the innovations noise.

Hence, if we constraint the matrices F, H in pseudo-canonical form type III we can identify the system parameters by identifying the parameters of its associated steady-state Kalman Filter. Based on these observations, we will present our proposed algorithm of system identification when the matrices are constrained in pseudo-canonical form type III in the next chapter.

3.6 Summary

In this chapter we presented several pseudo-canonical forms developed in literature. We showed that these forms are a necessary condition for identifiability of a system but in order to be a sufficient condition too we must examine the identification procedure. We showed that we can not use the Expectation-Maximization algorithm for identification of an innovation representation, other alternatives should be considered. One such alternative is the identifiability of the steady-state Kalman Filter. Based on this observation, we laid the foundation to present our proposed system identification algorithm in the next chapter.

Chapter 4

Maximum Likelihood Estimation of Identifiable State-Space Models

4.1 Introduction

In the previous chapter we described some of the most common pseudo-canonical forms presented in literature and examined the identifiability of these forms under the Expectation Maximization Algorithm. We established that applying EM on the innovation representation is a complicated task that cannot yield unique estimation of the system parameters. We overcome this issue by reducing the association of a steady-state Kalman Filter with a state-space system to one-to-one and then identifying the parameters of the steady-state Kalman Filter, thus identifying the model. In this chapter we will describe out proposed algorithm for system identification of systems whose matrices are constrained to be in canonical form type III.

4.2 Description of the Algorithm

This far we have established that our method first attempts to estimate the parameters of a steady-state Kalman Filter and through it identify the parameters of the system. To ensure that the association of the steady-state Kalman Filter with the system is unique, the matrices must be constrained to follow pseudo-canonical form Type III. Consider the linear time-invariant system:

$$x_{k+1} = Fx_k + w_k \tag{4.2.1a}$$

$$y_k = Hx_k + v_k \tag{4.2.1b}$$

$$\mathbf{E}[w_p w_q^T] = Q\delta_{pq} \quad \mathbf{E}[v_p v_q^T] = R\delta_{pq} \tag{4.2.1c}$$

where w_k and v_k are zero-mean Gaussian noises with covariance Q and R respectively, also $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$. Furthermore we make the following assumptions:

- F is stable
- (F, H) is observable pair
- (F, K) is controllable pair, where K is the steady-state Kalman Gain given by (3.5.2b).

Then, the steady-state Kalman Filter for the above system 4.2.1 is:

Steady-State Kalman Filter

$$e_k = y_k - H\hat{x}_{k|k-1} \tag{4.2.2a}$$

$$\Sigma_{e_k} = H\Sigma_{k|k-1}H^T + R \tag{4.2.2b}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \left(\Sigma H^T \Sigma_{e_k}^{-1}\right) e_k \tag{4.2.2c}$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} \tag{4.2.2d}$$

$$\tag{4.2.2e}$$

where Σ is given by the Discrete Algrebraic Riccati Equation:

$$\Sigma = F\left[\Sigma - \Sigma H^T \left(H\Sigma H^T + R\right)^{-1} H\Sigma\right] + Q \tag{4.2.3}$$

and the steady-state Kalman Gain is given by

$$K = F\Sigma H^T \left(H\Sigma H^T + R\right)^{-1} \tag{4.2.4}$$

As we can see in the steady-state form of the Kalman Filter, we can pre-compute some quantities to reduce the computational cost.

The parameters we need to estimate for a steady-state Kalman Filter and a state-space model are:

- The parameters of a steady-state Kalman Filter given by (4.2.2: F, H, Σ_{e_k}, K
- The parameters of a state-space model given by (4.2.1): F, H, Q, R

As we can see the matrices F, H are common parameters in the steady-state Kalman Filter and state-space model. We remind that the matrices F, H must be in type III

form, presented in Subsection 3.2.3, and formulates F, H such that:

$$F = \begin{bmatrix} F_{11} & 0 & \cdots & 0 \\ F_{21} & F_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mm} \end{bmatrix} \tag{4.2.5a}$$

$$F_{ii} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ \times & \times & \times & \times & \times \end{bmatrix}, F_{ij} = \begin{bmatrix} & 0 & \dots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \times & \times & \cdots & \times \end{bmatrix} \tag{4.2.5b}$$

$$h_i = [0, \dots, 0, 1, 0, \dots, 0] \text{ if } p_i > 0 \tag{4.2.5c}$$

$$h_i = [\times, \times, \dots, \times, 0, \dots, 0] \text{ if } p_i = 0 \tag{4.2.5d}$$

F being a lower triangular block matrix, F_{ii} a $p_i \times p_i$ matrix filled with zeros with ones in its superdiagonal and its last row filled with free parameters $\times$ while F_{ij} a $p_i \times p_j$ matrix filled with zeros and its last row filled with free parameters $\times$. h_i in the case where $p_i > 0$ has the 1 in the $1 + p_1 + \ldots + p_{i-1}$ column and in the case where $p_i = 0$ the free parameters occupy the first $p_1 + \ldots + p_{i-1}$ columns.

Now, since the mapping of the steady-state Kalman Filter with the system is one-to-one, if the parameters of the Kalman Filter converge to the "true" values then the parameters of the system will converge to their "true" values and vice versa. To estimate the parameters of the system through the EM Algorithm, consider that the states of the system described by equations 4.2.1 is not hidden and we want to find the Maximum Likelihood estimates of the system parameters θ given $Y = [y_0, y_1, \dots, y_N]$ and $X = [x_0, x_1, \dots, x_N]$. In this case the ML estimates of θ are obtained by maximizing:

$$\begin{aligned} L(X, Y, \theta) = &- \sum_{k=1}^{N} \left\{ log|Q| + (x_k - Fx_{k-1})^T Q^{-1} (x_k - Fx_{k-1}) \right\} \\ &- \sum_{k=0}^{N} \left\{ log|R| + (y_k - Hx_k)^T R^{-1} (y_k - Hx_k) \right\} + \text{constant} \end{aligned} \tag{4.2.6}$$

since, without loss of generality, w_k and v_k were assumed uncorrelated white Gaussian noise sources and the base of log is e. This problem can be treated as one with incomplete data with the state vector playing the role of missing observations thus enabling us to apply the EM Algorithm. In this case the auxiliary function of the EM becomes [26]:

$$\mathcal{Q}(\theta^{i+1}, \theta) = \mathbf{E} \left\{ L(X, Y, \theta^{i+1}) | Y, \theta^i \right\} \tag{4.2.7}$$

which is the conditional expectation of $L(X, Y, \theta)$ defined by equation 4.2.6 given the observed data Y and the current parameter estimations θ^i. As we have mentioned in

Chapter 2 the EM algorithm for the exponential family as is our case under the Gaussian assumption reduces to computing the conditional expectations of the complete data sufficient statistics during the E-step and using these in place of the complete-data sufficient statistics in the M-step in order to compute the ML estimates [25].

We remind that in [26], [5] it is proven that if $[e_1(\theta)\ e_2(\theta)\ \cdots\ e_N(\theta)]$ has full rank and θ, Σ have no common parameters then the quantity:

$$L = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{k=1}^{N} e_k^T(\theta)\Sigma^{-1}e_k(\theta)$$

is maximized by

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \log\left|\left[\frac{1}{N}\sum_{k=1}^{N} e_k(\theta)e_k^T(\theta)\right]\right| \\ \hat{\Sigma} &= \frac{1}{N}\sum_{k=1}^{N} e_k(\hat{\theta})e_k^T(\hat{\theta})\end{aligned} \tag{4.2.8}$$

We can replace the first term in 4.2.8 with:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \left|\left[\frac{1}{N}\sum_{k=1}^{N} e_k(\theta)e_k^T(\theta)\right]\right| \\ &= \arg\min_{\theta} \operatorname{tr}\left[\log\frac{1}{N}\sum_{k=1}^{N} e_k(\theta)e_k^T(\theta)\right]\end{aligned} \tag{4.2.9}$$

where $\operatorname{tr}A$ express the trace of matrix A. The above equality was proven in [10].

Under the assumption that F, H, Q, R do not have common parameters we can examine the terms of (4.2.6) separately. Let us consider the first term of (4.2.6). The difference with the classical approach presented in Section 2.3 is that here θ does not include all the elements of F but only the free parameters that appear the predetermined positions of the type III pseudo-canonical form. Hence we have to maximize the first term of (4.2.6) only with respect to these free parameters that appear in F and not all the elements of F, this problem is one of patterned matrix derivative (more on patterned matrix derivatives can be found in [55], [71], [89]).

In our case in (4.2.8), $e_k(\theta) = x_k - Fx_{k-1}$. We will expand 4.2.9 for every element f_{ij} in F and then take the partial derivatives and set to zero only with respect to the elements that are the free parameters (since the rest are constant terms, zeros or ones):

$$\begin{aligned}J =& \operatorname{tr}\left[\log\frac{1}{N}\sum_{k=1}^{N}(x_k - Fx_{k-1})(x_k - Fx_{k-1})^T\right] = \\ &\log\left(\frac{1}{N}\sum_{k=1}^{N}\left(x_{k,1} - f_{11}x_{k-1,1} - f_{12}x_{k,2} - \ldots - f_{1n}x_{k-1,n}\right)^2\right) + \\ &\log\left(\frac{1}{N}\sum_{k=1}^{N}\left(x_{k,2} - f_{21}x_{k-1,1} - f_{22}x_{k,2} - \ldots - f_{2n}x_{k-1,n}\right)^2\right) + \ldots\end{aligned} \tag{4.2.10}$$

where the second subindex l in $x_{k-1,l}$ and $x_{k,l}$ denotes the position in the vector. It is easy to observe in 4.2.10 that each term depends only on elements of the same row of F. By construction ,F contains free elements to the left-most part of each row (the rest being zero) and by assumption each element is independent from any other, thus we can take partial derivatives to the elements of each row, hence ignoring the terms of J that contain elements of other rows than the one under differentiation. Assume that the free parameters in the first row are only f_{11} and f_{12} then by construction the rest elements of the row are zero. Taking partial derivatives with respect to the free parameters we have:

$$\frac{\partial J}{\partial f_{11}} = 0 \Rightarrow \sum_{k=1}^{N} x_{k,1}x_{k-1,1} - \sum_{k=1}^{N} f_{1,1}x_{k-1,1}^2 - \sum_{k=1}^{N} f_{1,2}x_{k-1,1}x_{k-1,2} = 0$$
$$\frac{\partial J}{\partial f_{12}} = 0 \Rightarrow \sum_{k=1}^{N} x_{k,1}x_{k-1,2} - \sum_{k=1}^{N} f_{1,1}x_{k-1,1}x_{k-1,2}^2 - \sum_{k=1}^{N} f_{1,2}x_{k-1,2}^2 = 0$$

By expressing the above result in matrix form we have:

$$\begin{bmatrix} f_{11} & f_{12} \end{bmatrix} =$$
$$\left[\sum_{k=1}^{N} \begin{bmatrix} x_{k,1}x_{k-1,1} & x_{k,1}x_{k-1,2} \end{bmatrix}\right] \left[\sum_{k=1}^{N} \begin{bmatrix} x_{k-1,1} \\ x_{k-1,2} \end{bmatrix} \begin{bmatrix} x_{k-1,1} & x_{k-1,2} \end{bmatrix}\right]^{-1}$$

We remind that we assume N is large enough to guarantee positive definiteness of the matrix $\sum_k x_k x_k^T$, which means that the upper block diagonal of the matrix is positive definite too and thus invertible [42], [66]. Now, if f_{13} is a free parameter and the rest elements of the row are zero, we have:

$$\begin{bmatrix} f_{11} & f_{12} & f_{13} \end{bmatrix} =$$
$$\left[\sum_{k=1}^{N} \begin{bmatrix} x_{k,1}x_{k-1,1} & x_{k,1}x_{k-1,2} & x_{k,1}x_{k-1,3} \end{bmatrix}\right] \left[\sum_{k=1}^{N} \begin{bmatrix} x_{k-1,1} \\ x_{k-1,2} \\ x_{k-1,3} \end{bmatrix} \begin{bmatrix} x_{k-1,1} & x_{k-1,2} & x_{k-1,3} \end{bmatrix}\right]^{-1}$$

By the same procedure, we can estimate the parameters of the other rows. By induction we have the general formula for each row i with $r \leq n$ free parameters:

$$\begin{bmatrix} f_{i1} & f_{i2} & \dots & f_{ir} \end{bmatrix} = \left[\sum_{k=1}^{N} x_k x_{x-1}^T\right]_{[i,1:r]} \left[\sum_{k=1}^{N} x_{k-1} x_{k-1}^T\right]^{-1}_{[1:r,1:r]} \tag{4.2.11}$$

where $[i, 1:r]$ denotes the first r elements of row i and $[1:r, 1:r]$ denotes the upper square $r \times r$ matrix. What is left now is the estimation of Q, which involves derivation with respect to all of its elements. By expanding the second term of 4.2.8 on our

estimation of F, $\hat{F}$ we have:

$$\begin{aligned}\hat{Q} = \frac{1}{N}\Bigg(&\sum_{k=1}^{N} x_k x_k^T - \left[\sum_{k=1}^{N} x_k x_{k-1}^T\right]\hat{F}^T \\ &-\hat{F}\left[\sum_{k=1}^{N} x_k x_{k-1}^T\right]^T + F\left[\sum_{k=1}^{N} x_{k-1} x_{k-1}^T\right]\hat{F}^T\Bigg)\end{aligned} \tag{4.2.12}$$

With exactly the same procedure we can estimate the free elements of H (the fact that is not square does not affect the derivation) and the covariance matrix R. To sum up the estimation formulas are:

$$\begin{bmatrix} f_{i1} & f_{i2} & \dots & f_{ir}\end{bmatrix} = \left[\sum_{k=1}^{N} x_k x_{x-1}^T\right]_{[i,1:r]} \left[\sum_{k=1}^{N} x_{k-1} x_{k-1}^T\right]^{-1}_{[1:r,1:r]} \tag{4.2.13a}$$

$$\begin{bmatrix} h_{j1} & h_{j2} & \dots & h_{jl}\end{bmatrix} = \left[\sum_{k=1}^{N} y_k x_x^T\right]_{[j,1:l]} \left[\sum_{k=1}^{N} x_k x_k^T\right]^{-1}_{[1:l,1:l]} \tag{4.2.13b}$$

$$\begin{aligned}\hat{Q} = \frac{1}{N}\Bigg(&\sum_{k=1}^{N} x_k x_k^T - \left[\sum_{k=1}^{N} x_k x_{k-1}^T\right]\hat{F}^T \\ &-\hat{F}\left[\sum_{k=1}^{N} x_k x_{k-1}^T\right]^T + \hat{F}\left[\sum_{k=1}^{N} x_{k-1} x_{k-1}^T\right]\hat{F}^T\Bigg)\end{aligned} \tag{4.2.13c}$$

$$\begin{aligned}\hat{R} = \frac{1}{N+1}\Bigg(&\sum_{k=0}^{N} y_y x_k^T - \left[\sum_{k=0}^{N} y_k x_k^T\right]\hat{H}^T \\ &-\hat{H}\left[\sum_{k=0}^{N} y_k x_k^T\right]^T + \hat{H}\left[\sum_{k=0}^{N} x_k x_k^T\right]\hat{H}^T\Bigg)\end{aligned} \tag{4.2.13d}$$

for which the sufficient statistics are:

$$\Gamma_1 = \frac{1}{N+1}\sum_{k=0}^{N} x_k x_k^T \tag{4.2.14a}$$

$$\Gamma_2 = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T \tag{4.2.14b}$$

$$\Gamma_3 \frac{1}{N}\sum_{k=1}^{N} x_{k-1} x_{k-1}^T \tag{4.2.14c}$$

$$\Gamma_4 \frac{1}{N}\sum_{k=1}^{N} x_k x_{k-1}^T \tag{4.2.14d}$$

$$\Gamma_5 \frac{1}{N+1}\sum_{k=0}^{N} y_k y_k^T \tag{4.2.14e}$$

$$\Gamma_6 \frac{1}{N+1}\sum_{k=0}^{N} y_k x_k^T \tag{4.2.14f}$$

Of course, we replace the sufficient statistics by their expected values to complete the EM Algorithm. The statistics at iteration i are:

$$\mathbf{E}\left\{y_k x_k^T | Y, \theta^i\right\} = y_k E\left\{x_k^T | Y, \theta^i\right\} \tag{4.2.15a}$$

$$\mathbf{E}\left\{y_k y_k^T | Y, \theta^i\right\} = y_k y_k^T \tag{4.2.15b}$$

$$\mathbf{E}\left\{x_k | Y, \theta^i\right\} = \hat{x}_{k|N} \tag{4.2.15c}$$

$$\mathbf{E}\left\{x_k x_k^T | Y, \theta^i\right\} = \Sigma_{k|N} + \hat{x}_{k|N}\hat{x}_{k|N}^T \tag{4.2.15d}$$

$$\mathbf{E}\left\{x_k x_{k-1}^T | Y, \theta^i\right\} = \Sigma_{k,k-1|N} + \hat{x}_{k|N}\hat{x}_{k-1|N}^T \tag{4.2.15e}$$

where $\hat{x}_{k|N}$, $\Sigma_{k|N}$, $\Sigma_{k,k-1|N}$ are computed by the RTS smoother. The above results conclude the EM procedure for the estimation of the matrices F, H, Q, R. Due to the one-to-one mapping (under type III pseudo-canonical form) of the system and the steady-state Kalman Filter if the system parameters converge to the "true" parameters then the filter parameters will converge to the "true" parameters too. To summarize the steps of our algorithm are:

1. Initialize F,H,Q,R
2. Solve the DARE equation 4.2.3 to find $\Sigma_{k|k-1}$ as $k \to \infty$, the Steady State Kalman Gain K and innovation noise Σ_{e_k}.
3. Apply the Steady-State Kalman Filter to the data set Y and then the RTS smoother
4. Collect sufficient statistics
5. re-estimate F,H,Q,R through the statistics gathered in previous step
6. return to step 2 and solve the DARE with the new matrices

The steps 2-4 could be considered the E-step of the EM while step 5 could be considered as the M-step if we project the EM algorithm on our procedure. In the next section we will present experimental results testing our identification procedure.

4.3 Experimental Results

In this section we will present some experimental results of our identification procedure. We estimate the system matrices with the EM Algorithm and check their convergence with respect to the "true" values. We also check the convergence of the steady-state Kalman Filter parameters, since if those parameters converge to the true values then the system parameters will converge to the true values too. We remind that the matrices

F, H follow type III pseudo-canonical form. The experiments were implemented in MATLAB R2009a. We examine the convergence by measuring the "distance" between the system matrices from which we have generated the data and the estimated matrix in each iteration of our algorithm. The distance is expressed by calculating the Frobenius norm of the difference of the true matrix and the estimated one. The Frobenius norm of a matrix A is given by :

$$\|A\|_F = \sqrt{\sum_{i-1}^{n}\sum_{j=1}^{m}|a_{ij}|^2}$$

In our experiments we have generated 10000 multi-dimensional observation vectors y_k from a given system and our aim is to try and identify the system parameters through the observation data only.

4.3.1 Experiment 1

Let's assume that we have the following parameter set: $p1 = 3, p2 = 0$ then by the construction formula of pseudo-canonical form of type III will be:

$$F = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.2 & 0.1 & -0.1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 \\ 0.3 & 0.3 & 0 \end{bmatrix}$$

where the free parameters occupy the last row of matrix F and the two first positions of the last row of matrix H. The covariance matrices Q, R are:

$$Q = R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

all the elements in Q, R are free parameters. For these matrices the associated steady-state Kalman Gain and Innovation covariance are:

$$K = \begin{bmatrix} -0.1328 & 0.4867 \\ 0.0601 & -0.0158 \\ -0.1589 & 0.0163 \end{bmatrix}, \Sigma_{e_k} = \begin{bmatrix} 3.7552 & 0.7927 \\ 0.7927 & 1.4125 \end{bmatrix}$$

The figures below show in each step the distance of the estimated matrices with the true system matrices.

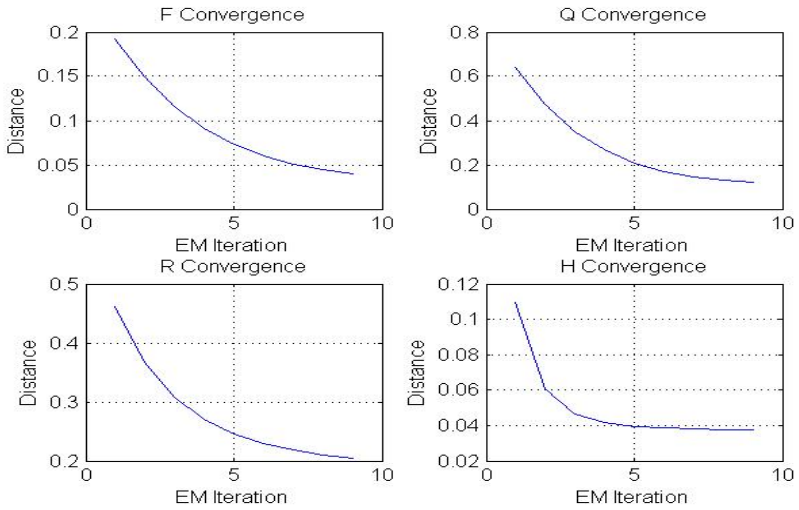

Figure 4.1: Experiment 1 - Convergence of the system matrices.

We also check the convergence of the innovation noise and steady-state Kalman Gain to ensure that the steady-state Kalman Filter converges to the true values thus ensuring the identification of the system.

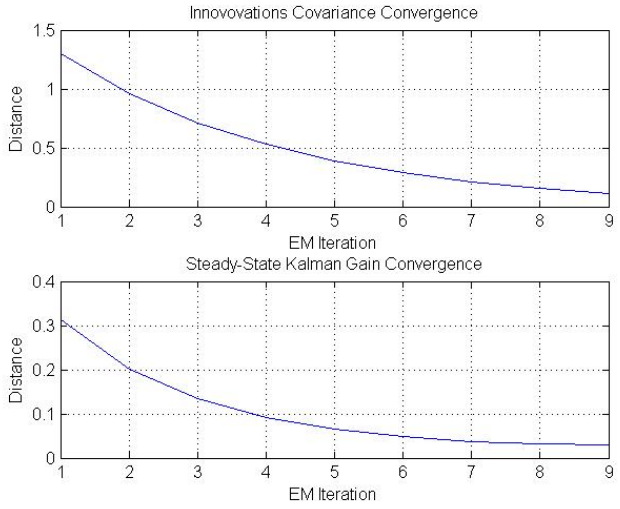

Figure 4.2: Experiment 1 - Convergence of the steady-state Kalman Gain and Innovations covariance.

As we can see from the figures in each step the Frobenius norm of the difference between the true matrix and the estimated approaches zero which means that the matrices converge to their true values.

4.3.2 Experiment 2

In this experiment we construct the pseudo-canonical based on the parameters p_i having values: $p_1 = 2, p_2 = 2, p_3 = 2$. For this set the matrices are:

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -0.2 & 0.1 & 0.2 & -0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0.2 & 0.1 & -0.1 & 0.3 & -0.2 & -0.1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where the free parameters occupy the first two elements of the second row, the first four elements of the fourth row and the last row of matrix F. H does not have free parameters and is a constant matrix. The covariance matrices Q, R are:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where all the elements in Q, R are free parameters. For these matrices the associated steady-state Kalman Gain and Innovation covariance are:

$$K = \begin{bmatrix} 0.0705 & 0.0010 & 0.0024 \\ 0.0812 & 0.0002 & 0.0008 \\ 0.0295 & -0.1176 & 0.0119 \\ -0.1359 & 0.1715 & -0.0061 \\ 0.0379 & 0.1155 & -0.0409 \\ 0.1456 & -0.1126 & -0.1268 \end{bmatrix}, \Sigma_{e_k} = \begin{bmatrix} 3.0358 & -0.0017 & 0.0293 \\ -0.0017 & 3.1347 & -0.1088 \\ 0.0293 & -0.1088 & 3.1487 \end{bmatrix}$$

As we can see from figure 4.3 the parameters of the system converge to their true values since the Frobenius norm of the distance approaches zero.

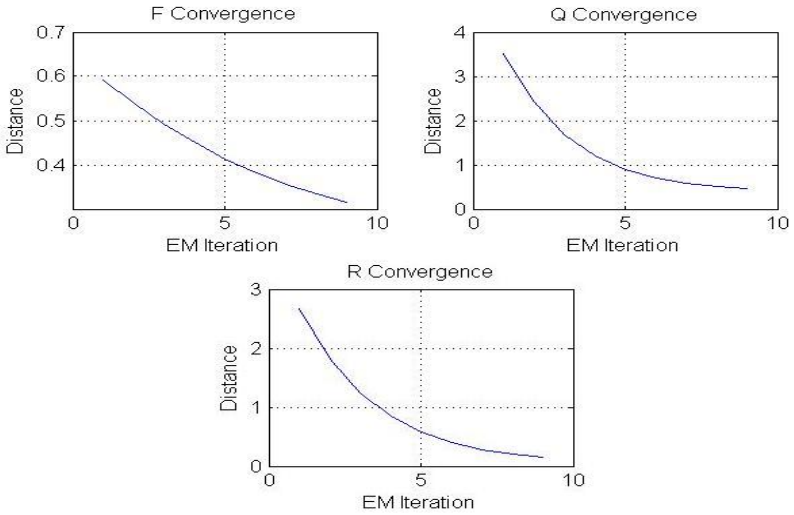

Figure 4.3: Experiment 2 - Convergence of the system matrices.

Not only the system parameters converge but it is obvious from figure 4.4 that the parameters of the steady-state Kalman Filter converge to their true values too thus ensuring the identification of the system.

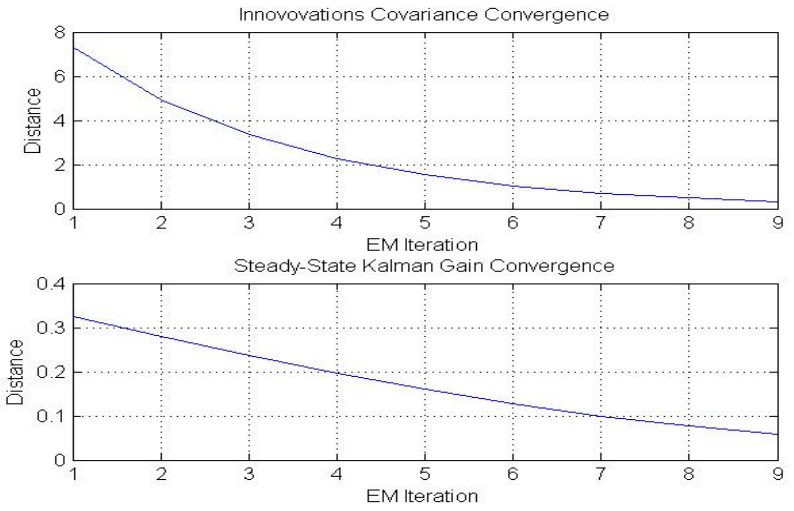

Figure 4.4: Experiment 2 - Convergence of the steady-state Kalman Gain and Innovations covariance.

In this experiment we have not included a diagram for the convergence of matrix H since it is a constant matrix and does not contain free parameters. Obviously all the matrices converge to their true values.

4.3.3 Experiment 3

In this experiment we assume that the parameters p_i have values: $p_1 = 1, p_2 = 1, p_3 = 1$. For these parameters the matrices become:

$$F = \begin{bmatrix} 0.1 & 0 & 0 \\ 0.2 & -0.1 & 0 \\ -0.2 & 0.1 & -0.1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where the free parameters appear in the first position of the first row, the first and second position of the second row and in the last row of matrix F, H is a constant matrix and does not contain parameters to be estimated. The covariance matrices Q, R are:

$$Q = R = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

and as before we consider all their elements to be free parameters. In this experiment the steady-state Kalman Gain and Innovation covariance are:

$$K = \begin{bmatrix} 0.0501 & 0.0002 & -0.0002 \\ 0.1000 & -0.0501 & 0.0001 \\ -0.0998 & 0.0507 & -0.0509 \end{bmatrix}, \Sigma_{e_k} = \begin{bmatrix} 4.0100 & 0.0200 & -0.0200 \\ 0.0200 & 4.0500 & -0.0500 \\ -0.0200 & -0.0500 & 4.0602 \end{bmatrix}$$

Figure 4.5 clearly depicts that the system parameters again converge to the true values of the system.

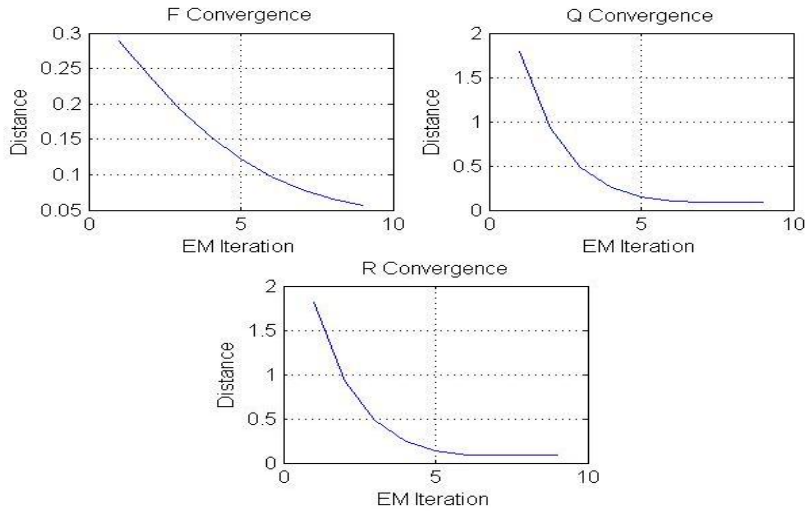

Figure 4.5: Experiment 3 - Convergence of the system matrices.

Again we present the diagrams that show the convergence of the steady-state Filter parameters in figure 4.6:

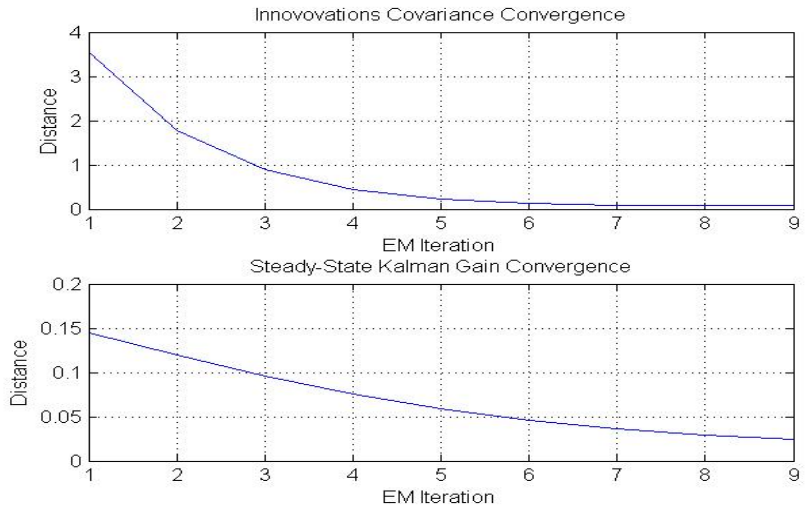

Figure 4.6: Experiment 3 - Convergence of the steady-state Kalman Gain and Innovations covariance.

We remind that in this experiment we do not present a diagram for matrix H since it is a constant and does not contain free parameters. We easily conclude from figures 4.5, 4.6 that the identification procedure approaches the true values in each iteration step.

4.3.4 Experiment 4

In this final experiment we present we have assumed that: $p_1 = 3, p_2 = 0, p_3 = 0, p_4 = 3$ then the matrices become:

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.1 & 0.2 & -0.3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -0.1 & 0.3 & -0.1 & 0.2 & -0.2 & 0.1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.3 & 0 & 0 & 0 & 0 \\ 0.3 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

In this case the free parameters are the first tree elements of the third row and the last row of F. Moreover, H contains free parameters in the first two positions of the second

and third row. The covariances matrices Q, R are:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, R = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

in this case the steady-Kalman Gain and innovation covariance are:

$$K = \begin{bmatrix} -0.1474 & 0.2332 & 0.1506 & -0.0054 \\ 0.0985 & -0.0143 & -0.0062 & 0.0024 \\ -0.0053 & 0.0680 & 0.0451 & -0.0011 \\ -0.0399 & -0.0006 & -0.0017 & 0.0046 \\ 0.1180 & 0.0174 & 0.0155 & -0.0790 \\ -0.0865 & 0.0564 & 0.0347 & 0.1121 \end{bmatrix}, \Sigma_{e_k} = \begin{bmatrix} 4.9479 & 1.0718 & 0.8128 & 0.1028 \\ 1.0718 & 2.5810 & 0.4230 & 0.0314 \\ 0.8128 & 0.4230 & 2.3091 & 0.0244 \\ 0.1028 & 0.0314 & 0.0244 & 5.2295 \end{bmatrix}$$

It is obvious from Figure 4.7 that the system matrices approach the true values with each iteration step.

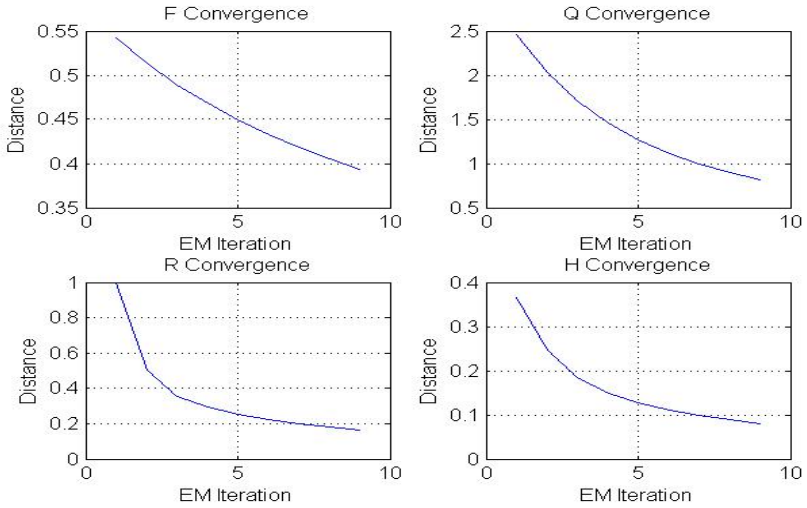

Figure 4.7: Experiment 4 - Convergence of the system matrices.

Moreover in Figure 4.8 we can see the the steady-state Kalman Filter parameters converge too. It is worthy to observe that in this experiment that it takes only few iterations for the matrices to converge.

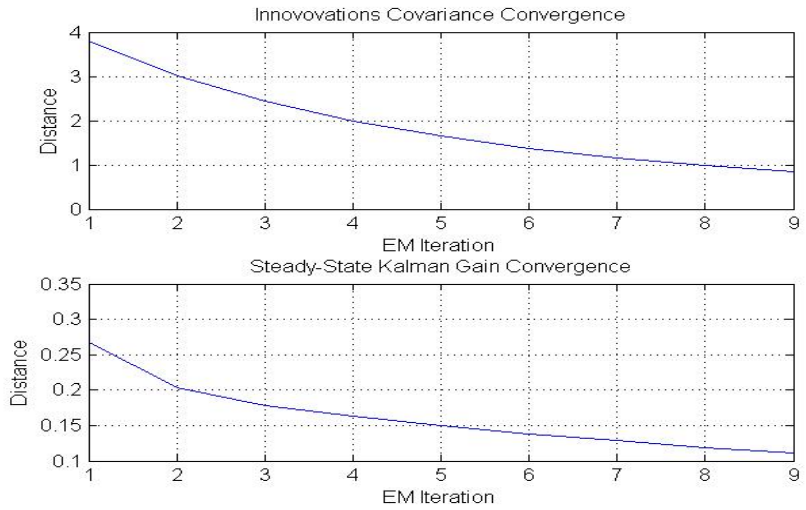

Figure 4.8: Experiment 4 - Convergence of the system matrices.

4.4 Conclusions

We have presented a robust method for EM identification of general state-space models in which the matrices follow pseudo-canonical form type III. We have presented the inherent problems of EM identification, when the system is in innovation representation and proved that we can apply the EM straight on the original form of the system under the assumptions we have made on the form of matrices. By exploiting the one-to-one mapping of a steady-state Kalman Filter to a state-space system when its matrices follow Type III pseudo-canonical forms we have managed to correctly identify the parameters of the system. In all our experiments we observed that the system matrices converge to the "true" system values only after few iterations. To the best of our knowledge, this is the first work that solves the problem of maximum-likelihood identification of linear state-space models for arbitrary state and observation dimensions. ML identifiability is ensured by using canonical forms and the solution is obtained using the EM framework. The experimental results show good convergence properties in all the cases that we examined.

Chapter 5

Extension and Future Research

5.1 Introduction

In this final chapter we will state our opinion regarding future work on identification through the Expectation Maximization Algorithm. We will also present some other ideas that attack the problem from different perspectives. Finally, we will make a brief review on the problem of nonlinear identification.

5.2 Future Work

In the previous chapter we described analytically the algorithm we developed for system identification and presented some experimental results that confirm our theoretical results and validate the procedure. Identification of systems through and EM framework was first introduced in [26] for the general case and to the best of our knowledge this is the first work that attempts to identify systems in canonical forms.

One interesting idea is to attempt the extension of our EM procedure to other system classes, such as closed-loop systems, bilinear systems, continuous-time systems, descriptor systems, periodic systems.

Moreover, we have not examined the mapping of the steady-state Kalman Filter to other canonical forms presented in literature. For those canonical forms that guarantee the one-to-one mapping of the steady-state Kalman Filter and the system we can attempt to extract Maximum Likelihood sufficient statistics.

Expectation Maximization could also be applied in identification of nonlinear models. Though the EM have been applied in nonlinear models [79], [81] the nonlinearity has been described with nonparametric methods. Our goal is to employ some parametric structure, like radial basis functions, to describe the nonlinearity and apply the EM algorithm for estimating the parameters of the system.

5.3 Other Innovative Approaches

System Identification is a challenging and hot topic that attracts many researchers which not only aim to improve current techniques but also propose alternative methods. De Cock for example in [24] tried to enrich subspace methods by introducing elements

for other scientific areas such as Information Theory, Statistics, Geometry among others and developed a new subspace identification method based on principal angles.

Ribarits in [77], trying to overcome the problem of canonical form parametrization, developed an identification method based on Data Driven Local Coordinates and Generalized Least Squares. The result was an iterative algorithm whose innovation lied to the fact that there was no need for canonical forms to be applied in system matrices.

A main focus of research on system Identification is a unification attempt of prediction error methods and subspace methods, trying to combine the advantages of both approaches under a single coherent theory [86].

5.4 Non-Linear System Identification

Though in our work our attention was focused on linear time-invariant systems, identification of nonlinear systems maybe the most active area in System Identification today [62]. Stan Ulam characterized nonlinear system identification as "non-elephant" zoology in an attempt to describe how huge this topic is.

To construct and estimate models on non-linear dynamic systems is an important and difficult task. It draws upon many different scientific areas such as: physical modeling ([17]), mathematical statistics ([43]), neural network techniques ([12]), learning theory and support vector machines ([95]), automatic control and system identification ([83]) and several others. Nonlinear models play important roles in many different application fields, and many specific problem areas have developed their own techniques and methodologies. Therefore, there is a vast amount of methods, concepts and results. It is not possible to give a short, comprehensive survey of the field but we will try to make a brief review of the problem and present some classic approaches.

The equations describing a nonlinear system in state-space form are:

$$\begin{aligned} x_{k+1} &= f(x_k) + w_k \\ y_k &= h(x_k) + v_k \end{aligned}$$

where f, h are nonlinear functions and w_k, v_k are noise vectors.

Identification of a nonlinear system involves estimating the covariance matrices of w_k, v_k but also defining the nature of functions f and h.

Indeed, one can find numerous approaches in literature with theories drawn from many scientific areas as Statistics, Time Series Analysis, Machine Learning, Artificial Neural Networks, etc. One approach is to try and "linearize the problem" by considering the problem as one of linear time-varying one [11]. Then one can apply the well-known theory of linear systems for identifying the model parameters.

Another approach which is usually applied only when the observation equation is nonlinear, is to model the nonlinearity with an artificial neural network. By training the neural network from the observations ([44]), one can define the nonlinearity and then by application of Extended or Unscented Kalman Filter estimate the states, from which point on various estimation methods can be employed for the determination of the parameters of the state equation [88].

Roweis in [79] attempted to apply the EM Algorithm for identification of nonlinear systems. The goal was to to integrate over the uncertain estimates of the unknown

hidden states and optimize the resulting marginal likelihood of the parameters given the observed data. The Extended Kalman Filters was used to estimate the approximate state distribution and they modeled the nonlinearities with Radial Basis Functions Neural Networks.

As we have already mentioned, Nonlinear System Identification is a vast research topic and there exist a number of algorithms and approaches to deal with the problem. Some texts dealing exclusively with the problem of nonlinear system identification are [68], [67], [15].

5.5 Summary

This chapter concludes this thesis. We outlined some interesting directions for future work based on identification through the Expectation Maximization Algorithm. Moreover, we presented some other innovative ideas that have emerged in the area of system identification. Finally, we attempted to give a brief review of nonlinear identification problem and present some approaches found in literature that attack the problem from different perspectives. In our opinion, even though there are numerous identification methods and approaches (both for linear and nonlinear identification), still the subject is far from closed and new ideas can have large contribution in the field.

Bibliography

[1] H. Akaike, "Stochastic Theory of Minimal Realization," *IEEE Tranactions on Automatic Control*, vol. AC-19, no. 6, pp. 667–674, 1974.

[2] H. Akaike, *Systems Identification : Advances and Case Studies*, ch. Canonical correlation analysis of time series and the use of an information criterion. Academic Press, 1976.

[3] H. Akaike, "Comments on Model Structure Testing in System Identification'," *International Journal of Control*, vol. 27, no. 2, pp. 323–324, 1978.

[4] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice Hall, 1979.

[5] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley & Sons, Inc, 2nd ed., 1984.

[6] A. Andrews, "A Square Root Formulation of the Kalman Covariance Equations," *American Institute of Aeronautics and Astronautics Journal*, vol. 6, pp. 1165–1166, June 1968.

[7] M. Aoki, *State Space Modeling of Time Series*. Springer, 2nd ed., 1990.

[8] W. F. Arnold and A. Laub, "Generalized Eigenproblem Algorithms and Software for Algebraic Riccati Equations," *Proceedings of the IEEE*, vol. 72, pp. 1746–1754, 1984.

[9] K. J. Åström and T. Bohlin, "Numerical Identification of Linear Dynamic Systems from Normal Operating Records," in *Theory of Self-Adaptive Control Systems*, pp. 96–111, 1965.

[10] Z. Bai and G. H. Golub, "Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices," *Annals of Numerical Mathematics*, vol. 4, pp. 29–38, 1996.

[11] B. Bamieh and L. Giarré, "Identification of Linear Parameter Varying Models," *International Journal of Robust and Nonlinear Control*, vol. 12, pp. 841–853, 2002.

[12] A. R. Barron, "Statistical Properteis of Artificial Neural Networks," in *Proceedings of the 28th IEEE Conference on Decision and Control*, pp. 280–285, 1989.

[13] J. Bellantoni and K. Dodge, "A Square Root Formulation of the Kalman-Schmidt Filter," *American Institute of Aeronautics and Astronautics Journal*, vol. 5, pp. 1309–1314, 1967.

[14] R. Bellman and K. J. Åström, "On Structural Identifiability," *Mathematical Biosciences*, vol. 7, pp. 329–339, 1970.

[15] J. S. Bendat, *Nonlinear System Analysis and Identification from Random Data.* Wiley-Interscience, 1990.

[16] J. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," tech. rep., Berkeley International Computer Science Institute, 1997.

[17] T. Bohlin, *Practical Grey-box Process Identification.* Springer-Verlag, 2006.

[18] S. Boyd and C. Barrat, *Linear Controller Design: Limits of Performance.* Prentice Hall Information and System Sciences Series, 1991.

[19] R. A. Boyles, "On the Convergence of the EM Algorithm," *Journal of Royal Statistical Society : Series B*, vol. 45, no. 1, pp. 47–55, 1983.

[20] P. Brunovsky, "On Stabilization of Linear Systems Under a Certain Class of Persistent Perbutations," *Differential Equations*, vol. 2, pp. 401–405, 1966.

[21] R. Bucy, "Canonical Forms for Multivariable Systems," *IEEE Transactions of Automatic Control*, vol. AC-13, pp. 567–569, 1968.

[22] R. Bucy and P. Joseph, *Filtering for Stochastic Processes with Applications to Guidance.* John Wiley, 1968.

[23] P. E. Caines, *Linear Stochastic Systems.* John Wiley & Sons, Inc, 1988.

[24] K. De Cock, *Principal Angles in System Theory, Information Theory and Signal Processing.* PhD thesis, Katholieke Universiteit Leuven, 2002.

[25] A. P. Dempster, N. M. Laird, and D. B. Durbin, "Maximum Likelihood Estimation from Incomplete Data," *Journal of the Royal Statistic Society.*

[26] V. Digalakis, *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition.* PhD thesis, Boston University Graduate School, 1992.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* John Wiley & Sons, Inc, 2nd ed., 2000.

[28] P. Dyer and S. McReynolds, "Extension of Square Root Filtering to Include Process Noise," *Journal of Optimization Theory and Applications*, vol. 3, no. 6, pp. 444–459, 1969.

[29] P. Eykhoff, *System Identification: Parameter and State Estimation.* John Wiley & Sons Ltd, 1974.

[30] P. L. Faurre, *System Identification: Advances and Case Studies*, ch. Stochastic realization algorithms, pp. 1–25. Academic Press, 1976.

[31] J. Frankel, *Linear Dynamic Models for Automatic Speech Recognition.* PhD thesis, University of Edinburgh, 2003.

[32] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital Control of Dynamic Systems.* Addison-Wesley Publishing Company, 2nd ed., 1990.

[33] D. C. Fraser, "A New Technique for the Optimal Smoothing of Data," tech. rep., M.I.T. Instrumentation Lab, 1967.

[34] A. Gelb, J. F. Kasper Jr., R. A. Nash Jr., C. F. Price, and A. A. Sutherland Jr., *Applied Optimal Estimation.* The MIT Press, 1974.

[35] M. Gevers, "A Personal View on the Development of System Identification," in *IFAC Symposium on System Identification*, 13th Proceedings, pp. 773–784, 2003.

[36] M. Glover and J. C. Willems, "Parametrizations of Linear Dynamical Systems: Canonical Forms and Identifiability," *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 640–646, 1974.

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations.* The John Hopkins University Press, 3rd ed., 1996.

[38] B. Gopinath, "On the Identification of Linear Time-Invariant Systems from Input-Output Data," *Bell Systems Technical Journal*, 1969.

[39] M. S. Grewal and A. P. Andrews, *Kalman Filter: Theory and Practice Using Matlab.* John Wiley & Sons, Inc, 2001.

[40] R. Guidorzi, "Canonical Structures in the Identification of Multivariable Systems," *Automatica*, vol. 11, pp. 361–374, 1975.

[41] N. K. Gupta and R. K. Mehra, "Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations," *IEEE Transactions on Automatic Control*, vol. AC-19, pp. 774–783, December 1974.

[42] D. A. Harville, *Matrix Algrbra from a Statistician's Perspective*. Springer, 1997.

[43] T. Hastie, R. Tibshirani, and J. Friedman, *he Elements of Statistical Learning*. Springer, 2001.

[44] S. Haykin, *Neural Networks*. Prentice Hall, 1999.

[45] S. Haykin, *Kalman Filtering and Neural Networks*, ch. Kalman Filters. John Wiley & Sons, Inc, 2001.

[46] B. Ho and R. Kalman, "Effective Construction of Linear State-variable Models from Input-Output Functions," *Regelungstechnik*, vol. 12, pp. 545–548, 1965.

[47] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

[48] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Prentice Hall, 6th ed., 2007.

[49] S. J. Julier, J. K. Uhlmann, and H. Durrant-Whyte, "A New approach for Filtering Nonlinear Systems," *Proceedings of the American Control Conference*, pp. 1628–1632, 1995.

[50] T. Kailath, *Linear Systems*. Prentice Hall, 1980.

[51] R. Kalman, "Mathematrical Description of Linear Dynamical Systems," *SIAM Journal of Control : Series A*, vol. 1, pp. 152–192, 1963.

[52] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, March 1960.

[53] P. G. Kaminski, A. E. Bryson Jr., and S. Schmidt, "Discrete Square Root Filtering: A Survey of Current Techniques," *IEEE Transactions on Automatic Control*, vol. AC-16, pp. 727–736, 1971.

[54] T. Katayama, *Subspace Methods for System Identification*. Springer-Verlag London, 2005.

[55] D. B. Kinghorn, "Integrals and Derivatives for Correlated Gaussian Functions Using Matrix Differential Calculus," *International Journal of Quantum Chemistry*, vol. 57, no. 2, pp. 141–155, 1998.

[56] V. Kucera, "The Discrete Riccati Equation of Optimal Control," *Kybernetika*, vol. 8, no. 5, pp. 430–447, 1972.

[57] J. Kybic, "Kalman Filtering and Speech Enhancement," Master's thesis, Czech Technical University, 1998.

[58] N. Levinson, "The Wiener RMS (root-mean-square) Error Criterion in Filter Design and Prediction," *Journal of Mathematical Physics*, vol. 25, 1947.

[59] F. H. Lewis, *Optimal Estimation With An Introduction to Stochastic Control Theory.* John Wiley, 1986.

[60] L. Ljung, *System Identification Theory for the User.* Prentice Hall PTR, 2nd ed., 1999.

[61] L. Ljung and T. Södesrstörm, *Theory and Practice of Recursive Identification.* The M.I.T. Press, 1987.

[62] L. Ljung and A. Vicino, eds., *IEEE Transactions on Automatic Control: Special Issue on Identification*, vol. AC-50, October 2005.

[63] D. G. Luenberger, "Canonical Forms for Linear Multivariable Systems," *IEEE Trans. Automatic Control*, vol. AC-12, no. 3, pp. 290–293, 1967.

[64] D. G. Luenberger, *Introduction to Dynamic Systems: Theory, Models & Applications.* John Wiley & Sons, 1979.

[65] D. Mayne, "A Canonical Model for Identification of Multivariable Systems," *IEEE Transactions of Automatic Control*, vol. AC-17, pp. 728–729, 1972.

[66] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra.* SIAM, 2000.

[67] O. Nelles, *Nonlinear System Identification.* Springer, 2000.

[68] T. Ogunfunmi, *Adaptive Nonlinear System Identification.* Springer, 2007.

[69] C. C. Paige, "Properties of Numerical Algorithms Related to Computing Controllability," *IEEE Transactions on Automatic Control*, vol. AC-26, pp. 130–138, February 1981.

[70] K. B. Petersen and M. S. Pedersen, "The matrix cookbook." http://matrixcookbook.com.

[71] D. S. G. Pollock, "Tensor Products and Matrix Differential Calculus," *Linear Algebra and its Applications*, vol. 67, pp. 169–193, 1985.

[72] V. Popov, "Invariant Description of Linear Time-Invariant Controllable Systems," *SIAM Journal of Control*, vol. 10, pp. 252–264, 1972.

[73] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 2003.

[74] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," in *Proceedings of Eurospeech*, 2003.

[75] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[76] H. E. Rauch, F. Tung, and C. T. Stribel, "Maximum Likelihood Estimates of Linear Dynamic Systems," *American Institute of Aeronautics and Astronautics Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[77] T. Ribarits, *The Role of Parametrizations in Identification of Linear Dynamic Systems*. PhD thesis, Technical University Wien, 2002.

[78] M. M. Rosenbrock, *State-Space and Multivariable Theory*. John Wiley, 1970.

[79] S. Roweis and Z. Ghahramani, *Kalman Filtering and Neural Networks*, ch. Learning Nonlinear Dynamics using the Expectation Maximization Algorithm. John Wiley & Sons, Inc, 2001.

[80] S. F. Schmidt, "The Kalman Filter: Its Recognition and Development for Aerospace Applications," *Journal of Guidance and Control*, vol. 4, no. 1, pp. 4–7, 1981.

[81] T. B. Schön, A. Wills, and B. Ninness, "Maximum Likelihood Nonlinear System Estimation," in *14th IFAC Symposium on System Identification*, 2006.

[82] D. Simon, *Optimal State Estimation*. John Wiley & Sons, Inc, 2006.

[83] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyona, P. Y. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear Black-box Modeling in System Identification: A Unified Overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[84] T. Södesrstörm and P. Stoica, *System Identification*. Prentice Hall International Ltd, 1989.

[85] D. Spain, *Identification and Modelling of Discrete, Stochastic, Linear Systems.* PhD thesis, Dep. of Elect. Engr. Stanford University, 1971.

[86] A. A. Stoorvogel and J. H. van Schuppen, "Approximation Problems with the Divergence Criterion for Gaussian Variables and Gaussian Processes," *Systems & Control Letters*, vol. 35, pp. 207–218, 1998.

[87] S. Theodoridis and K. Koutroubas, *Pattern Recognition.* Elsevier, 2nd ed., 2003.

[88] R. Togneri and L. Deng, "A Structured Speech Model Parameterized by Recursive Dynamics and Neural Networks," in *Proceedings of Interspeech*, 2007.

[89] D. Tracy and K. G. Jinadasa, "Patterned Matrix Derivatives," *The Canadian Journal of Statistics*, vol. 16, no. 4, pp. 411–418, 1988.

[90] E. Tse and J. Anton, "On the Identifiability of Parameters," *IEEE Transactions on Automatic Control*, vol. AC-17, pp. 637–646, 1972.

[91] E. T. S. Tse, H. L. Weinert, J. J. Anton, and R. K. Mehra, "Model Structure Determination and Identifiability Problems in System Identification," annual, Office of Naval Research, 1973.

[92] P. Van Overschee and B. De Moor, "Subspace Algorithms for the Stochastic Identification Problem," *Automatica*, vol. 29, no. 3, pp. 649–660, 1993.

[93] P. Van Overschee and B. De Moor, "N4sid - subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems," *Automatica*, vol. 30, no. 1, pp. 75–93, 1994.

[94] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications.* Kluwer Academic, 1996.

[95] V. Vapnik, *Statistical Learning Theory.* Wiley, 1998.

[96] M. Verhaegen and P. Dewilde, "Subspace Model Identification, Part 1: The Output-Error State-Space Model Identification Class of Algorithms," *International Journal of Control*, vol. 56, no. 5, pp. 1187–1210, 1992.

[97] M. Verhaegen and P. Dewilde, "Subspace Model Identification, Part 2: Analysis of the Elementary Output-Error State-Space Model Identificationalgorithm," *International Journal of Control*, vol. 56, no. 5, pp. 1211–1241, 1992.

[98] E. A. Wan and R. Van Der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation," in *Proceedings of Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control*, IEEE, October 2000.

[99] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* The M.I.T. Press, 1949.

[100] C. Wu, "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.